A DESCRIPTIVE ANALYSIS ON CREDIT CARD DATA SET

A MID TERM PROJECT
ON
STAT 451/551 PREDICTIVE ANALYTICS I

Date : 3/13/2019

SUBMITTED BY:
BASANTA CHALISE
DIVYA SINHA
SANTOSH CHAPAGAIN
WEI GU

SUBMITTED TO:
DR. THOMAS BRANDENBURGER

List of Figures

List of Tables

Content

1. Introduction of the Dataset

The data contains information on customers' credit card monthly statement for 2010 of an anonymous business firm located in Sioux Falls. Specifically, it has 97,465 observations on 26 different variables. The data features 9997 different customer for which the observation on the following variables were recorded.

Table 1: Variables and their Description for given data set

Varibale Name	Description
DebtDimId	Customer identification number
Open Date	Date the credit card was opened
Row Num	A calculated field to index row number for each customer
Last Statement Date	Date of the last monthly statement or bill
Cycle Date	Date that the monthly billing cycle starts again
Months On Book	Number of months the credit card has been open
External Status	Status of the account
Days Deliq	Number of days the monthly bill is 'past due'.
Credit Limit	Limit to how much credit can be borrowed by the customer
Opening Balance	Balance on the card at the beginning date of the statement cycle
Ending Balance	Balance on the card at the ending date of the statement cycle
Over limit Amount	Amount credit limit has been exceeded.
Actual Min Pay Due	Amount of required minimum payment
Total Min Pay Due	Amount of required minimum payment
Net Payments During Cycle	Net Payments during this billing statement cycle
Net Purchases During Cycle	Net Purchases during this billing statement cycle
Net Cash Advances During Cycle	Net cash advances during this billing statement cycle
Net Premier Fees Billed During Cycle	
Net Behavior Fees Billed During Cycle	
Net Concessions Billed During Cycle	
Closure Reason	Reason the account was closed
Month End Date	Last day of the month
Last Payment Date	Date of last payment
Quarterly Fico Score	Quarterly FICO score of the customer
Behavior Score	Behavior Score
Good Customer Score	Good Customer Score

On this data set, we have account information on each customer at the end of every billing cycle and we are interested on finding the customer that are good for the credit card company. The definition of good for each customer is based on the account status. If the account status for the customer is open, then they are good customer while the customer that has account status other than open are bad customer. For every transaction made by the customer, they will incur some balances to pay and they have time for paying the debt. If the customer made a payment on time, then the days delinquency is zero for them while for those customers that doesn't make timely payment, they have certain days delinquency and thus have different external statuses: E for revoked, F for Frozen, I for Interest Prohibited, and Z for Charged Off. Similarly, the account status is also dependent on different variables including Net Payments During Cycle, Actual Min Pay Due, Quarterly FICO Score, Behavior Score and so on. The ultimate task is determining who are a good or bad customer based on their behavior for the period the data has been collected. We will begin our

analysis by exploring the data and producing the descriptive statistics and visualization on variables that seems more important in predicting the account status.

2. Data Exploration

The table 1 shows the variables along with the description of the variables for which the data was collected for every customer. The description for three variables were missing. The observations include different data types including continuous, categorical and date variables. The provided data set has several missing variables denoted by NA. It doesn't have any values for those account which has External Status as Open.

As part of data pre-processing, we need to remove those rows that has missing variables. Initial data exploration reveals that there were 5963 observation has NA values and thus are dropped. Similarly, for data redundancy we check if there are multiple rows that has same value or not and found that there were 14 of the observations that has duplicate entry. These observations were again dropped for further analysis. Thus, the data pre-processing process yield at total of 91486 observations for 9997 different customers. There were several NULL Values in Good Customer Score and these values were retained in the data set and will be converted into categorical variable based on the score range on second and third stage of analysis. These NULL values were excluded while calculating the summary statistics and the summary statistics for the data is shown in table below and are discussed below based on the data types.

Table 2: Summary Statistics for Data Set

Variables	Min	1st Quart	Median	Mean	3rd Quart	Max
Open Date	10/4/1996					2/10/2010
Last Statement Date	1/5/2010					12/29/2010
Cycle Date	3/4/2010					12/31/2010
Months on Book	1	11	20	30	40	170
Days Delinquent	0	0	0	18	30	270
Credit Limit	$0	$250	$300	$327	$350	$2,500
Opening Balance	-$2,078	$186	$263	$264	$342	$3,026
Ending Balance	-$2,078	$180	$262	$261	$344	$3,026
Over Limit Amount	$0	$0	$0	$20	$11	$2,826
Actual Minimum Pay Due	$0	$20	$25	$36	$40	$720
Total Minimum Pay Due	$0	$20	$25	$53	$50	$3,026
Net Payment During Cycle	-$1,122	$0	$30	$57	$63	$2,834
Net Purchases During Cycle	-$270	$0	$0	$4	$46	$2,416
Net Cash Advances During Cycle	$0	$0	$0	$1	$0	$542
Net Premier Fees Billed During Cycle	-$235	$7	$9	$14	$14	$203
Net Behavior Fees Billed During Cycle	-$119	$2	$5	$11	$14	$132
Net Concessions Billed During Cycle	-$612	$0	$0	$1	$0	$334
Last Payment Date	1/1/1900					12/31/2010
Quarterly FICO Score	0	521	581	566	632	811
Behavior Score	13	614	648	612	670	724
Good Customer Score	0	690	727	749	797	1000

i) Analysis of External Status and Closure Reason

We are interested in finding the customer that are good and bad based on their behavior. Customers with open external status are good customers while the customer whose account status is other than open are treated as bad as they have different status as mentioned earlier. The bar chart below shows the customer number with respect to account status.

