2019

Consumer credit risk modelling using machine learning algorithms: a comparative approach

Nyangena, Brian Okemwa
Strathmore Institute of Mathematical Sciences (SIMs)
Strathmore University

Consumer Credit Risk Modelling Using Machine Learning Algorithms: A Comparative Approach

Nyangena Brian Okemwa

Submitted in partial fulfillment of the requirements for the Degree of Masters of Mathematical Finance at Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June, 2019

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the dissertation contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Brian Okemwa Nyangena

Signature:.....

Date:.....

Approval

This dissertation of Brian Okemwa Nyangena was reviewed and approved by the following :

Dr. Lucy Muthoni
Lecturer, Strathmore Institute of Mathematical Science.
Strathmore University

Mr. Ferdinand Othieno
Dean, Strathmore Institute of Mathematical Science.
Strathmore University

Prof. Ruth Kiraka
Dean, School of Graduate Studies.
Strathmore University

List of Abbreviations

AUC Area Under the Curve

CAMELS Capital Adequacy, Asset Quality, Management, Earnings, Liquidity, and Sensitivity

CBK Central Bank of Kenya

FSD Financial Sector Deepening

IFRS International Financial Reporting Standards

kNN k-Nearest Neighbour

PD Probability of default

PR Precision-Recall

ROC Receiver Operating Characteristic

SMOTE Synthetic Minority Over-Sampling Technique

SVM Support Vector Machine

Abstract

Consumer credit risk scoring involves the assessment of the risk that is associated with a customer that applies for credit. The ability to confidently identify customers that will repay the credit and those that will not is therefore, an important aspect for any institution. The purpose of this study is to model consumer credit risk using machine learning models and compare the results to the traditional logistic model. The aim is to identify whether there is improved performance in the classification of default among customers when machine learning algorithms are used. Additionally, the study aims to identify how different customer characteristics affects their default experience.

The data used was obtained from Kenya Metropol between 2014-2017 and had customer details such as age, loan amount, marital status and sex among others, during this period. 5 models are used to model the default experience namely: Logistic regression, Random Forest, Support Vector Machine, Gradient Boosting and Multi-layer Perceptron Neural Network. The efficiency of the models was assessed using the following metrics; Accuracy, Precision, Recall, F1-score and Precision-Recall curve. Due to the imbalanced nature of credit data set, the F1-score, which is a weighted average of the Precision and Recall, was eventually used as the metric to determine the best model for credit scoring.

The findings showed that Random Forest performed the best, having an F1-score of 0.307. The machine learning algorithms outperformed the logistic model and showed an improved performance in the classification of default, especially in identifying false positives. It was also established that male customers had a higher default probability, younger customers were more likely to default and single customers defaulted more than married customers.

Contents

List of Figures vii

List of Tables viii

Acknowledgement ix

Dedication x

1 Introduction **1**

1.1 Background To The Study 1

1.1.1 Credit Risk 1

1.1.2 Machine Learning and Credit Risk Modelling 2

1.2 Problem Statement 3

1.3 Objectives of the study 4

1.4 Research Question 5

1.5 Significance of the Study 5

2 Literature Review **6**

2.1 Introduction 6

2.1.1 The Current Credit Risk Management Framework 6

2.2 Empirical Literature 7

3 Research Methodology **11**

3.1 Introduction . . . 11

3.2 Study Design . . . 11

3.3 Traditional Models . . . 12

3.3.1 Logistic Regression . . . 12

3.4 Machine Learning Models . . . 13

3.4.1 Neural Networks(Multi-Layer perceptron) . . . 13

3.4.2 Support Vector Machine (SVM) Model . . . 14

3.4.3 Gradient Boosting . . . 16

3.4.4 Random Forest . . . 16

3.5 Performance Measures . . . 18

3.5.1 Confusion Matrix . . . 18

3.5.2 Accuracy . . . 19

3.5.3 Precision, Recall and F- measure (F1 score) . . . 19

3.5.4 Receiver Operating Characteristic Curve (ROC), Precision-Recall Curve and AUC . . . 20

4 Data Analysis and Discussion **22**

4.1 Data Overview . . . 22

4.2 Findings . . . 24

4.2.1 Feature Analysis . . . 24

4.2.2 Model Analysis and Discussion . . . 25

5 Conclusion and Recommendations **30**

5.1 Conclusion 30

5.2 Limitations 31

5.3 Recommendations 32

References **33**

Appendix 36

List of Figures

3.1 ROC curve for different threshold values between 0 and 1. 21

3.2 Precision-recall curve for various thresholds between 0 and 1. 21

4.1 PR Curves 27

4.2 Precision-recall curve for the random forest model after optimisation. 28

4.3 Precision-recall curve for the random forest model after optimisation. 29

List of Tables

3.1 Confusion Matrix describing the performance of classification models 19

4.1 Summary of the performance evaluation results of the models 25

Acknowledgement

First, I would like to thank the Almighty God for the good health and strength to see me through my studies.

Secondly, a special thanks to my supervisor Dr. Lucy Muthoni for her continuous support and guidance. Lastly, to my classmates, friends and my family, thank you for always being there for me and for your support. God bless you all.

Dedication

This is for my mother, Josephine Nduku Ndivo for believing in me and for being my pillar of strength through my whole education. I also dedicate this to my wife, Stella, for being by my side through this journey. Lastly, I'd like to thank my friends and family for always cheering me on. God bless you all.

Chapter 1

Introduction

1.1 Background To The Study

1.1.1 Credit Risk

Credit risk is one of the most important and traditional risk known in the financial markets. It is also known as counter party risk and is defined as the likelihood of a borrower not being able to honor their debt obligations under specific agreed terms as stipulated by the lending institution (Klein, 1992).Duffie, Pan, and Singleton (2000) also described it as the risk of default or that resulting from reduction in the market value which is caused by fluctuations in the quality of the counter-party or issuer.

This makes credit risk modelling an important field in financial risk management. It has also attracted this much attention following the financial crisis and the recent regulatory concerns of Basel II extended by Basel III. This has led to the prevailing climate of very tight credit and increased interest by banks to reduce credit risk losses. Additionally, in the Kenyan economy, there has been a recent surge of peer to peer lending and microfinance institutions where credit risk management is of utmost importance to ensure profitability in the competitive scene. The ability of an institution to distinguish between good and bad customers is crucial for its success and performance over a period of time.

The problem of credit risk has been a real puzzle for researchers as it stems from a backdrop of asymmetric information that then causes adverse selection and moral hazard issues. The

paper by Casu et al. (2006) acknowledges the fact that all transactions and contracts are based on information and during financial intermediation. A number of problems can arise such as the issue of all participants being not perfectly informed or certain transactions having more information which may not be available to both parties. This creates an asymmetric flow of information that make financial agreements difficult to enter into which can lead to the inefficiency of intermediation. According to Andries (2008), models exhibiting adverse selection are characteristic of one side not having the information while performing the transaction while in moral hazard models, the issue arises after the transactions, such as the lenders not being able to observe the borrower's actions, which affect their probability of default.

1.1.2 Machine Learning and Credit Risk Modelling

To avoid these issues, credit institutions sometimes practice credit scoring and credit rationing to ensure these risks are minimized and their ability to maximize profits increased. Credit scoring models have become a norm in most lending institutions globally to reduce these risks associated with moral hazard and adverse selection. Credit scoring is the use of statistical models to transform a set of relevant data into a numeric transform capable of guiding credit decisions for a financial institution (Anderson, 2007). These models categorize the applicants as either good or bad through the use of characteristics such as income, age and marital status. Credit scoring helps in reducing the cost of credit by helping reduce the chances of default through assessing creditworthiness of a customer and in some instances detecting fraud and it additionally, can be able to monitor existing loan accounts and thus prioritize the collection of repayment. Today, almost all credit institutions conduct some form of credit scoring before disbursing a line of credit or loans to individuals and corporates alike.

Traditional statistical methods have been used to develop models for credit scoring such as Linear Discriminant Analysis (LDA) or linear regression through models such as the logistic, logit, probit or tobit models with logistic models being the most common models in use. These are mostly parametric models and recently, research has explored the use of non-parametric models such as Artificial Neural Networks (ANN), random forests and gradient boosting algorithms among others. Some studies have shown that these non-parametric models outperform the traditional models for example, West (2000) and Blanco, Pino-Mejías, Lara, and Rayo (2013). Other studies have shown the traditional models still having better results such as the study by Komarad (2009) who found Logistic regression had a better performance than ANN's.

In light of the Kenyan credit industry with the increase in microfinance Institutions and peer to peer lending institutions, there is need to conduct research into these non-parametric models that explore complex relationship between the factors affecting the probability of default. The industry stands to benefit from the use of non-parametric techniques since not a lot of literature exist tackling the Kenyan microfinance and peer to peer lending, let alone the banking credit industry. Wagacha and Othieno (2015) conducted a study in the Kenyan market to determine the transition probabilities to different credit status using semi Markov processes. This paper however, did not conduct a comparative study into the efficiency of the existing classical models used in determining probabilities and those suggested for the non-parametric relationships. The key concern of most lending institutions is the need to increase the accuracy of the scoring for credit decisions and an improvement, even in the smallest of fractions, can translate into a significant gain for the company in future. This therefore means that the choice of model is a key factor in determining the success of an institution and due attention needs to be accorded to new and better models that may help reshape how the credit industry has been operating over the years.

