**36106 Machine Learning Algorithms**

**and Applications - Autumn 2022**

**Assessment task 2**

**- Part A:** Building a Classification Model (Group)

**Group 7 – Supervised Learners:**

Dinh Thi Tran (14382497),

Ivan Cheung (13975420),

John Rho (24509337),

Ryan Yeo (14328254)

**Table of Contents**

[1. Introduction 3](#_Toc1142998233)

[2. Exploratory Data Analysis (EDA) & Data Preparation 3](#_Toc1859166)

[2.1 The Dataset 3](#_Toc1743510252)

[2.2 Data Preparation & EDA 3](#_Toc1936668194)

[3. Modelling 6](#_Toc401450452)

[3.1. Random Forest (RF) 6](#_Toc1144509187)

[3.2. Support Vector Machine (SVM) 7](#_Toc903235707)

[3.3. Logistic Regression (GLM) 7](#_Toc420653673)

[3.4 Gradient Boosting Machines (GBMs) 7](#_Toc610042720)

[4. Modelling Summary 8](#_Toc949492599)

[5. Summary of GBM Model Optimisations 9](#_Toc876491713)

[6. Deployment to Kaggle 10](#_Toc1639842047)

[7. Ethical Considerations 10](#_Toc1034834974)

[Appendix 11](#_Toc231656563)

[Appendix 1: Data Dictionary 11](#_Toc253109693)

[Appendix 2: Table 1. Variables summary 12](#_Toc1056047172)

[Appendix 3: Statement of contributions of each member 13](#_Toc340566011)

[Appendix 4: Random Forest model selection 14](#_Toc1469410075)

[Appendix 6: figure 7 ROCs – GBM 15](#_Toc608073944)

[Appendix 7: Figure 8. Variable importance chart – base GBM 16](#_Toc1598271434)

[Appendix 8: Submission records – Kaggle 17](#_Toc609186973)

[Appendix 9: ODI Data Ethics Canvas 18](#_Toc939282729)

[Appendix 10: References 19](#_Toc721164531)

## 1. Introduction

The business has set forward a goal of predicting if existing customers are likely to default on their credit payments in the month. The ability to accurately determine this information will provide valuable knowledge to the business, allowing for better projections of business liabilities and opportunities to mitigate loss through proactive strategies. To achieve this outcome a data modelling approach will be undertaken using historic data on customer credit card payments and default records. The objective will be to produce a probability of a customer defaulting based on patterns of payments in the previous months. This report has been prepared in compliance with the CRoss Industry Standard Process for Data Mining (CRISP-DM) to provide a structured analysis of team’s discoveries during the project life cycle.

## 2. Exploratory Data Analysis (EDA) & Data Preparation

### 2.1 The Dataset

The main data available to the team, "AT2\_credit\_train”, consists of credit card payment history for 6-month period and it contains 23,101 rows of anonymised customer details and payment information (limit balance, past payment history, etc). The target variable (dependent variable) is named “default” and it has only two values which are “yes” = 1 and “no” = 0. This target variable separates customers into two classes, **class 0** and **class 1.** Our aim is to predict the target class probability based on customer details and payment information (independent variables). A data dictionary was provided (Appendix 1) outlining the valid inputs for each variable.

### 2.2 Data Preparation & EDA

Below is the summary of data pre-processing steps taken to handle data issues prior to modelling. (See Appendix 2 for more details):

* AGE: filtered age to be less than or equal to 75 (limit determined by values in the validation set).
* LIMIT\_BAL: removed observations with negative or zero balance.
* EDUCATION: merged 0 with 4 (both referred to “others”), 6 with 5 (both referred to “unknown”). Hence, the EDUCATION variable takes integer values from 1 to 5 corresponding to graduate school, university, high school, others and unknown (respectively).
* MARRIAGE: merged 0 (others) with 3 (divorce) as both 0 and 3 have very small number of observations. Thus, MARRIAGE take values 1,2,3 corresponding to married, single and others.
* SEX: removed values which are **not** 1 (male) and 2(female) as only 1 and 2 appear the data dictionary.
* ID: dropped this column as it provides no additional information for modelling.
* PAY\_AMTX: changed to factor with levels 0 and 1 as these variables only take two values in the dataset.

|  |  |  |
| --- | --- | --- |
| **Default Class** | **Observations** | **Observations (%)** |
| **N (class 0)** | 16961 | **73.49%** |
| **Y (class 1)** | 6119 | **26.51%** |

*Table 1. Observations of the default class, showing imbalance in the dataset.*

We found that **73.49% of the records falls in class 0 and 26.51% in class 1 (Table 1).** This shows that the data is mildly imbalanced for the dependent variable.

We have also found that the median and the mean for limit balance (LIMIT\_BAL) for group 0 is also higher than group 1 (Figure 1).

