Executive 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. The goal of our project is to build the best credit scoring system that people could use to score their own risk profiles using attributes people know about their own credit behavior and usage.

Historical data is provided on 250,000 borrowers. The factors to be taken into account include annual fees, interest rates, loyalty provisions, cost of capital, write-off provisioning, transaction fees, retention incentives, balance transfers, the cost of borrowing funds on the money markets, operational costs, expected losses, the expected time to default, and economic factors. All these come into play into all sorts of decisions, and will be modeled.

In a credit card environment, riskier customers are likely to be most profitable. For example, the customers who have maxed out their limit and are making minimum payments each month pay a lot of interest. However, they are also more likely to write-off, and when they do write-off they generally take the whole limit. The least risky customers - those who pay off all their balance within the month and collect lots of loyalty points - are likely to be very unprofitable, as their lending needs to be funded (but is not offset by interest income), and their frequent flyer points need to be provisioned for.

In short, banks need to take on risky customers to make some money. The challenge is deciding how close to the cliff you want to go in terms of your 'risk appetite' without having too many customers fall over that cliff. Hence, the objective of the model is predicting the probability that somebody will experience financial distress in the next two years considering the factors mentioned above.

CONTENTS

OVERVIEW OF THE PROBLEM

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 becoming heavily demanded triggered due to the recent 2008 financial crisis 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. With this project, we tried to come up with models which can predict customer default using cutting edge analytics models such as decision tree, logistic regression and Neural Network so as to increase profit of the bank. The data is based on real consumer loan data. The data mining models include CART decision tree, neural networks and traditional logistic regression. In addition, we will conduct uni-variate and multivariate analysis of data to identify insights into banking customers.

GOAL OF THE PROJECT

The goal of the model is to find out who might be suffering delinquency in the next two years. On the other hand, the goal is to build a model that lenders such as banks can use to help making the best financial decisions.

Our team is going to use classification to analysis this problem. To classification the customers into two classes: People who experienced 90 days past due delinquency or worse in 2 years and those who do not. According to the researches we found, classification is a better solution for this problem. Moreover, classification serves an excellent tool to simplify our goal and problem statement.

METHODOLOGY

In order to evaluate models, we created a dashboard making some assumption. We first assume that the bank will provide a single loan to each customer. This loan is more like education loan and borrower don't need to pay the interest during the first two years. After two years, borrower needs to pay principal and interest all at once. Second, we assume there are only two kinds of people, one who are going to pay the whole loan and another who are not going to pay anything. Third, we define interest is 11% each year, so it is 12.1% for two years and every one loan 10,000 dollars.

Based on our assumption, we can calculate our profit by:

(Number of customer paying loan * 2100) – (number of customer not paying loan*10,000)

Following this equation, we evaluate three models (decision tree, logical regression, and neural network) and decide which one is the best.

We started off with logistic regression as it helps us decide which variables are important based on the p-value. We got Monthly income, Revolving Utilization Of Unsecured Lines, Debt Ratio, Number of times 90 days late or worse and Number of Open Credit lines and Loans as the key variables. In logical regression model, we used JMP to make the model automatically based on the R square. Then we continued modifying this model base on p value. We found that the profit is going to be highest when we select all variable. We built a model using that and using accuracy and also the profit calculated from the dashboard to compare various models.

We followed the same methodology for both decision tree and Neural Network. Both the method gave us the best models. It was clear that the logistic regression is not the proper algorithm for our project. The decision tree model from JMP provided us with the maximum profit and after many iteration we could not get more profit than the JMP model.

We applied Neural Network to build a model to predict those people who are going to experience 90 days past due or worse in the next 2 years. Variables used to construct the model include Revolving Utilization Of Unsecured Lines, Age, Number Of Time 30-59 Days Past Due Not Worse, Monthly Income, and Number Of Open Credit Lines And Loans. Except for those five, we also applied three more variables: Number Of Times 90 Days Late, Debt Ratio, and Number Of Dependents. We got out best Neural Network model using 8 variables and 17 nodes. Apart from accuracy, we used R-square to compare the model initially.

DATA

The dataset is obtained from kaggle.com entitled "Give Me Some Credit" and sponsored by a bank. The dataset contains default/non-default event of 150,000 borrowers. To decide whether accept to a loan application, the bank looks on the credit history of borrowers. Besides the predicted variable namely **SeriousDlqin2yrs**, the 10 independent variables comprise 2 types of common data sources: demographic data (age, income, number of dependents) and transactional data which includes number of open credit lines and other approved loans. At a glance, we less and more try to predict the relation between 10 independent variables and the predicted variable. For example: Since the debt ratio is defined as the division of total debt over total monthly income, we think that the higher the debt ratio is the lower chance the bank will accept the borrower's loan application. In other words, it is a high chance that debt ratio might have significant influence on default. In general, to help understand the data, we use different techniques such as Statistical techniques which includes statistical summaries or correlation analysis and Graphical techniques with the illustration of histogram or ROC curve (refer Appendix 1.1).

The given dataset of 150,000 is very huge. After discussion with the instructor and with the team member, we decided a dataset of 20,000 for training and testing is enough for the study and building a reliable model. Choosing the smaller dataset was a tricky part and the whole process is described below in detail.

After brainstorming and analysis of the data, we selected the following as the important variables in determining the credit score of customers: Age, Debt Ratio, Monthly income, Number of open credit lines and loans, Number of times 90 days late.

1. Debt Ratio:

Debt Ratio is a financial ratio that indicates the percentage of a company's assets that are provided via debt. It is the ratio of total debt (the sum of current liabilities and long-term liabilities) and total assets (the sum of current assets, fixed assets, and other assets such as 'goodwill'). **Debt ratio =Total Liability / Total Assets**

The higher the ratio, the greater risk will be associated with the firm's operation. For the purpose of the project it is defined as Monthly debt payments, alimony, living costs divided by monthly gross income.

The data pertaining to debt ratio were highly ridded with outliers. The statistical properties are greatly influenced by the outliers. The mean value of the debt ratio is 300.91 after the 0.5% of the outliers are removed from the existing data set. The same is visible in the quantile statistics as well.

2. Monthly income:

Monthly Income of the a person is inversely related to the chances of a person facing a delinquency situation. Person with higher income will have less chance of getting into trouble. But on the other hand a person with higher income but with a high debt ratio and many number of credit lines and mortages will have higher probability of delinquency.

The mean income of the given customer database was 6670.05 but this dataset was marred with both missing values and outliers. There were 29,732 data which were missing which is close to 20% of the total data. We had to find a good way to impute these missing values which doesn't

affect the overall reliability of the remaining dataset. There are many methods to deal with missing values. Few of them have been discussed in detail in the appendix. Outliers is another big concern with this variable which were dealt with using statistical tools.

3. Number of Times 90 days late:

Once a defaulter, always a defaulter. A person who has defaulted to pay the due for more than 90 days is an indication that the person is going through a rough financial patch and is highly likely that he will have serious financial trouble in 2 years. This is the most important variable of all the variables given. As learnt from decision tree model, a person who has defaulted for more than 90 days for even once, he/she will definitely default whatever be his/her financial condition or debt ratio or any other factor.

4. Number Of Time 30-59 Days Past Due Not Worse:

This variable depicts the number of times borrower has been 30-59 days past due but no worse in the last 2 years. Person with high Revolving Utilization of Credit Lines and who has defaulted more than once are at a risk of defaulting.

