ASSIGNMENT2A – CREDIT CARD DEFAULT MODEL

Pell Mell Group:

Peter McNamara

Credit Default Report

Introduction

This report will discuss building a model which customers are likely to default on their credit card repayments next month. The data comes from a Taiwan credit card company. A file of historical default data has been provided to act as input to the model training. Approximately 24% of customers have defaulted in the historical data so the business is keen to identify customers likely to default, to better manage their response.

This report will explore the historical data provided and then discuss the predictive models that have been built. Predictions on future customers defaulting will be loaded to the Kaggle website for evaluation.

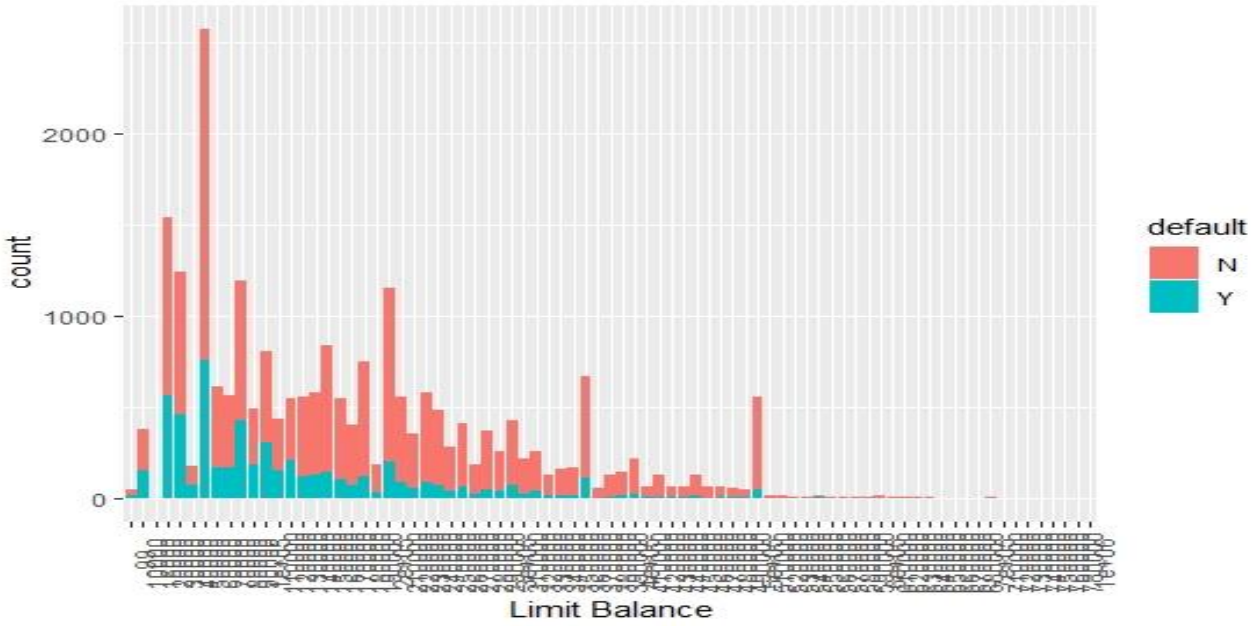
Exploratory Data Analysis

A file of historical default data has been provided, to be used for training the predictive models. This data set has 23,098 observations of 17 variables. The 10 numeric Payment and Statement Amount columns have had Principal Component Analysis applied to them. This is a statistical technique to convert columns that may be correlated to uncorrelated principal components. For modelling purposes, the columns can then be used beside one another without needing to cater for relationships between the variables.

