

Final_Submission

by Chisom Nwankwo

Submission date: 05-May-2023 05:25PM (UTC+0100)

Submission ID: 204970409

File name: 45129_Chisom_Nwankwo_Final_Submission_493224_1163485289.pdf (667.91K)

Word count: 4637

Character count: 28550



Prifysgol Cymru
Y Drindod Dewi Sant
University of Wales
Trinity Saint David

School of Applied Computing

ACCA7015

Report on the 'Default of Credit Cards' dataset

By

Chisom Vanessa Nwankwo

ABSTRACT

Credit card default arises due to severe delinquency on credit card repayments. It is an issue that banks still struggle to tackle. This emphasises the need to put checks and controls in place to reduce the number of credit card defaulters. With the default on credit cards dataset downloaded from the UCI machine learning repository, this study attempts to understand the factors contributing to a default in credit repayment and identify the key factors that predict a default. The dataset is based on the data obtained from a Taiwanese bank on their credit card clients. Two supervised learning models capable of predicting default were developed to determine these key factors contributing to the default. Data analysis tools like Orange and Microsoft Excel were employed in this study. To prepare the training data, it was pre-processed using Orange's widgets by removing outliers, one-hot encoding the categorical values, normalisation, etc. The task carried out was a binary classification of imbalanced datasets, so oversampling of the minority class (or undersampling of the majority class) was necessary. After pre-processing, the models were trained and used to make predictions. Reducing the number of false negatives in the models is important, so this factor was considered when comparing the models. The models developed were logistic regression, support vector machines, random forest, decision tree and neural network models. The results show that the random forest classifier performs better than the other supervised machine learning models. However, it is still not reliable enough for real-world situations. Other ensemble machine learning algorithms like gradient boosting algorithms will be considered for future works as they generally perform better than random forest models on imbalanced datasets.

TABLE OF CONTENTS

CHAPTER ONE: INTRODUCTION	5
1.1. Background Information	5
CHAPTER TWO: DATASET INFORMATION	7
2.1. Origin of the Dataset	7
2.2. Content of the Dataset	7
2.3. Suitability of the Dataset for Analysis.....	8
2.4. Objectives of Analysis.....	8
CHAPTER THREE: METHODOLOGY	9
3.1. Tools Used in the Study.....	9
3.2. Data Analysis.....	9
3.2.1. Primary Preprocessing	9
3.2.3. Classification.....	18
CHAPTER FOUR: RESULTS AND DISCUSSION.....	22
4.1. Reflection	22
4.1.1. Performance Evaluation.....	22
4.1.2. Confusion Matrix	23
CHAPTER FIVE: CONCLUSION.....	25
5.1. Summary and Recommendations	25
REFERENCES	26

TABLE OF FIGURES

Figure 3.1: Target Class	10
Figure 3.2: Frequency of Males and Females	11
Figure 3.3: Age distribution	12
Figure 3.4: Credit limit and probability of default.....	12
Figure 3.5: Credit limit of Defaulters	13
Figure 3.6: Credit Limit and Age of clients	14
Figure 3.7: Credit above 500000	15
Figure 3.8: Credit below 500000	16
Figure 3.9: Education and default.....	17
Figure 3.10: Marital status and default	18
Figure 3.11: Scree diagram of the PCA	19
Figure 3.12: Class balance	20
Figure 3.13: Evaluation Result	21
Figure 4.1: Logistic Regression	23
Figure 4.2: Support Vector Machine	24
Figure 4.3: Random Forest	24
Figure 4.4: Decision Tree.....	24
Figure 4.5: Neural Network	24

LIST OF TABLES

Table 4.1: Performance Scores.....	22
Table 4.2: Confusion matrix.....	23

CHAPTER ONE: INTRODUCTION

1.1. Background Information

Credit cards are granted to customers by banks and allow customers to borrow money to pay for goods and services. A stipulated date is also given to the customer at which they are required to repay the money borrowed and used. Customers can also be allowed an overdraft, which allows them to spend over their credit limit. Banks sometimes allow it when the credit available to a customer reaches zero. In this case overdraft protection is granted, preventing the customer from embarrassment due to 'insufficient funds' when purchasing items [1]. However, it increases the credit repayment amount of the customer as they will be required to pay for overdrafts. A customer's credit card becomes delinquent due to overspending of credit allowed to that customer and failing to repay at the stated date. The delay in payment also causes banks to charge fees to the account, making the debt even worse. This is known as credit card delinquency. As the debt continues to accumulate, it becomes a credit card default. Credit card default is due to continued breach of credit repayment agreement between the customer and the bank. This reduces their credit scores and affects their ability to borrow money later. In the past, some banks have been known to over-issue credit cards to customers who are unqualified to increase their market share and profit. These customers then end up overspending and become delinquent [2]. The essence of a credit default prediction is to aid financial institutions in determining whether or not to issue credit cards to a customer.

Generally, there are fewer credit card defaulters than non-defaulters [3]. The classification task into defaulters and non-defaulters is called a binary classification task. The positive class is the default class, and the negative class is the non-default class. The study will use the 'Default on credit cards' dataset from the UCI machine learning repository to obtain more insight as to why default occurs, which could help banks put checks and controls in place to prevent the default before it happens.

Previous studies on this dataset show that ensemble machine learning models perform better than traditional machine learning models. A study by [4] compared data mining methods such as Logistic Regression, SVM, neural network and LightGBM (light gradient boosting machine). It showed that the LightGBM model performed better than the others and is very useful with big data. Experiments by [2] presented artificial neural networks as the best classifier, compared to k-nearest neighbours, naïve Bayesian, decision trees, discriminant analysis and logistic regression models. In the paper by [5], the gradient-boosted tree algorithm performed better than other traditional classifiers. The study also showed that oversampling produced better results than undersampling. In forecasting credit card default [6], the model proposed for classification and prediction is the incremental random forest algorithm which

performs relatively better than other models. Also, a review of 76 works of literature by [7] revealed that deep learning and ensemble models performed better than single machine learning and statistical models. The ensemble models comprised Random Forests and boosting algorithms.

CHAPTER TWO: DATASET INFORMATION

2.1. Origin of the Dataset

The dataset was collected by Yeh, I-Cheng, and Lien, Che-hui in 2009 and analysed in their paper [2]. It was downloaded from the UCI machine learning repository. The dataset contains information about 30,000 credit card customers of a Taiwanese bank from April to September 2005. It consists of the credit limit for each customer, history of repayment of credit balance, bill statements for each month, repayment status for each month, demographic factors such as age, marital status, etc., and the default on their credit card payment. The study compared the predictive accuracies of the probabilities of credit card default among the bank's customers using various classification techniques. The real probability of default is unknown. The researchers felt that it was more important to determine the technique which produced higher performance scores than others rather than simply classifying customers with a single technique as credible or not credible. It seems a plausible theory from the perspective of risk management in banks, as they would be more interested in a model that predicts defaults with higher accuracies to prevent defaults.

2.2. Content of the Dataset

Table 2.1 describes the attributes present in the dataset.

