Creditworthiness of Customers

Group 1 – Ryan Fritton, Vasishtha Mangavelli, Rajat Mishra, and Sanjay Ramji

28 April 2020

## Executive Summary

To better align the company objective of increasing overall profits by reducing the likelihood of defaults, we have undertaken a project to create useful models that will predict whether a customer is going to default. In order to support this objective, we have created two models. The first model uses a logistic approach to predict a customer’s creditworthiness by utilizing know variables. The second model uses the same dataset as the first model, but this model takes the random forest approach, which is a higher fidelity version of decision trees. If we can be successful in predicting customers that will default, we can successfully selloff those accounts before they default or avoid similar customer profiles in the future.

## Background / Context

Domain

The domain of this business analysis project is determined by the business domain of the client company. In this case, the client is a consumer finance company that offers various credit products and services to customers with a wide range of creditworthiness. More specifically, the analysis pertains to decisions of interest impacting the company’s consumer credit card line of business. Accordingly, the dataset that the analysis will be based on includes information that is relevant to the domain. As discussed below, the decisions of interest are also highly relevant to the field.

Brief Description of the Scenario

A consumer finance company would like to increase the profitability and reduce the default risk within its consumer credit card line of business. The company has decided to leverage predictive business analysis techniques to generate information from data on hand to help identify opportunities for risk reduction and increased profit. The strategy being utilized to enhance profitability is to identify the most credit-worthy customers and market to them higher profit type products. The strategy to reduce the risk of default is to identify those customers that are indicated to have a high risk of default and sell them off to reduce the risk in their portfolio.

Decisions of Interest

The decisions of interest in the analysis are guided by the overall business objective and its supporting strategies. In particular, the decisions of interest must, therefore, help the business determine who the most credit-worthy customers are, and who the least credit-worthy customers are. The main decision of interest in both the models utilized is whether the customer is going to default or not default. The models can classify the customers based on a multitude of independent; therefore, it will be a classification model.

The purpose of the decision variable is whether a customer will default or not would be to classify consumers as either being credit-worthy or not-credit worthy. That is, whether the consumer finance company would like to keep the customers or whether they would like to sell the customers off, as part of its strategy. This decision would be based on predictive analysis that would identify whether the consumer presents a risk of default and would likely be a binary dependent variable (Yes =1, No = 0). If the consumer is predicted to default (Yes =1), then the business would likely decide to sell that customer off.

In those cases where the predictive model determines that the customer is not likely to be at risk of default, the business would probably decide to retain the customer. For this class of customers, stemming from the highest-level decision, further analysis would then be conducted to inform even more granular-level decisions of interest. These more granular decisions could include whether to offer a lower interest rate, whether to increase the credit line, whether or not the customer would be credit-worthy for a term loan of the credit product.

The decision of interest would help the consumer credit finance company determine which actions or offers to target to which customers. The population of credit-worthy customers is not a homogenous group, as they vary in terms of demographic profile, amount of credit borrowed, size of their payments, and their history of paying on time. Based on variations in these aspects, it follows that the strategies for increasing profitability would vary accordingly. Offers that are better suited to a class of candidates should be offered to only that type and not others. Doing so could either result in sub-optimal profit generation or increased portfolio risk. This outcome would, therefore, result in the same scenario the business is in today and would not have solved the stated business problem.

 Decision Makers

Post the data mining processes, the business objectives, and the individual results will be analyzed by the Financial and Business Operations Executives who are the key decision-makers for this process. The sales executives can use this information for target marketing to maximize their revenue. In addition to the executives, who will have the final say, other decision-makers will be the domain experts who will verify that the data and model make sense and that it is useful. The domain experts' input will be the primary data utilized by the executives to determine if the models are implemented. The purpose of the domain expert in the decision-maker model is not only to make sure the model makes sense, but also that we do not waste the executive decision-makers time with proposals that do not add value.