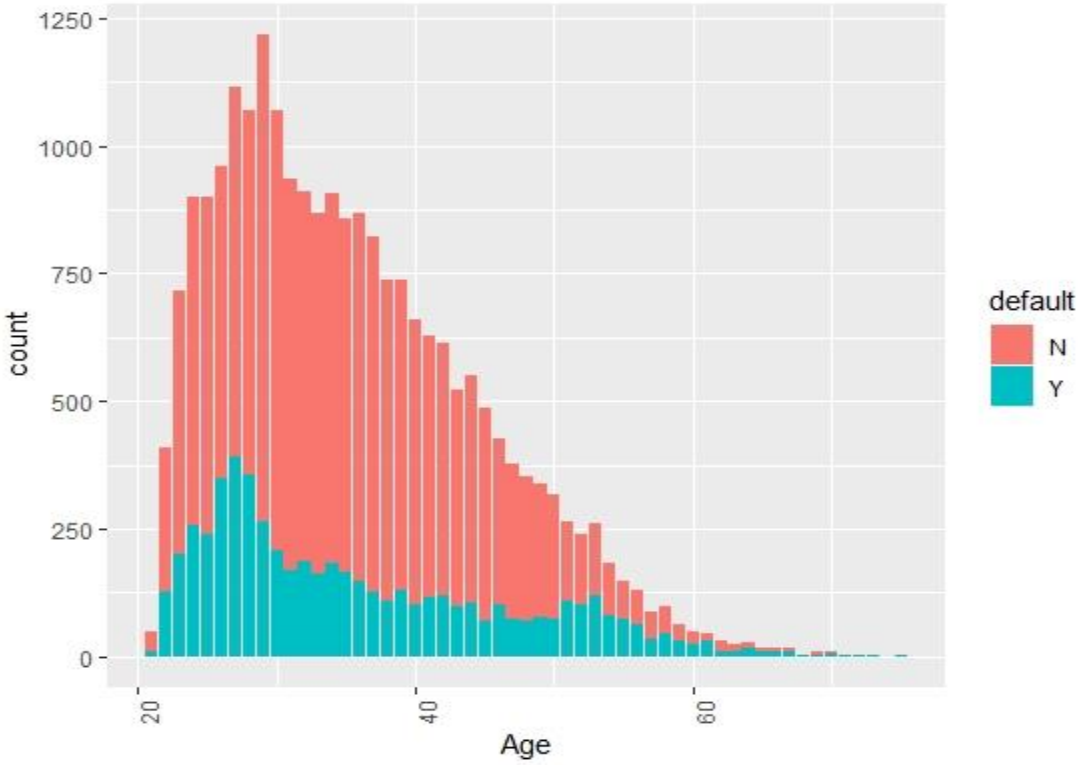
The ID column is a customer identifier that can be ignored for modelling as it does not. The default column contains whether the customer has defaulted on their credit card, 'Y' being that they have defaulted and 'N' they have not defaulted. Approximately 24% of customers have defaulted.

Delving into the LIMIT_BAL variable, there are 50 borrowers with a balance limit of -99. Intuitively, a credit card should not have a negative credit balance. However, a credit card can be used as a debit card, so a negative balance may signify the cardholder is having a personal credit balance on the card. This feature requires clarification from the domain expert.

The plot below shows default rates by limit balance and demonstrates that there are high occurrences of defaults in the lower bands. For balances up to 120,000 NT the default rate is 34% and up to 40% for 10,000 NT. However, even though there are limited counts on defaults in the higher balance bands, the absolute amounts of non-repayment of these higher bands carry more significant amounts than the lower bands (e.g., a default of 1 Million NT default bears a much larger financial loss to the lender than 10 occurrences of 10,000 NT defaults).



