National College of Ireland

Minimizing Credit Risk In Peer-to-Peer Lending Business Using Supervised Machine Learning Techniques

MSc Research Project
Data Analytics

Akeem Ayantola
Student ID: X18168680

School of Computing
National College of Ireland

Supervisor: Dr. Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Akeem Lekan Ayantola
Student ID:	X18168680
Programme:	Data Analytics
Year:	2020
Module:	MSc Research Project
Supervisor:	Dr. Vladimir Milosavljevic
Submission Due Date:	23/04/2020
Project Title:	Minimizing Credit Risk In Peer-to-Peer Lending Business Using Supervised Machine Learning Techniques
Word Count:	9451
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	22nd April 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Minimizing Credit Risk In Peer-to-Peer Lending Business Using Supervised Machine Learning Techniques

Akeem Ayantola
X18168680

Abstract

One of the major challenges facing the retail finance market including banks is the issue of credit risk. This process involves the evaluation of individual customers' historical transaction data to determine their credit worthiness. Hence, the decision-making process is largely influenced by the outcome of the credit risk evaluation. Data mining techniques have been applied on financial dataset and this has shown significant results. In this report, the dataset from a peer-to-peer company (LendingClub Corporation) has been utilized. An oversampling Technique (Resample) was applied to overcome the class imbalance in the dataset. Similarly, five machine learning algorithms have been implemented to train and build an efficient classification model for evaluating credit risk. These models include Logistic regression, Random Forest, Decision Tree, Naïve-Bayes, and Ada-Boost. A feature importance plot was implemented where a new dataset containing only the top ten important features were obtained . All five models were retrained using this new dataset and resulted obtained. The results were evaluated on their Accuracy,Recall,precision,F1-Score and Auc-Score.The highest proportion of Charged-Off clients were correctly classified by Decision Tree and the result showed that Decision Tree performed best with an accuracy score of 87% and 96% recall respectively.There was improvement in the computation time and the model's performance especially with Random forest in the second experiment. This research will further enhance the operations of peer-to-peer lenders to identify potential defaulters and further minimize the risk associated with such decsions.

1 Introduction

There has been a rapid development in the retail finance sector which has resulted in the emergence of both medium and small-scale consumer credit businesses. With the continued growth and adoption of technology in online businesses which continues to provide more opportunities for consumer credit businesses. However, while there have been increased opportunities provided by the growing demand for credits and loans there are also growing concerns about the risks. Credit risk remains one of the major challenges facing banks and other consumer credit business who give out loans and other credit facilities.

Credit Risk as described by Dahiya et al. (2015) refers to the probability (Risk)involved in granting a credit or loan facility to a potential defaulter thereby resulting in his inability to pay back the loan. Banks and consumer credit businesses are constantly faced with

the challenges of managing such credit risk. Therefore, the evaluation of credit risk plays a major role in decision making process in the financial sector. Banks have developed cumbersome and stringent procedures for evaluating the probability of credit return for individuals. This is achieved by utilizing individuals' historical data to evaluate their probability of credit return. Therefore, customers are categorized into groups (Good or Bad) based on their probabilities of credit worthiness. Loan applicants above the set benchmark (high probability) are regarded or classified as being credit worthy(Good) while those whose probabilities of return falls below the benchmark are classified as Not credit worthy(Bad). While the banks have put in place various screening steps alongside stringent conditions which demand for collaterals, other consumer credit businesses have streamlined the process with little or no demand for collateral. This has made the consumer credit businesses more vulnerable and susceptible to financial losses arising from loan/credit default.

According to a publication by Shen et al. (2019) before the adoption of technology in the financial sector, Banks and other financial institutions relied on the use of traditional risk evaluation systems which consists of defined rules and policies that must be adhered to by the officer in charge. The efficiency of the systems are left at the discretion of the officer in charge. These systems and approaches were grossly inaccurate, inefficient and were always marred with biased probabilities.

The advancement in the computing processing capacity coupled with big data has seen its adoption and application in many sectors. The financial sector has also witnessed its own share of the development with the development of various decision support systems, these systems have capacity to process huge tons of data that are generated daily in the financial sector and this forms the basis on which informed and accurate decisions are made. Various researches have been conducted in the financial sector with the application of data mining techniques to develop risk evaluation systems with the probabilities of the system's accuracies being the major metric to either grant or reject such request. Since the adoption of this system, There has been a great turn around in the processing time and the response time for the decision making.

However, concerns have been raised about the biases in the accuracies of this systems which has been attributed to the imbalance in the class distribution of most financial dataset Shen et al. (2019). This phenomenon contributes greatly to Type-II error and it usually occurs when there is an uneven number of cases between the positive cases an d the negative cases of the target variable i.e. when the minority class is greatly outnumbered by the majority classes, the result tend to be influenced by this imbalance and hence the result is in favor of the majority class. Such accuracies does not reflect a balanced consideration of the two classes. The impact of this usually result to misclassification of loan applicants whereby loan defaulters would be wrongly classified as good creditors therefore resulting into financial losses for the financial institution. To overcome this challenge, this research has employed an oversampling technique (Resample).This technique has been widely used in financial dataset analysis to overpopulate the minority class such that we can have a balanced class distribution in the target variable.

Similarly, another notable setback for the application of machine learning in risk evaluation is the issue of high dimensionality of the financial dataset. For this reason, this research has implemented a feature selection technique to obtain the important variables that have the most impact on the target variable and use it to optimize the developed models. The models implemented include Logistic regression, Random Forest, Naive Bayes, decision Trees and Ada Boost.

1.1 Research Question

One of the primary goals of this research is to improve the risk evaluation decision making processes in banks and other businesses in the finance retail market. This can be achieved by optimizing the credit risk evaluation decision support systems used to screen loan applicants by way of analyzing individual historical data and generating a probability of credit return. The output of these systems is a binary classification (Good or Bad) of the applicants according to their credit worthyness. To achieve this goal, dataset was obtained from a peer-topeer lending business. This dataset was analyzed and trained with five classification models. These models were chosen and used to answer the research question based on their proven efficiencies as gathered from the literature.

RQ - *"To what extent can the efficiency of a supervised machine learning model be optimised to evaluate the probability of default in peer-to-peer lending business?"*

Various classification algorithms have been applied to finance dataset especially for profit scoring problems. The accuracy of these model is very key in the decision-making process. Therefore, a percentage increase in the probabilistic accuracy of a classification model will translate to a decrease in a potential loss that may arise from granting a loan to a loan defaulter.

1.2 Research Objectives

In order to provide answers to the research question above, the following objectives will be followed. This includes;

- A comprehensive and critical review of related work was carried out. This was achieved through a review of literatures as contained in articles, conference publications and reputable journals in the field of data mining and machine learning.
- Obtain historical transaction data from a credible open source.
- Prepare the dataset for use by performing data preprocessing.
- Generate preliminary insights from the dataset by performing EDA (Exploratory Data Analysis).
- Implement oversampling technique (Resample) to overcome the imbalance in the class distribution.
- Implement the supervised classification models using the five classification algorithms fitted with the training and testing dataset.
- Evaluate and compare results from each model using some evaluation metrics (Accuracy, precision,Recall,F1-score, and Auc-score).
- Implement a feature selection technique to obtain the important features in the dataset.
- Implement the classification models fitted with the newly generated datasets and Optimize the hyper-parameters of the models using Gridsearch.

- Evaluate and compare the performance of the models based on the evaluation metrics stated in both instances.

2 Related Work

Over the years, the evaluation of credit risk has been one of the major challenges facing banks and other financial institutions. Various systems and approaches have been adopted to overcome these challenges. Credit risk evaluation models are designed to evaluate applicant's individual characteristics as obtained from their historical transaction data. this is used to evaluate their probability of credit risk (i.e. if they are credit worthy or not). These systems are very important in the decision making processes in banks and other financial institutions. High efficiency of these systems can help to increase profitability and also reduce losses which could arrive from inaccurate decisions taken by granting loans to individuals that would not pay back.

Credit risk evaluation systems are used to evaluate the level of risk by individual customer based on their historical data. The probability of risk generated from the models are used to classify applicants into high risk or low risk category. Applicants with high probability of risk are classified into the 'bad' class while those with lower probability of risk are grouped into the Good' classes. Credit risk evaluation systems needs to perform at the highest level of accuracy possible hence the continued interests in ways to optimize the systems through various researches.