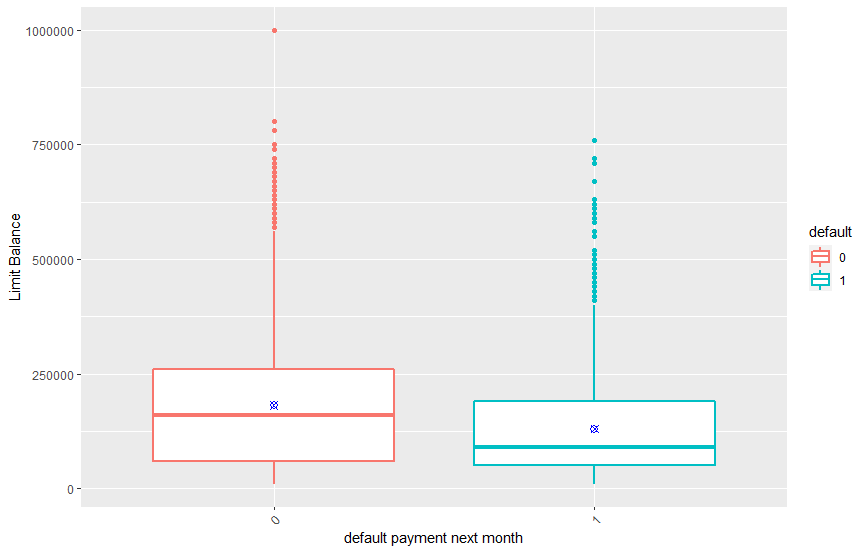
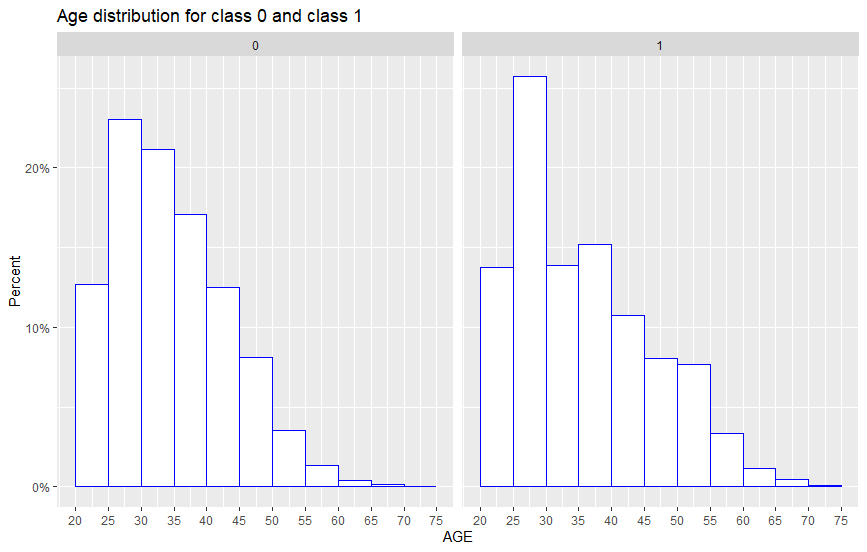
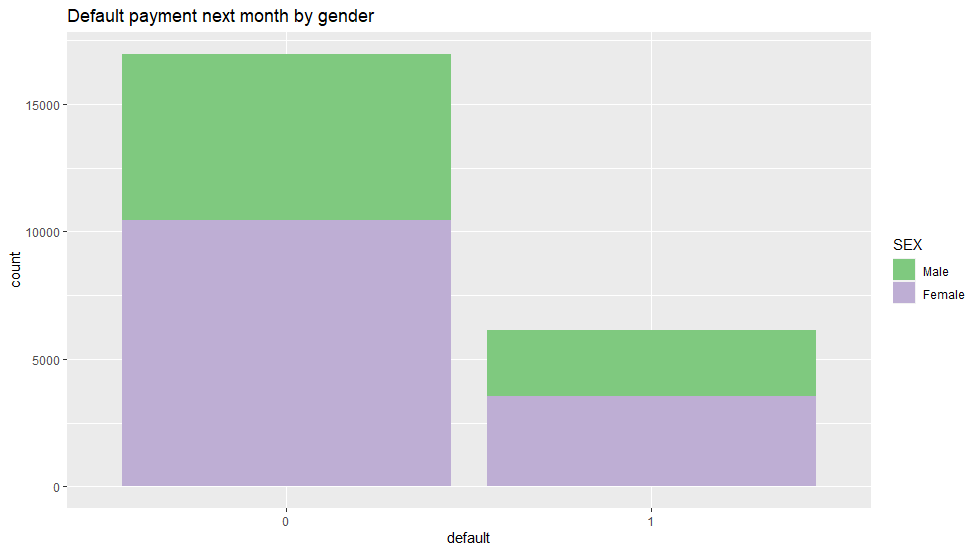
  
*Figure 1: Plot of limit balance (amount of given credit) for each class. The blue dot represents the mean of limit balance.*

