Data analysis project

Simple | Smart | Speed

HOOLS

CONTENTS

I. Summary:

Banks play a crucial role in market economies. They decide who can get finance and on what terms and can make or break investment decisions. For markets and society to function, individuals and companies need access to credit.

Credit scoring algorithms, which make a guess at the probability of default, are the method banks use to determine whether or not a loan should be granted. This project requires us to improve on the state of the art in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years.

In this project, historical data on 150,000 borrowers is analyzed. We implemented various classification algorithms for predicting whether some-body will experience financial distress in the next two years by looking on various parameters like monthly income, age, number of open credit lines and loans etc. An assessment of accuracy of implemented models is also done. Also, data cleaning approaches used are explained in this report. The model can be used to help in making best financial decisions when granting loans.

II. Project motivation:

The problem that we are attempting to solve is a classical one for analytics. It is commonly desired for banks to accurately assess the probability of default for their customers so that they can manage their loan risk better. With a better model, they can take calculated Risk in lending out to customers thus improving the certainty of their profit. They can also tailor make interest rate to cover for the level of risk they are exposed from the loan. Such models are heavily in demanded triggered due to the 2008 financial crisis followed by competitive forces and globalization effects which underscores the importance of business intelligence in financial risk.

However, developing such models still remains a challenge, especially with the growing demand for consumer loans. Data size is huge, and we are often talking about panel data. Such initiative necessitates bank to invest in proper data warehouse technologies to support development and active updating of such models. Also, it is common to see banks using score cards or simple logistic regression models to evaluate customer risk. This paper will attempt to use traditional analytic model such as logistic regression.

The data is based on real consumer loan data. The data mining models also include decision tree, neural networks and traditional logistic regression. In addition, we will conduct univariate and multivariate analysis of data to identify insights into banking customers.

The details of the various techniques and data cleaning approach used are given in subsequent sections in this paper.

III. What is the problem?

- The goal is to predict delinquency on debt.

- To do so we will examine a database of 150 000 customers.

- There are numerous models that can be used to predict which customers will default. The chosen model could be used to decrease credit limits or cancel credit line for current risky customers to minimize potential loss. Also we could better screen which customers are approved for the credit.

- This is a basic classification problem with important business implications.

- In this problem each customer has a number of attributes like (age, monthly income, debt ratio … etc.). we will use the customer attributes to predict if they were delinquent.

- We will use the file cs.training to get a sense of how well our model performs.

IV. Data description:

We have chosen “Give Me Some Credit” from kaggle.com and plan to predict the probability and/or decision as to whether a person will experience financial distress (90 day delinquency or worse) in the next two years. The dependent variable is SeriousDlqIn2Yrs and is binary. All of the independent variables will be evaluated and some or all of them used as independent variables in our generated prediction model.

Our initial training dataset has 150,000 observations and 10 variables. The variables are described in the table below:

Variable Name	**Description**	**Type**
SeriousDlqin2yrs	Person experienced 90 days past due delinquency or worse	Y/N
RevolvingUtilizationOfUnsecuredLines	Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits	percentage
Age	Age of borrower in years	integer
NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
DebtRatio	Monthly debt payments, alimony, living costs divided by monthly gross income	percentage
MonthlyIncome	Monthly income	real
NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	Integer
NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

We do believe that the quality of the data is far from perfect due to 3 major factors:

- Incompleteness: we have missing values for one or more records.
- Noise: Data may contain erroneous or anomalous values such as outliers.
- Inconsistency: Data may have some discrepancies due to changes in coding system.

To explore our dataset we used first IBM SPSS, but due to some variable-type problem and huge data set, we will continue the processing of Data using SAS.

V. Data Exploration & transformation

UNIVARIATE ANALYSIS

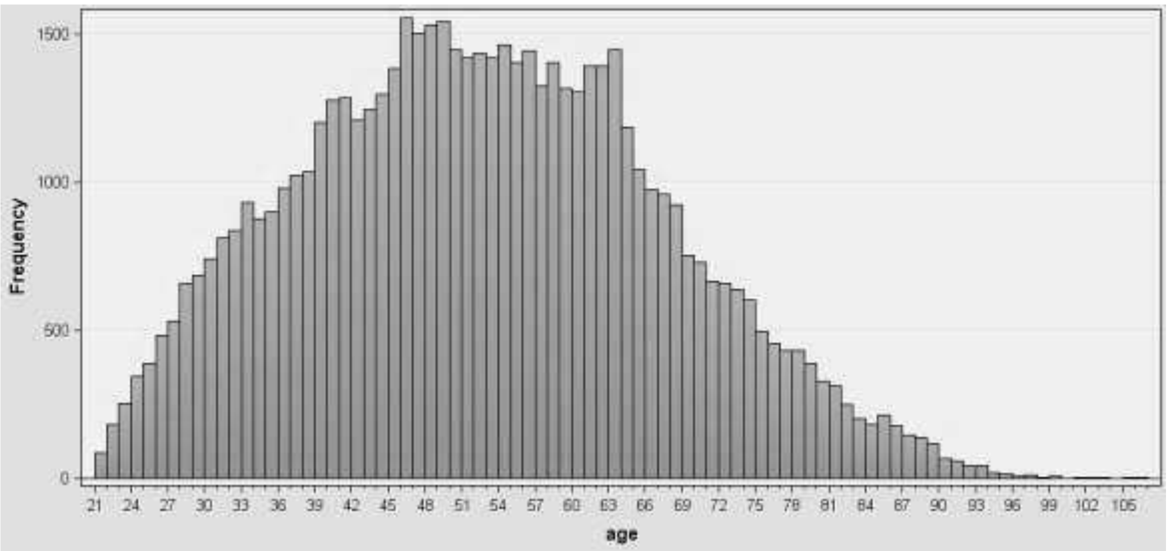
We also went on to graph each variable to get a better understanding of its distribution and outliers.

The following is a summary of each variable with its initial histogram and observations, an explanation of the data clean up that was completed to remove outliers and/or missing values, and our observations and a histogram of the variable after transforming it.

VARIABLE: AGE

Description: Age of borrower in years

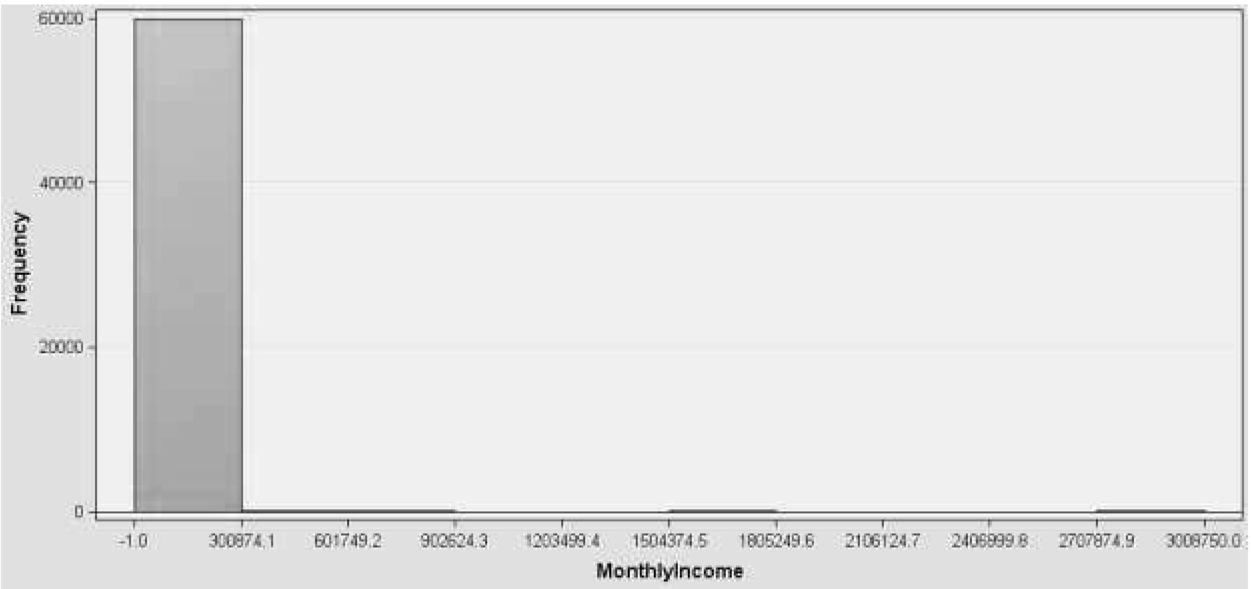
Initial Observations: The data looks fairly well distributed and gives a Poisson or normal distribution. This variable doesn't appear to need any cleanup.



VARIABLE: MONTHLYINCOME

Description: Monthly income

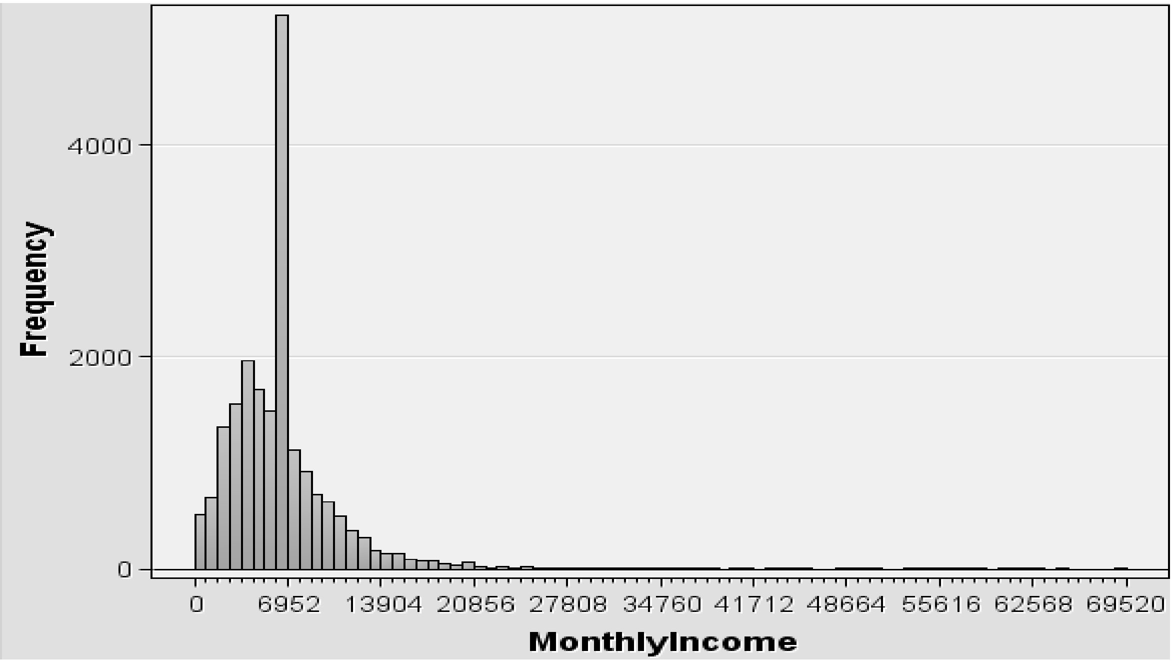
Initial Observations: The data looks to have some outliers at the top end that need to be removed. After this, we may be able to get a better distribution. The frequency of the outliers is very low, so we will look at a way to remove these. The -1 values were answered N/A, and we will need to also do something with these values.



Data Cleanup Explained: the -1 value was set to missing initially, and a MonthlyIncome_NA_IND field was added to capture the N/A responses for that field.