2.1 Credit Risk Modelling

There have been a rapid development in credit risk modelling thereby becoming an important factor in the financial sector especially in the management of financial risks.

Dahiya et al. (2015) developed a hybrid credit risk evaluation model. The main goal of their work was to optimize the model by overcoming the imbalance distribution in the dataset by introducing an under sampling technique. Similarly, the hybrid model was achieved by combining two single models together. The algorithms used in this research are Naïve Bayes, RBF, Logistic regression, Decision Tree c4.5 and MLP. Separate dataset was used to evaluate the single model and it showed that the MLP model outperformed the other models in terms of accuracy. Similarly, the hybrid model was created by combining MLP with MLP. The overall outcome of the models revealed that the hybrid model (MLP with MLP) outperformed all other singular models with an accuracy of 86percent. However, the main limitation of this study is that the under sampling technique employed to balance the dataset could result to the loss of information and also this was done manually without using any recognized statistical technique.

Zhu et al. (2018) proposed a hybrid Relief-CNN deep learning credit scoring model for consumer credit business. The model was developed with the aim of leveraging on the high efficiency and outstanding ability of the deep learning techniques to achieve a higher accuracy in credit risk management. Convolutional Neural Network algorithm was utilized as the base model while the feature selection algorithm Relief. The feature selection algorithm acted as a catalyst to lessen the computational weight of the base model. This was used to identify and obtain a subset of the variables which are most relevant to the target variable. The model was empirically evaluated and compared with other traditional classifiers using dataset from a Chinese finance company. The results of the experiment showed that the hybrid model with 69% Auc-score outperformed both

logistic regression and random Forest classifiers with 52% and 60percent respectively. However, the researchers failed to address the imbalance in the dataset thereby making the probabilistic accuracy biased in the direction of the majority class.

Wei et al. (2019) proposed a novel double layered ensemble credit scoring model built on backflow learning integrated with noise filtering. The model was aimed at amplifying the strength of the base classifiers using the backflow learning such that the classifiers can relearn the misclassified cases in the dataset noise adaptive technique was integrated to the model using the isolation forest algorithm. This was used to identify and boost the noise in the training data set by recalculating the outliers in the training dataset. Five base classifier algorithms were integrated in the ensemble Modell, they are XG boost, LDA, Random forest, Decision Tree and support vector machine. The model was tested on three different dataset and the outcome showed a satisfactory performance compared with the traditional classifiers. However, the multi layered approach further reduced the interpretability of the model and also the model proved to be time consuming and computationally expensive due to the quadratic increase in the number of combinations for the base classifiers.

Various ensemble and hybrid credit risk evaluation models have been developed with proven abilities in credit scoring. However, little attention has been paid to the combining different classification output combination techniques with underlying feature selection and data filtering methods. Ala'raj and Abbod (2016) proposed a consensus approach for classifier selection for a hybrid credit scoring model. The approaches were based on defining new combination rule for combining two preprocessing methods using GNG and MARS and also using a consensus approach to combine classifiers. The aim of this approach was to optimize the predictive ability of the model. The study made us of five base classifiers which include Neural Network, Support vector machine, Naïve Bayes, Random Forest and Decision Trees. The study reported a slight improvement in the predictive accuracy of the hybrid model over the base classifiers. The main limitation of this study just like other ensemble and hybrid models was that it was too complex to interpret with the output being a floating point ranking rather than a fuzzy matrix.

Xia et al. (2018) proposed a novel approach to optimizing the predictive ability of credit risk systems by developing an ensemble stacking model. The model is an integration of the bagging algorithm with the stacking method. For the base classifiers, Random forest, GPC, Support vector machine and Xgboost were utilized due to their ability to maintain a balance between the accuracy and efficiency metrics. The outcome of the model showed a great improvement in the performance of the BStacking model over the traditional classifiers with an improved accuracy from 72% to 2% respectively. It is also noted that interpretability was sacrificed for accuracy in the model.

In an attempt to overcome the challenges with ensemble models as highlighted by Xia et al. (2018). A novel ensemble approach was proposed by Florez-Lopez and Ramon-Jeronimo (2015) to improve the interpretability and accuracy of ensemble models by merging decision trees with a correlated adjusted decision forest. This model embraces the diversity of the algorithms which limits the number of decision rules to a manageable size hence not compromising the interpretability of the model. The resulting model was trained and fitted on the German financial dataset and the result evaluated based on accuracy and Auc score. The outcome showed that the there was a drastic reduction in both the Type I and Type II error with 0.43 and 0.1 respectively.

Mai et al. (2019) carried out a study to include the use of textual data in the analysis of credit risk management. This study was aimed at improving the predictive accuracy

of the decision support models. The strategy proposed the use of deep learning model to predict bankruptcy using features extracted from textual data. The model consists of a convolutional neural network ask the base classifier and a mixture of both structured and unstructured data. Features were extracted from the unstructured data via the layers of the neural network.AC was used to evaluate the performance of the model in order to determine which of the underlying architectures performed best. The result showed that the CNN architecture outperformed the average embedding model with 0.78 and 0.71 Auc-scores respectively. However, the limitation of this study was that the dataset used for the experiment was unstructured and thereby making it less interpretable. Also, the accuracy of the model was not evaluated hence raising concerns about the performance of the model.

Similarly, Masmoudi et al. (2019) developed a novel Bayesian network based model for predicting loan default payment using latent variable. The aim of this model was to establish the relationship between the applicant's attributes and the terms of the loan contracts .The authors believed this would be a major pointer in calculating the probability of defaulting a loan payment. The model is capable of evaluating the probability of default while taking into consideration all other factors. The learning of the dataset parameters is handled by the expectation maximization algorithm defined in the algorithm which is also used to calibrate the model. The authors conclude that the Bayesian Network model performs better when applied to large datasets compared to traditional classifiers.

In a bid to further improve the predictive ability of credit scoring models, attempts were made by Zhang et al. (2019) to develop a multistage hybrid model. The model strategy was to create an optimal subset of features and classifiers which would then be integrated in an ensemble model. Therefore, the feature selection selected optimal features and likewise the classifier selection to select the optimal classifiers. The Genetic algorithm was highly enhanced with multiple layers of filters to enhance the feature selection. The authors believed that the performance of the model was significantly improved due to the multi-layer selectors to select optimal features for the model. However, the drawback for this study is that the model was not tested on a small dataset and the efficiency of the heuristic algorithm can be optimized.

A comparative study was carried out by Wang et al. (2011) to evaluate the impact of the three ensemble methods(Stacking, Bagging and Boosting) on the predictive abilities of some classifiers. The authors proposed an ensemble model with Decision Tree, Support vector machine, Artificial Neural Network and LRA as the base classifiers. Outcome of the model when tested on a Chinese bank dataset revealed that base classifiers with bagging ensemble methods perform better in terms of accuracy. In this case the decision Tree with the bagging ensemble method had a significant improvement in its predictive accuracy compared to the Svm, LRA,NN models respectively. However, there was an increase in the Type I error of the Support Vector classifier.

Similarly, Feng et al. (2018) proposed a novel approach to managing credit risk with a research study aimed at reducing or eliminating Type I and Type II error. The proposed model would select classifiers based on their relative cost of both Type I and Type II error. A soft probability function is used to calculate the interval probability of default between the selected classifiers. The experiment showed that the approach improved the predictive accuracy of the model by returning a lesser Type and Type II error. However, the model is not economical in terms of the computation time and also it lacks interpretability like other ensemble models

2.2 Overcoming Imbalance Credit Dataset

Imbalance class distribution has remained one of the major drawbacks in the advancement of credit risk evaluation. Significant researches have been carried out adopting different techniques to overcome the problem.

A quantitative credit risk model was proposed by Shen et al. (2019).An optimized BP neural network as the base classifier was used to build an ensemble with Adaboost. This was optimized by identifying the best optimum deviations and best weights in the base classifier (BP neural network) using a PSO (Particle Swarm Optimization) algorithm. The novel ensemble model employed an oversampling technique (SMOTE) to deal with the imbalance dataset. The proposed model was tested on real dataset from the Australia and Germany and its performance evaluated based on the Auc score and accuracy. The outcome revealed that the model outperformed seven other traditional classifiers with and accuracy of 78percent and 90percent on the two datasets respectively.

