Developing a Credit Scoring Model Using Social Network Analysis

Ahmad Abd Rabuh

UP821483

This thesis is submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy in the School of Business and Law at the University of Portsmouth

November 2020

Author's Declaration

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Signature: 

Name: Ahmad Abd Rabuh

Date: the 30th of April 2020

Word Count: 63,809

Acknowledgements

As I witness this global pandemic, my greatest gratitude goes to my mother, Hanan, for her moral support and encouragement through the good, the bad and the ugly. During the period of my PhD revision, I received mentoring and help from my future lifetime partner, Kawthar. Also, I have been blessed with the help of wonderful people from Scholars at Risk, Rose Anderson and Sarina Rosenthal, who assisted me in the PhD application and scholarship processes. Additionally, I cannot be thankful enough to the Associate Dean of Research, Andy Thorpe, who supported me in every possible way during my time at the University of Portsmouth. Finally, many thanks to my supervisors, Mark Xu and Renatas Kizys for their continuous guidance.

Developing a Credit Scoring Model Using Social Network Analysis

Table of Contents

List of Tables

Table of Figures

Abstract

This research examines the effects of adding social network attributes on credit scoring. Many lenders have realised the potential of borrowers with thin financial files who lack sufficient credit history. To overcome this information asymmetry problem, there has been a trend in examining the behaviour of borrowers. In many cases, such behaviour is influenced by peers within social circles of borrowers. This influence imposed by social circle is explained with the concept of homophily in sociology and network science disciplines. In this research, reducing information asymmetry is the first of two aims; whereas, increasing financial inclusion is the second aim. Achieving the aforementioned aims is done by finding meaningful information on social data of those who are unbanked or underbanked to measure how such data would affect their credit scores. Examples of such data are network types and sizes. Nine exploratory in-depth interviews were conducted with professional bankers and regulators to explore the effects of social networks on performance of borrowers. Additionally, a dataset containing loans given by a European lender to 307,000 borrowers was used to confirm and explain the effects of social network attributes on credit scoring. Alternative data made of social and behavioural artefacts were identified in the aforementioned dataset. Also, traditional data that are used in financial institutions were identified. A Mann-Whitney hypothesis test revealed that, at 1% significance level, bad social network types are higher at the sample group of defaulters than the sample group of transactors who repay their loans. Thereafter, a preliminary tree-based Bayesian analysis and a machine learning technique in applying Logistic regression model were completed on the dataset. Results have shown that one of the two social network types tested, defaulting ties, has a significant relationship with the probability of default and, accordingly, the credit score of borrowers. The aforementioned variable had a coefficient of 0.22 in two test trials when social data was added to financial and behavioural separately and 0.18 in the last test performed on all types of data combined. The area under curve (AUC) produced by the model was 0.58. In evaluating the applicability of social data in lending practices, the best explainable dataset, which included social network variables, was evaluated by running machine learning classification algorithms and achieving 0.68 accuracy level using XGBoost classifier. This research contributes, empirically, to the understanding credit scores using new variables (i.e. social network types). Finally, the study provides theoretical framework and evidence from the industry on when social data become important and justifies selecting a social credit score in such cases.

Keywords: Credit scoring, financial inclusion, homophily, information asymmetry, social networks.

CHAPTER 1: INTRODUCTION

In the era of globalisation and connectedness, economies compete fiercely on attracting investments, successful business ideas, and entrepreneurs into their ecosystems. Business models across industries have been revolutionised where digital solutions are manifested within those companies that aim to lead with a competitive advantage. In finance, there has been a trend in relying on artificial intelligence (AI) by banks to make decisions or, at least, classify investors and borrowers. Data on the aforementioned clients has been abundant in an unprecedent way. Also, data has been collected from many sources to reflect on consumers' behaviour and whereabouts or what they do in their social lives.

In order for treasury departments in governments, through their regulators such as the Bank of England (BoE) and Financial Conduct Authority (FCA), to attract innovative businesses and investments, there must be a developed credit ecosystem and infrastructure in place that facilitates loans. Therefore, regulations in developed economies, such as the one in UK, allowed the use of behavioural modelling in credit risk assessment and scoring (FCA, 2018). Complying with credit risk assessment regulations, such as the accords of Basel II committee, allowed banks, for example, to release more liquidity for lending by lowering their risk coefficients. Such banks have an 'Advanced Internal Rating Based (A-IRB)' accreditation (Sousa, Gama, & Brandão, 2016). Also, relying on data outside the financial context (also known as 'alternative data') allowed lending recommendations for customers and better judgement by lenders (Kumire, 2019).

The efficient market hypothesis (EMH) introduced by Eugene Fama in the 1970s emphasises that, in order for a financial market be efficient, assets needs to reflect all information available in it (Fama, 1991). Similarly, in credit, the availability of information is deemed necessary at both ends – the lenders' and the borrowers'. The inequality of information available in credit with the aforementioned parties has caused a phenomenon called 'information asymmetry'. Lenders not only avoided those who do not present enough information, but they also prefer not to search for information whenever the costs of search exceeded the interest generated from the loan (Yan, Yu, & Zhao, 2015). For lenders, credit was merely approved for those with long financial histories and verified income. On the other hand, lenders had been reluctant to approve credit to those with information opacity, lack of proper financial reporting and audits, low credit amount as opposed to high transaction costs. Additionally, there has been many cases of misrepresentations by some rating agencies (Kshetri, 2016; Yan et al., 2015). For borrowers, those who had negative

information about their creditworthiness avoided sharing such information while those who had positive information outside the financial contexts could not share the same. As a result, promising small business and creative individuals who are new to workforce were, financially, excluded and denied credit (Redrup, 2017). Moreover, those who, initially, were profitable and repaid their loans were showing affordability issues later when their credit facilities were extended. According to MasterCard, half of the adults, globally, have no access to banking services let alone credit facilities (McEvoy & Chakraborty, 2014). Some researchers indicated that good quality borrowers considered associating with those with good financial records as a way to signal their good quality (Lin, Prabhala, & Viswanathan, 2013).

The inequality of information available at the borrowers' and lenders' ends resulted-in two effects: moral hazards and adverse selection, where the former happens as a borrower acts in a way that increases the probability of default after obtaining the loan (Guttentag & Herring, 1984). On the other hand, the latter refers to unobservable poor factors that could contribute negatively to the credit decision at the time of the underwriting process (Berndt & Gupta, 2009). Such factors represent the interactions and differences between applicants and their probability of taking loans (Experian.co.uk, 2013).

As a result, lenders tend to charge high interest rates whenever information is limited on borrowers. However, raising the interest rates would discourage rational borrowers from taking up the offer. Also, it would expose borrowers who are desperate for the money and are risk-takers to higher default probabilities. In summary, raising interest rates would cost the lender more default than the promised returned interest (Guttentag & Herring, 1984)

In response to the information asymmetry phenomenon, economies have suffered from financial exclusion. Excluding and blocking potentially-profitable individuals and small-medium enterprises from accessing banking-type services such as credit would make economies suffer a huge opportunity cost. Excluded individuals are not, necessarily, limited to low-income earners as they can be new entrants to the job market such as fresh graduates, immigrants with high skills and calibre, and many others left out due to a thin or a non-existing financial profile or history (McEvoy & Chakraborty, 2014). Not only that, but also financial exclusion is arguably a main cause of social inequality (McKillop, Ward, & Wilson, 2007)For example, the COVID-19 pandemic has created a mounting demand for nursing professionals, medical supplies and preventive kits. Medical students, in the end of their course of studies in disease control or clinical

pharmacy, were expecting immediate appointments at hospitals and clinics (BMJ, 2020). In a developed economy, those are assessed according to the coursework and projected income, by companies such as SoFi as will be discussed later, although those have no or very thin financial records. Similarly, for excluded start-ups and SME's, those who had been researching in the medical sector on vaccines and seeking the approval of their product by a regulator, such as Food and Drug Association (FDA), would be successful borrowers had licensing news been accessed by lenders. Therefore, the need for an economy that transmits new information freely has been the focus driven by the efficient market hypothesis (Mittal & Goel, 2012; Potì & Siddique, 2013).

In credit industry, this particular need was resembled in a credit score that measures the likelihood of a potential borrower repaying a loan on time by incorporating as much recent information as possible. Such a score would use, not only financial data represented by traditional transactions, but also behavioural and social data that are described as alternative by McEvoy and Chakraborty (2014), Pokorná and Sponer (2016), and FCA (2017).

Innovative models that utilised alternative data were adopted by new businesses such as Peer-to-Peer websites (P2Ps). In the case of P2Ps, the idea was to establish a direct link between lenders and borrowers. Credit terms such as loan amount, duration, and interest rate are agreed upon thereafter (Gonzalez & Loureiro, 2014).

One thing that has been rarely discussed in research and not-widely-adopted in lending is social network analysis. Social networks guide individuals down paths that might not be of their own choices. Sociologists argue that, in some cases, social networks could shape up destinies if individuals have persistent connections within the same social network for long periods (Currier). In principle, social networks and social media are two different concepts. Although they may overlap (social networks can be extracted from social media platforms), the distinction has not been clearly contrasted as the next chapter will demonstrate.

In credit risk assessment, the choice of model is another issue that challenges credit risk assessors. Many parametric models are linear and provide lenders and borrowers of good understanding about the reasons behind a credit risk score. A simple ancillary that complements parametric models and helps in creating a cut-off score is the scorecard that can be applied mathematically without the need for a sophisticated system.

The big data revolution has been a game changer in credit risk scoring and assessment models. As data storage allowed storing large volumes of data, the possibility of training a model on a large

dataset then testing it on a hold-out sample became possible. Finally, there has been many data structures, types and forms such as text, numbers, audio, video, images, web-scraped clicks, times, dates, etc. The aforementioned variety resulted in heterogenous datasets with high dimensions (Liu & Schumann, 2005). The 'curse of dimensionality' was another challenge to basic classification techniques (Han, Pei, & Kamber, 2011) such as parametric models in addition to few non-parametric models.

Therefore, the introduction of machine learning and artificial intelligence as advanced models was prominent in last two decades. Many researchers argued their superior suitability. For example, it was argued by Dash, Kremer, Nario, and Waldron (2017) that machine learning and deep learning algorithms can make better and faster underwriting decisions in credit. When predicting risk scores, machine learning models can learn iteratively from both financial and non-financial data (Turner & McBurnett, 2019). Although, those outperformed parametric models in many cases, they were less intuitive and provided little or no information about what causes a credit risk score to be low or high. Finally, they were criticised for their biases and led to a stream of research in the area of 'explainable AI' (Bussmann, Giudici, Marinelli, & Papenbrock, 2020).

Dynamic models that run over states of times (usually instalment day) and consider changes in behaviour over the time of loan. Those can be an automated version of the previously-mentioned two types or can be made of a completely different model such 'Cox Proportional Hazard' (Tong, Mues, & Thomas, 2012). Finally, credit analytics will be introduced which takes other considerations in credit risk assessment such as the likelihood that someone has changed jobs from geo-spatial data. Credit analytics combines different models with different types of data. It, sometimes, uses unsupervised machine learning for clustering whenever possible before running different models on different clusters.

When looking at social network data specifically, firms can estimate which employee is more connected by monitoring all social communications whether on WhatsApp or via e-mail and telephone (Leonardi & Contractor, 2018). The main advantage of online platforms is that the information they contain is 'democracised', or, in other words, bias-free because those networks empower the individuals and give them a platform to express their ideas freely (Yan et al., 2015). As a result, FinTechs had become proactive, instead of reactive/responsive, by using data available in multiple sources to target customers who are more likely to ask for a product or service (Yan et al., 2015). Consistently, emerging popular sources such as messaging and chatting as well as social

media platforms in addition to customer's reviews and ratings were proposed as a valid and effective source of data to be evaluated in risk analytics (Dash et al., 2017). This has caused a shift is the credit industry from minimising the chance of defaulting customers to maximising the profit from borrowers (Lyn C Thomas, 2009).

Financing institutions aim for an easy and a simple credit application process (Lyn C Thomas, Edelman, & Crook, 2002). Accordingly, the necessity for banks to adopt a novel big data analytics approach, such as social network analysis of different graph sources, is plausible. This research explains different credit assessment models adopted and justify why lenders should drive their systems into analytics-oriented systems and seek alternative data sources such as those extracted from different social networks.

1.1. Research Questions

In this research, the following questions are addressed: (1) Does analysing borrowers' social networks determine their credit scores more accurately? (2) Do bad borrowers who end-up defaulting have larger bad social networks? (3) Can lenders infer whether a borrower is likely to repay or default based on the type and the size of one's social network? And, finally, (4) how can social network data be incorporated within the traditional credit risk assessment?

The methods used, in this research, trangulate both qualitative and quantitative data to answer those questions. The importance and validity of the questions were confirmed with banking professionals. The first question is answered using a hypothesis test, Mann-Whitney U non-parametric test, for comparing distributions and medians. The second question is answered by using a Bayesian tree-based preliminary analysis. In such analysis, classes of the social networks were analysed to check if any evidence and additional information can be captured by knowing the class that a borrower belongs to. In addition to that, a logistic regression model was run on various subsets of a large dataset to measure the benefits (in both: explaining and/or classifying accurately) of adding social data to financial and/or behavioural data. The aforementioned model is used to answer the third research question. Also, a machine learning technique is used to add rigor to the procedure. Finally, the performances of machine learning models were compared on the best explainable subset in order to present the researchers with the best classification algorithm as well as the best selection of data subset.

1.2. Aims and Objectives

The main objective of this research is to provide lenders with a credit risk assessment model that

considers not only traditional financial data, but also alternative social network data by illustrating how such data can influence borrowers' behaviour and their credit scores. The aforementioned assessment will adjust credit scores based on social networks' types and sizes. Achieving the aforementioned objective will be attained through the following goals:

i. Reviewing and evaluating existing credit assessment metrics, models, and tools (systems) in order to identify gaps in credit risk scoring and assessment.
ii. Presenting the concept of incorporating social data, specifically social networks, as a key determinant for credit risk modelling from the literature.
iii. Conducting in-depth interviews with banking professionals to identify current factors and discuss the plausibility of using alternative data, particularly social networks, to be considered for credit risk assessment.
iv. Applying theoretical and applicable concepts of social network science from other disciplines onto credit risk assessment discipline.
v. Presenting evidence that social networks have different distribution between two independent groups of borrowers – repayors and defaulters.
vi. Developing a model to examine how different network data types and sizes would affect credit scores and highlighting improvements in predictability as well as explainability.
vii. Creating a guideline for credit risk scoring based on what data is available with lenders.

By achieving the above-mentioned goals and objectives, the research accomplishes its two major aims:

A. Reduction in information asymmetry; and
B. Increase in financial inclusion.

1.3. Contribution

Literature in credit has been categorised, mainly, in two streams. The first stream focused on how to find new measures and models to estimate creditworthiness. The second stream aimed at identifying new sets of variables in addition to the existing demographic, economic, and psychological variables (L. Wang, Lu, & Malhotra, 2011). The second stream of research was a result of the financial exclusion phenomena where many causes resulted-in limiting credit facilities to small-medium enterprises (SMEs) as well as to individuals (Kumire, 2019). In addition to the two main streams, a third less-common stream focused on 'credit rationing' or analysing the

reasons behind a borrower's denial of credit and refusal (Blumberg & Letterie, 2008). This research is aligned with the last two streams. Also, it reviewed the first stream within the literature review chapter. Theoretically, this research serves as the only academic research to have empirical findings on the effects of social network types on credit scores and outcomes of loans provided by a financial institution.

This research will contribute to extend the study of Lin et al. (2013), who proposed the use of types of groups in peer-to-peer (P2P) lending. Their study used dummy types and added a hierarchy to each type. In this study, there are two types of social networks that represent real data collected by a lender and those can complement the aforementioned studies. Also, this research will contribute to regulating creditworthiness in terms of allowing more inclusion as lenders can use a novel data source, social network, to estimate borrowers' affordability. Recently, the FCA allowed for behavioural methods to be accounted for by lenders when calculating credit worthiness and credit scores since behavioural biases may influence the credit outcome in addition to the circumstantial and economic factors (FCA, 2017, 2018).

In practical terms, this research contributes to the creation of a social-network-based credit model where a guideline would be followed on the cases where social data matters the most and other. Specifically, the effects of the bad social network ties are explored, confirmed, and tested to provide an explanation. Finally, the use of logistic regression model on social network data provides lenders with inference and explainability which contributes to the compliance of the regulatory requirements in transparency, fairness, and accountability in the modelling process.
In summary, this research will develop a model for credit scoring. The said model will contribute to solving the problem of information asymmetry and promoting financial inclusion through new sources of data - social networks.

1.4. Organisation of Chapters

The remaining part of this research will be in the following structure. Chapter 2 reviews the literature on credit scoring. It explains the ecosystem of credit risk resources, concepts and techniques from different perspectives. Also, it sets the scene for researchers and practitioners who want to establish a comprehensive understanding of how the credit system works. The debate on whether technology leads regulations on credit risk or if regulators oppose developing automated systems and to what extent those embrace the use of alternative data can be found in section 2.1.

Thereafter, explaining how lenders make their credit-related decisions is demystified. Particularly, on what basis lender decide on pricing criteria and differences between nature, static or dynamic, and types, financial or behavioural, of data provided by each source, whether internal or external, is explained in section 2.2. Then, a review is laid down on common types of credit risk models and their properties and limitations is presented. The aforementioned review concludes with a comprehensive summary of previous studies where models adopted in each study are highlighted along with classification accuracies in section 2.3. A review of network science theories and structures that have anecdotal and theoretical influence on credit risk follows that in section 2.4. Chapter 2 concludes with identifying the gap.

In chapter 3, a review on practical systems and tools used by different stakeholders in the credit industry is conducted. Peer-to-peer platforms are, critically, reviewed in section 3.1. Thereafter, micro lenders and LendTechs are highlighted with their role in using alternative data to include the unbanked in section 3.2. Additionally, digital banks who rely on online banking heavily and capitalise on their clients' behavioural data generated by variable sources is discussed in 3.3. Finally, third-party assessors that lenders refer to for a wider range of data and huge pool of trends and models such as credit bureaus and other credit referencing agencies are discussed in section 3.4.

Details on methodology and how data was prepared and collected to enhance creditscoring using social data are discussed in chapter 4. In this chapter, the framework guides the flow of the study where design is serving the logical sequence of steps in research, while citing similar studies, can be found in section 4.1. Research philosophy reflects the mindset that the author followed while addressing the research questions regarding social network variables and their effects as explained in section 4.2. A qualitative approach represented by exploratory in-depth interviews with banking professionals and regulators is described in section 4.3. Finally, the description of the dataset that was used for quantitative modelling and tests as both data and methods are introduced can be seen in 4.4.

In chapter 5, results from qualitative study as well as statistical introduction of the data, its dimensions and pre-processing span from section 5.1 to sub-section 5.2.4.3. The results of the main three tests/model, Mann-Whitney, Bayesian analysis, and Logistic regression, are reported in section 5.2.4. Evaluation of the results can be seen in section 5.3. Consequently, the findings

are discussed in 5.4.

In chapter 6, a summary of results' findings, a reflection on how this research advances both academic and the industry, and a motivation for further research that can overcome challenges seen in this research are discussed in sections 6.1, 6.2, and 6.3 respectively to conclude this research.

CHAPTER 2: LITERATURE REVIEW

In its simplest definition, credit scoring is the process that estimates how much of the money lent will be lost due to default or delinquency (i.e. arrears). The decision associated with credit applications, or in other words: underwriting decision, has evolved from a binomial variable known as a credit decision: accept or reject to a multinomial variable based on a credit assessment classification process that puts customers into categories such as: high-risk, medium-risk, or low-risk. Eventually, credit decisions yielded a continuous variable known as a credit score and, in some cases, were tied with ongoing consumers behaviour manifested in financial transactions completed in the recent past – a process known as risk-based pricing (Lyn C Thomas, 2009).

In the following sections, a general framework will initiate the subject of credit scoring with a highlight on its regulation. Then, the research will discuss what behavioural finance is and what it entails of considering as reasonable criteria to be included in borrowers' credit scoring. After highlighting how behavioural finance affects borrower's actions, a section on social networks and what recent developments and applications of social network aspects are will be presented. Also, variations in credit risk modelling and a comparison between parametric models and non-parametric ones will be discussed in light of applicability and inference. Finally, the gap found in the literature will be presented.

2.1. Regulations on Assessing Creditworthiness

The consumer credit act (CCA) 1974 was implemented in the UK to regulate firms on how to make assessments of individuals for credit. The enforcement of the regulation is overseen by the Financial Conduct Authority (FCA). The main aim is to guide and monitor the process assessing creditworthiness every time an individual applies for credit or, in the case of credit cards or overdraft, every time a credit limit is increased significantly (FCA, 2018).

Enabling borrowers to buy goods and services and repay over time has a positive impact on the economy. However, regulators such as the FCA have realised that some borrowers are sub-prime or vulnerable. Hence, macro-economic studies were shared with lenders to examine the impact on borrowers working in a specific sector for example. In addition to that, prudential policies that aim to control liquidity and reserves were enforced on lenders such as stress tests to limit irresponsible lending (FCA, 2018).

The rapid emergence of technologies in financial services has driven regulators to align their

regulations with the emerging FinTech models. When it comes to financial inclusion, a number of solutions were provided to facilitate the move of those unbanked into the financial system. One of the regulators is the Financial Action Task Force (FATF), which is a government body based at the Office for Economic Cooperation and Development (OECD) in Paris. It has issued a guideline to demonstrate risk-based approaches that promote integrity while including the unbanked using technology (De Koker & Jentzsch, 2013).

For the banked consumers, Financial Conduct Authority (FCA) in the UK has been facilitating the move into open banking where banks share their databases with trusted third parties to allow building powerful models that can learn from larger number of transactions. FCA (2018) explained that while ensuring that the accuracy of credit scores will guarantee that lenders will be repaid, affordability assessment ensures that borrowers do not get in trouble, financially, for repaying their loans. This will contribute to minimising the financial distresses across the UK economy (FCA, 2018). With the new capabilities, such as open banking, found in developed countries, comprehensive credit reporting (CCR) has become a requirement in the era of digital disruption (Redrup, 2017). If challenged, lenders should be able to justify their credit decisions (FCA, 2018; "General Data Protection Regulation," 2016).

2.1.1. International Financial Reporting Standards 9

The international financial reporting standards (IFRS) 9 is a comprehensive set of accounting and regulatory disclosures that affect the measurement of loan allowances for banks. In general, banks do report their consumer credit portfolios differently due to different measurements and varying forward-looking assumptions of expected credit losses. Therefore, IFRS 9 implemented an expected credit loss guideline for banks to follow in order to account for risky loans and recover their values in the case of impairment or default. The main highlights in IFRS 9 included accounting for unused credit in medium-risk unsecured, also knowns as stage 2, loans. This would allow for the release of more reserves for lending. In addition to that, behavioural measures are allowed to be used in estimating the lifetime of a loan using loss given default models (EY, 2018). In summary, the idea of using alternative data has been absorbed and generally accepted in accounting standards.

2.1.2. Markets in Financial Instruments Directive II

The markets in financial instruments directive (MiFID) II was implemented in January 2018 and

sought protecting investors in stock market investments. The directive regulates credit institutions that are tasked with raising capital. According to point 1.2.2 in the PS17/5 policy statement of the directive, there is a responsibility on the credit broking firm to investigate the investor's affordability to buy stocks using a credit agreement. Therefore, companies like Schorders do protect investors' wealth from default.

2.1.3. General Data Protection Regulation

The idea behind the general data protection regulation (GDPR) was to present privacy and personal data of the European nationals and people living within the European Economic Area (EEA). The major concern is that credit scoring has been viewed by lenders as a pragmatic process recently. In other words, the main objective is to get a score and act upon it rather than understanding the score and explain it (Lyn C Thomas, 2009). Nevertheless, GDPR states that entities are obliged to limit the collection to the specific data required and make sure that their data policy explain how this data is going to be used, for what reason, and for how long. Specifically, in point 71 of the directive, it is clearly mentioned that "The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects … based solely on automated processing … such as automatic refusal of an online credit application" ("General Data Protection Regulation," 2016). Therefore, many banks have given the rights to their customers who disagree with an automated lending decision to ask for a staff review (HSBC, 2018) in order to avoid any violation of GDPR.

2.1.4. Payments Services Directive II

Technological advancements in credit had led the creation of LendTech (or lending technology) as one of the venues of FinTechs (or financial technology firms). FinTechs are technology-driven companies facilitated by open banking regulation known as 'Payments Services Directive 2 (PSD2)' where access to bank transactions has been granted to authorised third-party providers (TPPs) to run analytics and come-up with predictive models (Kumire, 2019). Some of those FinTechs produced predictive and prescriptive models to help borrowers avoid arrears or paying high interest. Such recommendation resulted-in savings of an average of £ 287 in the UK per borrower a year tallying a hefty national saving of £ 2.7 bn for the borrowers to consume (Reynolds & Chidley). This suggests that, by influencing borrowers' behaviour, their financial responsibility and credit outcome may change.

With the introduction of open banking, Payments Services Directive II (PSD2) was enforced by the Financial Conduct Authority (FCA) in the UK to regulate third party providers (TPPs) who are licensed to access bank accounts through application programming interfaces (APIs). Open banking facilitates a faster processing of payments since TPPs are authorised to receive payment directly on their platforms. Also, the existence of open banking allowed LendTechs to access the different accounts of individuals. There are two types of TPPs – account information service providers (AISPs) and payment initiation service providers (PISPs). AISPs offer personal finance management solutions and advise on where to get a loan from (Kumire, 2019) such as the case of Habito.

2.1.5. Basel Accords

Basel Accords specify the risk and capital requirements for banks (Sousa et al., 2016). Basel II accords incentivised financial institutions to implement the most appropriate and accurate models in order to reduce the risks of credit portfolios (Brown & Mues, 2012). This has posed a challenge on banks to not only produce an individualised-credit risk rate, but also to come up with a composite score for each of their credit product (Lyn C Thomas, 2009). The financial credit crisis resulted-in low interest rates and low, or in many cases lack of, securitization or collaterals and, in extreme cases, no documents required for loans (Taffler, 2017). Therefore, Basel committee emphasised, in its recommendations, on the necessity of strengthening internal controls that help banks avoid risks including credit risk. Consequently, Basel II recommendations allowed Advanced Internal Rating Based (A-IRB) accreditation holders to lower their credit risk coefficients (Sousa et al., 2016)

A-IRB foundation and advanced Basel capital accords have been criticised in credit risk modelling as it ignores the behavioural aspect of risk scoring and bases its calculation on corporate lending (Lyn C Thomas, 2009). Behavioural finance will be discussed later in this chapter in sub-section 2.2.4

2.2. Credit Scoring

Lending is a decision problem associated with, correctly, understanding and estimating the factors that determines one's repayments whether financial, macro-economic or situational. Usually, regulators set the guidelines for including major factors when assessing borrowers' creditworthiness. In the UK, the FCA considers the following factors as essential in the basic credit

scoring: the nature and amount of the loan, the costs/interest rate of the loan, the number and amounts of instalment repayments, and the potential consequences for non-repayment such as default charges (FCA, 2018). Some researchers have categorised the data needed for producing a credit risk score into basic information, repayment ability, life stability, credit record, and guarantees (Zhang, Jia, Diao, Hai, & Li, 2016). Credit risk assessors aim at identifying certain desirable characteristics that may improve the scoring accuracy whether classification or scoring. In doing so, one can look at historical data and decide whether knowing an additional characteristic would help in deciding (in the case of classification) or estimating (in the case of scoring) better or not (Lyn C Thomas, 2009). The more lenders have access to borrowers' financial histories and situation, the more willing they are to approve or extend credit for those (Peón, Antelo, & Calvo, 2016).

Credit terms and conditions have, always, been a function of the risk associated with credit through a score. Pricing this risk, however, was, initially, based on the revenue generated from a performing loan as in its repayments or instalments (Lyn C Thomas, 2009). Alternatively, the time value of money rule used in pricing a loan associates interest rate that discounts future cash flows, loan instalments, to the current present value of the loan amount requested plus the margin required (Rajan, Seru, & Vig, 2010) In investments, that discount rate is known as the internal rate of return (IRR). Therefore, it is concluded that the higher the internal rate of return, the more profitable a loan is. Most lenders assess their borrowers periodically and update the credit scores accordingly. In addition, whenever a credit is renewed, a full adjustment to credit scoring takes place as well (HSBC, 2018).

Estimating default possibilities is a critical process in the underwriting decision of whether accepting or rejecting a loan application. In addition to that, estimating a possible arrear or many of the same would require a higher interest rate needed to account for delays in payments. Also, projecting when the delinquency would happen could suggest a more-suitable tenure of the funding period (Banasik, Crook, & Thomas, 1999).

2.2.1. Scope of Assessment

In consumer lending, applicants are the main subjects to be assessed. Credit scoring not only investigates credit scores and its associated probability of default, but, also cares about affordability. High affordability risk and its consequences on applicants not repaying or barely-repaying loans would affect their financial situation adversely. On the other hand, rejecting many

applicants due to suspecting high affordability risks would cause financial exclusion. This has led to investigate other income streams within the household where applicant resides. For example, income of parents in the case of fresh graduates living with parents or spouse income in the case of couples (FCA, 2018). Specifically, the performance of joint accounts is taken into consideration when application is submitted. Therefore, many credit scoring advisors encourage applicants to associate their accounts with another person who has a healthy credit score. A joint account can be highly beneficial in such a case (Hayes, 2019). It is, however, up to the lender to decide whether such an income can be available to borrowers when financial difficulty arises (FCA, 2018).

In the case of guarantor lending, assessing the potential obligations that might fall on the guarantor's behalf, in case of arrears or default by the borrower, is suggested by FCA. Though such an assessment does not have to take rigorous measures as it does with the applicant (FCA, 2018).

2.2.2. Credit Pricing

Initially, evaluating interest rates was based on the expected return of portfolio of loans. Under the expected monetary value (EMV) theory, probability distribution and pay-off, i.e. pricing of interests, determine risk scores to be assigned for different loan portfolios. For example, a loan of £100 that is offered to a borrower within a portfolio characterised with a 5% default rate at a 10% interest rate should be accepted based on EMV because the expected outcome would be

– (0.05*100) + (0.95*10) = - (5) + 9.5 = + 4.5 (Lyn C Thomas, 2009).

The aforementioned process was not granular enough as it did not target individuals. Instead, it targeted pooled loans who share similar conditions overall. In such a pool, Lin et al. (2013) indicated that borrowers with high quality will be discouraged from applying as they would be feeling over-charged or at least equated to those with lower quality within their portfolio. Therefore, risk-based pricing was proposed to evaluate individuals based on their own characteristics and circumstances. The aforementioned strategy yielded in customised interest rates, loan values, duration of the loan (Experian.co.uk, 2013), and many other features such as grace periods, guarantors, type of credit, frequency of repayments, amount of repayments, total amount payable, total charge for credit, and whether charges are fixed or variable (FCA, 2018).

As a result, pricing of loans has become an individualised process that relies on borrowers' specific information retrieved at the time of application. In literature, there has been differentiations between hard information which can be retrieved at the time of application by the lender or credit

referencing agencies (CRAs) and soft information that are non-financial and can be sourced from the applicant's surroundings either by the lender or by third-parties (Lin et al., 2013). The former types are common; whereas, the latter are innovative and considered by lenders as an alternative source. Both types contribute to a credit score which will be discussed in the modelling section of this chapter (see section 2.3). Before discussed the models, the above-mentioned two groups of data will be discussed in the upcoming sub-sections (see sub-sections 2.2.3 through 2.2.5).

2.2.3. Traditional Criteria

Mainly, lending is based on translating customer financials into trustworthiness (Lyn C Thomas, 2009). For Lyn C Thomas (2009) and Brockett and Golden (2007), traditional scoring criteria focuses on socio-economic characteristics. Such characteristics include income, marital status, nationality, sex, number of children, age, profession, sector, residential status, employment type, time at present job, and loan specific features such as duration, amount, and purpose (Steenackers & Goovaerts, 1989). Other information on credit historical performances and payment behaviour (whether paid on time or had arrears) and existing debt obligations are, also, retrieved at the time of application (Kruppa, Schwarz, Arminger, & Ziegler, 2013). In addition to that, when applying for mortgages or assets such as property, location and vehicle ownerships are taken into consideration (HSBC, 2018).

In credit scoring, data is collected, internally, from two sources – application data and transactional data (Lyn C Thomas, 2009). Application data reflects demographics, or individual characteristics, (Kruppa et al., 2013) as well as the applicant's financial circumstances at the point of lodging the application. Examples of such data are age, income, purpose of the loan, current address, number of dependents, marital status, number of bank accounts and credit cards held (Banasik et al., 1999; L. C. Thomas, Edelman, & Crook, 2017), years with bank, employment category (L. Thomas, Banasik, & Crook, 2001), residency type/ownership of home (Banasik et al., 1999; Freedman & Jin, 2017), credit history (in case of any previous delinquencies or defaults), and debt-to-income ratio (Freedman & Jin, 2017). In the case of mortgages, lenders are interested in the location of the property as well (Kshetri, 2016). In terms of income, it is not limited to salaries and wages, but any cash inflow that may come from savings and another income of a person living in the same household. Similarly, expenses are classified into eligible "non-discretionary" expenses and ineligible voluntarily expenses. The former do matter in credit and those represent any contractual or statutory obligation to be made. Disposable income is the income minus non-discretionary

expenses (FCA, 2018).

In addition to demographics, application data takes marketing and macro-economic factors into consideration. In marketing, lenders treat a borrower that responds to a targeted campaign differently to one who approached for credit (L. C. Thomas et al., 2017). As for macro-economic variables, market research and economic conditions are studied (L. C. Thomas et al., 2017) to estimate economic indicators such as changes in consumer price index, average interbank lending rate, annual return on log of FTSE 100, and percentage growth in GDP (Malik & Thomas, 2010). Sometimes, accounting for macro-economic variables comes from regulators. For example, the UK government, through FCA, instructed lenders to allow a holiday pay for loans of those whose affordability were affected by the corona virus pandemic without impacting their credit scores negatively (HM-Treasury, 2020).

On the other hand, lenders consider statistics derived from transactions within customers' accounts such as average balance, maximum credit payment, number of times over credit limit (Lyn C Thomas, 2009), average transactions value, number of cash withdrawals, credit limit changes, rate of total jumps in proportion to months in arrears (Leow & Crook, 2016). In addition to application and transactional data, lenders rely on credit referencing agency (CRA) data such as credit bureaus. In their study, Malekipirbazari and Aksakalli (2015) identified application data at peer-to-peer website, lending club, as annual income, credit age, delinquencies, employment length, home ownership, inquiries, loan amount, loan purpose, open accounts, total accounts, and term. Additionally, they identified transactional data by calculating the debt-to-income (DTI) ratio, income to payment ratio, revolving[1] utilisation rate, revolving to income ratio. Finally, FICO credit score is used as an external metric in the credit scoring process (Malekipirbazari & Aksakalli, 2015).

In his neural network credit model, West (2000) produced a criteria for modelling credit based on German banking data. The table below lists the variables along with their weights (see Table 1). It is worth to mention that he was able to extract the weights/coefficients of the aforementioned variables through a 'clamping' practice. This was achieved by controlling all variables except one and measuring the impact of ± 5% variations in the tested variable.

As for external sources of data, lenders rely on external scores provided by credit referencing agencies (CRAs) commonly-known as credit bureaus (Kshetri, 2016; Rajan et al., 2010). The

[1] Refers to credit that does not have a fixed number of payments such as credit card or overdraft.

aforementioned third parties provide credit reports based on borrowers' financial and non-financial histories with other financial institutions, retail stores, and service providers. Examples of the aforementioned entities include other banks, department stores, energy suppliers, and county courts (Kshetri, 2016). Kruppa et al. (2013) lists unpaid bills, requests to pay issued by court orders, enforcement procedures, and uncovered checks as bad indicators coming from such entities.

Banks provide credit bureaus with repayment history on previous loans, types of previous loans, and the number of accounts opened and closed recently (Redrup, 2017).

Variable	Weight
Account longevity	0.113
Credit history	0.082
Employment classification	0.078
Checking account status	0.069
Asset owned	0.056
Years in residence	0.054
Other existing loans	0.053
Housing classification	0.053
Amount of loan	0.051
Purpose of loan	0.051
Years employed	0.040
Savings account status	0.038

Table 1: neural network loan criteria (West, 2000)

CallCredit, which was acquired by TransUnion in 2018, considered the following criteria in the UK: balance of credit card statement, total repayment as (a) percentage of the balance and (b) absolute value, percentage and number of minimum payments completed during the last 3 months, the ratio of minimum payments to voluntary payments, number of months since last payment, number of months since last payment on more than 50% of credit card accounts, total payments now as a percentage of total payments 3 months ago, number/value of cash advances during the last month/3 months for both, and ratio of value of cash advances during the last month/3 months to the statement balance (CallCredit, 2008).

Finally, credit bureaus have access to public sector databases and can verify whether a lender is registered for an electoral vote (Banasik et al., 1999), debt-to-income ratio, and the number of credit inquiries with other lenders in the last 6 months (Lin et al., 2013).

Those who have rights to vote are registered in an address, which, in turn, demonstrates more life stability to the lenders. On the other hand, those with county court judgements, such as bankruptcy, are less likely to be funded (HSBC, 2018). The existence of open banking facilitated the creation of a lot of credit bureaus around the world (Kumire, 2019; Redrup, 2017). The big three credit bureaus are Equifax, Experian, and TransUnion. In addition to the big three, Fair Isaac Corporation (FICO) is common in the US and is made of five components: (a) payment history (0.35), amounts owed (0.30), length of credit history (0.15), credit mix (0.10), and new credit (0.10). A FICO score

ranges between 300 - 850 (Wei, Yildirim, Van den Bulte, & Dellarocas, 2015; "What's in my FICO scores,"). FICO's scores range between 300 and 850 where the higher the score is, the better (Hayes, 2019). Some P2P lenders such as Prosper base their ratings on FICO scores. For example, AA rating translates to FICO score ≥ 760, A = 720 – 759, B = 680 – 719, C = 640 – 679, D = 600 – 639, E = 560 – 599, and HR = 520 – 559 (Lin et al., 2013).

Usually, lenders rely on a credit grade, category or score to determine one's probability of default (PD). Assessments are, mainly, focused on revolving line utilisation, debt-to-income ratio or the infamous FICO score. The Fair Isaac Corporation (FICO) score is based on 5 major components: payment history, accounts owed, length of credit history, credit mix, and new credit (more details and explanations can be found on FICO in sub-section 3.4.1).

It is, however, worth to note that legislators at the Federal Trade Commission and Consumer Financial Protection Bureau in the US prohibit the use of sex, age, race, and religion in credit scoring (Guo et al., 2016; Lyn C Thomas et al., 2002).

The dynamic nature of the aforementioned variables led to the introduction of risk-based pricing in which temporal, or time-series, analysis is used. In risk-based pricing, each borrower is quoted a tailored offer based on one's recent history and financial projections (Experian.co.uk, 2013). For example, borrowers who demonstrate low-risk are charged low annual percentage rates (APRs), allowed a longer grace period, and given a higher credit limit (Rusli, 2013). However, the absence of recent histories and financial projections for the unbanked and underbanked (see sub-section 2.2.4.2) causes lenders to charge higher rates and fees for loans (McEvoy & Chakraborty, 2014). As a result, those rates will discourage borrowers who are rational, responsible, and pay back on time as it is, fundamentally, an over-priced deal. Instead, such conditions would only attract those who are risk-takers and think they even deserve to be charged higher (Lin et al., 2013).

The above-mentioned situation has triggered the use of alternative data in order to understand borrowers' behaviour and, accordingly, project their financial exposure. In the next sub-section (see sub-section 2.2.4), behaviour will be discussed within the financial context, in general, and credit industry, in specific.

2.2.4. Behavioural Finance

The criteria discussed in the previous section was considered 'traditional' by Brockett and Golden (2007). It was reported by Redrup (2017) that credit bureaus such as FICO, Experian and CRISP (see sub-section 3.4.1) are using non-traditional data. The aforementioned shift from economic

forecasting to behavioural scoring was justified in literature by the coercive distorting taking place by governments (Nye, 2014). For example, in their creditworthiness assessment consulting paper, the FCA highlighted that a borrower with behavioural biases or low financial capability may end up defaulting on an affordable loan (FCA, 2018).

In general, researchers classified criteria into three categories. The first category is loan characteristics such as loan size, maturity, and product bought using the loan. The second category is household characteristics such as demographics. Finally, the third category was related to the borrowing behaviour which would demonstrate proxies to handling cash and liquidity (Stango & Zinman, 2006). In fact, L. Wang et al. (2011) asserted that demographic variables have less explanatory power than those that focus on attitudes and personalities.

Researchers argued that, despite meeting the financial criteria such as account balance and application data such as a desirable marital status, abrupt incidents such as divorce, job termination, diseases can lead to financial distresses (Sousa et al., 2016). The aforementioned circumstances could be much anticipated when examining borrowers' behaviour (Sousa et al., 2016) through 'alternative' data (McEvoy & Chakraborty, 2014). Alternative data were found in the models of many innovative lenders (Lazarow, 2017). Before reviewing behavioural metrics used in credit scoring, behavioural finance and its implications on the financial sector will be reviewed.

In literature, a lot of terms were used interchangeably while referring to the science that seeks the understanding of human mindsets when it comes to consumers' behaviour such as 'neuroscience in finance' (Ackert & Deaves, 2009). A clear example of how human brain and neurons are affected financially is the illusion of income that would some consumers have when having a credit card (L. Wang et al., 2011). Similarly, 'emotional finance' introduced the notion of unconscious behaviour when a financial decision is made (Taffler, 2017). 'Behavioural finance' is the most commonly-used term and is a science that studies how the individual's brain reacts to different situations to come up with a financial decision. It got its importance as economists realised that individuals do not act, all the time, with rationality when in complexed environments (Simon, 1959).

Data collected revealed that financial information at the bank or the lender is not updated in real time with the credit bureau. Also, there has been inconsistency in reports produced by credit bureaus with how recent the information on a particular applicant is retrieved based on domestic

practices followed within the premises of the bank or the lending institution (Guo et al., 2016).

Behavioural finance had been presented as an explanation for significant financial events. In investment banking, the scandal of energy company, Enron, and how it used a 'Ponzi' scheme was motivated by a corruptive behaviour that drove the empire into bankruptcy (Hamilton & Francis, 2003; Rajan et al., 2010). More globally, the financial credit crisis in 2008 manifested in offering sub-prime mortgage-backed securities (MBS's) was caused by lowering credit scoring requirements in order to achieve maximum profit irresponsibly (Peón et al., 2016). Also, in financial markets, it was argued by Mittal and Goel (2012) that public sentiment, predicted from investor's blogs and online behaviour, does have a direct relationship with the market prices of equity and financial assets.

As mentioned earlier, during the financial crisis in 2008; however, borrowers kept applying for mortgages and investing in MBS's. Those MBS's were backed or guaranteed by sub-prime high-risk mortgages that had loose credit terms and poor checks. The purchasing behaviour was driven by speculative motives and the desire to sell mortgage agreements and MBS's in the short term to make profit. Making profit in the mortgage credit industry was perceived to be persistent over time due to a gambler's mentality followed by the hot hand fallacies (Rabin & Vayanos, 2010).

2.2.4.1. Information Asymmetry

As one of the behavioural finance phenomena, information asymmetry happens when two parties have different access to the hidden information concerning their agreement (Ackert & Deaves, 2009). For example, a job candidate may accept a non-competitive offer while hiding from the employer the fact that one has permanent illness and requires an extensive medical treatment which would cost the employer an expensive insurance. Sophisticated borrowers believe that lenders do not genuinely provide the full information on the loan terms and conditions in an attempt to make an unmoral profit. Therefore, those act in a similar way. It is believed that those extrapolate their own profit-seeking behaviour and reflect it on lenders (Berndt & Gupta, 2009). Similarly, this research acknowledges that a borrower hiding information that signals his or her poor possible repayment ability or willingness would provoke an information-asymmetric decision which would eventually cause an adverse selection or a moral hazard case. Also, lenders having limited resources and tend to avoid the costs of collecting data on borrowers especially when loan values are small. The aforementioned situation was depicted by Yan et al. (2015) in the below figure (see Figure 1). The costs of negotiating, administrating and enforcing restrictive conditions on the loans

are high. Therefore, lenders tend to ignore those especially when the value of loan is relatively small (Guttentag & Herring, 1984).

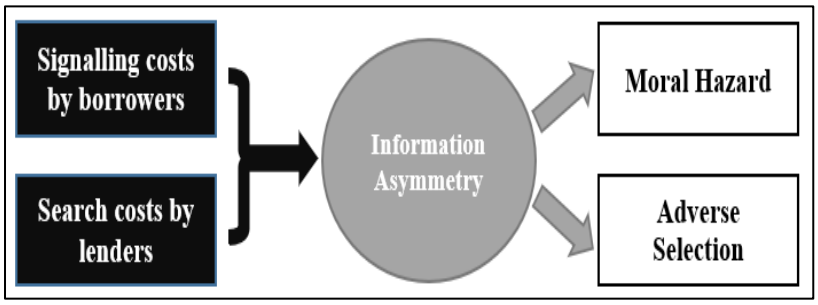

Figure 1: causes and effects of information asymmetry in credit

2.2.4.2. Financial Inclusion

Financial inclusion is defined as ensuring access to financial services at an affordable cost in a transparent manner. Financial services include, but not limited to, savings account, credit, and transfers. It aims to extend financial services to those who are using cash as a medium for exchange (De Koker & Jentzsch, 2013). The push for avoiding cash has been a vital decision not only from an economic point of view, but also from a medical and humanitarian point of views. With the current corona virus pandemic breakout, the world health organisation (WHO) has advised consumers to avoid using cash and adopt contactless methods of payments. They justified their advice with the fact that notes and coins can pick up all kinds of viruses (Finextra.com, 2020). Obviously, individuals who are financially excluded are unable to respond to the aforementioned call. According to De Koker and Jentzsch (2013), exclusion happens due to one or more of three reasons: affordability, eligibility, and geographical barriers. Globally, around 1.7 billion, or half of the adults, have no access to financial services including credit (Fitzgerald, 2018). While trying to achieve high market share in credit, lenders focused on cross-selling or extending credit to those existing customers who are, typically, well-off. Meanwhile, those with great potential and great skills remain undiscovered. This may include fresh graduates, homemakers returning to work force, or immigrants. The aforementioned is typically known as the underbanked or, in extreme cases, unbanked. Usually, those tend to reach out to family and friends (McEvoy & Chakraborty, 2014). It was argued by De Koker and Jentzsch (2013) that financial inclusion contributes to an economic growth as well as reducing poverty. Moreover, financial inclusion enables anti-money laundering (AML) and countering of financial terrorism (CFT) functions by being able to track the movement of funds across channels.