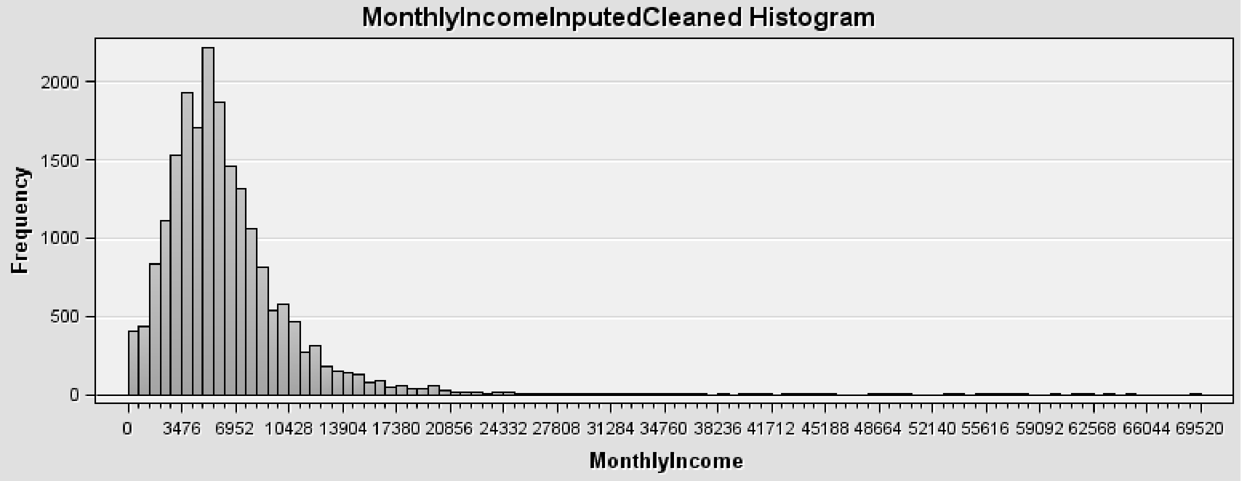
We then set the MonthlyIncome values to the mean value of all of the MonthlyIncome observations. After this was completed, MonthlyIncome values above the 99.9th percentile were dropped.

Observations: After the changes to MonthlyIncome explained above, we have more of a Poisson distribution. There is a spike at the mean value, but overall the quality improved.



The distribution above would probably be skewed by the high frequency of mean values. In addition, the data distribution does not look that real. After looking at several options, we did the following: We used Base SAS to impute MonthlyIncome with a linear regression using the other dependent variables and stepwise selection method for variable selection.

This imputation for the missing values seems to work well, and we hope it helps the prediction model. The resulting chart is shown below:

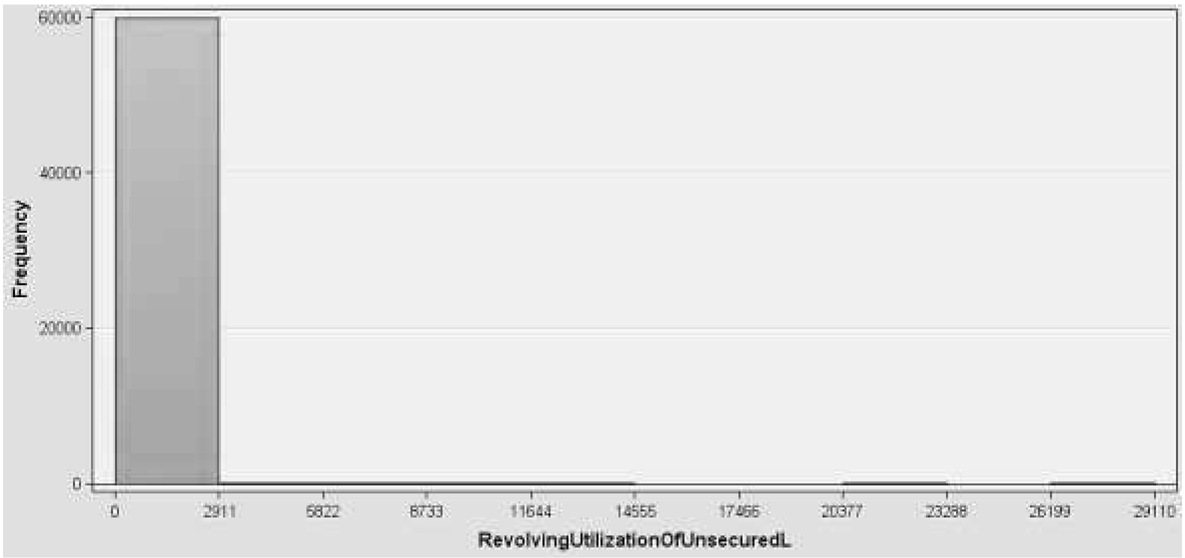


Not as many rows were pulled in for the above chart, so that may be why the frequency is lower, but the distribution was much better and seems to match the natural distribution of the data above. We worked a bit with proc mi and looked at misanalyse but decided that the linear regression was something that we could do and understand better.

VARIABLE: REVOLVINGUTILIZATIONOFUNSECUREDLINES

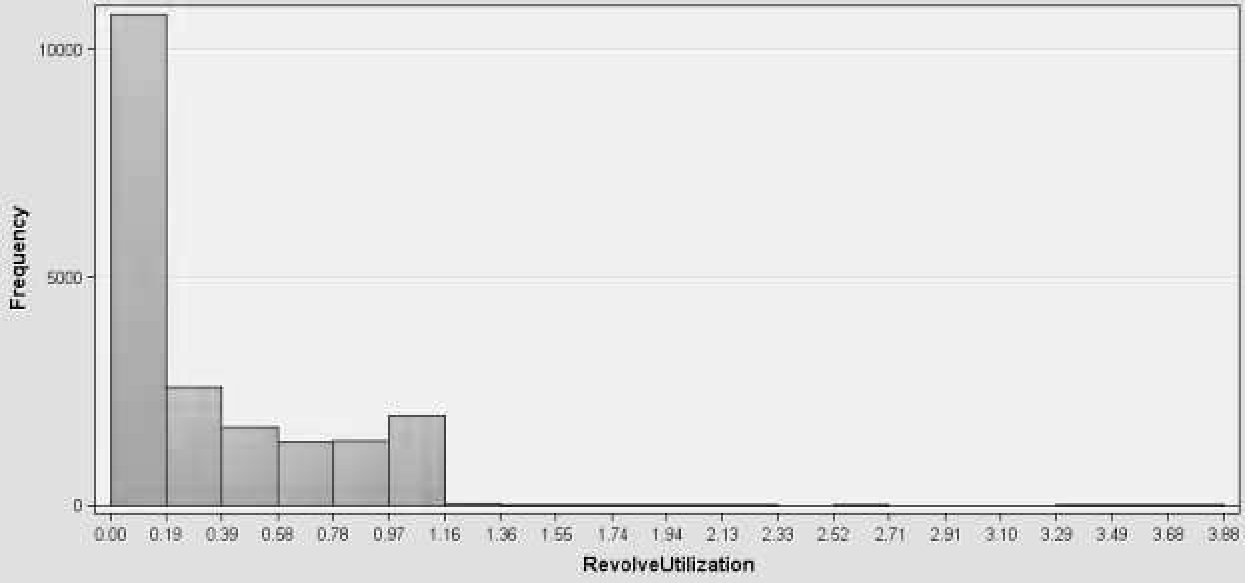
Description: Total balance on credit cards and personal lines of credit except real estate and no installment debt like car loans divided by the sum of credit limits

Initial Observations: This variable contains some outliers. Since the variable should be the ratio of unsecured loans that are being utilized, values between 0 and 1 should make up most of the observations. Some values might be higher than that if a line of credit was closed while a balance still exists.



Data Cleanup Explained: An assumption was made that the values in RevolvingUtilizationOfUnsecuredLines that were above 10 were entered in error. The real value was meant to be less than 1. The values over 10 were adjusted by moving the decimal place to the correct position and an indicator variable was created.

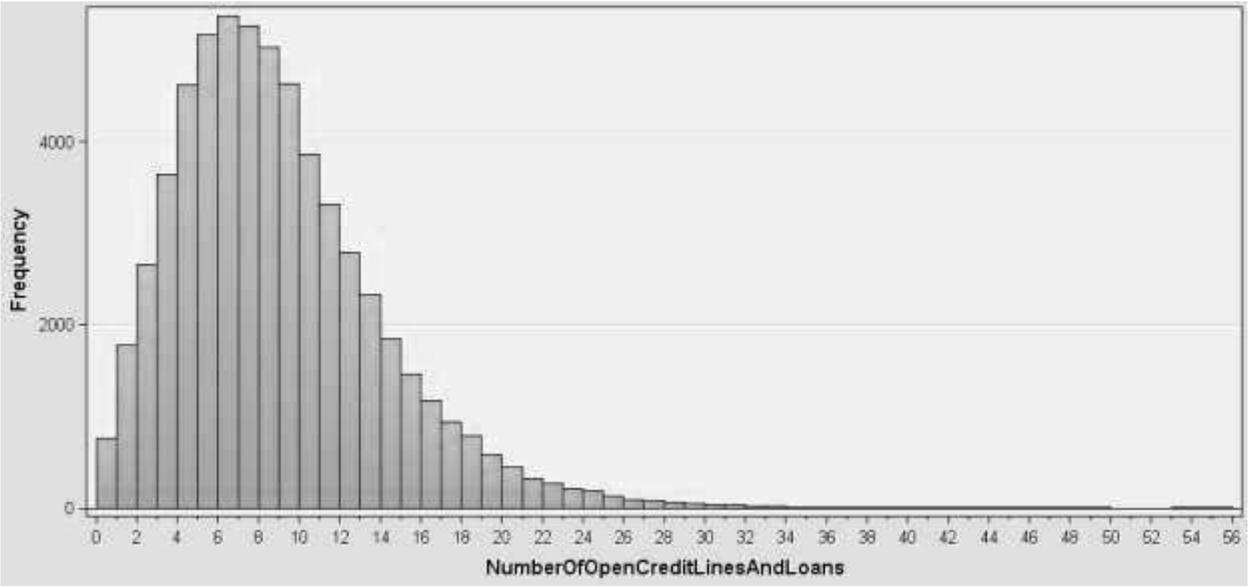
Observations: After the changes to RevolvingUtilizationOfUnsecuredLines explained above, most of the values are between 0 and 1 with a decrease in values above 1.



VARIABLE: NUMBEROFOPENCREDITLINESANDLOANS

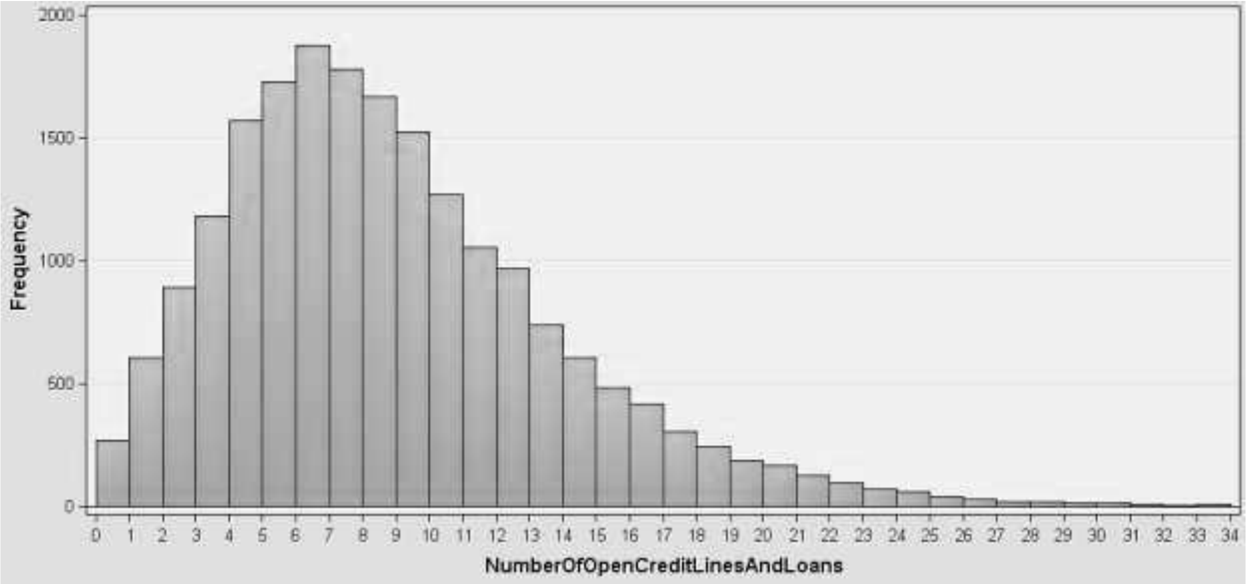
Description: Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)

Initial Observations: The data looks fairly well distributed, but there seems to be a few outliers.