## Business Understanding

Business Objective

For our data analytics project, the business objective is two-fold. First, it is our objective to use the data and analytical tools to identify individuals whom we can sell more products to, which could be additional credit, new credit cards, or credit promotions to increase their spending. The second objective is to identify high-risk individuals who present enough risk where it is more beneficial to sell their accounts than continue to hold that risk. Once this criterion has been determined, and a prediction model can effectively project these high-risk individuals, we can input this model to avoid providing credit to these high-risk individuals in the future. Essentially, the business opportunity is to either increase the value derived from profitable customers or reduce risk by jettisoning our less profitable and riskier customer.

By implanting the predictive models in the future, we believe that we will increase the profitability of the company while lowering the risk at the same time. Effectively, we will have a model that determines which individuals we may be able to sell products to due to a low risk of default. At the same time, we can identify high-risk customers or potential customers to avoid, which will increase our profit margin while lowering our risk.

Situation Assessment

In the current situation, we are looking to increase profitability within a lending company by reducing risk and increasing profitability. Currently, the company believes there is too much undue risk of default that the company is presently absorbing, but there is not a model in place to assist with reducing this risk. Right now, we cannot predict who will be good customers and who will be less than ideal customers. As a result of this project, the company will be able to reduce risk and increase profitability.

Data Mining Goals

As far as our specific data mining goals, the overarching goal is obviously to create a predictive model that supports the objective of reducing risk and increasing profitability, but there are more specific goals to get to this end. Our first goal is to segment the data and like for relationships and correlations. Understanding what relationships may exist is valuable because these can be confirmed with the domain experts to determine if there is a real relationship.

Our second data mining goal is to be able to classify the population into defaulters and non-defaulters based on the independent variables. From there, our model will be able to predict the defaulters and non-defaulters that we had previously classified. The main objective from a data mining perspective is to be able to predict this for the end business goal of reducing risk and increasing profitability.

## Data Understanding

Data requirement

The essential requirement of the data is to be high-quality, complete, relevant, and accurate enough to successfully predict the outcome of whether the customer is creditworthy enough for us to target them for additional products or sell off their accounts. The data must be useful about the customers, too, so we can apply our decision-making technique and get the desired result. To create a useful model, we should at least know the following things about customers;

* credit limit
* gender
* education level
* marital status
* age
* previous payment records
* credit utilization.

Another requirement of the data is that it needs to represent an accurate representation of the population. If the data had come from only individuals who had defaulted on their credit, then predicting future default would be impossible because the model would treat every attribute as a predictor of default, which would be a useless model in practice. This is very important for our business goals because a diverse population will allow our model to accurately predict defaulters and assist in reducing our risk and increasing profitability.

We also need to ensure we have enough data to create a model confidently. Assuming we use ten times the number of variables as a guide to determining whether the data set is large enough, we would need 240 rows of data to be able to create an accurate model. In this case, we have 30,000 rows of data, so there is more than enough data to not only create a model but to validate and test that model, as well. The ability to validate and test the model may be just as important as any other requirement of the data.

Describe Data

The data set we are utilizing contains 24 variables, 23 variables being explanatory variables to base our prediction model on and one decision variable to which will essentially be our output of our model. The decision variable (defaulters/non-defaulters) will be binary (Yes =1, No = 0). This approach can be utilized a second time to alter the model to determine if we should sell off accounts, as well. Below is a breakout of the explanatory variables utilized in this data set.

X1: Amount of the given credit (contains both individual consumer credit and his/her family (supplementary) credit.)

 X2: Gender (1 = male; 2 = female).

 X3: Education (1 = Graduate school; 2 = University; 3 = High school; 4 = others).

 X4: Marital status (1 = Married; 2 = Single; 3 = others).

 X5: Age (year).

 X11 – X6: History of past payment. Repayment status from April to September. If paid duly then assigned -1 otherwise for every month delay assigned given number 1,2, 3….

 X17-X12: Monthly Bill statement generated from April to September (credit utilization)

 X23-X18: Amount paid for the given bill from April to September.

Sources

 The data set we are utilizing to create our model is from the UCI machine learning repository located [here](https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients). The data is based on research into the default payments in Taiwan and was studied to create predictive arguments to estimate the probability of default in the future. The data was originally compiled by I-Cheng Yeh, and it is from the Department of Information Management, Chung Hua University, Taiwan and the Department of Civil Engineering, Tamkang University, Taiwan.

