2020

Optimization of Home Mortgage Mover Predictive Model Applying Geo-Spatial Analysis and Machine Learning Techniques

Natalia Riscovaia

Technological University Dublin

Recommended Citation

Riscovaia, N. (2020). Optimization of Home Mortgage Mover Predictive Model Applying Geo-Spatial Analysis and Machine Learning Techniques. *A dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of M.Sc. in Computing (Data Analytics).* doi:10.21427/hx2d-n902

Optimization of Home Mortgage Mover Predictive Model Applying Geo-Spatial Analysis and Machine Learning Techniques

Natalia Riscovaia

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (Data Analytics)

16 June 2020

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Natalia Riscovaia

Date: 16 June 2020

Abstract

In the last decade digital innovations and online banking services have significantly changed customers banking preferences and behaviour. Banking industry is going through the changes and developments in the provision of banking services that are affecting the structure and the organization of the bank network. However, private home loan, referred as *Home Mortgage* hereinafter, continue to remain among the products, that customers prefer to have personal interaction about with professional advisors prior making the decision to apply for the loan with financial institution.

This work aims to analyse whether the physical presence of the bank branches and their distance from customers residential addresses are significant factors for home mortgage customers in their decision to apply for home loan with the financial institution, referred as *Bank A* hereinafter. The machine learning techniques are combined with spatial analysis to build the Home Mover prediction models in the rural counties of Ireland for binary classification task. The population base is limited to the customers already holding at least one mortgage product with *Bank A* and residing in the rural counties of Ireland, for a number of reasons outlined in the context of this work.

Different machine learning algorithms across spatial and non-spatial feature engineering techniques are tested to evaluate whether spatially-conscious machine learning models outperform non-spatial models when integrated with prediction techniques such as Random Forest, Gradient Boosting and Gradient Boosting Ensemble models.

Keywords: Banking, Home Mortgages, Classification, Predictive Models, Machine Learning, GIS, Geo-Spatial Analysis, Gradient Boosting, Random Forest, Ensemble Models.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Sean O'Liery, for his help and guidance throughout the dissertation.

I would like to thank my employers for sponsoring this Masters, facilitating me with access to the necessary data, and providing invaluable support. I would also like to express my sincerest thanks to my colleagues who provided constant support and shared their extensive expertise and knowledge, throughout my studies and the dissertation writing process.

My deepest gratitude goes to my family and friends. To my parents for their unconditional love, guidance and support that helped me getting to this stage in my life. To my husband and my beautiful boys for always being by my side, caring about me and making me lough. To my friends for their patience and sense of humour that carried me through the moments of weakness and never let me to give up.

Without doubt my family and friends showed that the geographical distance can never define the true sense of love, care and support.

Contents

List of Figures

List of Tables

List of Acronyms

ABT	Analytical Base Table
ANN	Artificial Neural Networks
GDPR	General Data Protection Regulation
GIS	Geographical Information System
GSD	Garda Sub District
EDs	Electoral Divisions
EDW	Enterprise Data Warehouse
LTV	Loan to Value
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
ROC	Receiver Operating Characteristics

Chapter 1

Introduction

1.1 Background

Predictive modeling with application of advanced machine learning techniques has significant impact on how most business areas operate in the modern reality, empowering companies to make automated insightful decisions. Financial institutions and banking industry utilises machine learning in a wide range of applications, such as fraud detection, risk assessment, automatic credit approval, customer segmentation analysis and customer churn prediction (Ilyas et al., 2020).

Machine learning techniques are also applied in propensity modeling. Statistical scorecard that aims to predict behaviour of the customer is commonly referred to as a propensity model, where the propensity is a natural tendency to behave in a certain way. In business, and particularly banking context, propensity is calculated by application of the machine learning techniques to a customer base predicting what are the future actions of the customers, based on their 'historical behaviour' (Sharma, Alford, Bhuian, & Pelton, 2009). Propensity modeling also plays a significant role in identifying new customers, who potentially will take similar actions. In machine learning context this is usually a classification problem, where the model aims to classify a customer behaviour with binary output 'yes' or 'no' to a certain action.

Geospatial Analysis and Geographic Information Systems(GIS) offer powerful solutions for location related analysis and decision making in different areas of application. Traditionally geospatial analysis are referred to as a spatial auto-correlation, looking at the things that are near to each other and tend to be like each other (Jankowski, Fraley, & Pebesma, 2014; Simionescu, 2015). In the banking and financial industry context, as these services are often location-based, linking spatial data in GIS is a way to analyze and drive more informed decisions by understanding customers locations and their transactions, at the same time optimizing and better managing branch networks and financial services.

This work aims to evaluate whether the integration of spatial features can be used in machine learning context, that will allow to optimize predictive model for home mover mortgage customers in the rural areas of Ireland.

1.2 Research Problem

In the changing reality of digitization and therefor changing banking habits of the customers the question arises: does the physical access to the branch location remain the important and driving factor for home mortgage sales.

To address this question, binary classification models will be built, based on the population of existing home mortgage customers residing in the rural areas of Ireland, aiming to predict what customers will likely decide to apply for the private home mortgage loan with *Bank A*.

Factors defining the criteria that the existing mortgage customer will decide to apply for their not first home loan (hereinafter referred to as *Home Mover Customers*) may differ from the 'First Time Buyers' - customers who apply for their first home loan. Additionally, factors defining the criteria that the existing mortgage customers will decide to apply for another home mortgage loan may differ in cities, commuter areas

and rural areas of Ireland. Hence, the population base will be limited to the existing mortgage customers residing in the rural counties of Ireland only.

The branch network availability in close proximity from the customers will be considered and performed by calculating the distance from customers residential address to the nearest branch locations. The fact that the population base is limited to the customers already holding mortgage with *Bank A* will allow accurately identify customers' residential addresses, stored in the Enterprise Data Warehouse.

1.3 Research Objectives

Home mortgages sales prediction may be challenging due to the nature of the factors that tend to change over time, that impact customers purchasing behaviour, such as customers demographics, financial state of economy, home mortgage rates and stability retail home market. Additionally these factors may vary for Home Movers, customers who apply for home loan to change the existing residential property, and First Time Buyers, customers who apply for their first home loan. Another factors that have an impact on home mortgage sales are the geographical location, defined by where customers reside and consider to purchase their property, and financial services availability, defined by how customers will apply for the home loan.

The aim of this research work is to evaluate whether the integration of the spatial features based on the distance from customers residential locations to the nearest branches, in machine learning modeling techniques have an effect on the predictive capability of the binary classification task to accurately predict the customers who will apply for the Home Mover loan in rural counties of Ireland.

1.4 Research Question

Is the distance of the bank branches from customers residential addresses integrated into demographic and transactional dataset a strong indicator of predictive significance of the binary classification model predicting customers who will apply for Home Mover mortgage loan in the rural counties of Ireland?

1.5 Research Hypothesis

To carry out the experiment, defined in the research objectives, the following Hypothesis is stated:

H_0: Distance of the bank branches from the customers residential addresses, integrated into the dataset comprised of demographic profiles of the customers along with customers spending behaviour can not improve the predictive significance of the binary classification model predicting customers who will apply for Home Mover mortgage loan in the rural areas of Ireland.

H_a: Distance of the bank branches from the customers residential addresses, integrated into the dataset comprised of demographic profiles of the customers along with customers spending behaviour can improve the predictive significance of the binary classification model predicting customers who will apply for Home Mover mortgage loan in the rural areas of Ireland.

1.6 Research Methodologies

Research methodology follows the below structure:

- This is primary research, data collected from different data sources to carry out the experiment.

- Quantitative method is employed, geographical, demographic, transactional data is be encoded and tested using mathematical and statistical approaches.
- As hypothesis and prediction are tested with experiment, the experiment is empirical in form.
- This research takes an inductive bottom-up approach, starting from specific observations to uncover new unseen patterns.

1.7 Scope and Limitations

The scope of this research work is to build Mortgage Mover sales prediction binary classification model, which has the capability to accurately identify the customers who likely will consider to change their residential property in the rural counties of the Republic of Ireland and apply for another home mortgage loan with *Bank A*. The 'non-spatial' model will be then optimised by integrating distance features based on the calculations performed in QGIS on the distance from the existing customers residential addresses to the nearest *Bank A* branch locations to evaluate whether the distance feature has a significance in the model.

The experiment will use historical data to build and test the models, which will then be validated against the previously unseen data.

There were some challenges and limitations encountered in this research:

- Although Artificial Neural Networks (ANN) were originally considered and could have been investigated for this problem, due to the nature in which ANNs automatically calculate weights in their models the *Bank A* refrains from utilising this approach inline with General Data Protection Regulation (GDPR) requirements. The General Data Protection Regulation (GDPR) - European regulation introduced to ensure data protection and privacy rights. Based on the GDPR regulations the data subjects are required to be interpreted in the way that are

compliant to this regulations and have meaningful explanations of automated decision making.

- Geocoding to convert customers' residential addresses to latitude and longitude did not return accurately all the entries, despite the multiple attempts. As a result, originally considered population base to carry out the experiment had to be reduced.

- Due to the private nature of the commercial data being used, this data could not be used outside of the company's systems that limited choice of available tools and software to carry out the experiment.

1.8 Document Outline

The next chapters of this research work are organised in the following order: Literature Review, Experiment Design & Methodology, Experiment Results & Evaluation, Conclusions and Future work.

1. *Chapter 2* will review the literature related to this study. Studies in predictive modeling followed by the studies in geospatial analysis applied in financial and banking sectors will be researched. Data imbalance issue and approaches to address it will be studied, as the common issue in financial institutions' data. The analysis of existing models and evaluation techniques will define the final design and evaluation of the experiment.

2. *Chapter 3* will provide an overview of the experiment design and methodology. Data pre-processing will be described, followed by the taken modeling approaches. Methods of testing and evaluation will be discussed and presented. The workings of distance calculations and map visualisations will be presented.

3. *Chapter 4* will discuss the implementation, evaluation and results of the experiment.

4. *Chapter 5* will summarise the findings of this research and make recommendations for future work and research in the area.

Chapter 2

Review of existing literature

2.1 Predictive Modeling in Financial & Banking Industry

2.1.1 Basic Concepts of Predictive Modeling

Frawley, Piatetsky-Shapiro, and Matheus (1992) define Data Mining as nontrivial extraction of implicit, previously unknown and potentially useful information from data. Data Mining contributes to solving business problems in banking and financial industry by finding patterns, associations and correlations which are hidden in large volumes of business information(Farooqi, Iqbal, & Rashid Farooqi, 2017).
The below figure presents Data Mining main tasks overview:

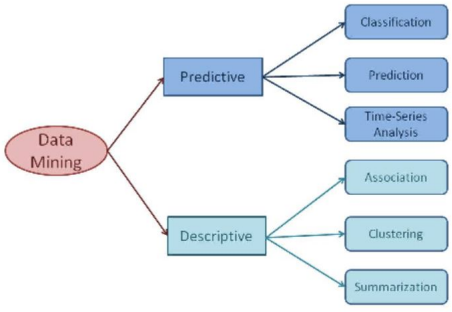


Figure 2.1: Data Mining Tasks

One of the data mining tasks is predictive modeling, that seek to be able to predict the future behaviour in a certain situation, given what has happened in the past. Most predictive models are trained using historic data, but after are tested against real world data to evaluate how well the models have performed. There is no area in the modern reality that wouldn't benefit from the insightful information that can be generated from the predictive modeling, including medical applications such as predicting cancer rates (Delen, Walker, & Kadam, 2005), sentiment analysis in predicting the outcome of elections (Tumasjan, Sprenger, Sandner, & Welpe, 2010), sports and betting industry (Silverman & Suchard, 2013) to name just a few.

Statistical scorecard, that is used to predict the behaviour of the customers is called a propensity model - a form of predictive model widely applied in the banking industry. Westreich, Lessler, and Funk (2010) defines propensity models as based on the "conditional probability of assignment to a particular treatment given a vector of observed co-variates".

The following dimensions are usually taken into account for propensity modeling:

1. Demographic characteristics that outline "who" the customer is based on the gender, age, geographic area, educational attainment, income level;

2. Transactional information outlines "what" a customer has purchased in the past as well as their estimated purchase capacity;

3. Psychographic information outlines "why" a customer purchases things in terms of attitude and opinions, as for example "likes" on Facebook ;

4. Personality information reflects person's characteristic patterns of thoughts, feelings, and behavior. This data is usually collected via surveys and can be difficult to acquire.

