UNIVERSITY OF SURREY

Report: Credit Card Approval Prediction Machine Learning

Submitted to Prof. Colin Fu

MACHINE LEARNING & VISUALISATION (MANM354)

Department of Surrey Business School

Report Prepared By:

Eashwar Thyagarajan - 6713922

Komal Luthra Verma - 6712133

Lucky Sikka - 6712909

Mansi Goyal - 6709502

Olayemi Agogbua - 6708359

Souvik Nandi - 6698258

MSc BUSINESS ANALYTICS

(September 2021-2022)

Total Word Count: 4744

INTRODUCTION

Credit cards have become one of the most popular modes of payment in the world today. It has therefore become pertinent for issuers of credit cards to identify the risks involved or the probability of the consumers failing to meet their obligations thereby incurring costs. To mitigate this risk, financial institutions utilise efficient credit risk evaluation tools such as credit scores which are based on data on the potential customers' sociodemographic status, credit reports and other criteria, before issuing credit. A credit scoring system reduces human errors and increases the accuracy and speed of credit risk determination which in turn reduces the time and process of granting loans, the costs and risks of granting loans, and increases the efficiency and transparency of the bank. The purpose of the credit scoring system, therefore, is to announce the quality of the credit customer requesting the loan and to predict their repayment behaviour.

The objective of this report is to develop and critically evaluate the machine learning models to predict the creditworthiness of credit card applicants, compare models based on certain performance metrics such as their accuracy, sensitivity and specificity and to devise a business strategy for profitability.

In the case where the target feature to be predicted is whether an applicant will default on a loan, the supervised learning algorithms are mostly used, and fall under the classification models. A wide range of classification methods such as Bayesian Network, Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, K-Nearest Neighbour, Decision Tree, Random Forest, Neural Network etc. are frequently used to detect financial risks such as the credit worthiness of loan applicants (Peng et al., 2011). The performance of the classifier may vary based on the data, performance measure and other circumstances therefore the selection of a suitable classifier remains an important task in the prediction of financial risk.

METHODOLOGY

The approach to the task was an inductive approach, with the aim of building predictive models that identify customers who should be approved for credit cards.

The six steps of the Cross-Industry Standard Process for Data Mining (CRISP-DM) iterative process were employed in developing the models for this credit card approval predictive analysis task. The steps are broken down in the methodology below.

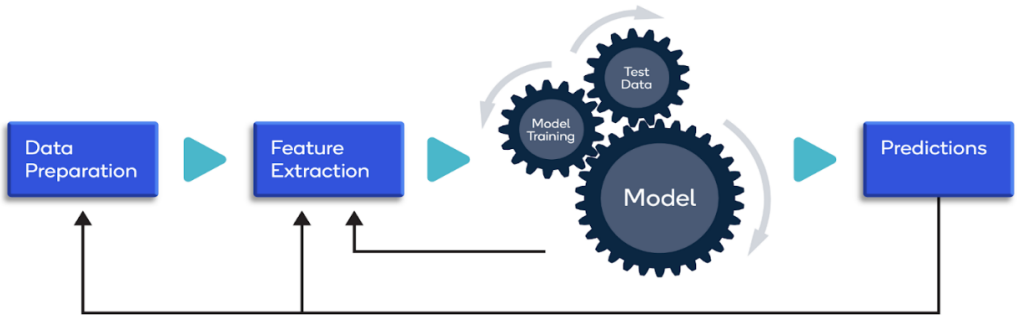

Figure 1: Machine Learning process

Data collection

The data used in this study were obtained from Kaggle.com. It consists of two datasets namely, credit_record.csv and application_record.csv. The application_record.csv has 18 variables and contains information on the customers' socio-economic status such as income, children and house ownership. The credit_record.csv has 3 variables and contains the payment status for each client. Both datasets are connected by the variable ID. Neither have a target variable and this is a task that was identified to be attended to during the pre-processing stage.

Data exploration was carried out to understand the features of the datasets using the data descriptor and identify the variable types to note the ones that may need to be converted. A quality assessment was then affected to check for missing values and assess the most appropriate method to treat them, to achieve an effective data mining process.

Data preparation

The main aim of this stage was to create the final dataset from the raw initial data that will be used to create the models. The major tasks carried out in this stage include cleaning of data, addressing missing values, determining a target variable, standardising the data and creating new features till it was satisfactory for modelling. The two datasets were initially merged using the unique ID. Insights discovery through the use of tools such as descriptive statistics, plots and graphs were useful in data cleansing by identifying missing values, addressing empty cells and observing the effects of one variable on another. This process used in combination with the correlation matrix also helped in feature selection and exclusion. From the insights gained, new data was constructed such as assigning points to each customer based on the time that has elapsed since payment was due, to easily categorise the credit status of the customers. New data was also constructed through the conversion of features such as days of birth, days

employed and civil marriage. Finally, some of the variables were formatted to facilitate further statistical analysis. More details on this stage are outlined in the pre-processing section.

Modelling

Machine learning was used to create predictive models firstly by splitting the data into training and test data. To achieve a detailed comparative analysis three categories of test and training data in the ratios of 70:30, 80:20 and 90:10 were created. The choice of models to be used were based on two key factors – the primary objective of the task which is to predict customers that should be approved for credit card applications combined with the available literature reviews on machine learning techniques used in finance to predict credit risk. Based on this, C5, Random Forest, Naïve Bayes and Neural Networks were deemed to be the most suitable models to achieve the desired objectives. More details on this as well as the rationale for each model selection and outputs are discussed in the modelling section.

Evaluation

The test and training data of each of the three categories defined above were run through each model and assessed based on their accuracy and their alignment with the outlined objectives and success criteria. The first level of accuracy was measured using Confusion Matrix, a fundamental term in machine learning (Kohavi and Provost 1998) and more specifically in credit scoring (Siddiqi 2006; Refaat 2011; Thomas 2009) to measure the accuracy of a model by comparing the number of true viable and non-viable credit card applicants with the number of predicted viable and non-viable credit card applicants for a certain cutoff score. Based on the business objective, of predicting applicants who should be approved for credit cards, a priority is to reduce the number false positives which is the applicants which the model predicts are creditworthy but are not as the risk in this is that credit cards may be given to applicants who will default and the company will lose money.

DATA UNDERSTANDING / VISUALISATION

Data visualisation is a useful tool in machine learning for identifying patterns and understanding trends in the data that are critical for making data-driven decisions.

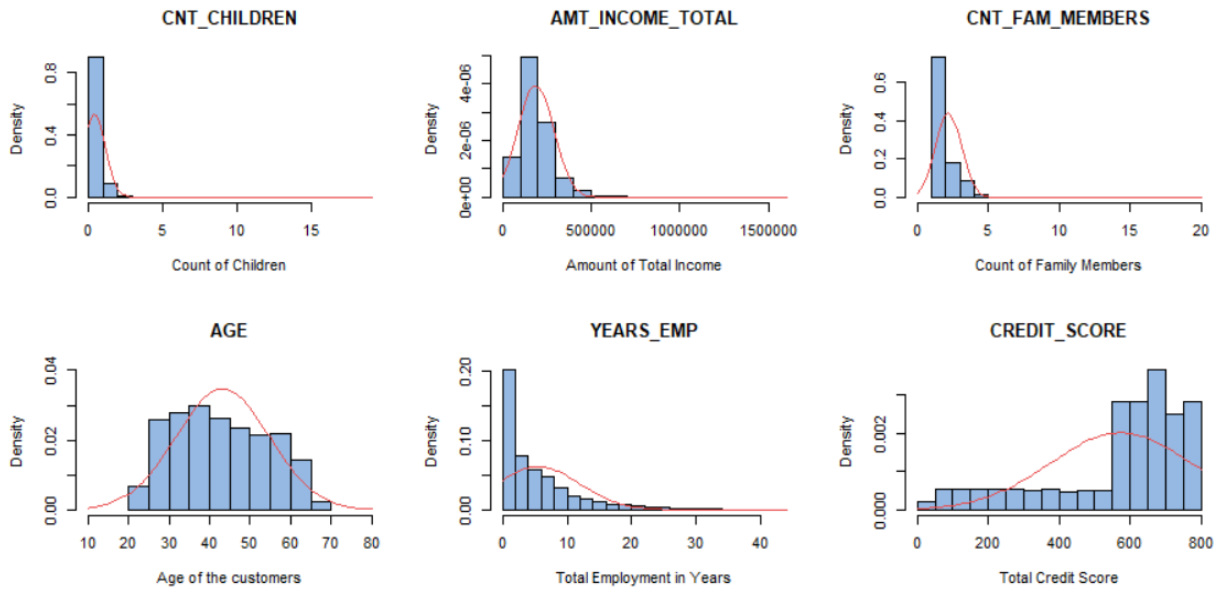

Figure 2: Histogram of Continuous Variables

The histograms plotted in figure 2 demonstrate the distribution of 6 continuous variables in the dataset. The variable CNT_CHILDREN gives the number of children each customer has. This variable is positively skewed with an outlier of 19. Most of the customers have 1 child. The majority have an income of less than 500,000, although the range of income is from 27,000 to 1,575,00. The number of family members for most of the customers is 2, however, there are very few customers having more than 5 members in their family. The age of the customers follows a nearly normal distribution ranging from 20 to 68 years. The years of employment a is positively skewed variable ranging from 0 to 43 years out of which 0 and have the most number of observations. The variable Credit score is negatively skewed. Most of the customers have credit a scores of more than 550.

