SI 618: Analysis of Credit Card Default

Motivation

Credit card defaults are one of the most significant issues that happen on a daily basis in commercial banks. My father has been working as a credit risk manager in a bank for over 20 years. What I heard from him motivated me to analyze the underlying factors that can be used to determine the exposure of credit risk. With these factors, I hope to predict what clients are more likely to default.

There are four research questions I look into:

1. What types of people like to use credit cards?
2. What characteristics of a client affect the amount of credit extended to him/her?
3. How does delay of payments differ by the characteristics of the clients?
4. How well can we predict the credit card default based on characteristics of the clients and his/her credit history?

Data Sources

I downloaded the data set stored in excel file from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#).
It is a data set of customers' default payments in Taiwan, including 30,000 instances containing 23 attributes and 1 response: amount of the given credit, gender, education, marital status, age, history of past payments from April to September in 2015, amount of bill statement from April to September in 2015, amount of previous payment from April to September in 2015, default payment next month.

Among the 24 explanatory variables, 9 are qualitative variables and 14 are quantitative variables. The response variable is qualitative.

The amount of the given credit (New Taiwan dollar) includes both the individual consumer credit and his/her family (supplementary) credit, falling into the range from 10000 to 1000000. For the education variable, 1 refers to graduate school, 2 refers to university, 3 refers to high school and 4 refers to others. Age ranges from 21 to 60 for 99% of our observations. History of past payments from April to September in 2015 has a measurement scale, with -1 representing paying duly, and from 1 to 9 representing delaying payments from 1 to 9 months correspondingly. The amount of bill statement from April to September in 2015, and the amount of previous payment from April to September in 2015 each contains 5 quantitative variables corresponding to 5 different months. Default payment next month is a binary variable with value 1 meaning default and 0 meaning non-default.

Methods and Analysis

Data Pre-processing

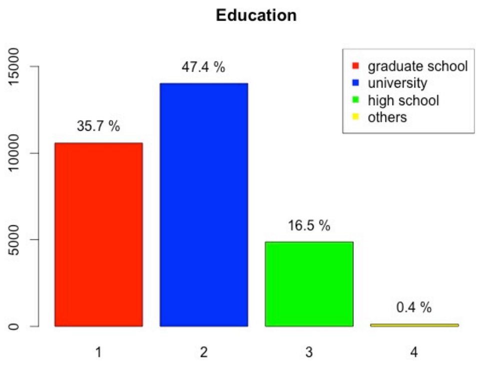
I eliminated instances with values that do not fit into the specified range. This includes instances that have values 0, 5 and 6 for education as well as 0 for marital status. After removal, our observations dropped from 30,000 to 29,601.

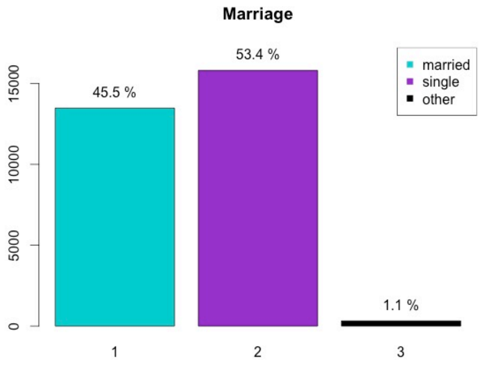
In addition, I replaced the 6 variables for history of past payments with its sum as a predictor. I also created 2 new predictors using respectively the average of the 6 variables for the amount of bill statements, and the average of the 6 variables for the amount of previous payments. In this way, we can reduce the number of variables to only 9 by taking the sum or average of the records from the past six months, allowing us to keep condensed yet thorough information just as the original data set. These 3 newly created variables could reflect the individuals' financial status and credibility.