The data used to build the model depends on what propensity is set to be predicted. For example, financial institutions use propensity-to-buy models to link customer

characteristics and product needs to the right solution. In gaining high propensity customers as leads, banks drive future marketing and sales initiatives to enhance marketing effectiveness and increase sales productivity.

In the last decade the machine learning has unlocked full potential of propensity modeling, although the concept of propensity modeling dates back to 1983 (Westreich et al., 2010). The most sophisticated propensity models use machine learning algorithms to predict what a customer is likely to do next by exploiting patterns in human behaviour.

2.1.2 Home Mortgages Predictive Modeling

In the recent years there has been a great focus on modeling in home mortgages that is motivated by the growing house market demand as well as risks associated with rise in mortgage arrears and default rates. At the same time, most of the research done in the mortgage prediction domain is kept internally at the financial institutions, due to the classified nature of the data and the sensitive nature of the results. As a result there is no large amount of the scientific research publicly available about home mortgage applications and sales prediction models. However, there are a number of papers that provide similar research in related domains, for example the prediction of retail market sales, as well as predicting mortgage defaults and arrears.

Findings of different researchers in the related areas were studied to identify the most common use cases, as well as existing challenges and future research possibilities (Martin, Miranda Lakshmi, & Prasanna Venkatesan, 2014; Henrique, Sobreiro, & Kimura, 2019; Huang & Yen, 2019; Niankara, 2019; Levin & Zahavi, 2001) .

Regression Models

The modeling of (linear) relationships between a dependent (target) and independent variable (predictor) can be performed by variant regression models. There is

large amount of different regression forms, each form has its own importance and a specific condition for the best suited application areas.

The basic Linear Regression models the relationship between one independent variable and one dependent variable(J Neter, 1996), using the following function:

$$Y = f(X) \tag{2.1}$$

In the below example from Figure 2.2, retrieved from (J Neter, 1996) the presented function is Y = 2X, where each item sold for $2, so fifty units equates to $100 in sales.

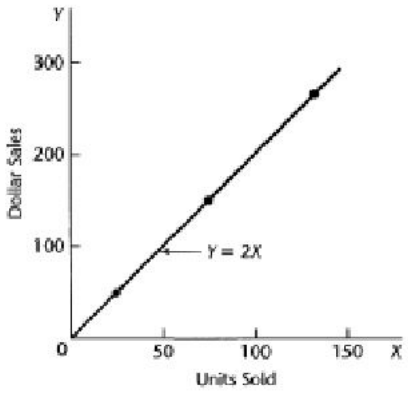


Figure 2.2: Simple Linear Regression

However, usually Linear Regression models have multiple variables, where data points will not lie on the function line, as shown in the above Figure 2.2. Figure 2.3 retrieved from (J Neter, 1996) presents a more typical Linear Regression model, where the data points will not necessarily lie directly on the regression curve. Data points will spread around the curve falling however within the probability distribution of Y, and therefor be close to the curve.

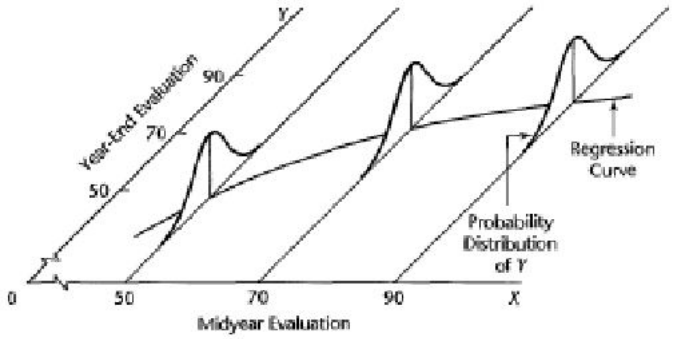


Figure 2.3: Multi Linear Regression

Logistic Regression is used to find the probability of the event, where the dependent variable is binary (0/ 1, Yes/ No). Logistic Regression is one of the most popular methods applied for propensity modeling, estimating propensity scores (Westreich et al., 2010), and classification problems. Logistic Regression is easy to implement in most of statistical packages (e.g. R, SAS) and is a familiar tool to most of the researchers in a variety of disciplines.

An example of application Logistic Regressions in Home Mortgage prediction modeling - Logistic Transition Matrix Approach (LTMA) that was proposed by (Molina Utrilla & Constantinou, 2011) and applied for mortgage default modeling. The proposed approach takes each borrower individually and calculates the probability of each mortgage moving from its current status to any other status, at the same time working out the probability of the mortgage staying in its current status. For example, if all the payments are up to date, then status of the mortgage can change from (1) current, to (2) 30 days in arrears or (3) paid off. If a mortgage is already 30 days in arrears then there are a number of other statuses it can change to including 60 days in arrears, or it can go back to being current with the repayments.

The below Figure 2.4, retrieved from (Molina Utrilla & Constantinou, 2011) shows that the proposed approach showed a good performance of the model:

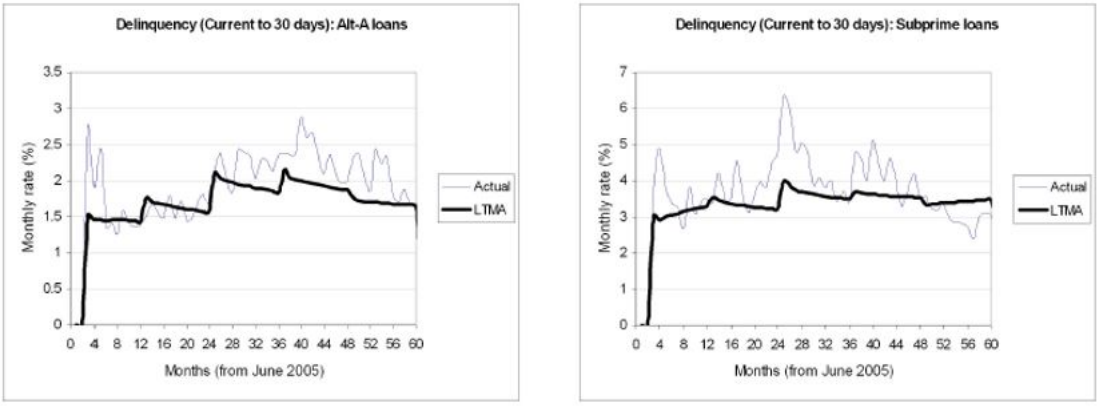


Figure 2.4: LTMA Performance

Stepwise Regression is another variant of regression techniques, that employs a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure.(*SAS/STAT 9.1 user's guide.*, 2004). In each step of the Stepwis Regression approach a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre-specified criteria. Agostinelli (2002) introduce a robust F-test and a robust stepwise regression procedure based on weighted likelihood in order to achieve robustness against the presence of outliers. Vélez, Ayuso, Perales-González, and Rodríguez (2020) in their research demonstrate a more sophisticated approach by fitting logistic regression models through a modified stepwise variable selection procedure, which automatically selects input variables while keeping their business logic, previously validated by an expert. In synergy with this procedure, a new method for transforming independent variables in order to better deal with ordinal targets and avoiding some logistic regression issues with outliers and missing data is also proposed. The combination of these two proposals with some competitive machine-learning methods earned the leading position in the NPS forecasting task of an international university talent challenge posed by a well-known global bank.

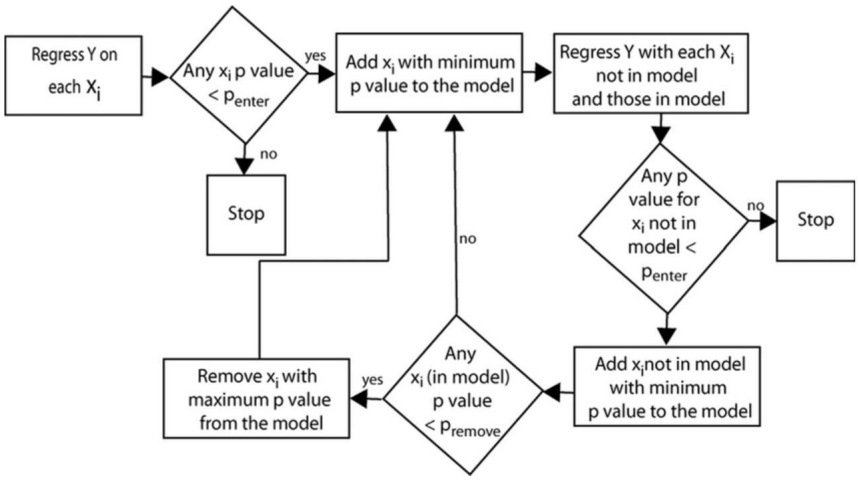


Figure 2.5: Stepwise Regression Process

Variant forms of regression models are considered as classical forms of modeling techniques for sales prediction models. However, in more recent years other approaches are taking over, such as Artificial Neural Networks, Decision Trees, Random Forest, Boosting models, to name just a few.

Artificial Neural Networks

Artificial Neural Networks (ANNs) have become popular modeling approaches for classification, clustering, pattern recognition and prediction in many disciplines, when researchers wanted to see if they could model and train computers to work in a similar way to the human brain (Abiodun et al., 2018; Celik & Karatepe, 2007). The main difference and advantage of ANNs in comparison to other machine learning techniques is in their parallel computing powers and ability to generalise on new data (Haykin, 1999).

The below Figure 2.5 retrieved from (Haykin, 1999) presents the ANN model architecture:

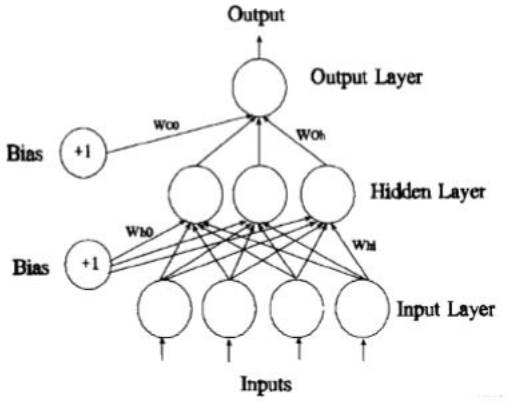


Figure 2.6: Artificial Neural Network

The versatility of ANNs means that they are almost universally applicable to most classification and predictive modeling tasks. Heo, Park, Kim, and Lee (2009) in their research proposed a two-step approach for predicting credit delinquents, by combining clustering with a multi-layer Neural Network. Authors have proposed the Multi-layer perceptron (MLP) methodology to build a better classifier. For the model created by (Heo et al., 2009) perceptron with three hidden layers was selected, as an optimal number of layers for generalisation, without overfitting the model. The model architecture is presented in the Figure 2.6 retrieved from (Heo et al., 2009).

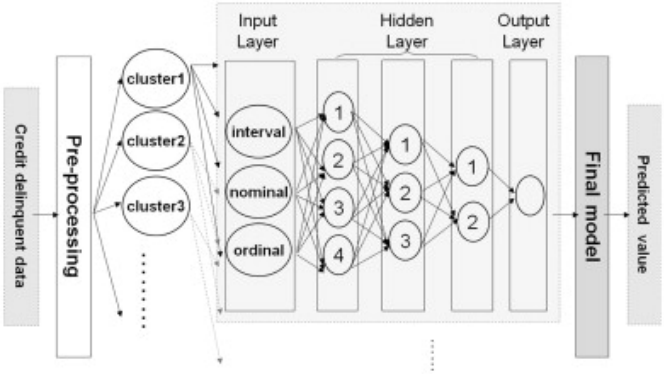


Figure 2.7: Perceptron Model Architecture

The proposed methodology of combination the clustering and Neural Network performed well with less computational cost, when compared with simple Neural Network model.

Scheurmann and Matthews (2005) build an ensemble of Neural Networks, to test the common theory that an ensemble of classifiers will perform better than a single classifier. Their findings state that in largely imbalanced datasets the proposed approach learns the majority class very well at the expense of the minority class, which is consistent with other resources.

