### Business problem

The main business objective of this project is to predict which customers are likely to default on their credit card repayments next month. This task is important in business practice as the default behaviour has a direct bearing on the financial health of the bank system, and effective prediction could reduce the financial risks. The high economic cost led by default behaviour has prompted researchers and practitioners to make efforts on default prediction.

### Available data

The datasets are based on the publicly available credit card default dataset from the UCI Machine Learning Repository. The training dataset contains 23101 observations of 25 variables as shown in Table 1. The testing dataset contains 6899 observations.

|  |  |
| --- | --- |
| Column | Description |
| ID | ID of each client |
| LIMIT\_BAL | Amount of given credit in dollars (includes individual and family/supplementary credit |
| SEX | Gender (1=male, 2=female) |
| EDUCATION | (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown) |
| MARRIAGE | Marital status (1=married, 2=single, 3=others) |
| AGE | Age in years |
| PAY\_X | Repayment status for the past X months. -1=paid on time; 1=payment delay for one month; 2=payment delay for two months etc |
| BILL\_AMTX | Bill amount for past X months |
| PAY\_AMTX | Payment amount for last X months |
| default | Default payment next month (1=yes, 0=no). This is the Target variable |

Table 1

### Data preparation

The result of missing value checks shows that there is no missing value in the dataset. However, there are some noises in the dataset. For example, the ages of some customers are over 200, which is obviously false. I have done basic data cleaning based on personal characteristics including age and gender. After basic data cleaning, 23084 observations remain in the training set and 26.52% of them will default next month.

### Insights from exploratory data analysis

Figure 1 is the correlation heatmap which shows correlation exists between variables. The distinct rectangles show that payment amounts in different months are correlated, and a similar pattern exists in bill amount and repayment status in past months.

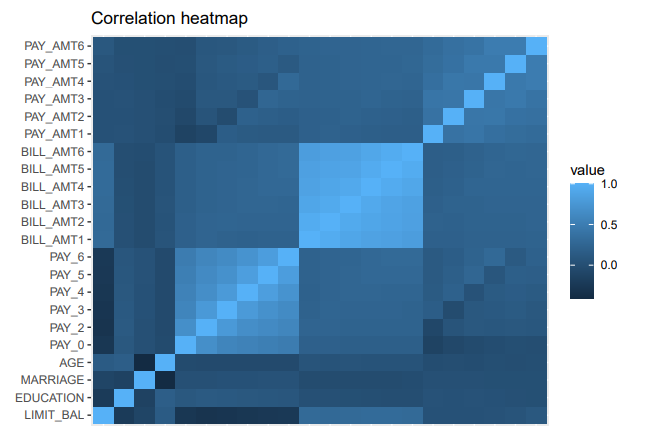


Figure 1[[1]](#footnote-0)

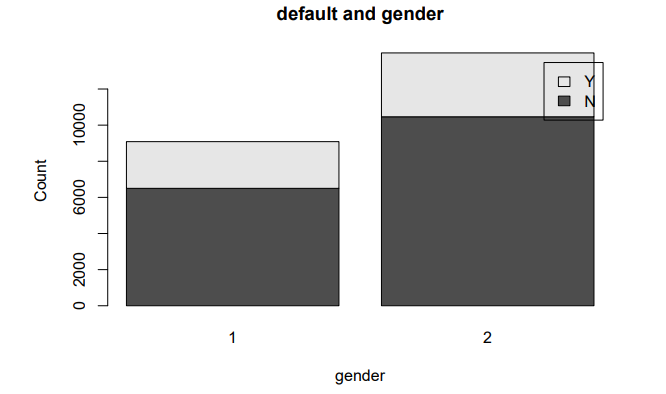


Figure 2

Figure 2 depicts the relationship between gender and default probability. It shows that there is no clear difference between men and women in the default behaviour.

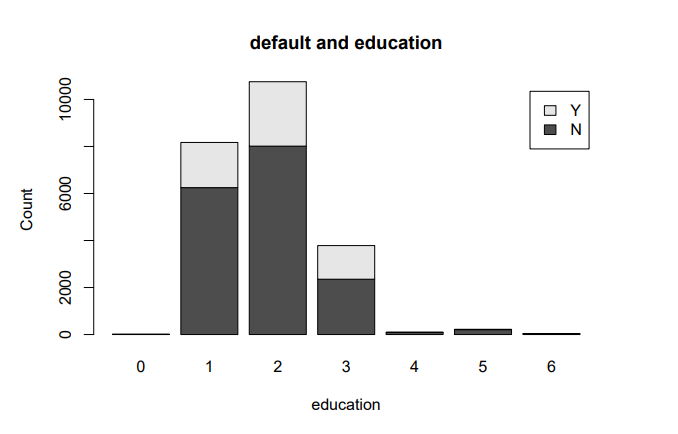


Figure 3[[2]](#footnote-1)

Figure 3 shows the relationship between education level and default probability. It seems that the probability of default will decrease with the education level.

