Model #101: Credit Card Default Model

Model Development Guide

George Brown

1. Introduction

In this capstone project the goal is to look at a dataset that contains credit card payment information from a set of Taiwanese clients. We will look at what variables will accurately predict the customers to default or not to default on their credit cards. There will be several parts to this project. First, we will look at the dataset and define the variables with a data dictionary. Then we will have a feature engineering section where we will create new variables from the original variables. In the following section there will be two parts. First, we will look at an exploratory data analysis for the different variables. The second part will be to do an exploratory data analysis from more of a model-based approach. We will then see how well we can predict default with four different models. We will use R, a programming language to run the models. For each model we will have a training set of data followed by a test set. We will look at many different metrics for each model and look at the variables that are considered most important. In the next section we will look at a comparison of all the model metrics and determine which model works the best based off the different results. Finally, we will wrap up the project by talking about the main results and any recommendations that we may have for future research and for different techniques.

2.1 Data Description and Data Dictionary

The dataset is from the UC Irvine Machine Learning Repository. The dataset was from research aimed at Taiwanese customers from 2005 and looked at the default of credit card clients. There are 30,000 observations with a total of 23 variables explanatory variables. This does not include the “ID” variable which is used to identify each unique observation. Each observation includes repayment status, bill statement and amount paid over a six-month period in 2005. There is also additional information about a person such as sex, education, marriage, and age. Below is a data dictionary of the 30 variables including ones that were created for the modeling which include the training, test, and validation variables. There is also a “DEFAULT” variable which tells us if the customer defaulted on their payment. Default is the response variable for this project. The goal is to find the variables that best predict default.

Table 1: Data Dictionary

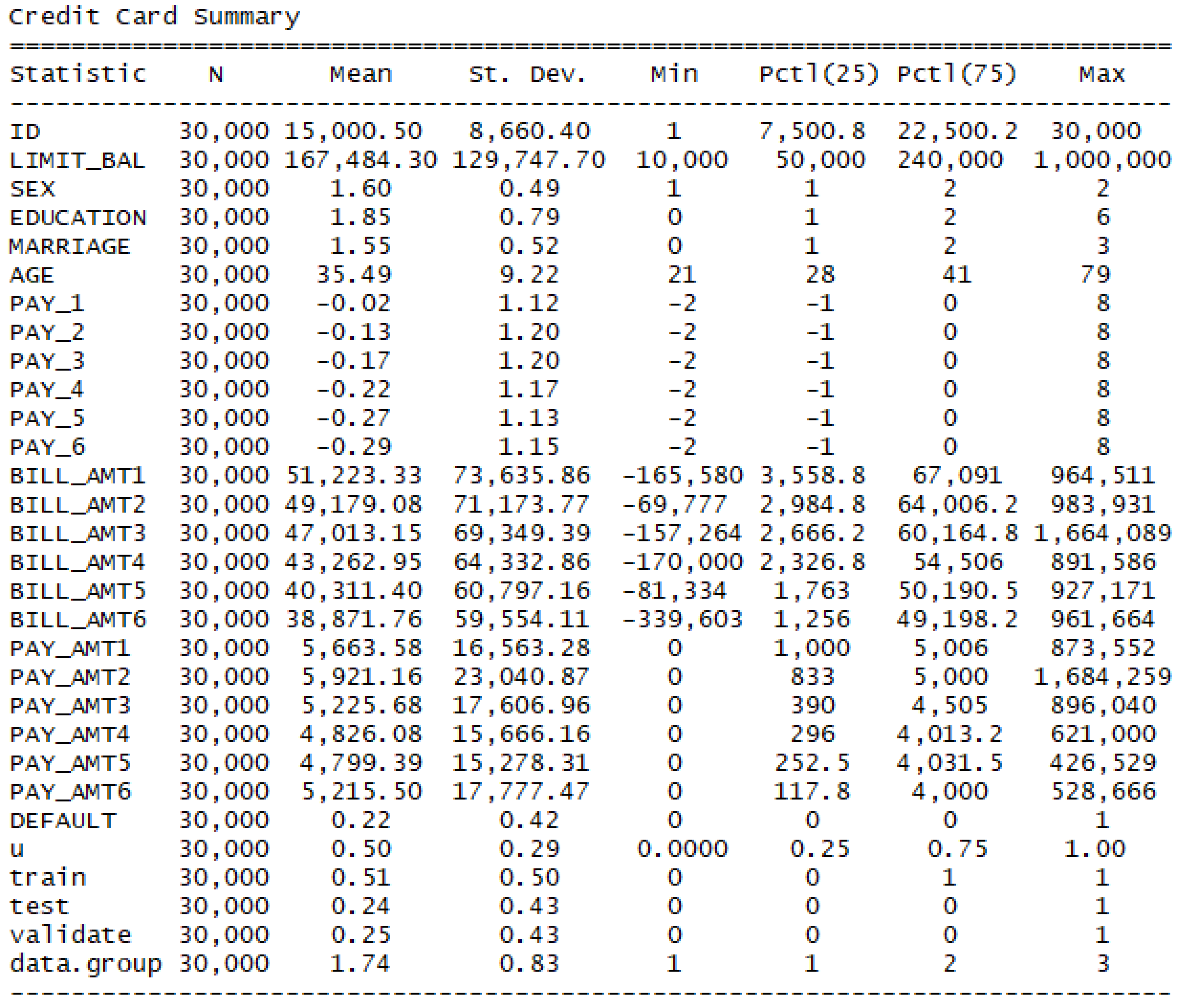
Number	Variable	Description
1.	ID	Identifies each row
2.	Limit Balance	Amount of the given credit (NT dollar)
3.	Sex	(1 = male; 2 = female)
4.	Education	1 = graduate school; 2 = university; 3 = high school; 4 = others
5.	Marriage	1 = married; 2 = single; 3 = others
6.	Age	Age (year)
7.	Pay_1(Formerly Pay_0)	Repayment status in September, 2005
8.	Pay_2	Repayment status in August, 2005
9.	Pay_3	Repayment status in July, 2005
10.	Pay_4	Repayment status in June, 2005
11.	Pay_5	Repayment status in May, 2005
12.	Pay_6	Repayment status in April, 2005
13.	Bill_Amt1	Bill statement in September, 2005
14.	Bill_Amt2	Bill statement in August, 2005
15.	Bill_Amt3	Bill statement in July, 2005
16.	Bill_Amt4	Bill statement in June, 2005
17.	Bill_Amt5	Bill statement in May, 2005
18.	Bill_Amt6	Bill statement in April, 2005

19.	Pay_Amt1	Amount Paid in September, 2005
20.	Pay_Amt2	Amount Paid in August, 2005
21.	Pay_Amt3	Amount Paid in July, 2005
22.	Pay_Amt4	Amount Paid in June, 2005
23.	Pay_Amt5	Amount Paid in May, 2005
24.	Pay_Amt6	Amount Paid in April, 2005
25.	Default	Response Variable (1=True; 0=False)
26.	u	Sorting for train/test/validate splits
27.	Train	Train Data Set
28.	Test	Test Data Set
29.	Validate	Validate Data Set
30.	Data.group	Train, test or validate group (1=train, 2=test, 3=validate

2.2 Data Quality Check

In this section we can look at the dataset. By looking at the data dictionary it appears one thing is out of place. Pay_0 is out of place. It should be Pay_1 so it follows the same patterns as the Bill_Amt and Pay_Amt. The summary of the table is below with the Pay_1 replacing Pay_0. The below chart shows the number of observations, the mean, the standard deviation, the minimum, the 25%, the 75%, and the maximum for all the different variables.

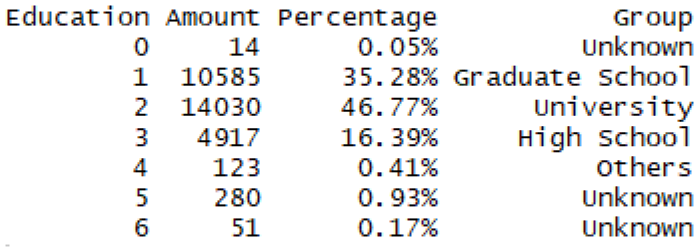
Table 2: Credit Card Summary



```
Credit Card Summary
==========================================================================
Statistic     N       Mean      St. Dev.     Min    Pctl(25) Pctl(75)    Max
--------------------------------------------------------------------------
ID          30,000 15,000.50   8,660.40      1     7,500.8  22,500.2  30,000
LIMIT_BAL   30,000 167,484.30 129,747.70  10,000    50,000  240,000  1,000,000
SEX         30,000    1.60       0.49        1        1        2         2
EDUCATION   30,000    1.85       0.79        0        1        2         6
MARRIAGE    30,000    1.55       0.52        0        1        2         3
AGE         30,000   35.49       9.22       21       28       41        79
PAY_1       30,000   -0.02       1.12       -2       -1        0         8
PAY_2       30,000   -0.13       1.20       -2       -1        0         8
PAY_3       30,000   -0.17       1.20       -2       -1        0         8
PAY_4       30,000   -0.22       1.17       -2       -1        0         8
PAY_5       30,000   -0.27       1.13       -2       -1        0         8
PAY_6       30,000   -0.29       1.15       -2       -1        0         8
BILL_AMT1   30,000 51,223.33  73,635.86  -165,580 3,558.8   67,091   964,511
BILL_AMT2   30,000 49,179.08  71,173.77  -69,777  2,984.8  64,006.2  983,931
BILL_AMT3   30,000 47,013.15  69,349.39  -157,264 2,666.2  60,164.8 1,664,089
BILL_AMT4   30,000 43,262.95  64,332.86  -170,000 2,326.8   54,506   891,586
BILL_AMT5   30,000 40,311.40  60,797.16  -81,334   1,763   50,190.5  927,171
BILL_AMT6   30,000 38,871.76  59,554.11  -339,603  1,256   49,198.2  961,664
PAY_AMT1    30,000  5,663.58  16,563.28      0      1,000    5,006   873,552
PAY_AMT2    30,000  5,921.16  23,040.87      0       833     5,000  1,684,259
PAY_AMT3    30,000  5,225.68  17,606.96      0       390     4,505   896,040
PAY_AMT4    30,000  4,826.08  15,666.16      0       296    4,013.2  621,000
PAY_AMT5    30,000  4,799.39  15,278.31      0      252.5   4,031.5  426,529
PAY_AMT6    30,000  5,215.50  17,777.47      0      117.8    4,000   528,666
DEFAULT     30,000    0.22       0.42        0        0        0         1
u           30,000    0.50       0.29     0.0000     0.25     0.75      1.00
train       30,000    0.51       0.50        0        0        1         1
test        30,000    0.24       0.43        0        0        0         1
validate    30,000    0.25       0.43        0        0        0         1
data.group  30,000    1.74       0.83        1        1        2         3
--------------------------------------------------------------------------
```

There appears to be a few variables that have data discrepancies. Education is only supposed to have values at 1, 2, 3, and 4 but the chart below shows answers at 0, 5, and 6. The assumption is that 0, 5, and 6 are in the “other” category. Below is a table for the education. The table values have the following meanings. 1= graduate school, 2= university, 3= high school, 4= others.

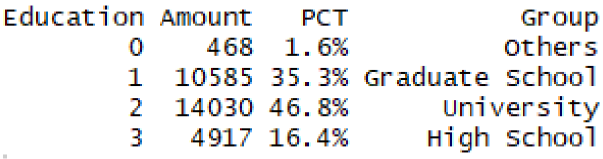
Table 3: Original Education Variable



```
Education Amount Percentage           Group
        0     14      0.05%         Unknown
        1  10585     35.28% Graduate School
        2  14030     46.77%      University
        3   4917     16.39%     High School
        4    123      0.41%          Others
        5    280      0.93%         Unknown
        6     51      0.17%         Unknown
```

When we remap the values, we change all values from 0,5, & 6 to 4 which is others. For simplicity we regroup the “4” to “0”. Groups 1,2, and 3 stay the same.

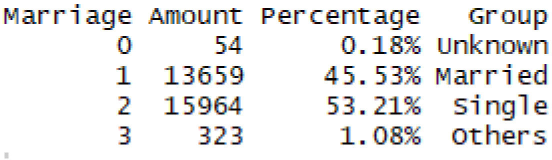
Table 4: New Education Variable



```
Education Amount   PCT           Group
        0    468  1.6%          Others
        1  10585 35.3% Graduate School
        2  14030 46.8%      University
        3   4917 16.4%     High School
```

Another discrepancy is marriage. There are only supposed to be values 1, 2, and 3, but there were also answers coming in at 0. The assumption for this is that the 0 is in the other category. The below chart shows marriage. The table has the following values. 1= married, 2 = single, 3= others

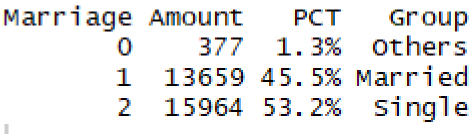
Table 5: Original Marriage Variable



```
Marriage Amount Percentage   Group
       0     54      0.18% Unknown
       1  13659     45.53% Married
       2  15964     53.21%  Single
       3    323      1.08%  Others
```

Next we regroup marriage. There are 54 answers of 0 which is an incorrect answer. We move those to the others category. After this there are a total of 377 in the others category and that is grouped at “0”.

Table 6: New Marriage Variable



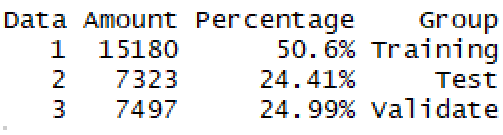
```
Marriage Amount   PCT   Group
       0    377  1.3%  Others
       1  13659 45.5% Married
       2  15964 53.2%  Single
```

The last discrepancy is there are -2 and -1 for the repayments. All numbers should be from 0 to 9. The assumption for this is the payments are on time and that they should be at 0 and not -1 or -2.

2.3 Data Observations

There are a total of 15180 observations in the training set, 7323 in the test set and 7497 in the validation set. The training set has about just over half of the observations while the test and validation sets are each at about a quarter. The training set needs to be at least half so we can make sure the models are accurate when we test them using the training data. This is important for when we train, test and validate the models. This should be an adequate number of observations for each group.

Table 7: Data Observation by group



```
Data Amount Percentage    Group
   1  15180      50.6% Training
   2   7323     24.41%     Test
   3   7497     24.99% Validate
```

3- Feature Engineering

In this section we will add new variables. We have many variables with this data set but adding new variables can help provide additional insight for the data and for our modeling. In section 3.1 we talk about the variables and define them.

3.1-New Variables and Definitions

There are 15 variables that we will talk about in this section. They are the following.