Data Cleanup Explained: NumberOfOpenCreditLinesAndLoans values above the 99.9th percentile were dropped.

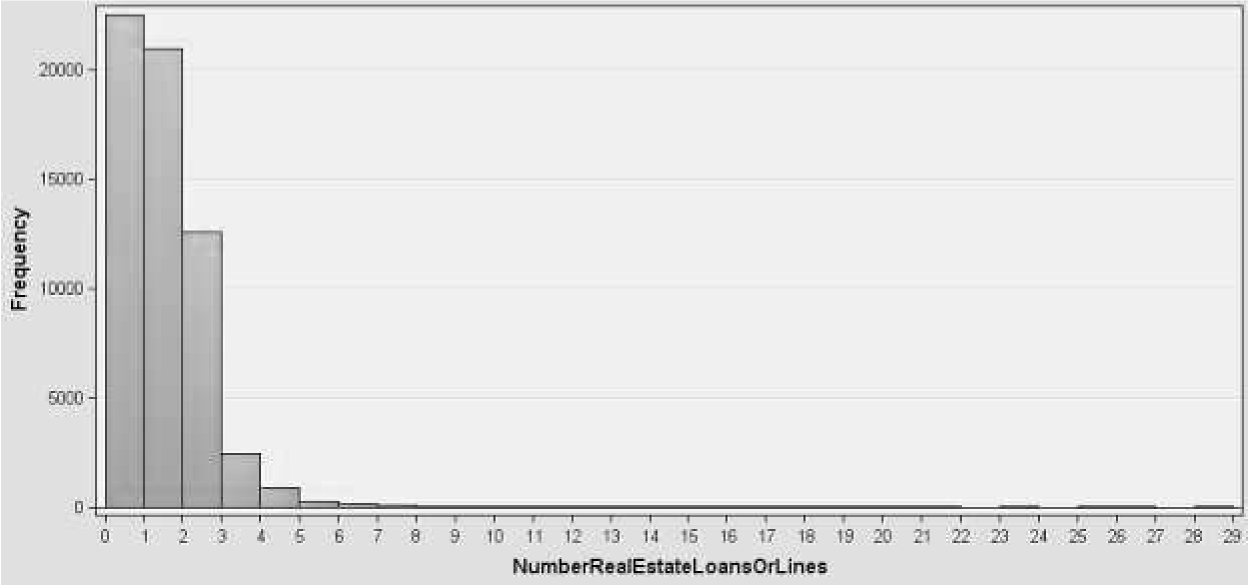
Observations: After the changes to NumberOfOpenCreditLinesAndLoans explained above, there are fewer outliers as shown in the graph below.



VARIABLE: NUMBERREALESTATELOANSORLINES

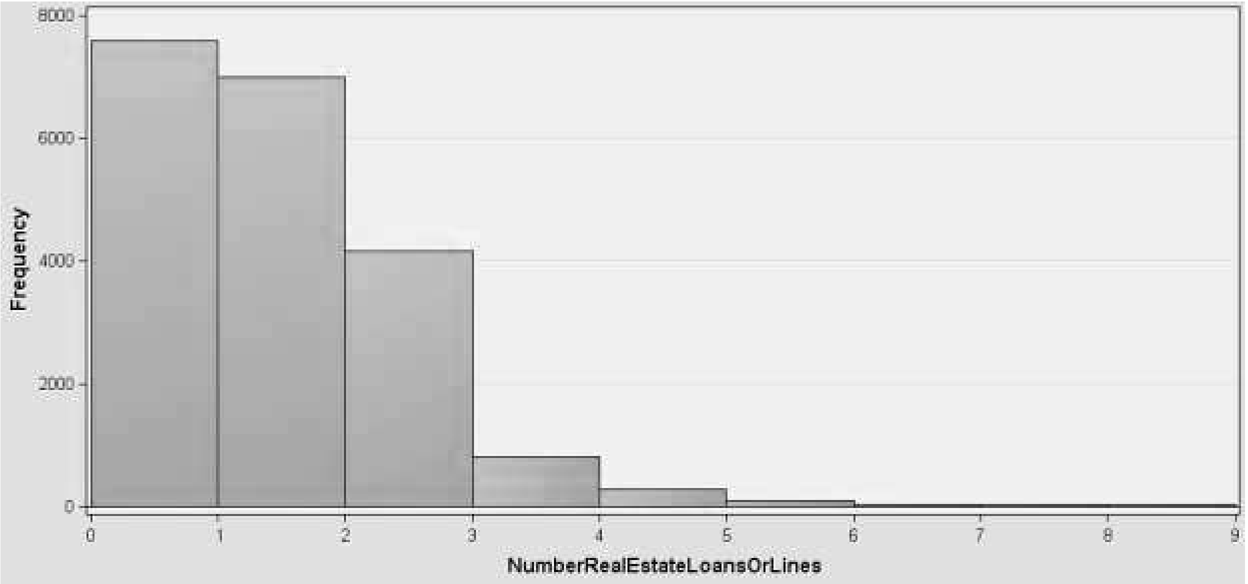
Description: Number of mortgage and real estate loans, including home equity lines of credit

Initial Observations: The data looks fairly well distributed, but there seems to be a few outliers with low frequency.



Data Cleanup Explained: NumberRealEstateLoansOrLines values above the 99.9th percentile were dropped.

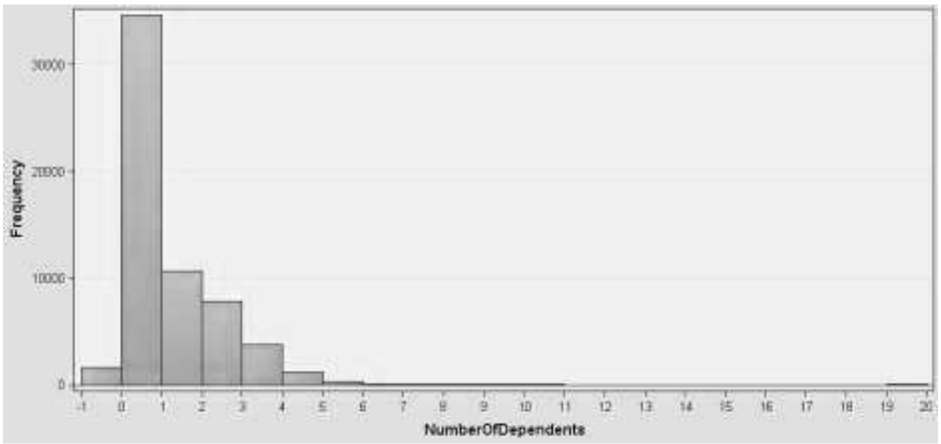
Observations: After the changes to NumberRealEstateLoansOrLines explained above, the data contains fewer outliers as shown in the graph below.



VARIABLE: NUMBEROFDEPENDENTS

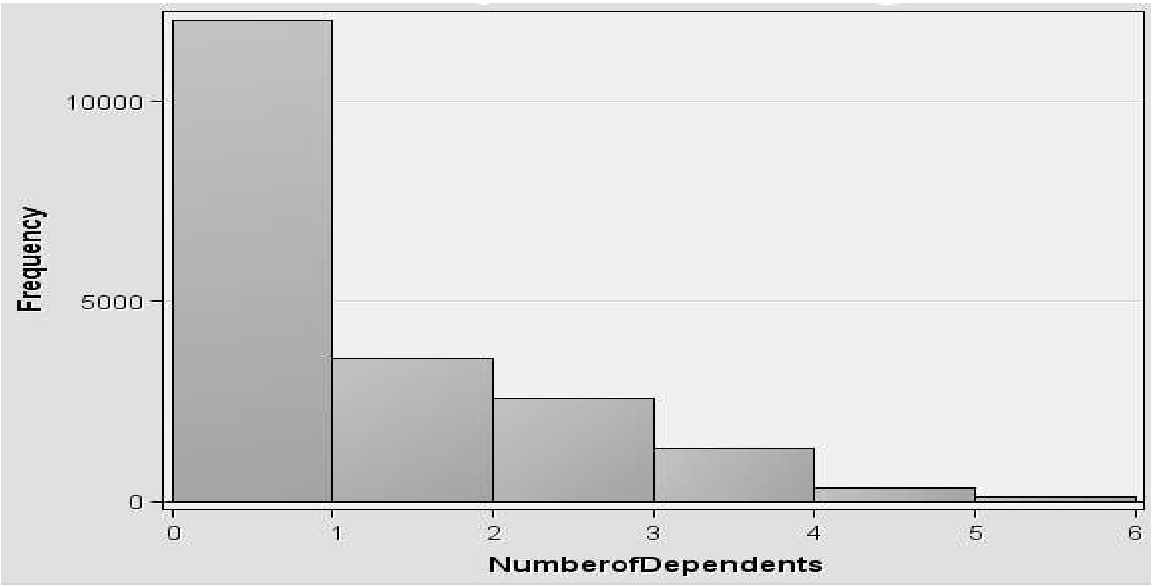
Description: Number of dependents in family, excluding themselves (spouse, children, etc.)

Initial Observations: The number of dependents has some -1 value, and these values were N/A in the initial dataset. We will do something with these values. There are also some outliers that will need to be dropped to improve the chart. Again, the frequency on these outliers is very low.



Data Cleanup Explained: the -1 value was set to missing initially, and a NumberofDependents_NA_IND field was added to capture the N/A responses for that field. We then set the NumberofDependents missing values to 0. We made the assumption that most of the people who answered N/A did so because they had no dependents. After this was completed, NumberofDependents values above the 99.9th percentile were dropped.

Observations: After the changes to NumberofDependents explained above, we have more of a power law distribution.