Attributes	Description	Data Type
ID	Each customer's ID number	Numeric
default_payment_next_month	Probability of default of customers	Boolean
LIMIT_BAL	Amount of credit given (in New Taiwan Dollars)	Numeric
SEX	Gender	Nominal
EDUCATION	The education level of each customer.	Ordinal
AGE	The age of the customers.	Numeric
MARRIAGE	Marital Status of each customer	Nominal
PAY_0 to PAY_6	The bill repayment status for the customers.	Ordinal
BILL_AMT1 to BILL_AMT6	Amount of bill statement for each customer, from April to September (in reverse order)	Numeric
PAY_AMT1 to PAY_AMT6	Amount of previous payments made by each customer from April to September (in reverse order)	Numeric

Table 2.1: Dataset attributes

In this study, 23 attributes (excluding the ID column) were used as the explanatory variables or features, while the last column, 'default_payment_next_month', represents the response variable or label. The values for each attribute include:

- a. ID column with unique values from 1 to 30,000 to identify each customer.
- b. default_payment_next_month: These values are based on the experiment carried out by the researchers. 1 represents Yes, 0 represents No.
- c. Limit Balance: The maximum and minimum values are 1,000,000 and 10,000.
- d. Sex: Male = 1, Female = 2.
- e. Education: Graduate degree = 1, University degree = 2, High school diploma = 3, Others = 4.
- f. Age: The maximum and minimum ages of the customers are 79 and 21.
- g. Marriage: Married = 1, Single = 2, Others = 3.
- h. PAY_0 to PAY_6 (x6 to x11):
- i. -1 = Duly pay, 1 = Delay in payment for one month, 2 = Delay in payment for two months, 3 = Delay in payment for three months, ... 9 = Delay in payment for nine months
- j. BILL_AMT1 to BILL_AMT6 (x12 to x17): The maximum bill amount is 1664089, and the minimum bill amount is -339603.
- k. PAY_AMT1 to PAY_AMT6 (x18 to x23): Maximum value is 1684259, while the minimum value is 0.

2.3. Suitability of the Dataset for Analysis

Most of the information is well labelled, so it is easy to understand the categories each data belongs to. It is large enough for analysis to give an insight into the area and does not contain missing values. While the data has some inconsistencies, it does not affect the results.

2.4. Objectives of Analysis

The dataset will be explored to:

- 1) Identify the key factors that determine the likelihood of a default on credit card repayment.
- 2) Predict default on credit card payment.
- 3) Compare performance among supervised learning models in predicting default.

CHAPTER THREE: METHODOLOGY

3.1. Tools Used in the Study

The data analysis was carried out using the Orange data mining tool and Microsoft Excel.

- a. Orange is an open-source data mining and machine learning tool that is useful for data visualisation and visual programming.
- b. Microsoft Excel is a spreadsheet software developed by Microsoft to enable the creation and modification of data. It also allows data visualisation.

3.2. Data Analysis

This section details the activities carried out to conduct the study.

3.2.1. Primary Pre-processing

Microsoft Excel was used to organise the data before loading it into Orange.

The activities include:

- a. Changing PAY_0 noticed in the data to PAY_1 as PAY_1 was previously not included. This activity prompted re-labelling some variable names, like PAY_5 to PAY_STAT_May, and BILL_AMT6 to BILL_AMT_Apr, to allow for readability since the attributes' names are in reverse order. For example, PAY_5 and BILL_AMT6 represent the payment status for the second month 'May' and the bill amount for the first month 'April', but PAY_1 represents the repayment status for the last month 'September'.
- b. The attribute EDUCATION contained 345 unknown values, such as 0, 5 and 6, which were merged with 4 for other levels of education besides from graduate, undergraduate and high school.
- c. The attribute MARRIAGE also contained 54 values for 0 which were unknown. Therefore, it was merged with 3 for other marital statuses besides from married and single (assumed to be divorced).
- d. The attributes PAY_STAT_N where N represents the month, contained unknown values such as -2 and 0, while the value '9' was missing. To fix this, the attributes were scaled between 0 and 9 by adding 1 to them, and merging values '0' and '-1' as 0 to represent Pay duly.
- e. I also dropped the ID column as it was unnecessary in the analysis.

Then the dataset was loaded into Orange using the 'File' widget.

3.2.2. Data Visualisation

Default class

Figure 3.1 shows the default class, the ratio of the non-defaulters to defaulters. The blue represents the non-default class, while the red represents the default class.

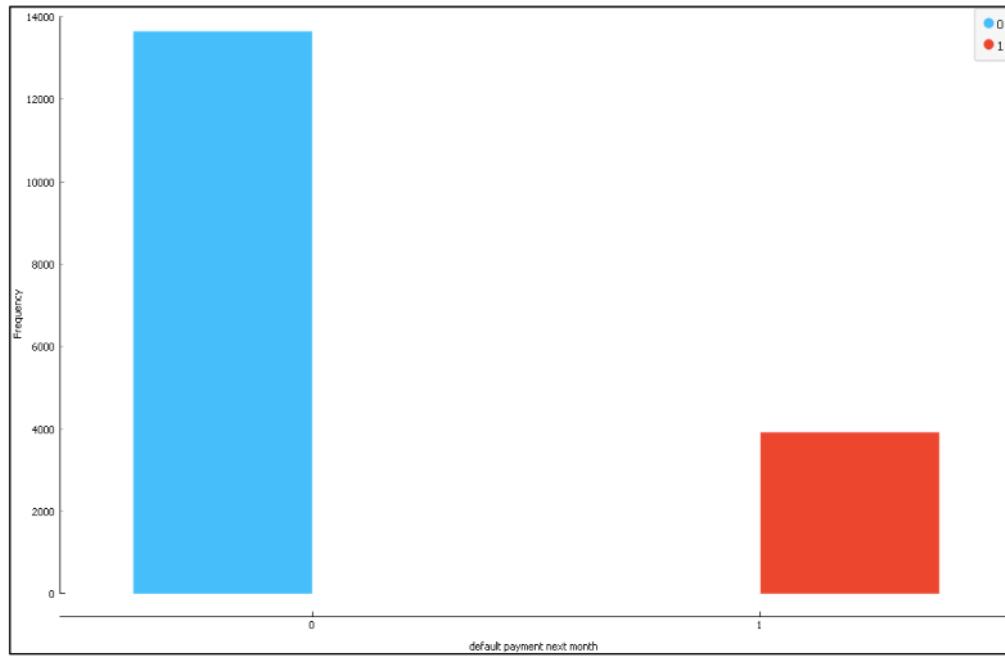


Figure 3.1: Target Class

The scatterplot widget was used to derive some informative projections shown below.

i. Gender Distribution

Figure 3.2 shows the distribution of males =1 (11888) and females =2 (18112) in the dataset. The diagram shows that there are more female credit card clients of the bank, with the female clients being over 60%. The ratio of male non-defaulters to defaulters is 75.83% to 27.13%, while that of females is 79.22% to 20.78%.

