2020

Prediction of Bank Customer Response to a Personal Loan Campaign

BY RAGHAD ALOTAIBI

Table of contents

Lists of Abbreviations

ADASYN	Adaptive Synthetic oversampling method
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
WoE	Weight of Evidence

List of Tables

List of Figures

Abstract

Title: Prediction of bank customer response to a personal loan campaign

Author: Raghad Alotaibi

Marketing campaigns are a very critical decision for a business that can be unsuccessful if the product or service delivered does not attract interested customers, thus the key objective for direct marketing prediction is to identify potential customers from an existing database so that marketers can design accurate strategies to increase sales and profitability. Machine learning methods have been successfully used in direct marketing to predict customer response in the banking industry. In this project, the logistic regression, random forest, and neural network algorithms are used to predict the response of bank customers to a personal loan campaign according to their demographic information and transactional patterns. Moreover, detecting the features interaction terms and examining their influence on the performance of the logistic regression model. The random forest method achieved the best performance between the applied methods and succeeded in identifying the interactions between the features in the personal loan data.

العنوان: التنبؤ باستجابة عملاء البنك لحملة تسويقية للقروض الشخصية

الكاتبة: رغد العتيبي

تعتبر الحملات التسويقية قرارًا بالغ الأهمية بالنسبة للأنشطة التجارية والتي من المحتمل أن تفشل إذا لم يحقق المنتج أو الخدمة المقدمة جذب العملاء المهتمين، وبالتالي فإن الهدف الرئيس للتنبؤ هو توجيه الحملة التسويقية للعملاء المهتمين في المنتج من خلال تحديدهم من قاعدة البيانات ودراسة سلوكهم حتى يتمكن المسوقون من تصميم استراتيجيات دقيقة لزيادة المبيعات والربح. تستخدم طرق التعلم الآلي بنجاح وفعالية في التسويق المباشر لتوقع استجابة العملاء في الصناعة المصرفية. في هذا المشروع، تم استخدام خوارزميات الانحدار اللوجستي (logistic regression) والغابة العشوائية (random forest) والشبكة العصبية (neural network) لتوقع استجابة عملاء البنك لحملة القروض الشخصية وفقًا لمعلوماتهم وأنماط معاملاتهم المصرفية. علاوة على ذلك، الكشف عن العلاقات بين المتغيرات ودراسة تأثيرها على أداء نموذج الانحدار اللوجستي. حققت خوارزمية الغابة العشوائية أفضل أداء بين الطرق المطبقة ونجحت في تحديد العلاقات بين المتغيرات في بيانات القرض الشخصي مما أدى إلى الاستهداف الصحيح لشريحة العملاء ذات العلاقة بالحملة التسويقية.

Section 1

Introduction

This section provides an overview of the problem and presents the proposed solution. Lastly, it gives a short introduction to the coming sections in the report.

1.1 Problem Statement

The most challenging thing for marketers is when their marketing campaigns are not working as planned according to their strategy, and the most likely trigger is when the campaign targets the wrong consumer group. Knowing the right customers for the right products will make the marketing campaign success in terms of efficiency, time, and cost-effective, thus eventually helping the firm's revenue increase.

Mathematical modeling adopted in the financial industry can provide strong support for business decision-making, designing the marketing campaign to be more profitable and decrease the costs by targeting the most valuable customers. Moreover, banking databases provide useful information about the customers' status in the bank for more understanding of their transactional patterns and banking details that will contribute to the final objective of the marketing process.

In this context, the objective is to predict the customer propensity of responding to the personal loan campaign based on customer information and transactional patterns. Having the marketing target as lending a personal loan will make it a classification problem of whether a customer will borrow the loan or not. The marketing campaign data are mostly imbalanced because the responders are a rare event in this problem.

The usual approach used for predicting the customer response is the logistic regression, due to its high learning efficiency and the inclusion of feature coefficients which helps with interpretability. The logistic regression assumes variables independency which may not be the case of all datasets that require feature combination.

Other machine learning algorithms include in the development process the effect of feature interactions and data nonlinearity, such as neural networks and the random forest algorithm. Unlike the logistic regression, the random forest does not have the assumption of independence of variables and catch the data nonlinearity and interactions between features. It has the advantage of model interpretability and the model architecture reduces overfitting.

1.2 Proposed Solution

The aim of this project is to predict the customer's likelihood of responding to a personal loan marketing campaign given a set of customer features using logistic regression, random forest, and neural networks machine learning techniques.

Additionally, detecting the feature interactions and examining its influence on the model performance and how it can affect the interpretability in machine learning techniques.

The effectiveness and feasibility of the models are evaluated by using the loan customer response data and balancing it using the Adaptive Synthetic oversampling method (ADASYN). The performance of the models is measured in terms of classification accuracy, Area Under the Curve (AUC), F1 score, precision, and the recall test statistics.

1.3 Structure of Report

The following organization of the project is divided into six sections, including this introductory section. A literature review associated with the algorithms used for response modeling will be covered in section 2. Section 3 will introduce the predictive modeling methods used and the model evaluation metrics together with the implementing programming language. Section 4 is about the data description and preprocessing analysis. The methods results will be presented in section 5. Finally, in section 6 the conclusions from the results and the project summary will be given.

Section 2

Literature Review

This Section will review the statistical and machine learning methods used in the previous studies and their performance as well as the feature interactions effects in the response prediction modeling.

Reviewed studies used different kinds of methods to predict the customer response to the marketing campaigns, the logistic regression is usually adopted for developing such models giving its ability to yield a probability that could be used by a business in deciding whether to reach a customer and it can show the weight of each variable in the model to indicate the variable with the highest impact on the target variable. Yu, et al. [1] used the logistic regression to forecast the response probability of target bank customers for different marketing campaigns. The stepwise regression was used to select the more important features for the customer response model, the variables selected by the method include the age and gender of a customer, the credit limit, online banking signed and the amount of consumption. The model was evaluated by the Type I and Type II error tests, with the probabilities of 7.32% and 1.02% respectively. The Kolmogorov-Smirnov (KS) statistics test was conducted on a development and validation groups to examine the discriminative ability of the model, and it showed a maximum discriminative rate at the 40^{th} percentile to be 53% for approximately both groups. The model showed success in identifying the valuable information and hidden relationships from mass customer data, the analysis discovered that the male customers with a credit limit higher than 300 thousand have more response rate than female customers with 24% difference.

[2], and [3] compared the performance of different machine learning techniques to predict customer response to a bank direct telemarketing campaign in a Portuguese bank. The data contain 17 attributes that describe information about the customers, such as age, job, marital status, education, and housing loan. The attributes were used to predict the customer subscription to a bank term deposit. In [2], the random forest showed the best performance with an accuracy of 86.8% compared to Multilayer Perceptron Neural Network (MLPNN), decision tree (C4.5) and logistic regression having accuracies 82.9%, 84.7%, and 83.5% respectively. It also performed the best in terms of AUC with 0.927 and a True Positive rate of 90.2%. In addition, the random forest outperformed logistic regression, neural networks and Support Vector Machine (SVM) models in [3] with AUC of 0.98 compared to 0.89 for logistic regression, 0.93 for neural networks and 0.87 for SVM when considering the Easy Ensemble under-sampling method.

The performance of four data mining techniques was tested in [4] for personal loan marketing. The authors applied logistic regression, decision tree, neural networks, and SVM to a Taiwanese bank dataset including information about customer's gender, age, transactions and the variation of expenditure in terms of credit cards. The logistic regression had 1.75 cumulative lift, the decision tree had 1.62 cumulative lift and SVM had 1.76 cumulative lift while the best

performing model is the neural networks had 1.77 cumulative lift. However, the logistic regression model had the highest Gini coefficient of 0.535 among the decision tree with 0.441, neural networks with 0.522, and SVM with 0.487 Gini coefficients. The findings of the responders from the study revealed that they preferred the automatic channels for their transactions. Responders also showed a difference with non-responders in terms of financial needs, for example, the ATM usage rate.

A study to improve direct mail targeting was conducted on four direct marketing datasets provided by the Direct Marketing Educational Foundation in [5]. The first dataset is for a non-profit organization and it has the donation after direct mail as the target variable with 27.42% response rate, the remaining three datasets are about catalog companies to predict the responders for the catalog mailing. The datasets contain social and demographic characteristics together with the amount and frequencies of previous donations or purchases. The authors implemented logistic regression, linear and quadratic discriminant analysis, naïve Bayes, neural networks, decision trees (CHAID, CART, and C4.5) and k-nearest neighbor algorithms. The logistic regression only performed the best on one dataset with 0.642 AUC and the linear discriminant analysis reports a similar AUC results of 0.641 on the same dataset. The neural networks performed the best for two data sets with AUC of 0.684 and 0.823. The CHAID tree performed the best for one dataset with the AUC of 0.862. The highest AUC result for the CART decision tree is 0.836 whereas the quadratic discriminant analysis, naïve Bayes, C4.5 tree and k-nearest neighbor classification algorithms had poor performance compared to other algorithms for all the datasets. However, the study considered other criteria of algorithm choice for marketing managers including, the ease of implementation, resolution time, the availability and convenience of the algorithm software packages also the interpretability.

In [6], the writers proposed a system to reach the interested customers in credit products in a retail bank in Poland and derive their patterns from their historical transfer, temporal information and transactional data using Classification and Regression Trees (CART), random forests and deep neural networks (Deep Belief Networks with the use of H2O). The CART model had 67.27% recall, 9.01% precision when considering the Boruta feature selection algorithm, and was most efficient regarding the computing power. The deep learning networks had a slightly higher precision of 16% and a recall of 11% compared to the random forest model that achieved 15% precision and 10.63% recall. The system was able to extract significant patterns from customers' historical transfer and transactional data and predict credit purchase likelihood.

Authors of [7] developed a logistic regression model for predicting customer behavior. The model is built using the features selected from the Portuguese bank telemarketing data by the mutual information (MI) and data-based sensitivity analysis (DSA) feature selection methods to boost the performance over false-positive hits. The used feature selection methods reduced the number of feature sets that affected this marketing sector's performance. In the case of a low false-positive ratio with nine selected features, the authors found that the DSA method is superior. MI method was slightly better when false-positive values are moderately high with 13 selected features across a wide variety of different features. The logistic regression model with features selected by the DSA method outperformed the logistic regression with MI selected features.

[8] proposed a profit-conscious ensemble selection (PCES) methodology that includes business objectives in customer targeting during model development. It contained a collection of 15 learning algorithms and was applied to 25 cross-sectional datasets about selecting customers for targeted marketing actions, 10 of datasets had a response modeling objective. The PCES framework is based on the ensemble selection machine learning paradigm and uses 15 different algorithms with different setting for algorithms parameters including logistic regression, decision trees, discriminant analysis (linear and quadratic), Naïve Bayes, random forest, support vector machines, gradient boosting, neural network, k-nearest neighbor and AdaBoost learning algorithms. Next, the models were selected using directed hill-climbing and finally, the models' predictions were aggregated to produce the final prediction. The PCES framework was applied to 25 marketing datasets that have churn prediction, profitability scoring and response modeling as marketing objectives. The proposed framework recommended more profitable target groups compared with the logit model and random forest. It also showed the benefit of integrating business goals early in the modeling process in terms of the quality of the prediction and decision support.

Another profit-driven study was established by developing a profit-based Artificial Neural Network (PNN) classification algorithm [9]. The experiment was implemented on two credit card fraud detection data set from two Turkish banks and a Portuguese bank direct marketing campaign data to maximize the net profit deriving from the application of the classification model. The proposed approach performance was compared to seven versions of artificial neural network classifier with altered error functions for the individual cost and profit sensitivity for each instance, the decision trees, and the Naïve Bayesian classifier. In the proposed approach on the bank direct marketing campaign dataset, the most profitable cases were found and the overall net income was maximized with a classification accuracy of 87.9%.

The neural network algorithm was used as a feature construction tool then the features are added as input variables in the logistic regression model for the classification of bankcard response in [10]. The authors investigated the neural network algorithm's ability to generate new features that have a relationship with the output variable since it is considered as a supervised algorithm on the credit card response dataset provided by the Atlantics Services Corporation. The proposed hybrid model was compared to the logistic regression without the newly created features by the neural network. The hybrid model with six selected features had three of the features created by the neural network algorithm and it achieved a classification accuracy of 0.823 on the training set and 0.817 on the validation set, AUC of 0.801 on the training set and 0.787 on the validation set and KS statistics of 0.458 on the training set and 0.442 on the validation set. Whereas the logistic regression model without the features generated by the neural network achieved a 0.815 accuracy on the training and the validation sets, 0.777 AUC on the training set and 0.756 AUC on the validation set. In terms of KS statistics, it achieved 0.417 on the training set and 0.395 on the validation set with the same number of selected features. Generally, the hybrid model performed better on both the training and validation sets with different numbers of selected features.

The features interactions give more information about how a combination of independent features contribute to the dependent feature when features interact with each other in a prediction model, the prediction can be expressed as the product of the feature effects because the effect of one feature depends on the value of the other feature.

The features interactions are not detected by the traditional statistical methods however it can be used to increase the performance of the prediction when added to the model. The logistic regression showed improved performance when adding feature interactions, [11] proposed the use of decision tree-based Chi-square Automatic Interaction Detection (CHAID) to detect the potential feature interactions and used them as the additional input variables in logistic regression to support the design of direct marketing campaign, the proposed model was applied to a credit customer response dataset provided by Atlanticus Services Corporation. The experiment was conducted with a different number of selected features on the pure logistic model (with no interactions terms) and the logistic regression with the interaction terms, the logistic regression model with 15 selected features was chosen to be the best model in the study due to the fact that collecting customer information in the credit analysis domain is time-consuming and expensive and because it reached the criteria of having at least 0.4 of the KS statistics on the validation set. This model had four feature interaction terms and it reached an accuracy of 81.30% on validation set compared to the pure logistic regression model with the same number of features with 80.86% accuracy on the validation set. Generally, the model with the interaction terms outperforms the pure logistic regression model in the classification accuracy, AUC and KS metrics for the different number of selected features.

To detect interactions between features, common decision tree based machine learning methods are used, for instance, CART [12], random forests [13], Dirichlet process forests [14], and Rulefit [15]. The random forest showed success in detecting the interactions between features in different fields, in [16] the authors used a variation of the random forest algorithm called iterative Random Forest (iRF) to identify high-order interactions in biological data. However, it was not commonly used to detect interactions among the features in the banking marketing domain and was only considered for its high performance without detecting the interactions in the model.