Data Quality

The given dataset extracted from the UCI Machine learning repository has 30000 records, 23 predictors, and one target variable. We have one unique identifier (ID), which is set to auto-increment as new records are added. We can see that there are no duplicates in the ID column, and there are no null values in the dataset, which gives an understanding that the quality of the data seems to be pretty good, clean, and straightforward. This can help create strong predictive models to achieve the business objectives.

This data appears to be of a wide variety of a population, which will result in a higher quality distribution. Since we have good variability, completeness of the data, and no duplicates, it appears we have an excellent quality data set to create a predictive model and to test our model against.

## Data Preparation

Data Selection

While it is best to use the smallest set of data possible to create an accurate model, we have chosen to use the entire dataset as described above. We determined that the data was already a sample size, and we did not want to dilute the data further than the previously created sample. While 30,000 responses may seem significant for a dataset, it is a tiny number of individuals compared to the millions of customers our company currently has.

Data Cleaning

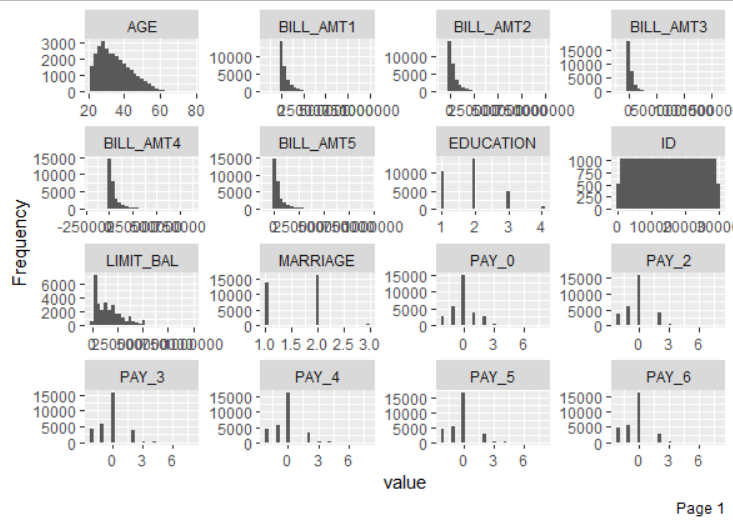
When preparing the data, the first thing we do is explore the data by running a simple View function. This confirms the data has properly been loaded. From there, we need to start cleaning the data. After viewing the data, we can see there is an unnecessary second header column that can be removed, which we remove. Next, we can see that the “default payment next month” column header may cause a problem in R, so we change the title to “default\_payment.” Then we utilize the sapply function to ensure the data is all in the same format, and we don’t have any characteristic data in the file that may cause it to be unreadable for our logistic and random forest models.

Finally, we must remove the incorrect responses from the Education and Marriage variables. Based on the data description, Education should only have responses 1, 2, 3, or 4, but some 0s, 5s, and 6s are showing up in the data.

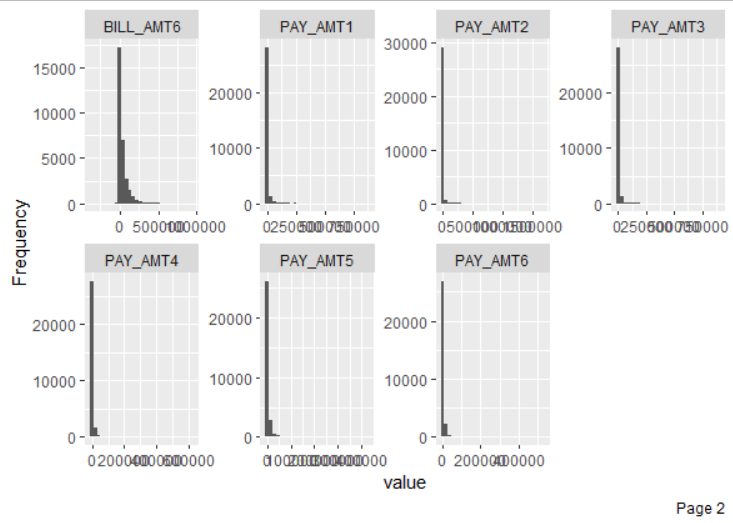
Since 4 is “other,” we place all the non-conforming responses in the 4 categories. Marriage has a similar problem where there should only be responses of 1, 2, or 3, but some 0s are in the data. We transform these 0s to 3s, which is “other.”