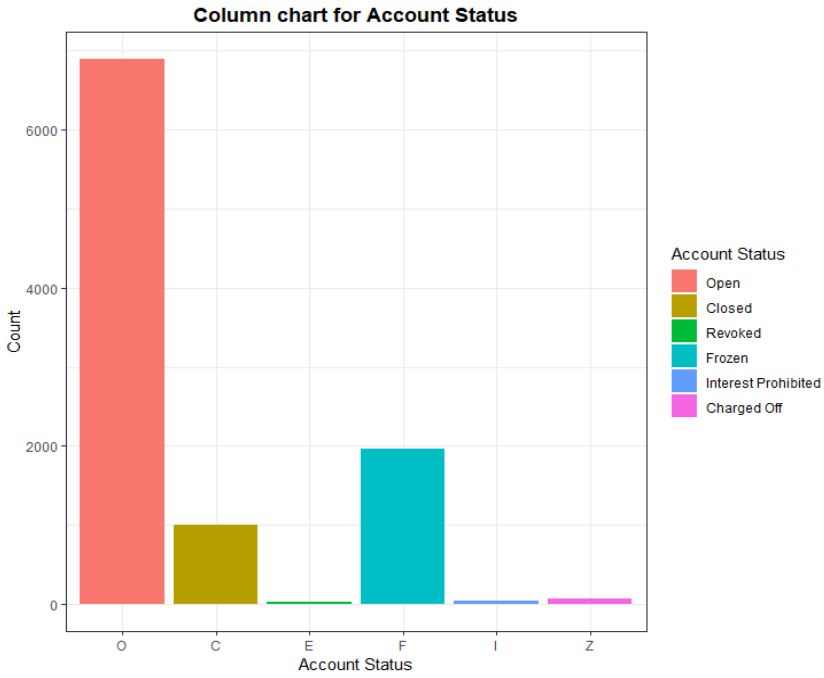

Figure 1: Bar Chart for account status

Table 3: Customer Number by External Status

O	C	E	F	I	Z
6894	1008	22	1961	39	73

It shows that there are 6894 customer who account was opened till last cycle date while there were 1008 customer whose account was closed. Similarly, there are 22 customers with revoked status, 1961 customer with Frozen status, 39 customers with Interest Prohibited status and 73 customers with Charged Off status.

For those customers whose account is closed, they have mentioned the reason for closing their account. Exploring the reasons and dealing with the them may entice them back will result some profit to the firm. To analyze the reason for closure, we have grouped the customer based on their reason and the bar chart below shows the count for different reasons.

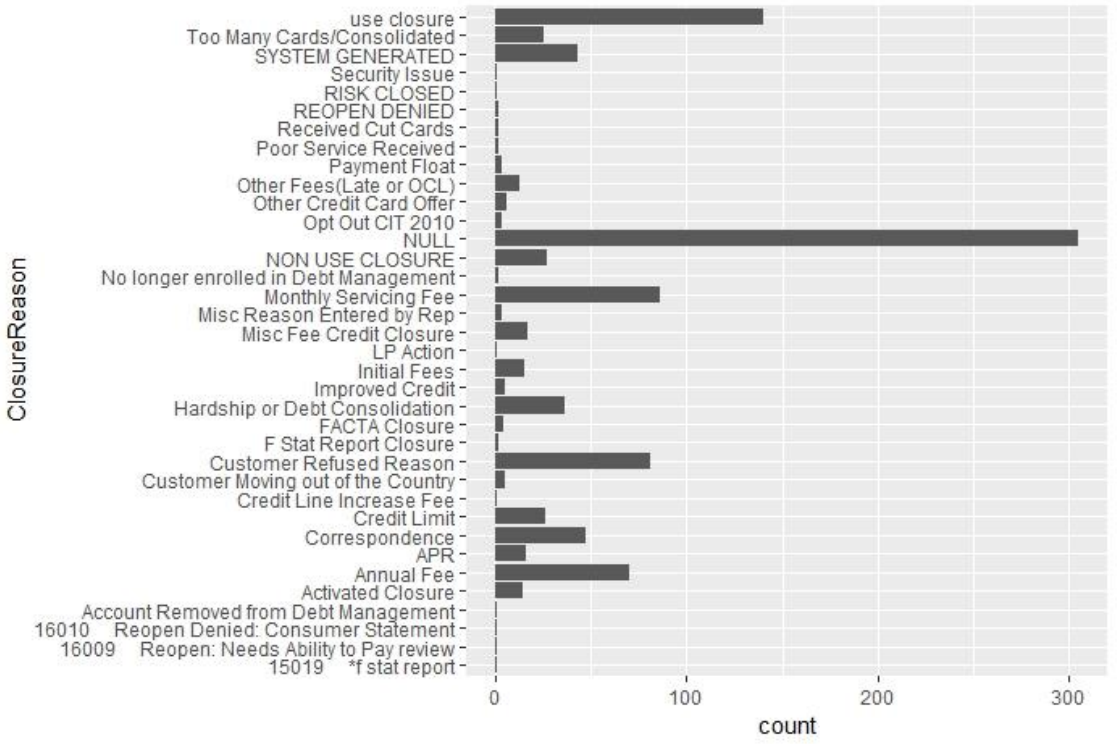

Figure 2: Bar chart for Closure Reason

The bar chart shows that most of customer didn't provide the closure reason. Some of the notable reason for closure include Monthly Servicing Fee, Customer Refused Reason, Annual Fee, Correspondence, APR, Hardship, Credit Limit and so on.