Overall, the reviewed studies used different statistical and machine learning algorithms to predict customer response to marketing campaigns. The logistic regression algorithm did not perform the best in the studies compared to the other machine learning approaches applied in them. However, the logistic regression performance was improved when considering the features interactions when developing the model and when adding features constructed by other algorithms also when using feature selection methods.

The random forest algorithm is one of the powerful machine learning techniques to predict the customer response, it achieved high accuracy results and was the best performing model in the reviewed studies. However, the random forest was not used for interaction detection in bank customer response datasets in the reviewed studies, thus it is used in this project for its high performance and to

detect the feature interactions and examine its effect when added to the logistic regression model.

The performance varied between studies for the other machine learning algorithms, for the neural network algorithm, the performance remained high in many studies as well as it is been used as a feature construction algorithm. The decision tree and SVM algorithms were frequently used to predict the customer response and showed good performances. The k-nearest neighbor and naïve Bayes were less common to be used as the predictive algorithm for the customer response and obtained poor results.

The class imbalance was common for this type of problem therefore the data balancing was needed for the machine learning algorithm for the most accurate classification of the responders. The studies used different balancing techniques such as oversampling and under-sampling and the studies used similar attributes to predict the customer response.

Section 3

Methodology

This section provides an overview of the predictive modeling methods used in this project. It includes a summary of each modeling technique together with the model evaluation metrics used to assess and compare the modeling techniques together with the programming languages and packages that have been used to implement the methods presented.

3.1 Models

As noted earlier the modeling objective is to predict the likelihood that a customer will respond to a personal loan marketing campaign given a set of customer features. Since the predictive outcome is categorical in nature and the project has a particular focus on feature interaction detection, there are a number of methods that could be considered for modeling. In this project three machine learning algorithms were adopted for investigation and performance comparison: logistic regression, random forests [13] and neural networks.

3.1.1 Logistic Regression

Logistic regression is a form of regression applied to problems with a binary target outcome. It estimates the probability of the target outcome based on a set of input features. In the logistic model, the log-odds for a value belonging to a class is a linear combination of one or more independent variables using the logit function:

$$z = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_k * x_k$$

Where p is the conditional probability of an observation belonging to a class, β_0 is the intercept term, and β_1 is the coefficient associated with the independent variable x_1.

The logit function is the inverse of the logistic function, logistic function (also known as the sigmoid function) is used to yield a probability output value between 0 and 1. Mathematically it transforms the values according to the formula given below:

$$y = \frac{1}{1 + e^{(-z)}}$$

As can be seen from Figure 3.1, the logistic function naturally contains the output y to lie in the range (0,1).

The classification is then made by defining a threshold and classifying the outcome probabilities into one of two binary target outcome categories according to whether the probability falls above or beneath the threshold. The logistic regression model parameters β, are estimated using the maximum likelihood probabilistic framework.

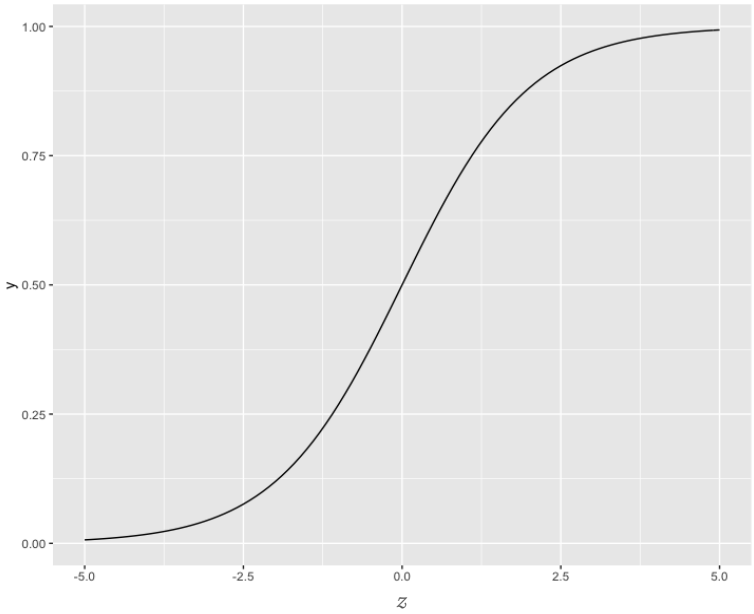

Figure 3. 1: The logistic function (sigmoid function).

3.1.2 Random Forest

Random forest is an ensemble method involving the use of multiple decision trees with varying depths and a final decision determined by aggregating the prediction result of the trees. It has broad applications in data mining and machine learning and can be used for classification and regression problems. It is recognized for its high performance and can be more interpretable than other methods.

The random forest grows the decision trees by generating each individual tree from a random sample of the dataset with replacement, in a process known as bootstrap aggregation (bagging). This process achieves diversity and uncorrelated behavior of each tree, also randomizing the input features to force more variation between the trees in the model. The resulting diversity of trees will capture more complex patterns of features than a single decision tree and reduce the risk of overfitting to training data, therefore increasing predictive accuracy.

The splitting of a node in decision trees is based on its impurity that can be measured using the Gini index or entropy for classification problems and variance in regression trees. The objective of this process is to choose an attribute that split records into pure classes and similar values of the dependent variable are grouped in the same set after the split.

The Gini index can be computed by subtracting 1 from the summation of the squared probability p_i of a given variable being wrongly classified when it is randomly chosen over n classes using the following equation:

$$Gini(T) = 1 - \sum_{i=1}^{n} {p_i}^2$$

Random forests can be used to select and rank variables by the variable importance measures that estimate the contribution from the variables to the class prediction. One of the measures is the mean decrease accuracy that is the average of the differences of the prediction error on each training sample using only trees in their bootstrap sample which did not have that training sample, this method is known as out-of-bag (OOB) error, and it is measured with and without the variable for all trees. Another measure is the mean decrease Gini that calculates the variable's total decrease in node impurity using the Gini index criterion. Feature interactions can also be detected using the random forest by studying the splitting behavior of variables.

Figure 3.2 shows an example of the random forest classifier representation, the random forests constructing process starts with the selection of random samples known as the bootstrap samples from the original data. Then a tree is grown for each of the bootstrap data set (Tree-1 until Tree-n), the variables are randomly selected for splitting at each node of the tree and each tree will provide a class prediction. After that, the information is aggregated for prediction as majority voting. Finally, the OOB error rate is computed using the data out of the bootstrap sample.

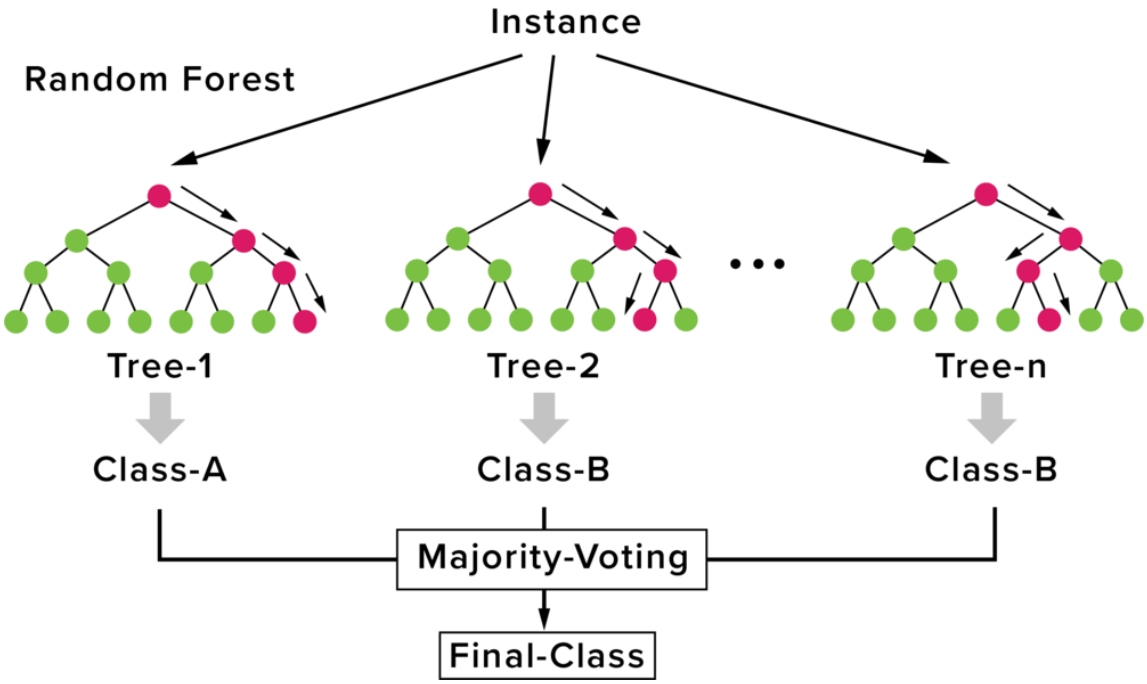

Figure 3. 2: Random forest classifier representation.

3.1.3 Neural Networks

Neural networks are a set of algorithms that its mechanism is derived from the biological brain parallelism between connected neurons, resulting in a highly fault-tolerant method, this characteristic is also known as "graceful degradation". Because of its distributed nature, a neural network keeps on working even when a significant fraction of its neurons and interconnections fail. Also, relearning after damage can be relatively quick [17].

The neural network architecture is made of three basic layers: the input layer, the hidden layer, and the output layer. The architecture defines how a network turns its input into an output based on the arrangement of neurons and the connection patterns between the layers, the activation function, and the learning methods. Figure 3.3 illustrates a neural network representation with an input layer L_1 that contains the input features, one hidden layer L_2 and the output layer L_3 with one node. The nodes labeled with "+1" refer to the intercept term.

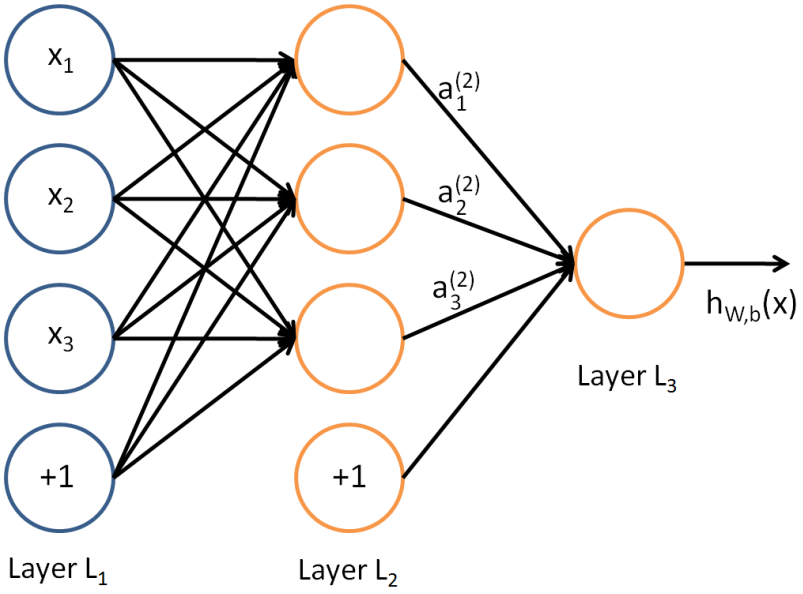

Figure 3. 3: Neural network representation [18].

The activation function converts the combined input signals of a neuron into one output for further transmission within the network. All inputs are weighted individually, added together and passed to the activation function. The sigmoid activation function is a common activation function used in the construction of neural networks. It has the properties of differentiability, monotonically increasing and its output is bounded between 0 and 1 [19].

To train the neural network, a learning algorithm is used. The learning algorithm is formulated in terms of the minimization of the error estimated from the loss function and it is used to tune the model's parameters for weights optimization. There are different types of learning algorithms with different computational speed and memory requirements, such as gradient descent and the quasi-Newton method.

3.2 Model Evaluation

In order to assess the classification models some performance metrics will be considered. The metrics assess model performance by comparing predicted output values with actual output values. The confusion matrix is used to evaluate the predicted values by the models against the actual values in the data based upon four fundamental measures. It is represented as a tabular form as shown in Table 3.1, where True Negative (TN) represents the number of negative values that where correctly classified by the model, False Negative (FN)

is the wrongly predicted negative values, False Positive (FP) is the number of values that are classified as positives when their actual values are negative, and True Positive (TP) that is the number of correctly classified positive values. The confusion matrix helps in finding the accuracy, precision and recall of the model and detects overfitting.

Table 3. 1: Confusion matrix for binary classification.

	True value	
Predicted value	1	0
1	*TP*	*FP*
0	*FN*	*TN*

3.2.1 Accuracy

Accuracy measures the performance of the model and whether it has been trained correctly by using the following equation:

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

It is the proportion of all predictions that are correct predictions with a value that ranges from 0 to 1.

3.2.2 Precision

The precision is used to measure the proportion of positive prediction that was actually correct as shown in the following equation:

$$\text{Precision} = \frac{TP}{TP + FP}$$

As such, precision is the proportion of all positive predictions that are actually positive and has a value in the range from 0 to 1, a model that produces no false positives has a precision of 1.

3.2.3 Recall (Sensitivity)

The recall returns the percentage of actual positives that the model captured correctly, it is beneficial in cases when the cost of false negatives is high. Recall is calculated as the following equation demonstrates:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity, on the other hand, is the percentage of correct negative predictions divided by the total number of negatives. It is also referred to as true negative rate and result a value between 0 and 1 using the following equation:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

3.2.4 F1 Score

The precision and recall are often inconsistent, that is improving one reduces the other, therefore the F1 score is often used to find a balance between the precision and recall which gives equal importance to them both. The equation of F1 score considers both the precision and recall of the test to compute the score and it is formulated as follows:

$$\text{F1} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

3.2.5 Area Under the ROC Curve

All of the above metrics depend upon the specific threshold value chosen to categorize predictions as either positive or negative and so the metrics will vary according to this choice of threshold. Since the choice of threshold is arbitrary then the metrics' reliance on choice of threshold could be considered a weakness. It is therefore perhaps preferable to use a metric which incorporates the effect of choice of threshold. This is traditionally carried out through the Receiver Operating Characteristic (ROC) curve which is a visualization of two of the key metrics for showing the performance comparison of classification models at all classification thresholds. That is, it illustrates the relationship between the true positive rate (recall/ sensitivity) and the false positive rate (1- specificity) that are calculated using the confusion matrix as follows:

$$\text{True positive rate} = \frac{TP}{TP + FN}$$

$$\text{False positive rate} = \frac{FP}{FP + TN}$$

With the true positive rate on the y-axis and the false positive rate on the x-axis the ROC curve is displayed in Figure 3.4. The diagonal line shows how the ROC curve would look for a random classifier (that is poor) and the other two lines shows how the ROC curve would look for models 1 and 2. A perfect predictive model will have a true positive rate of 1 and a false positive rate of 0, the closer the model’s line to the diagonal line the less accurate the model is in which case is seen that model 2 is preferred to model 1.