5. Revolving Utilization:

Revolving Utilization is counted by the 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. '0' values exist indicates that the borrowers have no money on credit cards or personal lines, but they are still able to own real estate. The minimum value is 0 and maximum value is 5.14. The mean value is 0.34 meanwhile the standard deviation is0.379

EXTRACT, TRANSFORM, LOAD, ENRICH (ETL)

The data was taken from Kaggle.com. Kaggle is a platform for predictive modeling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

The data provided had to be changed a lot. The data was not perfect; it was incomplete, noisy and inconsistent. It had many missing values and also was marred with outliers. Nevertheless, we prepared the data by doing some transformations to allow for the best models to be developed.

Our team realized that the dataset is huge to work with, it will be difficult to deal with so many variables for such huge data. So we decided to subset a smaller set of data of 20,000. Also, we realized that the dataset had 6.68% which had delinquency rate, in order to build a reliable model we need the percentage to be higher. So, first we sorted the data into 1's and 0's and took a subset of 2200 and 19800 of 1's and 0's respectively in order to have a delinquency/ default rate of around 10%. We took extra data because we knew that the data set has many missing values and also outliers. After that we divided the subset into 10,000 training and testing dataset.

ANALYSIS

The process of building a model and its analysis was based on the sole fact of increasing the profit in the excel dashboard as per our assumption and accuracy of the prediction. The two measures were preceded by using R-square and Lift as the initial motivation of improving the model. After going through initial phase of model building using Logistic Regression, we found out key variables.

We started off with logistic regression as it helps us decide which variables are important based on the p-value. We got Monthly income, Revolving Utilization Of Unsecured Lines, Debt Ratio, Number of times 90 days late or worse and Number of Open Credit lines and Loans as the key variables. In logical regression model, we use JMP to make the model automatically base on the R-square. Then we continuing modify this model base on p-value. We found that the profit is going to be highest when we select all variable. We built a model using that and using accuracy and also the profit calculated from the dashboard to compare various models.

We followed the same methodology for both the Decision tree and Neural Network. Both the methods gave us the best models. It was clear that the logistic regression is not the proper algorithm for our project. The decision tree model provided us with the maximum profit and after many iterations taking into consideration the insight got from the logistic regression we chose monthly income, debt ratio, number of times 30-59 days late, number of times 60-89 days late, number of times more than 90 days late, revolving utilization of credit lines and number of open credit lines and loans as the variables to build decision tree.

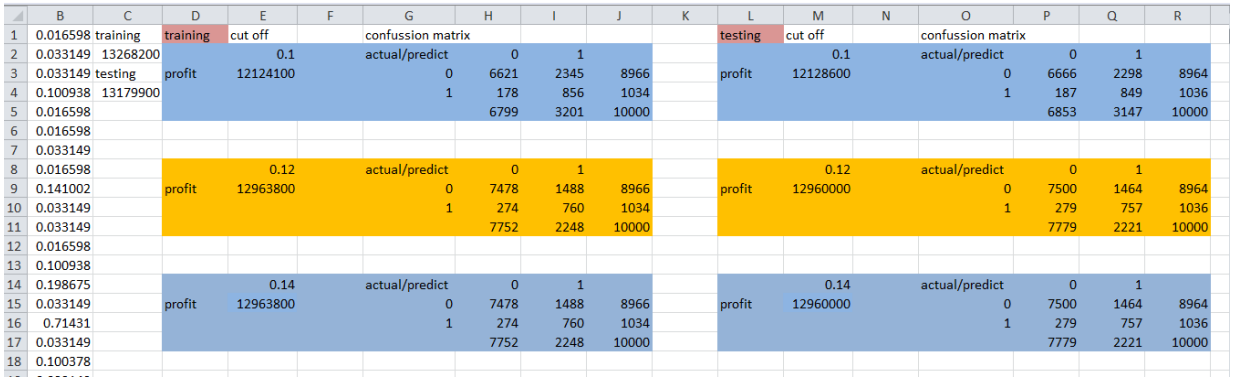
We applied Neural Network to build a model using the variables to construct the model include Revolving Utilization Of Unsecured Lines, Age, Number Of Time 30-59 Days Past Due Not Worse, Monthly Income, and Number Of Open Credit Lines And Loans. Along with the above five, we also applied three more variables: Number of Times 90 Days Late, Debt Ratio, and Number of Dependents. We got out best Neural Network model using 8 variables and 15 nodes. Apart from accuracy, we used R-square to compare the model initially.

The neural network gave us the best model with a profit of $13,572,800 with training data and $13,226,400 with testing data. The decision tree gave us a profit of $13,268,200 and $ 13179900 with training and testing data respectively.

Monetarily neural network gives us a better model but the difference between neural network's training and testing data is more than 350k and decision tree is less than 100k. This means that decision tree has a better validation compared to neural network. Hence, we chose decision tree as our best model.

PERFORMANCE MEASURES

The key performance index for our project is the Cost-Benefit and ROC. Lift and R-square value were the precedence to the main KPI.



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	0.016598	training	training	cut off		confussion matrix					testing	cut off		confussion matrix			
2	0.033149	13268200		0.1		actual/predict	0	1				0.1		actual/predict	0	1	
3	0.033149	testing	profit	12124100		0	6621	2345	8966		profit	12128600		0	6666	2298	8964
4	0.100938	13179900				1	178	856	1034					1	187	849	1036
5	0.016598						6799	3201	10000						6853	3147	10000
6	0.016598																
7	0.033149																
8	0.016598			0.12		actual/predict	0	1				0.12		actual/predict	0	1	
9	0.141002		profit	12963800		0	7478	1488	8966		profit	12960000		0	7500	1464	8964
10	0.033149					1	274	760	1034					1	279	757	1036
11	0.033149						7752	2248	10000						7779	2221	10000
12	0.016598																
13	0.100938																
14	0.198675			0.14		actual/predict	0	1				0.14		actual/predict	0	1	
15	0.033149		profit	12963800		0	7478	1488	8966		profit	12960000		0	7500	1464	8964
16	0.71431					1	274	760	1034					1	279	757	1036
17	0.033149						7752	2248	10000						7779	2221	10000
18	0.100378																

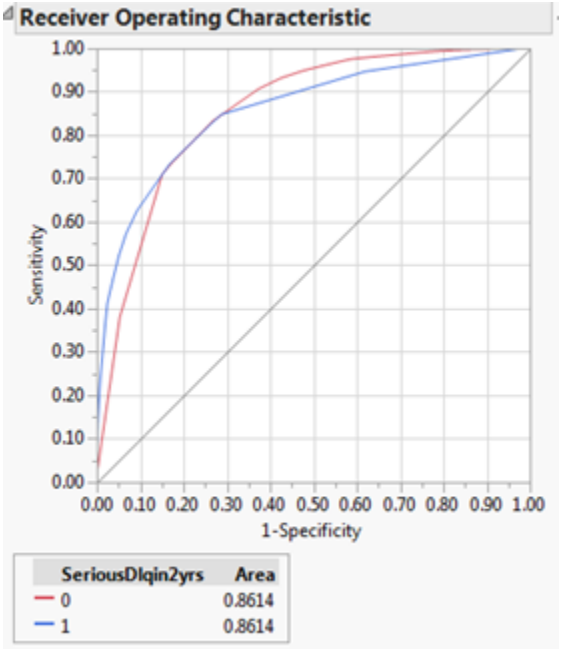
The above screenshot capture both the profit maximize and the confusion matrix which is used to calculate accuracy of the prediction of the model.