ii) **Account Opening Data and Months Card has been Used**

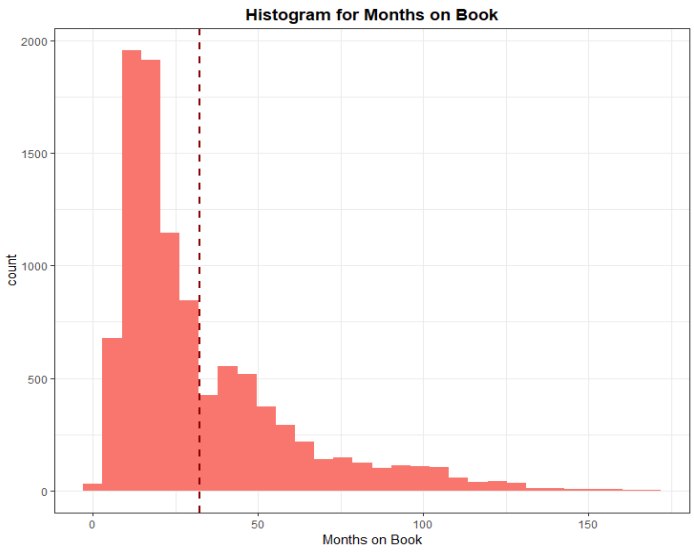

Figure 3: Histogram for Months on Book

From the summary statistics we noticed that the oldest account was dated on 10/4/1996 and the latest account was opened on 2/10/2010. Based on the data set, there were several customers whose external status was closed for several months and they have finally paid off their balances after several months. For analyzing the months card has been used we have considered the status for latest for each customer. The result obtained is presented on the histogram below.

From histogram we can see variable is skewed toward right with lower number of customers using card for more than 150 months from the data account has been opened while there are large number of customers using the card for less than the average month since the account has opened. The average month that the card has been used is around 30 months for all the customers.

iii) Credit Limit

Another most important variable in this analysis is credit limit. Several customers have different economic status and thus will have different credit limit. The credit limit does provide more information about the spending behavior of each customer. For the given data set, we looked after the credit limit for every status despite the external status of their account. The result is presented in the figure below.

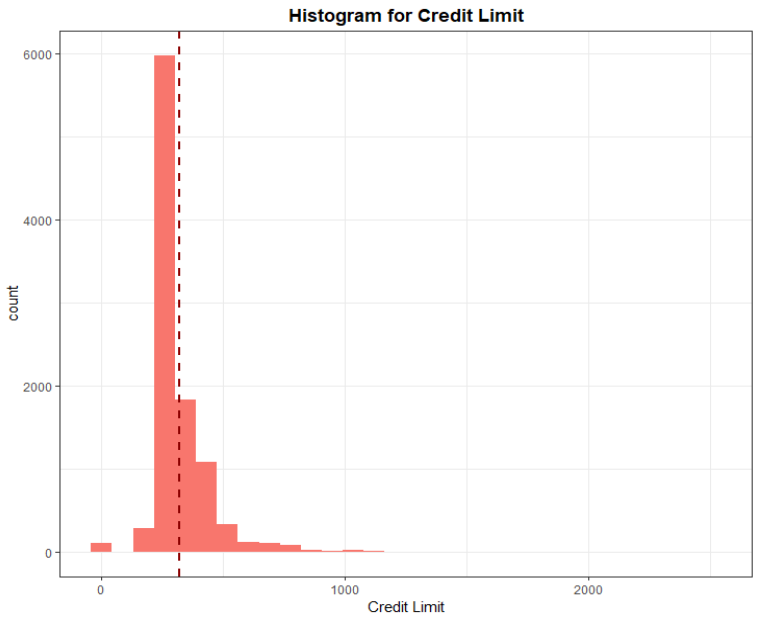

Figure 4: Histogram for Credit Limit

The histogram suggests the variable credit limit has normal distribution with mean credit limit at around $300 which is shown by dotted vertical line on the figure. We further investigated whether the credit limit has increased for customer on the given period and found that none of them has increased credit limit by the end of Year 2010 meaning that their limit stays same from bringing of 2010.

We suspect the account status is too dependent on the Over limit Amount i.e the amount that is being spent more than the credit limit. The descriptive statistics for the Over limit Amount shows average over limit is $20 with maximum of $ 2826 and a minimum value of 0. The boxplot summaries the over limit amount for different external status. The box plot shows there are numerous observations beyond the interquartile range for all external status except status of Z. Further it suggests that, the mean over limit amount for F status is higher followed by E and I. The mean over limit amount for closed and open status is quite close to 0 which shows that these customers were more cautious while spending the amount.