More specifically, the Area Under the Curve (AUC) is used as an index of accuracy and a performance metric for ROC curve. It falls between 0 and 1 with a higher number indicating better model predictive power. In Figure 3.4, the AUC of the random classifier is just 0.5 and the AUC for model 2 is greater than that for model 1.

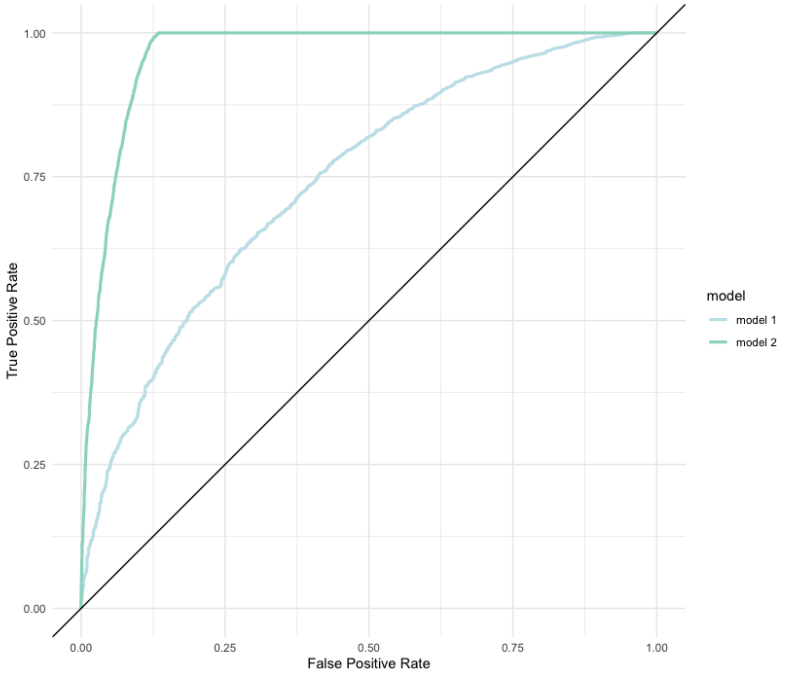

Figure 3. 4: ROC curve example for two models.

3.3 Implementation

In this project, the R programming language was used for data exploration and visualization, data pre-processing, model build and model evaluation for the logistic regression and the random forest models, Table 3.2 describes packages used in R. Neural network modeling was implemented using Python programming language with the packages ScikitLearn, Pandas and matplotlib for data visualization.

Table 3. 2: Description of packages used in R.

Package	Description
dataPreparation	Data preparation.
caret	Classification and regression training.
imbalance	Preprocessing algorithms for imbalanced datasets (ADASYN oversampling).
glm	Logistic regression modeling.
randomForest	Implements random forest algorithm for classification and regression.
iml	Interpretable machine learning. Interprets the behavior and explains predictions of machine learning models.
ggplot2	Data visualization.

Section 4

Data & Analysis

The purpose of this section is to describe the data used in this project, it is a personal loan marketing campaign banking dataset to predict the propensity of customers responding to the campaign and accepting the personal loan. Relevant variables are described in detail, the data pre-processing and data partition is explained at the end of this section.

4.1 Data Description

The personal loan campaign dataset is obtained from Kaggle website [20], it includes the records of 20,000 customers and 40 features that are related with the customers' demographic information, such as age, gender and occupation. The data contains their bank details and transactional patters for example the number of credit and debit transactions and the amount withdrawn from ATM. The details of the data attributes and the data type are described in Table 4.1. The target variable TARGET denotes binary value 0 to customers who did not respond to the personal loan campaign and 1 otherwise.

Table 4. 1: Personal loan campaign dataset description.

Attribute	Description	Data type
CUST_ID	Customer ID	Categorical
TARGET	Customer response	Boolean
AGE	Age of the customer in years	Numerical
GENDER	Gender	Categorical
BALANCE	Average monthly balance	Numerical
OCCUPATION	Occupation	Categorical
AGE_BKT	Age bucket	Categorical
SCR	Generic marketing score	Numerical
HOLDING_PERIOD	Ability to hold money in the account	Numerical
ACC_TYPE	Account type	Categorical
ACC_OP_DATE	Account open date	Date
LEN_OF_RLTN_IN_MNTH	Length of relationship in months	Numerical
NO_OF_L_CR_TXNS	Number of credit transactions	Numerical
NO_OF_L_DR_TXNS	Number of debit transactions	Numerical
TOT_NO_OF_L_TXNS	Total number of transactions	Numerical
NO_OF_BR_CSH_WDL_DR_TXNS	Number of branch cash withdrawal transactions	Numerical
NO_OF_ATM_DR_TXNS	Number of ATM debit transactions	Numerical
NO_OF_NET_DR_TXNS	Number of net debit transactions	Numerical
NO_OF_MOB_DR_TXNS	Number of mobile banking debit transactions	Numerical
NO_OF_CHQ_DR_TXNS	Number of check debit transactions	Numerical
FLG_HAS_CC	Has Credit Card	Boolean
AMT_ATM_DR	Amount withdrawn from ATM	Numerical

AMT_BR_CSH_WDL_DR	Amount cash withdrawn from branch	Numerical
AMT_CHQ_DR	Amount debited by check transactions	Numerical
AMT_NET_DR	Amount debited by net transactions	Numerical
AMT_MOB_DR	Amount debited by mobile banking transactions	Numerical
AMT_L_DR	Total amount debited	Numerical
FLG_HAS_ANY_CHGS	Has any banking charges	Boolean
AMT_OTH_BK_ATM_USG_CHGS	Amount charged by way of the other bank ATM usage	Numerical
AMT_MIN_BAL_NMC_CHGS	Amount charged by way minimum balance not maintained	Numerical
NO_OF_IW_CHQ_BNC_TXNS	Amount charged by way inward check bounce	Numerical
NO_OF_OW_CHQ_BNC_TXNS	Amount charged by way outward check bounce	Numerical
AVG_AMT_PER_ATM_TXN	Average amount withdrawn per ATM transaction	Numerical
AVG_AMT_PER_CSH_WDL_TXN	Average amount withdrawn per cash withdrawal transaction	Numerical
AVG_AMT_PER_CHQ_TXN	Average amount debited per check transaction	Numerical
AVG_AMT_PER_NET_TXN	Average amount debited per net transaction	Numerical
AVG_AMT_PER_MOB_TXN	Average amount debited per mobile banking transaction	Numerical
FLG_HAS_NOMINEE	Has Nominee	Boolean
FLG_HAS_OLD_LOAN	Has any earlier loan	Boolean
random	Random number	Numerical

Histogram plots of six numeric features with the vertical dotted line denoting the median value are represented in Figure 4.1. The targeted customers are highest in the age group between 30-35. The maximum length of relationship with the bank was 221 months and customers with the highest ability to hold money in the account were in the range of 5-20 in holding period. The frequency distribution of the total number of transactions by customers was at its highest between 10-15 transaction, the customers had most of their credit transactions in the range between 0-15 and the number of debit transactions not more than 5 transactions.

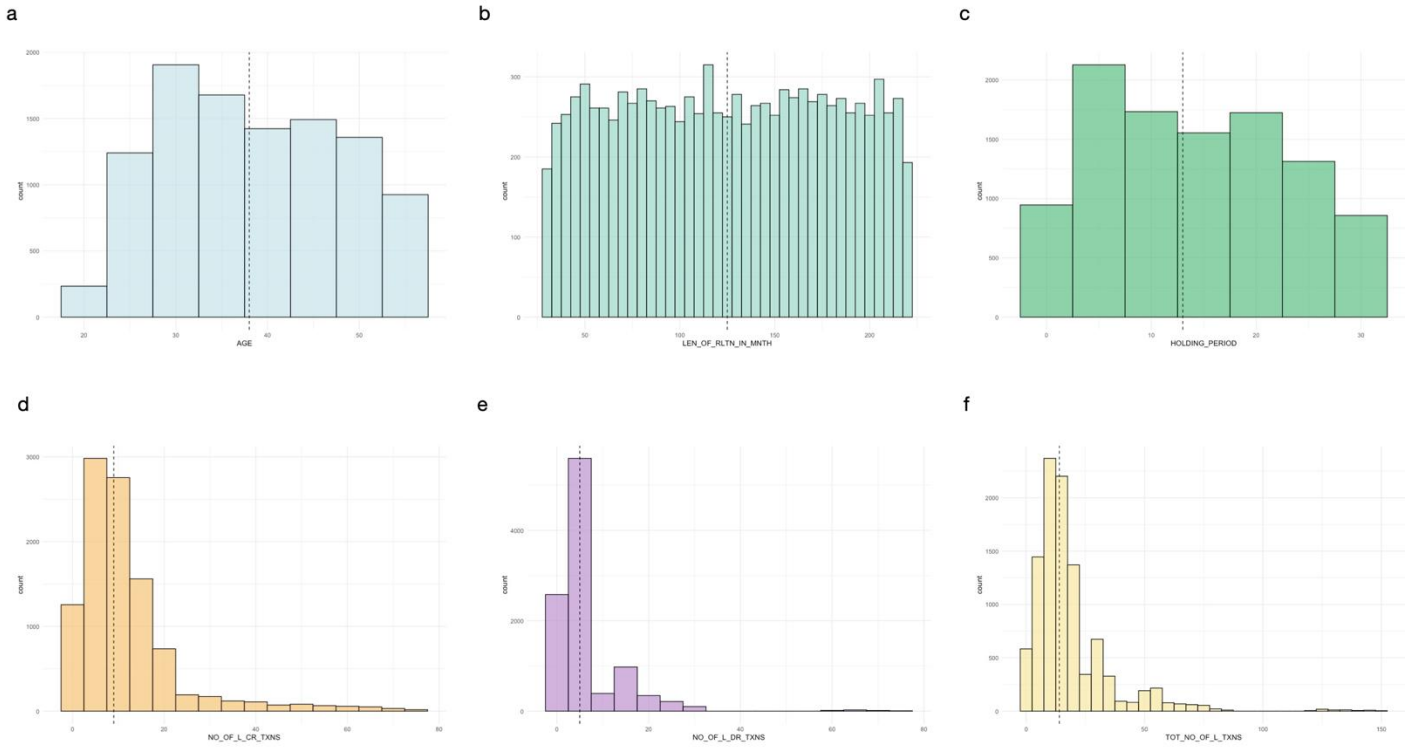

Figure 4. 1: Histograms of numeric features. **a** Histogram of AGE feature, **b** Histogram of length of relationship in months feature, **c** Histogram of holding period feature, **d** Histogram of the number of credit transactions feature, **e** Histogram of the number of debit transactions feature and **f** Histogram of the total number of transactions feature.

The dataset contained more non-responder customers than responders, as Figure 4.2 shows, most of the bank customers are male customers with higher number of personal loan campaign responders than female customers.

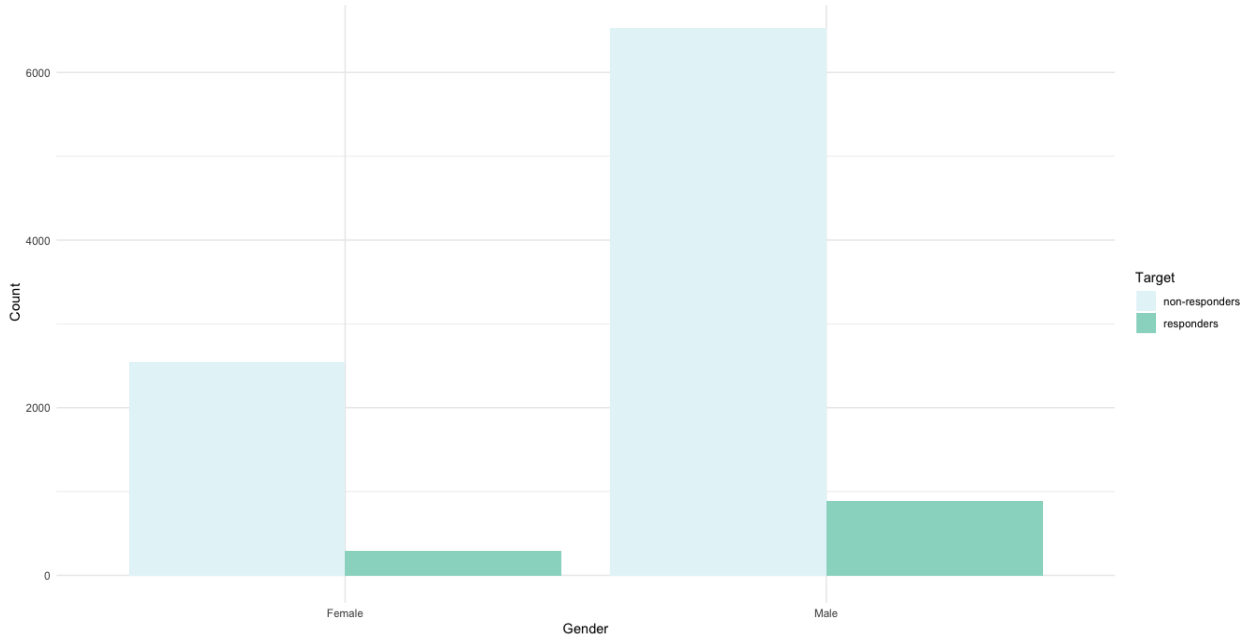

Figure 4. 2: Responders by Gender feature.

Customers with professional occupations had the highest number of responders between the other occupations with 323 responders as demonstrated in Figure 4.3, self-employed occupation comes in second place with a slight difference of 321 responders. The number of responders that have a salaried occupation is 273 and among the non-professional self-employed (SENP) customers 261 were responders to the marketing campaign.