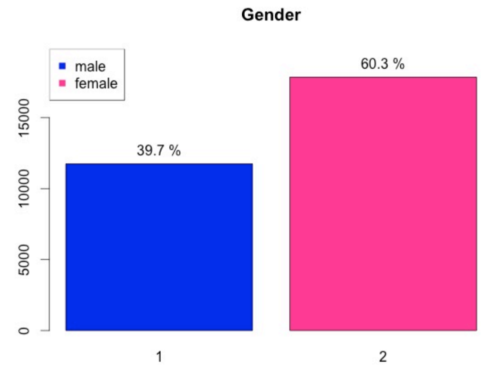
Analysis

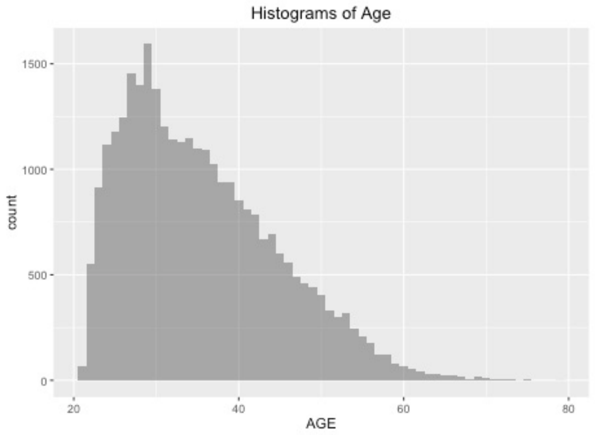
Question 1:

To investigate the types of people that use credit cards in Taiwan, I looked at the characteristics variables of the clients, including gender, education, marriage, and age. I renamed the levels of factor variables gender, education and marriage, changing the numeric labels to character labels. To see the population makeup of the credit card clients, I created three pie charts and a histogram.