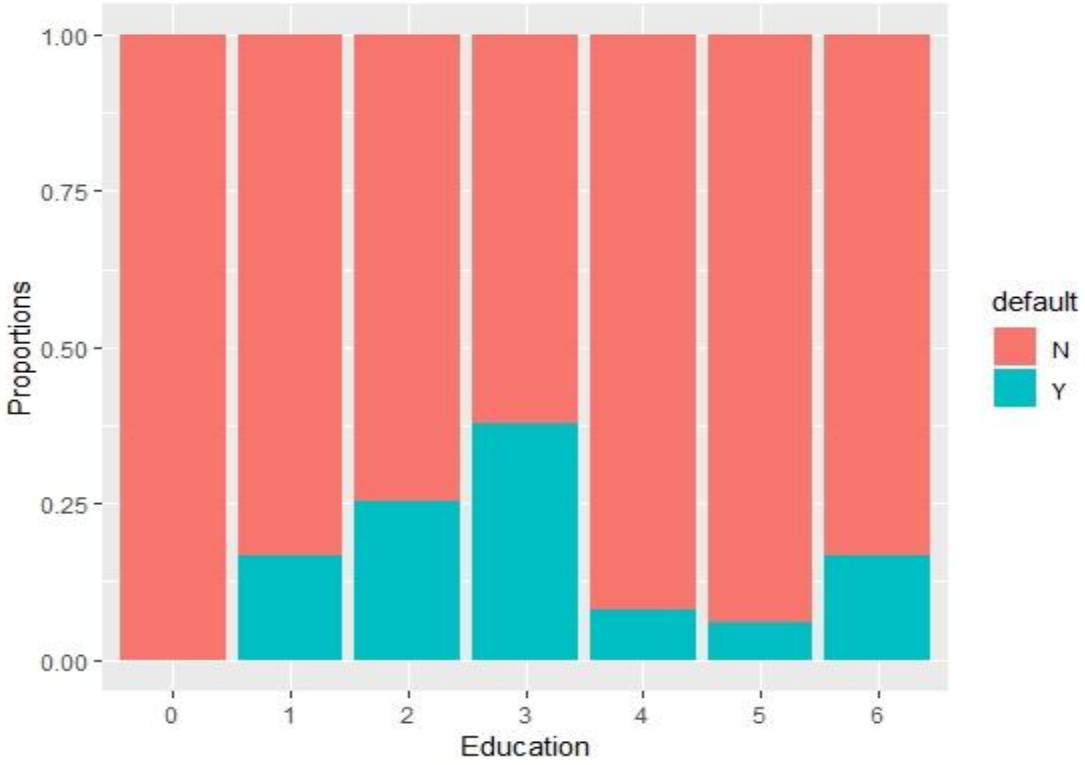
For Age, there seems to be a data entry error which results in 50 borrowers whose ages are over a hundred. Ages span from 21 to 75 and then from 128 to 141 which is unrealistic. Ignoring for now the values above 100, the graph below shows the default rates per age under 100. Default rates are highest amongst customers in their 20's or over 50.



There are also 3 categorical variables in the dataset for Sex, Marriage and Education. Gender is split relatively evenly between Male and Female for both overall counts and default rates. However, there are 3 observations with errors in this column, having values of cat, dog and dolphin.

Marriage has an extra categorical value of 0, that is not in the data dictionary for the dataset. For other values, borrowers split relatively evenly between married and single. There is only a small number of "others" in the marital status. All three classes have similar default rates – there is no such thing as "financial transmitted disease" for married couples in this training dataset.

Finally, for Education there are 2 'unknown' classes and again a small number of records that have a '0' category. The unknown and other classes have few respondents. High school leavers have the highest default rate of 38% compared to graduate school of 17%, as shown in the graph below that shows percentage of default rates per education category.



Data Cleaning

Several columns in the training data had unexpected values. For each of these, a decision had to be made to how to treat these probable erroneous values.

Age ranges from 21 to 75 and then there are a small number of entries between 128 and 141 which is incorrect for an Age field. It was decided that these were input errors where a 1 was added onto the front, so 100 was subtracted from the age in these cases to bring them back to between 31 to 45.

For Sex, three entries had values of 'cat', 'dog' or 'dolphin' rather than '1' (male) or '2' (female). All observations were defaulting, so without sacrificing the data of these entries, these incorrect values were given a value of 0 (unknown) so that the other variables could be used.

Without more information on the limit_bal column, it is difficult to determine whether a value of -99 is valid or not. Therefore, it has been left as is, as this field is an ordinal value.

Finally, Marriage and Education columns have '0' values that are not specified in the data dictionary. Education also has 2 unknown categories (4 and 5). To simplify the model, 0, 4 and 5 were combined into an unknown category for Education and 0 and 3 were combined into an unknown category for Marriage.

Modelling

From the team's deliberations, a decision was made to pursue three popular and powerful statistical modelling techniques: Random Forest (RF), Gradient Boost Machines (GBM) and Support Vector Machines (SVM), to achieve the goals of the project.