The accuracy of =a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier, i.e.

$$Accuracy = \frac{\text{TP+TN}}{\text{P+N}}$$

Accuracy can also be calculated as area under the ROC curve. Several softwares are available for that but we restricted ourselves in using ROC as only a visual aide to evaluate a particular model.

	Rsquare	Accuracy	(Training-Testing)	Cut-off
Decision Tree	33.20%		<100,000	
Neural Network	45.10%		>350,000	0.16



BUSINESS INSIGHTS

Data mining is not just a tool to feed in number and get an output. The main soul of the data mining procedure is to get an insight into the problem with a whole together a different perspective.

The decision tree model provides with a pictorial representation into the problem. From the decision tree, we can observe that a customer who has defaulted even once in the time horizon of more than 90 days, he/ she will default no matters whatever be his financial condition or other variable. A customer using 70% of his available credit line has a high probability of defaulting.

RevolvingUtilizationOfUnsecuredLines is an influential variable, while it goes down; the probability of not default goes down. It make sense, because if it is high, mean one has more credit and more willing to pay for the loan. Age, Number of Dependents or Number of open credit lines and loans are not critical factor as long as the Revolving Utilization is within limit and the customer doesn't default for more than 90days.

The striking part of the observation is that monthly income doesn't hold that much importance as someone would have imagined. Higher income is less likely to default. Higher income indicates those who are obviously able to pay for the loan. And usually indicates the one having a stable job, assets, securities, etc. Number of Open credit linea and loans is one more such variable that strike as an anomaly, Number of loan should have a direct relationship with delinquency but such observation were not seen during the project.

NumberOfTime30-59DaysPastDueNotWorse, NumberOfTime60-89DaysPastDueNotWorse, NumberOfTimes90DaysLate is similar but focus on different degree of 'not worse'. So, higher degree indicates higher probability of defaulting to pay loans. It makes sense as the higher degree indicates that one didn't pay the loan is far from the due day. Those people usually seem to default.

Higher age is less like to default as the people in the age range of 35-60 years usually have higher income and securities. So, they will pay for the loan.

Higher debt ratio is less likely to default. It don't make sense because higher debt ratio indicate more debt to one's income, if one has many debt, then he will not be able to pay the loan in the future.

Higher the NumberRealEstateLoansOrLines and NumberOfDependents it is more likely that they will default. It make sense as the numberrealestatelonaorline indicates the amount of loan one has and numberofdependents indicates the number of people dependent on the peron applying or taking the loan.

IMPROVEMENTS

As our studies mainly use financial variables as independent variables, future studies may aim at collecting more variables that are important, e.g. business environment, in improving the credit scoring accuracies.

To reach a more general conclusion, further experiments need to be conducted on other larger and better data sets. In addition, it would be meaningful to show the impact of the studied feature selection methods on other classification criteria, for example, the area under the receiver operating characteristic curve.

The assumption taken in order to compare the model has to be improved in the further studies. The salvage value of the bad loan and margin of the bank and other relevant cost has to be considered in order to better evaluate the models.

CONCLUSION

Modeling techniques like traditional statistical analyses and Data Mining techniques have been developed in order to successfully deal with the credit scoring problems. Decision Tree and logistic regressions are the most commonly used statistical credit scoring techniques, but are often criticized due to their model assumptions. On the other hand, the Data Mining approach is becoming a real alternative in credit scoring tasks due to its generalization capability, and outstanding credit scoring capability. However, it is also being criticized for its long training process, inability to identify the relative importance of potential input variables, and certain interpretative difficulties. This project has tried to investigate the accuracy of the classification models for credit scoring applications.

Although the conclusions in the project need to be validated with other data sets in future researches, a general conclusion can be drawn from this study: the automated data mining feature selection technique provides an effective method for selecting the most predictable features and Data Mining credit scoring models.

BIBLIOGRAPHY

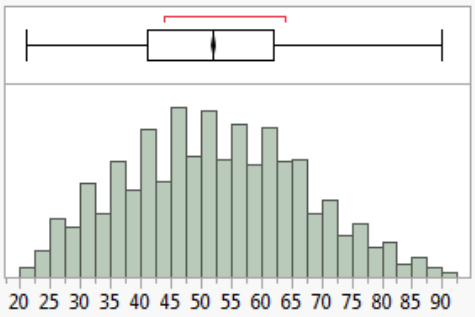
- A Two-step Method to Construct Credit Scoring Models with Data Mining Techniques, Hian Chye Koh, Associate Professor and Dean, School of Business, SIM University (Singapore) - International Journal of Business and Information, Volume 1 Number 1, 2006 pp 96-118.
- A comparison between statistical and Data Mining methods for credit scoring in case of limited available data,
- Engelmann, Bernd, Hayden, Evelyn, and Dirk Tasche, Measuring the Discriminative Power of Rating Systems, Discussion paper Series 2 Banking and Financial Supervision, No.01, 2003.
- Handling Missing Values in Data Mining, Bhavik Doshi, Department of Computer Science, Rochester Institute of Technology, Rochester, New York 14623-5603
- Statistical and Data Mining Methods in Credit Scoring, Alireza Hooman1, Govindan Marthandan1,Sasan Karamizadeh2,1Multimedia University(MMU) Faculty of Management(FOM) Cyberjaya,Malaysia, 2Advanced Informatics School(AIS), Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia
- Business Intelligence, Song JianYong - A0073004X, Abuzar S/O Yakaram - A0083234L, Li Yong - A0091027M, Ling Chi Tao - A0086081E, Nguyen Thi Bich Van - A0074274A

APPENDIX

1. Key Summary Statistics:

1.1 Distributions:

Age



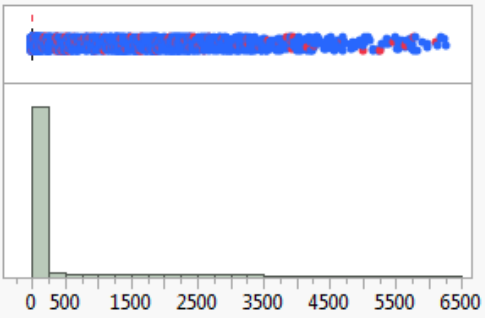
Quantiles

100.0%	maximum	90
99.5%		87
97.5%		81
90.0%		71
75.0%	quartile	62
50.0%	median	52
25.0%	quartile	41
10.0%		32
2.5%		26
0.5%		23
0.0%	minimum	21

Summary Statistics

Mean	51.9058
Std Dev	14.508411
Std Err Mean	0.1450841
Upper 95% Mean	52.190194
Lower 95% Mean	51.621406
N	10000

DebtRatio



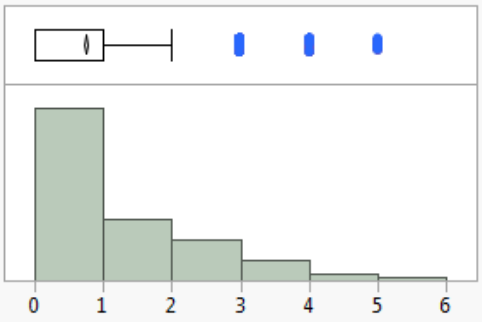
Quantiles

100.0%	maximum	6250
99.5%		5000.91
97.5%		3279.85
90.0%		1203
75.0%	quartile	0.85186
50.0%	median	0.36399
25.0%	quartile	0.16973
10.0%		0.02796
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	300.91046
Std Dev	870.21546
Std Err Mean	8.7021546
Upper 95% Mean	317.96843
Lower 95% Mean	283.85249
N	10000