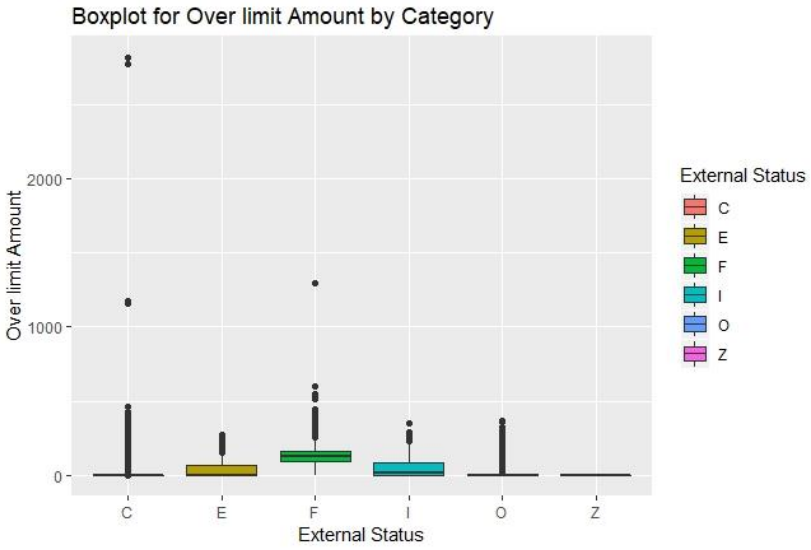

Figure 5: Box plot for Over limit Amount by category

iv) Delinquency Days

The account status and the Score (FICO and Behavior) are most related with the behavior of the customer. One of the most promising variables used for determining the status is Days Delinquency i.e number of days the monthly bill is 'past due'. On the given data set, we can the days delinquency are multiple of 30. On the data set, there are some customers that has higher days delinquency of around 270 while most of them has 0. The mean delinquency day is 18. We tried to explore the possible association of days delinquency with the external status by plotting the box plot between these two variables. The box plot is shown below.

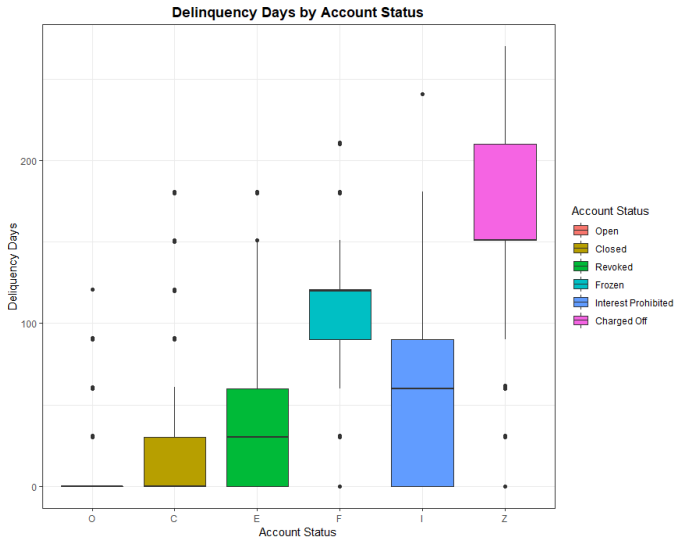

Figure 6: Delinquency Days by Account Status

From boxplot, the average delinquency for the open account status is 0 which indicates that these customers are quite good throughout the time they have used card while the delinquency days for customer with status other than open are quite high which indicates that days delinquency strong relationship for predicting the external status. The average days for status Z is high followed by F , I and E.

v) Score

Credit Score, Behavior score and Good Customer score are another important variable that has meaningful relation between the status. The box plot between the Quarterly FICO score and Behavior score is plotted to see the trend of these scores on each account status.

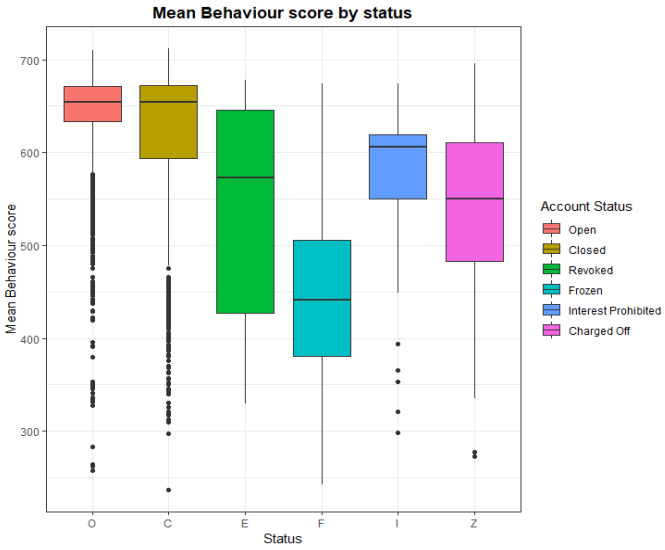

Figure 7: Box plot for Mean Behavior Score by Status

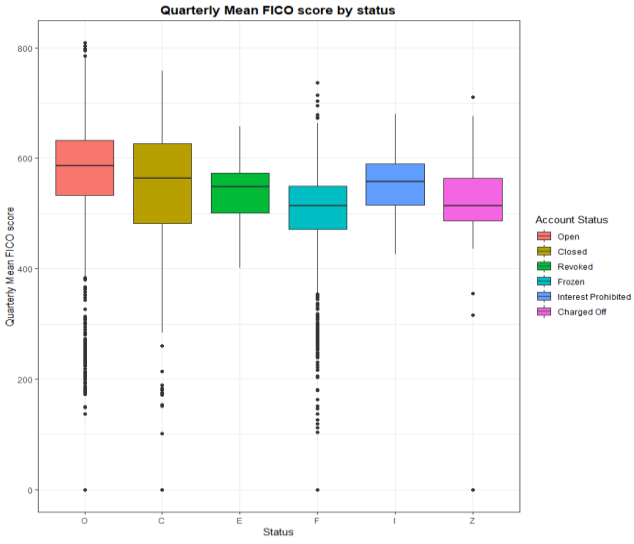

Figure 8: Quarterly Mean FICO Score by Status

The box plot shows that the mean behavior score for closed and open external status are high and has a mean value around 650 for both while the mean values for other status are comparatively low. This suggest that the behavior score is related with the account status too.