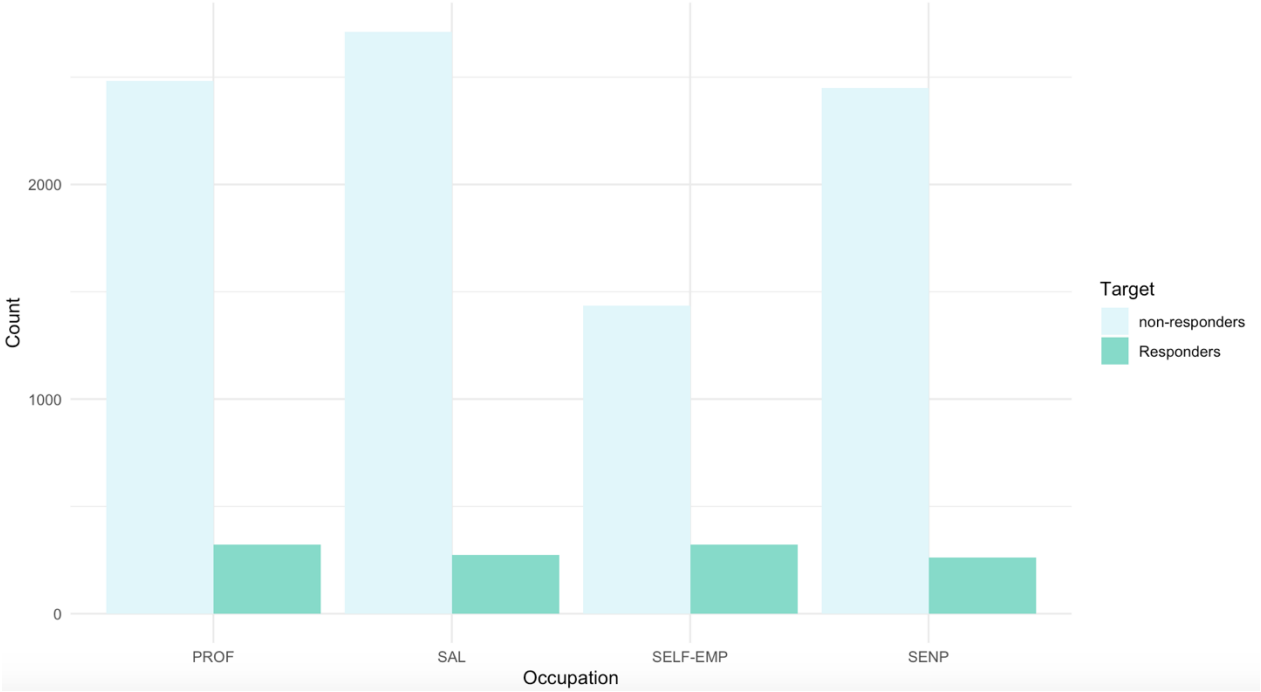

Figure 4. 3: Responders by Occupation feature.

As seen in Figure 4.4, most of the customers did not have a credit card with a number of responders of 681 in this group. Whereas the number of responders within customers with a credit card is 497.

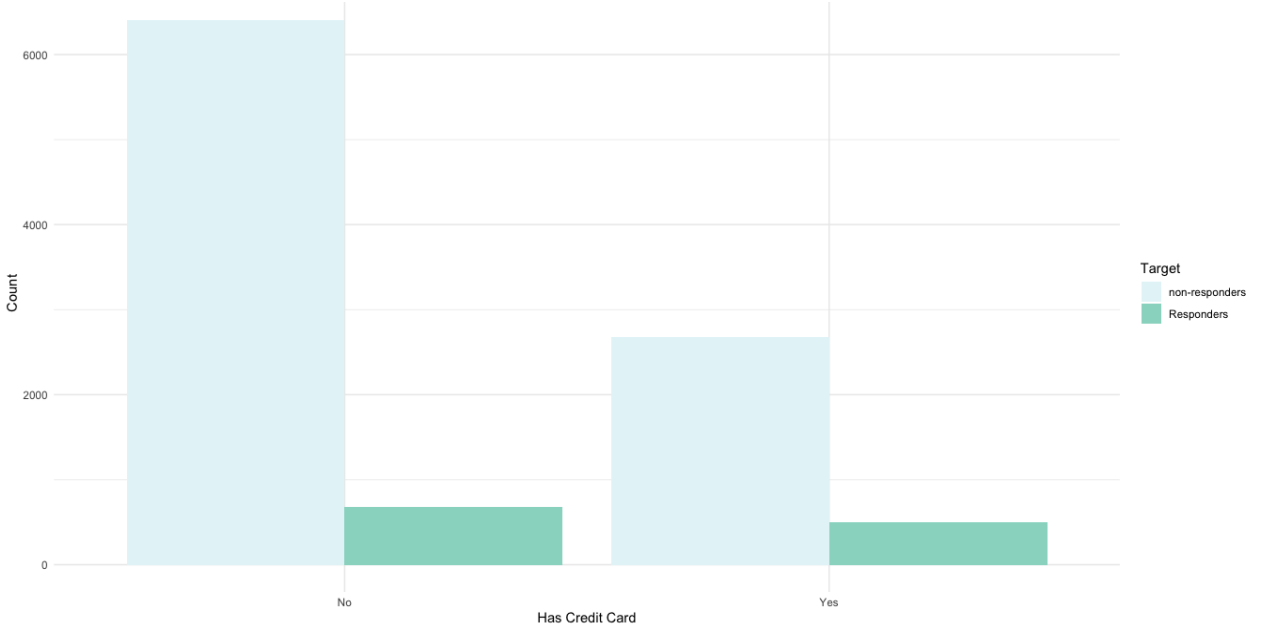

Figure 4. 4: Responders by Has Credit Card feature.

4.2 Data Pre-Processing

There are 29 continuous, 10 categorical features and 1 date feature in the dataset, and it does not contain any missing or null values. In order to prepare the data for the analysis, some pre-processing techniques are applied. The duplicated records were removed, the gender feature had 96 records with the value "other" and they are removed as well. The features account opening date and age bucket are removed because they are redundant and refer to the same information as the length of relationship in months and the age features. The customer ID and random features are deleted because they are not required in model building, the AMT_OTH_BK_ATM_USG_CHGS feature had high percentage of zero values (99.3%) and was excluded from the data set. Also, the features amount charged by way inward and outward check bounce had zero values percentage of 95.8% and 95.6% respectively and were excluded as well. The final number of features in the dataset is 33 and 10255 records.

4.2.1 Value Transformation

Most of the machine learning algorithms can not handle categorical variables until they are treated and converted to a numerical value. Additionally, numerical variables often affect the predictive model differently according to their ranges. Therefore, it is scaled to fall under common range to increase the efficiency of the predictive model.

Two methods are used in this project to encode categorical attributes as numbers. The first method is the one-hot-encoding method, in this method each category is converted to a new binary variable with values 0 and 1 denoting the absence or presence of the feature. The second method is Weight of Evidence (WoE) Encoding, it represents the categorical values with their numerical codes in the form of weights of evidence. The weight of evidence tells the predictive power of an independent variable in relation to the dependent variable. It can be calculated as:

$$WoE(X) = \ln\left(\frac{P(X = x_j \mid Y = 1)}{P(X = x_j \mid Y = 0)}\right)$$

The WoE for a categorical attribute X, is the natural log for each category level (x_j) of the ratio of belonging to class 1, divided by the ratio of belonging to class 0, this method is applied to continuous variables as well after discretization. The WoE data encoding method showed an improvement to the classification performance when applied to financial datasets [21]. In this context, the data encoded using this method is only applied to the logistic regression and neural network models.

Numerical attributes are normalized by using the min-max normalization method for rescaling the range of numerical data to the range [0, 1]. It is defined as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Where x is the original numerical value, and x' is the normalized value.

The normalized continuous data is only applied to the neural network and logistic regression machine learning models.

4.2.2 Data Balancing

Most of the customers are not applying for the loan, resulting in an imbalanced target variable with a 11.4% ratio of responders. In this context the minority class samples are of great importance, and the goal of the machine learning algorithm is to classify them as accurately as possible. Therefore, the dataset is balanced by the Adaptive Synthetic oversampling (ADASYN) method.

ADASYN is an algorithm that generates synthetic data based on a weighted distribution for different minority class examples according to their learning complexity [22]. It focuses on producing samples next to the original samples which are wrongly classified using the Euclidean distance-based k-nearest neighbor classifier.

Machine learning algorithms supported by ADASYN balancing technique showed a high performance compared to other oversampling methods such as Synthetic Minority Oversampling Technique (SMOTE) when applied to financial datasets [23].

The dataset comprises of 18253 records after balancing, with 98.9% imbalance ratio. Figure 4.5 shows an example of the data distribution before balancing with the total number of transactions variable on the x-axis, the balance variable on the y-axis and the colored dots denote each class. It can be seen that there are few observations of class 1 (responders) and more of class 0 (non-responders). Figure 4.6 represents the data after balancing using ADASYN algorithm, with more samples for the minority class.

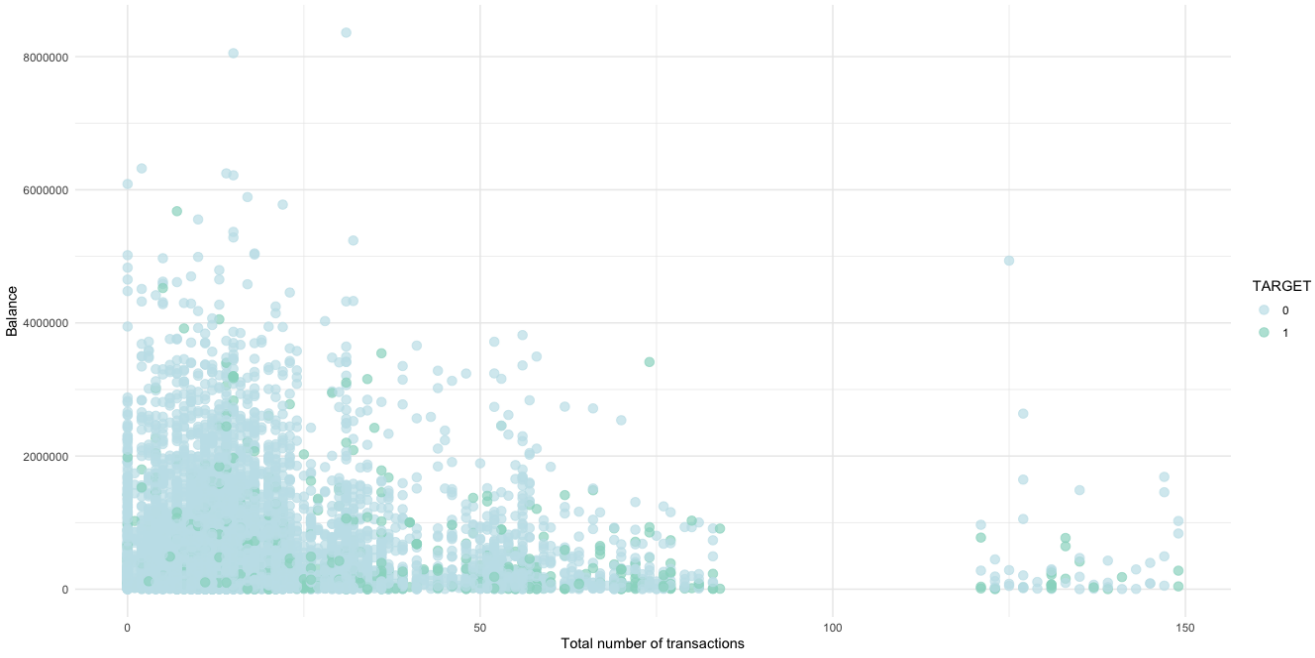

Figure 4. 5: Scatter plot of the original dataset before balancing.

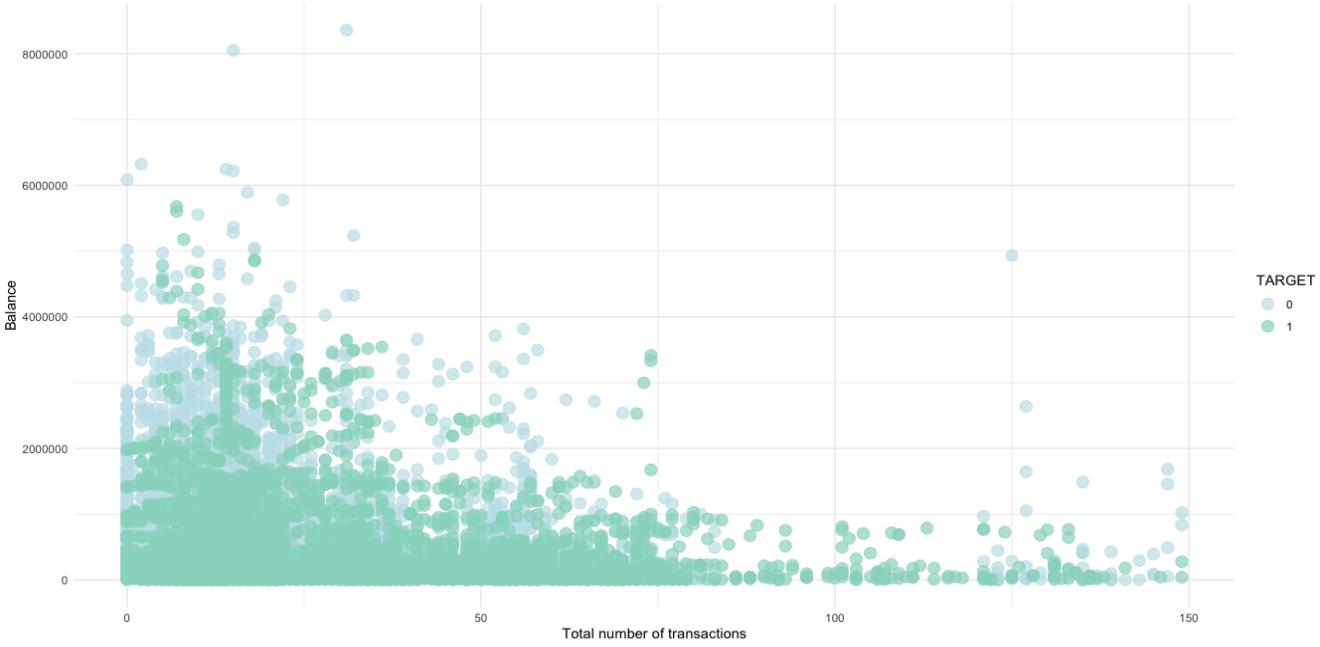

Figure 4. 6: Scatter plot of the modified dataset after balancing.