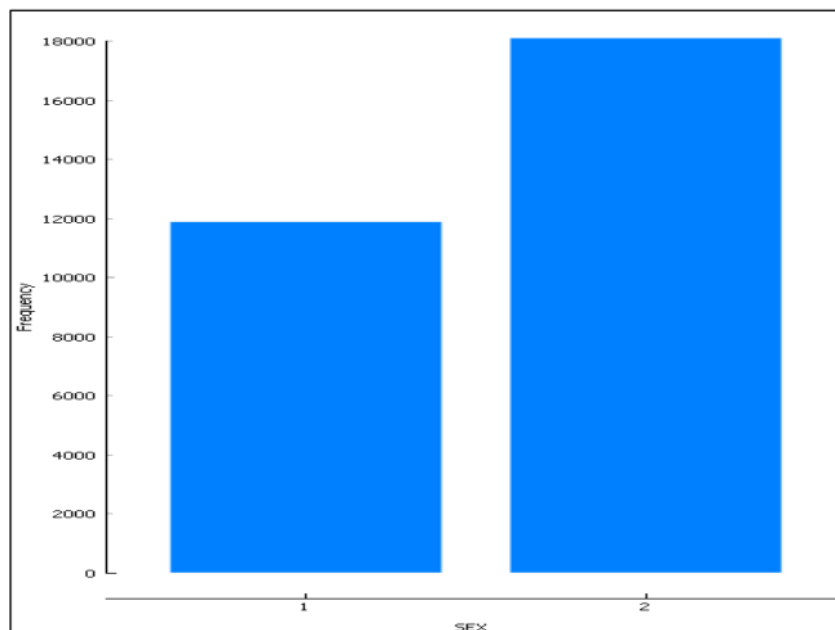


Figure 3.2: Frequency of Males and Females

ii. Age Distribution:

Figure 3.3 shows the frequency of males and females across the age groups: 20 to 25, 25 to 30, 30 to 35, ...75 to 80. Women are generally more than men, but in the age groups 55-60, 60-65, 65-70, and 75-80, the men are more, signifying that there are more older male than female credit card clients of the bank..

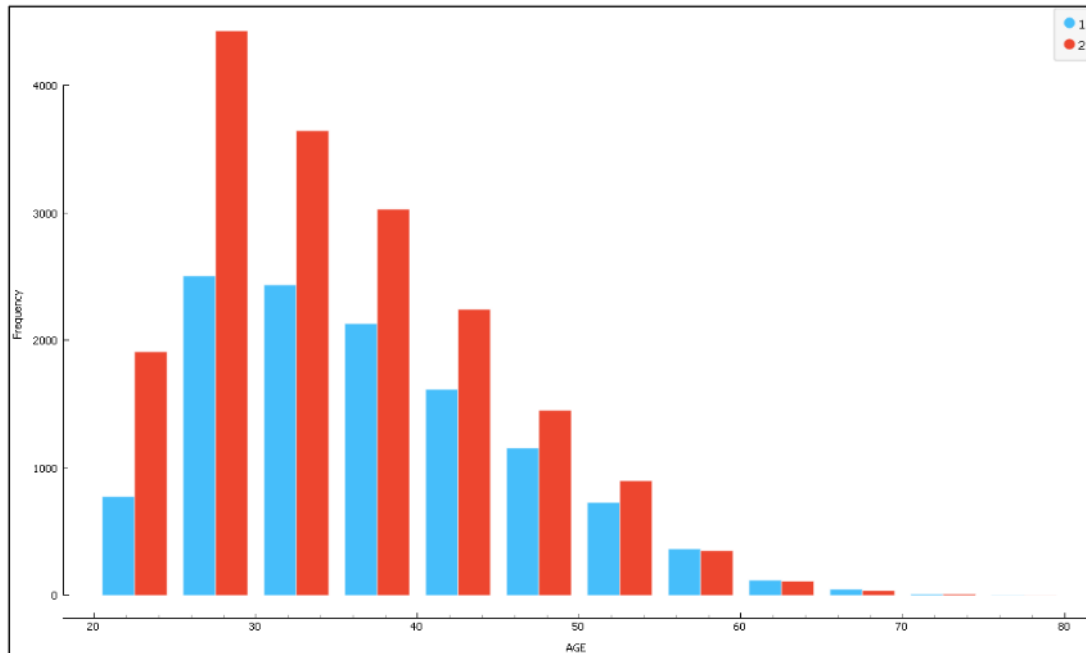


Figure 3.3: Age distribution

iii. Credit Limit:

Figures 3.4 and 3.5 explain the relationship between the credit limit and default.

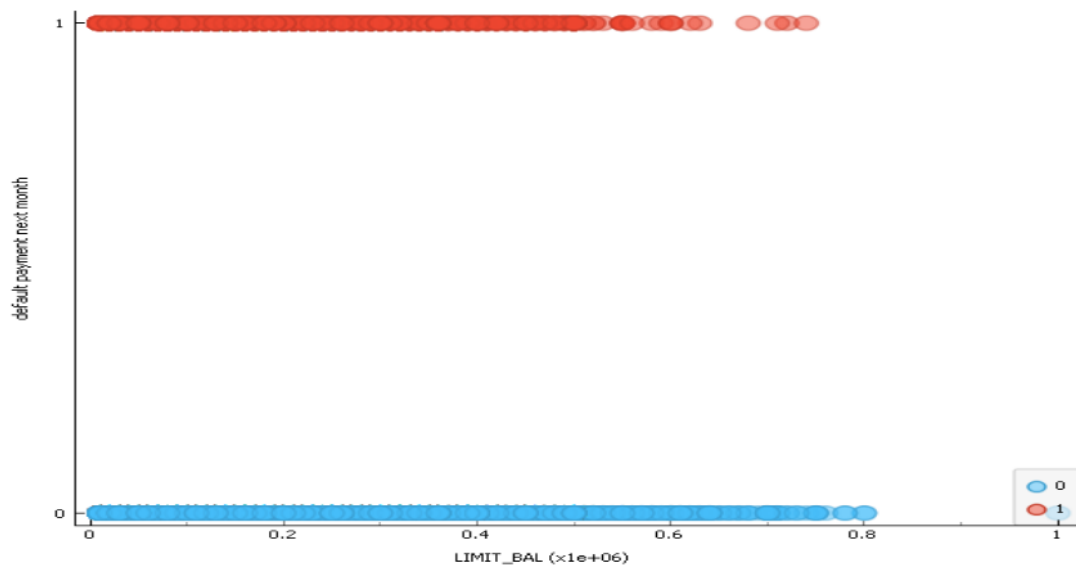


Figure 3.4: Credit limit and probability of default

A closer look at the defaulters shows that the higher the credit limit, the lower the default. Therefore, more people with lower credit are likely to default, especially with a credit limit between 10 000 (minimum credit) and 200,000. With a credit limit greater than or equal to 700,000, there are only two defaulters (0.03%) of the customers.