Zhu et al. (2019) proposed a Random Forest model for predicting loan default. The model employed an over sampling technique to overcome the imbalance in the dataset synthetic Minority Over sampling Technique was adopted to repopulate the minority instances in the target features. This was achieved by adopting the nearest neighbor algorithm for which the minority instances are repeated K number of times for every instance of the majority instance encountered. This would be done according to the specified proportion. The performance of the model was tested on a lending dataset and compared with Logistic regression, Decision Tree and Support Vector Machine classifiers. The result showed that the accuracy of the model was significantly improved by the balancing done. The Random forest performed best with an accuracy of 98percent over Logistic regression, SVM and DT which had 73%,75% and 95% respectively.

Kim et al. (2015) identified imbalanced dataset as a major factor degrading the efficiency of loan default classifiers. Therefore, the authors proposed a GMboost approach to overcome the imbalance. The Geometric mean of the error rate between the two classes is evaluated and considered in calculating the accuracy of the model. The model was applied to predict bankruptcy and the outcome compared with AdaBoost and cost sensitive boosting showed a significant improvement. However, the drawback observed in this study is that the boosting algorithm degrades the accuracy of the classifiers when there are outliers in the dataset.

In a different dimension to optimize credit risk models, Papouskova and Hajek (2019) investigated the impact of two important parameters (Loss Given Default and the Exposure at Default) on the probability of default. To this effect a Double stage integrated class-imbalance ensemble model was proposed to estimate the probability of default. The experiment of the proposed model on a credit dataset showed that the Random Forest had the highest performance as a base classifier with 0.80 R square and 38percent Mean Absolute Error. The drawback of this model is that the Loss Given Default was not calculated but rather a random number was assigned to it.

Wang et al. (2018) proposed a novel behavioral credit scoring model aimed at dynamically predicting the monthly probability of default for individuals. The model considers the behavioral characteristics of individuals in estimating their probability of default. The proposed EMRF(Ensemble Mixture Random Forest) model utilizes a random forest to evaluate the probability of default. An evaluation of the model was carried out on a Chinese peer-to-peer lending dataset and the results compared with Logistic regression, Cox proportional hazards model and standard mixture cure model. The outcome showed

that the model was able to predict the Probability of default over time. It also showed significant in terms of predicting when customers are likely to default on their loan.

While several researches are aimed at improving the precision of risk models at the expense of interpretability, Xu et al. (2017) focused their study on improving the performance and interpretability of credit risk models. The authors proposed an improved model based on RIPER algorithm. The model employed oversampling to balance the class distribution using SMOTE. Hence the resulting SRIPER(Smote Repeated Incremental pruning for Error Reduction) model is generated by inheriting the combination rules generated by the ripper and the SMOTE. The proposed model was applied on a Taiwanese credit card dataset and compared with other existing models. The outcome showed that proposed model had a significant increase in accuracy over the existing models (DT, RF, SVM) and RIPPER.

2.3 Handling High Dimensionality of Credit Dataset

Feature selection is an important data preprocessing technique that is used to evaluate the relative importance of every attribute in a data mining task. This would significantly have impact on the efficiency of the model by eliminating the redundant variables thereby reducing the training time, reduce overfitting and improve the model accuracy.

A comparative study on the combination of supervised and unsupervised learning was carried out by Bao et al. (2019).This study was aimed at improving the efficiency of improving credit risk models. A wide range of diverse classification algorithms(Decision Tree, SVM, Random Forest, KNN, Artificial Neural Network, Gradient Boosting decision Tree and logistic regression) were adopted as the base classifiers in the model. A consensus model was built on the base classifiers using the unsupervised methods. The model was tested on the Chinese peer to peer credit dataset and evaluated with accuracy, recall, precision and especially the Type I & II errors. The outcome revealed that the GDBT model permed best with the highest MCC score of 0.66.

The study carried out by Dahiya et al. (2016) aimed at improving the accuracy of credit risk evaluation model proposed a hybrid MLP(Multi-Layer perceptron) Neural Network model. The study developed three models, The ordinary MLP model, the Feature selection MLP and the Bagging MLP model. The three models were tested on the Australian and German credit dataset to predict the probability of default and the outcome showed that the accuracy of the ordinary MLP model was significantly improved by the Feature selection technique and further improved by the hybrid model. The MLP had an accuracy of 73% while the FS MLP and Hybrid MLP had 74% and 80% accuracies respectively.

Chornous and Nikolskyi (2018) proposed a business specific feature selection technique to improve the classification accuracy of credit scoring models. The hybrid ensemble model was built on four classification algorithms (SVM, KNN, Decision Tree and Gradient Boosting Machine) as the base classifiers. The features were selected based on the relative voting of the Chi-square coefficient, mean decrease Gini and the information Gain. The study concluded that the efficiency of model is improved by the increased ability to identify more user defined information.

Ma et al. (2018) proposed a double edged approach to data cleaning to optimize credit risk model. The proposed approach was based on multi-dimensional and multi-observation data cleaning method. The model is comprised of the LightGBM (Gradient Boosting Machine) and the Xgboost. The study reported that the LightGBM uses its

features for data and voting parallelism to select important features for the model. Features which were less relevant to the target variable were deleted. The proposed model was tested on a US peer to peer lending dataset and the LightGBM model outperformed the Xgboost with an accuracy of 80.1% and Error rate of 19.9%.

In the study carried out by Xia et al. (2017),a Bayesian hyper parameter optimized Boosting Decision tree model was proposed for credit scoring. The model was aimed at improving the predictive accuracy of the credit scoring model by optimizing the hyper-parameter of the base learners and also improve interpretability of the ensemble model. This was achieved by introducing a model-based feature selection technique to select the important attributes and remove the redundant ones from the dataset. Top ten features were selected according to the ranking of their importance scores. The model was tested on three dataset and the outcome revealed that that there was improvement in the predictive e accuracy of the model. Similarly, the feature selection also improved comprehensibility of the model with the importance ranking score.

Tripathi et al. (2018) recognizes the fact that high data dimensionality or redundant variables may degrade the predictive accuracy of a credit scoring model. Therefore, the authors proposed a cluster-based feature selection combined with an ensemble classifier model. The cluster of features are created by evaluating their correlation coefficient and then ranked and selected based on their ratio value closest to one. Heterogenous ensemble was used for the selection of the base classifiers consisting of Naïve Bayes, PNN, MLFN, RBFN and Decision Tree. The model was evaluated on three different dataset and compared with the existing feature selection methods. The result showed that the clustered feature selection approach improved the model and therefore outperformed the existing models.

Ha et al. (2019) proposed an advanced credit risk evaluation model capable of assessing the returns and risk of individual loans on a peer-to-peer lending network. Important features are selected based on the ranking of their Root Mean Square Error (RMSE) evaluated by the restricted Boltzman Machine. This enabled the elimination of the redundant feature and hence reducing the computational complexity of the model. Artificial Neural Network, K-Nearest Neighbor, Random Forest, Logistic regression and Linear Discriminant analysis were all used as base classifiers. The evaluation of the model was carried out on three different credit datasets, the Australian, Lending club and German datasets. The outcome revealed that the LDA model with optimum twelve feature subset performed best with 86.09percent over the Australian dataset as against the ordinary model with 85.8percent.

The study carried out by Addo et al. (2018) was aimed at modelling credit risk using deep learning techniques. The model employed feature selection to reduce the high dimensionality of the dataset by selecting the top ten attributes for the modelling process. The experiment was carried out on both deep learning algorithm(ANN) and other classification algorithm (Xgboost, Logistic regression and Random Forest).The outcome showed that Xgboost had the best performance over the optimum feature subset with an accuracy score of 98% while ANN had 79%.The study revealed that despite having the top features does not guarantee the best result, therefore it depends on the stability of the base model.