1.2 Problem Statement

With the passing of time, and advancements in technology, there is a lot that can now be done with the computing power of computers that was not possible in the earlier years. Some countries, especially the developed nations, have adapted quickly into using these models for their credit scoring. Khandani (2010) looked into the use of machine learning in assessing credit risk in a major commercial bank and was able to construct out of sample forecasts that improved the classification rates of defaults by credit card holders.

Additionally, many researchers and institutions have been taking innovative steps through microfinance institutions, peer to peer lending and social entrepreneurship which then creates a new area that needs exploring. This is in light of the variety of data available to us outside the traditional features used by banks such as social media data and the unstructured data collected from various forms such as mobile activity or geographical data among others. Vigano (1993) is considered as one of the first people who focused on the development of a model for credit scoring within the microfinance context. She considered borrower particular features to help give a complete picture of the borrower. Her research was then built on by Schreiner (2004)

who began a pilot program for developing a credit scoring model for the Burkina Faso and Bolivian microfinance market. These two papers are important since they were among the few papers that tackled the microfinance industry rather than the banking industry which was being covered a lot by the literature at the time.

The credit industry in Kenya has been growing over the past decade. Mobile lending has also been on the rise over the last decade and the scene was opened up by Safaricom with the introduction of Mpesa. This then led to the rapid growth of the digital loans market with recent numbers from the FSD Kenya (2018) report showing that 35% of Kenyans have borrowed from at least on digital lender. And of these borrowers, 47% repaid their loan late at one point while 12 defaulted on their loans. In the banking industry, non-performing loans have been on the rise and according to the Central Bank of Kenya (CBK) September 2018 report, the ratio of gross non-performing loans to gross loans went up to 12.52% in September 2018 from 11.97% in June, continuing the rising trend every quarter (CBK, 2018).

This is indicative of the fact that something has to change and this is probably due to the screening process undertaken by the banks before giving loans. The traditional models are being used to date and research has shown that there exist non-parametric models that can have a better performance because of taking into account the complex nonlinear relationships among the key variables determining default probability. Additionally, these models can be able to conduct feature selection process which can help determine which factors have the highest impact of default from thousands of possible factors.

This paper presents a comparison of the credit scoring models between the classical model commonly used, Logistic Regression, and the machine learning models namely Multi-layer Perceptron Neural Network model (MLP), Support vector Machine (SVM), Random Forests and Gradient Boosting Method. This is done in the context of the Kenyan Credit industry with an aim of ascertaining if machine learning models are better at predicting probability of default and if so, which is the most efficient.

1.3 Objectives of the study

The main objective of the study is to fit machine learning algorithms to a dataset from a Kenyan setting and identify whether machine learning algorithms are better at consumer credit

modelling compared to the commonly used logistic regression model.

1.4 Research Question

1. Which model is the most efficient at modelling credit risk using data set from Kenyan setting.

2. What is the default experience based on the various input variables for the Kenyan dataset?

1.5 Significance of the Study

The findings of this study focus at understanding and improving the way consumer credit risk modelling is done for the Kenyan market with an aim of adding to the literature in this field and also informing the various stakeholders. The credit market in Kenya is growing more so, on the microfinance and digital lending space. The findings of this study will help guide institutions reduce the cost of extending bad loans and the opportunity cost of denying credit to customers that would otherwise be profitable.

Banks on the other hand are much highly regulated and within the Basel II and Solvency II framework in addition to the new IFRS 9 regulation, they are required to have internal rating models to help in managing credit risk. With the increase in non-performing loans, it is clear that banks need a better credit scoring method to assess the customers and reduce defaults especially in the light of diminishing interest rate earnings on loans due to loss of market share to microfinance and digital lenders. This is where the findings of this research will prove significant because, a slight percentage improvement in the predictability of the models can translate to large sums of savings for the banks which will reflect as profits to them. The study will also form a basis for other research on the same area which will improve on some of the aspects that would not have been explored by this paper.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews some of the literature that has been done on the topic. It covers the history of credit risk models and the evolution into the new models that have emerged in covering this field. A survey of the existing theoretical and empirical literature will be assessed on the need for effective credit risk modelling.

2.1.1 The Current Credit Risk Management Framework

In 2010, the CBK pointed to the used of CAMEL ratings by banks and other financial institutions to analyze the financial soundness. Valle (2013) acknowledges the fact that prior studies, such as that by Fredrick (2012), CAMELS rating alongside other available data have the ability to predict the likelihood of bank failure. Furthermore, the paper by Allen et al. (2004) points to two approaches of credit risk modelling, the option structural approach that was introduced by Merton (1974) and the reduced form approach which utilises intensity based models in estimating stochastic hazard rates. They both argue that Reduced form models (RFM) focuses mainly on the accuracy of probability of default (PD) which is much better than using an intuitive economic approach.

The new Basel Capital Accord places the responsibility on banks to have sound internal credit risk management practices when assessing their capital adequacy (Wagacha & Othieno, 2015).

The CBK recommends banks must have a good enough assessment procedure to ensure they are able to asses the risk profile of the borrower and they advocate looking at the credit rating report obtained from any licensed credit bureau (CBK, 2013).Non-banking financial institutions are also in the lending business and as such, their risk also needs to be managed especially with the rise of mobile loans and digital lending. The growing use of the digital loans is due to the quick access of funds and the zero requirement for collateral and the use of alternative credit scoring models which use information on mobile money transactions when determining eligibility of a borrower. The use of such alternate data and data from the Credit Reference Bureau (CRB) can really fit well with the machine learning models which are robust in examining the relationships among variables. This increased competition means that the industry has become a margins' industry such that a small increase in efficiency in their processes will improve the chances of success. If they are to compete with banks, then the credit scoring mechanism must be efficient and automatic thus reducing the cost of credit analysis. Rhyne and Christen (1999) emphasizes the need of efficient credit scoring models and Schreiner (2004) affirmed that the study done in Bolivia and Columbia reduced costs by up to $75,000 per year in those countries. This is something that can easily translate to the Kenyan market as it is of similar characteristics with these markets. According to the FinAccess (2016) report, 27% of adults are borrowers of digital loans with the male/female ratio being 55/45%. This is clearly a big market and being able to asses the likely hood of of default will prove important to the success of the institutions.

2.2 Empirical Literature

Based on the study by Finlay (2011), one of the initial research into credit risk modelling or scoring using quantitative methods on consumer credit risk assessment was done by Durand (1941). He used a quadratic discriminant modelling method to categorize the applicants as either good or bad. Thereafter, Ohlson (1980) then published a paper that introduced the logistic regression modelling for credit risk analysis. The papers he borrowed from were White and Jarrow and Turnbull (1995) and additionally the paper by Santomero and Vinso (1977) also informed his study since they developed a way of estimating failure using probability.

Given the nature of credit risk scoring, there is a lack of adequate research papers documenting the performance of commercial consumer credit risk modelling. The body of research that exists currently focus on two key areas; the prediction of a company's insolvency and the

modelling of individual credit risk. The study by Altman and Bland (1994) uses both the linear discriminant analysis and neural networks to investigate close to a thousand Italian firms for corporate distress. This was among the first papers to look into corporate risk modelling. At the time, the authors came to the conclusion that neural networks were not a clearly superior technique when compared to the traditional models such as discriminant analysis and concluded that linear discriminant analysis is comparable to neural networks when it came to decision accuracy.

The study by Coats and Fant (1993) however, came to a different conclusion when comparing the two techniques when looking at distressed firms between the years 1970-1989. The findings indicate neural networks to be better at modelling as compared to linear discriminant analysis especially when looking at the financial distress of companies. The same data is then used in a later study by Lacher et al. (1995) and also came to a similar conclusion arguing that neural networks predict financial health of companies more accurately. Neural network comparison is also conducted by Desai et al. (1996) where the authors investigate the use of logistic regression, discriminant analysis and multi-layer perceptron neural networks using data from 3 Credit unions in the US. They work based on the assumption of a balanced dataset and they came to the conclusion that neural networks outperform but logistic regression wasn't far behind. However, the assumption of their study was not consistent with the kind of data usually obtained in the real world as it is usually heavily imbalanced. The performance of models and how well they do comparatively is very much dependent on the nature of the data being used in the modelling. Though ANN's and the logistic models use different notations, there are some few similarities such as ANN's use connection weight as a term while logistic used coefficient for the weights. The activation function is what determines the weights for ANN's but in the latter, the link function gives the coefficient.

It is noted that the better performance by ANN's is based on the fact that they have a strong learning ability and assumes nothing in regards to the relationships among the input factors. It also does not have a requirement on a priori because it has the ability to learn these relationships based on the data. In fact, this is the advantage of the machine learning models since they don't need a priori but rather learns the patterns on their own. This gives the models a higher robustness and flexibility in modelling misspecification. DeTienne and Chirico (2013) opines that machine learning models, more specifically ANN's, is a more appropriate model for prediction of classification problems rather than for the explanatory problems.