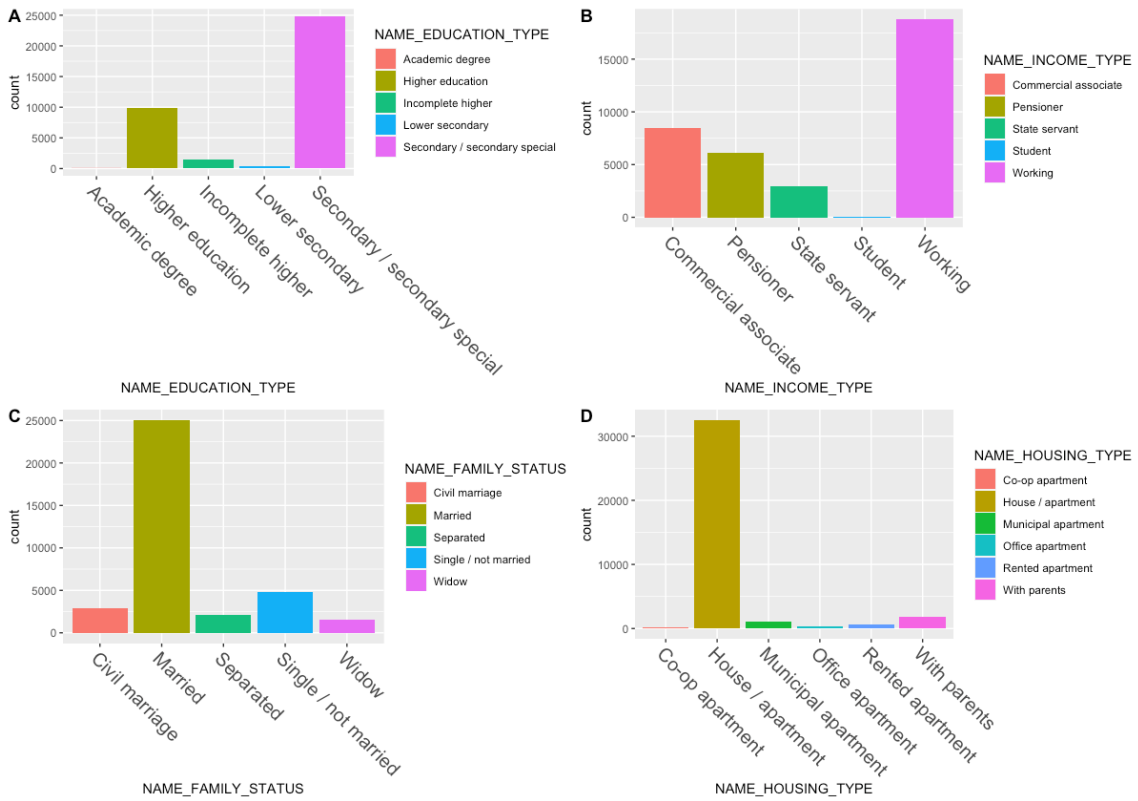

Figure 3: Bar-plot for Categorical Variables

The bar charts in **Figure 3A** display the count of the various educational levels of customers in the given dataset. The distribution reveals that the majority of the customers have attained secondary/secondary special qualifications. This stands at about 25,000 customers which is more than twice the number of customers with higher education qualification, the second-largest category in the dataset. Customers with lower secondary, incomplete higher and graduate degrees account for less than 15% of the entire dataset

The customers' varied sources of income are represented in the bar chart in **Figure 3B.** Working customers account for approximately half of the dataset, followed by commercial associates, Pensioners and state employees account for the balance of the dataset and the number of students in the dataset is negligible.

Figure 3C displays the relationship status of customers in the dataset. Married customers represent a majority of the dataset with a count of 25,000. The balance 30% customers are single/not married, civil marriage, separated, and widowed.

Figure 3D displays the types of accommodation the customers reside in. Over 30,000 customers dwell in homes or apartments while the balance of approximately 2,000 live in municipal apartments, office apartments, co-op apartments or with parents.

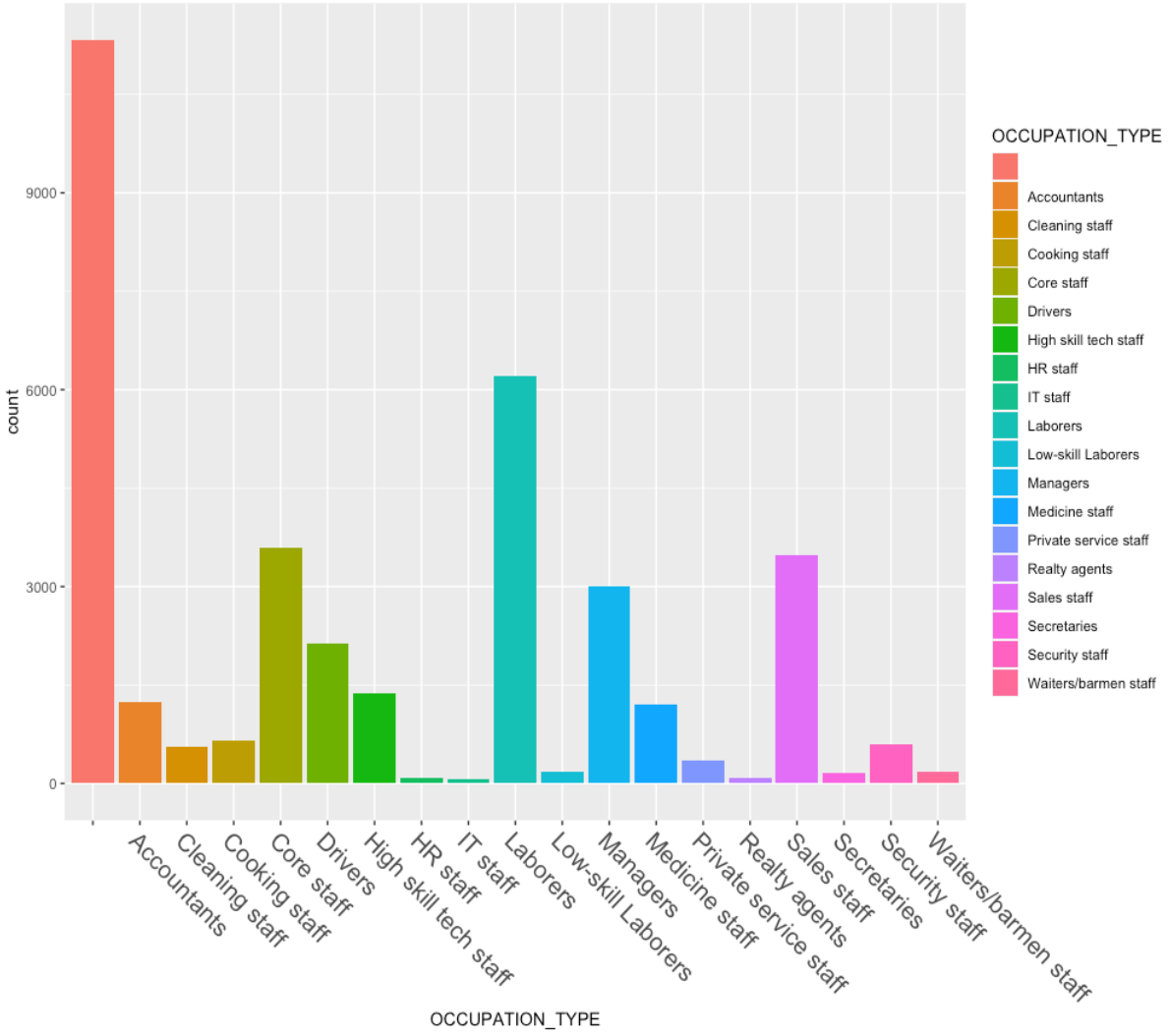

Figure 4: Bar plot for Occupation Type

The bar plot in figure 4 displays the various occupation types of the customers in the datasets. With over 11,000 customers, over 30% of the data are not labelled, indicating a high percentage of missing values in this variable. From the data that is labelled, it can be observed that labourers represent a majority of the customers which may be correlated with the high number of mid-level educational qualifications. Other occupations with high representation in the data are Core staff, Sales staff and managers. The rest of the data is a mix of blue-collar and white-collar jobs.

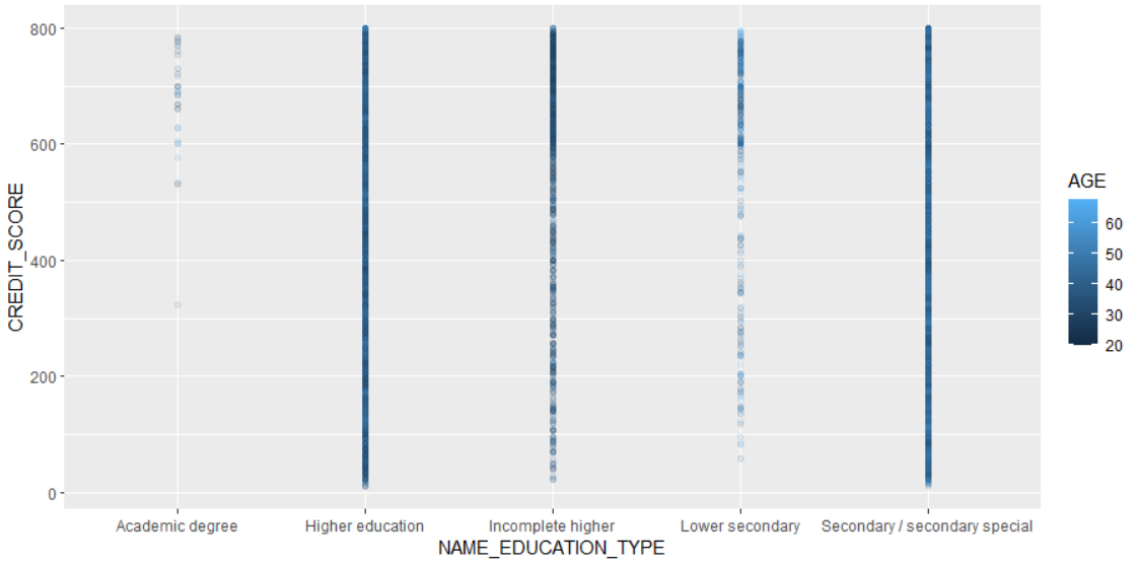

Figure 5: Scatterplot for relation between Credit Score, Education Type and Age

The scatter plot in figure 5 demonstrates the credit score distribution on various education types in the dataset. Customers with higher education and secondary/ secondary special education types have varied credit scores which range from very low to very high. Customers with academic degrees have maintained consistently high credit scores. Customers with lower secondary and incomplete higher have both with mostly high credit scores, although there are many instances of low credit scores as well.