1. **Random Forest** (Lantz, 2015)

RF focuses on ensembles of decision trees, combines the principles of bagging with random feature selection to add diversity to the decision tree models. After the ensemble of trees is generated, the model uses a vote to combine the trees' predictions. Since the ensemble uses only a small, random portion of the full feature set, random forests can handle extremely large datasets and avoid the so-called "curse of dimensionality" which might cause other models to fail, but RF's error rates for most learning tasks are on par with other statistical learning methods.

2. **Gradient Boost Machines** (Walia, 2018)

GBM is another famous ensemble learning technique in which it learns in a sequential process so that each subsequent model improves the previous one. GBM boosts the performance of a simple base-learner by iteratively shifting the focus towards problematic training observations that are difficult to predict. Once the information from the previous model is fed to the next model, every new tree added to the mix will do better than the previous tree because it will learn from the mistakes of the previous models and try not to repeat the mistakes. Therefore, this technique will eventually convert a weak learner to a strong learner which is better and more accurate in generalization for unseen test data.

3. **Support Vector Machines** (Lantz, 2015 and James, 2017)

SVMs embodies an extremely powerful algorithm enabling it to model highly complex relationships for both regression and classification settings. The goal of SVM is to create a flat boundary called a hyperplane, which divides the space to create homogeneous partitions on either side.

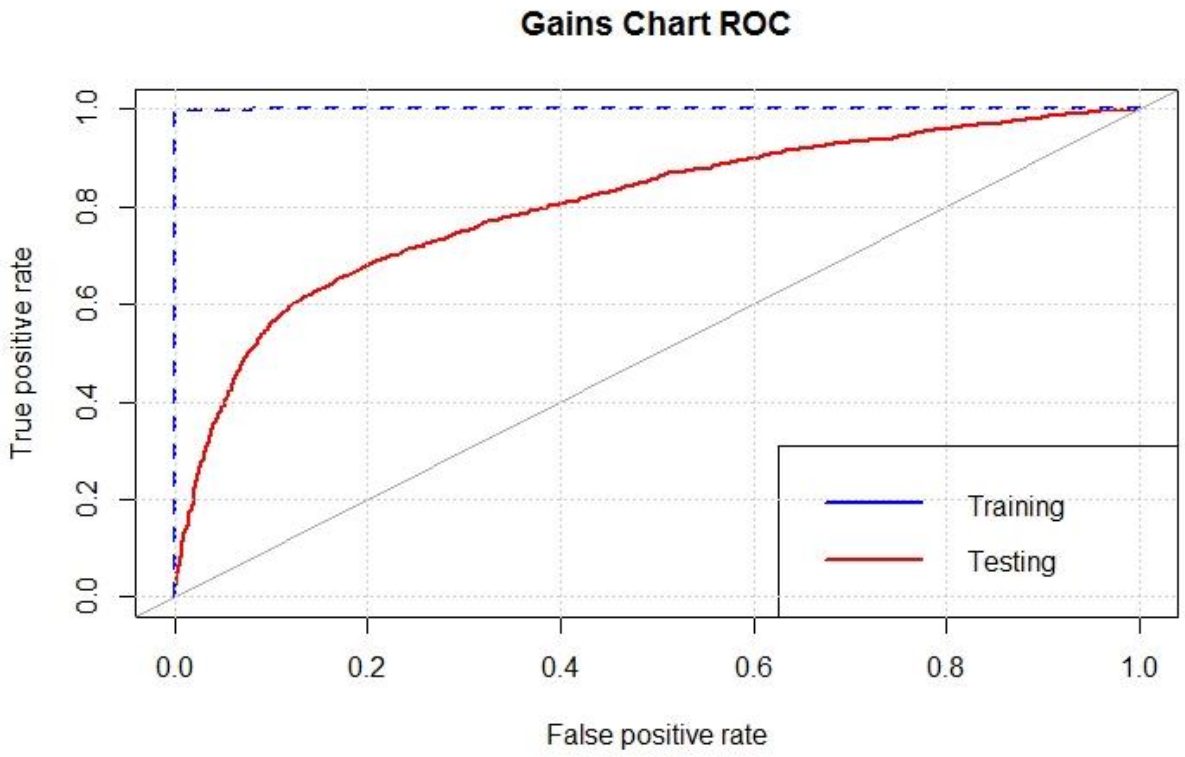
Evaluation

The training data supplied was split into 80% for training and 20% to test the model against. This was repeated several times against random splits to check that the models gave consistent results against varying input data.