- Age Below 25- Age is already a variable, but bins will be created using different age groups. This bin will be for age groups that under 25. The utilization for this variable is 0 to 100
- Age 26-40. This variable will be created for age groups 26-40. The utilization for this variable is 0 to 100
- Age above 40- This variable will be created for age groups above 40. The utilization is from 0 to 100.
- Average bill amount-This will be calculated by averaging the monthly bill amount over six months. The utilization for this is 0 to 100
- Average Payment Amount-This will be calculated by using the monthly payment amount over six months. The utilization for this is 0 to 100.
- Payment Ratio-This is dividing the payment by the bill amount. The utilization for this 0 to 1.
- Average Payment Ratio-This is the average of all the payment ratios. The utilization for this is 0 to 1.
- Total sum payment Ratio-The total of all the ratios added together. This is scaled from 1 to 100.
- Utilization Sum-This is how much of the credit line the customers are using. This is calculated by bill amount/limit balance. The is scaled from 1 to 100
- Average Utilization-This is the average utilization over the six months. This is scaled from 0 to 1
- Balance growth over 6 months- This is looking at the balance growth due to continued spending with only partial payments every month. This is scaled from 1 to 100.
- Utilization Growth over 6 months- This is looking to see if the balance is getting close to the credit limit. This is scaled from 0 to 1.

- Max Bill Amount-This is the maximum amount billed over 6 months. This is scaled from 0 to 100.
- Max Payment Amount-This is the payment amount over the six months-The utilization for this is 0 to 100.
- Total Bill amount-The total amount of the six total bill amounts over the six-month period. The utilization for this is 0 to 100.
- Max Delinquency-This is the max of the repayment variables. The utilization for this is 0 to 100.
- Total Payment Amount-The total amount of the six total payment amounts over the six-month period. The utilization for this is 0 to 100.

Table 8: Chart for Variables and exact formulas

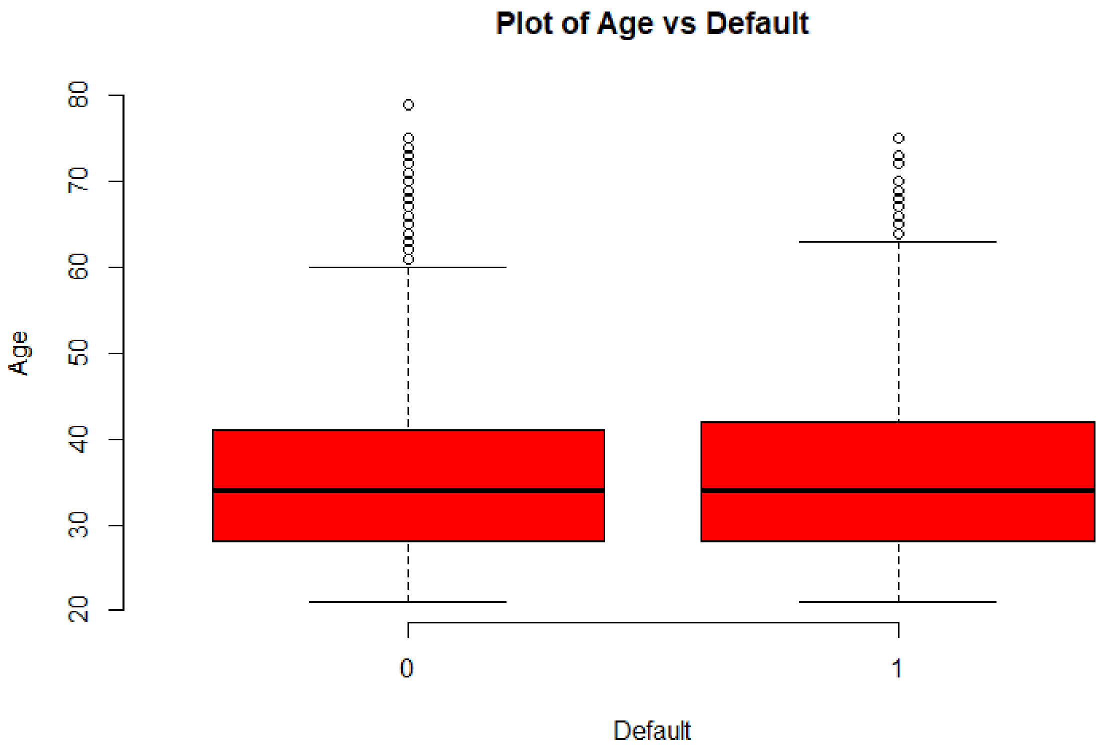
Variable	Formula
1. Age Bin (Below 26, 26-40, and above 40)	Grouped by age of clients
2. Average Bill Amount	Take the total bill amount and divide by 6
3. Average Payment Amount	Take the sum of the total payments and divide by 6
4.Payment Ratio	Take the payment amount for month 1 and divide by the bill amount for month 2. Keep going for each additional month. So, the second ratio would be take the payment amount for month 2 and divide by bill amount for month 3.
5. Average Payment Ratio	Take total payment ratios and divide by 5
6. Total Sum Payment Ratio	Add up the payment ratios over the 6 months
7. Utilization Sum	Bill amount (for each month)/Limit Balance. Gives utilization for each month. Then add up all six months together.
8. Average Utilization	Divide the utilization sum by 6.
9. Balance Growth over 6 Months	Limit Balance-Bill Amount for month 6 and subtract the limit balance-bill amount for month 1.
10.Utilization Growth over 6 Months	Take the first bill and subtract the sixth bill
11. Max Bill Amount	Take the maximum bill over the 6-month period.

12. Max Payment Amount	Take the maximum paid amount in the 6-month time frame
13. Total Bill Amount	Add up bill amounts for six months
14. Max Delinquency	Look at the payment variable. A 0 is good. Anything above a 0 means the payment has been delayed. 1 for 1 month, 2 for 2 months and so on. Take the maximum
15. Total Payment Amount	Add up payment amounts for six moths

3.2 Feature Engineering Age Bins

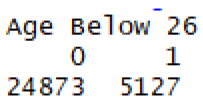
The one variable that we put into bins is ages. There are three separate bins 25 and under, 26-40, and 41 and above. The boxplot below shows the initial plot of Age before the binning versus Default. From the age perspective the outliers occur from people that are over 60. There are a few more outliers from that age group for people that did not default.

Figure 9- Age vs Default Boxplot



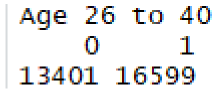
As we can see, below age 26 and above age 40 had far less people than the middle age group 26 to 40. If you add up the amount of people in the first and third group there are more in the middle age group than the other two age groups combined.

Table 10: Age Bin 1



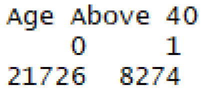
```
Age Below 26
    0     1
24873  5127
```

Table 11: Age Bin 2



```
Age 26 to 40
    0     1
13401 16599
```

Table 12: Age Bin 3

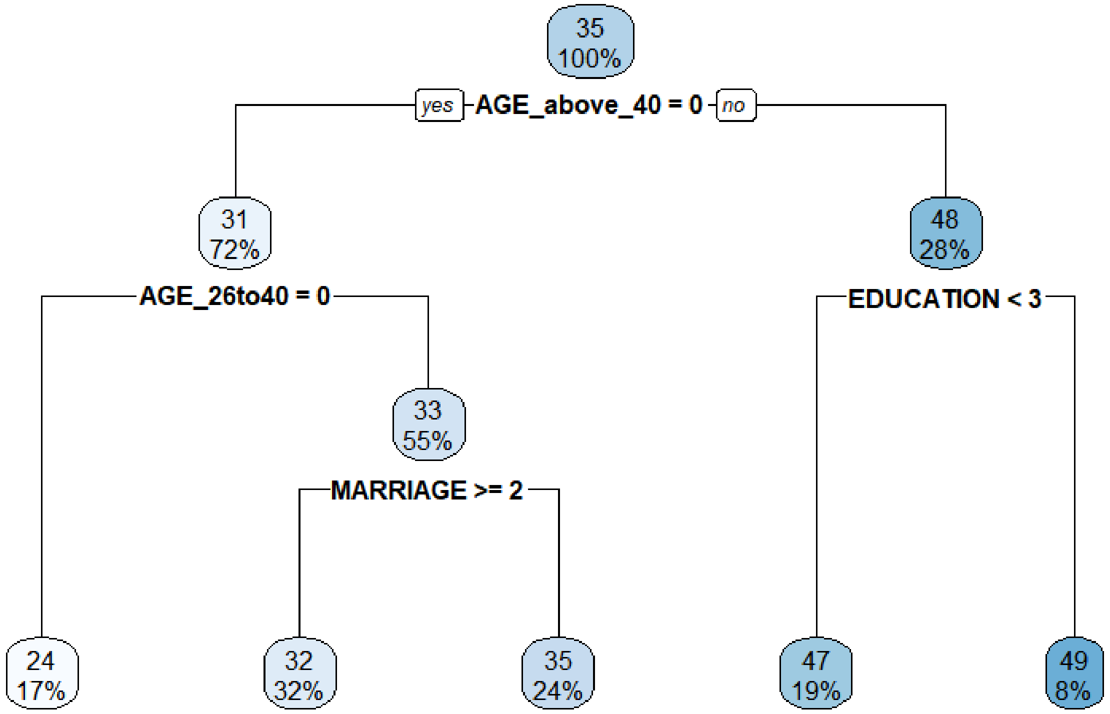


```
Age Above 40
    0     1
21726  8274
```

Finally, we look at a decision tree as it pertains to age. There are a couple things that we can take away from the decision tree.

- Ages 32 to 35 are more likely to be single
- Ages 47 and above are more than likely to have an education above a high school level.

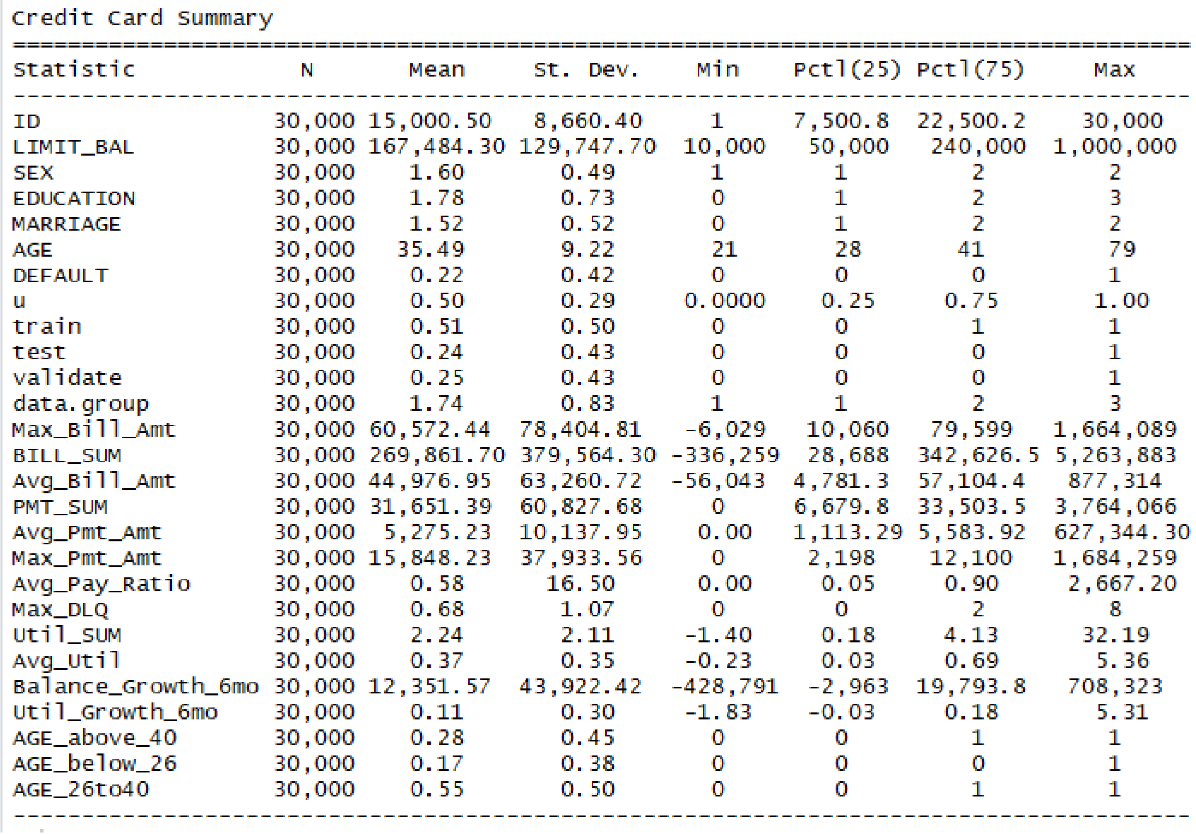
Figure 13: Age Decision Tree



4.1-Traditional Exploratory Data Analysis-

The first thing we want to do in the exploratory data analysis is look at our data including our featured engineered variables. We have removed monthly payment, bill and repayment variables and added in variables that consider these variables over the six-month period. Below is a chart that shows the new data set that we will do out EDA and our modeling from.