NumberOfDependents



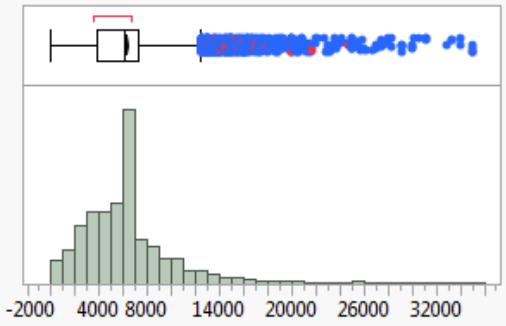
Quantiles

100.0%	maximum	5
99.5%		4
97.5%		3
90.0%		2
75.0%	quartile	1
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.7479
Std Dev	1.0632303
Std Err Mean	0.0106323
Upper 95% Mean	0.7687415
Lower 95% Mean	0.7270585
N	10000

MonthlyIncome



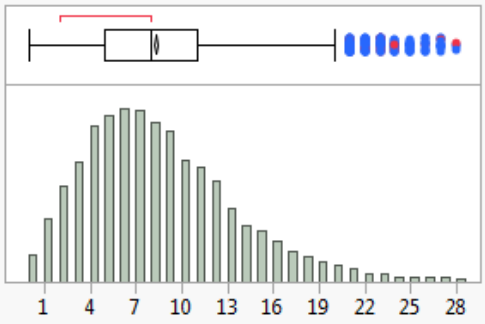
Quantiles

100.0%	maximum	35000
99.5%		24663.4
97.5%		16000
90.0%		10500
75.0%	quartile	7267.25
50.0%	median	6166
25.0%	quartile	3800
10.0%		2200
2.5%		736.1
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	6265.7858
Std Dev	3838.5926
Std Err Mean	38.385926
Upper 95% Mean	6341.0299
Lower 95% Mean	6190.5417
N	10000

NumberOfOpenCreditLinesAndLoans



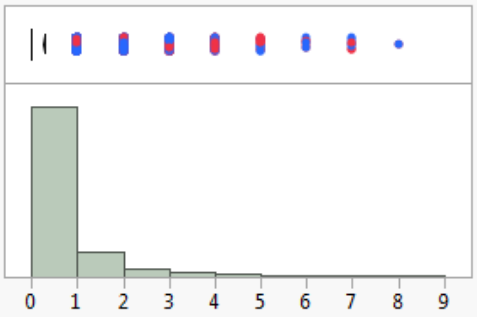
Quantiles

100.0%	maximum	28
99.5%		25
97.5%		20
90.0%		15
75.0%	quartile	11
50.0%	median	8
25.0%	quartile	5
10.0%		3
2.5%		1
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	8.293
Std Dev	4.8841311
Std Err Mean	0.0488413
Upper 95% Mean	8.3887388
Lower 95% Mean	8.1972612
N	10000

NumberOfTime30-59DaysPastDueNotWorse



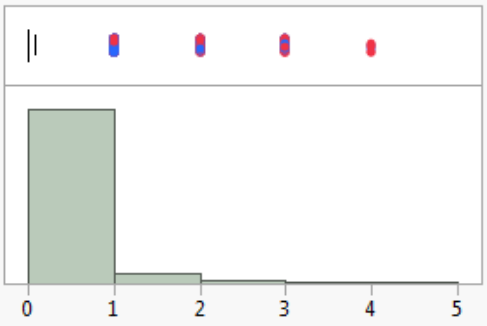
Quantiles

100.0%	maximum	8
99.5%		4
97.5%		2
90.0%		1
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.2693
Std Dev	0.7269322
Std Err Mean	0.0072693
Upper 95% Mean	0.2835493
Lower 95% Mean	0.2550507
N	10000

NumberOfTime60-89DaysPastDueNotWorse



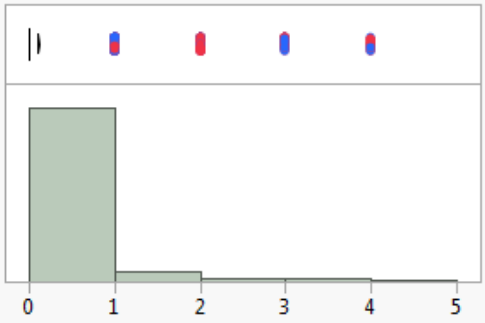
Quantiles

100.0%	maximum	4
99.5%		2
97.5%		1
90.0%		0
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.0747
Std Dev	0.3407808
Std Err Mean	0.0034078
Upper 95% Mean	0.08138
Lower 95% Mean	0.06802
N	10000

NumberOfTimes90DaysLate



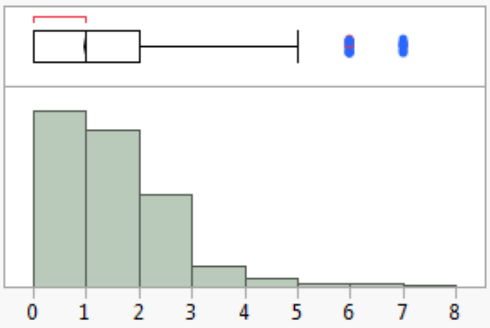
Quantiles

100.0%	maximum	4
99.5%		3
97.5%		1
90.0%		0
75.0%	quartile	0
50.0%	median	0
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.0996
Std Dev	0.4288336
Std Err Mean	0.0042883
Upper 95% Mean	0.108006
Lower 95% Mean	0.091194
N	10000

NumberRealEstateLoansOrLines



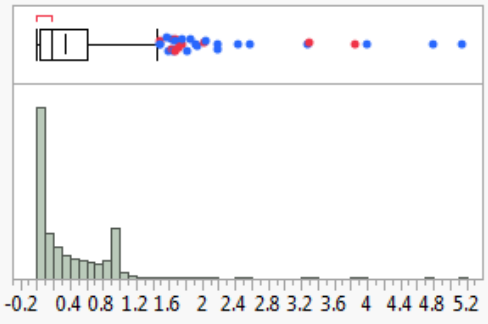
Quantiles

100.0%	maximum	7
99.5%		5
97.5%		3
90.0%		2
75.0%	quartile	2
50.0%	median	1
25.0%	quartile	0
10.0%		0
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.9787
Std Dev	1.0277898
Std Err Mean	0.0102779
Upper 95% Mean	0.9988467
Lower 95% Mean	0.9585533
N	10000

RevolvingUtilizationOfUnsecuredLines



Quantiles

100.0%	maximum	5.14874
99.5%		1.24587
97.5%		1
90.0%		1
75.0%	quartile	0.61286
50.0%	median	0.17063
25.0%	quartile	0.03256
10.0%		0.00388
2.5%		0
0.5%		0
0.0%	minimum	0

Summary Statistics

Mean	0.3423683
Std Dev	0.3794371
Std Err Mean	0.0037944
Upper 95% Mean	0.3498061
Lower 95% Mean	0.3349306
N	10000

