Credit Card Defaults Prediction

Pramir KC

Simranjeet Kaur Gahir

Master of Digital Innovation, Dalhousie University

STAT5620 Data Analysis

Prof. Joanna Mills Flemming

April 19, 2022

Abstract

Although credit cards have existed in the banking industry for the past several decades, in recent times their usage has witnessed a significant jump. The primary reason for this has been growth in income of the people resulting in changing their spending habits (Fulford & Schuh, 2018). Credit cards provide the increased spending power to the consumers leading to higher transaction volumes. This in turn also results in credit card defaults as consumers tend to spend more than their earning capacity. For consumers, their credit score takes a nosedive which makes it difficult for them to get loans. For banks, they face huge financial losses and therefore, credit card default prediction can help them to minimize their losses. The main aim of this experiment is to predict the probability of the customer to default by analyzing the customer's demographics and credit card history. For this purpose of prediction, logistic regression and random forest classification have been implemented to achieve the best performance results. Apart from this, key features leading to default have been identified using random forest feature importance scores. After experimentation with models, it has been found out that random forest performed comparatively better than logistic regression in terms of capturing the credit card defaults.

Keywords: Credit Card Default, Classification, Logistic Regression, Random Forest, Variable Selection

Table of Contents

List of Figures

List of Tables

Introduction

Credit card default is a major problem for the banks and financial institutions across the world (Li, 2020). For example, for the fourth quarter of 2021, the credit card default rate was 1.62 % for American banks. The delinquency rate increases significantly during uncertain times like the financial crisis or COVID pandemic. During the first quarter of the 2009 financial crisis, the delinquency rate increased to 6.61 (Fred Economic Data, 2022). As the percentage of the population using the credit card is increasing every year for developed countries, even a 1% of credit card default rate puts a huge financial loss to the banks (Best, 2021). Thus, if the banks can accurately determine the variables that contribute highly to credit card default and subsequently identify the customers who are more likely to default, they could devise a targeted strategy which could save a large sum of money that gets lost every year.

The research question that we want to address is whether we could predict customers that will default on their credit card loans based on some statistical models? Moreover, we want to identify key explanatory variables that contribute to credit card default based on our credit card default dataset (Yeh & Lien, 2009). To address our research question, we applied two most widely used methods for classification: Logistic Regression (LR) and Random Forest (RF).

LR is one of the most widely used statistical methods for binary classification for low dimensional data (dataset where the number of explanatory variables is significantly smaller than the sample size). (Maalouf, 2011). In this parametric model, the independent variable vector (X) is linked to the probability of outcomes (binary) modeled by the dependent variable (y) by a logit function. The response probability, p = Pr(Y =1|x), is modeled with the logistic regression of the form:

$$logit(p) = log(p/1\text{-}p) = \alpha + \boldsymbol{\beta}'x$$

where α is the intercept parameter and the $\boldsymbol{\beta}$s are slope parameters for independent variable vector, X (Hosmer et al., 2013).

In Random Forest, an ensemble of uncorrelated trees from CART methods are generated using bootstrap data from original samples. Moreover, when a tree is split, only some of the variables are selected for splitting (Couronné et al., 2018). As a result, Random Forest allows for the reduction of variance compared to a single decision tree (Couronné et al., 2018). Mathematically, RF can be represented as a collection of tree-structured classifiers {h (x, k), k = 1,...}. Here, k represents independent vectors that are identically distributed. Moreover, each tree is allowed to vote for the most popular class at input x (Breiman, 2001).

RF performs well for classification problems; however, it is very difficult to understand the intricate decision mechanism as the number of trees is very large (Malato, 2021). One of the main advantages of the RF model is the built-in mechanism to compute the variables importance. A variable that increases the pureness of the leaves gets the higher variable importance. This process is repeated for each decision tree in RF and an average value is calculated and finally normalized. Thus, when all the variable importance scores are added, it sums up to 1 (Malato, 2021). For the feature selection, Recursive Feature Elimination using k-fold cross Validation can be used. In this method the importance of each feature is calculated, and the least important features are removed from the current set. This process is

repeated until the best set of features are obtained which maximizes the performance through cross-validation (Malato, 2021).

Data Analysis and Pre-processing

The data on the credit card usage come from University of California Irvine's (UCI) machine learning repository (Yeh & Lien, 2009). This dataset contains six months data of credit card customers of a bank in Taiwan. Specifically, the data includes information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients April 2005 to September 2005. The sample size of the data contains 30,000 observations and 25 variables in total.