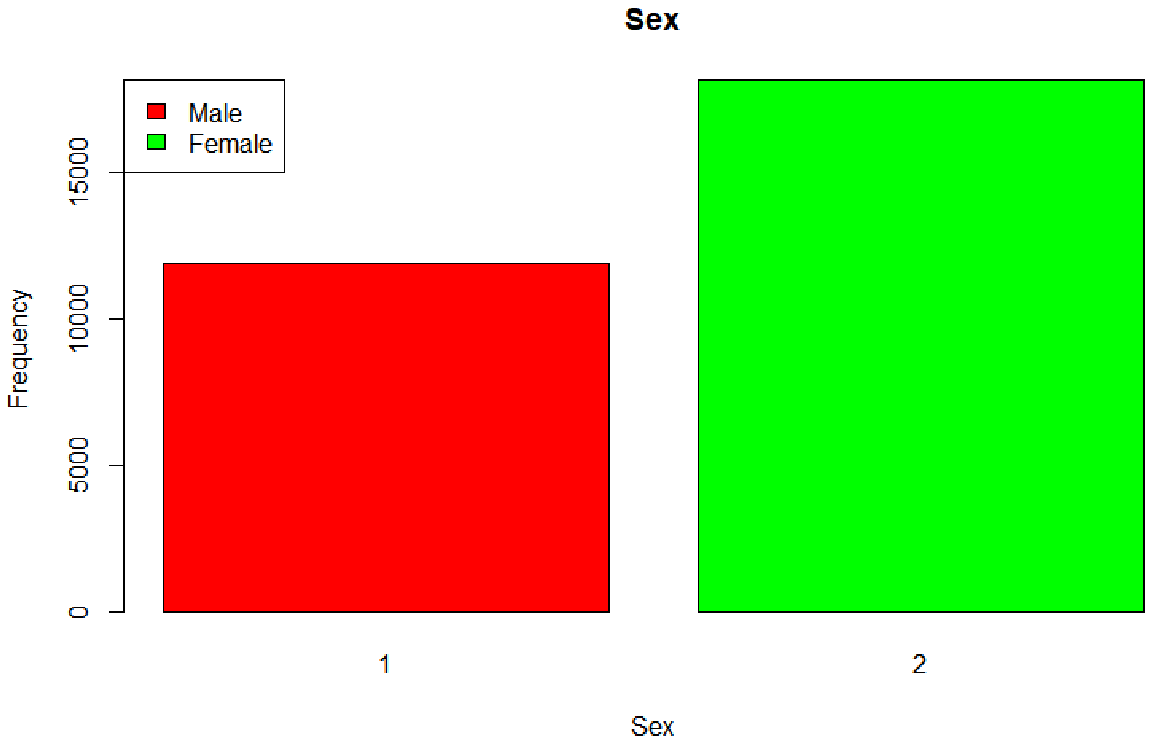
Table 14: Credit Car Summary



```
Credit Card Summary
==================================================================================
Statistic              N       Mean      St. Dev.     Min    Pctl(25) Pctl(75)     Max
----------------------------------------------------------------------------------
ID                   30,000 15,000.50   8,660.40     1      7,500.8  22,500.2    30,000
LIMIT_BAL            30,000 167,484.30 129,747.70 10,000    50,000    240,000  1,000,000
SEX                  30,000    1.60       0.49       1         1         2         2
EDUCATION            30,000    1.78       0.73       0         1         2         3
MARRIAGE             30,000    1.52       0.52       0         1         2         2
AGE                  30,000   35.49       9.22      21        28        41        79
DEFAULT              30,000    0.22       0.42       0         0         0         1
u                    30,000    0.50       0.29     0.0000     0.25      0.75      1.00
train                30,000    0.51       0.50       0         0         1         1
test                 30,000    0.24       0.43       0         0         0         1
validate             30,000    0.25       0.43       0         0         0         1
data.group           30,000    1.74       0.83       1         1         2         3
Max_Bill_Amt         30,000 60,572.44  78,404.81   -6,029    10,060    79,599   1,664,089
BILL_SUM             30,000 269,861.70 379,564.30 -336,259  28,688  342,626.5 5,263,883
Avg_Bill_Amt         30,000 44,976.95  63,260.72  -56,043   4,781.3  57,104.4   877,314
PMT_SUM              30,000 31,651.39  60,827.68     0      6,679.8  33,503.5 3,764,066
Avg_Pmt_Amt          30,000  5,275.23  10,137.95   0.00    1,113.29 5,583.92 627,344.30
Max_Pmt_Amt          30,000 15,848.23  37,933.56     0      2,198     12,100  1,684,259
Avg_Pay_Ratio        30,000    0.58      16.50     0.00      0.05      0.90    2,667.20
Max_DLQ              30,000    0.68       1.07       0         0         2         8
Util_SUM             30,000    2.24       2.11     -1.40      0.18      4.13     32.19
Avg_Util             30,000    0.37       0.35     -0.23      0.03      0.69      5.36
Balance_Growth_6mo   30,000 12,351.57  43,922.42  -428,791  -2,963   19,793.8   708,323
Util_Growth_6mo      30,000    0.11       0.30     -1.83     -0.03      0.18      5.31
AGE_above_40         30,000    0.28       0.45       0         0         1         1
AGE_below_26         30,000    0.17       0.38       0         0         0         1
AGE_26to40           30,000    0.55       0.50       0         0         1         1
----------------------------------------------------------------------------------
```

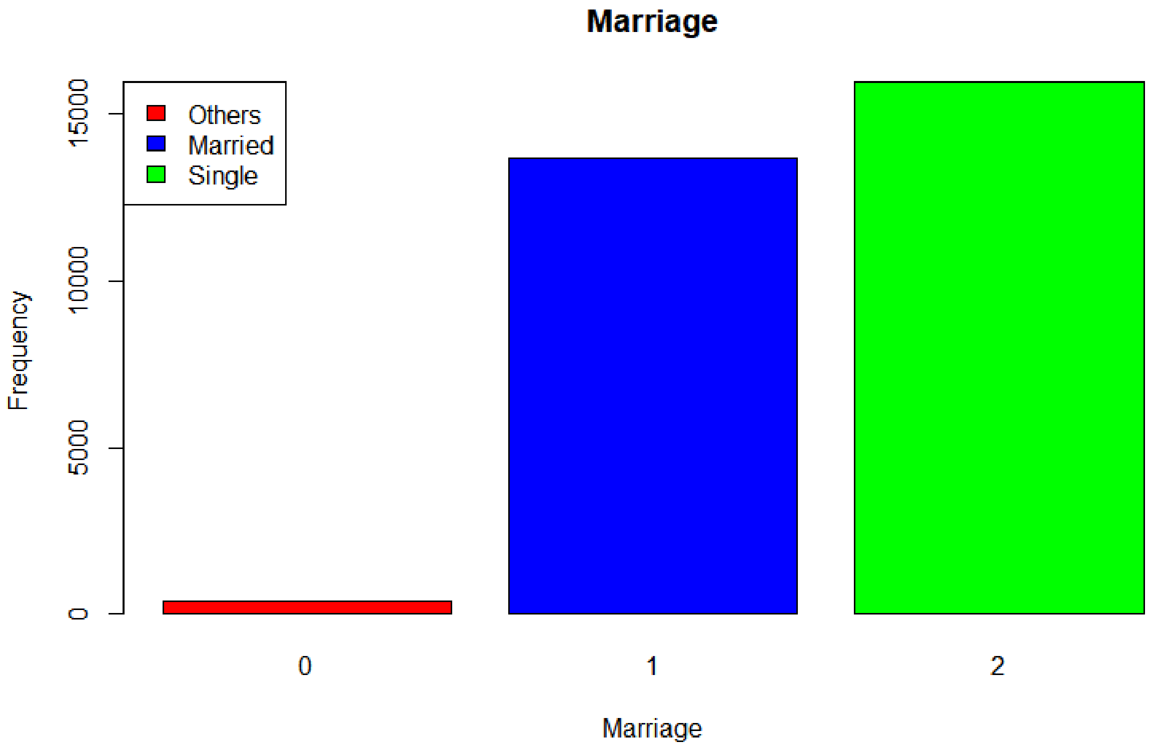
In the plots below we look at different categorical variables in the dataset. In the plot below there are a good amount more females than males. Of the 30,000 it only looks like about 12,000 are males.

Graph 15: Barplot of Male versus Female



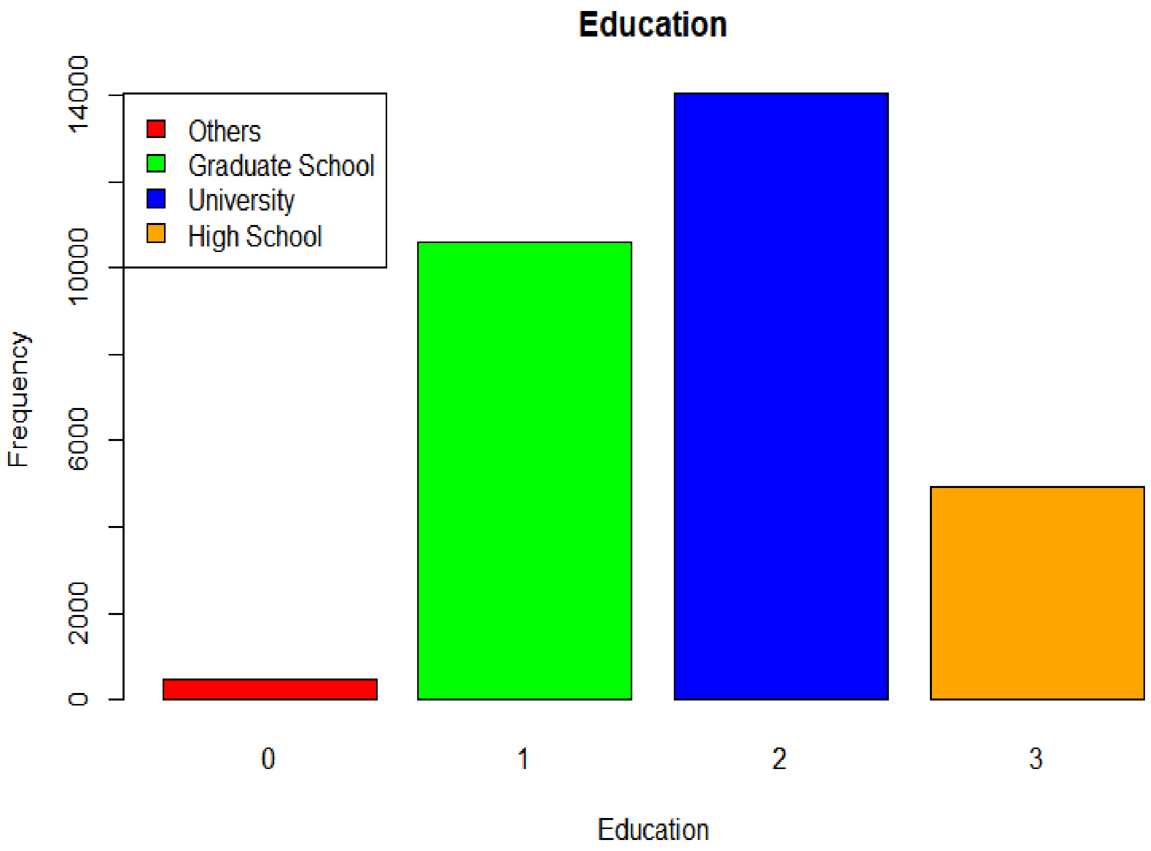
Another categorical variable to look at is the marriage variable. In the plot below the majority are single but married is not very far behind. There are very few people in the “others” category.

Graph 16: Barplot of Married vs. Not Married



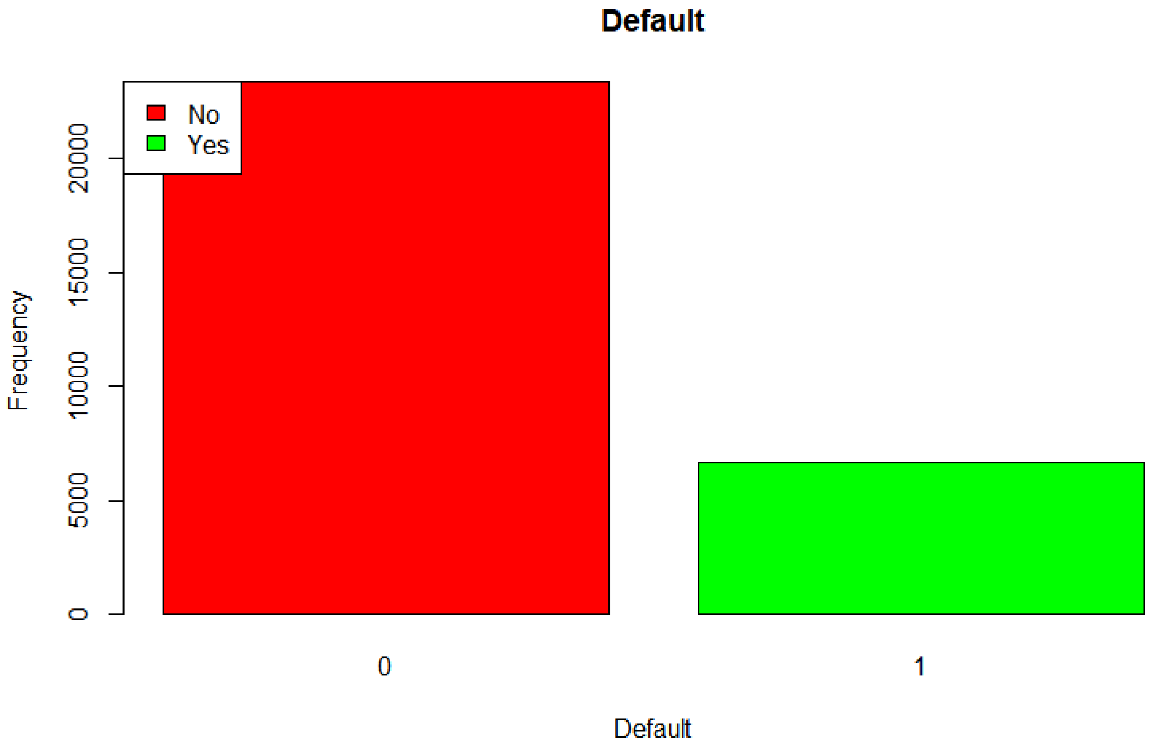
Education also seems like it could be an important categorical variable. In the plot below most people’s education went through a normal undergraduate school. This appears to be at about half. About 1/3 went to graduate school and the amount of people who went to high school was a distant third. There were very few people in the “others” category.

Graph 17: Barplot of Education



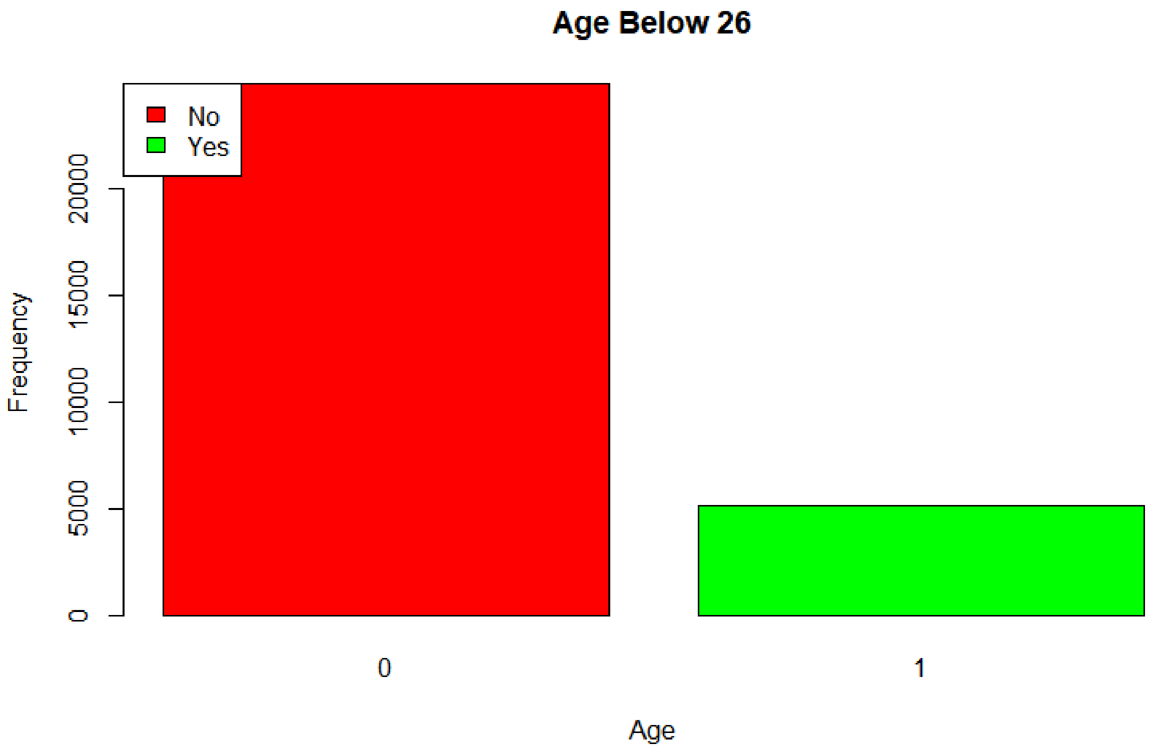
The goal of this project is to be able to accurately predict the default. The below plot shows the default variable, which is the response variable. There were far more people who did not default on their payment than those that did. With that being said there were still 6,636 people who did default on their payments.