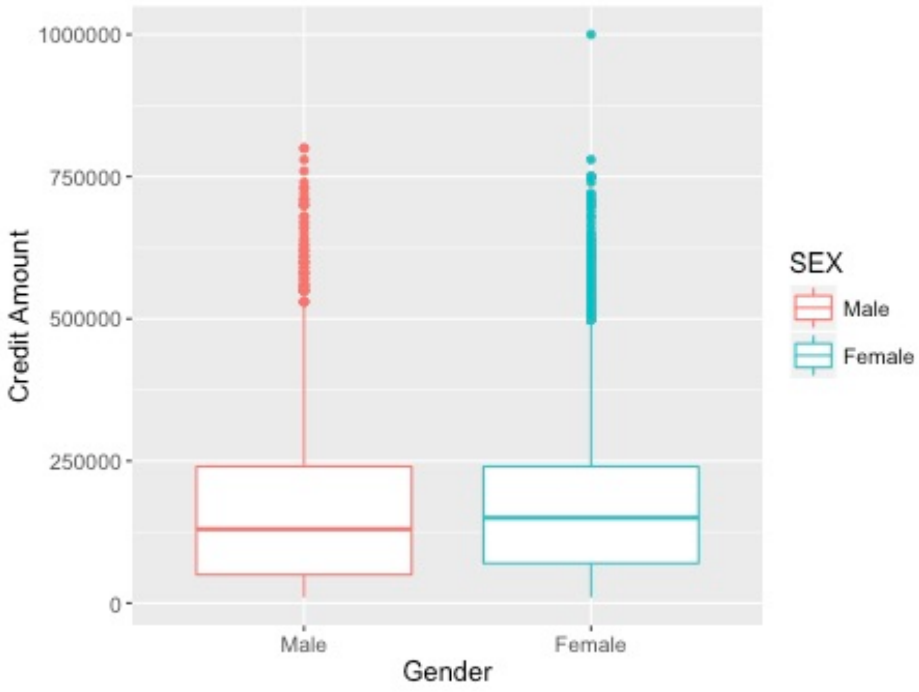


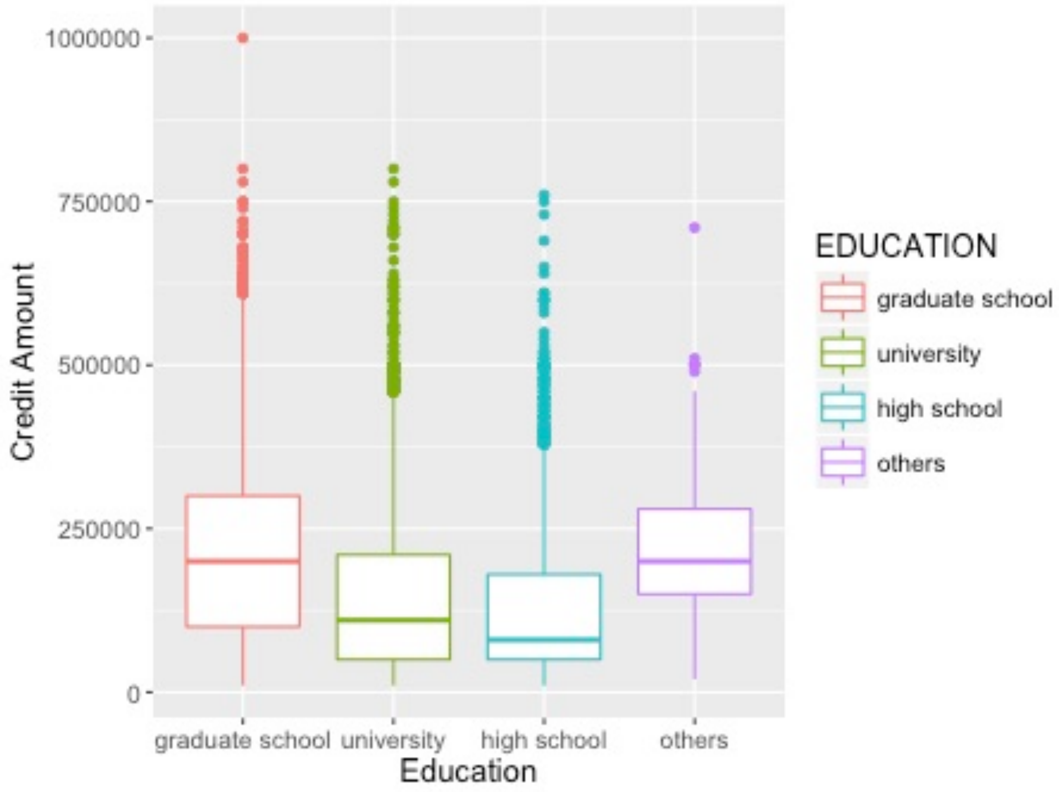
Some important findings are:

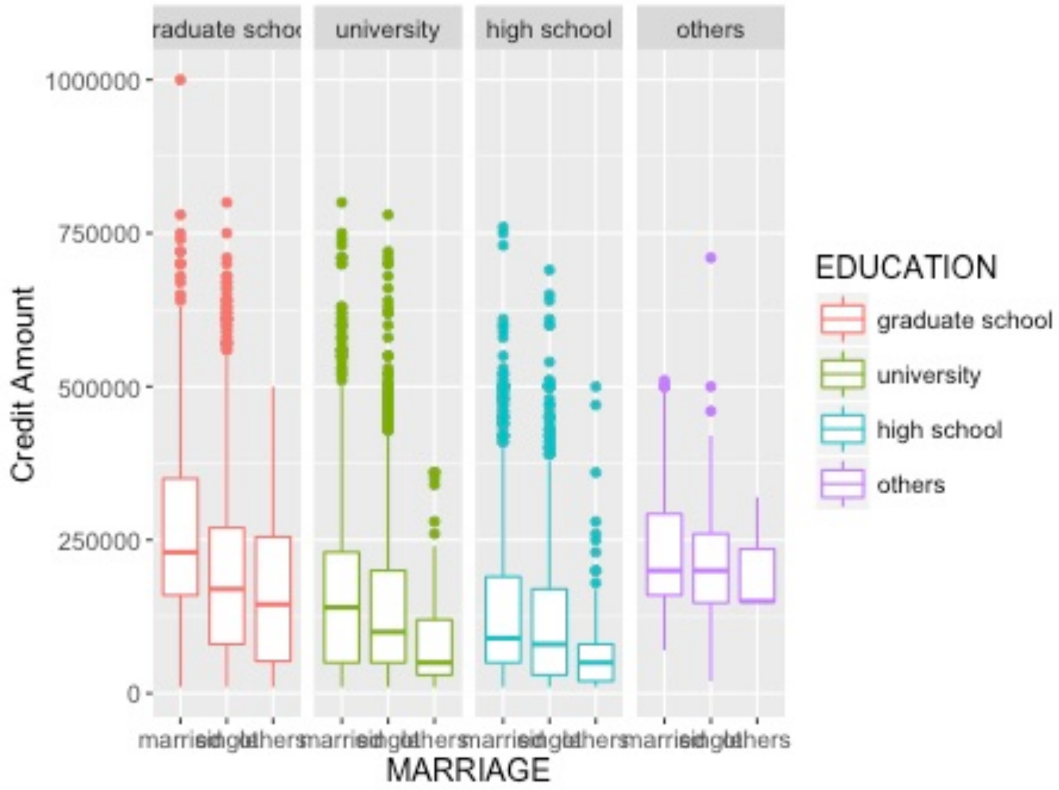
1. Most clients are between 20 and 50 years old, with a vast majority being in their 20s and 30s;
2. There are 20% more male clients than female clients;
3. 35.7% of the clients have graduate degrees, and 47.4% have college degrees. These proportions are much higher than the population proportions;
4. There are 8% more single clients than married clients.