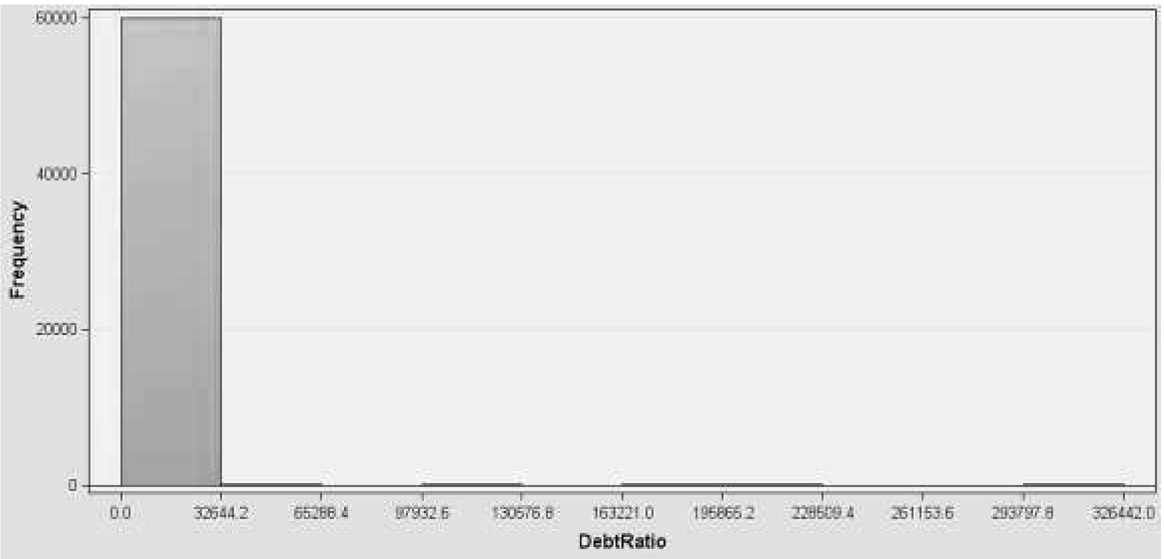


VARIABLE: DEBTRATIO

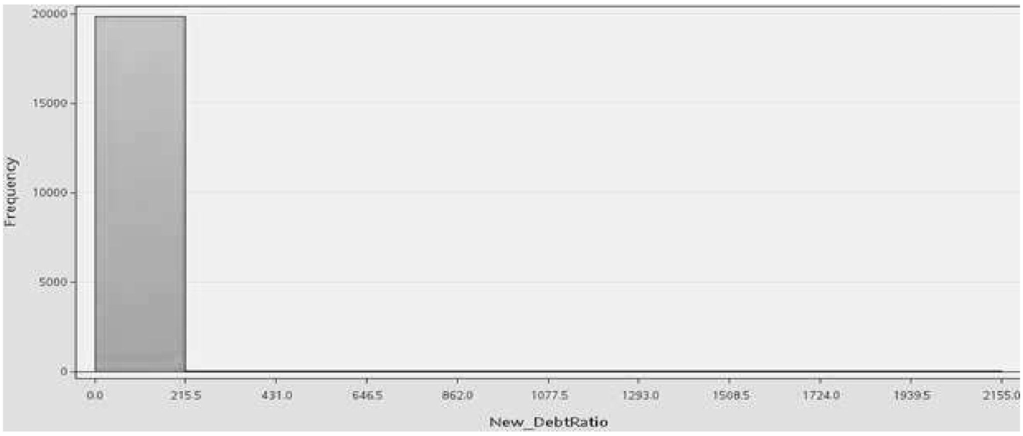
Description: Monthly debt payments, alimony, living costs divided by monthly gross income. The debt ratio is defined as the ratio of total debt to total assets, expressed in percentage, and can be interpreted as the proportion of an individual's assets that are financed by debt.

$$\text{Debt Ratio} = \frac{\text{Total Debt}}{\text{Total Assets}}$$

Initial Observations: The initial observation of the data suggested that we trim the New DebtRatio using 99.7 percentile on the top, which is 3 standard deviations, and do no trimming on the bottom.



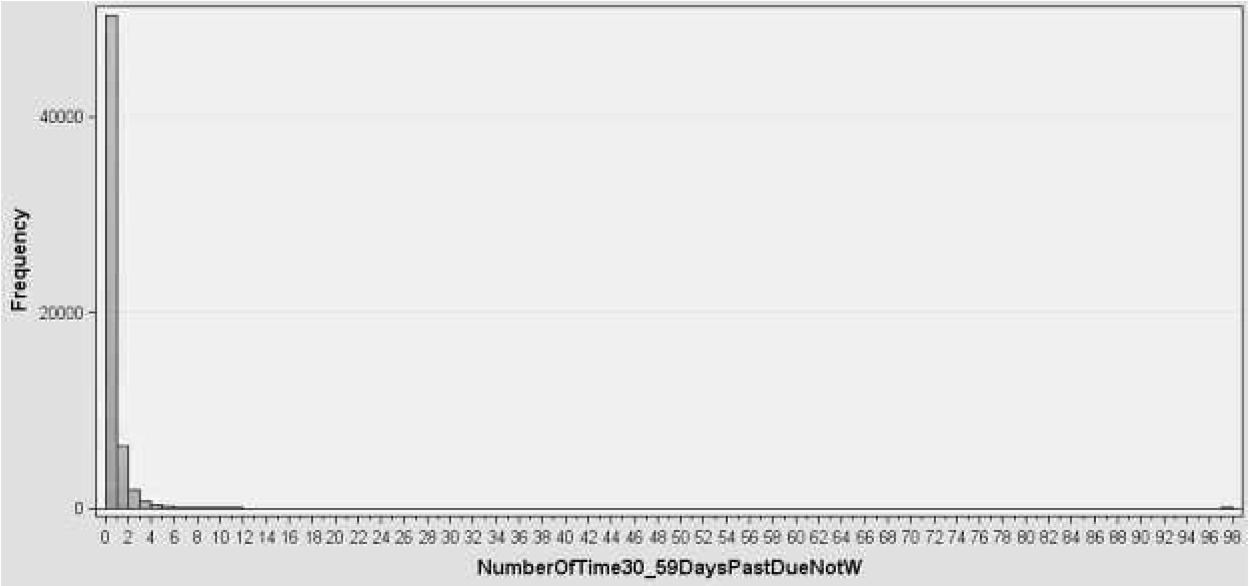
Data Cleanup Explained: The New_DebtRatio was calculated using the original values from DebtRatio, with imputations only where MonthlyIncome values were imputed. The imputation used the formula DebtRatio / MonthlyIncome. This was done in Base SAS after the cleansing of MonthlyIncome. For the specific case of the imputed MonthlyIncome being zero, the Original DebtRatio was kept. After this was completed, New_DebtRatio values above the 99.7th percentile were dropped.



VARIABLE: NUMBEROFTIME30-59DAYSPASTDUENOTWORSE

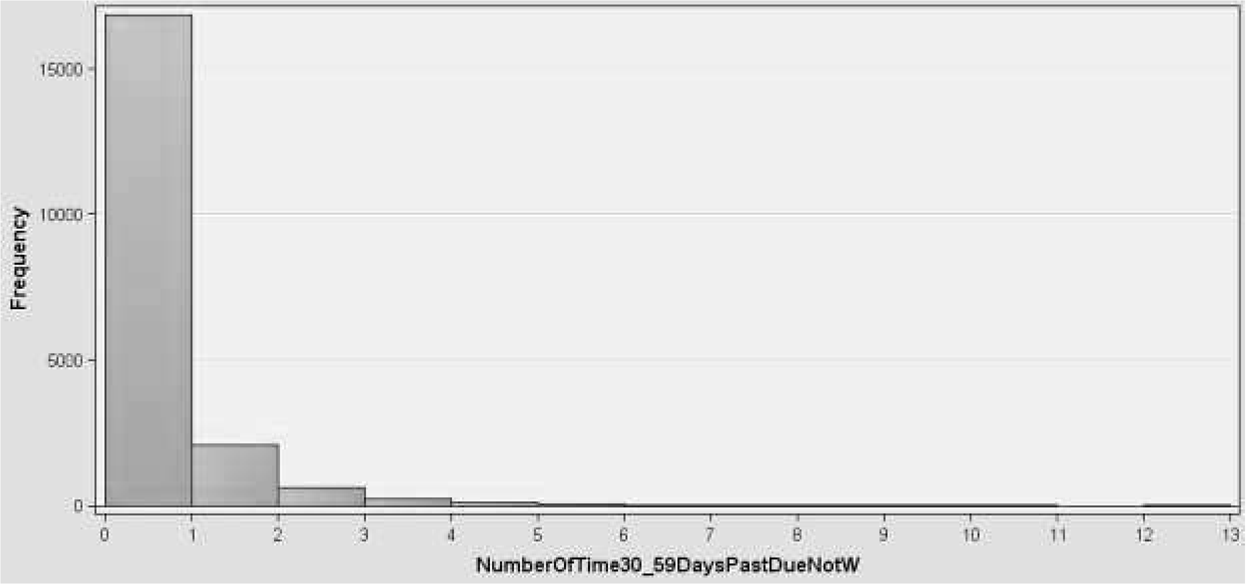
Description: Number of times borrower has been 30-59 days past due, but not worse, in the last 2 years.

Initial Observations: There are a few values of 96 and 98 in the dataset. This is probably code for did not want to answer or missing. Remove them as outliers.



Data Cleanup Explained: Changed the values of 96 and 98 to missing values, and then replaced these missing values with the mean of the variable.

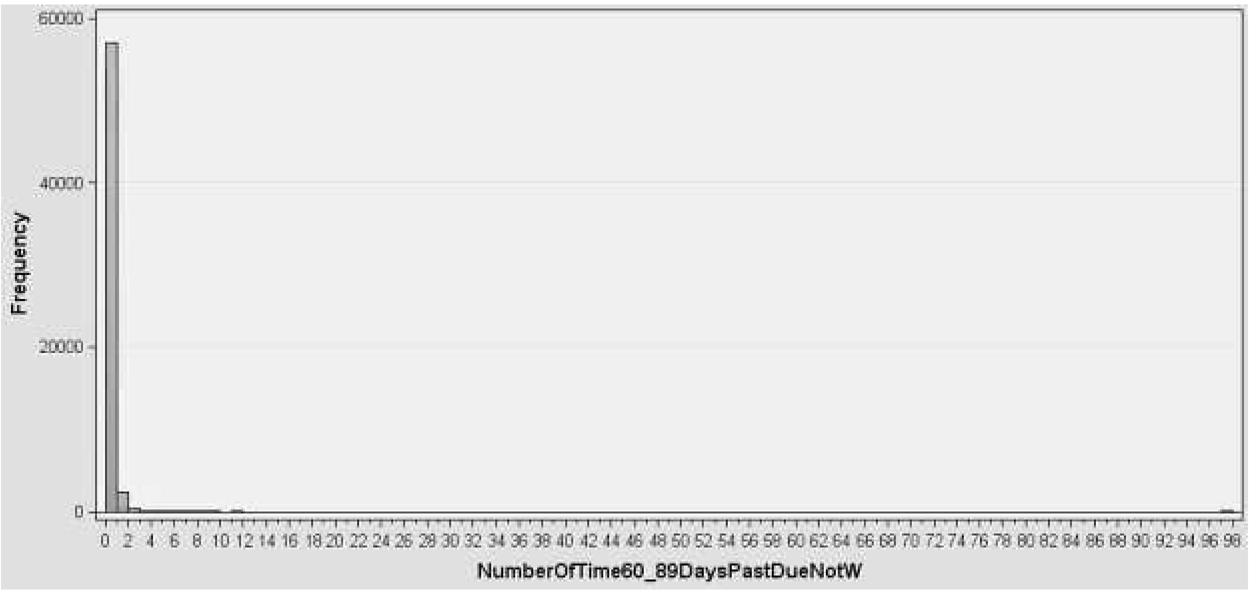
Observations: After removing the values 96 and 98 and replacing them with the mean, the data range is reasonable and quite a few of the observations are 0.