The use of decision trees, also known as decision learning, has gained traction over the years ever since it started gaining traction due to the paper by Mitchell (1997). They work based on the approach of evaluating each instance using a statistical test and then determining how well this separates a set of training dataset. This can be done either using the Gini Impurity of the information gain. The use of these decision trees alongside ANN's for credit card application was done in the study by Bagozzi et al. (1992) and this was on a single data partition. The authors found the two models to have a similar level of accuracy in their prediction. On the very same year, Jensen (1992) conducted a credit scoring just using neural networks and his study found a correct rate of classification of 76-80% which was accompanied by a 16% false positive rate and a 4% false negative rate and this was also done on a single partition of data with a test on 50 examples. The study by Addo et al. (2018) find that tree based models are more stable in prediction as compared to ANN's and logistic regression. They based the study on a data set that had as much as 181 labeled features that were obtained from the financial statements of companies for the year 2016/2017. They tried finding the most efficient of 7 models including random forests, gradient boosting, ANN's and logistic regression. They found that tree based algorithms outperformed the rest based on AUC and RMSE.

Random forest models were proposed by Breiman (2001)and then he later did another paper in 2004 where he built a predictor ensemble model with a number of decision trees which grow in a randomly selected subspace of the dataset used. Addo et al. (2018) based their research off improving what was done by Khandani (2010) who looked into a single bank and used CARTs to create a measure of consumer credit default and predicting the transition probabilities. Butaru et al. (2016) also conducted a similar study and looked into random forests, decision trees and logistic regression for 6 banks and found Random forest to perform marginally better compared to the decision trees.

Okay et al. (2008) looked at credit risk modelling using SVMs in classifying credit and then determining the default probability. The study compares the classification aspect by comparing the SVM with a logistic regression. The author suggests a cascade model which will be anchored in SVM to help in classification and then found a methodology of using SVM to classify credit risk. This data for the research was based of e-businesses in Turkey. Another study from Tunisia by Karaa and Krichene (2012) also looked at a comparison of ANN and SVM to determine whether companies can be predicted to be bankrupt or solvent. The data was on a Tunisian bank for the period 2002-2006 where they had extended credit to industrial

companies. The study found MLP neural network had a better performance than the SVM models with classification rates of 90.2% and 70.1% respectively. However, this is not the case based on other studies such as Shin et al. (2005) who also looked at SVM and compared them to neural networks of the back propagation form. The study found that the SVM had a better performance compared to the ANN when looking at corporate bankruptcy. Though this was observed to be the more accurate case when the training set size gets smaller.

Charpignon et al. (2014) from Stanford published their paper on credit risk modelling using a data set of around 100, 000 customers that was given in a Kaggle challenge. In their findings, they found that gradient boosting technique had a really good predictive accuracy with an AUC of around 85%. Cao and Yu (2018) also used the "Give me some Credit" Kaggle data set but used 8 models including gradient boosting, neural networks, MLP and SVM and assessed the models based on 3 factors; accuracy, AUC and the logistic loss. The study found that X gradient boosting model had the best performance compared to the rest. The study however did recommend using a bigger dataset so as to improve the accuracy of the predictions.

Chapter 3

Research Methodology

3.1 Introduction

This chapter looks into the models that will be used to analyse the Kenyan data set, the study design and then how these models will be evaluated for performance.

3.2 Study Design

This study aims to use machine learning models to contribute to the literature existing on credit risk modelling in Kenya. The study fits a logistic regression model alongside multi-layer perceptron neural network model (MLP), support vector machine (SVM), random forests and gradient boosting methods to a Kenyan data set to find the most effective learning at predicting credit default. For the implementation, the programming was done in Python for various tasks;

- Python 3.1 was used with the library Pandas performing the data preprocessing and Matplotlib and Seaborn were used for the visualisation
- For the machine learning models, the library Scikit-learn was used.

3.3 Traditional Models

3.3.1 Logistic Regression

Different models were historically used in the categorizing of input vectors based on two groups and this is the main agenda of statistical inference when dealing with credit scoring issues. Most of the variables in this study are categorical because of the particular characteristics of loan applicant data. Therefore, logistic regression is considered to be the most appropriate when dealing with such kind of data. Logistic regression is broadly used when it comes to the analysis of multivariate data that contains binary responses which is the case with this study. Thus, it offers powerful method that is equivalent to the multiple regression or ANOVA for continuous responses. The likelihood function for mutually independent variables $Y_1,\ldots,Y_n$ with binary scale outcomes is a member of the exponential family with $(\,log(\,\frac{\pi_1}{1-\pi_1})\,)$,...,$(\,log(\,\frac{\pi_n}{1-\pi_n})\,)$ as a canonical parameter, that is, π_j is a probability that Y_j becomes 1.

The logistic model is based on the assumption that there exist a linear relationship between a canonical parameter and the vector of explanatory variables $x_j$and that there exist no interaction between these variables while modelling as shown below;

$$log(\frac{\pi_j}{1-\pi_j}) = \mathbf{x}_j^\tau \beta \tag{3.1}$$

This linear relationship then then leads to a non linear relationship between the probability Y_j equals 1 and the vector of explanatory variables (Butaru et al., 2016):

$$\pi_j = exp\Big(\mathbf{x}_j^\tau \beta\Big) / \Big(1 + exp\Big(\mathbf{x}_j^\tau \beta\Big)\Big) \tag{3.2}$$

Linear regression models have the advantage of being quick and easy to train with their results being easy to interpret and make decisions on. This is why they are commonly used in the classification analysis since they give results that are easy to interpret and are also relatively effective. The main difference between a linear function and a logistic function is that the latter is ditchotomous. Once this is accounted for, the general assumption for both models are similar with there not being any constraints on the homoscedasticity and normality of the variables being used in the analysis (Abdon & Felipe, 2011).

3.4 Machine Learning Models

3.4.1 Neural Networks(Multi-Layer perceptron)

Neural networks (NN) are models that are based off the functionality if the human brain and they poses the flexibility to model any non-linear relationship between input variable and the response variable Bishop et al. (1995). There exist various architectures of the NN model however this study focuses on the the most commonly used one which is the Multi-layer perceptron (MLP). MLP is composed of an input layer that contains all the neurons for the explanatory variables, a hidden layer which consists of hidden neurons of any number and an output layer, which in our credit risk modelling case is one neuron. Each of the neurons processes its input/explanatory variables and then relays the output value to the neurons that follow in the subsequent layer and a weight, W_i is assigned during training to each connection between the neurons. The output value of the hidden neuron i is calculated by applying an activation function $f^{(1)}$ to the weighted inputs and the bias term $b_i^{(1)}$ such that;

$$h_i = f^{(1)}\left(b_i^{(1)} + \sum_{j=1}^{M} \mathbf{W_{ij}} x_j\right) \tag{3.3}$$

Where the W is representative of a weight matrix where W_{ij} denotes the weight connecting the input j to hidden neuron i. The analysis in this paper makes a binary prediction therefore, the activation function in the output layer will be a sigmoid activation function $f^{(2)}(x)$

$$f^{(2)}(x) = \frac{1}{1 + e^{-x}} \tag{3.4}$$

to give a response probability given by;

$$\pi = f^{(2)}\left(b^{(2)} + \sum_{j=1}^{n_h} \mathbf{v_j} h_j\right), \tag{3.5}$$

where n_h is the number of hidden neurons and the v is the weight vector with v_j being the weight connecting the hidden neuron j to the output neuron. The weights of the network are usually randomly initialised during model estimation and then iteratively adjusted to minimise a certain objective function such as the sum of squared errors with over-fitting being taken into account through a regularisation term. The iterative process could be a simple gradient descent

or a more complex method such as the Quasi-Newton method. The choice of the number of hidden neurons can be determined using a grid search based on a validation data performance.

3.4.2 Support Vector Machine (SVM) Model

This model has been described clearly by Kecman (2001) and Schölkopf et al. (2000). Consider a training set of instance-label pairs (x_i, y_i), i = 1,2,. . .,m where $x_i \in R^n$ and $y_i \in \{+1, -1\}$. The SVM model then finds the optimal separating hyperplane which has the maximum margin by solving the following optimization problem for the weight w such that;

$$\begin{array}{ll} \underset{w,b}{\text{Min}} & \frac{1}{2} w^{\mathrm{T}} w \\ \text{subject to:} & y_i \left(\langle w \cdot x_i \rangle + b \right) - 1 \geqslant 0 \end{array} \tag{3.6}$$

Solving the quadratic problem which is an optimisation one, we find the saddle point of the lagrange function given by;

$$L_p(w, b, \alpha) = \frac{1}{2} w^T.w - \sum_{i=1}^{m} (\alpha_i y_i (\langle w.x_i > +b) - 1) \tag{3.7}$$

where α_i is the Lagrange multipliers, thus $\alpha_i \geq 0$.