4.2.3 Data Partition

After the pre-processing, the dataset consists of 33 variables (42 variables when categorical data is encoded using one-hot-encoding method) and 18253 records. It is divided into a training set to train a model with 70 percent of the data, and a testing set to test the trained model with 30 percent of the data. This process aims to create a model that generalizes well to new data that the test set alternates it.

Section 5

Results

This section will present the performances of the models using the methods and evaluation metrics described in Section 3, trained on the preprocessed data described in Section 4. Moreover, this section includes the performance of the logistic regression model with interaction terms, the performance of models on the imbalanced data, and an overall comparison of the machine learning algorithms.

5.1 Models

5.1.1 Logistic Regression

The logistic regression model was developed using three data representations, first, the one-hot-encoding for categorical variables and a raw continuous variable. Second, one-hot-encoding for categorical variables and normalized continuous variables using min-max normalization method. The last data representation is the WoE encoded data.

The logistic regression model on the one-hot-encoded data is built using 24 variables (including the target) that have a Variance Inflation Factor (VIF) values less than 10 to prevent the multicollinearity in the regression model. The confusion matrix of the model with the first data representation is shown in Figure 5.1, the model was able to classify the responders (target value of 1) with a higher number of TPs 2292 than FPs that are 507 as well as predicting the non-responders with TNs of 2184 and a smaller number of FNs.

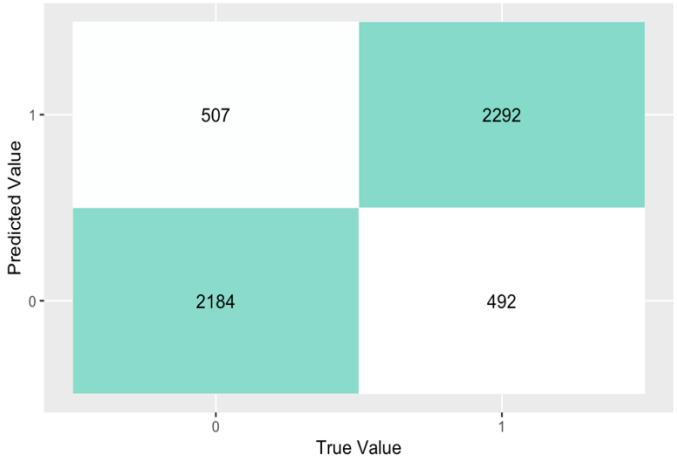

Figure 5. 1: Confusion matrix of logistic regression model on one-hot-encoded data.

Table 5.1 summaries the performance of the logistic regression model in terms of classification accuracy, precision, recall, F1 score and AUC. The model achieved an accuracy of 81.7% and a higher recall of 82.3% than precision of 81.8%. The AUC of model is 0.8901 and the ROC curve is shown in Figure 5.2.

Table 5. 1: Logistic regression model performance on one-hot-encoded data.

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.8175342	0.8188639	0.8232759	0.8210639	0.8901

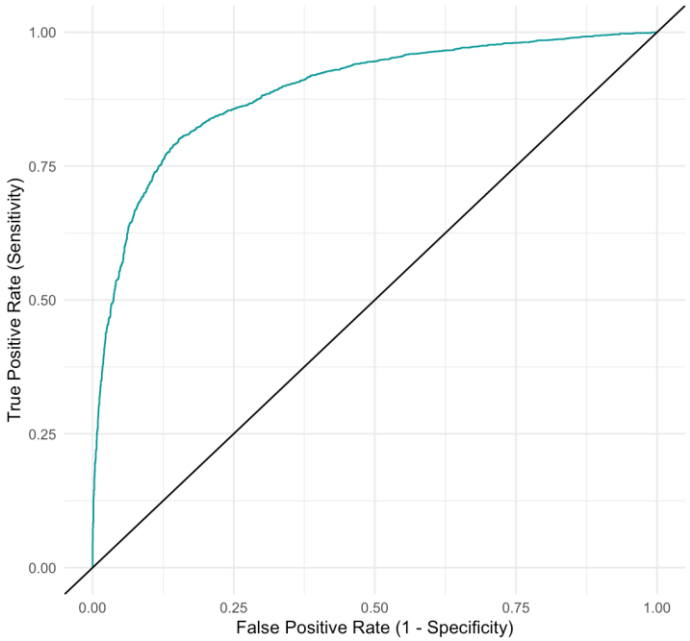

Figure 5. 2: ROC curve of logistic regression model on one-hot-encoded data.

As for the WoE encoding, the continuous variables were divided into groups each containing at least 5% of observations and with number of groups not higher than 6 then they are replaced with the WoE data representation. For the categorical variables the WoE replaced the variable levels. The logistic regression model trained on WoE encoded data is built using 29 variables (including the target) that have a VIF value less than 10. The confusion matrix of the model is shown in Figure 5.3, the model had a number of TPs of 1889 and FPs of 978.

Figure 5. 3: Confusion matrix of logistic regression model on WoE encoded data.

Table 5.2 shows the results of the logistic regression model on WoE encoded data. The model has a 65.9% accuracy, 65.8% precision, 67.9% recall. The AUC is 0.7181 and the ROC curve for the model is shown in Figure 5.4.

Table 5. 2: Logistic regression model performance on WoE encoded data.

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.6599382	0.6588769	0.6792521	0.6689093	0.7181

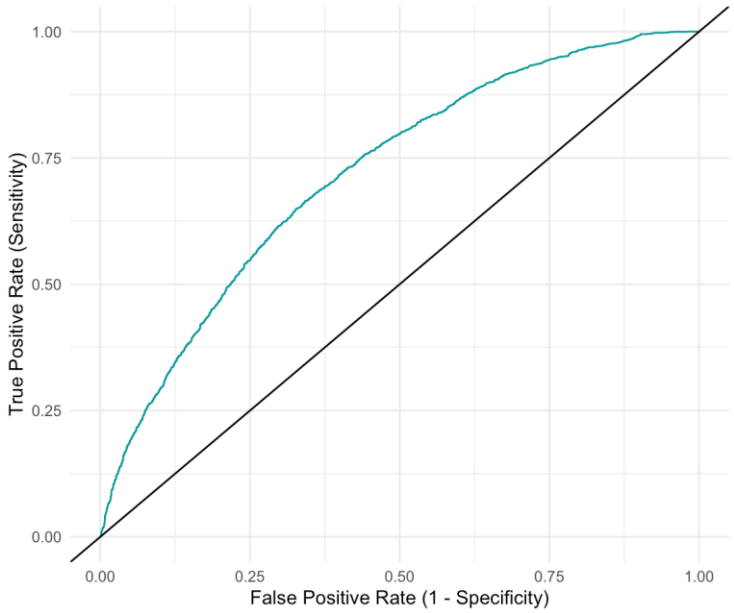

Figure 5. 4: ROC curve of logistic regression model on WoE encoded data.

The last data representation used to train the logistic regression model is the one-hot-encoding method for categorical variables and normalized numerical variables using the min-max normalization method. 24 variables (including the target) are used to build this model that have a VIF value less than 10. The variables selected for this model are the same as the variables used in the first logistic regression model on the one-hot-encoded data. Figure 5.5 shows the confusion matrix of the logistic regression model on the normalized data. The model predicted 2270 of responders and 2217 of non-responders correctly, and had a number of FNs of 529 and a number of FPs of 459.

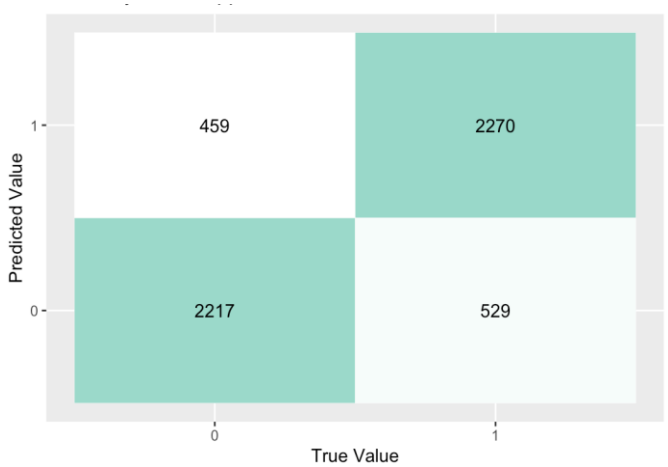

Figure 5. 5: Confusion matrix of logistic regression model on normalized data.

The model achieved an 81.9% in classification accuracy and 83.1% in precision and 81.1% in recall as seen in Table 5.3. As for the AUC, the model scored a 0.8858 and the ROC curve of the model is shown in Figure 5.6. This model performed slightly higher than the first model with the unnormalized data resulting in normalizing the numerical variables to be in a common range can increase the performance for logistic regression.

Table 5. 3: Logistic regression model performance on normalized data

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.8195434	0.8318065	0.8110039	0.8212735	0.8858

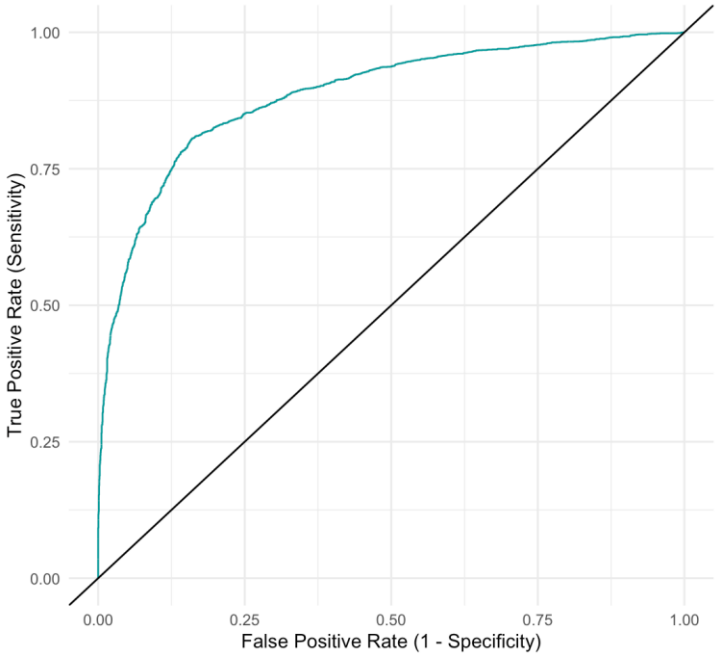

Figure 5. 6: ROC curve of logistic regression model on normalized data.

5.1.2 Random Forest

The random forest classifier used in this project is developed using 500 decision trees, the number of variables tried at each split equal 6 and the node impurity is measured by the Gini index. In addition, the categorical attributes used in this model were encoded using the one-hot-encoding method.

The confusion matrix in Figure 5.7 shows the performance of the random forest in predicting the responders. The model was able to correctly predict the responders with 2483 TPs and a few FPs of 32. The random forest had an OOB estimate of error rate of 5.48%.

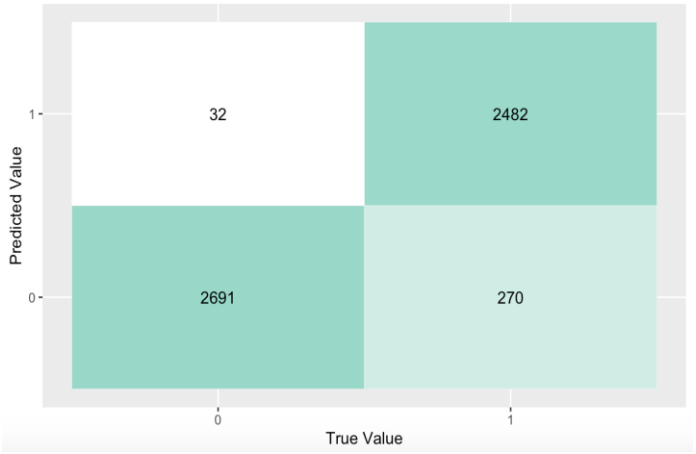

Figure 5. 7: Confusion matrix of random forest classifier.

The results of the random forest model are summarized in Table 5.4, the accuracy of the random forest model reached 94.4% and the model's precision is 98.7%. The recall of the model is 90.1% and it has the AUC of 0.9793, the ROC curve of the random forest classifier is shown in Figure 5.8.

Table 5. 4: Performance of random forest classifier.

Model	Accuracy	precision	recall	F1	AUC
Random Forest	0.9448	0.9872713	0.9018895	0.9426510	0.9793

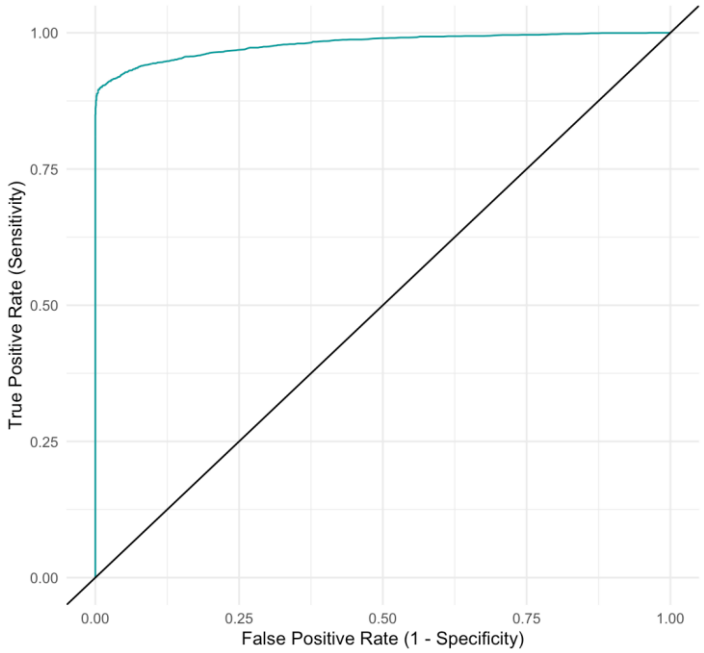

Figure 5. 8: ROC curve of random forest classifier.