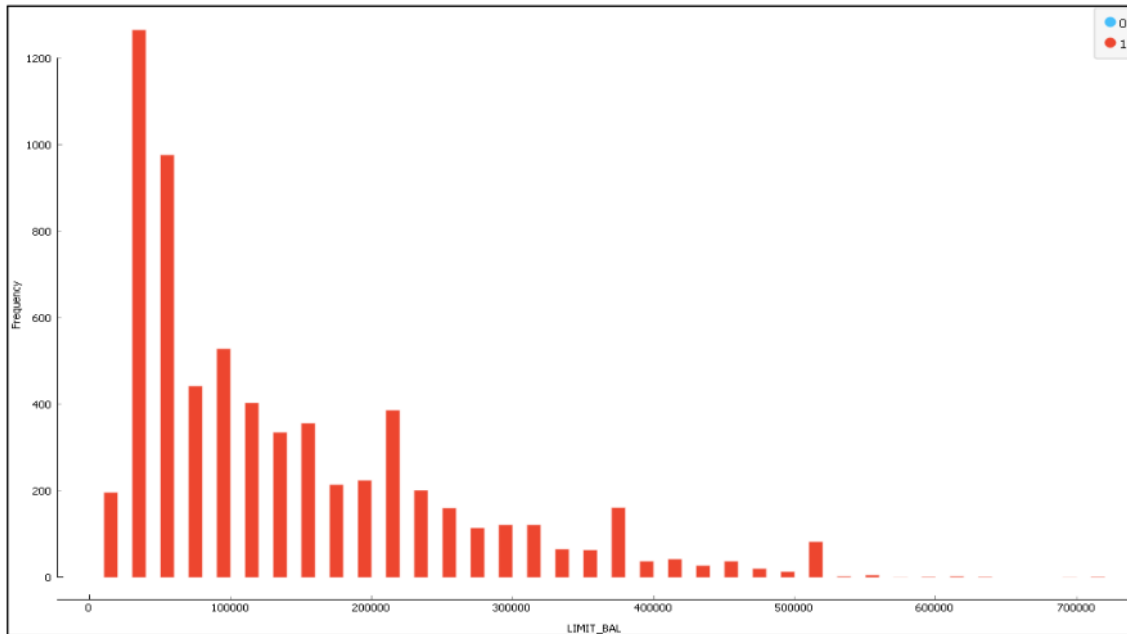


Figure 3.5: Credit limit of Defaulters

iv. Credit Limit and Age

In Figure 3.6, each customer's credit limit and ages are shown. Since there are more people with credit limits of less than 500,000 NT dollars and fewer defaulters above that amount, it will be used as a focal point. The minimum age of customers with credit above this amount is 28, while the maximum age is 67 years old. Also, only one person between the ages 40 and 50 was allowed a credit of 1 million NT dollars.

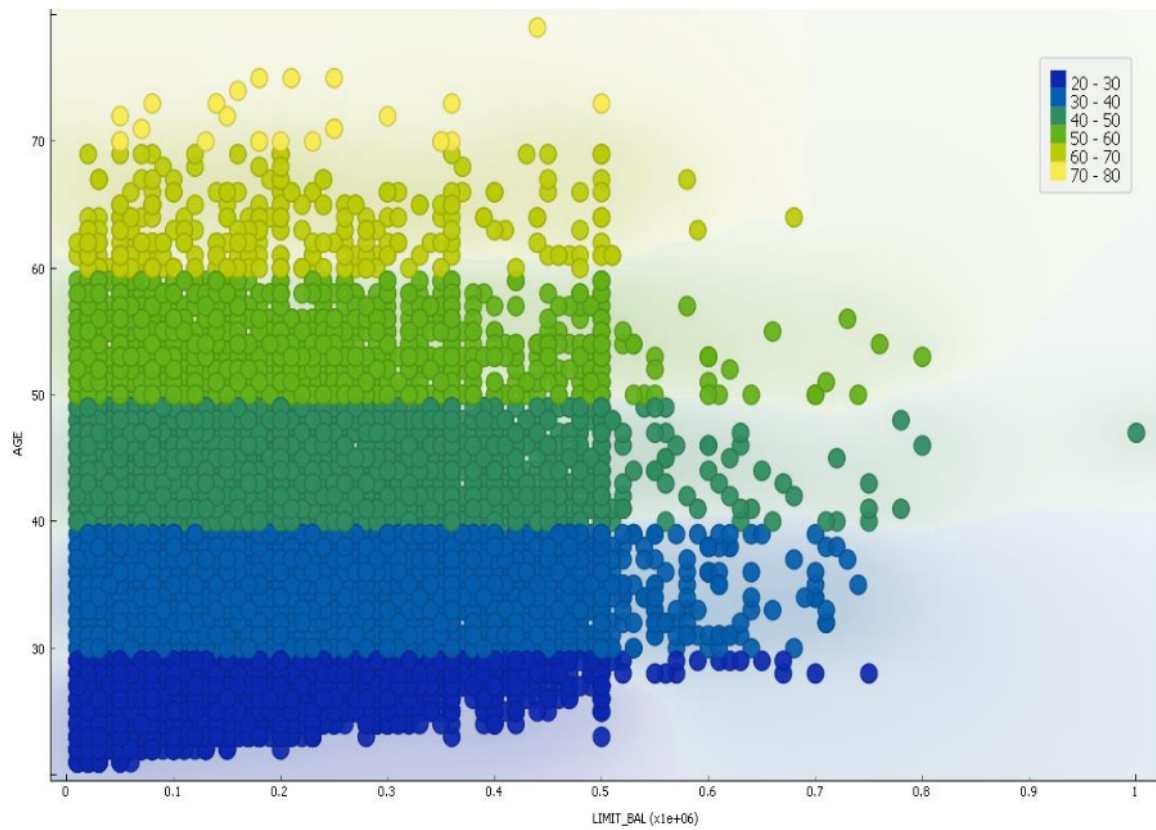


Figure 3.6: Credit Limit and Age of clients

Among the clients with credit greater than 500000, only 22 are seen to default, with the highest number of defaulters being people between 40 and 42. This is depicted in Figure 3.7.