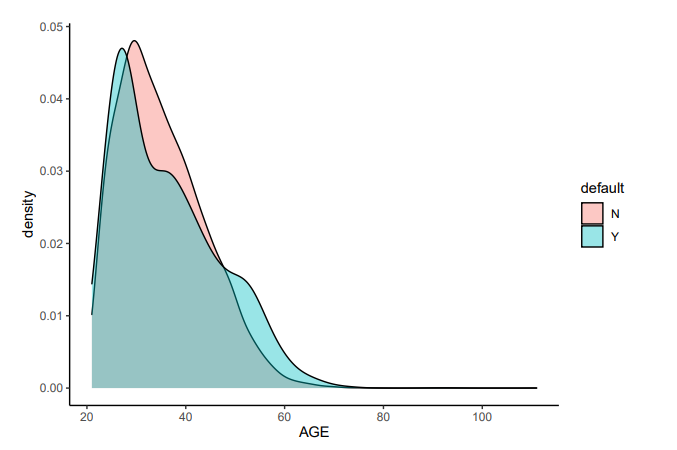


Figure 4

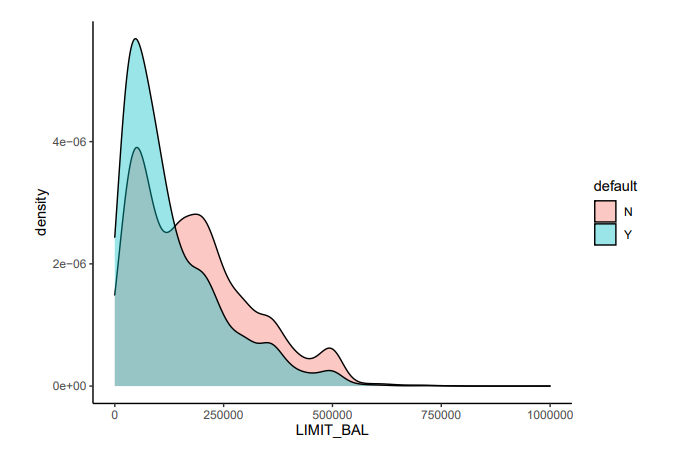


Figure 5

Figure 4 shows the relationship between age and default behaviour. The age of customers that will not default is more concentrated. Age instead of gender could have predictive power on the default prediction. Figure 5 shows the relationship between the amount of given credit and default behaviour. Customers who are given more amount of given credit are less likely to default. This is because that people with a high reputation will enjoy higher credits, and the act of keeping promises will also further improve their reputation.

### Modelling

R programming based on the RStudio platform will be used. The training set would be divided into two parts: the training part and the validation part. The training part accounts for 75% and the validation part accounts for 25%. Machine learning models would be trained in the training part and tested in the validation part, then the performance of different models in the validation part would be compared and the best model could be selected. The validation part is also used to tune the hyperparameters. For example, the random forest classifier has many hyperparameters such as the max depth of the tree, the minimum number of samples, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node. The best combinations of hyperparameters could be found given enough time. But considering the limit in time and computational power, only a few combination s will be tested. A better combination of model hyperparameters could be found given more time.

Three kinds of models are trained and tested in this project: linear (logistic) model, naïve bayes model, and random forest model. The logistic model estimate the probability of one event (out of two alternatives). It is an extension of the linear regression in the classification problem. Some assumptions should be held before implementing the logistic regression: the target variable should be binary and the desired outcome is represented by the factor level 1; the independent variables should be independent of each other; the sample size should be large enough. In this project, there are some multi-collinearities in the model, meaning this model may not generalize well to the practice. Naïve Bayes model is easy to build and useful especially for large data sets. It even outperforms other more sophisticated model though it is simple. Naïve Bayes classifier applies Bayes’ theorem with strong independence assumptions between features (From the correlation heatmap we see some features are correlated, which means the Naïve Bayes classifier may not perform well in the real practice). A random forest classifier consists of multiple decision trees which could make predictions based on a set of nested if-statements on input features. The output of random forest classifier is the class selected by the majority of decision trees (majority voting principle). Random forest could solve the overfitting problem existing in the simple decision tree model. Random forest is often considered as a black box in the business practices since this classifier could generate effective predictions based on a wide range of complicated input features while requiring little configuration. These three models are most widely in the classification problem and have great business significance. The performance of these classification models will be compared in the test set. The model with best performance will be chosen to make predictions on the test set on Kaggle.