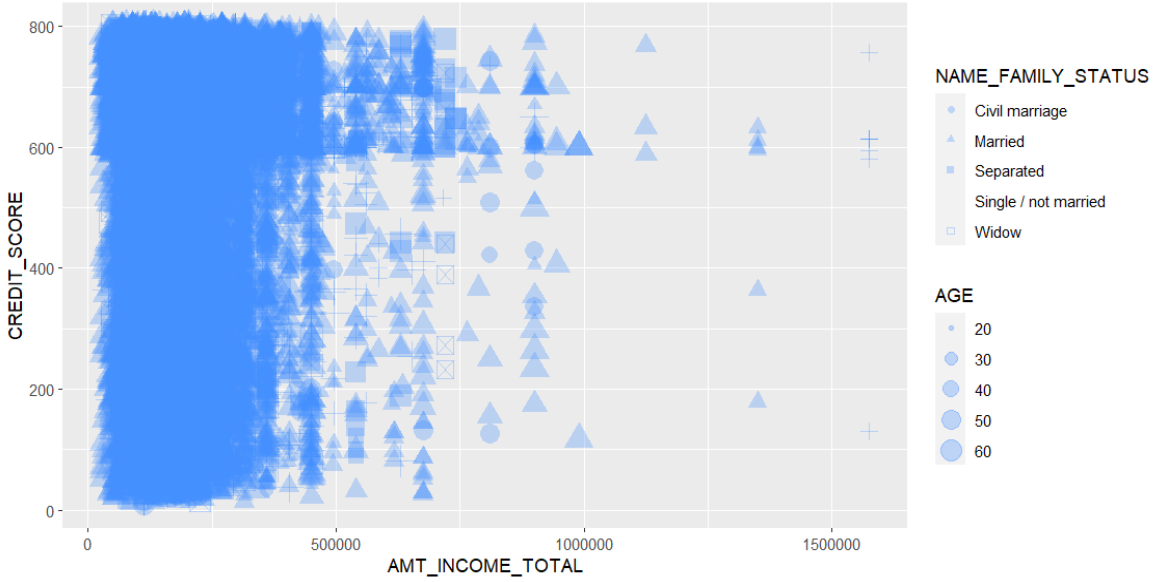

Figure 6: Scatterplot for relation between Credit Score, Income, Age & Family Status

The scatter plot in figure 6 illustrates the relationship between credit score, age, relationship status and income. All separated and widowed customers' incomes are less than 750,000.

Customers with relationship status as civil marriage and income of more than 750,000 have maintained higher credit scores in most of the cases. Higher age relates to high income, which is fairly intuitive. Single/ not married customers with incomes in the range of 750,000 to 1,500,000 represent a very small portion of the data. It is evident that customers with higher income have mostly maintained high credit scores.

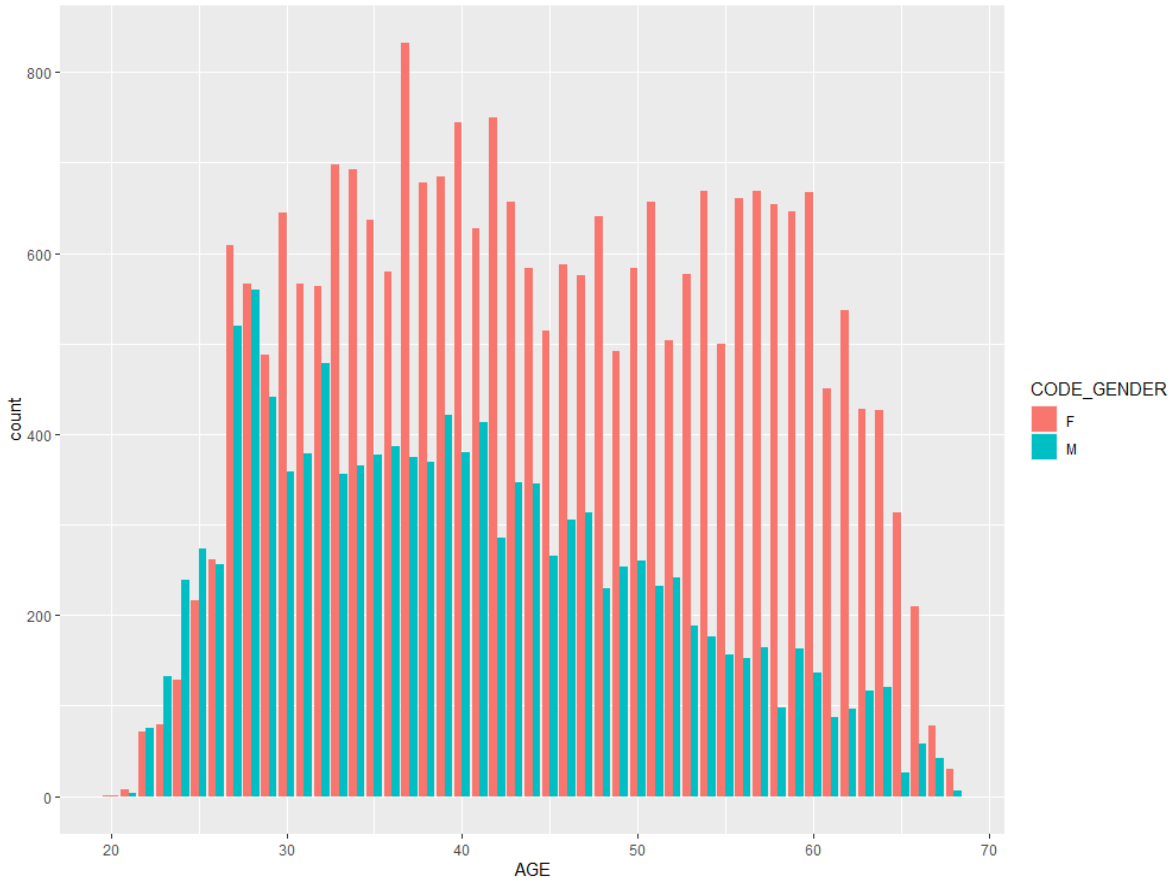

Figure 7: Barplot for Age and Gender distribution

The bar plot above in figure 7 depicts the age range of the customers as well as the total number of male and female customers who fall into the various age groups. It can be observed from the graph, that the customers' age range from 20 years to 70 years, with a critical mass falling between 27 years and 65 years. The graph also shows that the total number of female customers is much higher than the total number of male customers in the majority of the age categories.

Correlation

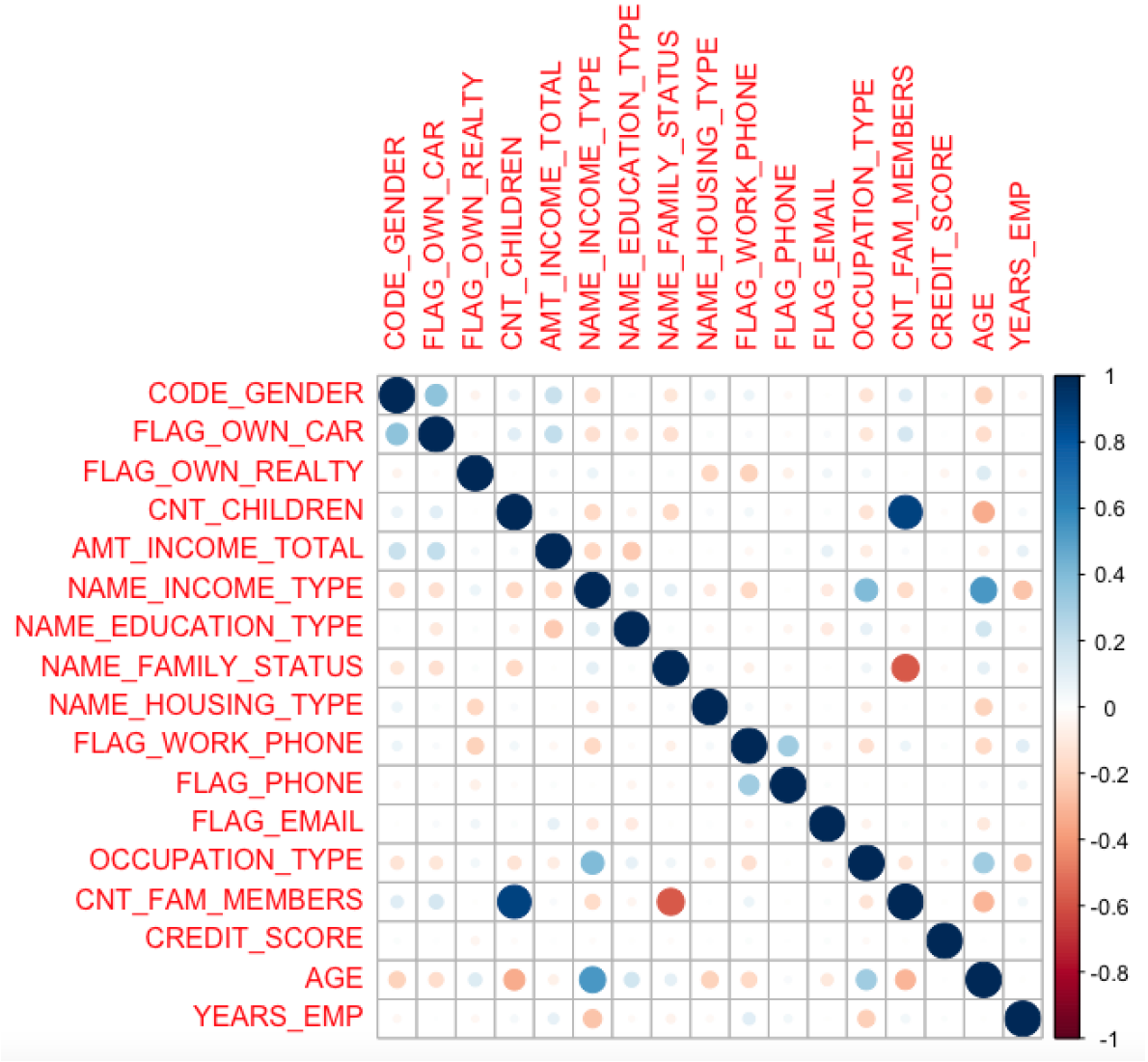

Figure 8: Correlation matrix

The correlation matrix in figure 8 shows the linear relationship between the features in the dataset. Some features having multicollinearity to each other have been removed to increase the precision of the models. The features having high correlation with more than one variable are removed first. CNT_FAM_MEMBERS is removed as it is positively correlated with CNT_CHILDREN and negatively correlated with NAME_FAMILY_STATUS. NAME_INCOME_TYPE has been removed as it has multicollinearity with AGE and OCCUPATION_TYPE. FLAG_OWN_CAR has been removed as it has multicollinearity with CODE_GENDER.