Graph 18: Barplot of Default Variable

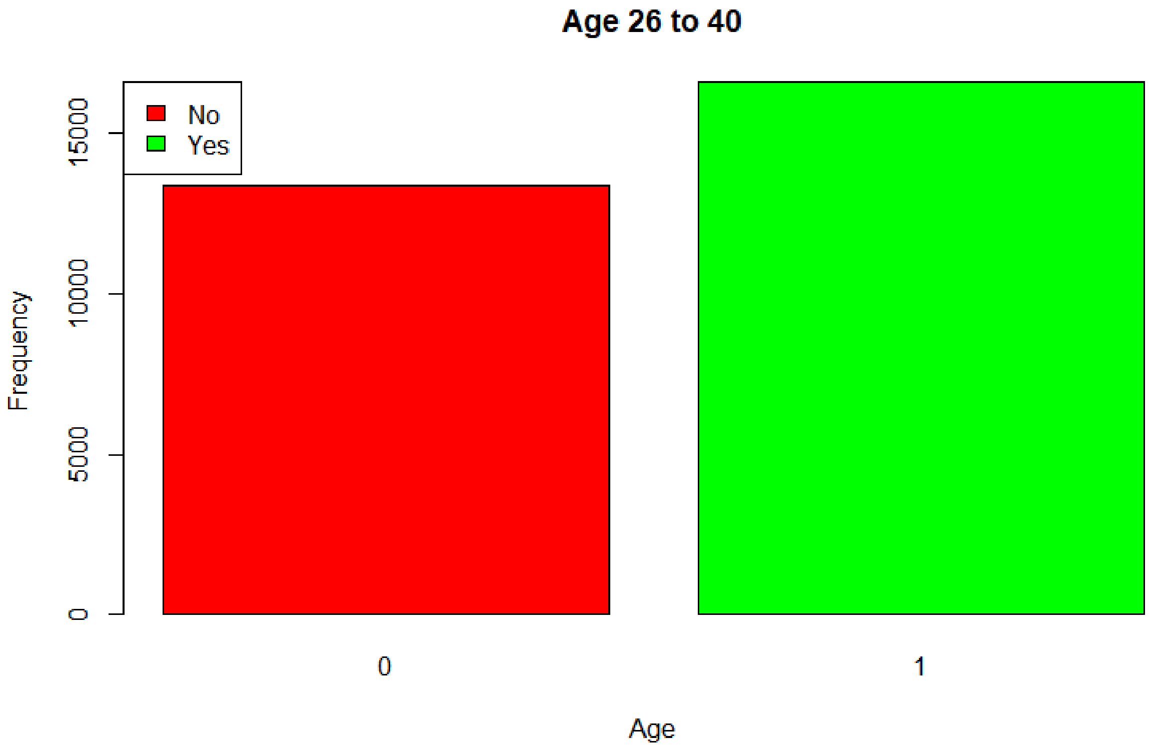


The below three plots below show our age categories. The green means that person is in that particular age group. As one can see below the 26 to 40 age group is the predominant age group for these observations.

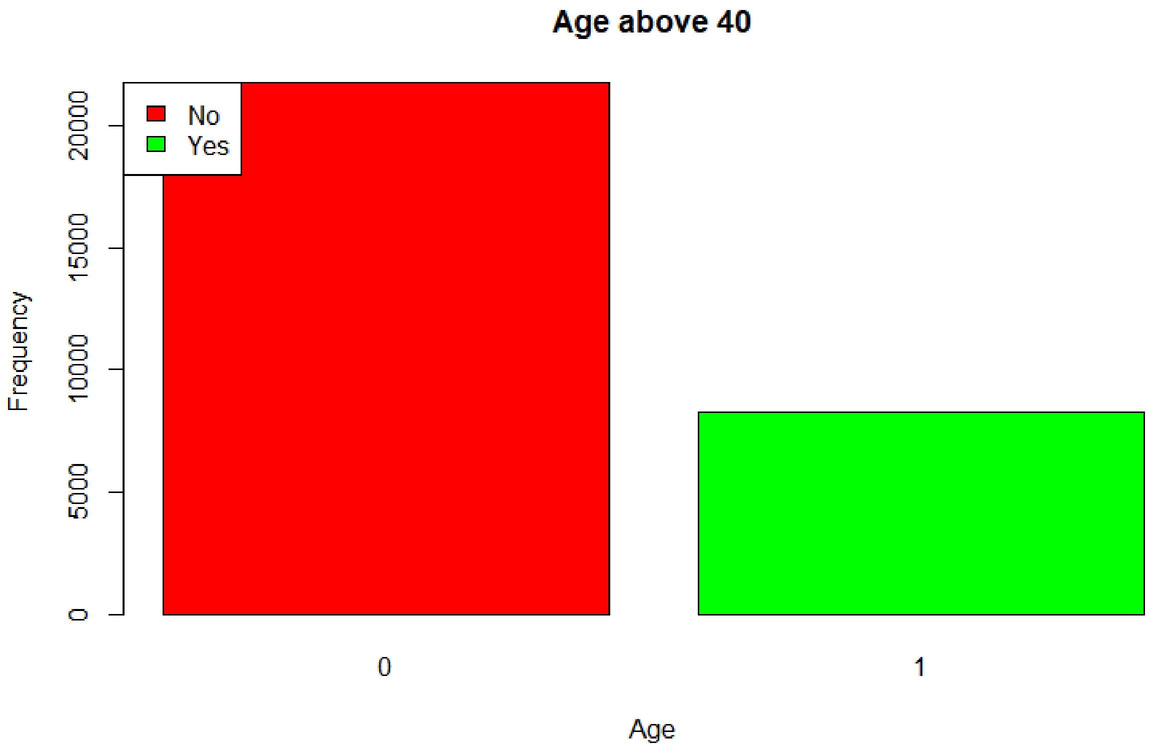
Graph 19: Barplot of Age Below 26



Graph 20: Barplot of Age 26 to 40

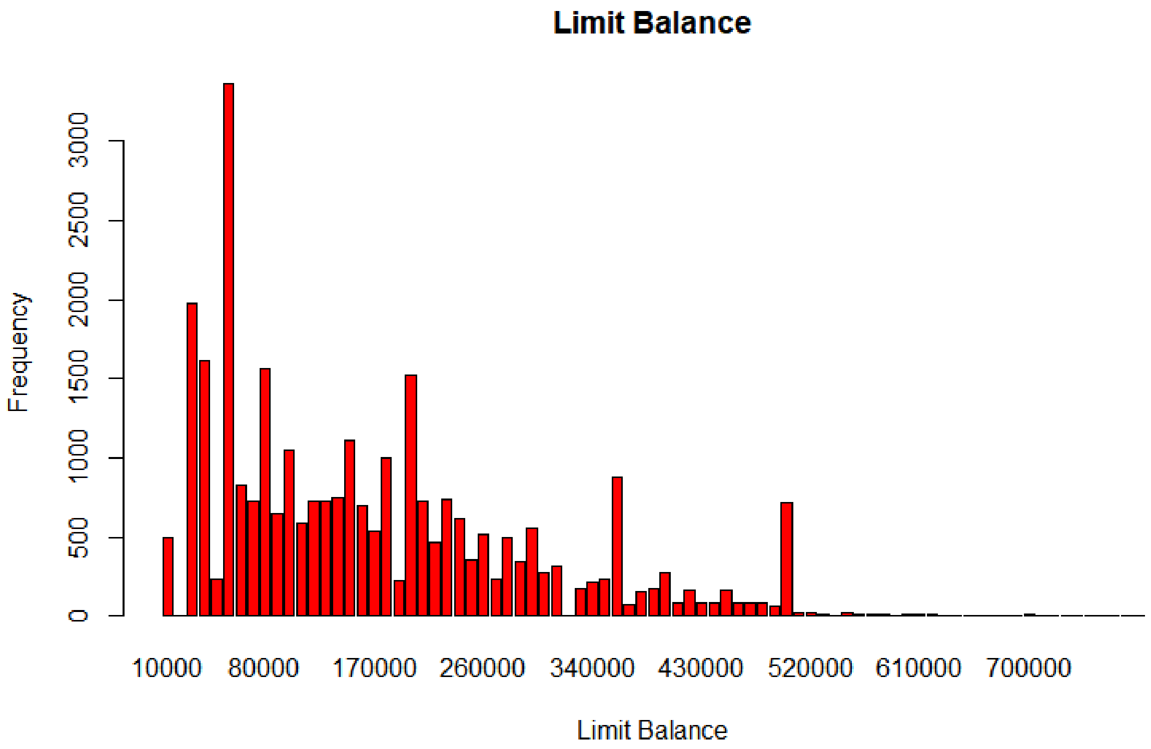


Graph 21: Barplot of Age Above 40



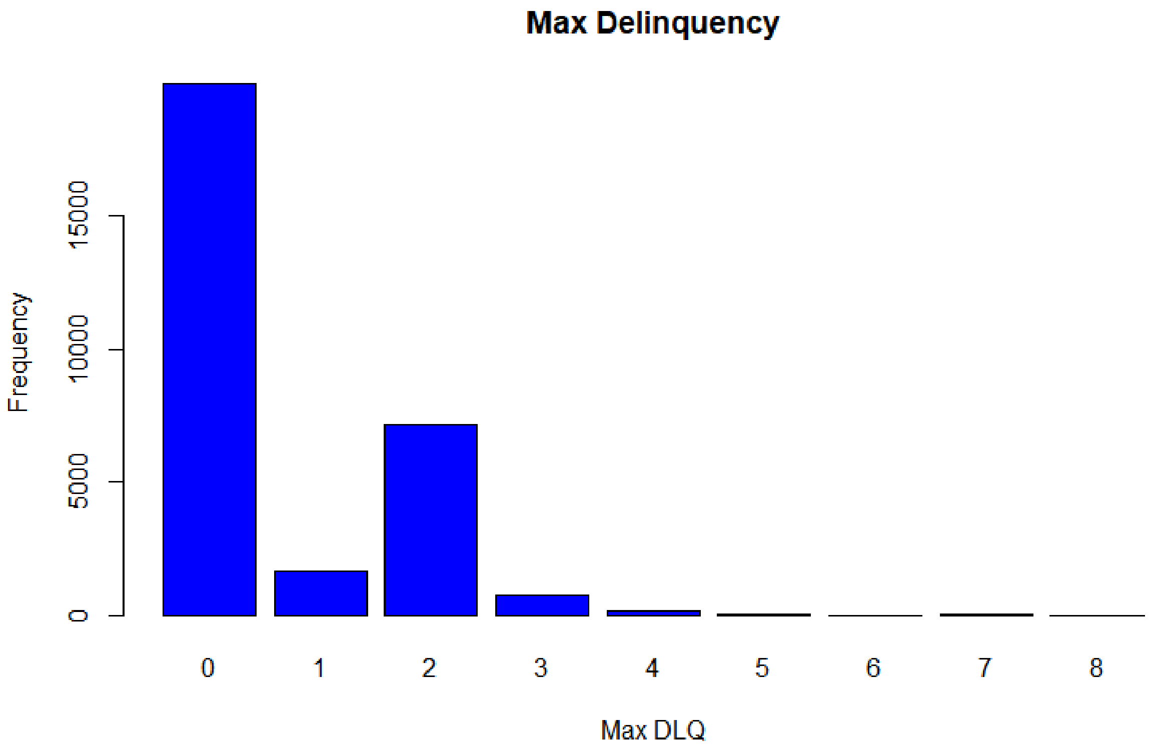
One of the other important variables is the limit balance. It could help us determine our response variable. The below plot shows the limit balance for the different people. Most of the people are between $0 and $100,000. However, there are also some people between $300,000 and $500,000. There is a wide variety of balance associated with the different observations.

Graph 22: Barplot of Limit Balance



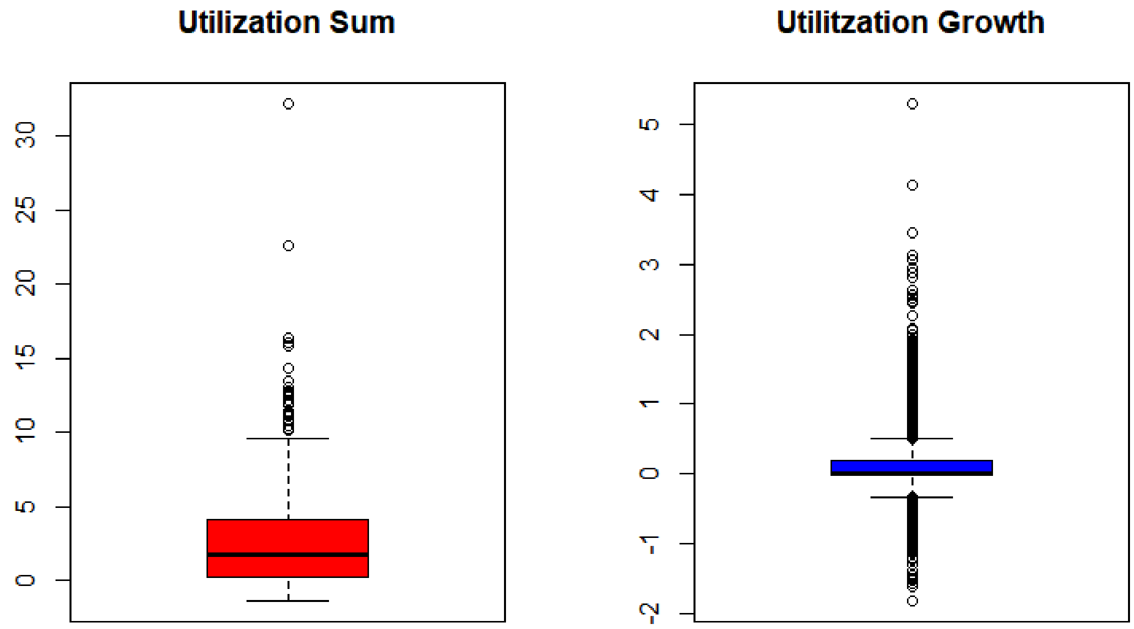
The plot below shows the max delinquency. 0 is good for max delinquency. However, anything later than 0 and the payment is late. Most people were not late on their payments. However, more people were 2 months late than were one month late which is interesting.