2.2.4.3. Personality Trait

In addition to the above-mentioned phenomena, behaviour on social media platforms has been studied by researchers. Activities and other dimensions were proposed to measure personality traits. The targeted traits were aligned with the five-factor OCEAN model proposed by many researchers. Those traits are: extraversion, neuroticism, agreeableness, conscientiousness, and openness (Bachrach, Kosinski, Graepel, Kohli, & Stillwell, 2012; Barrick & Mount, 1991; Kosinski, Stillwell, & Graepel, 2013). Bachrach et al. (2012) used data from Facebook to identify the five factors. Data varied from likes of photos and pages of products, sports, musicians, books, restaurants, to number of friends and density of networks (Kosinski et al., 2013). The table below (see Table 2) summarises the findings of Bachrach et al. (2012).

OCEAN's Psychometric Model					
Actions	Users' own actions			User's surrounding actions	
Reflections	▪ Number of published photos/videos. ▪ Number of events created. ▪ Number of groups joined. ▪ Number of likes (and other reactions). ▪ Number of status updates ▪ Number of shares.			▪ Number of times a user was tagged in a photo/video. ▪ Size/density of users' networks.	
Metrics	Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticism
Definition	The tendency to experience new things.	Following an organised pattern rather than a spontaneous lifestyle.	The need for an external motivation.	The extent to caring about the social relationships.	Strong emotional feeling such as depression, anxiety, guilt, anger, etc.
Correlations	(+) Number of likes / groups / status updates.	(-) Number of likes / groups. (+) Number of uploaded photos/videos.	(+) Number of likes / groups.	(+) Number of tagged photos.	Weak correlation with the number of likes and friends.

Table 2: estimating personality traits using social media psychometrics/behaviour

The findings of researchers were explained using psychological inference. For example, it was explained that, despite having less friends and likes on social media, highly-neurotic individuals would still spend longer times on chatting websites since avoiding face-to-face interactions will help concealing negative feelings. Conversely, with introverts who used to be on traditional social services such as blogs and forums, those use current social media channels less because their identities are revealed (Correa, Hinsley, & De Zuniga, 2010). Other behavioural models are intrusive by nature as discussed by Kosinski et al. (2013). Those can predict sexual orientation,

ethnicity, political views, religion, personality, intelligence, satisfaction with life, substance use whether drugs, cigarettes, or alcohol, and parents' marital status.

The main drawbacks of understanding personalities of individuals are misrepresentations and cheating of those who are answering the psychometric test questions or sharing fake activities on social media (Kosinski et al., 2013). In behavioural lending sub-section (see sub-section 2.2.4.6), the connection between personal qualities and credit-related behaviour will be established.

In an extension to that crisis, Taffler (2017) argued that investors seek that behaviour to achieve arousal feelings. In addition to the previous two theories, agency theory highlights how an agent tends to seek own interests despite the client's best interest being at risk. Therefore, a behavioural solution based on a cost-benefit pay-off matrix needs to be developed to asses motives (Ackert & Deaves, 2009).

Research has been ongoing on the influence of personality and day-to-day behaviour on financial decisions (Ackert & Deaves, 2009), job performance (Barrick & Mount, 1991; Muñoz de Bustillo & De Pedraza, 2010) as well as loan repayment (Arráiz, Bruhn, & Stucchi, 2017). The following sub-sections explore those three areas where behavioural finance literature existed.

2.2.4.4. Financial Decision-Making

The basic theory of financial decision making was the 'Expected Utility Theory' and it assumed rationality when making a financial decision and aim at maximising their expected utility from any transaction (Taffler, 2017). William Sharpe introduced the Capital Asset Pricing Model (CAPM) in an attempt to justify people's risk-taking behaviour with the return they are expecting from a financial decision. In his model, he introduced the common systematic risk known as 'Beta' (denoted β) as the risk every financial investor would bear. He argued that any other specific-risk can be eliminated with diversifying techniques (Ackert & Deaves, 2009).

Prospect theory described how investors are, usually, risk-averse when it comes to gain yet risk-takers when they bid for a loss (Ackert & Deaves, 2009). From a gender perspective, Muñoz de Bustillo and De Pedraza (2010) claimed that women have less risk-seeking characteristics and, therefore, worry more about every decision that take. In terms of financial decision-making, it was argued by Taffler (2017) that financial decision makers are imperfect and make judgemental and intuitive mistakes. Neuroscientists described feelings during the decision-making process where a person making the decision focuses on the exciting feelings while repressing anxiety. This was presented as a 'psychodynamic' analogy by (Taffler, 2017).

In addition to that, it was argued by contemporary researchers that behavioural finance focuses on heuristics (a set of pre-determined rules) and cognitive biases (personal tastes and perception) while ignoring social interactions (Hirshleifer, 2015; Taffler, 2017). Conversely, other researchers argued that judgements are affected by news transmitted and opinions made by social ties. More specifically, based on how a group of individuals connect with each other to represent their friendships in a social network (M. Newman, 2010).

2.2.4.5. Job Performance

When discussing over-inflated mortgage prices and sub-prime MBS's, researchers relied on accurate valuation of the aforementioned financing vehicles. Evaluating loans given to consumers is, mainly, tied with the present value of future cashflows, or instalments, discounted at the interest rates agreed upon at the time of application (Hagenau, Liebmann, & Neumann, 2013; Rajan et al., 2010). One of the main assurances of continuous payment of instalments is a steady income based on borrowers' job stability and employment status. Therefore, the need for looking at a borrower's career and job stability has become inevitable.

Researchers argued that younger employees are risk-seekers, who indulge in adventurous decisions and activities as opposed to older employees who are risk-averse and worry more about consequences (Manski & Straub, 1999). Accordingly, researchers suggested that behaviour is the underlying factor of indicating job performance, stability, and unemployment that drive fundamental and traditional variables such as age (Muñoz de Bustillo & De Pedraza, 2010). Other researchers concluded that behavioural aspects such as job withdrawal, absenteeism, and turnover are confounding variables when correlating job satisfaction with job performance. Not only that, but also, they indicated that someone's attitude produces a consistent pattern of behaviour (Judge, Thoresen, Bono, & Patton, 2001). The aforementioned argument was illustrated by an interchangeable relationship as seen in the integrative model below (see Figure 2). The aforementioned figure indicates that both of personality, through morals and cognition, and optimistic behaviour is existed in the relation between job satisfaction and job performance. The former moderates the relationship when performance causes employees' satisfaction; whereas the latter mediates in both cases whether satisfaction causes performance or vice versa. Moderators tend to strengthen the influence of one variable on the other. Mediators, on the other hand, tend to be essential for the relationship to complete (Judge et al., 2001).

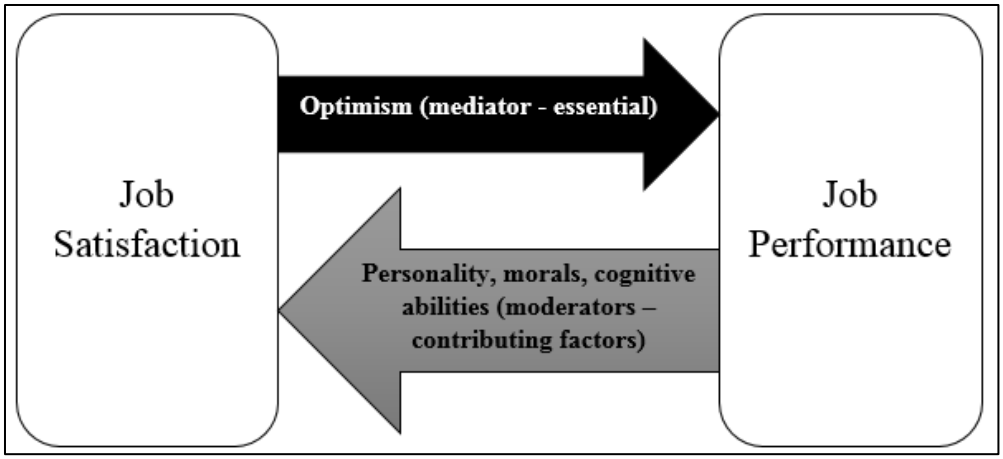

Figure 2: integrative model built on measure from behavioural finance

The table below (see Table 3) summarised the mean observed correlations between the five-factor dimensions and jobs investigated by Barrick and Mount (1991). Moreover, it was stated by Bachrach et al. (2012) that openness, for example, means being artistic and imagining. Accordingly, people who have an open personality and work in creative domains such as painters, musicians, or other jobs that require different decision-making process on a daily basis do excel in their jobs. Conversely, people who are less open to new experiences and unusual ideas are more likely to excel in jobs that follow a strict guideline such as administrative jobs and operators in the manufacturing processes.

Job Category	Extraversion	Neuroticism	Agreeableness	Conscientiousness	Openness
Professional	-.09	-.13	.02	.20	-.08
Police	.09	.10	.10	.22	.00
Managers	.18	.08	.10	.22	.08
Sales	.15	.07	.00	.23	-.02
Skilled/Semi-skilled	.01	.12	.06	.21	.01
Average	.13	.08	.07	.22	.04

Table 3: correlations between personality traits and job types (Barrick & Mount, 1991)

2.2.4.6. Behavioural Lending

Earlier this decade, risk-based pricing was introduced as a strategy in credit risk scoring where a fixed flat interest rate was replaced by different rates that are charged according to customers' risk exposure. Not only risk-based credit scoring allows to determine interest rate, but also decisions

to extend credit or not and by how much can be made. Additionally, the duration of credit facility and when to collect from delinquent accounts are also applications of risk-based modelling (Yeh & Lien, 2009).

Nevertheless, it was suggested that this strategy does not work well if no reliable information was available on differentiating good borrowers from bad ones (Experian.co.uk, 2013). That is why, lenders in the UK had expressed their interest in governmental sources such as HRMC tax data, council tax, rental repayment information, and statistics reported by the office for national statistics (ONS) in the UK (FCA, 2018). Nevertheless, in this era of big data analytics, there are many sources and features that make discriminating bad borrowers from good ones possible (Dash et al., 2017).

Banks and lenders, in general, have been considering the use of external data that is beyond the structured numerical form. In addition to analysing behavioural data collected internally, those have reached out to variable sources to estimate the behaviour of their customers and assess their credit risks. From checking how often individuals recharge their phone batteries (Fitzgerald, 2018) to plotting credit card transactions on a map, lenders used such data to estimate whether an applicant would have bad or good loan performance (Dash et al., 2017). Also, estimating one's increase of income can be by predicting an educational achievement during the repayment time in the future (FCA, 2018)

McEvoy and Chakraborty (2014) laid down a strategy for a successful credit risk modelling using alternative data. Their strategy consists of 4 main steps: (1) lenders need to identify the most suitable data sources for the targeted population. For example, high internet penetration suggests social data. (2) It is recommended to start with a training data set that has a small sample and a clear decision-rule map before applying on a larger dataset. (3) ensure long term access to data and that data will not be presented differently in the future for a systematic pre-process procedure. Finally, (4) running a pilot before a market-wide implementation.

In practice, the idea of borrowers' behaviour and their influence on loan outcomes had led to risk-based pricing strategies. Behavioural scores such as 'delphi' scores have been developed differently within credit bureaus in the UK. They are a combination of behavioural indices and financial credit scores. Those indices may reflect the tendency of a customer to apply for credit, known as 'propensity', collection indices that investigate the likelihood that a customer is going to pay for one's delinquent account within 3 months, and screening indices that estimates which

campaign would trigger a customer to apply for a loan based on one's behaviour in the recent past (Experian, 2009). With the scalable powers of big data infrastructure, machine learning models and devices that collect data from many connected sources (i.e. internet-of-things), lenders were able to come up with many models that outperformed the traditional financial credit scoring methods (Dash et al., 2017).

As a result, lenders sought alternative sources to traditional data found in the financial institutions. For example, lenders look for device identifiers when application is completed online to verify locations of borrowers, asses creditworthiness, and prevent fraud or money laundering operations (HSBC, 2018). Behavioural scoring has been found in many progressive lending institutions that considered e-commerce data, social rental (McEvoy & Chakraborty, 2014), psychometrics tests (Arráiz et al., 2017; Klinger, Khwaja, & LaMonte, 2013), social media data (Bachrach et al., 2012; Guo et al., 2016; Kosinski et al., 2013; Masyutin, 2015; Rusli, 2013; Weke & Ntwiga, 2016), web analytics, telecommunication and mobile phones (Bjorkegren & Grissen, 2015; FCA, 2017), and social network circles (Lin et al., 2013).

Lenders collect and investigate behavioural data to come up with an alternative. This is because, when it comes to mobile phone usage, there are 1 billion people around the world who own mobile phones yet do not have bank accounts as of 2013 (De Koker & Jentzsch, 2013). Also, Dash et al. (2017) gave examples on unconventional data coming from non-financial contexts such as government statistics, utility bill payments, supermarket loyalty card transactions and geospatial data that in many cases go back to earlier days than those of the credit bureaus (McEvoy & Chakraborty, 2014). Meanwhile, the nature of credit given, sometimes, suggests using behavioural data. For example, Kruppa et al. (2013) discussed the case of a store that sells home appliances through instalments. They used a dataset where behavioural attributes were existing. Those were distance travelled to the store, requested time for delivery, and whether delivery is to be made to the buyer's residential address or to another address. Another example is ANZ bank in Australia that gave mortgages to borrowers with customised credit limits, interest rates and tenures based on a borrower's behavioural risk (Birch, 2018). Not only that, but also many lenders highlighted the availability of behavioural data in unstructured forms such as chat transcripts, voice, e-stores that provide feedback/reviews on customers (texts) and data from social media platforms. Such data were considered by lending companies and peer-to-peer (P2P) social lending platforms such as the likes of Prosper, Lenddo, Entrepreneurial Financial Lab, Zopa, lending club, Financiera Confianza,

and many more.

Initially, behavioural scoring targeted the prediction of borrower's ability and willingness to repay a loan based on behavioural aspects (Lyn C Thomas, 2009). Lenders and credit risk scorers/assessors are creative in measuring behaviour as a predictor of credit risk score. In the case of mobile phone usages, the regularity and payment patterns of mobile phones charges whether pre-paid or billed, when considered by bankers, has contributed to reducing the defaults of credit given by banks in developing countries by 41 per cent (Bjorkegren & Grissen, 2015). Also, it was noted that average phone-calls' durations, time of the calls, calls origination would indicate someone's credit worthiness (McEvoy & Chakraborty, 2014). Additionally, the regularity of charging a smart phone's battery was thought of an indicative behaviour of an applicant's responsible borrowing (Fitzgerald, 2018)

Meanwhile, some cautious lenders praised the idea of estimating traditional criteria using proxies that are found within behavioural data in order to match or complement a variable under scrutiny. For instance, when considering age, many lenders realised that traditional models exclude younger borrowers. Therefore, lenders such as SoFi and NeoFinance designed models that can predict which youngsters are likely to be high earners in the future (Rusli, 2013) based on their behaviour as students or young professionals. As a result, they contributed to financial inclusion. Similarly, Affirm Inc. collects meta data about identities from Facebook and Gmail in order to verify the person's history in national datasets (Rusli, 2013). The big five personality traits discussed earlier in this chapter (i.e. agreeableness, consciousness, neuroticism, openness, and extroversion) were measured using social media data. Activities such as likes and subscribed pages predicted traditional credit scoring variables such as marital status, ethnicity, and gender. Also, it was suggested by Masyutin (2015) that income level can be estimated by looking at countries that a social media user visits. This would indicate to lenders the income bracket of a borrower based on the price indices of those countries.

On the other hand, other variables estimated were not, traditionally, correlated with credit outcomes such as political affiliations (Bachrach et al., 2012). Additionally, Kosinski et al. (2013) extended the work of Bachrach et al. (2012) to add to the predictions religion and information on social networks.

When applying such models, SoFi investigated immigrants' social and behavioural activities to allow access to finance. Larger credit bureaus such as Equifax and Experian followed the lead to

make use of the available data points (Rusli, 2013). Behaviour can be estimated using psychometrics or personality traits (Klinger et al., 2013). Psychometric tests are questionnaires used to analyse one's knowledge, abilities, attitudes towards life situations and personality traits (McEvoy & Chakraborty, 2014). Personality and psychometrics influence two behaviours that are associated with credit whether directly or indirectly. The former is the repayment behaviour and the latter are the stability of circumstances whether a borrower's job, marriage, involvement in offences, etc. Those were used by third party credit assessors such as Core Metrix. By using a psychometric tool, EFL classified 1,993 loan applications made to the 5th largest bank in Peru and achieved higher accuracy among banked customers while providing loans to 9% more of those who had previously been rejected to get access to finance (i.e. unbanked) (Arráiz et al., 2017).

In addition to that, VisualDNA used cognitive biases and emotional stability traits to estimate credit risks. Entrepreneurial Finance Lab (EFL) used different traits such as optimism, self-confidence, autonomy, acumen, and opportunism (McEvoy & Chakraborty, 2014) to facilitate credit access for SMEs in developing countries. Those SMEs were assessed based on: (a) ability, by measuring personality and intelligence of managers; and (b) willingness, through integrity of managers as well, to repay a loan. Recently, a psycho-behavioural variable known as 'sensation-seeking' was introduced as an indicator of excessive risks a consumer presents to both lenders and car insurance providers (Brockett & Golden, 2007).

Also, behaviour can be estimated by looking at a customer's response to a loan offer letter. Models targeted a triggered behaviour towards credit were known as 'propensity' models (Experian.co.uk, 2013). In such a case, a customer, who has many offers would not respond immediately and may not take-up the offer. However, a desperate borrower would take-up any offer regardless of its price and conditions. Therefore, banks have been approaching with an initial offer terms that are subject to change based on behavioural responses from the borrowers (Lyn C Thomas, 2009).

In light of the above, researchers have started to build models to measure credit solely on behavioural data collected from social media platforms. Social media data provide extensive and indicative texts, activities, pictures, videos, blogs, interests, and social network ties (Han et al., 2011). In 2018, with the introduction of the powerful open banking financial initiatives, Birch (2018) put his confidence in behavioural data and asserted that LinkedIn may have as good accuracy of credit scoring as financial data does with banks do. The study went further to argue the plausibility of using data collected by Microsoft, the parent company of LinkedIn, and the other

‘GAFA’ members (i.e. Google, Apple, Facebook, Amazon) such as search inquiries, locations, and shopping patterns.

In social media, behavioural metrics suggested by Masyutin (2015), for credit scoring calculations, were marital status (categorised based on classes in love, engaged, it’s complicated, etc.), political views, age (matched from social media), sex, number of days since last visit, number of days since the first post (on social media), number of job places, number of subscriptions, number of user’s posts with photos, number of posts with video, number of children, major life events attended (career and money, entertainment, fame and influence, research and science, etc.), and major qualities (such as creativity, humour, etc.). Also, Weke and Ntwiga (2016) highlighted three advantages when using social media data in assessing credit risks: first, by capturing valuable borrowers who have limited financial history and, thus, overcoming the adverse selection effects. Secondly, understanding the real customer needs and, thus, reducing moral hazard effects. Finally, matching information provided by borrowers with their social media profiles to overcome information asymmetry. The use of social media has been found in practice, as well, where lenders and assessors found it useful in many ways. For example, MasterCard Advisors’ report emphasised that checking data on social media is a way to verify identity. In addition to that it can be used to estimate one’s income level. Thirdly, information on employment status can be extracted from social media (McEvoy & Chakraborty, 2014). When verifying jobs and stability with employers, Lenders can use professional social media platforms such as LinkedIn. By doing so, those can check employment history of an applicant. Not only that, but they can perform a ‘cross-pathing’ exercise where they seek a colleague who works for the same employer at the same period of time and cross-check that colleague’s financials and history (Bradbury, 2011).

In other research on credit scoring of entrepreneurs in SMEs, three main behavioural factors were found to have contributed to the outcome loans given to SMEs: personality, intelligence, and integrity. Personality was explained in the last paragraph of the personality trait sub-section (see sub-section 2.2.4.5) and earlier in this sub-section where many models, such as the OCEAN five-factor model, can help estimating it. Intelligence is important when it comes to the management’s ability to manage wealth and foresee financial commitments that may arise. Finally, integrity forecasts the SME’s willingness to repay its obligations (Arráiz et al., 2017).

In literature, factors discussed by Zhang et al. (2016), such as life stability, can be measured through a borrower’s behaviour and, thus, can be an input for credit risk modelling. Attitudes

proposed by Judge et al. (2001) can help predicting patterns associated with different situations with regards to behaviours when repaying loans. Also, Nye (2014) highlighted why the incorporation of the behavioural finance variables was adopted by credit referencing agencies. Below are behavioural aspects and personality traits found in the literature, which would relate to behaviour of borrowers in theory.

Extroversion is the tendency to be a sociable, gregarious, assertive, talkative, active (Barrick & Mount, 1991), and enthusiastic (Hirsh & Peterson, 2009). Also, it is defined by Bachrach et al. (2012) as seeking stimulation from the external environment represented by friendships and happy environment. Extroverts are easily-distracted and are more influenced by opinions from the external environment (Zafar & Meenakshi, 2012). Therefore, in the times of financial crises or bad financial performance of people within the social networks of extroverts, there is a higher credit risk associated with those. On the other hand, when social ties perform well, extroverts get inspired by such an environment and demonstrate less credit risks. Overall, extroverts seem to be motivated by the positive reward of repayment such as an offer to refinance with better conditions after sometime of successful on-time payment. To the contrary, introverts are quiet and prefer reading to having large friendship ties (Zafar & Meenakshi, 2012). Accordingly, by determining that a borrower is an introvert, lender may limit the credit scoring to financial traditional data. This is because the information gained from one's social network would be, arguably, meaningless.

Neuroticism is represented by emotions linked to negative effects (Robinson, Ode, Moeller, & Goetz, 2007) and instability measured by the degree of anxiety, depression, anger, embarrassment, and other emotions that are expressed by the person openly (Bachrach et al., 2012; Barrick & Mount, 1991). It is manifested by volatility and withdrawal behaviour as well (Hirsh & Peterson, 2009). One of the causes of neuroticism is loneliness (Correa et al., 2010). It was concluded by Hirsh and Peterson (2009) that those who exhibit a withdrawal behaviour and are highly-neurotic tend to cooperate when agreements are made. This is because those worry about the consequences of defecting on such agreements. Overall, the more severe the consequences are, the more likely a neurotic person would repay (Hirsh & Peterson, 2009). Nevertheless, some researchers argued that neurotic individuals, also, exhibit excessive buying habits and might be subject to impulsivity (discussed later in this section) which is problematic (Otero-López & Villardefrancos, 2013).

Agreeableness measures how courteous, flexible, trusting, good-natured, tolerant (Barrick & Mount, 1991), compassionate, and polite (Hirsh & Peterson, 2009) a person is. Usually, an

agreeable person weighs maintaining positive social relationships very heavily (Bachrach et al., 2012). Hence, it is argued that an agreeable borrower who is surrounded by irresponsible social ties is more likely to behave similarly in order to please those ties. Conversely, an individual with high agreeableness is more likely to be a good borrower if one's social ties are financial exemplars.

Conscientiousness reflects industriousness and orderliness (Hirsh & Peterson, 2009) as it highlights a preference for an organised lifestyle rather than a spontaneous one (Bachrach et al., 2012). Conscientious individuals are dependable and have the will to achieve high goals. Dependability is represented by careful, thorough and responsible persona. Achievements happen when a planned, organised, persistent, and a hardworking person is in charge (Barrick & Mount, 1991).

Openness refers to being imaginative, curious about different cultures, intelligent and artistic (Barrick & Mount, 1991). As explained in the job performance previous sub-section (see 2.2.4.5), workers characterised with high openness scores are thought of as artistic and talented. Nevertheless, those do not necessarily succeed in tasks that require routine and strict guidelines. Therefore, determining the level of openness to unusual ideas along with the type of jobs that a borrower has can assist in enhancing the credit scoring process.

Materialism reflects the belief on the importance of a good to someone's life, the success that is perceived by this person (or by others) when owning such a good, and the happiness it brings by its possession. Materialism has shown high correlations with neuroticism and is a main driver of compulsive buying (Otero-López & Villardefrancos, 2013) as well as entering instalment plan agreements (L. Wang et al., 2011). In light of the aforementioned properties, materialism signals irrational spending and, thus, bad financial management. Therefore, when borrowers exhibit a tendency to spend on luxurious goods or goods that are perceived by people as indicative of doing well, lenders need to activate a close monitoring mechanism by ensuring that spending of the loan is on its sole purpose.

Overconfidence refers to estimating higher performance or outcome than the reality which leads to precise prediction of the uncertain future. Overconfidence explain many cases of business failures (Peón et al., 2016). Accordingly, researchers argued that overly confident managers enter new markets relying on debt (Ackert & Deaves, 2009). Overconfidence can be measured by asking questions to loan applicants about different facts within surroundings given a true/false option or a multiple choice one. The respondent will then be asked of how confident he is of his answers

and a comparison will be done between the percentage the borrower gave and the actual forecast completed by an economic research group to check confidence levels. Usually, overconfident individuals are less prepared when incidents and exceptional circumstances arise (McEvoy & Chakraborty, 2014).

Optimistic bias (or unrealistic optimism) happens in credit when consumers are optimistic about their repayment abilities. For example, in credit cards, if the card holder is settling the monthly balance, there will be no interest charges incurred. Accordingly, annual percentage rate (APR) becomes irrelevant to many of the borrowers who believe they can clear outstanding balances easily. Therefore, those who are unrealistically-optimistic tend to not worry about APRs and apply for loans with terms and conditions that are not in their best interests. In response, lenders should examine borrowers' past expectations on their payment patterns and compare it with the actual performance before deciding on a new or an extended loan (S. Yang, Markoczy, & Qi, 2007).

Egocentrism is a combination of both optimistic bias and overconfidence. In other words, lenders would, unrealistically, expect the most optimistic scenario to happen in the future. Meanwhile, those would falsely-believe that they are able to cope with any alternative scenario. Egocentric individuals tend to fail to, adequately, estimate risks (Kruger & Burrus, 2004). As with overconfidence and optimistic bias, lenders need to measure the level of egocentrism of loan applicants. This can be done by comparing the borrower's own beliefs on the likelihood of desired and undesired events with the probabilities reported in market research. Also, measuring the preparedness of borrowers to the extreme undesirable events would be a positive indicator of a rightfully-confident borrower.

Temporal Discounting is seen in young adults when they aim to buy desirable consumer products with a loan given at a high interest rate. Such a loan could be unaffordable on the long-run. Some young adults who incur debt on their credit card appear to be ignorant about future consequences of their indebtedness (Omar, Rahim, Wel, & Alam, 2014). The reasons behind such a behaviour are feelings of financial deficit, attitude towards borrowing, and financial involvement and knowledge. In summary, young adults suffer the 'tunnel vision' effect when applying for a loan to buy desirable goods. Also, they hope to break free from the deficit feeling (Gärling, Michaelsen, & Gamble, 2020). In order for lenders to avoid temporal discounters, they have to investigate the borrower's long-term plan and foreseeable future liabilities. Accordingly, translating those into projected cash outflows factoring-in inflation rates is the right way forward.

Prospect theory asserts that investors tend to treat gains different to losses which would lead to an overweight of small probabilities (Peón et al., 2016) resulting-in euphoric expectations by borrowers (Guttentag & Herring, 1984). In order to avoid adverse selection, lenders must question whether the borrower is aware of the likelihoods of undesirable events along with their consequences.

Sensation-seeking is usually characterised in four components: thrill in adventure, new experience, boredom susceptibility and disinhibition as in the release from all restrictions (Ackert & Deaves, 2009). In credit context, examining the purpose of the loan can be indicative. When a borrower states that the purpose of the loan is to buy stocks as a first-time buyer, both of adventurous thrills and debuts bring that sensation in for a sensation-seeker. In terms of breaking off from restrictions, lenders have to pay attention when loans are given to borrowers, who state that they are planning to leave their jobs and start a business from the loan amount. Also, sensation-seeking can be thought of as an arousal that one feels when assuming risk (Taffler, 2017). Therefore, making sure that loans are used and/or invested in a low-risk portfolio is essential.

Impulsivity (or impatience) is defined as predisposition to make choices favouring immediate benefits rather than remote ones due to lack of control. It is thought of as a biological factor that develops within someone's personality. Impulsive borrowers can be identified by their behaviour when shopping online using their credit cards (Henegar et al., 2013). Accordingly, when the aforementioned borrowers apply for new loans or extend their current credit, lenders should take that into consideration.

Homophily (or assortative mixing) represents the formation of friendship ties over common grounds such as age, nationality, income, language, educational level, and many other characteristics. Estimating the degree of homophily can be done by knowing the types of each node (vertex) in someone's network then calculating the proportion of edges that links the individual with a specific type out of all edges (M. Newman, 2010). Understanding borrowers' friendship ties that make up one's social network and its effects on credit scores is the main focus of this research.

Status-quo bias (or endowment effect) happens when individuals perceive goods to be more valuable than what their real values are worth once owned. As a result, individuals, in general, or borrowers suffer from 'comfort-seeking' and tend to stick to only one good or investment (Ackert & Deaves, 2009). Initially, people who have this kind of bias are risk-averse and they tend to stick

because they believe this is a safe option (Rasmussen, 1998). However, sticking with current states can be risky when borrowers seek buying or investing in a good or an asset that is outdated and has low functionality or earning potential in the era of rapid development. Ackert and Deaves (2009) emphasised that extrapolation needs to be adopted as opposed to mean analysis. In other words, lenders need to use time-series analysis of loans and records of lenders to ensure that the borrower is responding well to changes in the market and technologies.

Anchoring (or herding behaviour) is related to pitching an aspect of a product that will affect consumers' behaviour dramatically in a negative way while disregarding other aspects. It was reported by FCA (2018) that anchoring affects credit card holders when it comes to their choices of repayment. Those would most likely choose to pay the contractual minimum repayment regardless of their affordability in the future. Therefore, there has been a discussion on whether to mandate removing the minimum repayment statement from the credit card contract, a move towards 'de-anchoring' (FCA, 2018).

Hindsight bias, on the other hand, is resembled with past experiences and skills that are idealised in memory and is somehow ineffective at present (Ackert & Deaves, 2009). This kind of biases are common in technical fields with people who fail to upgrade their knowledge. Lenders who base their decision on the income streams of a technician or a project manager would most-likely be vulnerable if no knowledge of recent achievements and certifications are provided.

Framing is a behavioural concept that contradicts rationality in expected utility theory. It is built on the assumption that individuals make financial decisions based on how a decision frame is being presented to the decision maker (Ackert & Deaves, 2009). For example, when loans to buy equity in government bonds are presented as a way to secure a stable income for retirement, borrowers would react differently to when those are presented as lower income stream than growth stocks that have high potential.

Integration is a way of sequential thinking where borrowers could potentially react based on the outcome of recent event. An individual with high integration levels is unable to segregate issues (Ackert & Deaves, 2009) and may make unreasonable decisions. Lenders should investigate the recent activities of borrowers and find out if the loan is thought of as a compensation to a loss.

Emotional stability is found to be negatively-correlated with the use of social media. The aforementioned relationship was clear with young women. Survey questions that indicated low stability are those who ask if life has been difficult or unrewarding. Most of unstable individuals

do suffer from loneliness (Correa et al., 2010). Obviously, an emotionally-stable borrower is desirable. Lenders needs to investigate whether a borrower is unstable emotionally. Kosinski et al. (2013) highlighted that emotional instability would encourage individuals to seek products that provide security. This would trigger a high risk when affordability is not been factored-in with the borrower. Lenders may detect such individuals by running behavioural models that detect dissatisfaction with life variables.

Cognitive bias in credit is a systematic tendency to underestimate the annual percentage rate by considering the instalment payments only. It has been argued that consumers with such biases are, mainly, affected by payment-based loan offers (Stango & Zinman, 2006). Lenders have to recognise if borrowers are applying for credit to satisfy an urgent spending or because they want to even out peaks and troughs of expenses across a period of time – a month, quarter, or a year (McEvoy & Chakraborty, 2014). On the other hand, lenders should refrain from attracting those identified with cognitive biases as economies may become vulnerable to financial crises (Guttentag & Herring, 1984). Both of cognitive bias and emotional stability are used to assess credit risk by a company called VisualDNA. The company estimates the aforementioned behaviour using psychometric tests (McEvoy & Chakraborty, 2014).

Autonomy in credit reflects the decision-making power (Lont, 2001). It is used by a third party credit assessors, EFL, as a main psychometric feature to calculate behavioural credit scores (McEvoy & Chakraborty, 2014). Simply, lenders look for borrowers who enjoy a high financial autonomy to ensure that decisions to borrow and spend the loan amounts would not be affected by other individuals.

Acumen is the collection of knowledge, skills, and experiences that transform into strategic actions of one's behaviour. Individuals with high acumen are able to administer their finances and keep their saving accounts healthy (Gargiulo, Pangarkar, Kirkwood, & Bunzel, 2006). It was reported by McEvoy and Chakraborty (2014) that EFL measures the levels of borrowers' acumen using psychometric tests. Additionally, Gargiulo et al. (2006) asserted that those who demonstrate high acumen would easily upgrade their knowledge as developing technologies emerge. Accordingly, lenders should seek borrowers with high acumen. Similarly, **opportunism** is an indicator that lenders deem important to identify one's strengths and abilities to solve problems in order to capitalise on those. The aforementioned strategy in credit scoring is existing with market credit assessors such as EFL (McEvoy & Chakraborty, 2014)

Self-esteem represents individuals' feelings about themselves in terms of value and degree of positivity of the self-concept (Omar et al., 2014). Low self-esteem, on the other hand, is connected with highly-neurotic individuals where those reach out to friends and social ties for mental support (Correa et al., 2010) while attempting to impress others with the purchase of luxury goods (Omar et al., 2014). It is, therefore, concluded that lenders should seek individuals with low levels of neuroticism and high self-esteem in order to ensure that loans are really for genuine reasons and not for approval among members of the society.

Gratification represents consumers' taste and, in literature, its influence on credit was discussed and found to be indicative of consumer credit risk (Heidhues & Kőszegi, 2010; L. Wang et al., 2011; Wood, 1998). Particularly, time-inconsistent tastes negatively-affect those who have immediate gratification by mis-predicting this those tastes over time (Heidhues & Kőszegi, 2010). In practice, EFL considered gratification as one of the variables to measure in their psychometric tests (McEvoy & Chakraborty, 2014). Overall, deferring gratification is thought to be a good indication in borrowers where spending on interests and preferences does not happen impulsively.

As for ways to collect the above-mentioned behavioural data, application programming interfaces (APIs) in social data mining were used to access data from social media platforms (Bachrach et al., 2012; Kosinski et al., 2013; Zielinski, Middleton, Tokarchuk, & Wang, 2013). For those who do not have social media profiles, data from phone providers and smart phones can be sourced as well. The aforementioned has been called digital footprints. Finally, psychometric tests can be conducted on those who are 'unbanked' and 'unphoned' (Fitzgerald, 2018). The advantages of real-time data and scalability can be complemented with accuracy especially when determining noise and selecting keywords, hashtags, and mentions in social media platforms subjectively (Zielinski et al., 2013).

In modelling, Masyutin (2015) contributed to behavioural modelling by designing two types of score cards of data collected from a Russian social media platform, Vkontakte. The first scorecard estimated a default due to inability to repay; while, the second estimated a default due to a fraudulent behaviour. He concluded that a social-media-data-driven model should serve as a complement to the traditional credit scoring model.

Banks with analytics teams are looking at credit bureau reports which incorporate data from various sources such as utility bills and electoral vote rights. In addition to that, those banks merge government statistics and alternative data. Government statistics vary from utility bill payment to

county court judgements. In their credit scoring guideline, HSBC stated that those with court judgements registered in their name have shown less commitment to payment on time (HSBC, 2018). On the other hand, alternative data vary from supermarket loyalty cards to geospatial data. Also, banks extract unstructured data such as customer reviews, chats and voice transcripts (Dash et al., 2017). In the last decade, social media platforms were considered as a new source of data. Those provided different structured and unstructured valuable data to lenders. In recruitment and human resource management, firms watch their employees' use of social media networks especially when it comes to professional behaviour on LinkedIn (Bradbury, 2011).

Finally, Masyutin (2015) raised concerns in business practices when sourcing behavioural data. He argued that companies who appoint competent data scientists face the challenge of changing the business models and the hierarchy of the business such as the organisational structure. On the other hand, assigning the behavioural data collection task to a third party would complicate the legalities and liabilities while risking the reputation of the business. Also, the privacy of the personal data of customers base is at stake. Some behavioural models may promote bias as they would rely on social media data to estimate gender or ethnicity for example (Kosinski et al., 2013). However, the majority of models had been successful in predicting traditional fundamental attributes that are used in financial models as mentioned earlier in this sub-section.

On some occasions, behaviour was manipulated and staged. This has caused issues to some lenders such as prosper and PDPai. In order facilitate loans to friends, individuals who were supposed to act on behalf of the lenders were subject to agency theory (Ackert & Deaves, 2009). For example, in Chinese P2P lending platforms, PDPai, high endorsements were, falsely, given to friends for socioeconomic benefits (Zhang et al., 2016). Similarly, In the US, group leaders benefited from the $12 reward on every loan given to their group members, so they endorsed those blindly on the P2P lending platform, prosper (Freedman & Jin, 2017). This has prompted researchers and innovative lenders to analyse networks further and rely on relational aspects existing in social network data. The structure, types and sizes of networks were investigated. Social networks will be discussed later in this chapter (see section 0).

2.2.5. Dynamic Criteria

The FCA has recommended that credit scoring runs on an ongoing basis. Initially, the suggestion was to run the scoring model whenever a borrower applies for a new loan, whenever an existing loan is to be extended, or whenever a limit is to be increased (FCA, 2018). However, some lenders

run the model on a periodic-basis. Although credit scoring criteria are, mainly, financial and socioeconomic when it comes to lending at banks, different lending channels such as peer-to-peer (P2P) platforms and challenger banks have adopted alternative criteria. In addition to the aforementioned lenders, credit referencing agencies rely on innovative data (McEvoy & Chakraborty, 2014). This allows examining many of the variables that are dynamic and would change over time.

Consequently, some researchers argued that models need to be dynamic (Lyn C Thomas, 2009). In Dynamic modelling, researchers realised that variables to be measured continuously using an effective model. In defining dynamic models, Sousa et al. (2016) asserted that streaming data will provide an input for sequential learning in these models. For example, Bellotti and Crook (2009b) used survival analysis to, dynamically, estimate economic variables that change throughout the economic lifecycles. Specifically, Leow and Crook (2016) focused on dynamic manifestations in bank accounts such as: average transaction value, number of cash withdrawals, credit limit changes, rate of total jumps proportion of months in arrears, and repayment amount and outstanding balance.

The continuous changes in borrowers' circumstances are thought to be the fundamental reason behind dynamic models (Banasik et al., 1999). Accordingly, financial distresses can be caused by abrupt life events such as divorce, unemployment, and disease or illness in addition to behaviour associated with bad wealth management and erratic spending (Sousa et al., 2016). In South Africa, lenders denied 50% of loan applications due to unconfirmed employment, suspicious of fraud, bad credit rating, and high debt burden ratio (Karlan & Zinman, 2009).

Additionally, researchers had incorporated the continuous activities that take place on social media platforms in credit risk scoring. For example, Vkontakte, a Russian social media platform, was the data source for the study of Masyutin (2015), who determined credit scores based on desirable activity. Moreover, other researchers like Bjorkegren and Grissen (2015) derived credit risk behaviour from telecommunication data.

The table below (see Table 4) provides an example of dynamic criteria cited by researchers in credit risk. Some researchers focused on methods to measure traditional data dynamically from the financial aspect. Others resorted to alternative data and applied dynamic models on those to predict their behavioural credit risk scores.

Traditional	Behavioural

Leow and Crook (2016)

- Average transaction value
- Number of cash withdrawals
- Credit limit changes
- Rate of total jumps
- Proportion of months in arrears
- Repayment amount and outstanding balance

Rusli (2013)

- Education
- Employment History

Óskarsdóttir, Bravo, Sarraute, Vanthienen, and Baesens (2019)

- Age
- Amount spent in the month prior to the loan

Masyutin (2015)

- Social media metadata (marital status in love, it's complicated, etc.)
- Activity (high number of videos posted)
- Medium (3-7) number of page subscriptions, etc.

Bjorkegren and Grissen (2015)

- Intensity and distribution of calls over time.
- Top-up and depletion

Rusli (2013)

- Followers
- Friends and their type(s)

Sousa et al. (2016)

- Divorce
- Unemployment
- Disease

Óskarsdóttir et al. (2019)

- Phone call durations
- The number of phone calls received from social ties with late payments

Table 4: examples of dynamic criteria in literature

In practice, many lending institutions adopted dynamic criteria. Those proposed scenarios of default symptoms represented by events. Such events would be characterised by severity. Examples of such systems can be Experian's data labs (see sub-section 3.4.1).

2.2.6. Summary of Credit Scoring Criteria

In light of the above discussion on traditional and behavioural criteria, the below diagram (see Figure 3) illustrates howscoring is being conducted by lenders. It is worth to mention that some of the components might be dismissed depending on the type and size of loan as well as the capability of the lender. For example, high-street legacy banks would capitalise on long transactional record as well as economists who would issue industry-specific macro-economic studies, so those might forgo the behavioural criteria or the credit referencing agency (CRA) data. Conversely, a new start-

up would rely mainly on application data internally and CRA externally. Also, it may collaborate with another psychometrics start-up on a behavioural model.

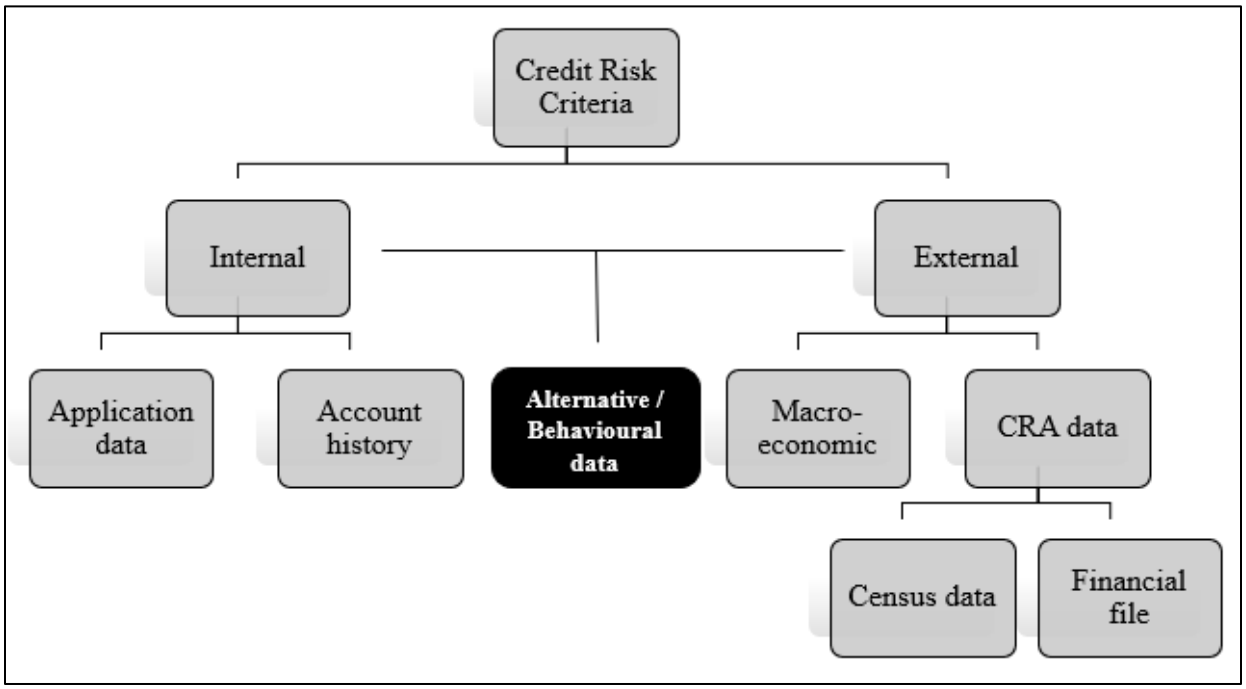

Figure 3: credit scoring criteria

As illustrated in the above diagram, behavioural data has been introduced in the literature as the 'alternative data' where lenders would rely on non-financial data collected about the borrowers. Such data can be collected, internally, such as whether a borrower has responded to a campaign email or who accompanied the borrower to the lender's offices. Also, lenders may buy an access to behavioural datasets such as access to telecommunication dataset through an agreement with the service provider. The model would be developed and behavioural credit score can still be produced internally.

Alternatively, behavioural data can be collected by a third-party credit assessor such as a credit bureau or any other credit referencing agency (CRA). A behavioural score would be produced externally based on geo-spatial data of the borrower or a personality trait conducted by a company that does psychometrics via tests or via social media activities.

It is worth to mention that CRAs have access to financial files from transactions completed with other financial institutions by the borrower. Also, they may have access to census data which includes crime rate in the specific post code where the borrower resides, electoral roll registry, and any court case against the borrower as explained earlier in this chapter.

2.3. Credit Risk Models

The issue of adopting the best credit risk models has been essential for lenders who want to asses credit risks. On the other hand, it is essential for borrowers to make sure that loans that are given match their affordability. In other words, repayment can be done while causing financial distresses, which would reflect negatively on individuals. The aforementioned combination of credit risk and affordability comprise the concept of creditworthiness (FCA, 2018). In credit risk, researchers discussed models that estimate three different components of credit risk. Some of those models targeted consumers' probability of default (PD); whereas, other models estimated loss given default (LGD), and, finally, the third type of models sought exposure at default (EAD) in credit (Tong, Mues, Brown, & Thomas, 2016). Basel accords requires calculating the expected loss (EL) of a loan by multiplying the aforementioned three values (Leow & Crook, 2016).

EAD applies to revolving credit such as credit cards and overdraft facilities. It is used to estimate the outstanding balance of an account at any point during the life of the loan (Leow & Crook, 2016). Modelling EAD requires measuring the changes of states between every payment's due date of which survival analysis was capable (Banasik et al., 1999). A cox proportional hazard model is an application of survival analysis that segments borrowers' states successfully (Ha & Krishnan, 2012).

LGD, on the other hand, estimates the amounts that the lender would lose in the event of default in an estimated point of time (Schuermann, 2004). Both EAD and LGD become relevant in the case of default (Leow & Crook, 2016).

According to Basel II committee's definition, a default happens when, at any point of time during the loan's tenure, 90 days' worth of instalments are in arrears within a 360-day period [2] (Puri et al., 2017). Accordingly, models that are designed to estimate PD, usually, are set for a binary target. In other words, there are only one of two outcomes – default or repayment in which a reject or an accept decision to be made (Gordy, 2000). In practice, a model that produces absolute accurate classifications does not exist because samples are often unrepresentative of the population analysed (Hsieh, 2005).