DATA PRE-PROCESSING

Data pre-processing was carried out to improve the quality of the data, promote the extraction of meaningful insights and make it suitable for the construction and training of machine learning algorithms. The dataset was acquired from https://tinyurl.com/mven9zp9 and consists of two csv files namely, the **application record** and **credit record**. The application_record.csv contains information about the customers' socio-economic status such as income, children and house ownership. This file contains 18 numeric and non-numeric variables which are as follows:

Table 1: Attribute description - Application record

Attributes	**Description**
ID	Unique Identification Number of the client
CODE_GENDER	Gender
FLAG_OWN_CAR	Car ownership status
FLAG_OWN_REALTY	Property ownership status
CNT_CHILDREN	Number of Children
AMT_INCOME_TOTAL	Annual Income
NAME_INCOME_TYPE	Income Category
NAME_EDUCATION_TYPE	Education Status
NAME_FAMILY_STATUS	Marital Status
NAME_HOUSING_TYPE	Type of accommodation
DAYS_BIRTH	Number of days since birth
DAYS_EMPLOYED	Number of days since employment
FLAG_MOBIL	Mobile phone ownership status
FLAG_WORK_PHONE	Work phone ownership status

FLAG_PHONE	Phone ownership status
FLAG_EMAIL	Email ownership status
OCCUPATION_TYPE	Occupation
CNT_FAM_MEMBERS	Family size

The credit_record.csv file consists of all payment history and default records for a given customer. This file comprises 3 fields which are explained below:

Table 2: Attributes description - Credit record

Attributes	**Description**
ID	Unique identification number of the client
Months_Balance	The month of the data collected is the beginning point; counting backwards, 0 represents the current month, -1 represents the previous month, etc.
Status	0: 1-29 days past due 1: 30-59 days past due 2: 60-89 days late 3: 90-119 days overdue 4: 120-149 days overdue 5: Past-due or delinquent loans, write-offs older than 150 days C: settled that month No loan for the month of X.

The 'months_balance' variable, which represents the duration of the loan, using the month the data was extracted as the baseline, was formatted as it contained positive and negative numbers, which complicated the analysis, by deriving a new variable using the total number of months payment due. To perform the supervised machine learning techniques, the target variable was created using the credit record dataset.

Credit points from 100 to 800 have been assigned to each customer based on their payment history according to the month's balance. Afterwards, Credit Score was calculated by averaging the credit points for each customer. A new variable named 'Credit_score' was then generated* using the credit points.

Using factors that apply in the real world, the 'Credit_score' variable was derived to distinguish between viable and non-viable customers using a threshold of ≥ 600 where < 600 is considered as 'non-viable' or 'no' and assigned 1 and ≥ 600 is considered as 'viable' or 'yes' and assigned

0. The application record and credit record datasets were then merged, using a primary key i.e. **ID**, to create one data frame. A target variable **CREDIT_STATUS** has been created as a categorical variable using the threshold of credit scores.

Other variables in the dataset that were transformed are the 'Days Birth' and 'Days Employed.' The 'Days Birth' variable was converted to the 'age in years' variable and rounded off to the nearest integer to ease comprehension and computation. The 'Days Employed' variable, which contained positive and negative integers, was transformed by replacing the positive numbers, which represent number of days unemployed with 0 and then firstly converting the negative integers, which represent the number of days employed, to years and then rounding off to the nearest whole integer.

Following the integration of the two datasets to form one dataframe and the creation of derived attributes, it was necessary to cleanse the data and identify missing values, and how to treat them. Failure to do so may result in inaccurate and erroneous conclusions and inferences drawn from the data which may adversely affect the accuracy of the models. Missing values were detected in the 'Occupation type' field. These were first treated labelled as 'NA', and then replaced with 'unavailable' to aid in the interpretation of the data.

Since, the dataset contains a combination of character, numeric and factor variables. For accurate numerical and statistical analysis, the non-numerical variables needed to be converted to numerical variables. This was done by converting the categorical variables in the dataset to factors and subsequently checking the level. This was implemented on 'Name_income_type', 'Name_education_type', 'Name_family_status', 'Name_housing_type' and 'Name_occupation_type' variables.

For variables with categorical data such as 'Yes' and 'No', a **one-hot encoding** was applied. This entails expressing each category variable with a binary vector that has one element for each unique label and marking the class label with a 1 for 'Yes' while all other components are marked with a 0 for 'No.' This was applied to 'Code_gender', 'Flag_own_car' and 'Flag_own_realty' variables. The 'Flag mobile' variable was observed to have a single value of '1' for all the rows, indicating that all the customers owned mobile phones. Therefore, dropped from the dataset, to avoid biases and unbalancing in the dataset. The Customer IDs were also removed from the data as it is not relevant in determining whether a customer is viable or not for a credit card. Also, due to the problem of multicollinearity, the features such as Flag own car, Name income type, Count Family member have been dropped from the

dataset. A new dataframe has now been created to run the models. The feature CREDIT_SCORE has been removed, as the target variable CREDIT_STATUS has been created from the Credit Scores. Also, the models are to be trained to predict good or bad applications depending on the socio-economic characteristics of the applicant. Therefore, credit score should be dropped from the dataset.

The 'Days_birth' and 'Days employed' variables were also excluded as they had been transformed to the 'age' and 'years employed' variables as mentioned above and were no longer relevant for the required correlation, and regression analysis.

Standardisation has been used to normalise the variables in the dataset within a given range of values. Also referred to as feature scaling, this technique narrows the range of variables available for comparison, allowing them to be compared on a level playing field or standardised basis.

The final step that takes place in data pre-processing before modelling for machine learning begins is splitting. To prepare a dataset for a machine learning model, it is necessary to divide it into two separate sets: a training and a testing set. The training set refers to a subset of a dataset that is used to train a machine-learning model. This is the portion of the dataset that will be fed into the machine learning model to discover and learn patterns. As the name suggests, it trains the model. The test set, on the other hand, is a subset of a dataset that is used to evaluate the performance of a machine learning algorithm. For the purpose of this task, the data set was split into three categories of training and test sets of 70:30, 80:20 and 90:10. More details of the splitting are discussed in the modelling section.

PREDICTION MODELS - CLASSIFICATION

In this task, we have used classification methods for prediction, which is referred to as supervised learning classification, as the predicted class or target variable is known.

To address the business problem of predicting whether a credit card application will be approved, a decision must be made on whether an applicant is creditworthy or not. The implications of errors in the model are that a creditworthy customer could be declined and a customer who is prone to defaulting could be approved for a credit card. While both scenarios have consequences on the business, the risk of approving a customer who is not creditworthy has more impact on the bottom line as this could result in loss of funds.

Four prediction models were developed and evaluated based on their accuracy, specificity, and sensitivity in a quest to find a solution to the above business problem.

The initial step toward creating models is the feature selection, it is the most critical step before creating the prediction models. It is one of the key components of feature engineering and identifies the most important features to employ as predictors in machine learning algorithms. Due to the problem of multicollinearity, the features such as Flag own car, Name income type, Count Family member have been dropped. The feature flag mobile was also dropped from the dataset as it contains only one value, leaving the dataset unbalanced and biased. The overall objective of this stage is to utilise the most relevant features to create the most efficient model.

Another crucial step adopted in data modelling is to normalize the range of independent variables or features which makes it easy for a model to learn and understand the problem. The dataset was then divided into test and training sets. We have utilised three splits of (training to test) dataset of 70:30, 80:20, and 90:10 to achieve the best outcomes. We have not used the ‘FOR’ loop for data splits as the Neural network takes a long time to run due to its complexity.

In the models that have been created, ‘0’ is considered as a positive class. So, in the confusion matrix, '0' is interpreted as Positive and '1' as Negative.

The following models have been created and evaluated

1. C5 - Decision Tree:

This is the first model created for the dataset as the algorithm is majorly used to estimate the likelihood whether a customer would default on a loan by employing predictive model generation with the customer's historical data. It helps in preventing losses by evaluating a customer's creditworthiness.

Another advantage of using decision trees is that the amount of data cleansing that must be done is reduced. So, the performance of the algorithm is least affected by the presence of missing values or outliers. However, in this case, the missing values have already been treated.

The accuracy for this algorithm is 69.60%, highest for 70:30 (training: test) ratio. However, accuracy is not the only indicator for evaluating the model, there are other parameters such as specificity and sensitivity, which determine the best model. The confusion matrix for the best model in C5 depicts, 3195 customers are predicted to be good customers, however, they are bad customers and would default on payment. Also, we are missing out on 130 good customers which the model predicted as bad.

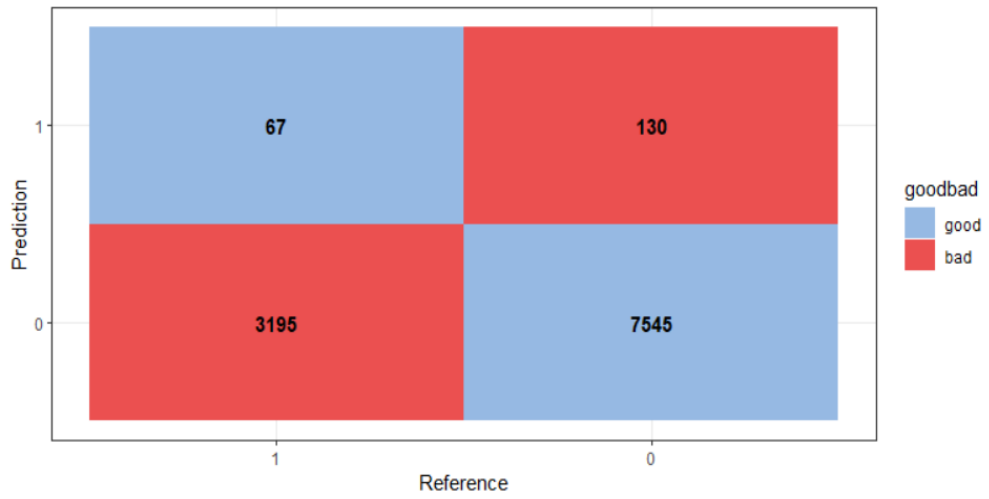

Figure 9: C5: Confusion Matrix (70:30)