To evaluate each model, the Area Under the Receiver Operating Characteristic Curve (AUC) was used. This is a measure of the true positive rate versus the false positive rate for the testing set. The closer the AUC percentage is to 100%, the better the model is at predicting for the testing set. The table below shows the average AUC for each model, as well as the Accuracy and the actual and predicted Positive rate.

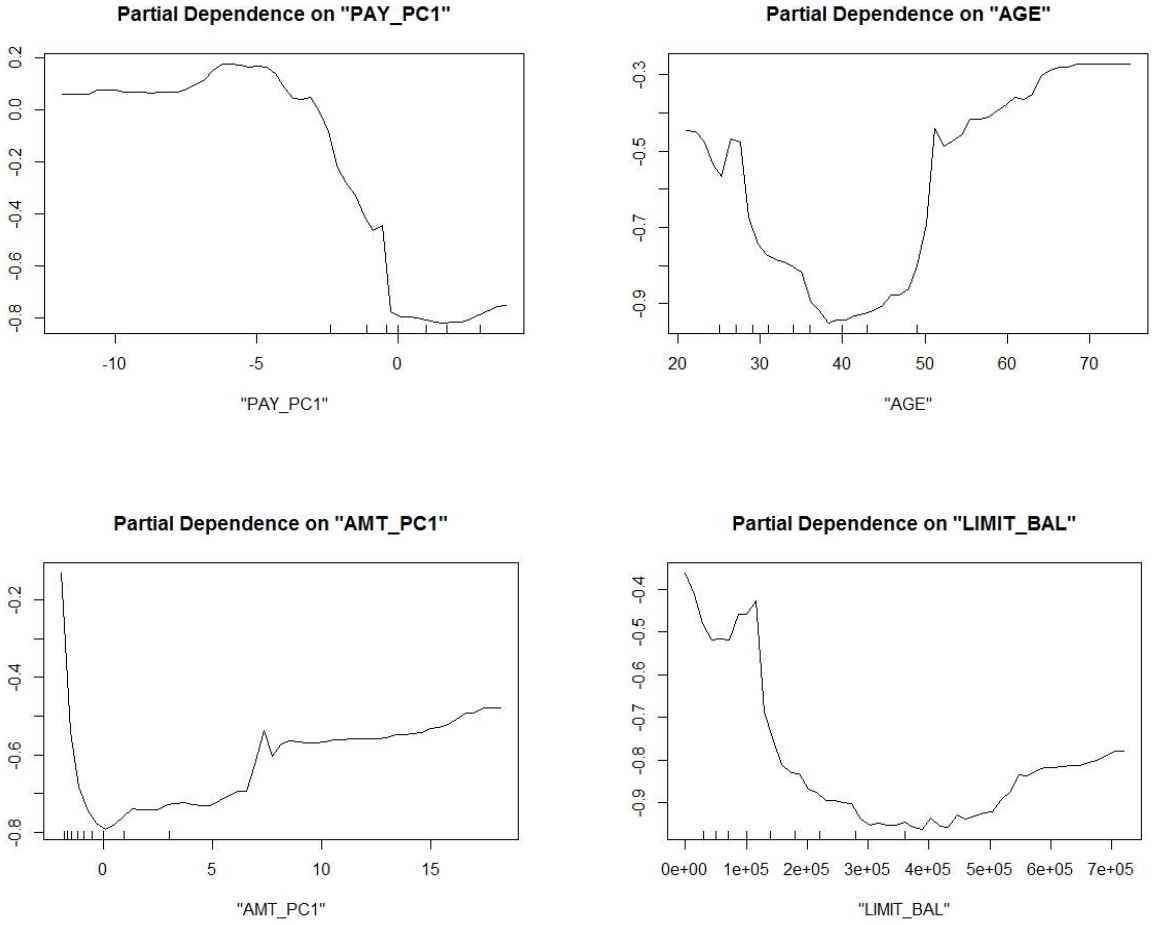
Model	*AUC (%)*	*Precision (%)*	*Recall (%)*	*f1 (%)*	*Accuracy*	*Actual Rate*	*Predicted Rate*
Random Forest	80.45	55.70	64.80	59.84	79.01	24.12	28.13
GBM	78.75	57.27	62.67	59.74	79.40	24.34	26.7
SVM	75.81	50.09	59.34	54.22	75.91	24.0	28.58

