ABSTRACT

Credit risk analysis is such an important part of any economy, as many economies have large amounts of existing debts from its citizens. The ability to properly manage these debts is invaluable in preventing loses from credit granting bodies such as banks, investment firms which can cripple an economy like the events of the last financial crisis. One of the many means of managing credit is credit scoring. Credit Scoring is the analytical process of computing the probability that a customer defaults on a line of credit or doesn't, with the digitization of data, this has been approached as a binary classification paradigm with capable machine learning algorithms.

However, the low default portfolio problem is still a paramount area of research in the credit scoring field. It refers to an imbalance that exists in credit scoring datasets where the number of defaulters is miniscule when compared to the number of non-defaulters, causing machine learning algorithms to perform unfavourably in their ability to optimally classify instances of both classes. In order to solve this problem a review was conducted on the subject matter, existing trends, paths of research and solutions where identified. Current solutions posited solve the problem but at a high complexity (black box) which limits their acceptance into the business industry.

This dissertation proposed an ensemble that solves this problem using a committee of learners' namely C4.5 Decision Tree, Support Vector Machine, K-Nearest Neighbours and Logistic Regression. The machine learning algorithms are combined in a stacking configuration, which is a parallel structure of algorithms with an algorithm as a meta learner. The ensemble was designed and tested using WEKA: a data mining environment for knowledge analysis, and two datasets from the UCI Machine Learning Repository. Its performance was evaluated using Accuracy, Area under the Receiver Operating Characteristic Curve, Area under the Precision Recall Curve and Mathew's Correlation Coefficient.

New uses for credit scores such as employment interviews and apartment renting have made it even more important to have a system that optimally computes customer's credit scores. The results obtained showed that the ensemble was able to sufficiently classify the minority class at a high accuracy when the dataset becomes skewed and makes a good trade-off between practicality and complexity. This ensemble can be further implemented as part of a credit score card and used by credit granting or scoring bodies.

Keywords: Credit Scoring, Low Default Portfolio, Machine Learning, Area under the Receiver Operating Characteristic Curve

Word Count: 391

CHAPTER ONE

INTRODUCTION

1.1 Background to the Study

Credit is an agreement based on trust, where a lender offers goods, services or something of value to receive repayment, usually with interest at a later agreed date (Investopedia, 2018). The credit industry is an important part of the economic ecosystem of developed countries with large economies such as the US (United States) and China. Credit may be in form of bank loans, credit cards, student loans, mortgages, cars loans. In the third quarter of 2016, the amount of consumer credit alone in the US rose to 1.2 trillion dollars (Carroll & Rehmani, 2016). The ability to properly predict and asses credit risk is one of the most profitable and difficult ventures. Assessing credit risk is such that a one percent increase in the predictive power of risk assessment will lead to a massive decrease in losses to credit granting bodies (Henley & Hand, 1997). To this end, credit ratings, credit scores and similar metrics for evaluating the risk a borrower may incur are developed typically by Credit Reporting Agencies (CRA)s such as FICO (Fair Isaac Corporation) and Moody's for credit granting bodies.

Credit scoring simply refers to classification of customers into bad credit and good credit with bad credit being, customers who are likely to default on the repayments of their credit within the stipulated period and good credit being, the customers who are likely to faithfully repay their credit within the allotted time. This kind of prediction was achieved historically by looking at a number of attributes grouped under the 5c's i.e. the character, capital, collateral, capacity, and conditions of the customer (Thomas, 2000). CRAs (Credit Reporting Agencies), apply statistical models such as Linear Discriminant Analysis (LDA), Multivariate Discriminant Analysis (MDA), Logistic Regression (LR), and expert knowledge in predicting credit risk. The Credit score generated for each customer affects how much is given to the customer in terms of credit risk and also how much interest rates will be charged. Credit granting bodies hardly reject customers who have scores, however low, these customers are just given steeper deals on their credit. If the credit scoring process is not objective and adequately accurate, then customers may be treated unfairly by credit granting bodies.

The size of digitized data in the credit industry and nature of the credit scoring problem: which is classifying the customer into good or bad credit classes, makes it a prime endeavour where machine learning techniques can be applied. Machine learning is a subfield of computer science that deals with developing programs that are able to learn from data instances and experience

the same way intelligent biological systems do, these programs are able to discover anomalies and learn patterns and features from data, that enable them to solve lots of problems in the digital age.

As relating to credit scoring, underlying features in credit datasets can be discovered, then used to classify customers into good and bad credit thus making predictions on the credit score of new customers. Since the early 2000s, non-parametric credit rating models have been developed based on machine learning algorithms such as Decision Trees (DT), K Nearest Neighbour (KNN), Artificial Neural Networks (ANN), Genetic Algorithms (GA), Support Vector Machine (SVM) and Naïve Bayes (NB) to mention a few. These algorithms have achieved a better level of accuracy when compared to statistical models used by Credit Rating Agencies.

In the last decade, hybrid models consisting of two or more diverse machine learning algorithms Chen, Ribeiro and Chen (2015) have been applied to the problem of credit scoring in a bid to increase the ability the properly classify customer's credit. Soft computing methods have also been able applied to the credit scoring problem as another approach Lin, Shih-Wen , and Tsai, (2017), leveraging the ability of such methods to tolerate ambiguity, increase interpretability and fault tolerance. More recently, feature selection (FS) algorithms such as Genetic Algorithms (GA), Particle Swarm Optimization (PSO), Rough Set Theory (RST), Principal Component Analysis (PCA), have also been proposed as a solution that can be in cooperated into existing systems for generating credit scores to customers. Hybrid ensembles and Ensemble methods/learning also known as committee of learners, where a group of learners are brought together by a specific method have been recently proposed as a means of addressing the credit scoring problem.

The motivation for this study stems from this approach, as the dissertation proposes an ensemble framework consisting of genetic algorithms for feature selection and an ensemble of 4 (four)algorithms namely; K-NN, SVM, LR, DT.

1.2 Problem Statement

The international nature of credit markets, increased the need for objective and accurate credit scoring and the CRAs which are responsible for assigning credit ratings to bodies and individuals alike have become indispensable in the current global markets. CRAs mostly employ statistical models such as Linear Discriminant Analysis (LDA), Multivariate Discriminant Analysis (MDA), Logistic Regression (LR), and human expert knowledge, but

these models make strong assumptions on the data, are time consuming, and are bias due to the subjective nature of human expert knowledge, this affects the accuracy of such systems. Since the 2008 financial crisis, the measurement and correct classification of credit risk has become more important than ever. The CRAs have come under intense scrutiny for inaccuracies in credit scores and ratings which led to many companies incurring bad risk.

Alternative methods have been broached since then to improve the accuracy and predictive power of credit scoring systems. Amongst these methods machine learning has been showing promising results with up to 80% accuracy on real-life data sets (Ozturk, Namli, & Erdal, 2016). Despite this progress, the machine learning solutions suffer from interpretability problems and could also be made to be more accurate, efficient with computational resources, robust and scalable. Ensembling which is a "strength in numbers approach" has emerged as a promising means of improving upon the successes of single classifier machine learning algorithms, many Ensembling models have already been proposed by researchers with promising results in tackling the existing problem.

Whilst accuracy and efficiency are intuitively understood problems, interpretability, robustness and scalability are more composite problems. For scalability and robustness, in this study is related to the Low Default Portfolio problem. The Low Default Portfolio problem in credit scoring is a severe case of the class imbalance problem (Kennedy, 2013). Class imbalance is a phenomenon that occurs in machine learning where a data set has an under sampled class. Standard classification algorithms assume that the classes existent in any data set are evenly distributed, Japkowicz (2000), as such when an imbalance exists, predictions made by these algorithms tend to fit the most common class (Drummond & Holte, 2005). Even if such algorithms have a good level of accuracy, they are not useful for real world applications, where imbalanced data sets exist.

A few studies in literature have attempted to resolve the Low Default Portfolio problem but that has been done at the expense of interpretability, efficiency and accuracy. Hence, this work aims at proposing a framework that can successfully tackle the low default portfolio problem, without an excessive sacrifice of accuracy, interpretability and practicality.

1.3 Objective of the Study

The aim of this study is to design an ensemble of machine learning techniques for credit scoring. The specific objectives are to:

I. Conduct an extensive review of literature and existing solutions in the field of endeavour to show mastery of the subject matter.
II. Obtain and pre-process credit data sets.
III. Design a framework for the proposed ensemble.
IV. Evaluate the framework developed in III.

1.4 Methodology Overview

The following procedures will be followed in order to achieve the above stated objectives.

I. Existing literature is reviewed, key concepts, milestone and closely related work are discussed to show a good understanding of the credit scoring and machine learning niche.
II. Publicly available data sets are obtained from the UCI Machine Learning Repository. The data sets are the German and Australian credit data sets, they are then pre-processed by: normalization of their attributes, analysing missing attributes values, performing preliminary data exploration and outlier detection for the purpose of understanding trends and optimizing the data set for analysis.
III. The proposed framework is developed using the stacking ensemble method, which combines a committee of learners namely; K-NN, C4.8 Decision Tree, and SVM as base learners in a parallel structure and Logistic Regression as the meta classifier responsible for processing the predictions of the base classifiers/learners. The ensemble is trained and tested using WEKA (Waikato Environment for Knowledge Analysis) which is a popular, free machine learning software developed at the university of Waikato and written in Java. The software is ideal for training and testing various machine learning algorithms.
IV. The framework will be evaluated against the single classifiers such as K-NN, MLP, SVM, Logistic Regression, ID3 Decision Tree which are the most frequently used single classifiers in literature, state of the art solutions like ensembles and hybrid feature selection methods which represent the current state of research in the credit scoring domain, using metrics such as Confusion Matrix, Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) Curve, Area under the Curve of the Precision Recall Curve (AUCPRC) and Mathew's Correlation Coefficient (MCC).

1.5 Significance of the Study

Credit scoring remains a paramount problem in the credit industry despite solutions that have been put forward, to prevent a reoccurrence of the events of the last financial crisis, more practical and efficient methods need to be proposed, methods that can be used by the lenders and borrowers to better understand the complex relationships that exist between the data used to develop credit scoring systems.

The acceptance of credit scoring in developed countries as a valid and objective means of evaluating a person's credit worthiness has led to some interesting uses for credit scores. For example, in the US credit scores have been used when negotiating house rents and also in the evaluation processes for certain kinds of jobs. Locally, credit score cards are not in use as pervasively in the Nigerian economy but a few Financial Technology (FinTech) companies such as Lendo have begun developing credit scores combining both application and behavioural scoring. This dissertation proposes a framework that maintains a high level of accuracy and computational efficiency without excessively compromising the interpretability and robustness and such can be implemented as part of credit score cards when credit scores are fully accepted by Nigerian financial institutions.

1.6 Scope of Study

This study is limited to application rating/classification of consumer credit using data from two commonly used and publicly available data sets. It will also be limited to the use of the following algorithms Decision Tree (DT), K-NN (K Nearest Neighbour), SVM (Support Vector Machines), Logistic Regression (LR).

1.7 Organization of dissertation

This study is organized as follows:

Chapter 1: This is the introductory part of the study, which includes the problem statement, aim and specific objectives of the dissertation, scope of the study amongst others.

Chapter 2: This presents an extensive review of literature, broad and closely related to the field of study.

Chapter 3: The framework for the proposed ensemble of machine learning algorithms is shown in this chapter, and analysis of previously developed systems.

Chapter 4: This chapter presents findings, and evaluation the developed system.

Chapter 5: The last chapter of this research will include a summary of the entire dissertation, along with recommendations for future studies.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter presents a synthesized review of literature academic and otherwise on the subject in focus. Section 2.2 talks about the fundamentals of credit scoring, types and categories. Section 2.3 visits statistical models that have been used to generate credit scores. Section 2.4 deals with machine learning, its fundamentals and how it has been used to approach the problem. Section 2.5 looks at soft computing methods, hybrids and ensembles that have been used for scoring credit of customers. Section 2.6 is a review of closely related work on credit scoring. Section 2.7 gives a synthesis or summary of the literature review conducted.

2.2 Credit Scoring

Credit as will be used in this dissertation refers to a contractual agreement between a borrower and a lender, that enables the borrower to access goods and services with the knowledge that the money required to buy said goods and services will be paid at a later date usually with interest. According to the Business (Dictionary), Credit also means the purchasing power created by financial institutions to customers usually through lines of credit such as credit card s. Credit scoring and Credit rating are used interchangeably in some literature, but there is a slight difference between the two terms, while Credit scoring refers to the means by which a credit rating body approved by the government such as the FICO (Fair Isaac Corporation), Shufac, statistically determines a three digit score that represents the creditworthiness of a customer (individual) based on their credit history and other data, Credit rating is a letter grade given to governments, cooperate organizations and businesses that represent their likelihood of defaulting on credit.

A recent review done by Louzada, Ara, and Fernandes defines credit scoring as “a numerical expression based on a level analysis of customer credit worthiness, a helpful tool for assessment and prevention of default risk, an important method in credit risk evaluation, and an active research area in financial risk management.” (Louzada, Ara, & Fernandes, 2016, p. 1) Historically, underwriting credit was done by cynical bank managers who accessed an individual’s creditworthiness by subjectively, Hand & Henley, (1997) taking into consideration the 5Cs (character, capital, collateral, capacity, and conditions of the customer) (Thomas, 2000). Customer credit scoring came into prominence when the RCC (Atlanta’s Retail Credit

Company) in 1899 were one of the first bodies to gather information about customers in a bid to formalize the credit scoring process. The information gathered on millions of Americans contained data on their credit history, capital, character, social, political and sexual lives (Trainor, 2015). Financial risk assessment in the form of customer credit risk was first attempted by Fitzpatrick (FitzPatrick, 1932). In the 1960's RCC revealed plans to digitized its records, this led to an outcry, many citing that digital evidence of such information was an infringement on civil liberties. In 1970, a milestone act was passed that made it mandatory for credit bureaus as they were now called to destroy all data on race, sexuality, disability and make their files publicly accessible. This was called the Fair Credit Reporting Act. The RCC after suffering public humiliation changed its name to Equifax and was soon joined by Experian and TransUnion, these three became the main credit reporting bureaus. These bureaus generated reports which were sometimes difficult to make sense of and were still analysed manually and subjectively, the sheer number of applicants at this point also necessitated a different approach to determining credit worthiness. Consequently, the FICO score introduced in 1989 became the industry standard to determine credit worthiness in the US (FICO, 2017). The FICO score is a three-digit number ranging from 300-850, which is derived by analysing statistically the credit report of a customer. The model for analysis is called a score card which is typically a statistical model, such as LDA (Linear Discriminant Analysis), LR (Logistic Regression), PR (Probabilistic Regression)

Credit Scoring structure differs from country to country, In Australia, Equifax which acquired Veda Advantage in February (2016), is the biggest provider of credit scores. Credit scores in Australia are not only used to decide whether to offer credit or not, but also used to determine limits of credit that is available on credit cards. In Germany, Schufa is the largest provider of credit scores which are used similarly as Australia's Equifax. The UK (United Kingdom) has a peculiar way of credit scoring as there is no universal credit score in the UK. Different lenders use their unique score cards to access credit risk of applicants and are not mandated to provide reasons for declining credit to customers. Experian and Equifax exist in the UK but their scores are not employed by lenders. FICO scores are still the most commonly used in the US today (Credit Score, 2018). A person's creditworthiness is computed and ranged between 300-850, the higher the score the lower the probability of default. According to Credit Sesame, on average, lending bodies consider that 750 and above (are excellent), 700-749 (are good), 650-699 (are fair), 550-649 (are poor), 550 and below (are bad). Customers with good scores have access to the best interest rates and payment plans, conversely costumers with bad score get

high interests and large down payments or are unable to access credit at all. FICO is tight lipped about how their score card works, but research has discovered that it considers the following factors (Credit Score, 2018):

I. Payment history of loans and credit cards, the number of on-time payments, number and severity of late payments (Experian, 2018). This has a 35% effect on the score.
II. Total amount owed, the amount of debts across all the accounts or loans. This accounts for 30% of the score, which represents how much of the available credit is being used.
III. Length of credit history, a good long history of good credit is bound to affect a score by 15%. Customers with long history of credit are considered to have lower probability of default as they have more data from analysis
IV. Types of credit, good mix of credit such as mortgage loans, revolving credit and car loans account for 10% of the score.
V. New credit, the number of credit lines opened and the recency of these accounts affects a customer's score by 10%, ultimately too many lines of credit can negatively affect one's score.

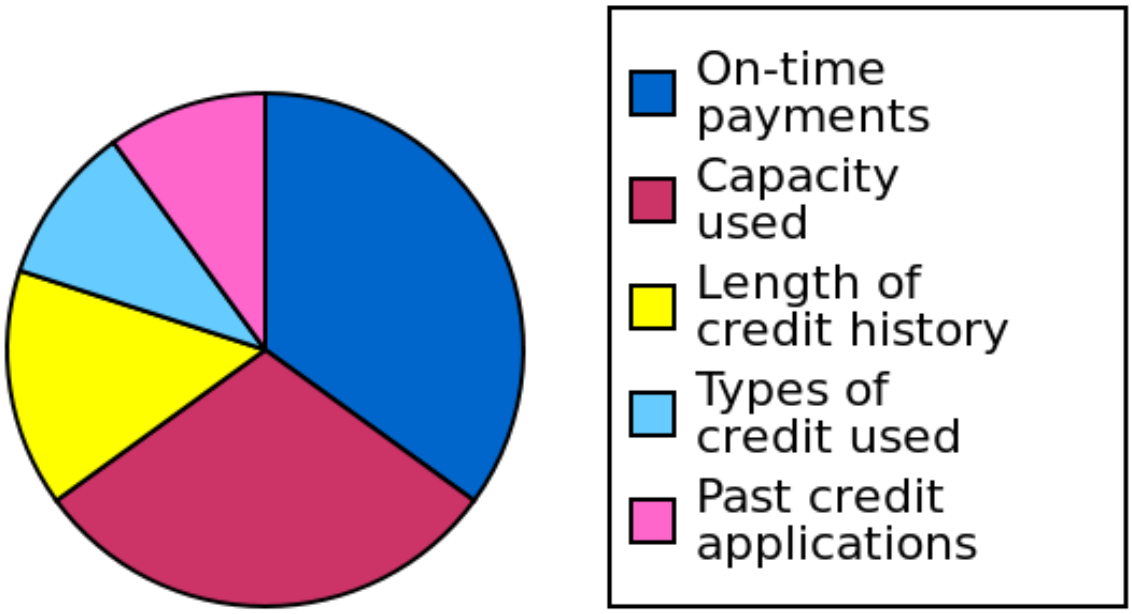

Figure 2.1 Chart of Relationship between Score factors. (Source myFICO, 2018)

2.2.1 Types of Credit

There are generally three types of credit according to (McBride, 2018):

I. Non-instalment Credit: This is a type of credit that requires principal borrowed to be paid back in full within a short period of time after it was borrowed, it does not allow for monthly payments of a certain figure over an extended period of time.