Based on the additional reviewed resources Artificial Neural Networks were identified as one of the best performing algorithms when it comes to human behaviour (Cai, Qian, Bai, & Liu, 2020; Tkáč & Verner, 2016; Tavana, Abtahi, Di Caprio, & Poortarigh, 2018), with the main focus on the forecasting banking risks and crisis. Cai et al. (2020) propose a risk evaluation model based on Back Propagation Neural Network model (BPNN) - a multilayer feedforward neural network, which has the characteristics of error reverse transmission and signals forward transmission. The results show that the BPNN performed well in solving highly non-linear problems, such as evaluation of risks of the supply chain. Tkáč and Verner (2016) have examined in their research the performance of neural networks in evaluating and forecasting banking crises. Two artificial neural network models were tested, one works with the banking data belonging to the same date and another works with cross sectional banking data. Both ANN models showed good performance, indicating that integration of these ANN models in the frame of banking and financial sectors can contribute towards knowledge to prevail a banking crisis.

Decision Trees & Random Forest

Decision Tree in its simplest description follow a divide-and-conquer approach to classification and regression problems and are used used to discover features and extract patterns in large databases that are important for predictive modeling (Myles, Feudale, Liu, Woody, & Brown, 2004). Decision Tree algorithms start with "root" node (usually representing the whole population), and break it down into progressively smaller sets

of data to grow a tree into "branches" and "leaves." .When the decision tree splits the data in to smaller groups, it is trying to separate the classes, so that ultimately each smaller group of records should belong to one class. Usually for most of the predictive models this means that the data is broken down into a positive and negative class, which will ultimately be assigned to each of the smaller groups or leaves as they are commonly known.

The below Figure 2.5 extracted from (Quinlan, 1986) presents two different Decision Tree design approaches - a simple and complex:

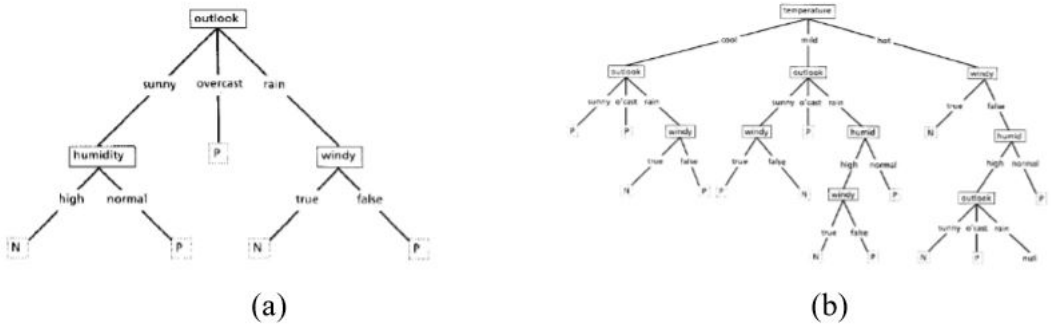


Figure 2.8: Simple Decision Tree (a) and Complex Decision Tree(b)

The drawback of the Decision Trees algorithms is in the general top down approach, where splitting the data at each of the nodes the algorithm only looks ahead one step. As a result the algorithm may not choose the optimal split for a node as each attribute is taken into account separately (Myles et al., 2004). Decision trees are also not very adept at handling missing values, or values that were not seen in the training dataset.

Random Forest approach helps to overcome this limitations. The Random Forest algorithm is based on the composing a forest of a myriad of different decision trees. The following main aspects define the Random Fores construction:

1. First, each tree is built on a random sample of the observations, according to the bagging method.

2. Second, for each tree of the forest, a random set of features is chosen to split nodes (feature sampling).

3. Finally, in order to use the model for prediction, the trees are aggregated by averaging the results when the outcome is numerical and by doing a plurality vote when predicting a class variable.

Popular use cases of Decision Trees and Random Forests modeling techniques in financial sector are customer segmentation analysis for marketing campaigns and loyalty / fraud detection (Ładyżyński, Żbikowski, & Gawrysiak, 2019; Tanaka, Kinkyo, & Hamori, 2016). Levin and Zahavi (2001) propose three different decision tree approaches for customer segmentation where targeting decisions should reflect customers' attitude toward the product / service involved (CHAID algorithm - a decision tree technique, based on adjusted significance testing, a variation of the AID algorithm, and a newly developed method based on genetic algorithm (GA)). Logistic regression model is used as a benchmark for the comparative analysis. The proposed decission tree approaches have benefits over the logistic regression as they are easy to understand and interpret, the output can be presented by means of rules that are clearly related to the problem stated in the research. Ładyżyński et al. (2019) propose a novel approach that can be adapted in banking industry, based on time series of customers' data representation for predicting willingness to take a personal loan. The system for identifying customers interested in credit products is based on classification with random forests and deep neural networks is proposed showed promising results proving that the system is able to extract significant patterns from customers' historical transfer and transactional data and predict credit purchase likelihood.

Boost Models

Boost are ensemble-based meta-learning algorithms that pool decisions from multiple classifiers and seek to improve the prediction power by training a sequence of weak models, each compensating the weaknesses of its predecessors(Adegoke, Chen, Banissi,

& Banissi, 2017). Roots for Boosting approach come from Probably Approximately Correct (PAC) learning framework that was first introduced by (Valiant, 1984) and is based on the question "Can a set of weak learners create a single strong learner?" The reasoning behind this technique is based on the logic that its much easier to find a number of rules that perform moderately well, as opposed to finding the one rule that performs exceptionally well (Schapire, 2002).

Most boosting algorithms are based on the approach where the weak classifiers are repeatedly fed different subsets of the training data, based on which the classification score for each of the iterations is given. Then the algorithm focuses on the most misclassified examples, and disregards the examples that have been classified correctly so that the algorithm is geared towards correcting these errors(Schapire, 2002).

The iterative approach achieves a better classification outcome for each of the weak predictors, and then the overall classification model is built by taking a weighted majority vote from each of the classifiers that provides a much stronger predictive capability when compared to the individual weak predictive models(Schapire, 2002).

The Adaptive Boosting algorithm (AdaBoost) is one of the most well-known boosting algorithms for binary classification problems (Bühlmann & Hothorn, 2007). The first AdaBoost algorithm was proposed by (Freund & Schapire, 1996). This method is based on the approach of adjusting the boosting during each iteration depending on the errors of the classifiers. In this way the algorithm adapts to the outputs of the model, and seeks to address the incorrectly predicted outcomes(Freund & Schapire, 1996).

Gradient Boosting approaches the problem differently from AdaBoost. Instead of adjusting weights of data points, Gradient Boosting focuses on the difference between the prediction and the ground truth. Gradient Boosting can be defined as a type of boosting where the objective is treated as an optimization problem and training is

done using weight updates by gradient descent (Friedman, 2001).

Boosting algorithms are popular choice in predictive modeling used in financial and banking sectors for different tasks like credit and loan assessment (Finlay, 2011) and direct marketing campaigns (Pan & Tang, 2014). Finlay (2011) proposed boosting algorithm called Error Trimmed Boosting or ET Boosting. Unlike the Adaboost, the ET Boosting algorithm does not apply weights to the observations, hence each observation is equally likely to be included in the construction of a classifier. AdaBoost and ET Boost were both applied to credit and loan datasets by (Finlay, 2011) to test the performance of the newly proposed methodology. ET Boost showed better performance than AdaBoost in all cases, although not enough evidence was provided that the algorithm can perform well in different scenarios, hence further investigation is required. Pan and Tang (2014) in their research carry out the experiment and compare different ensemble models, with application of bagged neural network, bagged logistic regression and gradient boosting to classify the customers who are likely to response to the marketing campaign. The Gradient Boosting performs slightly worse than bagged neural network model, although conclusion suggests that the model performance needs to be tested on different datasets in the area of bank direct marketing for fair evaluation.

2.2 Geospatial Intelligence in Business and Financial Sector

2.2.1 Basic Concept of Geo-spatial Science

Geo-spatial science in recent years has undergone important developments, presenting the solution through the visual geospatial displays to explore data and through that exploration to generate hypotheses, develop problem solutions and construct knowledge. (Kraak, 2003; do Nascimento & Eades, 2008). At the same time the spatial mining and machine learning techniques address the issue of computationally expen-

sive high density geo-spatial data by uncovering the unseen patterns in spatial data, and discarding the features that don't present insightful value. (Jankowski et al., 2014; Davidson, Drury, Lopez, Elmore, & Margolis, 2014; Wang & Yuan, 2014; Zhou, Li, Deng, Yue, & Zhou, 2018; Chen & Chen, 2010).

The problems in geospatial science have unique challenges that are rarely found in traditional machine learning applications, requiring novel approaches and methodologies(Kiely & Bastian, 2019). Deep collaboration between machine learning and geosciences is becoming more and more important for synergistic advancements in both disciplines. (Karpatne, Ebert-Uphoff, Ravela, Babaie, & Kumar, 2017)

2.2.2 Geospatial Analysis in Banking & Financial Industry

Geopsatial intelligence also know as location analytics is the process of deriving meaningful information from geospatial relationships to solve a specific problem. In conjunction with GIS - geographical information systems, location analytics present the insightful information through data visualisation for interpretation and discovery of unseen valuable knowledge (Yee, Ting, & Ho, 2018). Any area and industry can benefit from application of spatial and location based analysis, particularly business, enterprise, retail, government, banking, insurance, healthcare, pharmaceutical, automotive, travel, transportation, postal services, disaster planning, public safety, crime, airport, manufacturing and many more.

Droj and Droj (2015) propose approach with calculation of a commercial property value by integrating spatial analysis based on location and economic factors. The value of the commercial properties is directly linked to their location, hence when this information is combined and integrated with economic, financial and accounting information of a commercial property some elements of the property valuation process can be automated, optimizing the process and ensuring a more accurate estimate of the market value. However, no quantitative evaluation was provided as part of the research, to support the findings. Roig-Tierno, Baviera-Puig, and Buitrago-Vera

(2013) propose the methodology based on combined geographic information systems (GIS) and the analytical hierarchy process (AHP) approach for the retail site location selection. The AHP methodology shows that the success factors for a supermarket are related to its location and competition, the proposed retail site location decision process was applied to the opening of a new supermarket. To optimize the model performance another multi-criteria decision models (scoring, MAUT) can be tested as part of future research.

The optimal bank branch location problems have been extensively considered in the facility location research. The maximal covering location problem (MCLP) deals with the problem of selecting an optimal location taking into account that each customer may have specific requirements and the facilities have to be located to cover these requirements. (Atta, Sinha Mahapatra, & Mukhopadhyay, 2018). Cheng, Li, and Yu (2007) investigates the problem of determining optimum number and locations of ATM's applying GIS techniques. Petukhov, Zaikin, and Bochenina (2019) proposed the approach where geospatial profiles of bank customers can be used for marketing campaigns or to estimate customers' paying capacity. Algorithm for optimizing bank branches network is based on geospatial activity profiles of bank's customers applying additional layer of transaction historical data. Not enough evidence provided on model performance, it was tested over short period of time based mainly on historical data, however further evaluation can prove that the proposed approach can optimise the processes of branch location and customer target groups. Kiely and Bastian (2019) in their research on spatially- conscious machine learning models, propose the technique that ingrates the use of machine learning predictive modeling with spatial lag features usually seen in geographically-weighted regressions (GWR). The two-step modeling process is proposed, where the first step aims to determine the optimal building types and geographies suited to the feature engineering assumptions, second step performs a comparative analysis across several state-of-art algorithms (generalized linear model, Random Forest, gradient boosting machine, and artificial neural network).

2.3 Dealing with Data Imbalance

The issue with imbalance in class distribution has become more pronounced with the application of machine learning to the real world problems(Chawla, 2009). These applications range from bioinformatics, pharmaceutical and medical research, telecommunications management, text classification and speech recognition, environmental problems like oil spill detection. In prediction and classification models the imbalanced dataset can reduce models' effectiveness and accuracy, that may as a result cause a number of issues. Chawla, Bowyer, Hall, and Kegelmeyer (2002) in their research consider a dataset as being imbalanced if the number of records associated with each class is not approximately equal within the dataset, with the ratio being close to 50:50 in two classes, which is rarely the case in the real world scenarios. As a result most prediction and classification models will work well with a dataset that is balanced, but model performance significantly decreases when the dataset is skewed towards one of the classes(He & Garcia, 2009). To overcome data imbalance various sampling techniques are proposed and applied.

Random undersampling and oversampling techniques.

