Spring 4-15-2020

A CREDIT ANALYSIS OF THE UNBANKED AND UNDERBANKED: AN ARGUMENT FOR ALTERNATIVE DATA

Edwin Baidoo

A CREDIT ANALYSIS OF THE UNBANKED AND UNDERBANKED

AN ARGUMENT FOR ALTERNATIVE DATA

Edwin Baidoo

A Thesis Submitted in Partial Fulfillment

For the Degree of

DOCTOR OF PHILOSOPHY

At the

College of Computing and Software Engineering

Analytics and Data Science Institute

Kennesaw State University

April 2020

Abstract

The purpose of this study is to ascertain the statistical and economic significance of non-traditional credit data for individuals who do not have sufficient economic data, collectively known as the unbanked and underbanked. The consequences of not having sufficient economic information often determines whether unbanked and underbanked individuals will receive higher price of credit or be denied entirely. In terms of regulation, there is a strong interest in credit models that will inform policies on how to gradually move sections of the unbanked and underbanked population into the general financial network.

In Chapter 2 of the dissertation, I establish the role of non-traditional credit data, known as alternative data, in modeling borrower default behavior for individuals who unbanked and underbanked individuals by taking a statistical approach. Further, using a combined traditional and alternative auto loan data, I am able to make statements about which alternative data variables contribute to borrower default behavior. Additionally, I devise a way to statistically test the goodness of fit metric for some machine learning classification models to ascertain whether the alternative data truly helps in the credit building process.

In Chapter 3, I discuss the economic significance of incorporating alternative data in the credit modeling process. Using a maximum utility approach, I show that combining alternative and traditional data yields a higher profit for the lender, rather than using either data alone. Additionally, Chapter 3 advocates for the use of loss functions that aligns with a lender's business objective of making a profit.

Index Terms – Profit Scoring; Unbanked; Underbanked; Alternative Credit Data; Likelihood Ratio Test; Unscorables

Acknowledgments

First and foremost, I would like to express my deepest gratitude to God for His unfailing love and sustenance. Through it all, He has been the constant factor. Thank you!

Additionally, I would like to thank my wife for her understanding. This journey was lonely at times, but I am grateful for her encouragement and kind words that lifted me up when the road ahead seemed murky, and the load too heavy to bear. I am equally thankful for the motivation and reason behind it all - my two year old daughter Sophia. Although it was challenging pursuing this dream, she was my companion when I was coding and writing the manuscripts.

I am massively indebted to my advisor and mentor Dr. Stefano Mazzotta for allowing me to work with him. His research acumen, insight, integrity and patience is unmatched. From the first day I met him, he has demanded of me the highest research standard and often asks poignant questions to which he would often say "just think about it." I am thankful that he shared his time with me, including some weekends.

I also want to thank Dr. Jennifer Priestley and Dr. Sherrill Hayes for being my strong supporters. Apart from my family, Dr. Priestley has encouraged and affirmed me, especially when the going was tough. I have cherished her advise, professional wisdom and her ability to always make time for me when I needed her. Talking to Dr. Hayes always seemed like a breath of fresh air because he made the task ahead seemed doable. A big thank you also goes to my committee members for their support and willingness to read the manuscript.

For pops - thank you

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Consumer banking services fall between two groups: traditional and alternative. Traditional banking includes mainstream banks, credit unions, and thrifts that operate within the parameters of the Federal Deposit Insurance Corporation (FDIC). They are governed by well defined federal and state regulations that dictate their banking activities and products offered.

Alternative banking[1] includes services that operate outside of the traditional banking system. Some of the products they offer include check-cashing, rent-to-own, pawnshops, tax refund anticipation loans, etc.

Alternative banking often serves two groups of consumers: unbanked and underbanked. Consumers who are *unbanked* have no formal relationship with a bank. That is, they have no bank account or a credit card. *Underbanked* consumers have a bank account but use alternative financial products to supplement their credit need. Statistically, (Aitken, 2017) points out that 4.5 billion adults

[1] also known as fringe banking

globally, representing 62% of the world's population, are either underbanked or unbanked. As a result, unbanked and underbanked borrowers are missing the opportunity to be part of the global economic stream. (Smith and Henderson, 2018) report that 53 million American consumers are not fully served by traditional financial institutions.

A major concern surrounding alternative banking is that they are not as regulated as traditional banking. This means that a consumer who relies on alternative banking may not be fully protected. Therefore, there has been a significant push by regulators to find opportunities that will allow unbanked and underbanked consumers to join, participate and interact with the mainstream economy[2]. This presents an opportunity for lenders to learn more about consumers within this segment of the credit spectrum in an effort to meet their economic need.

The structure of this dissertation is as follows: Chapter 1 provides a discussion about unbanked and underbanked consumers and the role alternative data plays in understanding them. Chapter 2 lays the foundation upon which I assess the statistical value in alternative data. In Chapter 3, I study profit scoring and argue that alternative data brings economic contribution to the credit scoring process. Chapter 4 contains the conclusion, where the discussions throughout the chapters are tied in together. The Appendix has results pertaining to how robust some assumptions made in Chapter 3 are.

[2]Defined as the general economic or financial market

1.1 Unbanked and Underbanked Consumers

Unbanked and underbanked consumers are those who do not participate and interact with the mainstream economy. They do not have a savings or checking account and also use some alternative banking products to supplement their financial needs. For example, the 2017 FDIC National Survey of Unbanked and Underbanked showed that nearly 20% of households had not used mainstream credit within the previous year.

In terms of demographics, unbanked and underbanked consumers constitute over 24% of all minority families and 5% of Caucasian families. They are more likely to have lower income, be less educated, unemployed, and reside in low to moderate income neighborhoods, (FDIC, 2017).

The decision to participate in the mainstream economy carries many benefits. (Rhine and Greene, 2013) cite the following:

1. Wealth creation and the accumulation of assets
2. Protection from theft and other discriminatory lending practices that may seek to prey on consumers
3. Having the proper channel to save and deposit funds as well as cashing checks. This ensures that consumers will not require any alternative financial services, such as check cashing, that may provide similar benefits at high rates, especially during financial emergencies

The accumulation of wealth and asset proves exceptionally useful during retirement age, when consumers can draw upon any saved reserves to supplement fixed income or cope with unforeseen financial shocks.

The major drawback for those who do not participate in the mainstream economy is that they are susceptible to financial shocks that stem from natural or man-made disasters because they do not have the protection of a depository institution to act as a safe haven for their assets. Moreover, communities with a functioning financial market become more resilient to financial shocks and can even take advantage of a growing economy, (Rhine and Greene, 2013). In the next section, I highlight some reasons why borrowers remain unbanked or underbanked from the individual and household dynamics.

1.2 Why Some Consumers Are Unbanked or Underbanked

(B. F. Hayashi and F. Hayashi, 2016), identify six reasons why some borrowers remain at the peripheral of the traditional banking system.

The first and main reason why borrowers do not have a checking or savings account is because of the high cost of maintaining the account. For many borrowers, the median overdraft and maintenance fee[3] proves to be too much. Also, a checking or savings account requires account holders to maintain a certain minimum dollar amount for a specific period of time. Borrowers are charged additional fees if they do not meet this threshold.

The second reason comes from borrowers's negative experience with a depository institution. For example, some borrowers perceive banks to clear checks in a way that leads to an overdraft fee. This unexpected fee often causes their account to plummet into the negative. Another reason, which was uniquely cited by

[3]Median overdraft fee is $30

immigrants, was that banks could not properly communicate with them because of a language barrier.

The third reason is related to borrower proof of identification. For example, some consumers do not have the proper identification card or social security number. This reason differs from the rest because it is the only explanation that is directly related to how banks use proof of identification as a risk management tool for their preferred candidate. Consumers who do not meet such qualifications will not be able to open an account even if they want to.

Privacy, the fourth reason, is in opposition to the third reason, because it is related to the consumer's preference. As corporate data breaches become more common, some individuals choose not to use banks. Instead, they use cash exclusively because it lacks paper trail, provides anonymity, and shields borrowers from any fears about data breaches.

The fifth reason concerns issues related to consumer convenience, such as banking location, hours, and physical accessibility. In certain geographic areas where consumers do not have direct access to banks, they are often propelled to find a nearby depository institution that may meet their banking needs. This often incurs an indirect cost in the form of transportation and related costs.

The sixth reason relates to features of the account generation process that unintentionally marginalizes certain groups of consumers. For example, some consumers with little financial literacy may find the account opening process to be cumbersome and complex. This is further compounded when complicated account fee structures are introduced.

Although the reasons above appear on an individual level, research shows that *household* dynamics provides crucial reasons why individuals become unbanked

or underbanked. In their study, (Rhine and Greene, 2013), suggest that changes in marital status, loss of employment, loss of income, and loss of health insurance contributes negatively to why individuals do not interact with the mainstream economy. The reasons offered are often interrelated. For example, in the event of a divorce, they find that one of the parties have a greater chance of being unbanked or underbanked.

When there is a significant decrease in consumer income, through a loss or reduction of employment, to the level where consumers are pushed below the poverty line, it can be expected that the likelihood of individuals becoming unbanked will increase, especially if there are fee structures associated with the account.

The introduction of technology into financial services, efinancial services, provide additional challenge for those who cite privacy as a reason not to participate in the mainstream economy. As a result, they may be excluded from many benefits stemming from the use of efinancial services. For example, in an effort to lower their costs of administering welfare programs, the federal government began to advocate the use of electronic payment as mechanism to receive funds such as Social Security Benefits (Hogarth and O'Donnell, 1999). In the next section, I explore the role of data in banking and how the advent of big data helps to understand unbanked and underbanked consumers.

1.2.1 Role of Data in Banking

A central pillar of the traditional banking system that allows people to participate and interact with the mainstream economy is through borrowing and repayment cycles. To accomplish this, banks record data on current customers

and rely on historical data from previous customers as well as external information from credit bureaus.

Because unbanked or underbanked consumers do not participate or interact with the mainstream economy, they have little to no borrowing and repayment records. Therefore, unbanked or underbanked consumers are sometimes referred to as *credit invisible* or *thin-filed*.

As a standard, an important tool for making lending decisions is the credit score. For thin-filed borrowers, the central issue is that they do not have sufficient information for credit bureaus to build a reliable credit score that informs a potential lender about their credit worthiness. However, being credit invisible in itself does not imply that a borrower is not credit worthy. In fact, (Smith and Henderson, 2018) documented cases where many individuals who are credit invisible emerged as credit worthy and even became homeowners.

1.3 Other Sources of Data

While the age of big data has brought about novel algorithms for credit modeling, it has also introduced a question of whether consumer default behavior can be gleaned from non-traditional data sources[4], known as *alternative data*.

(Óskarsdóttir et al., 2019) suggest that the "best investment in better credit scoring models ... is to leverage innovative big data sources instead." The importance of this observation is that while unbanked and underbanked borrowers may not have sufficient traditional economic information, they may generate alternative data that could be used to measure their creditworthiness.

[4]See (Anderson and Hardin, 2014), (Wei et al., 2016) and (Onay and Öztürk, 2018)

For example, (Agarwal et al., 2018) assert that the widespread use of cellphones provide new socio-behavioral variables that can be used to determine the credit-worthiness of a potential borrower. They proposed a way of using phone-based variables with existing demographic and past financial behavior as markers of "financial trouble." Their data consisted of a combination of 82.2 million banking transaction records with 350 million phone logs derived from 180,000 individuals spanning 2 years. They found that phone-based data contain important signals that can be used to gauge an individual's credit worthiness.

With the advent of social media, there has been a focus on its use for credit modeling. (G. Guo et al., 2016) mined data from Weibo, a twitter-like platform in China with an intention to "identify credit related evidence hidden in social data." They conducted an analysis which consisted of more than 7.3 million tweets generated by 200,000 users. They found that incorporating social media data into the credit modeling process outperformed traditional methods by as much as 17%.

(Berg et al., 2018) analyzed the effectiveness of using *digital footprints*, information consumers leave online when they access or register on a website, to predict default. Using approximately 250,000 transaction data from a German E-Commerce company, they found that information contained in digital footprints complements information contained in traditional financial data, which helps give a broader description of the borrower default behavior.

Having recognized this, the Board of Governors of the Federal Reserve System of the Consumer Financial Protection Bureau (CFPB) issued a statement[5] where they acknowledged that the "use of alternative data may improve the speed and accuracy of credit decisions and may help firms evaluate the creditworthiness

[5]See (CFPB, 2019)

of consumers who currently may not obtain credit in the mainstream credit system."

Following a similar sentiment, (Smith and Henderson, 2018), discussed the question of whether thin-filed consumers can be credit worthy even without a credit score. Using the Equifax database, they "followed two samples of thin-file individuals with no credit scores for at least four years in order to develop a timeline indicating when they obtained sufficient credit to qualify for a credit score."

Their conclusion was that thin-filed consumers proved themselves to be creditworthy, with credit scores ranging from "below 520 to 740 and above." Additionally, the average time it took them to obtain a credit score and be part of the mainstream economy was between three to four years. Within the four year span, it was reported that the majority received credit scores within the first and second years. However, the use of alternative data may have shortened the wait time.

It is evident that researchers and regulators are advocating for a new approach to credit default modeling that takes advantage of alternative data sources. While alternative data may prove useful, researchers and regulators have called for caution in order to better understand variables that can *legally* be used as a discriminator for default risk. Additionally, alternative data has the potential to migrate individuals who are underbanked and unbanked into the mainstream economy.

1.3.1 Alternative Credit Data

According to (Aitken, 2017), the goal of helping individuals who are credit invisible can be boiled down to an exercise of making them visible. This attempt

can be achieved in two ways:

1. The first involves finding ways to identify credit worthy individuals who would otherwise be difficult to identify through traditional methodologies. This speaks largely to the use of innovative statistical credit scoring methodologies that can properly discriminate between borrowers who default and those who do not.

2. The second involves recording non-traditional financial behaviors that could be used to model credit. These "nontraditional behaviors" can include local public records, social networking patterns, academic achievement records, mobile phone usage, non-financial payment histories, and psychometric test results.

Historically, credit models were built using traditional financial data. This is defined as data that is "managed in the core credit files of the nationwide consumer reporting agencies" (Experian, 2015). Elements of traditional financial data includes trade-line information such as debt repayment history, current and historical account status, credit limit and credit usage information. Other elements include credit inquiries and public records such as bankruptcies.

At the opposite end, alternative credit data refers to data used in the credit modeling process that is not an element in the traditional data. The only criteria is that they must meet the guidelines of the Fair Credit Reporting Act (FCRA) - that is they must be disputable, correctable and displayable.

Within the last decade, there have been attempts to incorporate elements of alternative data in decision making. For example, Experian has helped popularized the use of RentBureau - a database that contains 24-hour rental payment

information. The significance of this database is that rental payment history can offer insights about loan default behavior. More importantly, it is a representation that some non-traditional data points can be harnessed as proxy for positive or negative credit behavior (Aitken, 2017).

Examples of alternative credit data includes mobile phone payment, cable TV payment, tax, and deed records (Experian, 2015). Figure 1.1 illustrates some of the differences between alternative and traditional data.

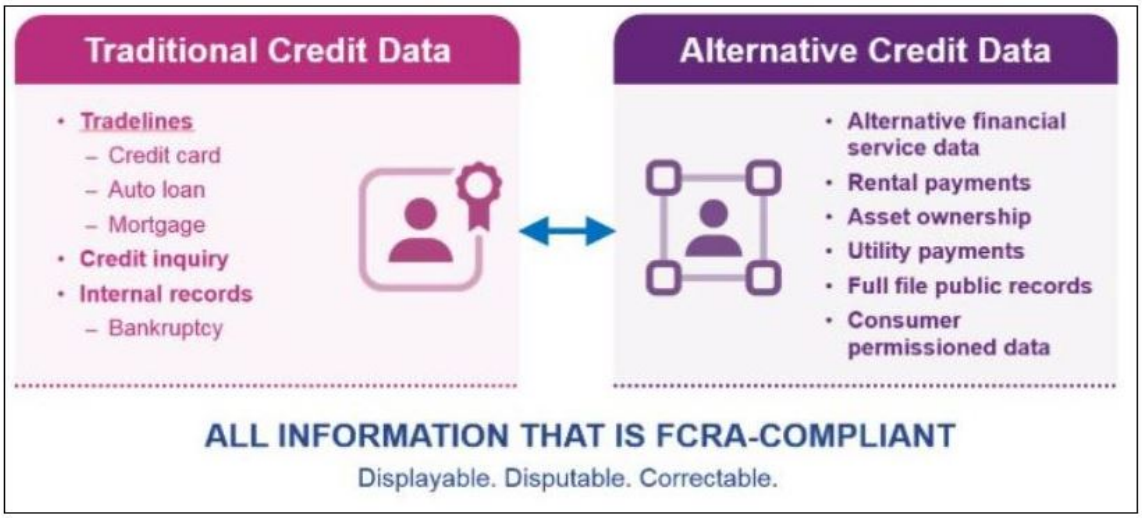


Figure 1.1: Traditional Credit Data vs. Alternative Credit Data (from Experian)

At the core of the issue is the predictive power of alternative data variables. Recently, TransUnion conducted a survey of more than 317 lenders concerning how they incorporated alternative data in their lending practices (TransUnion, 2015). They revealed that using alternative data

1. Opened opportunities in new markets
2. Allowed them to reach more credit worthy individuals

3. Situated them to be more competitive

The same survey indicated that nearly 64% of lenders saw tangible benefits within the first year of using alternative credit data. In the next section, I describe the data used in this analysis.

On the regulatory side, the Consumer Financial Protection Bureau (CFPB), the Board of Governors of the Federal Reserve System and four other regulatory bodies have recognized the importance of extending credit to thin-filed individuals in an effort to assimilate them into the mainstream economy. This creates an opportunity where reliable credit underwriting procedures can be combined with traditional and novel credit modeling methodologies[6]. In the next section, I provide a discussion on credit scoring methodologies.

1.4 Credit Modeling

Credit scoring[7] is the application of statistical methods to predict borrower default probabilities. It consists of building various statistical models known as scorecards that can sufficiently measure and predict a borrower's default risk. Given a historical data, a good scorecard will aim to discriminate between borrowers who will default and those who will not. To make a decision to extend or decline credit, a lender uses a cutoff or threshold value[8]. Borrowers with default probability less than the threshold will be given credit, while those with a probability score greater than the threshold will be denied credit, (Thomas, 2009).

[6] See (CFPB, 2019)
[7] Or default risk modeling
[8] The value could be domain specific or from past business experience

Since statistical techniques were first applied to consumer lending in the 1950's, the credit granting process has had two components (Feelders, 2003). The first component involves a quantitative method by which lenders can *legally* discriminate between good and bad borrowers. By definition, good borrowers comply with the terms of the loan contract and do not default, while bad borrowers violate all or some portion of the loan contract and ultimately default. The second component consists of the lender's observation of the borrower from loan origination to the time of loan maturity. Here, the lender observes the borrower's financial or payment behavior for the purpose of cross selling other products or most importantly to determine if they will default or not.

The use of statistical techniques to model default risk in consumer lending is rich and continues to grow. (Cyert, H. J. Davidson, and Thompson, 1962) used markov chains to analyze "doubtful accounts." First, they binned a lender's account receivables by age and then modeled the loss expectancy rate within each bin. From there, a lender would now be able to set aside an allowance in anticipation to potential losses given a default event.

To understand the dynamics between variables on borrower application forms, (Sewart, Pete, 1998) applied concepts in graph theory to describe the association between variables taken from credit card application forms. Using directed and undirected graphs, they showed conditional dependence to improve the understanding of the relationships between credit variables. Moreover, they were able to model the joint-distribution of variables on the borrower application form.

(David J. Hand and Kelly, 2002) explores the concept of super-scorecards, a classification ensemble of individual credit scoring models that yielded superior results when compared to its components. An attractive feature of this ap-

proach is that an arbitrary number of standalone models could be used in the construction of the super-scorecard. Also, the proposed model is more flexible but retains the ease of interpretability of the standard linear scorecard.

In the era of big data, machine learning models have been used in credit scoring. For example, using a Classification and Regression Tree (CART) model on combined consumer banking transaction and credit bureau data, (Khandani, Kim, and Lo, 2010) were able to detect non-linear relationships that a traditional logistic or discriminant model will not be able to find. The model's accuracy allowed them to predict default events three to twelve months in advance. (Lessmann et al., 2015) and (Baesens, Rösch, and Scheule, 2016) provide excellent overviews of machine learning models in the credit space.

