National College of Ireland

Project Submission Sheet – 2019/2020

School of Computing

Student Name:	Ian Smith		
Student ID:	19178816		
Programme:	MSc. Data Analytics	**Year:**	2020
Module:	Domain Applications of Predictive Analytics		
Lecturer:	Vikas Sahni		
Submission Due Date:	7th August 2020		
Project Title:	Domain Applications of Predictive Analytics Project Report		
Word Count:	2850		

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 encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: Ian Smith

Date: 07/08/20

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

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

Domain Applications of Predictive Analytics Project Report

Ian Smith
School of Computing
National College of Ireland
Dublin, Ireland
x19178816@student.ncirl.ie

***Abstract*—A bagged decision tree model was used to predict whether a loan would become seriously delinquent or not. An accuracy, AUC and recall of 78%, 0.77 and 0.76 respectively were obtained for the model indicating that the decision tree classifier performed reasonably well when predicting serious loan delinquency. Downsampling was also carried out to aid the model's performance due to imbalanced data. Due to the model's high recall score it can be used as an efficient tool for targeting risky loans by lending institutions and financial firms.**

Keywords: Loan Delinquency, Decision Tree, Classification, Downsampling, Bagging

I. Introduction

Successfully implementing a serious delinquency prediction model can provide great benefits to banks and other lending institutions. Identifying risky loans and acting in time can reduce the amount of loss which can lead to increased revenue for those organisations. This report details how a bagged decision tree model was implemented in order predict whether a loan will become seriously delinquent or not.

The layout of this report is as follows. Section II will provide a review of past techniques and problems in the domain of default prediction. One technique will be picked and justification given as to why that particular technique was picked. Section III will discuss how the PDCA method was implemented in order to get the data ready for the implementation of the machine learning model. Section IV discusses the results of the model in a quantitative and qualitative aspect.

II. Applicable Techniques

This section provides a discussion on different predictive analytics techniques applied to the domain of default prediction. Relevant literature will be examined and a technique will be chosen for this project with justification given as to why the technique was chosen.

The most common predictive analytics model used for predicting the default probability and hence developing a credit score is logistic regression. A logistic regression model was compared to a decision tree model with the decision tree achieving the lowest Type 2 error (even though the type 2 errors for all models was high) [1]. The authors in [1] also found that the data quality and structure affected the outcome of the model. As a result of this care must be given when carrying out data cleaning, transformation, etc... so as to avoid reducing the quality of the data which will affect the model's performance.

Other algorithms have been examined in the literature such as bagged k-nearest neighbors and random forests. These have been compared to a non-tuned and tuned logistic regression model in [2]. The random forest algorithm performed the best overall achieving an AUC and Brier Score of 0.959 and 0.071 respectively. However, what the random forest makes up for in performance power it loses in interpretability so gaining a full understanding of how the model determined the result is more difficult.

Algorithms such as random forest are referred to as ensemble techniques because they combine the results of multiple models to try and improve results. Ensemble techniques such as random forest and extreme gradient boosting (XGBoost) were used in [3] along with logistic regression, SVM and neural networks to develop a credit scoring model. XGBoost was the highest performing algorithm with an AUC and F1-Score of 0.7822 and 0.5496 respectively. The computational time/cost for the XGBoost model was

substantially shorter than the trained neural network. This is a key factor that was not mentioned in other papers. Lending institutions will have to deal with a large amount of data and despite advances in computing power in recent years training and implementing a computationally expensive model will eat into the firm's resources.

Another ensemble method that has been examined by researchers is the bagged decision tree. [4] found that bagged credal decision trees performed very well on imbalanced datasets. The authors also used a corrected pairs t-test and a Wilcoxon test to compare the credal decision tree and the bagged credal decision tree in the same dataset and between different datasets which allowed for proper verification of results. Both tests were statistically significant showing that the bagged tree performed better than the non-bagged tree achieving a higher accuracy of 80%. More evaluation metrics such as AUC or recall could have been used to determine how well the model performed as the accuracy score is not always reliable. If the class distribution of a variable is imbalanced the classifier could potentially learn the data from the majority class better and fail to generalize to data from the minority class.