Models vary from static and parametric to dynamic and non-parametric based on how explainable

[2] In addition to failing to pay 3 months' worth of instalments within a year, there are other less-common cases of default associated with banks, mostly, such as: (a) borrower is unlikely to repay anymore, (b) bank is considering the loan as a loss, (c) borrower's liabilities are restructured with the loan becoming a loss, (d) bank calls the loan, (e) selling the loan with a loss, and (f) the need to write-off the loan (Puri, Rocholl, & Steffen, 2017)

scores or decisions have to be and the number of attributes to be included in the analysis (C. Chen et al., 2018). The most commonly-used models for PD are logistic regression and linear discriminant analysis (Y. Yang, 2007) due to their abilities to fit a binary distribution with the logistic regression's sigmoid function (James, Witten, Hastie, & Tibshirani, 2013) and the discriminant analysis' pre-defined set of input vectors that are linked to target classes (Y. Yang, 2007). A logistic regression model allows for the building of scorecards from which a linear programming technique can choose the most optimised scorecard (L. Thomas et al., 2001). On the other hand, machine learning models had performed, exceptionally, well with their iterative learning feature and ability to handle high dimensions. The normal procedure in machine learning is to split the dataset into a large part (70-80%) of training set then test the model on the remaining (20-30%) holdout part (Han et al., 2011). In a nutshell, they improve with more data points and can handle more features. The drawback of machine learning models is their limited interoperability when it comes to understanding the PD score. Examples of the aforementioned models are neural networks, linear discriminant analysis, k-nearest neighbor (West, 2000), incremental kernel learning (Y. Yang, 2007; Zhang et al., 2016), naïve Bayes algorithms (Peluso, Mira, & Muliere, 2015), support vector machines (Bellotti & Crook, 2009b; Huang, Chen, & Wang, 2007), random forests, XGBoost (Dash et al., 2017; Zięba, Tomczak, & Tomczak, 2016), and genetic algorithms (Huang et al., 2007).

The aim behind designing a credit risk model is to increase its classification accuracy, in detecting low credit-worthy applicants, in addition to adding speed and simplicity to its performance (Liu & Schumann, 2005). While maximising accuracy, a model should still be transparent and an automated decision has to be justified whenever challenged. The aforementioned condition is enforced by lenders in the European Union and the UK while acknowledging the importance When measuring its accuracy, a model should minimise the expected loss (EL) score that is the product of probability of default (PD), loss given default (LGD), and exposure at default (EAD) ratios (Tong et al., 2016). A good model would normally produce a score card that produces a large area under receiver's operating characteristic (AUROC) curve. ROC curve will be used in this research to assess the performance of the credit risk model (see section 4.4.7.2). Also, it is worth to note that credit scores gave flexibility to lenders. For some lenders, a score might be acceptable whereas for others it may not be. This depends on the risk tolerance of the lenders and their objectives set by their credit officers (HSBC, 2018). Lyn C Thomas (2009) identified three objectives within the

revolution of credit risk modelling. He highlighted the lender's increased emphasis on profit maximisation by cross-selling. Also, he added that market penetration can be achieved by trying to attract new borrowers and retain current ones by stopping churning/attritions as much as possible. Finally, deciding what an appealing return on equity would a borrower consider by estimating the acceptance response rate.

As an extension to the profit maximisation modes, Lin et al. (2013) proposed panel models in peer-to-peer (P2P) lending. Those models measure which lender is more likely to respond to a borrower's request and by what proportion such a lender would cover the requested amount.

PD models assist in scoring customers and, in addition to providing a score, explainable models produce scorecards that can detail the effects of each class within the variable. The table below (see Table 5) illustrates an example of a scorecard based on eight variable – accommodation, years with employer, having credit card, savings account, other accounts, occupation, previous account, and credit bureau (Jensen, 1992). It justifies and explains a credit score for every potential borrower based on the input. The score is benchmarked to a cut-off score and decision is made accordingly. The aforementioned case reflects a binary outcome; however, the current practice uses many cut-off points in which the product and interest rate would defer according to the category of the credit score

Years with employer		**Accommodation**		**Credit Card(s)**	
Under 1 year	15	Own	45	Card	19
1-2 years	22	Rent	18	No card	0
3-9 years	26	Other	24	**Previous Account**	
10-12 years	29	**Savings Account**		Unsatisfactory	0
13 years or more	36	Yes	36	New	55
Occupation		No	0	Satisfactory	87
Professional	29	**Other Bank Accounts**		**Credit Bureaus**	
Office staff	25	Cheque & saving	50	No file	15
Production	15	Current	31	Derogatory	-33
Sales	22	Deposit	32	Satisfactory	24
Other	15	No account	5	Outstanding	30

Table 5: example of a scorecard (Jensen, 1992)

2.3.1. Parametric Models

The information used in assessing applicant's credit risk used to be collected during application, where data collected is called 'application data' whereas the score is known as 'application score. Models used on these types of data were static (Lyn C Thomas, 2009). Parametric models are highly-credited for their interoperability such as logistic regression, discriminant analysis (Y. Yang, 2007) and linear programming.

Parametric models make assumption that a function that describes the relationship between the predictors (or the independent variables) and the response (or the dependent variable) can be estimated using parameters due to its linearity. These types of models are based on the method of least squares (James et al., 2013).

2.3.1.1. Logistic Regression

Logistic regression models are the most common credit scoring model (Crook, Edelman, & Thomas, 2007; Kruppa et al., 2013; Lyn C Thomas et al., 2002). Initially, credit risk problem was perceived as a binary issue of predicting good and bad loans (Sousa et al., 2016). The logistic regression predictive analysis was proposed as a method to describe and predict the relationship between a binary or a 'dichotomous' (Lee, Chiu, Lu, & Chen, 2002) dependent variable (i.e. the loan outcome) and nominal, ordinal, interval or ratio independent predictors (i.e. the different criteria adopted by the lender). The model aims to predict whether the applicant would default on the loan requested or not and assist in choosing between the two decisions ("What is Logistic Regression,"). Logistic Regression is, also, known to be used to build a scorecard model (Banasik et al., 1999), where odds of good candidates to bad ones are calculated at each point of input characteristic (Lyn C Thomas et al., 2002).

When scorecards were produced using logistic regression, it was noted that ordinal results were produced which suggested a scale that would enable the lender to create a score for each of its customers (Lyn C Thomas et al., 2002). Linear regression is used to predict a continuous quantitative variable such as a score was based on the method of least squares invented by Legendre and Gauss. On the other hand, logistic regression predicts a qualitative discrete variable such as good versus bad or up versus down where it was based on discriminant analysis proposed by Fisher in 1963 (James et al., 2013). Therefore, logistic regression was used to estimate the response variable (default or no-default) since this variable violates the normality assumption required for linear regression (Yeh & Lien, 2009).

However, and with the addition to other types of behavioural data, those models suffered from the

‘curse of dimensionality’ or, in other words, handling many characteristics of borrowers provided by many entities and computing capabilities overwhelm those models that are unable to compute the optimal strategy. Some researchers referred to the aforementioned models as parametric models (Sousa et al., 2016; West, 2000). In addition to the inability to process high number of features or dimensions and when these models had limited sample sizes, they were incapable of proving linear relationships or handling non-linear relationships (Y. Yang, 2007). Linear parametric models were thought of as viewing a painting or photo of a borrower at the time of application and comparing it with one’s photos during the repayment in the forthcoming period (Lyn C Thomas, 2009). Simply, adding more colours to the painting or more pixels to the photo increases the resolution, but would complicate the comparison between paintings or photos within those models and make the task overwhelming.

The main advantage of logistic regression models is their simple probabilistic formula of classification. Also, providing more detailed information on the creditworthiness as opposed to non-parametric models (Kruppa et al., 2013). On the flipside, logistic regression suffers when classifying non-linear and interactive explanatory variables (Yeh & Lien, 2009). Kruppa et al. (2013) referred to the aforementioned interactive caveat as the ‘multicollinearity’ problem caused by high correlations between independent variables.

In this research, Logistic Regression model will be adopted in an attempt to classify customers and predict their default using the maximum likelihood procedure that is discussed in details in the methodology chapter (see sub-section 04.4.5.2) and applied thoroughly in the results section of the results, findings and discussion chapter (see sub-section 5.2.4.3).

2.3.1.2. Discriminant Analysis

Discriminant analysis (DA) sought the probability of a customer being good at a specific score (Banasik et al., 1999). Based on those probabilities, discriminant variables are decided and rules for classification are created. DA assumes that, for every class of the response variable (e.g. default and no-default), the explanatory variables are either distributed in a multivariate normal/Gaussian distribution (Yeh & Lien, 2009) or quadratic distribution where quadratic terms dictate the boundary in explanatory variables (Baesens et al., 2003). The former is known as linear discriminant analysis (LDA); whereas, the latter is called quadratic discriminant analysis (QDA) (Baesens et al., 2003). Also, LDA assumes that covariances are equal between those variables; whereas, QDA uses a quadratic equation to separate the surface between populations (Lee et al.,

2002).

DA allows for a common variance-covariance can be initiated (Yeh & Lien, 2009). DA was criticised in literature because it assumes that the discriminating variables are interval. Finally, it assumes a normally-distributed set of discriminating variables which allows a multivariate linear analysis (Desai, Conway, Crook, & Overstreet Jr, 1997). On the other hand, LDA has the ability to deal with high dimensions by reducing those while still providing an understanding of the data (F.-L. Chen & Li, 2010).

DA was adopted widely due to its simplicity. Not only that, but also it overperformed a more sophisticated non-parametric models in some cases such as the results presented in the work of Yobas, Crook, and Ross (1997). Both logistic regression and discriminant analysis techniques were adopted heavily in static modelling. They were successful when handling low-dimensional data with linear nature aiming at a credit decision as an outcome, but limited with more sophisticated banking data and wide range of possible outcomes (Sousa et al., 2016). They were referred to as parametric static models (West, 2000).

2.3.1.3. Linear Programming

Linear programming (LP) was investigated by many researchers in credit scoring and classification as an alternative to linear and quadratic discriminant models (Hardy Jr & Adrian Jr, 1985; Shi, Peng, Xu, & Tang, 2002). The technique is known as an optimisation technique which aims to maximise the minimum distances between borrowing observations and the cut-off value or score. Alternatively, LP model can aim to minimise the sum of deviations among one proposed class of borrowers (He, Liu, Shi, Xu, & Yan, 2004; Shi et al., 2002).

LP models were, first, developed from discriminant models to target a two-class problem -good versus bad borrowers (He et al., 2004). With the development of additional classes, Shi et al. (2002) introduced a multi-criteria LP technique. The aim was to identify the best solution to separate classes from each other (He et al., 2004). On the other hand, fuzzy linear programming was introduced by He et al. (2004) aiming at minimising the separation of observations within a category. Nevertheless, as one can imagine, LP was criticised to be ineffective when it comes to computation times. In addition to the aforementioned caveat, the model is said to be somewhat subjective to users' selection of groups and deciding on critical values. (Shi et al., 2002).

Generally speaking, parametric and linear models do not adjust automatically to changes in the data structure and require rebuilding from the scratch. In other words, they do not work well with

data of poor quality (Y. Yang, 2007). Moreover, they require long training hours and huge memory when learning from large training data sets (Han et al., 2011).

2.3.2. Non-Parametric Models

There has been a dramatic increase in the need for reliance on advanced computing and processing when making financial decisions recently. Particularly, machine learning and its classification powers and techniques. The term non-parametric was given to models that do not assume linear relationships between variables (Brown & Mues, 2012). Some of those are K-nearest neighbour, decision trees (Brown & Mues, 2012), random forests, XGBoost (Dash et al., 2017), incremental kernel learning (Y. Yang, 2007; Zhang et al., 2016), and Naïve Bayes (James et al., 2013; Lyn C Thomas, 2009)

The use of natural language processing (NLP) as well as geospatial analysis were suggested by Dash et al. (2017) in a new model where unconventional data overlay and intersect with banking traditional data. The use of analytics has helped banks in many ways. First of all, analytics and non-parametric machine learning models helped in automating the underwriting process, thus saving sales and administrative costs. Secondly, analytics helped in predicting who will accept what offer and, in doing so, achieving higher revenues. Thirdly, it helped in combining several credit risk scores, whether produced from financial transaction or external credit bureau or behavioural dataset which gave more indication about a borrower's performance.

Nevertheless, machine learning models that are known for their ability to process unstructured data or high dimensions is challenged with over-fitting or the tendency to learn from a subjective dataset that is not necessarily frequent or common (Han et al., 2011) with lenders.

2.3.2.1. Decision Trees

Decision Tree (DT) model is a more sophisticated model than linear regression is. It allows for non-linear relations between predictive variables (Serrano-Cinca & Gutiérrez-Nieto, 2016). It is called trees because branches never join back together. It is made up of decisions (internal nodes), chances of an event happening (branches), and a pay-off at the end of each and every branch (external leaf). A decision tree is formed using a splitting attribute that can discriminate applicants as purely as possible according to the targeted classes. In this research, the target variable is binary (i.e. loan outcome) and, thus, the tree is called a 'binary tree' and measuring the best split, in this case, can be estimated using the "Gini-index" (Han et al., 2011)

Decision trees, for example, were found more effective, when used to classify P2P loan data, than

the traditional logistic regression and neural network due to their static nature and reliance on long historical dataset (Zhang et al., 2016). A decision tree illustrates pay-off events where decisions are associated with an expected monetary value (EMV) criterion based on these pay-offs (Lyn C Thomas, 2009). DTs use the divide and conquer strategy where the most promising attribute is used to split the tree (G. Wang, Ma, Huang, & Xu, 2012). It was explained by Lyn C Thomas (2009) that these types of models help the lender understands the sequence of decisions followed-by processes. A later development of this philosophy helped in the introduction of survival analyses which will be discussed in a subsequent section, where dynamic scoring is adopted.

Mathematically-speaking, decision variables where trees should be branched are those who demonstrate highest information gain for one of the classes (Baesens et al., 2003). Information gain reflects minimal entropy, where entropy can be calculated, manually, by applying the term $-P_1\ log_2\ (P_1) -\ P_0\ log_2\ (P_0)$ or, more generally - $\sum_{i=1}^{n} P_i\ log_2\ (P_i)$, on one of the attributes for n observations (F.-L. Chen & Li, 2010). In practice, the C4.5 algorithm has been used to run the decision tree model using machine learning (Baesens et al., 2003).

Decision trees have the advantage of neither requiring domain knowledge nor parameter settings and, thus, it is appropriate for exploratory analysis. In addition to that, they work on multi-dimensional data points and are intuitive (Han et al., 2011) as it can handle interactive effects among explanatory variables (Yeh & Lien, 2009). However, the drawbacks in decision trees are found in the split preference on variables that have many distinct values and classes (Baesens et al., 2003). For example, redundant attributes and noise can lead to over-fitting, unstable, and bad accuracy (G. Wang et al., 2012). Also, it requires standardisation which is not possible for string and nominal variables in general (Baesens et al., 2003). Finally, the problem of generalisation was brought up by Yeh and Lien (2009) leading to a possible over-fitting problem.

2.3.2.2. Neural Networks

Neural networks (NNs) are non-linear models that have pattern recognition classification capabilities (Malhotra & Malhotra, 2003). This modelling technique allows to form non-linear relationships between variables and applies a map network to derive a classification label (Baesens et al., 2003). The aforementioned relationships are represented by successive mathematical equations, thus machine learning, between input variables in one layer and output variables in a succeeding layer (Yeh & Lien, 2009). Neural networks (NNs) consist of layers that contain processing elements or neurons. NN algorithms train and learn iteratively about the relationship

between the neurons (Desai et al., 1997). The functions that link neurons of layers can vary from logistic to exponential functions (Crook et al., 2007) The iterative process stops if fixed number of iterations are reached, an error reaches a pre-specified minimum, or when neurons of a network reach a stable state and "learning effectively ceases" (Desai et al., 1997). Neural networks are known for their complexity in terms of high training times and computation power requirements. On the other hand, they allow the understanding of unstructured data such as audio, image, video while limitations with structured data (Ray, 2018). When credit risk model creators realised the shortcoming of linear regression, they tried using neural networks which allowed them to employ full non-linearity inherent in order to better-fit the data of the customers (McBurnett). They were, initially, introduced in credit scoring when used to train on 125 labelled credit applications with outcomes of: delinquent, charged-off, or paid-off borrower (Jensen, 1992).

The most commonly-used structure in credit scoring is the multi-layer perceptron NN (Baesens et al., 2003; Crook et al., 2007).The table below (see Table 6) lists five neural network structures that were discussed by West (2000) in terms of results and performances:

Neural Network Structure	Properties
Multilayer Perceptron (MLP)	The most used architecture. Global response of the neuron according to the weight matrix Outperformed logistic regression, decision trees and K-nearest neighbour. Not superior, mathematically, to discriminant analysis, but, clearly better when it comes to predicting distressed companies.
Radial Basis Function (RBF)	Successful in symmetric problems. Partitioning of the problem domain into one Gaussian unit.
Mixture-of-Experts (MOE)	Local response of the neuron to learn specific parts of the problem. Learning is localised within the neuron
Learning Vector Quantization (LVQ)	Nearest neighbour prototyping
Fuzzy Adaptive Resonance (FAR)	Dynamic patterns and prototypes based on the strength of the feedback resonance

Table 6: neural network structures and properties (West, 2000)

It was proven that neural networks classify bad loans more accurately than parametric and linear

methods due to the asymmetric distribution mentioned earlier in this paper (West, 2000). For example, Desai et al. (1997) reported a higher performance by NN than LDA in credit scoring (Baesens et al., 2003). More dominantly, when it comes to default predictions, it was mentioned by Huang et al. (2007) that NNs are more accurate, adaptive, and robust than other parametric and non-parametric credit risk models. This was concluded after NN performed the best followed by LDA, logistic regression, DT, and finally K-Nearest Neighbor (Huang et al., 2007). Also, NNs are "generalizable to other machine learning techniques" (McBurnett). The main criticism of NNs is their poor performance when incorporating irrelevant attributes and small datasets as in observations and dimensions (Ong, Huang, & Tzeng, 2005).

2.3.2.3. Naïve Bayes

The main principle of Naïve Bayes (NB) theory is the class conditional independence. In other words, NB assumes that the effect of an attribute on a class of the target is independent of the effect of other attributes (Yeh & Lien, 2009). Bayesian models are dynamic models that include individual behaviour over multiple time periods (Erdem & Keane, 1996). The Bayes classifier is built on the conditional probability of estimating the likelihood of the events of repayment or defaulting based on the series of prior events happening. A mathematical representation would be like P(Y=G/X=x). This term represents the conditional probability of being a good customer (denoted G) if the borrower is characterised with a set of predictors equal to x (i.e. a condition). This is also known as a 'prior' probability. In other words, the calculation is based on a sequence of events that are examined in a retrospective view of historical data (Lyn C Thomas, 2009). For example, in marital status, a predictor (M) with two possible values – 0: single and 1: married[3] would be used by the classifier if one of its classes, say single, gave a probability higher than a random walk (P(G/M=0)> 0.5) (James et al., 2013).

Bayesian rules, on the other hand, are applied in a 'posterior' outlook. In other words, they try to predict the probability that a characteristic (or a set of characteristics) had happened and resulted in an outcome. In PD models, either G or B are expected of a loan.

$$P(x/G) = P(x) \,.\, P(G/x) \,/\, P(G) \qquad (1)$$

Where:

x is a set of predictors/attributes such as divorced, owning a home, and aged between 45 and 55

[3] In reality, there are many more possible classes for this characteristic (variable) such as divorced, widowed, de facto, etc.

for example,

P(G/x) is the prior conditional probability and serving as a descriptive statistic derived from historical data. In other words, the number of borrowers ended-up being good after being identified as x holders, and

P(x/G) is the posterior probability analysing retrospectively and serving as a predictive statistic (how likely would a good borrower be a x holder).

Estimating the probabilities in Naïve Bayesian models happens by counting frequencies if the attributes are numerical discrete or ordinal features. On the other hand, normal or kernel density distributions can be used to estimate the probabilities of numerical continuous attributes (Baesens et al., 2003). Naïve Bayes classifier is commonly-used to assess different classes of an attribute and what information those add to the general odds expected by using the log information odds. This is known as weights of evidence (WoE) approach.

NB classifier is useful when it comes to providing a theoretical justification of the probabilities. However, it is criticised for its reliance on the class conditional independence assumption (Yeh & Lien, 2009).

The use of Bayesian analysis is explained later in this research (see Sub-section 4.4.5.2). Also, the findings of the analysis are reported and discussed (see Sub-section 5.2.4.2). Finally, the Bernoulli Naïve Bayesian model was used to evaluate the validity of the new dataset produced in this research along with other machine learning models in Section 5.3.

2.3.2.4. K-Nearest Neighbor

K-Nearest Neighbor (KNN) classifier approaches a dataset by randomly-picking observations from distinct classes then considering the closest k similar observations in terms of Euclidean distances as a similarity measure. Training happens by iteratively selecting different starting points until the distances are minimised. Additional advanced distance measures were, also, used KNN algorithms (Baesens et al., 2003). The main advantage of KNN is its simplicity (Malekipirbazari & Aksakalli, 2015) and non-requirement of establishing predictive model to run the classification algorithm. Nevertheless, KNN, like other advanced machine learning models, does not produce a simple classification formula. Also, its accuracy is highly affected by the cardinality of the neighbourhood of classes (Yeh & Lien, 2009).

2.3.2.5. Genetic Algorithms

Genetic Algorithms (GA) are a product of the genetic programming environment. In such an

environment, a tree-based structure dominates and comprises of function and terminal sets. The function set is where arithmetic, conditional, or Boolean terms are found. On the other hand, the terminal set contains inputs and constants (Ong et al., 2005), put simply X and y. GA represent set of algorithms that classify good and bad debts using machine learning on a training set of data followed by a validation stage of the classification rules or testing the algorithms on a hold-out or "out-of-sample data set" (Sousa et al., 2016). In summary, the process starts by generating possible solutions based on a population followed by new generations until the best solution is reached (Desai et al., 1997). Simply, GA models pass through four stages: creation of a population, evaluation, selection, and reproduction where the last three stages reiterate (Šušteršič, Mramor, & Zupan, 2009). GA models proved to be ideal for estimating behavioural characteristics. They are used in predicting exposure at default (EAD) rates from their means and dispersions (Tong et al., 2016)

2.3.2.6. Support Vector Machines

Support Vector Machine (SVM) separated the data into two regions in the p-dimensional space where p number of attributes are considered (Malekipirbazari & Aksakalli, 2015). classifier starts with a non-linear function $\varphi(\cdot)$ where each vector is supported by two hyperplanes. The function maps the observations to a high (possibly infinite) number of dimensions (features). The hyperplanes contribute to a clear discrimination between the targeted two classes (Baesens et al., 2003). Nevertheless, it was clarified by Malekipirbazari and Aksakalli (2015) that having a hyperplane with a considerable margin is rare and the common case happens when soft margin is used. This would result in a slack of some observations falling in the wrong region causing a slack in the performance. The iterative function contributes to optimising both slacks (minimisation) and margins (maximisation) to achieve the highest accuracy (Malekipirbazari & Aksakalli, 2015).

It was argued by Crook et al. (2007) that SVM is the most accurate credit scoring model despite its reliance on data quality that is noise-free. Unlike other machine learning non-parametric models, the main advantage of SVMs is that it can produce higher accuracy when dimensions are low and the number of attributes are limited (Huang et al., 2007). Additionally, researchers such as Y. Yang (2007) clarified that SVM classifier can learn from a moving window of changed environments. The aforementioned property was called 'kernel learning'. However, when deploying SVM algorithms, challenges arise in choosing the optimal data and setting the best kernel parameter (Huang et al., 2007).

2.3.2.7. Gradient Boosting

Extreme Gradient (XG) Boosting algorithm is an ensemble that minimises the error term. Each of its ensembles fits into the Pseudo residual of the previous tree's prediction to achieve minimisation. The XG boosting algorithm requires tuning the parameters when it comes to the number of iterations and maximum branch size used in the splitting rule (Brown & Mues, 2012).

In FICO, researchers adopted a modified version of the gradient boosting algorithms known as stochastic gradient boosting. The reason behind adopting such a modification is to be able to transparently-detect the major players when it comes to criteria (Fahner, 2018). The main advantages of XGBoosting were summarised in four features: (1) its generalized applicability on different loss functions, (2) its sequential structure which enhance its interpretation, (3) reduction of variance and bias of classification problems, and (4) its sensitivity to costs when it comes to accepting a bad borrower (Xia, Liu, & Liu, 2017)

2.3.2.8. Random Forests

Random forests (RF) is a technique that generates trees based on selecting attributes randomly (Breiman, 2001; Brown & Mues, 2012). Specifically, for the k^{th} tree, a random vector is generated and is independent of previous vectors – ones for 1^{st} through k-1^{th} tree. The vectors have the same distributions and each tree will create a classifier. After a large number of trees are generated, they vote for the most popular class (Breiman, 2001). This model requires setting two parameters – the number of trees and the number of attributes to grow on each tree (Brown & Mues, 2012). The main advantage of this model is its simple and fast implementation procedure (Kruppa et al., 2013).

2.3.3. Summary of Previous Results

The below table (see Table 7) summarises the accuracy of classification results of parametric and nonparametric models found in the literature. It is worth to note that results may differ based on sample size, bias, and dimensions of the attributes within the datasets.

	Discriminant Analysis		Logistic Regression	Decision / Classification Trees	Linear Programming	Neural Networks	Naïve Bayes	K-Nearest Neighbors	Support Vector Machines	Gradient Boosting	Random Forests	Genetic Algorithms
	Linear	Quadratic										
Srinivasan and Kim (1987)	0.875	-	0.893	0.932	0.861	-	-	-	-	-	-	-
Steenackers and Goovaerts (1989)	-	-	0.766	-	-	-	-	-	-	-	-	-
Boyle (1992)	0.775	-	0.775	0.750	0.747	-	-	-	-	-	-	-
Jensen (1992)	-	-	-	-	-	0.760	-	-	-	-	-	-
Henley (1995)	0.434	-	0.433	0.438	-	-	-	-	-	-	-	-
Yobas et al. (1997)	0.684	-	-	0.623	-	0.624	-	-	-	-	-	0.645
Desai et al. (1997)	0.665	-	0.673	0.673	-	0.664	-	-	-	-	-	-
West (2000)**	0.726	-	0.763	-	-	0.750	-	0.676	-	-	-	-
Lee et al. (2002)	0.714	-	0.735	-	-	0.737	-	-	-	-	-	-
Shi et al. (2002)	-	-	-	-	0.601	-	-	-	-	-	-	-
Baesens et al. (2003)	0.744	-	0.744	0.748	0.748	0.750	-	0.748	0.748	-	-	-
Malhotra and Malhotra (2003)	0.693	-	-	-	-	0.720	-	-	-	-	-	-
He et al. (2004)	-	-	-	-	0.879	-	-	-	-	-	-	-
Ong et al. (2005)	0.808	-	-	0.784	-	0.817	-	-	-	-	-	-
Huang et al. (2007)*	-	-	-	-	-	-	-	-	0.760	-	-	0.779
Y. Yang (2007)	-	-	-	-	-	-	-	-	0.727	-	-	-
Tsai and Wu (2008)**	-	-	-	-	-	0.790	-	-	-	-	-	-
Yeh and Lien (2009)	0.430		0.440	0.536	-	0.540	0.530	0.450	-	-	-	-
Bellotti and Crook (2009b)	0.781	-	0.779	-	-	-	-	0.756	0.783	-	-	-
Šušteršič et al. (2009)	-	-	0.761	-	-	0.793	-	-	-	-	-	-
F.-L. Chen and Li	0.761	-	-	0.737	-	-	-	-	0.754	-	-	-

(2010)*												
Brown and Mues (2012)***	0.756	0.630	0.634	0.619	-	0.721	-	0.618	0.829	0.721	0.762	-
G. Wang et al. (2012)	0.726	-	-	0.721	-	0.733	-	-	-	-	0.775	-
Kruppa et al. (2013)	-	-	0.748	-	-	-	-	0.685	-	-	0.959	-
Kou and Wu (2014)*	-	-	0.975	-	-	0.463	0.938	-	-	-	-	-
Malekipirbazari and Aksakalli (2015)	-	-	0.545	-	-	-	-	0.701	0.633	-	0.780	-
Fahner (2018)	-	-	-	-	-	0.895	-	-	-	0.899	-	-

Table 7: summary results of accuracy of credit risk models in literature

* based on German credit data (not Australian)
** based on German credit data (neither Australian nor Japanese) and considering MLP for NN
*** results from a behavioural dataset with 30% bad class representation

2.3.4. Dynamic Modelling

With the dramatic growth in consumer credit, the necessity of dynamic models has risen to come up with prompt decisions. Decisions made based on dynamic modelling included what interest rate to charge, whether to extend credit or not, by how much would credit be given, for how long should credit be given, when to collect from delinquent accounts, etc. (Yeh & Lien, 2009). Moreover, such decisions are reviewed periodically throughout the periods of loans by lenders. In other words, the future behaviour of a borrower no longer depends on one's characteristics at the time of the application. Instead, it reflects the recent performance. That is why PD takes a new dynamic dimension (Z. Wang, Jiang, Ding, Lyu, & Liu, 2018).

A dynamic model is a sequential learning process (Sousa et al., 2016) that uses non-parametric tests to analyse streamlined, continuous data (West, 2000). As a result, a dynamic model would capture a change in a borrower's address, job status. For example, factors such as average transaction value, number of cash withdrawals, amount of cash withdrawals, credit limit changes, rate of total jumps, proportion of months in arrears, repayment amount and outstanding balance are dynamic in nature and were used in predicting credit cards default using a survival model (Leow & Crook, 2016)

Experian has spotted that transactions could, continuously, be monitored to predict someone's proximity to default through dynamic assessment and evaluation of transactions and interactions with one's bank. Similarly, Equifax insisted on the use of traditional data, i.e. within the banking context. One of its spokespersons had explained that behavioural patterns could be extracted since someone's account would tell if someone holidays all the time or whether he supports a specific football team (Redrup, 2017).

Entrepreneurial Financial Lab (EFL) transformed default from its traditional definition of 90-days of arrears to a more dynamic concept to estimate the proximity to default. In doing so, entrepreneurial borrowers were classified into normal (up to 8 days of arrears), potential problems (9-30 days of arrears), poor payments (31-60 days of arrears), doubtful (61-120 days of arrears), and lost (exceeding 120 days of arrears). This can, clearly, illustrate how a customer's score would change continuously as the days of arrears accumulate (Arráiz et al., 2017).

In summary, the focus of recent credit risk modelling has shifted from merely predicting whether a borrower would default using classification algorithms to when the default is likely to happen using mixture surviving models (Z. Wang et al., 2018). In the following sub-section, Cox

Proportional Hazard and Markov Chains models will be discussed. Such models were discussed in the literature within the studies of Banasik et al. (1999) and Lyn C Thomas (2000). They were, also, applied in the works of Leow and Crook (2016), Ha and Krishnan (2012), and Z. Wang et al. (2018).

2.3.4.1. Markov Chains

Markov chains are one type of survival models that is based on probabilistic methods applied using data science techniques. It consists of set of transitions determined by probability distributions. Markov property is mainly characterised with memoryless or, in other words, the inability to learn from long historical trends. Instead, it produces an N x N transition matrix where N possible states are predicted with different probabilities based on preceding state where teach state's probabilities add up to 1 stochastically (Soni, 2018)

Proportional Cox Hazard model is an application of survival analysis/Markov chains. It estimates the time till default through a hazard ratio (β) which indicates the relationship of an attribute (x) with the probability of default (PD) at the time of the payment (t). Whenever the ratio is larger than 1 ($\beta > 1.0$), the attributes is positively-correlated with PD. Conversely, a smaller-than-1 ratio ($\beta < 1.0$) reflects a negative correlation with PD. Obviously, a small hazard ratio is more desirable (Lin et al., 2013).

Survival analysis was used by Bellotti and Crook (2009a) and Malik and Thomas (2010) on dynamic macroeconomic variables and by Leow and Crook (2016) on behavioural metrics derived from account activities to project cash inflows and, accordingly, instalments. This type of models had risen because of the need to account for any possible reduction in borrowers' income during the period of the loan. Also, it is highly essential to determine whether it is likely that eligible "non-discretionary" expenses are going to increase or not

This approach was adopted by Zielinski et al. (2013) in their focused crawling stage and it resulted in an 'adaptive crawling algorithm'. The sliding window gives credit scoring a dynamic nature as per the framework presented in the work of Sousa et al. (2016). In practice, Lodex introduced a similar perspective to survival analyses basing its predictive model on income predictors.

2.3.5. Credit Analytics

The concept of predictive modelling led the birth of credit analytics. Machine learning techniques are adopted to cluster customers with similar features whenever credit history is missing. The aforementioned technique is unlabelled. Modelling in credit has been subject to limited types of

transactional financial data. However, the introduction of attributes that are derived from behavioural finance has led the use of behavioural models by credit rating agencies (Nye, 2014). On the other hand, anomaly detection used in cyber security helps in finding outliers and anomalies and label those as risky customers and, thus, charge a higher price on lower credit limits. Labelling target class (also known as supervised machine learning) helps in building knowledge from the training dataset, then validating the knowledge on a small testing technique. The aforementioned ways help in predicting the future based on a set of inputs by using predictive models.

Nevertheless, Hsieh (2005) argued that, in addition to bearing enormous costs, having labelled data limits the number of observations incorporated for data mining and, accordingly, causes a sample that is unrepresentative of the population of borrowers. Therefore, unsupervised learning technique was recommended by Hsieh (2005) using a K-Mean Clustering model which would first separate the unlabelled borrowers based on similarities into clusters. Thereafter, finding the most suitable parametric or non-parametric model for each one of the clusters would, arguably, achieve a higher accuracy (Hsieh, 2005).

In looking for alternative data that utilises credit analytics and data mining, it has been stated by Dash et al. (2017) that leading banks rely on the frequency of shopping and amounts spent to estimate the ability to repay debt by consumers. This has been adopted by a major central American bank. The bank, also, uses geo-spatial analytics when to approximate the consumer shopping behaviour based on passing through shops, department stores, and outlets. In addition to location, some lenders analysed texts and photos that are submitted at the time of the application. For example, Lin et al. (2013) found that texts that have more words, short sentences, use more numerical figures, mention money or its synonyms, have more assertive and less tentative words seem to be predictive of loan outcomes. On the other hand, images used in support of loans were turned into variables that can identify either gender, race, or age. Finally, social networks can be used in credit scoring (Freedman & Jin, 2017; Lin et al., 2013; Wei et al., 2015; Weke & Ntwiga, 2016) as it will be explained in the next section of this chapter.

2.3.6. Depiction of Credit Models

The below diagram (see Figure 4) summarises the literature on both data types and models used in credit scoring. It depicts the current trend and competitive advantages that digital lending can offer.

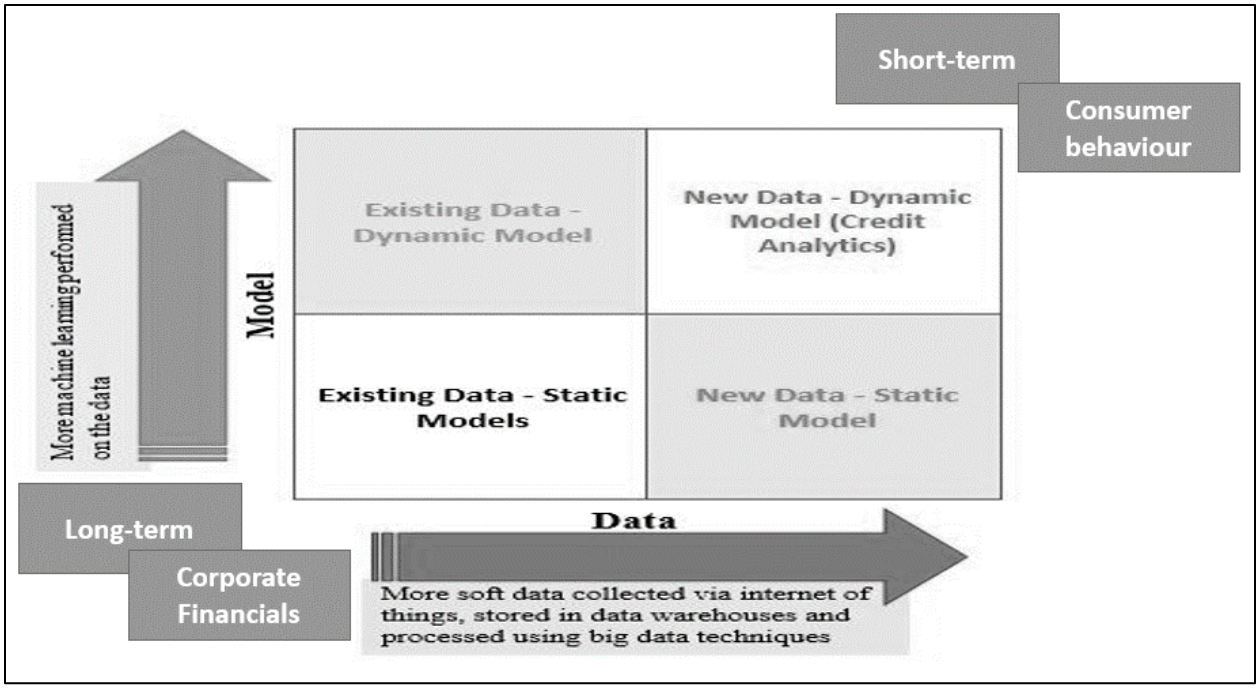

Figure 4: trends in credit scoring models and data used

2.4. Networks

Large scale real world complex networks have been widely explored during the last decade (Albert & Barabási, 2002; M. E. J. Newman, 2003; J. Yang & Leskovec, 2015). The term complex network refers to any large, dynamic, random graph that corresponds to a complex system, where the nodes of the network represent the individuals and the edges symbolize the relations between them (M. E. Newman, Strogatz, & Watts, 2001). Examples of real-world complex networks include World Wide Web (WWW), biological networks, communication networks, citation networks, social networks, semantic networks, etc.

2.4.1. Introduction to Social Networks

Social networks represent social interactions and friendships as part of network theory (Henegar et al., 2013). Recently, social media platforms e.g., Twitter, Facebook have reached major popularity by the involvement of large number of people and the exchange of information between them (Bachrach et al., 2012). Despite the differences in the interpretation of vertices and edges, complex networks display appreciable topological similarities and therefore it is important to study those topological properties that ensure the similarities. Community structure is an important topological property of complex networks and in recent years, detecting communities is of great importance in sociology, biology and computer science, where systems are often represented as

graphs (J. Yang & Leskovec, 2012). A community is defined as a subset of vertices that are densely connected in a relatively sparse neighbourhood (Chattopadhyay, Basu, Das, Ghosh, & Murthy, 2020).

In banking, social effects has been studied by many researchers. One of the challenges of implementing online banking was the lack of face-to-face interactions with bank staff (Pikkarainen, Pikkarainen, Karjaluoto, & Pahnila, 2004). In credit, it has been reported by HSBC (2018) that cross-checking the financial links with others related to the borrower is exercised. This is known as 'financial association'. In addition to the financial-related interactions, social influence by circles of friends and family ties was found to be affecting borrowers' financial behaviour (Wei et al., 2015). When explaining temporal discounting behaviour, Gärling et al. (2020) indicated that young adults borrow at an expensive rate due to peer influences through social media, one of the social network forms. It has been justified with homophily which translates to people acting similarly within a group (M. Newman, 2010; Wei et al., 2015). As a result, a relational aspect of the score has emerged (Lin et al., 2013) and the notion of social scoring has been followed and implemented by many lenders such as Lenddo (Wei et al., 2015). It is believed that social scoring would be of the biggest value to countries with developing economies due to the financial exclusion manifestation in those countries. Nevertheless, in developed economies social data can present additional data points and, thus, complement credit scores with another dimension to derive a faster decision (Redrup, 2017).

2.4.2. Structure of Social Networks

A network is a body of connected data that is evaluated using a graph. A graph is a visualisation technique that uses nodes (or vertices) and edges to represent a network (see Figure 5). A directed graph (diagraph) is a directional relationship between nodes where each edge has a direction as opposed to an undirected graph where bi-directional relationships exist between nodes (M. Newman, 2010).

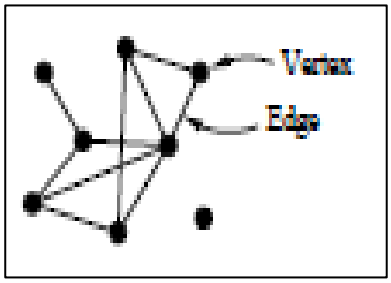

Figure 5: a small network consisting 10 edges and 8 nodes

Research on social networks was proven to be successful in psychology (Correa et al., 2010) sociology, health (J. Yang & Leskovec, 2012), human resources (Leonardi & Contractor, 2018), credit scoring (Wei et al., 2015), peer-to-peer lending (Freedman & Jin, 2017; Lin et al., 2013), and customer relationship management (Ascarza, Ebbes, Netzer, & Danielson, 2017).

Social networks carry three types of information: (1) the subjects represented by nodes, where, in our case, those are borrowers and new applicants on the professional social network, (2) the relationship between those individuals represented by the edges where borrowers decide to whether connect bi-latterly or follow others uni-latterly (directed network), and (3) the pattern of a network or a small part of a network like a node growth, for example, and how likely it will attach or detach to another node (M. Newman, 2010). In accordance with the homophily concept introduced by Wei et al. (2015), borrowers who are close to the centre of a network follow the same behaviour and are influenced by the most centralised node. This can be exhibited clearly in an egocentric network structure.

Networks have many statistics that may provide inference in credit scores such as: degree (number of edges), average degree, network order (number of nodes), betweenness, closeness, neighbours, degree, and centrality. Degree is defined by the number of edges, or in the case of credit scoring: friendship connections, which an individual has within a social network. A network's degree measures the level of connectedness within the network. Obviously, the average degree represents how social are the members of a network in general in comparison with another network. In a directed network, each node would have in-degree and out-degree where the former describes the number of edges that are directed to the node, in our case: LinkedIn incoming invitations, and the latter describes edges directed from the node toward other nodes, or outgoing invitations. Additionally, centrality refers to how a node or a vertex, representing an influential individual, that is powerful within a network and necessary for its cohesiveness (M. Newman, 2010). Usually, a successor node would represent a back-up to the central node (i.e. the influential individual) in a way that ensures the flow of influence in case of the removal of that central and influential individual; whereas, betweenness refers to how often an individual, represented by a node, lies between others in a network (Bradbury, 2011). Neighbours describe adjacency in the network's nodes.

In this research, the types and sizes of connections between the lenders represented by nodes in the social network will be evaluated based on credit outcomes.

2.4.3. Community Detection

Modules, motifs, and communities are other terminologies that refer to dense sub graphs. The issue of community discovery closely corresponds to the idea of data clustering in a system. When dealing with large and complex social networks, researchers proposed models that detect

communities based on how similar nodes are to their common neighbours (Ahn, Bagrow, & Lehmann, 2010; Ravasz, Somera, Mongru, Oltvai, & Barabási, 2002). Identifying communities in a network starts with partitioning a graph into set of disjoint subgraphs having similar properties within the graph (Coscia, Giannotti, & Pedreschi, 2011). Clustering algorithms partition a data set into several groups such that the data points in the same group are close to each other and the points across groups are far from each other. The task of community discovery is to segregate a network into groups of vertices having high density of edges within groups, and low density of edges between groups (Amelio & Pizzuti, 2014). A metric is required for such real-world network clustering to quantify the existence of a node in a particular community, which is known as node similarity.

2.4.4. Social Network Models

In the earlier studies, researchers have proposed different models for community discovery by using existing distance functions e.g., Jaccard distance, Hub Promoted Index etc. to find similarity between nodes (Ahn et al., 2010; Lü & Zhou, 2011; Ravasz et al., 2002; Zhou, Lü, & Zhang, 2009). Those models are built using algorithms such as the general stochastic model in telecommunication (Wan, Peng, Wang, & Yuan, 2016), community affiliation model, which introduced the idea of overlapping and nested communities (J. Yang & Leskovec, 2012) while the work of Coscia et al. (2011) focused on defining types of communities within networks then proposing the right technique to identify such communities. Based on the definition of a community, they proposed using either a clustering technique, relation summary network with Bregman divergence (RSN-BD), MRGC for multi-dimensional communities, or SocDim model which explores modularity then applies a discriminative classifier such as SVM.

2.4.5. Social Networks in Credit

Studying the structures and types of social networks can be of great value to lenders and borrowers. According to Manski (1993), endogenous social effects are transferred through a group of social ties. A later study by Wei et al. (2015) referred to his findings using the term ‘homophily’ by breaking its components down into social utility and posterior credit score utility.

Some researchers suggest that having friends is considered a signal of good quality in borrowers (Lin et al., 2013). Others, emphasised that attitudes of credit card holders depend on their feelings which can be informed by their social interactions and friendships. For example, social data has been useful when estimating creditworthiness (Lazarow, 2017) The term ‘financial socialisation’

was defined as “the process of acquiring and developing values, attitudes, standards, norms, knowledge, and behaviours that contributes to the financial viability and well-being of the individual” (Henegar et al., 2013). One main application of financial socialisation was the use of a mobile phone network where callers and receivers were connected by an edge. The edge carried a weight or a value representing the call duration (Óskarsdóttir et al., 2019)

In general, the internet has been a virtual place where people meet and expand their social circles (Correa et al., 2010). In practice, micro-lender Affirm Inc. considers the number of personal connections a borrower has and so does Lenddo which asks those connections to endorse the borrower and monitors how long it takes for them to do the endorsement (Rusli, 2013). Also, on utilising social media data, Lenddo checks the number of followers and collects information on those network connections (Rusli, 2013; Wei et al., 2015). Another example, is NeoFinance, a loan provider, which uses LinkedIn’s number and quality of connections of borrowers to estimate their career trajectories and success. It combines social network analysis with social data found on social media to derive job roles/seniority, length of employments as well as the industry and geographic location to come up with a credit score (Rusli, 2013).

When incorporating the networks dimension, the more social connections a borrower has, the better indication the score gives because of homophily. It reflects the notion of a person wanting to create a network with a similar type of friends. Therefore, network-based credit scoring would work best in collectivist cultures and in low-income countries where financial history does not reflect reality (Wei et al., 2015).

Apart from social media, telecommunication data has been used by many lenders from a social network point of view. Initially, incorporating data on phone bills from telecommunication service providers was limited to analysing payment patterns and top-ups. Nevertheless, other variables such as minutes per call, time of the day the mobile was used, and durations were the basis of forming reliable social networks (Brockett & Golden, 2007). Other credit scoring agencies went to analyse calls initiated as a proportion of all number of calls and analysed the network’s size and strength where the larger array of numbers called, the higher credit score First Access gives to its borrowers in Tanzania (McEvoy & Chakraborty, 2014). Individuals who have high in-degree or high out-degree with a social network are deemed influential since the former represents a case where likable individuals are approached whereas the latter represents someone who enjoys access to many societies and who has got many social skills (M. Newman, 2010).