Table 1:Variables Description

Variable	Variable Definition	Variable values	Type
ID	ID of each client		Continuous
LIMIT_BAL	Amount of given credit in NT dollars		Continuous
SEX	Gender of the client	1=male, 2=female	Categorical
EDUCATION	Educational status of the client	1=graduate school, 2=university, 3=high school, 4=others, 5,6=unknown	Categorical
MARRIAGE	Marital status	1=married, 2=single, 3=others	Categorical
AGE	Age in years		Continuous
PAY_0	Repayment status in September	-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, ... 9=payment delay for nine months etc.	Categorical
PAY_2	Repayment status in August		Categorical
PAY_3	Repayment status in July		Categorical
PAY_4	Repayment status in June		Categorical
PAY_5	Repayment status in May		Categorical

PAY_6	Repayment status in April		Categorical
BILL_AMT1	Amount of bill statement in September	NT dollar	Continuous
BILL_AMT2	Amount of bill statement in August	NT dollar	Continuous
BILL_AMT3	Amount of bill statement in July	NT dollar	Continuous
BILL_AMT4	Amount of bill statement in June	NT dollar	Continuous
BILL_AMT5	Amount of bill statement in May	NT dollar	Continuous
BILL_AMT6	Amount of bill statement in April	NT dollar	Continuous
PAY_AMT1	Amount of previous payment in September	NT dollar	Continuous
PAY_AMT2	Amount of previous payment in August	NT dollar	Continuous
PAY_AMT3	Amount of previous payment in July	NT dollar	Continuous
PAY_AMT4	Amount of previous payment in June	NT dollar	Continuous
PAY_AMT5	Amount of previous payment in May	NT dollar	Continuous
PAY_AMT6	Amount of previous payment in April	NT dollar	Continuous
default.payment.next.month	Default payment	1=yes, 0=no	Categorical

Exploratory Data Analysis (EDA)

Before experimenting with the variable and model selection, pre-processing is performed to prepare the data for further analysis and modeling. The unknown values of variable *EDUCATION* with status 5 and 6 have been merged with status 4 (‘others’). Similarly, *MARRIAGE* is handled by masking unknown values with the ‘others’ category. Since the ID column has no relationship with the predictor variable so it has been dropped from the dataset. This dataset contains no null and missing values.

The response variable is a binary variable i.e., *default_payment* (renamed) which depicts 0 or ‘no’ for non-defaulters and 1 or ‘yes’ for defaulters.

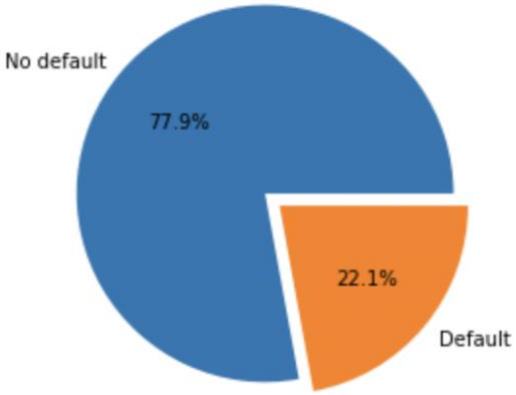

Figure 1: Distribution of Response Variable

From Figure 1, it is depicted that our dataset is highly unbalanced where only 22.1% clients are defaulters. Hence, out of total observations, 23364 are non-defaults while 6636 are found to have default status.

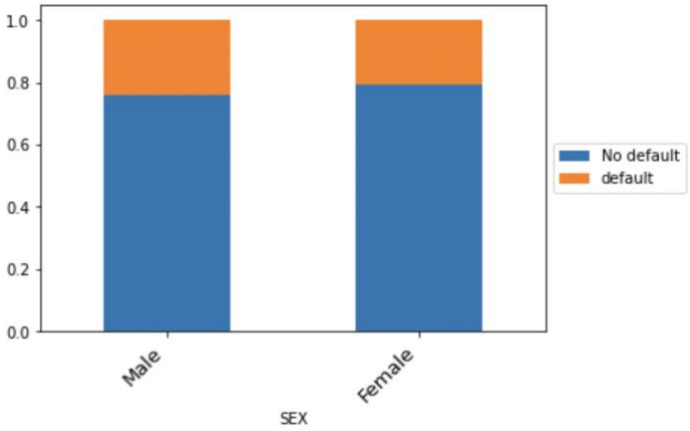

Figure 2:Distribution of default w.r.t gender

Figure 2 shows that males (24%) have engaged in slightly more default cases as compared to females (20%). However, whether it's male or female, overall the proportion of defaults is relatively low being inline with the overall data.

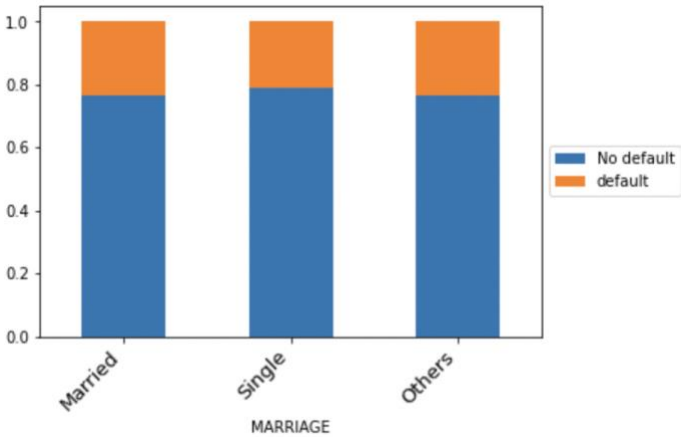

Figure 3: Distribution of default w.r.t Marital Status

From Figure 3, it is observed that in terms of marital status, the number of single person defaulters (20%) are slightly less as compared to married people (23%).