A lot of work has been carried out on comparing the performance of neural networks to other methods as seen in [3]. The performance of a model using a neural network was compared with one implementing Multiple Discriminant Analysis (MDA) when classifying consumer loan applications [5]. It was found that the neural network performed very well and managed to minimise Type 1 errors (false positives) effectively. However, the downside of using neural networks is similar to that of the random forest model in [2] - the lack of interpretability. Neural networks are often referred to as 'black box' methods because it is difficult to ascertain how they arrived at their result.

Lack of interpretability presents a real problem when explaining the results of a model. To combat this, more interpretability models are used. Both decision trees and logistic regression are more interpretable and understandable than more sophisticated models such as Artificial Neural Networks (ANN) or Support Vector Machines (SVM). Consumers as well as lending institutions tend to prefer easy-to-interpret models [6]. Consumers are often interested to see why their credit score is low or high and what are the strongest determinants of a high or low score. If the model is hard-to-interpret the consumer may become frustrated because of this. Another benefit of these models is that they can be easier to implement than more sophisticated ones.

One of the key issues in default prediction is dealing with imbalanced datasets. The reason for the imbalance is that there are often more applicants whose loans did not default than those that did. These datasets that contain a much smaller number of defaults are referred to as low default portfolios (LDPs) and are usually classified as low risk [7]. The performance of different classification algorithms faced with imbalanced data has been examined by researchers and it was found that algorithms such as random forest and XGBoost can perform better in the face of imbalanced datasets [7]. Algorithms such as decision tree, Quadratic Discriminant Analysis (QDA) and knn did not perform as well when presented with imbalanced data.

However, there are certain techniques or re-sampling strategies that can be implemented to deal with imbalanced datasets. These techniques can be implemented to prevent certain algorithms such as logistic regression, Linear Discriminant Analysis (LDA) and decision tree being biased towards the majority class (non-defaults). Down-sampling, up-sampling, Synthetic Minority Over-Sampling Technique (SMOTE) and other methods have been implemented to an imbalanced dataset in [8]. Different algorithms were implemented on each of the re-sampled datasets and it was found that overall LDA performed the best. However, logistic regression performed the best when SMOTE was implemented. Ultimately there is no one sized fits all solution for dealing with imbalanced datasets.

After reviewing the above literature the technique that will be implemented for this report will be bagged decision trees coupled with downsampling. The reason for this is that logistic regression has been used multiple times by lending institutions to develop credit scoring models and by using decision trees it provides an alternative way to develop a model. While decision trees by themselves are often prone to overfitting utilising bagging might reduce the variance in a model leading to greater generalization. Also, by utilising the downsampling technique this will hopefully prevent the bagged decision tree from being biased towards the majority class which, for this dataset is non-default. By combining both the bagged

decision tree with the downsampled data the model results should hopefully improve and lead to better results for consumers and the lending institutions.

III. Methodology

The methodology followed was Plan Do Check Act which was described in the project design document. The aim is to develop a delinquency prediction model using the decision tree algorithm and determine the performance of that model.

A. Data Cleaning and Transformation

This section details the cleaning and transformation steps that were carried out to prepare the data for use by the bagged decision tree. The reasons behind each step and their implication are discussed.

The unprocessed dataset from Kaggle[1] contained 150,000 rows and 11 columns. The dataset was inspected for missing and duplicate values. Null values from the "number of dependents" variable were dropped from the dataset as there were not that many of them (3924). The 29,731 null values from the "Monthly Income" dataset were filled with the median as the mean is too easily influenced by outliers and dropping all of these rows could have resulted in a substantial loss of information. Figure 1 clearly

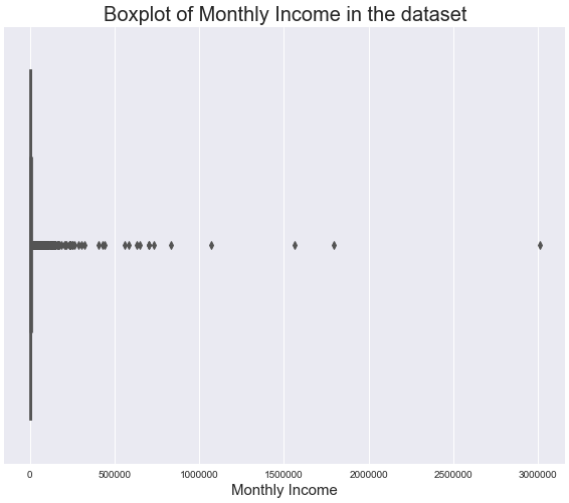

Figure 1. A boxplot of monthly income showing extreme skewness

shows an extreme positive skewness for the Monthly Income variable. The decision to fill the missing values with the median income value was taken due to this extreme skewness. The reasons for this extreme skewness may be due to data collection and/or input error or due to extremely wealthy individuals in the dataset. Also, any records above the 90th percentile which corresponding to a monthly income above 10,833 were dropped from the dataset.