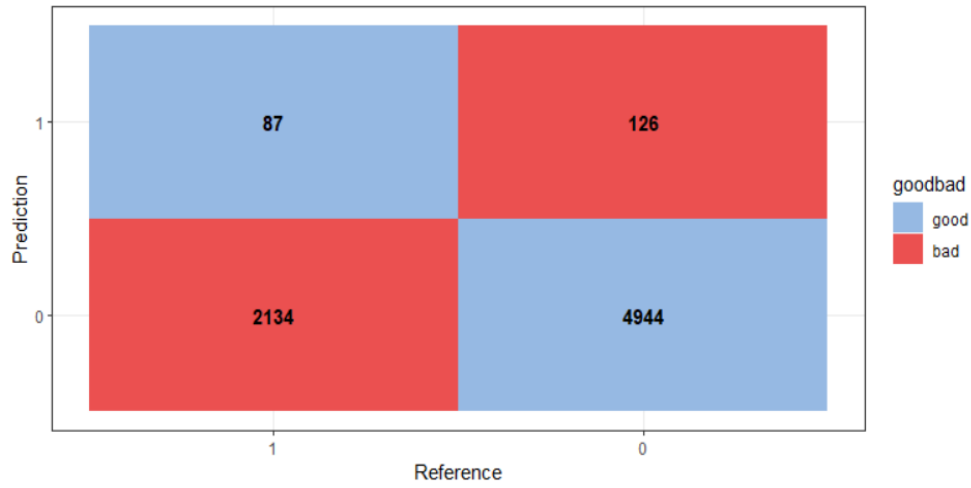

Figure 10: C5: Confusion Matrix (80:20)

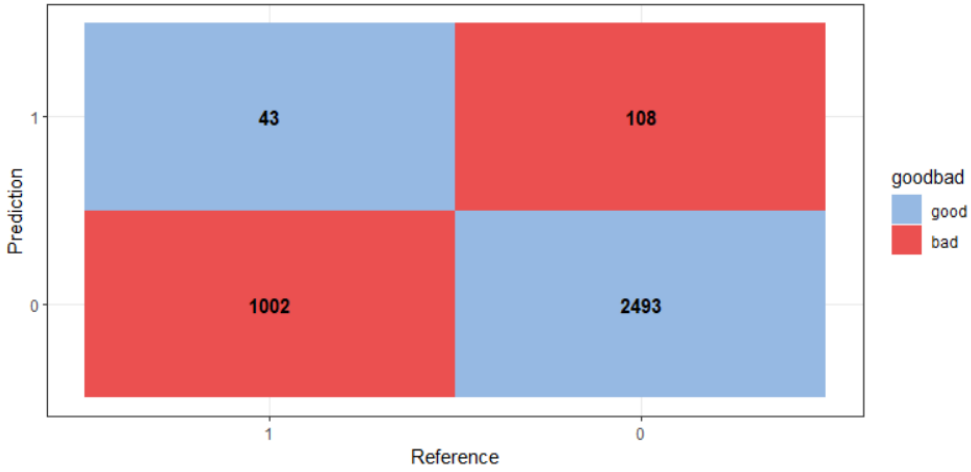

Figure 11: C5: Confusion Matrix (90:10)

Table 3: Performance metric for C5

Training: Test data	Accuracy	Sensitivity	Specificity
70:30	69.6%	98.30%	2.05%
80:20	69%	97.51%	3.92%
90:10	69.56%	95.84%	4.11%

2. Random Forest:

When using the Random Forest method, determining the relative relevance of each feature to the prediction is a simple and straightforward process. The characteristics to be eliminated from the model can be determined by looking at the relevance of the features, to confirm the ones that have little or no contribution to the prediction process. This is significant as one of the fundamental rules in machine learning is that the more features a model possesses, the greater the likelihood that it will suffer from overfitting, and vice versa.

The highest accuracy for this model is 71.34% for 90:10 (training: test) split with a sensitivity of 100%. The confusion matrix for the best model in Random forest depicts, 1045 customers are predicted to be good customers, however, they are bad customers and would default on payment. However, in this model, we are not missing out on any good customers.

70:30

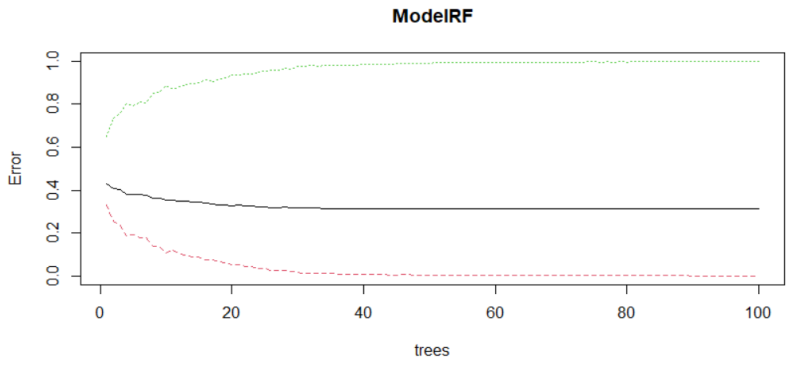

***Figure 12*: Random forest OOB graph (70:30)**

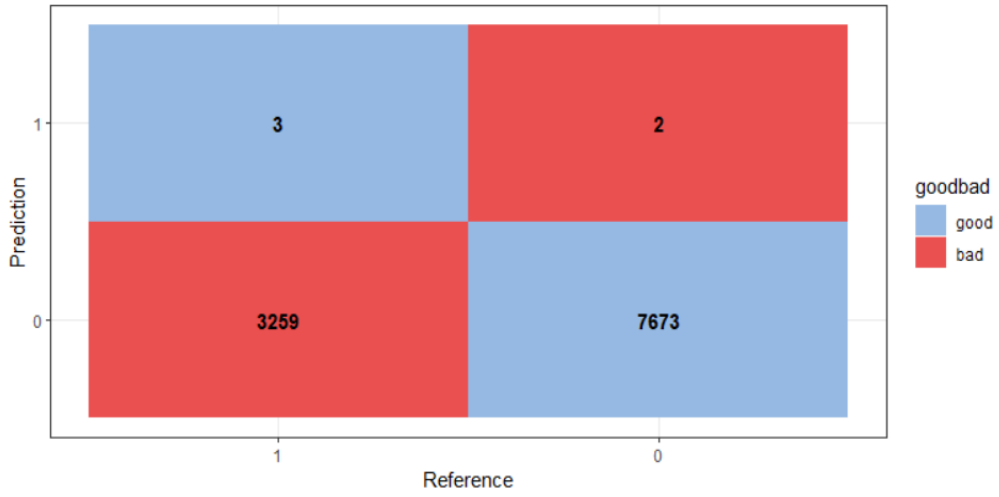

Figure 13: Random Forest: Confusion Matrix (70:30)

80:20

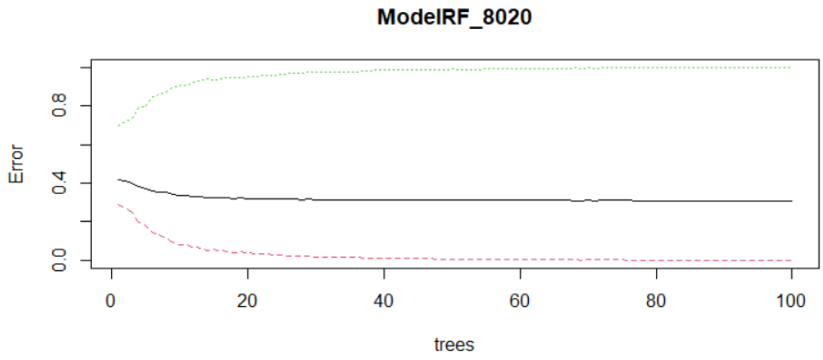

Figure 14: Random forest OOB graph (80:20)

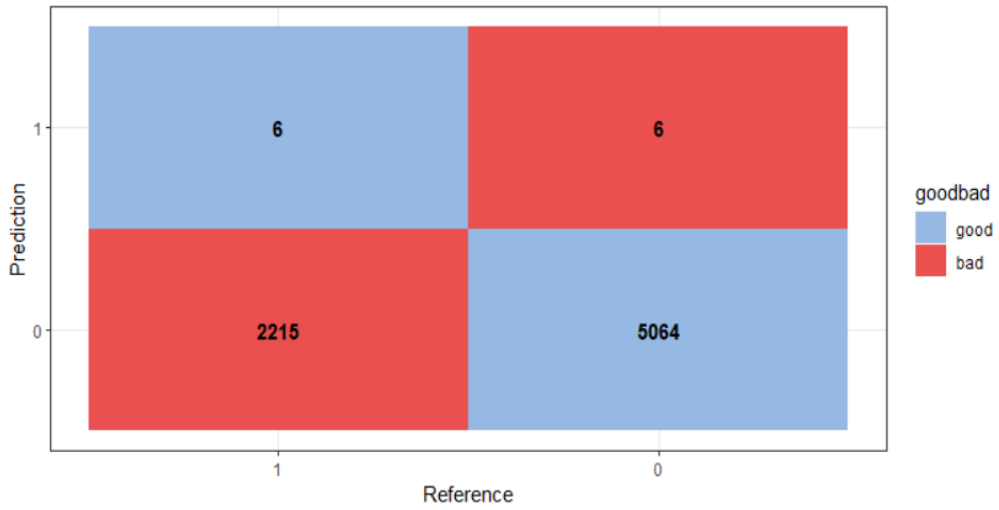

Figure 15: Random Forest: Confusion Matrix (80:20)