Owing to the fact that borrowers and lenders do not often share the same objective, (Keeney and Oliver, 2005) modeled credit default by using Cooperation Negotiation Analysis and Efficient Frontier Curves to develop a model that identifies and integrates both borrower and lender preferences. For example, a borrower's preference for a lower loan price may be matched with a lender's preference for profit or market share. The outcome is a win-win product that has the potential to significantly decrease probability of default. In the next section, I provide a discussion on the economic value for a lender to consider alternative data in the credit modeling process.

1.5 Profit Scoring

Although traditional credit scoring methods have largely focused on default risk, it represents only one aspect of the entire credit granting process. In most applications, the main objective of the lender is to maximize profit given

default probabilities and other borrower characteristics, (Thomas, 2000). In this regard, a default-centric approach *alone* to credit modeling may not be a sufficient indicator for profit, although there may be a strong correlation between higher levels of default probabilities and lower profit and vice versa.

Recent studies suggest that there has been a gradual shift from traditional probabilistic based approach to profitability based approach[9]. This is largely driven by the observation that traditional loss functions underlying some default based credit models does not effectively capture the lender's objective. (S. Finlay, 2010) puts it better when he describes existing loss functions as being at best a "crude approximation" to the real objective of the lender, which is to identify customers who will contribute to some profitability metric.

Often, the economic loss function is robust and well established in literature. For example, using the Internal Rate of Rate of Return (IRR), (Serrano-Cinca and Gutiérrez-Nieto, 2016) sought to predict expected profit within the context of peer-to-peer lending. Using over 40,000 loan transaction records, they found that a lender "applying a profit scoring system ... outperforms the results obtained by using a traditional credit scoring system." Other economic loss functions that have been used include Customer Lifetime Value (CLV)[10], Net Present Value (NPV)[11], and the Return On Investment (ROI)[12]. Therefore, the choice of an economic loss function plays an important role in profit scoring.

The overarching theme throughout the dissertation is to investigate the use of alternative data as a viable source of information for credit scoring for underbanked and unbanked individuals. I study the topic from two different perspectives. The first, studies the *statistical value* of alternative data. Using the AUC

[9] See (D. J. Hand and Henley, 1997), (S. M. Finlay, 2008), (Devos et al., 2018), (Paula et al., 2019) and (Kozodoi et al., 2019)
[10] See (Barrios, Andreeva, and Ansell, 2014)
[11] See (Lieli and White, 2010)
[12] See (Maldonado et al., 2017)

as the goodness of fit metric, along with bootstrap hypothesis testing, I test the value of alternative data in credit scoring. Here, I consider the use of alternative data in *absence* of traditional financial data. This is important because it imitates current challenges lenders face in credit scoring when the borrowers are underbanked or unbanked.

The second, examines the *economic value* of alternative data using the profit scoring approach. I extend the work of (Lieli and White, 2010) to the underbanked unbanked population and consider whether there is an economic value for a lender to invest in alternative data. Here, I also focus on a sub-theme of using loss functions that reflect the economic objective of the lender.

Chapter 2

Scoring The Unscored: A Statistical Approach

Abstract

The purpose of this chapter is to examine the effect of alternative data in the credit decision process. This is especially focused on unbanked and underbanked consumers who do not have sufficient traditional credit data. The consequence of not having sufficient economic information often determines whether borrowers who are credit-invisible will receive a higher price for credit or be denied entirely. In terms of regulation, there is a strong interest in how to incorporate non-traditional data into the credit building process. Additionally, regulators are interested in how to migrate underbanked and unbanked borrowers into the mainstream economy.

Using traditional and alternative auto loan data, statements can be made about which alternative data variables contribute to borrower default behavior. I find that alternative variables can complement traditional variables to gauge default behavior. For example, when an unbanked or underbanked consumer does not have a valid home phone number the probability of default increases by almost 4%. Also, when an unbanked or underbanked borrower does not have a valid home address, the probability of default increases by 6.37%. This is something that lenders can look to as a positive proxy for credit worthiness.

Index Terms – Unscorables, Alternative Data, Credit Invisible, Unbanked, Thin Files, Machine Learning.

2.1 Introduction: An Overview Of Classification Models

In this section, I provide a brief overview of the machine learning classification models that underlies Table 2.5. Machine learning models can be grouped into three main branches: supervised learning, unsupervised learning and reinforcement learning. In supervised learning, a learner or a model, is provided with a training data that contain features and *known* target labels. The objective of the learner is to study the underlying pattern in the data so as to predict the target labels of unseen data. Most classification and regression models fall into this category.

For unsupervised learning, a model is provided with training data with no target labels. Here, the goal of the learner is to decipher natural relationships in the data. An example of this approach is principal components analysis and cluster analysis, (Hinton and Sejnowski, 1999).

In reinforcement learning, an agent or a learner must take a series of actions that will maximize a given reward. The agent is not given instructions on which actions to take but learns it through a system of reward and penalty. The key difference between reinforcement learning and other forms of learning is that in reinforcement learning, the agent *cannot* learn from examples but instead must study their environment, (Sutton and Barto, 1998). The outline of this chapter is as follows: Section 2.2 and Section 2.3 describes the data, modeling methodology and the results. This is followed by the conclusion in Section 2.4.

2.1.1 Random Forest

Proposed by (Breiman, 2001), a random forest model belong to a class of classification algorithms that are also called ensemble methods. The rationale behind ensemble methods is to train multiple standalone *base* models or classifiers in order to make a prediction. The final prediction of unseen labels or classes is made by aggregating over the predictions of the base models. For many classification ensemble methods, the method of aggregation is done through a simple majority voting.

Since the base model for a random forest model is the decision tree[1], it means the final prediction is the mode of the prediction of each base tree. Each decision tree model is built using a sample with replacement from the training data such that increasing the correlation between each tree model will result in an increase in the error rate of the random forest model.

The aim of a random forest model is to partition the dataset into smaller "pure" groups by splitting on multiple variables. Purity is defined as the uniformity of a class label within each split. Unlike some classification models that use a single linear decision boundary for prediction, tree based models divide the feature space every time a decision is made to split on a variable. The result is that the feature space becomes rectilinear, containing purer observations, (Loh, 2011).

In light of the goal of tree-based models, finding the best variable to split on is crucial. Typically, the best feature to split on is the one that optimizes some purity or impurity measure. For random forest, an example of an impurity measure is the gini index. The gini index is calculated as follows:

[1] For a discussion on the decision tree algorithm, see (Quinlan, 1986) and (Esposito, Malerba, and Semeraro, 1997)

$$Gini(t) = 1 - \sum_{i=1}^{i=N} P(C_i|t)^2 \tag{2.1}$$

where N corresponds to the number of classes within the data. C_i corresponds to the class label associated with the i^{th} observation in the data and t is a condition of the variable. Therefore, a decision would be made to split on a variable if it has the maximum gini index (Fawagreh, Gaber, and Elyan, 2014). For an in-depth mathematical treatment, see (Pereira et al., 2017), (Denil, Matheson, and De Freitas, 2014), (Chou, 1991) and (Lomax and Vadera, 2013).

Random forests have been applied in problem domains such as document classification, employee turnover, speech recognition, remote sensing and healthcare[2] with significant results. For example, (Loh, 2014) remarks that on average, the accuracy of a best decision tree model is 10% less than that of a random forest model. They are known to be more robust than decision trees and are more preferred compared to other classification models, (Pereira et al., 2017). Compared to decision trees, they are more robust to overfitting. Compared to other classifiers they are easy to interpret, implement and can handle large datasets, (Gehrke, Ramakrishnan, and Ganti, 2000).

2.1.2 Support Vector Machines

For a binary classification task, let $X = X_{TR} \cup X_{TS}$ be the entire dataset, where X_{TR} is the training data and X_{TS} is the testing data. Then, for X_{TR} with N samples, define $\{x_i, y_i\}_{i=1}^{i=N} \in X_{TR}$ to be a single sample with $x_i \in \mathbb{R}$ and $y_i \in \{+1, -1\}$. The objective of support vector machines is to find a

[2]See Jain, Duin, and Mao (2000), Pereira et al. (2017), Denil, Matheson, and De Freitas (2014), and Gao, Wen, and C. Zhang (2019)

decision boundary or a hyperplane, $f(x)$, in X_{TR} that can provide the maximum separation between the two classes, such that prediction errors, ε, are minimized. This hyperplane is given by the following equation:

$$f(x) = w \cdot x + b \tag{2.2}$$

where $(\cdot)$ is the dot product and $w, b \in \mathbb{R}$. The goal of minimizing the prediction error, ε, is achieved when the *norm* of w is minimized. This can be stated as the following optimization problem

$$\begin{aligned} &\text{minimize} \frac{1}{2}||w||^2 \\ &\text{subject to} \begin{cases} f(x) - y_i & \leq \varepsilon \\ y_i - f(x) & \leq \varepsilon \end{cases} \end{aligned} \tag{2.3}$$

To control the model from overfitting, (Vapnik, 2000) introduced a regularization hyperparameter, $c > 0$. The result is that 2.3 is modified to include slack variables ψ_i and ψ_i^*. This is shown in 2.4.

$$\begin{aligned} &\text{minimize} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\psi_i + \psi_i^*) \\ &\text{subject to} \begin{cases} f(x) - y_i & \leq \varepsilon + \psi_i^* \\ y_i - f(x) & \leq \varepsilon + \psi_i \\ \psi_i, \psi_i^* & \geq 0 \end{cases} \end{aligned} \tag{2.4}$$

A given point (x_i, y_i) which is geometrically closer to or on $f(x)$ is called the *support vector*. Support vectors are important because they are used for the

prediction of unseen data. This is one of the advantages of support vector machines, namely that they do not use all of the training data but few points that lie on the optimal separating boundary. For an in-depth treatment of support vector machines, see (Madeo, Lima, and Peres, 2012), (Smola and Scholkopf, 2004) and (Anguita et al., 2010).

While support vector machines work well on linearly separable spaces, it requires the notion of kernels, for non-linear classification. A kernel is a similarity measure that maps a data point from a non-linear space to a high dimensional linear space, such that the new data points are linearly separable[3]. For a discussion on the role of kernel functions in support vector machines see (Cuentas, Peñabaena-Niebles, and Garcia, 2017), (Goh and Lee, 2019), and (Sun and Liang, 2015).

Since its inception, support vector machines have been applied to many disciplines. For example, in the field of engineering, (L. Zhang et al., 2013) used support vector machine to predict product degradation and its remaining useful life. In accounting, (Baranes and Palas, 2019) utilized support vector machines to forecast future quarterly earnings of a company. They reported a 5% increase in accuracy over the benchmark model. In agriculture, support vector machines have been used detect livestock quality, including rotten or healthy vegetables, fruits and even identify the onset of diseases in some poultry, (Nurhanna and Othman, 2017). In healthcare, it has been used to detect long term type-2 diabetes, (Abbas et al., 2019).

Like decision trees and random forest, support vector machines are robust to overfitting through the regularization parameter c. Also, it has capabilities to handle non-linear classification by means of a kernel function. While robust

[3]See (Somvanshi et al., 2017)

kernel functions exists, it is possible to customize or even create a new kernel function to tackle new problems. However, unlike decision trees support vector machines are not easily interpreted.

2.1.3 k-Nearest Neighbor (kNN)

The k-nearest neighbor classifier is one of the most simplest but effective classification algorithm in data mining and pattern recognition, (Z. Guo et al., 2019). The intuition is to classify unseen data by observing the class label of its nearest neighbor. Formally, let $x_j \in X_{TS}$ be an observation of the testing data with $\{(x_k, x_l, ..., x_\zeta)\} \in X_{TR}$ being its neighboring points. Also, let $\{(y_k, y_l, ..., y_\zeta)\} \in \{+1, -1\}$ be the classification label of the neighboring points. For x_j to be classified as $y_j \in \{+1, -1\}$, the k-nearest neighbor algorithm locates k points in $(x_k, x_l, ..., x_\zeta)$ and retrieves their classification labels. The label for x_j is the *mode* of the labels of the k points in $(x_k, x_l, ..., x_\zeta)$. It is better for k to be an odd number in order to handle cases of a tie.

For the k-nearest neighbor classifier to work, two conditions must be met. The first is a metric to compute distance in order to find neighboring points. While the popular metric is the euclidean distance, there are other metrics such as City-block, Chebychev, Minkowski, etc[4]. The last condition involves a function that assigns a class to the new unseen data. The widely used choice is the mode. For a more involved discussion on the k-nearest neighbor classifier, see (Laaksonen and Oja, 1996).

The most important parameter for k-nearest neighbor is the value of k, as they directly influence the prediction. Changing the value of k has the potential to

[4]See (Imandoust and Bolandraftar, 2013), (Padraig and Delany, 2007) and (Ali, Neagu, and Trundle, 2019)

produce different class labels. Because there is no way to select the optimal k, one approach is to continuously select k values and record an improvement in some selected goodness of fit metric. The value of k that gives the maximum improvement will be selected.

k-nearest neighbor algorithms have been described as a lazy learner because it does not learn any underlying pattern within the data until an instance of the testing data is introduced for classification. Usually this means that the training data is stored in memory at run-time to await any testing data. This may be problematic if the dataset is large, however, the growth of computational power often allows this to be possible.

Because of its simplicity, k-nearest neighbor classifier has been used in different fields. In education, (Intan, Ghani, and Salman, 2020) used it to predict whether children are ready, doubtful or not ready for elementary school. (Anggraini and Tursina, 2019) also used it to predict public sentiment to a change in educational policy in Indonesia. In the next section, I describe the data, the preprocessing steps and the methodology.

2.2 Data Description and Cleaning

2.2.1 Data Preprocessing

The purpose for this research revolves around the question of whether alternative data can be used as a source of data for credit decisions in the era of big data, when traditional data is not available or observed? For individuals with little to no credit information, the unbanked and underbanked, I will investigate if alternative data helps in assessing their credit worthiness.

The data was provided by "Company Z" for the exclusive purpose of research on the condition of anonymity. It consists of non-financial payment streams, social footprints and elements of traditional credit data. More specifically, it relates to the application of subprime automobile transactions. There is a total of 23,981 applications and 823 variables or features.

As part of the preprocessing stage, the data was stripped of any personally identifiable information such as names, date of birth, SSN, address and telephone numbers. Additionally, all variables whose meaning could not be found in the data dictionary were excluded from the data. The next step dealt with the treatment of missing information.

Missing data is a part of almost all research studies. Although their treatments are seldom the sole focus of substantive research, they introduce an element of uncertainty into the model building process while affecting other statistical mechanisms such as the mean, variance and standard deviation, (Mandel J, 2015). Therefore, the treatment of missing values plays an important part of any data analysis.

According to (Acuna and Rodriguez, 2010), the treatment of missing values fall under two categories: variable deletion or imputation. Under the first treatment, a variable will be excluded from the analysis if its percentage of missing values exceed some predetermined threshold. The drawback to this method is that there is no objective optimal threshold.

Under the last treatment, missing values are replaced with some value through the use of some robust statistical method. The advantage of this method is that it is statistically motivated and backed by numerous existing computational techniques. The disadvantage is that some methods may require heavy computational overhead cost.

The current analysis employs a mixture of both methods. Following (Dong and Peng, 2013), variables with more than 25% missing observations were excluded from the analysis. This brought the number of variables to 271, of which 180 belong to traditional data and 91 contained alternative data.

The rest of the missing observations were imputed using the median, because it is less sensitive to any outliers in the data. Also, in their analysis of various imputation techniques, (Sessa and Syed, 2017) concluded that using the median for imputation proved to be effective.

2.2.2 Variable Reduction

The next step was to reduce the number of variables for modeling purposes. When a model is built using a large number of variables, the relationship between the dependent and the explanatory variables becomes difficult to ascertain. This is further compounded when some of the explanatory variables are redundant. It destabilizes model parameters, increases computational time and confounds interpretation, (Nelson, 2010).

Additionally, working in high dimensional spaces presents two unique challenges: first, geometric properties and interpretations are far removed and counter-intuitive to the traditional two and three dimensional spaces. Second, current data analysis tools, including various optimization and learning algorithms, are designed to run on and interact with lower dimensional features[5].

There are various algorithms used to select a subset of variables from some high dimensional data. An example is the variable clustering procedure, (Svolba, 2017). The idea is to group variables in clusters such that those with similar

[5]See (Verleysen and François, 2005)

information, measured by the correlation, belong to the same cluster, while those with different information are in other clusters. It works by maximizing the variance explained by the cluster components across all clusters[6].

The main drawback of the variable clustering algorithm is that the stopping criteria depends on the specified input of the user. The algorithm used in this study provide two options for the stopping criteria:

1. It stops depending on the number of maximum clusters
2. It stops based on the percentage of explained variance

Because both criteria are specified by the user *before* any analysis, I used the "elbow method" to bypass the issue of the stopping criteria being dependent on the user. The elbow method is a graphical representation of the relationship between the percentage of variance explained across clusters, plotted on the y-axis, versus the number of clusters generated, plotted on the x-axis.

As the relationship progresses, there will be an angle, the elbow, in the graph, after which there is little or no change in the percentage of variance explained. The number of clusters corresponding to the elbow will be chosen as the number of clusters to use in the variable reduction process. This method was used in (Purnima and Arvind, 2014) to ascertain the optimal k in their k-means algorithm on sensor nodes used to detect variation in environmental temperature and pressure.

Figure 2.1 shows the result of the elbow method on the alternative data

[6]See (Sarle, 2014)

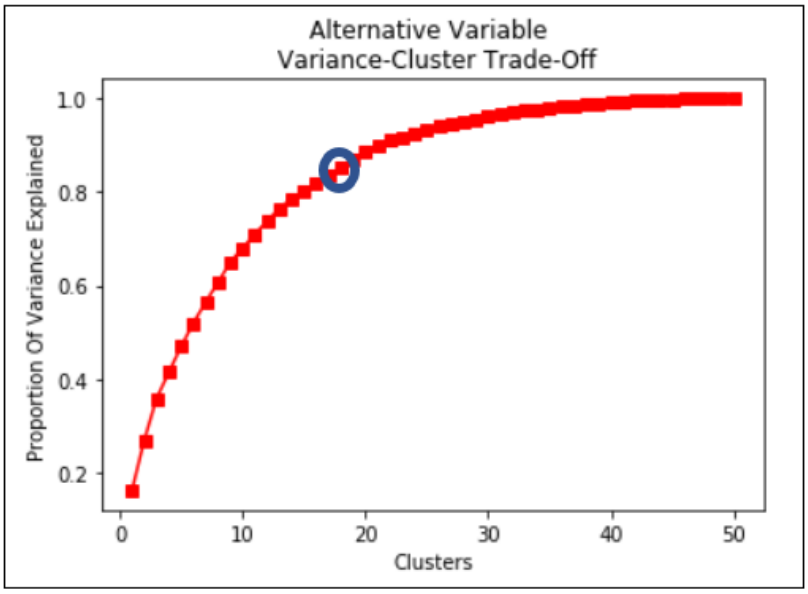


Figure 2.1: Alternative Data Variable Clustering

Figure 2.2 shows the result of the elbow method on the traditional data

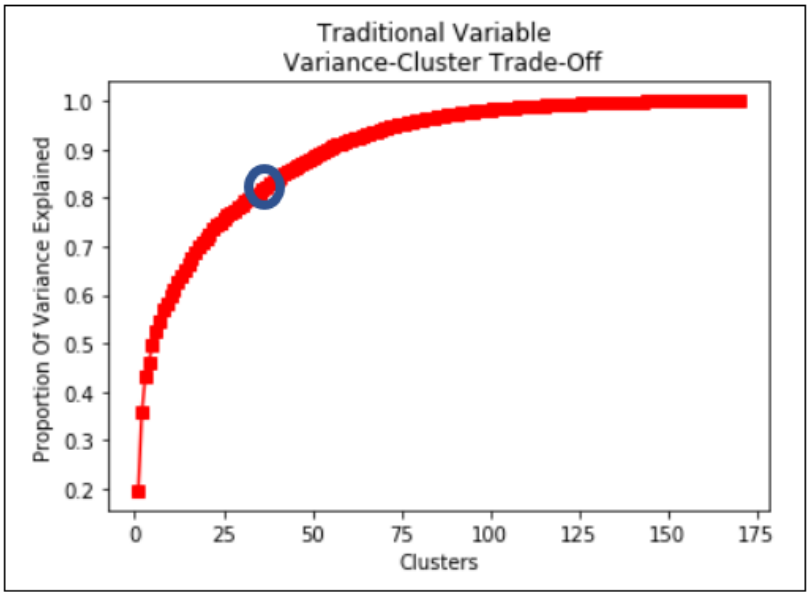


Figure 2.2: Traditional Data Variable Clustering

For the alternative data, the optimal number of clusters is 16, which explains roughly 84% of variation in the data. For the traditional data, the optimal number of clusters is 32, which explains roughly 80% of variation in the dataset. The obtained variables and their definitions are in Table 5.1 and Table 5.2 of the Appendix. From the variable clustering methodology, it can be seen that the alternative data is more parsimonious, compared to the traditional data.