Random undersampling technique is based on the approach where common cases are reduced, and oversampling technique is based on the approach where rare cases are duplicated seek. These techniques seek to address the imbalance in a dataset and adjust the class distribution.

Random undersampling technique is based on the approach where random sample of records is taken from the original majority class within the dataset. The model is built then based on this sample rather than the all the records from the majority class. This has the effect of balancing the ratio of minority to majority class items.

Random oversampling technique is based on the approach where additional minor-

ity class items are added in to the dataset to even out the balance in the classes. Random oversampling takes a random sample of the minority class and adds these records to the minority class along with the existing records. Depending on the level of oversampling this approach has an effect of duplicating some/all of the minority class records in the dataset.

While the oversampling increases the overall number of records in the dataset, undersampling decreases the total number of records in the dataset (He & Garcia, 2009)

The two approaches can be effective in many application cases of addressing the class imbalance issue. However the pitfall of these techniques is that they tend to be biased, where random undersampling can potentially remove certain important examples, and random oversampling can lead to overfitting(Chawla et al., 2002).

Informed undersampling techniques.

The informed undersampling addresses the issue of random sampling and seeks to use sampling to decrease the number of majority class items in a dataset in a different way to random undersampling(Liu, Wu, & Zhou, 2009). EasyEnsemble and BalanceCascade are two popular informed undersampling techniques, however there are many other approaches available (He & Garcia, 2009). Liu et al. (2009) evaluate the EasyEnsemble approach that samples several subsets from the majority class, trains a learner using each of them, and combines the outputs of those learners by creating an ensemble, which is boosted using the AdaBoost algorithm to create a classifier. BalanceCascade takes a supervised approach to the sampling of the majority class, and works in a very similar way to EasyEnsemble to work out which records from the majority sample should be included.

Ensemble modeling in general aims to alleviate the effect of data imbalance by using multiple models in unison to obtain better predictive performance. The main

motivation for combining classifiers is to improve their generalization ability; while each classifier will make errors, they will not necessarily make the same errors(Ali, Shamsuddin, & Ralescu, 2015)

Synthetic sampling techniques.

Synthetic sampling methods aim to resolve the lack of minority samples in the dataset by adding new samples into the dataset for the minority class. Chawla et al. (2002) proposed the synthetic sampling method - Synthetic Minority Oversampling Technique (SMOTE),where the new data items added to the dataset are synthetic samples that have been created to be similar, but not identical to the pre-existing minority data items.

The below figure extracted from (He & Garcia, 2009) shows how a synthetic example is created for the original data item x_i. The six nearest neighbours for x_i are identified in Figure (a) and then the new synthetic data item is created between x_i and x^i in Figure (b). If more than one synthetic item was required the new item would be created between x_i and another of its nearest neighbours.

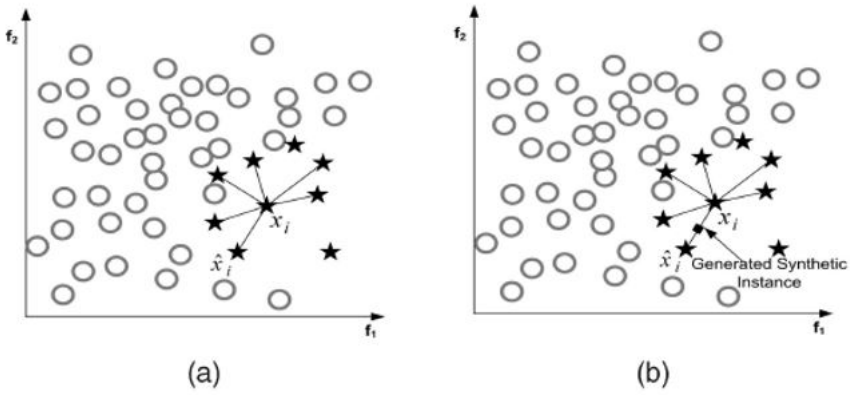


Figure 2.9: Synthetic data creation using SMOTE

There certain limitations with the this method. The main drawback to using SMOTE is that the generated model may overgeneralise due to the synthetic data that has

been added in, which can lead to the issue of overlapping between classes(He & Garcia, 2009). Given that there are certain limitations to SMOTE, a number of alternative adaptive synthetic sampling methods have been proposed. Han, Wang, and Mao (2005) in their research proposed Borderline-SMOTE one of the Adaptive Synthetic Sampling methods that looks to focus on the data items in the minority class that are closer to the majority class, as opposed to all of the items in the minority class. The reasoning for this approach is that the data items in the minority class that are most difficult to classify, are the items that are closest to the majority class.

The below figure extracted from (Han et al., 2005) presents an example of a dataset with an imbalance towards the majority class in Figure(a), Figure(b) presents the 'borderline' points that have been selected by the Borderline-SMOTE algorithm to use for producing the synthetic data items, (c) shows the dataset with the new synthetic items added in. These new synthetic items are primarily focused around the border of the two classes, so this should help with the generalisation of the model for the 'borderline' items in the minority class.

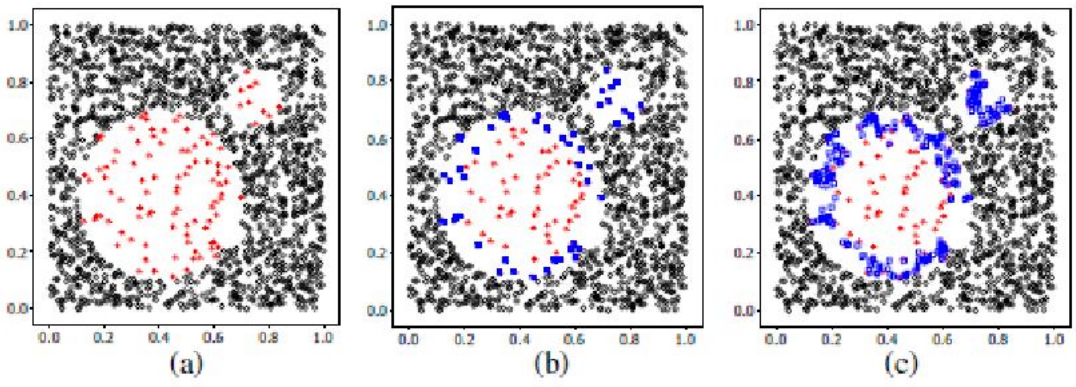


Figure 2.10: Borderline-SMOTE Example

The below figure extracted from (He & Garcia, 2009) presents the data point A,that is said to be in the 'DANGER' set, which is the dataset where the items have both classes as neighbours. This is one of the items that will be used to create synthetic

data, whereas points B and C will not

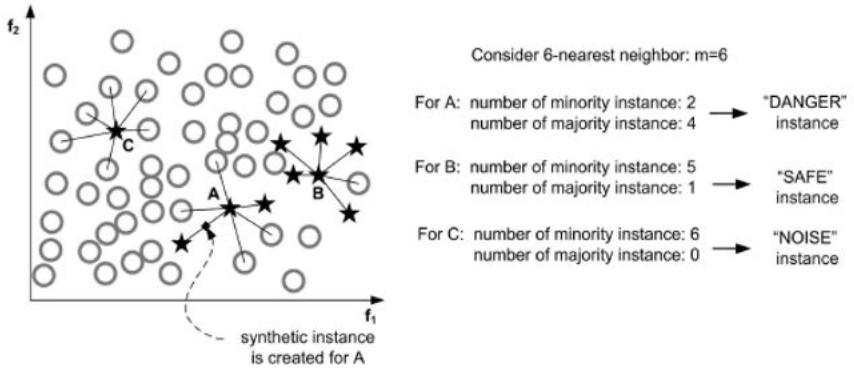


Figure 2.11: Synthetic data creation using Borderline-SMOTE

Another synthetic sampling technique is ADASYN, that seeks to generate more relevant sampling data rather than just generating synthetic samples for all data items in the minority class. In many ways this approach is similar to Borderline-SMOTE as it creates synthetic data based on the items in the dataset that are likely to be the hardest to classify. The ADASYN method uses a weighted distribution to work out which data items need to have more synthetic items created, based on the distribution of the minority class. data items that would prove For the more difficult to classify data items more synthetic example are created, and for those that are easier to classify there are less synthetic examples created. As a result the class imbalance issue is addressed in the whole dataset, and also compensates for skewed distributions within the minority class itself (He & Garcia, 2009).

He and Ghodsi (2010) in their research are highlighting the main issue with SMOTE method - biases of the classifier towards the minority class to improve the detection rate of the rare objects.

2.4 Model Evaluation Methods

In the same way as there are many different approaches and techniques for building predictive models, there are many different ways to assess and evaluate these models.

The below table represents a confusion matrix, that can be generated to any classification or prediction model. It shows the number of records that were correctly classified as as either true positives (TP) or true negatives (TN), as well as the records that were misclassified as either false negatives (FN) or false positives (FP).

	Predicited Negative	**Predicted Positive**
Actual Negative	TN	FP
Actual Positive	FN	TP

Table 2.1: Confusion Matrix

The overall accuracy rate is one of the main methods used to quantify the accuracy of a model, that is defined as the number of outcomes that were correctly predicted, over the total number of outcomes in the dataset:

$$Accuracy = \frac{TP + TN}{TN + FP + FN + TP} \tag{2.2}$$

The misclassification rate is inverse to the accuracy rate, defined as the number of false positives and the number of false negatives divided by the total number of records in the dataset:

$$Misclassification = \frac{FP + FN}{TN + FP + FN + TP} \tag{2.3}$$

For example, if out of 100 items in the dataset 90 are classified correctly as either TP or TN, and 10 misclassified, then the accuracy rate of the model will be 90%, while the misclassification rate will be 10%.

The drawback of accuracy rate and misclassification rate is, although traditionally they were most commonly used measures to gauge the performance of a model, they tend to to overlook deficiencies in the imbalanced datasets. Sun, Wong, and Kamel

(2009) highlight that in classification with the class imbalance problem, accuracy is no longer a reliable measure since the rare class has very little impact on the accuracy as compared to that of the prevalent class. If in the example of the dataset given previously, there were ninety items correctly classified, but all of these ninety belonged to one class, and the ten that were classified incorrectly all belonged to the opposite class, then the model would not actually have performed to 90% accuracy as the accuracy rate would suggest.

The Precision value of the model is the number of TP over by the total number of TP & FP. Precision can be viewed as the accuracy of all the records that have been classified as positive by the model:

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

Recall measure is also known as Hit Rate or Sensitivity. This measure is a more accurate reflection of the true accuracy of the model, where the TP class is the class which is the most important in the prediction. The recall rate gives a better reflection of how the positive target class is being classified, which is a good indicator of how the predictor is performing, especially in the case of imbalanced datasets:

$$Recall = \frac{TP}{TP + FN} \tag{2.5}$$

Specificity is defined as a measure of how well the model has performed at classifying the negative class:

$$Specificity = \frac{TN}{TN + FP} \tag{2.6}$$

With many possible predictive models available to choose from, metrics that can compare the models and select the best one are necessary for making final decisions set for a specific task. For example, (Pradhan, 2013) in their comparative study apply some commonly used metrics like Receiver Operating Characteristics (ROC) curve, Cumulative Gains Chart and Lift Chart to select the best performing model. All of these

provide metrics by trading off desirable outcomes (i.e. correct predictions) against undesirable outcomes (false positives or false negatives). These metrics are obtained by running the model on the training data set (used to create the model) or on an out-of-sample validation set. ROC Curve plots True Positives along the y-axis and False Positives along the x-axis. Visually, the higher the curve is above the 45 degree line, and the closer it is to the top left corner, the better the model.

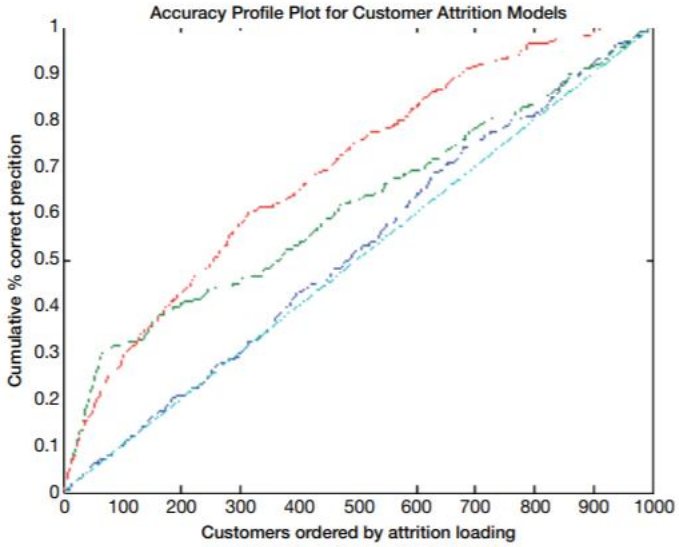


Figure 2.12: ROC Curve Example

Cumulative Gains Chart and Lift Charts are commonly used by marketing departments as they allow for direct visual comparison and interpretation of results with respect to marketing campaigns.