90:10

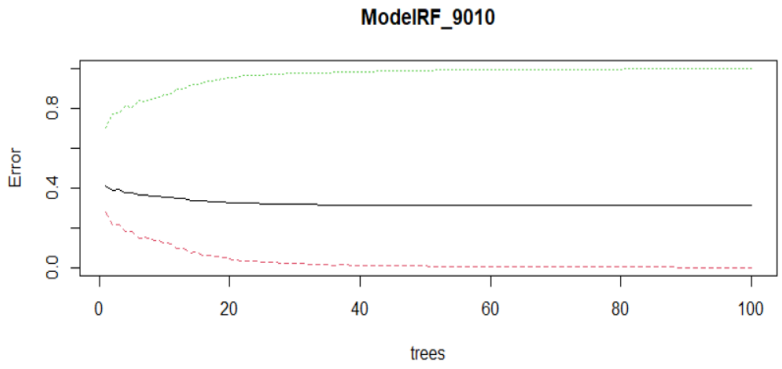

Figure 16: Random forest OOB graph

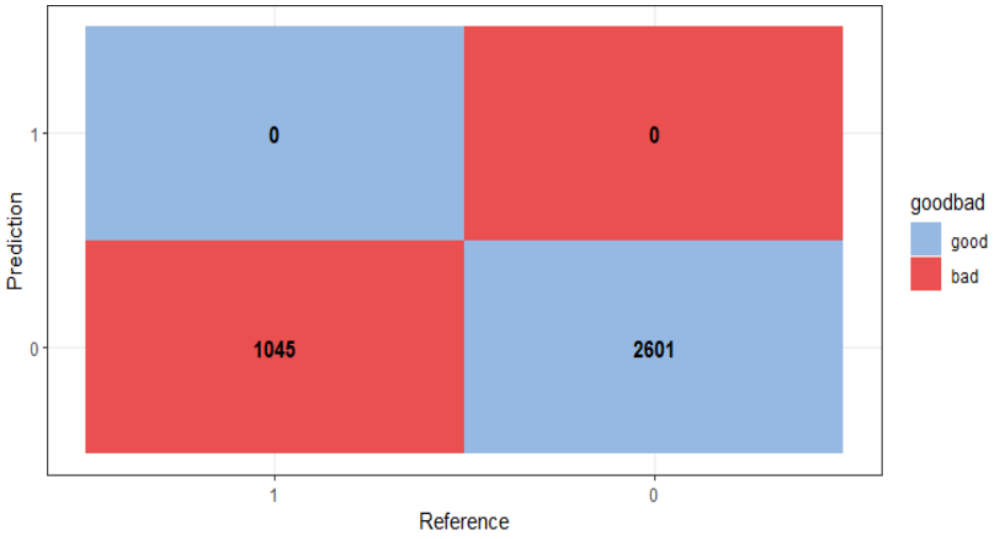

Figure 17: Random Forest: Confusion Matrix (90:10)

Table 4: Performance metric for Random Forest

Training: Test data	Accuracy	Sensitivity	Specificity
70:30	70.16%	99.96%	0.03%
80:20	69.4%	99.74%	0.13%
90:10	71.34%	100%	0%

3. Naïve Bayes

The Naive Bayes algorithm is one of the most common machine learning methods that makes use of Bayesian methods. It derives its name from its characteristic of making some "naive" assumptions about the data. Specifically, it treats all the features in the dataset independently and equally based on importance. Although not all features are crucial when applying in the real world, Naïve Bayes performs moderately well in most of the cases where the assumptions are breached. Naïve Bayes is however usually a first choice for consideration for classification learning tasks due to its adaptability and accuracy across various scenarios.

The highest accuracy for this model is 71.39% for 90:10 (training: test) ratio with a very high sensitivity. The confusion matrix for the best model in Naive Bayes depicts that 1039 customers are predicted to be good customers; however, they are bad customers and would default on payment. Here, we are missing out only on 4 good customers which the model predicted as bad.

70:30

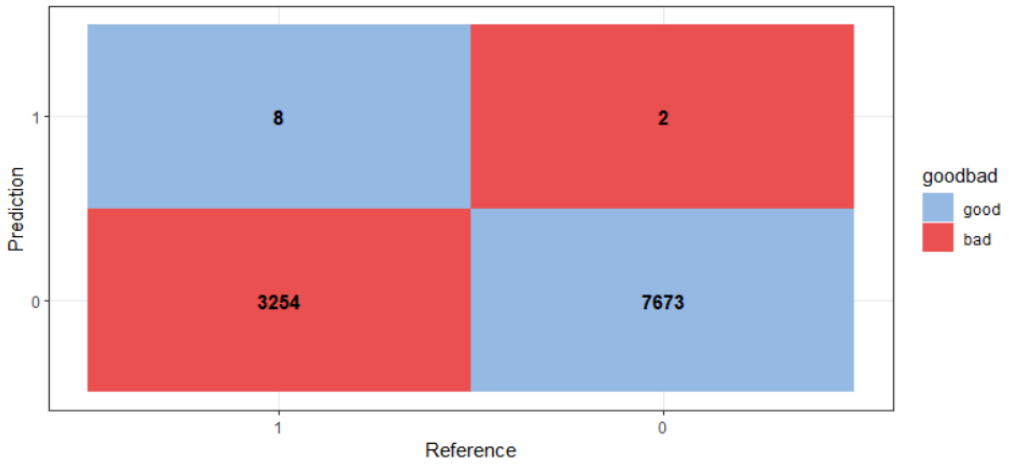

Figure 18: Naive Bayes: Confusion Matrix (70:30)

80:20

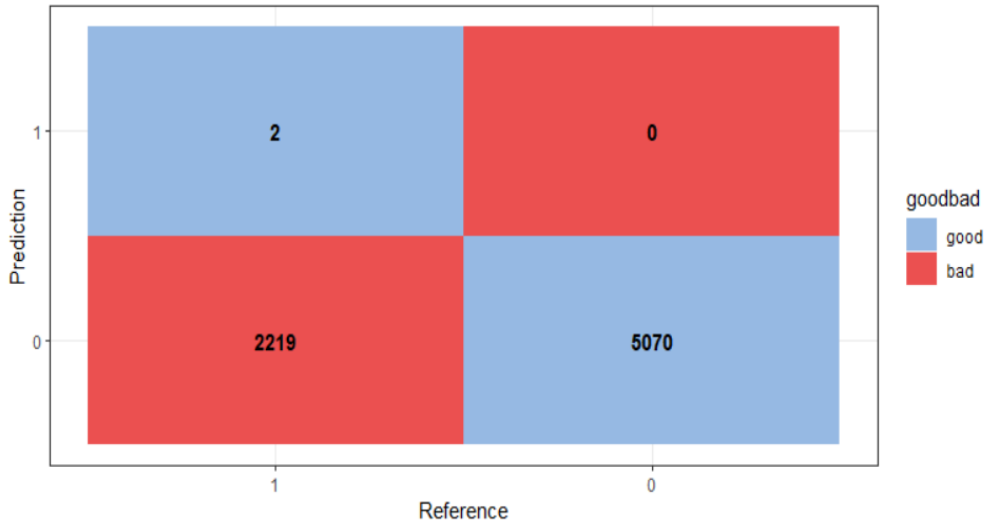

Figure 19: Naive Bayes: Confusion Matrix (80:20)

90:10

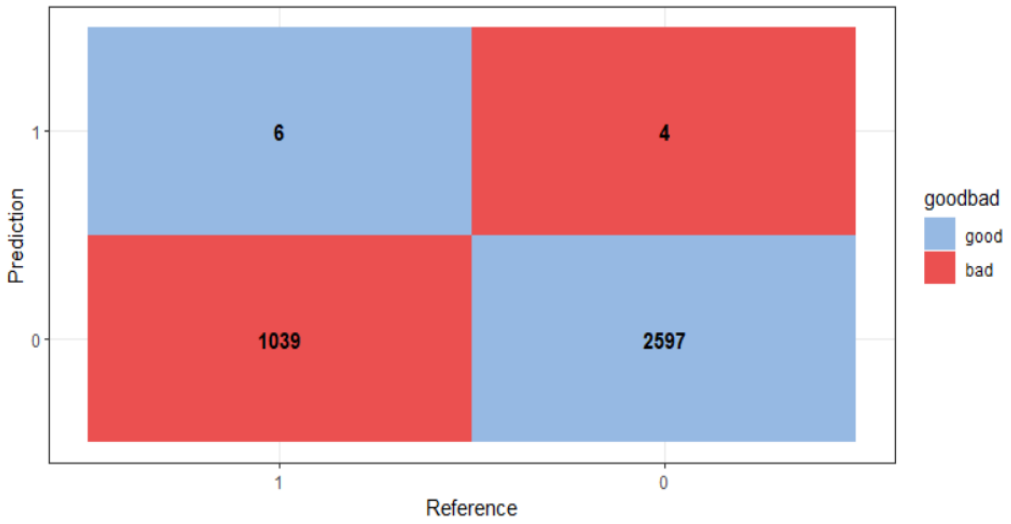

Figure 20: Naive Bayes: Confusion Matrix (90:10)

Table 5: Performance metric for Naïve Bayes

Training: Test data	Accuracy	Sensitivity	Specificity
70:30	70.23%	99.97%	0.24%
80:20	69.57%	100%	0%
90:10	71.39%	99.84%	0.57%

4. Neural Network

Artificial Neural Networks (ANN) mimic the way the brain creates a processor by using a system of interlinked cells providing solutions to machine learning problems. Artificial Neural Network models are made up of a layer that provides input, a layer that is hidden, a results layer and a function that activates the model. ANNs are usually recommended for learning tasks such as classification, estimation of functions and pattern recognitions. Despite their slowness and processing requirements, they are a popular option for credit risk and loan default prediction applications. The algorithm's main advantages are its ability to represent more complex patterns than other algorithms and its flexibility to numeric and classification prediction tasks.

The accuracy for this model is 30.6% on the 80:20 (training: test) ratio, implying that the model underperforms on the chosen dataset. The confusion matrix for the best model in the Neural network depicts 7 customers are predicted to be good customers, however, they are bad customers and would default on payment. Also, we are missing out on 5053 good customers which the model predicted as bad.