In social networks, people tend to associate with others who share similar features. The aforementioned features can represent demographics such as gender, race, religion, nationality, or other discrete characteristics. On the other hand, other social features or factors can be scalar characteristics such as age, income, or education (Henegar et al., 2013). This is known as 'homophily' or 'assertive mix' (M. Newman, 2010).

The introduction of homophily by Wei et al. (2015) and Freedman and Jin (2017) highlighted that individuals form social ties with others for a social utility and a posterior credit utility. Their study was inspired by the work of Manski (1993) who found a model that explains communities formation based on individual as well as socio-economic factors. An influential person in a network has a high degree of centrality; whereas, a high betweenness of a node in a network ensures the flow of ideas and information (Bradbury, 2011). Lin et al. (2013) tested data provided by prosper.com and confirmed that having friends makes it easier for borrowers to get funded in P2P lending with lower APRs. They justified their findings with the term 'social stigma', which costs borrowers when their friends default. In that case, a borrower with no friends is most likely a risky borrower because no one wants to risk having a social stigma cost created by a friendship connection

There have been many efforts by P2P platforms to establish distinct groups of borrowers where group leaders apply pressure on their group members to pay on time. Meanwhile, lenders to different borrowers from different groups discuss their experiences within their own circle of lending individuals such as in the case of prosper.com (Freedman & Jin, 2017). However, some P2P platforms allow peer-scoring where group members tend to falsify lenders, occasionally, such as the case of Chinese platform, PDPai. Therefore, some researchers suggested describing social networks with new factors such as prestige, forum currency, membership score, contribution (Zhang et al., 2016), size, type, and composition of the social network (Freedman & Jin, 2017). On the other hand, other researchers created a hierarchy of friends with Lin et al. (2013) launching an effort to measure the strength of a social connection in 5 different levels. In addition to that, they went further by identifying types of networks into alumni, geographical, military, medical, demographic, hobbies, business and religion social networks on Prosper P2P platform.

2.4.6. Other Use Cases of Social Networks

2.4.6.1. Social Networks in Recruitment

Social networks, extracted from blogs, wikis, bookmarks, social media platforms, and media

sharing websites indicate one's character and work ethics and has been utilised by recruiters for the purpose of employee selection (Roth, Bobko, Van Iddekinge, & Thatcher, 2016). In research, Bachrach et al. (2012) used properties of social networks such as: density and size to define correlations with personalities of Facebook users. In practice, managers, through social networks analysis and interactions among their employees, were able to identify key individuals and potential silos within their companies (Leonardi & Contractor, 2018). Networks can be formed based on frequency of exchanged e-mails, phone-calls, social media interactions, and many other behavioural aspects (McEvoy & Chakraborty, 2014). As a result, managers can motivate, increase efficiency and innovation of their employees through identifying who interacts with others and who does not (Leonardi & Contractor, 2018).

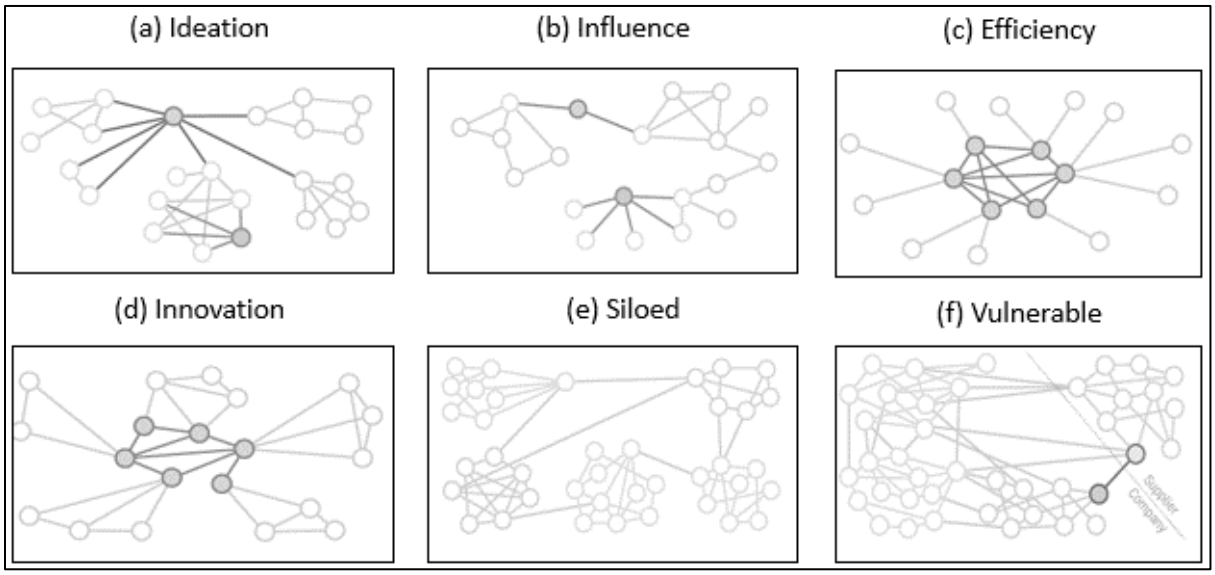

Figure 6: different structures of social networks and their behavioural properties

In the above figure (see Figure 6), a person exposed to different teams or departments can come up with innovative ideas. This is illustrated in graph (a). In graph (b), it is noted that a person that is connected to key individuals who are well-connected within their teams is a very influential person. A company's CEO may trust such a person to advocate for a new strategy. Graphs (c) and (b) demonstrate efficient and innovative teams respectively. The former illustrates great communication and relations between the team members; whereas, the latter consists of champions different departments. Graph (e) is an example of weak social network effects between and within communities. Finally, graph (f) indicates a possible disruption of work in business world or a defect of a group from a community in social terms.

2.4.6.2. Social Networks in Customer Relationship Management

It has been concluded by Ascarza et al. (2017) that companies that target influential and well-connected individuals within social networks with marketing campaigns gain a social multiplier. They estimated that the chances of the first-degree-connection individuals to buy the product is 28% despite not being targeted because of the influence of a targeted social tie (Ascarza et al., 2017)

2.4.6.3. Social Networks in Mobile Telecom Companies

Directed social networks can be formed from mobile inward/outward calling patterns. The behaviour of the nodes of a network can inform whether a node would defect from the network or not. This was used to predict churning in telecommunication companies (Óskarsdóttir et al., 2017)

2.4.6.4. Social Networks in Peer-to-Peer Websites

There have been efforts by peer-to-peer (P2P) lending platforms to establish distinct groups of borrowers where group leaders apply pressure on their group members to pay on time. Meanwhile, lenders shared their experiences in dealing with previous borrowers on some of those platforms (Freedman & Jin, 2017). Although groups in such platforms provide credit rating for their members, their ratings tend to falsify lenders in some occasions and, according to Zhang et al. (2016), few AA rated borrowers ended up defaulting in the Chinese PDPai P2P platform. Moreover, Freedman and Jin (2017) revealed that endorsements given by group leaders on Prosper.com to prospective borrowers within their groups are misleading and aim for the financial incentives sometimes. In other research, the term 'social stigma' was introduced by Lin et al. (2013) to highlight costs that borrowers suffer when their friends default.

Therefore, incorporating social network analysis has been suggested in credit scoring. New factors such as prestige, forum currency, membership score, contribution, and group would help in classifying borrowers more accurately based on their social networks (Zhang et al., 2016).

2.4.7. Systems in Social Network Analysis

A group of academic researchers founded the social media research foundation. They created an add-in[4] to Microsoft Excel called NodeXL. The tool does not require any coding or programming experience. Instead, it requires understanding of the basics of network structure (i.e. nodes, edges, and groups/communities) in addition to statistical formats of networks such as adjacency matrices,

[4] Optional menus and features added to an existing software.

edge lists, and other statistics to analyse and visualise a network. NodeXL focuses on creating social networks from social media data. In doing so, it accounts for privacy concerns and applies its collection to only individuals who have public accounts. The tool not only is capable of visualising a network graph, but also analysing one. Metrics such as centrality, successor, betweenness, clustering coefficient, and diameter can be extracted. Finally, an advanced version of the tool allows creating community detection or influencer detection algorithms that can be applied to other graphs ("The home of NodeXL,"). The figure below (see Figure 7) illustrates how NodeXL applies its graph visualisation on two types of networks: a mention and a retweet.

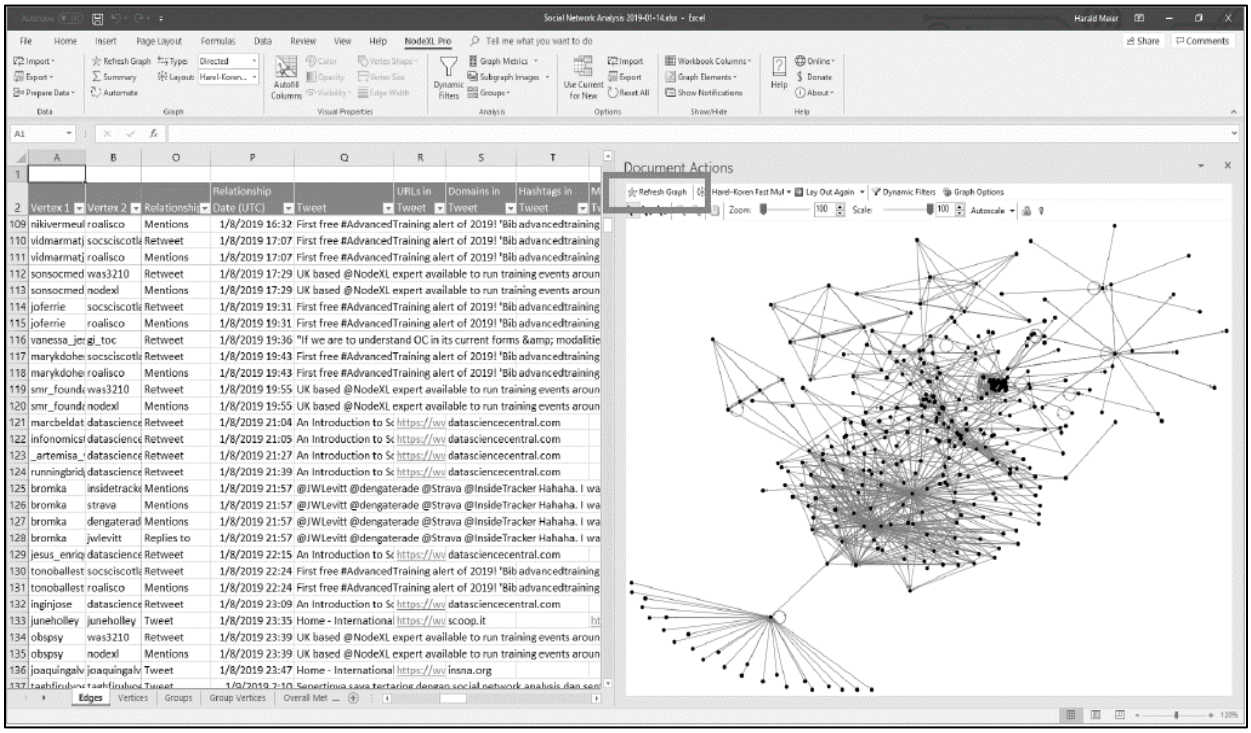

Figure 7: example of NodeXL visualisation function

In addition to NodeXL, VennMaker is another easy-to-use software that does not require programming knowledge and allows performing social network analysis. The only difference with VennMaker is its stand-alone design. Using VennMaker, a graph is built easily using drag and drop features where each node can be placed within one of the zones. Each zone has a different level of closeness to the ego. The closer a node is to the ego, the higher degree of centrality one has (see Figure 8). Also, edges and nodes can carry information such as type for the former (friend, colleague, family, neighbour, etc.) and demographic info for the latter (name, age, etc.). Additionally, statistics such as density or centrality can easily be retrieved from the dashboard. Similar to NodeXL, directed network can be differentiated from undirected network in

VennMaker. Finally, the software has been used for client-centred consulting.

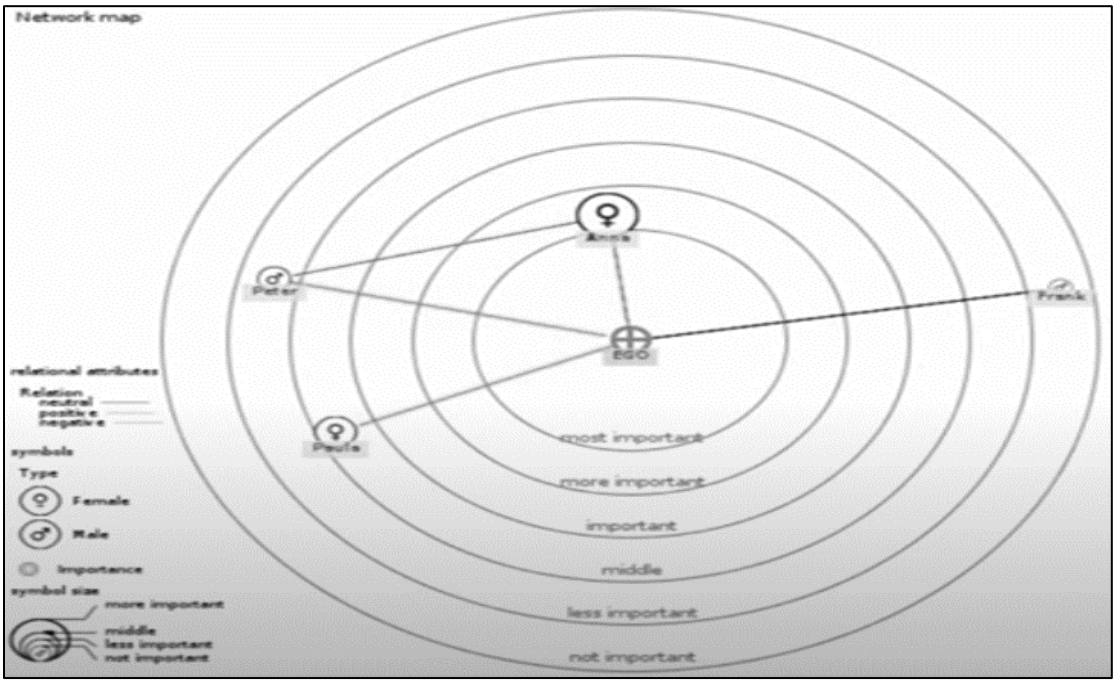

Figure 8: example of visualisation using VennMaker

In addition to the aforementioned two systems used for social network analysis, a web-based platform called 'FNA' is specialised in financial networks where types of networks represent financial transactions (credit and debit). Obviously, since each transaction has an originator and a recipient, FNA builds directed networks mainly. The online system has a great functionality of uploading datasets online. Also, it can visualise how a network changes over time (e.g. loan exposures between countries – see Figure 9). Such temporal networks show the trend in trade and credit between countries over time and can be used for future prediction.

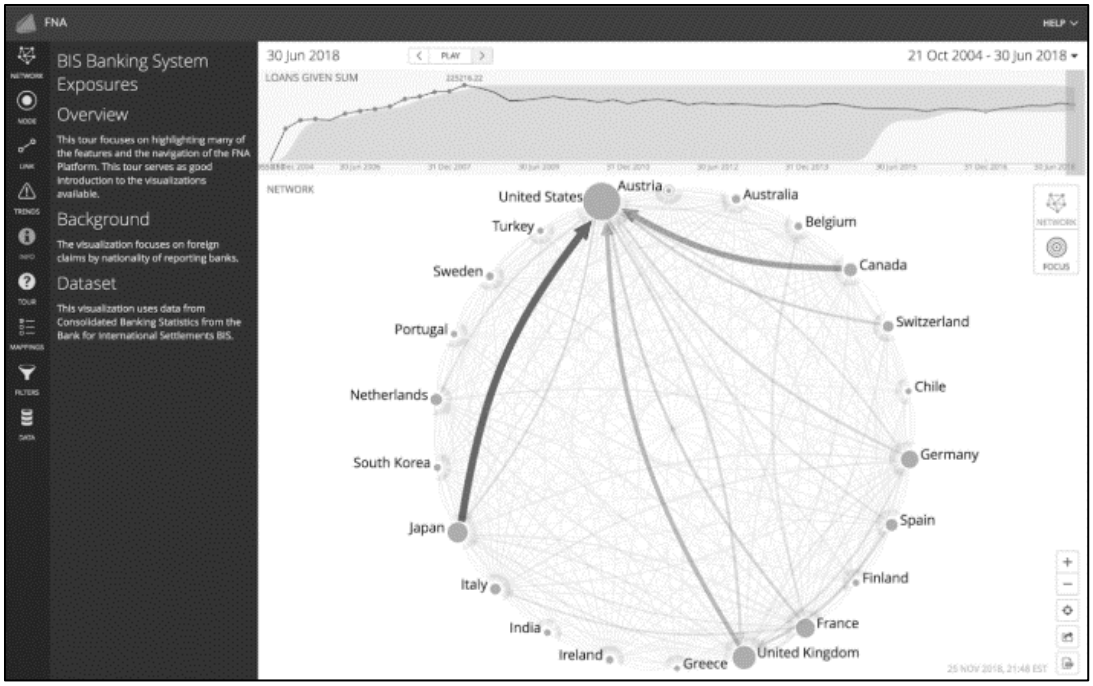

Figure 9: time-series visualisation of loan exposures between countries (fni.fi)

Additionally, FNA was used to visualise networks using a SWIFT Alliance dataset. The graph below (see Figure 10) visualise transactions between clients, whether consumers or corporate.

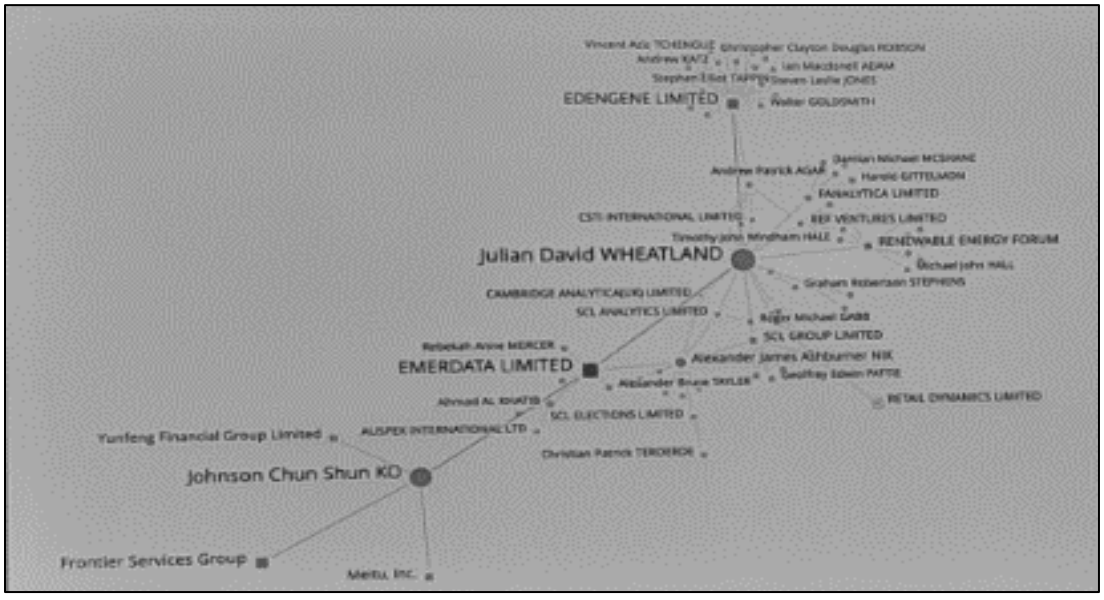

Figure 10: banking transfers between clients via SWIFT

Social networks can be formed from different sources. Bradbury (2011) listed the types of data that can be collected from LinkedIn (see Table 8). In the aforementioned table, suggested analyses were added to data types based on literature discussed earlier in this chapter. It is noted from the

below table that data listed in 17 and 18 indicate the structure of a network while 19 and 20 would indicate the type and direction of edges.

	Description	Data Type	Suggested Analysis
1.	First Name	Strings	ID matching
2.	Last Name	Strings	
3.	Current Title	Nominal / Ordinal	Cluster analysis / job hierarchy
4.	Past Title	Categorical	Educational level (Dr, Prof., or Mr./Ms.)
5.	Title	Categorical	
6.	Current Company	Hierarchy	Financial analysis
7.	Past Company	Nominal	Career growth (intra-industry versus inter-industries shifts)
8.	Company	Nominal	Industrial sector and company structure (government, public-private partnership, private listed, private equity, limited liability, sole proprietor, etc.)
9.	School	Nominal	Prospects and exposure
10.	Country	Nominal	Countries' credit risk by the big 3: Moody's, FITCH & S&P's
11.	Zip Code	Categorical	Neighbourhood risk (Cameo by TransUnion)
12.	Radius	Continuous	Geospatial analysis
13.	Industry	Categorical	Sector analysis
14.	Interested In	Categorical	Psychometrics using Singular Value Decomposition (SVD)
15.	Job Search	Integers	Job stability and satisfaction with life (SWL)
16.	Endorsements	Texts	Job satisfaction using sentiment analysis and text mining
17.	Joined LinkedIn	Date String	Network growth
18.	Number of Connections	Integer	Social network analysis
19.	Strength of network	Continuous	Social network analysis – centrality
20.	Number of mentions	Integer	Influence score
21.	Number of shares	Integer	Psychometrics – five factor OCEAN model
22.	Number of posts	Integer	Psychometrics – five factor OCEAN model

Table 8: LinkedIn data, types and proposed analyses

2.5. Summary

In this chapter, factors that influence the credit industry were, thoroughly, reviewed. In addition to that, structures, properties, capabilities, and applications of social network theory were highlighted. The organisation of this chapter was designed in a top-down view where literature started from regulations and reached to detailed variation from behavioural data, i.e. social

networks.

Regulations in the UK were recently updated through a consulting paper issued by the FCA in 2018. The updated procedures included behavioural data and allowed for adjustment in credit scores based on those. Also, lenders were required to justify their credit scores if challenged by borrowers. Therefore, they need to be cautious of the modelling technique and the accuracy of such behavioural data as well as the validity of the method of collection of such data to avoid biases. In addition to that, there was a clear trend in protecting borrowers' financial wellbeing by requiring lenders to not only to estimate probabilities of default, but also affordability scores.

In assessing creditworthiness, traditional data found in financial sector were not sufficient in economies where cash transactions exist abundantly. Also, the limitation was extended to new individuals that are joining the workforce continuously either through immigration, ease of recruitment of fresh graduates, or mothers returning to work after several years of being at home. Such traditional data was turned into dynamic nature and lenders started to run models that look at recent past.

Meanwhile, lenders who targeted those financially-excluded individuals aimed at different data sources such as mobile phones, social media, telecommunication, web browsing, and psychometrics to estimate borrowers' behaviour given different work environment or personality trait. At first, those were used for verification and matching purposes by credit bureaus. Later, such data was complementing traditional data that are missing such as estimating one's income by looking at geo-spatial data. In a more developed and independent manner, behavioural modelling became prominent and LendTech companies built their models solely on behavioural data.

As for models, the massive increase in data sources accompanied by computing capabilities facilitated the dependence on artificial intelligence or machine learning models. Such models can process faster and be more accurate than linear parametric models. However, explaining credit scores remains a challenge for those models. Also, subjectivity led to some models being biased against a group of people unintentionally. Very famous cases were discussed among in the industry of how considering last name attribute in a machine learning model would predict their scores based on racial backgrounds or how considering the length of credit card use where majority of credit card holders were male caused the algorithm to decide based on male behaviour. On the other hand, credit analytics is a concept that combines dynamic models with alternative data to come up with innovative credit scores.

Finally, network structures and theories had been reviewed and the mathematics behind their statistics were proven to be successful in many industries such as personnel management and peer-to-peer lending. It was then established from theoretical research that borrowers tend to have homophily whenever social utility exceeds the posterior credit utility that is driven by socio-economic motives according to Wei et al. (2015) who justified the former utility by homophily discussed in the works of Manski (1993).

2.6. Gap

This research acknowledges the problem of financial inclusion, highlighted by McEvoy and Chakraborty (2014), which leaves economies with almost half of the adults' population, globally, unbanked. On the other hand, assuming an inclusive policy is adopted by banks and financial institutions, the effects of information asymmetry, triggered by Yan et al. (2015), cannot be ignored. They highlighted the effects as moral hazard and adverse selection consequences.

In literature, previous research focused on improving the accuracy of classification of credit scoring models. Those, rightfully, argued that a slight improvement of such an accuracy would create massive savings on behalf of the lenders (Huang et al., 2007). The effects of such focus made lenders turn down opaque borrowers and, instead, extend limits and durations for those existing good cases (FCA, 2018). Therefore, another stream of research emerged and focused on investigating new sets of demographic and socio-economics variables that can be used for credit scoring purposes (L. Wang et al., 2011). Researchers who adopted this stream over-stated the argument of the model fitness by explaining the variation between credit scores. In addition to the aforementioned two main streams, there has been few research studies that focused solely on what values or categories are desirable within the set of variables adopted. This was an extension of the second stream that investigated types of data. The research study of Blumberg and Letterie (2008) represented this small stream of research within the consumer credit scoring discipline. This research is one of the fewest, if not the only one, to have combined the three streams where several machine learning models are tested in the evaluation process while six different combinations of data are used for testing using machine learning technique on the infamous logistic regression model. Finally, classes (or categories) within the social data are analysed using the classical posterior probability Naïve Bayesian method to determine the evidence that each one of those has. Therefore, in this research, the current trend of using analytics and data generated from Internet-of-Things (IoT) and other big data sources is scrutinised and discussed from financial,

psychological, behavioural, social, and technological perspective. Meanwhile, the suitability of current models was thoroughly discussed in light of the current advent of big data influence in finance and decision sciences. Therefore, this research aimed at using real data and investigating a specific effect of certain type of data – the social network attributes.

There has been few research studies in the area of the effects of social networks on credit scoring such as the works of Lin et al. (2013), X. Chen, Zhou, and Wan (2016), Óskarsdóttir et al. (2019) and Freedman and Jin (2017). In modelling, the former used a linear probability model; whereas, the lattest applied cox proportional hazard model. In practical application, all of the aforementioned three studies discussed the implications from a per-to-peer's point of view (prosper.com). When it came to data, Lin et al. (2013) created social network types of dummy data. Meanwhile, Óskarsdóttir et al. (2019) investigated the effects of incorporating call network data along with traditional data to predict credit scores.

Moreover, other studies targeted social network effects in credit using a statistical and mathematical approach without having an empirical evidence such as the case of Wei et al. (2015). In their study, homophily was represented in equations along with its 'hypothetical' effect on credit scores and probabilities of default in different scenarios. To our knowledge, social networks were only examined, empirically, by few authors from a lender's perspective such as the work of Óskarsdóttir et al. (2019). This marks this research as one of few of its kind to examine the effects of social network types and sizes from a lender's perspective on credit scores.

On another note, the research of Weke and Ntwiga (2016) reviewed models that can be used for estimating credit scores using social media data and argued the plausibility of such a strategy for a lender in theory. Meanwhile, Masyutin (2015) used actual real social media data and tested dynamic attributes. Those attributes were, mainly, behavioural and reflected on one's personality. Out of the aforementioned attributes, one referred to social networks indirectly – the pages subscribed to or 'number of subscriptions.'. A social network could have been formed of those subscribed to the same page as a network type. Both studies of Weke and Ntwiga (2016) and Masyutin (2015) failed to distinguish between social networks and social media when discussing to their findings. This research highlighted the differences clearly and explained how each of the two categories would influence the credit scores of borrowers during the assessment process.

In this research, the aim is to meet the needs of governments, banks and borrowers, who are, collectively, seeking the same objective (i.e. fair credit scoring to ensure easy-access to financial

services and business development) through reducing information asymmetry and increasing financial inclusion. Although some research has been ongoing on managing credit risk using big data introducing credit analytics, no study has provided a guideline based on empirical findings on what sources to consider and when would a specific source matter the most.

CHAPTER 3: PRACTICAL SYSTEMS AND TOOLS

It has been noted by researchers and financial data analysts that developing countries lack rich financial historical data (McEvoy & Chakraborty, 2014). However, individuals in those countries, usually, interact using smart phones creating digital exhausts (e-mails, SMS texts, whatsapp messages, iMessages, etc.). Also, they create friendships over social media networks and share their life experiences on those platforms. They surf the web and behave differently based on their interests and personalities. For example, web behaviour has been spotted and interpreted by Gonzalez and Loureiro (2014) on peer-to-peer lending websites.

Therefore, many companies decided to target those countries, with underdeveloped banking systems, as well as targeting individuals coming from those countries that are lacking financial and banking histories when arriving in a developed country. The aforementioned companies were, most of the time, small start-ups that took advantage of the open banking initiative proposed by regulators.

The following sub-sections (3.1 through 3.4) discuss different groups of financial institutions (i.e. peer-to-peer, micro-lenders, banks including digital banks known as 'LendTechs' as well as third-party assessors known as 'credit referencing agencies') that influence the volume of debt in a financial system in light of the data mining and machine learning technologies trending as pioneering business models with FinTechs and LendTechs.

3.1. Peer-to-Peer

Peer-to-peer (P2P) lending platforms are alternative channels that borrowers could reach up to in their attempts to get financed from lenders or investors without the need to go through a bank or a financial institution as intermediaries (Zhang et al., 2016). Put simply it connects lenders with borrowers directly without banking intermediation to save administrative costs (Z. Wang et al., 2018; Yan et al., 2015). P2P platforms were, first, introduced in 2005 and since then they oversaw a significant growth (Lin et al., 2013). In 2018, P2Ps facilitated loans worth of £ 3 bn. in the UK (Kumire, 2019).They are characterised with low transaction costs, less collateral requirement such as a guarantor. However, they represented a much higher risk on the lender's behalf and, as a result, a much higher interest rate (Zhang et al., 2016). In the UK, P2Ps are regulated by the FCA and are required to provide lenders with assessments of borrowers' creditworthiness. Some of the challenges to implementing an extensive assessment was justified that P2Ps are popular for their

fast processing and such a requirement will slow the application process remarkably (FCA, 2018). Also, there are other caveats to P2Ps such as no reserve requirements imposed on those platforms to meet lenders required returns in the case of bankruptcies. Therefore, P2Ps have been resorting to accessing open banking through application programming interfaces (APIs) and performing big data analytics in order to enhance the selection process (Kumire, 2019). The remaining part of this section will discuss real life examples of major global P2P FinTech companies.

Ali Finance is a subsidiary of the giant e-commerce firm – Alibaba, AliFinance collects e-commerce data from all Chinese internet shopping firms such as tmall.com, taobao.com, alibaba.com, and alipay.com to perform business intelligence and predict borrowers' credit risk score (Yan et al., 2015). It accepts future receipts as collateral and reported a default rate of below 1% (McEvoy & Chakraborty, 2014).

Lending Club is a subsidiary of Trans Unions, a credit rating agency, and is the largest peer-to-peer platform in the US. It targets financing SMEs to stimulate the economy and create more jobs. The company had piloted its SME financing project with Alibaba before offering it to public at affordable rates in an easy and flexible process (Alois, 2015).

Prosper is, arguably, the largest P2P platform for consumer lending in the US (Freedman & Jin, 2017; Lin et al., 2013). It is built on idea of creating groups of lenders and borrowers who interact with each other and endorse each other within a point-based system (Freedman & Jin, 2017).

Freedom finance is a P2P platform that has provided lenders with an access to real-time credit score check on borrowers to allow continuous assessment of their loan's repayment. In doing so, it has partnered with a world's top three credit bureau, Equifax, and a 3rd party assessor, AccountScore (Kumire, 2019).

Growth Street gets the financial history of a customer through open banking and forcasts financial strength and cash flow for the future. It has partnered with Starling digital bank to facilitate borrowers' monitoring of their credit facility in real-time (Kumire, 2019).

Zopa was established in 2005 in the UK, Zopa was the first ever P2P lending platform built on the idea of connecting lenders to borrowers directly through an online webpage (Yan et al., 2015). Zopa facilitates the financing of small loans from £ 1,000 to £ 25,000 over 1 – 5 years at a customised annual percentage rate (APR) through its simple-to-use calculator (see Figure 11). The loan is given for one of three purposes: car finance, home improvement or debt consolidation (zopa.com). Zopa does not require borrowers to upload their documents.

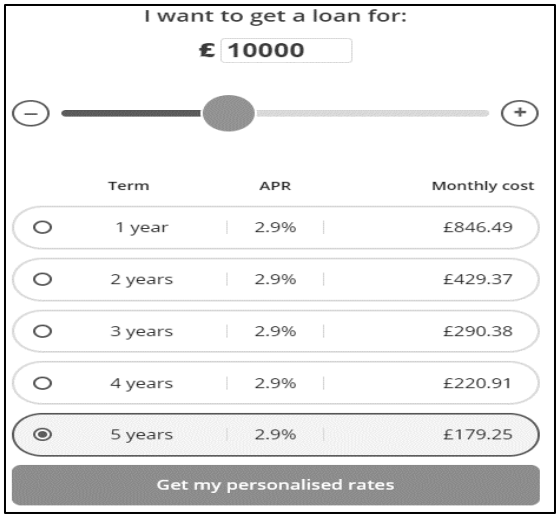

Figure 11: Zopa's loan calculator

Instead, it has created a verification tool, TrueLayer, that asks borrowers for permission to scrape their data from open banking database (Kumire, 2019). Customers can re-apply for new loans if their last application is at least 6-month old. Additionally, customers can consolidate an existing loan if one avails a better rate as long as total credit does not exceed £ 25,000. On the other hand, the company offers an 'innovative finance individual savings account (IFISA)' for those lenders who do not want to get involved in choosing borrowers and instead are happy for Zopa to invest their money in successful credit applications on their behalf and earn a tax-free interest on this investment as an incentive for lenders to save their money and facilitating investments in the economy. Finally, Zopa relies on a 3rd party credit assessor, ClearScore, in assessing its borrowers (zopa.com).

The main advantages of P2Ps are their flexibility, transparency, low costs and quick processing decisions. However, borrowers in P2Ps are usually those who were rejected by traditional lenders. Unless those are new entrants to the market or experiencing a major life event, those have poor financial health and, thus, can pose a high risk on the lenders. Additionally, since P2Ps charge low fees, they do minimum searches and do not evaluate many sources of data, which makes it easier for borrowers to misrepresent their behaviour or recent financial activities. Finally, high administrative and prosecuting fees make full amount unrecoverable (Pokorná & Sponer, 2016).

3.2. Non-banking Lenders

Lending was, first, offered by non-banking institutions when manufacturers in the 1850s such as Singer Sewing offered their products in a relaxed instalment terms to retail. Thereafter in the

early 1920s, the mass production of Henry Ford's automobiles witnessed the finance houses' inception (L. C. Thomas, 2009). With the revolution of big data analytics, non-banking lenders have emerged and become more innovative. Eventually, they were introduced as a type of financial technology firms or 'FinTechs' (Yan et al., 2015)

Affirm is a FinTech that developed its own algorithm in assessing individuals' credit risk through their phone numbers which leads to social media data and marketing information on their smart phones. The company has funded 1.5 million loans with US$ 1 billion, which accounted for 126% more people than the industry average as of December 2017. It has been reported that Affirm derived more than 70,000 features that can predict a credit outcome (Redrup, 2017).

The communication app company, WeChat, connects 600 million users across China. It provides loans through a service called Weilidai for up to $30,000 and relies, when making a credit decision along with setting its underwriting terms, on the information and content generated by the user on the app in addition to credit checks. The company has been famous for its quick decisions (Wade, Shan, & McTeague)

Market Invoice is backed by Barclays and Santander as major shareholders. It provides micro loans to settle invoices or start a business venture/line within a SME (Kumire, 2019). Omidyar developed an online tool that focuses on telecommunication data in terms of dialling and payment patterns. Specifically, cignifi mines minutes per call, time of the day an applicant's phone is used, call duration, whether calls and made or received, locations, SMS and data usage, and the frequency of top-up (McEvoy & Chakraborty, 2014). SoFi is a start-up lending firm in the US that gives loans to students using FICO and its own model that gives projections on a student's future earnings using one's university, course, employment likelihood and potential income versus earnings (Redrup, 2017).

Klarna provides shoppers with an instalment option. It is ranked as the most valuable fintech in Europe as of August 2019 (Kumire, 2019). Klarna provides credit to online shoppers from specific retailers with options to pay after delivery or pay in instalments (klarna.com). It assesses shoppers' credit scores by accessing data from more than 4,300 banks in 14 European countries through their XS2A application programming interface (API).

Those lenders have been using behavioural data mainly. They are innovative and been successful in utilising machine learning on large datasets. However, they face two main issues. First, borrowers are hesitant to allow them access to their day-to-day activities let alone the data on their

social circle. Second, they rely on consumers in developing countries where regulation on technological infrastructure is lacking and getting operating licenses are challenging.

3.3. Digital Banks

In open banking era where banks share their transactional data with trusted third-party providers (TPPs), digital banks offer exceptional experience through their tailored products, they are widely-criticised for not providing robust security when it comes to data privacy and compliance with regulators (Oakley, 2018). Starling is a challenger bank that has focused on using open banking to allow credit access to SMEs in partnership with a FinTech called SumUp. Starling provide a marketplace platform in collaboration with other financial services providers such as insurers and mortgage providers (Kumire, 2019). Revoult is a British FinTech that has more than 1 million users in the UK and 2.25 million users worldwide. It provides its clients with current accounts that work anywhere in Europe (Oakley, 2018)

3.4. Credit Referencing Agencies

In addition to in-house credit scoring within the lenders' premises, credit risk assessment has been done as part of the financial services sector. Particularly, when big data scalable algorithms were introduced, a lot of FinTechs touched upon the financial inclusion and information asymmetry dilemma. Cignifi was one of those companies who capitalised on mobile phone telecom data to help retailers, telecom operators, lenders, and insurers provide credit to those mobile phone users. The company uses AI technology to run behavioural models aiming at classifying customers into categories. The categories are, but not limited to, churning customers, best offer, pre-to-post-paid conversation, and customer lifetime value ("cignifi.com,").

Credit Referencing Agencies (CRAs) are considered by lenders as third-party assessors. Banks and financial institutions disclose information on account holders to the CRA they are registered with. In addition to that, borrowers who had their loan application rejected by a lender can inquire about the reasons. They have the right to contact the CRA directly and request a copy of their credit report, usually against a fee (HSBC, 2018).

Kabbage is based in the US and providing loans to SMEs in Canada, Mexico, and the UK. It has signed a partnerships with Amazon, UPS and Intuit to gather data about online shoppers and develop a credit scoring model based on sales, shipment, and customer feedback data (McEvoy & Chakraborty, 2014). The aforementioned model is a white-label built-in model and, in Europe,

Santander UK is able to assess a loan application within minutes with Kabbage's model. Simply, it matches customers' data with other sources such as social media platforms (Dash et al., 2017; Kumire, 2019).

The two credit scoring companies, Lenddo and Entrepreneurial Financial Labs, have announced a merger in October of 2017 by aligning their objectives to: approve more people, reduce default cases, and make real-time credit decisions. Their ultimate goal is to enable 1 billion unbanked/underbanked individuals and SMEs to gain access to credit (EFLGlobal, 2017; Fitzgerald, 2018). The company has three types of products: verification, insights and scoring ("Data-Driven Decisions for Financial Services," 2018). Below are brief descriptions on the services provided by the 2 companies prior to the merger.

Founded in 2006, the Entrepreneurial Financial Lab (EFL) was funded by Google to address information asymmetry problem with entrepreneurs and unbanked businesses. In 2010, EFLGlobal partnered with banks in Latin America such as Pichincha Bank in Ecuador (EFLGlobal, 2017). The funded project used a pilot on allowing credit to consumers by retailers and shops. Later and as of 2014, the company operated in more than 20 countries and its score was very successful in Kenya where top quartile scorers were 7 times less likely to default than bottom quartile ones (McEvoy & Chakraborty, 2014). When assessing credit risk, EFL's psychometric test was designed after collecting behavioural data on those who defaulted in the past as well as those who own a SME and earn high profit versus the ones who own low profits (Arráiz et al., 2017). LenddoEFL uses 12,000 variables in its modelling and produces a score in three minutes (Fitzgerald, 2018). The company used psychometric tests in developing countries, where financial files are thin or not existing, such as Peru to predict entrepreneurs' repayment patterns (Arráiz et al., 2017). In 2016, EFL agreed with FICO to extend financial inclusions to unbanked individuals from Russia, Turkey & Mexico (EFLGlobal).

Lenddo, on the other hand, estimated that there are more mobile users than individuals who are over 16 with bank accounts (4.8 billion mobiles versus 3.4 billion individuals respectively). Therefore, leveraging big data would be advantageous. Its model uses personality data (psychometrics), behavioural data, psychometrics, browser data, e-mail data, social network data, mobile data (see Figure 12). The aforementioned can be classified as alternative data that lenders have been collecting.

In addition to that, data provided by credit bureau on utilities and telecommunication data provided

by service providers in order for the bank to combine with their transaction data available within their systems. In other words, the company blends its collected data with the data provided by its third-party to innovate and increase predictive powers of its models.

In psychometrics, LenddoEFL based their scoring on five metrics namely: gratification, confidence, risk tolerance, conscientiousness, and honesty. As for the social network data, the standard deviation of message counts per day is considered as well as the percentage of message interactions with top e-mail provider out of all messages, the percentage of recipients who are also contacts, the median length of e-mail threads, intra thread median e-mail response time, and frequent contacts who have been contacted more than a threshold number of times. With regards to the mobile data, the company checks the browser history, the calendar entries, call log, contacts, installed apps, location hourly, phone model, and text messages.

Lenddo requires consents from borrowers to collect their data from social media accounts such as Facebook, Gmail, Twitter, LinkedIn, Yahoo and Microsoft Live in addition to data from mobile, telecommunication, e-mail correspondence (Redrup, 2017), psychometrics and behaviour (see Figure 12) in an attempt to better-assess consumers' credit risk by extracting attributes such as education, employment history and number of followers (Wei et al., 2015).

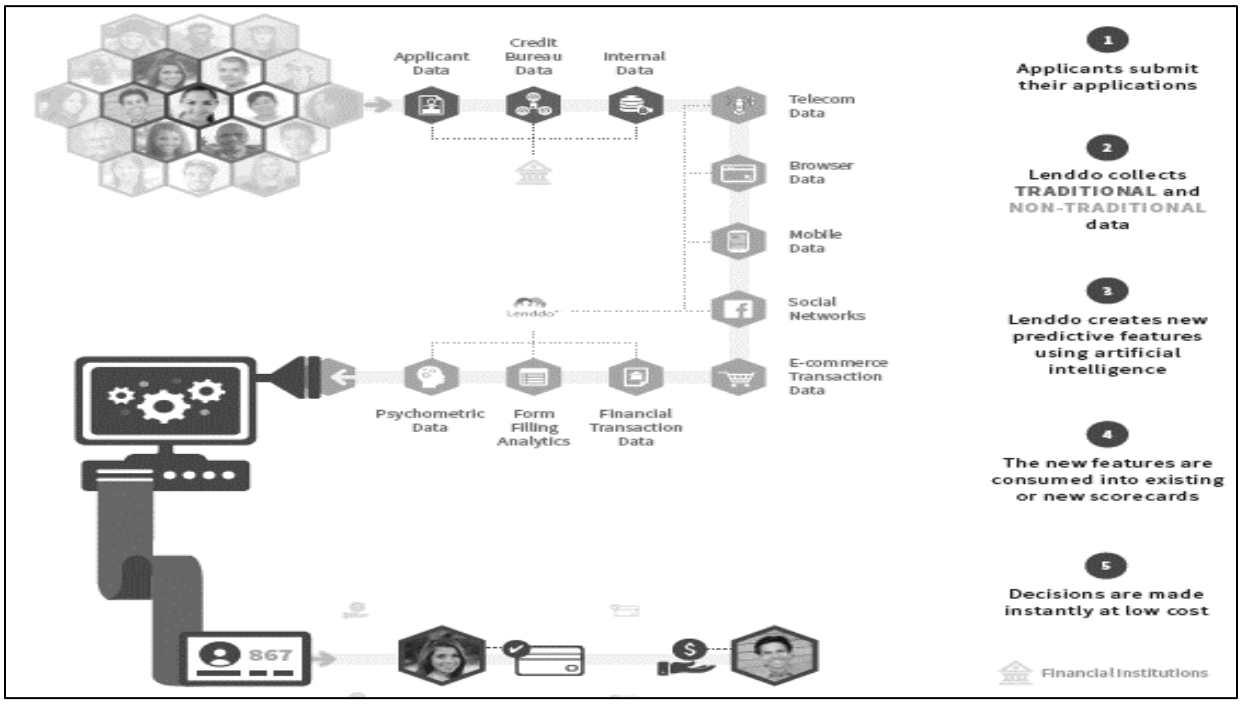

Figure 12: Lenddo's Business Model

Although Lenddo's focus is on the Asian/Pacific market and specifically The Philippines, it has

branched out to Latin America, such as Columbia, and Africa where more financial inclusion is required (Redrup, 2017). The CEO of Lenddo, stated that Facebook, Twitter, and LinkedIn made finance go back to the basics. Lenddo uses an algorithm that learns after every time an instalment is paid or surpassed without fulfilment and notifies group members if one of their connected peers failed in repayment so their scores drops (Rusli, 2013). Lenddo used social media data to predict credit scores. For example, they used length of time an active social account was held for to validate an online identity and, thus, increase the score. They also looked at how influential a potential borrower is by looking at the number of followers as well as the network a borrower has in terms of their Lenddo scores (McEvoy & Chakraborty, 2014).

Lenddo managed to secure a market place for depositers and borrowers in Australia in cooperation with Lodex (Redrup, 2017). Another example is ClearScore, which is a UK-based FinTech founded in 2015 that provides monthly credit score out of 700 to individuals. By entering your details such as who you bank with, your address within the UK for the last 3 years, your job and your residency type (owner, private tenant, council tenant, or lodging), your account will be created and matched with utility bills and other credit facilities you have under your name. Normally, information is pulled from Equifax agency report and explained.

DemystData is based in the U.S. It ran pilot studies in Canada and Mexico before they supported banks, microfinance, and insurance companies to assess their clients based on social data. Its model was successful in doubling inclusion to 25% for short-term loans in the U.S, U.K, Indonesia, and Thailand. Their model cross-checks online identities with documented identity of the applicant. It, also, seeks employment information in order to, eventually, estimate income from online sources (McEvoy & Chakraborty, 2014).