1.2 Correlation Analysis:

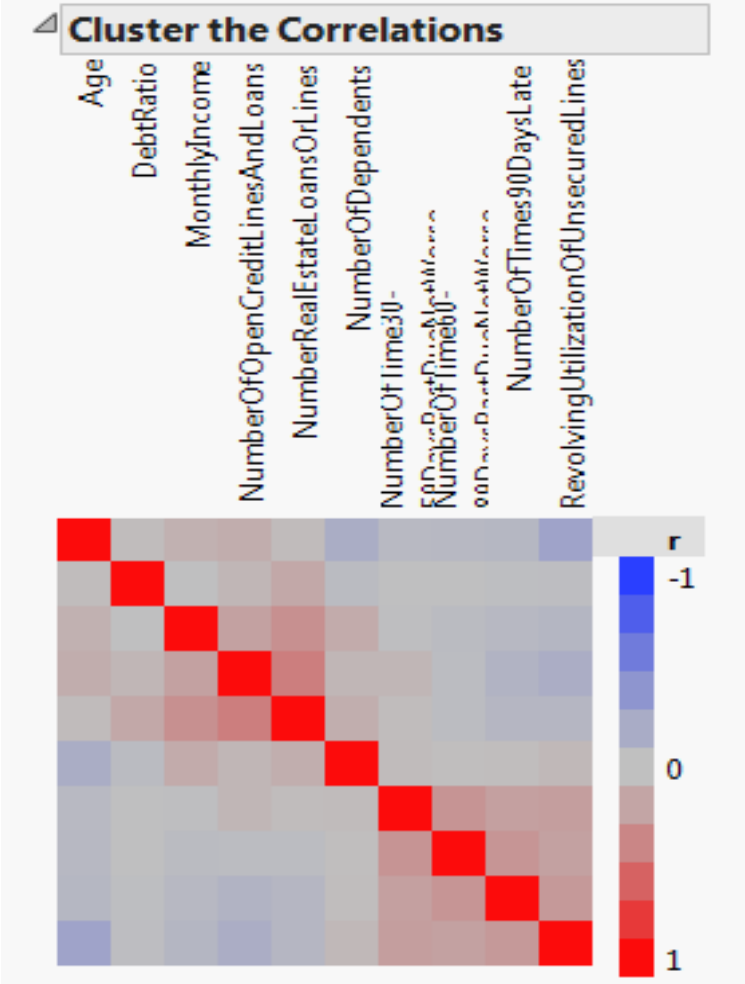
Row	Age	DebtRatio	MonthlyIncome	NumberOfDependents	NumberOfOpenCreditLinesAndLoans	NumberOfTime30-59DaysPastDueNotWorse	NumberOfTime60-89DaysPastDueNotWorse	NumberOfTimes90DaysLate	NumberRealEstateLoansOrLines	RevolvingUtilizationOfUnsecuredLines
Age	1	0.0290182	0.11725	-0.191254029	0.145053436	-0.071909963	-0.079285523	-0.094922185	0.036767	-0.275237
DebtRatio	0.0290182	1	-0.007	-0.056295367	0.077618315	-0.011181471	-0.008935342	-0.025530767	0.17781	-0.026816
MonthlyIncome	0.1172526	-0.007022	1	0.159582771	0.22951634	-0.022057107	-0.051596051	-0.084372206	0.335806	-0.104023
NumberOfDependents	-0.191254	-0.056295	0.15958	1	0.081669805	0.042429768	0.018029288	0.025683432	0.137764	0.066872
NumberOfOpenCreditLinesAndLoans	0.1450534	0.0776183	0.22952	0.081669805	1	0.076081479	-0.044757119	-0.135981899	0.444946	-0.184179
NumberOfTime30-59DaysPastDueNotWorse	-0.07191	-0.011181	-0.0221	0.042429768	0.076081479	1	0.31442624	0.237655783	0.03271	0.2488352
NumberOfTime60-89DaysPastDueNotWorse	-0.079286	-0.008935	-0.0516	0.018029288	-0.044757119	0.31442624	1	0.309052581	-0.042	0.2262619
NumberOfTimes90DaysLate	-0.094922	-0.025531	-0.0844	0.025683432	-0.135981899	0.237655783	0.309052581	1	-0.1007	0.2823622
NumberRealEstateLoansOrLines	0.0367666	0.1778103	0.33581	0.137763959	0.444945502	0.032709729	-0.041999484	-0.100698535	1	-0.103056
RevolvingUtilizationOfUnsecuredLines	-0.275237	-0.026816	-0.104	0.066871976	-0.184178993	0.248835196	0.226261876	0.282362202	-0.10306	1

Univariate Simple Statistics

Column	N	DF	Mean	Std Dev	Sum	Minimum	Maximum
Age	10000	9999.00	51.9058	14.5084	519058	21.0000	90.0000
DebtRatio	10000	9999.00	300.910	870.215	3009105	0.0000	6250.00
MonthlyIncome	10000	9999.00	6265.79	3838.59	6.27e+7	0.0000	35000.0
NumberOfDependents	10000	9999.00	0.7479	1.0632	7479.00	0.0000	5.0000
NumberOfOpenCreditLinesAndLoans	10000	9999.00	8.2930	4.8841	82930.0	0.0000	28.0000
NumberOfTime30-59DaysPastDueNotWorse	10000	9999.00	0.2693	0.7269	2693.00	0.0000	8.0000
NumberOfTime60-89DaysPastDueNotWorse	10000	9999.00	0.0747	0.3408	747.000	0.0000	4.0000
NumberOfTimes90DaysLate	10000	9999.00	0.0996	0.4288	996.000	0.0000	4.0000
NumberRealEstateLoansOrLines	10000	9999.00	0.9787	1.0278	9787.00	0.0000	7.0000
RevolvingUtilizationOfUnsecuredLines	10000	9999.00	0.3424	0.3794	3423.68	0.0000	5.1487

Multivariate Simple Statistics

Column	N	DF	Mean	Std Dev	Sum	Minimum	Maximum
Age	10000	9999.00	51.9058	14.5084	519058	21.0000	90.0000
DebtRatio	10000	9999.00	300.910	870.215	3009105	0.0000	6250.00
MonthlyIncome	10000	9999.00	6265.79	3838.59	6.27e+7	0.0000	35000.0
NumberOfDependents	10000	9999.00	0.7479	1.0632	7479.00	0.0000	5.0000
NumberOfOpenCreditLinesAndLoans	10000	9999.00	8.2930	4.8841	82930.0	0.0000	28.0000
NumberOfTime30-59DaysPastDueNotWorse	10000	9999.00	0.2693	0.7269	2693.00	0.0000	8.0000
NumberOfTime60-89DaysPastDueNotWorse	10000	9999.00	0.0747	0.3408	747.000	0.0000	4.0000
NumberOfTimes90DaysLate	10000	9999.00	0.0996	0.4288	996.000	0.0000	4.0000
NumberRealEstateLoansOrLines	10000	9999.00	0.9787	1.0278	9787.00	0.0000	7.0000
RevolvingUtilizationOfUnsecuredLines	10000	9999.00	0.3424	0.3794	3423.68	0.0000	5.1487



We have plotted a scatter plot matrix for all the variables and tried to find the correlation among them. We can observe from the scatter plot matrix and the correlation table, the positive and negative correlations among various variables.