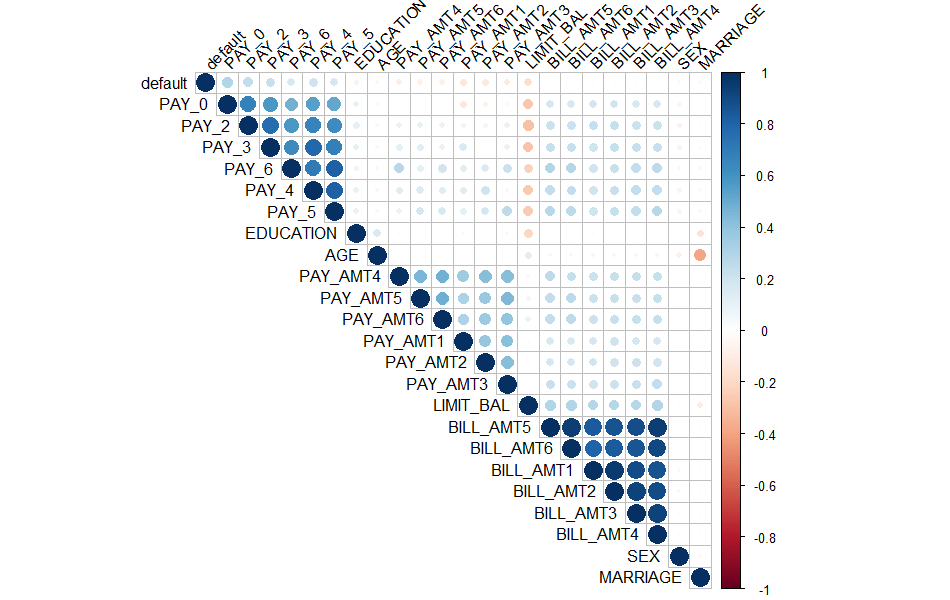
Figure 2 below clearly shows the differences in the age distribution for class 0 and class 1. The age band 25-30 had the highest proportions amongst other age bands in both classes. The 30-35 was the second highest in the class 0 while the proportion was significantly lower for the same group in class 0.

  
*Figure 2: Age distribution for the two classes.*

In terms of AGE, EDUCATION and MARRIAGE there is a similar trend in both groups. The proportion of male and female customers in group 0 is roughly the same as the gender proportion in group 2. This indicates that SEX might not be an important variable in predicting class 0 or 1.

  
*Figure 3: Gender distribution for class 0 and class 1. Male is encoded by “1” and Female is encoded by “2” in the dataset.*

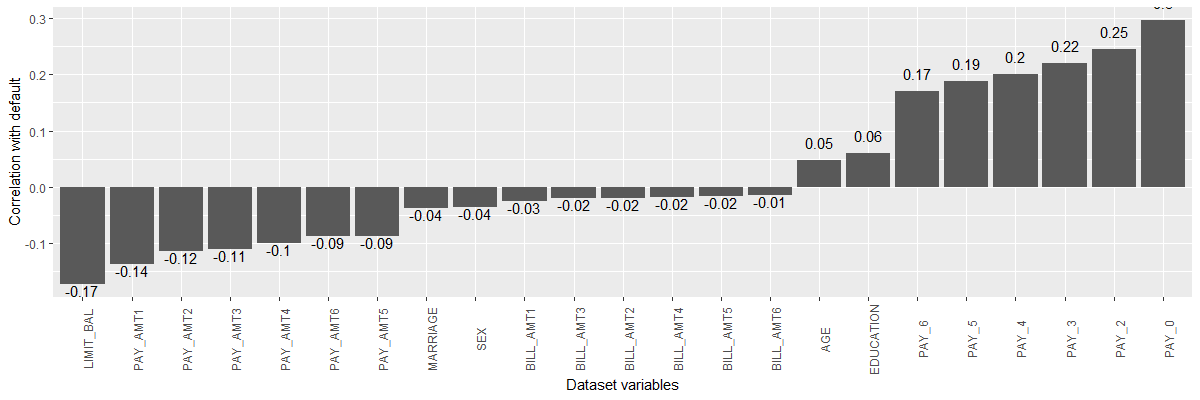
Furthermore, a correlation analysis was performed on the dataset (Figure 4).



*Figure 4: Correlation of variables*

The results show multicollinearity where several variables with strong correlations. This was expected within the family of variables: PAY\_X, PAY\_AMTX and BILL\_AMTX which are history of past payment, amount of bill statement, amount of previous payment for the past X months, respectively.

The correlation between default and any independent variables was further examined (Figure 5) to handle the existence of strong multicollinearity that may influence the model training. The results in Figure 5 reveal no variable with strong correlation with default values but it shows that the default variable has the highest correlation with PAY\_0 (the history repayment of the current month). It might suggest that PAY\_0 is one of the most important predictors.

  
*Figure 5: Correlation scores between default variable and each other variable within the dataset.*

## 3. Modelling

As the desired outcome is a model that can correctly identify if a customer will default (Y/N), a classification model is considered the most appropriate approach to this challenge.