An optimal saddle point is manditory since L_p must then be minimised with respect to the variable w and b, and introducing the Karush Kuhn- Tucker (KKT) condtition for the optimum contrained fucntion, L_p is transformed to the dual Lagrangian $L_D(\alpha)$:

$$\begin{array}{ll} \text{Max}_x & L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i v_j \left(x_i \cdot x_j \right\rangle \\ \text{subject to:} & \alpha_i \geqslant 0 \quad i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \end{array} \tag{3.8}$$

Maximize $L_D(\alpha)$ with respect to the non negative α_i to obtain the optimal hyperplane with parameters w* and b*, hence the optimal hyperplane function $f(x) = sgn(\langle w * .x \rangle + b*)$ can be derived as

$$f(x) = \text{sgn} \left(\sum_{i=1}^{m} y_i \alpha_i^* \left\langle x_i, x \right\rangle + b^* \right) \tag{3.9}$$

In a normal set up , only a sample subse of the Lagrange multipliers α_i tends to be greater than 0. These are the vectors that are, geometrically, closest to the optimal hyperplane. The

respective training vectors having nonzero α_i are known as support v vectors, as the optimal decision hyperplane $f(x, x*, b*)$ depends exclusively on them. The concepts above can then also be extended to the linear generalised SVM. In terms of these introduced slacked variables, the problem of finding the hyperplane that then gives the minimum number of training errors has the formal expression given below:

$$\begin{array}{ll} \mathrm{Min}_{w,b,\xi} & \frac{1}{2} w^{\mathrm{T}} w + C \sum_{i=1}^{m} \xi_i \\ \text{subject to:} & y_i \left(\langle w \cdot x_i \rangle + b \right) + \xi_i - 1 \geqslant 0 \\ \text{subject to:} & \xi_i \geqslant 0 \end{array} \tag{3.10}$$

C is a penalty parameter based on the training error and is chosen by the user while ξ_i is the non negative slack variable. The SVM can then be solvable using the Lagrangian method that is almost similar to the method for solving optimisation problems in the special case. The dual variables Lagrangian must be maximised based on;:

$$\begin{array}{ll} \mathrm{Max}_{\alpha} & L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle \\ \text{subject to:} & 0 \leqslant \alpha_i \leqslant C \quad i = 1, \ldots, m \text{ and } \sum_{i=1}^{m} \alpha_i y_i = 0 \end{array} \tag{3.11}$$

$L_D(\alpha)$ must then be maximised with respect to non-negative α_i under the constraints $\sum_{i=1}^{m} \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$. The penalty parameter C, which now is the upper bound on α_i, is determined by the user. The nonlinear SVM then maps the training sample from the input space onto a higher dimensional feature space through a mapping function $\Phi(x_i)$.

$$\begin{array}{l} \left(\Phi(x_i) \cdot \Phi(x_j) \right) := k(x_i, x_j) \\ L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k(x_i \cdot x_j) \end{array} \tag{3.12}$$

From the linear generalized case steps, we obtain decision function of the following form:

$$\begin{aligned} f(x) &= \mathrm{sgn} \left(\sum_{i=1}^{m} y_i \alpha_i^* \langle \Phi(x), \Phi(x_i) \rangle + b^* \right) \\ &= \mathrm{sgn} \left(\sum_{i=1}^{m} y_i \alpha_i^* \langle k(x, x_i) \rangle + b^* \right) \end{aligned} \tag{3.13}$$

The Radial Basis Function which is a known kernel function is given as follows(Butaru et al., 2016);

$$k\left(x_i, x_j\right) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right) \tag{3.14}$$

3.4.3 Gradient Boosting

This is an ensemble algorithim which improves on the accuracy of a predictive function by incremental minimisation of the error term (Friedman, 2001). It makes use of a base learner, most common being a decision tree, and after this base learner is grown, each of the trees in the series is fit to the residuals with an aim of reducing the error from earlier trees. This gives rise to the following model;

$$F(x) = G_0 + \beta_1 T_1(x) + \beta_2 T_2(x) + ... + \beta_n T_n(x) \tag{3.15}$$

Where G_0 is equal to the first value for the series ,$T_1 ... T_n$ are the trees that are fitted to the pseudo residuals and $\beta_i$are the coefficients for the tree nodes calculated by the algorithm. The detailed explanation of the model has been well covered in the paper by Friedman (2001). This algorithm usually requires some tuning in the iteration number and the maximum branch size which is used when considering the splitting rule. Its a form of ensemble technique that utilises a technique called bagging which makes it improve its predictive ability and it can handle class imbalance really well which makes it ideal for the credit risk modelling.

3.4.4 Random Forest

There have been numerous contributions to improve decision trees and on of the most notable on is the use of random forest algorithms. This model stems from the work of Geman et al. (1992), who was trying to solve the problem of trees instability and it was Breiman (2001) who finally defined Random Forests and his paper has formed the basis for the improvements on this algorithm. this algorithm enhances predictive accuracy and addresses the instability problem of decision trees such as the Classification and Regression Trees (CART) which Breiman (1996) had found to be a problem with decision trees.

To first understand random forest, and understanding of CART would be necessary. CART was introduced by Breiman et al. (1984) and it is a statistical technique in which a dependent or "output" variable which is either continuous or discrete, is related to a group of independent or "input" variables through a recurring sequence of simple binary relations hence the referred to as a "tree"Breiman et al. (1984). The collection of recurring relations partitions the multi-dimensional space of independent variables into different regions in which the dependent variable is generally assumed to be constant, in which case it is a classification tree or linearly related, which is a regression tree, to the independent variables.

This model is preferred as it overcomes the disadvantages of standard models where the dependent variable is forced to fit in a single linear model in the entirety of the input space. CART models are simply applied to problems with high-dimensional feature spaces.For instance, suppose we have N observations of the dependent variable $\{y_1, ..., y_N\}$ and its corresponding d-dimensional feature vectors $\{x_1, ..., x_N\}$. Estimation of the parameters of the CART model is done on the training data set by recursively selecting features from $x \in \{x_1, ..., x_D\}$and parameters $\{L_j\}$ that minimize the residual sum-of-squared errors. It is important to create a "pruning criterion" for stopping the expansion of the tree in order to avoid over fitting the training data (Breiman, 1996).

RF builds on the same principle as CART and has the ability of considering a very big number of predictors even when J¿N while still maintaining efficiency. After a large number of trees have been generated, a voting process for each tree happens to find the most popular class. This collective voting process is what constitutes a RF algorithm. In RF, for each tree and at every node, the algorithm chooses the most relevant splitting point and variable in order to reduce the errors an the model uses every variable to give predictions unlike CART where the variables not selected do not interfere with the response. Therefore the RF model provides a prediction that is based on all explanatory variables. To improve the hierarchy of the predictors, RF gives a ranking based on 2 main measures. The first is the Mean Decrease Accuracy and the Mean Decrease impurity given by the following equations (Breiman, 2001);

$$MDA(X_j) = \frac{1}{B}\sum_{1}^{B}(e_{OBB} - e_{OBBj^i}) \tag{3.16}$$

Where e_{OBB} is the error rate on the out-of-bag sample.

$$MDA(X_j)_{classification} = \frac{1}{B}\sum_{1}^{B}\sum_{t\in T_b} I(j_{t^*} = j)\left(\frac{N_t}{N}\delta i(s,t)\right) \tag{3.17}$$

A detailed explanation of how these errors are sued in analysing the RF algorithm can be found in (Breiman, 2001). This model has the ability to learn with class imbalance due to how the voting process works and therefore, has the potential of performing well when dealing with credit data which is usually very unbalanced with the input and response variables.

3.5 Performance Measures

When running these models, it is important to split the data into 2 sets: training and test data. This allows us to be able to run the model on the training data and get the relevant parameters and then use the results from here to see how the model will perform under the test data. It is common practice to split the data into a 70/30 % split respectively though this is not a strict requirement and any percentage split can be used (Al-Shayea et al., 2010). This will be further discussed in the following section on data analysis. Once the model is built, it is important to test how good the model is in its prediction. There exist different measures to asses the efficiency of a model and quantify the quality of the predictions. These measures include Confusion matrix, Accuracy, Precision, Recall, F1 Score and the AUC.

3.5.1 Confusion Matrix

A confusion matrix is a table which is commonly used in describing the performance of classification models on a test data set where the true values are known.The table 1 shows how the matrix looks like;

The four parameters contained in the table are what are used in the calculation of some of the measures used to asses the performance of the model so it is important to understand what each of them mean.

True Positives (TP)- These are the correctly predicted positive observations by the model based on the actual data, which in our case implies the correct number of defaults that have been predicted by the model and were defaults in the actual data.

	Predicted Class		
3*Observed Class		Class=1	Class=0
	Class=1	True Positive	False Negative
	Class=0	False Positive	True Negative

Table 3.1: Confusion Matrix describing the performance of classification models

True Negatives (TN)- These are the correctly predicted negative values by the model based on the actual data, which in our case implies the correct number of non-default that have been predicted by the model and were non-default in the actual data.

False Positives (FP)- These are the wrongly predicted positive values by the model based on the actual data, which in our case implies the observations predicted as default by the model but were non-default in the actual data.

False Negative (FN)- These are the wrongly predicted negative values by the model based on the actual data, which in our case implies the observations predicted as non-default by the model but were defaults in the actual data.