Munkhdalai et al. (2019) carried out a consumer finance survey to generate input to do a comparison between human expert-based approaches and machine learning approaches. Due to the high dimensionality of the data, feature selection techniques were applied on the dataset. First, correlation, hypothesis test and Random forest feature importance

were applied. Then a proposed novel Random forest NAP was applied as well. The dataset obtained from each feature selection method was applied a model built on Logistic regression, Support Vector Machine, Xgboost and Neural Network. The outcome showed that the gradient boosted model and the Neural network model trained on the NAP selected features performed best with an accuracy of 86%, while the logistic regression and the SVM models had 79% and 58% accuracies respectively.

3 Methodology

The Main focus of this research is to optimize the efficiency of the credit risk systems used in evaluating the credit worthiness of loan applicants. This would be achieved by ensuring a faster processing time in estimating the probability of default for individual applicant. The computational complexities associated with ensemble techniques can be overcome by ensuring that only the important attributes that have direct impact in estimating the ability of repaying back a loan are used for the process. This would help to reduce redundancies, save computation cost and also improve the performance of the model. Informed and accurate decision making is very key to the profitability or loss of businesses in the retail financial market, hence the need to maintain the balance between accuracy and interpretability in choosing the machine learning techniques for the model. The scope of this research is focused on the Peer-to-peer lending businesses where loans and credits are provided to applicants with virtually no collateral other than a verbal agreement and the informations provided by the applicant. To achieve the goal of this research, CRISP-DM (Cross Industry Standard process for data Mining) methodology was adopted to ensure a controlled and organized processes. This aligns with the objectives and requirements of this research. The five stages of the methodology are Business understanding, Data understanding, Data Preparation, Modelling and Evaluation. This is highlighted in the figure below;

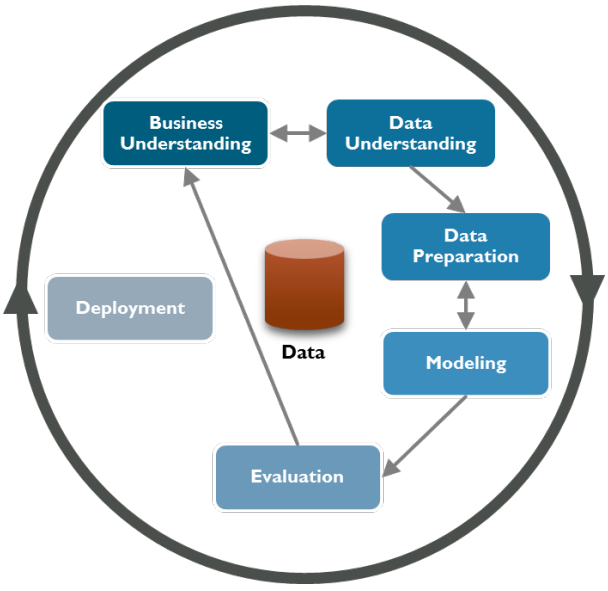

Figure 1: Adopted Credit risk Evaluation Model

3.1 Business understanding

Peer-to-Peer is an online lending platform whereby individuals and businesses can invest and access loans directly without going through the traditional banks. This form of lending has gained popularity due to the low interest rates it offers to borrowers and the guarantee of high return on investments for the lenders Wang et al. (2018).This is made possible by the fact that the businesses operates virtually online and therefore less overhead cost is incurred when compared with the overhead cost incurred by regular banks. Peer-to-peer lending is exposed to high risk of loan default due to the absence of collateral from the borrower compared to the traditional banks.

Lending Club is the biggest Peer-to-peer lending company in the United States of America with its's corporate headquarters located in California. The company is the pioneer lender to be traded on the secondary market of the US Security and Exchange commission where its loan offers are publicly traded. As at 2015,over 16billion dollar worth of loans had been transacted over its platform.

The lending process has been simplified to ensure that an application is completed in minutes. Lending Club serves as an intermediary between the borrower and the investor . The borrower completes an application form containing the amount, purpose and other all other necessary information that are required, then the company leverages on Data Mining technology to evaluate and screen the borrowers request. The various levels of risk ranges from A - G and the interest rate is fixed according to the level of risk, i.e. The higher the risk associated with the borrower the higher the interest rate. The loans are funded by investors who buy into Lending Club Bonds. Therefore, the investors also share in the liabilities of the company.

3.2 Data understanding

In accordance with the goal of this research to evaluate the probability of default and accurately classify applicants based on their ability to pay back a loan. This would be achieved by employing a supervised learning approach that builds on important attributes that can be used to evaluate the credit worthiness of an individual. Dataset that meets the requirement of this research was downloaded from LendingClub [1] . These were historical loan dataset generated between 2007 – 2018 from lending Club. The zipped folder contained two csv files; "accepted_2007_to_2018Q4.csv" contained accepted loan data from the year 2007 to 2018 and it contained approximately over Two million rows of data and 151 attributes, while the second file "LCDataDictionary.xls" contained the data description. The data preprocessing and transformation was done in jupyter notebook using python. The data was prepared to meet the business requirement of the research.

3.3 Data Preperation

This simply refers to the data preprocessing activities that were carried out on the raw dataset. The data preparation activities carried out on this data include cleaning whereby unwanted entities like symbols were removed from the dataset. Also, missing values were replaced with mean values for continuous variables while Range values were used to replace missing categorical values. Transformation of categorical values was also carried to ensure that all features were in a numerical format usable for the model building.

[1]LendingClub: `http://www.lendingclub.com/info/download-data.action`

Removal of constant features and empty columns was done as part of the data preparation. A detailed explanation of all data preparation activities is given in the implementation section.

3.4 Modelling

In a bid to achieve one of the objectives of this research to develop a supervised learning model for predicting loan defaulters and Non- defaulters, five machine learning algorithms have been used, they are Decision Tree, Random Forest, Logistic Regression, Naive Bayes and AdaBoost. These models have been utilized because of their proven efficiencies and comprehensibility as gathered from related work in the literatures. This agrees with Florez-Lopez and Ramon-Jeronimo (2015) that highlighted that simple machine learning models are easy to understand and hence they are preferred and easily adopted by financial managers.

AdaBoost : This is a powerful boosting algorithm popularly used for binary classification Kim et al. (2015). Boosting algorithms as one of the family of ensemble learners have strong learning ability and are able to generate new and accurate learners by linearly combining several weak learners. This is good for imbalanced dataset, where opportunity is given to minority class samples. Boosting algorithms ensure that every new classifier is built on the outcome of the previous classifier such that the misclassified samples are learnt and reclassified by the newly generated classifier.

Decision Tree : These are hierarchical classifiers with simple structure that enhances interpretability Mantas et al. (2016).In decision Trees, each Node represent an attribute and beach branch represent the value of the variable/attribute. This made it suitable for both binary and multi-class classifications using hierarchical representation Dahiya et al. (2015). Decision Trees have a strong ability to reduce overfitting in model through the post-pruning process.

Random Forest : This belongs to the ensemble learning family of algorithms and can be applied on both Regression and classification problems. Bao et al. (2019) suggested that Random forest can be regarded as an updated version of decision tree. A random Forest Model utilizes the bootstrap algorithm to generate sample subset from the training subset and then trains the decision trees on this subset.it combines both bagging and random subspace technique in generating the subset for the base classifiers. Random forest also follows the operation approach of decision tree by construction multiple Trees and aggregating the mean of the individual tree. This algorithm is powerful in dealing with high dimensionality dataset and also quite efficient in handling missing values and imbalanced dataset Zhu et al. (2019)

Logistic Regression : This has been widely used on classification problems and it explains the relationship of a set of variables on a categorical output variable Dahiya et al. (2015).According to research logistic regression is regarded as the basic standard for credit risk evaluation problems. The growing popularity in the adoption of logistic regression can be attributed to the simplicity of its implementation.The logistic regression function is used to calculate the relationship between the predictor variables and the outcome rather assuming a linear relationship as it is in linear regression. Logistic calculates the conditional probability of belonging to a particular class by evaluating the log value of the probability ratio between the two possible outcomes in a binary classification problem.

Naive Bayes : This is one of the most widely used algorithms in statistical pattern recognition. The algorithms utilizes the Euclidean distance in estimating the distance

between the new instance and a previously trained sample. Naive Bayes has largely been applied in constructing credit scoring models and operates by creating a new instance of the sample subset according to the nearest subset of the previously labeled instance. The algorithm is efficient in dealing with high dimensional data.