Lodex targets the unbanked students and refugees, who lack financial history in Australia. Those borrowers would have the behaviour of a credit-worthy person and Australian banks would use the data set issued by Lodex in their scoring for such cases. Unlike its parent company's, Lenndo, vision of increasing the degree of financial inclusion, Lodex aims at achieving faster decisions within Australia. Lodex produces a social score, a financial potential then provides borrowers with different sources of funding whether from peers or from banks in Australia. Lodex aims at estimating an income predictor in a way that is very similar to survival analysis

Kreditech is a German company that helps microfinancing firms in assessing unbanked customers to allow micro credit payments of EUR 150 on average. It emphasises on the location that an

applicant submits a loan application from through matching IP addresses with home and work locations. It, also, considers the time spent on filling out the application form (McEvoy & Chakraborty, 2014). The company uses creative web analytics techniques that can analyses online behaviour such as mouse movements (Yan et al., 2015). It operates in Poland, Spain, Russia, Czech Republic, and Canada (McEvoy & Chakraborty, 2014)

It was reported that, using psychometric tests, VisualDNA has helped some retail banks reducing 23% of default rates while, in other cases, succeeded in including more than 50% of applicants who had very limited financial history and records (McEvoy & Chakraborty, 2014).

Dongong uses a methodology that passes through eight stages: analysing environment for debt repayment, the ability for wealth creation, sources of debt repayment, ability for debt repayment, credit grading, review of grading, simulation test, and credit grades (en.dongong.com, 2016).

Although few CRAs have applied social network analysis in their credit scoring models such as Lenddo, those have not separated the behavioural modelling from the social modelling. In addition, those are unable to operate in developed economies where privacy laws are in place such as the GDPR

3.4.1. Credit Bureaus

Credit bureaus act as credit referencing agencies with wide access to data on individuals. Many countries rely on its credit bureau to collect information about the individuals living on their lands. In Latin America, the average credit bureau covers only 39.3 per cent of the adult population (Arráiz et al., 2017). Credit bureaus serve lenders who pay subscription fees and express their intent to supply their own data as part of a reciprocity agreement (CallCredit, 2008).

The “big three” agencies are Equifax, Experian and TransUnion. In addition to that, regulators have their own agencies that collect domestic data from governmental entities and utility providers. For example, a water supply company may share its users’ data with credit reference agencies shall those users fail to pay on time ("New Data Protection Regulation," 2018).

The drawbacks on credit bureaus in developing countries are the lengthy process to approve a change or an alteration in their modelling. In addition to that, the time required to collect and compile historical financial data in order to operationalise the traditional model. Also, it is a common behaviour between banks not to share fundamental information among their rivals. Finally, in the case of a borrower applying for the first time for a loan, one’s score will, merely, be a reflection of demographics (Arráiz et al., 2017). To the contrary, whenever a borrower applies

for a loan with any lender, a record will be kept with the credit bureau regardless of what the outcome of the application is (HSBC, 2018). Retaining the number of inquiries and how recent those inquiries were is known as the 'financial footprints' of a borrower. In this research, financial footprints were taken into consideration when conducting the quantitative testing in Chapter 5.

In addition to financial footprints, it is stated financial history and transactional data is important and may exhibit positive indicators such as good account standing as well as negative indicators when late payments, arrears and bankruptcies surface on the recent account activities (Arráiz et al., 2017)

Although, reports by credit bureaus had been fundamentally decisive in allowing access to credit, those have been 'patchy' in some countries where financial inclusion has been a challenge. Therefore, lenders complemented their scoring by using alternative data (McEvoy & Chakraborty, 2014)

3.4.1.1. TransUnion

Previously known as 'CallCredit', TransUnion is a registered credit referencing agency (CRA) that is one of the biggest 3 credit bureaus around the world. It collects public and financial data on individuals from all over the world. Also, TransUnion developed CAMEO which is a geo-demographic classification tool in the UK that classifies post codes based on many aspects such as income, unemployment, and welfare. TransUnion's application programming interface (API) allows its users, mainly lending institutions, to check borrowers' credit-related aspects. Those include public data such as: electoral rolls (current and historical), public information, search information, CAMEO geo-demographic data, gauge (a score that uses public data to rank order consumers according to their relative risk), address links that are indirectly related to a borrower, and alias links (other names that the borrower has been known by). Public information includes court judgements, decrees, sequestrations [5], trust deeds, individual voluntary arrangements (IVAs) [6], bankruptcy, and administrative orders such as debt relief orders (HSBC, 2018; TransUnion, 2019).

In addition to public data, the CRA gives access to financial data through its services such as MODA, SHARE information service, DataDNA, and CIFAS. MODA is a closed group of companies that give a summary of accounts and credit roll-overs for borrowers on a granular level.

[5] Seizing possession of owners' properties for the benefit of creditors.
[6] Arrangements to avoid bankruptcy.

Any extension and overdue payments can be checked by the service. SHARE information service provides financial data, default, delinquencies. DataDNA is a unique identifier of a borrower. It is used for matching, reconciliation, and identification. CIFAS is the UK's fraud prevention system that raises alerts whenever a transaction is considered suspicious or ingenuine (TransUnion, 2019).

3.4.1.2. Equifax

As a credit bureau, Equifax provides its credit scores and reports to its clients who wish to provide credit services to their clients. For example, assisting P2P websites such as in the case of Freedom Finance (Kumire, 2019) resembles how investors (i.e. lenders) would trust the platform since it is empowered by one of the big 3 credit bureaux. The comprehensive credit reporting (CCR) is Equifax's trademark where information on credit inquiries, payment defaults, arrears or infringements, dates of accounts opened and closed, credit limits, types of credit accounts, and a 24-month repayment history of the applicant is provided for lenders at fees ("What is Comprehensive Credit Scoring (CCR)?,").

3.4.1.3. Experian

Along with TransUnion and Equifax, Experian is the third member of the big 3 credit bureaus around the world. Operating in 40 countries, its reports are widely-adopted by many banks internationally (Redrup, 2017). Its credit score ranges from 0 to 999 (see Figure 13) and can be produced in a generic form online whether by lenders or by the individuals who wish to check their score based on common criteria and shared data among the other two reporting agencies. In addition to that, the score can be customised for a specific product for a specific lender where certain factors play an important role in that particular loan. Experian issues recommendations for borrowers to improve their scores such as registering for electoral vote at their addresses, closing unused accounts, reducing overall credit, building-up a longer credit history before applying for new credit and, obviously, avoid missing any payment. The innovative company complements its activities with B2B and B2C solutions where products to monitor clients of a business customer can be served as well as solutions to individuals to give them an overview of how they are spending and what is upcoming and probably an advice to avoid some kinds of risky transactions when instalments are due.

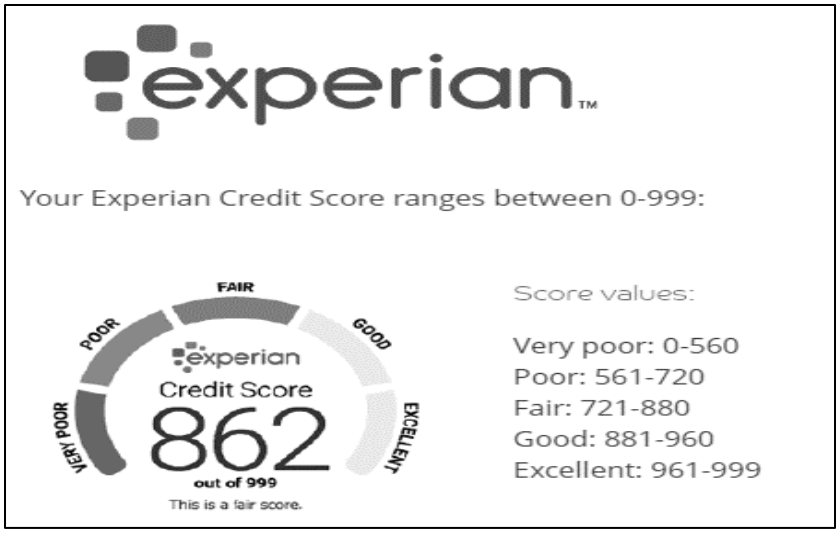

Figure 13: Experian Credit Score

The figure below (see Figure 14) depicts the dashboard of a system developed by one of the leading credit bureaus, Experian. This system is sold to banks in order to evaluate their clients' credit card behaviour.

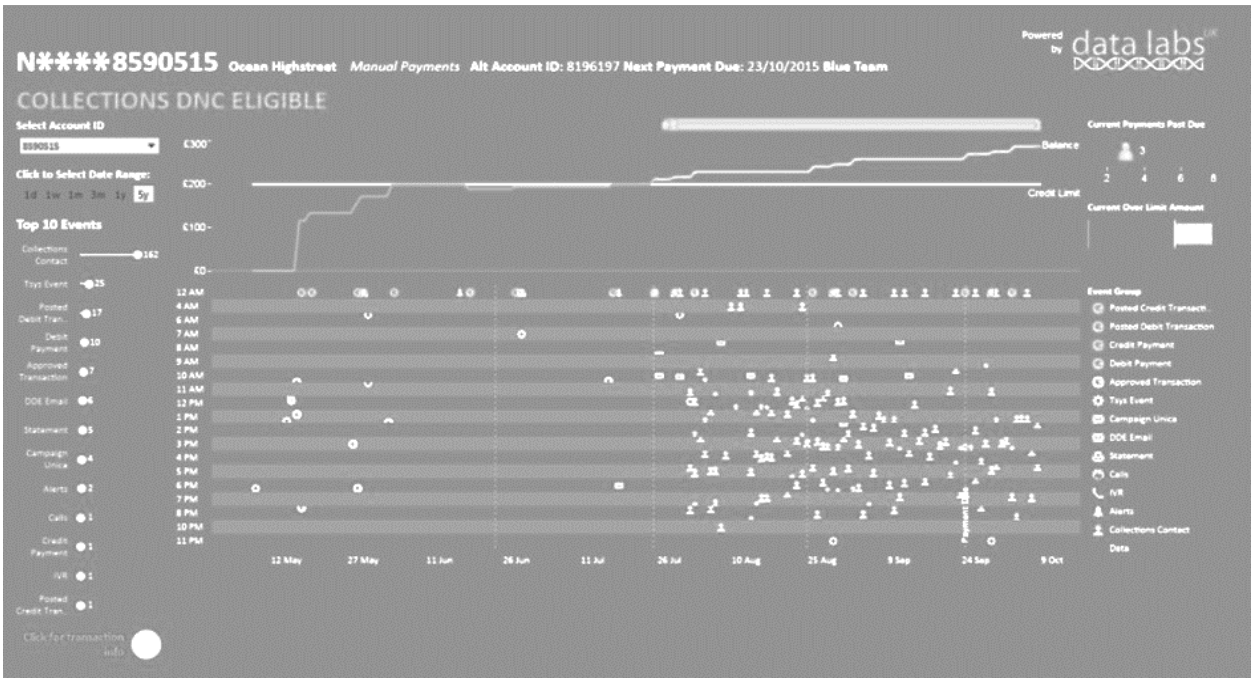

Figure 14: dashboard of the dynamic Experian data labs system

The aforementioned system allows banks and lenders, in general, to set their own criteria while it monitors such criteria and captures the number of times those criteria are met through events such as responding to credit campaigns, a credit payment, a debit payment, an alert sent (such as for

reaching maximum credit limit), or calls for collection. Each of the aforementioned events corresponds to a point-system or a score that reflects the severity of the borrower's financial health. Also, a visualised timeline of the event per day per hour is plotted with the distribution of events and their owners (the name of the team that responded to or initiated an event). Finally, the credit risk is quantified not only in values (i.e. amount exceeding credit limit), but also in frequencies (i.e. the number of times a client had days past due).

3.4.1.4. FICO

Bill Fair and Earl Isaac built their first credit risk model in 1958 (Y. Yang, 2007). The Fair Isaac Corporation (FICO) came up with a robust model that was adopted by many banks internationally. Although FICO components are traditional (see section 2.2.3), it has been known for its predictive power due to the use of dynamic statistical algorithms that adjust continuously (Redrup, 2017). In the below-illustrated figure (see Figure 15), FICO score is shown to be calculated based on five components. FICO's scores range between 300 and 850 where the higher the score is, the better (Hayes, 2019)

Figure 15: FICO score criteria

The components are explained below:

Payment History is weighted at 35%, the largest component, and it examines the applicant's performance on previous loans or commitments. For example, whether or not an applicant has been paying a utility bill to a provider on time or having arrears in the past. Not only the frequency, but also the significance of such an arrear or delay i.e. by how much was the applicant delinquent and for how many days. Additionally, the recency of such incidents i.e. how recent the delinquency was.

Clearly, payment history is a dynamic component as it changes with every payment due date. As mentioned, the older the arrear or delay in payment is, the less important and considered it becomes. Typically, an unpaid or late payment for a bill drops after 7 years. In the case of a bankruptcy, filing for such a case gets purged after 10 years.

Amounts owed, at 30%, is very similar to the traditional debt-to-burden ratio (DBR) where checks whether an applicant has been maxing out on limits with other revolving credit products such as credit cards and overdraft or the case of non-revolving loans such as mortgages and cash loans. For example, in the case of having an existing mortgage, customers who have already settled more than 30% of the property value will get a higher score than those who have just started paying back a similar mortgage value.

Length of credit history is weighted at 15% where the longer period an account is existing for, the longer records of previous payments it has and, thus, the better it is for the lender. It is calculated by taking the average number of months an applicant had credit facilities for.

New credit counts how many applications were made by an applicant, which is called in credit risk management 'footprints' taking into consideration the number of credit applications made with other lenders in the recent period (hour, day, week, month, quarter, year, etc.). Clearly, it is most effective when a lender is signed up with one or more credit bureaux who can provide such comprehensive insights.

Finally, **credit mix** refers to the credit products that an applicant, currently, owns, where the more variety of different types of credit products a borrower has, the better view of one's performance is demonstrated as opposed to one single type of credit such as having three credit cards and applying for a fourth, which would score very low on this particular component.

CHAPTER 4: METHODOLOGY AND DATA

In this chapter, the aim is to evaluate the plausibility of the claim that the use of social networks in credit scoring adds value to credit scores. The gap found in literature of lacking empirical evidence on social network effects for lenders is addressed in this chapter. This research acknowledges the need for bankers and professional working in the lending industry to reflect on social network effects on borrowers. Also, it accounts for the trending dynamic modelling concepts, machine learning techniques and models while capitalising on big data available in transactions found within open banking, IoT applications, forums, societies, web-based platforms, financial circles, alumni, and other unstructured data (texts, images, videos, etc.) exchanged in the web that can identify different types of social networks of borrowers.

4.1. Research Design and Framework

In this research, the design was built in a sequential principle where methods are products of their preceding. Specifically, the critical review of literature resulted in extracting concepts that are thought of as relevant to credit scoring. Therefore, such concepts were used in designing in-depth exploratory questionnaire (see Appendix 2) which will be discussed later in this chapter. The findings of this questionnaire sparked an interest in testing whether a relationship exists between social network variables and credit outcomes. Later, findings of this test were the basis for analyses, which in-turn prompted the model to explain the relationship between social network variables and the probability of default. Finally, the model was evaluated by state-of-the-art machine learning algorithms that are adopted by a major credit bureau. The diagram below (see Figure 16) illustrates the design of the research.

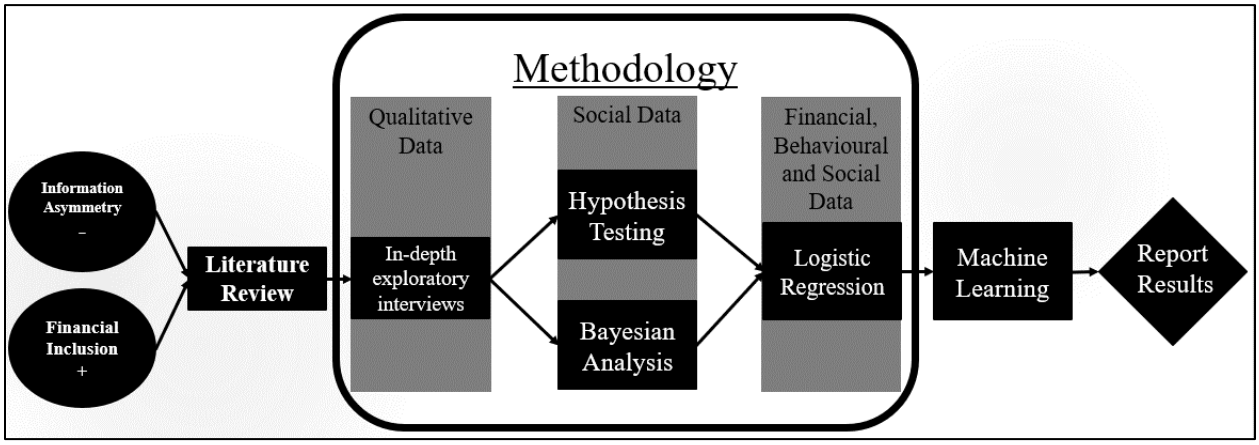

Figure 16: research design

Initially, a primary data collection was conducted using the qualitative in-depth interview method. The aforementioned interviews with industry professionals were completed in period between December 2017 and January 2019 in the city of Doha, State of Qatar. Insights from analysing the interviews were the starting point of a quantitative approach to the problem. The aim of the interviews is to answer the first research question on whether analysing social networks of borrowers helps determining credit scores more accurately.

Accordingly, four different quantitative methods were applied on a large dataset provided by a European lender. The first method, hypothesis testing. By achieving the aim of the hypothesis testing method, the second research question on the social network differences, if any, between the two groups is answered.

A relationship between two types of social networks and the probability of default (PD) was sought using a Bayes tree-based analysis. Part of the third research question was answered by highlighting a desirable size of social network for a good borrower.

Thereafter, a logistic regression model applied in a machine learning technique aimed at explaining the contribution of each significant variable to the variations in PD. The choice of logistic regression will be explained and justified later in this chapter in the model selection sub-section (see sub-section 4.4.5.3). In summary, an explanatory approach aimed at evaluating the degree to which two types of social networks affected PD and, accordingly, credit scores. By applying the aforementioned model, the remaining part of the third research question was answered and the distinction between types and sized was quantified.

Finally, adding social network data to behavioural and traditional data was evaluated by using contemporary machine learning models that are applied in a world-class credit bureau. This method illustrated how can social network data be added and, accordingly, provided an answer to the fourth research question.

The below figure illustrates both ideas presented in section 4.1 and this section (see Figure 17). Throughout this chapter, the methods will be explained and justified.

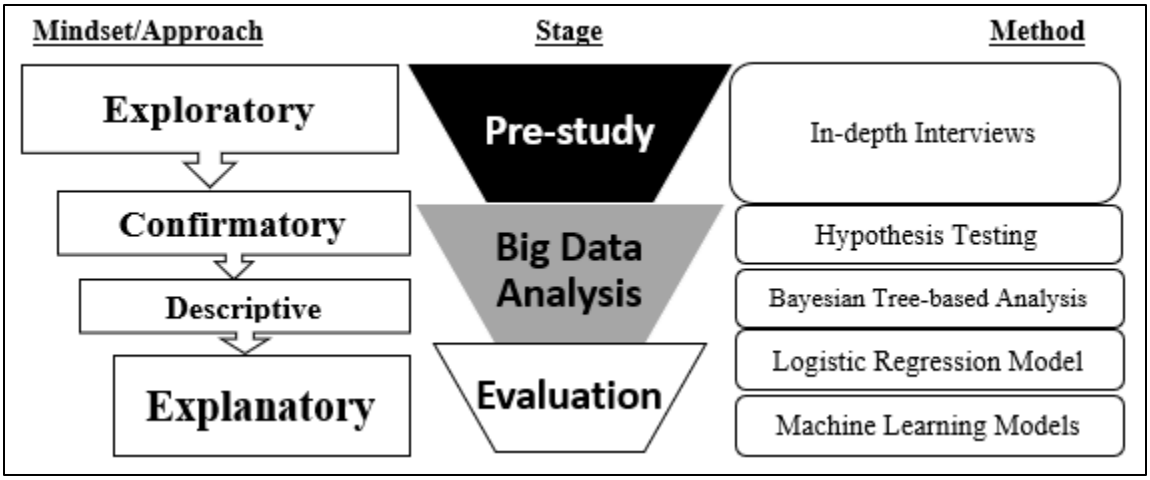

Figure 17: research framework

4.2. Research Philosophy

In this research, an inductive approach was initiated to explore what relational variables could be used to form borrower's social networks. The author collected subjective opinions from lenders who had been working on credit risk variables for a very long time. In order to objectify the aforementioned opinions, the author used a comprehensive dataset to provide confirmation of the relationships. The dataset included a variety of columns that contrasted by nature and it was deemed to be large. Therefore, literature on how to handle big data and pre-process such large datasets was applied. Machine learning techniques were used in an explainable logistic regression equation to demonstrate the effects of different variables. Finally, this research has interests in applying the findings to practice. Therefore, an evaluation using state-of-the-art machine learning models was performed. The stages that correspond to the aforementioned philosophies are discussed in the following section (see section 4.1). In summary, this research was initiated in an exploratory philosophy followed by a confirmatory then explanatory philosophies.

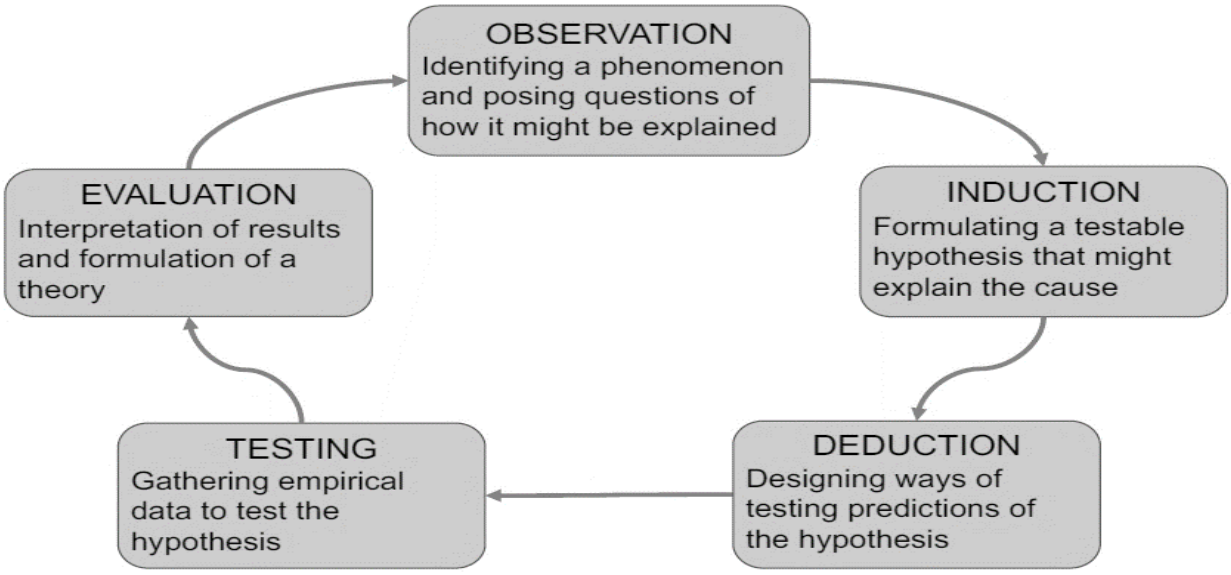

Figure 18: scientific research philosophy

Also, the research followed a scientific approach where observations of professionals were sought. Thereafter, an inductive technique introduced hypotheses that were to be tested. A deductive technique allowed selecting the appropriate test among other alternatives. Determining the right hypothesis entails testing the degree to which the hypothesized independent variable affects the independent. Finally, evaluating the performance of the new model is carried out using state-of-the-art machine learning models that are used in the industry. The above figure (see Figure 18) illustrates the scientific approach of this research.

4.3. Qualitative Method: Interviews

Pre-study exploratory in-depth interviews were arranged to collect primary data. A total of nine interviews took place in Doha, Qatar between December 2017 and January 2019. The interviews aimed at exploring the professionals' level of acceptance of using alternative data when assessing credit risk. More specifically, the idea of adopting social network analysis. This section will help answering the first research question on whether professionals deem social network analysis important to estimate credit scores more accurately.

4.3.1. Sample Selection

In this sub-section, the link between the pre-study interviews that were conducted in Qatar and the big data analysis collected from a European lender is demystified. Specifically, relying on the findings of pre-study interviews in Qatar to justify the selection of data types and adopting big data analyses on a dataset provided by a European lender can be challenged. The aforementioned

challenge can be addressed with the below justifications.

The selection of Qatar in the pre-study was motivated by the social network requirement from a practical and a theoretical point of views. According to Redrup (2017), social credit scoring is most effective when technological advancements are existing and accessible by borrowers. Also, it is complemented by a collectivist society / community structure, where people value interactions with each other and follow each other's leads. Usually, such a structure can be found in developing economies.

In fact, Qatar's recent adoption of technology has been remarkable with its national information and communication technology (ICT) plan highlighting initiatives such as a national broadband network, free internet in public spaces, and the international connectivity through satellites. In thriving for a knowledge-based economy, Qatar has pledged to grant equal access to technology in an attempt to bridge the digital divide and achieve digital inclusion (ICTQatar, 2015).

Meanwhile, Qatar's financial market and economy remains small and it is considered as a developing nation. The advantage of such as a small economy was the ability to gauge the responsiveness to adoption of social network analysis technology in credit scoring by main players of the economy. Those were the financial institutions, banking regulators, and national technology organisations.

Although borrowers were not interviewed, statistics showed that 76% (73% confirmed using their smart phones) of the population is active in social networking activities including posting and interacting on social media platforms, blogs, forums, and instant messaging. As of 2014, 71% of the population owned a smart phone (ICTQatar, 2014). This would create a great case for lenders who are trying to penetrate a small market that had recently started its own national credit bureau. Social networks can be formed using many sources that were discussed in chapter 2 in this research. Therefore, choosing Qatar satisfied the criteria of the aims and objectives of this research.

4.3.1.1. Background

The Qatari market is a developing market, where no advanced technological infra-structure is in place to gather real-time alternative data. The size of the market is relatively small where individuals can be identified based on reputation or family businesses as well as personal interactions with banks' senior managers and boards. Nevertheless, the high exposure of the economy of Qatar due to its fiscal policy made it important for banks to finance those who came into the country with a lot of planned projects and business ideas. A key focus of the interviews

was to understand the role of behavioural and social data and validate the views by testing a model using those types.

On the infrastructural technology part, the national big data research institute has plans to integrate a national database. In such a case, predictive models can have more data points. As a result, financial sector will be more protected against frauds or defaults.

In Qatar's retail banking sector, there are two types of individuals or borrowers: borrowers who receive salaries from their employers on a monthly basis (known as 'non-secured' applicants) and borrowers who demonstrate other sources of income whether holding equity, running a business, owning properties, etc. (i.e. 'secured' applicants).

The regulator in Qatar for local banks sets the criteria for non-secured lending with huge emphasis on capacity represented by salary. There is a maximum designated Debt-to-Burden ratio (DBR) and a maximum tenure for citizens and residents of the country. In addition to that, there is a maximum margin on the common interest rate set by the central bank.

As for the secured lending, banks have total flexibility to charge the borrower any interest rate (profit rate in Islamic banking) based on one's risk exposure. Also, the structure of the loan whether its long-term or short-term is decided solely at the bank's discretion based on the purpose of the loan. For example, financing an urgent need goes as short-term; whereas, financing a real-estate project usually is considered as long-term. Effectively, banks go to great lengths in interviewing each potential borrower and ask for guarantees as well as evaluating one's properties, reputation, and other sources of income to secure the loan given.

4.3.1.2. Other Considerations

There are two other notable remarks in Islamic banking on the procedures of underwriting a loan: first, the purpose of the loan has to be sharia-compliant. Also, procedures in granting a loan within Islamic banking entail full ownership of mortgaged house by the bank and a contracted re-sale of the same over the period of the loan. As soon as the mortgage is settled, ownership moves to the borrower.

In terms of criteria, the bank, first, qualifies the loan in terms of eligibility as a Sharia-compliant loan. In other words, the purpose of the loan needs to be in line of Islamic practices. For instance, a loan to build a brewery that manufactures beers would be disallowed by Islamic bank no matter

how profitable the loan would be since consuming alcoholic beverages is against Islamic practices and is prohibited by Sharia law. In the next chapter, the systems and weighting criteria of the three banks that hosted the interviews will be highlighted.

4.3.2. Data Collection

In this research, interviews were semi-structured and had an in-depth and open-ended settings. Those were similar to the settings found in literature when bankers were questioned about the implementation of a technological aspect, online banking, in Thai banks (Rotchanakitumnuai & Speece, 2003). An interview guide was used to steer the discussions and a questionnaire was used to collect certain more specific information on key points (see Appendix 1). The interviews were conducted with individuals from both financial institutions and regulatory organisations. In both cases, the selection of interviewees was based on seniority level where senior managers and decision-makers, who work in the credit scoring area, were selected. The decision to select senior professionals in banking was inspired by the work of Rotchanakitumnuai and Speece (2003), who selected managers from the Thai banking sector. They justified such a decision with the nature of a developed economy in Thailand as well as the need for an in-depth and open-ended discussions since the economy is not well-developed yet. As highlighted earlier, the interviews were conducted in Qatar, which has a similar economy to that described in the aforementioned study. Interviewed bankers along with their employers requested that their names be anonymised as well as their employers' identities.

In terms of banking professionals, the author conducted four interviews with bankers who worked in Islamic banking. Also, two bankers who worked for a commercial bank were interviewed as well as one banker from a state-owned bank. As mentioned earlier, the author interviewed two professionals who worked for regulatory organisations. The first individual worked for a national credit referencing agency (CRA) whereas the second individual worked in a national big data research institute to implement a nation-wide strategy for data pooling and warehousing. A table highlights the anonymised interviewees, their positions, and their anonymised employers can be found below (see Table 9). The type of bank was kept in order to argue the inclusivity of all views on the importance of alternative credit scoring.

Interviewee	Job Title	Employer
Banker 1	Head of Credit and Market Risk	Islamic Bank

Banker 2	Chief Risk Officer	Islamic Bank
Banker 3	Operations Manager	Islamic Bank
Banker 4	Head of Retail Banking	Islamic Bank
Banker 5	Senior Credit Manager	Commercial Bank
Banker 6	Retail Credit Manager	Commercial Bank
Banker 7	Head of Portfolio Management	State-owned Bank
Regulator 1	Data Quality Manager	National Credit Referencing Agency
Regulator 2	Principal Scientist	National Big Data Research Institute

Table 9: interviewees list

A semi-structured interview questionnaire was designed (see Appendix 1). The questionnaire consisted of ten open-ended questions that enabled the possibility to follow-up. Also, it presented a final question in a grid design where interviewee rate on a scale from 1- 5 four different social and behavioural alternative features proposed for credit scoring. Interviews were focusing on four integral parts: (1) the current credit scoring tools of individual borrowers, (2) the degree of innovation and sophistication in data collected, (3) the role external parties play in the aforementioned process, and (4) the evaluation of proposed social and behavioural metrics and whether they are worth of introducing to the credit risk methodologies. The following sub-sections describe how the banking system in Qatar approaches credit scoring functions and what role does technology play in that. Thereafter, a discussion on how behavioural and social networks are viewed in the credit risk context will be provided. The results of the qualitative analysis of this primary data is discussed in the findings and discussion chapter (see chapter 5). The primary data assisted in answering the first research question whether social network analysis is important in credit risk scoring. In light of the above, this research proceeded to test the plausibility of the aforementioned claims.

4.4. Quantitative Testing for Credit Score Modelling

After confirming the importance of social data with industry professionals in credit scoring, a dataset was collected from a European lender on loans given to consumers in the South East Asian region. The data not only incorporated financial aspects, but also behavioural and social aspects. When it came to behavioural data, attributes like owning a car, proximity of residence to workplace, or who accompanied the applicant to the bank were examples of how those behave. On the other hand, social aspects were focusing on any ties that an applicant has with financially-

troubled peers. There were two main types of those peers – delinquent and defaulting peers. Despite the bank agreeing to share the dataset, it did not specify the mechanism of how those social variables were collected.

Three quantitative methods were adopted to examine a large dataset that contained financial, behavioural, and social attributes. The aim of this stage is to prove, initially, that borrowers who repay their loans have a statistically-different social network size to those who do not repay their loans and default. The aforementioned would help in answering research question (2) on the differences of networks between the two groups.

In addition to that, the Bayesian tree-based analysis provided information what social network sizes are deemed desirable in a borrower and answered research question (3) partly. Then, the logistic regression answered the remaining part and introduced coefficients for the social network data.

Finally, evaluation of the performance was run through state-of-the-art machine learning models that are widely-adopted in the industry to answer the last research question (4).

4.4.1. Credit Scoring Model

The idea of examining performance based on different subsets of the data was found in the literature. For example, Šušteršič et al. (2009) tested the performance of logistic regression model and neural networks in consumer credit scoring based on three versions of the data for the same borrowers. The three subsets had 21 variables in one case, 21 in the second case, and 18 in the third version (Šušteršič et al., 2009). Similarly, in this research the logistic regression model will evaluate borrowers by examining different natures of their characteristics. Specifically, the model will run on financial data, behavioural data, and social data separately. Therefore, financial modelling, behavioural modelling, and social modelling will be reported. It will, thereafter, run on combinations of those to conclude on the added performance (predictability and interoperability) of social network data on both the financial and behavioural data. The results will be discussed in the next chapter (see chapter 5).

Financial data refers to the financial health of an applicant measured by one's assets and liabilities; whereas, behavioural data reflects choices made by the applicant due to heuristics (pre-determined rules in mind) and cognitive biases (personal tastes and perceptions) as per Hirshleifer (2015) and Taffler (2017). Finally, social data has a relational aspect (Lin et al., 2013) where the borrower is affected and influenced by the connections with other friends within a network (Wei et al., 2015).

Also, it was concluded by Lin et al. (2013) friends with different roles and identities have different influence on credit outcomes.

In addition to classifying the dataset into three different aspects, the dataset had attributes that are cross-sectional while maintaining few variables that can vary over time (i.e. time-series). Finally, based on the credit risk scoring models discussed in the literature, the variables were also put into two groups according to their suitability for either static or dynamic modelling. The table below (see Table 10) highlights examples of the data and their classifications along with proportionate existence (in percentages) within the dataset.

	Cross-sectional	**Time-series**	**Financial**	**Social**	**Behavioural**	**Static**	**Dynamic**
Characterisation	Multi-dimensional at a single point of time.	Trended over time and can be forecasted.	Financial records and banking information.	Relational and driven by groups and influencers.	Personalised actions driven by heuristics and biases.	Do not change throughout the life of the loan.	High variability requires continuous update.
Support in Percentage	95%	5%	53%	2%	45%	42%	58%
Example	Purpose of the loan	Instalment payments	Average account balances in the last 6 months	Number of friends who defaulted on loans	Car's ownership	Credit amount applied for	Address of the borrower

Table 10: data views and classifications

4.4.2. Data Selection

A dataset by a European lender on loans given in the South East Asian markets was acquired through an online repository. The dataset had 307,511 loan applications described in 120 attributes representing the columns. In the remaining sections, sub-sections, and paragraphs, the terms attributes, columns, and dimensions will be used interchangeably while referring to the same aspect of the dataset. In addition to the attributes, each loan had its own unique applicant number and a binary label representing whether the loan was repaid (0) or not (1) i.e. the case of default. The dataset represented application data (collected at the point of applying) and not temporal (time-series). Nevertheless, it contained historical information aggregated in some instances such as the

number of times an applicant changed their mobile phone number in the last couple of years.

In order to overcome the entity identification problem (Han et al., 2011), the attributes needed to be described and demystified to perform cleaning, extraction, aggregation, integration, and loading into a model to derive meaningful inferences and results. A table that describes each one of the attributes (columns), its definition, its data type and its nature (financial, behavioural or social) is available in Appendix 3

The dataset describes 307,511 applicants through 120 attributes (columns). Each one of the said attributes has two (binary) or more classes in the case of nominal attributes and a discrete number of values or continuous in the case of numeric attributes. Attributes with similar nature will be grouped and described in an attributes sub-section (see sub-section 4.4.3); whereas attributes with high details will be discussed and aggregated in the next chapter (see sub-section 5.2.1.2) in order to enable a comparison with the literature in previous sections 2.2.3 through 2.2.5 as well as the industry and practice in chapter 3. Finally, attributes with high number of classes that can be reduced to a lesser number by grouping the two or more of the classes in a meaningful aggregate will be discussed in the dimensionality reduction section 4.4.4.1 and reduced in data cleaning within next chapter (see sub-section 5.2.2.3)

4.4.3. Data Description

The 120 attributes (columns) of the application dataset were grouped into 16 categories that describe different aspects of a borrower's profile. Below are the categories created in this research that cover the attributes:

i. **Loan features:** amount and type (cash, revolving, or consumer POS) of loans requested. Also, for commodity loans (to finance buying assets), what is the price of the underlying commodity and how much financing of its value is requested, and, finally, the purpose of the loan.

ii. **Applicant's net worth:** property ownership, car ownership, income and type of income as opposed to current loans' values owed by the applicant.

iii. **Circumstances when lodging the application:** who accompanied the client when visiting the lender, at what time was the online application filled, on which day of the week, channel of the application (online, offline, call centre), loan application initiated by the customer or by lender via cross-selling marketing activities.

iv. **Life conditions:** education, marital status, number of children and number of other dependents.

v. **Demographics:** age and gender.

vi. **External factors:** external credit scores provided by other credit bureaux and financial footprints (the number of times that the applicant had applied for a loan with other lenders and was rejected) within the recent periods (during last hour, last day, last week, last month, last quarter and last year).

vii. **Accommodation conditions and standards:** the living arrangement of the residence (owned, rented, living with parents, etc.). Also, the size of the apartment, the existence of a communal area, living area, the age and state of the building such as number of lifts, number of entrances, the existence of an emergency door, etc.

viii. **Life stability:** length of employment with current employer, since when the applicant is registered for an electoral vote. Also, looking at time at the current living address, time since having the same ID that is applied with, time since owning a car for, work proximity to living address, time having the same phone number.

ix. **Accessibility/reachability:** did applicant provide valid contact phone number(s) (home, mobile, work, etc.) and e-mail address? Were they working at the time of the application?

x. **Data update/validity:** has the borrower updated the current address and/or work-related data? The date of the current credit bureau report on the applicant?

xi. **Census:** region density in comparison with other regions, region rating and city rating.

xii. **Career-related:** occupation type and organisation type.

xiii. **Social**: the number of friends who had been struggling to repay their loans and the number of friends who in fact went bankrupt and defaulted on a loan.

xiv. **Credit mix:** the different types (credit card, overdraft, vehicle, etc.), currency (on/off-shore) credit an applicant has.

xv. **Past credit performance:** the past number of overdue incidents and their amounts, if loan was rescheduled for early settlement or suppression, the credit outcome (paid, unpaid, ongoing, etc.), length of credit.

xvi. **Withdrawal Pattern:** from credit used, what amount is spent on purchases and what is withdrawn in cash and other uses from the total amount. How many times did the borrower withdraw cash using the credit card?

4.4.4. Data Preparation

In this research, data mining techniques are explained and their implications on the model results are clarified from both theoretical and practical points of view. Thereafter, applying such techniques will be justified and reflected in chapter 5.

4.4.4.1. Dimensionality Reduction

The aim from dimensionality (or numerosity) reduction is to replace original data by smaller size with the least possible information loss. In some extreme cases (such as aggregating the sales of 12 months into a year), there will be no loss of information and the process is described as a 'lossless' reduction. In contrary, reduction that involves approximation would cause some information loss (such as the case of combining HR staff and Administrative staff into one class called staff) and, in this case, it is a 'lossy' reduction (Han et al., 2011). In the next chapter (see chapter 5), dimensions of the dataset were reduced using three different techniques. When it came to attribute construction technique, advanced statistical and data mining techniques were performed and the analysis was systematic. In some research studies, such as the case of Lin et al. (2013), a derived column called 'number of friend lenders' was produced by adding 'real friend lenders' to 'potential friend lenders'. Similarly, a new column was constructed from adding 20 columns originally found in the dataset (see sub-section 5.2.2.1).

In data aggregation (see sub-section 5.2.2.2), many classes within attributes were grouped in fewer categories according to their similarity. Finally, data cleaning was completed by removing highly-correlated variables and dropping missing values within an attribute whenever those represent small proportion (less than 1%) as well as attributes with large composition of missing values (see sub-section 5.2.2.3).

4.4.5. Models and Tests

In this part, the aim is to establish the link between social data and credit scores through explaining the inference of the former on the latter. This will be done in three folds. First, the argument that borrowers who end-up defaulting have a larger bad social network will be tested using a hypothesis testing strategy, which addresses the second research question.

Second, the above-mentioned social networks will be examined more closely to provide evidence that some of the sizes and types (based on classes) inform credit risk assessors more on loan outcomes. The comparison will be based on the case that if the information on the social network was not available. This will be proven statistically using the tree-based information odds and

weights of evidence approaches in the Bayesian analysis. Finding that evidence will contribute to answering part of the third research question.

Third, a logistic regression model will be used in order to estimate the contribution of social network columns towards explaining the credit scores. Achieving this goal will answer the third research question thoroughly.

Finally, an evaluation on how successful is adding the social network columns using state-of-the-art machine learning models will be completed. This is going to answer the fourth research question.

4.4.5.1. Hypothesis Test on Social Data

Considering that social network variables are ordinal (that is having one bad social tie is worse than having none) and both histograms show non-normal skewed distributions, a non-parametric test was performed to compare the distributions of bad social ties (delinquent and defaulting) between both groups of good and bad borrowers. It is worth to note that the samples are independent from each other.

In this research, the five-step approach used in hypothesis testing is followed. Firstly, the null hypothesis is constructed based on a one-tailed theory that those who default on loans (bad borrowers) have less than or equal bad[7] social ties to those who repay their loans. Secondly, the alternative hypothesis states that those who defaulted on their loans have higher number of bad social ties than those who repay their loans. Thirdly, a one-tailed Mann Whitney U test is conducted at level of significance (α) of 0.05. Fourthly, U-statistic will be calculated then compared with the critical value in order to get the results of the test. Finally, the conclusion will be made by either rejecting the null hypothesis or the alternative hypothesis.

In light of the above, hypotheses are presented as:

H_0: *Borrowers who default have less than or equal number of social ties to those who repay.*

H_a: *Borrowers who default have more social ties than those who repay.*

It is worth to note that, by social ties, the author is referring to bad types – those who are in arrears

[7] The hypothesis test will be run on both types: delinquent social ties and defaulting social ties. Also, the null would contradict what the alternative hypothesis is aiming to prove (that is a defaulting borrower has statistically higher number of bad social ties than a repaying borrower).

or default.

Selecting Mann Whitney U test was due to the following reasons:

i. Variables measured are ranked on a scale (number of delinquent and defaulting social ties)
ii. Samples are independent
iii. Samples are sufficiently large since n1 & n2 ≥ 30 (the number of good borrowers is 279,766 whereas the number of bad borrowers is 24,659).

Moreover, in order to calculate U-statistic, the lower of either U_1 or U_2 (for groups 1 and 2 respectively) should be selected. Calculating U for a group 1, for example, is shown below:

$$U_1 = n_1 n_2 + \frac{n1\,(n1+1)}{2} - R_1 \qquad (2)$$

Where:

R_1: the sum of the ranks of all observations in group 1 after combining the 2 sample groups and ordering them ascendingly.

The test was conducted in IBM SPSS 25. The dependent variable in this case was set to be the number of social ties while the outcome of the loan was deemed an independent variable. Although this assumption reverses the original assumption that outcome is influenced by social ties, it can still reveal differences in the social behaviour of those who defaulted within the ex-ante (Lin et al., 2013) effects.

4.4.5.2. Bayesian Analysis on Social Data

The Bayesian formula was applied in the classical study on credit rationing by Guttentag and Herring (1984). As discussed earlier within the literature review chapter (see sub-section 2.3.2.3), Yeh and Lien (2009) praised using Naïve Bayes analysis when it comes to theoretical justification. As such, the aforementioned analysis will justify incorporating social attributes in the classification model. In this section, the different social network sizes within the social network attributes (columns) will be detailed in terms of inferences using the Bayes formula. Results consist of posterior conditional probability, information odds, and the weight of evidence concepts of the aforementioned classes.

In order to study the influence of bad social ties on borrowers, one need to differentiate between the bad loan performers. The below diagram (see Figure 19) illustrates the types of borrowers. In this research, there are two social attributes - one that describes the number of delinquent social ties with observed arrears and another attribute that describes the number of social ties who actually defaulted. Both belong to a revolver or, in other words, a bad borrower. Those attributes

were represented in four columns which were reduced in the dimensionality reduction process while cleaning the data to two columns later in this research (see sub-section 5.2.2.3). As a result, only columns that were measuring the social ties within a 30-day period for both cases were retained. Meanwhile, the social ties measured during the 60-day period were removed due to the high-correlated behaviour.

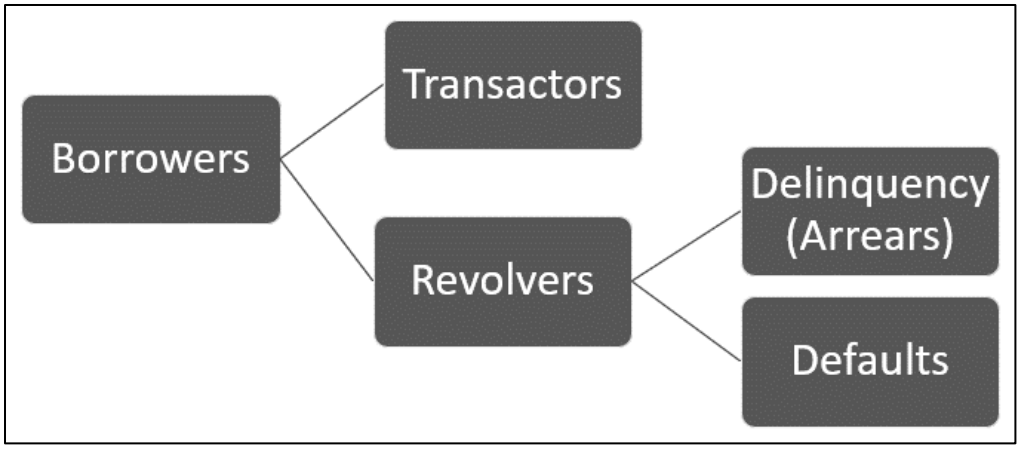

Figure 19: types of borrowers

The aim from investigating each class on its own is to highlight those classes that can give us a better discriminative power between good and bad borrowers than the population odds shown in equation 13. A probability tree that explains part of the classes is demonstrated in the below figure (see Figure 20). The below illustration applies to both social columns with one difference that the observed social ties attribute (column) contains 20 classes as opposed to 6 classes only in the defaulted social ties after removing outliers and an anomaly.