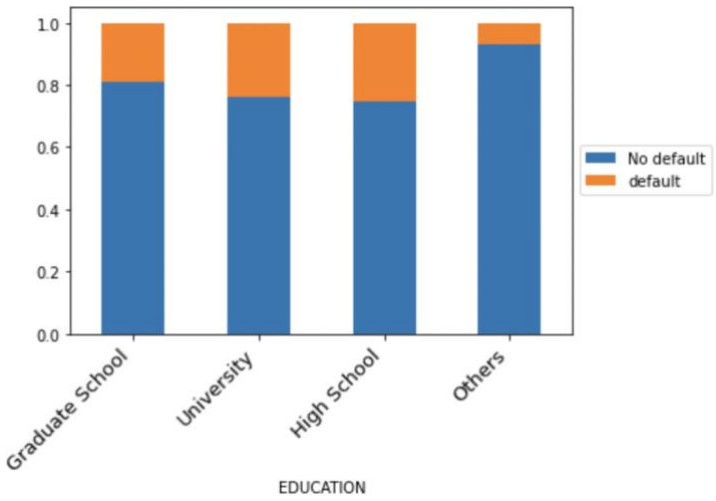

Figure 4:Distribution of default w.r.t Education

Credit card holders who have studied till high school or people who went to university for education default more on their credit card bills as compared to graduate degree holders. Due to lack of information about the data, we do not know why the other group has comparatively less number of defaulters.

```
50000     3365
20000     1976
30000     1610
80000     1567
200000    1528
Name: LIMIT_BAL, dtype: int64
```

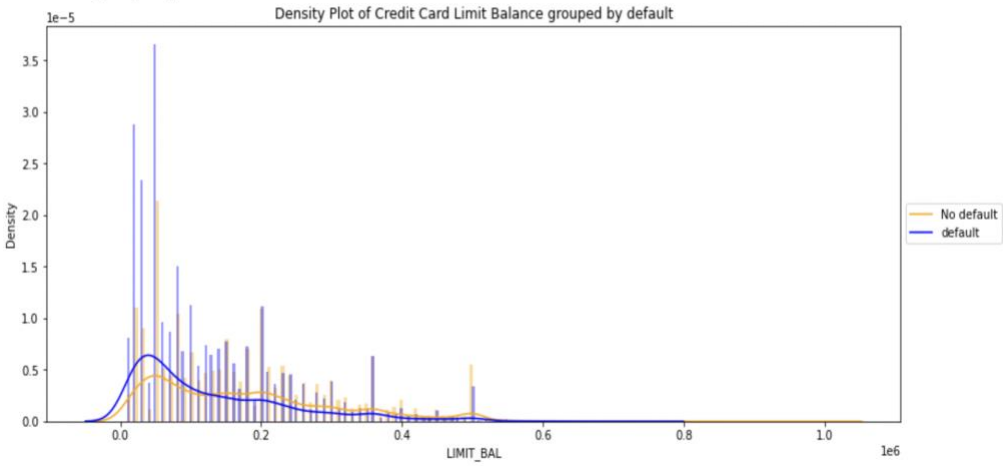

Figure 5:Density Plot of Credit Card Limit Balance grouped by default

Figure 5 shows that the largest number of credit card holders have a credit limit of 50,000, 20,000 and 30,000 NT dollars. Most defaulters have a credit limit in the range of 0-100000 NT dollars as the probability density for defaults is more than non-defaults in this segment.

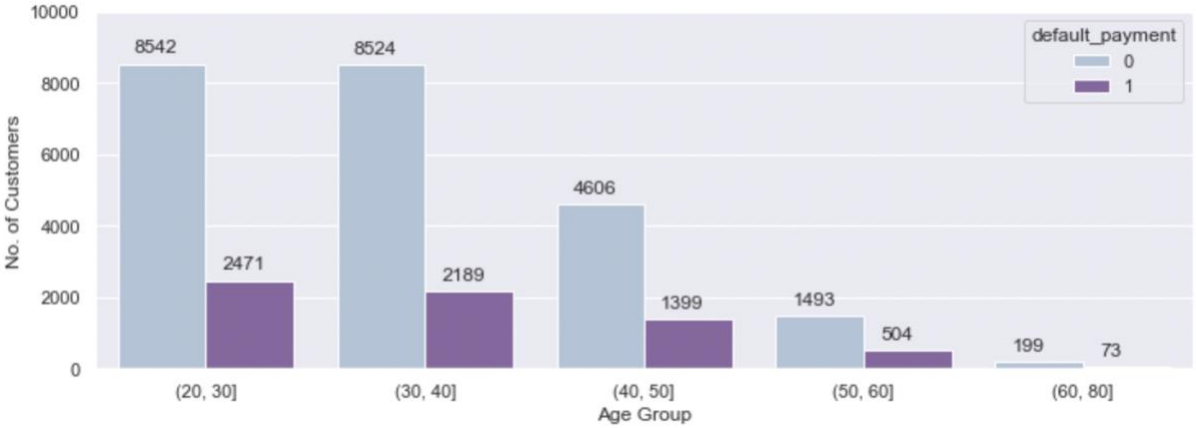

Figure 6:Customers count in different Age brackets

Figure 6 demonstrates that the majority of the credit card owners as well as defaulters fall in the age group of 20 to 40. Hence, younger people are more likely to default than the middle-aged and old generation.