3.5 Evaluation

A comparative analysis was carried out to evaluate the performance of the developed models on the dataset. To evaluate the model, the dataset was split into training and testing set using the ratio of 70:30 respectively and this would help to reduce overfitting of the models. The model learnt from the training set data and later tested its performance with the test set data. The goal of this research is aligned with the detection of potential loan defaulters and non-defaulters. Therefore emphasis was placed on limiting misclassification error which could result to huge financial loss i.e. Type II error will result to loss of capital if a defaulter is granted a loan and unable to pay back likewise a Type I error will result to loss(opportunity cost) when a Non-defaulter is denied a loan. The overall performance of the models were evaluated using the following metrics; Accuracy (total number of correctly classified samples), recall (the ratio of the correctly classified loan defaulters) ,F1 score (the ratio of precision to recall), AUC (the overall summary of the model's performance) and finally the Precision (The True positive)

4 Implementation

The implementation of the model was carried out in Jupyter notebook using python programming language. This was adopted due to its efficiency and flexibility in handling High dimensional data. Several inbuilt python libraries were used for the preprocessing and in some cases, libraries were installed to complete the given task. The detailed description of the implementation process is discussed in this section of the report

4.1 System Environment Setup

This simply describes the various software and hardware tools that were utilized in the various stages of this project. Jupyter Notebook was used for the implementation of python codes which was installed on a windows Operating system. Some pre-installed libraries like seaborn, pandas, NumPy etc. were imported while new library like the imbalance-learn had to be installed. Some other minimum requirement for the systems are Ram,2GB,processor,2.4Ghz.All other installations are discussed in the configuration manual.

4.2 The Process Flow

The diagram below is a graphical representation of the implementation structure of this project.All the implementatin was carried out in Jupyter notebook using python programming language.

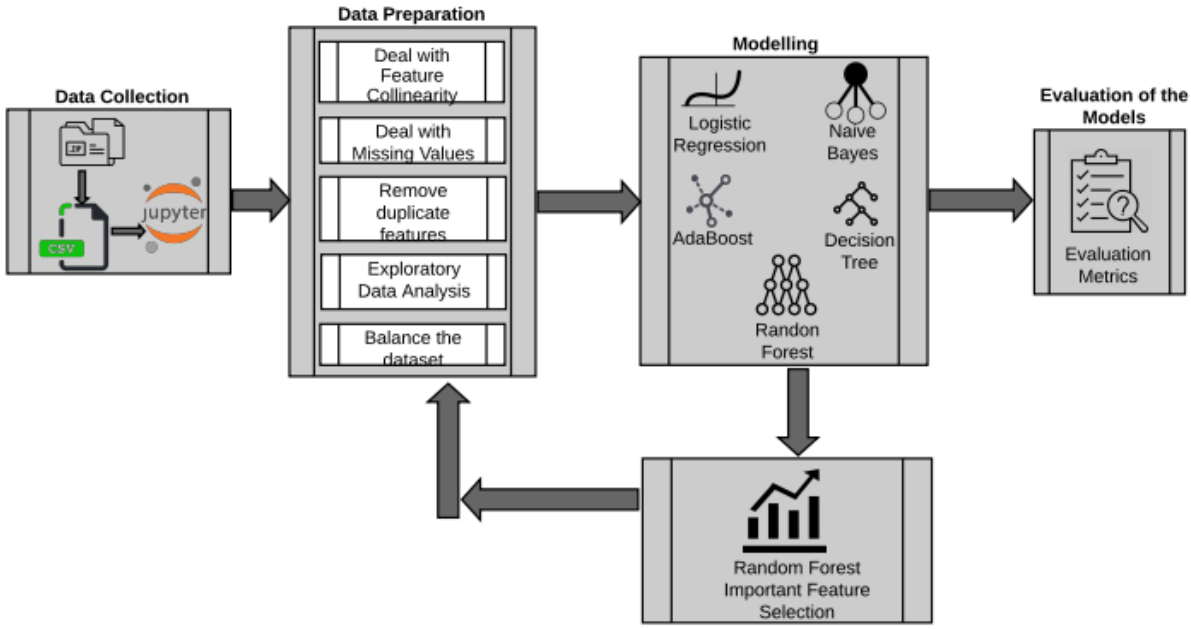

Figure 2: The process Flow Diagram

4.3 Collection Of Credit Dataset

Careful consideration of ethics was maintained in the process of obtaining the dataset used for this project. Therefore, data was obtained from a credible open source repository. The dataset was downloaded from lendingClub in a zipped format due to the large size of the file. The csv file containing the accepted loan dataset weighed over 1.8GB containing approximately two million rows and one hundred and fifty one features. The data was extracted and saved in the working directory for this project.

4.4 Dataset Preparation

Data preprocessing was carried out to clean the data and ensure there is uniformity in the models' input . The data processing was done in Jupyter Notebook using python libraries. The dataset was loaded into jupyter using pandas library. However due to the huge size of the data, only records for that relates to the class of interest (Charged Off and Fully Paid) were extracted and this sums up to approximately 1.2million rows. similarly, the Data dictionary containing the variable description was also loaded. This is to enable easy understanding of the variables.

The first action carried out on the data was to ensure that all the features matches the ones given in the data dictionary. This was done using python's regular expression to remove white space and replace them with underscores. Also features whose names were misspelled but mean the same thing were replaced with the correct one. The dataset with correct feature was saved in a new data frame as "loan_extracted" containing 151 features and Approximately 1.2million rows.

Secondly, the data types of the categorical variables were checked to ensure they were coded appropriately. Also features "emp_length" and "id" were transformed to float type because they contained only numbers. Also, a function was defined to check for missing values in the dataset.This function returns a table of features with the percentages of

missing values in each feature. All features with missing values above 48% were dropped from the dataset.This was done because of the huge size of the data still available. With this done the number of remaining features was prunned to 93.

Similarly,a function was defined to search for constant features i.e features that contained a single values across the dataset,this would not make any significant impact on the machine learning model we try to build.Five features were dropped as result of this and the number of features reduced to 88.The dataframe was also checked for duplicate rows. i.e rows containing thesame values as other rows.Each of the remaining features were examinedindividually and in relation to the loan status (Target variable) feature.Also some features like the "loan_amnt" were transformed to the log value due to its varying figures.

Multicollinearity test of the variables were carried out in two stages. First was a Pearson correlation coefficient test on the numerical features and the first pair of each numerical feature with correlation coefficient score higher than 0.9 were dropped. Secondly to reduce multicollinearity among the categorical features, the Cramer's v correlation coefficient was generated which was evaluated through the chi-square contingency table. The second feature of each pair of categorical features having an absolute value above 0.9 was dropped as well.These combinedpreprocessing activities helped to remove unwanted features from the dataset.

4.5 Exploratory Data Analysis

To understand the relationship between the features in the dataset and also to gain more insights from the data, preliminary exploratory analysis was carried out on the dataset. This included both univariate and bivariate analysis on the features. The exploratory analysis on the Loan amount revealed that the loan amount ranges between One thousand dollar and forty thousand dollars with the average distribution being around Fourteen thousand dollars.