Prepare Data

After cleaning the data, there are still a lot of variables to sift through, so it is valuable to look at these closer. Some of the variables are shown in the below visualizations that are also within the attached PowerPoint. These visualizations capture some of the relationships between variables and how certain variables relate to the outcome variable (defaulters). These visualizations help interpret the data in means that it is easy to understand. As shown in the plots below, histograms of 23 predictor variables were generated to provide an overall view of the frequency distribution of the predictor variables.

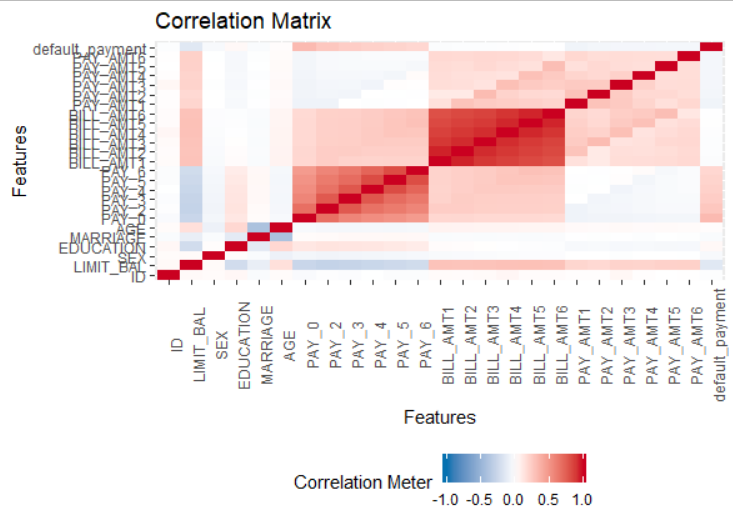


*Figure 1a*



*Figure 1b*

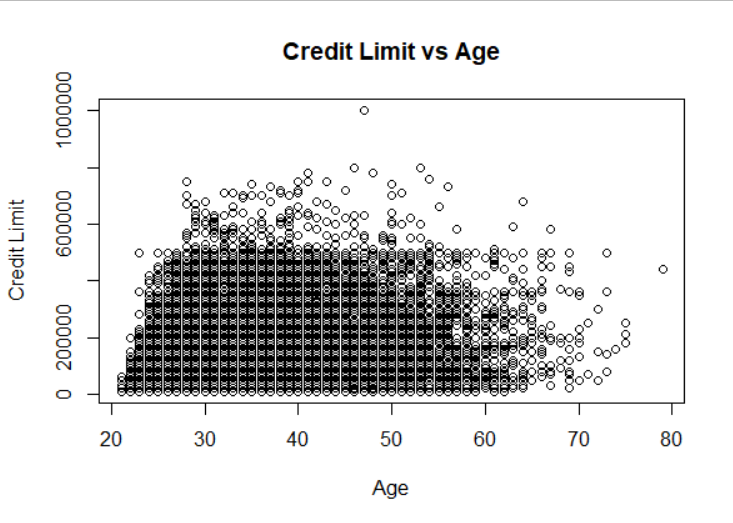
Having gained a sense of the frequency distribution of the variables, it is helpful to further look into what correlations exist among the predictor variables. The heatmap generated in Figure 2 provides an easy and informative view off the degree of correlation within the dataset’s variables.