Graph 23: Barplot of Max Delinquency



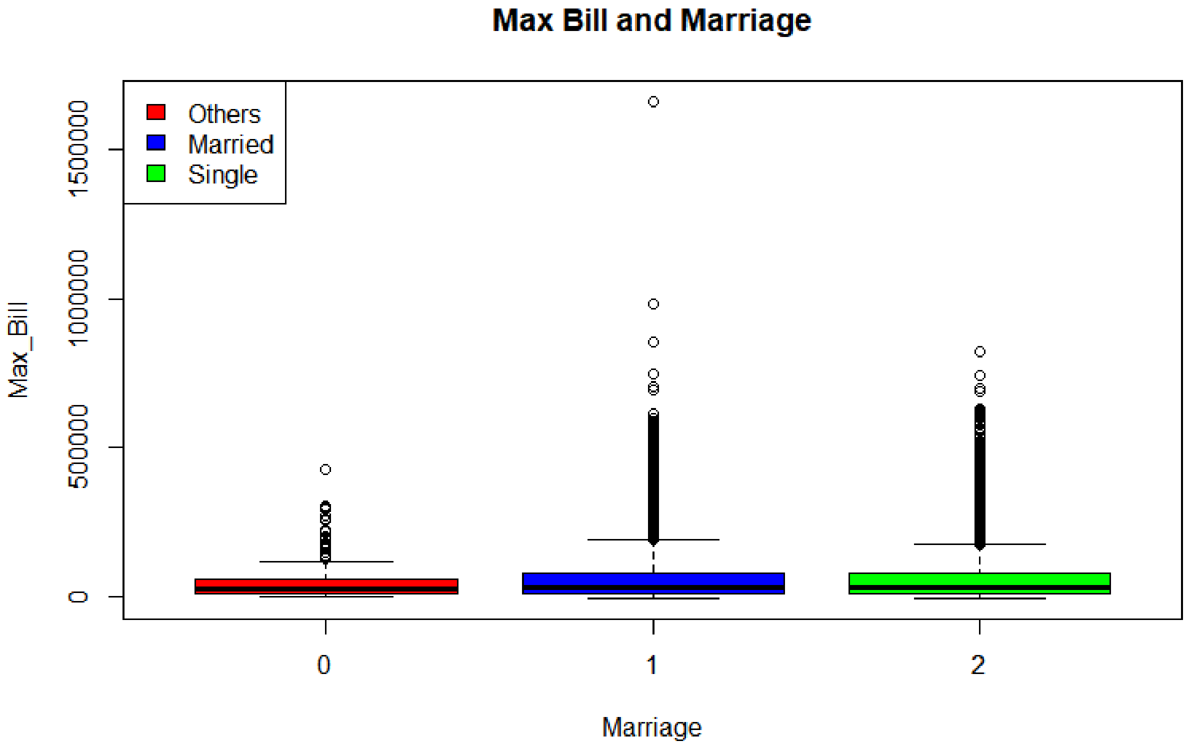
Below are the boxplots for two of our featured engineered variables, utilization sum and utilization growth over six months. The utilization sum had much more of a normal distribution with a few outliers on the high end. Utilization growth had a lot of outliers on both the high end and the low end.

Graph 24: Boxplots of Utilization Sum and Utilization Growth



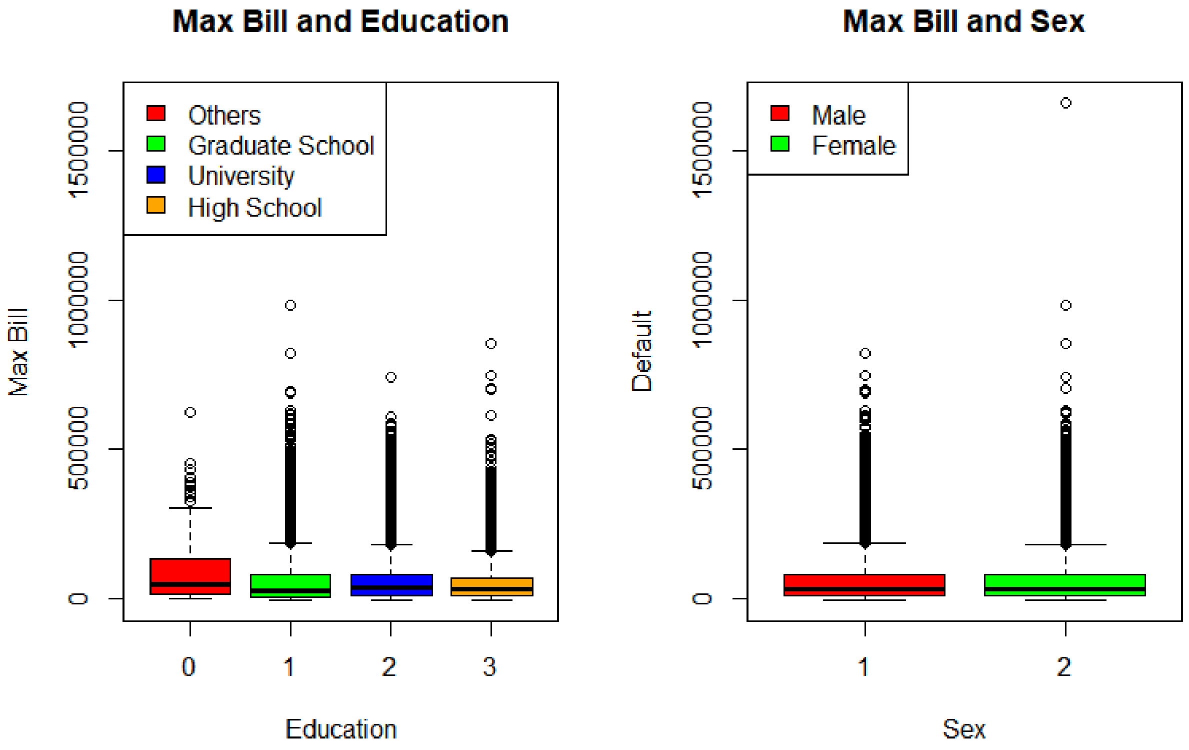
The plot below showed the max bill and marriage comparison. Max bill is one of the featured engineering variables and marriage is one of the original categorical variables. It looks like for people married and single they had very similar results with similar inter quartile ranges and outliers on the high end. The "others" category had far fewer outliers, but also far fewer observations.

Graph 25: Boxplot of Max Bill and Marriage



Below are two plots that look at the maximum bill and education as well as the maximum bill and sex. The common theme between the education was there were no outliers on the low end and there were a good number of outliers on the high end. For the sex and max bill amount graph there were similar type results as both boxplots were similar with a good number of outliers on the high end.

Graph 26: Boxplot of Max Bill with Education and Sex

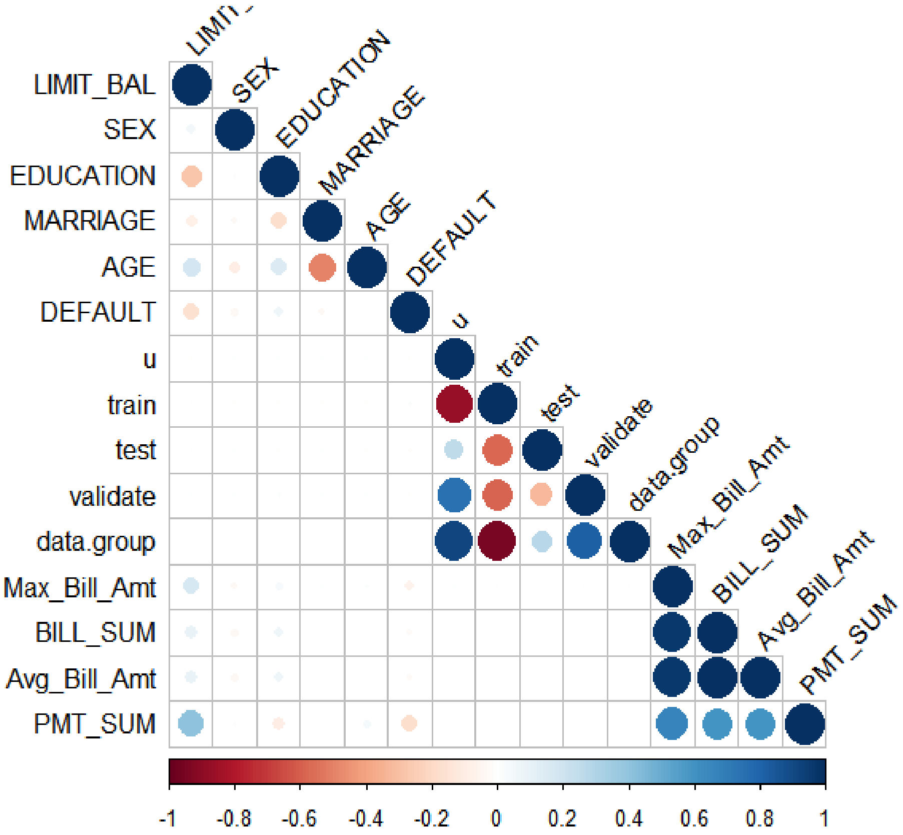


4.2-Model Based EDA

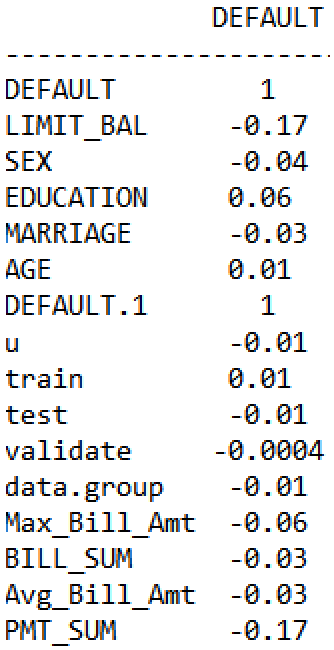
4.2.1-Model Based Correlations

The next step in this section is to look at the exploratory data analysis from a modeling perspective. The first step is to look at the correlations as they relate to the default variable. Below is the correlations and correlations plots for all the variables. The variables with the highest +- could be good indicators for predictors of default as they correlate with the default variable the most.

Graph 27: Correlation Plot 1 with Default

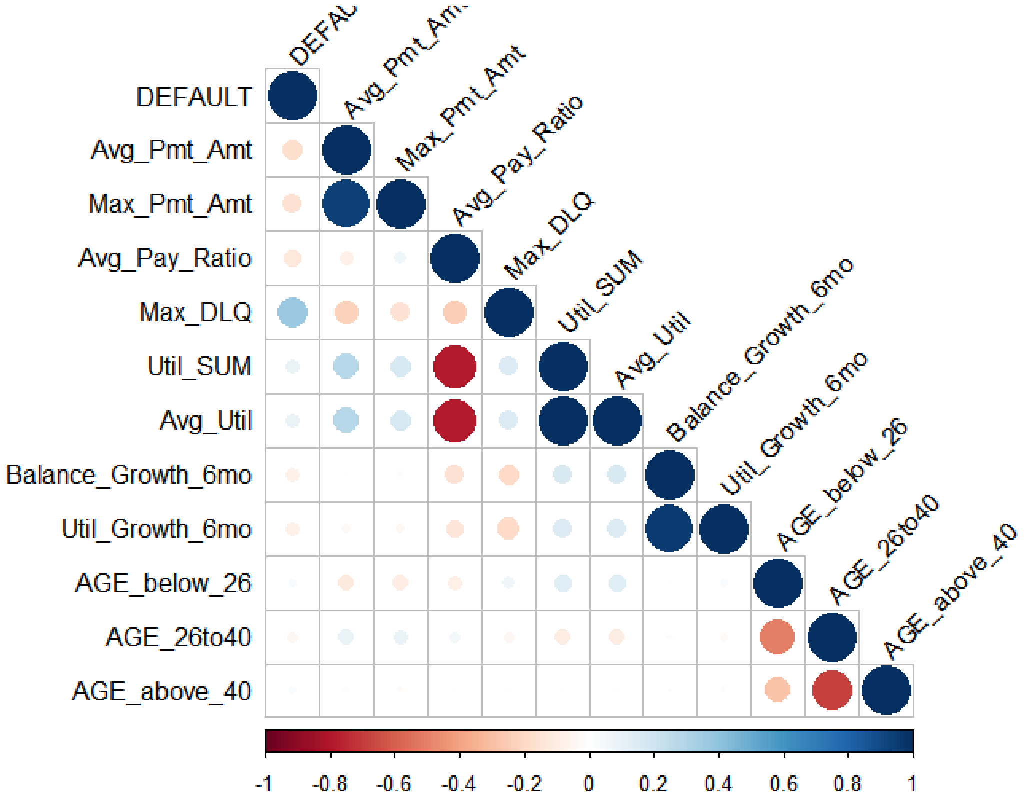


Graph 28: Correlation Table 1 with Default

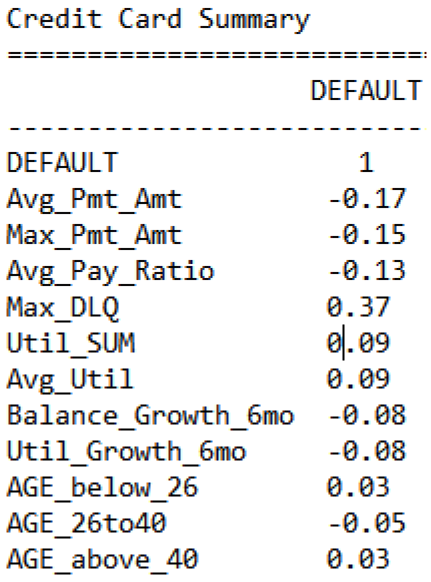


```
                DEFAULT
---------------------
DEFAULT           1
LIMIT_BAL       -0.17
SEX             -0.04
EDUCATION       0.06
MARRIAGE        -0.03
AGE             0.01
DEFAULT.1         1
u               -0.01
train           0.01
test            -0.01
validate       -0.0004
data.group      -0.01
Max_Bill_Amt    -0.06
BILL_SUM        -0.03
Avg_Bill_Amt    -0.03
PMT_SUM         -0.17
```