The box plot for quarterly FICO score suggest that mean quarterly FICO score for good customer i.e. whose account status is open are high compared to other account status. Similarly, the average score is consistently low for customer whose status is other than open.

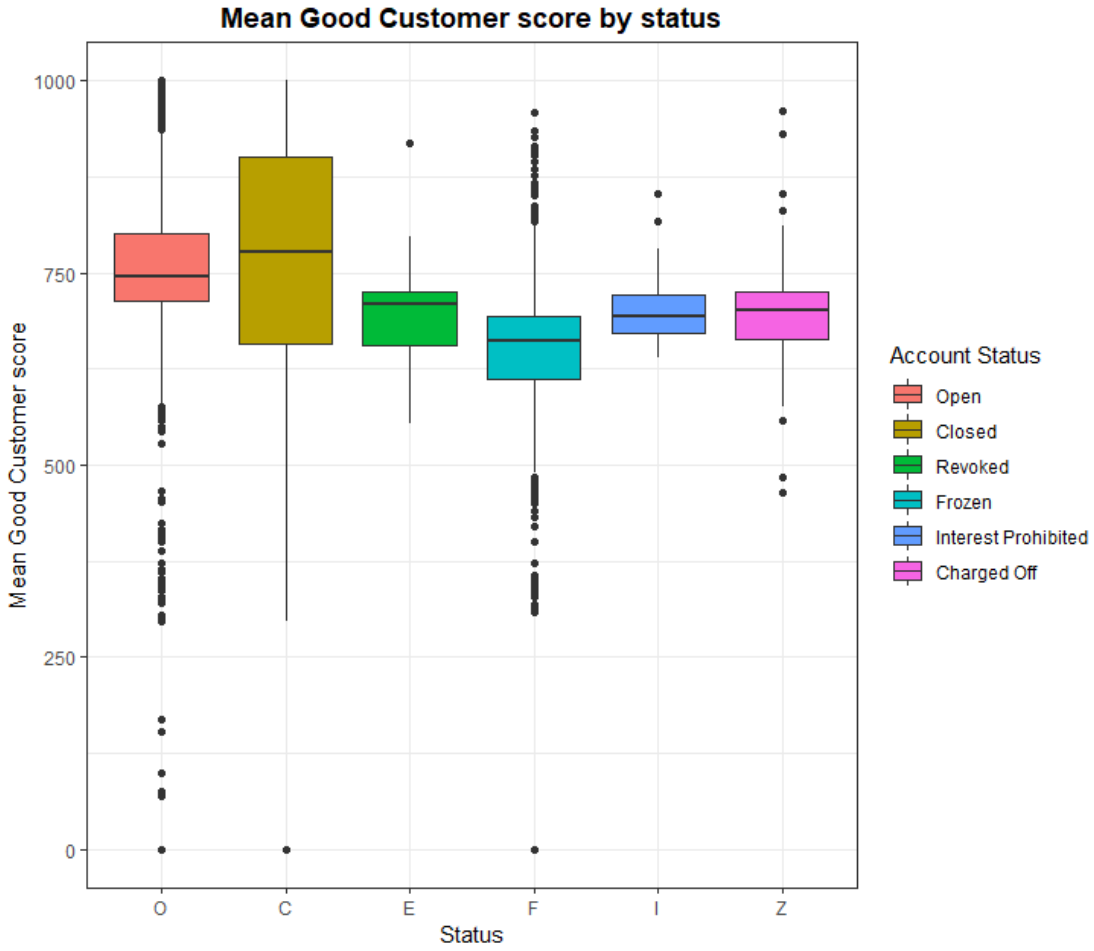

Figure 9: Mean Good Customer Score by Status

vi) Histogram and Correlation between the Numerical Variables in the data set

For Numerical variables, the correlation between the variables were examined by calculating the correlation and result is summarized on the graph shown below.