2.5 Gaps in the Literature

Different modeling techniques were reviewed as part of the Literature Review task, including variant Regression techniques (Agostinelli, 2002; Vélez et al., 2020; Molina Utrilla & Constantinou, 2011; Westreich et al., 2010), Neural Networks (Celik & Karatepe, 2007; Cai et al., 2020; Haykin, 1999; Moro, Cortez, & Rita, 2014; Westreich et al., 2010; Tavana et al., 2018; Tkáč & Verner, 2016), Decision Trees (Myles et al., 2004; Tanaka et al., 2016; Quinlan, 1986; Westreich et al., 2010) and Boosting Models (Huang &

Yen, 2019; Bühlmann & Hothorn, 2007; Pan & Tang, 2014; Friedman, 2001). It can be concluded that each each of the reviewed approaches have their strengths as well as their weaknesses, and there can't be a simple "one model fits all" scenario. Much of the reviewed literature suggest that Artificial Neural Networks perform well in financial sector modeling tasks, however due to its 'black box' nature, the practical application of this approach in financial institutions may be challenging due to the GDPR regulations. Generally a good practice in predictive modeling despite the area of application is testing a number of different approaches for each classification or prediction problem. Much of the reviewed literature in financial industry relate to the customers' segmentation analysis and prediction of defaults and arrears, and limited sources available on home mortgage sales and loan prediction analysis, due to the sensitive nature of information.

The problem of collaboration between machine learning and geosciences (Kraak, 2003; do Nascimento & Eades, 2008) and application of GIS techniques in different business areas (Droj & Droj, 2015; Roig-Tierno et al., 2013; Atta et al., 2018; Cheng et al., 2007; Petukhov et al., 2019) have been studied and proposed many researchers. Optimal site selection is identified as one of the most common problems where the GIS is applied (Roig-Tierno et al., 2013; Cheng et al., 2007; Atta et al., 2018). Roig-Tierno et al. (2013) propose combined geographic information systems (GIS) and the analytical hierarchy process (AHP) methodology for the retail site location selection.

Literature on working with imbalanced data was also reviewed as well as the best techniques in dealing with imbalanced datasets, which is a common issue in financial institutions (Ali et al., 2015; Han et al., 2005; Sun et al., 2009; Chawla, 2009; Liu et al., 2009; Chairi, Alaoui, & Lyhyaoui, 2012; Ilyas et al., 2020). Different methods were reviewed, that can be applied for dealing with imbalanced datasets, including over and under sampling, synthetic sampling techniques, as well as the ensemble modeling. Some of these methods will be implemented in this research project to address the imbalance that currently exists in the dataset.

To the author's best knowledge there was no previous research conducted, where GIS techniques are integrated in the home mortgage sales prediction modeling task, that takes into account changing nature of the customers banking habits. The suggested approach is also novel in its practical implementation within the financial institution the data is retrieved from.

Chapter 3

Experiment design and methodology

3.1 Introduction

The experiment design and methodology that is carried out as part of the research project is presented in this chapter.

Data sources, data collection, prepossessing and transformation are described in some detail. However, due to the sensitive nature of this data, it cannot be described and presented in full detail.

Details of the machine learning algorithms selection and implementation, as well as detailed steps of how the experiment is carried out as part of modeling phase are outlined.

Methods that address the imbalance in the dataset are considered and the approaches are covered in respect to the given experiment.

The evaluation methods used to assess the applied algorithms and models built are presented.

3.2 Experiment Overview

The experiment is carried out to to design, build and evaluate a predictive model that should be capable to predict with high degree of accuracy what existing mortgage customers within the financial institution are likely to consider to apply for another mortgage in the rural areas of Ireland.

Taking into account an increase in internet and digital banking development and changing banking habits of the customers the second phase of the development is set to identify whether the distance to the branch locations from the residential addresses of the customers has an impact on the decision to purchase the mortgage with the given financial institution. The distance feature is calculated and the second phase of the experiment carried out to evaluate whether it has an impact on the predictive model.

Phase 1 - Base Mortgage Mover Predictive Model:

The Model development process is implemented following similar to Crisp-DM methodology approach:

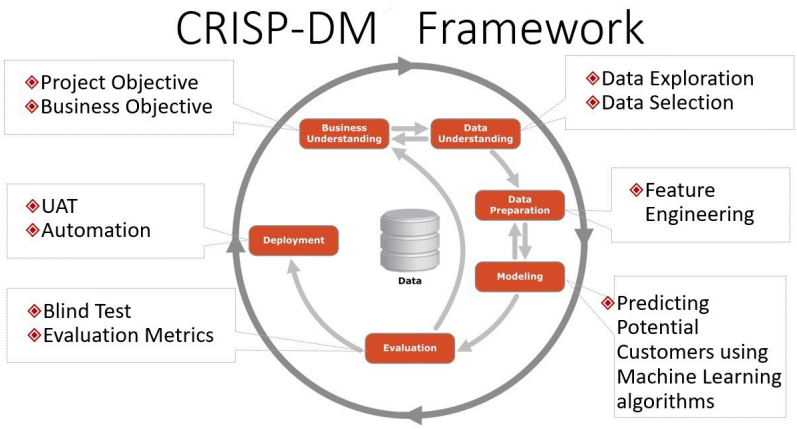


Figure 3.1: Crisp-DM Methodology

1. Business Understanding - understand what is set to accomplish from a business

perspective.

2. Data Understanding - defining what data the model should learn from (demographic, products held, transactions, risk factors, online banking usage, balances).

3. Data Preparation - ABT (Analytical Base Table) is built by extracting a large number of Customer attributes from various sources.

4. Attribution Reduction - all attributes can't be used in the models, so similar attributes are grouped.

5. Modeling - several models are developed based on the literature review, using a variety of machine learning techniques and different attributes.

6. Evaluation - the performance of each model is evaluated using a blind test, the best performing model is selected for the second phase of the development.

Phase 2 - GIS Implementation:

1. Distance Calculation - calculate distances from the Branch locations to the customers residential addresses, based on the ABT from the phase 1 customer base, draw visual maps

2. Update ABT, having all the features as in phase one with addition of distance feature.

3. Build models based on the models architecture from the phase 1 with the distance feature added to the model

4. Evaluate the performance of the models with the distance feature included.

In case if spatially conscious models outperform non-spacial models, further research can be carried out in optimisation of Branch locations and Home Mortgage services.

3.3 Software Selection

Software used to carry out the experiment are: SAS Enterprise Guide & SAS Enterprise Miner, QGIS.

The data selection, initial exploration and Analytical Base Table built in the proprietary Data Warehouse. Modeling approaches tested and evaluated using proprietary software SAS Enterprise Miner and the best performing model is deployed. Distance calculation, creation of the visual maps performed in QGIS.

3.4 Data Collation and Preparation

The data is collected from the enterprise data warehouse for all mortgages in scope for the experiment. The data used for building the models is based on the period dates from 28/02/2019 to 28/02/2020. Date range is defined by looking back at the period of 12 months. The ABT table was created for the purposes of this research on 28th of February 2020.

To limit data, excluding Dublin and Dublin commuter areas, the following counties are not included in the ABT: Dublin, Wicklow, Meath, Louth and Kildare, based on the county name in the address line of the customers residential addresses.

The following criteria is taken into account and data is limited to, for the Mortgage Mover ABT:

- Customer currently holds mortgage product with the financial institution
- Current mortgage length
- Loan to Value (LTV) Ratio, that is calculated as following:

$$LTV = \frac{CurrentLoanBalance * -1}{CurrentPropertyValue} \tag{3.1}$$

- Current accounts balance ratios (3 months average)
- Joint current account credit balance ratios (3 months average)
- Saving rate (3 months average)
- Floor Area
- Number of bedrooms
- Property status (e.g - self built) and property type (e.g - detached house)
- Hardware spent ratios (3 months average)
- Home improvement loans (count & amount)
- Difference between current income and income at application

At the high level target customers for the mortgage mover propensity model have the following characteristics:

Base Population Selection:

- Active customer with active current accounts
- Personal and Staff Customers
- Certain limit to the credit grade, set by the financial institution internal rules
- Customer has minimum 6 months of history
- Age 30-50, not retired and not deceased
- Derived income at least 45K
- Customer can not be in negative equity (customer with LTV more or equal to 100 is included)

Recency:

- Must have not made a mortgage application within last 12 months

Population Eligibility Criteria:

- Must have an active private home loan with the financial institution, and have held the mortgage for at least two years.

The prior size of the dataset generated consists of 19737 records, with 511 targets & 19226 non-targets identified.

The below table represents an example of the data items to be included in the table, due to the sensitive nature of the data the full dataset can not be provided:

Column Name	Data Type	Description
Custom No	Integer	Unique customer identifier
Appl No	Integer	Unique mortgage identifier
Target Ind	Integer2	Mover or Non-Mover, binary target identifier
Period Date	Date	Measured period date
Credit Grade	Varchar	Customer's credit grade
County Name	Char	Address county name to identify Rural areas
Address	Char	Current address for the geocoding & distance calculations
Age	Double	Age of the customer
LTV	Double	LTV ratio
Appl Source	Char	Channel the existing mortgage purchased through
Months since opened	Double	How long since the existing mortgage purchased
Property Status	Char	Existing property status, e.g new self build
Propery Type	Char	Existing property status. e.g. detached house
No Bedrooms	Double	How many bedrooms the existing property holds
Energy Ratings	Double	Energy ratings of existing property floor area
Floor Area	Double	Existing property floor area

Table 3.1: Data Items Example to be Included in the Model

The below flowchart presents the data modeling process:

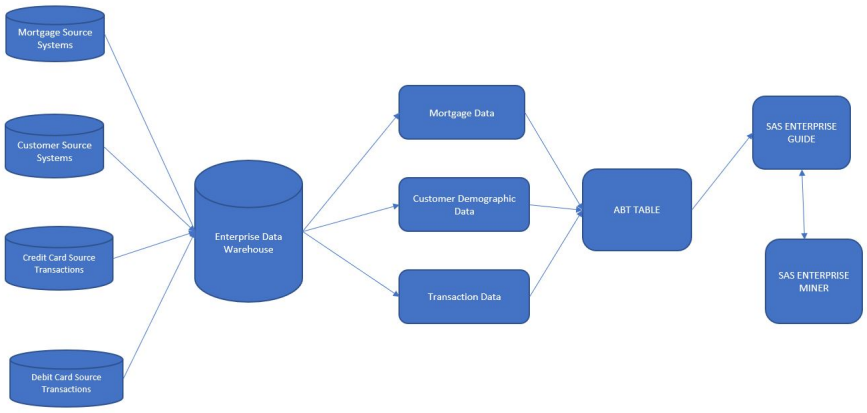


Figure 3.2: Flowchart of Data Modeling

3.5 Model Selection and Development

Based on the literature reviewed in Chapter 2 of this research the following models are selected and built to evaluate the performance of the task set: **Random Forest, Gradient Boosting and Ensemble Gradient Boosting**. As previously mentioned Neural Networks were originally considered for this research as part of the experiment, however due to the GDPR regulations the financial institution refrains from utilising this approach due to its 'black box' nature in calculating weights. There are other types of models, like for example SVM and Decision Trees, that could be considered for this work, but the three selected models are readily available in the enterprise software, and are well known and recognised by many researches.