2.3 Modeling Methodology

2.3.1 Nested Logistic Regression With The Likelihood Ratio Test

This section is principally concerned with two objectives. The first is to determine if alternative data can be used for credit decisions in the absence of traditional data. To that end, I will investigate whether alternative data carries its own predictive power. That is, in the absence of traditional data, can alternative data stand? The second objective is to observe some alternative data variables that may carry information pertaining to borrower default behavior.

For the first objective, I will use a Nested Logistic Regression model along with the Likelihood Ratio Test. The Likelihood Ratio Test, with its intuition based on the likelihood function, is a hypothesis test that tries to ascertain whether a restricted statistical model explains a data as well as a fully unrestricted statistical model[7]. First, the logistic regression function for the unrestricted model, URM, is defined as follows:

[7] See (Godfrey, 1996)

$$\frac{e^{\boldsymbol{X\beta}}}{1+e^{\boldsymbol{X\beta}}} \tag{2.5}$$

where $\boldsymbol{X} = x_1, x_2, ...x_k, x_{k+1}..., x_n$ corresponds to variables in the combined data and $\boldsymbol{\beta} = \beta_1, ..., \beta_k, \beta_{k+1}..., \beta_n$ represents a vector of real valued parameter in the parameter space of the combined data.

Let $x_1, x_2, ...x_k$ correspond to variables in the traditional data and $x_{k+1}..., x_n$ represent variables in the alternative data.

Let $\beta_1..., \beta_k = 0$ be the restriction imposed on the parameter space of the combined data. The result is that a model, RM_{alt}, constructed on this restricted space contains only variables in the alternative data. Then, the hypothesis underlying the likelihood ratio test is stated as follows:

$$\begin{aligned} H_0 &: \beta_1, ..., \beta_k = 0 \\ H_1 &: \beta_1, ..., \beta_k \neq 0 \end{aligned} \tag{2.6}$$

Similarly, let $\beta_{k+1}..., \beta_n = 0$ be the restriction imposed on the parameter space of the combined data. The result is that a model, RM_{trad}, constructed on this restricted space contains only variables in the traditional data. Then, the hypothesis underlying the likelihood ratio test is stated as follows:

$$\begin{aligned} H_0 &: \beta_{k+1}, ..., \beta_n = 0 \\ H_1 &: \beta_{k+1}, ..., \beta_n \neq 0 \end{aligned} \tag{2.7}$$

If the restriction is valid, that is, if the restricted model, RM_{alt} or RM_{trad}, explains the data as well as the unrestricted model, URM, then there should not be a large change in their respective log likelihood values. If L_U and L_R represents the log likelihood value of the unrestricted and restricted models

respectively, then the test is based on the difference of the likelihood values, $L_U - L_R$.

The logistic regression result for the unrestricted model, URM, is provided in Table 2.1. It should be noted that Table 2.1 contains only statistically significant variables and that URM consists of variables from the combined data.

Variables	**Estimates**	**Std. Error**	**Wald Chi.Sq**	**Pr > Chi.Sq**	**Marginal Effect (%)**
Intercept	1.6221	0.2611	38.5826	<.0001	-
VI01	-0.0207	0.00336	38.0401	<.0001	-0.3482%
AOT01	0.00972	0.00457	4.5194	0.0335	0.1635%
NI01	0.0697	0.00951	53.6648	<.0001	1.1721%
UT01	-0.1594	0.0299	28.3176	<.0001	-2.6808%
AT01	-0.0611	0.0192	10.0961	0.0015	-1.0281%
CN01	-0.0281	0.00913	9.4742	0.0021	-0.4726%
AP01	0.000226	0.000056	16.0066	<.0001	0.0038%
PA06	-0.0501	0.0129	14.9896	0.0001	-0.8423%
IV06	-0.0738	0.0238	9.5706	0.002	-1.2410%
AO01	0.00115	0.00034	11.5084	0.0007	0.0194%
RF01	-0.1927	0.0704	7.4854	0.0062	-3.2425%
CLA01	0.00367	0.00157	5.4385	0.0197	0.0617%
NBK01	0.000142	0.000064	4.8611	0.0275	0.0024%
RT01	-0.203	0.0992	4.1865	0.0407	-3.4153%
ADN	0.3783	0.038	99.0643	<.0001	6.3647%
LN01	-0.4079	0.1469	7.7049	0.0055	-6.8617%
SJY	-0.1245	0.0346	12.921	0.0003	-2.0944%
PCV02	0.2239	0.0455	24.1651	<.0001	3.7668%
BK01	0.2385	0.0635	14.1277	0.0002	4.0127%

Table 2.1: Logistic Regression Result for Unrestricted Model

Similarly, the logistic regression result for model RM_{alt} is shown in Table 2.2. As with the unrestricted model above, model RM_{alt} show variables that are statistically significant.

Variables	**Estimates**	**Std. Error**	**Wald Chi.Sq**	**Pr > Chi.Sq**	**Marginal Effect (%)**
Intercept	1.8105	0.252	51.6273	<.0001	-
RT01	-0.293	0.0975	9.024	0.0027	-5.0243%
ADY	0.3267	0.0347	88.396	<.0001	5.6031%
LN01	-0.3877	0.1451	7.1409	0.0075	-6.6482%
SJY	-0.1767	0.0337	27.4335	<.0001	-3.0297%
NMN	-0.0962	0.0429	5.0254	0.025	-1.6505%
PCV01	0.1974	0.0414	22.721	<.0001	3.3845%
BK01	0.3063	0.0619	24.4827	<.0001	5.2530%

Table 2.2: Logistic Regression Result for Alternative Data

In like manner, the logistic regression result for model RM_{trad} is shown in Table 2.3. As with both URM and RM_{alt}, Table 2.3 show variables that are statistically significant.

Variables	Estimates	Std. Error	Wald Chi.Sq	Pr > Chi.Sq	Marginal Effect (%)
Intercept	1.2236	0.0524	545.2487	<.0001	-
VI01	-0.0209	0.00332	39.3921	<.0001	-0.3535%
AOT01	0.0094	0.00455	4.2572	0.0391	0.1593%
NI01	0.0718	0.00948	57.3623	<.0001	1.2165%
UT01	-0.1549	0.0297	27.173	<.0001	-2.6250%
AT01	-0.0579	0.0192	9.1063	0.0025	-0.9807%
CN01	-0.0286	0.00907	9.9388	0.0016	-0.4845%
AP01	0.000251	0.000057	19.7232	<.0001	0.0043%
PA06	-0.0535	0.0129	17.2532	<.0001	-0.9059%
IV06	-0.0754	0.0238	10.0479	0.0015	-1.2775%
AO01	0.00126	0.000337	14.023	0.0002	0.0214%
CLA01	0.00407	0.00156	6.8216	0.009	0.0690%
NBK01	0.00015	0.000065	5.249	0.022	0.0025%

Table 2.3: Logistic Regression Result for Traditional Data

From the results in Table 2.1, it can be seen that alternative variables make up around 35% of all significant variables that serve as a good predictor for default. This means that alternative variables can have sizable explanatory power along with traditional variables to gauge default probability for individuals who may not have sufficient data to be scored using traditional methodology.

The extent to which this explanatory power holds is even more telling when the results are interpreted using the marginal effect. For an additional increase in bankruptcy, the probability of loan default increases by 4.01%.

For an additional increase of individuals whose home phone number is not valid, the probability of default increases by almost 4%. Also, for an additional group of borrowers who do not have a valid address, the probability of default increases by 6.37%.

This is important because a phone number and a valid home address can be seen as a sign of stability. Having the same working phone and a valid home

address, especially for a long period of time, may be an indicator of a consistent payment history. This is something that lenders can look on as a positive proxy for credit worthiness.

Surprisingly, last name changes in the last 60 days does not increase the probability of default. This can be attributed to the fact that name changes can take place for various reasons. A popular example is marriage, where an individual may choose to take on the name of their spouse.

In terms of traditional variables, an additional family residential house decreases the probability of default by -3.42%. A residential family house is additionally seen as a sign of stability, especially when there are children involved.

2.3.2 A Discussion on Incremental Variables

In many statistical settings, a common problem is to select a subset of the variables that inform model performance. Although having many variables may improve model performance[8], they also introduce computational costs and increase model complexity.

For a logistic regression, the maximum likelihood estimator, by design, does not factor in the number of parameters. In other words, the maximum likelihood estimator does not penalize the model for complexity.

With this in mind (Akaike, 1974) and (Schwarz, 1978) developed a criteria to measure whether performance of models are due to incremental variables or not. This was done by attaching a penalization term to the maximum likelihood estimator. In the case of (Akaike, 1974), for each model i, let k and $M_i(x_1, \ldots, x_n)$ be its dimension and maximum likelihood for variables $x_1, \ldots, x_n$.

[8] Assuming additional models are not noise

(Akaike, 1974) chooses the value of k for which $log\left[M_i(x_1,\ldots,x_n)-k\right]$ is the largest. (Schwarz, 1978) offered a different alternative by selecting a model for which $log\left[M_i(x_1,\ldots,x_n)-\frac{1}{2}k*log(n)\right]$ is the largest. Although they differ in their penalization term, research shows that for nested models, BIC is superior[9]. A lower BIC indicates that a model is preferred.

Because the likelihood ratio test depends on the value of the log-likelihood, a summarization of all models, their log likelihoods, c-stat/AUC and BIC are reported in Table 2.4. Looking at the AUC alone, one may be forced to the conclusion that the unrestricted model explains the data better than the two restricted models[10]. However, it can also be the case that the superior performance is a function of the incremental number of variables.

Since the BIC penalizes models with more parameters, I use it to test this objection. From the results in Table 2.4, the unrestricted model has the smallest BIC, followed by the traditional data model and the alternative data model, respectively. This means the incremental variables in the unrestricted model carry information that helps to explain default behavior. In other words, the increase in model performance, is due to information rather than noise.

Models	**Description**	**Log Likelihood**	**c-stat/ AUC**	**BIC**
URM	Number of variables: 48	-12,378.845	0.6134	25,252
RM_{trad}	Number of variables: 32	-12,464.22	0.5982	25,261
RM_{alt}	Number of variables: 16	-12,607.781	0.5619	25,387

Table 2.4: Nested Logistic Regression

[9] See (Wang and Liu, 2006)
[10] See the DeLong test in Section 5.2

2.3.3 Likelihood Ratio Test Set-Up

Let $x_1, x_2, x_3, \ldots, x_{32}, x_{33}, \ldots, x_{48}$ represent the variables in model URM and $\beta_1, \beta_2, \beta_3, \ldots, \beta_{32}, \beta_{33}, \ldots, \beta_{48}$ correspond to the coefficients estimated through Maximum Likelihood Estimation in the unrestricted model. Further, let $x_1, x_2, ...x_{32}$ correspond to variables in the traditional data and $x_{33}..., x_{48}$ represent variables in the alternative data. Then, model RM_{trad} is nested in model URM by restricting $\beta_{33}, \ldots, \beta_{48} = 0$.

For the likelihood ratio test, let L_U and L_{trad} represent the log-likelihood value for URM and RM_{trad}. Then the hypothesis test is stated as follows:

$$\begin{aligned} H_0 &: \beta_{33}, ..., \beta_{48} = 0 \\ H_1 &: \beta_{33}, ..., \beta_{48} \neq 0 \end{aligned} \tag{2.8}$$

The likelihood ratio is defined as

$$\kappa = 2(L_U - L_{trad}) \tag{2.9}$$

Where, $\kappa \sim \chi^2$ with degrees of freedom equal to the number of restricted variables imposed. For the traditional data,

$$\begin{aligned} L_U &= -12,378.845 \\ L_{trad} &= -12,464.220 \\ \kappa &= 2(-12,378.845 + 12,464.220) = 170.75 \sim \chi^2(df = 48 - 33 + 1 = 16) \end{aligned}$$

This gives a $p - value < 0.00001$, which is statistically significant. I therefore reject the null hypothesis, indicating that among individuals who are credit invisible, the restricted model based on the variables in the traditional data alone explains credit default behavior as well as the unrestricted model, based on the combined data.

Now, a similar restriction will be placed on the variable space of the combined data. Except this time, model RM_{alt} is nested in model URM by restricting $\beta_1, \ldots, \beta_{32} = 0$. Then, the hypothesis test is stated as follows:

$$\begin{aligned} H_0 &: \beta_1, ..., \beta_{32} = 0 \\ H_1 &: \beta_1, ..., \beta_{32} \neq 0 \end{aligned} \tag{2.10}$$

For the alternative data, the likelihood ratio test is as follows

$$\begin{aligned} L_U &= -12,378.845 \\ L_R &= -12,607.781 \\ \kappa &= 2(-12,378.845 + 12,607.781) = 457.872 \sim \chi^2(df = 32 - 1 + 1 = 32) \end{aligned}$$

This gives a $p - value < 0.00001$, which is statistically significant. I therefore reject the null hypothesis, indicating that among individuals who are credit invisible, the restricted model based on the variables in the alternative data alone explains credit default behavior as well as the unrestricted model, based on the combined data.

Putting the two results together, per the log-likelihood ratio test, restricting the dataset to traditional or alternative variables alone explains default risk as well

as the combined data. In other words, each dataset have their own explanatory power and are therefore not "substitutes" of each other.

2.3.4 Testing Performance Metrics Of Machine Learning Based Models

In this section, the focus is on devising a way for machine learning models to make a statistical statement about a model's performance in a manner similar to the Log-Likelihood Ratio Test. Specifically, this section explores the use of AUC as the goodness of fit metric in order to make more inference regarding alternative and traditional data. The AUC was selected because it is widely used across many classification settings[11].

Machine learning models struggle to make statistical statements about model performance. For a classification task, metrics such as accuracy, recall and AUC exists to measure a model's ability to discriminate between classes. However, a major drawback is that there is no uniform way to statistically test these metrics, especially when they are derived from different models or classifiers.

While there are no shortage of machine learning based models, this section will focus on the use of Random Forests, k-Nearest Neighbor and Support Vector Machines because of their widespread use in credit modeling (Lessmann et al., 2015).

The drawback of these models is that they do not provide a robust statistical test that can differentiate between model performance. Therefore, I design a way to test their performance in a manner similar to the likelihood test. The following steps provides a summary.

[11] See (Natole, Ying, and Lyu, 2019) and (Zhu, H. B. Zhang, and Huang, 2017)

1. Pick the performance metric of interest - AUC
2. Define restricted and unrestricted hypothesis
3. Bootstrap n sub-samples from data
4. Select traditional and alternative data
5. Run machine learning model n times on traditional and alternative data
6. Collect the metric for each estimation and compute difference
7. Run a paired non-parametric test on the observed metric difference

Bootstrap sampling has been widely adopted for the use of hypothesis testing[12]. In a bootstrap sampling, S random samples of size B are drawn from the original population with replacement, each of which is used to compute some test statistic of interest.

(R. Davidson and MacKinnon, 2007) assert that bootstrap sampling has two drawbacks: the choice of an optimal sample size B and the number of repetition used in the analysis. Addressing these challenges can be infeasible because their computational cost can be very high, in terms of the time it takes to run the algorithm.

For a discussion on the relationship between the optimal sample size, repetition amount and algorithm run time, see (Donald W.K Andrews and Buchinsky, 2000) and (Donald W.K. Andrews and Buchinsky, 2001).

However, the standard rule of thumb is that the sample size and repetition size should be sufficiently large. Because there is no definition for "largeness" in statistics, I selected a repetition of 1,000 along with a sample size of 1,500. All

[12]See (Kuhn and Johnson, 2013)

models tested were done using five-fold cross validation to check the robustness of this choice.

Following the steps outlined above, the performance of interest chosen is the AUC score. The table below shows the average AUC for all the models tested across three different datasets.

Data	**Metric**	**Model**	**Average AUC**
Traditional Data	AUC	Logistic Regression	0.5683
		Random Forest	0.5677
		Support Vector Machines	0.5677
		k-Nearest Neighbor	0.5338
Alternative Data	AUC	Logistic Regression	0.5439
		Random Forest	0.5484
		Support Vector Machines	0.5418
		k-Nearest Neighbor	0.5232
Combined Data	AUC	Logistic Regression	0.5798
		Random Forest	0.5721
		Support Vector Machines	0.5764
		k-Nearest Neighbor	0.5350

Table 2.5: Bootstrap AUC Summary

From Table 2.5, the average bootstrapped AUC results for the combined data are higher than those of alternative and traditional data across all models tested. This is followed by average bootstrapped AUC results for the traditional data. This seems to agree with the results in Table 2.4, showing that a better credit scoring model may be achieved when both alternative and traditional data are combined.

However, it is also evident that the contribution of alternative data cannot be overlooked considering that the percentage difference, across all models, between traditional data and alternative data is roughly 2%. Similarly, the percentage difference, across all models, between combined data and alternative data is roughly 3%.

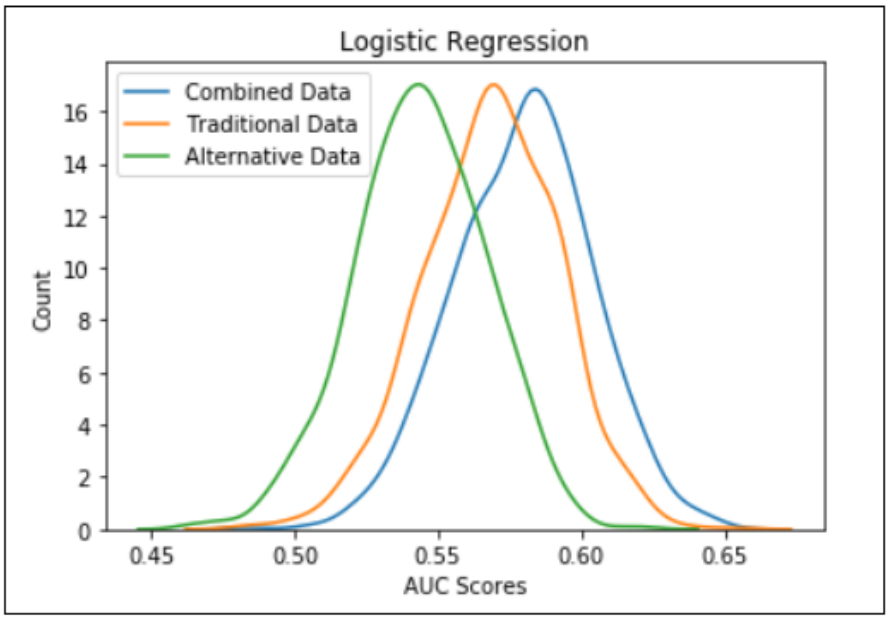


Figure 2.3: Logistic Regression AUC Distribution

2.3.5 AUC Bootstrap Distribution

This section analyzes the bootstrapped distributions of the datasets across all models. Figure 2.3 and Figure 2.4 show the distributions of the bootstrap AUC for the Logistic Regression and Random Forest models.

The bootstrap AUC metrics generated from the Logistic Regression model using all three types of data shows a distinctive difference. The average AUC measurements from the combined data is higher than that of the traditional and alternative data.

This is the general pattern that is also reflected in the results of the SVM and kNN models below in Figure 2.5 and Figure 2.6. In all cases, the mean bootstrapped AUC for the alternative data is less than that of the traditional and combined data.