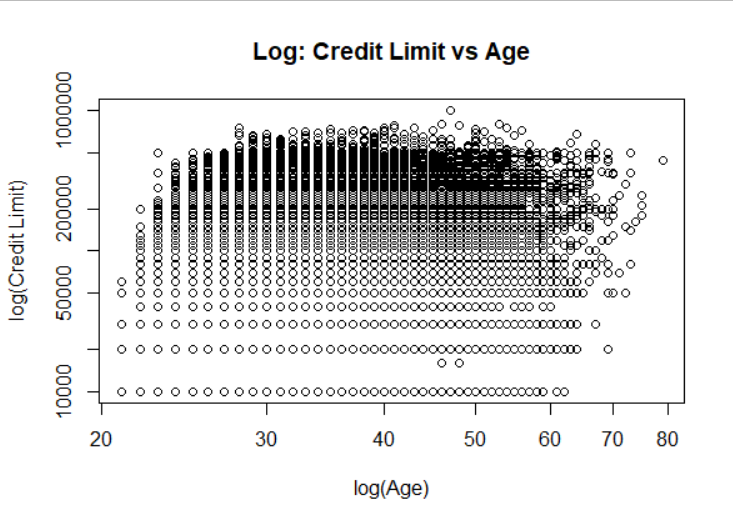
*Figure 2*

Based on the view in the heatmap, we observe strong correlations among the payment statuses and among the bill amounts. This suggests a strong likelihood of multicollinearity across these predictor variables, and informs the variable selection decision.

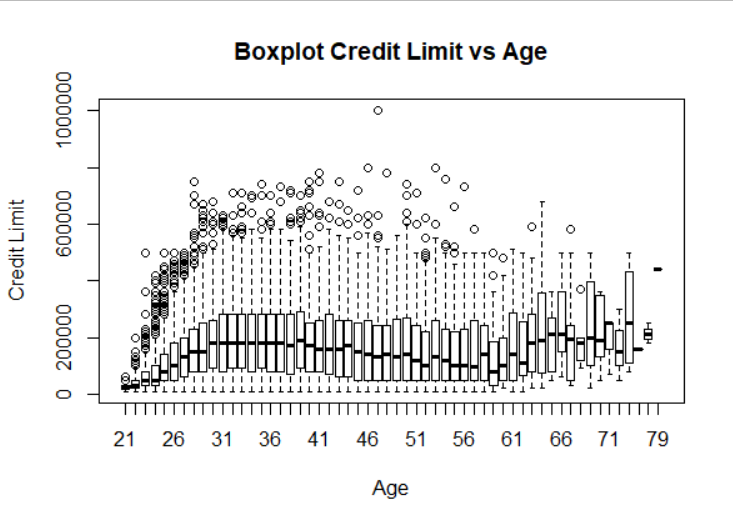
Proceeding with data visualization, we plot LIMIT\_BAL, which is the Credit Limit, against Age to examine any general pattern. The plot in Figure 3 shows that higher amount of credit given is correlated with older ages. However, this visualization is quite congested between Credit Limit values between 0-50,000. To better visualize this interval, we plot the log of the two variables, as shown in Figure 4. From this plot we see more clearly that the increase in Credit Limit flattens out at around Age 30. The highest density of Credit Limit above 20,000 seems to occur above age 40.



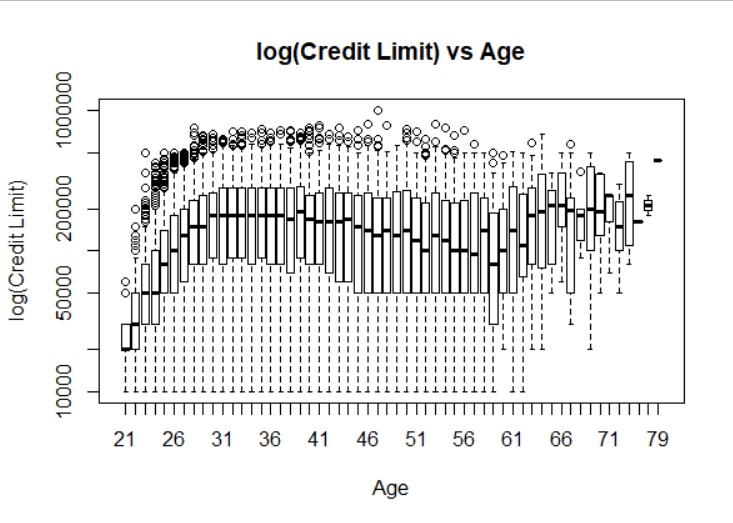
*Figure3*

*Figure 4*

To get a better idea of the averages and interquartile ranges of Credit Limit vs Age, we use the boxplots in Figures 5 and 6. Figure 5 shows Credit Limit on a standard scale where we see that our observations from the scatter plot were correct; around Age 30 the highest outliers and average Credit Limit values begin to flatten out. In Figure 6 we also use the log-scaled Credit Limit to get a better look at the means and interquartile ranges. We observe that the mean Credit Limit is highest between approximately ages 30 and 45.