The following flowchart presents the experiment modeling process:

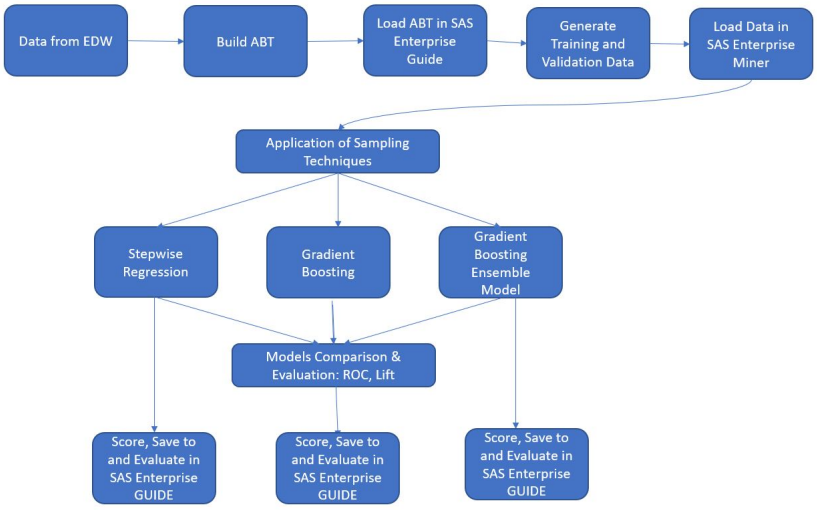


Figure 3.3: Flowchart of Modeling Process

Initially it was considered to include the Stepwise Regression in the experiment, as presented in the chart. However, upon further research and models evaluation process it was decided to replace Stepwise Regression model with Random Forest model. 19737 records from the dataset are split in train and test datasets. Each of the models produced is trained using the training data available. The models individually decide which data variables are most important as part of the training phase, and then the test dataset is used to validate the model output.

The initial observations show the following distribution in the source mortgage application channel - out of total population the majority of the customers applied for their existing mortgage through the Branch channel. This confirms the anticipation that in order to purchase a mortgage product customers in majority of the cases prefer personal interaction, despite the changes banking habits.

The detailed breakdown is provided in the below table:

Application Source	Target Id	Total Customers
Branch	1	402
Branch	0	16024
Broker	1	1
Broker	0	8
Direct & Online	1	65
Direct & Online	0	1101
Staff Business	0	43
Unknown	1	53
Unknown	0	2050

Table 3.2: Application Source Breakdown

The dataset of 19737 records includes 511 targets and 19226 non-targets retrospectively, having only 2.6 % of targets in the population. This indicates that data is heavily imbalanced. To address the data imbalance issue discussed in the Chapter 2 a number of methods are tested and implemented to allow model accurately predict the minority class.

3.6 Spatial Features Calculations and Maps visualisation

To carry out spatial part of the experiment the following steps are executed:

- Shapefile of Ireland Electoral Divisions (EDs) is loaded in QGIS
- Latitude ans longitude of the branch locations are extracted from data warehouse and generated in the CSV format. The branch locations are uploaded and displayed on the map.

- Customers residential addresses are geocoded and displayed on the map. Not all the addresses are generated from the first attempt and additional steps are executed, to overcome this complications.

    1. Records with all blank addresses identified and discarded, leaving the population size: 17633, with 459 targets and 17174 non-targets
    2. Address line 1 and 2 concatenated in cases where adress line did not contain valuable data on its own, for example having only number record
    3. Updated records are extracted from EDW and stored in CSV format
    4. CSV file is uploaded to QGIS,the MMQGIS is installed - the collection of QGIS vector layer operations plugins is used for geocoding a CSV file containing address data.
        (a) Two options can be selected: one for geocoding the address file using either Google or OpenStreetMap geocoding web services or by geocoding from a street layer.
        (b) OpenStreetMap option was selected to geocode the adresses
        (c) Despite multiple attempts, braking down adress file in smaller size csv files (8 scv files where populated) some of the records still not all the addresses could be geocoded, 16402 have been generated (83% of initial records), with 15974 non-targets and 428 targets.
    5. This population will be used in both versions of the ABTs for non-spatial modeling phase, and modeling with spatial features included to have fair evaluation of the 2 approaches.

- QGIS is very powerful in analyzing spatial relationship between features. Distance Matrix tool available in QGIS allows to carry out the nearest neighbor analysis. Distances from residential addresses to the closest branch locations are calculated employing distance matrix tool.

- Calculated distances are extracted and added to the updated ABT 2 table, that is used in the second modeling phase. ABT 1 table is updated, containing only

the same unique identifier customers' records as in ABT 2.

The below map presents branch locations and residential addresses overview example. The map view is presented on the high level, due to the sensitive nature of the data:

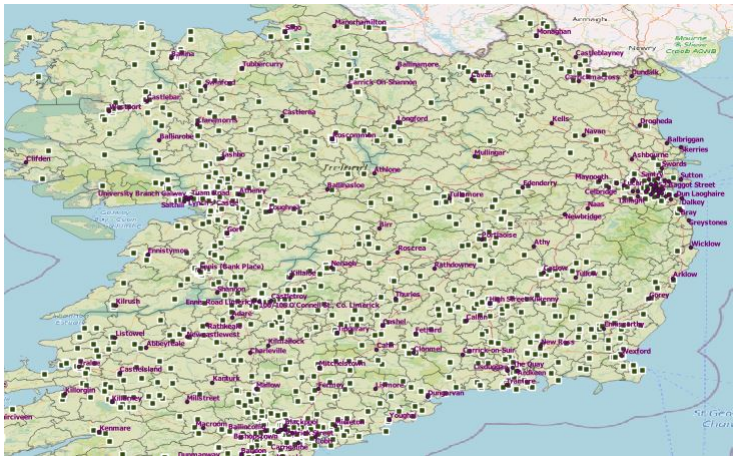
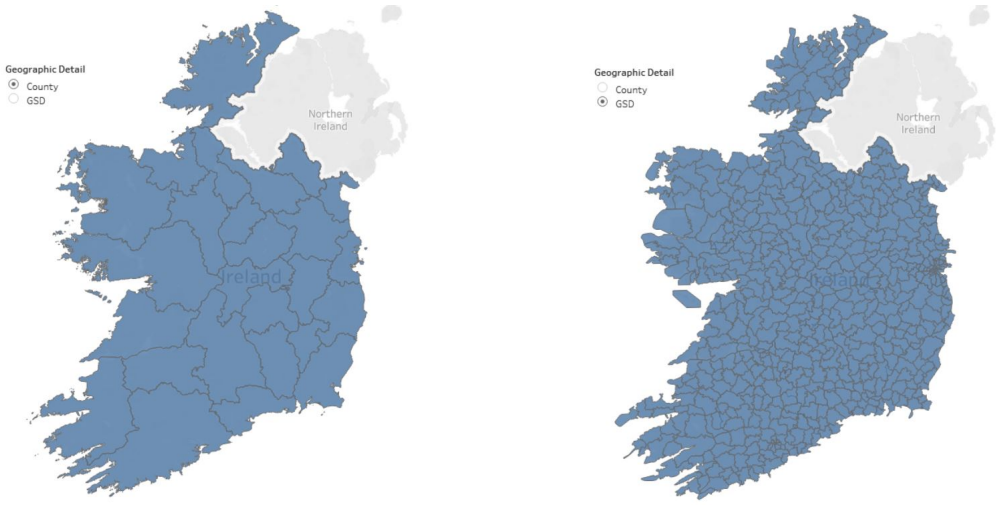


Figure 3.4: Branch & Addresses View

Upon further research it was decided to bring in the ABT 2 in the second modeling phase of the experiment another spatial feature based on the Garda Sub District - GSD metric that customers can be assigned to.

To assign customers to the GSD codes, the script was implemented in R, that essentially looks in the EDW Customer Location Reference table for instances where a GSD appears anywhere in the address fields. It also takes into account smaller area names which appear in the GSD itself, as well as some potential spelling mistakes used in the addresses (links to the Variation column from the EDW Customer Location Reference table). Then the addresses from Location Reference table are inner joined with addresses in ABT2, GSD is appended to ABT 2 for the identified customers.



(a) County Level Map

(b) GSD Level Map

Figure 3.5: County and GSD Maps of Ireland

Advantages of applying GSD (Garda Sub District) level for customers segmentation based on their residential addresses are comprised of the following facts:

1. GSD presents the lowest level available of the organisational structure employed by An Garda Síochána. The comparison view is presented in the Figure 3.5, County vs GSD level.

2. GSD level of granularity, comparing to higher levels of granularity available in Ireland is as following: 6 Regions - 25 Divisions - 109 Districts - 563 Garda Sub-Districts.

3. CSO (Central Statistical Office) reports figures on the GSD level, for example Census Data.

4. GSD allows to segment the country via granular boundary level.

Chapter 4

Results, evaluation and discussion

Detailed information on the experiment implementation and models evaluation are provided in this Chapter.

4.1 Experiment Implementation

The ABT 1 and ABT2 consist of 16402 observations. Each observation represents a customer and is described by both categorical and numerical attributes. The target variable represents whether the customer will apply for the home mortgage mover loan or not, with 15974 non-targets and 428 targets identified. 2.6% of targets from overall population indicate that dataset is highly imbalanced and the issue will need to be addressed. Models are trained on 310 variables for the Phase 1 of the experiment, and 312 variables for the Phase 2 of the experiment. Both tables are loaded in SAS Enterprise Guide, where the initial analysis data transformation are performed. Next the two ABT tables are uploaded in SAS Enterprise Miner.

As previously outlined the following models are trained: Random Forest, Gradient Boosting and Gradient Boosting Ensemble Model in SAS Enterprise Miner. Different settings where tested, and the final models were trained following the same structure for the two Modeling phases, 1 - with no spatial features included, 2 - with spatial features included for the fair evaluation of the models performance.

The below Figure 4.1. presents the final models training interface, that is identical for the two models:

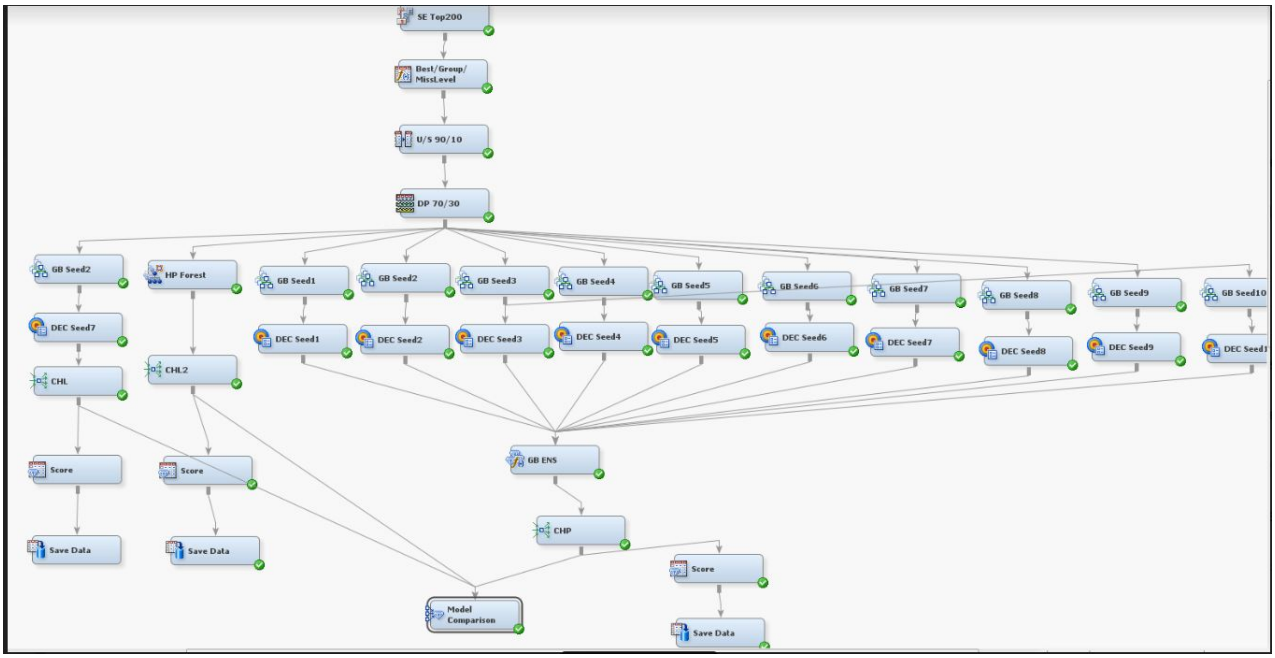


Figure 4.1: SAS Enterprise Miner Models Training Interface

As presented in the Figure 4.1, the following steps were implemented to train all the models in Phase 1 and Phase 2 of the experiment:

- SAS StatExplore node is connected to the ABT 1 table in the first modeling step, and ABT 2 table in the second modeling phase, and 200 top features were selected as the most significant in order of their predictive power.
- Trasform Variables node is connected to the StatExplore to group rare variables and treat missing values as level.
- Random undesampling technique is applied, with 90%:10% (3852 non targets, 428 targets) ratio to address the imbalance in the dataset, oppose to the initial distribution of 97.4%:2.06% (15974 non targets, 428 targets).
- Sample size of 4280 observations is split into training and validation data with 70:30 ratio, 2994 of overall are observations are used for training, and 428 for

validation of the models. As a result of random undersampling, number of target variables for models training is 299, and validation is 129; number of non-target variables for training is 2695, and validation is 1157.