Question 2:

I first plotted extended credits versus every single characteristic variable and retained the ones that show interesting patterns or evident differences across different categories. Then I plotted given credits versus each pair of variables and see whether the clients can be further divided into subgroups.

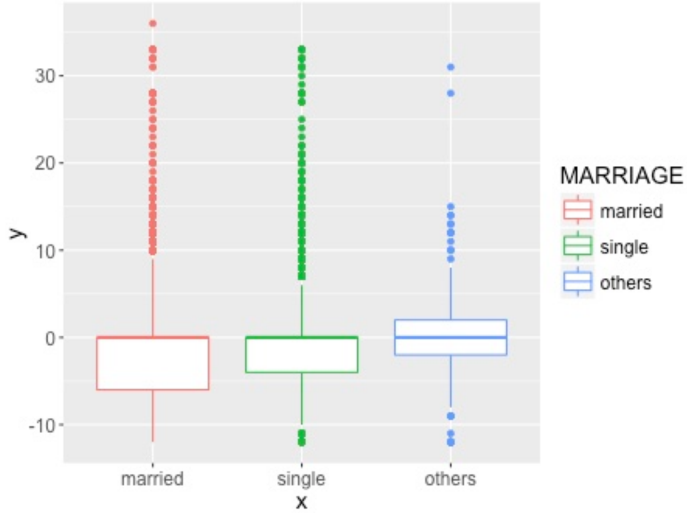


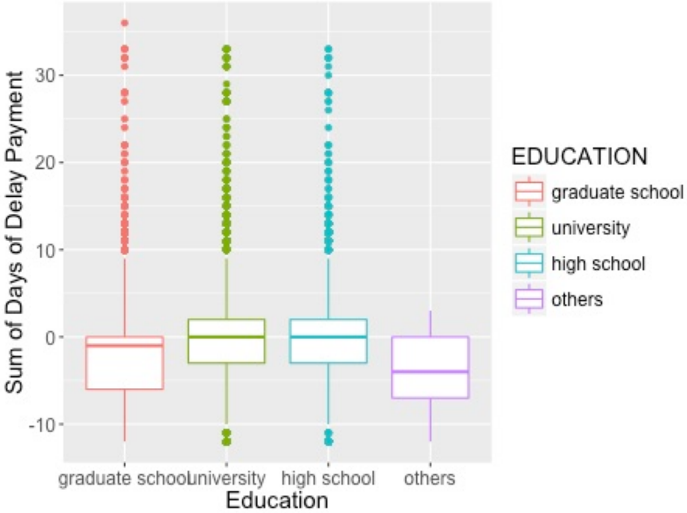

Credit Amount
1000000
750000
500000
250000
0
graduate school
university
high school
others
Education
EDUCATION
graduate school
university
high school
others


raduate schoo
university
high school
others
Credit Amount
1000000
750000
500000
250000
0
married
single
others
married
single
others
married
single
others
married
single
others
MARRIAGE
EDUCATION
graduate school
university
high school
others

I found that extended credits do not differ by gender. This is a bit unexpected, because of the fact that, on average, men earn higher income than women in Taiwan. Another noteworthy pattern is that, people with higher degrees were given more credits according to the median statistics. However, we need more data to make further conclusion, because income, for example, might be a confounding variable here. In addition, married clients were given more credit than single clients with regards to the median in every subcategory of education. In summary, we can safely infer that, married people with higher levels of education can have higher extended credits.