II. Instalment Closed-End Credit: A good example of this type of credit is a car loan, the sales price for the car is paid for in instalments over an agreed period of time. In essence, closed ended instalment requires that a service is paid for with a regimented reduction of a stipulated amount from the buyer's income over a period of time.

III. Revolving Open-End Credit: This is the kind of credit you find with credit cards. A customer will be allocated an amount of principal or credit which he/she can spend leisurely, the principal is paid back in part at a stipulated time which is typically a month. The customer continues to enjoy credit as long as his/her account is not closed, this attribute makes the credit revolving.

2.2.2 Types of Credit Scoring

I. Application scoring: this is the scoring that takes place when a customer initially applies for any type of credit (Credit Scoring, 2018). It evaluates social, demographic and financial data about the prospective customer to discover the probability of repayment of a given period (Kennedy, 2013). Most literature in credit scoring focuses on this type of scoring because of the availability of data, accessibility and ability for extensive comparative analysis, as opposed to behavioural scoring. This study will focus on application scoring.

II. Behavioural scoring: this attempt to quantify customer behaviour after credit has been granted. It is done in relation to customer portfolio management, since the better you know your customers, the better you can predict whether they will default on their payments over a period of time (Score, 2018). Behavioural scoring requires that socio-economic behaviour of customers be monitored closely so that customers that change behaviour can be flagged for fraud or enable the lender to take actions to minimize risk incurred (Kennedy, 2013). Data used for this type of scoring is called alternative data, as it does not directly affect the credit score of the individual. Alternative data is difficult to come by and getting them can easily lead to privacy and ethical problems.

2.3 Statistical Models

Statistical models were the first models to be applied in building score cards and they are still in use today. The two most commonly used by Credit scoring agencies are discussed here.

2.3.1 Linear Discriminant Analysis

Fisher's Linear Discriminant Analysis (LDA) introduced in (1936) is technique for discriminating between two or more groups of samples. It can also be used for dimensionality reduction, but in the credit scoring field it is used for binary classification. It was first applied to credit risk prediction by Durand in (1941), and then later used in the first score card called the Altman models which generated the Z- score. (Altman, 1968) LDA works by applying a discriminating function on sample data, typically of the form:

$$Z = \propto +\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i \qquad 2.1$$

where Z represents the discriminant value, $\propto$ the intercept, β_i is the linear contribution effect of the ith data instance X_i, i ranges from i= 1,2,..n. It also tries to maximize the separating distance between the two clusters and minimize then distance between data points in each cluster.

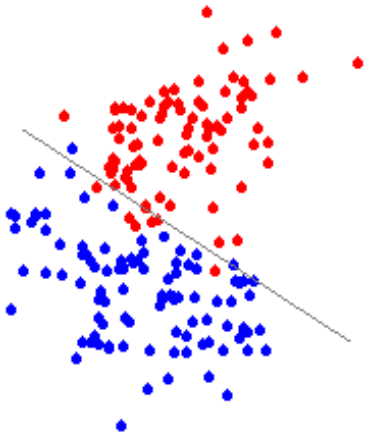

Figure 2.2 Linear Discriminant Analysis. (Source: Fundamentals of Statistics,2012)

This is called the score function and is given by:

$$M = \frac{distance\ between\ sample\ means\ of\ two\ groups}{(sample\ variance\ of\ each\ group)^{1/2}} \qquad 2.2$$

It is a good method, due to its simplicity and can model data with similar accuracy of complex methods. A good way of evaluating the effectiveness of the separation is by calculating the Mahalanobis distance (Δ) between the two groups created by the LDA as noted by Bhatia, Sharma, Burman, Hazari, and Hande (2017), a distance of more than three (3) represents a

healthy separability between the two groups.

$$\Delta^2 = \beta^{\tau}(\mu_1 - \mu_2) \quad 2.3$$

Where,$\beta^{\tau}$is the Coefficients Vector, and $(\mu_1 - \mu_2)$is the difference between the two means. LDA makes the following assumptions on data (Louzada, Ara, & Fernandes, 2016):

1. The covariance matrices of each classification subset are equal.
2. Both groups are follow a multivariate normal distribution.

2.3.2 Logistic Regression

Logistic Regression is means of binary classification where the probability of a data instance belonging to one of the classes is determined the logistic transform of a linear combination of variables. It was proposed by Berkson (1944) and remains arguably the most commonly used algorithm in the credit scoring field, individually or as part of some kind of hybrid (Hand & Zhou, 2009). LR works checking the probability of some event y (0/1) happens based on some feature x. It also known as Binomial Logistic Regression, Logit Regression or Logit Model. Logit Regression uses the odds ratios concept to calculate the probability. It can be defined formally as:

$$Odds = \frac{P(y=1|x)}{1-P(y=1|x)} \quad 2.4$$

A natural logarithm of the odds ratio is taken to create the logistic curve. The new equation becomes:

$$Logit\big(P(x)\big) = \ln\left(\frac{P(y=1|x)}{1-P(y=1|x)}\right) \quad 2.5$$

The logit of probability is also linear with respect to x, thus;

$$Logit\big(P(x)\big) = a + bx \quad 2.6$$

In order to find the probability of a single data instance, we combine the equations 2.5 and 2.6 and remove the logarithmic function. This produces:

$$\frac{P(y=1|x)}{1-P(y=1|x)} = e^{a+bx} \quad 2.7$$

And finally,

$$P(y=1|x) = \frac{e^{a+bx}}{1+e^{a+bx}} \quad 2.8$$

This equation gives logistic regression its ‘S’ like curve when it models data, where ‘a’ and ‘b’ represent the gradients for the function. Logistic Regression was chosen for this study because of its high interpretability and high accuracy in modelling credit scoring data.

2.4 Machine Learning

Machine learning is the ability of a computer program to learn features and make predictions on previously observed data instances. In the last 20 years machine learning has come leaps and bounds due to high rates of digitalization, development of new powerful hardware and software components for computing systems, the internet and more distributed and robust network infrastructure. Machine learning is classically defined by Tom Mitchel in his book as “A computer program is said to learn from experience E with respect to some class of tasks T and a performance measure P, if its performance at tasks T, as measured by P, improves with experience E”. (Mitchell, 1997, p. 2)

2.4.1 Types of Learning

There are generally four paradigms of machine learning paradigms.

I. Supervised Learning: as the name implies supervised learning deals with a type of learning where there is a teacher or supervisor which corrects the learning algorithm or adjusts its output error until it fits the desired or target output. The algorithm develops a model that bests fits the relationship between input and output, Krenker, Bešter, and Kos (2011) for classification problems such as credit scoring. Supervised learning uses error as the feedback in a closed-loop feedback system, and the learning only terminates if the error (e) has become adequately small (Du & Swamy, 2014).
II. Unsupervised learning: this is a type of learning with only input data (Kennedy, 2013). The algorithm is typically left to explore the data in other to find important features for clustering problems and to find anomalies for novelty detection problems.
III. Semi-Supervised learning: this is a flavour of supervised learning where not all of the input has target values or labels.
IV. Reinforcement learning: is involved with classic AI (Artificial Intelligence) and robotics, where a learning agent uses its actuators to interact with its environment with the aim of developing a policy that is the most cost-efficient solution to a specific problem the environment poses. The learning agent is assisted with ‘reward’ for good decisions and ‘punishment’ for bad decisions.

2.4.2 Classic Machine Learning problems

Most practical problems that can be solved with machine learning techniques are grouped under the following:

I. Binary Classification: this is arguably the most studied and frequently referenced problem in machine learning according to Smola & Vishwanathan (2008), credit scoring can also be defined as a binary definition problem. Put simply binary classification is problem that involves separating data instances into two classes, either a member of a class or not a member of the class.

II. Multiclass Classification: this type of classification is an extension of binary classification, where new instances of data are classified as belonging to members of several classes based on the features of data. For example, in text mining, the aim may be to classify documents into what part of an organization they originate from (Human Resources, Top Management, Operations, Information Technology).

III. Regression: this kind of problems involve predicting some real number value based on some attributes that affect the said value. The attributes are usually given different weights depending on how much they affect the target value. An example is estimating the value for a used car; attributes like brand, year, capacity, mileage, engine and other details will have their effect on the worth of the car (Alpaydın, 2010).

IV. Novelty Detection: also called anomaly discovery, is related with unsupervised learning. It is a problem that requires an algorithm to discover anomalies or outliers in the data space. The definition for what an anomaly is subject to the analyst although, typically anomalies occur scarcely in data search space.

2.4.3 Machine Learning Approaches in Credit Scoring

Machine learning is seen by the research community as good solution to the credit scoring problem, the ability of machine learning algorithms to work with large data sets with high dimensions and make accurate predictions on new data has galvanized researchers are businesses alike to sorting for machine learning solutions. Below the most prominent and frequently single machine learning classifiers are discussed.

i. Artificial Neural Networks (Multi-Layer Perceptron):

The profound ability of artificial neural networks (ANN) to fit models that represent complex relationships between input data and target output have informed their

application in the field of credit scoring (Brown & Mues, 2012). Neural networks are mathematical constructs that are modelled to loosely adopt the structure and information flow of the human brain (Bishop, 1995). In the brain there is an enormous number of interconnections between neurons sending information to each via synapses. Similar to the human brain the single unit of computation is the neuron.

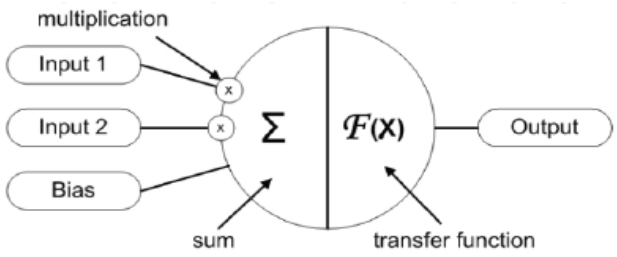

Figure 2.3 An artificial neuron. (Source: Andrej Krenker, Janez Bester and Andrej Kos 2011)

The artificial neuron takes in a weighted input usually with bias and performs simple arithmetic on the input, after which it compares the value to a transfer function to determine whether it propagates the signal to the next layer. ANNs are organized as interconnected neurons in layers.

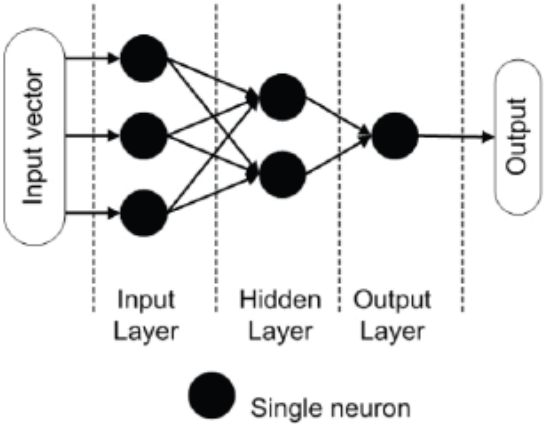

Figure 2.4 Multi-layered Feedforward Network. (Source Introduction to Artificial Neural networks 2011)

First the input layer receives the inputs with their respective weights, perform some computation before relaying the information to the next layer called the hidden layer, which can contain any number of neurons and layer depth. Lastly, the output layer presents the output which could be a regressed value or a binary classification. There exist various types of neural nets, due to the number of ways one can vary a network's depth, information flow and architecture. In credit scoring, the Multi-Layer Perceptron

(MLP) is the type most frequently employed in literature see (Altman, 1994; Baesens, et al., 2003; Blanco, Pino-Mejias, Lara, & Rayo, 2013; Rafiei, 2011; West, 2000) with excellent results. MLP have several inadequacies as they present a black box solution, take lots of time to train, and can easily overfit data, which is why they have not been selected for this study.

ii. K-Nearest Neighbours:

K-Nearest Neighbours (K-NN) is memory-based reasoning algorithm, Brown and Mues (2012) that is used for classification and clustering problems. K-NN is lazy Bronshtein (2017), non-parametric, algorithm that works by representing input data as points in a hyperplane of n dimensions. Where n is the number of features that the input data contain, such that input with 2 features are represented in 2-dimensional space, data with 3 features are represented in 3-dimensional space and so on. Geometric calculations are then performed on the data points as represented in the hyperplane, popularly, Euclidean distance is performed, enabling data to be classified relatively to the input around it. In classifying new input, the K-NN algorithms looks at the K nearest data points to the new input and allows them to vote on the label of the new input, it assumes that data points in the same space must be similar (Daumé, Geometry and nearest nieghbours, 2013). There is no standard value for K, but it is known that overfitting is caused by small values of K and underfitting by large values of K (Zhang, Introduction to machine learning, 2016). The figure below shows an instance of K set to equal 3.

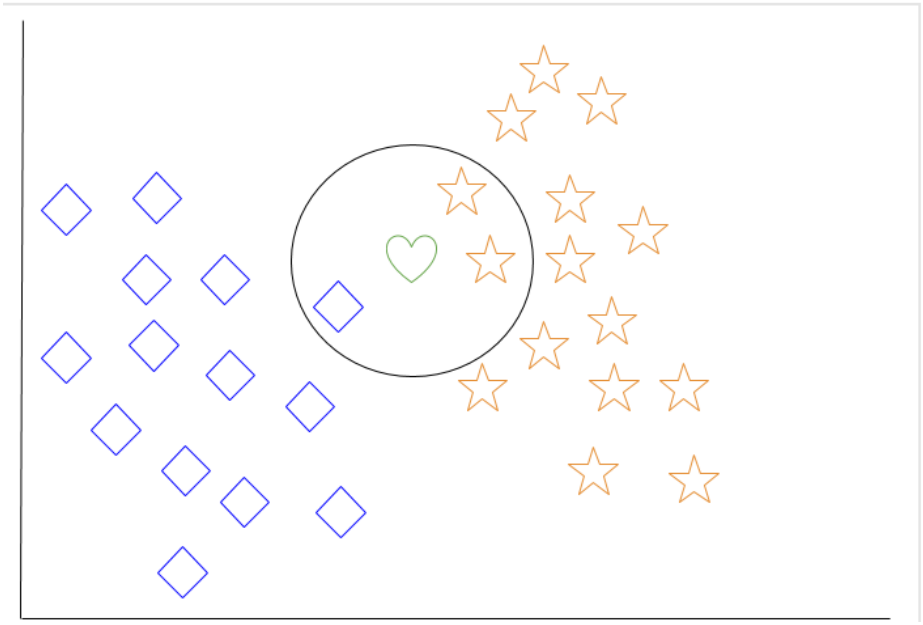

Figure 2.5 KNN, Where K is 3. (Source: Tagliaferri, 2017)

Euclidian distance calculated for the distance between points at different dimensions "D" is given formally as:

$$d(\boldsymbol{a}, \boldsymbol{b}) = [\sum_{d=1}^{D}(a_d - b_d)^2]^{1/2} \quad 2.9$$

Where **a** and **b** are two points in the hyperplane. The K-NN is simple and intuitive yet yields surprisingly accurate results (Daumé, Geometry and nearest nieghbours, 2013). K-NN as a single classifier suffers from the curse of dimensionality and is not memory efficient. It is chosen for this study for its performance, simplicity and interpretability.

iii. Decision Trees:

Decision Trees (DT) use the age-old notion of "divide and conquer" to solve classification and regression problems. It is a classic and natural technique of learning (Daumé, Decision Trees, 2013). DT has been applied in several works in credit scoring, showing good results as a single classifier and as part of an ensemble. The aim of a tree is to predict a target value (continuous or binary) for an input based on what the tree has learnt from previous input streams. It learns by splitting a data set into smaller subsets using features of the data and splitting rules such as Gini Index, Information Gain and Entropy (Quinlan, 1993). A decision tree after it has learnt is a hierarchical model that has decision nodes, terminal nodes according to Alpaydın (2010) and sometimes probability nodes (Decision tree: Introduction, 2015). These nodes show the induction process the tree follows in predicting the target value. The figure below shows a sample tree that predicts/classifies homes in a city (Breslow & Aha, 2000).

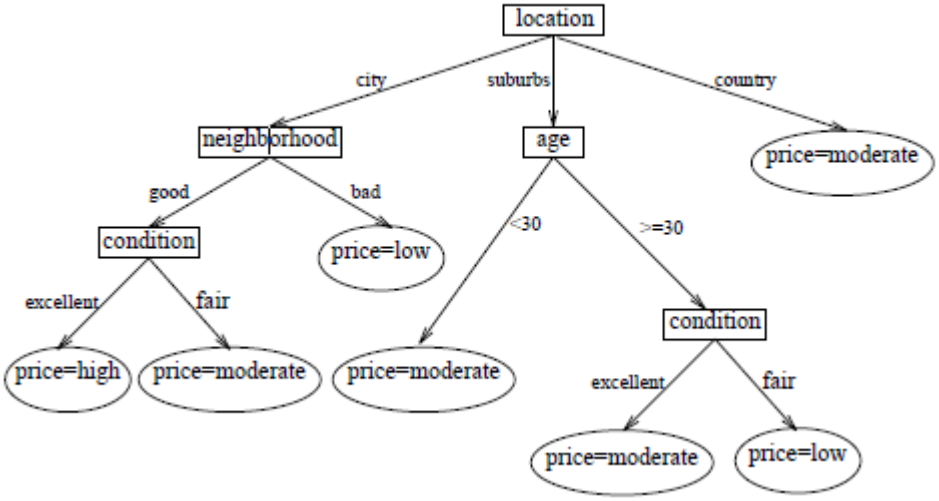

Figure 2.6 A Decision Tree, (Source Breslow & Aha, 2000)

Depth of the trees is important in the learning process as overly large trees can lead to overfitting and small trees can lead cause underfitting. There are different types of trees, some are, ID3, CART, C4.5, C5, J4.8. Decision Trees have been used extensively for credit scoring in various capacities, see (Brezigar-Masten & Masten, 2012; Delen, Kuzey, & Uyar, 2013; Yobas, Crook, & Ross, 2003). It's output an induction can be explained using sequential uf-then statements which makes decision trees arguably the most interpretable of the machine learnning tecnhiques.

iv. Support Vector Machines:

Support Vector Machine (SVM) is a powerful algorithm used for classification and regression. It works by finding the optimal hyperplane that best separates two classes of data present in a data set (binary classification) (Stecanella, 2017). For example, consider a data set with two features (x, y) like in the figure below, the SVM works by finding the optimal hyperplane that has the greatest margin between both classes.