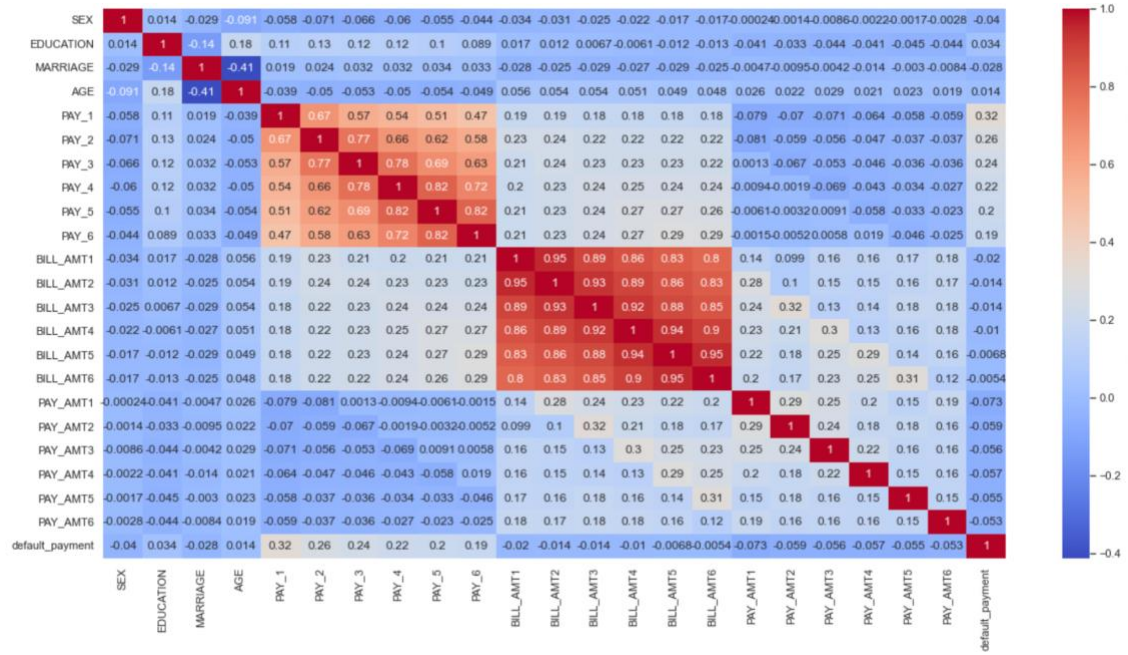

Figure 7:Correlation between all variables

From the correlation matrix, Bill_Amt variables are highly correlated with the adjacent months and the correlation decreases as the Bill_Amt are further months apart. For example, the amount of bill statements in September is highly correlated with August and so on and the correlation decreases between the months of September and July. Similarly, the repayment status of a month (pay_ variable) also follows the same pattern. This observation was not surprising to us. We expected a strong correlation between the amount of previous payment (Pay_AMT) in one month with its adjacent month, however there was no such correlation.

Methods

Logistic Regression

Logistic Regression is a statistical method to predict a binary outcome (Riffenburgh & Gillen, 2020). In other words, this model is mostly used for binary classification problems. Here, the logistic regression models the probability of the categorical response variable i.e. default_payment. The objective is to predict the probability of credit card default based on the credit card holder's demographics and credit history.

Before fitting the logistic model to the data, data has been prepared by binning the continuous variables to categorical data (Lund, 2016). Binning is a categorization technique in which a continuous variable is transformed to certain small sets of groups. Continuous variables like Age, Bill_Amt1 to Bill_Amt6, Pay_Amt1 to Pay_Amt6 have been preprocessed to create bins. In logistic regression, the outliers can influence the result of the analysis and lead to incorrect inferences (Sarkar et al., 2011). Hence, outliers were removed from the bins and rebinning is performed to accommodate the distribution of categorical data. Since the ID variable is not correlated to the response variable so it has been removed from the dataset. One hot encoding is the next step performed in data preparation. One hot encoding is the process of converting categorical values into binary vectors for the better prediction of the model (Cerda et al., 2018) (Refer Appendix A for more information).

After preprocessing is done, we split the dataset into a training and testing set. Training data is used to develop the model while testing data is used to confirm how the model performs in predicting the response variable. On dividing the dataset in the ratio of 70:30 given that 70% constitute the training data and 30% contain the testing data, the logistic model is fitted to the training data. Thereafter, prediction probabilities are generated based on the test data. We aimed to decide on the cut-off probability by calculating performance metrics on various probabilities.