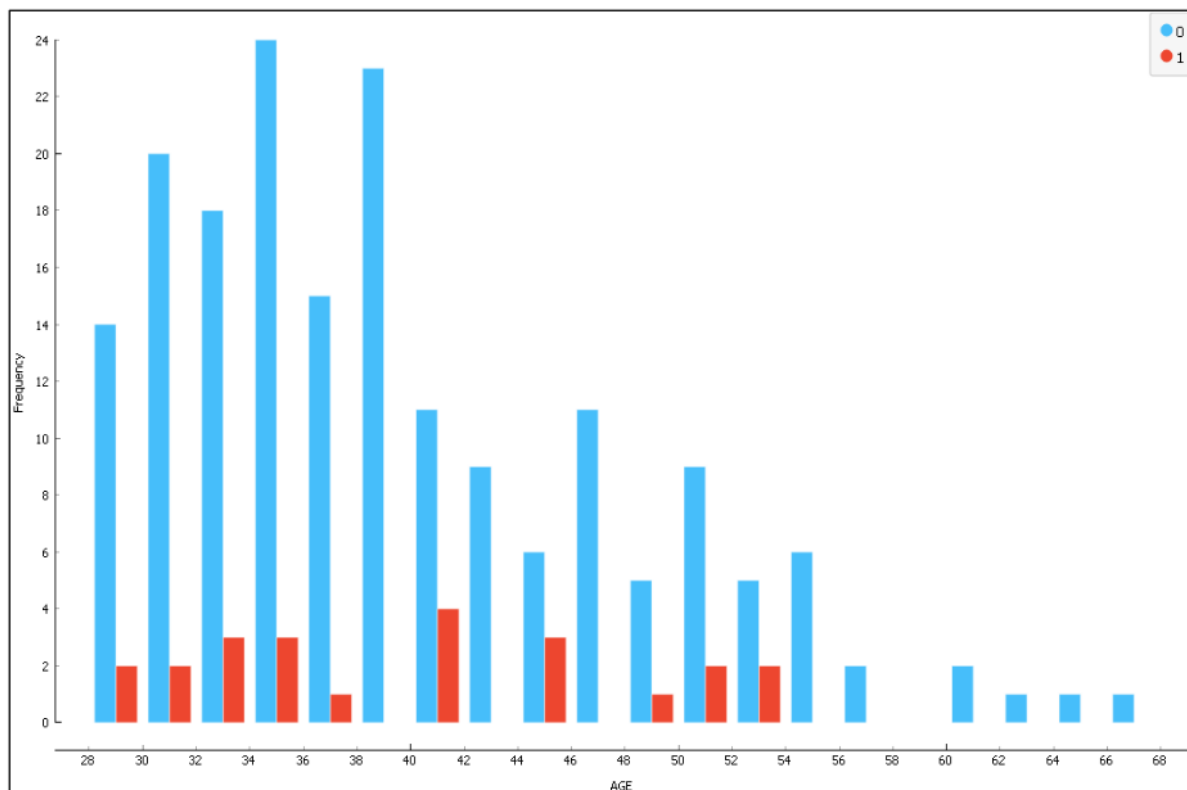


Figure 3.7: Credit above 500000

For people with credit limits less than or equal to 500000, the highest probability of default is observed in men and women aged 25-30, more precisely, at age 29. This is depicted in Figure 3.8, where blue represents males, and red represents females.

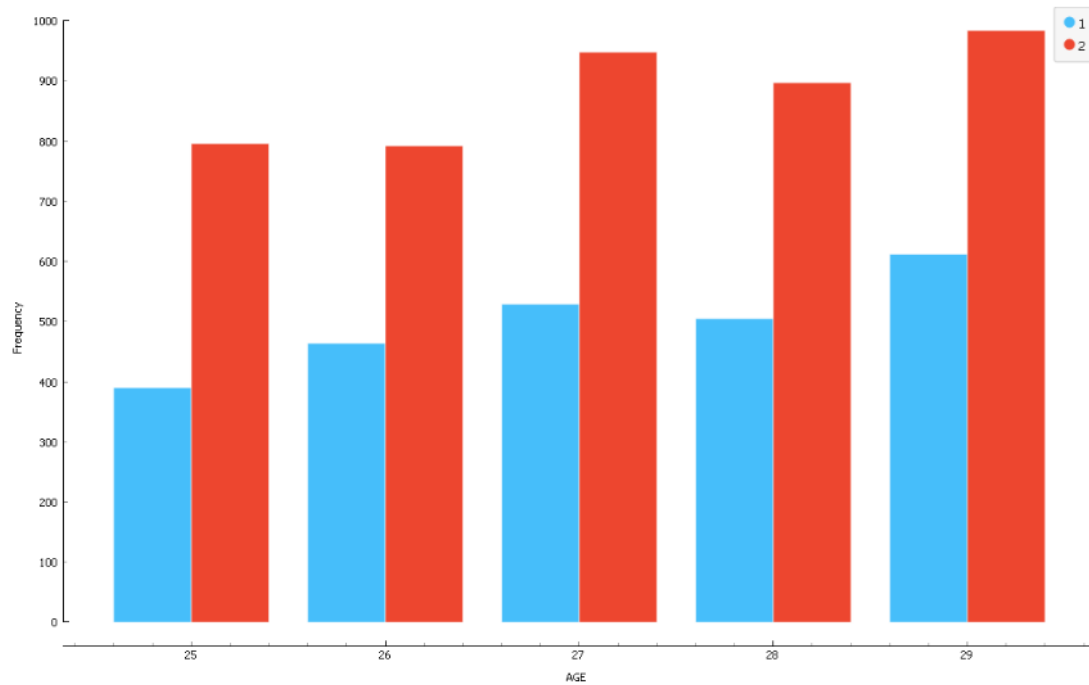


Figure 3.8: Credit below 500000

v. Education and default

The diagram in Figure 3.9 depicts the relationship between the education levels of the customers and default. The diagram shows there are more non-defaulters at the post-graduate and undergraduate levels.

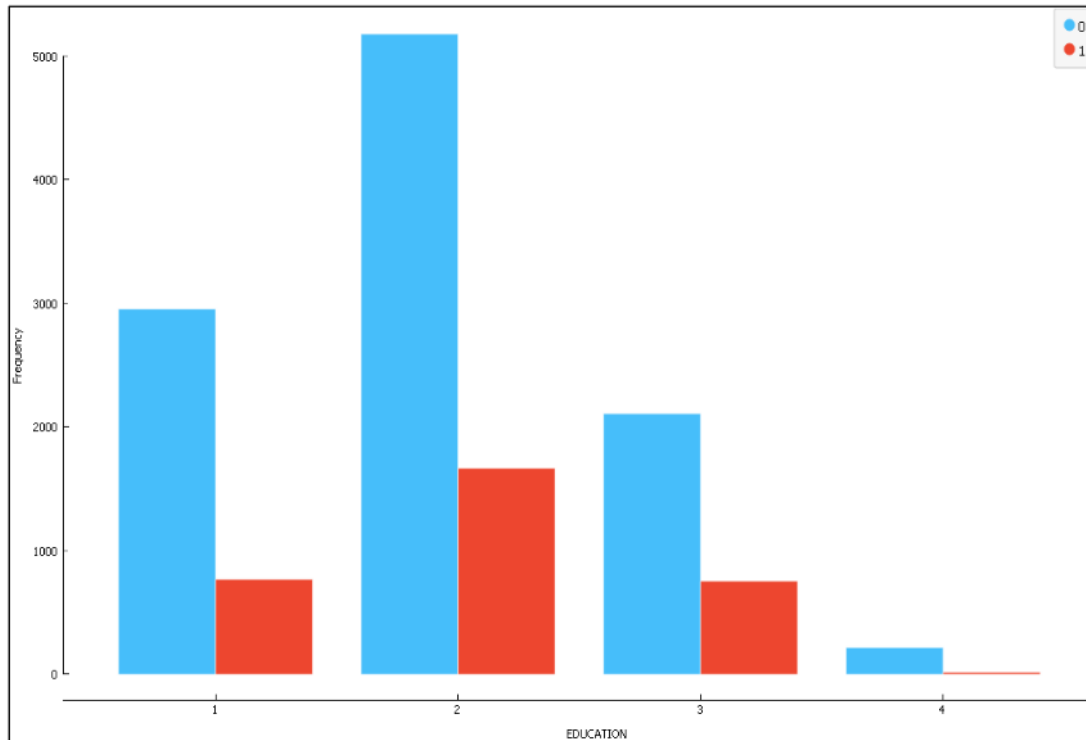


Figure 3.9: Education and default

vi. Marital Status and Default

According to the dataset, the single clients (including divorced) are mostly non-defaulters, meaning that married persons are more likely to default, with 43.04% single and divorced non-defaulters compared to 34.84% married non-defaulters. Among the married persons, 38% are married men, with 9.85% being defaulters, while 62% are married women, with 13.62% being defaulters. Comparing the total number of married women and men with their probability of default, and the age group with a high likelihood of default, we can conclude that married men between the ages of 25 to 30 are more likely to default. The diagram is shown in Figure 3.10.