Random forest classifiers can reveal the feature importance that indicates how much each feature contributes to class prediction. Figure 5.9 shows the feature importance using the random forest for loan campaign response data in mean decrease Gini. The attribute has credit card equal 0 (FLG.HAS.CC.0) ranked the

most important attribute in explaining the target variable in the model followed by the salaried occupation attribute (OCCUPATION.SAL). Whether a customer has an old loan or not appeared to have medium importance in predicting the response to the new loan. In addition, features like account type, amount debited by the net and mobile banking transactions, and the number of the net and mobile banking debit transactions were not important in predicting the responders to the personal loan campaign.

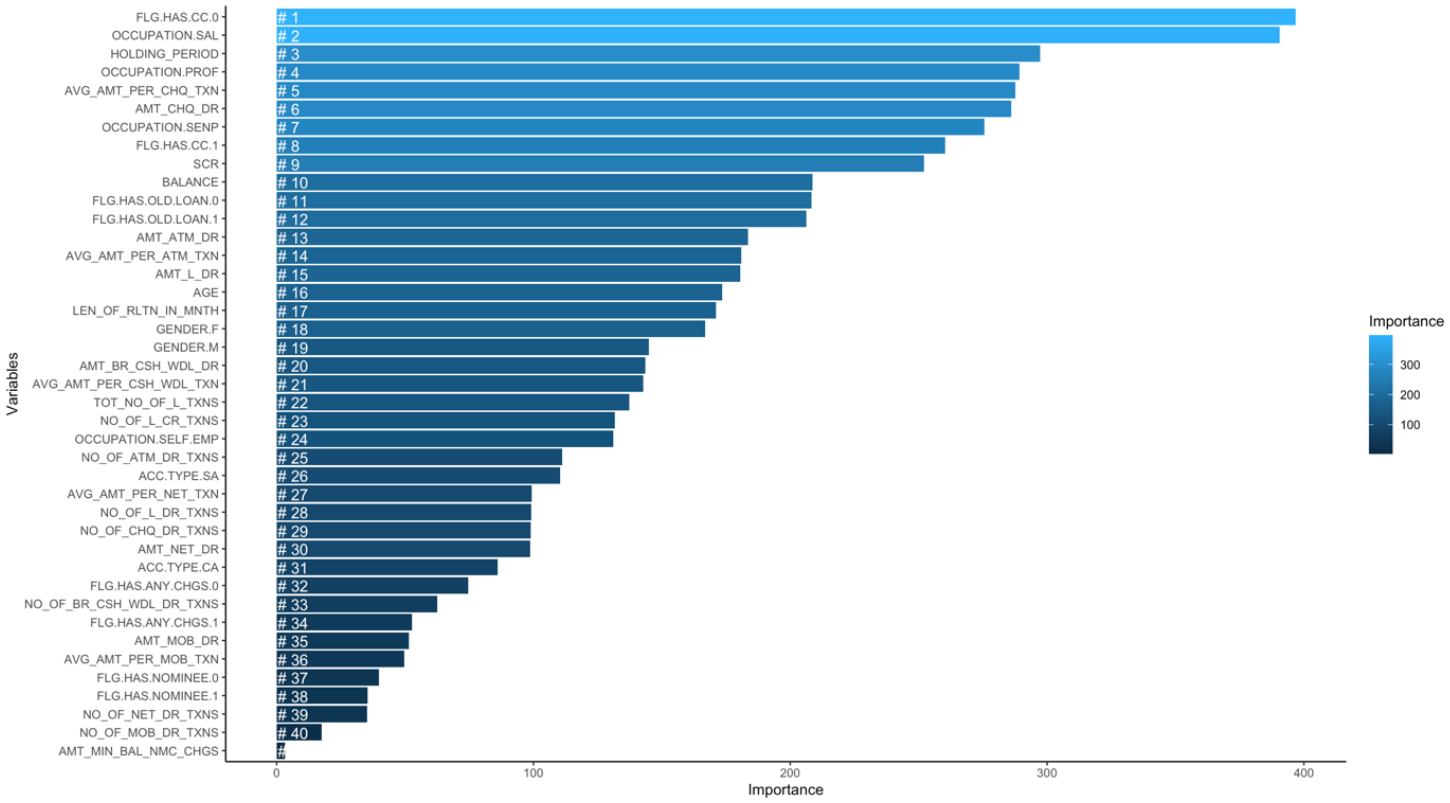

Figure 5. 9: Feature importance using random forest classifier in mean decrease Gini.

Along with the feature importance the feature interaction can be extracted from the random forest to increase the model's interpretability. The interaction between two features is the change in the prediction that occurs by varying the features after considering the individual feature effects. To estimate the interaction strength the H-statistic introduced by Friedman and Popescu [15] is used. It measures how much of the variation of the prediction depends on the interaction of the features.

The interaction strength using H-statistic for each feature with all the other features for a random forest predicting the response of a loan campaign is represented in Figure 5.10 for both target variable outcomes (0 and 1). The salaried occupation has the highest relative interaction effect with all other features, followed by the has credit card equal 0 (FLG.HAS.CC.0) variable. Female gender variable had more interaction strength compared to male gender variable, the flag has old loan also showed difference between its two possibilities with a stronger interaction to (FLG.HAS.OLD.LOAN.1) variable. Overall, the interaction strength was 0.0 for most of the features.

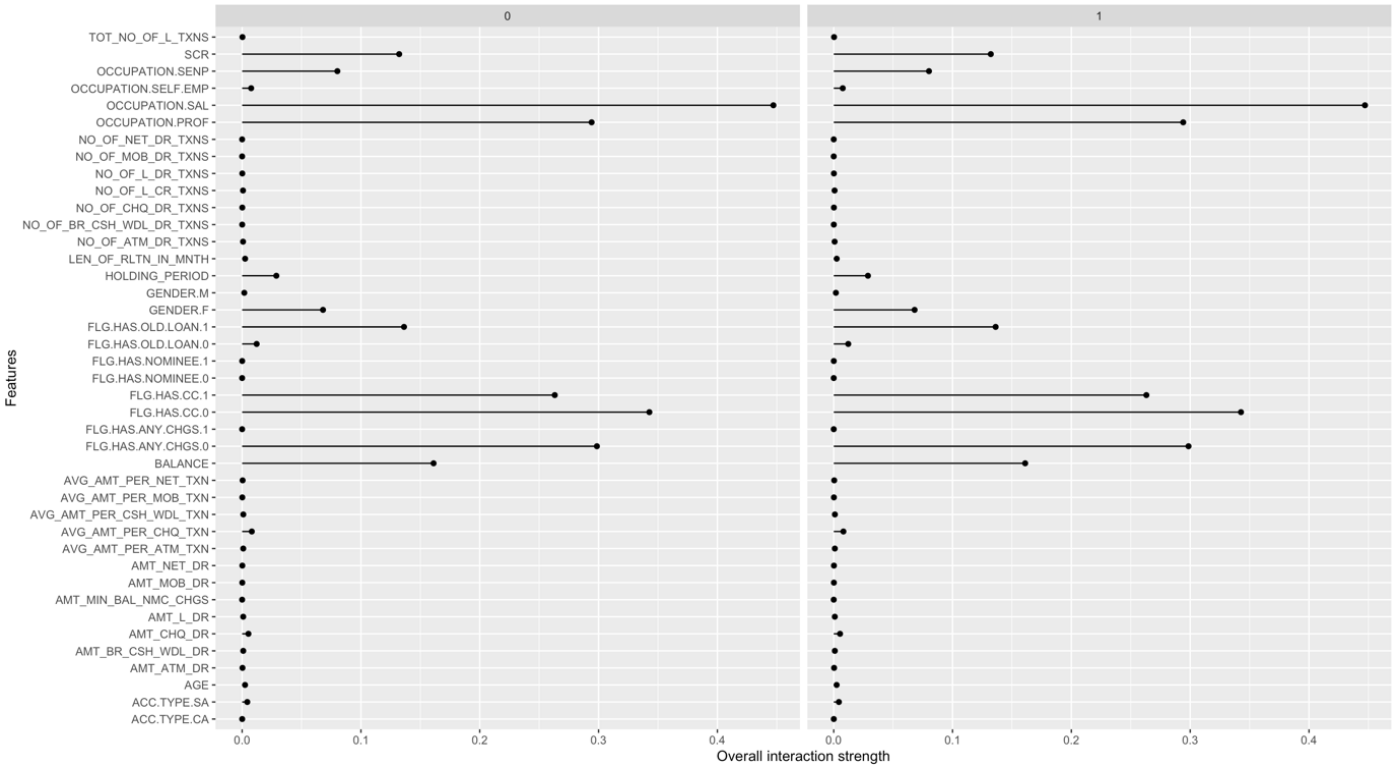

Figure 5. 10: The interaction strength for each feature with all other features.

After looking at the feature interactions of each feature with all other features in Figure 5.10, the salaried occupation (OCCUPATION.SAL) feature is selected due to its high interaction strength to find the two-way interactions between it and the other features in the dataset. The two-way interaction defines whether and to what degree two features in the model interact with each other, the strength of the two-way interactions is showed in Figure 5.11. The strongest interaction is between the variables salaried occupation and account type equal savings (ACC.TYPE.SA). There is a strong interaction between salaried occupation and the holding period and between the flag has an old loan equal to 1 (FLG.HAS.OLD.LOAN.1) and salaried occupation. The amount charged by way minimum balance (AMT_MIN_BAL_NMC_CHGS) and the number of mobile banking debit transactions (NO_OF_MOB_DR_TXNS) features appeared to have the weakest interaction strengths with the salaried occupation between all features in the dataset. The two-way interaction terms are used for further investigation of the performance of the logistic regression when added to the model.

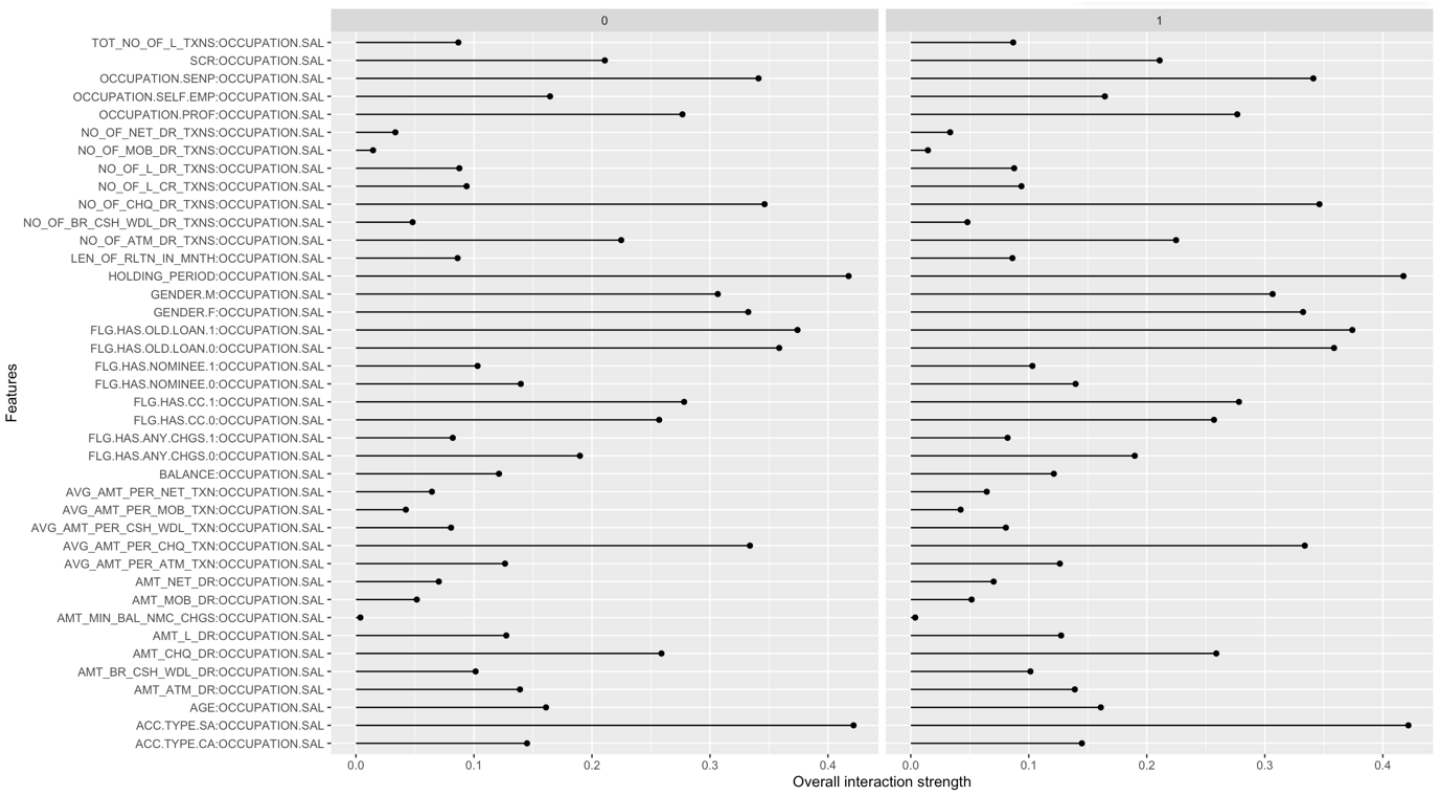

Figure 5. 11: The two-way interaction strengths between salaried occupation and each other feature.

5.1.3 Neural Networks

The Neural network model was developed using the logistic (sigmoid) activation function and the quasi-Newton solver for weight optimization. The model was trained using two data representations, first, one-hot-encoding for categorical variables and normalized numerical variables using the min-max normalization method. The second data representation is by using the WoE encoding method. Three neural network models were constructed with different topologies in this project, the first with one hidden layer and 3 nodes in it, another one with one hidden layer with 5 layers in it and the last neural network model with two hidden layers with 3 and 5 nodes in them. Table 5.5 shows the performance of the neural network model with different numbers of hidden layers and nodes on the one-hot-encoded and normalized data. The model with two hidden layers performed the best in terms of the accuracy, precision, recall and F1 score evaluation metrics with an accuracy of 92.3%, a precision of 99.5%, a recall of 85.2% and F1 score of 91.8%. However, in terms of AUC the model with one hidden layer and 5 nodes performed the best between the models with an AUC of 0.95610.