Performance metrics define the performance and evaluation of any statistical model. Accuracy score, Recall, Confusion Matrix and Flagged rate are the certain performance metrics that have been used to evaluate the prediction performance of the logistic model. Accuracy is the ratio of the number of correct predictions to the total number of input samples.
Accuracy is calculated as:

$$\text{Accuracy} = \frac{Number\ of\ correctly\ classified\ testing\ samples}{Total\ number\ of\ testing\ samples}$$

However, accuracy is a governing metric if the data are balanced. Since data is unbalanced in our case, accuracy is not a preferable metric. In order to define the governing metric, we created Confusion Matrix which shows for each class the number of the data that are correctly classified for that class.

Predicted Class

True Class		
	True Positives (TP)	*False Negatives (FN)*
	False Positives (FP)	*True Negatives (TN)*

Figure 8:Confusion Matrix

where,
TP = It is a positive sample classified as positive by the model in which all the predicted defaults are actual defaults.

TN = It is a negative sample classified as negative by the model in which all the predicted non-defaults are actual non-defaulters.

FN = It is a positive sample classified as negative in which the actual defaults are predicted as non-defaults by the model.

FP = It is a negative sample classified as positive in which the actual non defaults are predicted as defaults by the model.

Instead of accuracy, Recall is considered to be the governing metric here for this dataset. Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions that could have been made.

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

Flagged rate is the derived metric created which depicts the predicted default over the total number of samples with respect to each cut-off probability.

$$Flagged\ Rate = \frac{TP + FP}{TP + FN + FP + TP}$$

	probablity	accuracy	recall	flagged rate
0.16	0.16	0.647	0.745	0.462
0.17	0.17	0.672	0.719	0.425
0.18	0.18	0.693	0.699	0.395
0.19	0.19	0.715	0.680	0.365
0.20	0.20	0.730	0.664	0.342
0.21	0.21	0.743	0.643	0.320
0.22	0.22	0.754	0.621	0.300
0.23	0.23	0.763	0.609	0.286
0.24	0.24	0.770	0.592	0.271
0.25	0.25	0.776	0.577	0.258
0.26	0.26	0.782	0.562	0.245

Figure 9:Performance Metrics w.r.t probabilities

Random Forest

For the Random Forest model, the explanatory variables (X) and the response variable (y) were split into training and testing dataset with an 80:20 split. Furthermore, the training dataset was split into a 75:25 ratio such that the final training, validation, and testing dataset had 18000, 6000, 6000 samples respectively. The categorical variables were encoded into a one-hot numeric array. The values of X variables were standardized using the fit_transform function where the output z-score was obtained using the formula: z = (x - u) / s (u is mean and s is standard deviation) (Pedregosa et al., 2011). The RandomForestClassifier model was fitted using train and validation data and tested using testing data. The validation and testing accuracy was obtained. The hyperparameters for our base RF model were kept as default. The classification score was obtained using the confusion_matrix function which consisted of precession, recall and f-1 score along with the confusion matrix (Pedregosa et al., 2011). The weighted average precession score values were used to assess our model performance.

Variable selection in random forest:

We used the sklearn tool feature_importnace_ which calculates the feature importance score based on the mean decrease in impurity (MDI) (Pedregosa et al., 2011). MDI is the measure of overall decrease in impurity within each tree and it is calculated as the mean or standard deviation of accumulation of impurity decrease at every single split for a given variable (Li et al., 2019). Subsequently, the feature importance is ranked in this method.We subsequently used recursive feature elimination(RFE) to select the most important variables for our RF model. RFE is used along with cross-validation to select the best number of variables (Pedregosa et al., 2011).

Hyperparameter tuning in random forest:

We used randomizedsearchcv tool from sklearn to find the best hyperparameter for our RandomForest model (Pedregosa et al., 2011). In randomizedsearchcv only a select number of parameters settings are chosen during sampling and this is performed with cross-validation for selecting the optimal parameters. The hyperparameter types and choices for parameter values/class for the optimization using randomizedsearchcv are depicted in table1 in appendix B. The parameters of the estimator used to apply these methods are optimized by cross-validated search over parameter settings (Pedregosa et al., 2011).

Results

Logistic Regression

For results, we focussed on Recall rate achieved at certain cut-off probability rather than accuracy because our dataset has binary labels and is unbalanced (Juba & Le, 2019). Model is tuned in such a way that the default capture rate is high and incorrect flagging of defaults is low. Technically, recall rate should be high and flagged rate should be low. Hence, on careful tuning, the cut-off probability of 0.22 has been selected to get optimal results. To conclude, on flagging 30% of the credit card users for default, our model is able to correctly predict 62% of actual defaults and provides the accuracy of 75% as inferred from Figure 9. These results are depicted in graphical representation in Figure 10.