Question 3:
I plotted sum of days of delay payment versus every characteristics variable. Then when making a more complex plot with faceting, I didn't find them quite useful because the statistics look homogeneous in different subgroups.





In these figures, positive numbers represent delay, whereas negative numbers represent early repayment. Judging from the median statistics, marriage or education does not affect whether clients would repay credits on time. Clients with graduate degrees tended to do slightly better at paying back credits on time than others.

Question 4:
To predict whether clients would default on their credits, I used logistics regression and random forest models. Logistic regression can construct a linear model for both numeric and categorical variables and make few assumptions about the data. I did variable selection to ensure higher predictive power. As to random forest, it has the advantages of automatic variable selection and interaction consideration. It also works for both numeric and categorical variables.

I used the following criteria to evaluate the results:

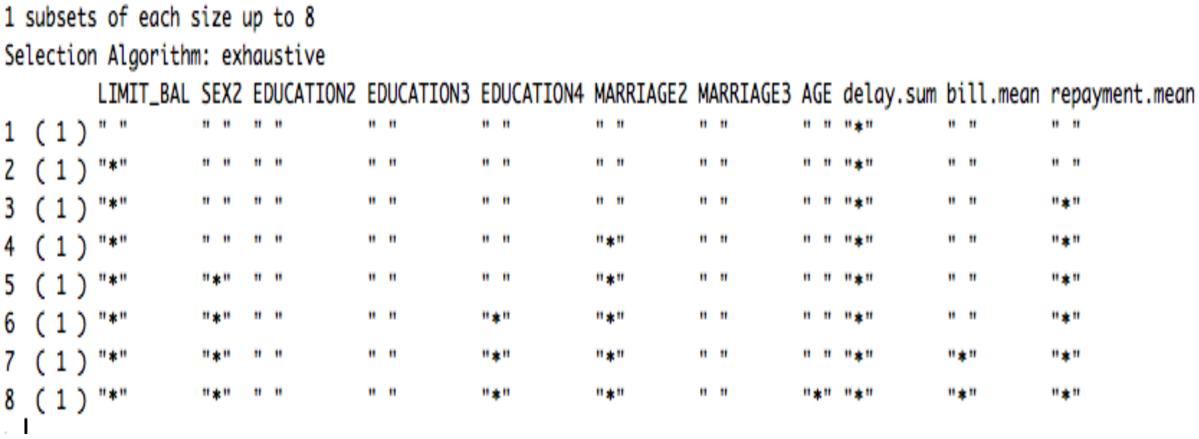
$$1.\,\text{Total Error Rate (TER)} = \frac{\text{False Positive} + \text{False Negative}}{Size\ of\ Testing\ Data}$$

$$= \frac{No.\ of\ non-default\ classified\ as\ default + No.\ of\ default\ classified\ as\ non-default}{No.of\ individuals\ in\ testing\ data\ set}$$

$$2.\,\text{False Negative Rate (FNR)} = \frac{\text{False Negative}}{Condition\ Positive} = \frac{No.\ of\ default\ classified\ as\ non{-}default}{No.\ of\ default}$$

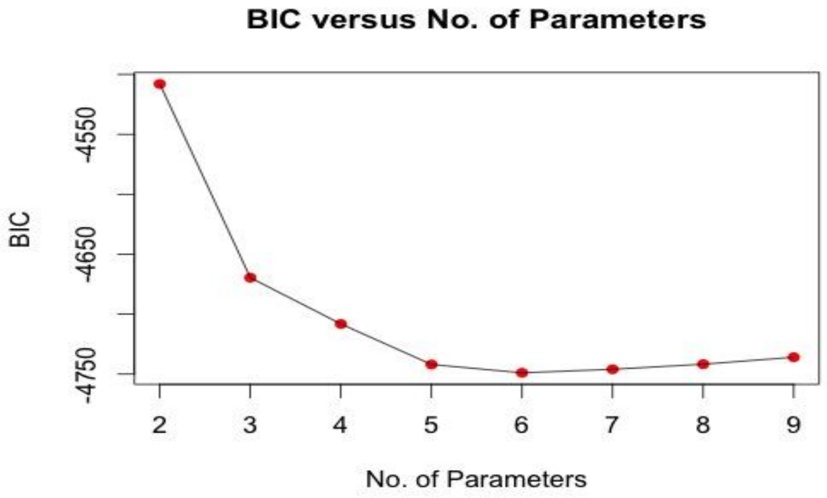
I employed 3-fold cross validation to better estimate error rate and verify the predictive power of our two models. I split the data into three parts. Then I left out the first part, fit the model on the remaining 2 parts, and computed the misclassification errors on the left out part. This procedure is repeated 3 times with each time taking out a different part. By averaging the 3 different misclassification errors I get an estimated validation (test) error rate for new observations.