- Random Forest, Gradient Boosting and Ensemble Gradient Boosting models are trained.
  - Random Forest model is trained by selecting the HP Node in SAS Enterprise Miner, with maximum number of 50 Trees in the forest and default seed size
  - Gradient Boosting model is trained, with 50 iterations and 1102 seed size. It was decided to bring in Decision node after initial testing of the models, that improved the performance of the model by defining profiles of the targets that produce optimal decisions.
  - Ensemble Gradient Boosting is composed of 10 Gradient Boosting models, all with 50 iterations but different seed sizes. Decision node is connected to each individual Gradient Boosting classifier, following the same architecture as in the simple Gradient Boosting model.
- All the models are compared and evaluated with ROC selection of statistics.
- Scores for each of the models are saved in SAS Enterprise Guide for future research purposes.

4.2 Experiment Evaluation

Phase 1 Models Evaluation

Confusion matrix is typically used to evaluate the performance of the classification models. However, as discussed in the Literature Review part of this research, due to the class imbalance, the overall classification rate is no longer appropriate because the minority class has less influence on accuracy as compared to the majority class.

The most relevant measures for the imbalanced data are precision, recall, F-measure, sensitivity, specificity, ROC curve, AUC, lift and cumulative lift charts. In the context of this work sensitivity, specificity, ROC curve, lift are chosen as the criteria in measuring the models performance. Sensitivity is the percentage of true positive instances, that are correctly classified by the model. Specificity is the proportion of the true negative instances, that are correctly detected by the classifiers. The ROC - a receiver operating characteristics curve presents the trade off between true positive and false positive rate of a classifier and graphically displays sensitivity versus 1-specificity.

The Models in the chart below are presented in the following order Random Forests(HP Forest), Gradient Boosting Ensemble (GB ENS) and Gradient Boosting (DEC Seed 7).

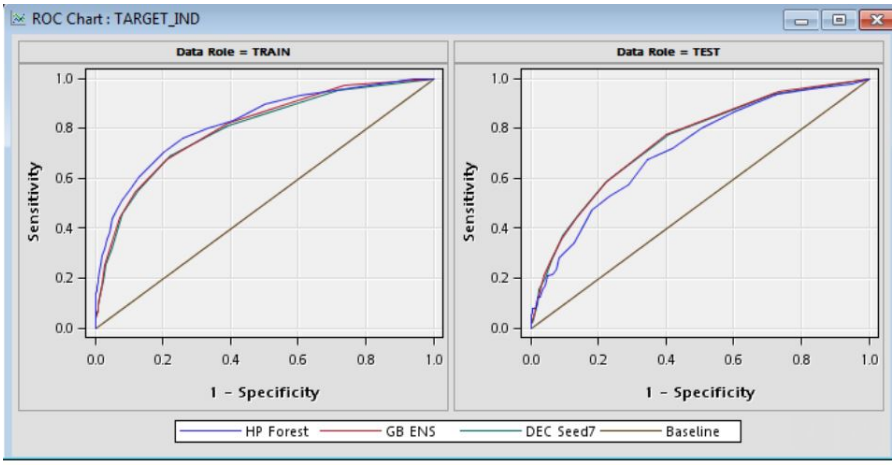


Figure 4.2: Phase 1 Model - ROC Chart

The Table 4.1 presents the ROC scores produced by the three models in training and validation sets. The best performing model according to this evaluation method on the validation set is Ensemble Gradient Boosting, followed by Gradient boosting model. The lowest score, indicating worse performance is produced by Random Forest model on the validation set. The threshold of more than 0.05 in ROC Curve Index indicates the overfitting error, that refers to a model that models the training data too well, in comparison to the previously unseen data. The highest threshold value is generated

by Random Forest model, that suggests that the model is overfitting.

Model Name	**ROC Test**	**ROC Train**	**Train /Test Threshold**
Ensemble GB	0.759	0.809	0.05
Gradient Boosting	0.757	0.804	0.047
Random Forest	0.715	0.824	0.109

Table 4.1: Phase 1 Model - ROC Evaluation

Another method considered for the models performance evaluation in the context of this research is Lift, that helps to choose between the competing models. Lift represents a measure of the effectiveness of a classification model calculated as the ratio between the results obtained with and without the model.

Model Name	**Lift Test**	**Lift Train**	**Train /Test Threshold**
Ensemble GB	4.873	3.55	1.319
Gradient Boosting	4.813	3.694	1.119
Random Forest	4.706	2.627	2.078

Table 4.2: Phase 1 Model - Lift Evaluation

The threshold of more than 1.00 in Lift measure indicates that the overfitting error may be present in the models. As per results provided in the Table 4.2 the overfitting is suspected in all three models, with the highest probability of overfitting in the Random Forest model.

Phase 2 Models Evaluation.

Models with addition of spatial features of Distance and GSD, are evaluated in the second phase of the modeling process. The evaluation methods are consistent with approaches applied in the first modeling phase.

ROC Chart illustrates better performance of all three models in comparison with phase 1 models.

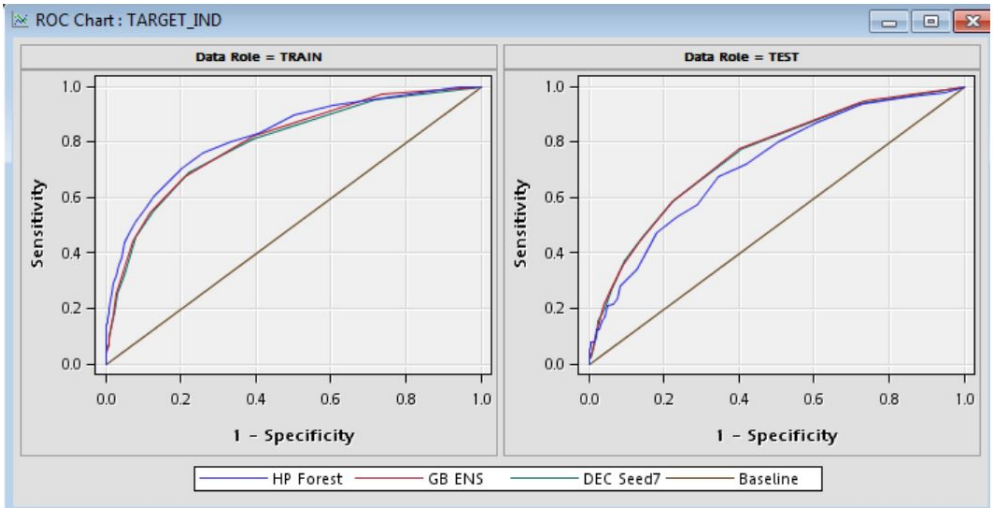


Figure 4.3: Phase 2 Model - ROC Chart

While Random Forest may still be overfitting, the threshold in train and validation scores for Gradient Boosting and Ensemble Gradient Boosting models are lower than in phase 1.

Model Name	ROC Test	ROC Train	Train /Test Threshold
Gradient Boosting	0.883	0.897	0.014
Ensemble GB	0.88	0.9	0.02
Random Forest	0.77	0.867	0.097

Table 4.3: Phase 2 ROC Models Evaluation

Results produced by Lift method also indicate that performance of the models that include the spatial features has improved when comparing with the phase 1 models, generating higher scores in validation set. The Random Forest model is still overfitting, while the threshold for Gradient Boosting and Ensemble model has decreased.

Model Name	**Lift Test**	**Lift Train**	**Train /Test Threshold**
Gradient Boosting	6.000	5.648	0.351
Ensemble GB	6.215	5.332	0.883
Random Forest	5.440	4.173	1.267

Table 4.4: Phase 2 Model - Lift Evaluation

4.3 Experiment Results & Discussion

This research is set out to evaluate if the integration of spatially conscious features in the mortgage mover binary classification model can improve the model performance. Based on the evaluation results provided as part of models evaluation process, the spatial features have significant impact on all three models that were trained as part of the experiment. Moreover, while the GSD feature was discarded from the models, as deemed not being significant in selecting top 200 features step of training the models, the Distance feature had the highest weight in all the cases. Gradient Boosting and Ensemble models showed better performance, when comparing with Random Forest models in Phase 1 and Phase 2 of the experiment. Performance of the Gradient Boosting and Ensemble Gradient Boosting had significantly improved with integration of spatial features of the models. The threshold that indicates overfitting in the models have decreased in Phase 2 of the modeling, where spatial feature were introduced. The Random Forest also showed an improvement with spatial features introduction, however the overfitting issue still remained present.

These experiment results can lead to further research, where different population base & models architecture can be tested.

Chapter 5

Conclusion

5.1 Research Overview & Problem Definition

The digitization has an impact on almost every area of the modern world, while direct banking development changes in many ways how the customers interact with financial institutions. As part of this research the scope was defined to seek whether the physical presence of the branches in close proximity from the customers locations has still an impact on home mortgage sales. This research sought to apply geospatial analysis and spatially conscious features in the context of the machine learning techniques. The Mortgage Mover binary classification models were built to carry out the experiment and evaluate the significance of spatial features. The customer base was limited to the rural areas of Ireland, taking into account that factors defining customers decisions to apply for the Mortgage Mover loan with *Bank A* may vary depending on the residential addresses of the customers based in the cities, commuting or rural areas of Ireland.

5.2 Design/Experimentation, Evaluation & Results

To design and carry out the experiment two analytical base table were built, that included the features that deemed to be significant in the context of the task set to build and optimize binary classification Mortgage Mover models. Experiment was designed and implemented in two phase approach.

- As part of the Phase 1 of the experiment three classification models Random Forest, Gradient Boosting and Ensemble Gradient Boosting were trained and evaluated with no spatial features included.
- For the second phase of the the experiment spatial features based on the customers distance to the nearest branch locations and GSD locations customers could be assigned where added to the second Analytical Base Model.

As a result of the experiment six models were evaluated, and the three models that included spatial features showed significance improvement in performance in comparison to non-spatial models.

5.3 Contributions and impact

The major contribution of this research is that, as far as the author is aware, it proposes novel approach to binary classification models, by calculating spatial features in GIS and integrating them in the context of home mortgage mover binary classification models, that aims to predict what customers will likely decide to apply for the mortgage mover loan with the *Bank A*. As a result of models evaluation the spatially conscious models have outperformed non-spatial models that indicates that physical presence of branch locations may still be a defining factor in customers decisions to apply for the Home Mortgage Mover loan with the *Bank A* in the rural areas of Ireland despite the digitization development and changing banking habits of the customers.

5.4 Future Work & recommendations

As part of future research and recommendations, future work could be carried out to apply the proposed approach on different customer bases and use different evaluation techniques to evaluate the model performance. The degree of class imbalance may play significant role on the model performance, as well as population size and selected parameters of the models.

Future research can be carried out to include 'commuter' counties of the republic of Ireland excluded in this work. For home mortgage first time buyers other models can be considered that include distance features based on the customers most frequent transaction locations, taking into account that registered in the system residential addresses of these customers may not be accurate. For example customer may register their parents address with the financial institution.

GSD feature excluded from the models presented in this work can be applied in other location based analysis for financial institution's network optimization and cost reduction strategy.

The effect of Covid 19 can be researched on customers banking behaviour in home loan application following similar modeling approach and taking into account additional factors that are closely monitored by the financial institution in the current environment.

At this stage it would be too early and ambitious to definitely conclude that the spatial features have significant impact on the mortgage mover classification models, however the results achieved indicate that taken approach may lead to the new directions of the future research.