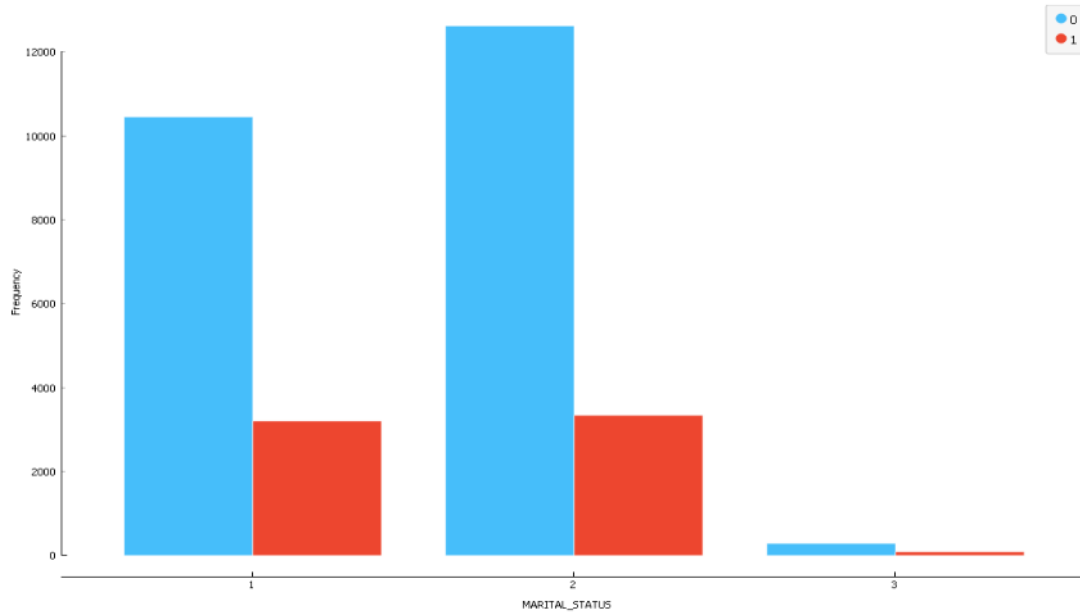


Figure 3.10: Marital status and default

3.2.3. Classification

After visualising the data to better understand it, the next task was to train classification models to determine the one with the best metrics and is best at predicting default with a higher level of accuracy. While accuracy is a good performance metric, it can sometimes be unreliable, so F1 score, Precision and Recall will also be used. The dataset was split 80:20 so that 20% would serve as new data for predictions to test the overall performance of the classifiers. Then this 80% data was used for training and evaluating the models. Before actual prediction, the data was pre-processed to make it suitable for the classification task with the following steps:

- 1) Removing Outliers: Outliers were removed using the Local Outlier factor, set to 13 Neighbors and 'Euclidean' as the metric. It produced a report of 21986 inliers and 2014 outliers, with a contamination of 0.1. The inliers were used in the training data.
- 2) One-hot encoding and Normalisation: Two 'select columns' and 'continuize' widgets were used to separate the categorical values to be encoded and the others containing the numeric values to be normalised. The numeric values were scaled between 0 and 1, as there is a high deviation between the numeric values. For instance, the BILL_AMT values and the age. This step is essential to prevent the classification algorithms from being biased towards attributes with higher values. After this, they were merged with the 'Merge Data' widget.

- 3) Training and Testing Data Split: The training data (80% of the overall dataset) was split, with 20% used as test data to evaluate the models with the Test and Score widget. The data was split using the Data Sampler widget.
- 4) Feature Selection and Reduction: For this step, the 'correlations' widget was used to determine the values that are highly correlated. 'Pearson correlation' was used to discover that the BILL_AMT attributes are highly correlated, with values above 80%. The Rank widget was also used to detect and select attributes with the highest ranking that provide more information about the dataset. The repayment status attributes were highly ranked, signifying their importance in predicting default. The PCA (principal component analysis) produced 10 components with a variance of 97%. The Scree diagram is shown in Figure 3.11.

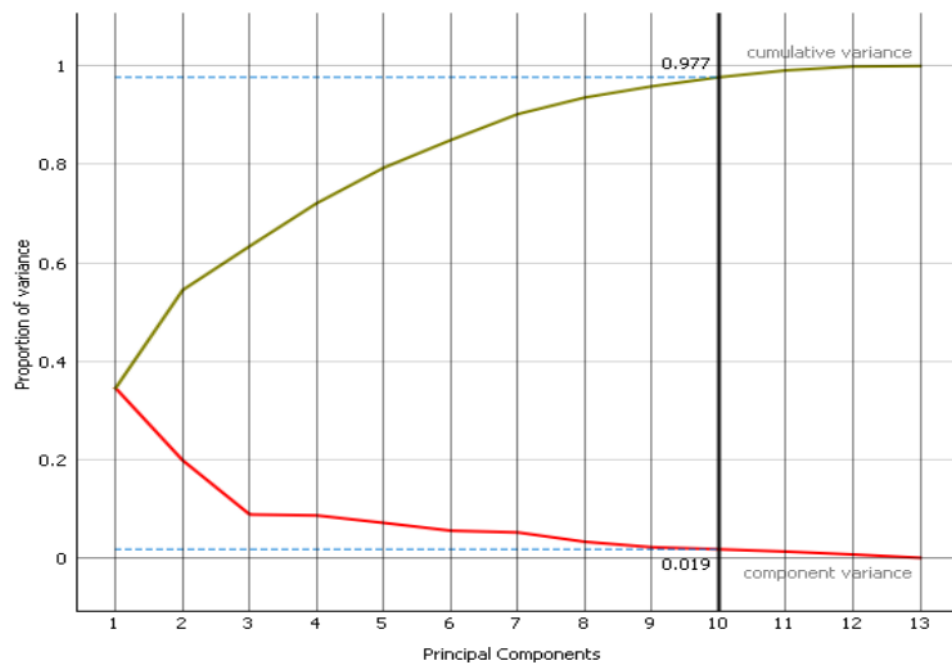


Figure 3.11: Scree diagram of the PCA

- 5) Oversampling: The data sampler widget was also used in oversampling the default class to handle the imbalance between the classes, i.e., the default and non-default classes. Oversampling of the default class was performed to prevent the models from being biased towards the majority class and ignoring the minority class. Figure 3.12 represents the default classes after oversampling.