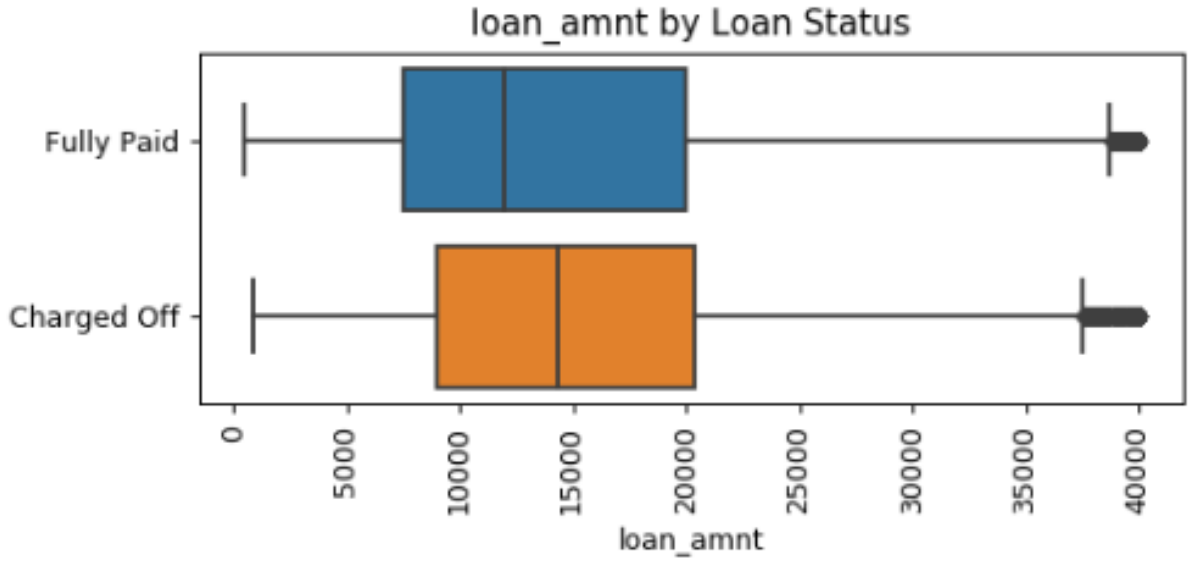

Figure 3: Loan Amount Distributed by Loan Status

Similarly, the boxplot analysis on the interest rate revealed that the rate for most of the loan applications ranges between eight and sixteen percentages respectively. This

also showed that the highest interest rate is a little above 30%.This further suggests that a higher percentage of the loans are high risk and they have a higher probability of defaulting or charging off.

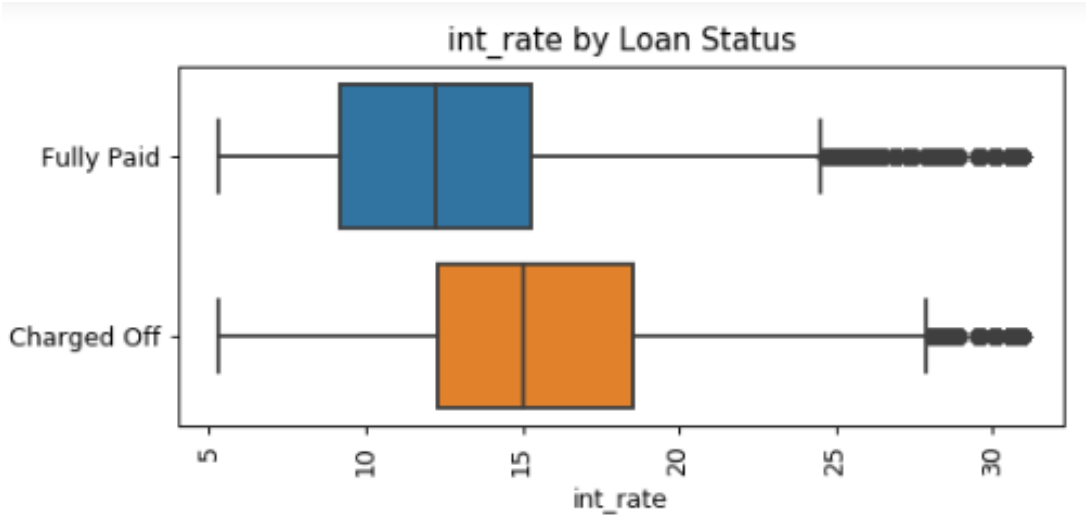

Figure 4: The Distribution of Interest rate by the loan status

Also, the target variable was explored. The distribution of classes in the target variable (loan_status) was explored through a bar chart and this revealed that only two classes have significant cases. This further showed that the data is highly imbalanced with a ratio of 80:20 in favour of the fully paid loans. Further encoding was don on the target variable to transform it into binary output. Similarly, the purpose and loan status analysis, revealed that debt consolidation is the most reason stated for requesting a loan while wedding is the least purpose for loan request. This would be discussed further in the section below.

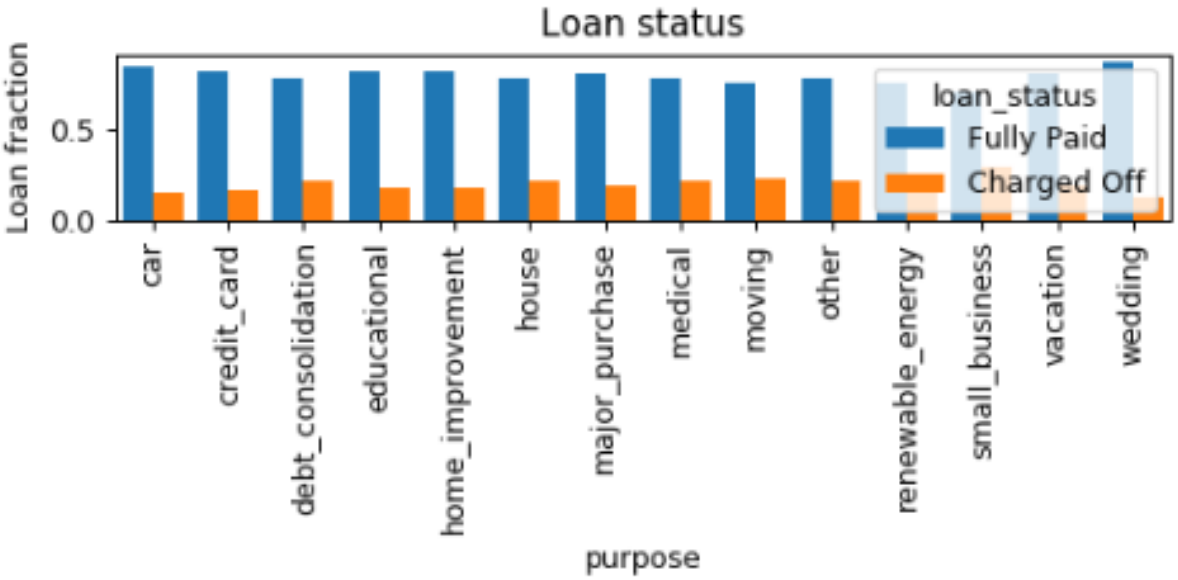

Figure 5: The Distribution of loan purposes

4.6 Feature Engineering

The insights gained from the preliminary exploratory analysis revealed that the target variable consisted of seven classes initially. Only two classes are relevant to the business understanding of this project i.e. the "fully paid' and "charged off". Therefore, only records related to these two classes were extracted for the remaining analysis in this project. Also , this revealed a significant class imbalance between the two classes with the charged off having only about 20% of the distribution. An Oversampling technique (Resample) has been employed to overcome the data imbalance, this would help to avoid bias where the majority class would dominate over the minority class.

Imbalance class distribution has been a major challenge with financial dataset analysis and several studies including Dahiya et al. (2016) successfully applied SMOTE technique to balance the dataset by creating multiple instances of the minority class using an underlying algorithm. The imbalanced-learn package was installed from the pypi repositories using the command "pip install -U imbalanced-learn". An oversampling technique (Resample) was applied on the dataset to get an equal distribution of "Fully paid" and "Charged-off".

Also, the remaining features were grouped into two based on their data types. The numerical variables were explored by evaluating the Pearson correlation for each pair of features. For each pair of highly correlated feature, the index feature is dropped. Similar action was carried out on the categorical features using the Cramer v correlation, this was applied using the chi-square method. The index feature of every highly correlated pair was dropped. This was done to avoid multicollinearity. Also, outliers were eliminated from features like "annual_inc" and other features with multiple categories were removed. The other categorical features were transformed to numbers. The data was checked again to ensure that all features were in the correct format and ascertain that there were no missing values.

4.7 Modelling

In preparation for the building of the models, the train-test-split method was imported from the model_selection library to split the dataset using a ratio of 70:30. This means that 70% of the data was used for training the models while the remaining 30% was used for testing the models. This was done to avoid overfitting the models. With the existing imbalance in the target variable, Resampling technique was applied on the training dataset to balance the class distribution in the target variable. This ensured that the models were trained with a fairly unbiased dataset that has equal number of "fully-paid' and "charged-off" instances.