References

Abiodun, O. I., Jantan, A., Omolara, A. E., Dada, K. V., Mohamed, N. A. E., & Arshad, H. (2018). *State-of-the-art in artificial neural network applications: A survey.* doi: 10.1016/j.heliyon.2018.e00938

Adegoke, V. F., Chen, D., Banissi, S., & Banissi, E. (2017). Predictive Ensemble Modelling - Experimental Comparison of Boosting Implementation Methods. In *Proceedings - uksim-amss 11th european modelling symposium on computer modelling and simulation, ems 2017* (pp. 11–16). Institute of Electrical and Electronics Engineers Inc. doi: 10.1109/EMS.2017.13

Agostinelli, C. (2002). Robust stepwise regression. *Journal of Applied Statistics*, *29*(6), 825–840. doi: 10.1080/02664760220136168

Ali, A., Shamsuddin, S. M., & Ralescu, A. L. (2015). Classification with class imbalance problem: A review. *International Journal of Advances in Soft Computing and its Applications*, *7*(3), 176–204.

Atta, S., Sinha Mahapatra, P. R., & Mukhopadhyay, A. (2018). Solving maximal covering location problem using genetic algorithm with local refinement. *Soft Computing*, *22*(12). doi: 10.1007/s00500-017-2598-3

Bühlmann, P., & Hothorn, T. (2007, 11). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477–505. doi: 10.1214/07-STS242

Cai, X., Qian, Y., Bai, Q., & Liu, W. (2020). Exploration on the financing risks of enterprise supply chain using Back Propagation neural network. *Journal of Computational and Applied Mathematics*. doi: 10.1016/j.cam.2019.112457

Celik, A. E., & Karatepe, Y. (2007). Evaluating and forecasting banking crises through neural network models: An application for Turkish banking sector. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2006.07.005

Chairi, I., Alaoui, S., & Lyhyaoui, A. (2012). Learning from imbalanced data using methods of sample selection. In *Proceedings of 2012 international conference on multimedia computing and systems, icmcs 2012* (pp. 254–257). doi: 10.1109/ICMCS .2012.6320291

Chawla, N. V. (2009). Data Mining for Imbalanced Datasets: An Overview. In *Data mining and knowledge discovery handbook.* doi: 10.1007/978-0-387-09823-4{_}45

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*. doi: 10.1613/jair.953

Chen, T. H., & Chen, C. W. (2010, 3). Application of data mining to the spatial heterogeneity of foreclosed mortgages. *Expert Systems with Applications*, *37*(2), 993–997. doi: 10.1016/j.eswa.2009.05.076

Cheng, E. W., Li, H., & Yu, L. (2007). A GIS approach to shopping mall location selection. *Building and Environment*, *42*(2). doi: 10.1016/j.buildenv.2005.10.010

Davidson, C., Drury, E., Lopez, A., Elmore, R., & Margolis, R. (2014, 7). Modeling photovoltaic diffusion: An analysis of geospatial datasets. *Environmental Research Letters*, *9*(7). doi: 10.1088/1748-9326/9/7/074009

Delen, D., Walker, G., & Kadam, A. (2005, 6). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine*, *34*(2), 113–127. doi: 10.1016/j.artmed.2004.07.002

do Nascimento, H. A., & Eades, P. (2008, 2). User Hints for map labeling. *Journal of Visual Languages and Computing*, *19*(1), 39–74. doi: 10.1016/j.jvlc.2006.03.004

Droj, L., & Droj, G. (2015). Usage of Location Analysis Software in the Evaluation of Commercial Real Estate Properties. *Procedia Economics and Finance*, *32*, 826–832. doi: 10.1016/s2212-5671(15)01525-7

Farooqi, R., Iqbal, N., & Rashid Farooqi, M. (2017). Effectiveness of Data mining in Banking Industry: An empirical study. *Article in International Journal of Advanced Computer Research*, *8*(5). Retrieved from `https://www.researchgate.net/publication/329518897`

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2). doi: 10.1016/j.ejor.2010.09.029

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). *Knowledge Discovery in Databases: An Overview* (Tech. Rep.).

Freund, Y., & Schapire, R. E. (1996). *Experiments with a New Boosting Algorithm DRAFT-PLEASE DO NOT DISTRIBUTE* (Tech. Rep.). Retrieved from `http://www.research.att.com/orgs/ssr/people/fyoav,schapireg/`

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine* (Vol. 29; Tech. Rep. No. 5). Retrieved from `http://www.jstor.orgURL:http://www.jstor.org/stable/2699986`

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Lecture notes in computer science* (Vol. 3644).

Haykin, S. (1999). *Neural networks: a comprehensive foundation by Simon Haykin* (Vol. 13) (No. 4). doi: 10.1017/S0269888998214044

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9). doi: 10.1109/TKDE.2008.239

He, H., & Ghodsi, A. (2010). Rare class classification by support vector machine. In *Proceedings - international conference on pattern recognition* (pp. 548–551). doi: 10.1109/ICPR.2010.139

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2019.01.012

Heo, H., Park, H., Kim, N., & Lee, J. (2009). Prediction of credit delinquents using locally transductive multi-layer perceptron. *Neurocomputing*, *73*(1-3). doi: 10.1016/j.neucom.2009.02.025

Huang, Y. P., & Yen, M. F. (2019). A new perspective of performance comparison among machine learning algorithms for financial distress prediction. *Applied Soft Computing Journal*. doi: 10.1016/j.asoc.2019.105663

Ilyas, S., Zia, S., un Nisa, Z., Campus, G., Umair Muneer Butt, P., & Pakistan Sukumar Letchmunan, G. (2020). *Predicting the Future Transaction from Large and Imbalanced Banking Dataset* (Vol. 11; Tech. Rep. No. 1). Retrieved from `www.ijacsa.thesai.org`

Jankowski, P., Fraley, G., & Pebesma, E. (2014). An exploratory approach to spatial decision support. *Computers, Environment and Urban Systems*, *45*, 101–113. doi: 10.1016/j.compenvurbsys.2014.02.008

J Neter, C. N. W. W., MH Kutner. (1996). Applied Linear Statistical Models. Fourth Edition. *Journal of Education*, *36*(3). doi: 10.1177/002205749203600311

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2017, 11). Machine Learning for the Geosciences: Challenges and Opportunities. Retrieved from `http://arxiv.org/abs/1711.04708`

Kiely, T. J., & Bastian, N. D. (2019, 2). The Spatially-Conscious Machine Learning Model. Retrieved from `http://arxiv.org/abs/1902.00562`

Kraak, M. J. (2003). Geovisualization illustrated. In *Isprs journal of photogrammetry and remote sensing* (Vol. 57, pp. 390–399). Elsevier. doi: 10.1016/S0924-2716(02)00167-3

Ładyżyński, P., Żbikowski, K., & Gawrysiak, P. (2019). Direct marketing campaigns in retail banking with the use of deep learning and random forests. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2019.05.020

Levin, N., & Zahavi, J. (2001). Predictive modelling using segmentation. *Journal of Interactive Marketing*, *15*(2), 2–22. doi: 10.1002/dir.1007

Liu, X. Y., Wu, J., & Zhou, Z. H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, *39*(2). doi: 10.1109/TSMCB.2008.2007853

Martin, A., Miranda Lakshmi, T., & Prasanna Venkatesan, V. (2014). An information delivery model for banking business. *International Journal of Information Management*. doi: 10.1016/j.ijinfomgt.2013.12.003

Molina Utrilla, J., & Constantinou, N. (2011). *Could the trigger to the subprime crisis have been predicted? A mortgage risk modeling approach* (Tech. Rep.). Retrieved from `http://ssrn.com/abstract=1616697`

Moro, S., Cortez, P., & Rita, P. (2014). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, *62*, 22–31. doi: 10.1016/j.dss.2014.03.001

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004, 6). *An introduction to decision tree modeling* (Vol. 18) (No. 6). doi: 10.1002/cem.873

Niankara, I. (2019). Panel and geospatial data for U.S. FDIC insured banks fiduciary activities and annual performance analyses over the periods 2016 to 2018. *Data in Brief*. doi: 10.1016/j.dib.2019.104358

Pan, Y., & Tang, Z. (2014). Ensemble methods in bank direct marketing. In *11th international conference on service systems and service management, icsssm 2014 - proceeding.* IEEE Computer Society. doi: 10.1109/ICSSSM.2014.6874056

Petukhov, A., Zaikin, O., & Bochenina, K. (2019). Analysis of the geospatial activity profiles of bank customers. In *Procedia computer science* (Vol. 156, pp. 245–254). Elsevier B.V. doi: 10.1016/j.procs.2019.08.200

Pradhan, B. (2013). A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Computers and Geosciences*, *51*. doi: 10.1016/j.cageo.2012.08.023

Quinlan, J. R. (1986). *Induction of Decision Trees* (Vol. 1; Tech. Rep.).

Roig-Tierno, N., Baviera-Puig, A., & Buitrago-Vera, J. (2013, 9). Business opportunities analysis using GIS: the retail distribution sector. *Global Business Perspectives*, *1*(3), 226–238. doi: 10.1007/s40196-013-0015-6

SAS/STAT 9.1 user's guide. (2004). [Verlag nicht ermittelbar].

Schapire, R. E. (2002). *The Boosting Approach to Machine Learning An Overview* (Tech. Rep.). Retrieved from `www.research.att.com/`

Scheurmann, E., & Matthews, C. (2005). Neural network classifers in arrears management. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 3697 LNCS).

Sharma, D., Alford, B. L., Bhuian, S. N., & Pelton, L. E. (2009). A higher-order model of risk propensity. *Journal of Business Research*. doi: 10.1016/j.jbusres.2008.06.005

Silverman, N., & Suchard, M. (2013). PREDICTING HORSE RACE WINNERS THROUGH A REGULARIZED CONDITIONAL LOGISTIC REGRESSION WITH FRAILTY. *Journal of Prediction Markets*, *7*(1), 43–52. Retrieved from `https://EconPapers.repec.org/RePEc:buc:jpredm:v:7:y:2013:i:1:p:43-52`

Simionescu, M. (2015). Predicting the National Unemployment Rate in Romania Using a Spatial Auto-regressive Model that Includes Random Effects. *Procedia Economics and Finance*. doi: 10.1016/s2212-5671(15)00281-6

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4). doi: 10.1142/S0218001409007326

Tanaka, K., Kinkyo, T., & Hamori, S. (2016). Random forests-based early warning system for bank failures. *Economics Letters*, *148*. doi: 10.1016/j.econlet.2016.09.024

Tavana, M., Abtahi, A. R., Di Caprio, D., & Poortarigh, M. (2018). An Artificial Neural Network and Bayesian Network model for liquidity risk assessment in banking. *Neurocomputing*. doi: 10.1016/j.neucom.2017.11.034

Tkáč, M., & Verner, R. (2016). *Artificial neural networks in business: Two decades of research.* doi: 10.1016/j.asoc.2015.09.040

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment* (Tech. Rep.). Retrieved from `www.aaai.org`

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11). doi: 10.1145/1968.1972

Vélez, D., Ayuso, A., Perales-González, C., & Rodríguez, J. T. (2020, 5). Churn and Net Promoter Score forecasting for business decision-making through a new stepwise regression methodology. *Knowledge-Based Systems*, *196*. doi: 10.1016/j.knosys.2020.105762

Wang, S., & Yuan, H. (2014, 10). Spatial data mining: A perspective of big data. *International Journal of Data Warehousing and Mining*, *10*(4), 50–70. doi: 10.4018/ijdwm.2014100103

Westreich, D., Lessler, J., & Funk, M. J. (2010). *Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression* (Vol. 63) (No. 8). Elsevier USA. doi: 10.1016/j.jclinepi.2009.11.020

Yee, H. J., Ting, C. Y., & Ho, C. C. (2018, 9). Optimal geospatial features for sales analytics. In *Aip conference proceedings* (Vol. 2016). American Institute of Physics Inc. doi: 10.1063/1.5055554

Zhou, G., Li, Q., Deng, G., Yue, T., & Zhou, X. (2018, 4). Mining co-location patterns with clustering items from Spatial data sets. In *International archives of the photogrammetry, remote sensing and spatial information sciences - isprs archives* (Vol. 42, pp. 2505–2509). International Society for Photogrammetry and Remote Sensing. doi: 10.5194/isprs-archives-XLII-3-2505-2018