For logistic regression, I first performed variable selection. The exhaustive selection algorithm was run using the “regsubset” function in package “leaps” to show the best combinations of variables for different subset sizes, as is displayed in the figure below.



```
1 subsets of each size up to 8
Selection Algorithm: exhaustive
         LIMIT_BAL SEX2 EDUCATION2 EDUCATION3 EDUCATION4 MARRIAGE2 MARRIAGE3 AGE delay.sum bill.mean repayment.mean
1  ( 1 ) " "       " "  " "        " "        " "        " "       " "       " " "*"       " "       " "
2  ( 1 ) "*"       " "  " "        " "        " "        " "       " "       " " "*"       " "       " "
3  ( 1 ) "*"       " "  " "        " "        " "        " "       " "       " " "*"       " "       "*"
4  ( 1 ) "*"       " "  " "        " "        " "        "*"       " "       " " "*"       " "       "*"
5  ( 1 ) "*"       "*"  " "        " "        " "        "*"       " "       " " "*"       " "       "*"
6  ( 1 ) "*"       "*"  " "        " "        "*"        "*"       " "       " " "*"       " "       "*"
7  ( 1 ) "*"       "*"  " "        " "        "*"        "*"       " "       " " "*"       "*"       "*"
8  ( 1 ) "*"       "*"  " "        " "        "*"        "*"       " "       "*" "*"       "*"       "*"
```

Then I calculated Bayesian Information Criterion (BIC) for each subset size. The figure below depicts the BIC versus the number of parameters p+1 included in a linear model. BIC around or less than p+1 indicates a good fit of the model. By the criterion of minimizing BIC, I decided to use 6 parameters, equivalently 5 predictors, because this was the lowest point observed in the figure.



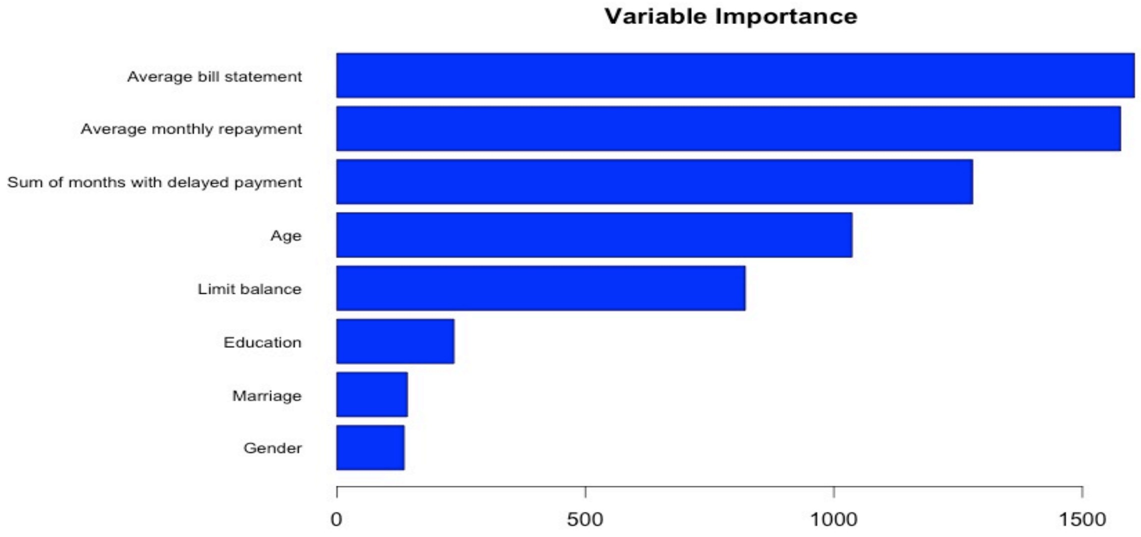
The 5 variables selected as predictors are: extended credit (limit balance), gender, marriage, sum of months with delayed payment, and average monthly repayment. I tried a grid of different threshold values for the classifier. I am concerned that clients who are likely to default would be classified as non-default type, so I want FNR to be as low as possible. There is, however, a tradeoff between TER and FNR. Thus when choosing the best classifier among different thresholds, I set a constraint that TER must be below 0.5. I want to be conservative in predicting individuals who are at risk for default without a whole lot of sacrifice of the overall classification accuracy. The result shows that the best threshold is 0.16. Under the best threshold, TER and FNR are 48% and 15% respectively.