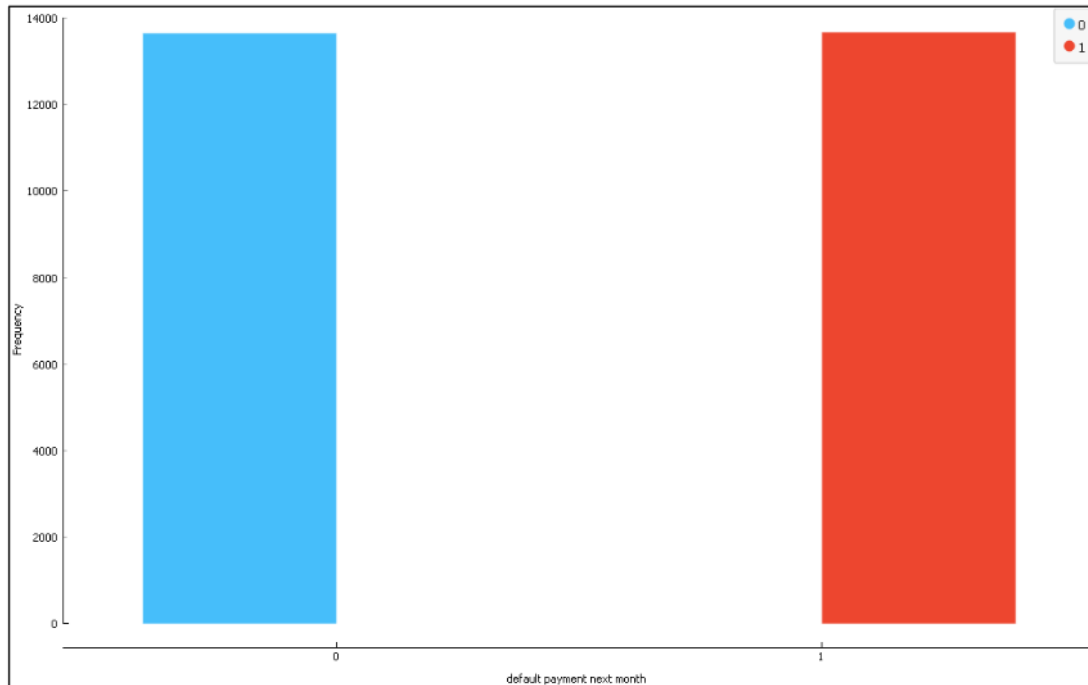


Figure 3.12: Class balance

- 6) Evaluating the models: The Logistic Regression, Support Vector Machine, Random Forest, Decision Tree, and Neural Network models were evaluated using the Test and Score widget on the testing data. The evaluation results in Figure 3.13 show that the random forest classifier is better than the other classifiers with this dataset.

Evaluation results for target (None, show average over classes) ▾					
Model	AUC	CA	F1	Precision	Recall
Tree (1)	0.890	0.871	0.871	0.876	0.871
SVM (1)	0.424	0.435	0.424	0.430	0.435
Random Forest (1)	0.970	0.917	0.917	0.920	0.917
Neural Network	0.809	0.727	0.726	0.730	0.727
Logistic Regression (1)	0.724	0.688	0.686	0.693	0.688

Compare models by: Area under ROC curve ▾ ☐ Negligible diff.: 0.1					
	Tree (1)	SVM (1)	Random Fo...	Neural Net...	Logistic Re...
Tree (1)		1.000	0.000	1.000	1.000
SVM (1)	0.000		0.000	0.000	0.000
Random Forest (1)	1.000	1.000		1.000	1.000
Neural Network	0.000	1.000	0.000		1.000
Logistic Regression (1)	0.000	1.000	0.000	0.000	

Figure 3.13: Evaluation Result

- 7) Prediction: The models above, i.e., the Linear Regression, Support Vector Machine, Random Forest, Decision Trees, and Neural Network, were trained and used in making predictions on new data which was derived from the original dataset. The new dataset consists of the last 6000 data points from the original dataset. This new dataset is also imbalanced, with 4734 non-default data points and 1266 default data points. The 'Predictions' widget was used to make predictions on the new data. Some of the models' hyperparameters were tuned as follows:

Neural Network: Number of neurons = 300, Activation Function = ReLu, Optimizer = Adam, Regularisation = 0.001.

Random Forest: Number of trees = 12.

(Every other hyperparameter retained its default settings).

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1. Reflection

This study attempts to predict the values for `default_payment_next_month` based on results from analysis by the dataset authors. Data analysis shows that the credit limit, repayment status and payment amount from April to September are the key predictors of default. Customers with lower credit limits are likely to default. Also, the higher the payment amount for each month, the lower the likelihood of a credit card default. Similarly, a higher delay in repayment of credit card bills from previous months increases the probability of default.

4.1.1. Performance Evaluation

In predicting the probability of a default on credit cards, five supervised learning models were trained with the pre-processed data and predictions were made on previously unseen data. The performance metric scores of the models are displayed in Table 4.1.

Model	CA (Classification Accuracy)	F1	Precision	Recall
Logistic Regression	0.748	0.762	0.786	0.748
SVM	0.393	0.441	0.602	0.393
Random Forest	0.796	0.787	0.781	0.796
Decision Tree	0.728	0.732	0.736	0.728
Neural Network	0.758	0.770	0.788	0.758

Table 4.1: Performance Scores

From the results, random forest is the best classifier, producing a relatively high Precision score, the highest Recall and F1 scores (a good metric for imbalanced data). The F1 score gives the harmonic mean of the recall and precision values. Precision determines the correctly predicted positives, in this case, the default values. Recall gives the total number of predicted positives (correctly predicted or not). The results also show that the Support Vector Machine model did not perform well in classification, both in evaluation with the test and score widget and in predicting default on new data. Accuracy, which is also a good metric, cannot be relied on in this dataset. If a model incorrectly classifies all the values for the minority class and correctly classifies the majority class, it could produce high accuracy scores, but it is actually inaccurate.

For the models to be of business significance, they should also produce low false negatives and high true positives, i.e., a reduced number of incorrect predictions for 'non-default' values, as it would be more

disadvantageous for a bank if the model predicts 'no default' but a default occurs. Hence, the confusion matrices for the models will be studied.

4.1.2. Confusion Matrix

The confusion matrix is also a good metric to determine the number of correct and incorrect predictions made by the models. The 'Confusion Matrix' widget was used to plot the confusion matrices for the models. Table 4.2 depicts the format for the confusion matrix.

	Predicted - 0	Predicted - 1
Actual - 0	TN (The model predicted 'no default', and the result is actually true).	FP (The model predicted 'default', but the result is actually false).
Actual - 1	FN (The model predicted 'no default', but the result is actually false).	TP (The model predicted 'default', and the result is actually true).

Table 4.2: Confusion matrix

Figures 4.1, 4.2, 4.3, 4.4, and 4.5 depict the confusion matrix for the machine learning models. The confusion matrix for the random forest model shows that it correctly classified 52% of default values and produced a low number of false negatives. While the logistic regression model had the lowest false negatives with 11.9%, it produced more inaccurate predictions for 1 (default) than the random forest classifier. Therefore, the confusion matrix also shows that the random forest classifier performs better than others.