### Evaluation methodology

These three models are trained in the training part and tested in the validation part. The confusion matrix are shown in Table 2, Table 3 and Table 4 respectively.

|  |  |  |
| --- | --- | --- |
| Logistic regression confusion matrix in the test set | | |
| Actual\Predict | 0 | 1 |
| 0 | 12171 | 545 |
| 1 | 3376 | 1221 |

Table 2

|  |  |  |
| --- | --- | --- |
| Naïve Bayes confusion matrix in the test set | | |
| Actual\Predict | 0 | 1 |
| 0 | 12495 | 3996 |
| 1 | 221 | 601 |

Table 3

|  |  |  |
| --- | --- | --- |
| Random forest confusion matrix in the test set | | |
| Actual\Predict | 0 | 1 |
| 0 | 12111 | 3096 |
| 1 | 605 | 1551 |

Table 4

Since the dataset is not extremely imbalanced so the accuracy score could be used as a evaluation metric[[3]](#footnote-2). The accuracy score for logistic regression is 0.7735228, for naïve bayes model is 0.7564258, for random forest is 0.789118. Random forest performs best out of three models. The accuracy score of 0.789118 means 78.91% of the random forest model predictions on whether the customer will default or not are correct. This result shows that the machine learning classifier performs better than the random guess, since previous result shows that 73.48% of the observations will not default, which means simply assigning everyone “non-default” will gain a accuracy of 73.48%.

We can then further explore which set of features have the strongest predictive power using the feature importance calculated by the random forest. This has business implications as the bank managers could better identify potential risks by paying extra attention to these important features. Figure 6 shows the feature importance calculated by the random forest. Random forest shows that Repayment status, Age and Amount of given credit are very important in predicting the default probability, and these features should be focused on. While other features such as marriage status, gender which thought to be important do not have high predictive power actually.

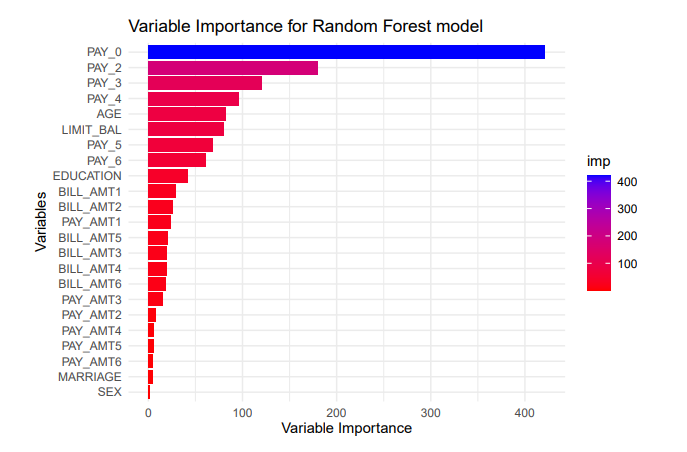


Figure 6

### Preliminary results

After tuning hyperparameters for the random forest model, a desirable set of predictions using the AT2\_credit\_test.csv is generated and submitted to Kaggle. The results are evaluated using the AUC measure (Area Under the ROC Curve) and I got a score of 0.78930.

|  |  |  |  |
| --- | --- | --- | --- |
| ID | default | ID | default |
| 5 | 0.266938 | 62 | 0.190692 |
| 7 | 0.15149 | 63 | 0.748539 |
| 17 | 0.479903 | 67 | 0.630125 |
| 20 | 0.306292 | 73 | 0.177745 |
| 24 | 0.163335 | 74 | 0.249624 |
| 31 | 0.131459 | 79 | 0.199634 |
| 42 | 0.197319 | 80 | 0.310724 |
| 46 | 0.294813 | 89 | 0.145713 |

Table 5

Part of final submitted table is shown as Table 5. This table has two columns: ID and predicted default probability. Customers with a high probability (higher than 0.5) will be considered as defaulters, and customers with a low probability (lower than 0.5) will not be considered to default.

### Consideration of ethical issues

The data in this report is sourced from legitimate open source sources.The data studied in this project is anonymous. Conducting research on the information for the participants will not violate the human rights or dignity of the study participants. Personal characteristics included in the dataset are very safe and not private as information like age and gender will not show the identity of the participants. The models and final results will not cause any physical, social, psychological or other types of harm.

1. the x axis is not labelled since the variable name is too long. The variables have the same order from LIMIT\_BAL to PAY\_AMT6 [↑](#footnote-ref-0)
2. 1=graduate school, 2=university, 3=high school. [↑](#footnote-ref-1)
3. For those extremely imbalanced datasets, other evaluation metrics such as recall, precision, F1 score may better describe the model performance. All these features could be calculated based on the confusion matrix. [↑](#footnote-ref-2)