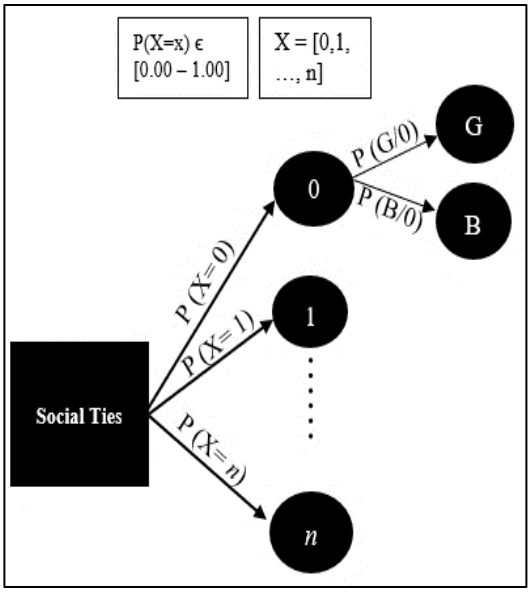

Figure 20: probability tree representation of social ties

The conditional probabilities shown in the probability tree figure (see Figure 20) are known as prior probabilities since they follow the normal forward-looking sequence of events. A corresponding question to those would be, what is the probability that an applicant with 0 observed social ties to be delinquent or default to end-up defaulting oneself (i.e. P(B/0))?

If all the conditional probabilities of being a good borrower are multiplied by their preceding events (classes) then added to each other, a total probability of being a good borrower would result.

$$P(G) = [(P(G/0) * P(0))] + (P(G/1) * P(1)) + \ldots + (P(G/n) * P(n))] \quad (3)$$

The information odds formula of a specific class, say x, is calculated by dividing the probability of this class when it is known that the borrower will repay (good borrower) i.e. P(x/G) by the probability of the same class if the borrower will default (bad borrower) i.e. P(x/B) as demonstrated in Equation 4. The aforementioned terms are considered 'posterior probabilities' since they tend to look backwards at what could have happened in the precedent of the current event.

$$\text{Information odds of class x in attribute X} = I(x) = \frac{P(x / G)}{P(x / B)} \quad (4)$$

Finally, when taking the natural logarithm of the information odds, the weights of evidence (W.o.E) is yielded for the corresponding class x.

$$W.o.E(X=x) = Ln(I(x)) \quad (5)$$

In the subsequent paragraphs, the aforementioned equations will be applied to the social data found in the full cleaned dataset.

4.4.5.3. Logistic Regression Model on Social, Behaviour and Financial Data

The selection of a logistic regression was inspired by studies that investigated financial inclusion matters such as the one completed by De Koker and Jentzsch (2013). The aforementioned study used a logistic regression model to investigate whether a consumer would incline toward the use of financial channels such as mobile payments instead of merely relying on cash. Similarly, a logistic regression model was adopted to predict the probability of default of the applicants. The

logistic regression formula estimates the probability of default (PD) whenever the outcome is default (i.e. Y=1) and the array of characteristic is x (X=x). The below formula (see Equation 6) is used to calculate PD given that it is a fraction between 0 and 1 inclusive.

$$p(Y = 1/x) = \frac{1}{1 + \mathrm{e} - (w_0 + w^T x)} \quad (6)$$

Where:

p (Y=1 / x) is the probability of default,

e is the natural exponent (where the base is e ≈ 2.718),

$\boldsymbol{w_0}$ is the scalar of the parameters vector or the intercept

$\boldsymbol{w^T}$ is the array of variables' coefficients estimated using the maximum likelihood procedure (Baesens et al., 2003)

In addition to the aforementioned justification, it was argued by Kruppa et al. (2013) that models that produce default probabilities (regression models) provide more detailed information about the creditworthiness of consumers than those that merely classify into binary or categories (i.e. machine learning and classification algorithms).

Having said that, the logistic regression model was developed using a machine learning technique. The technique starts by training on 70 per cent of the split data and applying the logit formula on the remainder 30 per cent to see its classification accuracy while producing regression coefficients (see sub-section 5.2.4.3).

In addition to the above-mentioned justifications, assumptions and requirements of the model were met in this research as recommended in the statistical guidelines of James et al. (2013). The following criteria was met in the dataset:

a. Target class is binary (default or repayment).
b. Data is free from missing values (after pre-processing)
c. The predictors are independent (correlations were tested and highly-correlated variables were represented by one component only)
d. There are more than 50 observations per variable (total number of balanced data observations is 49,320 which is higher than 50 times 48 attributes by far).

Therefore, a logistic regression model is considered and the treatment of the data, according to this selection, is explained in the next sub-section (see sub-section 4.4.6).

4.4.6. Data Treatment

In the literature of credit risk modelling and machine learning techniques, there are common

treatments when it comes to handling large datasets. Class imbalance is a practice applied to sensitive binary problems. Additionally, splitting datasets is the norm whenever iterative learning is applied to datasets. Finally, matching attributes with the specific modelling is another treatment that was considered. In the below three sub-sections, those topics are explained in further details.

4.4.6.1. Train/Test Split

When performing the logistic regression model, 70 % of the data were used to train the model and learn from the variances and biases in the data while leaving 30% for testing as a hold-out sample. This technique is common in machine learning as endorsed by Brown and Mues (2012) and is adopted by many consumer credit risk studies such as the one of Kruppa et al. (2013).

Therefore, the dataset was split using the function **train_test_split()** after identifying the target variable (y) and the independent variables (X) in Python 3.

Splitting the dataset after separating the target variable from the remaining independent variables have resulted-in 4 different data objects: X_train, y_train, X_test and y_test. After training the model on the X_train and y_train objects, the performance of the model was measured when predicting y (i.e. producing an array object $\hat{y}$) using the X_test object and benchmarking it against the actual y_test object to estimate the accuracy and other statistical measures of the model. The aforementioned statistical measures are discussed in the confusion matrix part (see Sub-section 4.4.7.1).

4.4.6.2. Sub-setting Attributes

Based on the definitions found in literature, columns of the dataset were classified into three groups – financial, behavioural, and social. The logistic regression model was run on those separately serving the modelling particular description and in combinations. The detailed classification of the attributes can be found in Appendix 3. The aforementioned strategy was inspired by the work of Brown and Mues (2012) who tested classifiers on multiple datasets.

Financial data refers to the financial health of an applicant measured by one's assets such as property, plants, equipment, intangibles as well as one's equity in corporates and funds. Also, financial health would highlight any outstanding debt and financial commitments of the borrower towards other individuals and/or entities. In addition to that, balances of other accounts formed by credit and debit transactions such as income or salary as opposed to regular expenses such as rents or mortgage payments, etc are also part of financial data that can be collected on a loan applicant.

On the other hand, behavioural data reflects choices made by the applicant due to heuristics (pre-

determined rules in mind) and cognitive biases (personal tastes and perceptions) as per the definitions introduced in the works of Hirshleifer (2015) and Taffler (2017) on behavioural finance (see section 2.1). Finally, social data has a relational aspect (Lin et al., 2013) where the borrower is affected and influenced by the connections with other friends within a network (Wei et al., 2015). It is worth to note that the dataset included other demographic data such as age and gender, which are, usually, included in the financial data modelling since those are, generally, found to be commonly-used by lenders.

4.4.7. Evaluation Methods

This sub-section explains what big data statistics were used to measure the success model applied to each variation of the data. It illustrates full equations that are processed by Python 3 software.

4.4.7.1. Confusion Matrix

In order to evaluate the logistic regression model, common measures used in the past have been considered and adopted. When it came to inference and showing impacts of variables, the goodness of the model fit is measured using Pseudo R^2. The aforementioned parameter translates to the explainability of logistic regression function. A clear example of such a focus is the work of Serrano-Cinca and Gutiérrez-Nieto (2016).

On the other hand, in order to measure the performance of the model, the evaluation method of G. Wang et al. (2012) is adopted. Simply, performance is evaluated using statistics derived from a confusion matrix (see Table 11) where those are used to compare the performance of the model on different subsets of the data (financial, behavioural, and/or social).

Accuracy = $\frac{TP+TN}{N\ test}$		True Outcome of Observations		Measured Statistics
		Actual Positives	Actual Negatives	
Model Prediction	Predicted Positives	True Positives (TP)	False Positives (FP)	Precision
	Predicted Negatives	False Negatives (FN)	True Negatives (TN)	
Measures Statistics		Recall / Sensitivity	Specificity	N_{test}

Table 11: confusion matrix

In the table above, true positives are actual defaults who were correctly detected by the logistic regression model used in this research; whereas, false positives are non-defaulters who were predicted as defaulters, mistakenly, in which case their applications would be rejected and an opportunity to make more gains from credit portfolios is missed. The aforementioned error is

equivalent to type I error in hypothesis testing which is rejecting a true null hypothesis i.e. suggesting that an applicant does not belong to the norm of people who commit to paying back on time.

On the other hand, false negatives are applicants who actually defaulted and were not picked by the model as risky borrowers. Instead, those were classified as good borrowers and, as a result, caused an adverse selection. This category has the highest cost as the lender could, potentially, lose the entire amount of the loan amount as well as its reputation and credit rating by regulators. The more false-negatives a model produces, the less sensitive it is to real risks. A recall (or sensitivity) statistic is used to describe how risky or safe a model is and it is shown below in Equation 8. This type of wrong classification is equivalent to a type II error in hypothesis testing which is rejecting a true alternative hypothesis and failing to reject a false null hypothesis i.e. suggesting that a borrower is a good borrower and providing the loan while failing to detect that one will default.

The following equations are used to calculate statistics based on data derived from the classification powers in Table 11 in order to evaluate the performance of the model over different subsets of the data (financial, behavioural, and/or social) in the subsequent sub-sections:

Precision is used to calculate the model's ability to detect those who are actually going to default out of all the ones that are predicted to default whether they actually will or will not.

$$\text{Precision} = \text{True Positives} / \text{Predicted Positives} = TP / (TP + FP) \quad (7)$$

Similarly, recall (or sensitivity) is used to estimate the model's ability identify those who are actually going to default out of all the actual defaulters. Accordingly, the higher it is, the safer the model is going to be.

$$\text{Recall (or sensitivity)} = \text{True Positives} / \text{Actual Positives} = TP / (TP + FN) \quad (8)$$

In addition to that, the specificity statistic investigates the model's ability to correctly identify people who are good borrowers out of a pool of all borrowers who are in fact good and, thus, a high specificity in a model can be of a good cause for increasing profit.

$$\text{Specificity} = \text{True Negatives} / \text{Actual Negatives} = TN / (TN + FP) \quad (9)$$

The last equation (i.e. specificity) is not going to be considered when evaluating the performance of the model on different subsets since credit scoring models focus mainly on the defaulting action (i.e. class '1') represented by the probability of default score.

The accuracy of a model calculates how successful a model is in predicting each class correctly

out of all observations that the model runs on.

$$\text{Accuracy} = \text{True Predictions / all observations} = (TP + TN) / N \quad (10)$$

Additionally, the F1 score is a common measure that combines both equations 7 and 8 when taking twice the fraction of the product of precision and recall divided by their sum:

$$\text{F1 Score} = 2 * \left(\frac{Precision * Recall}{Precision + Recall} \right) \quad (11)$$

Finally, log loss function penalises every data point that was classified falsely. In other words, the lower the score is, the more reliable the model is in classifying the observations correctly. It is calculated using the below formula for 2-class classification problem:

$$\text{Log Loss} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log P_i + (1 - y_i) \log (1 - P_i)] \quad (12)$$

Where:

(N) represents the number of observations;

(y_i) represents a binary indicator of whether an observation belongs to one class;

(P_i) represents the correct prediction of that same class

4.4.7.2. Receivers Operating Characteristics (ROC) Curve

ROC curve measures how efficient the model is in creating a cut-off score which would ideally produce the least number of incorrect classifications. It is equivalent to the probability of correctly classifying two randomly selected borrowers one from each class – defaulting and repaying (Kosinski et al., 2013). The vertical line represents the default outcome; whereas, the horizontal line represents the repayment one. In ROC procedure, scores for borrowers are arranged ascendingly then plotted on the grid system. When a model produces a perfect cut-off score for classifying the 2 target classes, it achieves a 1.00 ROC Area Under Curve (AUC).

CHAPTER 5: RESULTS, FINDINGS AND DISCUSSION

5.1. Interview Findings

In all three types of banks interviewed (state-owned, commercial, and Islamic), data is only collected for those who express their interest in getting loans only. In other words, no credit scores exist for individuals who have existing accounts and established relationship financial profile with the bank as long as those do not apply for credit facilities. The aforementioned phenomena represent an opportunity lost for the lender to attract the safe borrowers who would pay back on time and increase the profits for the lending institution (i.e. the bank in this case).

The interviews with Banker 1 and Banker 2 revealed that social network data can be useful when assessing credit risk. At the commercial bank, a behavioural model provided by an international CRA was adopted. In addition to that, Banker 5 and Banker 6 believed that if this practice is adopted locally, it would make huge improvements since behaviour differs from one region to another. Also, Banker 7 praised the use of logistic regression due to its explainability and simplicity. Finally, both regulators (Regulator 1 and Regulator 2) emphasised on the necessity to expand beyond traditional data by introducing alternative data.

As per the Banker 2, the overwhelming task of collecting customers data in the form of employment contracts, agreements, registration documents and e-mails from government authorities confirming the ownership of a land or a property is a lengthy process and requires a lot of resources

In addition to the manual correspondence and limited scoring to those applying for credit, the data, once collected, is entered to the system manually again which is exposed to human errors/biases. Not only that, but updating the score happens annually. Moreover, the system does not request it from the user, but instead a user voluntarily accesses the core banking system and updates it as per the internal policy and procedures recommendations.

Finally, the commercial bank derived its model from Experian to assess individuals' credit scores. Such a score is mapped into the international categories, which considers variables that are common in the European and North American regions. For example, to be registered for electoral vote is an important aspect in the aforementioned regions, but not that important in the middle east region. a report can be produced.

Overall, managers and professionals in both types of banks in the middle east praised the idea of a local credit bureau especially when they expressed that there used to be a problem with transparency between banks due to rivalry and competition on market share. They feel that a credit bureau helped in giving reliability and a broader view of customers.

Bankers in Qatar expressed that, for the small size of the market and its low population, they are capped with a certain amount of loans they can give. Moreover, they believe that regulators in Qatar are very conservative in the country's credit compliance policies such as the restriction on credit card limits to twice as much as the amount of the applicant's salary. As such, bankers found the idea of relying on alternative data interesting and plausible. Specifically, homophily was rated highly among respondents and social network analysis was thought of as a successful solution anecdotally.

5.1.1. Description

5.1.1.1. The Process of Loan Application

Both the Islamic and commercial banks have confirmed that interviewing the borrower is important. There is a workflow that is followed in both banks. The commercial bank has an online system that tracks the application request. The Islamic bank, on the other hand, has three stages: (a) credit officer interviews the applicant and gathers all documents needed as well as filling the application form then passes it to the manager, (b) credit analyst receives the application and does not interact with the client to ensure unbias, and (c) the credit department makes its final recommendation to the head of retail banking, who decides in cooperation with senior management (according to the amount) on the loan application.

Scores updating seemed inconsistent in both Islamic and commercial banks. It was confirmed that reviewing the model's criteria takes place annually to make sure it is, relatively, consistent with the scores of the CRA. Banker 1 indicated "this is a worry and prone to human error".

5.1.1.2. Third Party and Intermediary

The commercial bank relies on a credit report provided by an international CRA in assessing its customers' credit worthiness. According to a manager in the commercial bank, it is very effective when the customer is active internationally or coming from another country because, aside from the basic checks done on the accounts in Qatar, there is no other way to access the applicants'

international finances but relying on a world-wide trusted credit bureau. The Islamic bank, on the other hand, relies, externally, on local external scores such as the ones provided by the national CRA, which was established in 2011 with the access to banking information from all local banks. Its report shows all incidents of delays or arrears whether in credit cards or loans with other local banks. Also, it flags delinquency or arrears and bounced checks that happened within two years from the date of the application. Communication with the central bank, usually, happens through a system where a bank user has a username and password to inquire about individuals using their document numbers. In addition to the CRA data, the Islamic bank created an in-house system with weights given to components of the model (see Table 12)

5.1.1.3. Data Used by Banks in Qatar

Financial transactional data had been used in Qatar with pre-classifications of some categorical data such as the sector/industry of the applicant. For example, government sector and civil service jobs were preferred as those infer stability and longevity in their trajectories. Regulator 1 confirmed the plan to extend their scoring model beyond the banking context by incorporating utility bill payments data of individuals.

5.1.2. Systems and Models

The Islamic bank's criteria are developed in-house with no automated system that, dynamically, assesses borrowers' financials and transactions. Initially, the bank relied on excel sheets. More recently, it had gone from excel sheets to, manually, entering values in the core banking system for each criterion, which then gets calculated and a score is shown to front office representatives. This is done by the credit officers, where qualitative and quantitative metrics get evaluated then entered. Thereafter, as per their given weights (see Table 12), the score is assigned to each potential borrower. The score of the in-house criteria would represent 50% of the final score while the other 50% is outsourced by a CRA who relies mainly in its system on FICO. The main components of FICO scoring model was discussed in the current lending systems (see sub-section 3.4.1.4). Clearly, there are issues in this process as the criteria might be similar which may cause double effect. In such a case, borrowers who happen to perform poorly in a traditional metric will be penalised massively.

Source	Internal scoring (50% weight)				External scoring (50% weight)
Nature	Qualitative	Points	Quantitative	Points	Credit Bureau
Metric	Reputation	1	Capacity (income, work stability & sector)	60	FICO scoring
	Educational level	2	Banking relationship (acc. turnover, overdrafts, etc.)	10	
	Age	7	Employment type (labour, middle-management, project engineer, senior, etc.)	10	
			Work experience (total years & time in current job)	10	

Table 12: Islamic bank credit scoring criteria

The traditional bank, on the other hand, bought a system developed by an international rating agency, Standard & Poor's (S&P's). S&P's provided its software with best practices and offers updates and localisation to customise the system according to the Qatari market (see Table 13). Similarly, Moody's system accepts entered data and calculates its dual risk rating based on the probability of default (PD) and loss given default (LGD) models built within the system. The factors that matter in the traditional bank are: character (customer info, history), capacity (all sources of income), collaterals (in the case of secured lending), conditions (to mitigate the risk) and capital (especially for high net worth customers and private bankers).

The national credit referencing agency (CRA) relies on FICO model in producing a score that ranges from 300 to 850 with 850 being the highest creditworthiness and 300 being the lowest. It is up to the bank to decide how to deal with the score provided by the credit bureau and whether to consider it as a first step to pass initial assessment in order to qualify for the internal scorecard (as in the case of the traditional bank) or combine it with the bank's internal model (as the case of the Islamic bank).

As a result of the internal score, banks are required by the regulatory entity to map their credit ratings of the clients to the internationally-recognised Standard & Poor's (S&P) and Moody's

categories. In order to give be able to give feedback to the credit bureau whenever requested to do so. In the Islamic bank the maximum weighted score is 100 and is equivalent to AAA in S&P's or Aaa in Moody's as briefly demonstrated by the bank in the table shown below (see Table 13)

S&P Rating	**Internal Score**
AAA	*98 – 100*
AA	*94 – 97.9*
A	*90 – 93.9*
BBB	*84 – 89.9*
BB	*79 – 83.9*
B	*72 – 78.9*
CCC	*60 – 71.9*
CC/D	*59.9 or below*

Table 13: mapping internal scores to S&P's international standard

Obviously, the lower the rate, the higher the risk and, accordingly, a higher rate and more guarantees will be needed.

Behavioural data was, arguably, a reason behind someone's job stability – a main factor that Banker 2 mentioned. In other words, job loss can be predicted by monitoring the behaviour that precedes such an event.

To summarise findings of the primary data collected, it is clear that competition is very limited in the State of Qatar since banks give loans mainly to those who transfer their salaries to the bank. This may drive credit officers to try and compete on selling loans to get performance appraisals. In doing so, they would push for financing unsecured high-risk customers to achieve their targets. One could argue that such officers may advise lender on what to hide from the bank that they work for. The aforementioned behaviour was explained, earlier in this research, in the agency theory as part of behavioural finance section.

Considering salaries is a main component of credit scoring and that requires looking into disposable income. Disposable income refers to the residual money to spend after deducting the liabilities on the borrower.

> *"people who work and get their salaries transferred to our bank get high credit scores since our model gives high weight to work status" (Banker 1)*

However, understanding how a borrower would spend this remaining part of a salary would require understanding someone's traits and behaviour.
In general, bankers disagreed that family ties would influence someone's spending patterns. Nevertheless, they confirmed that borrowers' company (i.e. friendship network) would indicate interests and financial responsibility.

> *"credit risk cannot be derived using genetics. Family ties do not give any indication of attitudes toward debt. Instead, it is inferred from the lifestyle of those that a borrower spends most of the time with" (Banker 2)*

In other words, friends social network needs to be taken into consideration when assessing someone's credit risk. In commercial banking, professionals had already spotted behavioural influence and were open to the idea of incorporating social data. Banker 6 confirmed the existence of behavioural fields within the web-based form provided by an international CRA.

"They [the international credit referencing agency] considered behavioural characteristics in their model" such as providing a work e-mail and the number of times a borrower visited the bank recently. The only challenge, according to Banker 5 and Banker 6, was "the lack of scalable big data infrastructure and connectivity" in Qatar, which makes it harder to integrate social and behavioural data with banking transactions.
Nevertheless, concerns of breaching privacy and ethical laws regulations were, also, raised by Banker 3 as bank collects such data.
Despite believing in alternative data (behavioural and social), Banker 7 opposed the new trend of using machine learning models. This is due to their "lack of interoperability and black box designs". Instead, using a logistic regression model along with subjective judgement of credit professionals provide the best formula to such a large bank.
For regulators, the focus has been on data quality and making sure that infra-structure enables

integrating different data for each borrower using national ID number. Regulator 1 declared "according to an independent assessor, the national credit referencing agency was rated at 99% in data quality and reliability". Such data was provided by energy suppliers, retailers, and government entities such as ministries. Finally, it has been revealed that a nation-wide project to integrate innovative data on individuals and store those in a data centre where banks and lenders can request access through a cluster or a computer node in real-time according to Regulator 6.
instead.

5.2. Modelling and Testing

In this section, Python 3 coding language was used for analysing and modelling the secondary data source. The aforementioned coding language was run on Jupyter Notebook 5.6.0 which is found on the Anaconda Navigator platform for developers.

5.2.1. Exploratory Data Analysis (EDA)

In this sub-section, highlights are given on the dataset in order to understand the context of the variables. When it comes to loan types, tenured indicate a maturity repayment date; whereas, revolving indicates a credit limit that renews periodically whenever a borrower settles the outstanding. The number of tenured loans was almost 10 times more than that of revolving loans (278,136 tenured as opposed to 28,250 revolving applications). The distinction becomes even clearer when comparing the values of tenured loans with the values of the revolving loans. The tenured values summed to almost 20 times more than the those of revolving loans (175 Bn. In comparison with 9 Bn.). The figures below (see Figure 21) illustrate the differences in numbers and visually.

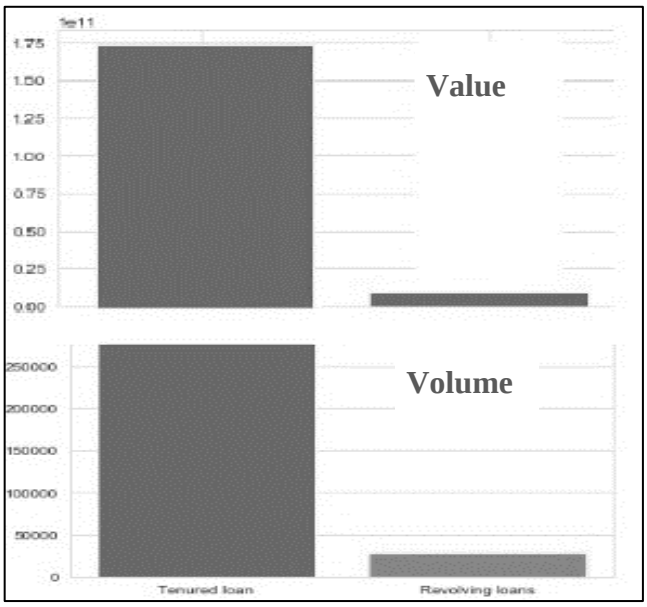

Figure 21: numbers and values of loans categorised by nature

The aforementioned figures were extracted using **value_counts(), groupby(),** and **sum()** functions in Python 3 (see **Error! Reference source not found.**).

Also, a visual representation of the values and volume of loan types are presented in Figure 21. Finally, the data is labelled with the outcome of credit applications given to 307,511 borrowers, where 282,686 borrowers had, successfully, repaid their loans as opposed to 24,825 who defaulted.

It is worth to note that the aforementioned figures are extracted before cleaning the data, which will be conducted later in this chapter. This variable represents the target or, in other words, the dependent variable.

In section 4.4.5.2, the ratio between the two main classes in the dataset will be introduced as a benchmark known as population odds. This benchmark will be used to identify those classes within attributes that add more useful information when classifying the target variable (i.e. probability of default). The below bar chart is created using the **countplot()** function found in *seaborn* library in Python 3 which visualises the targeted label (see Figure 22).

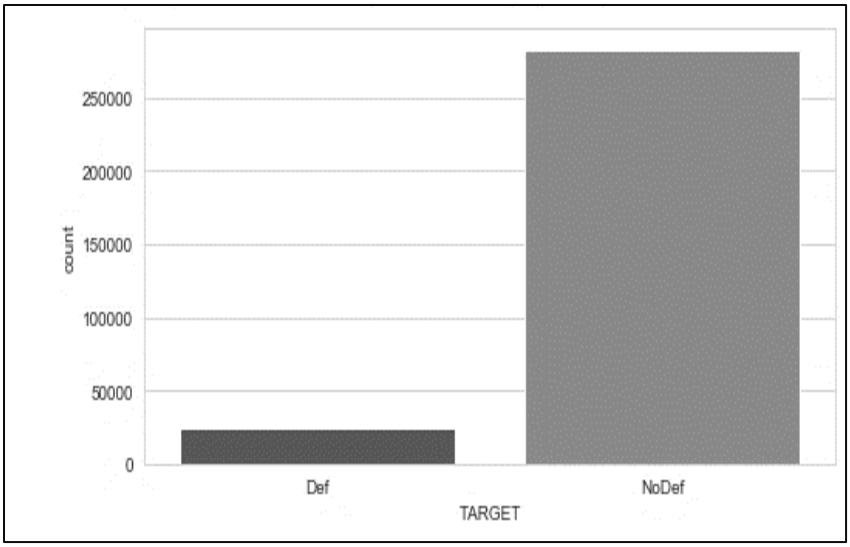

Figure 22: a bar chart for targeted label

Regarding the information odds and weights of evidence found in the observed delinquency within 30 days, it is noted from Figure 29 that when either none or one social tie has an observed delinquency or arrears, there is an additional information that can be inferred surpassing the original information of the population. However, the classes between 15 and 18 social ties, also, presented an additional information with information odds above 1 and positive weights of evidence. Therefore, this social network column does not necessarily indicate the outcome of a loan depending on whether having no social ties with observed delinquency or having many social network ties. The rationale behind this finding is that friends might be observed to be delinquent because they forgot to pay on time or they are out of town or moved the funds to another account or had no access to their bank accounts at the time of the scheduled payment.

On the other hand, the odds and weights of evidence in the defaulting social ties indicated that a borrower with no defaulting social tie has a 4.3% more evidence of being a good borrower than another borrower who has exactly the same features, but an unknown defaulting social network.

5.2.1.1. Descriptive Statistics

The table below (see Table 14) summarises the key attributes in the dataset and gives an overview on the sample of borrowers. The total number of borrowers considered after cleaning the data is 304,427. In the subsequent sections, the data cleaning and dimensionality reduction processes will be explained thoroughly.

	Annual Income	**Credit Amount**	**Annual Instalment**	**Age**	**Bad Social Ties**	
					Arrears	**Default**
Mean	168,667	599,560	27,145	44	1.41	0.14

Std. Deviation	237,927	402,137	14,478	11	2.28	0.44
Min.	25,650	45,000	1,615	20	0.00	0.00
Q1	112,500	270,000	16,573	34	0.00	0.00
Median	147,600	517,266	24,939	43	0.00	0.00
Q3	202,500	808,650	34,641	53	2.00	0.00
Max.	117,000,000	4,050,000	258,025	69	20.00	6.00

Table 14: descriptive statistics of key attributes

5.2.1.2. Classes within Attributes

In terms of taxonomy, some of the attributes (columns) had many detailed classes such as the attributes: loan types, purpose of the loan, type of the company that the applicant works for, type of income streams, and occupation type. The classes of the aforementioned columns are listed in the next few examples for the sake of comparison with the literature level of details.

In 'loans types' column, there are 13 different types - consumer credit, credit card, car loan, mortgage microloan loan for business development, loan for working capital replenishment, cash loan (non-earmarked), real estate loan, loan for purchase of equipment, loan for purchase of equity (margin loan), interbank credit, mobile operator loan. When performing data aggregation as part of dimensionality reduction, the aforementioned loan types will be clustered into two main classes – tenured loans and revolving loans (see sub-section 5.2.2.2). Also, insights on the total value and size of the aforementioned two classes will be discuss in exploratory data analysis section (see section 5.2.1).

As for the 'loan purpose', the following classes were extracted from the dataset to justify the applicant's needs for financing: repairs, urgent needs, buying a used car, building a house or annex, everyday expenses, medicine, payments on other loans, education, journey, purchase of electronic equipment, buying a new car, wedding/gift/holiday, buying a home/land, business development, gasification/water supply, buying a garage, hobby, money for a third person, refusal to name the goal, other, or unknown

Regarding 'type of company' attribute, the following classes were used to differentiate between workplaces: business entity, self-employed, medicine, government, school, trade, kindergarten, construction, transport, industrial, security, housing, military, bank, agricultural, police, postal, ministry, restaurant, services, university, hotel, electricity, insurance, telecom, advertising, realtor, culture, mobile, legal services, cleaning, and religious entities. The aforementioned classes will be reduced clustered (according to similarity and types within one category) in data aggregation sub-

section (see sub-section 5.2.2.2).

In addition to that, the dataset introduced the following classes for the 'type of income streams': working, commercial associate, pensioner, state servant, unemployed, student, businessman, and maternity leave (see Figure 23)

Working	158774
Commercial associate	71617
Pensioner	55362
State servant	21703
Unemployed	22
Student	18
Businessman	10
Maternity leave	5

Figure 23: income stream types found in the application dataset

Regarding 'occupation type', the dataset introduced 18 job titles of which some are very similar and, in sub-section 5.2.2.2, those will be reduced to 4 main job types through data aggregation techniques.

The occupations/job titles found in the dataset were: labourers, sales staff, core staff, managers, drivers, high-skill tech staff, accountants, medicine staff, security staff, cleaning staff, private service staff, low-skill labourers, waiters/barmen staff, secretaries, realty agents, HR staff, and IT staff. As highlighted, the aforementioned were categorised into (1) low income, (2) entry level, (3) middle management', and (4) managers.

5.2.2. Data Wrangling

5.2.2.1. Attribute Construction

The dataset had 20 columns of binary attributes that showed whether or not the applicant had provided 20 different documents or not. Instead of having 20 different attributes that refer to each one of those documents individually, an aggregator was created as a new column counting how many documents the applicant did submit and the value would range from 0 if not submitting any document at all to the application through 20 if all documents were submitted to support the loan application. In the dataset, the largest number of documents provided was 4 documents. The new column was constructed under the name 'no_of_docs_provided' and was added to the dataset. Columns called 'FLAG_DOCUMENT_x' where x is a number ranging from 1 to 20 were deleted. The dataset dimensions were reduced after removing the 20 columns using the **drop()** function in Python 3 from 123 columns (made up of 120 attributes + the ID number of the applicant + the label + the above-mentioned constructed column) to 103. In order to display the dimensions, a **shape** function was called on the dataset.

5.2.2.2. Data Aggregation

In data aggregation, the aim is to group a large number of classes within an attribute (column) into a smaller number of classes based on similarities among the grouped classes. Classes were aggregated using a Python 3 **where()** function in *numpy* library

Loan Type

Loan types in the dataset were classified into 13 classes. Despite the many loan types (mortgage, car, equity, etc.), loans were mainly revolving or tenured. The former indicates the credit renews every time the borrower settles the debt such as the case of credit cards and overdrafts. The latter, on the other hand, refers to loans that are settled on a repayment schedule in order to completely be repaid by the tenure. Therefore, the aforementioned classes were put into two major groups of loans: revolving loans and tenured loans. As previously defined in section 5.2.1, the main difference is that revolving loans represent continuous credit renewal whenever a payment is made to settle outstanding balances; whereas, tenured loans are tied to a repayment schedule.

In order to aggregate the classes, full mapping of the 13 classes was performed as per the below table (see Table 15). The resulting classes were mainly tenured loans as opposed to revolving loans with 278,136 applications made for the former and 28,250 for the latter

Occupation Type

Similarly, the occupation list contained 18 job types some of which are similar and related in terms of work nature. Therefore, job type classes that exhibited similarity were aggregated in a single class to strengthen the inference of the 'occupation type' attribute (column). Accordingly, jobs that required physical efforts were deemed as a 'low-income' occupation type. Whereas, office-type jobs were aggregated into an 'entry-level' bearing in mind that 'laborers' class was included in this aggregation given that the dataset had a 'low-skill laborers' class in which case it is believed that laborers would be paid a premium for their quality of performance. The third aggregation included three classes – 'High skill tech staff', 'accountants', and 'private service staff'. The aforementioned classes were aggregated in a 'middle management' class which exhibits direct interactions with decision-makers. Finally, 'managers' remains unchanged and represents the highest-paid class. Table 16 maps original classes to their respective aggregated category. This attribute (column) will be treated as an ordinal data type when its classes are transformed (see sub-section 5.2.3.1) since salary scale from low to high would inform the model more than merely separating categories of salaries.

Loan Types	
Original classes	Aggregated classes N (%)
Consumer credit	Tenured loan 278,136 (91%)
Car loan	
Mortgage	
Microloan loan for business development	
Cash loan	
Real estate loan	
Loan for purchase of equipment	
Loan for purchase of equity (margin loan)	
Interbank credit Mobile operator loan	
Credit card	Revolving loan 28,250 (9%)
Overdraft	
Loan for working capital replenishment	

Table 15: aggregating loan types

Occupation Type	
Original classes	Aggregated classes N (%)
Low-skill Laborers	Low-income 39,266 (19%)
Cleaning staff	
Waiters/barmen staff	
Security staff	
Cooking staff	
Drivers	
Laborers	Entry-level 126,129 (60%)
Sales staff	
Core staff	
HR staff	
IT staff	
Secretaries	
Medicine staff	
Realty agents	
High skill tech staff	Middle management 23,742 (11%)
Accountants	
Private service staff	
Managers	Managers 21,231 (10%)

Table 16: aggregating occupation type

(** no aggregation for 'managers' class)

The resulting list of occupation types became much simpler to be encoded, scaled, and processed by the logistic regression model in section 4.4.5.3. The number of 'managers' class remains constant as it does not get aggregated unlike the rest of the classes.

Finally, organisation type column had included a massive 58-item list of types of organisations.

The list had a built-in taxonomy where some industries had types themselves. For example, there were 3 types of business entities – business entity: type 1, business entity: type 2, and business entity: type 3. Similarly, trade entities had 7 types; whereas, industrial organisations had 13 types. Finally, transportation companies had 4 types and those were merged into one organisation type called 'transportation'. The Python 3 code mentioned earlier (refer to the first paragraph in this sub-section 5.2.2.2) was used in reducing the classes of 'organisation type' attribute from 58 classes to 32 using the simple taxonomy found in the dataset.

5.2.2.3. Data Cleaning

ID's Column

IDs of applicants are unique. In other words, each of the 307,511 observations (rows) belong to 307,511 unique customers. The main benefit of an ID column is when joining other datasets to from different sources to the current application dataset using the identifier column, in which case it will be called a primary key (P. P.-S. Chen, 1976). Nevertheless, this is not the case and it is believed that the ID's column is not necessary in this analysis. Thus, it is going to be removed as well taking the number of attributes down to 102. The **drop()** function is used in Python 3 to remove this column while setting the parameter 'axis' to 1 to inform the algorithm that the removal will happen vertically.

The dataset was trimmed to retain only applicant with 20 observed network ties or less in column (1) and with 6 defaulted cases or less in column (2) by slicing the original dataset as shown in the below Python 3 code. As a result, 104 applicants were excluded since their social ties values were extremely high.

The new descriptive statistics still convened the positive skew with data. However, the distribution looked more realistic and showed a gradual pareto declining shape. These frequencies in the figures below will be discussed and analysed in more details according to the target class (default and non-defaulted loans) in the modelling topic under the information odds and weights of evidence section (see section 4.4.5.1)

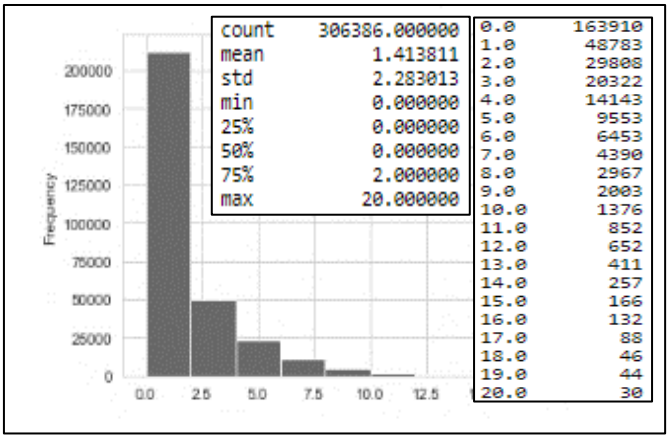

Figure 24: histogram of delinquent social ties

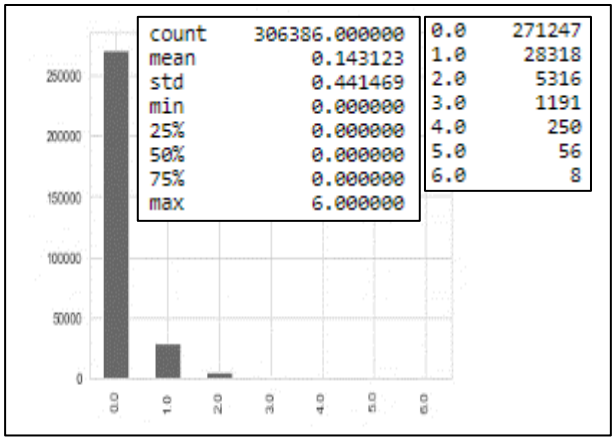

Figure 25: bar chart of defaulted social ties

Removing Missing Values (NA's)

The current dataset consists of 97 attributes (columns). It was clear that some of the columns had a large number of missing values, which limits the contribution of those columns to the model used. Most of the previous procedures performed in this sub-section (5.2.2.3 Data Cleaning) aimed at reducing the dimensionality of the dataset horizontally (i.e removing attributes or columns) except in the last procedures where 104 applicants were removed due to having outliers and anomalies within their social data columns. The said procedure resulted-in a vertical reduction (removing rows).

When dealing with missing values (NA's), there will be two options: (a) dimensionality reduction and (b) data pre-processing. In dimensionality reduction, either vertical or horizontal reduction will take place. In contrast, pre-processing refers to replacing NA's with either mean, median or mode of the attribute based on the data type and distribution (see sub-section 5.2.3.1).

Taking the decision on whether to reduce or pre-process and, if reduce, whether to drop columns (horizontal reduction) or delete observations in rows (vertical reduction) is going to be decided based on categorising the variables in 3 groups:

Group A: attributes with NA's representing more than a third of their values (>33.3%) will be removed (horizontal reduction).

Group B: attributes with NA's representing more than 1 percent but less than a third will be kept given that those NA's will be replaced with a central value (mean, median or mode) based on the variable type

Group C: attributes with NA's representing 1 percent or less will be kept given that those observations with NA's be removed (vertical reduction).

Also, the below attributes represent those with missing values and they add up to 63 attributes. A full list of the 120 attributes of the original dataset can be found in the table in Appendix 3.

Group	Attributes (columns)	Missing Values	% of Total Values
Group A	Communal area of the building – median of the neighbourhood	214162	69.9
	Communal area of the building – mean of the neighbourhood	214162	69.9
	Communal area of the building – mode of the neighbourhood	214162	69.9
	The non-living area of the apartment – median of the neighbourhood	212812	69.5
	The non-living area of the apartment – mean of the neighbourhood	212812	69.5
	The non-living area of the apartment – mode of the neighbourhood	212812	69.5
	State of the building ownership	209609	68.4
	Living area of the apartments – median of the neighbourhood	209513	68.4
	Living area of the apartments – mean of the neighbourhood	209513	68.4
	Living area of the apartments – mode of the neighbourhood	209513	68.4
	Minimum number of floors in the building – median of the neighbourhood	207968	67.9
	Minimum number of floors in the building – mean of the neighbourhood	207968	67.9
	Minimum number of floors in the building – mode of the neighbourhood	207968	67.9
	Age of the building – median of the neighbourhood	203830	66.5
	Age of the building – mean of the neighbourhood	203830	66.5
	Age of the building – mode of the neighbourhood	203830	66.5
	Owned Car Age	202194	66
	Area of land the building is built on - median of the neighbourhood	182014	59.4

Area of land the building is built on - mean of the neighbourhood	182014	59.4
Area of land the building is built on - mode of the neighbourhood	182014	59.4
Basement area – median of the neighbourhood	179384	58.5
Basement area – mean of the neighbourhood	179384	58.5
Basement area – mode of the neighbourhood	179384	58.5
Scores of credit referencing agency (CRA) 1	172769	56.4
The non-living area of the building – median of the neighbourhood	169181	55.2
The non-living area of the building – mean of the neighbourhood	169181	55.2
The non-living area of the building – mode of the neighbourhood	169181	55.2
Number of elevators/lifts in the building – median of the neighbourhood	163409	53.3
Number of elevators/lifts in the building – mean of the neighbourhood	163409	53.3
Number of elevators/lifts in the building – mode of the neighbourhood	163409	53.3
Material used in walls of borrower's accommodation	155887	50.9
Apartment area – median of the neighbourhood	155597	50.8
Apartment area – mean of the neighbourhood	155597	50.8
Apartment area – mode of the neighbourhood	155597	50.8
Number of entrances in the building – median of the neighbourhood	154375	50.4
Number of entrances in the building – mean of the neighbourhood	154375	50.4
Number of entrances in the building – mode of the neighbourhood	154375	50.4

	The living area of the building – median of the neighbourhood	153904	50.2
	The living area of the building – mean of the neighbourhood	153904	50.2
	The living area of the building – mode of the neighbourhood	153904	50.2
	House type	153845	50.2
	Maximum number of floors in the building – median of the neighbourhood	152575	49.8
	Maximum number of floors in the building – mean of the neighbourhood	152575	49.8
	Maximum number of floors in the building – mode of the neighbourhood	152575	49.8
	Years since constructing the building – median of the neighbourhood	149568	48.8
	Years since constructing the building – median of the neighbourhood	149568	48.8
	Years since constructing the building – median of the neighbourhood	149568	48.8
	Total area of the building	148005	48.3
	Existence of emergency exit(s)	145335	47.4
Group B	Job type category	96018	31.3
	Scores of credit referencing agency (CRA) 3	60671	19.8
	Credit inquiries within the last hour	41336	13.5
	Credit inquiries within the last day	41336	13.5
	Credit inquiries within the last week	41336	13.5
	Credit inquiries within the last month	41336	13.5
	Credit inquiries within the last quarter	41336	13.5
	Credit inquiries within the last year	41336	13.5
Group C	Borrower companions at the time of application	1292	0.4
	Scores of credit referencing agency (CRA) 2	656	0.2
	Value of underlying asset	278	0.1
	Annual instalments (annuity)	12	0

	Number of family members	2	0
	Number of days since changing contacts	1	0

Table 17: columns with missing values along with absolute and relative frequencies

The horizontal reduction resulted-in removing the 49 attributes (columns) that are found to have high NA's percentages and, thus, were put in group A (see Table 17). As a result, the cleaned data set had 48 columns left.

After dealing with columns in group A, the next step was to isolate those attributes (columns) with insignificant missing values found in group C by reducing the dimensions vertically. The NAME_TYPE_SUITE column had the highest number of NA's within group C (i.e. 1,292). Those missing values were removed using the **dropna()** function and it is worth to mention that, by removing those 1,292 observations, the remaining columns lost many of their NA's as the accompanying type column was cleaned. As a result, information loss was minimal.

The resulting dataset had 48 columns and 304,427 rows. As stated earlier, the original data set had 122 columns and 307,511 applicants.

5.2.3. Pre-processing

When pre-processing the data set, missing values (NA's) found in columns within group B were replaced with representative values of the attributes (columns) that those belong to. Thereafter, a class balancing procedure was completed to eliminate any over-representation bias and allow fair learning from both labels (default and no-default cases) to come up with the most representative precision, recall, accuracy and F1 scores. Finally, splitting the data set into training and testing parts where the model derived from the training will be evaluated using the testing data that were held out to compare between a predicted and an actual label. The train/test splits procedure was explained in sub-section 4.4.6.1. In the following sub-sections, each of the aforementioned process will be explained and illustrated through figures, functions, or algorithms.

5.2.3.1. Filling Missing Values

After removing attributes with two thirds or more missing values (NA's) and removing observations with missing values that represent 1% or less of the total number of observations (applicants), the dataset still had 8 columns with less than a third of their values as NA's (See Table 17). Those attributes (columns) were classified as group B in sub-section 5.2.2.3. The NA's in the aforementioned attributes/columns will be filled according to the data type of the column as follows.

Numerical – Continuous

For numerical (quantitative) continuous variable, a decision to be made on whether a mean or median will be used to fill in missing values based on the distribution's skewness as follows: for skewed distribution, the median would replace the missing value and, for symmetric distributions, the mean will be used. The remaining 8 columns with NA's included one numerical (quantitative) continuous variable (i.e. Scores of credit referencing agency (CRA) 3).

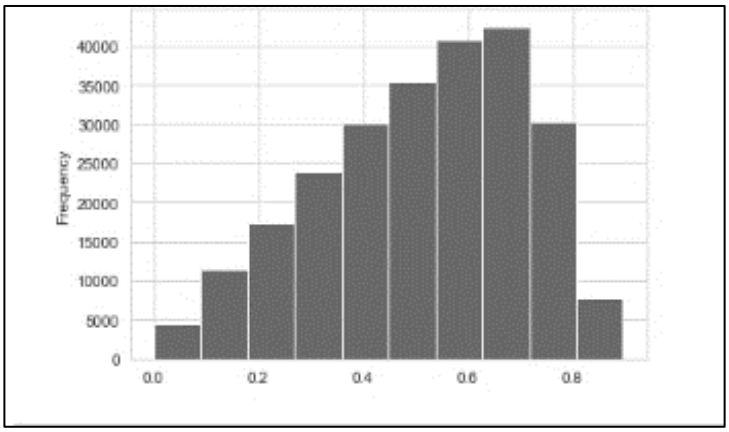

Figure 26: histogram distribution and skewness of a continuous variable

Plotting a histogram in Python 3 coding language using the **plot()** function and the **kind='hist'** argument revealed a negative skew in the distribution and a skewness score of -0.409 was extracted using the **skew()** function and the **skipna=True** argument (in order to avoid NA's that would have been treated as zeros) in Python 3 (see Figure 26).