Graph 29: Correlation Plot 2 with Default



Graph 30: Correlation Table 2 with Default



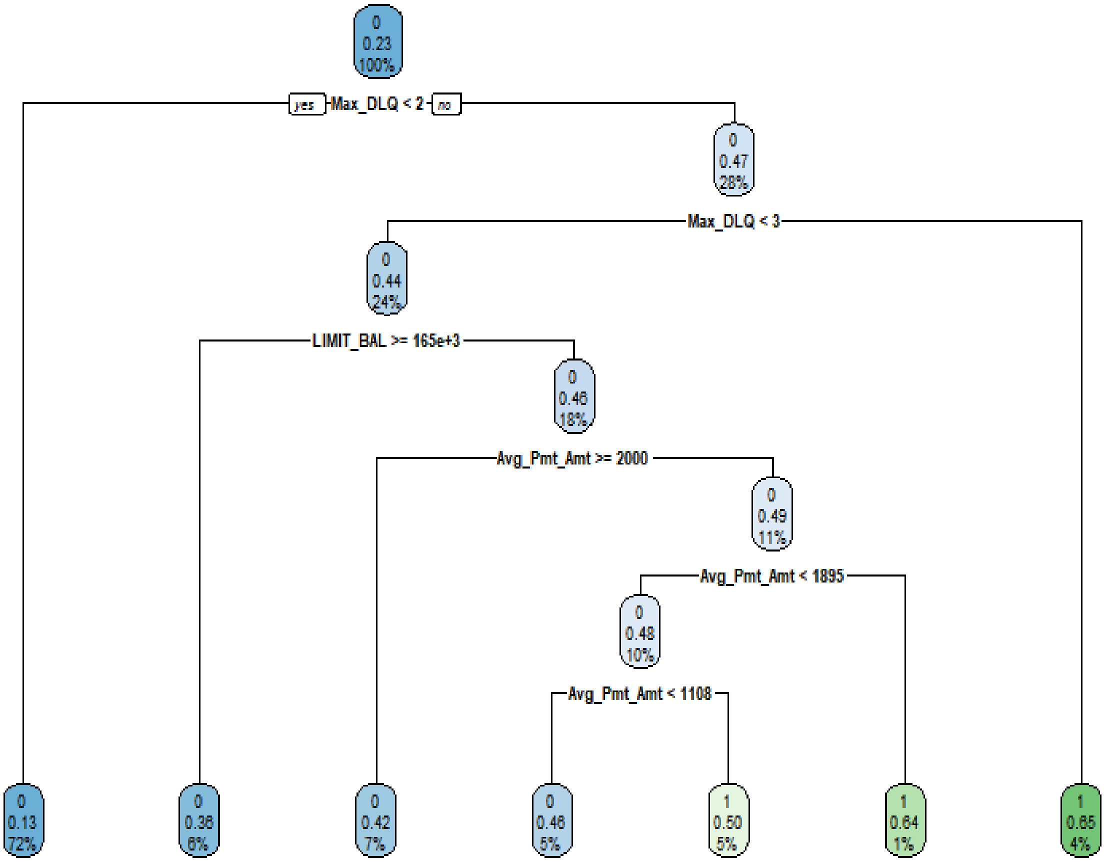
```
Credit Card Summary
==========================
                   DEFAULT
--------------------------
DEFAULT               1
Avg_Pmt_Amt         -0.17
Max_Pmt_Amt         -0.15
Avg_Pay_Ratio       -0.13
Max_DLQ             0.37
Util_SUM            0.09
Avg_Util            0.09
Balance_Growth_6mo  -0.08
Util_Growth_6mo     -0.08
AGE_below_26        0.03
AGE_26to40          -0.05
AGE_above_40        0.03
```

The next step is to fit a decision tree. In order to do this, let us look at the variables with the best +- correlations with default. There were four variables that had greater than a +- .17 correlation. The following variables will be used. Limit balance is at -.17, payment sum is at -.17, max delinquency at .37, and average payment amount at -.17. Next we will use these variables to plot a decision tree.

4.2.2-Model Based Decision Tree

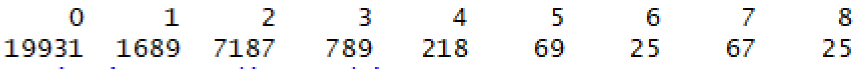
Looking at the below decision tree it looks like max delinquency is the best predictor for default. It looks like balance limit and the average payment amount are important as well, but max delinquency looks like the best predictor.

Figure 31: Important Variable Decision Tree



Below is the number of occurrences for max delinquency. The amount of people who did not default is about 2/3 of the total.

Figure 32: Table for Max Delinquency

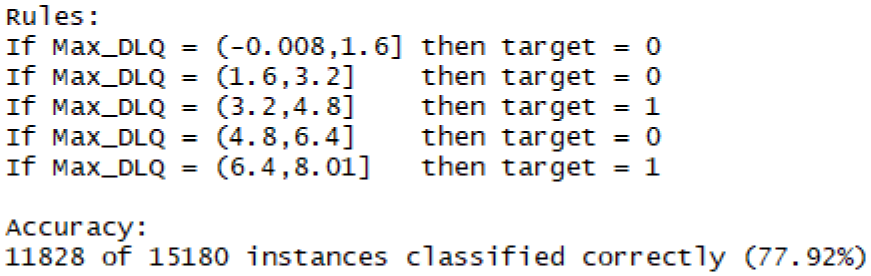


```
    0     1     2     3    4     5    6    7    8
19931  1689  7187   789  218    69   25   67   25
```

4.2.3-One R on Decision Tree

Using the one R function we can tell that max delinquency may be the most important variable for predicting default. When we ran the function on the decision tree it appears that it may be the best predictor of default. This may be a good thing to explore further as we begin our modeling in the next section.

Table 33: One R information on Decision Tree



```
Rules:
If Max_DLQ = (-0.008,1.6] then target = 0
If Max_DLQ = (1.6,3.2]    then target = 0
If Max_DLQ = (3.2,4.8]    then target = 1
If Max_DLQ = (4.8,6.4]    then target = 0
If Max_DLQ = (6.4,8.01]   then target = 1

Accuracy:
11828 of 15180 instances classified correctly (77.92%)
```

5. Predictive Modeling: Methods and Results

In this section we will create four different models. In the first section we will create a random forest model. In a random forest model, it consists of many decision trees. The decision trees then predict by committee. This makes resulting final trees more accurate than any individual tree. (Koehrsen). The second model is the gradient boosting model. A gradient boosting model is a machine learning model for a regression model which produces prediction models. It then builds on the prediction models and produces a main model that is produced by combining the prediction models (Grover). In both these models we will use all the engineered variables for our credit card data. In the third model we will use a logistic regression model. A logistic regression model uses predictor variables to see if they are good at predicting a response variable. In the fourth model we use a Naïve Bayes model. This is a probabilistic classifier that determines the probability of features occurring in their classification and it returns the most likely classification (Devins).

The entire list of variables will be used in the Random Forest and Gradient Boosting models. From there we will choose the most important variables for the Logistic Regression and Naïve Bayes Models.

Variables for Random Forest and Gradient Boosting Models-

Response Variable

- Default

Predictor Variables

- Limit Balance
- Sex
- Education
- Marriage
- Max Bill Amount
- Bill Sum
- Average Bill Amount
- Payment Sum
- Average Payment Amount
- Max Payment Amount
- Average Pay Ratio
- Max Delinquency
- Utility Sum
- Average Utility
- Balance Growth over 6 Months
- Utility Growth over 6 Months
- Age Below 26
- Age 26 to 40
- Age above 40

In this section we will be looking at many metrics for the models.

- True Positive Rate-To calculate this we add all the people that actually defaulted. We look at the number of true positive cases and false negative cases together. From there we divide the number of true positive case by the total number.
- AUC-ROC Curve- We will also look at the AUC and ROC curve for these models. This is a curve that will tell us the true positive rate versus the false positive rate. The higher the number the better.
- Precision-To calculate precision we take the True Positive and divide by the true positive + the false positive.
- Recall- Recall is calculated by taking the true positive and dividing by the true positive + the false negative.
- Sensitivity-This is the same as the recall formula. It is also known as the true positive rate.
- Specificity-This is also known as the true negative rate. It is calculated by taking the true negative and dividing by the true negative + the false positive.
- F1 Score- It is a measure of the model's accuracy. It is the average of the precision and the recall scores for the model.
- Model Accuracy-This is if the % of the target value that matched the predicted value (Drakos)

The table below is what we will be looking at for out model results:

Table 34: Model Performance Indicators

Performance Model	Predicted Non-Default	Predicted Default
Actual Non-Default	True Negative	False Positive
Actual Default	False Negative	True Positive

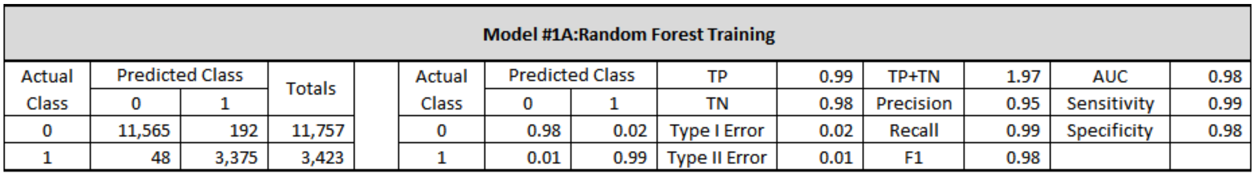
5.1 Random Forest

Random Forest is the first model that we are using. In this method we are going to use all the variables in our dataset. One of the goals of this model will be to figure out which variables are the most important. We will run the training set and then the test set and compare the results. Finally, we will list the variables that were the most influential on the model.

We can see below that for the training model there was a true positive rate for this model at .99. This model was very good as the AUC, Sensitivity, Specificity, Precision and Recall scores all scored at least a .95. Of the 15,180 predictions only 240 scored incorrectly.

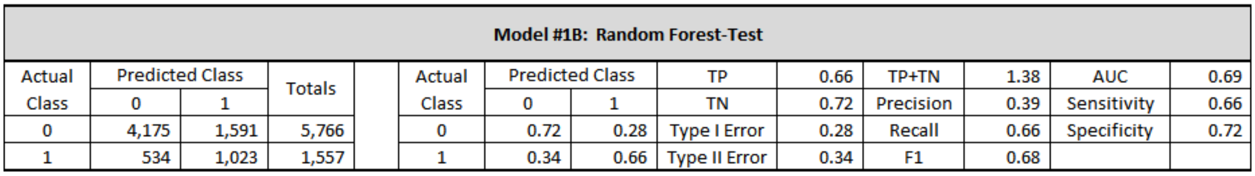
For the test model it did not score as well, but it was not terrible. The Sensitivity, Recall, and Specificity scores were all between .66 and .72. The precision score was much lower at .39. With the test model worse than the training model this could be a case of overfitting where the training model performs much better than the testing model because the parameters were too strict on the initial training model.

Table 35: Random Forest Model Training and Test Metrics:



Model #1A:Random Forest Training													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.99	TP+TN	1.97	AUC	0.98
Class	0	1			Class	0	1	TN	0.98	Precision	0.95	Sensitivity	0.99
0	11,565	192	11,757		0	0.98	0.02	Type I Error	0.02	Recall	0.99	Specificity	0.98
1	48	3,375	3,423		1	0.01	0.99	Type II Error	0.01	F1	0.98		

Table 1: Confusion matrix and classification metrics for Model #1A : Random Forest-Training

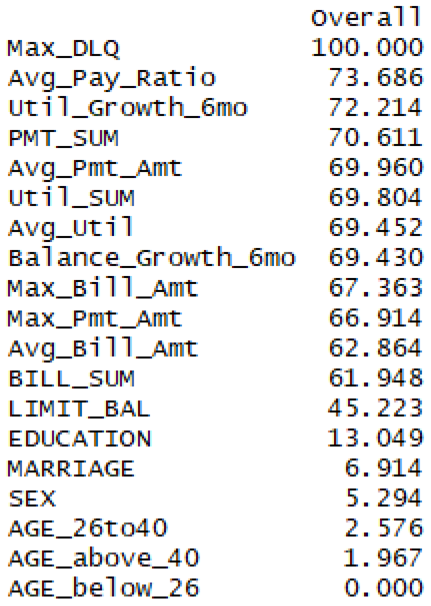


Model #1B: Random Forest-Test													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.66	TP+TN	1.38	AUC	0.69
Class	0	1			Class	0	1	TN	0.72	Precision	0.39	Sensitivity	0.66
0	4,175	1,591	5,766		0	0.72	0.28	Type I Error	0.28	Recall	0.66	Specificity	0.72
1	534	1,023	1,557		1	0.34	0.66	Type II Error	0.34	F1	0.68		

Table 2: Confusion matrix and classification metrics for Model #1B : Random Forest-Test

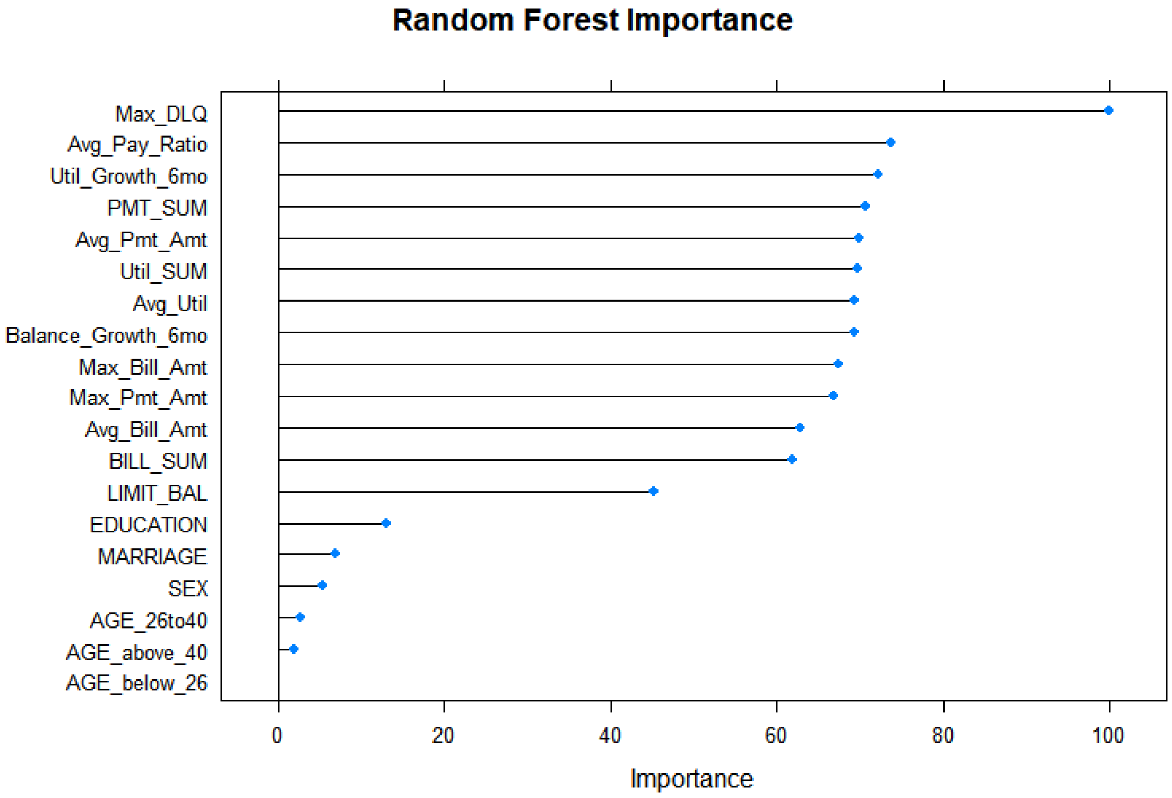
The final thing we want to do for the Random Forest model is look at the important variables that contributed to the model. There are several variables that are important. Max Delinquency appears more important than all the rest with a score of a 100. Average Pay Ratio, Utility Growth for 6 Months, Payment Sum, Average Payment Amount, Utility Sum, Average Utility, Balance Growth over 6 Months, Max Bill Amount, Max Payment Amount, Average Bill Amount, and Bill Sum were also all fairly important with scores between 61 and 74. One other interesting note was that the three age bins appear to have very little impact on the models as they were the lowest three scores of the variables.