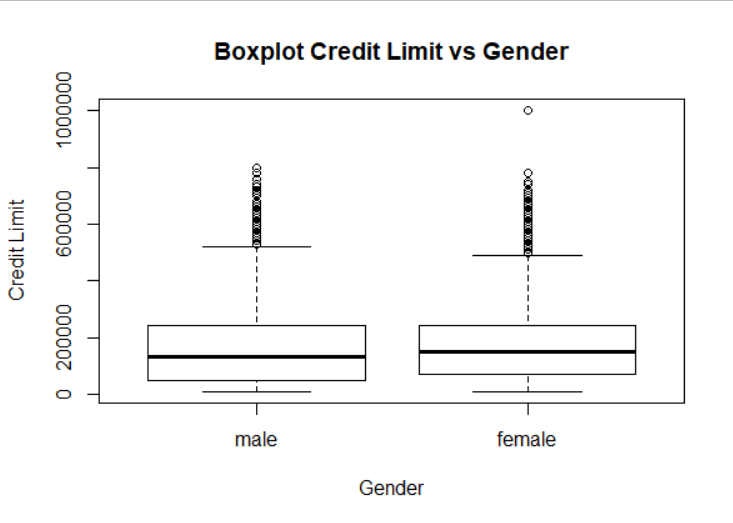


*Figure 5*

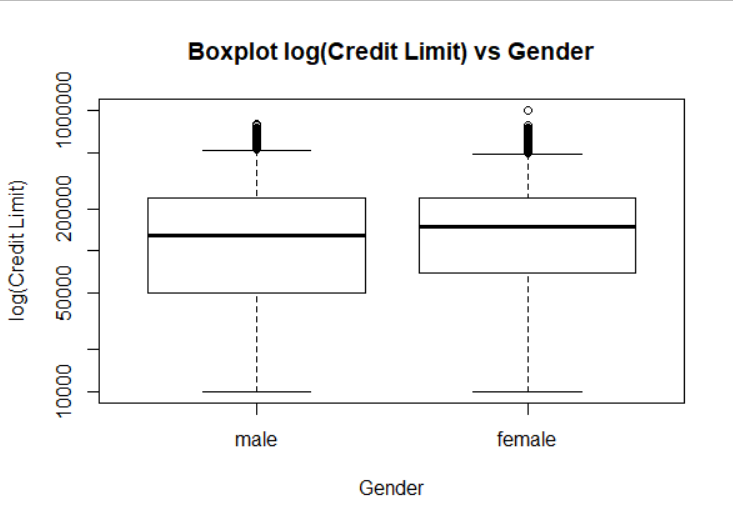


*Figure 6*

Another characteristic to examine Credit Limit against is the gender of the borrowers using the variable SEX. Figures 7 and 8 show two versions of boxplots of Credit Limit vs Gender. Figure 5 shows that the highest Credit Limit is correlated with female borrowers and that the mean and interquartile ranges are both higher for females. Figure 8 uses the log of Credit Limit to again better view the mean and ranges. We observe that there is a wider interquartile range for males, who also have a lower 1st-quartile value, whereas the interquartile range for females is more concentrated around the mean.