70:30

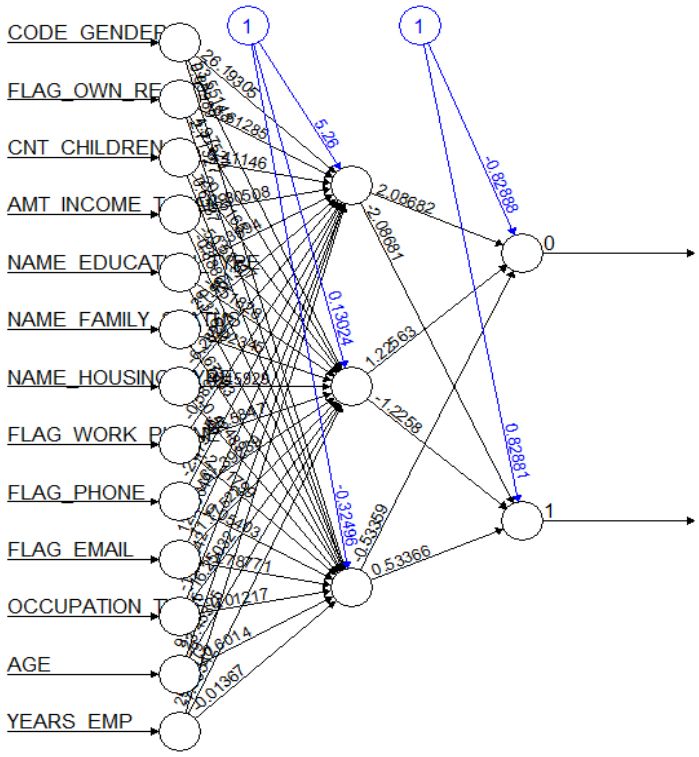

Figure 21: Neural Network (70:30)

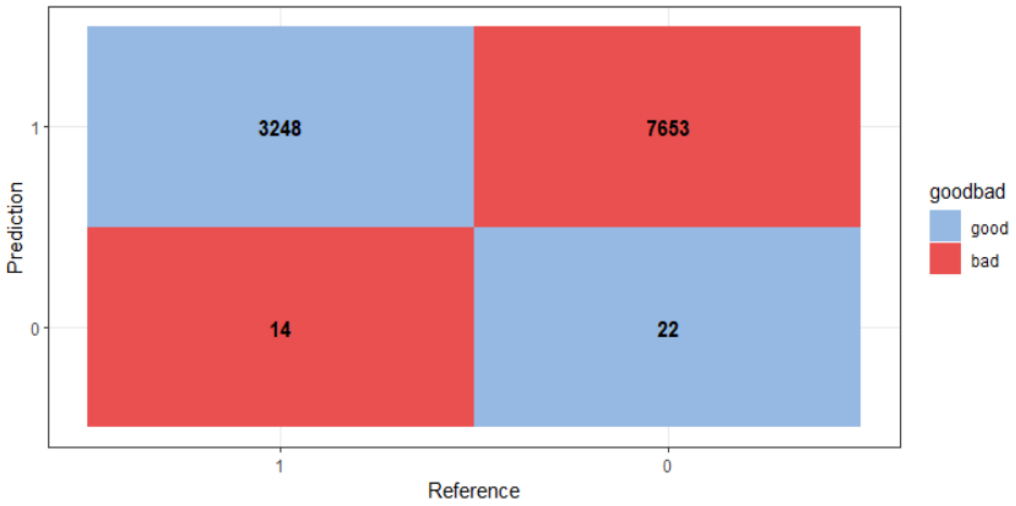

Figure 22: Neural Network: Confusion Matrix (70:30)

80:20

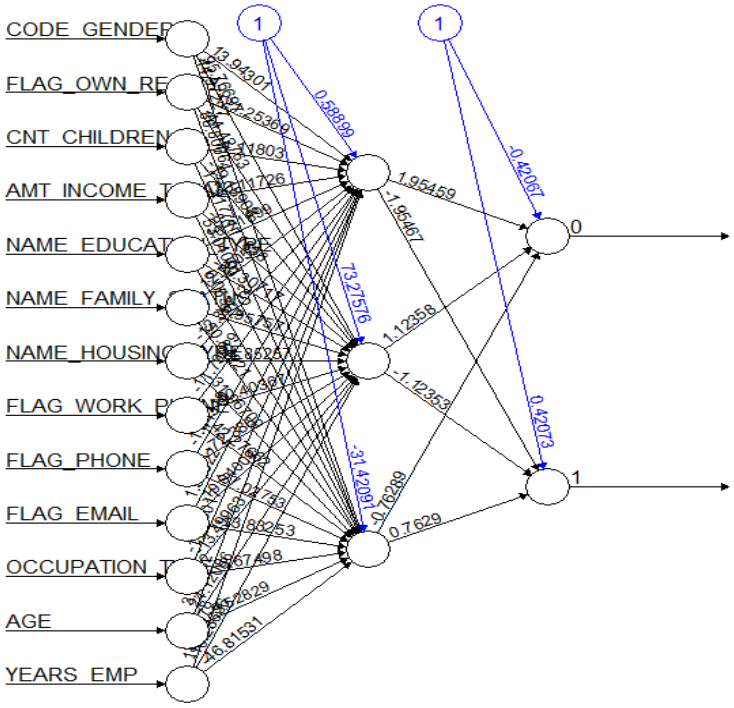

Figure 23: Neural Network (80:20)

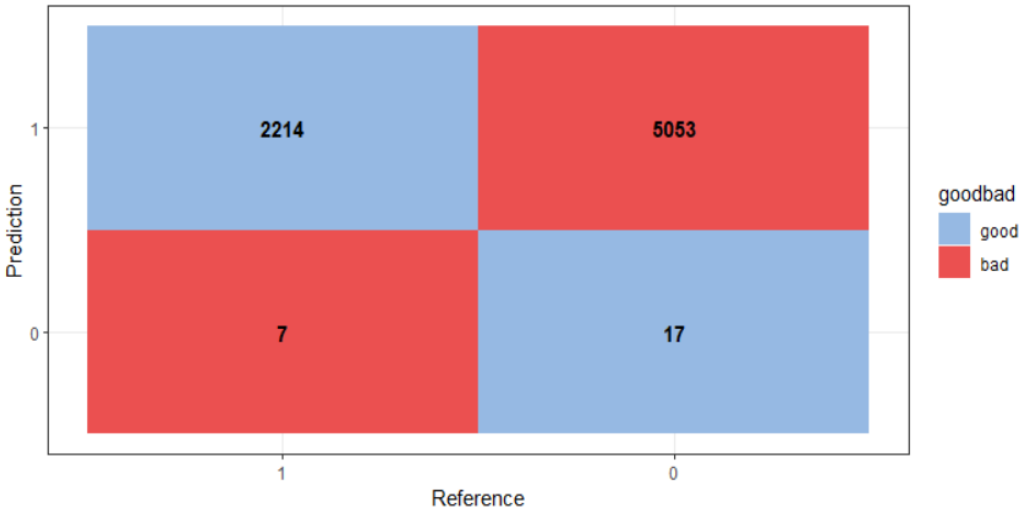

Figure 24: Neural Network: Confusion Matrix (80:20)

90:10

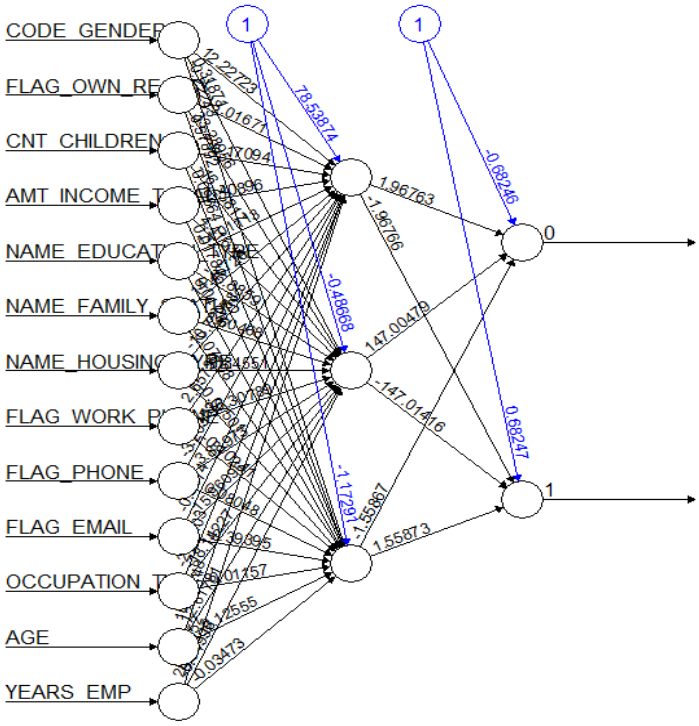

Figure 25: Neural Network (90:10)

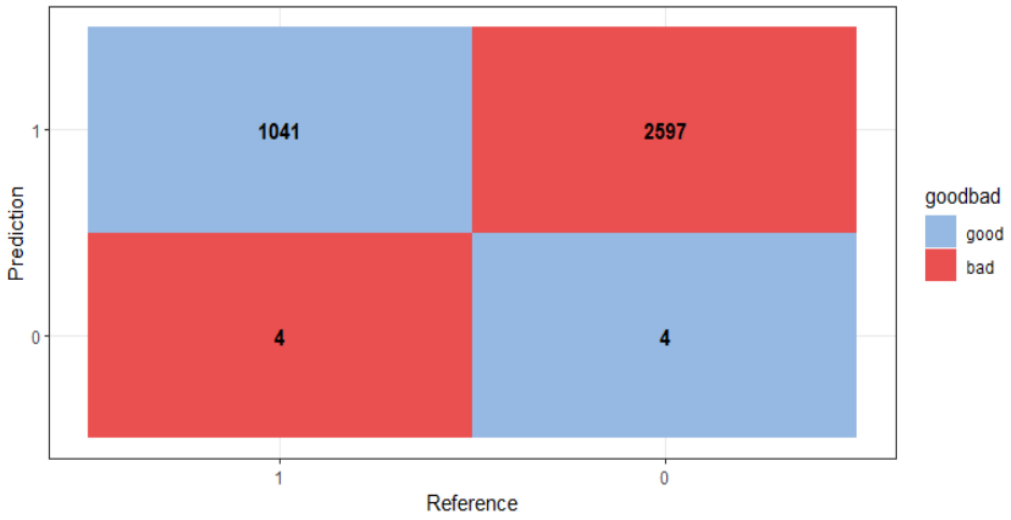

Figure 26: Neural Network: Confusion Matrix (90:10)

Table 6: Performance metric for Neural Network

Training: Test data	Accuracy	Sensitivity	Specificity
70:30	29.9%	0.28%	99.57%
80:20	30.6%	0.33%	99.68%
90:10	28.66%	0.15%	99.61%

MODEL EVALUATION

Table 7: Models comparison

Models	Accuracy	Sensitivity	Specificity
C5 Decision tree			
70:30	69.6%	98.3%	2.05%
80:20	69%	97.51%	3.92%
90:10	69.56%	95.84%	4.11%
Random Forest			
70:30	70.16%	99.96%	0.03%
80:20	69.4%	99.74%	0.13%
90:10	71.34%	100%	0%
Naïve Bayes			
70:30	70.23%	99.97%	0.24%
80:20	69.57%	100%	0%
90:10	71.39%	99.84%	0.57%
Neural Network			
70:30	29.9%	0.28%	99.57%
80:20	30.6%	0.33%	99.68%
90:10	28.66%	0.15%	99.61%

Table 7 compares the results obtained from each of the models created for predicting the outcome of the stated problem. The model performance is measured based on the metrics such as Accuracy, Sensitivity and Specificity. The effectiveness of the model and these parameters can also be calculated using the Confusion matrix.