Table 36: Importance for Random Forest Variables



```
                    Overall
Max_DLQ             100.000
Avg_Pay_Ratio        73.686
Util_Growth_6mo      72.214
PMT_SUM              70.611
Avg_Pmt_Amt          69.960
Util_SUM             69.804
Avg_Util             69.452
Balance_Growth_6mo   69.430
Max_Bill_Amt         67.363
Max_Pmt_Amt          66.914
Avg_Bill_Amt         62.864
BILL_SUM             61.948
LIMIT_BAL            45.223
EDUCATION            13.049
MARRIAGE              6.914
SEX                   5.294
AGE_26to40            2.576
AGE_above_40          1.967
AGE_below_26          0.000
```

Figure 37: Plot of Random Forest Variables

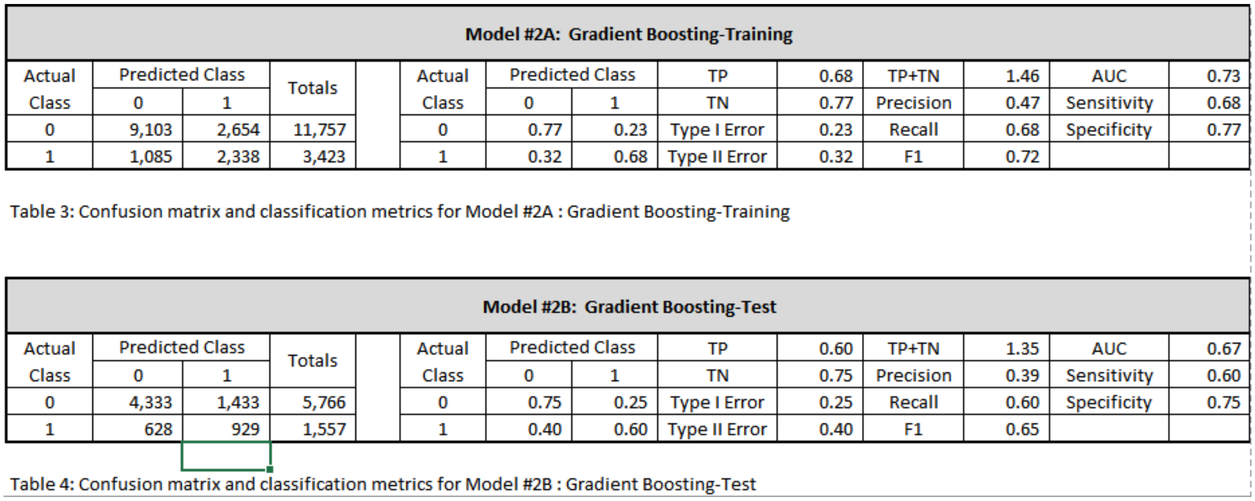


5.2 Gradient Boosting-

The second model we are going to use is the gradient boosting model. We are going to use the same variables as we did for the random forest model. After we run the models, we are going to see which variables had the greatest impact. The gradient boosting training model came in with the following measurements below. The training model had a true positive rate of .68, and an AUC of .73. The recall and specificity were also fairly high at .68 and .77 respectively. The precision was low at .47

For the test model it produced slightly worse results, but not by much. The true positive rate was at .60 and the AUC curve was at .67. The specificity was pretty high at .75. The precision was once again the lowest metric for the test model as well at .39.

Table 38: Gradient Boosting Model Training and Test Metrics



Model #2A: Gradient Boosting-Training													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.68	TP+TN	1.46	AUC	0.73
Class	0	1			Class	0	1	TN	0.77	Precision	0.47	Sensitivity	0.68
0	9,103	2,654	11,757		0	0.77	0.23	Type I Error	0.23	Recall	0.68	Specificity	0.77
1	1,085	2,338	3,423		1	0.32	0.68	Type II Error	0.32	F1	0.72		

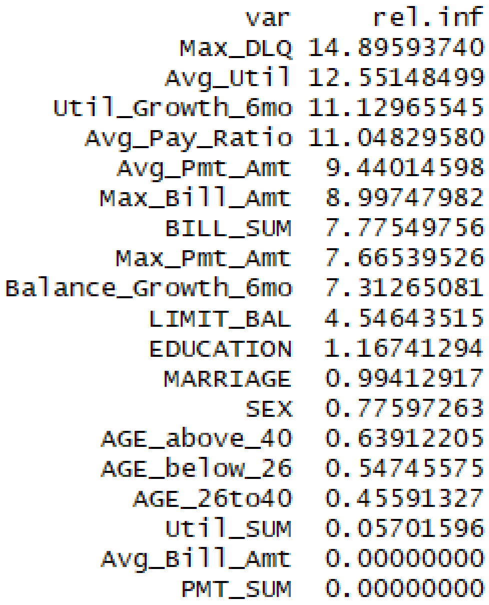
Table 3: Confusion matrix and classification metrics for Model #2A : Gradient Boosting-Training

Model #2B: Gradient Boosting-Test													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.60	TP+TN	1.35	AUC	0.67
Class	0	1			Class	0	1	TN	0.75	Precision	0.39	Sensitivity	0.60
0	4,333	1,433	5,766		0	0.75	0.25	Type I Error	0.25	Recall	0.60	Specificity	0.75
1	628	929	1,557		1	0.40	0.60	Type II Error	0.40	F1	0.65		

Table 4: Confusion matrix and classification metrics for Model #2B : Gradient Boosting-Test

The next step is to look at the variables that were important to the gradient boosting model. The most important variable was again max delinquency at 14.89. Other important variables included average utility, utility growth over 6 months, average pay ratio, average payment amount, max bill amount, bill sum, max payment amount, and balance growth over 6 months. All these variables scored between 7 and 13.

Table 39: Importance for Gradient Boosting Variables



```
               var     rel.inf
           Max_DLQ 14.89593740
          Avg_Util 12.55148499
   Util_Growth_6mo 11.12965545
     Avg_Pay_Ratio 11.04829580
       Avg_Pmt_Amt  9.44014598
      Max_Bill_Amt  8.99747982
          BILL_SUM  7.77549756
       Max_Pmt_Amt  7.66539526
Balance_Growth_6mo  7.31265081
         LIMIT_BAL  4.54643515
         EDUCATION  1.16741294
          MARRIAGE  0.99412917
               SEX  0.77597263
      AGE_above_40  0.63912205
      AGE_below_26  0.54745575
        AGE_26to40  0.45591327
          Util_SUM  0.05701596
      Avg_Bill_Amt  0.00000000
           PMT_SUM  0.00000000
```

Below are a list of the variables that are considered important in the random forest and gradient boosting models. All the highlighted variables have an importance in both models. There are a total of 9. We will use the variables that are important for both models as our variables that we use in models 3 and 4. These two models are the Logistic Regression and Naïve Bayes Models. Below is a chart that shows the final variable list for models 3 and 4.

Table 40: Variable Important to Random Forest

Random Forest
1. Max Delinquency
2. Average Pay Ratio
3. Utility Growth over 6 Months
4. Payment Sum
5.Average Payment Amount
6. Utility Sum
7. Average Utility
8. Balance Growth over 6 Months

9. Max Bill Amount
10. Max Payment Amount
11. Average Bill Amount
12. Bill Sum

Table 41: Variables Important to Gradient Boosting

Gradient Boosting
1. Max Delinquency
2. Average Utility
3. Utility Growth Rate per 6 Months
4. Average Pay Ratio
5. Average Payment Amount
6. Max Bill Amount
7. Bill Sum
8. Max Payment Amount
9. Balance Growth Over 6 Months

Table 42: Variables used for models 3 and 4

Variables used for models 3 and 4
1. Max Delinquency
2. Average Utility
3. Utility Growth Rate per 6 Months
4. Average Pay Ratio
5. Average Payment Amount
6. Max Bill Amount
7. Bill Sum
8. Max Payment Amount
9. Balance Growth Over 6 Months

5.3-Logistic Regression Model-

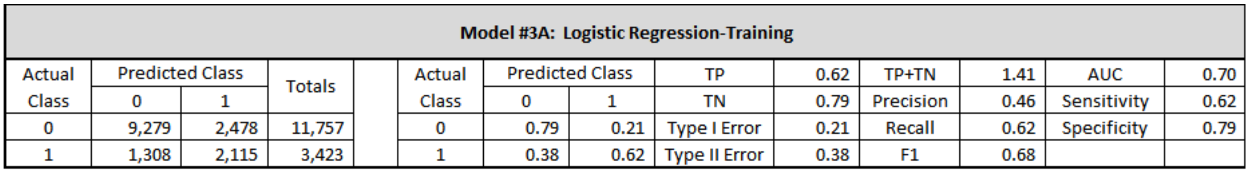
The third model is a logistic regression model. This model will only consist of the top variables by importance. The table above showed what variables are considered important from the Random Forest and Gradient Boosting Models.

We will use a stepwise method for this model. The stepwise model either adds or removes predictor variables one at a time depending on the variable's significance. It either adds the most significant variable or removes the most insignificant variable.

The training model had a true positive score of .62. The AUC came in at .70. The recall was at .62, the specificity was at .79, and the precision was at .46. This was an okay model, but not a great one.

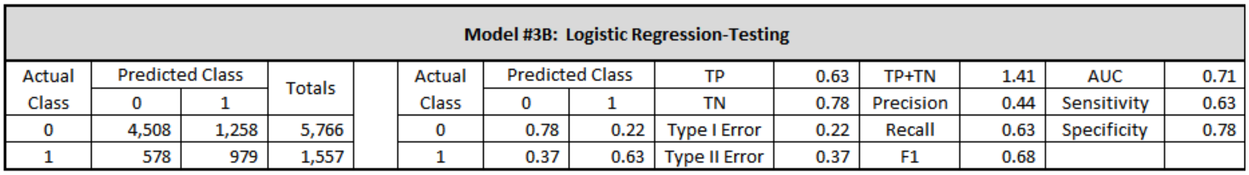
The test model for the logistic regression performed about like the training model did. In fact, all the scores were just about the same. The true positive rate was .63, the AUC was at .71, the recall was at .63 and the precision had a score of .44.

Table 43: Logistic Regression Model Training and Test Metrics



Model #3A: Logistic Regression-Training													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.62	TP+TN	1.41	AUC	0.70
Class	0	1			Class	0	1	TN	0.79	Precision	0.46	Sensitivity	0.62
0	9,279	2,478	11,757		0	0.79	0.21	Type I Error	0.21	Recall	0.62	Specificity	0.79
1	1,308	2,115	3,423		1	0.38	0.62	Type II Error	0.38	F1	0.68		

Table 5: Confusion matrix and classification metrics for Model #3A : Logistic Regression-Training

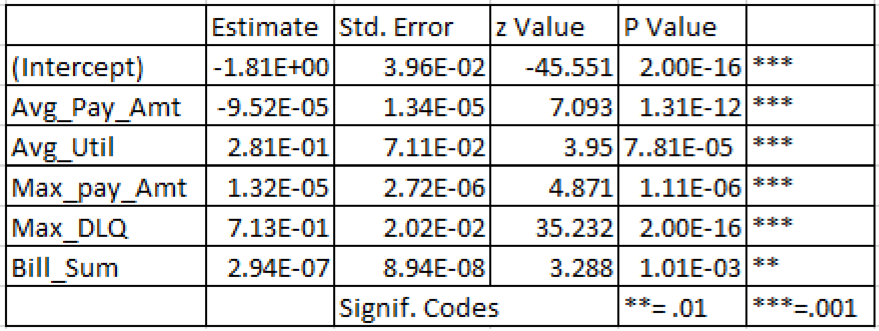


Model #3B: Logistic Regression-Testing													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.63	TP+TN	1.41	AUC	0.71
Class	0	1			Class	0	1	TN	0.78	Precision	0.44	Sensitivity	0.63
0	4,508	1,258	5,766		0	0.78	0.22	Type I Error	0.22	Recall	0.63	Specificity	0.78
1	578	979	1,557		1	0.37	0.63	Type II Error	0.37	F1	0.68		

Table 6: Confusion matrix and classification metrics for Model #3B : Logistic Regression-Test

Finally, we look below at the logistic model and all the variables. The variables with the ** indicate that the P value was less than .05 which means they are statistically significant. The bill sum variable falls into this category. The *** means the P value is less than .001 meaning that they are even more statistically significant. Four variables fall into this category and they are average payment amount, average utility, max payment amount, and max delinquency. There are also four variables that were removed by the stepwise function because they were insignificant to the model. The variables of average pay ratio, balance growth over 6 months, utility growth over six months and max bill amount are not considered important when predicting default.

Table 44: Output of Logistic Regression Model



	Estimate	Std. Error	z Value	P Value	
(Intercept)	-1.81E+00	3.96E-02	-45.551	2.00E-16	***
Avg_Pay_Amt	-9.52E-05	1.34E-05	7.093	1.31E-12	***
Avg_Util	2.81E-01	7.11E-02	3.95	7..81E-05	***
Max_pay_Amt	1.32E-05	2.72E-06	4.871	1.11E-06	***
Max_DLQ	7.13E-01	2.02E-02	35.232	2.00E-16	***
Bill_Sum	2.94E-07	8.94E-08	3.288	1.01E-03	**
		Signif. Codes		**= .01	***=.001

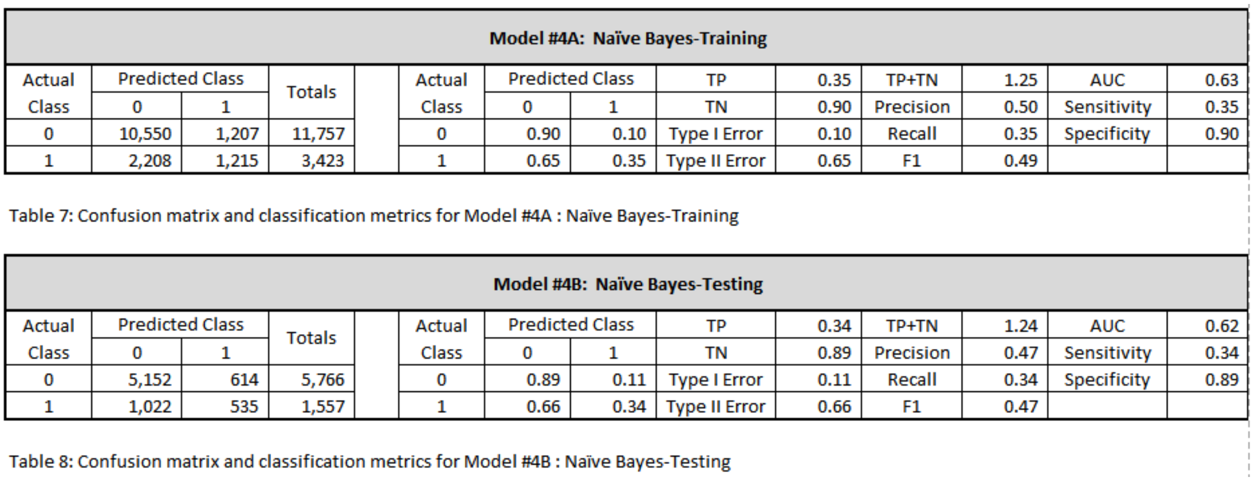
5.4 Naves Bayes Model-

In the fourth model we will look at the Naves Bayes Model. We will look at the same nine variables that we looked at in the logistic regression model and look at how the model can predict the target value.