3.5.2 Accuracy

One of the most commonly used measure for performance is accuracy and this gives the ratio of the correctly classified observations to the total observations. This measure does not take the distribution of the class into account and thus it is considered a poor measure for model evaluation on a case of imbalanced data. It is given by the following formula (Al-Shayea et al., 2010);

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.18}$$

3.5.3 Precision, Recall and F- measure (F1 score)

In classification problems, many a times one class is usually of special interest. In the cases where the class of interest, also known as the positive class, is heavily outnumbered, the dataset is considered to be imbalanced. Our study deals with such a case where the number of defaults

is usually outnumbered. The accuracy measure mentioned above does not do well with such cases as mentioned. For instance, imagine a dataset that has 2 classes where the positive class is outnumbered by a ratio of 100:1, by classifying all the classes as majority class, the model will achieve a score of 99% and when exposed to a different dataset that is balanced, the model will do very poorly.

June, 2019

This is where the measures that follow come in (Liu et al., 2018);

1. Precision is the ratio of all the correctly classified observations from the positive class among all observations classified by the model as the positive class.

$$Precision = \frac{TP}{TP + FP} \tag{3.19}$$

2. Recall is the ratio of correctly classified observations from the positive class among all samples from the positive class in the actual data.

$$Recall = \frac{TP}{TP + FN} \tag{3.20}$$

3. F-Measure is the harmonic mean between the Recall and Precision and is given by;

$$F - Measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{3.21}$$

3.5.4 Receiver Operating Characteristic Curve (ROC), Precision-Recall Curve and AUC

The ROC is a graph of the FP and TP for a number of various different threshold values between 0 and 1. The shape of this curve contains a lot of information that can help in determining how good a model is. The AUC can be used as a summary of this this curve for various possible thresholds. Figure 1 shows an example of an ROC curve.

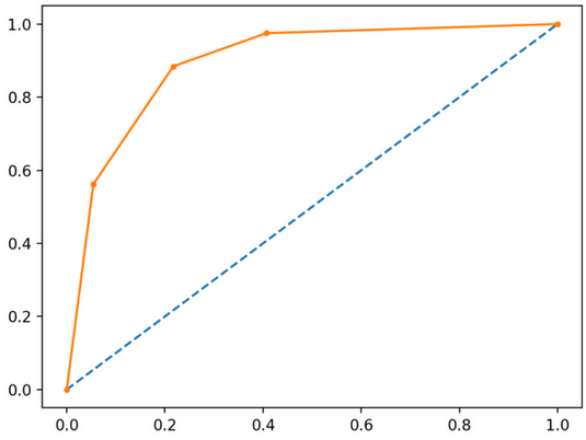


Figure 3.1: ROC curve for different threshold values between 0 and 1.

When there is a class imbalance, the precision-recall curve is preferred over the ROC and this plots the precision against the recall for various thresholds just like the ROC. The summary of this curve is also given by the AUC.

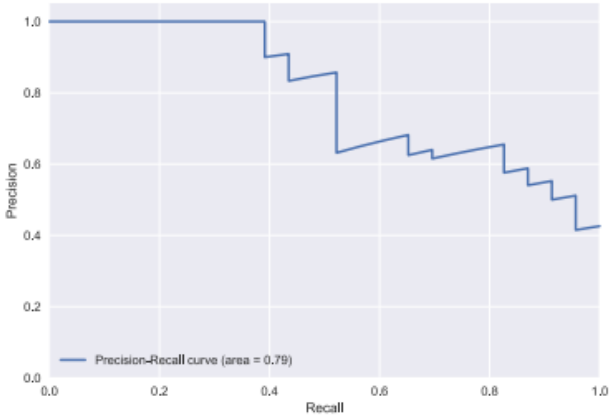


Figure 3.2: Precision-recall curve for various thresholds between 0 and 1.

Studies on credit risk modelling suggest that precision-recall curve is much more informative for evaluating performance of binary classifiers since they are imbalanced datasets and a measure that captures this is more appropriate than the ROC curve (Saito & Rehmsmeier, 2015).

Chapter 4

Data Analysis and Discussion

4.1 Data Overview

The data that was used for the analysis in this study was obtained from Kenya Metropol for the years 2014-2017. The data consisted of close 20,299 entries of individuals who had taken loans from different banks and were tracked over the period to observe whether they defaulted on the loan or not. The various features captured by the data included the gender, marital status, loan amount, age bracket, product type, time since loan origination, loan amount group, and default status.

Before the data was used for analysis, it had to be cleaned and organised in a way such that analysis was possible. All categorical variables were converted to their numeric equivalent with each category being assigned a number. The other numeric values such as age were converted into float values and the data was checked for missing values. There were no missing data found however, if there were, it would have been better to substitute the missing values with the class/feature mean or median . The whole dataset was then re-scaled through a Z-score standardisation since not all data sets are on the same scale though this only proved important for the logistic regression specifically.

The data was then divided into a training set and a test set based on a 70/30 percentage split respectively and the split was performed randomly across the data set. Various sampling techniques exist to deal with imbalanced datasets like the one used in this study such as SMOTE,kNN or Tomek-links, however, the study decided to just go with a random over-

sampling approach for the response variable due to time and computational constraints. The training set was what was used for prediction with the models then the test data set was used to evaluate how effective the models would be in predicting.

4.2 Findings

4.2.1 Feature Analysis

To better understand the features used, the table that follows provides a sumary of the features used in analysis

		Active	Defaulter
Gender	F	= 7857 (82.61%)	= 1654 (17.39%)
	M	= 7989 (74.05%)	= 2799 (25.95%)
Age Bracket	>54	= 2150 (86.14%)	= 346 (13.86%)
	18-33	= 4652 (71.95%)	= 1814 (28.05%)
	34-43	= 4908 (78.73%)	= 1326 (21.27%)
	44-53	= 4136(81.05%)	= 967 (18.95%)
Loan Groups	A	= 533 (86.67%)	= 82 (13.33%)
	B	= 848 (84.38%)	= 157 (15.62%)
	C	= 1527 (84.32%)	= 284 (15.68%)
	D	= 1473 (78.94%)	= 393 (21.06%)
	E	= 1657 (75.66%)	= 533 (24.34%)
	F	= 9808 (76.55%)	= 3004 (23.45%)
Product Name	Current Account	= 2419 (78.36%)	= 668 (21.64%)
	Loan Account	= 10877 (83.75%)	= 2111 (16.25%)
	Credit Card	= 2550 (60.37%)	= 1674 (39.63%)
Status	Divorced	= 43 (75.44%)	= 14 (24.56%)
	Married	= 10483 (78.60%)	= 2854 (21.40%)
	Single	= 5311 (77.11%)	= 1577 (22.89%)
	Widowed	= 9 (52.94%)	= 8 (47.06%)
Amount Group (‘000)	0-50	= 9808 (76.55%)	= 3004 (23.45%)
	50-100	= 1657 (75.66%)	= 533 (24.34%)
	100-250	= 1473 (78.94%)	= 393 (21.06%)
	250-500	= 1527 (84.32%)	= 284 (15.68%)
	500-1,000	= 848 (84.38%)	= 157 (15.62%)
	>1,000	= 533 (86.67%)	= 82 (13.33%)

It is observed that around 25.95% of male defaulted while 17.39% of females defaulted.It was also observed that 21.40% of married people defaulted while 22.97% of single people defaulted which is just slightly more than the latter. The 18-33 age bracket had a 28.05% default rate, the 34-43 bracket had a 21.27% default rate, the 44-53 bracket had a 18.95% default rate while the >54 bracket had a 13.86% default rate. The younger people seemed to default more and this could be due to lack of stable income stream among other factors. The loan category above 1 million had a 13.33% default rate which was the lowest rate and the rate kept rising with each category, however, the 50,001-100,000 category had the highest default rate at 24.34% which was higher than the 0-50,000 bracket which stood at 23.45% category. The overall default rate for the whole data set was 21.94%

4.2.2 Model Analysis and Discussion

The metrics that were highlighted in Chapter 3 were used to evaluate the model performance. Finding a universal measure for evaluating performance is a very difficult and subjective process depending on the task at hand. In practice, the optimal evaluation model would be a profit function, which would be a function of precision and recall, that would have to be optimized. The profit would be estimated as a trade off between the TP (profit) and the FP (cost), both of which are captured by the F-measure or the precision recall AUC. However, for this study, the F-measure is chosen as the metric to evaluate the models.There after, the model with the best F measure can be tuned and validated to try provide a better prediction and subsequently, better evaluation measure.