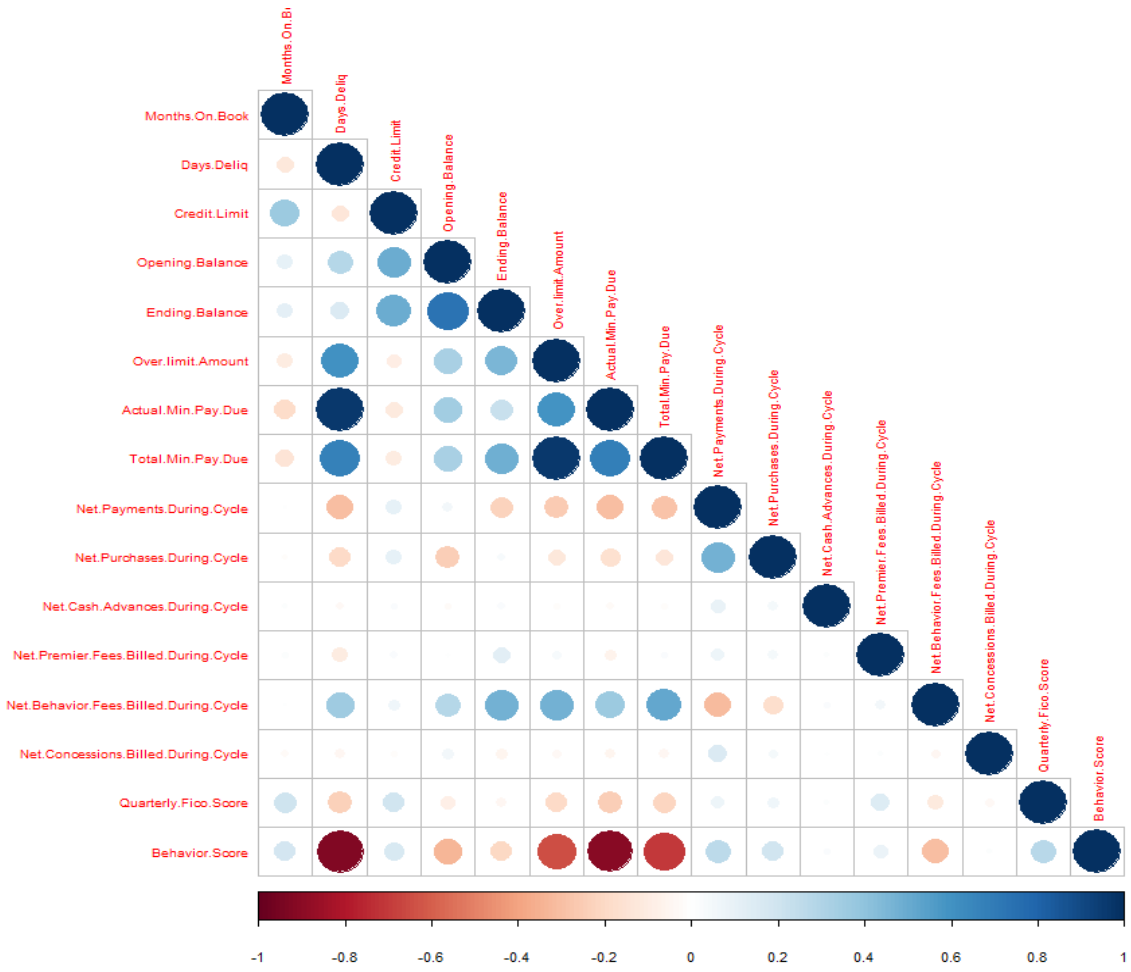

Figure 10: Correlation plot for Continuous Variables

Figure 10 shows how these variables are correlated. The darker and larger circle means two variables are more strongly correlated. The blue color represents positive relationships, and red color represent negative relationship.

The correlation matrix indicates that Behavior score is natively related to “Days Delinquency”, “Over limit amount”, “Actual minimum pay due”, and “Total minimum pay due”. “Net Behavior Fees Billed During Cycle” is positively related to “Days Delinquency”, “Opening Balance”, “Ending Balance”, “Over limit amount”, “Total limit amount”, “Actual min pay due”, and “Total min pay due”. Additionally, “Total minimum pay due” is positively related with “Days Delinquency” and “Over limit amount”. “Ending balance” is positively related with “Opening balance”. “Net payment during cycle” is natively related to “Days Delinquency”, “Actual minimum payment due”, “Total minimum payment due” and “Over limit mount”.

The histogram for all of the numerical variables are plotted and is shown in figure below.