The below Naïve Bayes training model had some good results and some okay results. It correctly predicted most of the true negatives with a score of 0.90, however the true positives were only predicted correctly at .35. The AUC score was at .63.

The testing model had similar scores as the training set did. It predicted .89 of the true negatives, but only .35 of the true positive results. The AUC score was about where the training model is at .62.

Table 45: Naïve Bayes Model Training and Test Metrics



Model #4A: Naïve Bayes-Training													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.35	TP+TN	1.25	AUC	0.63
Class	0	1			Class	0	1	TN	0.90	Precision	0.50	Sensitivity	0.35
0	10,550	1,207	11,757		0	0.90	0.10	Type I Error	0.10	Recall	0.35	Specificity	0.90
1	2,208	1,215	3,423		1	0.65	0.35	Type II Error	0.65	F1	0.49		

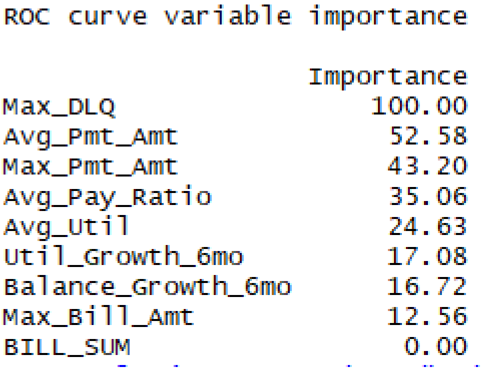
Table 7: Confusion matrix and classification metrics for Model #4A : Naïve Bayes-Training

Model #4B: Naïve Bayes-Testing													
Actual	Predicted Class		Totals		Actual	Predicted Class		TP	0.34	TP+TN	1.24	AUC	0.62
Class	0	1			Class	0	1	TN	0.89	Precision	0.47	Sensitivity	0.34
0	5,152	614	5,766		0	0.89	0.11	Type I Error	0.11	Recall	0.34	Specificity	0.89
1	1,022	535	1,557		1	0.66	0.34	Type II Error	0.66	F1	0.47		

Table 8: Confusion matrix and classification metrics for Model #4B : Naïve Bayes-Testing

Below is a list of the variables for the Naïve Bayes model with their importance for the model. By far the most important variable was Max Delinquency with a score of 100. Average payment amount, max payment amount, and average pay ratio were also important. One important thing was that bill sum had a score of 0 and was the least important variable for this model.

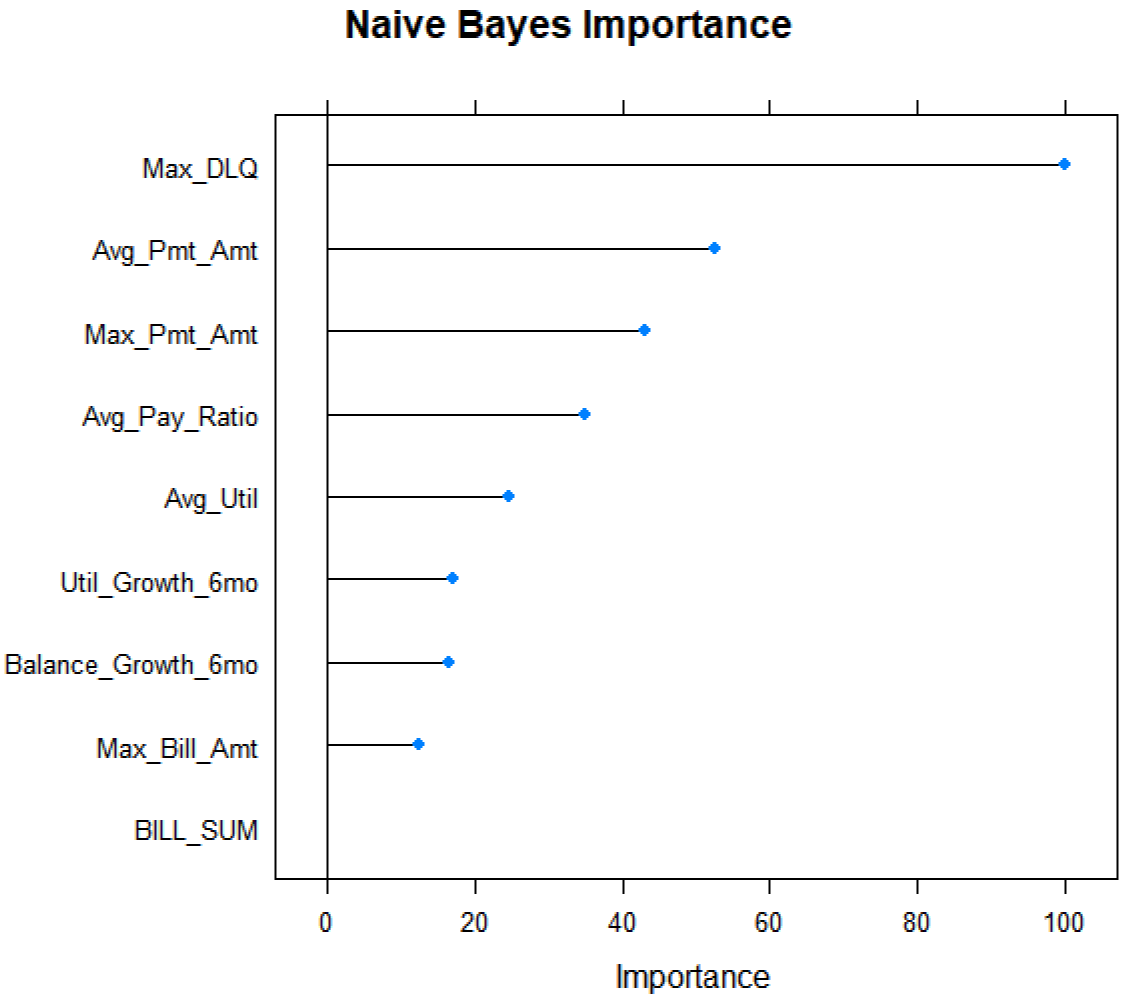
Table 46: Importance for Naïve Bayes Variables



```
ROC curve variable importance

                   Importance
Max_DLQ                100.00
Avg_Pmt_Amt             52.58
Max_Pmt_Amt             43.20
Avg_Pay_Ratio           35.06
Avg_Util                24.63
Util_Growth_6mo         17.08
Balance_Growth_6mo      16.72
Max_Bill_Amt            12.56
BILL_SUM                 0.00
```

Figure 47: Plot of Naïve Bayes Variables



6.1 Comparison of Results Variable Importance

In this section we will look again at all four models and see which variables are the best predictors of default. The below chart shows the variables that were considered the most important. These variables were used in all four models. If we look at models 3 and 4, the Naïve Bayes and the Logistic Regression models they both had the same order for the top three variables. Max Delinquency was first, average payment amount was second, and max payment amount was third. These variables were important for all four of the models. The most important variable in every model was Max Delinquency. This variable seems like the best predictor of default.

Table 48: Most Important Variables

Important Variables in all four models
1. Max Delinquency
2. Average Payment Amount
3. Max Payment Amount

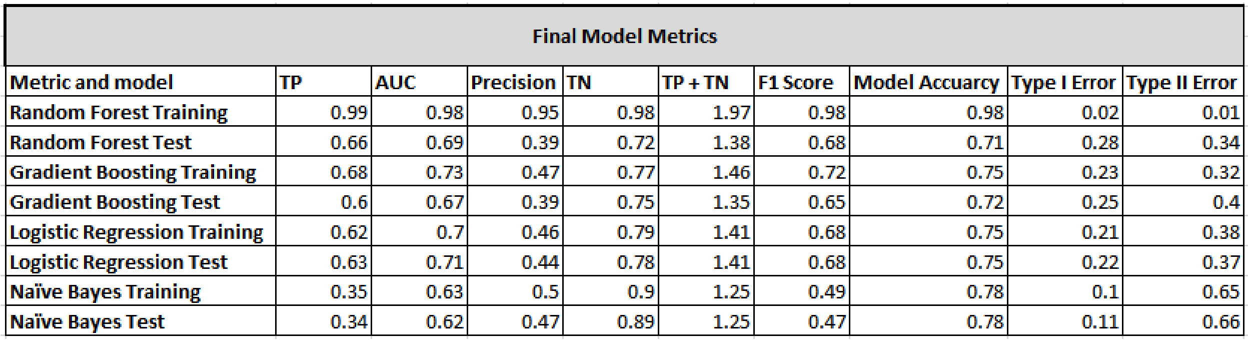
6.2 Comparison of Metrics for Models

In this section we want to take a closer look at all the models and the metrics that were associated with them. We will look at the following metrics.

- True Positive Rate- (also known as recall and sensitivity)
- AUC
- Precision-
- True Negative Rate- (This is also known as specificity)
- True Positive + True Negative
- F1 Score
- Model Accuracy-
- Type 1 Error
- Type 2 Error

Below is a chart with the comparison of all four models and how they performed for all the metrics.

Table 49: Final Model Metrics



Final Model Metrics

Metric and model	TP	AUC	Precision	TN	TP + TN	F1 Score	Model Accuarcy	Type I Error	Type II Error
Random Forest Training	0.99	0.98	0.95	0.98	1.97	0.98	0.98	0.02	0.01
Random Forest Test	0.66	0.69	0.39	0.72	1.38	0.68	0.71	0.28	0.34
Gradient Boosting Training	0.68	0.73	0.47	0.77	1.46	0.72	0.75	0.23	0.32
Gradient Boosting Test	0.6	0.67	0.39	0.75	1.35	0.65	0.72	0.25	0.4
Logistic Regression Training	0.62	0.7	0.46	0.79	1.41	0.68	0.75	0.21	0.38
Logistic Regression Test	0.63	0.71	0.44	0.78	1.41	0.68	0.75	0.22	0.37
Naïve Bayes Training	0.35	0.63	0.5	0.9	1.25	0.49	0.78	0.1	0.65
Naïve Bayes Test	0.34	0.62	0.47	0.89	1.25	0.47	0.78	0.11	0.66

There are a couple important things that we should take from these results

- The random forest training model performed very well on all the metrics. All the metrics were at least .95% accurate.
- The random forest test model did not perform nearly as well as most of the scores were around the .6 and .7 range. The model accuracy went from a .98 on the training model to a .71 on the test model.
- The random forest model could be a case of overfitting with the training model since the test model did so much worse.
- The gradient boosting training model performed pretty well with a 75% model accuracy. Most of the metrics were in the .6 or .7 range.
- The gradient boosting test model performed about the same as the training. It came up with a 72% model accuracy rate.
- The logistic regression training model produced a decent model with an accuracy of .75. Most of the metrics were around the .7 range.
- The logistic regression test model produced a model that had almost identical results to the training model. The model had the exact same accuracy at 75%. All the results were within 1% or 2% of the training set.
- The Naïve Bayes training model produced a pretty good model at 78%. It was very good at predicting the true negatives at .90 and not that great at predicting the true positives at .35.
- The Naïve Bayes testing model also had just about the same exact results as the training model. It had the same accuracy at 78%. Most of the metrics were only off by about 1% or 2%. It had similar results in predicting the true positive rate and the true negative rate as the training model. It was at .89 for the true negative rate and .34 for the true positive rate.

All the models produced at least decent results with the accuracy being at least .71 for the lowest model. By far the most accurate was the random forest training model. The other 7 all had similar results as it pertained to model accuracy. If I had to rank the models. I would rank them in the following order.

1. Random Forest-Even though the test model did not do as well, the training model did great and if you look at combined accuracy this model produced the highest results.
2. Naïve Bayes-This model produced the second highest results as it pertains to model accuracy with both the training and the test being at .78. It was very good at predicting true positive results and not great at predicting true negative results.
3. Logistic Regression-This training and test models both produced similar results, and both had model accuracy at .75.
4. Gradient Boosting-This was not a bad model, but it seemed to perform slightly worse than the other three. The combined accuracy of both models was the lowest combined % with the training being at 75% and the test being at 72%.

7. Conclusion

For this capstone project we used 30,000 customers from Taiwan to predict defaults from credit cards. The information was from 2005 and was over six months. There were categorical variables such as age, sex, marriage and education. There were also numerical variables that were either used or variables that we combine in the feature engineering section to make new variables. Some of the variables included max delinquency, average pay ratio, average payment amount, and max bill amount. The dataset was grouped into three groups. About half of the data was in the training set, about a quarter was in the test set, and the final quarter was in the validation set. For our modeling we made four different models which included the random forest model, the gradient boosting model, the logistic regression model, and the naïve bayes model. The random forest model produced the highest accuracy. The training model came in at 98%. The rest of the models produced accuracy that was in between 70% and 80%. It was interesting that even though the models produced different true positive and true negative rates the final accuracy was about the same for seven of the eight models. I think the fact that all the models were at least 70% accurate shows that are results were pretty good.

I think the capstone project was very valuable. I think one thing that could be done to try and better the models is to look at a couple other models and see how they perform. I think it would be good to look at neural networks and support vector machines. I think different variables could be created in the feature engineering section which could lead to better predictors of default. I think it would be good in the future to take a set of different featured engineering variables and see how they perform compared to the current set of variables.

Bibliography

Devins. (2019, July 16). Introduction to Naive Bayes Classification. Retrieved from https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

Drakos, George (2018, September 12). How to select the Right Evaluation Metric for Machine Learning Models: Part 3 Classification Metrics. Retrieved from https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-3-classification-3eac420ec991

Grover. (2019, August 01). Gradient Boosting from scratch. Retrieved from https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d

Koehrsen, William . (2017, December 27). Random Forest Simple Explanation. Retrieved from https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d