VARIABLE: NUMBEROFTIME60-89DAYSPASTDUENOTWORSE

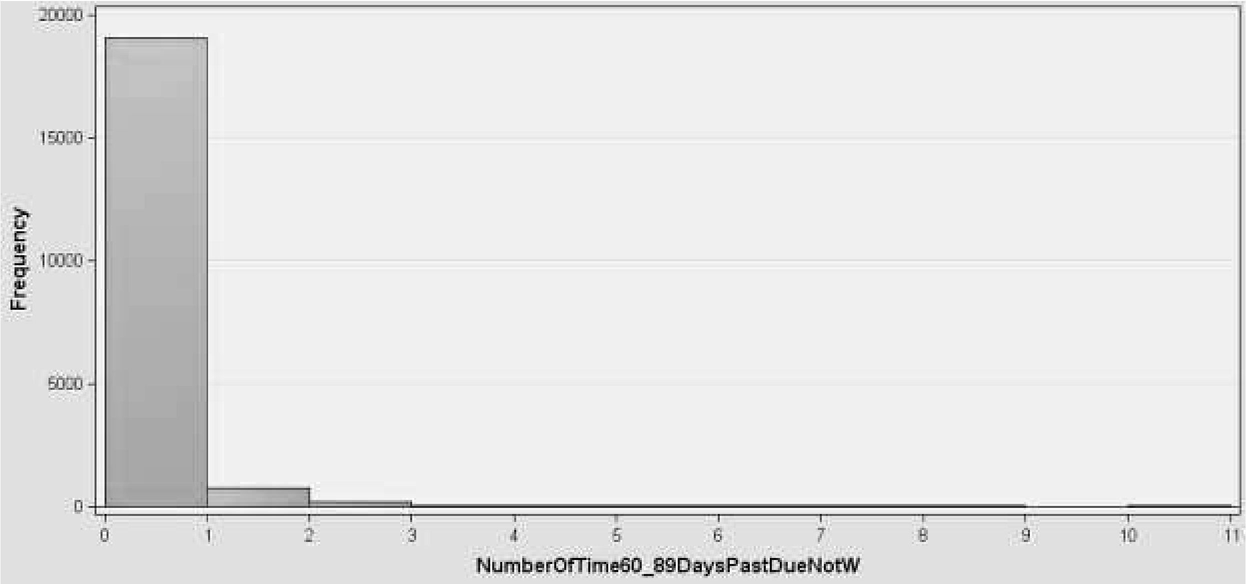
Description: Number of times borrower has been 60-89 days past due, but not worse, in the last 2 years.

Initial Observations: There are quite a few values of 96 and 98 in the dataset. This is probably code for did not want to answer. Remove them as outliers.



Data Cleanup Explained: The values of 96 and 98 were changed to missing values and then replaced with the mean of the variable.

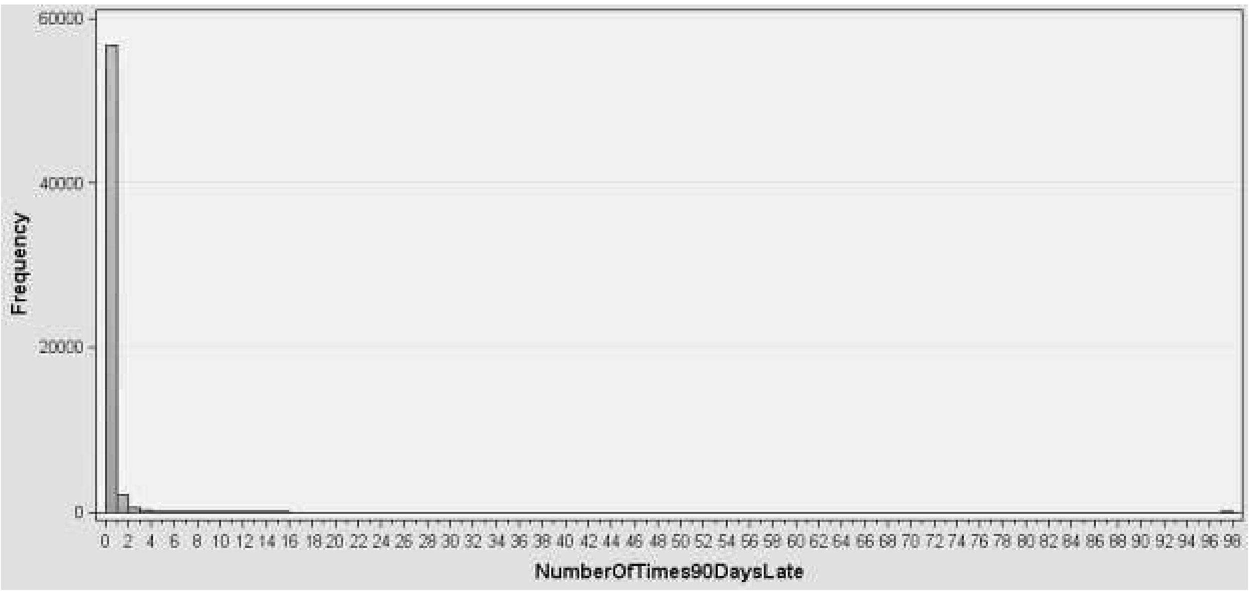
Observations: After removing the values 96 and 98 and replacing them with the mean, the data range is reasonable and quite a few of the observations are 0.



VARIABLE: NUMBEROFTIMES90DAYSLATE

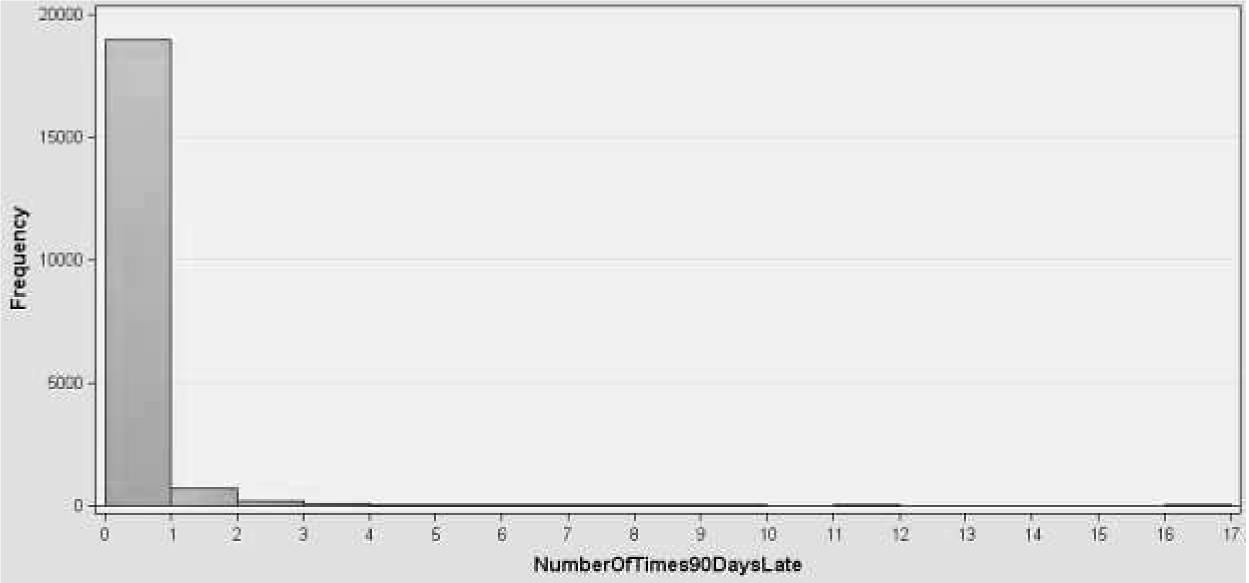
Description: Number of times borrower has been 90 days or more past due.

Initial Observations: There are quite a few values of 96 and 98 in the dataset. This is probably code for did not want to answer. Remove them as outliers.

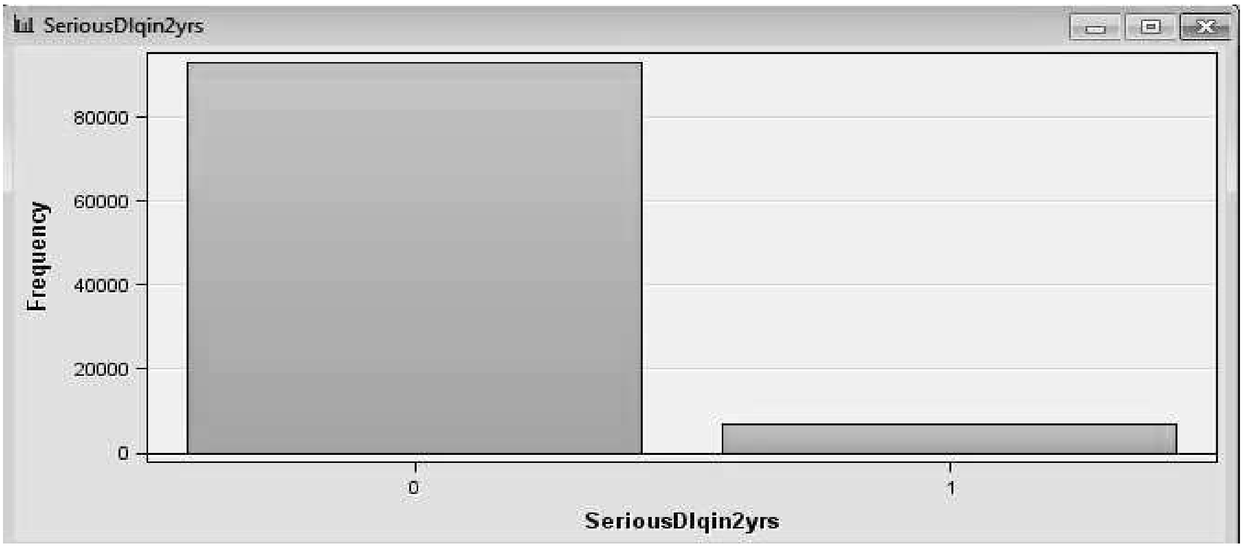


Data Cleanup Explained: The values of 96 and 98 were changed to missing values and then replaced with the mean of the variable.

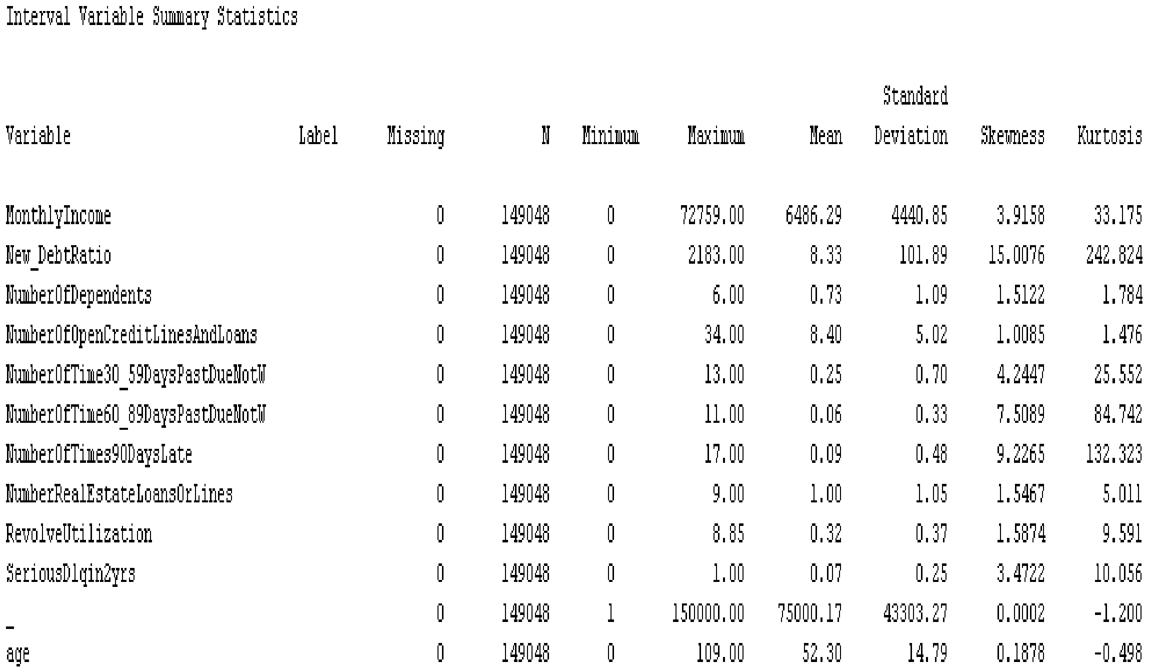
Observations: After removing the values 96 and 98 and replacing them with the mean, the data range is reasonable with most of the observations being 0.