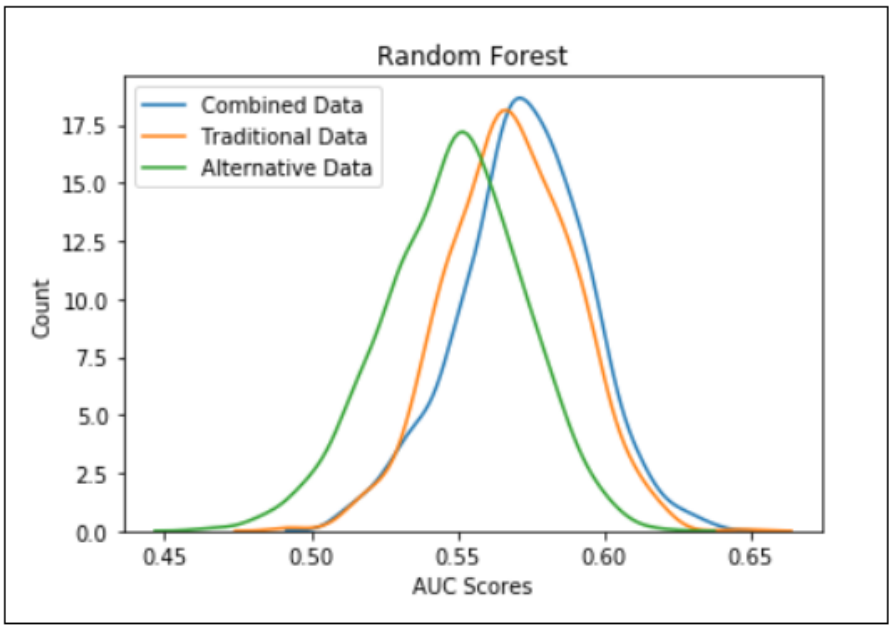


Figure 2.4: Random Forest AUC Distribution

This is not implausible; if I take the AUC as a measurement of how well a classifier discriminates between two groups based on the information presented in the group, a more established information - as represented by the traditional data - will have a greater AUC than a less established credit information - as represented by the alternative data.

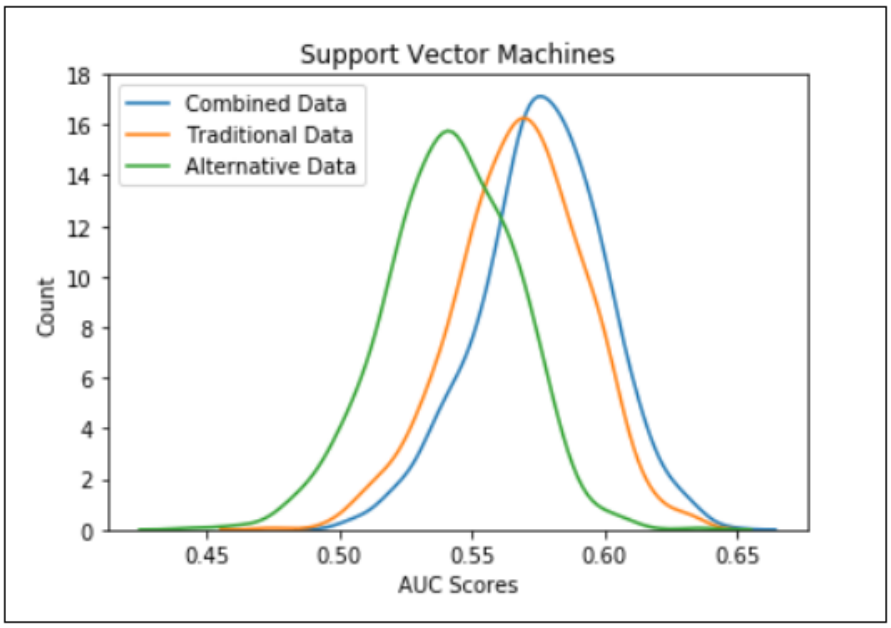


Figure 2.5: SVM AUC Distribution

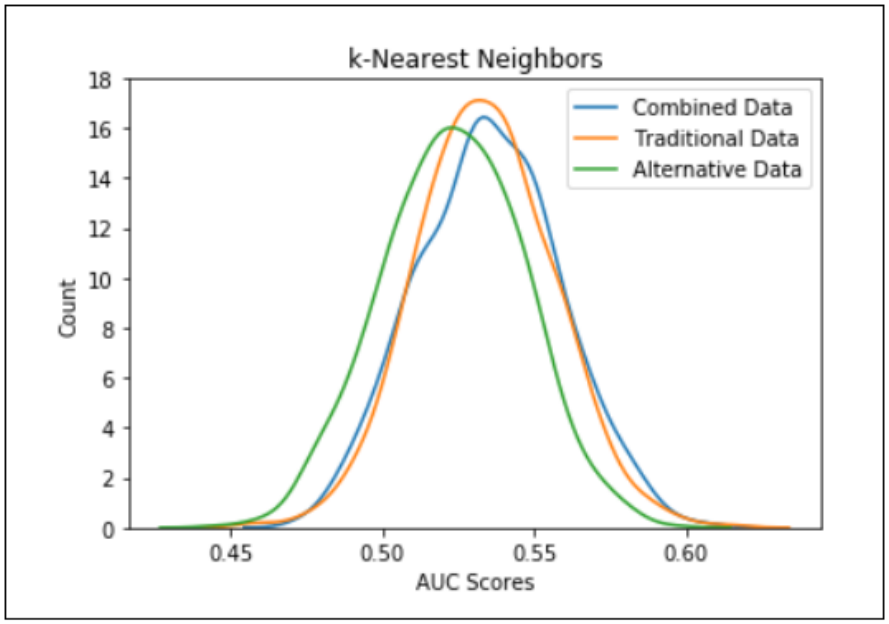


Figure 2.6: kNN AUC Distribution

The key takeaway from Figure 2.3 through Figure 2.6 is that both traditional and alternative data provide their own unique information in the credit building process.

2.3.6 AUC Bootstrap Hypothesis Testing

The previous section answered the question of the difference between AUC metrics generated by the three types of dataset using four different classification models. This was largely done through the construction of AUC distributions. This section looks at the same problem by following the procedure highlighted in Subsection 2.3.3. Here, the non-parametric test chosen is the Wilcoxon Signed Rank test because it is widely used in analysis that involves the AUC[13] and also because it does not impose distributional assumptions about the underlying data.

Most importantly, the Wilcoxon Signed Rank test was chosen over the Mann-Whitney-Wilcoxon test because the bootstrapping procedure introduces dependence structure in the collected AUC metric. It should be noted that the Wilcoxon Rank Test is the paired version of the Mann-Whitney-Wilcoxon test. Developed by (Wilcoxon, 1945), the Wilcoxon Signed Rank test is used to ascertain whether two groups of paired measurement come from the same distribution or not.

In other words, it tests whether there is a difference between two groups of paired measurement. Under the null hypothesis, the two groups are considered to come from the same distribution, while the alternative hypothesis states that the two groups do not come from the same distribution.

[13]See Colak et al. (2012)

The Wilcoxon Signed Rank test follows three assumptions of the Mann-Whitney-Wilcoxon test[14]. However, because it is the paired version of the Mann-Whitney-Wilcoxon test, there is an additional assumption about the dependence structure of the data[15]. The first assumption is that the two groups are randomly sampled. Although the data under consideration is proprietary, it was randomly sampled from a larger population to reduce sampling error.

The second assumption dictates that each observation be independent. For the data under consideration, there are no duplicate observations. That is, each individual appears once in the data. The third assumption says that the scale of the population being tested be either continuous or ordinal. Since the values of the AUC are continuous, this assumption is met as well.

The fourth assumption relates to the dependence nature of the data structure. Because the AUC metrics were computed using bootstrapped samples, dependency is introduced in the data. Putting it all together, the underlying assumptions of the Wilcoxon Signed Rank test has been met.

For each machine learning model, I use the Wilcoxon Signed Rank test to test the difference in bootstrapped AUC between the combined and traditional data as well as the difference between the combined data and alternative data. For the purpose of clarity, let CD be "combined data", TD be "traditional data" and AD be "alternative data". The null and the alternative hypothesis can be stated as follows

$$H_0 : AUC_{CD} - AUC_{TD} = AUC_{CD} - AUC_{AD}$$

$$H_1 : AUC_{CD} - AUC_{TD} \neq AUC_{CD} - AUC_{AD}$$

[14] See (Nachar, 2008)
[15] See (Wilcoxon, 1945)

Model	Hypothesis	p-Value
Logistic Regression	$H_0 : AUC_{CD} - AUC_{TD} = AUC_{CD} - AUC_{AD}$	$1.040 * 10^{-89}$
	$H_1 : AUC_{CD} - AUC_{TD} \neq AUC_{CD} - AUC_{AD}$	
Random Forest	$H_0 : AUC_{CD} - AUC_{TD} = AUC_{CD} - AUC_{AD}$	$1.321 * 10^{-61}$
	$H_1 : AUC_{CD} - AUC_{TD} \neq AUC_{CD} - AUC_{AD}$	
Support Vector Machine	$H_0 : AUC_{CD} - AUC_{TD} = AUC_{CD} - AUC_{AD}$	$1.057 * 10^{-87}$
	$H_1 : AUC_{CD} - AUC_{TD} \neq AUC_{CD} - AUC_{AD}$	
k-Nearest Neighbor	$H_0 : AUC_{CD} - AUC_{TD} = AUC_{CD} - AUC_{AD}$	$3.580 * 10^{-19}$
	$H_1 : AUC_{CD} - AUC_{TD} \neq AUC_{CD} - AUC_{AD}$	

Table 2.6: AUC Bootstrap Hypothesis Testing

The small p-values in Table 2.6 gives a strong indication to reject the null hypothesis. In this case, it means that there is a difference in the bootstrapped AUC between the combined and traditional as well as the difference between the combined and alternative data. This confirms the case that each data brings its own different information in the credit modeling process that helps determine who is credit worthy and who is not.

2.4 Conclusion

This chapter of the dissertation is motivated by the use of non-traditional economic data points that may be useful for building *default based* credit models when traditional borrower information is scarce. The consequences of not having enough financial information is significant in two main ways. First, lack of information often determines whether individuals who are credit invisible will receive a higher price of credit or be denied credit entirely.

Second, there is a strong interest from regulatory agencies regarding the use of alternative data points that carries similar economic implications that are useful for lending decisions. From a regulatory perspective, the purpose is to gradu-

ally build credit models that will inform policies on how to slowly incorporate individuals who are credit invisible into the general financial network.

Because alternative credit data have been shown to help in this space, I combined it with traditional credit data to ascertain the credit worthiness of borrowers. Using the likelihood ratio test through the logistic regression, I find that alternative variables can complement traditional variables to gauge default behavior for those who are credit invisible.

For further analysis of this finding, I used three machine learning models - random forest, support vector machines and k-nearest neighbor - to describe default behavior for individuals who are credit invisible. To make any statement about the complimentary nature of the alternative and traditional data, I established a way to statistically test the goodness of fit statistic that is similar in spirit to the likelihood ratio test.

For this, I constructed bootstrap distributions of the goodness of fit statistic, as well as fitting a function to that distribution. The conclusion from the machine learning models is that the alternative and traditional data each bring unique elements to the modeling process that contributes positively to modeling default behavior for individuals who are credit invisible.

Additionally, I am able to relate which alternative variables carry information pertaining to credit default. For example, when an unbanked consumer does not have a valid home phone number the probability of default increases by almost 4%. Also, when an unbanked borrower does not have a valid home address, the probability of default increases by 6.37%.

Chapter 3

Scoring The Unscored: A Profit-Based Approach

Abstract

This study is an extension of Chapter 2. Here, I look at borrower profitability by extending the maximum utility approach of (Lieli and White, 2010) to the subprime lending space. In Chapter 2, I established a *statistical* significance and the role of non-traditional credit data in modeling borrower default behavior for individuals who are credit invisible. In this chapter, I analyze the *economic* importance of incorporating alternative data in the credit modeling process. Using a maximum utility approach, I show that there is an economic value in alternative data. Additionally, this chapter advocates for the use of loss functions that aligns with a lender's business objective of making a profit.

Index Terms – Profit Scoring, Profit Maximization, Utility Maximization

3.1 Introduction

The goal in the previous chapter was to shed some light building credit models for individuals who are credit invisible. The proposed solution was to use non-traditional financial data to serve as a proxy for default behavior. The underlying assumption was that a lender's decision to accept or reject a borrower is *solely* influenced by the borrower's default probability.

However, in this chapter, I suggest that the lender is willing to consider and integrate profit into their credit building process. More specifically, I investigate if non-traditional data can be used to construct models that are *economically* viable for the lender. If the conclusion is in the affirmative, then it give lenders economic incentive to invest and use non-traditional data in their credit granting policies. Additionally, it gives policymakers another tool to use as they consider how to solve the problem of assimilating borrowers who are credit invisible into the mainstream economy.

A lender's credit granting policies directly influence their profit or loss: a more restrictive policy will approve fewer borrowers and may generate fewer default losses but also less revenue. However, a less restrictive policy will approve more borrowers, generating higher revenues but also higher losses. It is therefore imperative for a lender to strike a healthy balance between default rates, losses given default and gains given loan repayment. This is true especially when the goal is to extend credit to individuals who are credit invisible.

Underlying all credit modeling methodologies are loss functions that will either be minimized or maximized depending on the problem of study and its application. For a classification problem, the most popular loss function is the Maximum Likelihood Estimate. Its popularity is due to its optimal properties

during parameter estimation[1].

While that approach is extremely useful, it only helps answer the question of whether a potential borrower will default or not in probabilistic terms. However, if I assume, as most businesses do, that the lender seeks to maximize economic profit or utility[2], then the lender's profitability objective is simply a consequence of a borrower's loan outcome of default or not default. For example, a lender can only assess profitability *after* the decision to accept or reject a loan.

The drawback in traditional credit default modeling is that it does not *directly* incorporate the lender's profitability objective into the model building process. The process in which that is done in credit modeling is called *profit scoring*. Profit scoring seeks to maximize a loss function[3] that is more in line with the lender's profitability objective[4].

Profit scoring poses two major drawbacks. The first concerns the issue of *how* to design the utility function - should it be designed on a micro or macro level? For example, in terms of a loan application, profit can be measured on a per loan basis or on a portfolio of loans. Should it capture direct or indirect profit? That is, should the function measure most or all intermediary nuances through which that profit was attained? This will include cost accounting for IT infrastructure, overhead, occupancy costs, etc.

The second drawback is a consequence of the business model. In this context, a profit metric is by definition, business-centric. That is, there is a need to be well versed in the business model and even in the industry as a whole in order to derive a custom utility function that can effectively measure profit. By this

[1] See (Myung, 2003)
[2] The term "profit" and "utility" will be used interchangeably.
[3] The term "loss function", "utility function" and "objective function" will be used interchangeably
[4] See (Thomas, 2009)

logic, a custom utility function designed to capture profits for an airline industry may not be suitable in the medical industry. Therefore it is also important to study default in the context of profits as opposed to probabilistic terms.

The outline of this chapter is as follows: Section 3.2 discusses the theoretical foundations of a lender who seeks to maximizes profit. Section 3.4highlights the data and the methodology while Section 3.4 presents the conclusion.

3.2 The Profit Objective Of The Lender

This section establishes the theoretical foundation needed for the profit scoring approach. I will argue that it is important to use a loss function that is appropriate for the application under study. Additionally, I will illustrate that a variable cutoff value can be derived from the lender's loss function, which will be useful to make lending decisions.

I follow the structure and notations presented in (Lieli and White, 2010) and make the following definitions

- Let $\lambda > 0$ represent the loan principal
- Let $r \in (0, 1)$ be the interest rate on the loan
- Let t be the maturity on the loan

Therefore, a loan is characterized by a vector $\ddot{X} = (\lambda, r, t)$. Assume further that the lender only issues a conventional loan that is payable in equal monthly installment and that in the event of default, the lender stands to recover a fraction of the principal. Define this recovery rate as $q \in [0, 1]$.

Let $\tilde{X}$ denote features from the sample data that is sufficient to predict borrower default risk. Then, I define the binary default risk variable as follows

$$Y = \begin{cases} -1 & \text{bad borrower} \\ +1 & \text{good borrower} \end{cases} \tag{3.1}$$

where a good borrower is defined to be those who did not default on the loan, while a bad borrower defaulted. It is important to realize that Y, the outcome variable is only observed at loan maturity. In other words, the lender does not definitively know Y at the time of the loan origination. I make the following definitions

- Let $D = \{A,\ R\}$ represent the lender's decision to *accept* or *reject* a loan application
- Let $\pi_{\{d \in D, y \in Y\}}$represent the lender's profitability metric as a consequence of their decision to accept or reject a loan application and their classification of the borrower at maturity as good or bad

Then, Table 3.1 gives an overview of the framework of the lender's profitability. The lender is only profitable when they accept a "good borrower", as measured in the accept/ no default quadrant of Table 3.1, and stands to lose money when they accept a "bad borrower", as measured in the accept/ default quadrant of Table 3.1. In the case where a borrower is rejected, the profit is defined to be zero. In other words, profit is calculated only for those applicants who were accepted.

Although loan profitability is conditioned on the lender's decision and the loan outcome, it is also a function of borrower default features in the sample data as

well as the characteristics of the loan contract, such as loan rate, amount and maturity.

	No Default (Y=1)	**Default (Y=-1)**
Accept (A)	$\underset{A,1}{\pi}\left(\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right)\geq 0$	$\underset{A,-1}{\pi}\left(\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right)<0$
Reject (R)	$\underset{R,1}{\pi}\left(\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right)=0$	$\underset{R,-1}{\pi}\left(\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right)=0$

Table 3.1: Lender's Profitability Scenarios

For a conventional loan with equal payments, I choose the Net Present Value (NPV) to play dual roles. First, it acts as the lender's profitability metric and secondly, it serves as the custom loss function. For such a loan, let CF_k represent the cashflow or the equal installment payable at time k, where $k \subseteq t$. Then, the NPV is defined as

$$\pi = \sum_{i=1}^{i=t} \frac{CF_k}{(1+r)^i} - \lambda \tag{3.2}$$

Rearranging Equation 3.2 gives the following

$$\pi = CF_k \sum_{i=1}^{i=t} (1+r)^{-i} - \lambda \tag{3.3}$$

Because the lender is only profitable when they accept a "good borrower", the accept/ no default quadrant in table Table 3.1 is calculated as follows

$$\underset{A,1}{\pi}\left(\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right) = CF_k \sum_{i=1}^{i=t} (1+r)^{-i} - \lambda > 0 \tag{3.4}$$

That is, the lender *always* stands to make a profit when they accept a good borrower. However, in the event of a default, the assumption is that the lender

can recover only a fraction of the loan. In this case, the lender can invest the recovered amount in a risk-free government note of the same maturity as the loan. If I denote the return on the government note as r_g, then it should be noted that $r_g < r$.

The loss associated with the default event is calculated as follows

$$\underset{A,-1}{\pi}\left(\tilde{X} = \tilde{x}, \ddot{X} = \ddot{x}\right) = q * \lambda(1 + r_g)^{-t} - \lambda < 0 \tag{3.5}$$

Rearranging Equation 3.5 gives the following

$$\underset{A,-1}{\pi}\left(\tilde{X} = \tilde{x}, \ddot{X} = \ddot{x}\right) = \lambda\left(q * (1 + r_g)^{-t} - 1\right) < 0 \tag{3.6}$$

Because the lender cannot observe Equation 3.1 at the time of the loan application, their best approach is to make a decision based on expected profit or loss. Let $\alpha = P(Y = 1|\tilde{X} = \tilde{x}, \ddot{X} = \ddot{x})$ denotes the probability of not defaulting on the loan. Then, the lender needs to make a decision such that expected profit is maximized. That is,

$$\underset{d\in\{A,R\}}{max} E\left(\underset{D,Y}{\pi}|\tilde{X} = \tilde{x}, \ddot{X} = \ddot{x}\right) = \underset{d\in\{A,R\}}{max}\left\{\alpha \underset{A,1}{\pi} - (1 - \alpha)\underset{A,-1}{\pi}\right\} \tag{3.7}$$

Therefore, a lender will accept a borrower's loan application if there is an expected economic gain. This is stated as follows

$$\alpha \underset{A,1}{\pi} - (1 - \alpha)\underset{A,-1}{\pi} > 0 \tag{3.8}$$

Here, α, known as the cutoff value, serves the purpose of regulating whom the lender will accept or deny. Solving for it gives the following equation

$$\alpha = \frac{\pi_{A,-1}}{\pi_{A,-1} + \pi_{A,1}} \stackrel{\text{def}}{=} c(\ddot{x}) \in (0,1) \tag{3.9}$$

Simplifying that gives Equation 3.10

$$\alpha = \left(1 + \frac{\pi_{A,1}}{\pi_{A,-1}}\right)^{-1} \stackrel{\text{def}}{=} c(\ddot{x}) \in (0,1) \tag{3.10}$$

In the traditional scoring approach, the cutoff value is motivated by past business experiences and it is often expressed as a constant number. However, as a direct consequence of using Equation 3.3 as the loss function, Equation 3.9 and Equation 3.10 presents a variable cutoff that is *function* of the loan characteristics. As a result, two important features emerge

1. The cutoff function is *directly* tied to the lender's profitability objective. Therefore, the lender controls whom to extend credit to based on their custom loss function

2. The cutoff function uses borrower loan characteristics. This means that borrowers with different loan contracts will also have different cutoff value

In addition to the benefits highlighted above, Equation 3.9 carries an intuitive economic interpretation: the variable cutoff per applicant is a ratio between possible losses when they default and potential gains when they do not default.