Four classification model approaches were identified for evaluation based on industry standards. These models are:

* Random Forest (RF)
* Support Vector Machine (SVM)
* Logistic regression (GLM)
* Gradient Boosting Machines (GBM)

For all models trained for this project, the provided training dataset has been partitioned into a train set and a test set using stratified sampling methods based on the dependent variable ratio of 8:2, accordingly. The train set has been further prepared for up-sampling and down-sampling to handle unbalance data during modelling process. We split the trainset to 10-folds (9 parts for training and 1 part for validation). ROC/AUC was chosen as the metric for model selection.

### 3.1. Random Forest (RF)

Random forest is one of the most popular machine learning algorithms for classical problems as its algorithm is easy to understand. Random forest also provides good outcome for prediction and the hyperparameter tunning process is simple.

For random forest models, there are several hyperparameters that can be tunned such as mtry (the number of variables randomly selected at each split), nodesize (the minimum size of terminal node) and ntree (the number of trees) (see ref. 1, 2, 3). We started with a default random forest model and then tunned some of the above hyperparameters. For some models, we also removed some unimportant variables (PAY\_AMTX) but it did not improve the performance of these models.

The final model for random forest was chosen with nodesize=5, mtry=8 and ntrees=300 where all the independent variables were included. It is important to note that due to computer capacity, we did not use higher numbers of trees in our models. More details for random forest models can be found in Appendix 4 and the R file.

### 3.2. Support Vector Machine (SVM)

SVM is a powerful algorithm for solving many problems, yet it is not the most effective method in some cases. SVM can be used for both regression and classification problems, and it is especially effective with high dimensional data. However, SVM models do not work well with a large data set and determining an optimal kernel is often challenging (BotBark 2019).

Altering the hyperparameters of the SVM model proved difficult as increasing the cost or introducing polynomial kernel input resulted in the model calculations stalling out.

|  |
| --- |
| Hyperparameters**:** |
| SVM-Type: C-classification |
| SVM-Kernel: linear |
| cost: 1 |
| Number of Support Vectors: 10242 |

*Table 2. SVM Hyperparameters*

### 3.3. Logistic Regression (GLM)

Logistic regression (GLM) is a type of regression analysis and used for binary classification. It is very fast at classifying unknown record; however, it cannot be used on non-linear problems, meaning that it has very little real world use cases (RanjounRout 2020).

Before the model was trained the ‘default’ observations were transformed to 0 (no) and 1 (yes) before the model was trained.

The first GLM model trained was using all variables available from the data set. From the output of this model, another model was trained with only variables that had a p value of 0.05 in the first GLM model. It was observed that the model trained with all variables performed better by having a lower AIC score.

The final GLM model, which was trained with all variables, delivered an AUC score of 0.7104.

### 3.4 Gradient Boosting Machines (GBMs)

Gradient boosting machines (GBMs) are an extremely powerful machine learning algorithm that are widely used in many winning Kaggle competitions. GBMs works well with both categorical and numerical variables and offers a great deal of flexibility for model optimisation with several hyperparameter tuning options. However, GBMs can overemphasize outliers and cause overfitting. Therefore, cross-validation must be implemented in order to neutralise the risk of overfitting (Singh 2018).

For the initial base GBM model, 10-fold cross-validation has been used. The hyperparameters were set as per below.

|  |  |
| --- | --- |
| Hyperparameter | Value |
| interaction.depth | 5 |
| n.trees | 100 |
| shrinkage | 0.1 |
| n.minobsinnode | 20 |
| *Table 3. Hyperparameters\_base GBM* | |

The team has considered various subsampling methods to deal with the unbalanced target class. Although the results were similar between different methods, the GBM model has returned the highest AUC of 0.7998 using the unbalanced data set compared with both down-sampling (0.7905) and up-sampling (0.7965) subsets (figure 7)[[1]](#footnote-1). Figure 8[[2]](#footnote-2) shows that ‘PAY\_0’, ‘AGE’, and ‘LIMIT\_BAL’ were the 3 most important variables in the GBM model.