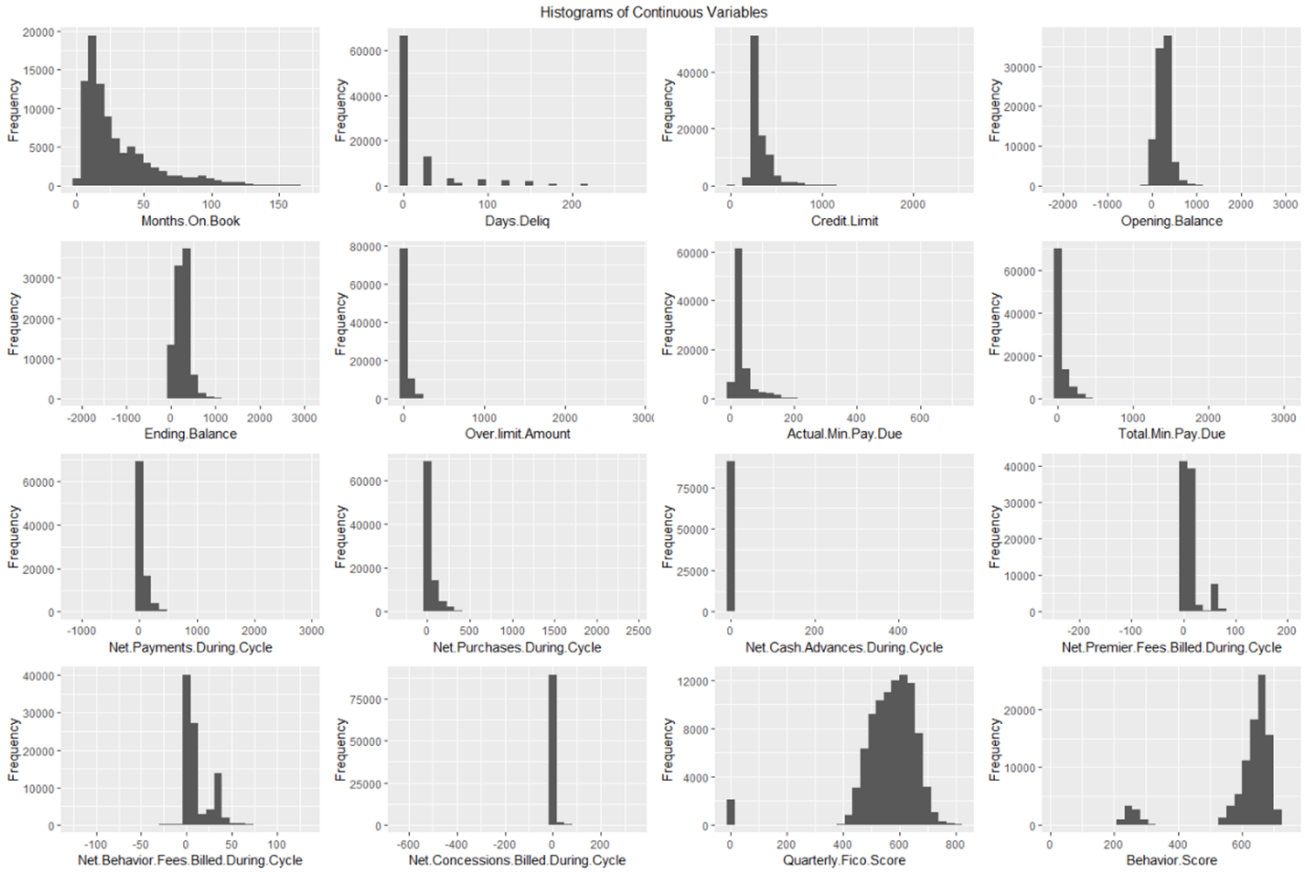

Figure 11: Histogram for Continuous Variables

- Opening Balance: Most of the accounts have a balance on the card at the start of a new cycle. This means that most accounts are paying the minimum on their cards and not paying off the card right away.
- Ending Balance: showing that most of the accounts have an ending balance between 0 and 500.
- Actual Minimum Payment Due: we see that most of the accounts have a minimum payment due of less than $50.00.
- Total Minimum Payment Due: Most of the accounts have a Total Minimum Payment Due that is less than $100.00.
- Net Payments During Cycle: Shows how much was paid on each account during the cycle. Most of the net payments were centered around zero.
- Net Purchases During Cycle: show the net purchases on the account during the cycle. Most of these accounts were used for net purchases under $200.00.
- Net Cash Advances During Cycle: shows the distribution of cash disbursements to the accounts over the course of the cycle. Most of the accounts had zero cash advances, but we were interested in the accounts that took out large cash disbursements.
- Net Premier Fees Billed During the Cycle: shows that most of the accounts were billed less than $100.00 for Premier Fees. We were not told exactly what the Premier Fees were, but we found it interesting that there is a small cluster of Premier Fees on the accounts that are around $-200.00.

- Net Behavior Fees Billed During the Cycle: showed that most of the accounts were billed for amounts between $-10.00 and $50.00. Since we were not given any information on how this account was prepared, we found it interesting that some accounts had negative balances for this variable.
- Net Concessions Billed During the Cycle shows that a majority of the accounts had a zero balance at the start of the cycle. There were a few accounts that either had a large negative balance or a large positive balance in the account at the end of the cycle.

vii) Date Related Variables

There are 6 variables related to date. First, the Open Date. The earliest date the credit card open was on 10/4/1996, and the latest date was on 2/10/2010. Second, the Last Statement Date. The date of last monthly statement or bill ranges from 1/5/2010 to 12/29/2010 which indicate that the data was from the Year 2010. Third, the Cycle Date. The date that the monthly billing cycle starts again was from 3/4/2010 to 12/31/2010. Fourth, the Months on Book. The number of months the credit card has been open ranged from 1 to 170, with a mean value of 30 and median value of 20. Fifth, the Last Payment Date. The date of last payment was started from 1/1/1900 to 12/31/2010. Sixth, the Days Delinquent, the number of days the monthly bill is 'past due' had a range from 0 to 270, with a mean value of 18 days.

We noticed an unusual observation on the last payment date that starts from 1/1/1900 while the cycle date and last statement date starts from Year 2010.

viii) Variables associated with Amount

With *Actual Minimum Pay Due*, we recorded a minimum value of $ 0 and a maximum value of about $157. The Opening Balance had a range from $-2077 to $3026, with mean value of $264. The Ending Balance had the same range of the Opening Balance, with a different mean value of $261. This indicate that there are customers who pay their bill ahead of time. The Actual Minimum Pay Due was from $0 to $720 with a mean value of $36. The Total Minimum Pay Due was from $0 to $3026 with a mean value of $53. The different between Actual Minimum Pay Due and Total Minimum Pay Due indicated that there were customers did not pay their minimum payment on time.

Regarding to *Net Payment During Cycle* variable, a minimum and maximum value of $-1122 and $2834 were recorded respectively. On average we had a mean value of $57. Another variable of interest to us was the *Net Premier Fees Billed During Cycle.* With this variable, we recorded a mean value of about 10.6 with a minimum and maximum values being $-235 and $203 respectively. *Net Behavior Fees Billed During Cycle* was not far from our consideration as well. We observed both the minimum and maximum values to be $-119 and $132 respectively and with an average value of about $11.

From the perspective of credit firm, we tried to see the trend net payment during cycle and sum of balance for all of customer for the months of year 2010 and the box plot is shown below and noticed that average payments during cycle has almost same mean for all months and so does the monthly balance for all months of year 2010.