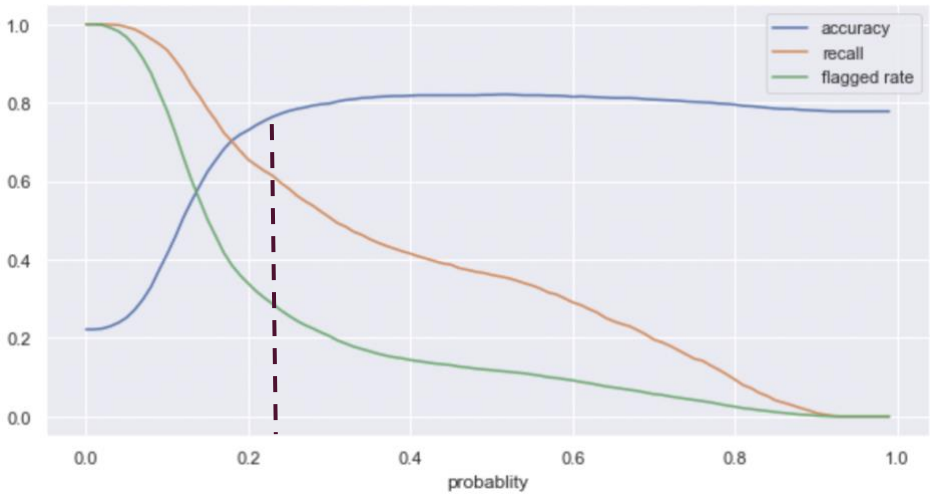

Figure 10:Performance metrics graphical representation

(at Cut-off probability - 0.22, Flagged Rate - 0.3, Recall - 0.62, Accuracy - 0.75)

Random Forest

From the baseline model, we performed variable selection based on feature score of mean decrease in impurity.

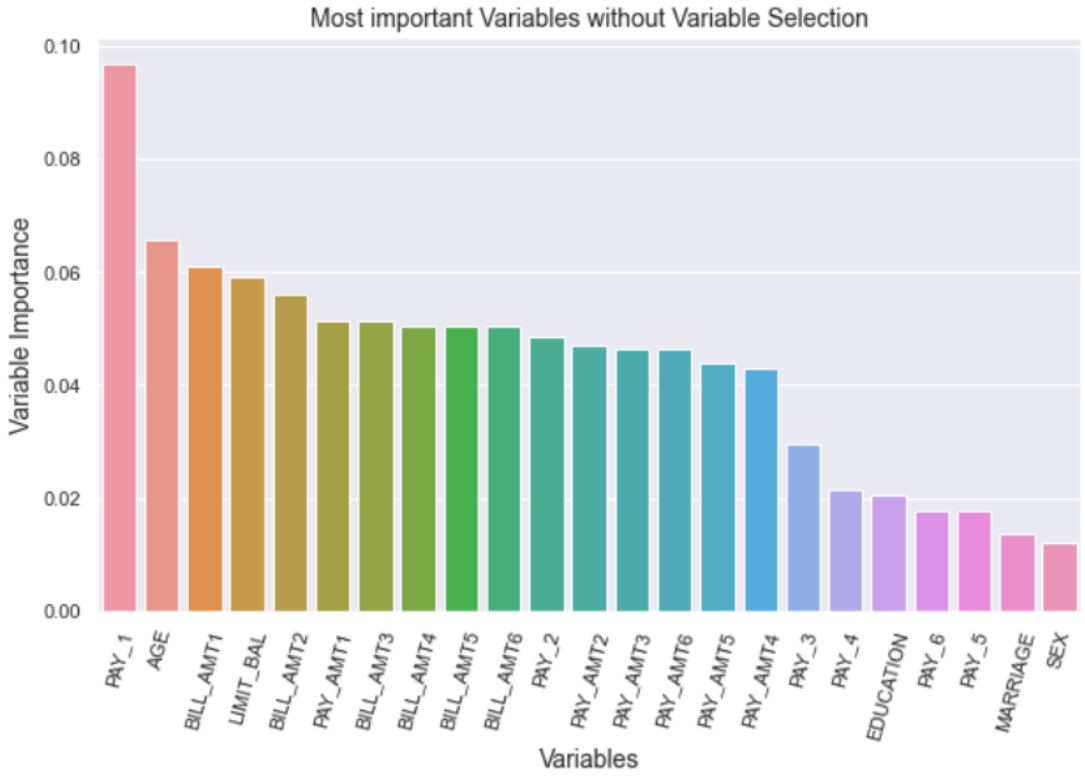

Figure 11:The explanatory variable importance shown from highest to lowest from left to right.

We eliminated 7 of the least important features based on recursive feature elimination (RFE). The figure with most important variables after RFE was performed is included in the appendix B.

When we compared the accuracy of the baseline model to the updated model, the accuracy increased from 82.27 % to 82.38 %. For the updated model, the weighted average for the recall score was obtained to be 82 %.

We applied the following best parameter obtained from hyperparameter tuning to our updated model.

Table 2:Optimized parameters for updated RF model

Hyperparameters type	Choices for parameters after hyperparameter tuning
Number of trees in the forest (n_estimators)	200

Min sample required to split internal node (min_samples_split)	8
Min number of samples required to be at a leaf node ('min_samples_leaf)	5
max_features=sqrt(n_features)	sqrt
Bootstrap samples used	False
maximum depth of the tree (max_depth)	4

When the hyperparameter tuning was applied to our updated model, we observed a small increase of model performance as the accuracy and recall score of 82.98 % and 83 % was obtained compared to the accuracy and recall score of 82.38 % and 82 % for our updated model. Thus, the hyperparameter tuning for our updated model helped in slight increase in our model performance.