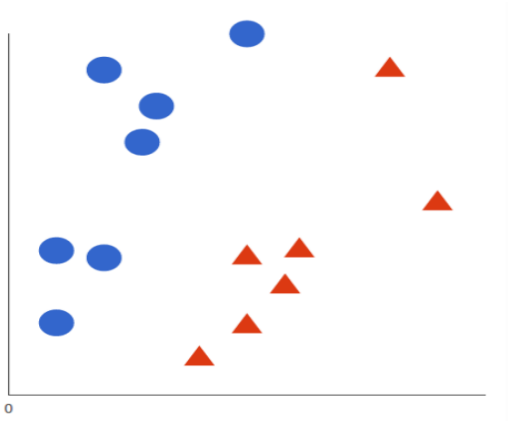

Figure 2.7 A dataset with two features. (Source: Stecanella 2017)

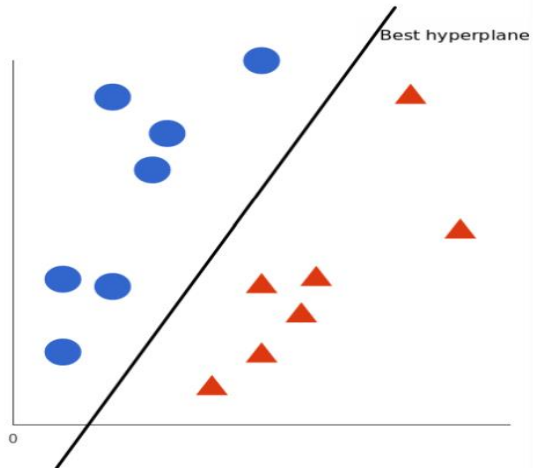

Figure 2.8 The output of a trained SVM. (Source: Stecanella 2017)

In the figure above the SVM has found the optimal hyperplane that divides the separate classes optimally. New data input is then classified into either side of the hyperplane. In reality however, the data sets are not as simple and contain non-linear relationships, the SVM classifies such data sets using a method kernelling. Kernelling involves transforming a complex data set into a dimension that enables it to be classified using an optimal hyperplane. For example, a non-linear data set in 2D can be transformed into 3D in other for the hyperplane to have any chance to classify the data optimally.

v. Genetic Algorithms:

Genetic algorithms belong to a larger class of algorithms called evolutionary algorithms. Genetic algorithms are meta heuristic search algorithms that work by performing functions similar to biological processes such as natural selection, mutation, reproduction on data streams to determine the healthiest solutions to a problem (Carr, 2014). A typical example is training a GA to find the optimal route between two cities

A &B given different routes. The algorithm begins an iterative process where it selects the possible routes between the two cities and then attaches a fitness value to the routes, after which evolutionary processes are carried out on the various solutions. This process generates a new set of solutions with usually better fitness values. The process repeats until a threshold fitness value is reached or the maximum level of generations is reached. GA is usually applied to reinforcement learning and unsupervised learning problems. In the credit scoring solutions, it is used as an optimization technique or as a feature selection method, because it is better than traditional methods at optimizing solutions to real world problems. The performance of genetic algorithms is not consistent across all problems as there is no guarantee that the optimal solution will be achieved, and the fitness value as defined by the programmer has a big effect on the overall performance of the algorithm.

2.5 Other Notable Approaches

There have been other notable approaches in improving credit scoring solution existing in literature, relevant ones as connected to this study are discussed below

2.5.1 Soft Computing methods

Soft computing was coined by Lotfi Zadeh (1994) the father of fuzzy logic. He defined it as "a collection of methodologies that aim to exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and low solution cost." (Zadeh, 1994, p. 77) Its principal constituents are fuzzy logic, neurocomputing, and probabilistic reasoning. Soft computing solutions leverage human like thinking in approaching problems that contain imprecise and incomplete information as it is sometimes found in credit scoring (Ramík, 2017). In literature Wang, Wang, and Lai (2005), used a fuzzy SVM to evaluate credit risk, other researchers have mostly employed one soft computing member for feature selection.

2.5.2 Feature Selection methods

Feature Selection methods are hybrids that involve the use of heuristic, search algorithms generally as a pre-processing stage in model creation (Koutanaei, Sajedi, & Khanbabaei, 2015). This is intended to improve classification performance because these methods can: 1. Decrease the noise in the Dataset; 2. Reduce computational costs; 3. Make updating the model easier; 4) Simplify the classification process (Salappa, Doumpos, & Zopounidis, 2007). Feature Selection techniques are able to pick out the features that greatly affect the ability of an

algorithm to classify new input while removing features that do not have an impact on the classification accuracy. Examples of feature selection techniques are Genetic Algorithm, Information gain ratio, Principal Component Analysis (PCA). Wang and Huang (2009), conducted a case study of credit card data using evolutionary algorithms, Wang, Hedar, Wang, and Ma (2012), used rough set and scatter search meta heuristic in FS were used for credit scoring. Oreski and Oreski (2014), proposed a method for identifying the optimum feature set to increase classification accuracy and scalability using GA and ANN. Examples in literature where these were used are (Chen & Li, 2010; Marques, Garcia & Sanchez, 2012).

2.5.3 Hybrid Models

Hybrid models are a blanket term for any combination of data mining techniques in credit scoring. The algorithms are combined usually for increasing predictive ability or reducing computation cost. According to Chen, Ribeiro, and Chen (2015), there are three structures to developing a hybrid model, as shown in the figure below.

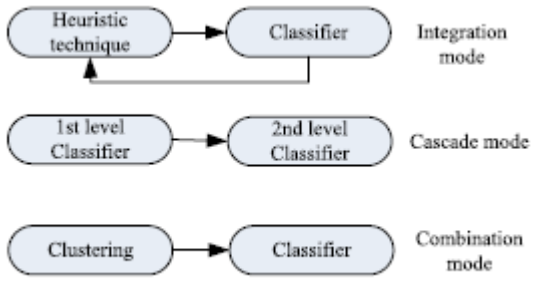

Figure 2.9 Structure of hybrid systems. Source (Chen, Ribeiro, & Chen., 2015)

The first type involves the use of heuristic algorithms for feature selection as discussed in the previous sub section. The second structure comprises of two or more algorithms where the first algorithm feeds learned data into the second as input and the second to the third in a cascading structure. In the third structure clustering is used as a pre-process stage for the eventual classification algorithm, the clustering aims at identifying outliers and noise in the input data in order to eliminate them. Another special type of hybrid called Ensembling is discussed in the following subsection. The following studies applied hybrids (Liu, Fu, & Lin, 2010; Pavlidis, Tasoulis, Adams, & Hand, 2012; Vukovic, Delibasic, Uzelac, & Suknovic, 2012).

2.5.4 Ensemble methods

Ensemble methods are special kind of hybrid that structure the combination of learners in a specialized way, they are also known as multiple classifier systems. The learners combined can

be homogenous that is learners of the same type or heterogenous that is learners of different types. There are majorly three types of ensembles used in credit scoring namely, Bagging, Boosting, and Stacking.

i. Bagging

Bagging or Bootstrap aggregating developed by Breiman (1996), involves two steps bootstrap and aggregating. The data set is bootstrap sampled such that it is randomized among the base learners (independent leaners) where some data instances may appear more than once in a data set and not appear at all in other data sets (Patil, Aghav, & Sareen, 2016). This process fosters the ability of the base learners to generalise, by making them diverse viz a viz the data set. The other part which is aggregating is how the output is attained and this is usually via some kind of voting usually majority voting or averaging for regression problems. Bagging is an Ensembling method that is based on parallel computation and as such has excellent robustness.

ii. Boosting

Boosting as opposed to bagging is a serial Ensembling method. It was popularised by the Adaboost developed by Freund and Schapire (1996), which was geared at converting weak learner into a strong learner. A weak learner is an algorithm that is a little better than a random guess. In boosting, a sequence of weak learners are trained sequentially on a bootstrap sampled data set, the misclassified samples of the first weak learner are fed into the learner after it such that the consequent learner is able to learn from the mistakes of the previous learner. This is done iteratively with an arbitrary number of weak learners hence boosting the overall classification ability of the ensemble. This process takes place until a suitable level of accuracy has been achieved. Boosting is a good method of Ensembling but that model can easily overfit the data and may not be as robust as bagging ensembles.

iii. Stacking

Stacking is a special form of Ensembling with combines several strong learners as base learners in a parallel structure similar to bagging. In stacking however, the combiner isn't a voting or averaging method but another strong learner called the meta classifier (Wolpert, 1992). The meta classifier is able to learn the behaviours and faults of the previous learners and improve on their performance (Patil, Aghav, & Sareen, 2016). In Stacking improper training method can easily lead to overfitting, thus a means of varying the data set for each base learner and especially the meta learner is required.

This is achieved albeit cross validation or leave one out procedure which increases the diversity and reduces the bias in the stacking ensemble (Zhou, 2012). In theory, stacking ensembles are able to simulate the performance and behaviour of the other types of ensembles depending on the type of meta level classifier chosen. In practice however, logistic regression is the most commonly used combiner algorithm. From research there is no one Ensembling method that is consistently better than the other, as their use and performance depends on the domain area.

2.6 Review of Related Work

S/No	Authors and Year	Work Title	Objective	Method	Result	Gap
1.	Baker (2015)	Consumer Credit Risk Modelling	The aim of the work was to compare in terms of scalability, stability and performance their variant of CART and Random Forest to the standard versions of the same algorithms.	Custom CART and Random Forest algorithms were developed using R.	Their Custom CART achieved a score 0.868 and the custom RF achieved a score of 0.895 when evaluated using the AUC	Overall, their custom algorithms took much longer to run when compared to the standard implementations. Scalability was not measured by varying the data set.
2.	Koutanaei, Sajedi, and Khanbabaei (2015)	A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring.	This work had several objectives which are 1. Conducting an extensive study of the credit scoring problem in relation to	The work used four feature selection algorithms namely, PCA, GA, Information Gain ratio and Relief Selection methods	Their results were that PCA was observed to be the strongest FS method. For the classifiers the best in terms of	The discovered solution is a black box solution and is prone to overfitting.

			different FS methods and classifiers. 2. Hybrid use of three general, ensemble, and FS based credit scoring techniques. 3.Evaluating the different parts of the hybrid individually.	to generate parameter for the proposed model. These parameters were first tested with SVM to evaluated their accuracy before the parameter are fed into the single and ensemble models.	accuracy and AUC was ANN-Adaboost with mean value of 91.1%.	
3.	Wang, Xu, and Pusatli, (2015)	A Survey of Applying Machine Learning Techniques for Credit Rating: Existing Models and Open Issues	The aim of this work was to review algorithms and ideologies in the credit scoring field in relation to two commonly used data sets.	The work reviewed credit rating accuracy for all classifiers used on this data sets. It went on to proposed research directions.	The results showed that feature selection methods and Ensembling methods had the best performance on the two data sets.	Hybrid Ensembles were not evaluated in the review. Also, Accuracy is not the only characteristic that can be used to evaluate classifier performance objectively.
4.	Turkson, Baagyere, & Wenya., (2016)	A Machine Learning Approach for Predicting Bank Credit Worthiness	The objective of this work is threefold. First to understand the nature of the data set and determine the kind of	15 algorithms including but not limited to: Logit, SVM, CART, K-NN, ANN, Bagging	15 algorithms were used and only two failed to perform favourably. Using linear regression, the work was able	The work also cannot classify thin file customers accurately. Linear regression also makes assumptions on the data which affects

			algorithms that will perform best on the data set. Secondly, to determine with of the 23 features are the most important for predicting customer default. Lastly, using linear regression, the credit worthiness of a customer is predicted.	were trained with all features and trained again using extracted features. Linear regression was later applied on the three most outstanding features.	to determine that only 5 features were required to classify accurately.	its robustness.
5.	Lanzarini, Monte, Bariviera, and Santana. (2016)	Simplifying Credit scoring Rules with LVQ (Learning Vector Quantization) +PSO (Particle Swarm Optimization)	The study had two objectives. First to benchmarking their solution against two classification techniques. Secondly, to show that the solution is more intuitive than existing solutions and promotes transparency of CRA	To achieve their objectives. A competitive neural network and an optimization algorithm were combined in this study for handling numerical and nominal data in this study.	Their hybrid model was able to generate simpler rules as compared to rules from C4.5 tree and PART.	The solution put forward was not accurate as the algorithms it was benchmarked against.

6.	Chang, et al. (2017)	An Innovative Framework for Building Agency-free Credit Rating Systems.	The work focused on developing a Credit Scoring process that was exempt of Credit Rating Agencies.	A hybrid model was developed that combined Duffe's Model, Logistic Regression and Random Forests in predicting the credit rating, rating migration of rating and rating stability of sovereign entities using historical data gathered over 5 years.	Random Forests achieved the best results at generating credit ratings when evaluated using AUC of ROC.	The model was not scalable and sensitive to noise. Secondly, the predictive quality of the model on migrating of rating was low.
7.	Wang, Zhong, Zhang, and Zou (2017.)	A New Classification Algorithm for the Bank Customer Credit Rating.	The main aim of the study was to develop an algorithm with superior efficiency as compared to existing solutions.	The algorithm used combined PCA for dimension reduction for a novel voting system of two SVM models and a Random Forest.	A high accuracy of 97.92% for the Chinese data set and a low accuracy of 80% on the German data set was achieved by their algorithm. The algorithm also fared competitiv	Their work was not robust as changes in the distribution of the datasets caused significant reduction in the classification accuracy of their solution.

					ely well when compared to other hybrid algorithms.	
8.	Chen, Dautais, Huang, and Ge. (2017)	Data Driven Credit Risk Management Process: A Machine Learning Approach	The work focused on developing a completely automated credit risk management system.	SVM and feature selection technique was used at several stages in the credit analysis stage in a bid to completely remove bias from the classification process.	They achieved a 64% Gini coefficient on their SVM model. They were able to completely remove human involvement in the credit management process.	The proposed solution was a black box solution. SVMs were applied throughout the system which affected the accuracy of the solution.
9.	Huang and Chen (2018)	Domain Adaptation Approach for Credit Risk Analysis.	The main of objective of the study was to compare domain adaption approach to existing data mining techniques.	Their approach was to train algorithms on similar domains with binary classification problems and applied the algorithms to peer to peer lending systems.	They achieved slightly better predictive accuracy than K-NN and Decision Tree algorithms.	The tools used in the domain approach weren't mentioned, also, the scalability and interpretability of the solution were not evaluated.
10.	Ding, Ma, and Zhou, (2018)	Implementation of dynamic credit rating	The objectives of this study were	The work proposed a model that	The solution put forward	The work was not interpretable due to the

		method based on clustering and classification technology.	to improve the credit rating process for online trading companies by making them more dynamic and to calculate the credibility of the ratings in an objective manner.	combined fuzzy set theory with time frames, fuzzy clustering to measure credit rating of users in the short and long term.	was able to adjust accurately to new information and score users dynamically over time frames.	complexity of the method used. Fuzzy clustering theory also introduced some degree of bias in the evaluation of user's credit.

2.6 Synthesis of Review of Related Work and Important Issues

In the review of related work, it can be seen that the prominent problems are those of scalability, robustness, efficiency, and accuracy. The solutions reviewed are recent and represent the new thought process and research direction of the researchers in the credit scoring domain which show the following:

I. Hybrid models: Ensembles and otherwise, are better in performance than any single classifier (Koutanaei, Sajedi, and Khanbabaei, 2015; Wang, Xu, and Pusatli, 2015).

II. Feature Selection process is useful for reducing the complexity of the credit scoring whilst optimizing the solution (Koutanaei, Sajedi, & Khanbabaei, 2015; Turkson, Baagyere, & Wenya, 2016; Wang, Zhong, Zhang, & Zou, 2017).

III. The dynamism of the solution and its ability to process credit scores with their changes relative to time makes the solution more objective (Ding, Ma, & Zhou, 2018).

IV. Scalability and Robustness of the solution have also become important due to the global nature of the credit market and the imbalance nature of real data sets.

Being informed by the gaps and trends in recent and closely related literature, this study proposes to address the pressing issues whilst leveraging achievements of the previously

proposed solutions. In this problem domain, and relating to the solution proposed in this study, there are two concepts that are being debated about. They are discussed below.

2.6.1 Diversity

Diversity in Ensembling is a contentious subject Brown, Wyatt, Harris, and Yao (2005), but many researchers in the past have agreed that the learners in an ensemble have to be diversified in other for any improvement in learning to occur (Cunningham & Carney, 2000; Krogh & Vedelsby, 1995; Lam, 2000). The definition and understanding of diversity is still unclear in literature and, the extent to which diversity is important and how to properly measure diversity are unclear and subjective. This has led new research in the area according Kuncheva and Whitaker (2003). Diversity in Melville and Mooney (2001.), is defined as the amount of "dissagreement" between the learners used in an ensemble, and proposed to measure the diversity of algorithms by their level of accuracy. This definition is not widely accepted as in practice the learners are trained with the same data set and tend to be highly correlated if they are sufficiently accurate.

Concusively, achieving diverse learners in practice is a difficult task considering the fact that the combined learners need to be sufficeintly accurate to be used and leaners do not produce different posterior probabilities. In the end, the success of ensemble learning rests on an effective tradeoff between the individual performance and diversity (Zhou, 2012).

2.6.2 Interpretability

Interpretability is another buzz word hotly debated in the machine learning community, since machine learning algorithms have been applied to practical problems the need for "interpretability" has vastly increased. Interpretability from literature can take on many definitions: it could mean the ability of an algorithm or model's functionality and structure to be complete described, it could mean the ability of a model to show causality relationship clearly between input and output, it could also mean the ability of a model to show the relationships between several feature in the input streams.

In reality, to achieve any form of interpretability is a difficult task. Considering that problems have high dimensional number of features and the data sets represent complex relationships between input and labels, even simpler techniques like Decision Trees and linear regression lose their interpretability, if applied to such problems. Ultimately, heavy data pre-processing and feature engineering are used to support simpler, more interpretable models, but this can

easily lead to a bias in the output of the models and generate uninterpretable features. As put by Zachary Lipton (2017) researchers could be left with a solution that is simpler but has uninterpretable features and an uninterpretable solution that uses the raw features in the data sets. So, the need to determine Why we want interpretability? What notion of interpretability applies? and What we are willing to sacrifice? Become paramount. (Zachary Chase, 2015)

CHAPTER THREE
METHODOLODY

3.1 Introduction

This chapter presents the methodology adopted in solving the problems outlined in the introductory chapter. A research methodology simply put is the systematic plan for conducting a research (Moffitt, 2018). A good methodology presents the taxonomic and theoretical analysis of the methods that are applied in a specific field of study. Section 3.1 describes the design of the research, giving an overview of the methodology in a step wise fashion. Section 3.2 presents the proposed framework in relation to the existing framework in literature. Section 3.3 discusses the experiment design, visiting the tools and evaluation methods to be employed in this research.

3.2 Research Design

A research design is a well-developed plan as to how the information gained from literature and methods used will address the research problem. It helps the researcher justify methods and tools used in an objective and unambiguous manner. A good research design enables conclusions drawn from research to appear cogent and well rounded.

I. Existing literature is reviewed, key concepts, milestone and closely related work are discussed to show a good understanding of the credit scoring and machine learning niche.