3.3 Parameter Estimation

3.3.1 Estimation of Likelihood Parameters

According to (Myung, 2003), parameter estimation falls under two broad categories: least-squares estimation (LSE) and maximum likelihood estimation (MLE). MLE is widely used and recognized because of its statistical properties. Additionally, MLE is often the basis for tests such as the chi-square, AIC, BIC and the G-squared test.

Following the notation of (Myung, 2003), let $X = x_1, \ldots, x_n$ and $W = w_1, \ldots, w_n$ represent a given data and real valued parameters. Then a probability density function (pdf), gives the probability of observing X given W. Mathematically, the pdf of observing each observation is expressed as if X is independently and identically distributed

$$f(X \Big| W) = f_1(x_1 \Big| w_1) * f_2(x_2 \Big| w_2) * f_3(x_3 \Big| w_3) * \ldots * f_n(x_n \Big| w_n)$$

Given X, a likelihood function seeks to find a pdf that is most likely to have generated the data. Let$L(W \Big| X)$ be the likelihood function, then the relationship between $L(W \Big| X)$ and $f(X \Big| W)$ is given as follows:

$$L(W \Big| X) = f(X \Big| W)$$

For maximum likelihood estimation, the idea is to find parameters W that maximizes the likelihood function. As an optimization problem, it is stated as

$$\underset{W}{argmax} \prod_{i=1}^{i=n} f_i(x_i|w)$$

However, for computational efficiency, the log of the likelihood is maximized as follows:

$$\underset{W}{argmax} \sum_{i=1}^{i=n} log\left(f_i(x_i|w)\right)$$

Let $y_i \in \{0, 1\}$, then for a logistic regression the probability distribution is given as

$$P(Y_i \Big| X_i) = \left(\frac{e^{\sum W \cdot X}}{1 + e^{\sum W \cdot X}}\right)^{y_i} * \left(1 - \frac{e^{\sum W \cdot X}}{1 + e^{\sum W \cdot X}}\right)^{1-y_i}$$

The likelihood function is

$$L(W \Big| X) = (1 - y_i) * log\left(\frac{1}{1 + e^{\sum W \cdot X}}\right) + y_i * log\left(\frac{e^{\sum W \cdot X}}{1 + e^{\sum W \cdot X}}\right)$$

$$L(W \Big| X) = \begin{cases} log\left(\frac{e^{\sum W \cdot X}}{1+e^{\sum W \cdot X}}\right) & y_i = 1 \\ log\left(\frac{1}{1+e^{\sum W \cdot X}}\right) & y_i = 0 \end{cases}$$

3.3.2 Estimation of Parameters That Maximizes Profit

Let $\tau(\tilde{x}, \ddot{x}; \theta)$ be some parametric model having $T(\tilde{x}, \ddot{x}; \theta) \in [0, 1]$ as the CDF. Although the exact nature of $T(\tilde{x}, \ddot{x}; \theta)$ is not known, I will interpret its values to be a conditional[5] probability. As in (Lieli and White, 2010), $T(\tilde{x}, \ddot{x}; \theta)$ is selected to be the logistic distribution because currently, it is the industry standard[6]. Then a profit maximizing lender will prefer to maximize profit based on a decision to accept or reject an applicant. This is stated as a linear programming problem in Equation 3.11

[5] conditioned on $\tilde{X}$

[6] See (Baesens, Rösch, and Scheule, 2016)

$$\underset{d\in\{A,R\}}{max}\, E\left(\underset{D,Y}{\pi}|\tilde{X}=\tilde{x},\ddot{X}=\ddot{x}\right)$$
$$\text{subject to } d=\begin{cases} R & T(\tilde{x},\ddot{x};\theta^*)\leq c(\ddot{x}) \\ A & T(\tilde{x},\ddot{x};\theta^*)>c(\ddot{x}) \end{cases} \tag{3.11}$$

Therefore, the lender seeks to find an optimum parameter θ^* that solves Equation 3.11. (Elliott and Lieli, 2013) show that, Equation 3.11 can be written as

$$\underset{\theta^*\in\Theta}{max}\, E\left\{b(Y+1-2c(\ddot{x}))*sign(T(\tilde{x},\ddot{x};\theta^*)-c(\ddot{x}))\right\} \tag{3.12}$$

where

$$sign(v)=\begin{cases} 1 & v>0 \\ -1 & v\leq 0 \end{cases} \tag{3.13}$$

and Y takes the same form as Equation 3.1 and $b=\pi_{A,-1}+\pi_{A,1}$ corresponds to the denominator of Equation 3.9. The benefit of Equation 3.12 is that it shows clearly the role of the variable cutoff and total profit as it pertains to the lender's profit maximizing objective. In the next section, I will discuss the data and methodology used for the discussion above.

3.4 Data and Methodology

The data used in this section has been described thoroughly in Section 2.2 of Chapter 2. However, loan specific information such as the amount, rate and

maturity were assigned to each borrower, in absence of observed data, based on relevant literature and plausibility of the underlying application. As such, no claim is made that they are realistic. Nevertheless, sensitivity analysis in the Appendix show that the results are robust.

According to (Adams, Einav, and Levin, 2009) credit scores and loan sizes are negatively correlated. That is, if the lender determines that the borrower is a low default risk, as measured by a high credit score, then the lender can be confident to give the borrower a higher loan amount. However loan amounts should also take into consideration the borrower's income and other debt(s). I do this by introducing the debt-to-income (DTI) ratio.

According to Experian[7], a DTI ratio should be at or below 40% of the borrower's income for auto loans. Therefore, the maximum allowable loan amount assigned to an applicant was 40% of their income. Loan amounts were assigned based on borrower's relative default risk classification, as measured by their credit score.

I constructed this ranking, shown in Table 3.2, by dividing the credit score into quartiles, with low risk individuals in the first quartile, followed by medium, etc. Using this table, a borrower is who has a low default risk (as measured by their credit score) will receive the maximum allowable loan amount, while a borrower who has a high default risk will receive the lowest loan amount.

Credit Quartile	**Default Risk**	**Loan Amount**
First	High	(10%) * Income
Second	Upper medium	(20%) * Income
Third	Lower medium	(30%) * Income
Fourth	Low	(40%) * Income

Table 3.2: Default Risk Classification For Loan Amount

Auto loans have fixed maturities of 36 months, 48 months, 60 months and 72

[7] https://www.experian.com/blogs/ask-experian/how-to-get-a-car-loan/#s3

months. Unlike loan sizes, borrower default risk, as measured by a credit score, and loan maturity are positively correlated[8]. That is, the greater a borrower's credit score, the more time they will be assigned to repay the loan because the lender trusts them to hold unto the loan for a longer period.

Similar to the loan sizes, I create borrower default risk by dividing their credit score into quartiles, with low risk individuals in the first quartile, followed by medium, etc. Loan maturity was assigned according to the default risk of the borrower. For example, borrowers with low default risk have 72 months to repay the loan, etc. This is shown in Table 3.3.

Credit Quartile	**Default Risk**	**Loan maturity**
First	High	36 months
Second	Upper medium	48 months
Third	Lower medium	60 months
Fourth	Low	72 months

Table 3.3: Default Risk Classification For Loan Maturity

The loan rate, was calculated as being 10% more than the risk free rate[9]. However, because of the subprime nature of the population under study, I used a loan rate of 11%, 15% and 20% above the risk free rate as well. The result of this analysis is shown in the Appendix.

Following (Lieli and White, 2010), I make the assumption that recovery rates and loan amount are inversely related. That is, a lender stands to recover more if the borrowed amount is relatively small. For (Lieli and White, 2010), the reason for this assumption was to "further emphasize the role of a variable cutoff." The table below shows the recovery rate for loan amount $10,000\lambda$.

[8] See (Kuvíková, 2015)

[9] https://www.treasury.gov/resource-center/data-chart-center/interest-rates/pages/textview.aspx?data=yield

Loan Amount	Recovery Rate
$\lambda \leq 0.4112$	0.95
$0.4112 < \lambda \leq 0.8223$	0.90
$0.8223 < \lambda \leq 1.2335$	0.85
$1.2335 < \lambda \leq 1.6446$	0.80
$1.6446 < \lambda \leq 2.0558$	0.75
$2.0558 < \lambda \leq 2.4670$	0.70
$2.4670 < \lambda \leq 2.8781$	0.65
$2.8781 < \lambda \leq 3.2893$	0.60
$3.2893 < \lambda \leq 3.7004$	0.55
otherwise	0.50

Table 3.4: Recovery Rates For Loan Amounts

3.4.1 Alternative Data Profit Results

The theoretical framework outlined in Subsection 3.3.2 requires the specification of a benchmark parametric model that will be used to predict the probability of default. In this section, I present the result of the logistic regression model for the alternative data in Table 3.5. It should be noted that the variables were presented and discussed in Table 2.2 of Chapter 2.

Variables	Estimates	Std. Error	Wald Chi.Sq	Pr > Chi.Sq
Intercept	1.4911	0.1494	99.636	<.0001
RT01	-0.3592	0.0859	17.4913	<.0001
ADY	0.3414	0.0341	100.164	<.0001
LN01	-0.411	0.1436	8.1906	0.0042
SJY	-0.1841	0.0334	30.349	<.0001
NMN	-0.1036	0.0413	6.2999	0.0121
PCV01	0.2164	0.0401	29.1804	<.0001
BK01	0.3082	0.0618	24.8376	<.0001

Table 3.5: Logistic Regression Results for Alternative Data

Next, I present the result of the profit based approach. To this end, 80% of the data was used for in-sample calculations while 20% was used for out-of-sample confirmation. The results of the following models were built and compared

1. Model I - a logistic regression model with a constant cutoff value
2. Model II - a logistic regression model with a variable cutoff based on Equation 3.10
3. Model III - a profit based model based on Equation 3.12
    (a) Because of the non-smooth nature of Equation 3.11, the Simulated Annealing algorithm was used to solve the optimization problem.

For each model the following metrics were also computed

1. Accept Ratio (AR) - Proportion of applications that were accepted
2. Reject Ratio (RR) - Proportion of applications that were rejected
3. P(A|G) - Proportion of good borrowers that were accepted
4. P(B|R) - Proportion of bad borrowers that were rejected
5. Profit - Average profit per application
6. CI.Profit - The confidence interval of the profits

Following, (Lieli and White, 2010), because I have no way of knowing the lender's explicit cutoff value, the constant cutoff value under Model I was chosen such that the acceptance ratio under Model I and Model II will be roughly similar. The entire exercise was repeated 250 times and their average results are shown Table 3.6

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.223	0.882	$154.60	$153.67, $155.52
Logistic Regression	Variable	0.992	0.008	0.224	0.914	$157.87	$156.97, $158.85
Profit Based	Variable	0.916	0.084	0.233	0.888	$192.18	$191.24, $193.14
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.222	0.880	$152.39	$148.87, $156.13
Logistic Regression	Variable	0.992	0.008	0.223	0.915	$155.85	$151.91, $159.47
Profit Based	Variable	0.916	0.084	0.231	0.874	$180.40	$176.87, $183.34

Table 3.6: Alternative Data Profit Scoring Result

From the result in Table 3.6 it can be seen that directly incorporating the lender's profitability objective into the credit scoring process has the potential to yield higher profits, compared to traditional methodologies. It is also worthwhile to note that compared to the other approaches, the profit based method accepted fewer applicants but recorded larger profit. Considering the credit-invisible nature of the borrowers, this means that from a risk management perspective, a lender does not have to be overexposed to be profitable. Additionally, in terms of the research question, it shows that there is measurable monetary value in alternative data and that it is in the lender's best interest to consider incorporating it into the lending practice.

3.4.2 Traditional Data Profit Results

In this section, I present the result of the logistic regression model for the traditional data. The variables used in the modeling process were presented and discussed in Table 2.3 of Chapter 2. The results are shown in Table 3.7

Variables	Estimates	Std. Error	Wald Chi.Sq	Pr > Chi.Sq
Intercept	1.1994	0.0445	727.7369	<.0001
VI01	-0.0194	0.00279	48.6375	<.0001
AOT01	0.0125	0.00407	9.506	0.002
NI01	0.0745	0.00691	116.1444	<.0001
UT01	-0.1518	0.028	29.4561	<.0001
AT01	-0.032	0.0168	3.6165	0.0572
CN01	-0.0408	0.0058	49.3962	<.0001
AP01	0.000282	0.000054	27.0795	<.0001
PA06	-0.0479	0.012	15.9544	<.0001
IV06	-0.0569	0.0151	14.1019	0.0002
AO01	0.00125	0.000332	14.1293	0.0002
CLA01	0.0059	0.00146	16.3265	<.0001
NBK01	0.000151	0.000053	8.244	0.0041

Table 3.7: Traditional Data Logistic Regression Results

Next, I present the result of the profit based approach outlined in Section 3.2. All metrics were explained in Subsection 3.3.1. It should be noted that the entire exercise was repeated 250 times and their average results are shown in Table 3.8

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.889	$188.66	$187.83, $189.49
Logistic Regression	Variable	0.926	0.074	0.231	0.890	$190.41	$189.49, $191.38
Profit Based	Variable	0.878	0.120	0.238	0.892	$200.99	$200.18, $201.81
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.890	$184.60	$181.27, $188.21
Logistic Regression	Variable	0.926	0.074	0.231	0.890	$186.91	$184.02, $189.85
Profit Based	Variable	0.879	0.120	0.236	0.877	$189.67	$186.75, $192.07

Table 3.8: Traditional Data Profit Scoring Result

As with the alternative data, a lender stands to gain more profit when they incorporate their profitability objective into the credit modeling process. I also observe the same phenomenon of accepting fewer applications but being more

profitable when compared to the other models.

3.4.3 Combined Data Profit Results

In this section, I present the result of the logistic regression model for the traditional data. The variables used in the modeling process were presented and discussed in Table 2.1 of Chapter 2. The results are shown in Table 3.9

Variables	**Estimates**	**Std. Error**	**Wald Chi.Sq**	**Pr > Chi.Sq**
Intercept	1.3986	0.1565	79.8297	<.0001
RT01	-0.2718	0.0875	9.6523	0.0019
ADY	0.3446	0.0344	100.0674	<.0001
LN01	-0.4088	0.1452	7.925	0.0049
SJY	-0.133	0.0341	15.1818	<.0001
NMN	-0.0517	0.0419	1.5249	0.2169
PCV01	0.1877	0.0407	21.3124	<.0001
BK01	0.2389	0.0632	14.2986	0.0002
VI01	-0.0189	0.00282	45.0946	<.0001
AOT01	0.0131	0.00408	10.3753	0.0013
NI01	0.0723	0.00692	109.1011	<.0001
UT01	-0.1558	0.0281	30.7083	<.0001
AT01	-0.0316	0.0169	3.5169	0.0607
CN01	-0.0393	0.00587	44.7416	<.0001
AP01	0.000269	0.000054	24.7097	<.0001
PA06	-0.0429	0.012	12.7325	0.0004
IV06	-0.0505	0.0152	11.0225	0.0009
AO01	0.0012	0.000335	12.7652	0.0004
CLA01	0.00551	0.00147	14.0756	0.0002
NBK01	0.000144	0.000052	7.6199	0.0058

Table 3.9: Combined Data Logistic Regression Results

Next, I present the result of the profit based approach outlined in Section 3.2. All metrics were explained in Subsection 3.3.1. It should be noted that the entire exercise was repeated 250 times and their average results are shown in Table 3.10

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.233	0.886	\$184.62	\$183.77, \$185.45
Logistic Regression	Variable	0.913	0.087	0.234	0.898	\$202.68	\$201.83, \$203.57
Profit Based	Variable	0.833	0.167	0.247	0.900	\$221.44	\$220.54, \$222.26
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.232	0.886	\$180.36	\$177.03, \$184.07
Logistic Regression	Variable	0.913	0.087	0.233	0.897	\$197.83	\$194.53, \$201.23
Profit Based	Variable	0.833	0.167	0.243	0.881	\$201.97	\$198.46, \$204.11

Table 3.10: Combined Data Profit Scoring Result

As in Subsection 3.3.1 and Subsection 3.3.2, there is an economic gain when the lender considers a profit metric as a basis for making credit decisions. Moreover, the combined traditional and alternative credit data has economic value that is seen in the *magnitude* of the profit. That is, there is a progression of profit value with alternative data having the least amount to the combined data having the highest amount. This is seen in both in-sample and out-of-sample data. A similar phenomenon was seen in Table 2.4 of Chapter 2 when the AUC was used as the performance metric.

3.5 Conclusion

This chapter of the dissertation takes a profitability approach into the question of constructing credit models when traditional borrower information is scarce. More specifically, it is concerned with credit models that directly incorporates the lender's profitability objective into the lending process.

This indirectly implies that minimizing the borrower default behavior may not be the same as maximizing the lender's profit. This is important because any

rational lender will seek to optimize or at least be concerned about profit when dealing with individuals who are credit invisible.

Using the NPV as the profitability measure, I show that directly incorporating the lender's profitability objective into the credit scoring process has the potential to yield higher profits, compared to existing methodologies. Compared to the other approaches, the profit based method accepted fewer applicants but recorded superior profit.

This is important because considering the credit-invisible nature of the borrowers, this means that from a risk perspective, a lender does not have to be overexposed to be profitable. Additionally, in terms of the research question, it shows that there is measurable monetary value in alternative data and that it is in the lender's best interest to consider incorporating it into the lending practice.

Moreover, the combined alternative and traditional credit data have a greater economic value that is seen in the *magnitude* of the profit. That is, the lender gains the highest profit when they incorporate both alternative and traditional data into the credit building process. Of course, more research is needed considering that some of the data was simulated.

Chapter 4

Conclusion

Chapter 2 of the dissertation is motivated by investigating the statistical usefulness of non-traditional data for credit modeling when traditional borrower information is scarce or unavailable. The consequences of not having enough financial information is significant in two main ways. First, lack of information often determines whether individuals who are credit invisible will receive a higher price of credit or be denied credit entirely.

Second, there is a strong interest from regulatory agencies regarding the use of alternative data points that can be used to predict default risk for lending decisions. From a regulatory perspective, the purpose is to gradually build credit models that will inform policies on how to slowly incorporate individuals who are credit invisible into the mainstream economy.

Because alternative credit data has been shown to help in this space, I contribute to the debate and further show other alternative data variables that can act as proxies for credit default risk. Using the likelihood ratio test, through a

nested logistic regression model, I find that alternative variables can complement traditional variables in order to gauge default behavior for those who are credit invisible. Additionally, alternative variables contain unique information that can help predict default risk.

For further analysis of this finding, I used three machine learning models, random forest, support vector machines and k-nearest neighbor, to describe default behavior for individuals who are credit invisible. To make any statement about the complementary nature of the alternative and traditional data, I developed a way to statistically test the goodness of fit statistic that is similar in spirit to the likelihood ratio test.

For this, I constructed bootstrap distributions of the goodness of fit statistic for machine learning models in order to make statistical statements about their performance. The conclusion from the machine learning models was that the alternative data brings unique elements to the modeling process that contributes to modeling default behavior for individuals who are credit invisible. This is in agreement with the results of the log-likelihood ratio test.

Also, I am able to relate which alternative variables carry information pertaining to credit default. For example, when an unbanked consumer does not have a valid home phone number the probability of default increases by almost 4%. Also, when an unbanked borrower does not have a valid home address, the probability of default increases by 6.37%.

Chapter 3 of the dissertation takes a profitability approach to the same research question. More specifically, it is concerned with credit models that directly incorporates the lender's profitability objective into the lending process. This is important because any rational lender will seek to optimize or at least be concerned about profit when dealing with individuals who are credit invisible.

The implication is that minimizing borrower default behavior may not be the same as maximizing the lender's profit.