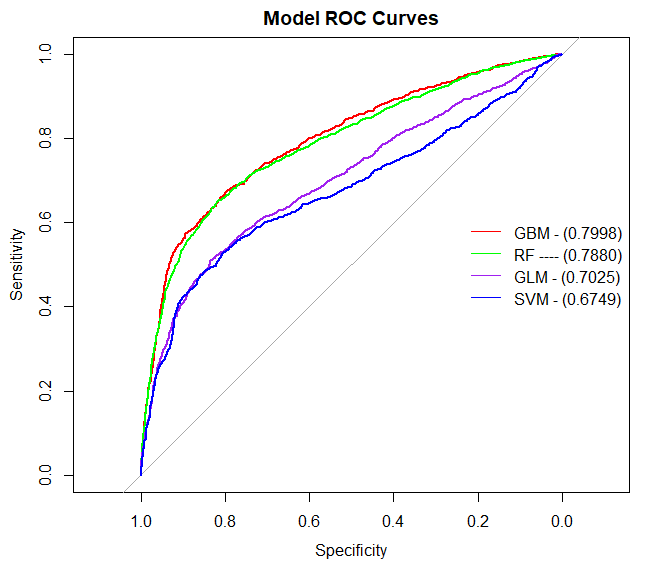
## 4. Modelling Summary

Following is a comparative summary of four models from the Section 3. The following tables summarise key evaluation matrices from all models considered.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Confusion Matrix Scores** | | | |
| **Models** | **Accuracy** | **Precision** | **Recall** | **F1** |
| **GBM** | **0.8137** | **0.7430** | **0.4538** | **0.5635** |
| Random Forest | 0.8076 | 0.7190 | 0.4497 | 0.5533 |
| GLM | 0.61824 | 0.6550459 | 0.2919052 | 0.4038462 |
| SVM | 0.5962138 | 0.7142857 | 0.2248569 | 0.3420398 |
| *Table 4. Confusion Matrix Summary* | | | | |

As can be seen from the *Table 3, GBM and Random Forest models have returned relatively higher overall scores compared with the GLM and SVM models.*

Furthermore, the ROC curves of the models are visualised in the *Figure 9. The plot shows a similar pattern where curves for GBM and Random Forest sit above the GLM and SVM curves.* The area under each ROC curve has been calculated and summarised in the chart which confirms that GBM has returned the highest AUC score amongst the other models.

  
*Figure 9. Model ROC Curves*

## 5. Summary of GBM Model Optimisations

As confirmed in the Section 4, GBM model and Random Forest model performed the best on the test set. Following further attempts to achieve improved Random Forest model performance, the team has concluded that that GMB outperformed the Random Forest model for the given data set. Hence, the team focused on optimising the GBM to acquire the best possible AUC.

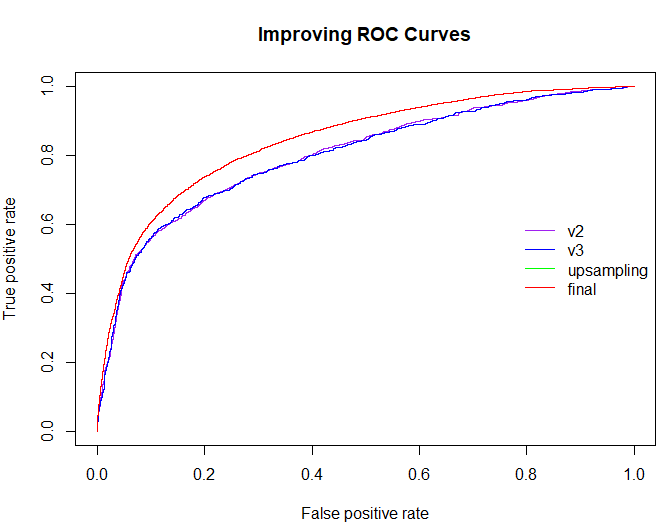
The team initially considered both random search and grid search. However, it was quickly observed that hyperparameters from grid search were delivering a GBM with higher AUC (0.79 and above) compared to AUC form hyperparameters defined by random search (0.75).

Below table is a summary of key hyperparameters and methods used in tunning process. (Grid Search)

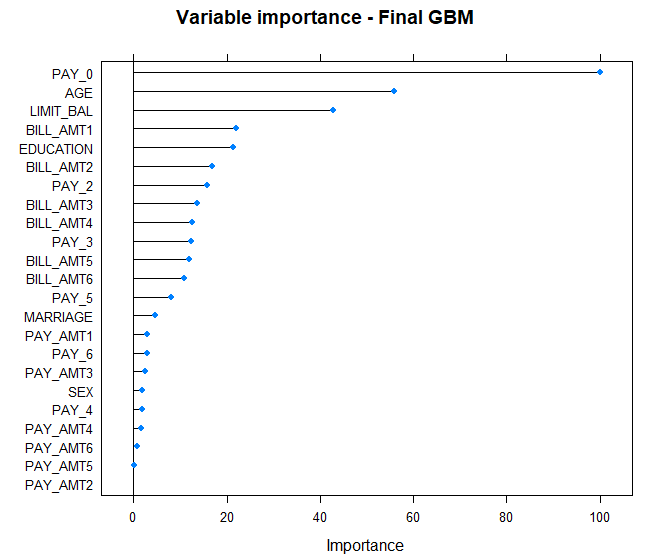
The references in the below correspond to the references in the attached R script file.

|  |  |  |
| --- | --- | --- |
| **Reference** | **Hyperparameters / Methods** | **AUC** |
| Version 2 | interaction.depth = 5, n.tree = 300, shrinkage = 0.05, n.minobsinnode = 15 | 0.8006 |
| Version 3 | interaction.depth = 5, n.tree = 500, shrinkage = 0.05, n.minobsinnode = 5,10 | 0.7997 |
| Up-sampling | interaction.depth = 5, n.trees=350,450,550,650, shrinkage= 0.05, n.minobsinnode=10,15 | 0.8442 |
| Final  (100% of available data) | interaction.depth = 5, n.trees=350,450,550,650, shrinkage= 0.05, n.minobsinnode=10,15 | 0.8442 |
| *Table 6. Optimisation Summary* | | |

The results from *Table 5* are visualised in the graph below. It shows that the AUC scores were relatively lower in the models using single numbers of trees, and the opposite was true for models using a combination of number of trees values. Although the up-sampling method and the final model using all of the available data have both produced the AUC score of 0.8442, a Kaggle submission result was slightly lower using the up-sampling method. This is reflected in the decision of team’s final model.