II. Publicly available data sets are obtained from the UCI Machine Learning Repository. The data sets are the German and Australian credit data sets, they are then pre-processed by: normalization of their attributes, analysing missing attributes values, performing preliminary data exploration and outlier detection for the purpose of understanding trends and optimizing the data set for analysis.

III. The proposed framework is developed using the stacking ensemble method, which combines a committee of learners namely; K-NN, C4.8 Decision Tree, and SVM as base learners in a parallel structure and Logistic Regression as the meta classifier responsible for processing the predictions of the base classifiers/learners. The ensemble is trained and tested using WEKA (Waikato Environment for Knowledge Analysis) which is a popular, free machine learning software developed at the university of Waikato and written in Java. The software is ideal for training and testing various machine learning algorithms.

IV. The framework will be evaluated against the single classifiers such as K-NN, MLP, SVM, Logistic Regression, ID3 Decision Tree which are the most frequently used single classifiers in literature, state of the art solutions like ensembles and hybrid feature selection methods which represent the current state of research in the credit scoring domain, using metrics such as Confusion Matrix, Area Under the Curve (AUC) of the ROC.

3.3 Frameworks

A conceptual framework for researcher presents his/her understanding of existing literature in a subject field (Reginiel, 2015). It is also an embodiment of the researcher's observations on the subject matter and structures how the research plans to approach what they have discovered from literature. As such, an existing framework which shows what the research intends to leverage is presented as shown below. The framework proposed for this study is immediately after it.

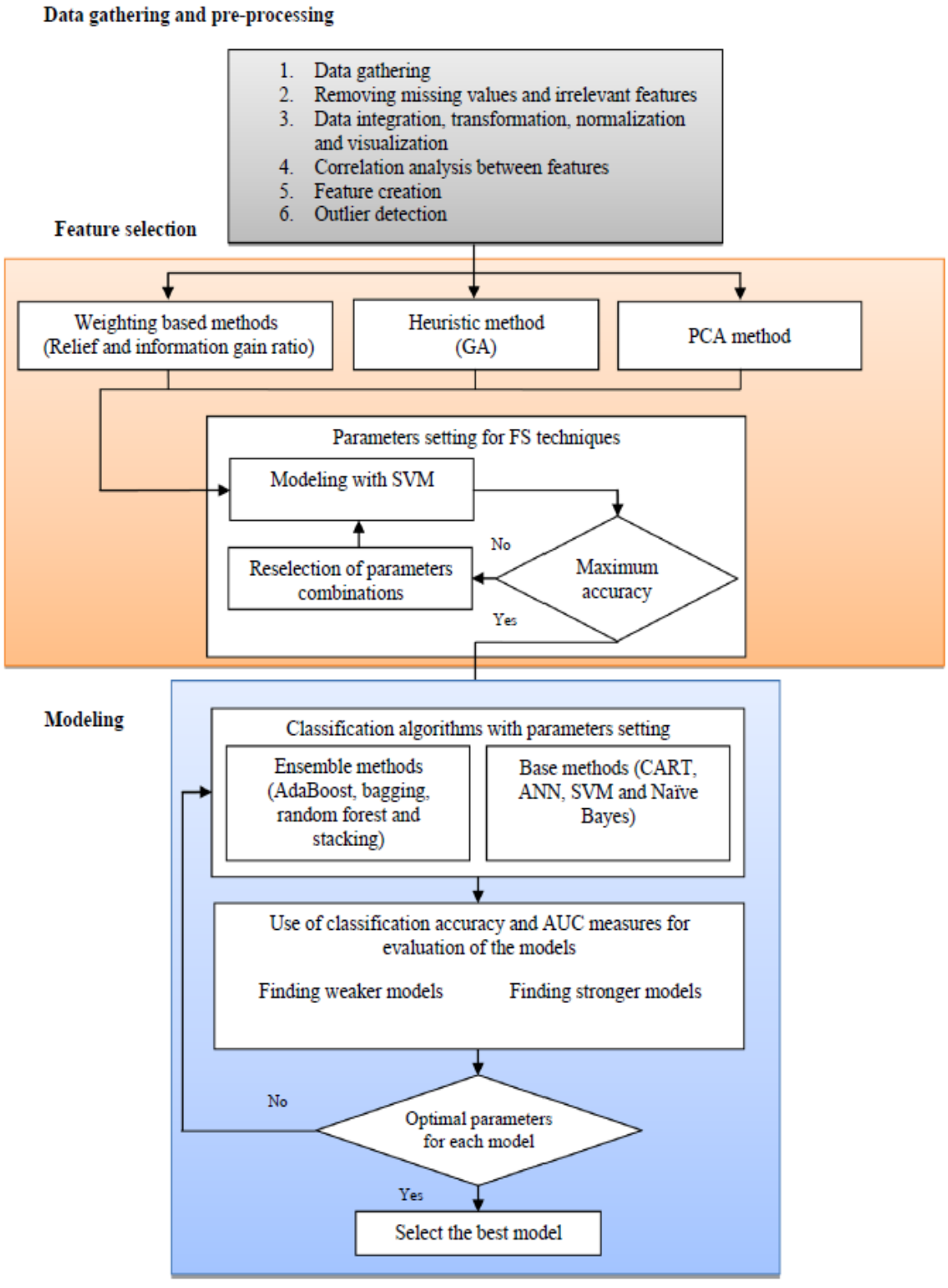

Figure 3.1 Existing framework. (Source Koutanaei, 2015)

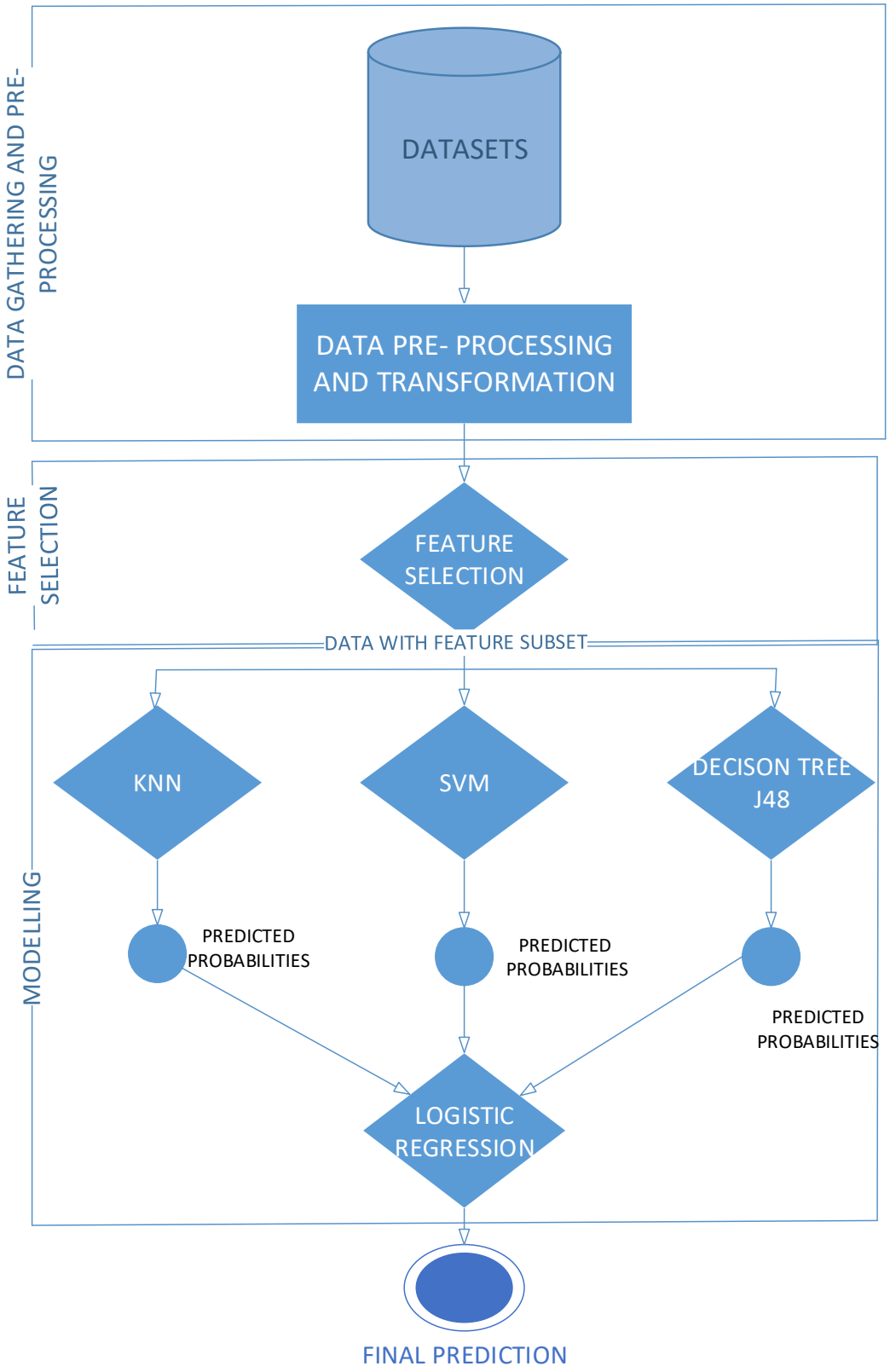

Figure 3.2 Proposed Framework

3.4 Experiment Design

Following the architecture of the proposed framework, the experiments conducted applies the methods specified. The experiment is conducted in Weka environment which contains all of the methods with default parameters, used in this study and also allows for objective result interpretation and comparison. The data sets used were accessed from the publicly available UCI Machine Learning Repository and have been used several times in literature, thus presenting a viable means for model comparison. The experiment is carried out with a 10-fold cross validation for all base learners. To test the robustness of the solution and its performance on class imbalance, the framework is evaluated on the data sets with imbalance ratio of 70%-30%, 80%-20%, 90%-10%. This imbalance is achieved by under sampling the minority class, which in this case is the defaulter class.

For machine learning endeavours the process for conducting experiments is a well-established one and it is followed in this research, albeit through a feature selection process. It begins with data pre-processing, then a model selection or fine-tuning process, after which the model is tested. This process is shown diagrammatically below with the flow of data.

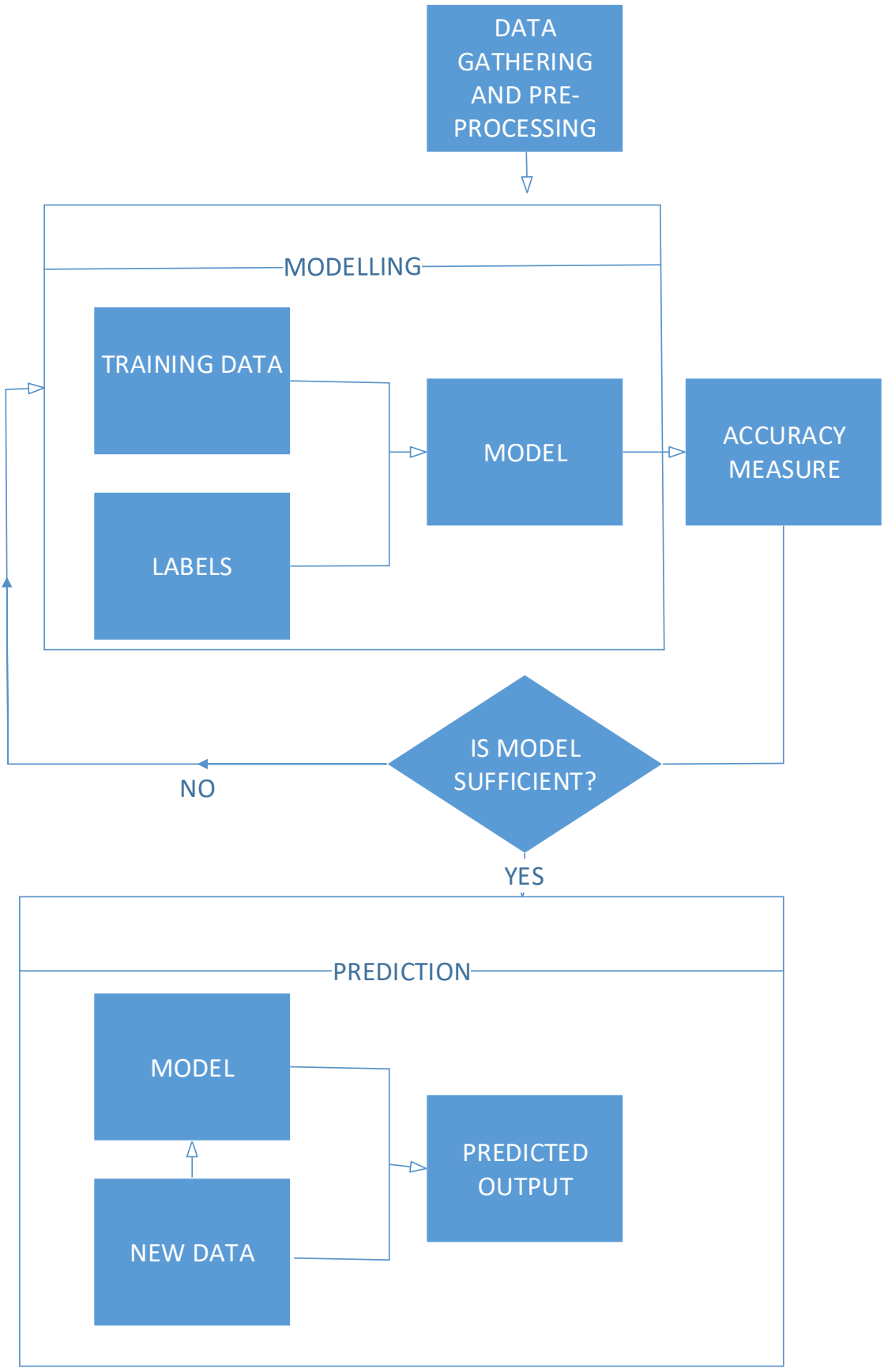

Figure 3.3 Experiment Design

3.4.1 Data Set Description

The Data set used for this study is gotten from the UCI Machine Learning Repository, they are the Australian, German and Taiwanese data sets. The German data set contains 1000 instances, of which 700 are good credit and 300 are bad credit. The data set has 20 attributes or features (7 numerical and 13 categorical attributes) although the data set has been revised to have 24 numerical attributes for algorithms that cannot process categorical data. The Australian data set contains 690 instances, 45% positive and 55% negative instances. The data set contains 14 attributes consisting of 6 numerical and 8 categorical attributes. This dataset was donated by Quinlan (Quilan, 1987). Features present in the data sets are generally divided into two: financial status features and general and demographic features.

There are 4 major validation procedures in literature namely hold out sample, K-fold Cross validation, Leave One Out, Train-Validation-Test. In this dissertation, K-fold Cross Validation was chosen. K-Cross Validation procedure involves splitting the training data sets into K folds and then training a classifier on K-1 folds in the training set. The last fold which was not used for training will be used to test the performance of the classifier. This can be done K times and then the accuracy averaged. This validation approach ensures that the data set is fully maximized and the algorithms performance is not affected by how the data set is split. K- fold is used for two reasons. Firstly, the data sets used in this study are relatively small so in order to fully maximize the data set, k-fold cross validation which uses the entire training data set for training and testing is used. Secondly, k-cross validation overcomes the shortcoming of the random split. Other validation methods, such as hold out approach, causes the performance of the classifiers to depend on how the entire data set was split.

3.4.2 Methods and Justification

The various methods or tools used for this study are described and justified below

i. Support Vector Machines

Support Vector Machines (SVM) is a machine learning algorithm that is suitable for many tasks, including classification and regression. Credit scoring being a classification task, can be explored with SVM. SVMs work by mapping data instances with non-linear relationships to a high dimensional space so that they become linearly separable and performs vector operations on them in a process called kernelling. In this high dimensional space, a hyperplane that best separates the two classes present in the data set is developed. SVM can have different types

of kernels, from literature there are three major types used depending on how complex the data set is. They are: Linear kernel, Radial Bias Function kernel and Least squared kernel. SVM has been chosen for this study as a base classifier because, from literature it is has been observed that SVM has good level of accuracy when applied to real world data sets and is arguably the most computationally efficient machine learning algorithms used in credit scoring.

ii. K-Nearest Neighbours

K- Nearest Neighbours algorithm (KNN) is a classification algorithm that functions by representing data instances as vectors positions in a hyperplane. During the test stage it classifies new input by averaging the votes of the K nearest data points to the new data input. Although this is a simple process, the algorithm achieves surprisingly good accuracy on most problem domains. It requires little or no training hence making it computationally efficient. In relation to the credit scoring problem, KNN has competitive performance when compared to other single classifiers with better computational efficiency. Diversity among the base classifiers also informed the decision to use K-NN

iii. Decision Tree

Decision Tree is an intuitive machine learning algorithm that predicts an outcome by splitting input values by their features using various rules. A decision tree is a hierarchical model that consists of decision nodes and terminal nodes, that show an inductive process from input with features to predicted value. There are several types of decision trees; ID3, CART, C4.5, C5.0. For this dissertation C4.5 is used, the variant of this available in Weka is called J4.8 because it is built in Java. In credit scoring, DT has good performance against other single classifiers and has been shown from literature to be immune to class imbalance. DT of all of the base classifiers has the highest interpretability, because the tree can be written as nested if-else statements. For the aforementioned reasons Decision trees have been chosen as a base classifier for this dissertation.

iv. Logistic Regression

Logistic Regression (LR) is a tool for binary classification that classifies data instances by calculating the probability that it belongs to either of the classes. The classes are determined by a transformation of the linear combination of the features in the input stream. LR is arguably the most commonly used tool in practice in the credit scoring, this is because it is interpretable and has competitive performance

when compared to more complicated algorithms. LR in the dissertation is used as the meta classifier.

v. Feature selection using Genetic algorithm

Feature selection is a special type of dimension reduction, that is concerned with finding the optimal subset of features from the entire feature set that enable classifiers to perform with good accuracy or equal accuracy as to when they are trained with the entire features of the data set. In other words, feature selection techniques help removes redundant and highly correlated features from the data set, there by optimising the classifier that is trained afterwards. There are many feature selection techniques, chief among them is Genetic algorithm (GA), Principal Component Analysis (PCA), Information Gain ratio, Relief Based methods. Genetic Algorithm is a feature selection algorithm that uses rules such as crossover, natural selection, mutation with are evolutionary concepts from biology to determine the set of features with the best fitness values. Interpretability being one of the problems this dissertation aims to address informed the use of genetic algorithm, because genetic algorithm produces real feature sets rather than the synthetic ones produced by PCA.

vi. Stacking

Stacking is a type of Ensembling where a number of classifiers are trained on a data set and their outputs are supplied as probabilities to another classifier called a meta classifier. Stacking is an interesting method of Ensembling that has famously been used to win several Kaggle and Netflix competitions. It allows for non-linear combinations of the probabilities from the base learners, which further enhances the performance of the model as a whole. An excerpt from Wolpert (1992), describes stacking as “a means of non-linearly combining generalizers to make a new generalizer, to try to optimally integrate what each of the original generalizers has to say about the learning set. The more each generalizer has to say (which isn’t duplicated in what the other generalizer’s have to say), the better the resultant stacked generalization.” It is encouraged for all the base classifiers to “span the space”, this means that the base classifiers used have to look at the data in different ways. This is achieved by using diverse classifiers, and diversity in this sense means that base classifiers should comprise of space searchers, Turing machine simulators, non-linear functions etc. The use of stacking in this dissertation is due to its good accuracy and its immunity to bias from individual classifiers and class imbalance.

vii. WEKA

WEKA Waikato Environment for Knowledge Analysis is an open source machine learning tool written in Java, that can be used for many machine learning tasks as varied as classification, clustering, regression, interpretation and visualization. All the techniques to the used in this study are present in the WEKA tool, with parameters that can be varied depending on the part of the experiment being conducted. WEKA with its intuitive design allows for models to be developed by simply understanding the use of several buttons on the workbench, which allows the researcher to focus on understanding the intricacies of the models rather than programming code. WEKA also contains explorer for data pre-processing, and consequently data visualization and evaluation, which means that models trained can be accurately compared and evaluated. The latest version of WEKA, version 3.8.2 was used for this dissertation. It was run on an i7-3630QM processor, 10GB RAM, Windows 10 64-bit machine. An image of the WEKA machine learning tool is shown below.