Table 5. 5: Neural network model performance on normalized data.

Network architecture	Accuracy	precision	recall	F1	AUC
One hidden layer with 3 nodes	0.920927	0.991539	0.850199	0.915446	0.95349
One hidden layer with 5 nodes	0.916179	0.969362	0.850199	0.915446	0.95610
Two hidden layers with (3,5) nodes	0.923849	0.995762	0.852375	0.918506	0.95533

Figure 5.12 represents the confusion matrix for the best performing neural network model with two hidden layers, the model a 2709 of non-responders and 2350 of responders correctly. The ROC curve of the neural network model with two hidden layers on normalized data is shown in Figure 5.13.

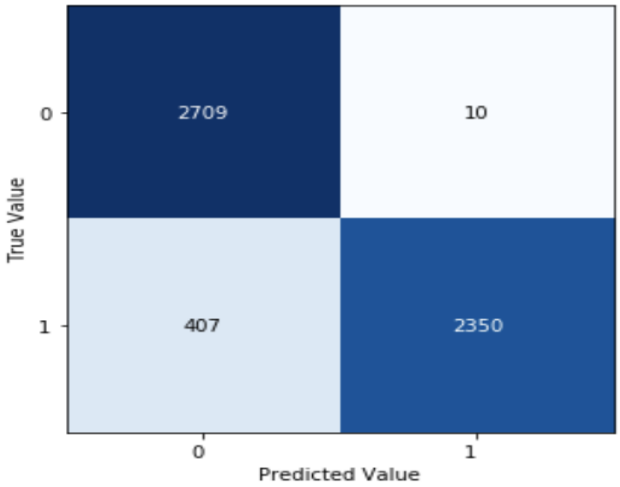

Figure 5. 12: Confusion matrix of neural network model with two hidden layers on normalized data.

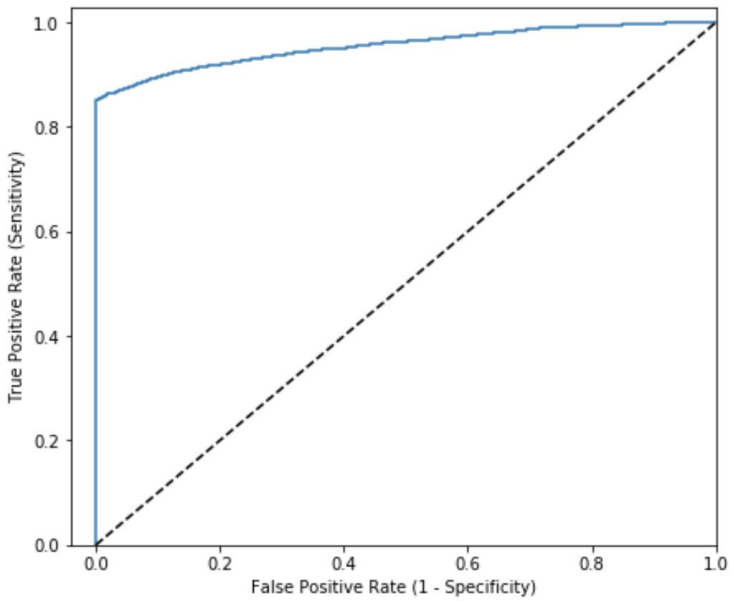

Figure 5. 13: ROC curve of neural network model with two hidden layers on normalized data.

The neural network performance decreased when trained on the WoE encoded data as shown in Table 5.6. Among the three neural network architectures the model with two hidden layers achieved the highest accuracy of 69.3%. The neural network model with one hidden layer and 3 nodes was able to predict most of the positive predictions with 69.5% precision and it resulted the highest AUC of 76.1% between the models, and the model with one hidden layer and 5 nodes had the highest recall of 81.1%. Overall, the neural network model with two hidden layers performed better compared to one hidden layer on both data representations.

Table 5. 6: Neural network model performance on WoE encoded data.

Network architecture	**Accuracy**	**precision**	**recall**	**F1**	**AUC**
One hidden layer with 3 nodes	0.692125	0.695922	0.712415	0.704072	0.76116
One hidden layer with 5 nodes	0.685942	0.657683	0.811460	0.726524	0.75595
Two hidden layers with (3,5) nodes	0.693035	0.672523	0.785284	0.724543	0.75835

The confusion matrix of the neural network model with two hidden layers are represented in Figure 5.14. The number of TPs is 2220 and the number of FPs is 1081, the model was able to correctly classify the non-responders with a higher number of TNs of 1591 than FNs of 607. The ROC curve for this model is shown in Figure 5.15.

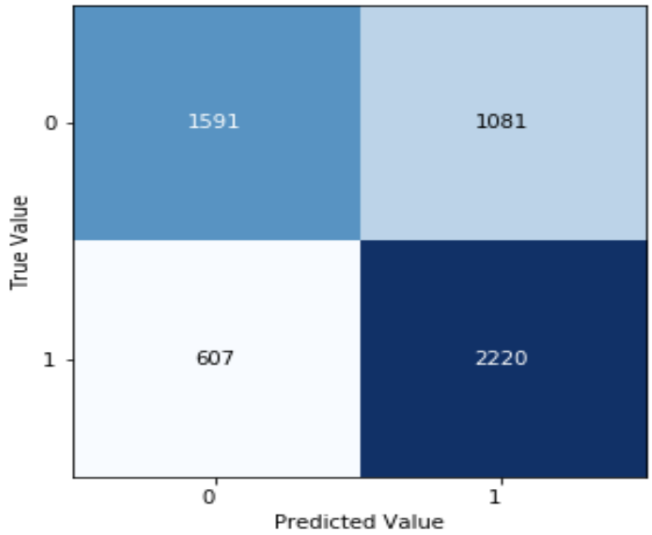

Figure 5. 14: Confusion matrix of neural network model with two hidden layers on WoE encoded data.

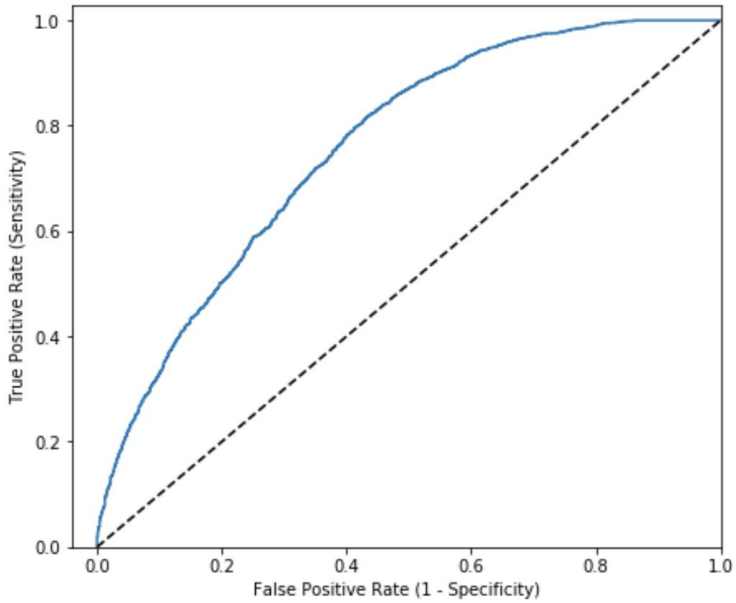

Figure 5. 15: ROC curve of neural network model with two hidden layers on WoE encoded data.

5.2 Performance of Logistic Regression Model with Interaction Terms

The performance of the logistic regression model with interaction terms was tested using the interactions detected from the random forest classifier. Three interactions were selected according to their strengths and added to the logistic regression models with the different data representations and on balanced datasets.

For the first logistic regression model, the same one-hot-encoded data and variables is used in addition to the three interaction terms: HOLDING_PERIOD*OCCUPATION.SAL, FLG.HAS.OLD.LOAN.0*OCCUPATION.SAL and NO_OF_CHQ_DR_TXNS*OCCUPATION.SAL. The confusion matrix for the model is shown in Figure 5.16, the model predicted 2297 of responders accurately and had a number of FPs of 508.

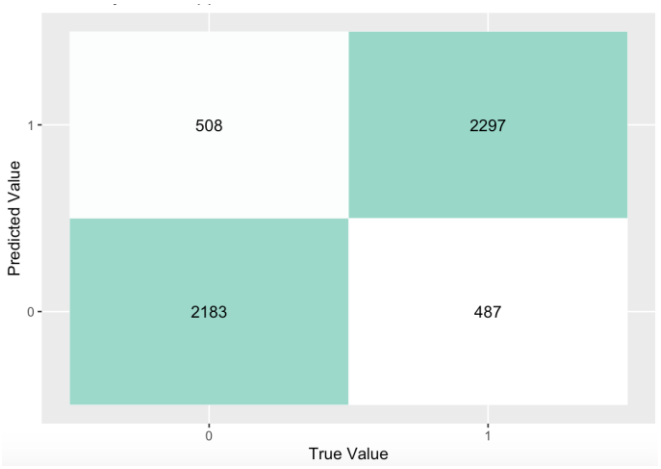

Figure 5. 16: Confusion matrix of logistic regression model with interactions on one-hot-encoded data.

Table 5.7 summaries the performance of the model with interaction terms. The model achieved an accuracy of 81.8% and a recall of 82.5% which is higher than the model without interaction terms. The model resulted a higher precision than the model without interactions however with a small difference of 0.0000309. Also, the AUC increased when the interactions were added to model, the ROC curve for the logistic regression model with interactions on one-hot-encoded data is shown in Figure 5.17.

Table 5. 7: Logistic regression model with interactions performance on one-hot-encoded data.

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.8182648	0.8188948	0.8250718	0.8219717	0.8902

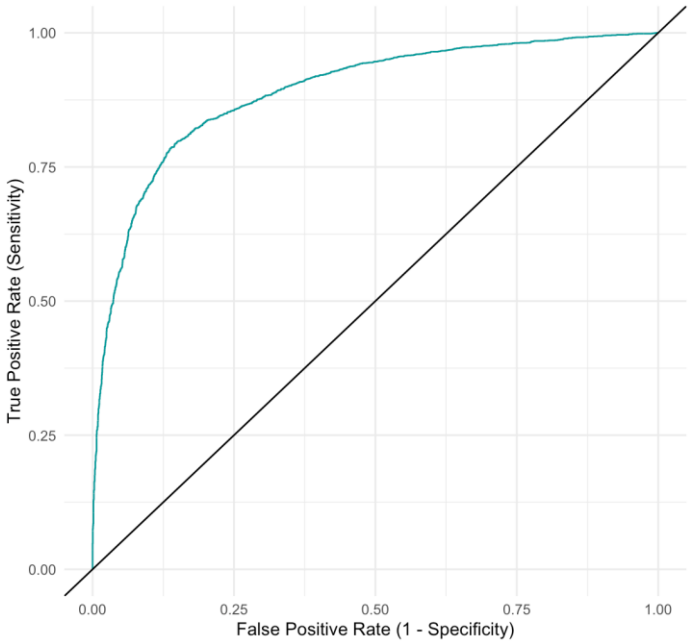

Figure 5. 17: ROC curve of logistic regression model with interactions on one-hot-encoded data.

The second logistic regression model is trained using the same WoE encoded data and variables but the interaction terms HOLDING_PERIOD*OCCUPATION.SAL, ACC_TYPE*OCCUPATION.SAL and FLG_HAS_OLD_LOAN*OCCUPATION.SAL were added to the model. Figure 5.18 represents the confusion matrix of the model with interaction terms on WoE encoded data, the model classified 1895 of the responders and 1746 of non-responders accurately.

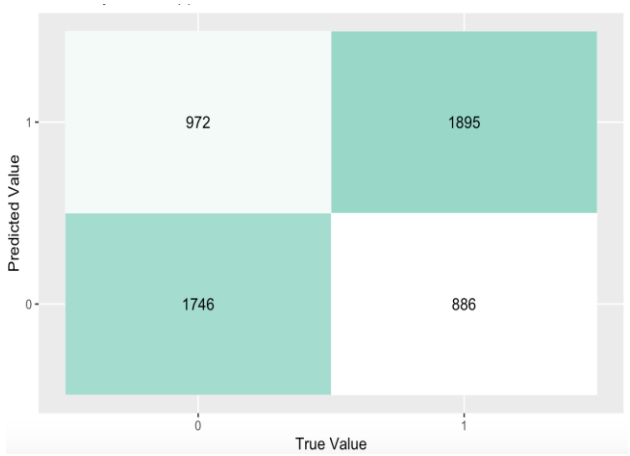

Figure 5. 18: Confusion matrix of logistic regression model with interactions on WoE encoded data.

The model with interactions reached an accuracy of 66.2% as shown in Table 5.8. The precision for the model is 66% and the recall is 68.1%, the model with interaction terms successfully achieved higher results of accuracy, precision, recall and F1 score by 0.002 increase on WoE encoded data compared with model without the interaction terms. The AUC for the model with interaction is 0.7186 that is higher the AUC of the model without interactions. The ROC curve of the logistic regression model with interactions on WoE encoded data is shown in Figure 5.19.

Table 5. 8: Logistic regression model with interactions performance on WoE encoded data.

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.6621204	0.6609697	0.6814096	0.6710340	0.7186

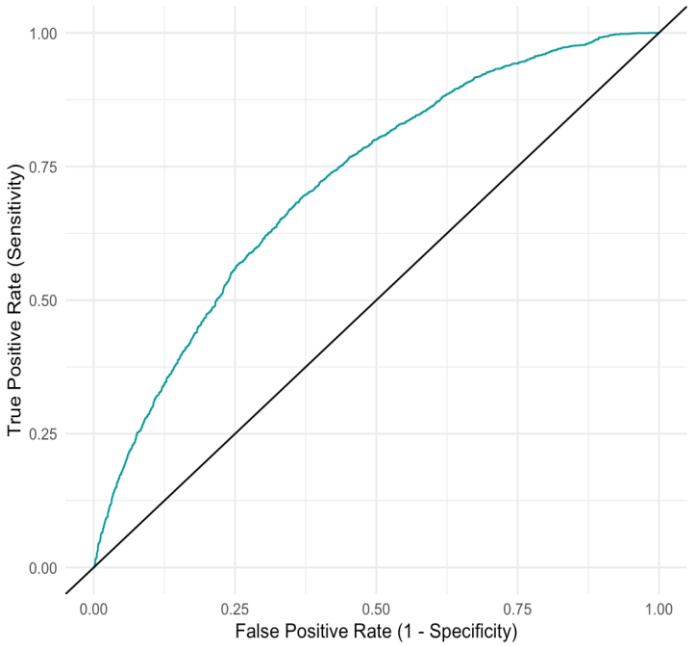

Figure 5. 19: ROC curve of logistic regression model with interactions on WoE encoded data.