Five classification algorithms were implemented in this project, they are logistic regression, Decision Tree, random Forest, AdaBoost and Naïve Bayes. To maintain data consistency, the dataset was scaled using Standardscaler() method. This was to ensure that no feature had arbitrary huge value and all features were recalibrated to have values between -1 and 1. For the first experiment, all the features in the preprocessed dataset was used and each model was fitted and trained with the training set of the data. This was to ensure the model learn some rules about the dataset. The Decision tree classifier was trained with the training dataset and i'ts criterion set to entropy. similar criteria was selected for the Random Forest. The Naïve Bayes model was also implemented through the GaussianNB method(). With the training completed successfully, each model was

fitted with the test data set and evaluated using several metrics (Accuracy, Precision, Recall, F1 score and AUC Score).

The second phase of the experiment was carried out to investigate the impact of using fewer features on the the performance of the classification models.Therefore a feature selection technique was implemented to generate a new subset of the dataset containing ten important features from the dataset. A Random Forest variable importance plot was applied to achieve this. These were preprocessed, cleaned and checked for any inconsistencies. The new dataset was split into training and testing set using a split ratio of 70:30 percent respectively. Each model was fitted and trained with the training set of the new data. Attempt was made to optimize the performance of the models by applying GridsearchCv technique. This was implemented to optimize the hyper-parameters of the models. The performance of all five models were evaluated on the test set using the evaluation metrics earlier stated.

5 Evaluation And Discussion

After building the five classification models with the training dataset across the two experiments, The performance of each model was evaluated based on the optimized hyper-parameters and 10 fold cross-validation was applied using five different evaluation metrics. This is to ensure a broader basis of comparison for the models. The models were evaluated based on Accuracy, precision, Recall, F1-score and Auc Score.

Accuracy is one of the most widely used metrics for evaluating classification models either multiclass or binary classification. This can be associated to it's ease of comprehension. Accuracy measures the proportion of True instances (TP + TN) that were predicted by the model among the overall instances predicted. i.e. ACC = (TP + TN) / (TP + FP + TN + FN)

Recall is another reliable performance metric that is used to evaluate the proportion of the actual positive instances that are predicted correctly as positive. i.e the proportion of actual defaulters that are predicted as defaulters. It is mathematically expressed as; REC = (TP) + (TP + TN).

Precision simply measures the performance of the model in terms of the proportion of the cases predicted as positive that are actually positive. i.e. the proportion of predicted defaulters that are actually defaulters. It is expressed mathematically as; PRE = (TP) / (TP + FP).

F1-Score measures the harmonic mean between the recall and the precision scores of a model. This metric helps to find a fair balance between the precision and recall of a model. It can be expressed as 2 * ((PRE * REC) / (PRE + REC)).

AUC Score is quite different from the other metrics. It measures the probabilities of the predictions instead of their absolute values. AUC reveals or measures the quality of our model prediction by showing how distinct a positive class is different from a negative class.

5.1 Performance of Models in Experiment 1

Models	Accuracy	AUC-Score	Precision	Recall	F1-Score
Logistic	0.66	0.71	0.66	0.65	0.65
R. Forest	0.66	0.72	0.65	0.69	0.67
D.Tree	0.87	0.87	0.81	0.96	0.88
AdaBoost	0.66	0.71	0.65	0.67	0.66
Naive B	0.65	0.70	0.66	0.61	0.63

Table 1: Models performance on the entire dataset after preprocessing

5.2 Performance of Models in Experiment 2

Models	Accuracy	AUC-Score	Precision	Recall	F1-Score
Logistic	0.64	0.70	0.65	0.62	0.64
R. Forest	0.69	0.76	0.67	0.73	0.70
D.Tree	0.87	0.87	0.81	0.96	0.88
AdaBoost	0.65	0.70	0.64	0.69	0.66
Naive B	0.64	0.70	0.65	0.58	0.61

Table 2: Models performance on the Important features

5.3 Discussion Of Results

As stated in the sections above, The goal of this research is to optimize the process of evaluating credit risk by improving the efficiency of the machine learning models in evaluating the probability of default. This objective was carried out in two phases of experiment. First was to balance the dataset and utilize all the available features in the dataset after preprocessing to build a supervised machine learning model and the second experiment was to obtain a subset of the main dataset containing only the ten most important features to retrain the model earlier developed in experiment one and then do a comparative analysis of the results.

The result obtained from the first experiment revealed that the Decision Tree model had the best overall performance among the five models.It had the highest proportion of charged-off clients correctly classifed (Recall - 0.96). It had the highest accuracy score of 0.87 and only closely followed by Random Forest with accuracy score of 0.66. Also in terms of the F1 score and Recall, Decision Tree still outperformed other models with a score of 0.88 and 0.96 respectively. However, it would be insufficient to conclude on the performance by considering only the accuracy, Therefore looking at the performance in terms of their respective AUC scores; Decision Tree outperformed other models and only followed by Random Forest with scores of 0.87 and 0.72 respectively. Also, it is worthy to note that the models in the first experiment took over two hours to execute completely.

The second phase of the experiment started by obtaining a new dataset which is a subset of the main dataset containing only the top ten most important variables. The

features were obtained from the feature importance plot of the Random Forest classifier. This new dataset was used to retrain the five models with the aim of improving the execution time and the performance of the models. The result obtained revealed that Decision Tree still performed better than the other models although there was no significant changes in its performance from the first experiment.It had the highest proportion of charged-off clients correctly classifed with a Recall of 96%.It had the best accuracy score of 0.87 and only distantly followed by Random Forest which had an accuracy of 0.69. While also not relying entirely on the accuracy score of the models, the AUC score revealed that the best performance was recorded by the Decision Tree model with a score of 0.87 and closely followed by Random Forest,logistic Regression and AdaBoost with 0.76,0.70 and 0.70 respectively. However, it was observed that only Random Forest had an improved performance in the second experiment with about 3% increase in accuracy there was no significant changes in the performance of other models as evaluated with the metrics,rather significant improvement was recorded in terms of the execution time it took to run the models in the second experiment. The models took less than thirty minutes to run successfully in the second experiment.

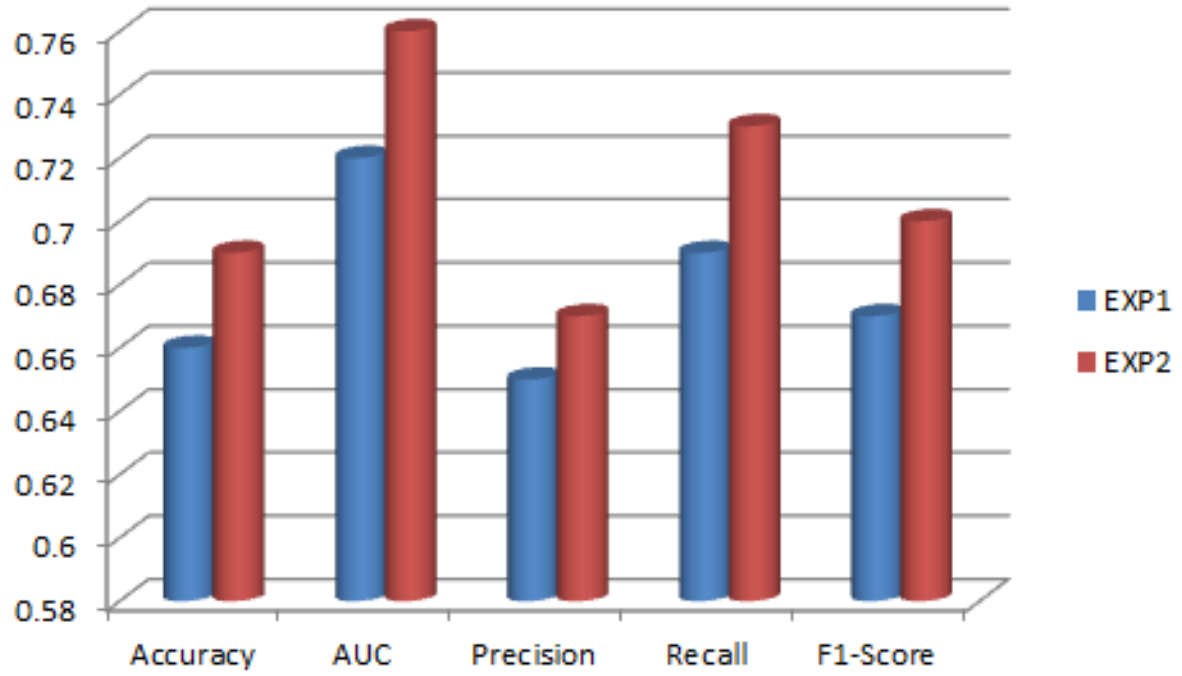

Figure 6: Comparison of Random Forest Performance in Experiment 1 & 2

6 Conclusion and Future Work

The set out goal for this project was to minimize credit risk in peer-to-peer lending business by utilizing machine learning models to evaluate the probability of default and identifying possible loan defaulters.This was aimed at improving the credit risk evaluation process in peer-to-peer lending businesses such that credit decisions can be determined faster and accurately in identifying potential defaulters. This was executed in two phases