		Predicted		
		0	1	Σ
Actual	0	88.1 %	57.0 %	4734
	1	11.9 %	43.0 %	1266
Σ		4226	1774	6000

Figure 4.1: Logistic Regression

		Predicted		Σ
		0	1	
Actual	0	72.0 %	83.8 %	4734
	1	28.0 %	16.2 %	1266
Σ		2480	3520	6000

Figure 4.2: Support Vector Machine

		Predicted		Σ
		0	1	
Actual	0	85.1 %	48.0 %	4734
	1	14.9 %	52.0 %	1266
Σ		4995	1005	6000

Figure 4.3: Random Forest

		Predicted		Σ
		0	1	
Actual	0	83.5 %	63.4 %	4734
	1	16.5 %	36.6 %	1266
Σ		4637	1363	6000

Figure 4.4: Decision Tree

		Predicted		Σ
		0	1	
Actual	0	87.9 %	55.5 %	4734
	1	12.1 %	44.5 %	1266
Σ		4329	1671	6000

Figure 4.5: Neural Network

CHAPTER FIVE: CONCLUSION

5.1. Summary and Recommendations

This study carried out a binary classification task on an imbalanced dataset. It used the Orange data mining tool to determine the probability of default of credit card clients from a Taiwanese bank and the classifier that best predicts this default. The data was split for training, testing and predictions. The training data was pre-processed with Orange's widgets to enable the machine learning models to perform well in classification, and then it was used to make predictions on new data.

The results show that the random forest model performed better than the other models; however, it could not produce performance metric scores above 80%. Also, it correctly predicted only 52% of credit card defaulters, meaning that it is still insufficient to be used in real-world situations. It is also evident from the data that the amount of credit allowed to customers and frequency of repayment of credit card bills highly contribute to the probability of a credit card default. That is, they are very good predictors of default.

For future works, the study will make use of gradient boosting algorithms which are found to generally perform better than random forest algorithms, and proper fine-tuning of hyperparameters to derive better results.

REFERENCES

- [1] “Overdraft Explained: Fees, Protection, and Types,” *Investopedia*, 2023. <https://www.investopedia.com/terms/o/overdraft.asp> (accessed Apr. 16, 2023).
- [2] I-Cheng. Yeh and C. Lien, “The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients,” *Expert Systems with Applications*, vol. 36, no. 2, pp. 2473–2480, Mar. 2009, doi: <https://doi.org/10.1016/j.eswa.2007.12.020>.
- [3] Y. Chen and R. Zhang, “Research on Credit Card Default Prediction Based on k-Means SMOTE and BP Neural Network,” *Complexity*, vol. 2021, pp. 1–13, Mar. 2021, doi: <https://doi.org/10.1155/2021/6618841>.
- [4] S. Yang and H. Zhang, “Comparison of Several Data Mining Methods in Credit Card Default Prediction,” *Intelligent Information Management*, vol. 10, no. 05, pp. 115–122, 2018, doi: <https://doi.org/10.4236/iim.2018.105010>.
- [5] T. M. Alam *et al.*, “An Investigation of Credit Card Default Prediction in the Imbalanced Datasets,” *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: <https://doi.org/10.1109/access.2020.3033784>.
- [6] X. Gao, J. Wen, and C. Zhang, “Forecasting of Credit Card Default Based on Incremental Random Forest,” *Journal of Physics: Conference Series*, vol. 1828, no. 1, p. 012122, Feb. 2021, doi: <https://doi.org/10.1088/1742-6596/1828/1/012122>.
- [7] S. Shi, R. Tse, W. Luo, and G. Pau, “Machine learning-driven credit risk: a systemic review,” *ResearchGate*, Jul. 16, 2022. https://www.researchgate.net/publication/362051812_Machine_learning-driven_credit_risk_a_systemic_review (accessed May 05, 2023).
- [8] “UCI Machine Learning Repository: default of credit card clients Data Set,” *Uci.edu*, 2016. <https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients> (accessed Apr. 15, 2023).
- [9] reel jojo, “Default on Their Credit Card,” *Kaggle.com*, 2023. <https://www.kaggle.com/datasets/reeljojo/default-on-their-credit-card/code> (accessed Apr. 13, 2023).
- [10] robertofranceschi, “GitHub - robertofranceschi/Default-Credit-Card-Prediction: Analysis and classification using machine learning algorithms on the UCI Default of Credit Card Clients

Dataset,” *GitHub*, Nov. 20, 2022. <https://github.com/robertofranceschi/Default-Credit-Card-Prediction> (accessed Apr. 13, 2023).

- [11] A. Band, “The default of Credit Card Clients Dataset: Classification & Evaluation,” *Medium*, Sep. 21, 2020. <https://ameyband.medium.com/default-of-credit-card-clients-dataset-classification-evaluation-2c46c858d981> (accessed Apr. 15, 2023).
- [12] T. M. Alam *et al.*, “An Investigation of Credit Card Default Prediction in the Imbalanced Datasets,” *IEEE Access*, vol. 8, pp. 201173–201198, 2020, doi: <https://doi.org/10.1109/access.2020.3033784>.
- [13] ainslie, “Credit Card Default Prediction & Analysis,” *Kaggle.com*, May 17, 2018. <https://www.kaggle.com/code/ainslie/credit-card-default-prediction-analysis> (accessed Apr. 16, 2023).
- [14] “Comparative Study of Classification Techniques on Credit Defaults,” *Comparative-Study-of-Classification-Techniques-on-Credit-Defaults*, 2019. <https://analytics-nuts.github.io/Comparative-Study-of-Classification-Techniques-on-Credit-Defaults/> (accessed Apr. 16, 2023).
- [15] “Orange Visual Programming — Orange Visual Programming 3 documentation,” *Readthedocs.io*, 2022. <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/index.html> (accessed Apr. 19, 2023).
- [16] Smiling Buddha, “Credit Card Default Prediction Classification Capstone Project,” *YouTube*. Dec. 03, 2022. Accessed: Apr. 29, 2023. [YouTube Video]. Available: <https://www.youtube.com/watch?v=EJNlrZ79Er0&t=363s>
- [17] Joos Korstanje, “The F1 score | Towards Data Science,” *Medium*, Aug. 31, 2021. <https://towardsdatascience.com/the-f1-score-bec2bbc38aa6#:~:text=Solving%20imbalanced%20data%20through%20metrics,of%20errors%20that%20are%20made>. (accessed Apr. 29, 2023).

Final_Submission

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

Instructor

80/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

GRADING FORM: DATA MINING / VIS A2_2022/3

CHISOM NWANKWO

80

KNOWLEDGE AND UNDERSTANDING

Dataset description (15 marks)



Credit card customer dataset taken from UCI repository examined and your rationale for exploring its features shown very clearly.

16

COGNITIVE/ INTELLECTUAL SKILLS

Primary and Secondary data analysis, reflection (35 marks)



Your choice of algorithms evaluated are robust and appraisal of each consistent.

30

PRACTICAL APPLICATION SKILLS

Primary and Secondary data processing (20 marks)



Pre-processing as applicable to your chosen dataset performed sufficiently well and reflection provided offers a fair knowledge of established techniques of enquiry underpinned by adequate forms of fundamental research.

17

TRANSFERABLE SKILLS

Data reporting and report writing (30 marks)



Format and content demonstrate an adequate portrayal of skills for life and professional practice.

17