On the bar chart we can see that approximately 93% of the observations are with value 0 and approximately 7%are with value 1.That means approximately 7% of the customers in this data set have experienced 90 days past due delinquency or worse.



The following table created from the DMDB node summarizes our final data set, with all variable transformations complete.

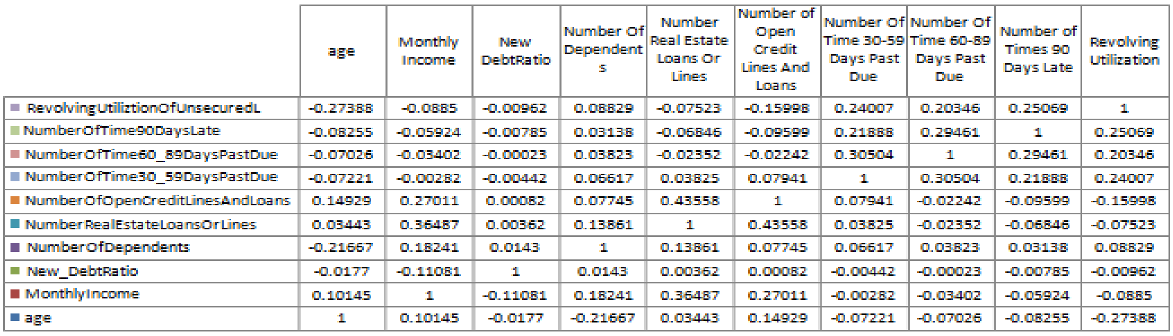


```
Interval Variable Summary Statistics


                                                                                         Standard
Variable                          Label    Missing           N   Minimum     Maximum        Mean   Deviation   Skewness   Kurtosis


MonthlyIncome                                    0      149048         0    72759.00     6486.29     4440.85     3.9158     33.175
New_DebtRatio                                    0      149048         0     2183.00        8.33      101.89    15.0076    242.824
NumberOfDependents                               0      149048         0        6.00        0.73        1.09     1.5122      1.784
NumberOfOpenCreditLinesAndLoans                  0      149048         0       34.00        8.40        5.02     1.0085      1.476
NumberOfTime30_59DaysPastDueNotW                 0      149048         0       13.00        0.25        0.70     4.2447     25.552
NumberOfTime60_89DaysPastDueNotW                 0      149048         0       11.00        0.06        0.33     7.5089     84.742
NumberOfTimes90DaysLate                          0      149048         0       17.00        0.09        0.48     9.2265    132.323
NumberRealEstateLoansOrLines                     0      149048         0        9.00        1.00        1.05     1.5467      5.011
RevolveUtilization                               0      149048         0        8.85        0.32        0.37     1.5874      9.591
SeriousDlqin2yrs                                 0      149048         0        1.00        0.07        0.25     3.4722     10.056
_                                                0      149048         1   150000.00    75000.17    43303.27     0.0002     -1.200
age                                              0      149048         0      109.00       52.30       14.79     0.1878     -0.498
```

BIVARIATE ANALYSIS

VARIABLE REDUNDANCY:

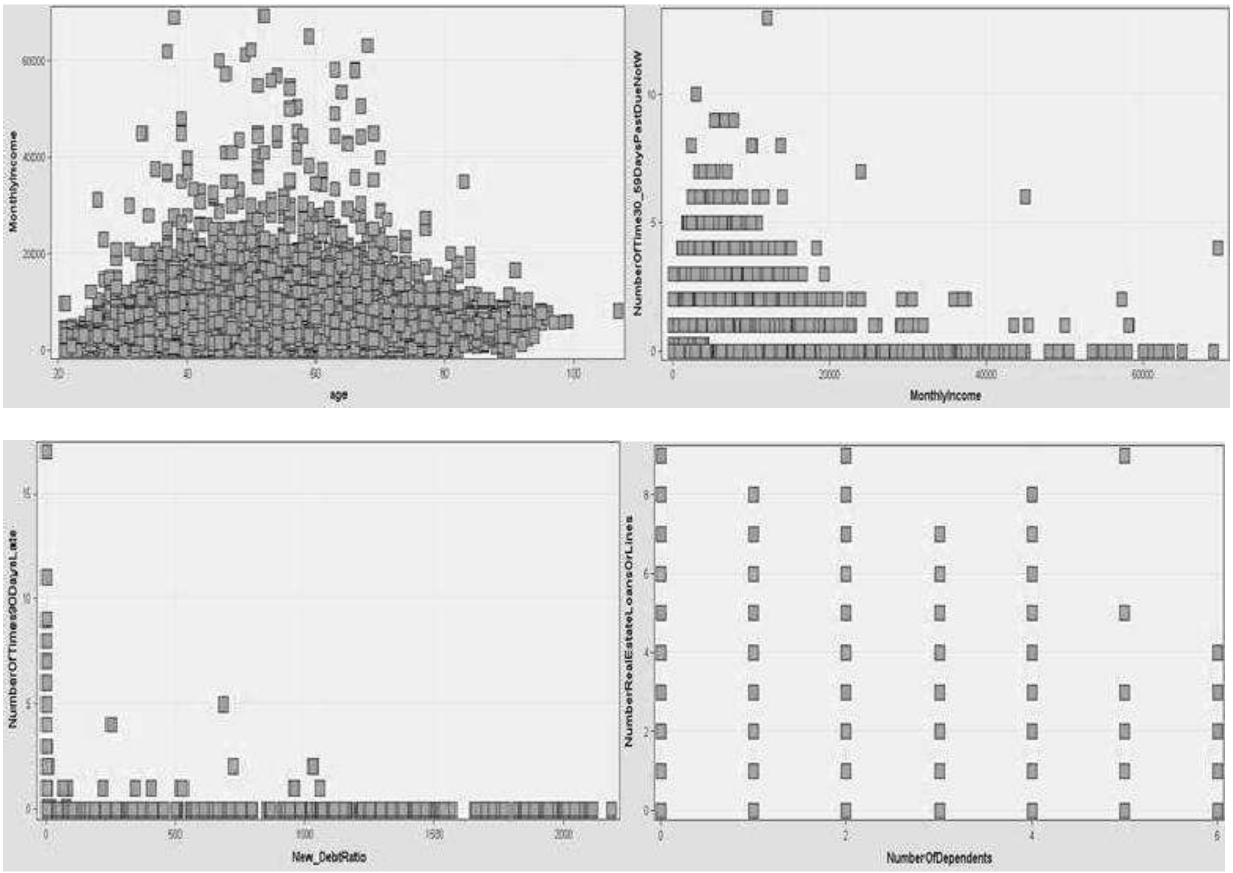
Here we are attempting to ensure that we do not have variables that are related to each other or essentially serve the same function. If this is the case, it is possible to eliminate one of them. This is needed for interpretation and not accuracy. If we have a linear regression, it is easier to interpret for the customer when we do not have multicollinearity or variables that are highly related to one another. To do this, we will look at the correlation coefficients of the variables with respect to each other. The number "1" represents perfect correlation in the table and identifies two completely redundant variables. We believe that we need to keep our coefficients under 0.7 in order to be able to interpret a linear/logistic regression model while using the statement holding all other variables constant. The correlation matrix is shown below:



	age	Monthly Income	New DebtRatio	Number Of Dependents	Number Real Estate Loans Or Lines	Number of Open Credit Lines And Loans	Number Of Time 30-59 Days Past Due	Number Of Time 60-89 Days Past Due	Number of Times 90 Days Late	Revolving Utilization
■ RevolvingUtiliztionOfUnsecuredL	-0.27388	-0.0885	-0.00962	0.08829	-0.07523	-0.15998	0.24007	0.20346	0.25069	1
■ NumberOfTime90DaysLate	-0.08255	-0.05924	-0.00785	0.03138	-0.06846	-0.09599	0.21888	0.29461	1	0.25069
■ NumberOfTime60_89DaysPastDue	-0.07026	-0.03402	-0.00023	0.03823	-0.02352	-0.02242	0.30504	1	0.29461	0.20346
■ NumberOfTime30_59DaysPastDue	-0.07221	-0.00282	-0.00442	0.06617	0.03825	0.07941	1	0.30504	0.21888	0.24007
■ NumberOfOpenCreditLinesAndLoans	0.14929	0.27011	0.00082	0.07745	0.43558	1	0.07941	-0.02242	-0.09599	-0.15998
■ NumberRealEstateLoansOrLines	0.03443	0.36487	0.00362	0.13861	1	0.43558	0.03825	-0.02352	-0.06846	-0.07523
■ NumberOfDependents	-0.21667	0.18241	0.0143	1	0.13861	0.07745	0.06617	0.03823	0.03138	0.08829
■ New_DebtRatio	-0.0177	-0.11081	1	0.0143	0.00362	0.00082	-0.00442	-0.00023	-0.00785	-0.00962
■ MonthlyIncome	0.10145	1	-0.11081	0.18241	0.36487	0.27011	-0.00282	-0.03402	-0.05924	-0.0885
■ age	1	0.10145	-0.0177	-0.21667	0.03443	0.14929	-0.07221	-0.07026	-0.08255	-0.27388

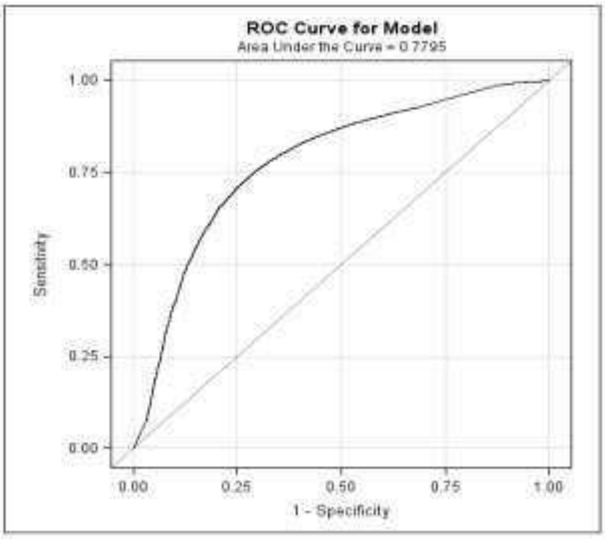
In the correlation matrix, we ignore the 1's as those are a variable that is correlated against it. The highest is NumberofOpenCreditLinesAndLoans and NumberRealEstateLoansOrLines at 0 .44. This is not too high, and for interpretability we should be able to interpret a linear equation if we do not take the log or execute another complex function on the features. It does not appear that there are any variables that need to be removed for multicollinearity.