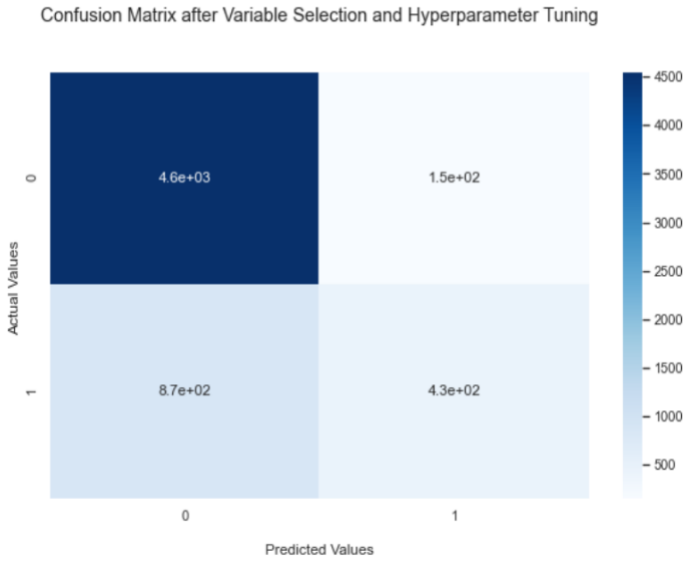

Figure 12:Confusion Matrix for the updated RF model after eliminating 7 least unimportant variables and performing hyperparameter tuning.

Conclusion

To predict the probability of the credit card default, Logistic Regression and Random Forest classification has been applied to the dataset. Through the experiments, we were able to achieve the accuracy of 75% by logistic model and 83% by random forest model.

Metric	*Logistic Regression*	*Random Forest Classification*
Accuracy	75%	82.98%
Recall	62%	83%

However, accuracy not being the governing metric here, we are more concerned about the recall rate. The reason behind this is our dataset is highly unbalanced in terms of classifying the response binary variable. Because of the low event rate, we don't want the incorrect flagging of non-defaulters by our model while predicting the credit card defaulters. Hence, recall has been considered as the better metric for evaluating the performance of our statistical models. Conclusively, Random Forest is a better model for this classification problem than logistic regression for the credit card default dataset because of its better recall rate. Additionally, Random Forest is robust in handling the outliers (Roy & Larocque, 2012) and requires no preprocessing in handling the continuous data. Other than that, variable selection is another important factor in considering the Random Forest classification model.

Acknowledgements

Simranjeet Kaur Gahir:
Data Cleaning , EDA & experimented with Logistic Regression (including binning and one hot encoding) implemented in Python. In Project Report, Abstract, Data & EDA, Logistic Regression, Appendix and Conclusion written in APA format.

Pramir KC:
Data Cleaning, EDA & experimented with Random Forest model. For the project report, I wrote the Introduction section , and the methods and results and appendix B for the Random Forest model.

References

Fulford, S. L., Schuh, S. (2018). Credit Cards and Consumption. Retrieved from: https://www.bis.org/events/eopix_1810/schuh_paper.pdf

Yeh, I. C., & Lien, C. H. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2), 2473-2480. https://doi.org/10.1016/j.eswa.2007.12.020

Riffenburgh, R. H., & Gillen, D. L. (2020). Logistic regression for binary outcomes. https://doi.org/10.1016/B978-0-12-815328-4.00017-6

Hosmer, D. W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied Logistic Regression. https://doi.org/10.1002/9781118548387.ch4

Lund, B., (2016). Weight of Evidence Coding and Binning of Predictors in Logistic Regression. https://www.mwsug.org/proceedings/2016/AA/MWSUG-2016-AA15.pdf

Sarkar, S.K., Midi, H., & Rana, S. (2011). Detection of Outliers and Influential Observations in Binary Logistic Regression: An Empirical Study, 11: 26-35. DOI: 10.3923/jas.2011.26.35

Cerda, P., Varoquaux, G., & Kégl, B. (2018). Similarity encoding for learning with dirty categorical variables. https://arxiv.org/pdf/1806.00979.pdf

Juba, B., & Le, H. S. (2019). Precision-Recall versus Accuracy and the Role of Large Data Sets. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 4039-4048. https://doi.org/10.1609/aaai.v33i01.33014039

Roy, M., & Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics*. https://doi.org/10.1080/10485252.2012.715161

Best, R. de. (2021, July 20). *Topic: Credit cards worldwide*. Statista. Retrieved April 18, 2022, from https://www.statista.com/topics/8212/credit-cards-worldwide/#dossierKeyfigures

Board of Governors of the Federal Reserve System (US). (2022, February 23). *Delinquency rate on credit card loans, all commercial banks*. FRED Economic Data. Retrieved April 18, 2022, from https://fred.stlouisfed.org/series/DRCCLACBS

Breiman, L. (2001). *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random Forest versus logistic regression: A large-scale benchmark experiment. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2264-5

Li, X., Wang, Y., Basu, S., Kumbier, K., & Yu, B. (2019). A Debiased MDI Feature Importance Measure for Random Forests. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada*.