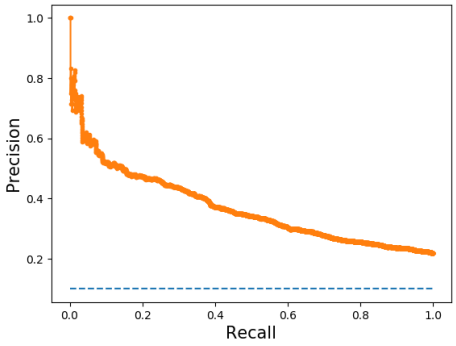
The best performing model for each class was decided using a stratified 5-fold cross validation process and the following results show the performance of the models using the metrics;

Model	Precision Score	Recall Score	F1_score	Accuracy	PR AUC
Logistic Regression	0.7778	0.0052	0.0104	0.7814	0.4006
Random Forest	0.4386	0.2058	0.2802	0.7680	0.4932
SVM	0.6373	0.0921	0.1609	0.7893	0.4643
Gradient Boosting	0.6012	0.1422	0.2300	0.7911	0.4658
MLP	0.5579	0.1369	0.2200	0.7869	0.4421

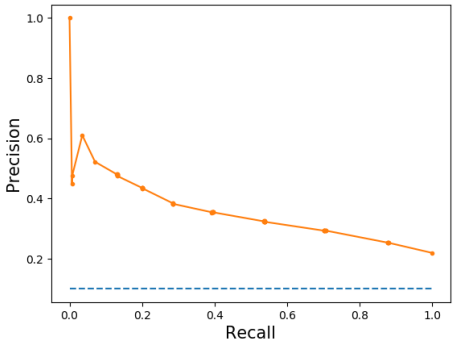
Table 4.1: Summary of the performance evaluation results of the models

Based on the table of results above, we calculated the various values for the metrics which could be used to assess our models. The Logistic regression has the highest accuracy score followed by the SVM model and the Random forest has the least precision score. On the other hand, the recall score for the random forest is the highest followed by the Gradient boosting model while that of the logistic regression is the least. This is due to the precision-recall trade off that was demonstrated when Precision-recall curve was mentioned in the previous chapter. This precision -recall curve is what is used to identify the optimal threshold between the two value for an optimum result. Hyper parameter optimization is the process that is used when trying to find the ideal threshold based the objective at hand.

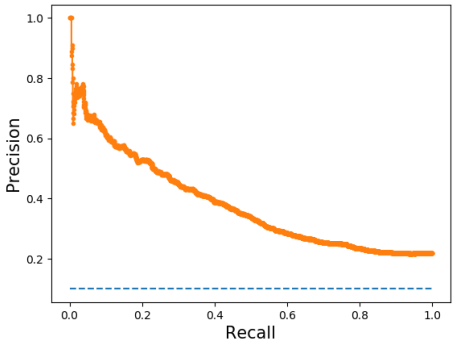
The F score is a measure that tries to capture this optimal combination and based on the Table 2, Random forest has the best F1- Score followed by the Gradient boosting model while the logistic regression has the least F score. All the models had a relatively strong accuracy score which is the ability of the model to capture the defaults. The Precision-Recall (PR) AUC for all the models are below 0.5 which is indicative that the models didn't do a good job capturing the false negatives and positives and thus are not skillful models however, it could also be that there wasn't enough features for the models to train adequately enough. The graphs for the PR curves of the models follow;



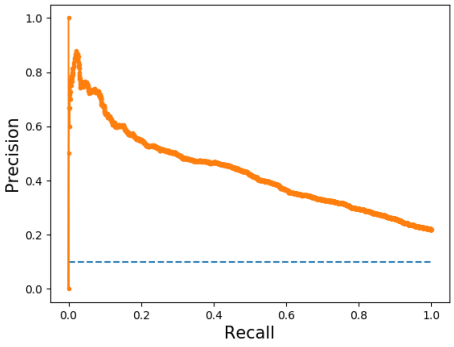
(a) Logistic Regression PR curve



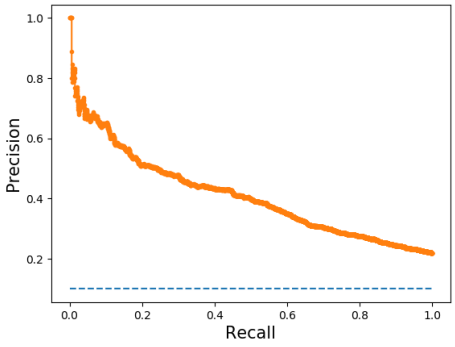
(b) Random Forest PR curve



(c) SVM PR curve



(d) Gradient Boosting PR Curve



(e) MLP PR Curve

Figure 4.1: PR Curves

When dealing with credit risk, we have to decide what kind of falses are tolerable for the institution. False positives will lead us to reject customers that would have otherwise been profitable customers but the models have wrongly classified them as being a high risk for the bank. False Negatives would introduce more risk to the bank by classifying customers as not risky then they end up being a higher risk of default and therefore bring losses to the company. Precision is what determines the former and Recall determines the latter. Since both instances are intolerable for the institution, it therefore means that we have to find the optimal weightage between these two measures to ensure that the bank minimises its cost and maximises its profit. The choice of weights is usually done through hyper parameter optimisation to determine the most optimum threshold between these two. The PR curve is a plot of all the possible threshold combinations possible between the Recall and the Precision. TO therefore determine the best threshold to optimize the performance of the model, hyper parameter optimisation and model tuning are necessary.

Due to computational and time constraints, hyper parameter optimization was not possible for all the models and therefore the study did optimisation for the two best performing models based on the F score and the PR AUC score which were; Random forest and Gradient Boosting. A gridsearch was done to determine the optimal parameters and there after model tuning was done on the two models. After model tuning the Random forest model improved its F1-score from 0.280 to 0.307 and the graph below shows the PR curve for the model after optimisation;

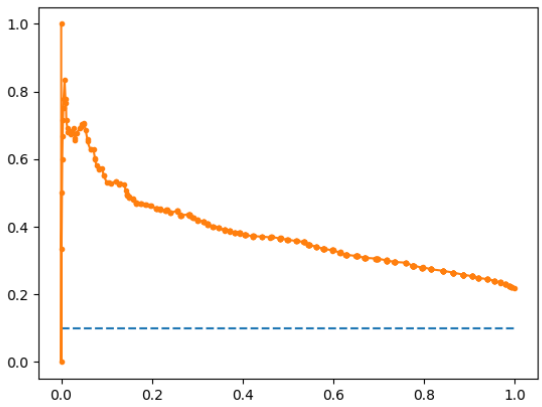


Figure 4.2: Precision-recall curve for the random forest model after optimisation.

Gradient Boosting Model, the F1-score improved from 0.230 to 0.271 and the graph below shows the PR curve after optimisation;

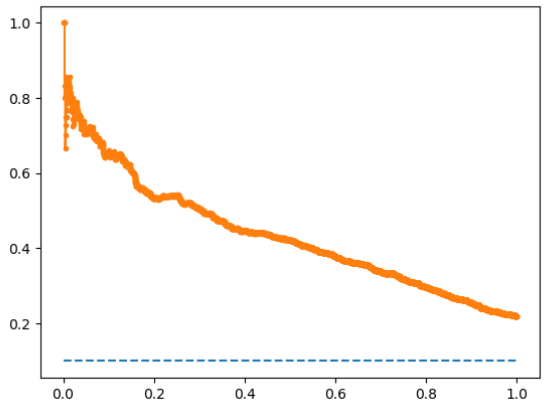


Figure 4.3: Precision-recall curve for the random forest model after optimisation.

Based of this, we observe that the Random forest model optimises with a better F1- score compared to the Gradient Boosting Model and from the results and analysis it can be inferred that the Random Forest model was the best fit among all the models in predicting credit risk.

Chapter 5

Conclusion and Recommendations

5.1 Conclusion

The study has shown that modelling credit risk is quite a challenging problem since what makes a high risk customer is not exactly a binary decision but rather a sliding scale. In conducting binary classification, a boundary has to be set and this is the challenging task, and the findings have shown this. Additionally, the imbalanced data set did make things much more difficult for the algorithms to learn, which is a common problem with imbalanced datasets as shown by studies such as (Khandani, Kim, & Lo, 2010) and (Cambria et al., 2013).

For the data used in this study, it was established that machine learning models had a better performance compared to the logistic regression and among the machine learning models, Random forest showed the best efficiency at modelling credit risk. This was followed by the gradient Boosting model while the logistic regression model performed the worst. Efficiency, in this study, was taken to be the ability of the models to optimize the companies profitability by minimising the cost of false negative and maximising the revenue through minimising the opportunity cost of false positives. the measure that was used to capture this was the F1-score which was a harmonic mean between the Precision and Recall. However, logistic regression did perform well when it came to the Accuracy measure, which is the ratio of the correctly classified defaults to the total observation, with an accuracy of 77.78% . The accuracy measure is however, not a good measure of performance as it does not take into account the cost of misclassification which is captured in the false positives and negatives which is why the score was not used.

Ensemble models generally performed better and this has also been found in other studies such as the study by (Butaru et al., 2016). The presence of data imbalance made overfitting an issue and Random forest, being a model that is robust against overfitting issues, could have been the reason why it performed the best among the other models. To optimize for a better threshold between Precision and Recall, the study conducted a hyper parameter optimization and model tuning and the Random Forest model still outperformed with the F1-score improving to 0.307 from an initial value of 0.280.

The study also established that male customers were more likely to default compared to their female counterparts with around a 26% and 17% chance respectively. Single customers were slightly more likely to default compared to married customers with 23% and 21% chances respectively. It was also determined that the 18-33 age bracket was the most likely to default while the >54 age bracket being the least likely. Furthermore, it was observed that the loan amounts between KES 50,000-100,000 were the most likely to be defaulted followed by the KES 0-50,000 bracket with the least likely group to default being >1,000,000.