Another way to check if the variables are correlated to each other is to create scatter plots with one variable plotted on the x-axis and another on the y-axis. This allows you to see if there are any patterns between the two variables. We plotted all variables against each other and found a few different types of patterns but no real correlations. The following are examples of some of the patterns we observed:

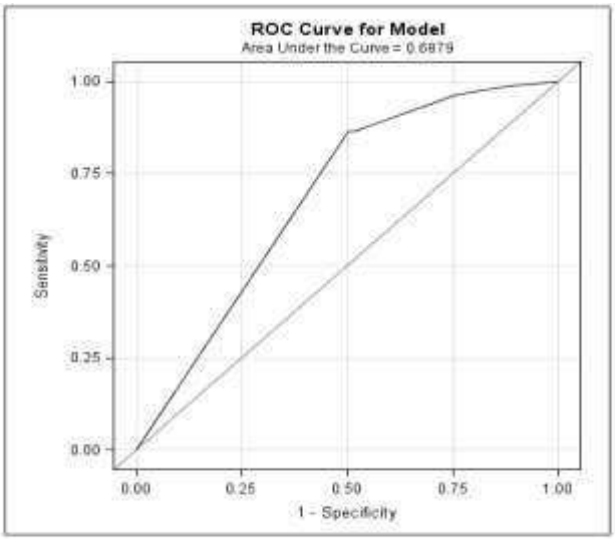


VARIABLE RELEVANCY:

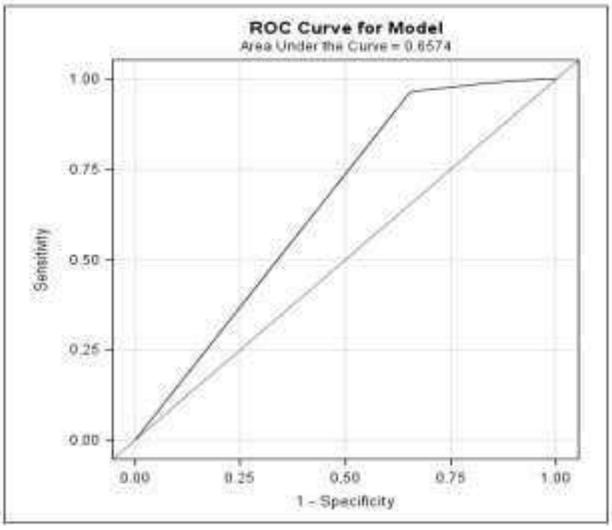
We are attempting to identify the relevancy of our variables on the outcome, which is SeriousDlqIn2yrs, a binary True False Value. To do this, we have run a logistic regression using each variable as a dependent variable and the binary target variable for the result. We then plot the ROC curve for each variable. The larger the area under the curve more relevant the variable on the target. The plots below are ordered by the area under the curve, with the largest starting first



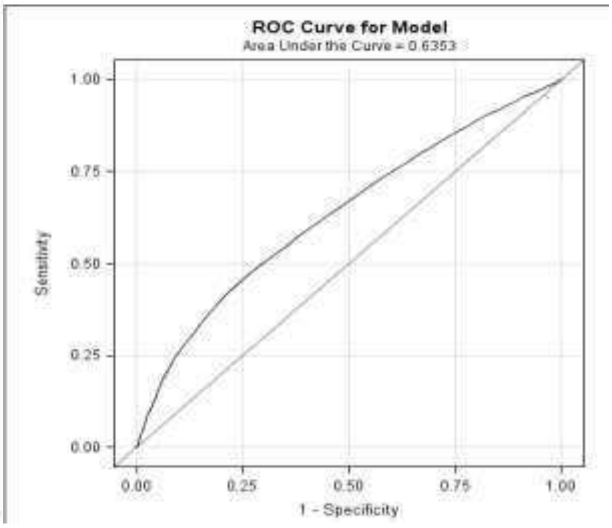
REVOLVINGUTILIZATION



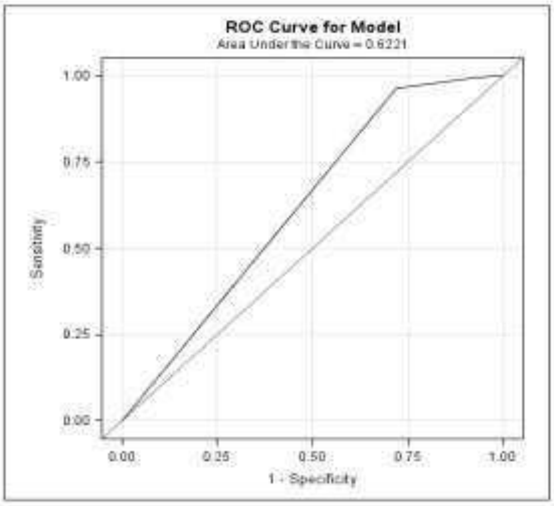
NUMBEROFTIME90DAYSLATE



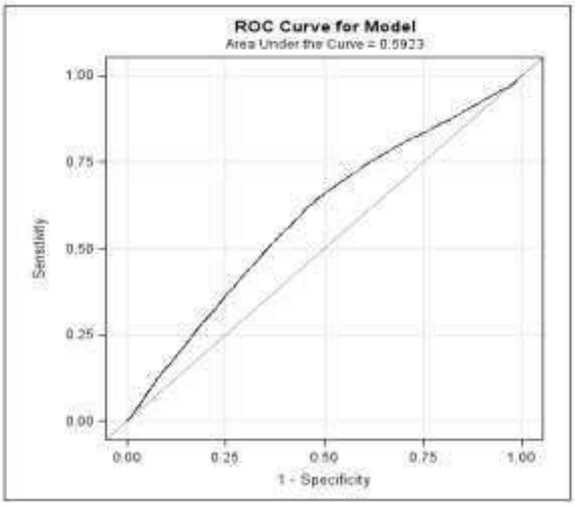
NumberofTimes30_59DaysPastDue



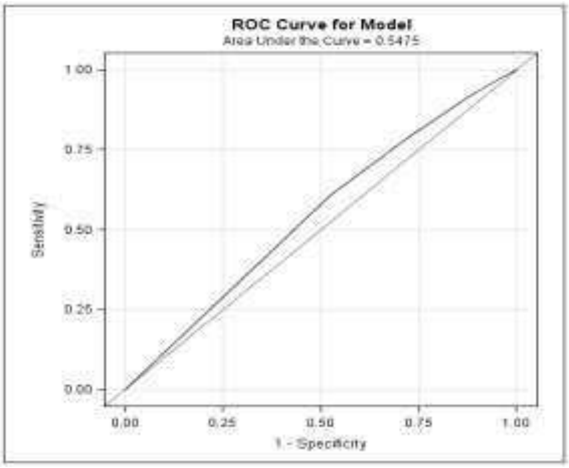
Age

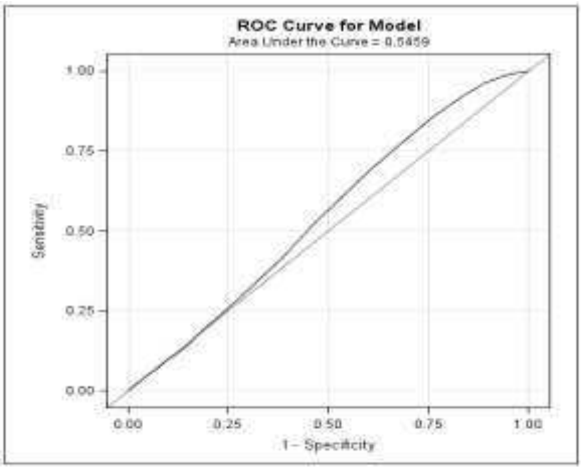


NumberofTime60_89DaysPastDue

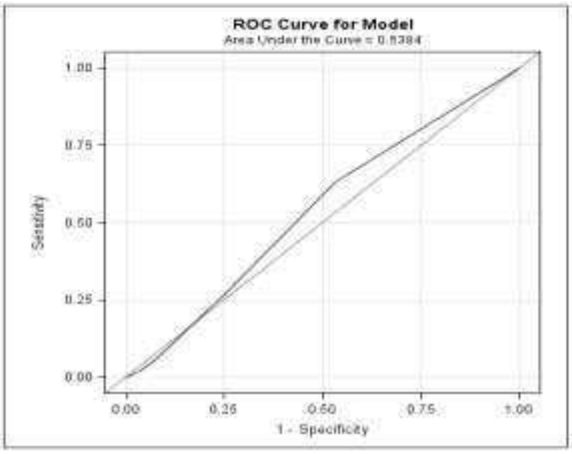


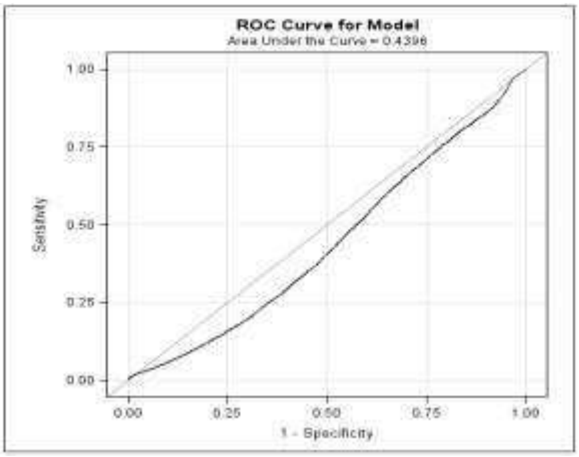
MonthlyIncome



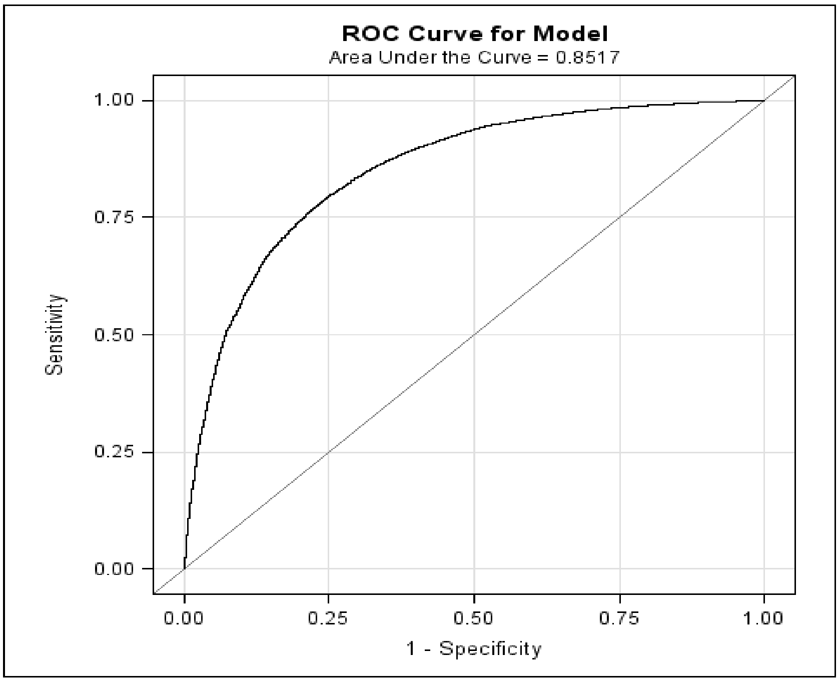


NumberofDependents **NumberofOpenCreditLinesAndLoans**





NumberOfRealEstateLinesorLoans **New_DebtRatio**



The ROC of all variables is quite good at 0.852. The individual variables seem to contribute something until the scores are below 0.5, so the New_DebtRatio is the least important according to the ROC AUC values.

MODEL SELECTION

This is a supervised learning problem with a True False binary target variable.