This shows that Random forest did the best, with over 80% AUC, whilst GBM was close behind with approximately 79%. The following graph shows the AUC for a training iteration for the Random Forest algorithm. The red line is the held out 20% of unseen test data run against the model.



The Random Forest algorithm allows for analysis of the variables that were most important to predictions. Investigating this shows the Pay_PC1, Age, Amt_PC1 and Limit_Bal were the 4 most important variables using the Mean Decrease in Accuracy measure provided. Below are the partial dependency graphs for these variables. These show how the prediction is influenced by each variable in isolation. For Age, for example, as noted in the Exploratory Data Analysis section customers are more likely to default in their 20's or over 50. This can be seen below in the J curve. The balance limit partial dependency plot validates that limits below 120,000 NT are more likely to default.

Pay_PC1 was the most important variable under random forest and for GBM. The graph below shows a marked drop off as the value goes from -5 to 0, so anything below this value has a strong default prediction and those above 0 tend to not default.



Final Predictions Lodged in Kaggle

The Random Forest technique was then used to generate predictions on the supplied test dataset. To build the model, the full training data set was used as input. Predictions were in line with preliminary training above with an AUC of 80.9%. The predicted default rate using the optimal percentage threshold was 23%, again in line with testing results.

Ethical Issues

The team subscribes to the Mission, Principles and Code of Professional Conduct of the Data Science Association in promoting data science to improve life, business and government. The guiding principles of the Association are:

- Setting standards for the ethical professional practice of data science.
- Assuring base-level data scientist competency.
- Advancing data science to serve core values of the scientific method and noblesse oblige.

- Helping to shape a better future - not just for the powerful, but also for the majority of people.

The major conflict of interest in this credit card project between the lender and the borrowers are **commerciality** vs **personal convenience**. It is our duty, as data scientists, to explicate this conflict to both parties that the lender must act with due care in exercise its responsible lending obligations. Yet at the same time, the lender must reject credit card applications, if necessary, and to educate customers about taking personal responsibility in management their financial affairs.

As witnessed from the above analysis, 24% credit card default is unacceptably high, which brings about adverse social and economic impacts on all ages of credit card users. Data scientists, as a growing professional group, should aim in making positive contributions to improve the credit card industry in whatever ways they can.

It should also be noted that correlation does not imply causation. The model is predicting which customers will default next month but this does not necessarily mean those customers will default. There are many more factors that influence a customers' ability to repay their debts that are outside the realm of this model. In validation of the model, the predicted default rate was broadly in line with the 24% actual rate. However, the Precision of the model (percentage of true positives compared to total predicted positives) was quite low at 55%, meaning that 45% of the predicted customers to default in testing did not default.

The predicted results can be used to highlight those customers in danger of defaulting. The financial institution can then proactively implement strategies for monitoring and intervening positively to reduce the default rate.

References

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Sheare and Rüdiger Wirth 2000, ***CRISP-DM 1.0 Step-by-step data mining guide***, CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA), and OHRA Verzekeringen en Bank Groep B.V. (The Netherlands)

Data Science Association, ***Code of Professional Conduct***, (http://www.datascienceassn.org/code-ofconduct.html) <accessed 1 October 2018>

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani 2017, ***An Introduction to Statistical Learning with Applications in R,*** Springer New York Heidelberg Dordrecht London

Brett Lantz 2015, ***Machine Learning with R Second Edition,*** *Packt Publishing Ltd.*

Anish Singh Walia 2018, ***Gradient boosting in R,*** https://datascienceplus.com/gradient-boosting-in-r/ <accessed 30 September 2018>