For random forest, I used a full set of variables as predictors, because random forest performs variable selection automatically. When building the decision trees, a random sample of 4 predictors is chosen as split candidates from the full set of 8 predictors for each time a split in a tree. The split is allowed to use only one of those 4 predictors. The 8 explanatory variables are: extended credit (limit balance), gender, education, marriage, age, sum of months with delayed payment, and average monthly repayment, average monthly bill statement.

For random forest method, the TER and the FNR are 20% and 67% respectively. Compared to the logistic regression, random forest has much lower TER at the expense of much higher FNR. This is expected since by using the majority vote to assign each individual to a particular class, random forest would produce a result that leads the lowest TER regardless of FNR.

I used two measures of variable importance to estimate the quality, or influence of the splits over each variable. The results are reported in the table below. The second column is based on the mean decrease of accuracy in predictions on the training data set when a given variable is excluded from the model. The third column is a measure of the total decrease in Gini Index that results from splits on a particular variable, averaged over all trees.

	Mean Decrease Accuracy	Mean Decrease Gini
Extended Credit (Limit Balance)	38.70	821.75
Sex	7.28	135.41
Education	13.14	235.6
Marital status	21.60	141.67
Age	30.40	1036.49
Sum of months with delayed payment	192.34	1278.95
Average bill statement	49.94	1604.66
Average monthly repayment	48.34	1576.71



The above figure is a bar chart of the variable importance in terms of mean decrease in the Gini index. The larger the decrease in the Gini Index, the better the split is on a particular variable, and thus the higher importance that variable is. From this perspective, the most important variables are (in decreasing order): average bill statement, average monthly repayment, sum of months with delayed payment, age and extended credit (limit balance).

For logistic regression with variable selection using BIC, the model contains limit balance, gender, marital status, sum of months with delayed payment, and average monthly repayment, a total of 5 predictors. For random forest, the model contains limit balance, sex, education, marriage, age, sum of months with delayed payment, and average of monthly repayment, average bill statement, a total of 8 predictors.

By measuring variable importance, I found that gender and marital status are the two least important variables, but nonetheless they were included in the logistic regression model. This may simply mean that different models use variables in different ways to predict results.

	Total Error Rate (TER)	False Negative Rate (FNR)
Logistic Regression	48%	15%
Random Forest	20%	67%

The table above depicts the misclassification rates of two methods. The contrasts between the two methods were due to seeking the best threshold for classifier in logistic regression that could produce the lowest possible FNR. The tradeoff between TER and FNR, however, resulted in a higher TER for logistic regression. For the purpose of identifying clients that are likely to default and reducing credit risk, we prefer logistic regression to random forest.

Both methods recognize limit balance, sum of months with delayed payment, and average monthly repayment as the most contributing/important variables. This informs us that they should be used in predicting credit card default of new clients in the future.

Limitation

When doing variable selection, I found that using $Adjusted\ R^2$ and $Mallows'C_p$ had different results than BIC. However, I simply employ BIC because it gave us a less complex model. Further analysis is needed to justify our use of BIC and a less flexible model.

Furthermore, because of the control of thresholds in logistic regression, I am unable to compare the accuracy of logistic regression to that of random forest. I would need to seek another way to evaluate the prediction.