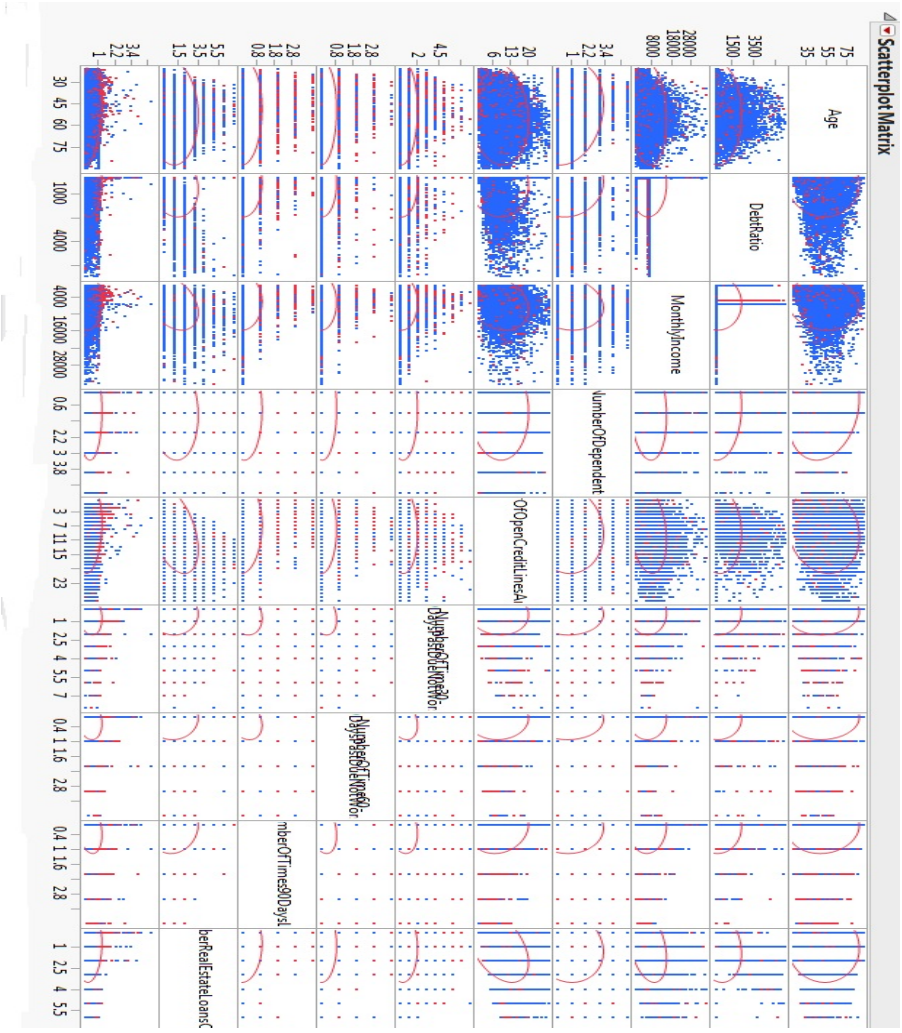
There is a strong positive correlation between the variables 'Number of times borrower has been 30-59 days past due but no worse in the last 2 years' and 'Number of times borrower has been 60-89 days past due but no worse in the last 2 years' and 'Number of times borrower has been 90 days or more past due'. This strong correlation shows that once a customer has defaulted, there is a high possibility that he will default in the successive payments.

Age and No. of dependents are negatively correlated. The correlation coefficient is -0.2171. This is along the anticipated lines because as the person grows old, the no. of dependents decreases.

Debt Ratio is total monthly debt payments, alimony, living costs divided by monthly gross income. So from the formula, we can deduce that there is an inverse relationship between debt ratio and monthly income which is visible in the scatter plot and the correlation matrix.

As observed from the scatter plot, there is also a strong correlation between Number of open credit lines and loans and Number real estate loans or lines with a coefficient of 0.4340. This is expected due to the fact that the variable 'Number of open credit lines and loans' is defined as Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards) which overlaps with that of Number of Real estate loans or lines.

Thus, we can say that there are variables which have direct or inverse correlation which will help us build a model considering only the significant variables.



1.3 Outliers

An outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set.

Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

We can use the following methods for dealing with outliers:

- **Mahalanobis Distance**:

The Mahalanobis Outlier Distance plot shows the Mahalanobis distance of each point from the multivariate mean (centroid). The standard Mahalanobis distance depends on estimates of the mean, standard deviation, and correlation for the data. The distance is plotted for each observation number. Extreme multivariate outliers can be identified by highlighting the points with the largest distance values.

- **Jackknife Distance:**

The Jackknife Distances plot shows distances that are calculated using a jackknife technique. The distance for each observation is calculated with estimates of the mean, standard deviation, and correlation matrix that do not include the observation itself. The jack-knifed distances are useful when there is an outlier. In this case, the Mahalanobis distance is distorted and tends to disguise the outlier or make other points look more outlying than they are.

- **T^2 Statistic:**

The T2 plot shows distances that are the square of the Mahalanobis distance. This plot is preferred for multivariate control charts. The plot includes the value of the calculated T2 statistic, as well as its upper control limit. Values that fall outside this limit might be outliers.

1.4 Missing Values

Missing Values and its problems are very common in the data cleaning process. Several methods have been proposed so as to process missing data in datasets and avoid problems caused by it. Missing data is a familiar and unavoidable problem in large datasets and is widely discussed in the field of data mining and statistics. Sometimes program environments may provide code for missing data but they lack standardization and are rarely used. Thus analyzing the impact of problems caused by missing values and finding solutions to tackle with them is an important issue in the field of Data Cleaning and Preparation. Many solutions have been presented regarding this issue and handling missing values is still a topic which is being worked upon. In this paper we discuss various hitches we face when it comes to missing data and see how they can be resolved.

For missing values, several methods can be employed to be solved.

- **K-nearest neighbor (KNN)**

For each data point which has missing value, KNN can be used to find k-neighbors. If a missing field is categorical, 1 to k bit encoding is needed for data processing since the distance between two categories should be the same. If missing field is quantitative, then KNN can be applied directly. However, data normalization might needs in order to give same weight for each feature.

This process needs extra programming to do pre-processing.

- **Regression**

We replace the missing value based upon linear regression value. A regression model is estimated to predict observed values of a variable based on other variables, and that model used to impute value in case where that variable is missing.

- **Maximum likelihood estimation**

Compare each tuple with missing data and tuple with complete data and maximize likelihood to estimate the missing data. The advantage of ML is data consistency and unbiased estimation. (ML has presumption that missing data is MAR (Missing at random).

- **Dummy Variable Method**

We can create a new column to record if any original column has missing data. If the tuple has missing data, we assign 1; otherwise, we assign 0 to the new column.