  
*Figure 10. Optimising ROC Curves*

In addition, the variable importance chart of the team’s final model shows that ‘PAY\_0’, ‘AGE’, and ‘LIMIT\_BAL’ were the 3 most important variables (Figure 11) which are consistent with earlier findings (in Figure 8*)*.

  
*Figure 11. Variable Importance – Final model*

## 6. Deployment to Kaggle

The final GBM model has been used to predict the target classes on the validation data set, ‘*AT2\_credit\_test.csv*’, and validation outputs were exported as csv with prediction probabilities. The team has made a total of 28 successful submissions with the top score of 0.79541.[[3]](#footnote-3)

## 7. Ethical Considerations

Ethical considerations for this project were approached using the Data Ethics Canvas (Appendix 9) framework developed by ODI (Open Data Institute).

## Appendix

### Appendix 1: Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Data Type** | **Column Description** |
| ID | Numeric | ID of each client |
| LIMIT\_BAL | Numeric | Amount of given credit in dollars (includes individual and family/supplementary credit) |
| SEX | Factor | Gender (1=male, 2=female) |
| EDUCATION | Factor | 1=graduate school,  2=university,  3=high school,  0=others,  4=others,  5=unknown,  6=unknown |
| MARRIAGE | Factor | Marital status (1=married, 2=single, 3=divorce, 0=others) |
| AGE | Numeric | Age in years |
| PAY\_X | Numeric | History of past payment.  The measurement scale for the repayment status is:  -2: No consumption;  -1: Paid in full;  0: The use of revolving credit;  1 = payment delay for one month;  2 = payment delay for two months;  . . .;  8 = payment delay for eight months;  9 = payment delay for nine months and above. |
| BILL\_AMTX | Numeric | Amount of bill statement (NT dollar) for past X months |
| PAY\_AMTX | Numeric | Amount of previous payment (NT dollar) for last X months |
| default | Factor | Default payment next month (1=yes, 0=no).  This is the Target variable |

### Appendix 2: Table 1. Variables summary

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Valid Observations** | **Invalid Observations** | **Comments** |
| LIMIT\_BAL | 23082 | 19 | Limit balances of 0 or negative value. |
| SEX | 23097 | 4 | Outside valid options |
| MARRIAGE | 23101 | - |  |
| EDUCATION | 23101 | - |  |
| AGE | 23080 | 21 | AGE > 75 |
| PAY\_X | 23101 | - |  |
| PAY\_AMTX | 23101 | - |  |
| BILL\_AMTX |  |  |  |
| DEFAULT | 23101 | - |  |

### Appendix 3: Statement of contributions of each member

The project team worked collaboratively throughout the project and work was distributed evenly across all members. Team collaboration was facilitated through Microsoft Teams (communication), GitHub (coding) and Office 365 (documentation). Model exploration work was divided between members with each member to explore one model approach. Following initial exploration, some members continued the optimisation of select models while other members contributed to EDA, code revisions and report documentation.

Ivan Cheung – 13975420

* Project management: created a github repo, Team leader
* Codes: data cleaning, SVM, GBM
* Report writing: Introduction, correlations of variables for section 2, Appendices 1, 2 and 3

John Rho – 24509337

* Project management: organised meetings
* Codes: data cleaning, logistic regression, GBM
* Report writing: Section 3.3 Logistic regression, section 5: Optimisation Summary

Dinh Tran – 14382497

* Codes: data cleaning (rewrote some codes to make them shorter), EDA, Random Forest, GBM (working with Ryan to get the final model using Dinh’s initial tunning hyperparameters)
* Report writing: Section 2 (EDA, data preparation), 3.1 Random Forest section, Appendices 3, 4
* Kaggle submissions

Ryan Yeo – 14328254

* Codes: data cleaning, GBM and working with Dinh to get the final model, trained the final model which was used for Kaggle
* Report writing: Sections 3.3 SMV and 3.4 GBM sections, section 4, section 5, section 6. Deployment to Kaggle
* Kaggle submissions
* ODI Data Ethics Canvas