Using the NPV as the profitability measure, it is shown that directly incorporating the lender's profitability objective into the credit scoring process has the potential to yield higher profits, compared to existing methodologies. Compared to the other approaches, the profit based method accepted fewer applicants but recorded superior profit.

This is important because considering the credit-invisible nature of the borrowers, it means that from a risk management perspective, a lender does not have to be overexposed to be profitable. Additionally, in terms of the research question, it shows that there is measurable monetary value in alternative data and that it is in the lender's best interest to consider incorporating it into the lending practice.

Moreover, the combined alternative and traditional credit data have a greater economic value that is seen in the *magnitude* of the profit. That is, the lender gains the highest profit when they incorporate both alternative and traditional data into the credit building process.

It should be noted that some data points in Chapter 3 were simulated. Although sensitivity analysis in the Appendix appears to be robust, the results should be interpreted with caution, especially when the out of sample results appears somewhat weaker.

Chapter 5

Future Research

Because of the age of big data, credit risk modeling is evolving. Current research suggests that incorporating non-traditional data sources have the potential to improve existing models. In most cases, novel methodologies will be required to deal with the influx of unseen data.

I am interested in exploring web and social media-based data to enrich the modeling process. Some of the questions I am interested in include the extent to which online relationships contribute to default. Specifically, I plan to use network-based metrics such as centrality, communities and cliques as variables in the model to ascertain their usefulness. This is important because it may even help identify a network of defaulters.

Another interest of mine is to explore the role of other profitability metrics for profit scoring. Although the current analysis used the NPV, there may be other metrics in the literature. The real contribution here is to explore optimization techniques that will be able to maximize profit.

Using any classification technique, an interesting topic to explore is to use a grid-search approach to find an optimum cut-off point and compare the corresponding profit to that of the profit-based approach. Additionally, the idea of using a loss function that minimizes expected loss rather than maximizing the expected profit is very interesting.

Also, I am interested in extending the work to multiple periods. For example, credit credits are known to have revolving balances. They do not follow the same loan structure as a conventional loan. Because they show multiple periods, the contribution lies in modeling multi-period profits. It will be a novel profit function that has the capability to capture profits from multiple period.

Appendix

5.1 Variable definitions

Table 5.1 gives the variable names and meaning for the result of the variable clustering procedure on the alternative data in Subsection 2.2.2

Alternative Variable	**Meaning**
RT01	Number of bank routing number changes in the last 90 days
ADY	Affirmative indication of a valid address
ADN	Negative indication of a valid address
HPH01	Number of home phone number changes in the last 30 days
WPH01	Number of work phone number changes in the last 30 days
LN01	Number of last name changes in the last 60 days
NHPHY	Affirmative indication of a match between name and home phone number
ZP01	Number of zip code changes in the last 60 days
SNY	Affirmative indication of validity of social security number
DSY	Affirmative indication that a deceased identity is found with name and/or SSN
SJY	Affirmative indication of evictions, liens, judgments and suits
NMN	Negative indication that name and address match
WSY	Affirmative indication that SSN was issued before date of birth
AC01	Number of bank account changes in the last 15 days
PCV01	Affirmative indication that home phone number is valid
PCV012	Negative indication that home phone number is valid
PD01	Affirmative indication of a match between phone and address
BK01	Affirmative indication of a bankruptcy

Table 5.1: Alternative Data Variables

Table 5.2 gives the variable names and meaning for the result of the variable clustering procedure on the traditional data in Subsection 2.2.2

Traditional Variable	Meaning
VI01	Number of vehicle inquiries within 24 months
CMD01	Number of non-medical collections assigned within 24 months
AOT01	Number of trades open over 6 months
NI01	Number of inquiries within 30 days
UT01	Number of utility inquiries within 12 months
CA01	Amount in collections
AT01	Number of trades opened within 12 months
CN01	Number of non-medical accounts in collections
AA01	Number of trades active within 12 months
CI01	Number of collection inquiries within 12 months
CI001	Number of inquiries within 6 months
AP01	Total payments
RB01	Revolving balance with limit opened within 12 months
CNL01	Number of non-medical accounts in collections greater than $100
VIQY	Number of vehicle inquiries within 14 days
TL12	Amount in collections assigned within 12 months
PA06	Number of trades rated 60 days past due
RVO	Bank revolving balance with limit
IT01	Number of telecommunications inquiries within 1 month
RR01	Residential type: renting
CNM01	Number of non-medical collections assigned within 12 months
IV06	Number of vehicle inquiries within 6 months
AO01	Months since oldest trade opened
BRV01	Bank revolving balance with limit opened within 24 months
CM01	Number of inquiries within 1 month
NC001	Number of collections assigned within 12 months on original balance greater than $100
RF01	Residential type: family
IC1M	Number of collection inquiries within 1 month
CLA01	Months since most recent collection assigned
RSM01	Residential type: military
NBK01	Non-bank revolving balance with limit
NBK001	Non-bank revolving balance with limit opened within 12 months

Table 5.2: Traditional Data Variables

5.2 DeLong Test for AUC Comparison

This section explores the use of the DeLong test alluded to in Subsection 2.3.2. The DeLong test[1] is a widely used method that provides a 95% confidence interval and standard errors of the difference between two or more AUCs. For comparison purposes, a visual representation of the ROC-AUC curves are provided in Figure 5.1

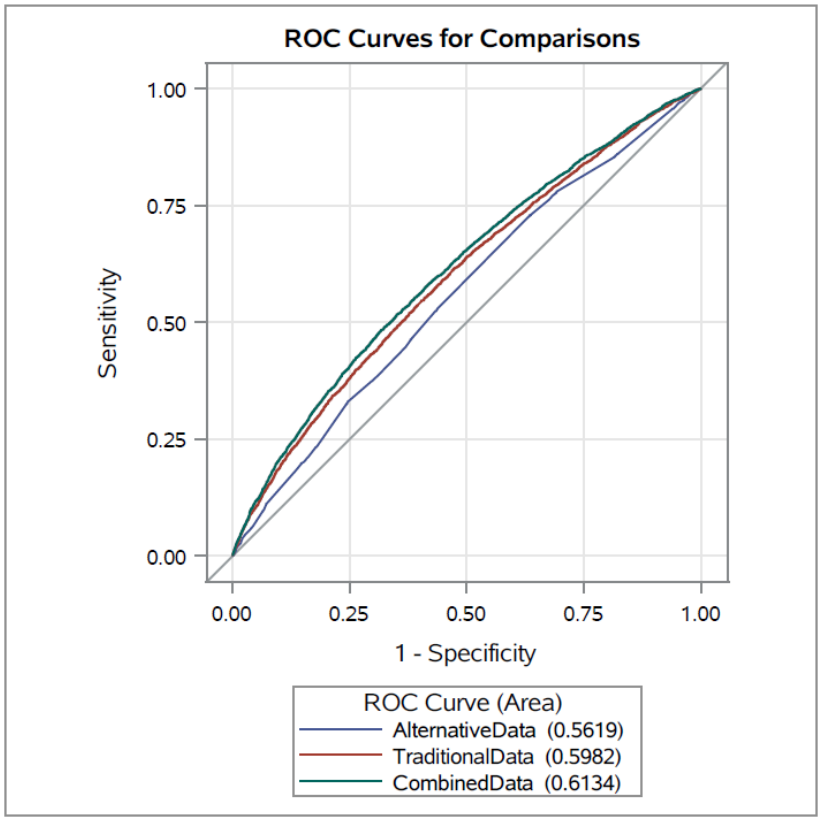


Figure 5.1: ROC Comparison

The standard error and the 95% confidence interval of the AUCs are also shown in Figure 5.2

[1]See (E. R. Delong, D. M. Delong, and Clarke-Pearson, 1988)

ROC Association Statistics							
	Mann-Whitney						
ROC Model	Area	Standard Error	95% Wald Confidence Limits		Somers' D	Gamma	Tau-a
AlternativeData	0.5619	0.00442	0.5532	0.5706	0.1238	0.1344	0.0428
TraditionalData	0.5982	0.00432	0.5897	0.6066	0.1964	0.1964	0.0679
CombinedData	0.6134	0.00428	0.6050	0.6218	0.2268	0.2268	0.0784

Figure 5.2: DeLong Test Confidence Interval

For all models, the lower limit of the confidence interval is greater than 0.5, implying that their performance is different from random guessing. Additionally, using the chi-square test with two degrees of freedom, it is statistically significant that the unrestricted model is different from at least one of the restricted models. This is shown in the table below.

ROC Contrast Test Results			
Contrast	DF	Chi-Square	Pr > ChiSq
Reference = CombinedData	2	367.4458	<.0001

Figure 5.3: DeLong Test Chi-Square

Also, from the table below it can be seen that the difference between the unrestricted and restricted models are statistically significant. This also confirms the analysis of the BIC and the log-likelihood ratio test. In terms of the research question, it means that alternative data and combined data carry unique information in the credit building process.

ROC Contrast Estimation and Testing Results by Row						
Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
AlternativeData - CombinedData	-0.0515	0.00403	-0.0594	-0.0436	163.2076	<.0001
TraditionalData - CombinedData	-0.0152	0.00275	-0.0206	-0.00982	30.5771	<.0001

Figure 5.4: DeLong Test Contrast Estimation

While the DeLong test is useful to test the difference in correlated AUCs, its value may be limited in the hypothesis testing outlined in Subsection 2.3.3. The focus of Subsection 2.3.3 is to find a way to test model metrics across different machine learning models. Here, the DeLong test may not be appropriate if I choose to test a different metric, say accuracy, precision or recall. In other words, the use of the DeLong test is a consequence of choosing the AUC as a metric, rather than a research question.

Also, the DeLong test is used to test difference in AUC metric for nested models. In Subsection 2.3.3, the goal is to test the difference of a difference in AUC (or any other metric) within the context of bootstrap sampling using machine learning models. Therefore, the DeLong test may not be applicable.

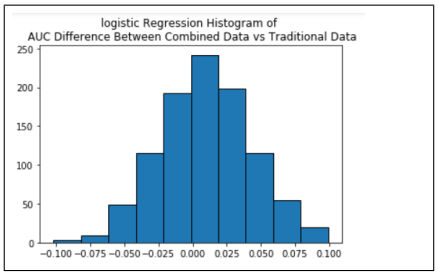
5.3 Testing Normality Assumptions

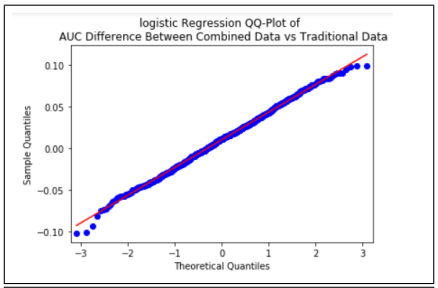
In this section, I explore whether a parametric test could be used to test the difference in AUC between the combined and traditional data, as well as the combined and alternative data for the hypothesis testing outlined in Subsection 2.3.3. Although non-parametric tests do not make distributional assumptions, it also has less power compared to their parametric counterparts.

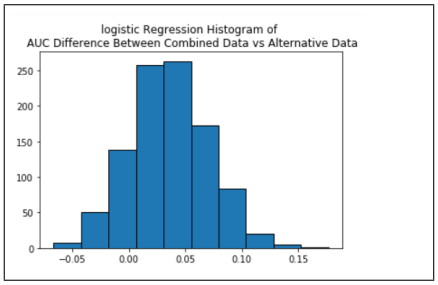
Normality tests fall under two broad categories: graphical and hypothesis tests. Under graphical tests, a researcher concludes that a sample is normally distributed by looking at histograms, Q-Q plots or box plots. While this may serve as a good starting point, it does not provide conclusive proof that the sample is normal. This is because by nature graphical interpretations are subjective and may require knowledge in statistics to be fully appreciated[2].

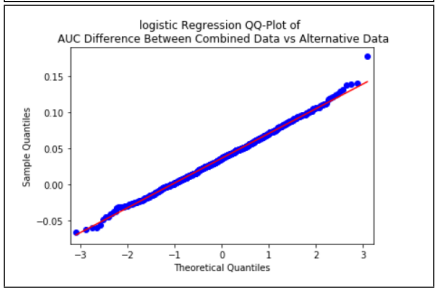
Hypothesis tests for normality assumptions tend to be more robust than their visual counterparts because they are often backed by statistical theory. Over the years, many tests have been developed to test for normality assumptions. For a good overview, see (Rani Das, 2016) and (D'Agostino and Stephens, 1987).

For this analysis, I combined both approaches. It should be noted that the goal is not to test whether AUCs generated from individual datasets are normally distributed, but whether the difference in AUC between datasets are normally distributed. For the visual approach, I used the histogram and the Q-Q plot. The results are shown below.





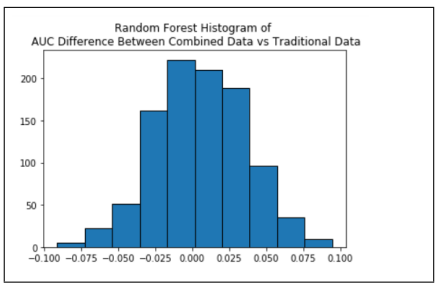


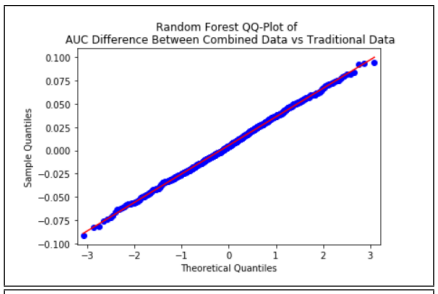


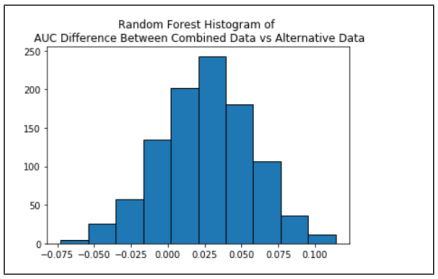
[2]See (Yap and Sim, 2011)

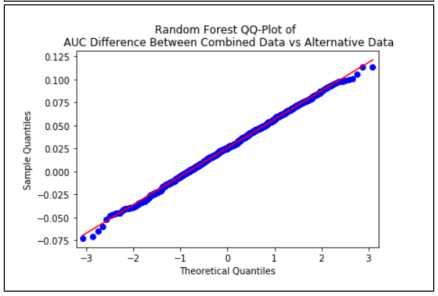
For the logistic regression, the histogram does not show large deviations from a classical normal distribution. Particularly, the histogram is not skewed. This is also confirmed by the Q-Q plot, except for small deviations on the tail. Using the visual approach, the conclusion is that the difference in AUC between the combined and traditional data is normally distributed. The same conclusion is also valid for the difference in AUC between the combined and alternative data.

For the random forest model, the histogram does not show large deviations from a normal distribution. Particularly, there is not abnormal tail behavior. This is also confirmed by the Q-Q plot, where the quantiles are shown to be on the line $y = x$, with slight deviation in the case of the combined and alternative data. Using the visual approach, the conclusion is that the difference in AUC between the combined and traditional data is normally distributed. The same conclusion is also valid for the difference in AUC between the combined and alternative data.



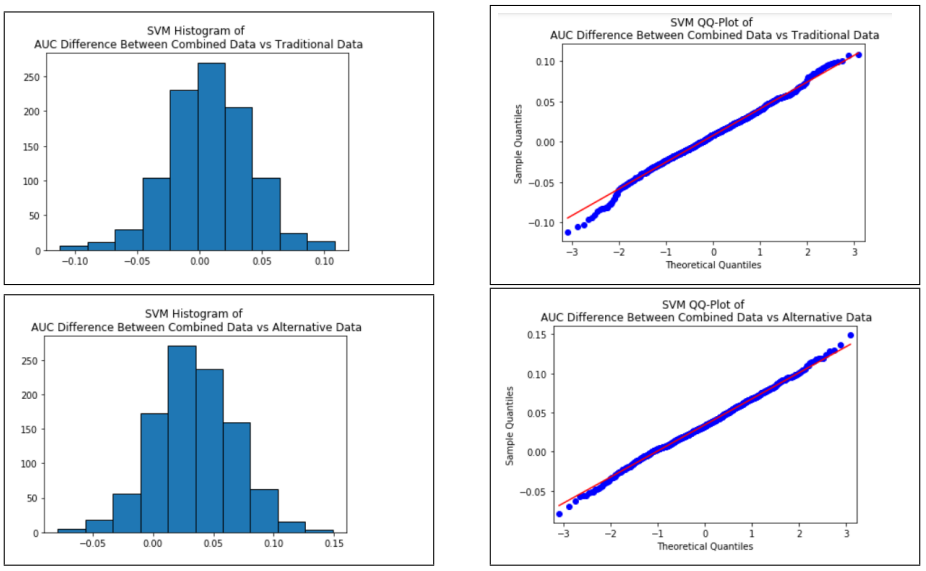




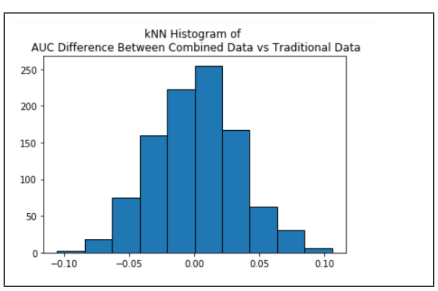


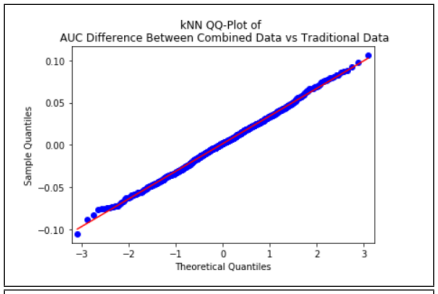
For the SVM model, the histogram does not show large deviations from a normal

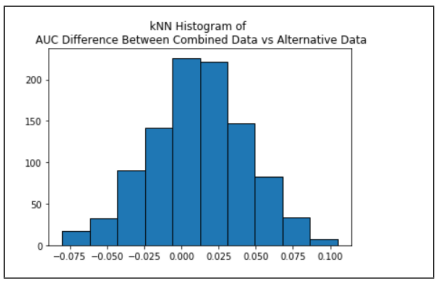
distribution. It is hard to determine if there is abnormal tail behavior. However, the Q-Q plot for the AUC difference between the combined and traditional data shows some skewed behavior. In this case, it is not conclusive that the difference in AUC between the combined and traditional data is normally distributed. The Q-Q plot for the AUC difference between the combined and alternative data does not show skewed behavior. In this case, I will conclude that the difference in AUC between the combined and alternative data is normally distributed.

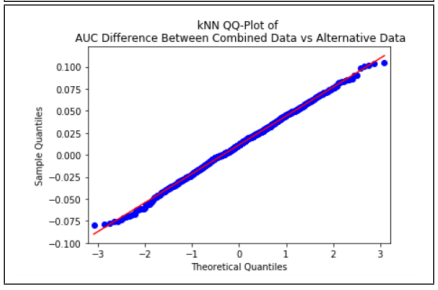


For the kNN model, the histogram does not show large deviations from a normal distribution and there is no abnormal tail behavior. The Q-Q plot confirms this as well, where the quantiles are shown to be on the line $y = x$. Using the visual approach, the conclusion is that the difference in AUC between the combined and traditional data is normally distributed. The same conclusion is also valid for the difference in AUC between the combined and alternative data.









While the visual tests for normality has been done, I also show the hypothesis test for normality assumptions. Although normality tests are many, research shows that the Shapiro-Wilk test is the preferred test[3]. For each model, I describe two hypothesis test. The null hypothesis of the first test is that the AUC difference between the combined and traditional data are normally distributed, while the alternative hypothesis states that they are not normally distributed.

The null hypothesis of the second test is that the AUC difference between the combined and alternative data are normally distributed, while the alternative hypothesis states that they are not normally distributed. The result of the hypothesis test in Table 5.3 confirms the result of the graphical test. With the exception of the difference in AUC between the combined and traditional data using SVM, I fail to reject the null hypothesis that the AUC differences are normal[4].