The 'Number of Open Credit Lines and Loans' variable was positively skewed (see Fig. 2). All of the values that were greater than the 95th percentile were removed which corresponded to any values that were greater than or above 18.

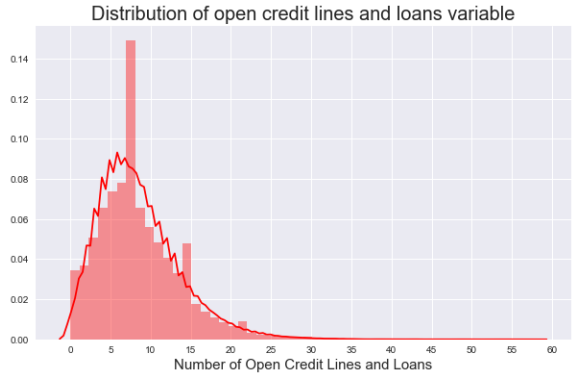

Figure 2. Distribution of the open credit lines and loans variable

The three variables that showed how many days someone was late (30-59, 60-89 and 90 days) all had extreme values which were dropped from the dataset. The number of times late was kept at less than 10 to remove the extreme values but preserve as much information as possible. A similar process was followed for the 'Number of Real Estate Loans of Lines' variable where the total number of real estate loans/lines was less than or equal to 10.

The final variable that was cleaned was the 'Revolving Utilisation of Unsecured Lines' variable. This variable was extremely positively skewed (see Fig. 3). All of the values that were greater than 0.99 (which corresponded to the 95th percentile) were removed. The reason for this skewness may be due to some individuals having particularly large balances and/or lines of credit. This would account for the massive lines of skewness in the plot.

The cleaned dataset had 146,076 rows and 11 columns.

B. Model Implementation

After data cleaning took place the model was developed and implemented. This subsection will describe the steps that were taken to implement the decision tree algorithm.

[1]https://www.kaggle.com/c/GiveMeSomeCredit/data

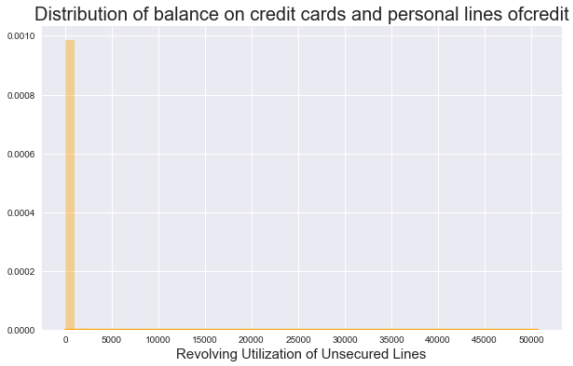

Figure 3. Histogram of the Revolving Utilisation of Unsecured Lines variable.

Before the algorithm was implemented the first step was to perform downsampling. Downsampling is a method for dealing with imbalanced data that involves creating a random subset of the majority class that matches that of the minority class. The 'SeriousDlqin2yrs' is a binary variable with 0 indicating that there is not a serious delinquency and 1 indicating that there is a serious delinquency. The majority class has 136,229 cases and the minority class has 9847 cases showing a significant imbalance. Fig. 4 provides a useful visual aid to show the imbalance.