- **Mean**

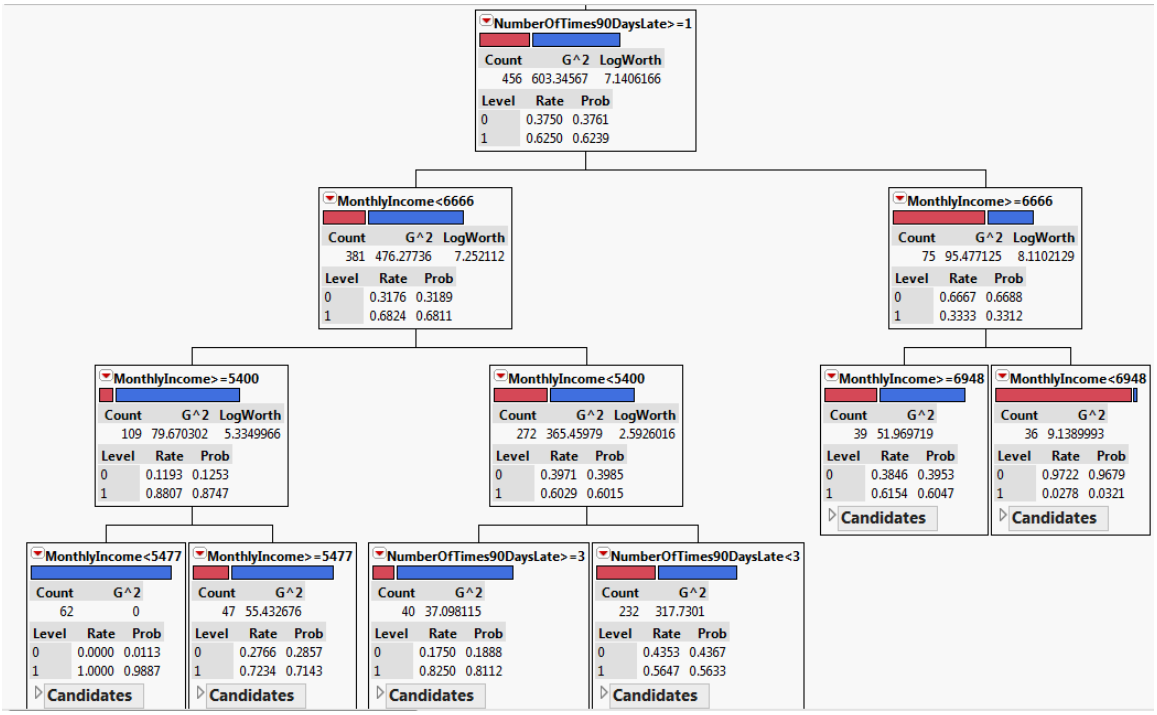
Another method is substituting a mean for the missing data. Using mean substitution makes only a trivial change in the correlation coefficient and no change in the regression coefficient.

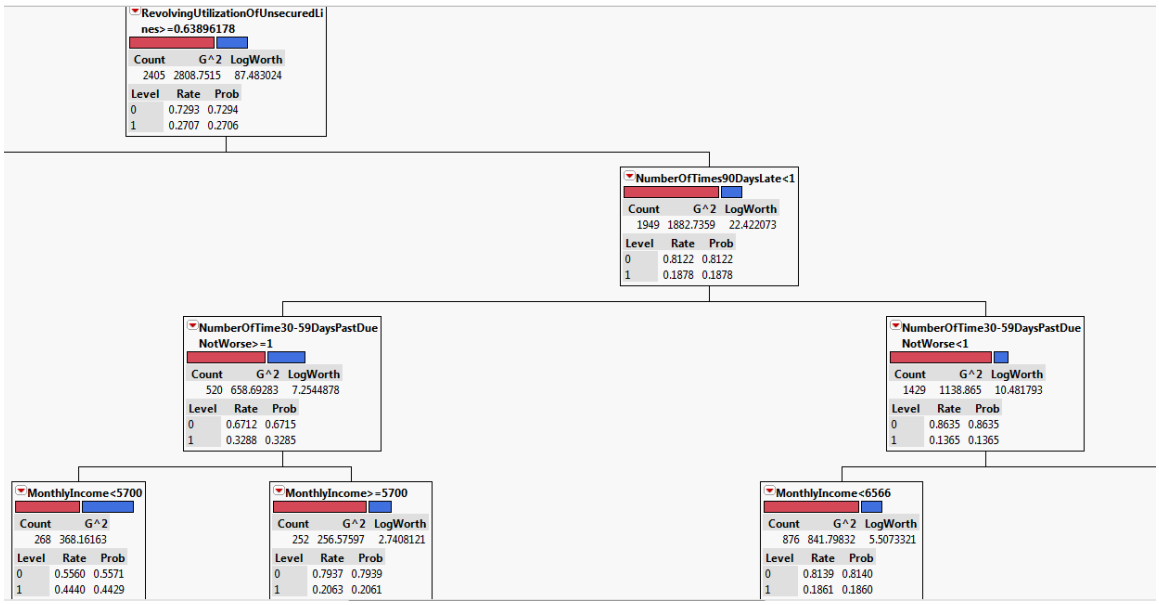
- **Deletion**

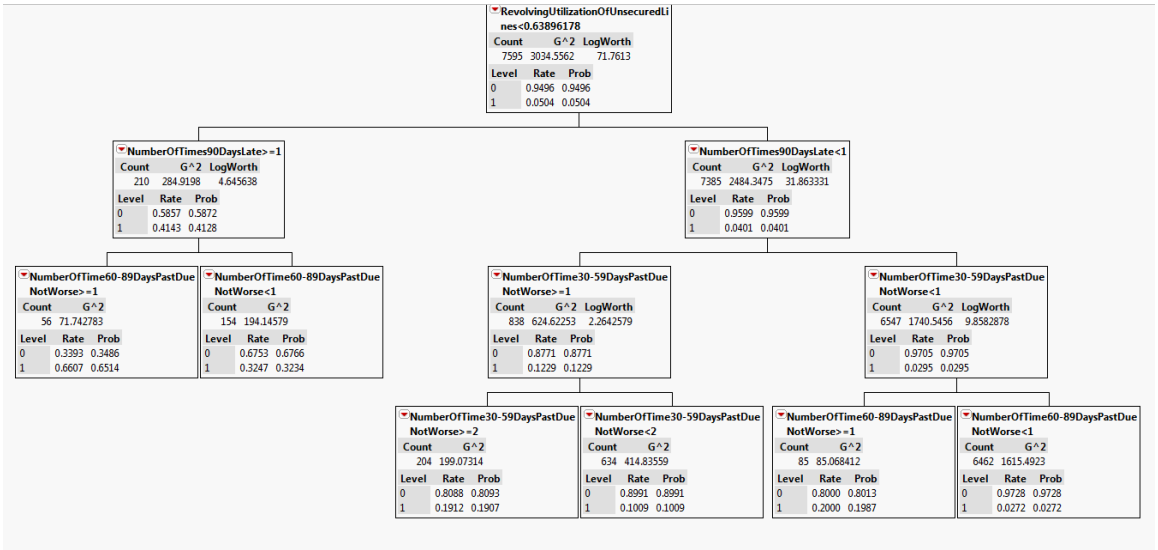
Another option is to delete cases with missing values. For every missing value in the dataset, you can delete the subjects with those missing values. Thus, you are left with complete data for all subjects. The disadvantage to this approach is you reduce the sample size of your data. If you have a large dataset, then it may not be a big disadvantage because you have enough subjects even after you delete the cases with missing values. Another disadvantage to this approach is that the subjects with missing values may be different than the subjects without missing values (e.g., missing values that are non-random), so you have a non-representative sample after removing the cases with missing values. Data cleaning and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss approaches to deal with missing data in different scenarios.

Data cleaning and preparation is the primary step in data mining process. We first identify different types of missing data and then discuss approaches to deal with missing data in different scenarios.