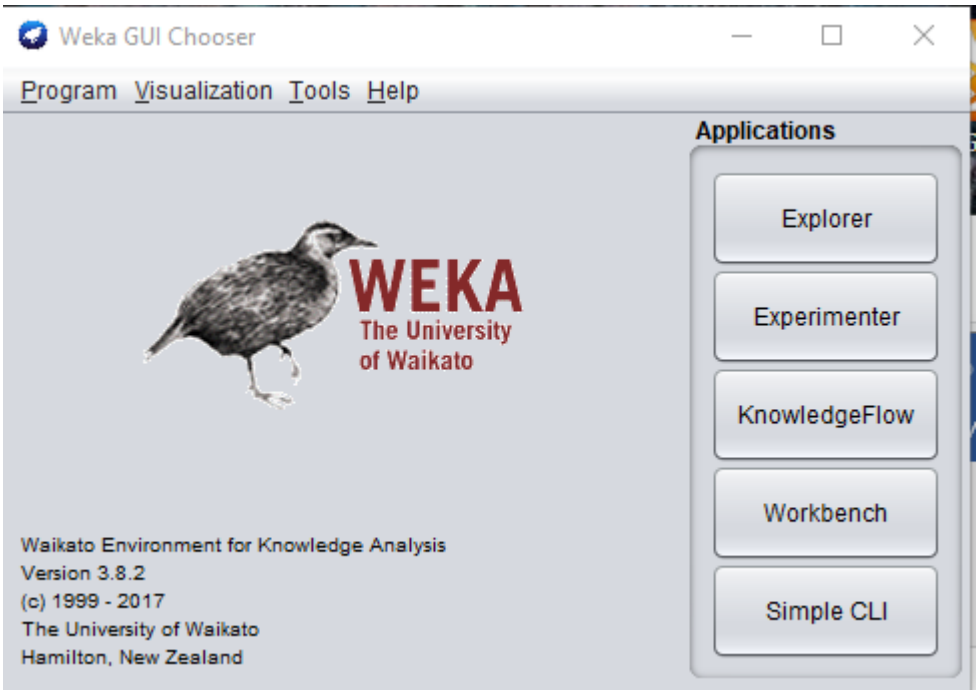

Figure 3.4 Weka GUI Chooser. (Source: Weka 3.8.2)

viii. Evaluation Methods

The evaluation methods used for dissertation is the AUC of the ROC curve and the confusion matrix, both of which have been applied in most literature in the credit scoring field.

Confusion Matrix: It is also known as the error matrix. It is a traditional means for evaluating the performance of a classifier on a specific data set where the true values are known as shown below.

		Model Predictions	
Data Set Values		True	False
	True	TP	FN
	False	FP	TN

From the above table, TP is the number of True Positives, FP the number of false positives, FN is the number false negatives and TN is the number of true negatives. These metrics are used to draw some conclusions from the table, such as, obviously, TP+FP+FN+TN=N, where N is the total number of data instances.

Accuracy (ACC) is the ratio of correct predictions made by the model when classifying input into class True or False. Defined as ACC = (TP+TN)/(TP+TN+FN+FP).

Sensitivity (R) also known as **Recall** is the fraction of class True that the model correctly classified as belonging to the class True. It is defined as R = TP/(TP+FN).

Specificity also called the **True Negative Rate** is the fraction of instances correctly classified by the model into class False among all instances belonging to the class False. Given by, SPE = TN/(TN+FP).

Precision is the ratio of true positives and the total number of instances labelled as positive. It is defined as TP/(TP+FP).

False Positive Rate which is also called the type two error is the fraction of class true cases that were misclassified as belonging to class false. It is given as FPR = FP/(FP+TN) = 1-Specificity.

Area under the Receiver Operating Characteristic Curve (AUROC): Area under the receiver operating characteristic curve is a means of measuring the quality of the ability of a learner to properly classify an input into one of two classes. The ROC is a graph that plots Sensitivity on the y axis against 1-Specificity on the x axis. On this graph a random classifier will create a diagonal and have an AUC of 0.5, meaning that it cannot successfully distinguish between the two classes as

shown in the figure below. AUC simply means the area of the graph which the curve covers.

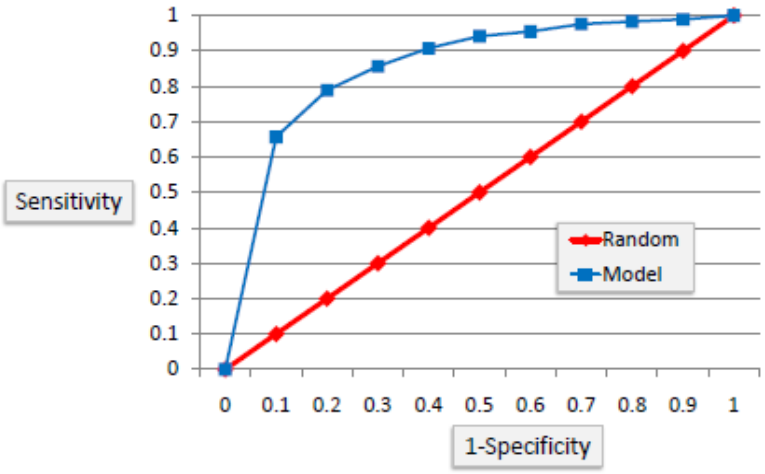

Figure 3.5 ROC Curve. (Source: Sayeed, 2018)

A good model will be expected to have an AUC close to 1 and curve like shown in blue above, a perfect classifier will have an AUC of 1 if there are no overlapping data instances in the data set. A poor model however, will have a ROC closer to the diagonal line in red, meaning it is no better than a random guess. AUROC is specifically useful because it is immune to class imbalance, that is, the graph and AUC will remain objective even when one class is relatively under sampled compared to the other.

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the findings and observations made from carrying out the project, these results are also discussed. Section 4.1 discusses the attributes, visualizing them and feature selection results. Section 4.2 presents and compares the performance of various single classifiers and the proposed ensemble on the data set in its original state, these classifiers are compare with the previously described metrics such as; AUCROC curve, confusion matrix, time spent. Section 4.3 shows the results when feature selection has been performed on the data set, the proposed ensemble is compared to single classifiers, hybrids and other ensembles using WEKA and the body of existing literature. Section 4.4 shows the results when the data set is artificially imbalanced by over sampling the majority class. Section 4.5 discusses the results observed and how they relate to the existing body of knowledge.

4.2 Attributes and Attribute Selection

Two datasets were used in this dissertation, the first is the Australian dataset which contains 15 attributes including the class attribute. The attributes in this dataset have been encoded from their original values: especially for the categorical attributes for ethical reasons, there are 6 numerical and 8 categorical attributes in the 15-total number of attributes. A few attributes values were missing in this data set, some were replaced with the mean value of the samples in that same class, some were replaced by using inference-based methods such as decision tree induction. The figure below shows the attributes visualized, the distribution of the instances with their attributes in relation to what classes they belong, the darker shade being the negative class instance, the lighter shade being the positive class instance.

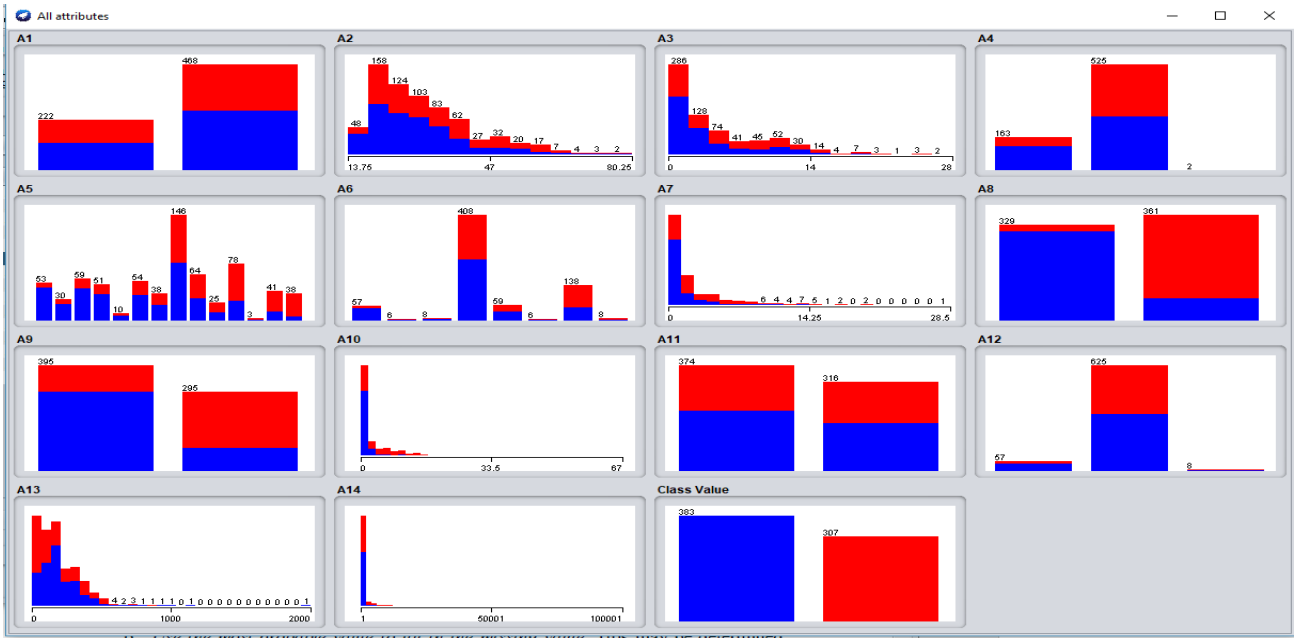

Figure 4.1 Visualized Attributes from the Australian data set. (Source: WEKA)

The attributes were further analysed by ranking the importance of the attributes using information gain ratio and a ranker search method; which qualitative compares the information in each attribute in the dataset to see which of the attributes are most relevant. This method was proposed by Ross Quilan and is typically used for decision trees when deciding what nodes to split on. Below are the rankings of the attributes in the Australian data set.

Average Rank	Attribute
1 + - 0	A8
2.1 + - 0.3	A9
2.9 + -0.3	A10
4 + - 0	A14
5 + -0	A7
6.4 + - 0.66	A3
7 + - 0.89	A13
8 + - 0.89	A4
9.6 + - 0.92	A5
10.4 + - 0.49	A6
10.6 + - 2.46	A2
11.6 + - 0.49	A12
12.9 + - 0.54	A11
13.5 + - 0.67	A1

Table 4.1 Ranked Attributes from the Australian data set. (Source: WEKA)

Feature selection performed on the dataset using genetic algorithm (ENORA) as described in the methodology, determined that 7 attributes were sufficient for the algorithms to classify the instances without significant loses in terms of accuracy and AUCROC curve. The results of the feature selection process were evaluated with a scheme independent evaluator called the

Correlation based Feature subset selection method (CFS Subset Eval) which checks the predictive ability of the features along the redundancy within them. It was also evaluated with scheme dependent methods like SVM and NB which are classifiers. These attributes were A4, A5, A8, A9, A10, A13, A14 of which 3 are numeric and 4 are categorical.

The second data set is the German data set, it has 20 attributes including the class value, 7 attributes are numerical whilst 13 are categorical. There are no missing values in this data set but it uses a cost matrix, which indicates that the cost of misclassifying a defaulter as good credit is 5, and the cost of misclassifying a good credit as a defaulter is 1. The figure below visualizes the attributes.

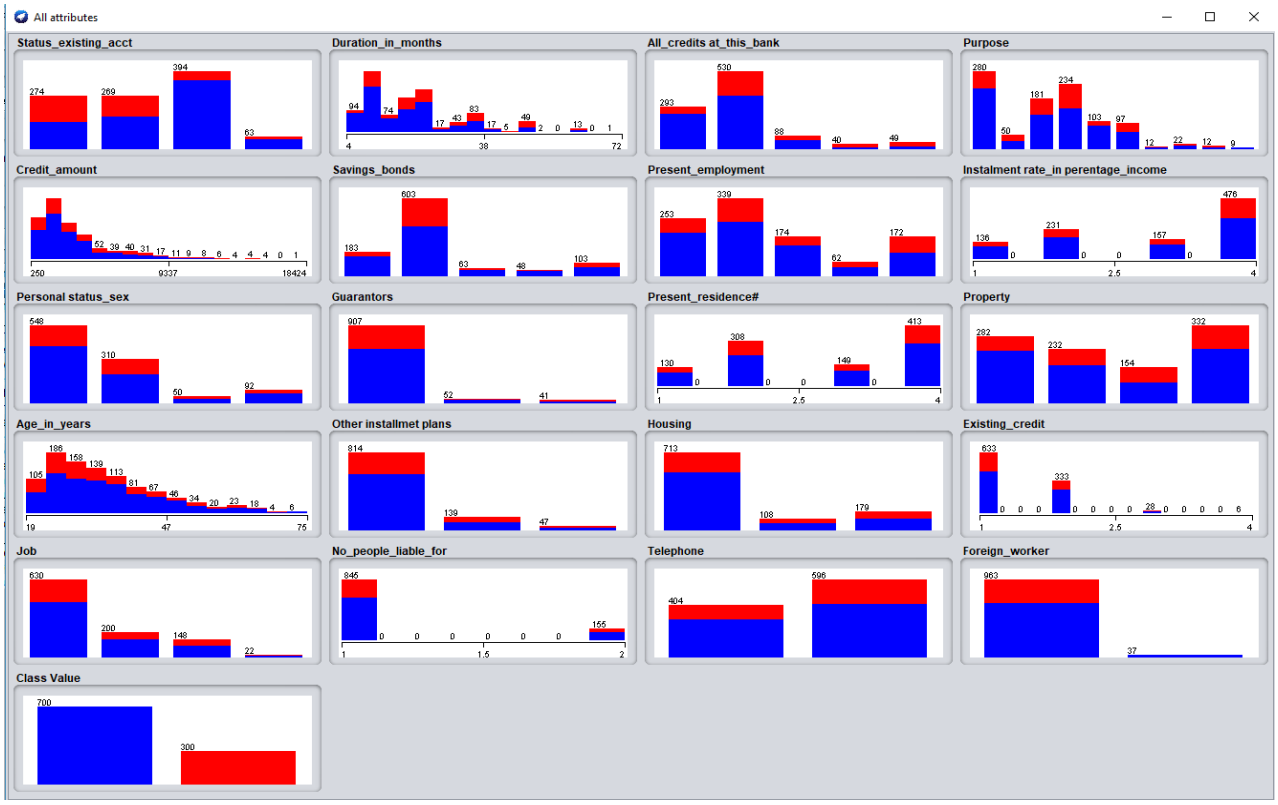

Figure 4.2: Visualization of German dataset attributes (Source: WEKA)

The attributes in the data set were ranked as shown in the table below after being evaluated by information gain ratio.

Average Rank	Attribute
1.1 + **-0.4**	1 Status_existing_acct
3.2 + - 1.08	3 All_credits at_this_bank
3.2 + - 1.4	5 Credit_amount
3.4 + -0.92	2 Duration_in_months

4.2 + - 1.33	20 Foreign_worker
6.3 + - 0.46	6 Savings_bonds
8.8 + - 0.98	15 Housing
9.5 + - 2.2	14 Other_installment_plans
9.7 + - 1.1	4 Purpose
10.2 + - 1.94	10 Guarantors
10.3 + - 5.62	13 Age_in_years
10.3 + -1.19	12 Property
12.5 + - 0.81	7 Present_employment
13.6 + - 0.49	9 Personal status_sex
15.1 + - 0.83	17 Job
15.3 + - 0.46	19 Telephone
17.5 + - 1.36	18 No_people_liable_for
17.9 + - 1.04	16 Existing_credit
18.6 + - 0.66	11 Present_residence
19.2 + - 1.25	8 Instalment_rate_in_percentage_income

Table 4.2: Ranked Attributes from the Australian data set. (Source: WEKA)

Feature selection was also performed on this data set using the technique described prior, and 7 (seven) attributes were selected as sufficient for accurate classification of attributes. The attributes were 1 Status existing account, 2 Duration in months, 3 All credits at this bank, 4 Purpose, 6 Saving bonds, 10 Guarantors, 14 Other instalment plans.

4.3 Performance of the Classifiers

The performance of the single classifiers and the proposed ensemble on the two datasets are presented in the tables below

Algorithms	**Mathew's Correlation Coefficient**	**AUC of the ROC Curve**	**AUC of the PRC**	**Time Spent Testing (ms)**	**Accuracy (%)**
Zero R	0.00	0.50 w	0.552 w	0.00	55.51(0.67) w
Naïve Bayes	0.55	0.89	0.905 b	0.47	77.26(4.20) w
SVM	0.71	0.86 w	0.851 w	0.16	84.88(3.92) w
MLP	0.64	0.89 w	0.890	0.94	83.17(4.63) w
LR	0.70	0.91	0.900	0.16	85.12(4.17)

KNN	0.61	0.90 w	0.890	3.91	85.68(3.83)
Proposed Ensemble	**0.73**	**0.92**	**0.898**	**4.22**	**86.20(3.68)**
C 4.5	0.70	0.88 w	0.857 w	0.00	85.17(3.56)

Key: All evaluations are made at a 0.05 level of significance

W = Significantly Worse

B = Significantly better

Values in brackets are Standard Deviations

Table 4.3 Performance of Single classifiers on the Australian data set. (Source: WEKA)

The results from the above table show that the proposed ensemble outperformed most of the single classifiers based on accuracy at a 0.005 level of significance and shared a similar level of performance with J48, Logistic Regression and K-Nearest Neighbours, which means that these algorithms were the best at predicting default/non-default on the Australian dataset. In terms of AUCPRC, Naïve Bayes has the best performance followed by the proposed ensemble, Logistic Regression and K-Nearest Neighbours. The AUCROC results follows the occurring trend as the Proposed Ensemble, K-Nearest Neighbours were the best performers which means they are the most scalable algorithms, if distribution remains the same on this dataset.