The study concludes that the machine learning models have a better performance in modelling credit risk when dealing with imbalanced datasets, like credit data sets, however, the models could have had a better performance had the number of features in the dataset been more. Additionally, more sophisticated sampling techniques such as SMOTE could have helped improve the imbalanced data set and improve performance.

5.2 Limitations

There are a number of factors that affected how the study was conducted and hence contributed to the limitations. The include;

1. The data used for the study did not have enough features to be able to better train the models to give more accurate results. With more features, the number of hidden feature interaction would have increased providing a better chance for the models to learn and provide more accurate predictions and improved measure metrics

2. Data preprocessing and preparation is usually the most consuming part of any machine learning project and more could have been done to better improve performance, especially

feature engineering but the lack of enough features hindered this process.

3. Hyperparameter optimisation is a very time consuming and computational intensive activity. Due to these constraints, not all models were fine tuned and time was only used in optimizing the best performing models.

4. The use of forecast horizon for default windows could also have been a good approach to pursue however the study only focused on the classification problem alone

5. Estimation of the potential profits would have been a good addition to this thesis but it fell beyond the scope of this study and since this is an important aspect of any business, this makes it a limitation for the study having not pursued it.

5.3 Recommendations

As noted in section 5.2, there are a number of recommendations this study suggests based on the limitations namely;

1. Analysis should be done on data with much more features such as the monthly transactional details of the accounts of the customers and this will enable feature engineering which may improve performance by having a more optimal subset. Additionally, the use of alternate data, which is a trending issue in the digital lending, would also be an interesting area to explore given the growth of mobile loans in Kenya.

2. Future works should also consider taking this a step further and model the default experience after classification into examining different window lengths to capture the month to month changes of the customer.

3. Future works could also consider exploring deep learning models and how well they may perform within the Kenyan context but the one drawback of these models is that they are black boxes which could be difficult to explain using theory.

References

Abdon, A., & Felipe, J. (2011). The product space: What does it say about the opportunities for growth and structural transformation of sub-saharan africa?

Addo, P., et al. (2018). Credit risk analysis using machine and deep learning models. *Risks*, *6*(2), 38.

Allen, L., et al. (2004). Issues in the credit risk modeling of retail markets. *Journal of Banking & Finance*, *28*(4), 727–752.

Al-Shayea, Q., et al. (2010). Neural networks in bank insolvency prediction. *International Journal of Computer Science and Network Security*, *10*(5), 240–245.

Altman, D. G., & Bland, J. M. (1994). Diagnostic tests. 1: Sensitivity and specificity. *BMJ: British Medical Journal*, *308*(6943), 1552.

Anderson, R. (2007). *The credit scoring toolkit*. Oxford University Press.

Andries, A. (2008). Theories regarding the banking activity. *Analele Stiintifice ale Universitatii "Alexandru Ioan Cuza" din Iasi*, *55*, 19–29. Retrieved from `http://anale.feaa.uaic.ro/anale/resurse/03`$_F$`12`$_A$*ndries.pdf*

Bagozzi, R. P., et al. (1992). Development and test of a theory of technological learning and usage. *Human relations*, *45*(7), 659–686.

Bishop, C. M., et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Blanco, A., Pino-Mejías, R., Lara, J., & Rayo, S. (2013). Credit scoring models for the microfinance industry using neural networks: Evidence from peru. *Expert Systems with applications*, *40*(1), 356–364.

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123–140.

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

Breiman, L., et al. (1984). Classification and regression trees. wadsworth int. *Group*, *37*(15), 237–251.

Butaru, F., Chen, Q., Clark, B., Das, S., Lo, A. W., & Siddique, A. (2016). Risk and risk

management in the credit card industry. *Journal of Banking & Finance*, *72*, 218–239.

Cambria, E., Liu, Q., Li, K., Leung, V. C., Feng, L., Ong, Y.-S., ... others (2013). Extreme learning machines. *IEEE Intelligent Systems*(6), 30–59.

Cao, E., & Yu, M. (2018). Trade credit financing and coordination for an emission-dependent supply chain. *Computers & Industrial Engineering*, *119*, 50–62.

Casu, B., et al. (2006). *Introduction to banking*. Pearson Education.

Coats, P. K., & Fant, L. F. (1993). Recognizing financial distress patterns using a neural network tool. *Financial management*, 142–155.

DeTienne, D. R., & Chirico, F. (2013). Exit strategies in family firms: How socioemotional wealth drives the threshold of performance. *Entrepreneurship Theory and Practice*, *37*(6), 1297–1318.

Duffie, D., Pan, J., & Singleton, K. (2000). Transform analysis and asset pricing for affine jump-diffusions. *Econometrica*, *68*, 1343–1376.

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

Finlay, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, *210*(2), 368–378.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Geman, S., et al. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, *4*(1), 1–58.

Jarrow, R. A., & Turnbull, S. M. (1995). Pricing derivatives on financial securities subject to credit risk. *The journal of finance*, *50*(1), 53–85.

Jensen, H. L. (1992). Using neural networks for credit scoring. *Managerial finance*, *18*(6), 15–26.

Karaa, A., & Krichene, A. (2012). Credit-risk assessment using support vectors machine and multilayer neural network models: a comparative study case of a tunisian bank. *Accounting and Management Information Systems*, *11*(4), 587.

Kecman, V. (2001). *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Klein, W. (1992). The present perfect puzzle. *Language*, 525–552.

Lacher, R., et al. (1995). A neural network for classifying the financial health of a firm.

European Journal of Operational Research, *85*(1), 53–65.

Liu, M., et al. (2018). A comparison of machine learning algorithms for prediction of past due service in commercial credit.

Merton, R. C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *The Journal of finance*, *29*(2), 449–470.

Mitchell, T. M. (1997). Does machine learning really work? *AI magazine*, *18*(3), 11–11.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of accounting research*, 109–131.

Okay, E., et al. (2008). Views on turkey's impending esco market: Is it promising? *Energy Policy*, *36*(6), 1821–1825.

Rhyne, E., & Christen, R. (1999). *Microfinance enters the marketplace*. Citeseer.

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, *10*(3), e0118432.

Santomero, A. M., & Vinso, J. D. (1977). Estimating the probability of failure for commercial banks and the banking system. *Journal of Banking & Finance*, *1*(2), 185–205.

Schölkopf, B., et al. (2000). New support vector algorithms. *Neural computation*, *12*(5), 1207–1245.

Schreiner, M. (2004). Benefits and pitfalls of statistical credit scoring for microfinance/ventajas y desventajas del scoring estadístico para las microfinanzas/vertus et faiblesses de l'évaluation statistique (credit scoring) en microfinance. *Savings and Development*, 63–86.

Shin, K., et al. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, *28*(1), 127–135.

Valle, A. C. (2013). *Credit risk modeling in a semi-markov process environment* (Unpublished doctoral dissertation). The University of Manchester (United Kingdom).

Wagacha, A., & Othieno, F. (2015). Semi-markov credit risk modeling for a portfolio of consumer loans in the kenyan banking industry.

West, W. M. (2000, Dec). Images and diagnoses. synovial sarcoma. *The West Indian medical journal*, *49*(4), 337, 346.