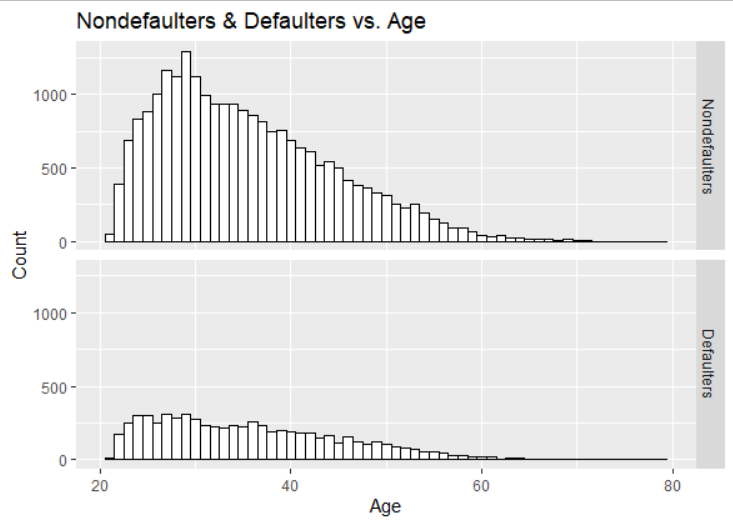


*Figure 7*



*Figure 8*

Having explored the relationship between Credit Limit against two key demographic characteristics, we now turn our attention towards visualizing some key characteristics of Non-defaulters and Defaulters. We generate histograms for Non-defaulters and Defaulters against Age. We see from these plots that there are more Non-defaulters than Defaulters. This is encouraging from the standpoint of the business problem we are solving since that means we can expect a greater share of our customers to be worthy of receiving offers. Looking at Age, we can see that the greatest density of customers is between approximately 25 and 45 years old.



*Figure 9*

There are many different pair-wise plots that can be undertaken given the large number of predictor variables in the dataset. However, a fundamental principal of data mining is to determine which variables are most useful to the business problem, by determining which ones have the biggest impact on the outcome variable. To limit the scope of the analysis, we will consider the visualization undertaken so far to be enough to move ahead with modeling.

One additional item we did address prior to creating any models, was the removal of Bill Amount variables except the first BILL\_AMT1. The logic behind this was to remove the chance of increased multicollinearity in the model since the amounts add much of the same data to the prediction, which would increase the error of any model.

## Modeling

Describe Data

The first thing necessary for the data utilized in the modeling aspect of this project is to split the data into training and validation sets. If we used the data as is, the prediction model created would not be advantageous because it would predict results that were in the model itself. To be able to test our model, we have split the data into 60 percent (18,000 instances) for the training data and 40 percent (12,000 instances) for the validation data.

Starting with the logistic model, we utilize all the independent variables as the visualizations above help us understand the data. From here, we can look at the information provided to determine which variables have the most reliable predictive power. The lowest p-values indicate that the variable is very relevant in the log equation. Below are the most relevant variables to run a log regression:

|  |  |
| --- | --- |
| **Variable** | **Pr(>|z|)** |
| LIMIT\_BAL | 0.000266 |
| EDUCATION | 0.000386 |
| MARRIAGE | 0.0000000000465 |
| PAY\_0 | 0.0000000000000002 |
| BILL\_AMT1 | 0.0000185 |
| PAY\_AMT1 | 0.0000167 |
| PAY\_AMT2 | 0.0000740 |

Based on the above values, the log equation could run utilizing these values and probably be as accurate or more accurate than using all the values due to reducing redundancies and potential multicollinearities. We have reduced the variables down to this small list for ease of use and to avoid overfitting going forward.

Decision Making Models

For this project, we have gone with two distinct models. The first model is a standard logistic regression used to predict defaulters and non-defaulters. This is a very reliable model that can have substantial predictive probabilities if the right variables are chosen. The second model is the random forest models. The random forest model is a decision tree model, but it has a higher fidelity than other decision trees without much extra work. These are two different models we can use to predict defaulters and non-defaulters to decrease risk and increase profitability within the company.

Rationale on Model Choice

When deciding on what model to utilize, it is necessary to think of the long-term usability, ease of use, and the accuracy of the model. Choosing to use a logistic regression to build a model was a relatively easy choice.

Logistic regressions are highly popular and powerful classification methods. Since they are highly popular, it means the end user is familiar with the jargon associated with a logistic regression, along with the general idea of how to create and use a logistic regression. Since our model is to be deployed in the company to potential users who may not be sound using R or other types of software, we thought it was valuable to utilize a method most users would be familiar with.

While a basic regression tree or classification tree was discussed as a possible model choice, a random forest model improves predictions versus the first two approaches. Random forests models combine the predictions/classifications from the individual trees to obtain improved predictions. This model is more out of the norm than a logistic regression, but it does have the potential to be a more accurate model, and that is why this model was chosen.