Algorithms	**Mathew's Correlation Coefficient**	**AUC of the ROC Curve**	**AUC of the PRC**	**Time Spent Testing (ms)**	**Accuracy (%)**
Zero R	0.00	0.50 w	0.309 w	0.16(1.56)	30.00 w
Naïve Bayes	0.42	0.73(0.04) w	0.458 w	1.09(4.01)	69.13(4.20) w
SVM	0.38	0.67(0.05) w	0.450 w	0.31(2.20)	74.73(4.05)
MLP	0.33	0.67(0.05) w	0.414 w	2.19(5.45)	70.12(4.63) w
LR	0.36	0.70(0.05) w	0.420 w	0.16(1.56)	63.64(4.35) w
KNN	0.31	0.65(0.05) w	0.411 w	10.00(8.46) w	71.87(3.61)
Proposed Ensemble	**0.36**	**0.77(0.05)**	**0.582**	**0.31(2.20)**	**74.72(4.00)**
C 4.5	0.30	0.65(0.05) w	0.402 w	0.16(1.56) w	65.98(4.47) w
MC-SVM (2014)	-	-	-	-	75.8

Key: All evaluations are made at a 0.05 level of significance

W = Significantly Worse
B = Significantly better
Values in brackets are Standard Deviations

Table 4.4 Performance of Single classifiers on the German data set. (Source: WEKA)

Accuracy of the algorithms on the German dataset were lower than 80%, with Chen & Li's SVM showing the best result of 75.2 which was at the same level with the proposed ensemble performance. The proposed ensemble had the best recorded values for AUCROC and AUCPRC with 0.77 and 0.582 respectively which was better than the performance of all the single classifiers on this dataset.

4.4 Performance of Classifiers on the data set after Feature Selection

The performance of the ensemble with the use of feature selection is presented below. The ensemble is compared with single classifiers, hybrids and other ensembles.

Algorithms	**Mathew's Correlation Coefficient**	**AUC of the ROC Curve**	**AUC of the PRC**	**Time Spent Testing (ms)**	**Accuracy (%)**
Single Classifiers					
Zero R	0.00	0.50 w	0.56 w	0.00	55.51 w
Naïve Bayes	0.70	0.90	0.92	0.00	77.42 w
SVM	0.73	0.86 w	0.86 w	0.00	85.59
MLP	0.71	0.91	0.91	0.63	85.57
LR	0.72	0.93	0.93	0.00	86.52
KNN	0.69	0.91 w	0.89 w	2.03	86.30
Proposed Ensemble	0.73	0.92	0.92	1.88	86.07
C 4.5	0.68	0.87 w	0.87 w	0.16	84.41
Hybrid Classifiers					
Random Forest	0.70	0.92	0.92	2.50	85.33
Fuzzy Rules (2015)	-	-	-	-	84.83
Bayesian Networks	0.70	0.92	0.92	0.47	85.14

Fuzzy Kernel (2014)	-	-	-	-	88.84
Ensembles					
AdaBoost	0.66	0.90	0.91	0.16	83.72
Bagged J48 Ensemble (2014)	0.72 -	0.92 -	0.92 -	0.00 -	85.94 87.23

Table 4.5 Performance of Classifiers on the Australian data set with Feature Selection. (Source: WEKA)

With Feature selection employed several algorithms shared a similar high AUCROC value, the Proposed Ensemble had a value of 0.92 which was comparatively equal to the performance of AdaBoost, Bagged J48 ensembles, Random Forest, Bayesian Networks which are hybrids, and single classifiers: Logistic Regression, Naïve Bayes. The proposed Ensemble has a good AUCPRC value which it shares with the aforementioned hybrids and Ensembles. The best accuracy in the table is from the work of Zhang, Gao, & Shi in 2014 which had an impressive value of 88.84 achieved with the use of penalty factors coupled with classifier fuzzy kernelling, the proposed ensemble achieves a competitive accuracy of 84.41.

Algorithms	**Mathew's Correlation Coefficient**	**AUC of the ROC Curve**	**AUC of the PRC**	**Time Spent Testing (ms)**	**Accuracy (%)**
Single Classifiers					
Zero R	0.00	0.50 w	0.30	0.16 b	30.00 w
Naïve Bayes	0.38	0.69 w	0.43	0.00 b	62.94 w
SVM	0.36	0.65 w	0.44	0.31 b	75.29
MLP	0.34	0.67 w	0.42	0.63 b	68.55 w
LR	0.37	0.69 w	0.42	0.16 b	61.91 w
KNN	0.27	0.64 w	0.39	4.53 b	66.35 w
Proposed Ensemble	0.35	0.74	0.55	6.25 b	75.28
C 4.5	0.32	0.67 w	0.40	0.00 b	63.26 w
Hybrid Classifiers					

Random Forest	0.29	0.67 w	0.53	7.03	61.37 w
Fuzzy Rules (2015)	-	-	-	-	73.51 w
Bayesian Networks	0.38	0.69 w	0.57	0.00 b	62.51 w
Fuzzy Kernel (2014)	-	-	-	-	73.20 w
Ensembles					
AdaBoost	0.33	0.72	0.49	0.08	67.10 w
Bagged J48	0.35	0.75	0.54	0.04	67.92 w
Ensemble (2014)	-	-	-	-	76.48

Table 4.6 Performance of Classifiers on the German data set with Feature Selection. (Source: WEKA)

On the German data set after Feature selection was applied, the proposed ensemble had an accuracy of 75.28 which is similar to the accuracy SVM and the ensemble developed in the work of Tsai, Hsu, & Yen (2014), these were the best algorithms in terms of accuracy. In terms of AUCROC, the best performers were the proposed ensemble and Bagged J48 which were sihnificantly better than the other algorithms on this dataset. Bayesian Networks and the proposed Ensemble were the best performers in terms of AUCPRC.

4.5 Performance of Classifiers after Over Sampling

The following charts shows the performance of the classifiers on the datasets after the majority class has been oversampled to create an imbalance of the two classes. The distributions were changed from their defaults to 70/30, 80/20 and 90/10 split between the non-default and default classes respectively.

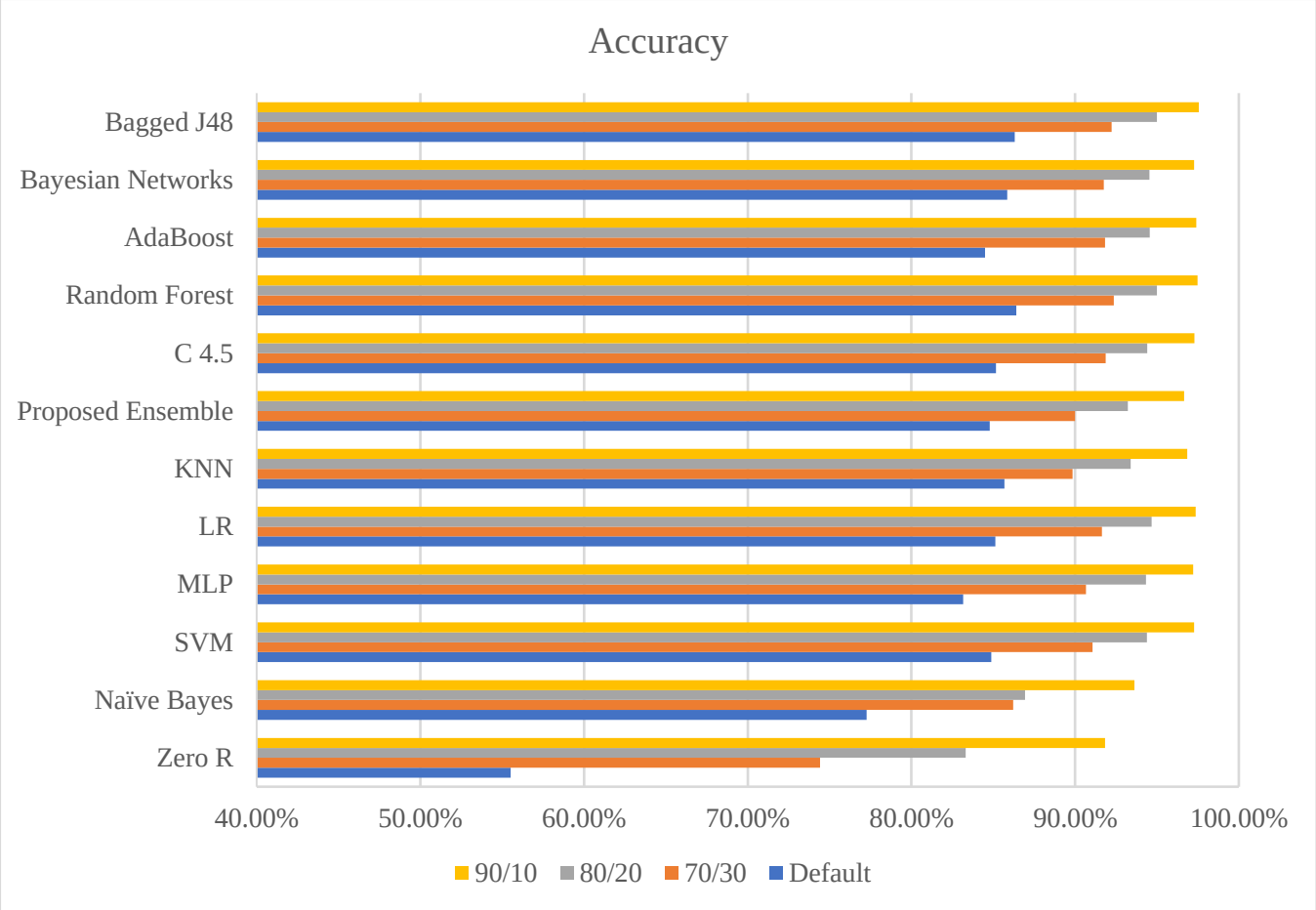

Figure 4.3 Accuracy on the Australian Dataset

The figure above shows the changes in performance of the classifiers on the different distributions of the Australian dataset. Zero R algorithm serves as the baseline and is used to show the minimum level of expected performance. The trend immediately visible from the chart is that the accuracy of the algorithms improves as the dataset becomes more skewed towards one class, this is expected as algorithms tend to fit one particular class and imbalance makes this extremely likely as discovered from literature. Of all the algorithms, the best three performers on average were; Random Forest, Bagged J48 and Bayesian Networks with scores of 92.81%, 92.78% and 92.36% respectively. The proposed Ensemble had an average accuracy of 91.16% which is significantly not different from the best performers.

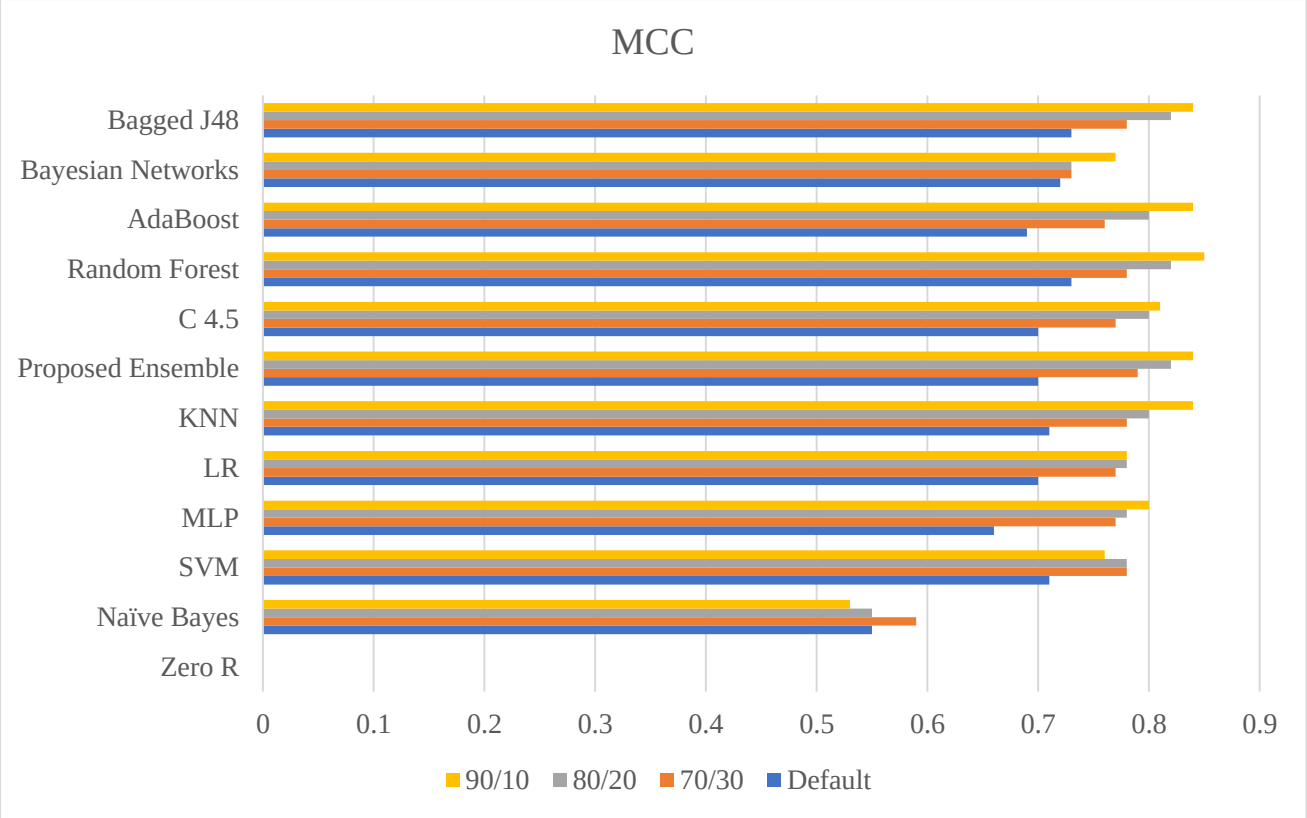

Figure 4.4 MCC on the Australian Dataset

With respect to the MCC of the algorithms on the different variants of the Australian dataset, most of the algorithms had a slightly better performance as the dataset became more imbalanced, with the exception of LR, SVM and Naïve Bayes whose performance became worse or plateaued. From the chart, the best three algorithms were: the proposed ensemble, Random forest and Bagged J48, which had average values of 0.8, 0.79, 0.79 respectively. Zero R has a zero score because it is a random guess.

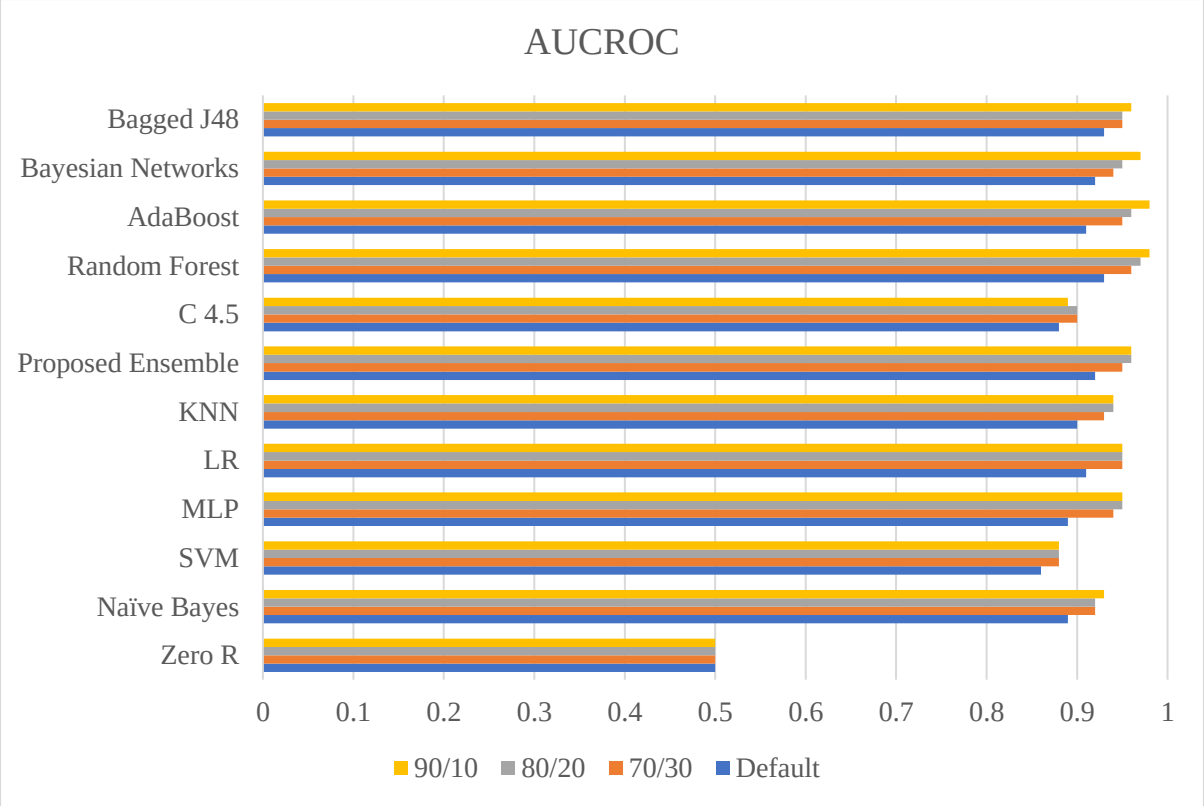

Figure 4.5 Area under the ROC Curve for the Australian Dataset

The chart above shows the performance of the algorithms with respect to AUC of the ROC Curve, most of the algorithms exhibited good performance with values close to 1 which is the score for a perfect classifier. Although, there is an increase in the performance of the algorithms as the data set becomes more skewed, the improvement is little and is caused by the increase in the number of instances. AUC of the ROC is a metric that measures amongst other things the scalability of an algorithm. The best three performers observed were: Bagged J48, Random Forest and the proposed Ensemble which means these are the most scalable algorithms on this dataset.