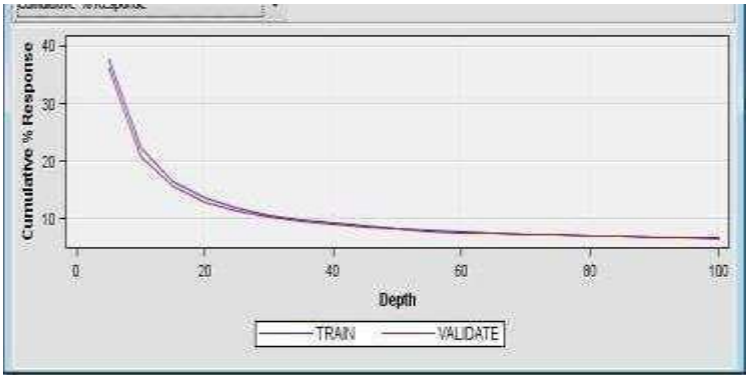
Model	Dependent Var. Type	Description
Logistic Regression	Binary/Probability	This is our base model that we always use initially for a binary or probability problem. We also use it with sequential selection to choose variables for some other models. It is linear and can separate linear data.

LOGISTIC REGRESSION:

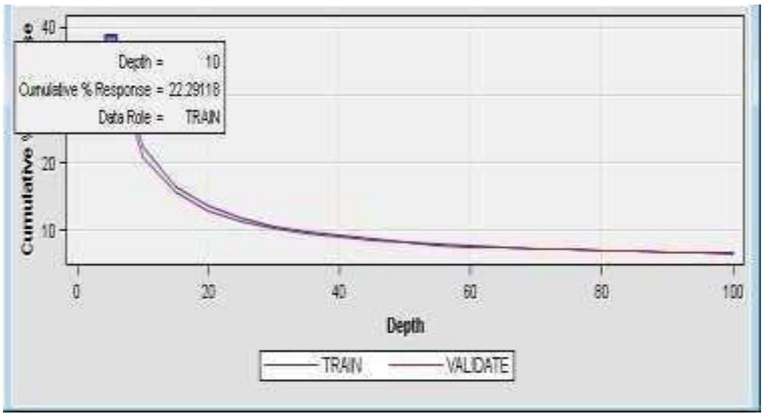
Fitting and Evaluating a Regression Model:

We are attempting to identify the relevancy of our variables on the outcome, which is SeriousDlqIn2yrs, a binary True False Value. To do this, we have run a logistic regression using each variable as a dependent variable and the binary target variable for the result.

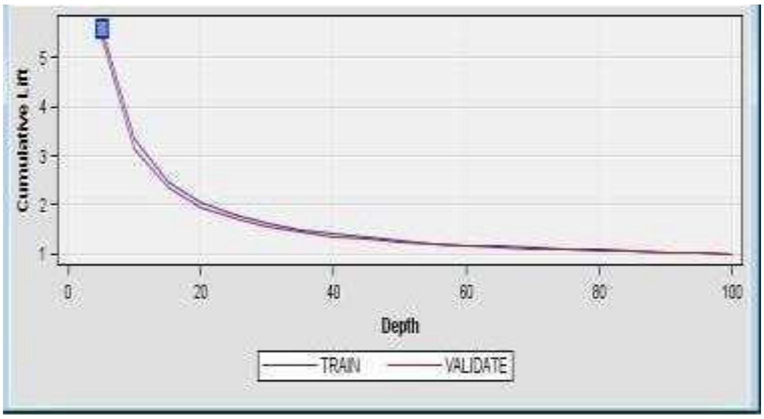
We then plot the ROC curve for each variable. The larger the area under the curve, the more relevant the variable on the target. Each point on the ROC curve represents a cutoff probability. Points closer to the upper-right corner correspond to low cutoff probabilities. Points closer to the lower-left corner correspond to higher cutoff probabilities. The performance quality of a model is indicated by the degree that the ROC curve pushes upward and to the left. This degree can be quantified as the area under the ROC curve.



The Cumulative% Response window does display smooth decrease .This might be taken as an indication of fitting subject to further evaluation

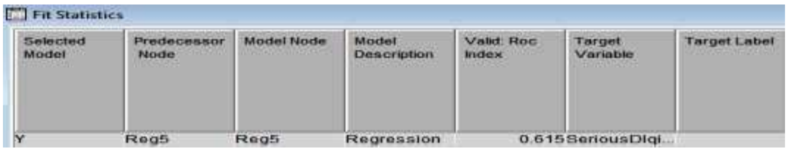


However for the first decile of data (top10%) approximately 22% of loan recipients experienced serious delinquency



Lift Charts reiterate the same information about a different scale. The overall delinquency was about 7%.The percentage of respondents in the first decile was 22% so 22/6.6=3.33

This indicates that response rate in first decile is ~ 3 times greater than response rate of population **RoC Index: 0.615 for normal logistic regression.***

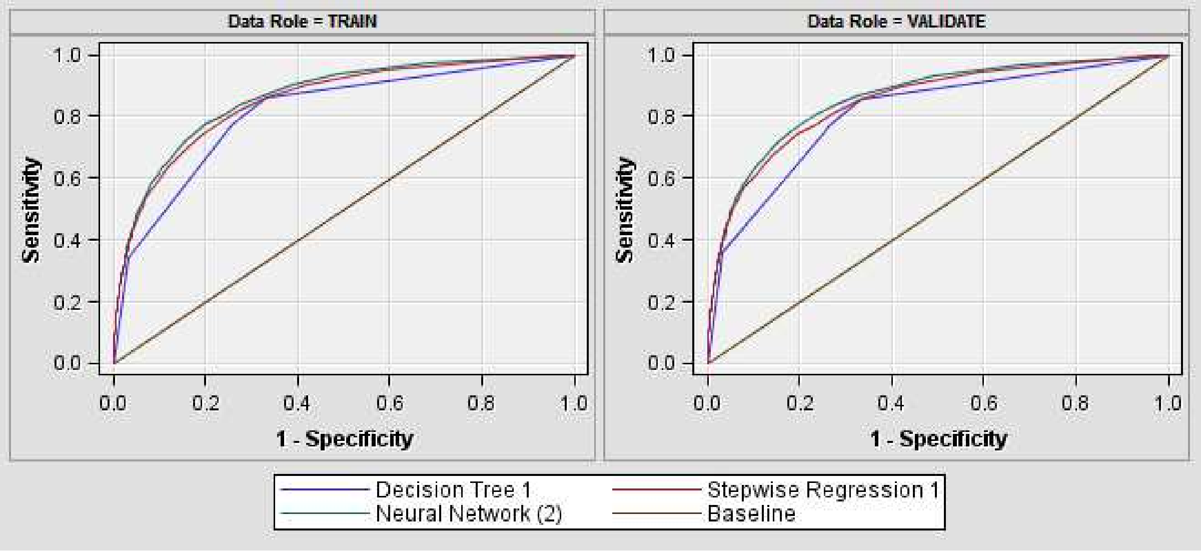


Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Valid: Roc Index	Target Variable	Target Label
Y	Reg5	Reg5	Regression	0.615	SeriousDlqi...	

VI. Assessment:

Assess the model's performance using the ROC curve and index for transformed variables:

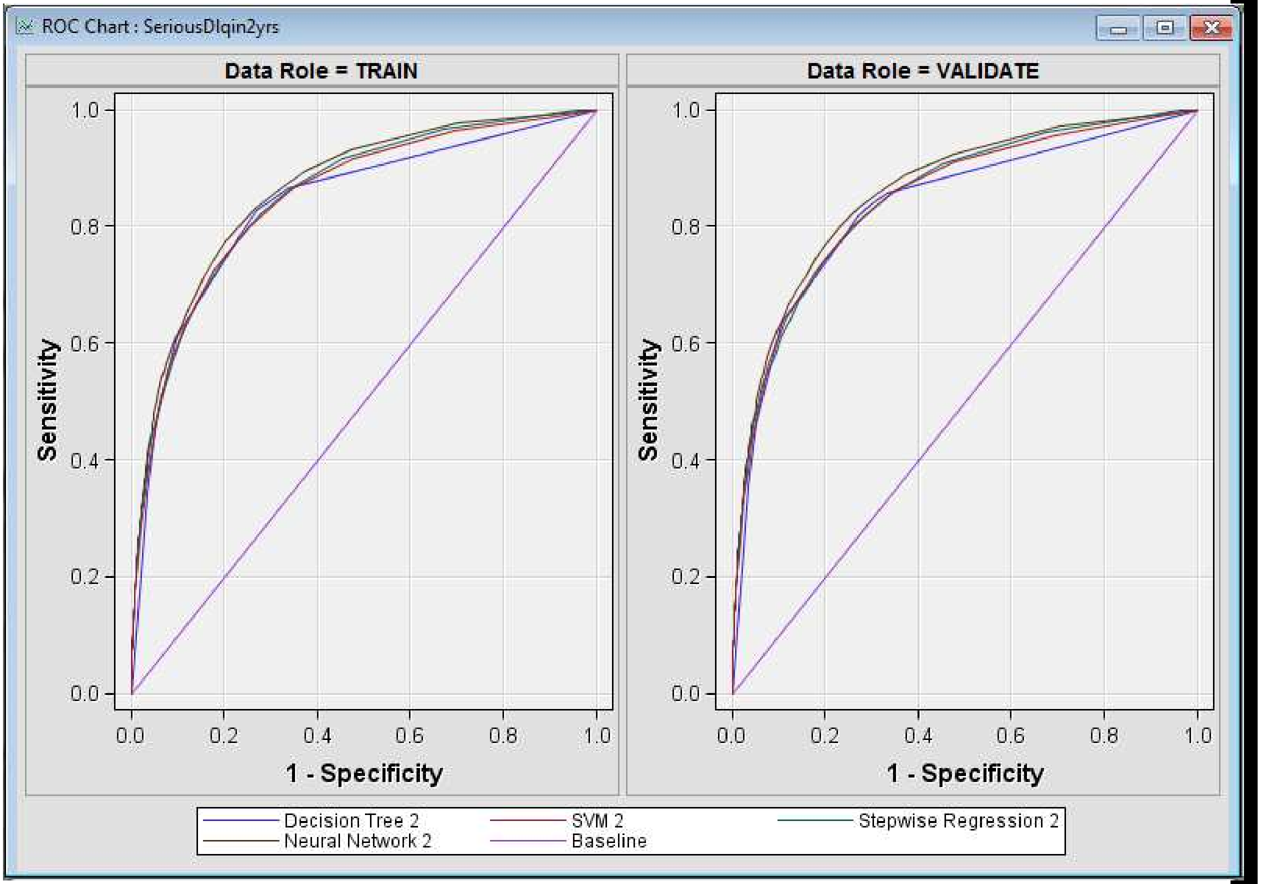


MODEL COMPARISON 1 – LOGARITHMIC TRANSFORMATION PERFORMED

Logistic regression : training =0.86

Validation = 0.85

Assess the model's performance using the ROC curve and index for non-log-transformed variables:



Model Comparison 2 – Logarithmic Transformation performed

VII. Findings:

After obtaining the sample data we started by exploring the data .We found that Histograms were very useful in Exploring and understanding the spread of the data.

Also, skewness of the data has significant effect on analysis. So, different methods need to be implemented to overcome such shortcomings of the dataset.

Also, dealing with missing values is an important step of data cleaning. Different methods like mean, percentile, regression were used to treat the missing values. Also other techniques like scatter plots, correlation were used to identify variable relevancy and importance. When modifying the data we followed tried logarithmic transformations, although they did not impact our findings. We further modeled our data using various classification techniques like logarithmic regression.

When changing the sampling method; we were able to get better results of about 78% of the True Positives (Those who will default) for the sample this is more accurate.

VIII. Conclusion:

In conclusion, we have shown that it is possible to use superior analytics algorithm through the use of ensemble methods to correctly classify customers according to their probability of default. We believe our neural network ensemble model is as close to the true model as possible. Such model can be easily updated within a single data and would have the capability to scale up for banking usage in the commercial world. We are confident that development and up-keeping of this model will help the bank gain extra profitability.