Since the distribution is not symmetric, the decision to replace 60,229 missing values with the median was taken. The median is 0.535 and the **fillna()** function was called on this value to be used as a replacement.

Numerical – Discrete

For numerical (quantitative) discrete variables (i.e. integers) such as the number of times an applicant submitted loan applications with other lenders during the last hour, day, week, month, quarter and year, the median number will be used to fill in the missing values. The median for the first five time periods (hour, day, week, month, and quarter) is zero; whereas, the median for the longest time period was '1'. The function **fillna()** in Python 3 was called on the median

Categorical (Nominal and Binomial)

As seen in the previous sub-section, 7 out of the 8 remaining attributes with NA's were treated with the median when filling those NA's. However, the last attribute (column) left was the type of occupation and there were 96,018 applicants with missing occupation type (see Table 17). Recall that in the dimensionality reduction section, occupations were aggregated into 4 main types when

performing data aggregation (see sub-section 5.2.2.2). The aforementioned aggregated types were low income, entry level, middle management, and managers types. Since neither arithmetic mean nor median can be calculated for qualitative categories, the mode will be used when replacing NA's as it is the most frequent type.

Job types are distributed as the following: 125,308 for entry level, 39,039 for low income, 23,583 for middle management, and 21,110 for managers (see pie chart in Figure 27). Therefore, the NA's were replaced by 'entry level' using the **fillna()** function in Python 3. The entry level category increased by 95,387 filled NA's.

The resulting dataset had no missing values in any of its 304,427 rows and 48 columns.

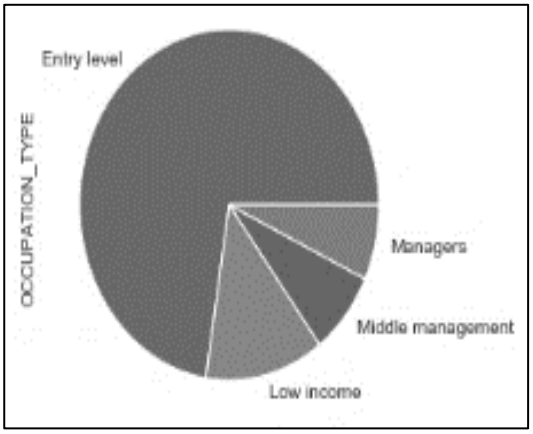

Figure 27: job types

5.2.4. Results of Models and Tests

5.2.4.1. Hypothesis Testing

When looking at social networks and across two outcome groups – default and repayment, a positive skew is exhibited in the distributions of both groups. The social network variables are the number of delinquent and the number of defaulting network ties. The aforementioned variables are not continuous, but instead are discrete and can be ranked. Although the repayment group has larger number of observations in comparison with the defaulting one (279,767 borrowers who repaid as opposed to 24,660 who did not), both distributions of those who repaid and who did not looked, relatively, similar in the case of delinquent social ties and the case of defaulting social ties. A descriptive analysis was performed on both social network types: delinquent and default ties. The mean and standard deviation were reported for both types among both groups of good and bad borrowers (see Table 18 and Table 19).

Type 1 of Social Network Ties		
	Good borrowers	Bad borrowers
Sample Size	279,767	24,660
Mean	1.41	1.49
Std. deviation	2.28	2.35
Min.	0.00	0.00

Type 2 of Social Network Ties		
	Good borrowers	Bad borrowers
Sample Size	279,767	24,660
Mean	0.14	0.19
Std. deviation	0.43	0.52
Min.	0.00	0.00

Q1	0.00	0.00
Median	0.00	0.00
Q3	2.00	2.00
Max.	18.00	30.00

Table 18: descriptive statistics for delinquent social ties

Q1	0.00	0.00
Median	0.00	0.00
Q3	0.00	0.00
Max.	6.00	14.00

Table 19 descriptive statistics for default social ties

The difference between the averages of the aforementioned two groups of borrowers could have happened by chance or could have happened due to the influence of borrowers' social network ties that resulted-in a borrower being in one of the two groups (repayment or defaulted). In other words, if the number of bad social ties for those who ended-up not repaying their loans is, significantly, higher than the number of ties of those who did repay, it can be argued that social data help in assessing credit risk and predicting credit scores.

The below results (see Table 20) were produced at both confidence levels 95% and 99%. It is proven that the distributions are different between both groups of defaulters and non-defaulters regardless of the network type.

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of 'delinquent social network ties' for group G: good borrowers is the same as or shifts to the right of distribution for group B: bad borrowers	Independent samples Mann-Whitney U test	0.000*	Reject the null hypothesis
2	The distribution of 'defaulting social network ties' for group G: good borrowers is the same as or shifts to the right of distribution for group B: bad borrowers	Independent samples Mann-Whitney U test	0.000*	Reject the null hypothesis

Table 20: results of Mann-Whitney U test

* at 95% and 99% confidence levels

5.2.4.2. Bayesian Analysis

The population is considered all applicants in the cleaned dataset before balancing classes i.e. 304,427 applicants and the population odds is calculated by dividing the probability of good borrowers (non-defaulting) i.e. P(G) by the probability of bad borrowers (defaulting) i.e. P(B) as per the what was briefly discussed in class balancing sub-section (see sub-section **Error! R eference source not found.**)

$$\text{Odds of the population} = O_{pop} = \frac{P(G)}{P(B)} = 279{,}767 / 24{,}660 = 11.345 \qquad (13)$$

As explained in the data cleaning sub-section under dimensionality reduction (see sub-section

5.2.2.3), classes within the observed social ties attribute (column) were reduced to 20 since the larger classes of applicants with 21 social ties or more contained only few applicants (each class had 5 applicants or less) and, in one case, one of the classes was an anomaly (an applicant that has 348 observed social ties with delinquency and arrears in comparison with the closest applicant of 47 similar ties).

The below figure (see Figure 28) illustrates a cross tabular view of the classes by the outcome of the loan where '0' denotes repayment (no default – good borrower) and '1' denotes default (bad borrower). In addition to the cross-tabulation, the adjacent horizontal bar chart illustrates percentages of the reduced classes ranging from those who had no observed delinquent social ties to those having 20 such ties.

Number of Delinquent Social Ties	Repayment Outcome	Default Outcome	Total
0	149967	12889	162856
1	44614	3873	48487
2	27180	2448	29628
3	18520	1666	20186
4	12795	1244	14039
5	8709	791	9500
6	5854	563	6417
7	3949	409	4358
8	2695	244	2939
9	1822	168	1990
10	1249	120	1369
11	765	81	846
12	585	60	645
13	363	44	407
14	233	24	257
15	153	11	164
16	123	9	132
17	81	6	87
18	44	2	46
19	39	5	44
20	27	3	30
Grand Total	279767	24660	304427

Table 21: odds of repayment based on delinquent social ties

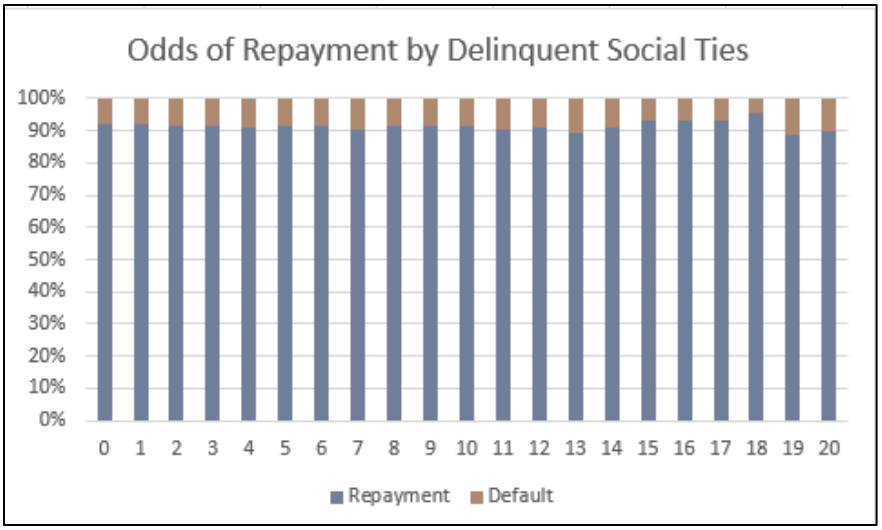

Figure 28: stacked bar chart of the odds of repayment based on delinquent social ties

Consequently, odds, information odds, and weight of evidence for each of the classes was performed to highlight those classes which have more discriminative powers and help in predicting the probability of default better. The summary depicted in Figure 29 below summarises the statistical inferences of each of the classes. A Python 3 code was run using the function **crosstab()** from *Pandas* library to apply cross-tabulation on the target and the observed social ties columns. In addition to the previous function, the function **DataFrame()** from *Pandas* library was used to construct the table whereas the function **concat()** from the same library was used to add columns to the table. Finally, the function **log()** from *numpy* library was used to apply a mathematical operation – taking natural logarithm of the information odds to calculate the weights of evidence (WoE).

SOCIAL_CIRCLE	G	B	Odds	Px	P(G/x)	P(B/x)	I(x)	WoE
0.0	149967	12889	11.635270	0.534959	0.920856	0.079144	1.025588	0.025266
1.0	44614	3873	11.519236	0.159273	0.920123	0.079877	1.015360	0.015244
2.0	27180	2448	11.102941	0.097324	0.917375	0.082625	0.978666	-0.021565
3.0	18520	1666	11.116447	0.066308	0.917468	0.082532	0.979857	-0.020349
4.0	12795	1244	10.285370	0.046116	0.911390	0.088610	0.906602	-0.098052
5.0	8709	791	11.010114	0.031206	0.916737	0.083263	0.970484	-0.029960
6.0	5854	563	10.397869	0.021079	0.912264	0.087736	0.916518	-0.087174
7.0	3949	409	9.655257	0.014315	0.906150	0.093850	0.851060	-0.161272
8.0	2695	244	11.045082	0.009654	0.916979	0.083021	0.973566	-0.026789
9.0	1822	168	10.845238	0.006537	0.915578	0.084422	0.955951	-0.045049
10.0	1249	120	10.408333	0.004497	0.912345	0.087655	0.917440	-0.086168
11.0	765	81	9.444444	0.002779	0.904255	0.095745	0.832478	-0.183348
12.0	585	60	9.750000	0.002119	0.906977	0.093023	0.859412	-0.151507
13.0	363	44	8.250000	0.001337	0.891892	0.108108	0.727194	-0.318561
14.0	233	24	9.708333	0.000844	0.906615	0.093385	0.855739	-0.155790
15.0	153	11	13.909091	0.000539	0.932927	0.067073	1.226014	0.203768
16.0	123	9	13.666667	0.000434	0.931818	0.068182	1.204645	0.186185
17.0	81	6	13.500000	0.000286	0.931034	0.068966	1.189954	0.173915
18.0	44	2	22.000000	0.000151	0.956522	0.043478	1.939185	0.662268
19.0	39	5	7.800000	0.000145	0.886364	0.113636	0.687529	-0.374651
20.0	27	3	9.000000	0.000099	0.900000	0.100000	0.793303	-0.231550

Figure 29: probabilities, odds, information odds and weights of evidence for delinquent ties

Similarly, the social ties who were declared bankrupt and, actually, defaulted on their loans during a 30 days-past-due (DPD) period were put in a cross-tabular format against the outcome of the applicant's loan's outcomes. Those social ties were reduced to 7 classes (from 0 ties to 6 ties) since the classes of 7 and included outliers and an anomaly. As explained in sub-section 5.2.2.3, the three applications with 7 defaulting ties or above were removed since the outliers/anomalies provided no odds information.

The odds for each class of social ties (from 0 to 6) were visualised in a horizontal, stacked bar chart (see Figure 30) in order to spot the trend in the ratios of good to bad borrowers across the classes.

Number of Defaulting Social Ties	Repayment Outcome	Default Outcome	Total
0	248523	20989	269512
1	25276	2862	28138
2	4667	616	5283
3	1032	150	1182
4	214	34	248
5	48	8	56
6	7	1	8
Grand Total	279767	24660	304427

Table 22: odds of repayment based on defaulting social ties

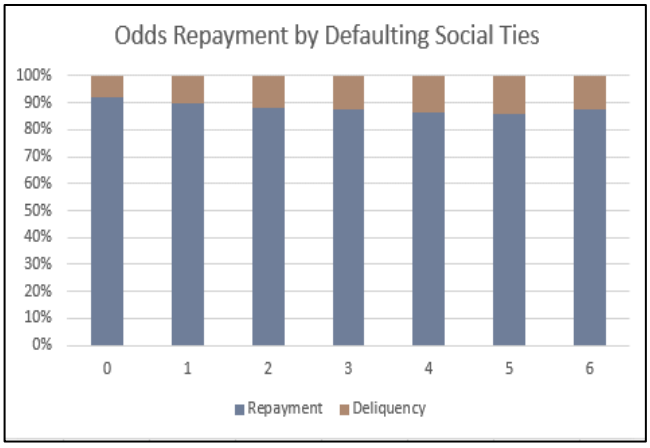

Figure 30: stacked bar chart of the odds of repayment based on defaulting social ties

Consequently, odds, information odds, and weight of evidence for each of the classes were performed to highlight those classes which have more discriminative powers and help in predicting the probability of default better. The summary depicted in Figure 31 below highlights the statistical inferences of each of the classes. A Python 3 code was run using the function **crosstab()** from *Pandas* library to apply cross-tabulation on the target and the observed social ties columns. In addition to the previous function, the function **DataFrame()** from *Pandas* library was used to construct the table whereas the function **concat()** from the same library was used to add columns to the table. Finally, the function **log()** from *numpy* library was used to apply a mathematical operation – taking natural logarithm of the information odds to calculate the weights of evidence (WoE).

SOCIAL_CIRCLE	G	B	Odds	Px	P(G/x)	P(B/x)	I(x)	WoE
0.0	248523	20989	11.840631	0.885309	0.922122	0.077878	1.043690	0.042762
1.0	25276	2862	8.831586	0.092429	0.898287	0.101713	0.778458	-0.250440
2.0	4667	616	7.576299	0.017354	0.883400	0.116600	0.667811	-0.403750
3.0	1032	150	6.880000	0.003883	0.873096	0.126904	0.606436	-0.500156
4.0	214	34	6.294118	0.000815	0.862903	0.137097	0.554794	-0.589159
5.0	48	8	6.000000	0.000184	0.857143	0.142857	0.528869	-0.637015
6.0	7	1	7.000000	0.000026	0.875000	0.125000	0.617013	-0.482864

Figure 31: probabilities, odds, information odds and weights of evidence for defaulted ties

The above figure (see Figure 31) suggests that lenders can gain 1.044 times more information had they knew a borrower has no defaulted social tie. A weight of evidence of score improvement by 4.3% emphasises this argument as seen in the aforementioned table.

5.2.4.3. Logistic Regression

A logistic regression model was run on different subsets and combination of the original dataset. In each trial, the results are shown with both classification and interpretation powers for comparison.

i. Financial Data

The selected columns are the ones available in the traditional core banking records and the meta data that give indication of the financial health of an applicant. Those columns are: (1) loan type, (2) gender, (3) annual income of the applicant, (4) loan amount, (5) instalment amount, (6) price of goods that are going to be bought using the loan, (7) income type whether a student, a pensioner, working, businessman or other types, (8) age, (9) type of job whether a low-income, entry level, etc., (10) type of organisation such as industrial, trading, transport, agricultural, financial, etc., (11) the number of supporting documents provided by the applicant, (12) and (13) external scores from credit rating agencies.

The logistic regression was run and produced a Pseudo R^2 of 9.5% that translates to the goodness of fit of the model Sigmoid function on to the data. The figure below shows the results of the logistic regression model.

The columns that demonstrated significant influence on the probability of default score were (1) loan type, (2) gender, (7) income type, (9) job type, (10) company type, (11) the number of documents provided, (12) and (13) the scores provided by external credit rating agencies. According to the above results, the logistic regression model sigmoid function is shown below. The sign of the coefficient explains the direction of the relationship between the explanatory variable and the target variable (probability of default). For example, the negative sign that precedes the coefficient of the number of documents provided means that the more documents submitted with the application the less probability of default is expected. Similarly, external credit scores have negative coefficients because the higher one's external score is, the lower one's probability of default should be. Interestingly, income stream type has a positive coefficient. The positive coefficient perhaps suggests looking into disposable income instead merely the type of income stream to explain the influence of income type on probabilities of default and credit scores.

PD =

$$\frac{1}{1+e^{-(0.63+1.09*Loan\ Type+0.44*Gender+0.13*Income\ Type-0.184*Job\ type+0.01*Org\ type-0.13*Docs\ provided-2.03*Score2-2.17*Score3)}}$$

A confusion matrix is produced in Python 3 from the *scikit-learn* library within the *metrics* folder

and the target variables were predicted then benchmarked to the 'y_test' object to produce the below results (see Table 23). It is worth to note that the total number of observations in the table represents the 30% test data that was held out of the balanced dataset of 49,320 applicants distributed equally between defaulters and non-defaulters.

TP = 4,821	FP = 3,829
FN = 2,544	TN= 3,602

Table 23: confusion matrix of logistic regression model based on financial data

Finally, the accuracy of the model is 0.57 and so is the F1 score. Also, the precision and sensitivity scores were 0.57. See Table 26 for comparisons between different results of the logistic regression model based on the nature of the dataset.

ii. Behavioural Data

When selecting the behavioural data, there were 33 columns in the subset of the full balanced dataset (32 behavioural attributes and the target dependent variable). The logistic regression exhibited a modest interoperability through a Pseudo R^2 of 5%.

In the above summary, it is noted that 19 attributes exhibited significant relationship with the probability of default (the outcome of the loan). The below confusion matrix (see Table 24) depicts the model's true and false classifications. Other subsets of the data are found in Table 26.

TP = 4,245	FP = 3,203
FN = 3,120	TN= 4,228

Table 24: confusion matrix of logistic regression model based on behavioural data

The accuracy, precision, and F1 score are 0.57; whereas, the sensitivity (recall) of the model is at 0.58.

iii. Social Data

The cleaned balanced dataset included 2 columns that have relational aspects. As a result, those columns represented social network data and were isolated for training and testing. The aforementioned columns are: (1) the number of delinquent social ties (or those who are in arrears or observed to have financial troubles), and (2) the number of defaulting social ties. Both networks were measured during the last 30-day period, which gives the network a dynamic nature. In financing terms, the social ties are deemed to be within a 30 day past due (DPD) group. It is worth to note that those who sustained their delinquency status beyond the 30-day period were included in new columns of 60-DPD delinquency and default. Therefore, there had been a lot of redundancy

manifested in high correlation coefficients that led to dropping the 2 social network 60-DPD columns. Finally, as discussed in the literature review chapter, arrears that are worth 90 days of instalments happening during a 360-day period make-up a default case by the definition of Basel II committee (see section 2.1.5).

The results summarised the performance of the logistic regression model. First, the model has a very low interoperability represented by a minimal Pseudo R^2 of 0.2%. In other words, this model is able to justify 0.2% only of the variations between different outcomes. Clearly, this is due to the low dimensions used in the model since it only relied on 2 independent variables. Second, it is noted that one of the variables, delinquent social ties, had a p-value (0.035) above the critical 0.05 threshold. In other words, there is no significant relationship between the dependent variable, probability of default, and this particular dependent variable. Third, contrary to the previous statement, the default social ties variable is significantly related to the independent variable with a p-value close to zero and, thus, within the significant region. It is noted that the coefficient of the aforementioned social variable is positive with a value of 0.22. This is consistent with the general convention of wisdom that the higher the number of defaulting social ties, the higher PD score is. Consistent with the model's interoperability, the classification accuracy, also, was weak with the model's accuracy being just above a random-walk model at 0.52. Moreover, the log loss function recorded a high loss of 16.75 while other metrics were extracted from the below confusion matrix (see Table 25). From the table, the precision is 0.51, but the sensitivity is 0.89. In other words, the model is very sensitive to any risky network and would classify most of the borrowers as potentially defaulters.

TP = 6,519	FP = 6,329
FN = 846	TN= 1,102

Table 25: confusion matrix of logistic regression model based on social data

5.2.5. Social Effects

In this section of chapter 0, social data will be used to contrast results before and after it is added. As emphasised by Brown and Mues (2012), only a fraction of percent improvements in accuracy could cause the cause a lot of savings. Before looking at the improvements of the performance of logistic regression model highlighted in Table 26 after using different variations of data, the results of ROC curve are assessed. First, it is discussed in section 4.4.7.2 that the highest possible AUC was 0.58 and this result was achieved using 3 different combinations of datasets. Two of those

combinations involved social network columns. Specifically, the combination of behavioural and social data included the least number of columns (i.e. 34 columns) as opposed to the other two combinations (47 and 45 columns). This indicates that social data was efficient in achieving similar results to those achieved with combinations that have higher dimensions. For example, when gigantic open banking datasets are uploaded into a parallel-processing unit, such as HADOOP, social data can be processed easier and achieve faster scoring results.

Statistically, the effects of social data will be explained when social network columns are added to different variations of subsets of the dataset. The effects of social data will be measured when added to financial data (see section 5.2.5.1), behavioural data (see section 5.2.5.2), and a combination of both types (see section 5.2.5.3)

5.2.5.1. On Financial Data

It is noted that the observed delinquent social ties had a p-value of 0.0066, which is smaller than the critical value of 0.05 (representing α or the level of significance). Consequently, a coefficient value of 0.012 was given to the first social network attribute (column). The relationship is considered weak. However, the relationship between social network and a probability of default is, clearly, manifested within the second social network attribute (column) when a p-value close to zero was recorded for the number of defaulted social ties at the same level of significance (α = 0.05). This aforementioned value corresponded with a higher coefficient of 0.224 for the number of defaulted social ties.

In the table of results, it is noted that, by adding only two social data columns, the interoperability increases by 0.2% (Pseudo R^2 rises from 9.5% to 9.7%). This is because the observed delinquent number of social ties and the number of defaulted social ties (i.e. the size of bad social ties) has recorded a moderate level of significance for the former and a high level of significance for the latter.

In light of the above-mentioned addition, the financial logistic regression model presented earlier in sub-section 5.2.4.3 can be adjusted by adding the term '0.012*observed delinquent social ties + 0.224*defaulting social ties' in its denominator's exponential part. The positive signs of both coefficients indicate a direct relationship between the size of a bad (delinquent and defaulting) social network and the probability of default variables, which is consistent with the convention of wisdom: the more delinquent or defaulting friends a borrower has, the higher chance one is going to default. On the other hand, the accuracy, recall, and precision did not improve when adding

social network factors to the financial traditional ones, which means that the model's predictions remained the same while being able to explain the credit scores more. The financial and social combination produced the same confusion matrix that was seen in Table 23.

Finally, when looking at the LR model run on the financial data, it is noted that financial assets and salaries are not detrimental and instead the job type and the nature of the income rather than its value are. In fact, it is noted that the variable 'income type' went from being insignificant and with zero coefficient value when no social data was used to being an explanatory variable and significantly-related with PD. This has been achieved when adding social data to the independent variables. We argue that a scalar homophily (M. Newman, 2010) happens in our sample and those who have similar life standards based on their earnings tend to, not only bond together, but also act financially in the same way.

5.2.5.2. On Behavioural Data

Adding the 2 social data columns to the existing 33 found in the behavioural dataset would give us a new dataset consisting of 35 columns. When isolating the loan outcome as a dependent variable (y), a LR model was run on the remaining 34 variables and produced the below results.

Unlike when added to financial columns, it is noted that the delinquent social ties variable is not showing any significant relationship with the outcome of the loan. This is because a p-value of the delinquent social ties is 0.48 which is well above the critical value of level of significance (α = 0.05). Therefore, observed delinquent social ties have no influence on borrowers' loan repayments whenever behavioural data is adopted. This can be explained that behavioural data can provide a more specific information on the borrowers themselves instead of relying on the behaviour of the borrower's social ties.

Conversely, there is a very significant relationship between the defaulting social ties and loan repayment. It is noted that the defaulting social ties variable had a p-value close to zero at a level of significance (α) equals to 0.05. The estimated coefficient of defaulting social ties is 0.22 and the positive sign of the coefficient reflects that the more defaulting social ties a borrower has, the higher probability of default one has.

Despite getting 1 less explaining variable when adding social data to behavioural data, in comparison with behavioural on its own, the interoperability of the model increased (Pseudo R^2 increased from 5% to 5.2%). Also, the log loss function was reduced from 14.76 to 14.62 while

accuracy precision, and F1 scores recorded an increase to 0.58 from 0.57. The only statistic that remains unchanged is the recall (sensitivity) of the model at 0.58.

Finally, a logistic regression model was run on sub-samples based on different job types from managerial to middle management to entry-level to low-income and a coefficient of 0.27 for the defaulting social ties showed a significant relationship with the outcome of loans which indicates that social relationships do affect low-income borrowers as those are influenced by their peers.

5.2.5.3. On Financial and Behavioural Data

When excluding the two social data columns from the 48 attributes (columns) that were, initially, pre-processed, the resulting data set contained financial and behavioural attributes that were modelled using a logit function and produced a Pseudo R^2 of 13.2%. 32 financial and behavioural attributes had shown significant relationships with the probability of default. Those attributes had varied levels of significance with some attributes having p-values very close to the critical value ($\alpha = 0.05$) while other attributes were well below the aforementioned critical value. For example, owning a real-estate property has a p-value of 0.0472 which considered significant with a lot of scepticism as it is very close to the critical value. On the other hand, the level of education recorded a close-to-zero p-value, which means that its relationship with the probability of default is significant (negative coefficient of - 0.128).

In order to test the effects on predicting the probability of default, the two social data attributes (columns) were added to the financial and behavioural columns and all 47 columns (loan outcome column was excluded as a dependent variable) were tested using the logistic regression model. The figure below summarises the results of the model.

Similar to the arguments of the previous two sub-section (see sections 5.2.5.1 and 5.2.5.2), the observed-delinquent-social-ties variable did not have a significant relationship with the probability of default as its p-value was well above the critical value of 0.05 (p-value = 0.93). Therefore, it is clear that the aforementioned variable reflects the behaviour of the social ties which is subordinated by the behavioural data of the borrowers themselves.

Conversely, the defaulting social ties variable presented a high level of significance at a critical value 0.05 with a p-value close to zero. The corresponding coefficient to such a significant variable was 0.176.

When comparing the results produced using financial and behavioural data with those produced

using all data including the social columns, the interoperability of the model increased by 0.1% (Pseudo R^2 rose from 13.2% to 13.3%). In addition to that, the log loss function showed a frictional improvement where the loss has slightly dropped from 14.571 to 14.569. after adding the social data. Finally, the precision of the model increased from 0.57 to 0.58 with all other scores (accuracy, recall and F1) remaining unchanged.

5.2.6. Summary Tables

When conducting a comparison between performances of the logistic regression model on different subsets of the dataset (see Table 26), it is clear that using the conventional financial data would only explain 9.5% of the variations between the probabilities of default. However, incorporating behavioural data and social network data improved interpret the scores further. A 0.2% increase was achieved by only adding the two social network columns to financial data.

More evidently, an increase in the model's precision and Pseudo R^2 was achieved when social data was added to behavioural data. The former statistic went up from 57% to 58% while the latter increased, again, by 0.2% from 5% to 5.2%. Another interesting feature is that the log loss of features is reduced by adding social data from 14.76 to 14.62.

Finally, when 45 financial and behavioural features were used, a Pseudo R^2 of 13.2% was achieved. Adding social data helped increasing that percentage to 13.3% as the model ran over 47 columns. Also, precision increased from 57% to 58%. It is worth to mention that adding 33 behavioural columns to the traditional financial data did not improve precision although it did improve accuracy, recall, and F1 scores. In contrast, adding two social network columns only to the traditional financial data did not change the model's precision, but increased the precision of the modelled behavioural data on its own and the combination of financial and behavioural data. As for the log loss, it was reduced in two out of three cases with the loss from financial and behavioural combination being frictional when adding two social network columns although not reflected in the table below because of rounding the results to the nearest hundredth.

		Financial	F + S	Behavioural	B + S	F + B	F + B + S
Predictability	**Recall**	**0.57**	**0.57**	**0.58**	**0.58**	**0.62**	**0.62**
	Precision	0.57	0.57	0.57	0.58	0.57	0.58
	F1 Score	0.57	0.57	0.57	0.58	0.58	0.58
	AUC	**0.57**	**0.57**	**0.57**	**0.58**	**0.58**	**0.58**

	Log loss	14.88	14.88	14.76	14.62	14.57	14.57
Interoperability	Pseudo R^2	9.5%	9.7%	5.0%	5.2%	13.2%	13.3%

Table 26: performances of the logistic regression model on different natures of the data

Variable Nature	Variable Name	Logistic Regression (LR) Model Coefficients					
		F	F + S	B	B + S	F + B	F + B + S
Financial	Loan type	1.09	1.06			0.55	0.54
	Gender	0.44	0.44			0.39	0.39
	Job type	-0.07	-0.07			-0.09	-0.09
	Organisation type	0.01	0.01			-	-
	Income type	-	0.13			-0.04	-0.4
	Number of documents provided	-0.13	-0.13			-0.28	-0.28
	Credit referencing agency 1	-2.03	-2.03			-2.06	-2.06
	Credit referencing agency 2	-2.17	-2.17			-2.74	-2.73
Social	Number of delinquent social ties		0.01		-		-
	Number of defaulting social ties		0.22		0.22		0.18
Behavioural	Car ownership			-0.17	-0.16	-0.29	-0.28
	Real estate ownership			0.02	-	0.04	-
	Number of children			0.10	0.10	-	-
	Companion of the applicant when applying			0.01	0.01	-	-
	Educational level			-0.18	-0.17	-0.13	-0.13
	Marital status			-	-	-0.05	-0.05
	Accommodation type			0.06	0.06	0.03	0.03
	Mobile number			-0.38	-0.38	-0.21	-0.21
	Work number			-0.20	-0.19	-0.17	-0.16
	Email			-	-	-0.13	-0.13

	Region’s rating		0.33	0.33	0.15	0.15
	Day of the week application submitted		-	-	0.01	0.01
	Time when application submitted		-0.01	-0.01	-	-
	Workplace not within living region		-0.17	-	-	-
	Workplace not within home region		-	-	-0.14	-0.13
	Living in a city far from home city		0.18	0.17	0.14	0.14
	Credit inquiries in the last year (footprints)		0.04	0.04	-	-

Figure 32: models' variables and coefficients table

The key finding of this table is that whenever social ties that form a network define as defaulting ties, the score needs to be adjusted by increasing the probability of default (PD) by 0.22 for every additional defaulting tie. A smoothing effect can be applied when traditional data is complemented by behavioural data. In such a case, an adjustment of the 0.18 for every additional defaulting tie would be ideal.

5.2.7. ROC Visualisation

The Receiver Operating Characteristics (ROC) was computed on the out-of-sample dataset to show the area under the curve (AUC) which corresponds to correct and efficient classifications. The highest ROC AUC achieved was 0.58. This curve was produced in 3 cases:

(1) When the pre-processed full dataset, made of 47 attributes, was utilised including financial, behavioural, and social network columns.
(2) When a dataset made of behavioural and social 34 columns was adopted.
(3) When a dataset made of behavioural and financial 45 columns was adopted.

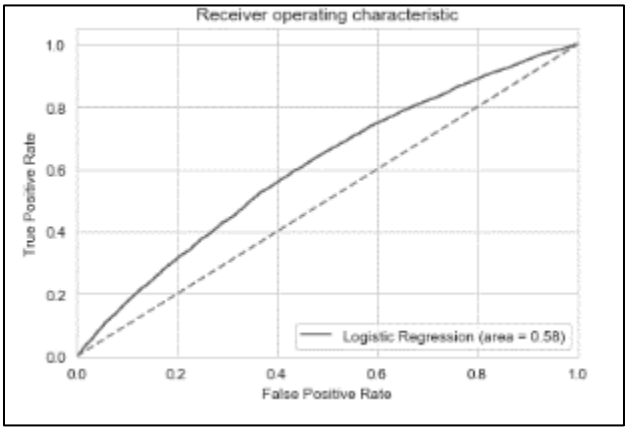

Figure 33: ROC curve for logistic regression model on full pre-processed dataset

5.3. Machine Learning

The applications of machine learning algorithms have been widely-adopted in many disciplines where classification is the main task of a business model. As part of artificial intelligence, machine learning models are characterised with the ability to improve the performance as more data points are collected. This happens as a part of an 'iterative' process. As seen in the previous part that discussed the social effects on credit scoring and predicting PD (see section 5.2). After confirming that social data represents an important element of the criteria used when assessing credit risk, machine learning algorithm was run on the complete dataset. The accuracy exceeded this of a logistic regression's, while log loss was reduced enormously as seen in the table below (see Table 27). The best performing classification algorithms were Gradient Boosting, which is consistent with the findings of Zięba et al. (2016), and Linear Discriminant Analysis (LDA).

	Traditional Scoring		Alternative Scoring	
ML Classifier	Accuracy	Log Loss	Accuracy	Log Loss
K-Nearest Neighbors	0.607	4.408	0.591	4.451

Support Vector Machine	0.664	0.614	0.668	0.609
Nu-Support Vector Machine	0.613	0.673	0.636	0.643
Decision Trees	0.581	14.46	0.583	14.405
Random Forests	0.634	0.952	0.636	0.824
AdaBoost	0.668	0.689	0.673	0.689
Extreme Gradient (XGBoost)	0.676	0.604	0.680	0.599
Naïve Bayes (Gaussian)	0.604	0.914	0.635	1.316
Linear Discriminant Analysis	0.667	0.613	0.677	0.604
Quadratic Discriminant Analysis	0.583	0.907	0.535	10.159

Table 27: machine learning comparable results

5.4. Discussion

The challenges found in traditional criteria when assessing credit risks arouse from static features which prompted dynamic scoring in modelling. Not only that, but also the limited sources of financial data providers and the dilemma of how to include the unbanked financially had led to considering alternative data sources. When alternative data was considered by non-banking lenders and P2Ps, both behavioural to social aspect were used. Based on theories discussed in behavioural finance, credit applications might be driven by a gambler's mentality or a hot-hand fallacy that assumes static persistent circumstances in the financial market and economic system.

However, this has not always been the case. Lenders started to incorporate not only behavioural measures, but also social criteria. Freedman and Jin (2017) demonstrated how Prosper's criteria changes over time. Prosper has always relied on financial data such as credit history, debt-to-income ratio, and other metrics. It introduced endorsements in early 2007 as a behavioural metric and rewarded group leaders every time a loan to any of group members is approved. However, this has prompted group leaders to act favourably to those who are socially connected to them. As a result, the screening process was inadequate. They concluded that social network attributes replaced behavioural metrics later that year - Sept. 2007.

Social networks, whether derived from social media platforms, community blogs, telecommunication, electronic interactions, social clubs, sports clubs, literature clubs, arts clubs, or academic citations provide insights (Correa et al., 2010) on similarities, known as homophily, and reveal common beliefs when it comes to financial responsibility (Wei et al., 2015).

When examining some social network structures that are seen in Figure 6, one can identify individuals full of bright ideas. Lenders should seek such borrowers whenever loan purpose is for starting a new business or in developing a product within an innovative sector. Also, whenever a borrower is located in a network that is influenced by a good borrower, one's credit score should

be improved and vice versa. An example of such a network can be seen in graph (b) of the same figure. Efficient and innovative social network structures can be advantageous in SME lending whenever loan is requested for projects or research and development respectively. Therefore, social network data is of a high value in those cases.

The idea of social networks not only provides the lender of a contextual understanding of the borrowers' environment, but also a dynamic dimension. In fact, as borrowers socialise and interact with each other, communities become clearer. This would overcome the problem discussed with bankers in Qatar where updating the credit scores takes place manually every year and a whole review happens in retrospect. Instead, a score would adjust whenever a borrower moves closer to the centre of a social network or whenever one defects from a community and joins another one. In addition to that, it is possible to run of temporal network analysis and find persistent connections over time to derive homophily. The use of temporal networks is suggested within the future agenda section (see section 6.3)

It is understood that social lending found in peer-to-peer platforms (Pokorná & Sponer, 2016) which facilitates the formation of social networks. Also, for lenders that have access to open banking data, the trade and transfer data can be extracted to for social network. Other sources can be social media. However, it is important that the information found on social media is somehow verified. Therefore, a professional social network like LinkedIn can be adopted by banks seeking to implement social network analysis with the least costs. In LinkedIn, the network is, primarily, 'undirected' i.e. connections between two individuals, each visualised in the LinkedIn network in a node, represents a symmetric relationship. Both individuals acknowledge the friendship and the professional relationship (i.e. connection). However, it is possible to extract information on who requested the connection in LinkedIn as a way to estimate the degree of homophily and whether a person is enforcing a connection to attain a strategic goal or merely because one has mutual interests with another person. Also, LinkedIn's platform has another type of 'directed' networks where LinkedIn members follow public figures. In such a case, the personality of perspective borrowers can be analysed by applying the singular value decomposition techniques found in the work of Kosinski et al. (2013).

In credit scoring, selecting individuals from a social network should be based on the type of connection with others in the network. As such, eliminating strategic connections, i.e. those aiming to improve credit scores through connecting with higher types (Wei et al., 2015), is permissible.

On the other hand, there should be an emphasis on those connections with high degree of homophily. By estimating the adjacency to the central node, an individual's social effects can be incorporated within the credit scoring model.

Based on the results of the hypothesis tests, it is noted that the sizes of both social network types, delinquents and defaulters, happen to be larger with borrowers who defaulted on their loans than those who did not. Initially, this indicated some relationship between both social network types and PD (or a credit risk score). However, the results of the Bayesian analysis and Logistic regression model confirmed only the significance of one of the types – the defaulting social network with PDs (or credit risk scores) while the delinquent social network did not reveal a significant relationship with PDs (or credit risk scores). The reason of this discrepancy is believed to be because of the set-up of the hypothesis test. The dependent variable in the Mann-Whitney U-test was set to be the social network size because the independent variable has to be binary (the outcome of loans) in order to compare the groups. In other words, the hypothesis test may infer that when borrowers default on their loans, those start to socialise with peers who are already in arrears and suffering from delinquency regardless of whether those eventually manage to repay the loan of not.

CHAPTER 6: CONCLUSION

It is concluded that having a network of delinquent social ties does not provide sufficient evidence regarding a borrower's loan outcome. However, when those observed social ties become actual defaulters, influence become obvious and those with high number of defaulting social ties tend to have a higher probability of default. This is because the observed delinquency could be incidental and not persist over time. On the other hand, a default is more likely to happen after following a pattern which would, eventually, prevail, thus, defaulting social ties can inform one's credit score. This is because the defaulting social ties variable has been significant in every case and it has been proven that having defaulting friends and family in a borrower's network correlates positively with the probability of default and, accordingly, negatively with one's credit score.

Furthermore, using social network data solely has not been infeasible due to its weak inference and classification powers. Meanwhile, by using only two social network variables, performance of credit scoring models was improved. This was manifested when the results of credit scoring models were explained using the social network data and, in some cases, were more powerful in classifying prospective borrowers.

As a result, it is believed that analysing borrowers' social circles and network types helps in understanding a score and, sometimes, in classifying a borrower more correctly. Furthermore, changes in those social networks over time (temporal dimension) and types of ties (adding many layers) is strongly recommended when evaluating borrowers' ability to repay a loan. Adding the aforementioned data would, essentially, improve credit scoring.

6.1. Summary of Findings

In this section, a theoretical framework and empirical results are discussed then conceptualised. A clear theoretical analysis was discussed based on the five-factor model. It was clear that introverts do not prefer to socialise with people and, when they do, they are less influenced by people' opinion. Instead, their actions are inspired by knowledge and reading. Therefore, a sophisticated model that investigates traditional financial variables of an introvert would suffice the credit scoring process. We call this process 'traditional credit scoring'.

Meanwhile, analysing a highly-neurotic personality in terms of behaviour would indicate someone's financial state-of-mind. A lender may conclude that such a personality will worry about consequences and there would be no need to look into the social network. Similarly, a

conscientious borrower does not give importance to the social network regardless of its size and type. Accordingly, running a behavioural model that reveal such a personality would be beneficial for credit risk assessors where conscientiousness is preferable to spontaneity. We call this process 'behavioural credit scoring'.

The interesting case of an openness trait requires that, based on the degree of how open a borrower is, matching jobs with aforementioned degree would guarantee a steady cashflow and income as a quantified form of job stability.

Finally, in agreeableness and extroversion, social networks play an important role in predicting the credit and financial outcomes of individuals with those traits. The aforementioned summary will be represented in the following section (see 6.2) in a decision-tree diagram (see Figure 36)

6.2. Implications and Contribution

The first implication of this research is its distinctive treatment of social network data by separating it from other behavioural metrics such as owning a car or living away from work or activity on social media platforms. This has explicitly paved the way for researchers and practitioners to rely on different models according to what type of alternative data the lender has. Ideally, lenders should identify whether the data reflects someone's personality and tastes or a borrower's assertive selection of individuals that one likes to mix with and be interactive with. The figure below (see Figure 34) is an extension to the figure presented in chapter 2 (refer to Figure 3)

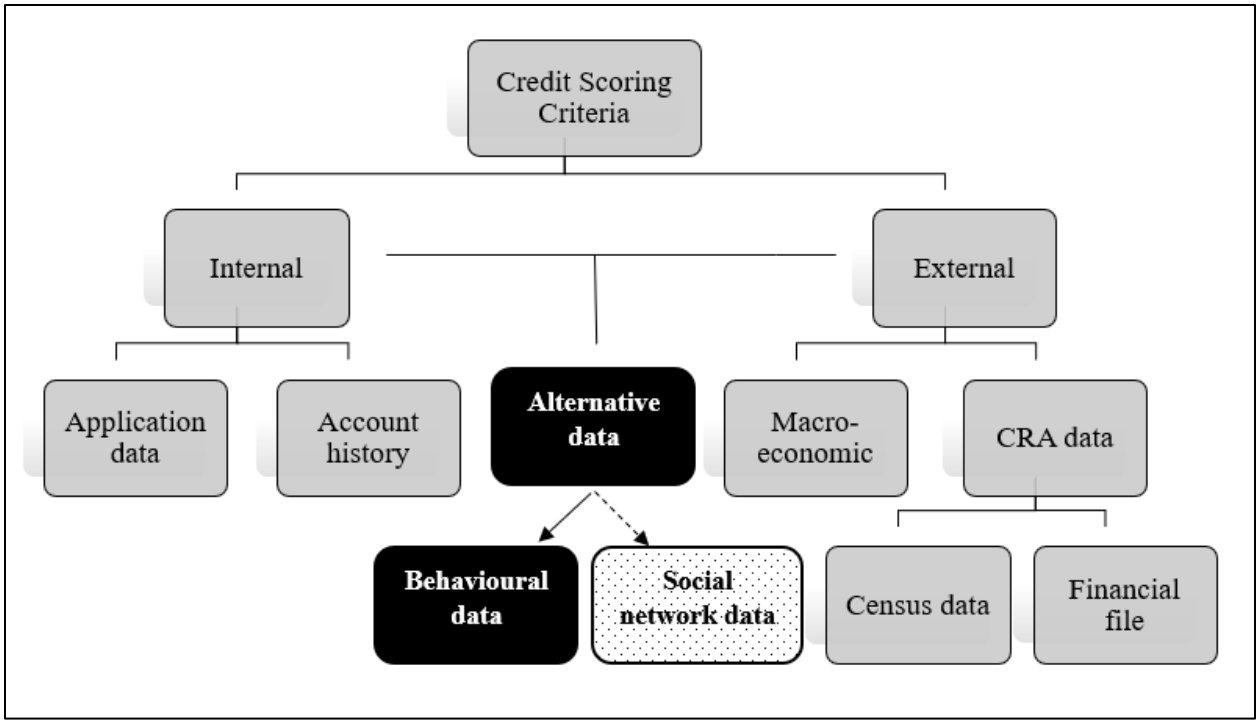

Figure 34: extension to credit scoring criteria

In credit, the types of borrowers discussed earlier in Chapter 4 (see Figure 19) can be predicted using social network analysis. Particularly, transactors[8] and revolvers (whether delinquent[9] or defaulters) would create a network cluster each and those can be identified using communities' detection algorithms. However, since the results and findings did not find a significant inference of delinquency on credit scores, those with arrears may end up in either community – transactors or defaulters. On another note, marketing teams that work for lenders aim at expanding the loan market share of their lending firms. Similarly, they are interested in identifying those borrowers who would churn[10] in order to try and reach out before defection happens.

[8] Borrowers who pay on time.
[9] Borrowers who have arrears (severity calculated using amounts and days-past-due).
[10] Move with a competitor.

Figure 35 is a simple visualisation of how using social network analysis can be used to define different groups of borrowers such as transactors, churning borrowers, and defaulters by identifying the most influential nodes. In response, lenders would attract, retain, and avoid those respectively.

For lenders, the question remains whether or not it is feasible to implement a social network analysis and to evaluate borrowers' credit scores.

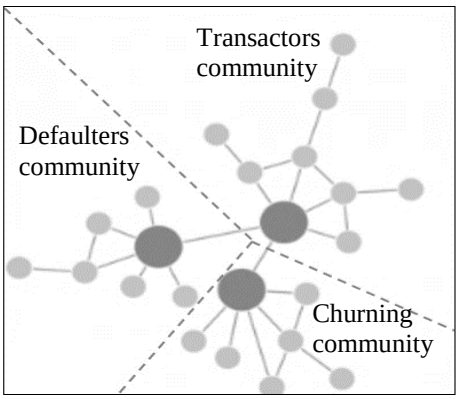

Figure 35: detection of borrowers' communities using social network analysis

The diagram below (see Figure 36) depicts the findings of this research and how social networks type would affect credit scores. As discussed in the findings of the dataset earlier in this chapter (see section 5.2), the types of social network do have an effect on credit scores given the more information on the types of behaviour.

As lenders adopt more sophisticated systems and models, the decision can be made based on the type of network and the behaviour of a borrower. For example, suppose that a network size is small and a borrower is highly-organised (high in conscientiousness) or is an introvert, relying on traditional financial data using a dynamic model would be sufficient.

In the case of a larger social network where a borrower is placed in the ego centre demonstrating a highly-influential position or if someone is defined as an artistic or an adventurous person who is open to new experiences (high in openness), relying on behavioural data becomes essential and the associated behavioural credit score scoring.