The last logistic regression model with interactions was trained on the same one-hot-encoded categorical variables and normalized numerical variables. The interactions terms are the same as the interactions used for the first logistic regression model: HOLDING_PERIOD*OCCUPATION.SAL, FLG.HAS.OLD.LOAN.0*OCCUPATION.SAL and NO_OF_CHQ_DR_TXNS* OCCUPATION.SAL. The confusion matrix for the model with interaction terms on normalized data is represented in Figure 5.20, the model resulted a number of FPs of 460 and a number of FNs of 519.

Figure 5. 20: Confusion matrix of logistic regression model with interactions on normalized data.

The results on the logistic regression model with interactions on normalized data are presented in Table 5.9. The accuracy of the model is 82.1%, the precision is 83.2%, the recall is 81.4% and the model resulted a F1 score of 82.3%. The AUC is 0.8861 and the ROC curve of the model is shown in Figure 5.21. As the previous logistic regression models with interaction terms, the results of the logistic regression model with interactions on normalized data increased the performance compared with the model without interactions on the same normalized data.

Table 5. 9: Logistic regression model with interactions performance on normalized data.

Model	Accuracy	precision	recall	F1	AUC
Logistic regression	0.8211872	0.8321168	0.8145766	0.8232533	0.8861

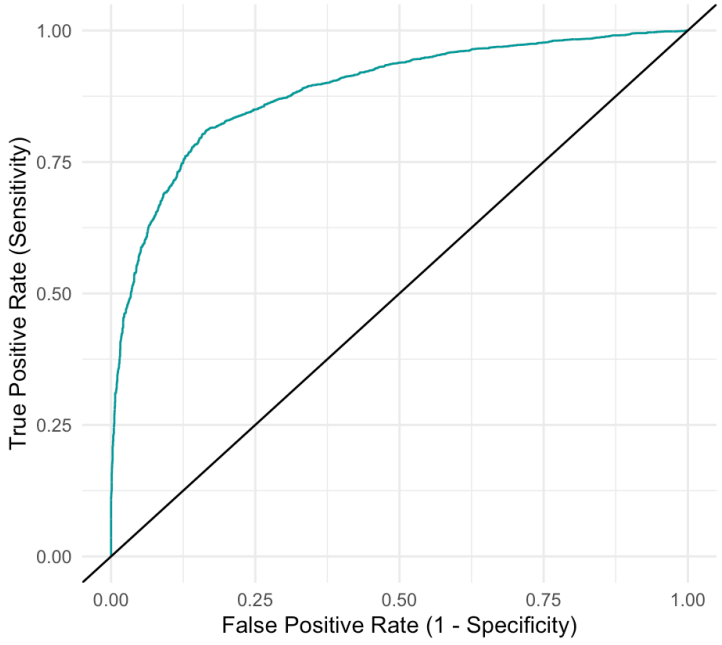

Figure 5. 21: ROC curve of logistic regression model with interactions on normalized data.

5.3 Performance on Imbalanced Data

To examine the effect of balancing the data on the models' performance, the algorithms were tested on imbalanced datasets, Table 5.10 summaries the results of logistic regression model with three data representations: one-hot-encoded data, WoE encoded data and normalized data. Although the performance of the logistic regression model on the imbalanced data is higher in terms of accuracy, the precision, recall and F1 have poor results compared to the balanced model on the same data representations. The case is different for the model trained on the imbalanced WoE encoded data, the accuracy, precision and AUC are higher than the balanced model but the recall and F1 score are lower in the imbalanced model. Between the three models the model trained on the WoE encoded data achieved the best performance on all evaluation metrics concluding that encoding the data by WoE can achieve high results even when the data is imbalanced.

Table 5. 10: Performance of logistic regression model with different data representation on imbalanced data.

Model	Accuracy	precision	recall	F1	AUC
Logistic Regression (one-hot-encoded data)	0.889466	0.571428	0.011730	0.022988	0.7099
Logistic regression (WoE encoded data)	0.889791	0.764705	0.037356	0.071232	0.732
Logistic regression (normalized data)	0.883615	0.647058	0.030303	0.057894	0.7022

The performance of the random forest decreased when trained on imbalanced data set as shown in Table 5.11. The model has an accuracy of 89.5% and precision of 86.1%. The recall for the model on imbalanced model highly decreased by 0.81 compared to the random forest model on the balanced data.

Table 5. 11: Performance of random forest classifier on imbalanced data.

Model	Accuracy	precision	recall	F1	AUC
Random Forest	0.895968	0.861111	0.0895953	0.1623036	0.756

The Neural network performance on imbalanced normalized data is summaries in Table 5.12. The performance of the neural network models with different number of hidden layers and nodes decreases in all metrics when trained on imbalanced normalized data compared to the model trained on the balanced

normalized data. Between the three models the neural network with two hidden layers have the best performance on the imbalanced dataset.

Table 5. 12: Neural network models' performance on the imbalanced normalized data.

Model	Accuracy	precision	recall	F1	AUC
Neural network (one hidden layer with 3 nodes)	0.886902	0.310344	0.026706	0.049180	0.69800
Neural network (one hidden layer with 5 nodes)	0.884302	0.306122	0.044510	0.077720	0.67871
Neural network (two hidden layers with 3 and 5 nodes)	0.887552	0.4	0.053412	0.094240	0.69401

On the other hand, the performance of the neural network models with different numbers of hidden layers and nodes on imbalanced WoE encoded data increased in classification accuracy compared to the neural network model on the same balanced data representation as shown in Table 5.13. However, the precision, recall, F1 and AUC metrics decreased when the models are trained on the imbalance dataset. The best performing neural network model on the imbalanced WoE encoded data is the model with two hidden layers.

Table 5. 13: Neural network models' performance on the imbalanced WoE encoded data.

Model	Accuracy	precision	recall	F1	AUC
Neural network (one hidden layer with 3 nodes)	0.887227	0.426470	0.086053	0.143209	0.73335
Neural Network (one hidden layer with 5 nodes)	0.886902	0.426666	0.094955	0.155339	0.71186
Neural Network (two hidden layers with 3 and 5 nodes)	0.893727	0.580645	0.106824	0.180451	0.73847

5.4 Comparison

The best performing algorithm for the personal loan campaign response in this project is the random forest. It achieved the highest classification accuracy, recall and AUC compared to the logistic regression and neural network algorithms. Also, balancing the data provided more responder observations to the model to train on therefore building a more robust model to predict the responders and the performance of the random forest classifier was affected when the data is imbalanced. Additionally, the random forest had the advantage of deriving the feature importance and detecting the feature interactions resulting in a more interpretable model that can be reliable in the marketing campaign design.

The logistic regression models were built using variables that achieved a value of VIF less than 10. The best performing logistic regression model in classification accuracy and precision is the model trained on normalized and balanced data. However, when the data is imbalanced, the models trained on the different data representations resulted in higher accuracies but the precision, the recall, F1 score and AUC are lower compared to when the models are trained on balanced data. The performance on WoE encoded data was the weakest when the data is balanced but it had the best performance in all metrics when the data is imbalanced compared to the models on other data representations on the imbalanced data. Adding the interaction terms detected from the random forest classifier to the logistic regression models enhanced their outcomes in all the evaluation metrics.

The neural network model showed better performance when the data is balanced and normalized with two hidden layers and it can be ranked as the second-best performing model after the random forest. In all data representations used to train the neural network model, the model with two hidden layers and with 3 and 5 nodes in each layer accomplished higher results than the models with a single hidden layer even when the data is imbalanced. For both the logistic regression and neural network models trained on the WoE encoded data the performance was less when the data is balanced and better when the data is imbalanced. Generally, balancing the data improved the performance for all the models in terms of precision, recall, F1 score and AUC evaluation metrics.

Section 6

Conclusions

In this section, the project summary and conclusions are presented.

Marketing campaign planning aims at maximizing the revenue with the efficiency of resource utilization and minimizing costs. Machine learning algorithms can support the campaign planning process in predicting the most likely customers to accept the offer and in discovering their patterns to gain the most effective way of the Return on Investment (ROI) and eventually the profit of the firms.

This project targets to predict the bank customer response to a personal loan campaign. It is shown that it is possible with an accuracy of 94.4% to predict responders using the random forest algorithm. Furthermore, the random forest showed that the flag has credit card equal to zero, salaried occupation, and holding period are the most important variables in predicting the customer response. The random forest also detected the interaction term ACC.TYPE.SA*OCCUPATION.SAL to be the strongest feature interaction in the personal loan campaign data using the H-statistic measure. It is worth mentioning that the H-statistic is expensive to evaluate as it iterates and assess the partial dependency over all data points. The random forest algorithm was successful in predicting the customer response with high accuracy and provide more information about the variables' importance and interactions providing a higher level of the model interpretability.

The logistic regression algorithm is used to derive a marketing response model for target marketing because it yields a probability that could be used by a business in deciding whether to reach a customer or not. The model was trained on different data representations and with the interaction terms extracted from the random forest classifier. The logistic regression model did not perform the best to predict the customer response in this context, usually, the logistic regression models fail in situations where the relationship between features and outcome is nonlinear or where features interact with each other therefore other machine learning algorithms can be used that includes the effects of interactions and nonlinear data modeling such as the neural network algorithm.

The neural network algorithm was trained on two different data representations, the data normalization and WoE encoded. Normalizing and balancing the data lead to the attainment of high classification accuracy of 92.3% in the neural network model with two hidden layers.

The results are based on the highly probable segment of customers who will reply to the campaign in accordance with their demographic information and transactional history. In this project, only the two-way feature interactions are detected and their impact is investigated when added to the logistic regression model. For further work, a higher order of interactions can be detected and its impact can be explored on the logistics regression model.

References

[1] Q. Yu, H. Jiang and X. Ma, "The Application of Data Mining Technology in Customer Relationship Management of Commercial Banks," in *2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 2018, pp. 1368-1373.

[2] J. Asare-Frempong and M. Jayabalan, "Predicting Customer Response to Bank Direct Telemarketing Campaign," in *In 2017 International Conference on Engineering Technology and Technopreneurship (ICE2T)*, 2017, pp. 1-4.

[3] V. L. Miguéis, A. S. Camanho and J. Borges, "Predicting direct marketing response in banking: comparison of class imbalance methods," *Service Business,* vol. 11, no. 4, pp. 831-849, 2017.

[4] J. Y. Shih, W. H. Chen and Y. J. Chang, "Developing Target Marketing Models for Personal Loans," in *2014 IEEE International Conference on Industrial Engineering and Engineering Management*, 2014, pp. 1347-1351.

[5] K. Coussement, P. Harrigan and D. F. Benoit, "Improving direct mail targeting through customer response modeling," *Expert Systems With Applications,* vol. 42, no. 22, pp. 8403-8412, 2015.

[6] P. Ładyżyński, K. Żbikowski and P. Gawrysiak, "Direct marketing campaigns in retail banking with the use of deep learning and random forests," *Expert Systems With Applications,* vol. 134, pp. 28-35, 2019.

[7] N. Barraza, S. Moro, M. Ferreyra and A. de la Peña, "Mutual information and sensitivity analysis for feature selection in customer targeting: A comparative study," *Journal of Information Science,* vol. 45, no. 1, pp. 53-67, 2019.

[8] S. Lessmann, J. Haupt, K. Coussement and K. W. De Bock, "Targeting customers for profit: An ensemble learning framework to support marketing decision-making," *Information Sciences,* 2019.

[9] A. Zakaryazad and E. Duman, "A profit-driven Artificial Neural Network (ANN) with applications to fraud detection and direct marketing," *Neurocomputing,* vol. 175, pp. 121-131, 2016.

[10] Y. Wang, X. S. Ni and B. Stone, "A two-stage hybrid model by using artificial neural networks as feature construction algorithms," *International Journal of Data Mining & Knowledge Management Process (IJDKP),* vol. 8, no. 6, November 2018.

[11] Y. Wang, X. S. Ni and B. Stone, "An Automatic Interaction Detection Hybrid Model for Bankcard Response Classification," in *2018 5th International Conference on Systems and Informatics (ICSAI)*, 2018, pp. 1111-1119.

[12] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, "Classification and regression trees," *CRC press,* 1984.

[13] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[14] J. Du and A. R. Linero, "Interaction Detection with Bayesian Decision Tree Ensembles," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019, pp. 108-117.

[15] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," *The Annals of Applied Statistics,* vol. 2, no. 3, pp. 916-954, 2008.

[16] S. Basu, K. Kumbier, J. B. Brown and B. Yu, "iterative Random Forests to discover predictive and stable high-order interactions," *Proceedings of the National Academy of Sciences,* vol. 115, no. 8, pp. 1943-1948, 2018.

[17] R. Beale and T. Jackson, Neural Computing - an introduction, CRC Press, 1990.

[18] "Standford UFLDL Tutorial. Multi-Layer Neural Network," [Online]. Available: http://ufldl.stanford.edu/tutorial/supervised/MultiLayerNeuralNetworks/.

[19] N. Kyurkchiev and S. Markov, Sigmoid functions: some approximation and modelling aspects., Saarbrucken: LAP LAMBERT Academic Publishing, 2015.

[20] "Loan Campaign response," [Online]. Available: https://www.kaggle.com/dineshmk594/loan-campaign#PL_XSELL.csv .

[21] B. Swiderski, J. Kurek and S. Osowski, "Multistage classification by using logistic regression and neural networks for assessment of financial condition of company," *Decision Support Systems,* vol. 52, no. 2, pp. 539-547, 2012.

[22] H. He, Y. Bai, E. A. Garcia and S. Li, "ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning," in *2008 IEEE international joint conference on neural networks (IJCNN 2008)*, 2008, pp. 1322-1328.

[23] L. E. B. Ferreira, J. P. Barddal, F. Enembreck and H. M. Gomes, "An Experimental Perspective on Sampling Methods for Imbalanced Learning from Financial Databases," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1-6.