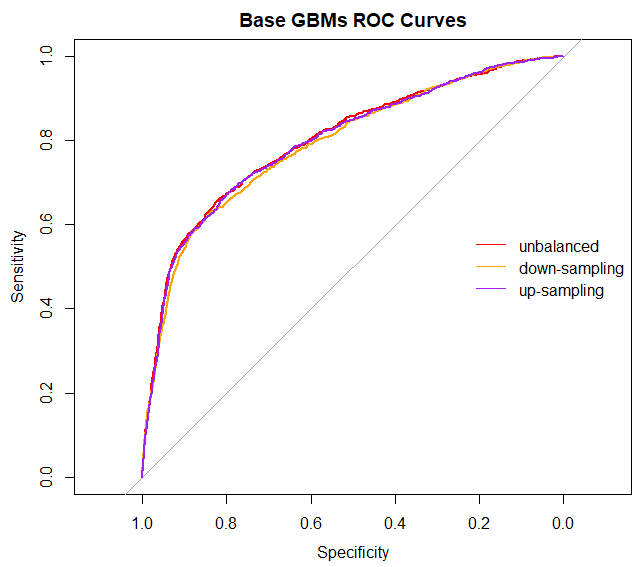
### Appendix 4: Random Forest model selection

We first tried a simple random forest model with ntree being 200 and other hyperparameters are taken as the default values. We obtained an AUC of 0.78544 on the test set.

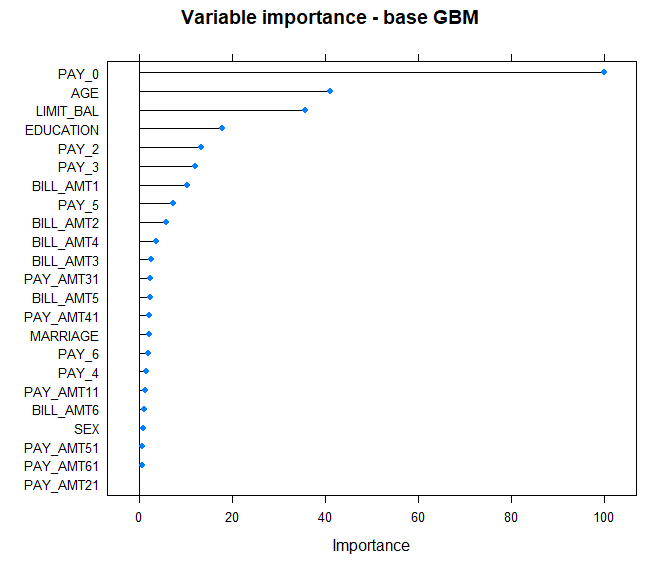
Secondly, we trained a model using the default grid search and obtained some information for variable importance. For this default grid search model, the PAY\_AMTX variables are the least important model, therefore we removed the variables and trained another model using the default grid search. These two models provided us with an AUC of 0.7863344 and 0.7856454, respectively when applying on the test set. This suggested that removing some unimportant variables did not improve the model performance in this case.   
We also trained two models with the number of trees being 200 and 300 and the recommend(default) mtry for the classification problem (see ref. 1, 2). In this case, mtry is taken to be the square root of the number of features in the training data set. By comparing the AUC on the test set, the model with ntree=300 performed better with an AUC of 0.788 compared to 0.786.

Furthermore, we used manual tunning to see if we could train a better model. We chose ntree to be 300, the nodesize to be 1, 5 annd 10. In addition, we tuned mtry in the range from 4 to 9. It is noted that by default nodesize is taken to be 1 for a classification problem. As results, on the test set we obtained AUC= 0.7866337 for nodesize=1 and AUC=0.788013 for nodesize=5. The corresponding mtry for these three models was chosen to be 9, 8 and 6, respectively.