of experiment. Due to the peculiar nature of financial dataset, which is highly imbalanced, an oversampling technique (Resample) was employed to overcome the imbalance class distribution in the target variable.This ensured that the models were trained with unbiased datasets with equal number of Charged-off and Fully_paid instances.

Five classification algorithms were employed to build a supervised machine learning model. These models are Random Forest, Decision Tree, AdaBoost, Logistic regression and Naïve Bayes. The result obtained showed that Decision Tree performed best with the highest proportion of "Charged-Off" correctly classified.Decision Tree had an accuracy score of 87% and 72% AUC respectively.The top ten features with the most impact on the variance of the target variable were identified and selected from a feature importance plot and this was used to retrain the five models.

The result obtained from the second experiment showed a significant improvement in the performance of Random Forest with about 3% increase in it's accuracy and AUC score.However, there was no significant changes in other models from the first experiment.Decision Tree still outperformed the other models with the highest proportion of charged-off clients correctly classified with a Recall of 96% and AUC score of 86% . Therefore, the hypothesis that the probability of default of default by an applicant can be evaluated quicker using fewer features has been validated by the results obtained from this experiment. This implies that credit decisions regarding the processing of loan applications can be concluded quicker and efficiently by focusing largely on the identified important features(open_acc,addr_state, credit_history, loan_amnt, annual_inc, sub_grade, revol_util, revol_bal, dti, int_rate) and filtering out unnecessary information which may constitute noise and increase computational cost. By doing so peer-to-peer lending businesses would reduce the risk and losses resulting from defaulting loans.

Recommendations for future work can be done to expand the scope of this work to explore other dimensionality reduction techniques like PCA (principal component Analysis) and LDA (Linear discriminant Analysis).Researches aimed at improving the accuracy of credit risk systems should be a continuous one till a near perfect system can be achieved since a percentage increase in a risk evaluation system will translate into more profit for a lending business. Also, further studies can be carried out on applying deep learning techniques on the problem with the aim of improving the predictive accuracy of the model. Similarly, the use of cost-matrix can be explored to add additional weight to the default class while implementing the models.

7 Acknowledgement

Gratitude to God Almighty for giving me the grace to complete this project. Also, my sincere appreciation goes to my supportive supervisor Dr. Vladimir Milosavljevic, for his relentless guidance and support throughout the duration of this project. I must also appreciate my family and friends who have stood by me and contributed immensely both morally and financially to my academic sojourn. A very big Thanks to you all.

References

Addo, P. M., Guegan, D. and Hassani, B. (2018). Credit risk analysis using machine and deep learning models, *Risks* **6**(2): 38.

Ala'raj, M. and Abbod, M. F. (2016). A new hybrid ensemble credit scoring model based on classifiers consensus system approach, *Expert Systems with Applications* **64**: 36–55.

Bao, W., Lianju, N. and Yue, K. (2019). Integration of unsupervised and supervised machine learning algorithms for credit risk assessment, *Expert Systems with Applications* **128**: 301–315.

Chornous, G. and Nikolskyi, I. (2018). Business-oriented feature selection for hybrid classification model of credit scoring, *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, pp. 397–401.

Dahiya, S., Handa, S. and Singh, N. (2015). Credit modelling using hybrid machine learning technique, *2015 International Conference on Soft Computing Techniques and Implementations (ICSCTI)*, IEEE, pp. 103–106.

Dahiya, S., Handa, S. and Singh, N. (2016). Impact of bagging on mlp classifier for credit evaluation, *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, IEEE, pp. 3794–3800.

Feng, X., Xiao, Z., Zhong, B., Qiu, J. and Dong, Y. (2018). Dynamic ensemble classification for credit scoring using soft probability, *Applied Soft Computing* **65**: 139–151.

Florez-Lopez, R. and Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. a correlated-adjusted decision forest proposal, *Expert Systems with Applications* **42**(13): 5737–5753.

Ha, V.-S., Lu, D.-N., Choi, G. S., Nguyen, H.-N. and Yoon, B. (2019). Improving credit risk prediction in online peer-to-peer (p2p) lending using feature selection with deep learning, *2019 21st International Conference on Advanced Communication Technology (ICACT)*, IEEE, pp. 511–515.

Kim, M.-J., Kang, D.-K. and Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction, *Expert Systems with Applications* **42**(3): 1074–1082.

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q. and Niu, X. (2018). Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning, *Electronic Commerce Research and Applications* **31**: 24–39.

Mai, F., Tian, S., Lee, C. and Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures, *European Journal of Operational Research* **274**(2): 743–758.

Mantas, C. J., Abellán, J. and Castellano, J. G. (2016). Analysis of credal-c4. 5 for classification in noisy domains, *Expert Systems with Applications* **61**: 314–326.

Masmoudi, K., Abid, L. and Masmoudi, A. (2019). Credit risk modeling using bayesian network with a latent variable, *Expert Systems with Applications* **127**: 157–166.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, J. Y. and Ryu, K. H. (2019). An empirical comparison of machine-learning methods on bank client credit assessments, *Sustainability* **11**(3): 699.

Papouskova, M. and Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning, *Decision Support Systems* **118**: 33–45.

Shen, F., Zhao, X., Li, Z., Li, K. and Meng, Z. (2019). A novel ensemble classification model based on neural networks and a classifier optimisation technique for imbalanced credit risk evaluation, *Physica A: Statistical Mechanics and its Applications* **526**: 121073.

Tripathi, D., Edla, D. R., Kuppili, V., Bablani, A. and Dharavath, R. (2018). Credit scoring model based on weighted voting and cluster based feature selection, *Procedia computer science* **132**: 22–31.

Wang, G., Hao, J., Ma, J. and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring, *Expert systems with applications* **38**(1): 223–230.

Wang, Z., Jiang, C., Ding, Y., Lyu, X. and Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending, *Electronic Commerce Research and Applications* **27**: 74–82.

Wei, S., Yang, D., Zhang, W. and Zhang, S. (2019). A novel noise-adapted two-layer ensemble model for credit scoring based on backflow learning, *IEEE Access* **7**: 99217–99230.

Xia, Y., Liu, C., Da, B. and Xie, F. (2018). A novel heterogeneous ensemble credit scoring model based on bstacking approach, *Expert Systems with Applications* **93**: 182–199.

Xia, Y., Liu, C., Li, Y. and Liu, N. (2017). A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring, *Expert Systems with Applications* **78**: 225–241.

Xu, P., Ding, Z. and Pan, M. (2017). An improved credit card users default prediction model based on ripper, *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, IEEE, pp. 1785–1789.

Zhang, W., He, H. and Zhang, S. (2019). A novel multi-stage hybrid model with enhanced multi-population niche genetic algorithm: An application in credit scoring, *Expert Systems with Applications* **121**: 221–232.

Zhu, B., Yang, W., Wang, H. and Yuan, Y. (2018). A hybrid deep learning model for consumer credit scoring, *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, pp. 205–208.

Zhu, L., Qiu, D., Ergu, D., Ying, C. and Liu, K. (2019). A study on predicting loan default based on the random forest algorithm, *Procedia Computer Science* **162**: 503–513.