Li, Z., & News, B. (2020, March 30). *A global consumer default wave is just getting started (2)*. Bloomberg Law. Retrieved April 18, 2022, from https://news.bloomberglaw.com/banking-law/a-global-consumer-default-wave-is-just-getting-started-in-china

Maalouf, M. (2011). Logistic regression in Data Analysis: An overview. *International Journal of Data Analysis Techniques and Strategies*, *3*(3), 281. https://doi.org/10.1504/ijdats.2011.041335

Malato, G. (2021, November 8). *Feature selection with Random Forest*. Your Data Teacher. Retrieved April 18, 2022, from https://www.yourdatateacher.com/2021/10/11/feature-selection-with-random-forest/

Yeh, I.-C., & Lien, C.-hui. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020

Appendix

Appendix A

One hot encoding is a technique to prepare data in order to facilitate better prediction. Some statistical tools like decision tree, random forest etc. can handle categorical data directly depending upon the implementation. However, others require categorical variables to be mapped to numeric or integer values.
In this method, each categorical value is converted into a new categorical column and is assigned a binary value of 0 or 1. The index is assigned the binary value 1 while all other values are 0.

Type	**One Hot Encoding** →	**Type**	**Red_Onehot**	**Green_Onehot**	**Blue_Onehot**
Red		Red	1	0	0
Green		Green	0	1	0
Blue		Blue	0	0	1

Appendix B

Table 3:Choices for parameters values for hyperparameter tuning

Hyperparameters type	Choices for parameter values/class for Optimization
Number of trees in the forest (n_estimators)	50, 100, 150, 200
Min sample required to split internal node (min_samples_split)	2, 5, 10
Min number of samples required to be at a leaf node ('min_samples_leaf)	2, 3, 4, 5
max_features=sqrt(n_features)	'auto', 'sqrt'

Bootstrap samples used	True, False
maximum depth of the tree (max_depth)	4, 6, 10, 12, 15

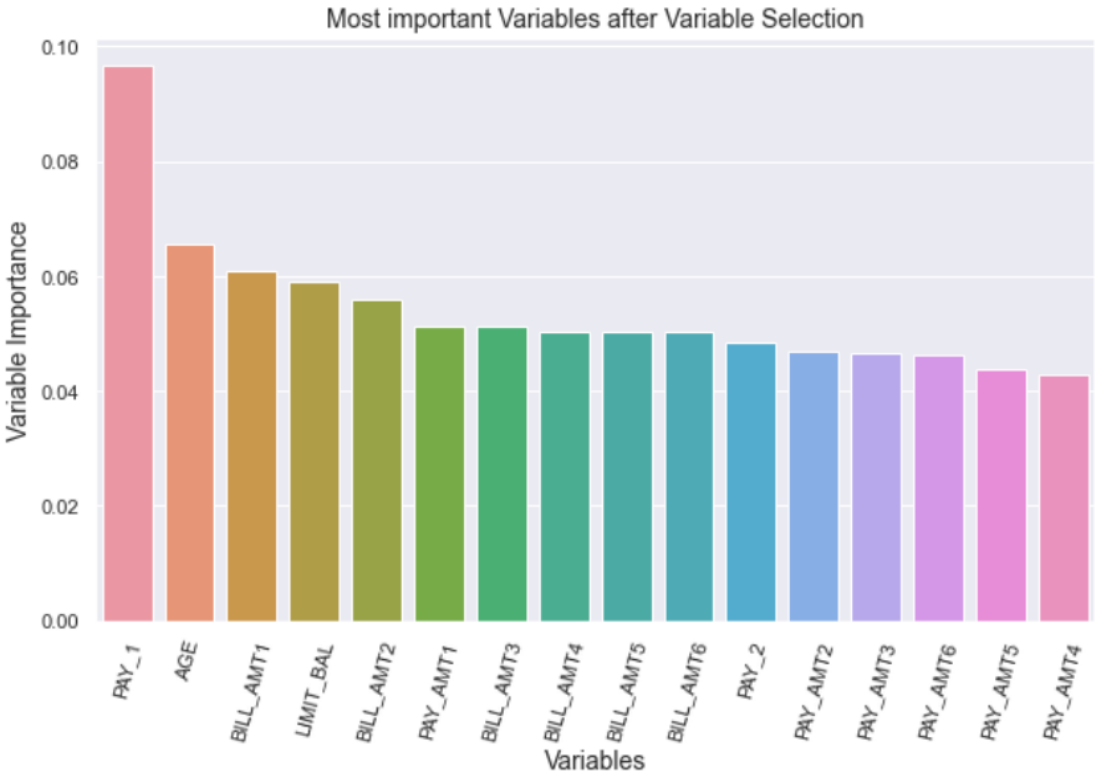

Figure 13:The explanatory variable importance shown from highest to lowest from left to right after removing 7 least unimportant features.