[3] See (Ghasemi and Zahediasl, 2012)

[4] Although I report the Shapiro-Wilk test, I also conducted a KS test, D'Agostino's k^2 test and Anderson-Darling test. Their conclusion were the same as the Shapiro-Wilk test

Model	Hypothesis	p-Value
Logistic Regression	$H_0 : AUC_{CD} - AUC_{TD} = N(\mu, \sigma)$	0.3872
	$H_1 : AUC_{CD} - AUC_{TD} \neq N(\mu, \sigma)$	
	$H_0 : AUC_{CD} - AUC_{AD} = N(\mu, \sigma)$	0.3151
	$H_1 : AUC_{CD} - AUC_{AD} \neq N(\mu, \sigma)$	
Random Forest	$H_0 : AUC_{CD} - AUC_{TD} = N(\mu, \sigma)$	0.5759
	$H_1 : AUC_{CD} - AUC_{TD} \neq N(\mu, \sigma)$	
	$H_0 : AUC_{CD} - AUC_{AD} = N(\mu, \sigma)$	0.6812
	$H_1 : AUC_{CD} - AUC_{AD} \neq N(\mu, \sigma)$	
Support Vector Machine	$H_0 : AUC_{CD} - AUC_{TD} = N(\mu, \sigma)$	0.0033
	$H_1 : AUC_{CD} - AUC_{TD} \neq N(\mu, \sigma)$	
	$H_0 : AUC_{CD} - AUC_{AD} = N(\mu, \sigma)$	0.4313
	$H_1 : AUC_{CD} - AUC_{AD} \neq N(\mu, \sigma)$	
k-Nearest Neighbor	$H_0 : AUC_{CD} - AUC_{TD} = N(\mu, \sigma)$	0.6706
	$H_1 : AUC_{CD} - AUC_{TD} \neq N(\mu, \sigma)$	
	$H_0 : AUC_{CD} - AUC_{AD} = N(\mu, \sigma)$	0.4051
	$H_1 : AUC_{CD} - AUC_{AD} \neq N(\mu, \sigma)$	

Table 5.3: Result of Normality Test

5.4 The Cost of Alternative Data

The use of alternative data in the credit space is becoming prevalent and has been shown to contribute positively to the credit lending process. For example, according to a report[5] by TransUnion, more than 317 lenders revealed that using alternative data opened opportunities in new markets, allowed them to reach more credit worthy individuals and situated them to be more competitive.

Like all data, the costs associated with alternative data may include recording, storing, analyzing and maintenance[6]. Whether the cost outweighs the benefit goes beyond the focus of this research. Although the cost associated with data may be many, I focus on computational cost. Therefore, I have conducted an empirical experiment of computational cost as a function of incremental

[5]See (TransUnion, 2015)
[6]See (Haug, Zachariassen, and Liempd, 2011)

variables in a model. Here, the assumption is that increasing the number of variables also increases the computational cost.

To achieve this, I construct a nested logistic regression model with 8,10 and 16 variables across all data, using the variable clustering algorithm used in Subsection 2.2.2. The results are shown in Table 5.4, Table 5.5 and Table 5.6. In all models, the highest AUC is seen when using the combined data followed by the traditional and alternative data.

Models	**Description**	**Log Likelihood**	**c-stat/AUC**	**BIC**
URM	Number of variables: 8	-12,449.060	0.6018	24,989
RM_{trad}	Number of variables: 8	-12,541.278	0.5856	25,173
RM_{alt}	Number of variables: 8	-12,592.499	0.5572	25,276

Table 5.4: Logistic Regression With 8 Variables

Models	**Description**	**Log Likelihood**	**c-stat/AUC**	**BIC**
URM	Number of variables: 10	-12,452.534	0.6011	25,061
RM_{trad}	Number of variables: 10	-12,525.827	0.5887	25,163
RM_{alt}	Number of variables: 10	-12,612.556	0.5521	25,336

Table 5.5: Logistic Regression With 10 Variables

Models	**Description**	**Log Likelihood**	**c-stat/AUC**	**BIC**
URM	Number of variables: 16	-12,430.282	0.6045	25,032
RM_{trad}	Number of variables: 16	-12,495.689	0.5931	25,163
RM_{alt}	Number of variables: 16	-12,607.781	0.5619	25,387

Table 5.6: Logistic Regression With 16 Variables

The key takeaway is that restricting the variable space – and therefore accounting for computational cost – results in a model where credit default behavior is maximally explained by a combined force of alternative and traditional data. This also confirms the result in the dissertation that there are elements in alternative data that contributes to the modeling of borrower default behavior.

5.5 Sensitivity Analysis

5.5.1 Profit Scoring - 500 Repetitions with loan rate 10% above risk free rate

This section presents the work in Subsection 3.4.1 through Subsection 3.4.3, except that entire exercise was repeated 500, 750 and 1000 times using a loan rate of 10% above the risk-free rate. Their results are shown below for the Alternative, Traditional and Combined Data.

5.5.1.1 Alternative Data - 500 repetitions with loan rate 10% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.223	0.882	$154.31	$153.60, $154.95
Logistic Regression	Variable	0.992	0.008	0.224	0.915	$157.64	$156.99, $158.30
Profit Based	Variable	0.914	0.086	0.233	0.887	$191.59	$191.20, $192.54
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.995	0.005	0.222	0.884	$153.50	$150.95, $156.22
Logistic Regression	Variable	0.993	0.007	0.223	0.913	$156.79	$154.20, $159.42
Profit Based	Variable	0.914	0.086	0.231	0.873	$180.34	$178.71, $183.37

Table 5.7: Alternative Data Profit Result - 500 Iterations

5.5.1.2 Traditional Data - 500 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.890	$188.54	$187.98, $189.09
Logistic Regression	Variable	0.926	0.074	0.231	0.890	$190.20	$189.54, $190.88
Profit Based	Variable	0.879	0.120	0.238	0.892	$200.67	$200.06, $201.14
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.888	$185.46	$182.95, $187.85
Logistic Regression	Variable	0.926	0.074	0.231	0.889	$187.73	$185.52, $189.74
Profit Based	Variable	0.880	0.120	0.236	0.877	$190.37	$188.48, $192.41

Table 5.8: Traditional Data Profit Result - 500 Iterations

5.5.1.3 Combined Data - 500 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.233	0.886	$184.44	$183.83, $184.99
Logistic Regression	Variable	0.913	0.087	0.234	0.898	$202.47	$201.90, $203.04
Profit Based	Variable	0.832	0.168	0.247	0.900	$221.34	$220.74, $221.90
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.232	0.884	$181.23	$178.86, $183.61
Logistic Regression	Variable	0.913	0.087	0.233	0.896	$198.60	$196.47, $200.99
Profit Based	Variable	0.833	0.167	0.243	0.880	$201.92	$198.99, $202.91

Table 5.9: Combined Data Profit Result - 500 Iterations

5.5.2 Profit Scoring - 750 repetitions with loan rate 10% above risk free

5.5.2.1 Alternative Data - 750 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.223	0.882	$154.62	$154.09, $155.15
Logistic Regression	Variable	0.992	0.008	0.224	0.915	$157.96	$157.47, $158.51
Profit Based	Variable	0.914	0.086	0.233	0.887	$191.87	$191.29, $192.38
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.995	0.005	0.223	0.883	$152.29	$150.06, $154.40
Logistic Regression	Variable	0.992	0.008	0.223	0.912	$155.61	$153.43, $157.75
Profit Based	Variable	0.914	0.086	0.231	0.874	$180.03	$178.71, $182.30

Table 5.10: Alternative Data Profit Result - 750 Iterations

5.5.2.2 Traditional Data - 750 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.890	$188.39	$187.90, $188.87
Logistic Regression	Variable	0.926	0.074	0.231	0.890	$190.07	$189.54, $190.58
Profit Based	Variable	0.879	0.121	0.238	0.891	$200.57	$200.21, $201.09
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.888	$185.94	$183.88, $187.92
Logistic Regression	Variable	0.926	0.074	0.231	0.889	$188.25	$186.60, $189.95
Profit Based	Variable	0.880	0.120	0.236	0.877	$190.52	$189.26, $192.29

Table 5.11: Traditional Data Profit Result - 750 Iterations

5.5.2.3 Combined Data - 750 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.233	0.886	$184.31	$183.87, $184.76
Logistic Regression	Variable	0.913	0.087	0.234	0.898	$202.33	$201.85, $202.85
Profit Based	Variable	0.833	0.167	0.247	0.900	$221.27	$220.78, $221.72
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.232	0.884	$181.86	$179.91, $183.93
Logistic Regression	Variable	0.913	0.087	0.234	0.896	$199.14	$197.03, $201.12
Profit Based	Variable	0.833	0.167	0.243	0.880	$201.75	$199.63, $202.99

Table 5.12: Alternative Data Profit Result - 750 Iterations

5.5.3 Profit Scoring - 1000 repetitions with loan rate 10% above risk free

5.5.3.1 Alternative Data - 1000 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.223	0.882	$154.71	$154.24, $155.16
Logistic Regression	Variable	0.992	0.008	0.223	0.915	$158.04	$157.59, $158.49
Profit Based	Variable	0.915	0.085	0.233	0.888	$192.24	$191.60, $192.58
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.994	0.006	0.223	0.881	$151.99	$150.10, $153.79
Logistic Regression	Variable	0.992	0.008	0.2232	0.912	$155.31	$153.56, $157.14
Profit Based	Variable	0.914	0.086	0.231	0.874	$180.12	$178.56, $181.69

Table 5.13: Alternative Data Profit Result - 1000 Iterations

5.5.3.2 Traditional Data - 1000 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.889	$188.36	$187.94, $188.77
Logistic Regression	Variable	0.926	0.074	0.231	0.890	$189.93	$189.45, $190.38
Profit Based	Variable	0.879	0.121	0.238	0.891	$200.63	$200.22, $201.01
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.926	0.074	0.231	0.887	$185.99	$184.32, $187.76
Logistic Regression	Variable	0.926	0.074	0.231	0.889	$188.18	$186.72, $189.57
Profit Based	Variable	0.880	0.121	0.236	0.877	$190.79	$189.10, $191.85

Table 5.14: Traditional Data Profit Result - 1000 Iterations

5.5.3.3 Combined Data - 1000 repetitions with loan rate 10% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.233	0.886	$184.31	$183.90, $184.75
Logistic Regression	Variable	0.913	0.087	0.234	0.898	$202.28	$201.86, $202.70
Profit Based	Variable	0.833	0.168	0.247	0.900	$221.34	$220.85, $221.69
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.913	0.087	0.233	0.885	$181.80	$179.98, $183.51
Logistic Regression	Variable	0.913	0.087	0.234	0.896	$199.08	$197.48, $200.73
Profit Based	Variable	0.832	0.168	0.243	0.880	$201.63	$200.21, $203.09

Table 5.15: Combined Data Profit Result - 1000 Iterations

5.5.4 Profit Scoring - Alternative Data with 250 Repetitions

This section presents the work in Subsection 3.3.1 through Subsection 3.3.3, except that entire exercise was repeated 250 times using a loan rate of 11%, 15% and 20% above the risk-free rate. The results are shown below for the Alternative, Traditional and Combined Data.

5.5.4.1 Alternative Data - 250 repetitions with loan rate 11% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.998	0.002	0.223	0.906	$219.66	$218.65, $220.70
Logistic Regression	Variable	0.997	0.003	0.223	0.870	$221.23	$220.21, $222.25
Profit Based	Variable	0.931	0.069	0.231	0.870	$251.03	$250.27, $252.39
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.999	0.001	0.222	0.897	$217.44	$213.33, $221.60
Logistic Regression	Variable	0.997	0.003	0.222	0.894	$218.94	$214.83, $223.26
Profit Based	Variable	0.931	0.069	0.230	0.879	$236.69	$233.46, $241.08

Table 5.16: Alternative Data Profit with 11% Loan Rate

5.5.4.2 Alternative Data - 250 repetitions with loan rate 15% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	1	0	0.222	NaN	$485.20	$483.83, $486.66
Logistic Regression	Variable	1	0	0.222	NaN	$485.20	$483.83, $486.49
Profit Based	Variable	0.958	0.041	0.230	0.908	$495.16	$493.75, $496.45
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	1	0	0.222	NaN	$482.46	$477.21, $488.12
Logistic Regression	Variable	1	0	0.222	NaN	$482.46	$476.99, $487.95
Profit Based	Variable	0.958	0.042	0.227	0.890	$486.06	$480.98, $491.47

Table 5.17: Alternative Data Profit with 15% Loan Rate

5.5.4.3 Alternative Data - 250 repetitions with loan rate 20% above risk free

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	1	0	0.222	NaN	$818.93	$817.00, $820.75
Logistic Regression	Variable	1	0	0.222	NaN	$818.93	$817.05, $820.79
Profit Based	Variable	0.994	0.006	0.223	0.973	$830.41	$828.51, $832.49
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	1	0	0.222	NaN	$815.41	$808.06, $822.51
Logistic Regression	Variable	1	0	0.222	NaN	$815.41	$808.43, $823.07
Profit Based	Variable	0.994	0.006	0.223	0.905	$816.15	$808.83, $822.84

Table 5.18: Alternative Data Profit with 20% Loan Rate

5.5.5 Profit Scoring - Traditional Data with 250 repetitions

5.5.5.1 Traditional Data - 250 repetitions with loan rate 11% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.949	0.051	0.229	0.893	$243.56	$242.61, $244.49
Logistic Regression	Variable	0.948	0.052	0.229	0.901	$248.01	$247.21, $248.91
Profit Based	Variable	0.902	0.098	0.236	0.901	$265.03	$263.97, $265.87
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.949	0.051	0.228	0.895	$239.25	$235.68, $243.60
Logistic Regression	Variable	0.948	0.052	0.229	0.900	$243.99	$240.54, $247.60
Profit Based	Variable	0.902	0.098	0.2331	0.880	$249.48	$246.51, $253.07

Table 5.19: Traditional Data Profit with 11% Loan Rate

5.5.5.2 Traditional Data - 250 repetitions with loan rate 15% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.982	0.018	0.225	0.899	$492.65	$491.37, $493.96
Logistic Regression	Variable	0.982	0.018	0.225	0.912	$498.98	$497.61, $500.24
Profit Based	Variable	0.971	0.029	0.227	0.944	$512.25	$511.11, $513.67
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.982	0.018	0.224	0.900	$486.18	$480.80, $491.47
Logistic Regression	Variable	0.982	0.018	0.225	0.913	$492.70	$487.79, $497.74
Profit Based	Variable	0.971	0.029	0.226	0.900	$494.64	$488.99, $498.98

Table 5.20: Traditional Data Profit with 15% Loan Rate

5.5.5.3 Traditional Data - 250 repetitions with loan rate 20% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.992	0.008	0.223	0.911	$822.34	$820.48, $824.10
Logistic Regression	Variable	0.992	0.008	0.224	0.931	$824.06	$822.25, $825.79
Profit Based	Variable	0.985	0.015	0.225	0.969	$839.58	$837.95, $841.53
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.992	0.008	0.223	0.911	$813.19	$806.26, $819.80
Logistic Regression	Variable	0.992	0.008	0.223	0.930	$815.38	$808.62, $822.29
Profit Based	Variable	0.985	0.015	0.224	0.919	$821.50	$815.27, $828.83

Table 5.21: Traditional Data Profit with 20% Loan Rate

5.5.6 Profit Scoring - Combined Data with 250 Repetitions

5.5.6.1 Combined Data - 250 repetitions with loan rate 11% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.939	0.060	0.229	0.889	$243.68	$242.75, $244.62
Logistic Regression	Variable	0.939	0.060	0.230	0.903	$254.60	$253.64, $255.50
Profit Based	Variable	0.861	0.141	0.243	0.907	$277.90	$277.21, $279.13
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.939	0.060	0.229	0.889	$238.97	$234.98, $242.67
Logistic Regression	Variable	0.939	0.060	0.230	0.901	$249.57	$245.62, $253.38
Profit Based	Variable	0.861	0.140	0.240	0.884	$255.53	$251.50, $258.72

Table 5.22: Combined Data Profit with 11% Loan Rate

5.5.6.2 Combined Data - 250 repetitions with loan rate 15% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.980	0.020	0.225	0.910	$492.74	$491.43, $494.04
Logistic Regression	Variable	0.980	0.020	0.225	0.915	$497.01	$495.67, $498.30
Profit Based	Variable	0.941	0.061	0.233	0.937	$517.35	$516.19, $518.84
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.980	0.020	0.225	0.911	$486.64	$481.71, $492.11
Logistic Regression	Variable	0.980	0.020	0.225	0.914	$490.08	$486.12, $496.42
Profit Based	Variable	0.941	0.060	0.229	0.893	$491.87	$486.23, $496.21

Table 5.23: Combined Data Profit with 15% Loan Rate

5.5.6.3 Combined Data - 250 repetitions with loan rate 20% above risk free rate

Model	Cutoff	AR	RR	P(A|G)	P(B|R)	Profit	CI.Profit
IN-SAMPLE RESULTS							
Logistic Regression	Constant	0.991	0.009	0.22356	0.906	$821.02	$819.20, $822.86
Logistic Regression	Variable	0.991	0.009	0.22371	0.923	$823.44	$821.72, $825.22
Profit Based	Variable	0.987	0.013	0.22488	0.968	$835.22	$833.17, $836.72
OUT-OF-SAMPLE RESULTS							
Logistic Regression	Constant	0.991	0.008	0.223	0.907	$811.94	$805.18, $819.25
Logistic Regression	Variable	0.991	0.009	0.223	0.926	$814.61	$808.15, $821.67
Profit Based	Variable	0.987	0.013	0.223	0.905	$815.82	$809.10, $822.92

Table 5.24: Combined Data Profit with 20% Loan Rate

Bibliography

[1] Hasan T. Abbas et al. "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test". In: *PLoS ONE* 14.12 (2019), pp. 1–12. ISSN: 19326203. DOI: 10.1371/journal.pone.0219636.

[2] Edgar Acuna and Caroline Rodriguez. "Classification, Clustering, and Data Mining Applications". In: *Classification, Clustering, and Data Mining Applications*. 1995. 2010, pp. 1–9. DOI: 10.1007/978-3-642-17103-1.

[3] William Adams, Liran Einav, and Jonathan Levin. "Liquidity constraints and imperfect information in subprime lending". In: *American Economic Review* 99.1 (2009), pp. 49–84. ISSN: 00028282. DOI: 10.1257/aer.99.1.49.

[4] Rishav Raj Agarwal et al. "Predicting financial trouble using call data - On social capital, phone logs, and financial trouble". In: *PLoS ONE* 13.2 (2018), pp. 1–18. ISSN: 19326203. DOI: 10.1371/journal.pone.0191863.

[5] Rob Aitken. "All data is credit data: Constituting the unbanked". In: *Competition and Change* 21.4 (2017), pp. 274–300. ISSN: 14772221. DOI: 10.1177/1024529417712830.

[6] Hirotugu Akaike. "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6 (1974), pp. 716–723. ISSN: 15582523. DOI: 10.1109/TAC.1974.1100705.

[7] Najat Ali, Daniel Neagu, and Paul Trundle. "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets". In: *SN Applied Sciences* 1.12 (2019), pp. 1–15. ISSN: 2523-3963. DOI: 10.1007/s42452-019-1356-9. URL: https://doi.org/10.1007/s42452-019-1356-9.

[8] Billie Anderson and J. Michael Hardin. "Credit Scoring in the Age of Big Data". In: *Encyclopedia of Business Analytics and Optimization* 18.1 (2014), pp. 549–557. DOI: 10.4018/978-1-4666-5202-6.ch049.

[9] Donald W.K. Andrews and Moshe Buchinsky. "EVALUATION OF A THREE-STEP METHOD FOR CHOOSING THE NUMBER OF BOOTSTRAP REPETITIONS BY COWLES FOUNDATION FOR RESEARCH IN ECONOMICS Evaluation of a three-step method for choosing the number of bootstrap repetitions". In: *Journal of Econometrics* 103 (2001), pp. 345–386.

[10] Donald W.K Andrews and Moshe Buchinsky. "A Three-Step Method For Choosing The Number Of Bootstrap Repetitions". In: *Econometrica* 68.1 (2000), pp. 23–51.

[11] Nenny Anggraini and Muhammad Jabal Tursina. "Sentiment Analysis of School Zoning System On Youtube Social Media Using The K-Nearest Neighbor With Levenshtein Distance Algorithm". In: *International Conference on Cyber and IT Service Management*. 2019, pp. 1–4. DOI: 10.1109/citsm47753.2019.8965407.

[12] Davide Anguita et al. "Model selection for support vector machines: Advantages and disadvantages of the Machine Learning Theory". In:

Proceedings of the International Joint Conference on Neural Networks (2010). DOI: 10.1109/IJCNN.2010.5596450.

[13] Bart Baesens, Daniel Rösch, and Harald Scheule. *Credit Risk Analytics: Measurement Techniques, Applications, and Example in SAS*. 2016. ISBN: 9781119278344.