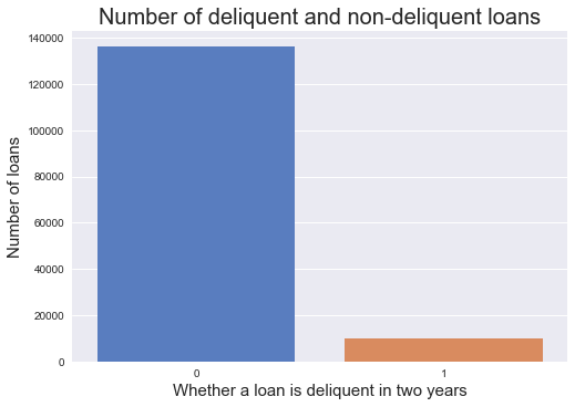

Figure 4. Breakdown of the majority and minority classes in SeriousDlqin2yrs

Before downsampling was implemented the data was split into a training and test set using a 75:25 split. It is vitally important to carry this out before implementing downsampling because if it is done afterwards then some observations may appear in both the training and test sets which could lead to overfitting and the model would not generalize well to new data.

After the train test split was carried out, downsampling was implemented on the target variable and the number of cases for both classes was 7441. The data was then fitted to the decision tree classifier and new predictions made based off of the test dataset.

IV. Results & Evaluation

The quantitative results of the bagged decision tree model as well as the qualitative interpretation of the results are given in this section.

A. Evaluation Metrics

The evaluation metrics that are included in this report (as set out in the project design) are accuracy, precision, recall, f1-score, area under curve (AUC) and the confusion matrix. The Receiving Operating Characteristic (ROC) curve and precision-recall curve will also be plotted. The values for the first five evaluation metrics are given below.

- Accuracy = 78%
- Precision = 0.20
- Recall = 0.76
- F1-Score = 0.31
- AUC = 0.77

The accuracy is defined as the total number of correct predictions divided by the total number of incorrect predictions. For this project it is the correct number of loans classified as defaulted or not-defaulted divided by the total number of false classifications. Therefore, the accuracy score of 78% indicates that the model did an above average job of correctly classifying loans as delinquent or non-delinquent. However, accuracy alone is not enough to determine the performance of a model.

Precision refers to the number of true positives divided by the number of true positives + the number of false positives. The precision for the decision tree model is 0.20. This means that of all the loans that were labelled as delinquent only about 20% were delinquent. Although this score is low it is better to have a low precision than low recall in the context of default prediction. It is better to incorrectly label non-risky loans as risky than miss out on the risky loans.

Recall (or sensitivity) refers to the number of true positives divided by the number of true positives plus the number of false negatives. The recall for the bagged decision tree model is 0.76 which means that the model correctly identified 76% of all delinquent

loans which is a good result. In the context of loan default prediction a high recall is better because it is more important to correctly identify seriously delinquent loan than to mislabel non-delinquent loans.

Improving precision often comes at the expense of recall and vice versa. The F1-score (or F-score) is the harmonic mean of the precision and recall. It is a way to balance the needs of both metrics. It is defined as follows:

$$(2 \times \text{Precision} \times \text{Recall}) \; / \; (\text{Precision} + \text{Recall})$$

A higher F1-score is better. In this model, the F1-score was 0.34 which is a poor result that is mainly due to low precision. However as previously mentioned the high recall score makes up for the low F1-score and precision value.

The confusion matrix for the model is shown in Fig 5. This model correctly predicted 26651 loans that did not result in a serious delinquency correctly and 1834 loans that did result in a serious default. The number of false negatives in the model was 572. This meant that the model incorrectly classified 570 out of 2406 loans as non-delinquent when they actually were seriously delinquent. This error rate among individuals who were seriously delinquent was approximately 24%. This model missed roughly one-quarter of individuals who were seriously delinquent. The lending institution would ideally like this result to be lower but the model is still identifying 3/4 of all loans that could be at risk of default.

The number of false positives was very high in this model with 7462 false positives. Although the false positive rate was very high this might not bother the financial institutions as much because classifying loans as not seriously delinquent when they are actually delinquent is much more costly than classifying loans as seriously delinquent when it actually isn't seriously delinquent.

		Predicted	
		No Default	Default
Actual	No Default	26651	7462
	Default	572	1834

Figure 5. The confusion matrix for the bagged decision tree model

The AUC is the overall performance of the classifier, summarized over all possible thresholds. The value of the AUC is approximately 0.77 which is a good result as an AUC value of 0.5 indicates that a classifier performed no better than chance. Therefore, this model performs decently well at classifying whether loans are seriously delinquent or not.