It is a combination of predicted and actual values:

True Positives - This is a creditworthy customer whose application should be approved

True Negatives - This not a creditworthy customer and the application should be rejected

False Positives - The model predicted it to be a creditworthy customer whose application should be approved but it is actually not a creditworthy customer

False Negatives - The model predicted it to be a customer who is not creditworthy and whose application should be rejected but it is actually a creditworthy customer

Models Performance Improvement

The precise evaluation of consumer credit risk is of the utmost significance to lending financial institutions. We have attempted to enhance the performance of the models and the same is explained below:

Random Forest - Although the default number of trees in the Random forest algorithm is 500, these many trees are not always required for optimal results. Therefore, we have tried running the model with different numbers of trees to find the optimum number of trees. It is evident from figures 12, 14 & 16 that the Out of Bag error (OOB) is getting stabilised after 30 trees. Therefore 100 trees have been used for this model.

Naïve Bayes - We have run the model with and without Laplacian smoothing. However, there is no significant difference in the performance of the model. The optimal result can be obtained, with 90% training data.

Neural Network - Increasing the number of hidden layers in the model from one to three has been carried out in an attempt to improve the model's performance, however the model's accuracy and other metric parameters remain unchanged. However, we tried increasing the hidden layers, but could not achieve the results due to the hardware constraint.

It is evident that, in terms of Accuracy, the two best performing models are Naïve Bayes and Random Forest with a training test ratio of 90:10 for both. However, the model performance cannot only be assessed in terms of accuracy, other metrics should be considered as well. Random Forest has a sensitivity of 100% and Specificity of 0%. Also, it has predicted all customers as good and none of them as bad. As this is evidence of possible overfitting, this model should not be selected as a good model. In view of all the above, Naïve Bayes is selected as the best model to achieve the objective for this dataset.

MODEL ANALYSIS

Table 8: Model Analysis

PARAMETERS (Model Accuracy / Revenue / Profit / Loss / Interest)	MODELS - BUSINESS INSIGHT											
	C5			Random Forest			Naive Bayes			Neural Network		
	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10
Accuracy	69.60%	69.00%	69.56%	70.16%	69.40%	71.34%	70.23%	69.57%	71.39%	29.90%	30.60%	28.66%
Sensitivity	98.30%	97.51%	95.84%	99.96%	99.74%	100.00%	99.97%	100.00%	99.84%	0.28%	0.33%	0.15%
Specificity	2.05%	3.92%	4.11%	0.03%	0.13%	0.00%	0.24%	0.00%	0.57%	99.57%	99.68%	99.61%
Average credit Limit (Each Customer)*	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00	£5,000.00
TP- No.of Correctly Approved Cards / Good Customers	7545	4944	2493	7673	5064	2601	7673	5070	2597	22	17	4
Total Credit Amount (£)	37725000	24720000	12465000	38365000	25320000	13005000	38365000	25350000	12985000	110000	85000	20000
Credit Card Annual Charges / Processing Fee*	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%	5%
Revenue from Correctly Aprroved / Good Customers(£)	**1886250**	**1236000**	**623250**	**1918250**	**1266000**	**650250**	**1918250**	**1267500**	**649250**	**5500**	**4250**	**1000**
FP - No. of Inccorectly Predicted Cards Approved - A	3195	2134	1002	3259	2215	1045	3254	2219	1039	14	7	4
Probability of Default (22%) In wrongly Predicted *	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%	22%
Number of Customer Defaulting- B	703	469	220	717	487	230	716	488	229	3	2	1
No of Customer Delaying the payment (A-B)	2492	1665	782	2542	1728	815	2538	1731	810	11	5	3
Loss from Incorrrectly Predicted (£)	**3514500**	**2347400**	**1102200**	**3584900**	**2436500**	**1149500**	**3579400**	**2440900**	**1142900**	**15400**	**7700**	**4400**
Number of Customer (Test data in each Model)	10937	7291	3646	10937	7291	3646	10937	7291	3646	10937	7291	3646
Interst Rate on Delay Payemnts *	21%	21%	21%	21%	21%	21%	21%	21%	21%	21%	21%	21%
Revenue(£)from Delayed Customer (Bad Credit Score)	2616705	1747746	820638	2669121	1814085	855855	2665026	1817361	850941	11466	5733	3276
PROFIT / LOSS (£)	988455	636346	341688	1002471	643585	356605	1003876	643961	357291	1566	2283	-124
Profit / Loss (Per Applicant In Future)	£90.38	£87.28	£93.72	£91.66	£88.27	**£97.81**	£91.79	£88.32	**£98.00**	£0.14	£0.31	-£0.03

*Note : All Assumptions are as per UK market standard

Table 9: Chosen Model Cost Matrix

PARAMETERS (Model Accuracy / Revenue / Profit / Loss / Intrest)	Naive Bayes (90:10)
Accuracy	71.39%
Sensitivity	99.84%
Specificity	0.57%
Average credit Limit (Each Customer)*	£5,000.00
TP- No.of Correctly Approved Cards / Good Customers	2597
Total Credit Amount (£)	12985000
Credit Card Annual Charges / Processing Fee*	5%
Revenue from Correctly Aprroved / Good Customers(£)	**649250**
FP - No. of Inccorectly Predicted Cards Approved - A	1039
Probability of Default (22%) In wrongly Predicted *	22%
Number of Customer Defaulting- B	229
No of Customer Delaying the payment (A-B)	810
Loss from Incorrrectly Predicted (£)	**1142900**
Number of Customer (Test data in each Model)	3646
Interst Rate on Delay Payemnts *	21%
Revenue(£)from Delayed Customer (Bad Credit Score)	850941
PROFIT / LOSS (£)	357291
Profit / Loss (Per Applicant In Future)	**£98.00**

*Note : All Assumptions are as per UK market standard

While accuracy, sensitivity and specificity are standard measures used to compare and measure the accuracy of machine learning algorithms, these outputs on their own are not sufficient to predict the impact on the business and aid decision making. To support the business in deciding on the model to select, a profit and loss analysis was carried out using the outputs from the training and test ratios of each model. An average credit limit usage of £5,000 and an annual charge/processing fee of 5% of the average credit limit, are some of the assumptions made in the computation.

The customers which are incorrectly predicted as good customers who may either delay the payments or default on the entire loans taken on the credit card, are referred as bad customers. As per the market standards, 22% of these bad customers are most likely to default on payments resulting in financial loss to the company. The balance 78% are assumed to be delaying payments and therefore generating revenue by paying the interest on the delayed payment. An annual interest of 21% (UK market standards) is charged on customers for delayed payments. These 78% bad customers who delay their payments turn out to be profitable customers for the institutions.

The Table in figure demonstrates the profit/loss calculations for each model. The same calculation for the chosen model has been depicted in the table 8. Based on the calculations and assumptions, Naive Bayes algorithm with a train and test ratio of 90:10 is generating a profit of £98 per applicant which is the highest among all models and thus is being suggested as the best model for credit card approval prediction.

CONCLUSION

To conclude, the performance metrics of machine learning algorithms such as Accuracy, Sensitivity and Specificity are not the only parameters to be considered to choose the best model for future prediction as each industry has their own Key Performance Indicators. In financial organisations, where credit cards are issued, revenue is generated from a combination of interest fees, annual charges and processing fee payments from customers. Therefore, an effective business strategy in this industry may require pursuing the most profitable route that comes with the lowest risk implications, which involves achieving the right balance between good and bad applications.

The profit and loss estimates are based on a number of assumptions, including an average credit limit, a specific default rate, the maximum credit limit the customer may default on, the maximum credit limit utilised, and generating interest owing to late repayments, among others. These computations may be improved by adding specifics from these areas.

References:

1. Bernardo, Dario, et al. "A Genetic Type-2 Fuzzy Logic Based System for the generation of Summarised Linguistic Predictive Models for Financial Applications." *Soft Computing*, vol. 17, no. 12, 13 Aug. 2013, pp. 2185–2201, 10.1007/s00500-013-1102-y. Accessed 15 May 2022.
2. Danenas, Paulius, and Gintautas Garsva. "Selection of Support Vector Machines Based Classifiers for Credit Risk Domain." *Expert Systems with Applications*, vol. 42, no. 6, Apr. 2015, pp. 3194–3204, 10.1016/j.eswa.2014.12.001. Accessed 24 Sept. 2020.
3. Kibria, Md. Golam, and Mehmet Sevkli. "Application of Deep Learning for Credit Card Approval: A Comparison with Two Machine Learning Techniques." *International Journal of Machine Learning and Computing*, vol. 11, no. 4, Aug. 2021, pp. 286–290, 10.18178/ijmlc.2021.11.4.1049. Accessed 4 Aug. 2021.
4. Lantz, Brett. *Machine Learning with R : Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*. Birmingham, Packt Publishing, 2015.
5. Peng, Yi, et al. "An Empirical Study of Classification Algorithm Evaluation for Financial Risk Prediction." *Applied Soft Computing*, vol. 11, no. 2, Mar. 2011, pp. 2906–2915, 10.1016/j.asoc.2010.11.028. Accessed 2 Oct. 2021.
6. Sugiyarto, Ipin, et al. "Performance Comparison of Data Mining Algorithm to Predict Approval of Credit Card." *SinkrOn*, vol. 4, no. 1, 5 Oct. 2019, p. 149, 10.33395/sinkron.v4i1.10181. Accessed 5 Nov. 2019.
7. Yousofi Tezerjan, Mostafa, et al. "ARF: A Hybrid Model for Credit Scoring in Complex Systems." *Expert Systems with Applications*, vol. 185, Dec. 2021, p. 115634, 10.1016/j.eswa.2021.115634. Accessed 1 Dec. 2021.