[14] Amos Baranes and Rimona Palas. "Earning movement prediction using machine learning-Support Vector Machines (SVM)". In: *Journal of Management Information and Decision Science* 22.2 (2019), pp. 36–53. ISSN: 15325806.

[15] Luis Javier Sánchez Barrios, Galina Andreeva, and Jake Ansell. "Monetary and relative scorecards to assess profits in consumer revolving credit". In: *Journal of the Operational Research Society* 65.3 (2014), pp. 443–453. ISSN: 01605682. DOI: 10.1057/jors.2013.66.

[16] Tobias Berg et al. "On the Rise of FinTechs-Credit Scoring Using Digital Footprints". 2018.

[17] Leo Breiman. "Random forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 08856125. DOI: 10.1023/A:1010933404324.

[18] CFPB. *Interagency Statement on the Use of Alternative Data in Credit Underwriting*. Tech. rep. 2019.

[19] Philip A Chou. "Optimal Partitioning for Classificationand Regression Trees". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.4 (1991).

[20] Ertugrul Colak et al. "Comparison of semiparametric, parametric, and nonparametric ROC analysis for continuous diagnostic tests using a simulation study and acute coronary syndrome data". In: *Computational and*

Mathematical Methods in Medicine 2012 (2012). ISSN: 1748670X. DOI: 10.1155/2012/698320.

[21] Sandra Cuentas, Rita Peñabaena-Niebles, and Ethel Garcia. "Support vector machine in statistical process monitoring: a methodological and analytical review". In: *International Journal of Advanced Manufacturing Technology* 91.1-4 (2017), pp. 485–500. ISSN: 14333015. DOI: 10.1007/s00170-016-9693-y.

[22] Richard M Cyert, H Justin Davidson, and Gerald L Thompson. "ESTIMATION OF THE ALLOWANCE FOR DOUBTFUL ACCOUNTS BY MARKOV CHAINS". In: *Management Science* (1962). ISSN: 0025-1909. DOI: 10.1287/mnsc.8.3.287.

[23] Ralph B D D'Agostino and Michael A Stephens. *GOODNESS-OF-FIT TECHNIQUES*. 1987. URL: https://www.gbv.de/dms/ilmenau/toc/04207259X.PDF.

[24] Russell Davidson and James G. MacKinnon. "Bootstrap tests: how many bootstraps?" In: *Econometric Reviews* 19.68 (2007), pp. 23–52. ISSN: 0747-4938. DOI: 10.1080/07474930008800459.

[25] Elizabeth R Delong, David M Delong, and Daniel L Clarke-Pearson. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach". In: *Biometrics* 44.3 (1988), pp. 837–845.

[26] Misha Denil, David Matheson, and Nando De Freitas. "Narrowing the Gap: Random Forests In Theory and In Practice". In: *Proceedings of The 31st International Conference on Machine Learning* 1998 (2014), pp. 665–673. ISSN: 9781634393973. arXiv: arXiv:1310.1415v1. URL: http://jmlr.org/proceedings/papers/v32/denil14.html.

[27] Arnout Devos et al. "PROFIT MAXIMIZING LOGISTIC REGRESSION MODELING for CREDIT SCORING". In: *2018 IEEE Data Science Workshop, DSW 2018 - Proceedings* (2018), pp. 125–129. DOI: 10.1109/DSW.2018.8439113.

[28] Yiran Dong and Chao Ying Joanne Peng. "Principled missing data methods for researchers". In: *SpringerPlus* 2.1 (2013), pp. 1–17. ISSN: 21931801. DOI: 10.1186/2193-1801-2-222.

[29] Graham Elliott and Robert P. Lieli. "Predicting binary outcomes". In: *Journal of Econometrics* (2013). ISSN: 03044076. DOI: 10.1016/j.jeconom.2013.01.003.

[30] Floriana Esposito, Donato Malerba, and Giovanni Semeraro. "A comparative analysis of methods for pruning decision trees". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19.5 (1997), pp. 476–491. ISSN: 01628828. DOI: 10.1109/34.589207.

[31] Experian. *The State of Alternative Data*. Tech. rep. 2015, pp. 1–20.

[32] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. "Random forests: From early developments to recent advancements". In: *Systems Science and Control Engineering* 2.1 (2014), pp. 602–609. ISSN: 21642583. DOI: 10.1080/21642583.2014.956265. URL: https://doi.org/10.1080/21642583.2014.956265.

[33] FDIC. *FDIC National Survey of Unbanked and Underbanked Households*. 2017.

[34] Ad J. Feelders. "An Overview of Model Based Reject Inference for Credit Scoring". In: *Sciences-New York* (2003).

[35] S. M. Finlay. "Towards profitability: A utility approach to the credit scoring problem". In: *Journal of the Operational Research Society* (2008). ISSN: 01605682. DOI: 10.1057/palgrave.jors.2602394.

[36] Steven Finlay. "Credit scoring for profitability objectives". In: *European Journal of Operational Research* (2010). ISSN: 03772217. DOI: 10.1016/j.ejor.2009.05.025.

[37] Xiang Gao, Junhao Wen, and Cheng Zhang. "An Improved Random Forest Algorithm for Predicting Employee Turnover". In: *Mathematical Problems in Engineering* 2019 (2019). ISSN: 15635147. DOI: 10.1155/2019/4140707.

[38] Johannes Gehrke, Raghu Ramakrishnan, and Venkatesh Ganti. "RainForest - A framework for fast decision tree construction of large datasets". In: *Data Mining and Knowledge Discovery* 4.2-3 (2000), pp. 127–162. ISSN: 13845810. DOI: 10.1023/A:1009839829793.

[39] Asghar Ghasemi and Saleh Zahediasl. "Normality tests for statistical analysis: A guide for non-statisticians". In: *International Journal of Endocrinology and Metabolism* 10.2 (2012), pp. 486–489. ISSN: 1726913X. DOI: 10.5812/ijem.3505.

[40] L.G. Godfrey. "Misspecification Tests and Their Uses in Econometrics". In: *Journal of Statistical Planning and Inference* 49.2 (1996), pp. 241–260.

[41] R. Y. Goh and L. S. Lee. "Credit Scoring: A Review on Support Vector Machines and Metaheuristic Approaches". In: *Advances in Operations Research* 2019 (2019). ISSN: 16879155. DOI: 10.1155/2019/1974794.

[42] Guangming Guo et al. "From footprint to evidence: An exploratory study of mining social data for credit scoring". In: *ACM Transactions on the Web* 10.4 (2016). ISSN: 1559114X. DOI: 10.1145/2996465.

[43] Zhenzhou Guo et al. "A k-Nearest Neighbor Algorithm Based on Homomorphic Encryption". In: *IEEE International Conferences on Ubiquitous Computing & Communications*. 2019, pp. 15–20. ISBN: 9781728152097. DOI: 10.1109/iucc/dsci/smartcns.2019.00032.

[44] D. J. Hand and W. E. Henley. "Statistical classification methods in consumer credit scoring: A review". In: *Journal of the Royal Statistical Society. Series A: Statistics in Society* 160.3 (1997), pp. 523–541. ISSN: 09641998. DOI: 10.1111/j.1467-985X.1997.00078.x.

[45] David J. Hand and Mark G. Kelly. "Superscorecards". In: *IMA Journal of Management Mathematics* (2002). ISSN: 1471678X. DOI: 10.1093/imaman/13.4.273.

[46] Anders Haug, Frederik Zachariassen, and Dennis van Liempd. "The costs of poor data quality". In: *Journal of Industrial Engineering and Management* 4.2 (2011), pp. 168–193. ISSN: 20138423. DOI: 10.3926/jiem.2011.v4n2.p168-193.

[47] By Fumiko Hayashi and Fumiko Hayashi. "Access to Electronic Payments Systems by Unbanked Consumers". In: *Federal Reserve Bank of Kansas City Economic Review* 101.3 (2016), pp. 51–76. ISSN: 01612387.

[48] Geofrey Hinton and Terrence Sejnowski. *Unsupervised learning: Foundations of neural computation*. 1999. DOI: 10.1016/s0898-1221(99)90165-7.

[49] Jeanne M Hogarth and Kevin H O'Donnell. "Banking Relationships of Lower-Income Families and the Governmental Trend toward Electronic

Payment". In: *Federal Reserve Bulletin* 85 (1999), pp. 459–473. URL: http://heinonlinebackup.com/hol-cgi-bin/get%7B%5C_%7Dpdf.cgi?handle=hein.journals/fedred85%7B%5C&%7Dsection=112.

[50] Sadegh Bafandeh Imandoust and Mohammad Bolandraftar. "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background". In: *Internatiional Journal of Engineering Research and Applications* 3.5 (2013), pp. 605–610.

[51] I. Intan, S.T.A.D. Ghani, and N. Salman. "Implementation of the K-Nearest Neighbor and Neural Network for Predicting School Readiness to Enter Elementary School". In: (2020), pp. 1–6. DOI: 10.1109/citsm47753.2019.8965346.

[52] Anil K. Jain, Robert P.W. Duin, and Jianchang Mao. "Statistical pattern recognition: A review". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.1 (2000), pp. 4–37. ISSN: 01628828. DOI: 10.1109/34.824819.

[53] R. L. Keeney and R. M. Oliver. "Designing win-win financial loan products for consumers and businesses". In: *Journal of the Operational Research Society* (2005). ISSN: 01605682. DOI: 10.1057/palgrave.jors.2601992.

[54] Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. "Consumer credit-risk models via machine-learning algorithms". In: *Journal of Banking and Finance* (2010). ISSN: 03784266. DOI: 10.1016/j.jbankfin.2010.06.001.

[55] Nikita Kozodoi et al. "A multi-objective approach for profit-driven feature selection in credit scoring". In: *Decision Support Systems* 120.April (2019), pp. 106–117. ISSN: 01679236. DOI: 10.1016/j.dss.2019.03.011.

[56] Max Kuhn and Kjell Johnson. *Applied predictive modeling*. 2013. ISBN: 9781461468493. DOI: 10.1007/978-1-4614-6849-3.

[57] Gabriela Kuvíková. "DOES LOAN MATURITY MATTER IN RISK-BASED PRICING? EVIDENCE FROM CONSUMER LOAN DATA". In: *WORKING PAPER SERIES. Center for Economic Research and Graduate Education* (2015).

[58] Jorma Laaksonen and Erkki Oja. "Classification with learning k-nearest neighbors". In: *IEEE International Conference on Neural Networks - Conference Proceedings* 3 (1996), pp. 1480–1483. DOI: 10.1109/icnn.1996.549118.

[59] Stefan Lessmann et al. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". In: *European Journal of Operational Research* (2015). ISSN: 03772217. DOI: 10.1016/j.ejor.2015.05.030.

[60] Robert P. Lieli and Halbert White. "The construction of empirical credit scoring rules based on maximization principles". In: *Journal of Econometrics*. 2010. ISBN: 0304-4076. DOI: 10.1016/j.jeconom.2009.10.028.

[61] Wei Yin Loh. "Classification and regression trees". In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.1 (2011), pp. 14–23. ISSN: 19424787. DOI: 10.1002/widm.8.

[62] Wei Yin Loh. "Fifty years of classification and regression trees". In: *International Statistical Review* 82.3 (2014), pp. 329–348. ISSN: 17515823. DOI: 10.1111/insr.12016.

[63] Susan Lomax and Sunil Vadera. "A survey of cost-sensitive decision tree induction algorithms". In: *ACM Computing Surveys* 45.2 (2013). ISSN: 03600300. DOI: 10.1145/2431211.2431215.

[64] Renata C.B. Madeo, Clodoaldo A.M. Lima, and Sarajane M. Peres. "A review on temporal reasoning using support vector machines". In: *Proceedings - 2012 19th International Symposium on Temporal Representation and Reasoning, TIME 2012* (2012), pp. 114–121. DOI: 10.1109/TIME.2012.15.

[65] Sebastián Maldonado et al. *Integrated framework for profit-based feature selection and SVM classification in credit scoring*. 2017. DOI: 10.1016/j.dss.2017.10.007.

[66] Schmitt P Mandel J. "A Comparison of Six Methods for Missing Data Imputation". In: *Journal of Biometrics & Biostatistics* 06.01 (2015), pp. 1–6. DOI: 10.4172/2155-6180.1000224.

[67] In Jae Myung. "Tutorial on maximum likelihood estimation". In: *Journal of Mathematical Psychology* 47 (2003), pp. 90–100. ISSN: 00222496. DOI: 10.1016/S0022-2496(02)00028-7.

[68] Nadim Nachar. "The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution". In: *Tutorials in Quantitative Methods for Psychology* 4.1 (2008), pp. 13–20. ISSN: 1913-4126. DOI: 10.20982/tqmp.04.1.p013.

[69] Michael Natole, Yiming Ying, and Siwei Lyu. "Stochastic AUC Optimization Algorithms With Linear Convergence". In: *Frontiers in Applied Mathematics and Statistics* 5.June (2019), pp. 1–9. ISSN: 22974687. DOI: 10.3389/fams.2019.00030.

[70] Bryan D Nelson. "Variable reduction for modeling using PROC VARCLUS". In: *Conference Proceedings SAS Users Group International*. 2010, pp. 1–3.

[71] A. A. Nurhanna and M. F. Othman. "Multi-class support vector machine application in the field of agriculture and poultry: A review". In: *Malaysian Journal of Mathematical Sciences* 11.S (2017), pp. 35–52. ISSN: 18238343.

[72] Ceylan Onay and Elif Öztürk. "A review of credit scoring research in the age of Big Data". In: *Journal of Financial Regulation and Compliance* 26.3 (2018), pp. 382–405. ISSN: 17400279. DOI: 10.1108/JFRC-06-2017-0054.

[73] María Óskarsdóttir et al. "The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics". In: *Applied Soft Computing Journal* 74 (2019), pp. 26–39. ISSN: 15684946. DOI: 10.1016/j.asoc.2018.10.004. URL: https://doi.org/10.1016/j.asoc.2018.10.004.

[74] Cunningham Padraig and Sarah Jane Delany. "k-Nearest Neighbour Classifiers". In: *Multiple Classifier Systems* 34.8 (2007), pp. 1–17. ISSN: 1860949X. DOI: 10.1007/978-3-030-03895-3_6.

[75] Daniel Abreu Vasconcellos de Paula et al. "Estimating credit and profit scoring of a Brazilian credit union with logistic regression and machine-learning techniques". In: *RAUSP Management Journal* 54.3 (2019), pp. 321–336. ISSN: 25310488. DOI: 10.1108/RAUSP-03-2018-0003.

[76] Bruno M. Pereira et al. "The role of point-of-care ultrasound in intra-abdominal hypertension management". In: *Anaesthesiology Intensive Therapy* 49.5 (2017), pp. 373–381. ISSN: 17312531. DOI: 10.5603/AIT.a2017.0074.

[77] Bholowalia Purnima and Kumar Arvind. "EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN". In: *International Journal of Computer Applications* 105.9 (2014), pp. 17–24. URL:

https://www.ijcaonline.org/archives/volume105/number9/18405-9674.

[78] J. R. Quinlan. "Induction of decision trees". In: *Machine Learning* 1.1 (1986), pp. 81–106. ISSN: 0885-6125. DOI: 10.1007/bf00116251.

[79] Keya Rani Das. "A Brief Review of Tests for Normality". In: *American Journal of Theoretical and Applied Statistics* 5.1 (2016), p. 5. ISSN: 2326-8999. DOI: 10.11648/j.ajtas.20160501.12.

[80] Sherrie L.W. Rhine and William H. Greene. "Factors that contribute to becoming unbanked". In: *Journal of Consumer Affairs* 47.1 (2013), pp. 27–45. ISSN: 00220078. DOI: 10.1111/j.1745-6606.2012.01244.x.

[81] Warren. S Sarle. *The VARCLUS Procedure*. 2014.

[82] Gideon Schwarz. "Estimating the Dimension of a Model". In: *Annals of Statistics* 6.2 (1978), pp. 461–464.

[83] Carlos Serrano-Cinca and Begoña Gutiérrez-Nieto. "The use of profit scoring as an alternative to credit scoring systems in peer-to-peer (P2P) lending". In: *Decision Support Systems* (2016). ISSN: 01679236. DOI: 10.1016/j.dss.2016.06.014.

[84] Jadran Sessa and Dabeeruddin Syed. "Techniques to deal with missing data". In: *International Conference on Electronic Devices, Systems, and Applications* (2017), pp. 1–4. ISSN: 21592055. DOI: 10.1109/ICEDSA.2016.7818486.

[85] Joe Whittaker Sewart, Pete. "Graphical Models In Credit Scoring". In: *IMA Journal of Mathematics Applied in Business & Industry* (1998), pp. 241–266.

[86] Marvin M. Smith and Christopher Henderson. "Beyond Thin Credit Files". In: *Social Science Quarterly* 99.1 (2018), pp. 24–42. ISSN: 15406237. DOI: 10.1111/ssqu.12389.

[87] Alex J Smola and Bernhard Scholkopf. "A Tutorial on Support Vector Regression". In: *Statistics and Computing* 14 (2004), pp. 199–222. ISSN: 0960-3174. DOI: 10.1023/B:STCO.0000035301.49549.88. arXiv: arXiv:1011.1669v3. URL: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1CAD92EF8CCE726A305D8A41F873EEFC?doi=10.1.1.114.4288%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf%7B%5C%%7D0Ahttp://download.springer.com/static/pdf/493/art%7B%5C%%7D3A10.1023%7B%5C%%7D2FB%7B%5C%%7D3ASTCO.0000035301.49549.88.pdf?auth66=1408162706%7B%5C_%7D8a28764ed0fae9.

[88] Madan Somvanshi et al. "A review of machine learning techniques using decision tree and support vector machine". In: *International Conference on Computing, Communication, Control and Automation*. IEEE, 2017, pp. 1–7. ISBN: 9781509032914. DOI: 10.1109/ICCUBEA.2016.7860040.

[89] Wei Sun and Yi Liang. "Comprehensive evaluation of cleaner production in thermal power plants using particle swarm optimization based least squares support vector machines". In: *Journal of Information and Computational Science* 12.5 (2015), pp. 1993–2000. ISSN: 15487741. DOI: 10.12733/jics20105590.

[90] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning*. 1998.

[91] Gerhard Svolba. *Applying Data Science: Business Case Studies Using SAS*. Cary, NC: SAS Institute Inc, 2017, pp. 325–336.

[92] Lyn C. Thomas. "A survey of credit and behavioural scoring: Forecasting financial risk of lending to consumers". In: *International Journal of Forecasting* (2000). ISSN: 01692070. DOI: 10.1016/S0169-2070(00)00034-0.

[93] Lyn C. Thomas. *Consumer credit models: Pricing, profit and portfolios*. 2009, p. 6. ISBN: 9780191715914. DOI: 10.1093/acprof:oso/9780199232130.001.1.

[94] TransUnion. "The State of Alternative Data". In: *TransUnion* (2015), pp. 1–20.

[95] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, 2000.

[96] Michel Verleysen and Damien François. "The Curse of Dimensionality in Data Mining and Time Series Prediction". In: *Iwann 2005: Computational Intelligence and Bioinspired Systems*. Vol. 3512. 2005, pp. 758–770. ISBN: 3540262083. DOI: https://doi.org/10.1007/11494669_93.

[97] Yanjun Wang and Qun Liu. "Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock-recruitment relationships". In: *Fisheries Research* 77.2 (2006), pp. 220–225. ISSN: 01657836. DOI: 10.1016/j.fishres.2005.08.011.

[98] Yanhao Wei et al. "Credit scoring with social network data". In: *Marketing Science* 35.2 (2016), pp. 234–258. ISSN: 1526548X. DOI: 10.1287/mksc.2015.0949.

[99] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83.

[100] B. W. Yap and C. H. Sim. "Comparisons of various types of normality tests". In: *Journal of Statistical Computation and Simulation* 81.12 (2011), pp. 2141–2155. ISSN: 00949655. DOI: 10.1080/00949655.2010.520163.

[101] Longlong Zhang et al. "Review of remaining useful life prediction using support vector machine for engineering assets". In: *International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering* (2013), pp. 1793–1799. DOI: 10.1109/QR2MSE.2013.6625925.

[102] Lin Zhu, Hong Bo Zhang, and De Shuang Huang. "Direct AUC optimization of regulatory motifs". In: *Bioinformatics* 33.14 (2017), pp. i243–i251. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx255.