### Appendix 6: figure 7 ROCs – GBM



### Appendix 7: Figure 8. Variable importance chart – basic GBM



### Appendix 8: Submission records – Kaggle

|  |  |  |
| --- | --- | --- |
| **Submission Time** | **Submission Comment** | **Submission Score** |
| 15 hours ago by UTS\_32130\_ryan | version 9 | 0.72241 |
| a day ago by UTS\_32130\_ryan | version 8 | 0.79541 |
| a day ago by UTS\_32130\_ryan | version 7 | 0.79242 |
| a day ago by UTS\_32130\_ryan | version 6 | 0.79526 |
| 2 days ago by UTS\_32130\_ryan | v3 | 0.78749 |
| 2 days ago by UTS\_32130\_ryan | fingers crossed | 0.78847 |
| 2 days ago by UTS\_32130\_ryan | try2 | 0.78936 |
| 2 days ago by UTS\_32130\_ryan | v4 | 0.79113 |
| 2 days ago by UTS\_32130\_ryan | v5 | 0.79235 |
| 5 days ago by UTS\_32130\_ryan | another one | 0.78749 |
| 5 days ago by Dinh Tran | add submission details | 0.78894 |
| 5 days ago by Dinh Tran | add submission details | 0.79088 |
| 8 days ago by Dinh Tran | add submission details | 0.79044 |
| 9 days ago by Dinh Tran | add submission details | 0.78285 |
| 13 days ago by UTS\_32130\_ryan | Please | 0.78525 |
| 13 days ago by UTS\_32130\_ryan | again and again | 0.78553 |
| 13 days ago by UTS\_32130\_ryan | Come on! | 0.7864 |
| 13 days ago by Dinh Tran | add submission details | 0.78322 |
| 13 days ago by Dinh Tran | add submission details | 0.78363 |
| 14 days ago by Dinh Tran | add submission details | 0.77727 |
| 14 days ago by Dinh Tran | add submission details | 0.78301 |
| 15 days ago by Dinh Tran | add submission details | 0.76968 |
| 19 days ago by UTS\_32130\_ryan | gbm\_age\_group | 0.77467 |
| 19 days ago by UTS\_32130\_ryan | without age group | 0.78505 |
| 20 days ago by Dinh Tran | add submission details | 0.78156 |
| 21 days ago by jkr9292 | add submission details | 0.71477 |
| 22 days ago by ivan\_uts\_mdsi | First pass of predictions from SVM using PAY\_X values only in training variables. | 0.67225 |
| 22 days ago by ivan\_uts\_mdsi | Second pass - trained on full df using multiple factors. default ~ LIMIT\_BAL + MARRIAGE + AGE + PAY\_DELAYS + PAY\_0 + PAY\_2 + PAY\_3 + PAY\_4 + PAY\_5 + PAY\_6 | 0.68407 |

### Appendix 9: ODI Data Ethics Canvas

|  |  |  |  |
| --- | --- | --- | --- |
| **Section** | **Comments** | **Actions and decisions** | **Responsible** |
| **1. Data sources** | The data has been sourced by internal company records. | - | - |
| **2. Rights around data sources** | The company holds the intellectual property rights for all data sources. | - | - |
| **3. Limitations in data sources** | The primary data reveals only a subset of financial information regarding an individual.  Customer records are provided as-is by individuals and may not be accurate where details of individuals have changed over time. | User recommendations from this work should seek further financial and personal information from the individuals prior to any action taken. | End User |
| **4. Ethical and legislative context** | The output of this project is for private use only and has no legislative implications.  The recommendation from this work makes assumptions of an individual’s financial capabilities based on limited information. Improper use of this information may have negative impacts on an individuals’ financial wellbeing and community standing (credit scores). | Users following recommendations by this work should have proper training in customer engagement and management. | End User / Department Heads |
| **5. Your reason for using data** | The data was used to train the modelling processes as described in this report. | - | - |
| **6. Positive effects on people** | The recommendations from this work are intended to identify customers at risk of defaulting on credit payments prior to the occurrence. Following up with at-risk customers would allow the business to put strategies in place, such as staggered payment plans to prevent clients from needing to default on payments.  This is expected to improve the financial stability of customers. | - | - |
| **7. Negative effects on people** | There are no expected negative impacts from this work. | - | - |
| **8. Minimising negative impact** | N/A | - | - |
| **9. Engaging with people** | The recommendations of this work will be provided to relevant users by report format on a fixed schedule. | - | - |
| **10. Communicating your purpose** | This work will be communicated with internal stakeholders by this report and follow-up information can be provided in presentations and meetings. | - | - |
| **11. Openness and transparency** | The methodology of this work has been communicated in the body of this report. The project team are open to further explanation of these methods where requested. | - | - |
| **12. Sharing data with others** | The data is not intended to be shared outside of the company. | - | - |
| **13. Ongoing implementation** | The project team will continue to monitor the effectiveness of the recommendations. | - | Project Team |

### Appendix 10: References

1. Boehmke.B & Greenwell. B (2020) Hands-on Machine Learning with R, available at <https://bradleyboehmke.github.io/HOML/>
2. James G., Witten D., Hastie T & Tibshirani (2013) *An Introduction to Statistical Learning with Application in R*, Springer available at <https://www.statlearning.com/>
3. Kuhn M. (2019) The caret Package, available at https://topepo.github.io/caret/index.html
4. Pham D.K. (2018) Caret Practice, available at <https://rpubs.com/phamdinhkhanh/389752>
5. StackExchange <https://stackexchange.com/>
6. Singh, H. (2018) *Understanding Gradient Boosting Machines*, *Medium*. Available at: <https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab> (Accessed: 21 May 2022).
7. BotBark (2019) *Top 5 Advantages and Disadvantages of Support Vector Machine Algorithm*, *Bot Bark*. Available at: <https://botbark.com/2019/12/19/top-5-advantages-and-disadvantages-of-support-vector-machine-algorithm/> (Accessed: 21 May 2022).
8. RanjanRout (2020) *Advantages and Disadvantages of Logistic Regression.* Available at: <https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/>
9. *The Data Ethics Canvas.* Available at: <https://theodi.org/article/the-data-ethics-canvas-2021/> (Accessed May 2022)

1. See Appendix. 6 [↑](#footnote-ref-1)
2. See Appendix. 7 [↑](#footnote-ref-2)
3. Please see Appendix. 8 [↑](#footnote-ref-3)