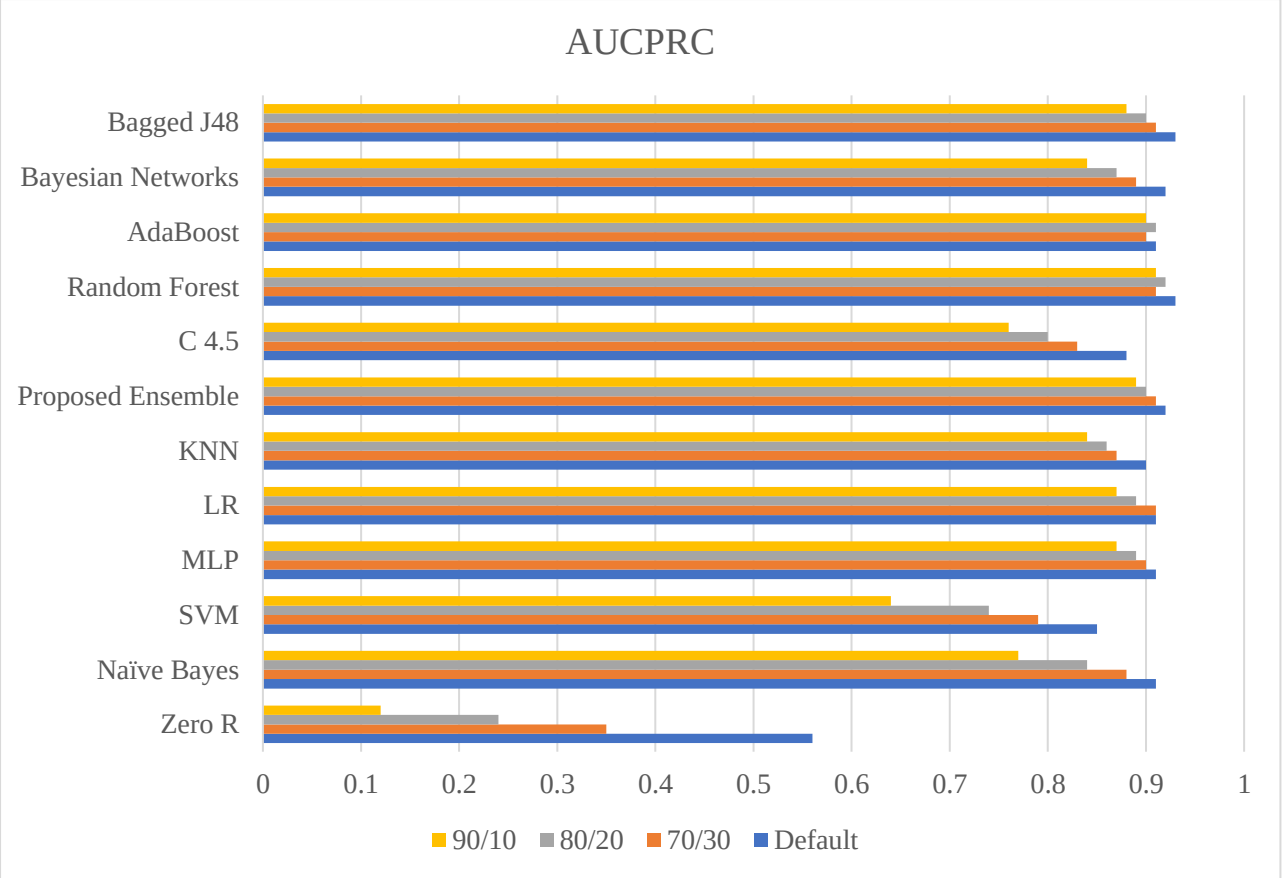

Figure 4.6 Area under the Precision Recall Curve for the Australian Dataset

AUC of the PRC is a good metric for determining performance of algorithms on an imbalanced dataset, as it evaluates the performance of the algorithms with relation to the class imbalance that exists in the dataset. From the chart most algorithms suffer some dip in performance as the majority class outweighs the minority because it typically becomes more difficult for the algorithms to optimally classify both classes. It was also observed that the fall in performance was worse for most of the single classifiers as compared to the hybrids or ensembles. The best three classifiers were: Random Forest, Ada Boost, the proposed Ensemble in that order.

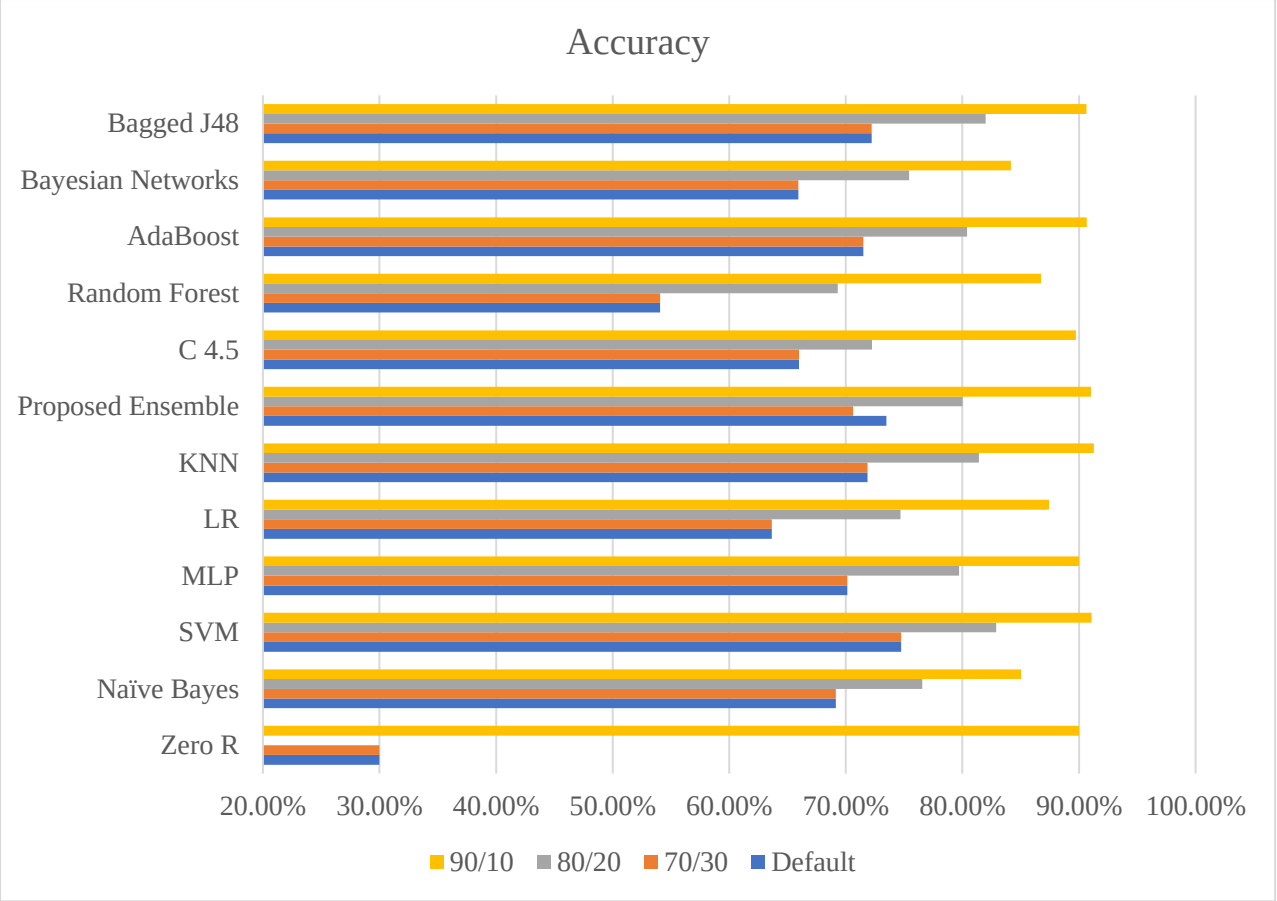

Figure 4.7 Accuracy on the German Dataset

The German dataset contains a cost matrix, which is essentially a re-weighting system that compels the algorithms to focus on the minority class. The cost matrix for this dataset assigns an error cost of 5 for every misclassified defaulter instance, but it was observed that there were spikes in performance especially on the 90/10 distribution of the dataset, this is because the cost matrix weighting system has been overrun by the sheer number of instances in the majority (non- defaulter class). Zero R for example reached an accuracy of 90% on the 90/10 split, by guessing the majority class, the best three classifiers were: SVM, KNN and Bagged J48. The Proposed Ensemble had an average accuracy of 80.55%

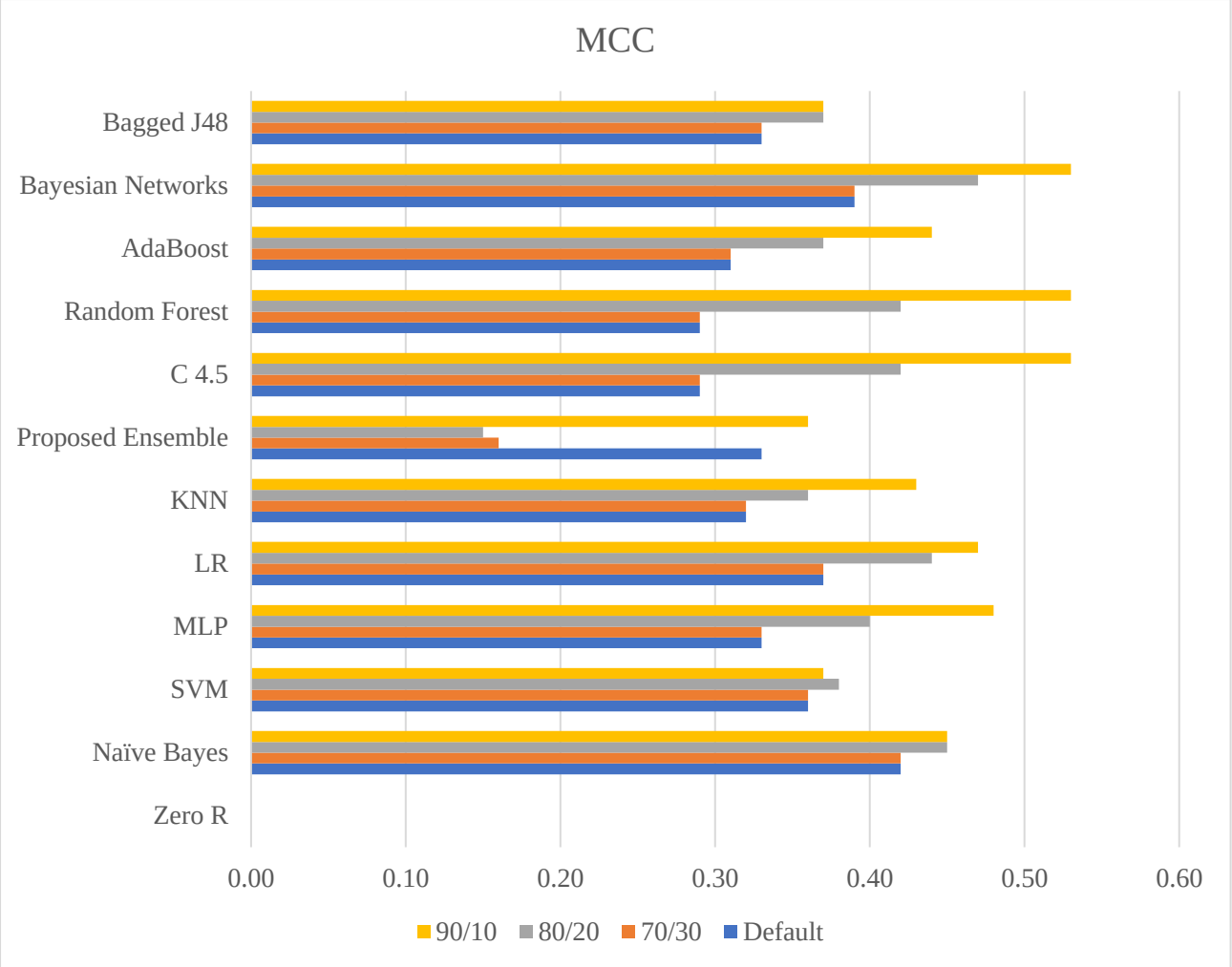

Figure 4.8 MCC on the German Dataset

In terms of MCC, the best machine learning algorithms were Bayesian Network, Naïve Bayes and Logistic Regression. Most of the algorithms had a reducing or plateauing performance as the data set became more imbalanced, showing their inability to properly classify the minority class. MCC measures how well an algorithm makes the trade-off between completeness and exactness, a random guessing algorithm such Zero R has an MCC of 0.

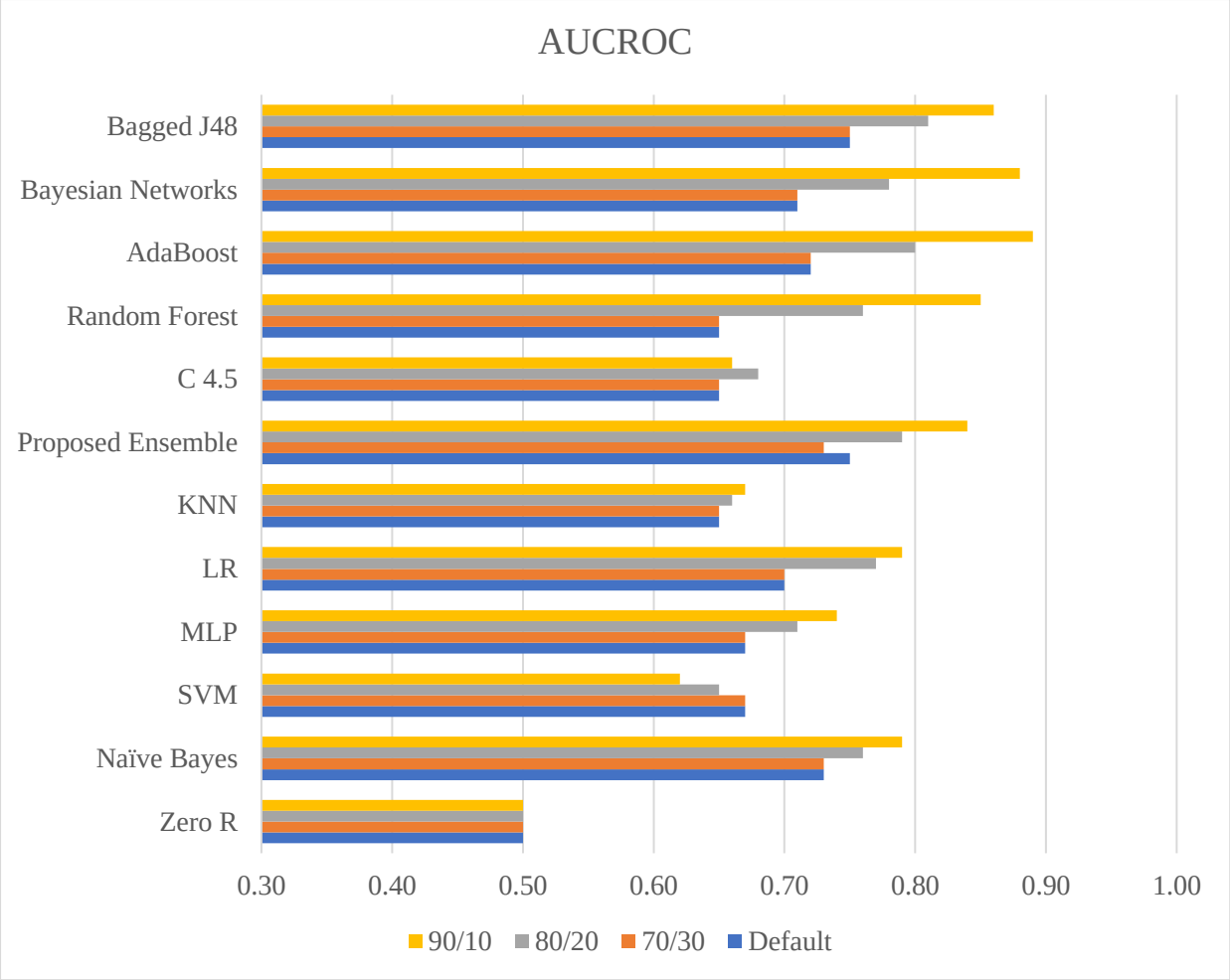

Figure 4.9 Area under the Receiver Operating Characteristic Curve for the German Dataset

AUC of the ROC Curve is a metric that describes the ability of classifiers to scale properly, if the dataset distribution remains slightly even. As seen from the figure above, the best performers were the hybrid classifiers because they are more robust and immune to outliers. The best three algorithms on this dataset in terms of AUC of the ROC Curve are the proposed ensemble, Bagged J48 and AdaBoost, and their performance improves with the skewed performance, unlike the single classifiers which suffer some lose in performance as the data set becomes more imbalanced.

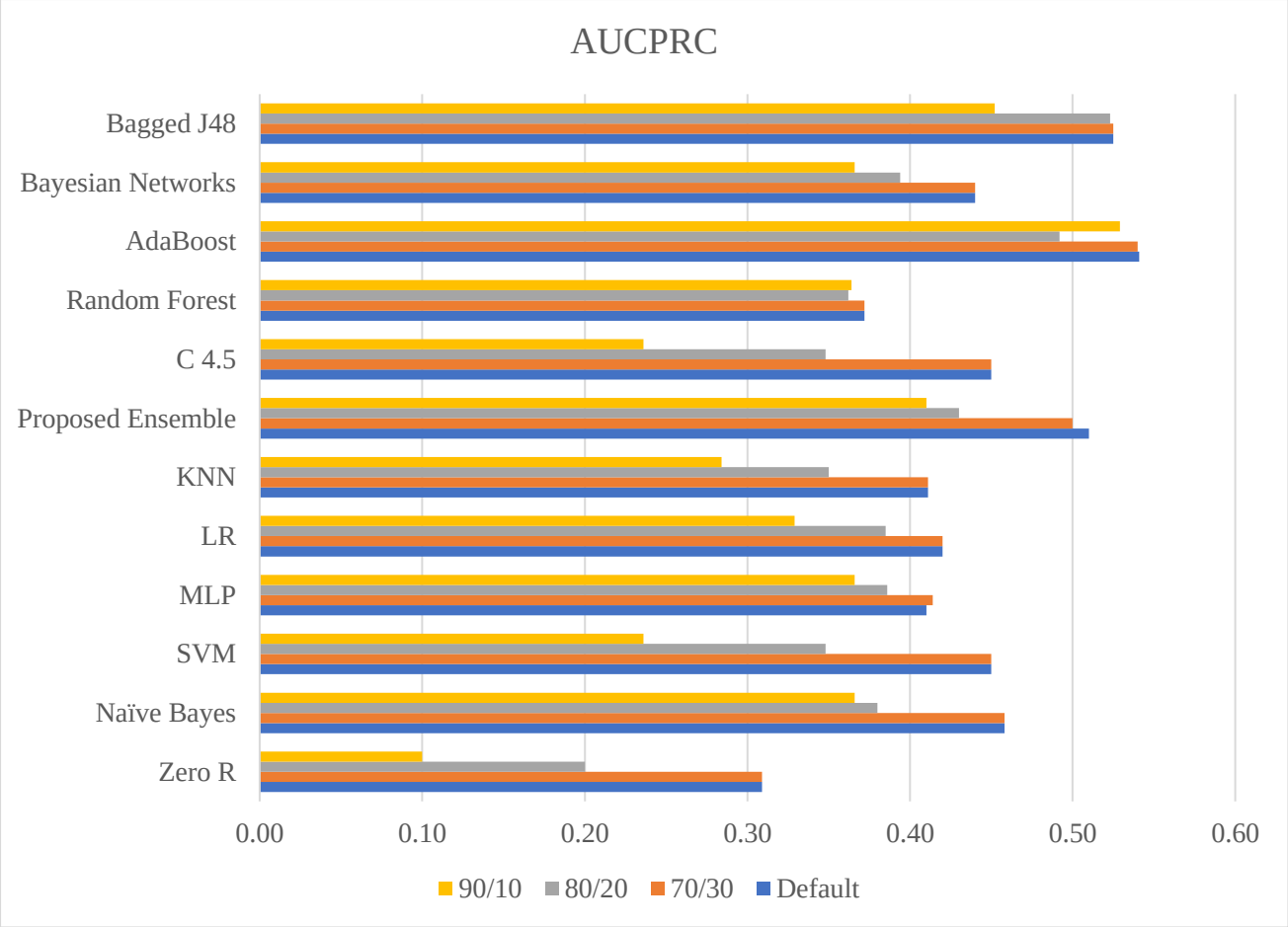

Figure 4.10: Area under the Precision Recall Curve for German Dataset

From the figure above, the trend with most of the algorithms is that the performance of the algorithms get worse as the dataset becomes more imbalanced. The AUCPRC measures the trade off between recall and precision of the algorithms, a metric that is not affected by class imbalance. The best three classifiers were the proposed ensemble, AdaBoost and Bagged J48 with values of 0.53, 0.51 and 0.46 on this dataset.