B. Visualisation

The ROC curve (Fig. 6) provides a useful visual representation of this. A perfect classifier would have the green curve approaching the top left hand side of the graph so this model performs moderately well in terms of classifying loans.

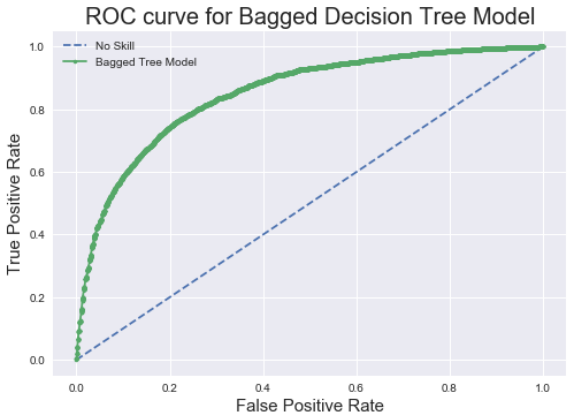

Figure 6. The ROC Curve for the bagged decision tree model. The curve shows that the model performed well at classifying delinquent loans.

Also, age is included in this model which may be present an ethical problem if a model was developed with it as one of its features. Therefore, it is recommended for further model developments that this predictor should not be included to comply with any regulations/laws as the model may discriminate to customers based on age.

V. Conclusion & Future Work

For this project, a bagged decision tree model was implemented to predict whether a loan will become seriously delinquent or not. The model performed well with an accuracy and recall score of 83% and 0.76 respectively. These scores show that the bagged decision tree model achieved moderately good results and could possibly be used as prediction tool by different lending and financial institutions.

For future work, researching more papers and implementing more advanced algorithms such as boosted decision trees or random forests might prove fruitful. It would also be possible to plot each individual bagged tree from both algorithms providing opportunity for interpretation. Also, the model should be

updated constantly to account for any changes in the market or economy. This is particularly important during the current covid-19 pandemic which is causing a tremendous deal of uncertainty and instability in markets across the globe.

REFERENCES

[1] I. Yeh and C.-H. Lien, "The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients", *Expert Systems with Applications*, vol. 36, pp. 2473–2480, Mar. 2009. DOI: 10.1016/j.eswa.2007.12.020.

[2] J. Kruppa, A. Schwarz, G. Arminger, and A. Ziegler, "Consumer credit risk: Individual probability estimates using machine learning", *Expert Systems with Applications*, vol. 40, no. 13, pp. 5125–5131, 2013, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2013.03.019.

[3] L. Marceau, L. Qiu, N. Vandewiele, and E. Charton, *A comparison of deep learning performances with other machine learning algorithms on credit scoring unbalanced data*, 2019. arXiv: 1907.12363, [Accessed: June 19, 2020.]

[4] S. Moral-García, C. J. Mantas, J. G. Castellano, M. D. Benítez, and J. Abellán, "Bagging of credal decision trees for imprecise classification", *Expert Systems with Applications*, vol. 141, p. 112 944, 2020, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.112944. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417419306621, [Accessed on: July 21, 2020.]

[5] R. Malhotra and D. K. Malhotra, "Evaluating consumer loans using neural networks", *Omega*, vol. 31, no. 2, pp. 83–96, 2003.

[6] B. W. Yap, S. H. Ong, and N. H. M. Husain, "Using data mining to improve assessment of credit worthiness via credit scoring models", *Expert Syst. Appl.*, vol. 38, no. 10, pp. 13 274–13 283, Sep. 2011, ISSN: 0957-4174. DOI: 10.1016/j.eswa.2011.04.147. [Online]. Available: https://doi.org/10.1016/j.eswa.2011.04.147, [Accessed: June 19, 2020].

[7] I. Brown and C. Mues, "An experimental comparison of classification algorithms for imbalanced credit scoring data sets", *Expert Systems with Applications*, vol. 39, no. 3, pp. 3446–3453, 2012, ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2011.09.033. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095741741101342X, [Accessed on: July. 10, 2020].

[8] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods", *Knowledge-Based Systems*, vol. 41, pp. 16–25, 2013, ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2012.12.007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S095070511200353X, [Accessed on: July. 10,2020].