Model Development and Output

Once the data has been prepared, and the models have been chosen, the model development is simple. We have already reduced the logistic regression down to just seven variables. In order to create the model, those seven variables are run through the glm function in R against the training data, as described within the attached R code file. Once we have the logistic regression, this is run against the validation (label test in the R code) data to determine precisely how accurate the model is at predicting a defaulter. The actual regression that results and the accuracy associated with that regression is shown below:

Logistic Regression Prediction:

p= 1 .

1+e^-(-1.447+LIMIT\_BAL\*-.1914 + MARRIAGE\*-.1196 + PAY\_0\*.7769 + BILL\_AMT1\*-.0871 + PAY\_AMT1\*-.2479 + PAY\_AMT2\*-2.418)

Confusion Matrix:

0 1

0 9107 1978

1 253 662

Accuracy:

0.8140833

Random Forest

Once the data has been prepared, and the models have been chosen, the model development is simple. We have taken all the predictors to build our model. Random Forest consists of a large number of individual decision trees that operate as an [ensemble](https://en.wikipedia.org/wiki/Ensemble_learning). Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model’s prediction. It also ranks individual predictors as per their importance in the model. In order to create the model, all variables are run through the random forest function in R against the training data, as described within the attached R code file. Once the model is complete, the below output from R shows the accuracy of the random forest model.

Confusion Matrix:

0 1

0 8901 1671

1 492 936

Accuracy:

0.8198

DSM Evaluation

Recommendation

Based on the review of both models above, the random forest model appears to be more accurate when run against the validation data than the logistic regression model. Since it is paramount that we are accurately identifying those individuals who are going to default, the more accurate model is the model that must be implemented going forward. If we can pick with almost 82 percent accuracy the accounts that are going to default in the future, this is going to be a significant increase in profits due to the reduced risk the company will be undertaken. This model also can predict those who will no default and allow the company to continue to market to those individuals until they show up on the potential default list if they become overstretched.

## Using the 82 percent accuracy rating we can make some assumptions on how much potential money we can save the company by reducing risk. Running the potential defaulters through an 82 percent accuracy rating, the company could potentially save $1,310,610,018 Taiwanese, which converts to approximately $43,250,130 in US currency. This is the amount that the model is able to accurately predict will be defaulted, and the potential savings that is out there by implementing this model. While it is highly unlikely we will be able to recover all of these funds, the company may be able to see these accounts off before they default for 75 percent of their value, which is going to be significantly higher than the company would get after the account defaults. This is potentially millions of dollars of added profit to the company’s bottom line.

Limitations

There are some obvious limitations with the logistic regression approach we have chosen to implement. First, it is not a fail prove model, and it does have a large number of false positives and false negatives over the 12,000-item validation set. If we implement this model over a million customers, we are going to have approximately 200,000 false positives and false negatives. This will lead to the company either selling off or closing accounts that are predicted to default, but would not actually default, or you could be providing additional credit to individuals who are predicted not to default but will end up defaulting. Both of these are realities in any predictive model, but a 20 percent failure rate seems much more significant when we increase the number to potentially 1 million customers.

Enhancements

Models can always be improved, and the models we have created here are no different. For the Logistical regression model, future iterations could reduce the number of predictor variables even further to minimize any potential multicollinearity. For the random forests model, a different number than of trees (500 trees in this model) to determine if there is a more efficient number. The risk with this improvement is that more than 500 trees will require additional computing power and time to process.

While these were the models chosen for the project, another improvement could be to try a different method entirely. A neural network model may be able to assist in this prediction very well, but the processing power and time required for the neural network to process 30,000 responses with 24 variables each may not be practical.

Closing

In closing, this project has demonstrated that a significant amount of data can be cleaned and prepared to create models that do perform reasonably well at predicting an outcome variable (defaulters). These models can assist this company in improving its profitability by reducing its risk of defaulting customers. While it can be argued that approximately 82 percent accuracy is not high enough, it cannot be argued that 82 percent accuracy at predicting the defaulters is much higher than if just random chance were applied to the data. We recommend we implement this model, which will result in an increase in net profit for the company and, therefore, this model should be implemented as soon as possible.