4.6 Discussion of Findings

From the conducted research several inferences were drawn, a few of which support theories and observations in literature, the inferences were:

i. Hybrids and Ensembles generally perform better than single classifiers: from the tables and charts, the results of the hybrids algorithms and the ensembles are better than the single classifiers using metrics such as AUC of the ROC curve, MCC and AUC of the PRC, which are the most important metrics in this problem domain.

ii. Feature selection is useful for improving the efficiency of algorithms; feature selection which reduces the dimensionality of a dataset, generally improved the computational

efficiency of the algorithms without significantly affecting performance, this is observed on most of the algorithms, although few improved significantly with feature selection.

iii. Better data improves algorithms; with parameter tuning there was some improvement in the performance of the algorithms on the datasets, but for marked improvements then better rather than more data is required. Better data refers to data that is properly pre-processed and does not contain redundant attributes and outliers

Other inferences drawn from the recorded results suggest that the proposed ensemble was significantly better than the algorithms it was compared against across the two datasets, it terms of scalability, and efficiency. The research also suggests that Random Forest and AdaBoost are good algorithms for the datasets used and able to scale and retaining practicality, they are also interpretable like the proposed ensemble.

From the study, the significance of the metrics used for evaluating performance of classifiers differed from one algorithm to the other. For instance, the improvement in the performance of the algorithms with respect to accuracy on both datasets is a misleading positive because, simple algorithms which do no learning such as Zero R have high values of accuracy simply because it predicts that all instances are non-defaulters, which will lead to it being correct approximately 90% of the time on the 90/10 split. In practical terms if deployed, such an algorithm will categorize all customers as non-defaulters which will lead to catastrophic loses for credit granting bodies. This observation suggests that although accuracy is the most reported metric of performance, accuracy is not the most objective representation of performance, and cannot be considered relevant in imbalanced datasets or certain problem domains.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECCOMENDATION

5.1 Summary

In this dissertation, the domain of credit scoring, problems and solutions posited and in use were discussed. Low default portfolio is a type of class imbalance problem that occurs in credit scoring datasets, which formed the basis of this research. Solutions have been put forward to solve this problem but the solutions are complex which stifles their acceptability for practical use. An ensemble of machine learning techniques is designed in this dissertation which makes a good trade-off between complexity and efficiency of the ensemble.

This ensemble in this work was configured with WEKA (Waikato Environment for Knowledge Analysis) and was tested on two datasets. Feature selection was done on the dataset to reduce the dimensionality and optimize the dataset, after which the datasets where imbalanced using SMOTE (Synthetic Minority Oversampling Technique), and the performance of the proposed ensemble, single classifiers and hybrids from literature were recorded over the various distributions of the dataset. The proposed ensemble was found to have the best performance across both datasets, closely followed by Random Forest and AdaBoost which are types of hybrids that use C4.5 tree. Their performance was evaluated with metrics such as Accuracy, Area under the Receiving Operating Characteristic Curve (AUCROC), Mathew's Correlation Coefficient (MCC) and Area under the Precision Recall Curve (AUCPRC), although some metrics have greater significance than others in the credit scoring domain.

5.2 Conclusion

In conclusion, credit scores in large economies have new interesting uses such as job interviews, apartment and rent negotiations as such optimising every facet of the industry has become important as customer scores now have impact in other areas of an individual's life. In addition, since credit scores are tied to global businesses the ability for any firm to fully control and manage expected returns from issuing credit is vitally important.

Low default Portfolio is a credit scoring problem that has been solved with complex algorithms that cannot be easily deployed, hence this work puts forward an ensemble which deals with this problem without increasing complexity and achieves good results on datasets it was tested on. The metrics used for evaluation describe different behaviour that the ensemble exhibits with changes in the distribution of the datasets, scalability is measured using the AUCROC,

stability is measured using the AUCPRC, and ability to deal with imbalance is measured with the MCC.

5.3 Recommendations for Further Studies

The following are recommendations to further refine the solution put forward in this work and other path of research to broaden the horizon on knowledge in this problem domain.

i. The dataset used for this dissertation contained static information, in practical environments, the datasets more dynamic as attribute values change with time hence, a dataset that captures this dynamism is require to fully test these algorithms on.

ii. There is unavailability of datasets sets due to ethical constraints and the competitive nature of banks and other loan granting bodies. This particularly affects research in behavioural credit scoring which requires personal day to day data concerning the customer to train machine learning algorithms. The ECOA act for example, which prohibits the use of certain demographics such as race, age, sex for credit scoring which potentially limits the ability of machine learning algorithms. This is especially important to thin file customers, that is, those customers with no previous credit or financial history. Better data that describes the function to be model is required to reduce error rates to the barest minimum.

iii. This work was limited to designing and evaluating the ensemble model, thus the scope can be enlarged to fully implementing the model using a language such as python which can be used for machine learning purposes and for web development.

iv. The scope of this work can also be broadened to handle big data streams and behavioural credit scoring for new customers which are problems that still exist in this domain but were not addressed in this work.

v. It is recommended that in further studies a new evaluation system is developed which weights the importance of the evaluation metrics used in this domain. Accuracy which is the most reported metric of performance is usually misleading and does not account for other factors existing in the dataset or experimental setup such as bias, imbalance, overfitting and algorithm completeness.

References

Alpaydın, E. (2010). *Introduction to machine learning. 2nd Edition.* London: The MIT Press.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankcruptcy. *Journal of Finance*, 589-609.

Australia, E. (2016, 02 25). *Equifax completes acquisition of australia's leading credit information company, veda group limited, for total consideration of USD$1.9 Billion*. Retrieved from Equifax Australia: (https://www.equifax.com.au/news-media/equifax-completes-acquisition-australias-leading-credit-information-company-veda-group

Baker, B. (2015). *Consumer credit risk modeling.* Cambridge.: MIT.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association, 39 (227)*, 357-365.

Bhatia, S., Sharma, P., Burman, R., Hazari, S., & Hande, R. (2017). Credit scoring using machine learning techniques. *International Journal of Computer Applications*, 1-4.

Bishop, C. M. (1995). *Neural networks for pattern recognition.* Oxford: Oxford University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 123-140.

Breslow, L. A., & Aha, D. W. (2000). Simplifying decision trees: A survey. *Semantic Scholar*, 1-47.

Bronshtein, A. (2017, April 11). *A quick introduction to k nearest neighbours algorithm*. Retrieved from Medium: https://medium.com/@adi.bronshtein/a-quick-introduction-to-k-nearest-neighbors-algorithm-62214cea29c7

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Journal of Information Fusion, 6(1)*, 1-28.

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

Carr, J. (2014). An Introduction to genetic algorithms. 1-40.

Carroll, P., & Rehmani, S. (2016). *Alternative data and the unbanked.* Oliver Wyman.

Chang, C.-H., Liou, W.-C., Hsu, W. W., Shih, J.-Y., Wu, C.-S., & Ho., J.-M. (2017). An innovative framework for building agency-free credit rating systems. *IEEE 10th International Conference on Service-Oriented Computing and Applications.* (pp. 181-189). Hong Kong: IEEE Computer Society.

Chen, C.-C., & Li, S.-T. (2014). Credit rating with a monotonicity-constrained support vector machine model. *Expert Systems with Application*, 7235–7247.

Chen, M., Dautais, Y., Huang, L., & Ge., J. (2017). Data Driven Credit Risk Management Process: A Machine Learning Approach. *International Conference on Software and Systems Process.* (pp. 109-113). Paris: Association for Computing Machinery.

Chen, N., Ribeiro, B., & Chen, A. (2015). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 1-23.

Chen, N., Ribeiro, B., & Chen., A. (2015). Financial credit risk assessment: a recent review. *Artificial Intelligence Review*, 1-23.

Daumé, H. (2013). Decision Trees. In H. D. III, *A Course in Machine Learning* (pp. 8-18). Self Published.

Daumé, H. (2013). Geometry and nearest nieghbours. In H. D. III, *A Course in Machine Learning* (pp. 29-40). Self Published.

Decision tree: Introduction. (2015, July 4). Retrieved from Tree Plan: http://treeplan.com/chapters/introduction-to-decision-trees.pdf

Delen, D., Kuzey, C., & Uyar, A. (2013). Measuring firm performance using financial ratios: a decision tree approach. *Expert Systems with Application*, 3970-3983.

Dictionary, B. (2017, December 21). *Credit*. Retrieved from Business Dictionary: http://www.businessdictionary.com/definition/credit.html

Dictionary, B. (2018, April 23). *Credit Scoring*. Retrieved from Business Dictionary: http://www.businessdictionary.com/definition/credit-scoring.html

Ding, S., Ma, Y., & Zhou, H. (2018). Implementation of dynamic credit rating method based on clustering and classification technology. *Cluster Computing.*, 1-11.

Drummond, C., & Holte, R. (2005). Severe class imbalance: Why better algorithms aren't the answer. *Machine Learning: ECML 2005* , 539-546.

Du, K., & Swamy, M. (2014). Fundamentals of machine learning. In K. Du, & M. Swamy, *Neural Networks and Statistical Learning.* (pp. 15-65). London: Springer-Verlag.

Durand, D. (1941). *Risk elements in consumer instalment financing.* New York: National Bureau of Economic Research.

Experian. (2018, April 12). *Score Basics.* Retrieved from Experian: https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/

FICO. (2017, December 22). *History*. Retrieved from FICO: http://www.fico.com/en/about-us/history/

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*, 179-188.

FitzPatrick, P. (1932). A comparison of the ratios of successful industrial enterprises with those of failed companies. *Journal of Accounting Research*, 598-605.

Freund, Y., & Schapire, R. (1996). Experiments with a new boosting algorithm. *Proceedings of the13th International Conference on Machine Learning* (pp. 148-156). Bari: Machine Learning.

Hand, D., & Henley, W. (1997). Statistical classification methods in consumer credit scoring: A review. *Royal Statistics Society*, 523-541.

Hand, D., & Zhou, F. (2009). Evaluating models for classifying customers in retail banking collections. *Journal of the Operational Research Society, 61*, 1540-1547.

Henley, W. E., & Hand, D. J. (1997). Construction of a k-nearest neighbour credit scoring system. *IMA Journal of Management Mathematics*, 305-321.

Huang, J., & Chen., M. (2018). Domain adaptation approach for credit risk analysis. *ICSIM2018* (pp. 1-4). Casablanca: Association for Computing Machinery.

Investopedia. (2018, April 1). *Credit*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/credit.asp

Investopedia. (2018, April 3). *Credit Score*. Retrieved from Investopedia: https://www.investopedia.com/terms/c/credit_score.asp

Japkowicz, N. (2000). The class imbalance problem: Significance and strategies. *International Conference on Artificial Intelligence.* (pp. 111-117). Chicago: ICAI.

Kennedy, K. (2013, Feburary). *CGI.* Retrieved from ARROW@DIT: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&cad=rja&uact=8&ved=0ahUKEwiriNnjxc7aAhUIOxQKHb3BANwQFggxMAE&url=https%3A%2F%2Farrow.dit.ie%2Fcgi%2Fviewcontent.cgi%3Farticle%3D1138%26context%3Dsciendoc&usg=AOvVaw1kSw4Y-PgAf9a_0zSLXhwW

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, 11-23.

Koutanaei, F. N., Sajedi, H., & Khanbabaei, M. (2015). A Hybrid Data Mining Model of Feature Selection Algorithms and Ensemble learning Classifiers for Credit scoring. *Journal of Retailing and Consumer Services.*, 11-23.

Krenker, A., Bešter, J., & Kos, A. (2011). Introduction to the artificial neural networks. In P. K. Suzuki, *Artificial Neural Networks - Methodological Advances and Biomedical Applications* (pp. 1-18). Shanghai: InTech.

Kuncheva, L. I., & Whitaker., C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 181–207.

Lanzarini, L. C., Monte, A. V., Bariviera, A. F., & Santana., P. J. (2016). *Simplifying credit scoring rules With LVQ+PSO.* University of Brazil.

Lin, W.-C., Shih-Wen, K., & Tsai, C.-F. (2017). Top 10 data mining Techniques in business applications: A brief survey. *Kybernetes*, 1-21.

Lipton, Z. C. (2017). The mythos of model interpretability. *ICML Workshop on Human Interpretability in Machine Learning.(WHI2016)* (pp. 1-9). New York: arXiv.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: A systematic review and overall comparison. *arXiv*, 1-50.

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to Credit scoring: A Systematic Review and overall comparison. *Elsevier*, 1-50.

McBride, C. (2018, April 13). *What are three types of consumer credit?* Retrieved from Capital Chron: http://smallbusiness.chron.com/three-types-consumer-credit-1886.html

Melville, P., & Mooney, R. J. (2001.). Constructing diverse classifier ensembles using artificial training examples. *Journal of Learning.*, 505-510.

Mitchell, T. (1997). *Machine learning.* Chicago: McGraw Hill.

Moffitt, K. (2018, April 28). *Research methodology: Approches and techniques.* Retrieved from Study.com: https://study.com/academy/lesson/research-methodology-approaches-techniques-quiz.html

Oreski, S., & Oreski, G. (2014). Genetic algorithm-based heuristic for feature selection. *Expert Systems with Applications*, 2052-2064.

Ozturk, H., Namli, E., & Erdal, H. I. (2016). Reducing overreliance on sovereign credit ratings: Which model serves better? *Computing Economics*, 59-81.

Patil, P. S., Aghav, J. V., & Sareen, V. (2016). An overview of classification algorithms and ensemble methods in personal credit scoring. *International Journal of Computer Science And Technology*, 183-188.

Quilan. (1987). Simplifying decison trees. *International Journal of Machine Studies #27*, 221-234.

Quinlan, J. R. (1993). *C4.5 programs for machine learning.* San Mateo, CA: Morgan Kaufmann.

Ramík, J. (2017, March 3). *Soft computing: Overview and recent developments in fuzzy optimization.* Retrieved from http://irafm.osu.cz/research_report/118_softco01.pdf

Reginiel, P. A. (2015, January 5). *Conceptual framework: A Step by Step guide on how to make one.* Retrieved from Simply Eduacate ME: https://simplyeducate.me/2015/01/05/conceptual-framework-guide/

Sadatrasoul, S., Gholamian, M., & Shahanaghi, K. (2015). Combination of feature selection and optimized fuzzy apriori rules: The case of credit scoring. *International Arab Journal of Information Technology*, 138-145.

Salappa, A., Doumpos, M., & Zopounidis, C. (2007). Feature selection algorithms in classification problems:an experimental evaluation. *Optimal Methods in Software*, 199-212.

Score, P. a. (2018, April 23). *Types of scoring*. Retrieved from Plug and Score: https://plug-n-score.com/learning/types-of-scoring.htm

Smola, A., & Vishwanathan, S. (2008). *Introduction to machine learning.* Cambridge: Cambridge University Press.

Stecanella, B. (2017, June 22). *An introduction to support vector machines*. Retrieved from Monkey Learn: https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

Thomas, L. (2000). A Survey of credit and behavioural scoring: Forecasting financial risk of lending to customers. *International Journal Forecast*, 149-172.

Trainor, S. (2015, July 22). *The Long, twisted history of your credit score*. Retrieved from TIME: http://time.com/3961676/history-credit-scores/

Tsai, C.-F., Hsu, Y.-F., & Yen, D. (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing Journal* , 977-984.

Turkson, R. E., Baagyere, E. Y., & Wenya., G. E. (2016). A Machine learning approach for predicting bank credit worthiness. *IEEE*, 81-87.

Wang, C., & Huang, Y. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications*, 5900-5908.

Wang, H., Zhong, J., Zhang, D., & Zou., X. (2017.). A new classification algorithm for the bank customer credit rating. *Ninth International Conference on Advances Computational Intelligence (ICACI)* (pp. 143-148). Doha: IEEE.

Wang, J., Hedar, A., Wang, S., & Ma, J. (2012). Rough set and scatter search meta-. *Expert Systems with Applications.*, 6123-6128.

Wang, X., Xu, M., & Pusatli, Ö. T. (2015). A Survey of applying machine learning techniques for credit rating: Existing models and open Issues. *ICONIP 2015, Part II* (pp. 122-132). Springer International Publishing Switzerland.

Wang, Y., Wang, S., & Lai, K. K. (2005). A new fuzzy support vector machine to evaluate credit risk. *IEEE Transactions on Fuzzy Systems, Vol. 13, No. 6.*, 820-831.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks, 5*(2), 241-259.

Zachary Chase, L. (2015, April 29). *The myth of model interpretability.* Retrieved from KDnuggets.: https://www.kdnuggets.com/2015/04/model-interpretability-neural-networks-deep-learning.html

Zadeh, L. A. (1994). Soft computing and fuzzy logic. *IEEE Software*, 48-56.

Zhang, Z. (2016). Introduction to machine learning. *Ann Transl Med*, 4-11.

Zhang, Z., Gao, G., & Shi, Y. (2014). Credit risk evaluation using multi-criteria optimization classifiers, kernel, fuzzification and penalty factors. *European Journal of Operational Research*, 335-348.

Zhou, Z.-H. (2012). *Ensemble methods foundations and algorithms.* London: CRC Press.