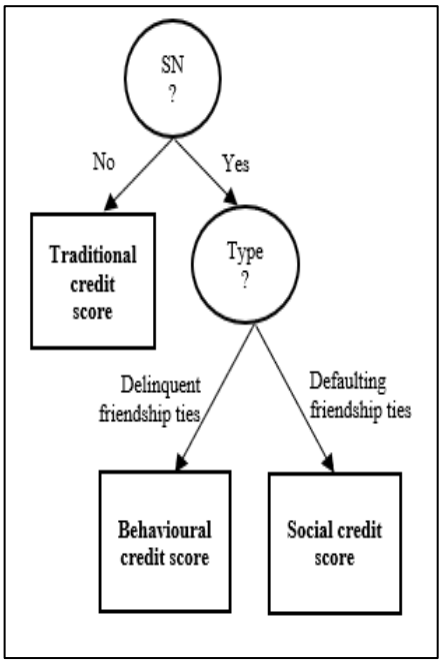

Figure 36: results conceptualised

Finally, whenever a borrower is part of a large social network and one is characterised with being agreeable, extrovert or a highly-neurotic person, social network analysis needs to be incorporated in the credit risk scoring as part of a social modelling contribution. In such a case, incorporating the statistics of the network (size, type, centrality, in/out-degree, etc.) would define the influence of the network on the credit risk score.

Whenever social ties that form a network define as defaulting ties, the score needs to be adjusted by increasing the probability of default (PD). As for the other social network type – delinquent ties, it does not have a direct effect on credit scores, but it is vital in such a case for a lender to assess the borrower's behavioural factors as well and not be limited to traditional ones.

6.3. Limitations and Future Agenda

In this research, the exploratory nature and the developing case of using social network analysis in financial modelling limited the number of interviews. Since the adoption of such a trend is a decision likely to be made by senior management members, the availability of such members and the limited time they could offer resulted-in a small sample size when it came to qualitative analysis. Moreover, the unstructured-interviewing style/format deterred the comparability of the results of interviews to a certain extent.

The main challenge of this research is the implementation of social scoring in developed countries in light of the privacy and confidentiality rules and regulations such as GDPR. However, when financial exclusion causes individuals to be denied an access to fund their business, those would give the lender consents to analyse the number and types of connections as well as the type of network structure they belong to.

In the case of looking at individual's financial interactions with others, open banking initiative has made it possible, theoretically, to track payment between individuals. Nevertheless, lenders need to notify their borrowers and their social circle that their financial transfers will be used for credit scoring purposes. Finally, social media platforms carry information on social networks ties and friendships and those offer APIs that collect data after notifying the user of what data is being collected and for what purpose. Some platforms, unfortunately, prohibit collecting data for credit rating purposes.

As for the social network columns found in the data set while performing quantitative analyses, apart from the definition of the columns, those lacked explicit details on how the data was collected (source and mechanism) as well as other relevant information. For example, the number of defaulting social ties is known; however, the total number of social ties is unknown and, thus, the density and proportion of the bad social influence within a network are yet to be determined. Also, the nature of those social ties is not defined. Moreover, the information on whether those social network ties are part of directed or undirected network is missing from the description.

In addition to that, there is no indication of any interaction between the observations of the dataset.

In other words, we were unable to determine whether those 300,000 borrowers belong to a certain community or social network let alone knowing if they interact with each other. This could have affected our methodology when conducting hypothesis testing as the decision was to consider the observations independent from each other subjectively.

The above-mentioned limitations would present researchers with a motivation for future research where networks can be formed in a graphical representation highlighting those nodes who happen to be bad as opposed to the normal financially-responsible social ties. Therefore, a research on influential social networks in credit can be an extension to this study. Defining such networks would indicate types of edges as well as other statistics (size, density, structure, adjacency, etc.).

Also, monitoring those ties across time would add a temporal layer of the analysis where dynamic social network scores are used. Furthermore, performing multi-layer network analyses between nodes would estimate the strength of an edge (or a bond in real-life terms) between two people. For example, when information on networks of financial transactions, common club memberships, friendship on social media, similar spatial/geo-locational data, and matching careers, credit model are available, a multi-layer network analysis can be performed. For example, the efforts of Lin et al. (2013) in classifying networks into alumni, medical, demographic, etc. can be extended by running a multi-layer social network analysis on the borrower. In such case, knowing the types, strengths, and sizes of one's networks would inform the outcome of the financial activity based on social ties.

Finally, in the case of abundant social data and networks, it is worth trying to run the model completely on social network data and see if lenders can sustain lending to those with no financial and behavioural records.

The works of Wei et al. (2015) and Arráiz et al. (2017) asserted that economies would benefit from implementing a social-network-based solution as an alternative tool to assess its individuals' credit risk. However, the infrastructure and the database management in developing countries remain a challenge for implementing alternative credit scoring systems. In addition to that, under-developed countries may follow the European Union in implementing a restrictive data sharing policy to protect the identity of its nationals similar to GDPR, so data localization and warehousing rules may hinder the implementation. In conclusion, it is believed that such a tool should be implemented in parallel with the central bank's vision of those countries.

References

Ackert, L., & Deaves, R. (2009). *Behavioral finance: Psychology, decision-making, and markets*: Cengage Learning.

Ahn, Y.-Y., Bagrow, J. P., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *nature, 466*(7307), 761.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics, 74*(1), 47.

Amelio, A., & Pizzuti, C. (2014). Overlapping community discovery methods: a survey *Social Networks: Analysis and Case Studies* (pp. 105-125): Springer.

Arráiz, I., Bruhn, M., & Stucchi, R. (2017). Psychometrics as a Tool to Improve Credit Information, S67.

Ascarza, E., Ebbes, P., Netzer, O., & Danielson, M. (2017). Beyond the target customer: Social effects of customer relationship management campaigns. *Journal of Marketing Research, 54*(3), 347-363.

Bachrach, Y., Kosinski, M., Graepel, T., Kohli, P., & Stillwell, D. (2012). *Personality and patterns of Facebook usage*. Paper presented at the Proceedings of the 4th Annual ACM Web Science Conference.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., & Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society, 54*(6), 627-635.

Banasik, J., Crook, J. N., & Thomas, L. C. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, 1185-1190.

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology, 44*(1), 1-26.

Bellotti, T., & Crook, J. (2009a). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society, 60*(12), 1699-1707.

Bellotti, T., & Crook, J. (2009b). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications, 36*(2), 3302-3308.

Berndt, A., & Gupta, A. (2009). Moral hazard and adverse selection in the originate-to-distribute model of bank credit. *Journal of Monetary Economics, 56*(5), 725-743.

Birch, D. (2018). Forget banks. In 2018, you will pay through Amazon and Facebook. Retrieved from https://www.wired.co.uk/article/banks-data-tech-giants

Bjorkegren, D., & Grissen, D. (2015). Behavior revealed in mobile phone usage predicts loan repayment.

Blumberg, B. F., & Letterie, W. A. (2008). Business starters and credit rationing. *Small business economics, 30*(2), 187-200.

BMJ. (2020). Covid-19: medical students to be employed by NHS as part of epidemic response. Retrieved from https://www.bmj.com/content/368/bmj.m1156

Boyle, M. (1992). Methods for credit scoring applied to slow payers, Credit Scoring and Credit Control,(edited by L. Thomas, J. Crook and D. Edelman): Oxford University Press.

Bradbury, D. (2011). Data mining with LinkedIn. *Computer Fraud & Security, 2011*(10), 5-8.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brockett, P. L., & Golden, L. L. (2007). Biological and psychobehavioral correlates of credit scores and automobile insurance losses: Toward an explication of why credit scoring works. *Journal of Risk and Insurance, 74*(1), 23-63.

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications, 39*(3), 3446-3453.

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable Machine Learning in

Credit Risk Management. *Computational Economics*, 1-14.
CallCredit. (2008). *APACS Data Integration*. Retrieved from
Chattopadhyay, S., Basu, T., Das, A. K., Ghosh, K., & Murthy, L. C. (2020). Towards effective discovery of natural communities in complex networks and implications in e-commerce. *Electronic Commerce Research*, 1-38.
Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk. *arXiv preprint arXiv:1811.12615*.
Chen, F.-L., & Li, F.-C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert Systems with Applications, 37*(7), 4902-4909.
Chen, P. P.-S. (1976). The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS), 1*(1), 9-36.
Chen, X., Zhou, L., & Wan, D. (2016). Group social capital and lending outcomes in the financial credit market: An empirical study of online peer-to-peer lending. *Electronic Commerce Research and Applications, 15*, 1-13.
cignifi.com. Retrieved from cignifi.com
Correa, T., Hinsley, A. W., & De Zuniga, H. G. (2010). Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior, 26*(2), 247-253.
Coscia, M., Giannotti, F., & Pedreschi, D. (2011). A classification for community discovery methods in complex networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal, 4*(5), 512-546.
Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research, 183*(3), 1447-1465.
Currier, J. Revealing the Unseen Forces That Guide Your Life. *Your Life and Network Effects*. Retrieved from https://www.nfx.com/post/your-life-network-effects/
Dash, R., Kremer, A., Nario, L., & Waldron, D. (2017). Risk analytics enters its prime. Retrieved from http://www.mckinsey.com/business-functions/risk/our-insights/risk-analytics-enters-its-prime?cid=other-eml-alt-mip-mck-oth-1706&hlkid=fab1a4e9e2c6433f95cba97c2560136a&hctky=9492246&hdpid=fe6acf27-aaf4-4041-b1ad-aaeb679b6e98
Data-Driven Decisions for Financial Services. (2018). In LenddoEFL (Ed.).
De Koker, L., & Jentzsch, N. (2013). Financial inclusion and financial integrity: Aligned incentives? *World development, 44*, 267-280.
Desai, V. S., Conway, D. G., Crook, J. N., & Overstreet Jr, G. A. (1997). Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics, 8*(4), 323-346.
EFLGlobal. About. Retrieved from https://www.eflglobal.com/about/
EFLGlobal. (2017). Lenddo and EFL Team Up Lead Financial Inclusion Revolution. Retrieved from https://www.eflglobal.com/lenddo-efl-team-lead-financial-inclusion-revolution/
Erdem, T., & Keane, M. P. (1996). Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science, 15*(1), 1-20.
Experian. (2009). *Delphi for Customer Management*. Retrieved from
Experian.co.uk. (2013). *Risk-based pricing: When does it work and does it not?* Retrieved from
EY. (2018). *IFRS 9 Expected Credit Loss*. Retrieved from
Fahner, G. (2018). Developing Transparent Credit Risk Scorecards More Effectively: An Explainable Artificial Intelligence Approach. *Data Anal, 2018*, 17.
Fama, E. F. (1991). Efficient market hypothesis. *The Journal of Finance, 46*, 383-417.
FCA. (2017). *Preventing Financial Distress by Predicting Unaffordable Consumer Credit Agreements: An Applied Framework*. Retrieved from July 2017:

FCA. (2018). *Assessing creditworthiness in consumer credit*. UK: Financial Conduct Authority.
Finextra.com. (2020). WHO Urges to Switch to Contactless to Slow Virus Transmission. Retrieved from https://www.finextra.com/newsarticle/35384/who-urges-switch-to-contactless-to-slow-virus-transmission?
Fitzgerald, R. (2018). How LenddoEFL Uses Data and Personality Analyses to Increase Access to Financial Services in Emerging Economies. Retrieved from https://www.cardrates.com/news/lenddoefl-helps-emerging-economies-access-financial-services/
Freedman, S., & Jin, G. Z. (2017). The information value of online social networks: lessons from peer-to-peer lending. *International Journal of Industrial Organization, 51*, 185-222.
Gargiulo, T. L., Pangarkar, A., Kirkwood, T., & Bunzel, T. (2006). *Building Business Acumen for Trainers: Skills to Empower the Learning Function*: John Wiley & Sons.
Gärling, T., Michaelsen, P., & Gamble, A. (2020). Young adults' borrowing to purchases of desired consumer products related to present-biased temporal discounting, attitude towards borrowing and financial involvement and knowledge. *International Journal of Consumer Studies, 44*(2), 131-139. doi:10.1111/ijcs.12552
General Data Protection Regulation, (2016).
Gonzalez, L., & Loureiro, Y. K. (2014). When can a photo increase credit? The impact of lender and borrower profiles on online peer-to-peer loans. *Journal of Behavioral and Experimental Finance, 2*, 44-58. doi:http://dx.doi.org/10.1016/j.jbef.2014.04.002
Gordy, M. B. (2000). A comparative anatomy of credit risk models. *Journal of Banking & Finance, 24*(1-2), 119-149.
Guo, G., Zhu, F., Chen, E., Liu, Q., Wu, L., & Guan, C. (2016). From footprint to evidence: an exploratory study of mining social data for credit scoring. *ACM Transactions on the Web (TWEB), 10*(4), 1-38.
Guttentag, J., & Herring, R. (1984). Credit rationing and financial disorder. *The Journal of Finance, 39*(5), 1359-1382.
Ha, S. H., & Krishnan, R. (2012). Predicting repayment of the credit card debt. *Computers & Operations Research, 39*(4), 765-773.
Hagenau, M., Liebmann, M., & Neumann, D. (2013). Automated news reading: Stock price prediction based on financial news using context-capturing features. *Decision Support Systems, 55*(3), 685-697.
Hamilton, S., & Francis, I. (2003). *The Enron collapse*: International Institute for Management Development.
Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.
Hardy Jr, W. E., & Adrian Jr, J. L. (1985). A linear programming alternative to discriminant analysis in credit scoring. *Agribusiness, 1*(4), 285-292.
Hayes, A. (2019). FICO Score. *Credit & Debt*. Retrieved from https://www.investopedia.com/terms/f/ficoscore.asp
He, J., Liu, X., Shi, Y., Xu, W., & Yan, N. (2004). Classifications of credit cardholder behavior by using fuzzy linear programming. *International Journal of Information Technology & Decision Making, 3*(04), 633-650.
Heidhues, P., & Kőszegi, B. (2010). Exploiting naivete about self-control in the credit market. *American Economic Review, 100*(5), 2279-2303.
Henegar, J. M., Archuleta, K. L., Grable, J., Britt, S. L., Anderson, N., & Dale, A. (2013). Credit card behavior as a function of impulsivity and mother's socialization factors. *Journal of Financial Counseling and Planning, 24*(2), 37-49.
Henley, W. E. (1995). *Statistical aspects of credit scoring*. The Open University.
Hirsh, J. B., & Peterson, J. B. (2009). Extraversion, neuroticism, and the prisoner's dilemma.

Personality and individual differences, 46(2), 254-256.
Hirshleifer, D. (2015). Behavioral finance. *Annual Review of Financial Economics, 7*, 133-159.
HM-Treasury. (2020). Support for those affected by COVID-19. Retrieved from https://www.gov.uk/government/publications/support-for-those-affected-by-covid-19/support-for-those-affected-by-covid-19
The home of NodeXL | Your Social Network Analysis Tool for Social Media. Retrieved from https://www.smrfoundation.org
HSBC. (2018). *Guide to Credit Scoring, Credit Referencing and Fraud Prevention Agencies*. Retrieved from
Hsieh, N.-C. (2005). Hybrid mining approach in the design of credit scoring models. *Expert Systems with Applications, 28*(4), 655-665.
Huang, C.-L., Chen, M.-C., & Wang, C.-J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications, 33*(4), 847-856.
ICTQatar. (2014). *Qatar's ICT Landscape 2014*. Retrieved from
ICTQatar. (2015). *Qatar's National ICT Plan*. Retrieved from
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.
Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial finance, 18*(6), 15-26.
Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological bulletin, 127*(3), 376.
Karlan, D., & Zinman, J. (2009). Observing unobservables: Identifying information asymmetries with a consumer credit field experiment. *Econometrica, 77*(6), 1993-2008.
klarna.com. About us - Klarna UK. Retrieved from https://www.klarna.com/uk/about-us/
Klinger, B., Khwaja, A., & LaMonte, J. (2013). *Improving credit risk analysis with psychometrics in Peru*. Retrieved from
Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences, 110*(15), 5802-5805.
Kou, G., & Wu, W. (2014). An analytic hierarchy model for classification algorithms selection in credit risk analysis. *Mathematical Problems in Engineering, 2014*.
Kruger, J., & Burrus, J. (2004). Egocentrism and focalism in unrealistic optimism (and pessimism). *Journal of Experimental Social Psychology, 40*(3), 332-340.
Kruppa, J., Schwarz, A., Arminger, G., & Ziegler, A. (2013). Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications, 40*(13), 5125-5131.
Kshetri, N. (2016). Big data's role in expanding access to financial services in China. *International Journal of Information Management, 36*(3), 297-308.
Kumire, J. (2019). Open Banking in the UK: what's happened so far. Retrieved from https://pulse.11fs.com/research-reports/2019/08/open-banking-in-the-uk-whats-happened-so-far
Lazarow, A. (2017). *What do emerging market consumers expect from InsureTech?* Retrieved from
Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245-254.
Leonardi, P., & Contractor, N. (2018). Better People Analytics. *Analytics*. Retrieved from https://hbr.org/2018/11/better-people-analytics
Leow, M., & Crook, J. (2016). A new Mixture model for the estimation of credit card Exposure at Default. *European Journal of Operational Research, 249*(2), 487-497.
Lin, M., Prabhala, N. R., & Viswanathan, S. (2013). Judging borrowers by the company they keep:

Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science, 59*(1), 17-35.
Liu, Y., & Schumann, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society, 56*(9), 1099-1108.
Lont, H. (2001). Negotiating Financial Autonomy: Women, Income and Credit in Urban Java1. *Women and Credit*, 203.
Lü, L., & Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: statistical mechanics and its applications, 390*(6), 1150-1170.
Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications, 42*(10), 4621-4631.
Malhotra, R., & Malhotra, D. K. (2003). Evaluating consumer loans using neural networks. *Omega, 31*(2), 83-96.
Malik, M., & Thomas, L. C. (2010). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society, 61*(3), 411-420.
Manski, C. F. (1993). Identification of endogenous social effects: The reflection problem. *The review of economic studies, 60*(3), 531-542.
Manski, C. F., & Straub, J. D. (1999). *Worker perceptions of job insecurity in the mid-1990s: Evidence from the Survey of Economic Expectations*. Retrieved from
Masyutin, A. (2015). Credit scoring based on social network data. *Бизнес-информатика*(3 (33)).
McBurnett, M. T., Matthew. In E. Corporation (Ed.).
McEvoy, M. J., & Chakraborty, T. (2014). Enabling financial inclusion through alternative data.
McKillop, D. G., Ward, A.-M., & Wilson, J. O. (2007). The development of credit unions and their role in tackling financial exclusion. *Public money and management, 27*(1), 37-44.
Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. *Standford University, CS229 (2011 http://cs229. stanford. edu/proj2011/GoelMittal-StockMarketPredictionUsingTwitterSentimentAnalysis. pdf), 15*.
Muñoz de Bustillo, R., & De Pedraza, P. (2010). Determinants of job insecurity in five European countries. *European Journal of Industrial Relations, 16*(1), 5-20.
Newman, M. (2010). Networks: an introduction: Oxford University Press, Oxford.
Newman, M. E., Strogatz, S. H., & Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E, 64*(2), 026118.
Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review, 45*(2), 167.
Nye, R. P. (2014). *Understanding and Managing the Credit Rating Agencies*. [N.p.]: Euromoney Books.
Oakley, P. (2018). Revoult Uncovers Potential Money-Laundering Activity on Its Platform. Retrieved from http://www.techx365.com/author.asp?utm_source=SendPulse&utm_medium=push&utm_campaign=1473426&section_id=605&doc_id=744784&f_src=techx365_sitedefault
Omar, N. A., Rahim, R. A., Wel, C. A. C., & Alam, S. S. (2014). Compulsive buying and credit card misuse among credit card holders: The roles of self-esteem, materialism, impulsive buying and budget constraint. *Intangible Capital, 10*(1), 52-74.
Ong, C.-S., Huang, J.-J., & Tzeng, G.-H. (2005). Building credit scoring models using genetic programming. *Expert Systems with Applications, 29*(1), 41-47.
Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., & Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing, 74*, 26-39.
Óskarsdóttir, M., Bravo, C., Verbeke, W., Sarraute, C., Baesens, B., & Vanthienen, J. (2017).

Social network analytics for churn prediction in telco: Model building, evaluation and network architecture. *Expert Systems with Applications, 85*, 204-220.
Otero-López, J. M., & Villardefrancos, E. (2013). Five-Factor Model personality traits, materialism, and excessive buying: A mediational analysis. *Personality and individual differences, 54*(6), 767-772.
Peluso, S., Mira, A., & Muliere, P. (2015). Reinforced urn processes for credit risk models. *Journal of econometrics, 184*(1), 1-12.
Peón, D., Antelo, M., & Calvo, A. (2016). Overconfidence and risk seeking in credit markets: an experimental game. *Review of Managerial Science, 10*(3), 511-552.
Pikkarainen, T., Pikkarainen, K., Karjaluoto, H., & Pahnila, S. (2004). Consumer acceptance of online banking: an extension of the technology acceptance model. *Internet Research*.
Pokorná, M., & Sponer, M. (2016). Social Lending and Its Risks. *Procedia - Social and Behavioral Sciences, 220*, 330-337. doi:10.1016/j.sbspro.2016.05.506
Potì, V., & Siddique, A. (2013). What drives currency predictability? *Journal of International Money and Finance, 36*, 86-106.
Puri, M., Rocholl, J., & Steffen, S. (2017). What do a million observations have to say about loan defaults? Opening the black box of relationships. *Journal of Financial Intermediation, 31*, 1-15.
Rabin, M., & Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The review of economic studies, 77*(2), 730-778.
Rajan, U., Seru, A., & Vig, V. (2010). Statistical default models and incentives. *The American Economic Review, 100*(2), 506-510.
Rasmussen, R. K. (1998). Behavioral Economics, the Economic Analysis Bankruptcy Law and the Pricing of Credit. *Vand. L. Rev., 51*, 1679.
Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., & Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *science, 297*(5586), 1551-1555.
Ray, S. (2018). Building your First Neural Network on a Structured Dataset (using Keras). Retrieved from https://medium.com/analytics-vidhya/build-your-first-neural-network-model-on-a-structured-dataset-using-keras-d9e7de5c6724
Redrup, Y. (2017). How email and smartphone data could help you get a loan. *Technology*. Retrieved from http://www.afr.com/technology/how-email-and-smartphone-data-could-help-you-get-a-loan-20171212-h02zi0#ixzz534zFfQmg
Reynolds, F., & Chidley, M. *Consumer priorities for open banking*. Retrieved from
Robinson, M. D., Ode, S., Moeller, S. K., & Goetz, P. W. (2007). Neuroticism and affective priming: Evidence for a neuroticism-linked negative schema. *Personality and individual differences, 42*(7), 1221-1231.
Rotchanakitumnuai, S., & Speece, M. (2003). Barriers to Internet banking adoption: a qualitative study among corporate customers in Thailand. *International Journal of Bank Marketing*.
Roth, P. L., Bobko, P., Van Iddekinge, C. H., & Thatcher, J. B. (2016). Social media in employee-selection-related decisions: A research agenda for uncharted territory. *Journal of Management, 42*(1), 269-298.
Rusli, E. M. (2013). Bad credit? Start tweeting. *WALL ST. J., Apr, 1*.
Schuermann, T. (2004). What do we know about loss given default?
Serrano-Cinca, C., & Gutiérrez-Nieto, B. (2016). The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending. *Decision Support Systems, 89*, 113-122.
Shi, Y., Peng, Y., Xu, W., & Tang, X. (2002). Data mining via multiple criteria linear programming: applications in credit card portfolio management. *International Journal of Information Technology & Decision Making, 1*(01), 131-151.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *The American Economic Review, 49*(3), 253-283.
Soni, D. (2018). Introduction to Markov Chains. *KDnuggets Blog*. Retrieved from https://www.kdnuggets.com/2018/03/introduction-markov-chains.html
Sousa, M. R., Gama, J., & Brandão, E. (2016). A new dynamic modeling framework for credit risk assessment. *Expert Systems with Applications, 45*, 341-351.
Srinivasan, V., & Kim, Y. H. (1987). Credit granting: A comparative analysis of classification procedures. *The Journal of Finance, 42*(3), 665-681.
Stango, V., & Zinman, J. (2006). How a cognitive bias shapes competition: Evidence from consumer credit markets. *Dartmouth College, Tuck School of Business*.
Steenackers, A., & Goovaerts, M. (1989). A credit scoring model for personal loans. *Insurance: Mathematics & Economics, 8*(1), 31-34.
Šušteršič, M., Mramor, D., & Zupan, J. (2009). Consumer credit scoring models with limited data. *Expert Systems with Applications, 36*(3), 4736-4744.
Taffler, R. (2017). Emotional finance: investment and the unconscious. *The European Journal of Finance*, 1-30.
Thomas, L., Banasik, J., & Crook, J. (2001). Recalibrating scorecards. *Journal of the Operational Research Society, 52*(9), 981-988.
Thomas, L. C. (2000). A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting, 16*(2), 149-172.
Thomas, L. C. (2009). *Consumer Credit Models: Pricing, Profit and Portfolios: Pricing, Profit and Portfolios*: OUP Oxford.
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2002). *Credit scoring and its applications*: SIAM.
Thomas, L. C., Edelman, D. B., & Crook, J. N. (2017). *Credit scoring and its applications* (Second edition. ed.): Society for Industrial and Applied Mathematics.
Tong, E. N., Mues, C., Brown, I., & Thomas, L. C. (2016). Exposure at default models with and without the credit conversion factor. *European Journal of Operational Research, 252*(3), 910-920.
Tong, E. N., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research, 218*(1), 132-139.
TransUnion. (2019). CallReport API Reference Guide.
Tsai, C.-F., & Wu, J.-W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34*(4), 2639-2649.
Turner, M., & McBurnett, M. (2019). Optimizing neural networks for risk assessment: Google Patents.
Wade, M. R., Shan, J., & McTeague, L. Strategies for responding to digital disruption. Retrieved from https://www.imd.org/research/insightsimd/strategies-for-responding-to-digital-disruption2/
Wan, C., Peng, S., Wang, C., & Yuan, Y. (2016). *Communities Detection Algorithm Based on General Stochastic Block Model in Mobile Social Networks*. Paper presented at the Advanced Cloud and Big Data (CBD), 2016 International Conference on.
Wang, G., Ma, J., Huang, L., & Xu, K. (2012). Two credit scoring models based on dual strategy ensemble trees. *Knowledge-based systems, 26*, 61-68.
Wang, L., Lu, W., & Malhotra, N. K. (2011). Demographics, attitude, personality and credit card features correlate with credit card debt: A view from China. *Journal of Economic Psychology, 32*(1), 179-193.
Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications, 27*, 74-82.

Wei, Y., Yildirim, P., Van den Bulte, C., & Dellarocas, C. (2015). Credit scoring with social network data. *Marketing Science, 35*(2), 234-258.
Weke, P., & Ntwiga, D. B. (2016). Consumer lending using social media data.
West, D. (2000). Neural network credit scoring models. *Computers & Operations Research, 27*(11), 1131-1152.
What's in my FICO scores. Retrieved from http://www.myfico.com/credit-education/whats-in-your-credit-score/
What is Comprehensive Credit Scoring (CCR)? Retrieved from https://ccr.equifax.com.au/what-is-ccr
What is Logistic Regression. Retrieved from http://www.statisticssolutions.com/what-is-logistic-regression/
Wood, M. (1998). Socio-economic status, delay of gratification, and impulse buying. *Journal of Economic Psychology, 19*(3), 295-320.
Xia, Y., Liu, C., & Liu, N. (2017). Cost-sensitive boosted tree for loan evaluation in peer-to-peer lending. *Electronic Commerce Research and Applications, 24*, 30-49.
Yan, J., Yu, W., & Zhao, J. L. (2015). How signaling and search costs affect information asymmetry in P2P lending: the economics of big data. *Financial Innovation, 1*(1), 1.
Yang, J., & Leskovec, J. (2012). *Community-affiliation graph model for overlapping network community detection.* Paper presented at the 2012 IEEE 12th International Conference on Data Mining.
Yang, J., & Leskovec, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems, 42*(1), 181-213.
Yang, S., Markoczy, L., & Qi, M. (2007). Unrealistic optimism in consumer credit card adoption. *Journal of Economic Psychology, 28*(2), 170-185.
Yang, Y. (2007). Adaptive credit scoring with kernel learning methods. *European Journal of Operational Research, 183*(3), 1521-1536.
Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications, 36*(2), 2473-2480.
Yobas, M., Crook, J., & Ross, P. (1997). *Credit scoring using neural and evolutionary techniques. Credit Research Centre, University of Edinburgh*. Retrieved from
Zafar, S., & Meenakshi, K. (2012). A study on the relationship between extroversion-introversion and risk-taking in the context of second language acquisition. *International Journal of Research Studies in Language Learning, 1*(1), 33-40.
Zhang, Y., Jia, H., Diao, Y., Hai, M., & Li, H. (2016). Research on credit scoring by fusing social media information in online peer-to-peer lending. *Procedia Computer Science, 91*, 168-174.
Zhou, T., Lü, L., & Zhang, Y.-C. (2009). Predicting missing links via local information. *The European Physical Journal B, 71*(4), 623-630.
Zięba, M., Tomczak, S. K., & Tomczak, J. M. (2016). Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications, 58*, 93-101.
Zielinski, A., Middleton, S. E., Tokarchuk, L. N., & Wang, X. (2013). *Social media text mining and network analysis for decision support in natural crisis management.* Paper presented at the ISCRAM.
zopa.com. Simple loans & investments | Zopa - The Feel Good Money compant. Retrieved from zopa.com

Appendix

Appendix 1. Ethics Approval for Primary Data Collection

29 March 2018

Ahmad Abd Rabuh
PhD Student
Faculty of Business and Law

Dear Ahmad

Study Title:	**Developing A Dynamic Credit scoring Model using Social Behaviour**
Ethics Committee reference:	BAL/2018/E494/RABUH

Thank you for submitting your documents for ethical review. The Ethics Committee was content to grant a favourable ethical opinion of the above research on the basis described in the application form, protocol and supporting documentation, revised in the light of any conditions set, subject to the general conditions set out in the attached document, and with the following stipulation:
The favourable opinion of the EC does not grant permission or approval to undertake the research. Management permission or approval must be obtained from any host organisation, including University of Portsmouth, prior to the start of the study.

Summary of any ethical considerations:
CONDITION. In your interview schedule form, complete the requested anonymisation of participants and organisations. Replace the date, time, name, job title and bank name fields with one field for 'Participant code', in exactly the same way as you have already done on your consent sheet. Forward the amended schedule to ethics-bal@port.ac.uk to confirm this condition has been met.

RECOMMENDATION. We recommend that you keep the situation about flight bans to Qatar under regular review, and make contingency plans as necessary.

Documents reviewed
The documents reviewed by Peter Scott [LCM] + BaL Ethics Committee

Document	*Date*	*Version*

Application Form	28/02/2018	1.6
Application Form	14/03/2018	1.7
Application Form	2403/2018	1.8
Peer / Independent Review	20/02/2018	
Questionnaire	20/02/18	2
Consent Form(s) (list if necessary)	12/03/18	1
Consent Form(s) (list if necessary)	24/03/18	1.1
Participant Information Sheet(s) (list if necessary)	12/03/18	1
Participant Information Sheet(s) (list if necessary)	24/03/2018	1.1

Statement of compliance
The Committee is constituted in accordance with the Governance Arrangements set out by the University of Portsmouth.

After ethical review

Reporting and other requirements
The attached document acts as a reminder that research should be conducted with integrity and gives detailed guidance on reporting requirements for studies with a favourable opinion, including:

- Notifying substantial amendments
- Notification of serious breaches of the protocol
- Progress reports
- Notifying the end of the study

Feedback

You are invited to give your view of the service that you have received from the Faculty Ethics Committee. If you wish to make your views known please contact the administrator, Christopher Martin.

Please quote this number on all correspondence: **BAL/2018/E494/RABUH**

Yours sincerely and wishing you every success in your research

Chair
Email:

Enclosures: *"After ethical review – guidance for researchers"*

Copy to: Mark Xu,
Renatas Kizys

Appendix 1

After ethical review – guidance for researchers

This document sets out important guidance for researchers with a favourable opinion from a University of Portsmouth Ethics Committee. Please read the guidance carefully. A failure to follow the guidance could lead to the committee reviewing and possibly revoking its opinion on the research.

It is assumed that the research will commence within 3 months of the date of the favourable ethical opinion or the start date stated in the application, whichever is the latest.

The research must not commence until the researcher has obtained any necessary management permissions or approvals – this is particularly pertinent in cases of research hosted by external organisations. The appropriate head of department should be aware of a member of staff's research plans.

If it is proposed to extend the duration of the study beyond that stated in the application, the Ethics Committee must be informed.

If the research extends beyond a year then an annual progress report must be submitted to the Ethics Committee.

When the study has been completed the Ethics Committee must be notified.

Any proposed substantial amendments must be submitted to the Ethics Committee for review. A substantial amendment is any amendment to the terms of the application for ethical review, or to the protocol or other supporting documentation approved by the Committee that is likely to affect to a significant degree:
(a) the safety or physical or mental integrity of participants
(b) the scientific value of the study
(c) the conduct or management of the study.

A substantial amendment should not be implemented until a favourable ethical opinion has been given by the Committee.

Researchers are reminded of the University's commitments as stated in the Concordat to Support Research Integrity viz:

- maintaining the highest standards of rigour and integrity in all aspects of research
- ensuring that research is conducted according to appropriate ethical, legal and professional frameworks, obligations and standards
- supporting a research environment that is underpinned by a culture of integrity and based on good governance, best practice and support for the development of researchers

- using transparent, robust and fair processes to deal with allegations of research misconduct should they arise
- working together to strengthen the integrity of research and to reviewing progress regularly and openly

In ensuring that it meets these commitments the University has adopted the UKRIO Code of Practice for Research. Any breach of this code may be considered as misconduct and may be investigated following the University Procedure for the Investigation of Allegations of Misconduct in Research.
Researchers are advised to use the UKRIO checklist as a simple guide to integrity.

Appendix 2. In-depth Questionnaire Template

Date: Time:

Name of Interviewee:
Job Title:
Bank Name:

Pre-study Interview Questions

1. What credit rating systems used in your bank? Are they different from product to product? E.g. personal loan, mortgage, business loan?

2. What are the current key factors and measures used in your credit rating/scoring systems?

3. How do you feel about these measures in terms of its reliability, accuracy, and forward prediction?

4. What factors are missing from your model and why?

5. Where can you get the data for these missing factors and measures? Do you face challenges when collecting the same?

6. Do you rely on 3rd party credit rating agencies? How do you evaluate the service provided by the 3rd party?

7. What other info can 3rd party rating agencies provide you with?

8. Do you, by any mean, believe that a borrower's social and behavioural factors would indicate his/her credit-worthiness? Why and why not?

9. What are the examples of a borrower's behaviour/lifestyle that indicates positive credit worthiness? What about the negative behaviour/lifestyle?

10. What other metrics do you think might improve credit rating scoring (financial education/literacy, political affiliation, etc.)?

11. How would you evaluate their relevance and reliability as additional measure in the credit scoring?

Collective Theme	Relevance					Reliability					Corresponding Academic Behavioural Factors
	1	2	3	4	5	1	2	3	4	5	
Social interactions with others											Openness
											Conscientiousness
											Extroversion
											Agreeableness
											Neuroticism
											Homophily
Financial interactions with others											Financial Integrity
											Temporal discounting
Outlook on investments (realistic/unrealistic) Bias											Optimistic Bias
											Hindsight bias
											Framing
											Ambiguity aversion
Mental realism (degree of illusion) Bias											Overconfidence
											Sensation-Seeking
											Impulsivity
											Egocentrism
											Status-quo bias
											Anchoring
											Integration

Appendix 3. Attributes Defined and Classified

Attribute	Classes	Data Type	Nature		
			Financial	Behavioural	Social
1. Account ID	Sequence	String	✓		
2. Outcome of Loan	NoDefault or Default	Binomial (categorical)	✓		
3. Type of loan	Consumer credit, car loan, mortgage, micro-loan for business development, cash loan, real estate loan, loan for purchase of equipment, loan for purchase of equity (margin loan), interbank credit, mobile operator loan, credit card, or overdraft.	Nominal (categorical)	✓		
4. Gender	Female, male, or unspecified	Nominal (categorical)	✓		
5. Car ownership	No or yes	Binomial (categorical)		✓	
6. Real estate ownership	No or yes	Binomial (categorical)	✓		
7. Number of children	0,1, … ,11	Integers (numerical)		✓	
8. Income	Number	Continuous (numerical)	✓		
9. Credit Amount	Number	Continuous (numerical)	✓		
10. Annual instalments (annuity)	Number	Continuous (numerical)	✓		
11. Value of underlying asset	Number	Continuous (numerical)	✓		
12. Borrower's companions at the time of application	Unaccompanied, family, spouse/partner, children, group of people, or other	Nominal (categorical)		✓	
13. Income type	Unemployed, student, pensioner, maternity leave, working, businessman, state servant, commercial associate	Nominal (categorical)	✓		
14. Education	Lower secondary, secondary, academic degree, incomplete higher, or higher education	Ordinal (categorical)		✓	
15. Marital Status	civil marriage, married, single/unmarried, widow, separated, or unknown	Nominal (categorical)	✓		
16. Housing type	Co-op apartment, house / apartment, municipal apartment, office apartment, rented apartment, or living with parents.	Nominal (categorical)		✓	
17. Population of the region proportionate to	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	

the nation population					
18. Age (in days)	Number	Integers (numerical)	✓		
19. Employment length (in days)	Number	Integers (numerical)	✓		
20. Account longevity (in days)	Number	Integers (numerical)	✓		
21. Days since ID/passport was issued	Number	Integers (numerical)	✓		
22. Owned Car Age (in years)	Number	Integers (numerical)		✓	
23. Provided mobile number	No or yes	Binomial (categorical)		✓	
24. Provided work number	No or yes	Binomial (categorical)		✓	
25. Provided home number	No or yes	Binomial (categorical)		✓	
26. Was mobile phone contactable?	No or yes	Binomial (categorical)		✓	
27. Provided other phone numbers	No or yes	Binomial (categorical)		✓	
28. Provided email address	No or yes	Binomial (categorical)		✓	
29. Job type	Low-skill labour, cleaning staff, waiters/barmen, security staff, cooking staff, drivers, laborers, sales staff, core staff, HR staff, IT staff, secretaries, medicine staff, realty agents, high-skilled tech staff, accountants, private service staff, managers.	Ordinal (categorical)		✓	
30. Number of dependents	Number	Integers (numerical)		✓	
31. Region rating	Number	Integers (numerical)		✓	
32. Region and City ratings	Number	Integers (numerical)		✓	
33. Day of the application	0,1,…,6	Integers (numerical)		✓	
34. Hour of the application	0,1,…,23	Integers (numerical)		✓	
35. Home region not work region	No or yes	Binomial (categorical)		✓	
36. Home region not living region	No or yes	Binomial (categorical)		✓	
37. Living region not work region	No or yes	Binomial (categorical)		✓	
38. Home city not work city	No or yes	Binomial (categorical)		✓	
39. Home city not	No or yes	Binomial		✓	

living city		(categorical)			
40. Living city not work city	No or yes	Binomial (categorical)		✓	
41. Organisation type	58 categories: business entity (3 types), industry (13 types), trade (7 types), transportation (4 types), bank, agriculture, advertising, cleaning, construction, culture, electricity, emergency, government, hotel, housing, insurance, kindergarten, legal services, medicine, military, mobile, other, police, postal, realtor, religion, restaurant, school, security, security ministries, self-employed, services, telecom, university, not-specified.	Nominal (categorical)	✓		
42. Scores of credit referencing agency (CRA) 1	Standardised fraction between 0.00 and 1.00	Continuous (numerical)	✓		
43. Scores of credit referencing agency (CRA) 3	Standardised fraction between 0.00 and 1.00	Continuous (numerical)	✓		
44. Scores of credit referencing agency (CRA) 3	Standardised fraction between 0.00 and 1.00	Continuous (numerical)	✓		
45. Apartment – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
46. Apartment area – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
47. Apartment area – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
48. Basement area - median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
49. Basement area - mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
50. Basement area - mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
51. Years since constructing the building – median of the neighbourhood	Number of years	Integers (numerical)		✓	
52. Years since constructing the building – median of the	Number of years	Integers (numerical)		✓	

neighbourhood					
53. Years since constructing the building – median of the neighbourhood	Number of years	Integers (numerical)		✓	
54. Age of the building – median of the neighbourhood	Number of years	Integers (numerical)		✓	
55. Age of the building – mean of the neighbourhood	Number of years	Integers (numerical)		✓	
56. Age of the building – mode of the neighbourhood	Number of years	Integers (numerical)		✓	
57. Communal area of the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
58. Communal area of the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
59. Communal area of the building – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
60. Number of elevators/lifts in the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
61. Number of elevators/lifts in the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
62. Number of elevators/lifts in the building – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
63. Number of entrances in the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
64. Number of entrances in the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	

65. Number of entrances in the building – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
66. Maximum number of floors in the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
67. Maximum number of floors in the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
68. Maximum number of floors in the building – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
69. Area of land the building is built on - median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
70. Area of land the building is built on - mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
71. Area of land the building is built on - mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
72. Living area of the apartments – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
73. Living area of the apartments – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
74. Living area of the apartments – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
75. The living area of the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
76. The living area of the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
77. The living area of the building –	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	

mode of the neighbourhood					
78. The non-living area of the apartment – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
79. The non-living area of the apartment – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
80. The non-living area of the apartment – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
81. The non-living area of the building – median of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
82. The non-living area of the building – mean of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
83. The non-living area of the building – mode of the neighbourhood	Fraction between 0.00 and 1.00	Continuous (numerical)		✓	
84. Apartment type	Studio, 1 B/R, 2 B/R, 3 B/R, 4 B/R or 5 B/R	Nominal (categorical)		✓	
85. State of the building ownership	owned by the borrower, owned by real estate developer, owned by another individual, state-owned or council, not registered, or not specified	Nominal (categorical)		✓	
86. House type	Block of flats, terraced house, stand-alone house, specific house.	Nominal (categorical)		✓	
87. Accommodation arrangement	Rent, shared-rent, living with parents or owner of property	Nominal (categorical)	✓		
88. Material used in walls of accommodation	Panel, stone, wood, mixed, block, monolithic, brick or other	Nominal (categorical)		✓	
89. Existence of emergency exit in the building	No or yes	Binomial (categorical)		✓	
90. Delinquent social ties during the last 30 days	Number	Integers (numerical)			✓
91. Defaulting social ties during the	Number	Integers (numerical)			✓

last 30 days					
92. Delinquent social ties during the last 60 days	Number	Integers (numerical)			✓
93. Defaulting social ties during the last 30 days	Number	Integers (numerical)			✓
94. Number of days since changing contact numbers	Number	Integer (numerical)		✓	
95. Providing document 1	No or yes	Binomial (categorical)	✓		
96. Providing document 2	No or yes	Binomial (categorical)	✓		
97. Providing document 3	No or yes	Binomial (categorical)	✓		
98. Providing document 4	No or yes	Binomial (categorical)	✓		
99. Providing document 5	No or yes	Binomial (categorical)	✓		
100. Providing document 6	No or yes	Binomial (categorical)	✓		
101. Providing document 7	No or yes	Binomial (categorical)	✓		
102. Providing document 8	No or yes	Binomial (categorical)	✓		
103. Providing document 9	No or yes	Binomial (categorical)	✓		
104. Providing document 10	No or yes	Binomial (categorical)	✓		
105. Providing document 11	No or yes	Binomial (categorical)	✓		
106. Providing document 12	No or yes	Binomial (categorical)	✓		
107. Providing document 13	No or yes	Binomial (categorical)	✓		
108. Providing document 14	No or yes	Binomial (categorical)	✓		
109. Providing document 15	No or yes	Binomial (categorical)	✓		
110. Providing document 16	No or yes	Binomial (categorical)	✓		
111. Providing document 17	No or yes	Binomial (categorical)	✓		
112. Providing document 18	No or yes	Binomial (categorical)	✓		
113. Providing document 19	No or yes	Binomial (categorical)	✓		
114. Providing document 20	No or yes	Binomial (categorical)	✓		
115. Credit inquiries within the last hour	Number	Integers (numerical)	✓		
116. Credit inquiries within the last	Number	Integers (numerical)	✓		

day					
117. Credit inquiries within the last week	Number	Integers (numerical)	✓		
118. Credit inquiries within the last month	Number	Integers (numerical)	✓		
119. Credit inquiries within the last quarter	Number	Integer (numerical)	✓		
120. Credit inquiries within the last year	Number	Integer (numerical)	✓		

FORM UPR16

Research Ethics Review Checklist

Please include this completed form as an appendix to your thesis (see the Research Degrees Operational Handbook for more information

Postgraduate Research Student (PGRS) Information			**Student ID:**	821483
PGRS Name:	Ahmad Abd Rabuh			
Department:	BAL - OSM	**First Supervisor:**	Mark Xu	
Start Date: (or progression date for Prof Doc students)	**01/02/2016**			
Study Mode and Route:	Part-time ☐	MPhil ☐	MD ☐	
	Full-time ☒	PhD ☒	Professional Doctorate ☐	

Title of Thesis:	Developing a Credit Risk Assessment Model Using Social Network Analysis
Thesis Word Count: (excluding ancillary data)	56,100 (excluding cover page, declaration, acknowledgements, references and appendix sections)

If you are unsure about any of the following, please contact the local representative on your Faculty Ethics Committee for advice. Please note that it is your responsibility to follow the University's Ethics Policy and any relevant University, academic or professional guidelines in the conduct of your study

Although the Ethics Committee may have given your study a favourable opinion, the final responsibility for the ethical conduct of this work lies with the researcher(s).

UKRIO Finished Research Checklist:
(If you would like to know more about the checklist, please see your Faculty or Departmental Ethics Committee rep or see the online version of the full checklist at: http://www.ukrio.org/what-we-do/code-of-practice-for-research/)

a) Have all of your research and findings been reported accurately, honestly and within a reasonable time frame?	YES ☒ NO ☐
b) Have all contributions to knowledge been acknowledged?	YES ☒ NO ☐
c) Have you complied with all agreements relating to intellectual property, publication and authorship?	YES ☒ NO ☐
d) Has your research data been retained in a secure and accessible form and will it remain so for the required duration?	YES ☒ NO ☐
e) Does your research comply with all legal, ethical, and contractual requirements?	YES ☒ NO ☐

Candidate Statement:

I have considered the ethical dimensions of the above named research project, and have successfully obtained the necessary ethical approval(s)

Ethical review number(s) from Faculty Ethics Committee (or from NRES/SCREC):	BAL/2018/E494/RABUH

If you have *not* submitted your work for ethical review, and/or you have answered 'No' to one or more of questions a) to e), please explain below why this is so:

Signed *(PGRS)*:		**Date:** 03 / 01 /2021