Appendix

The Python Code

```
# Import the required Packages
from sklearn.model_selection import train_test_split
from sklearn.metrics import roc_curve, auc, confusion_matrix, classification_
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import GridSearchCV
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.model_selection import KFold
from sklearn.ensemble import GradientBoostingClassifier, RandomForestClassifi
from sklearn.neural_network import MLPClassifier
from sklearn import preprocessing, metrics
import pandas as pd
import matplotlib
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.manifold import TSNE
from IPython.core.interactiveshell import InteractiveShell
import warnings
warnings.filterwarnings('ignore') # to supress warnings
from sklearn.preprocessing import StandardScaler
from sklearn.manifold import TSNE
warnings.filterwarnings('ignore') # to supress seaborn warnings
pd.options.display.max_columns = None # Remove pandas display column number l
#InteractiveShell.ast_node_interactivity = "all" # Display all values of a ju
train = pd.read_csv("survival_dataset.csv")
train=pd.DataFrame(train)
```

```
#print(train.isnull().sum())

train.gender=train.gender.astype('category')
train.age_bracket=train.age_bracket.astype('category')
train.original_amount=train.original_amount.astype('float')
train.loan_groups=train.loan_groups.astype('category')
train.status=train.status.astype('category')
train.time_diff=train.time_diff.astype('category')
train.age=train.age.astype('float')
train.log_amount=train.log_amount.astype('float')
train.product_name=train.product_name.astype('category')
print(train.default_dummy.mean()*100)
print(train.head())
#print(train.isnull().values.any())
#print((train.isna().sum().sum()/(train.shape[0]*train.shape[1])))

def plot0(col1, col2, tittle, xticks, train):
    dt = train.groupby(col1).agg([np.mean])*100.0
    dt = dt[col2].reset_index()
    f, ax = plt.subplots(figsize=(5, 5))
    sns.barplot(x=col1, y="mean", data=dt)
    ax.set(xlabel="", ylabel="Defaulter %")
    ax.set_title(label=tittle, fontsize=15)
    ax.set_xticklabels(xticks, fontsize=11)
#Crosstab

group_crosstab = pd.crosstab(train.default_dummy, train.loan_groups, margins=
new_index = {0: 'Non-default', 1: 'Default', }
new_columns = {1 : '>1,000,000', 2 : '500,001-1,000,000', 3 : '250,001-500,00
group_crosstab.rename(index=new_index, columns=new_columns, inplace=True)
print(group_crosstab/group_crosstab.loc['All'])

#graph
```

```
col1 = "loan_groups"
col2 = "default_dummy"
tittle = "% of Defaulters by Loan Group in '000"
xticks = [">1,000", "500-1,000", "250-500", "100-250", "50-100", "0-50"]
plot0(col1, col2,  tittle, xticks, train)

age_crosstab = pd.crosstab(train.default_dummy, train.age_bracket, margins=Tr
new_index = {0: 'Non-default', 1: 'Default', }
new_columns = {1 : '18-33', 2 : '34-43', 3 : '44-53', 4 : '>54'}
age_crosstab.rename(index=new_index, columns=new_columns, inplace=True)
print(age_crosstab/age_crosstab.loc['All'])

#graph
col1 = "age_bracket"
col2 = "default_dummy"
tittle = "% of Defaulters by Age"
xticks = ["18-33", "34-43", "44-53", ">54"]
plot0(col1, col2,  tittle, xticks, train)

cor = train.corr()
plt.figure(figsize=(18,18))
graph=sns.heatmap(cor, cbar = True,  square = True, annot=True, fmt= '.2f',an
            xticklabels=cor.columns.values,
            yticklabels=cor.columns.values)
print(graph)
plt.show()


#Crosstab
sex_crosstab = pd.crosstab(train.default_dummy, train.gender, margins=True, m
new_index = {0: 'Non-default', 1: 'Default', }
new_columns = {1 : 'Male', 2 : 'Female'}
```

```
sex_crosstab.rename(index=new_index, columns=new_columns, inplace=True)
print(sex_crosstab/sex_crosstab.loc['All'])

#Bar Chart
col1 = "gender"
col2 = "default_dummy"
tittle = "% of Defaulters by Sex"
xticks = ["Male", "Female"]
plot0(col1, col2, tittle, xticks, train)

#Crosstab
marital_crosstab = pd.crosstab(train.default_dummy, train.status, margins=Tru
new_index = {0: 'Non-default', 1: 'Default', }
new_columns = {1 : 'Married', 2 : 'Single'}
marital_crosstab.rename(index=new_index, columns=new_columns, inplace=True)
print(marital_crosstab/marital_crosstab.loc['All'])

#Bar Chart
col1 = "status"
col2 = "default_dummy"
tittle = "% of Defaulters by Marital Status"
xticks = ["Married", "Single",]
plot0(col1, col2, tittle, xticks, train)

defaulters = train[train.default_dummy == 1]
non_defaulters = train[train.default_dummy == 0]
defaulters["Defaulter"] = defaulters["age_bracket"]
non_defaulters["Non Defaulter"] = non_defaulters["age_bracket"]
f, ax = plt.subplots(figsize=(12, 6))
ax = sns.kdeplot(defaulters["Defaulter"], shade=True, color="r")
ax = sns.kdeplot(non_defaulters["Non Defaulter"], shade=True, color="g")
plt.show()
```

```
#Predictive modelling

x = train.drop(['default_dummy'],axis = 1)
y = train.default_dummy

# rescale the metrics to the same mean and standard deviation
scaler = preprocessing.StandardScaler()
x = scaler.fit(x).transform(x)

# Further divide the train data into train test split 70% & 30% respectively
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, stra

# list of different classifiers we are going to test
clfs = {
'LogisticRegression' : LogisticRegression(),
'RandomForest': RandomForestClassifier(),
'SVM': SVC(),
'GradientBoosting': GradientBoostingClassifier(),
'MLPClassifier': MLPClassifier(),

}
# code block to test all models in clfs and generate a report
models_report = pd.DataFrame(columns=['Model', 'Precision_score', 'Recall_sco

for clf, clf_name in zip(clfs.values(), clfs.keys()):
    clf.fit(x_train, y_train)
    y_pred = clf.predict(x_test)
    y_score = clf.score(x_test, y_test)

    # print('Calculating {}'.format(clf_name))
    precision, recall, thresholds = precision_recall_curve(y_test, y_pred)
```

```
    t = pd.Series({
        'Model': clf_name,
        'Precision_score': metrics.precision_score(y_test, y_pred),
        'Recall_score': metrics.recall_score(y_test, y_pred),
        'F1_score': metrics.f1_score(y_test, y_pred),
        'Accuracy': metrics.accuracy_score(y_test, y_pred),
        'AUC': metrics.roc_auc_score(y_test, y_pred),
        'AUC1': metrics.auc(recall, precision),
    })

    models_report = models_report.append(t, ignore_index=True)

print(models_report)


# Function to optimize model using gridsearch
def gridsearch(model, params, x_train, x_test, y_train, y_test, kfold):
    if __name__ == '__main__':
        gs = GridSearchCV(model, params, scoring='accuracy', n_jobs=-1, cv=kf
        gs.fit(x_train, y_train)
        print('Best params: ', gs.best_params_)
        print('Best AUC on Train set: ', gs.best_score_)
        print('Best AUC on Test set: ', gs.score(x_test, y_test))


# Function to generate confusion matrix
def confmat(pred, y_test):
    conmat = np.array(confusion_matrix(y_test, pred, labels=[1, 0]))
    conf = pd.DataFrame(conmat, index=['Defaulter', 'Not Defaulter'],
                        columns=['Predicted Defaulter', 'Predicted Not Defau
    print(conf)
```

```
# Function to plot roc curve
def roc(prob, y_test):
    y_score = prob
    fpr = dict()
    tpr = dict()
    roc_auc = dict()
    fpr[1], tpr[1], _ = roc_curve(y_test, y_score)
    roc_auc[1] = auc(fpr[1], tpr[1])
    plt.figure(figsize=[7, 7])
    plt.plot(fpr[1], tpr[1], label='Roc curve (area=%0.2f)' % roc_auc[1], lin
    plt.plot([1, 0], [1, 0], 'k—', linewidth=4)
    plt.xlim([0.0, 1.0])
    plt.ylim([0.0, 1.0])
    plt.xlabel('False Positive rate', fontsize=15)
    plt.ylabel('True Positive rate', fontsize=15)
    plt.title('ROC curve for Credit Default', fontsize=16)
    plt.legend(loc='Lower Right')
    plt.show()


def model(md, x_train, y_train, x_test, y_test):
    md.fit(x_train, y_train)
    pred = md.predict(x_test)
    prob = md.predict_proba(x_test)[:, 1]
    print(' ')
    print('Accuracy on Train set: ', md.score(x_train, y_train))
    print('Accuracy on Test set: ', md.score(x_test, y_test))
    print(' ')
    print(classification_report(y_test, pred))
    print(' ')
    print('Confusion Matrix')
    confmat(pred, y_test)
    roc(prob, y_test)
```

```
    return md


# Use gridsearch to fine tune the parameters
gb = GradientBoostingClassifier()
gb_params = {'n_estimators': [100,200,300],'learning_rate' : [0.01, 0.02, 0.0
gridsearch(gb, gb_params,x_train, x_test, y_train, y_test,5)


# feature selection with the best model from grid search
gb = GradientBoostingClassifier(learning_rate= 0.02, max_depth= 7,n_estimator
model_gb = model(gb, x_train, y_train,x_test, y_test)
mod_l= GradientBoostingClassifier(learning_rate= 0.02, max_depth= 7,n_estimat
mod_l.fit(x_train,y_train)
prob = mod_l.predict_proba(x_test)
prob = prob[:, 1]
y_hat = mod_l.predict(x_test)
precision, recall, thresholds = precision_recall_curve(y_test, prob)
f1 = f1_score(y_test, y_hat)
auc = auc(recall, precision)
ap = average_precision_score(y_test, prob)
print('f1=%.3f auc=%.3f ap=%.3f' % (f1, auc, ap))
# plot no skill
plt.plot([0, 1], [0.1, 0.1], linestyle='--')
# plot the precision-recall curve for the model
plt.plot(recall, precision, marker='.')
# show the plot
plt.show()
mod_l= RandomForestClassifier(n_estimators=100)
mod_l.fit(x_train,y_train)
prob = mod_l.predict_proba(x_test)
prob = prob[:, 1]
y_hat = mod_l.predict(x_test)
```

```
precision, recall, thresholds = precision_recall_curve(y_test, prob)
f1 = f1_score(y_test, y_hat)
auc = auc(recall, precision)
ap = average_precision_score(y_test, prob)
print('f1=%.3f auc=%.3f ap=%.3f' % (f1, auc, ap))
# plot no skill
plt.plot([0, 1], [0.1, 0.1], linestyle='--')
# plot the precision-recall curve for the model
plt.plot(recall, precision, marker='.')
# show the plot
plt.show()
```