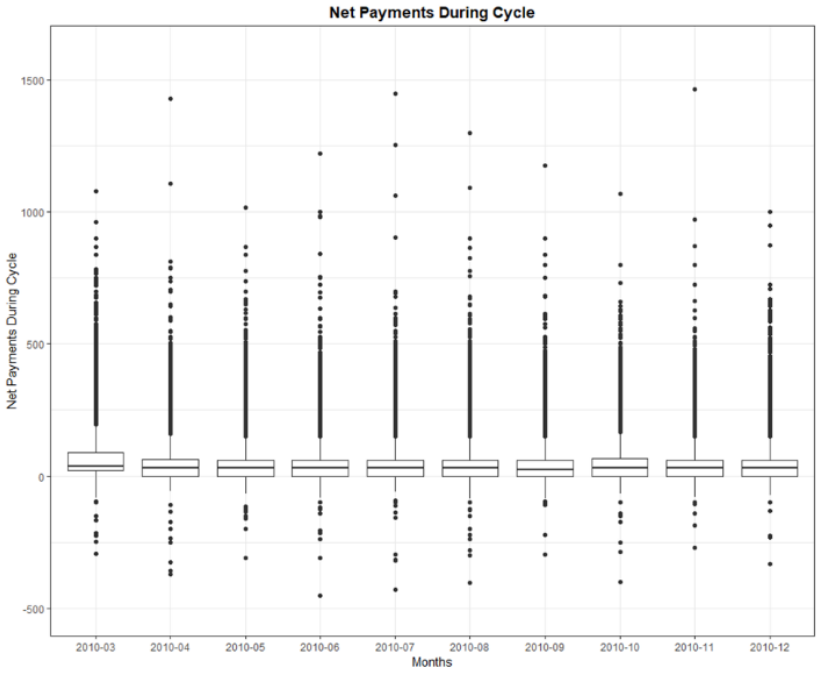

Figure 12: Net Payments During Cycle for Months of 2010

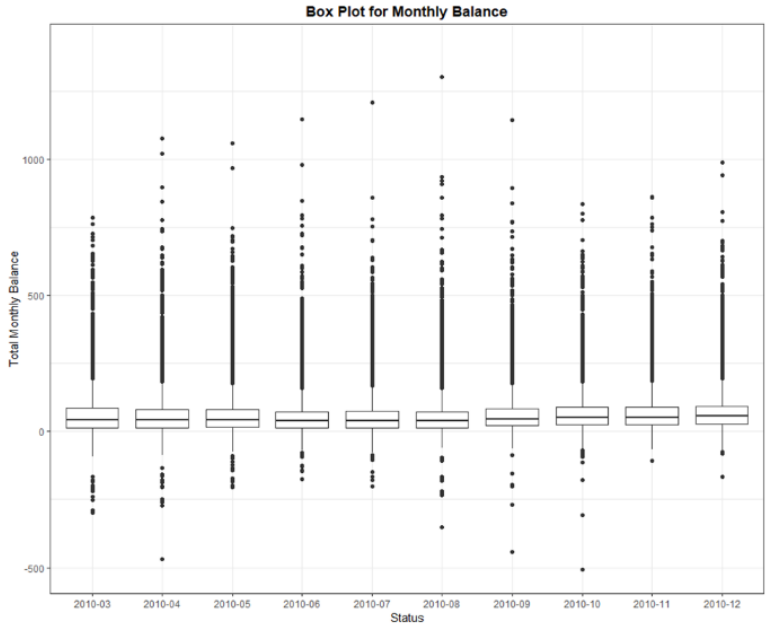

Figure 13: Monthly Balance for months of 2010

3. Conclusion

The data cleaning process was carried out on the data set and descriptive and exploratory analysis was carried out on the data set. The External Status variable seems to be our variable of interest as we are predicting the customer into a class of good and bad customer based on the account status. Several exploratory analyses including bar chart, boxplot, histogram etc. was carried out on the data set and it was found that majority of the customer has open external status followed by closed status. Through this preliminary exploration, several variables like Credit Limit, Months on Book, Delinquency Days, Net Payments, Quarterly FICO Score, Behavior Score, over limit Amount etc. variables exhibits some sort of relation with the external status and thus can be considered more significant for discerning the account status or for predicting the good and bad customer. We found "Net Behavior Fees Billed During Cycle" is positively related to "Days Deliq", "Opening Balance", "Ending Balance", "Over limit amount", "Total limit amount", "Actual min pay due", and "Total min pay due" while Behavior score is natively related to "Days Delinquency", "Over limit amount", "Actual minimum pay due", and "Total minimum pay due". This clearly shows how the payment behavior are affecting the score and thus the external status of customer. Several variables have correlated either positively or negatively and thus the variables with strong correlation can be dropped from model building process as introducing these one or more correlated variables will introduce the problem of multicollinearity.

4. Questions (based on dataset exploration):

- In the variable "last payment date" the earliest payment date is 1900-01-01. There are 571 such dates. Is that an error during data collection? Same with the 208 observations whose last payment date was 1980-01-01.

- Why some ID don't have row1 in "Row Num" column?

- Second, we described most of 12 variables related to dollar amount included in figure 1, and we have a question on 5 variables which shows negative dollar amount. They are Net Payment During Cycle, Net Purchases During Cycle, Net Premier Fees Billed, Net Behavior Fees Billed, and Net Concession Billed. All those values are net values, why do they have negative values?

- Moreover, what are the meaning of "Net Premier Fees Billed During Cycle", "Net Behavior Fees Billed During Cycle", "Net Concessions Billed During Cycle"? How are they related to other variables?

- What is the meaning of "Quarterly Fico Score", "Behavior Score", "Good Customer Score"? How are they related to other variables?