2. **Best Model:**


NumberOfTimes90DaysLate>=1
Count G^2 LogWorth
456 603.34567 7.1406166
Level Rate Prob
0 0.3750 0.3761
1 0.6250 0.6239
MonthlyIncome<6666
Count G^2 LogWorth
381 476.27736 7.252112
Level Rate Prob
0 0.3176 0.3189
1 0.6824 0.6811
MonthlyIncome>=6666
Count G^2 LogWorth
75 95.477125 8.1102129
Level Rate Prob
0 0.6667 0.6688
1 0.3333 0.3312
MonthlyIncome>=5400
Count G^2 LogWorth
109 79.670302 5.3349966
Level Rate Prob
0 0.1193 0.1253
1 0.8807 0.8747
MonthlyIncome<5400
Count G^2 LogWorth
272 365.45979 2.5926016
Level Rate Prob
0 0.3971 0.3985
1 0.6029 0.6015
MonthlyIncome>=6948
Count G^2
39 51.969719
Level Rate Prob
0 0.3846 0.3953
1 0.6154 0.6047
Candidates
MonthlyIncome<6948
Count G^2
36 9.1389993
Level Rate Prob
0 0.9722 0.9679
1 0.0278 0.0321
Candidates
MonthlyIncome<5477
Count G^2
62 0
Level Rate Prob
0 0.0000 0.0113
1 1.0000 0.9887
Candidates
MonthlyIncome>=5477
Count G^2
47 55.432676
Level Rate Prob
0 0.2766 0.2857
1 0.7234 0.7143
Candidates
NumberOfTimes90DaysLate>=3
Count G^2
40 37.098115
Level Rate Prob
0 0.1750 0.1888
1 0.8250 0.8112
Candidates
NumberOfTimes90DaysLate<3
Count G^2
232 317.7301
Level Rate Prob
0 0.4353 0.4367
1 0.5647 0.5633
Candidates


RevolvingUtilizationOfUnsecuredLines>=0.63896178
Count G^2 LogWorth
2405 2808.7515 87.483024
Level Rate Prob
0 0.7293 0.7294
1 0.2707 0.2706
NumberOfTimes90DaysLate<1
Count G^2 LogWorth
1949 1882.7359 22.422073
Level Rate Prob
0 0.8122 0.8122
1 0.1878 0.1878
NumberOfTime30-59DaysPastDueNotWorse>=1
Count G^2 LogWorth
520 658.69283 7.2544878
Level Rate Prob
0 0.6712 0.6715
1 0.3288 0.3285
NumberOfTime30-59DaysPastDueNotWorse<1
Count G^2 LogWorth
1429 1138.865 10.481793
Level Rate Prob
0 0.8635 0.8635
1 0.1365 0.1365
MonthlyIncome<5700
Count G^2
268 368.16163
Level Rate Prob
0 0.5560 0.5571
1 0.4440 0.4429
MonthlyIncome>=5700
Count G^2 LogWorth
252 256.57597 2.7408121
Level Rate Prob
0 0.7937 0.7939
1 0.2063 0.2061
MonthlyIncome<6566
Count G^2 LogWorth
876 841.79832 5.5073321
Level Rate Prob
0 0.8139 0.8140
1 0.1861 0.1860



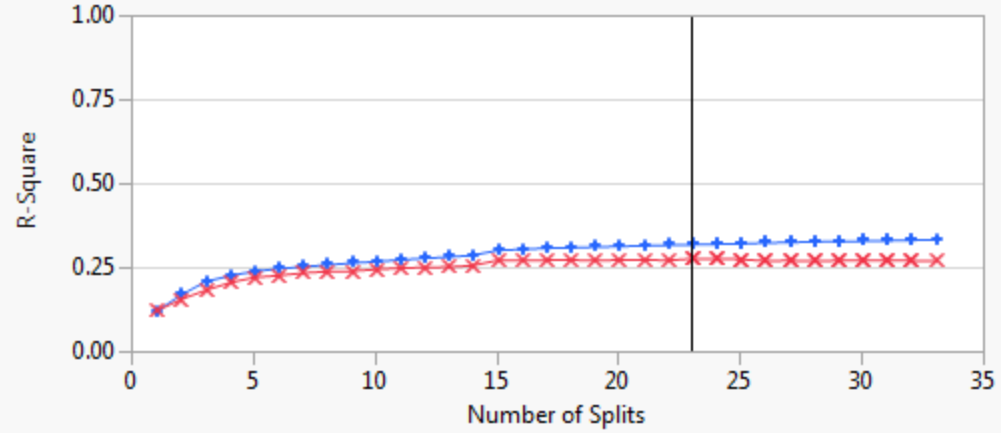
Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3260	0.2829	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.4012	0.3532	(1-(L(0)/L(model))^(2/n))/(1-L(0)^(2/n))
Mean -Log p	0.2241	0.2387	∑ -Log(ρ[j])/n
RMSE	0.2515	0.2608	√ ∑(y[j]-ρ[j])²/n
Mean Abs Dev	0.1268	0.1312	∑ |y[j]-ρ[j]|/n
Misclassification Rate	0.0807	0.0902	∑ (ρ[j]≠ρMax)/n
N	10000	10000	n

Confusion Matrix

Actual	Predicted		Actual	Predicted	
Training	**0**	**1**	**Validation**	**0**	**1**
0	8765	201	0	8719	245
1	606	428	1	657	379

Split History

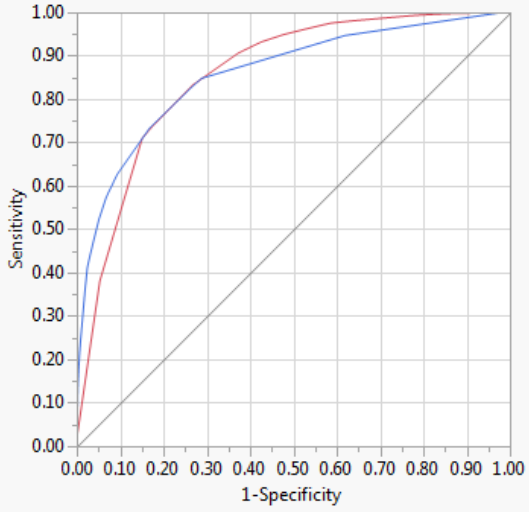


Validation Data in Red

Column Contributions

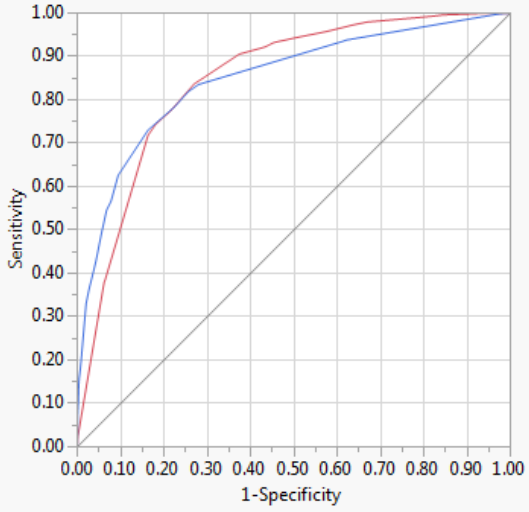
Term	Number of Splits	G^2	Portion
RevolvingUtilizationOfUnsecuredLines	1	806.49128	0.3714
NumberOfTimes90DaysLate	3	598.590471	0.2757
MonthlyIncome	9	325.290923	0.1498
NumberOfTime30-59DaysPastDueNotWorse	4	226.557219	0.1043
DebtRatio	2	98.114515	0.0452
NumberOfTime60-89DaysPastDueNotWorse	3	71.6789748	0.0330
Age	1	27.9107908	0.0129
NumberRealEstateLoansOrLines	1	16.8264535	0.0077
NumberOfDependents	0	0	0.0000
NumberOfOpenCreditLinesAndLoans	0	0	0.0000

Receiver Operating Characteristic

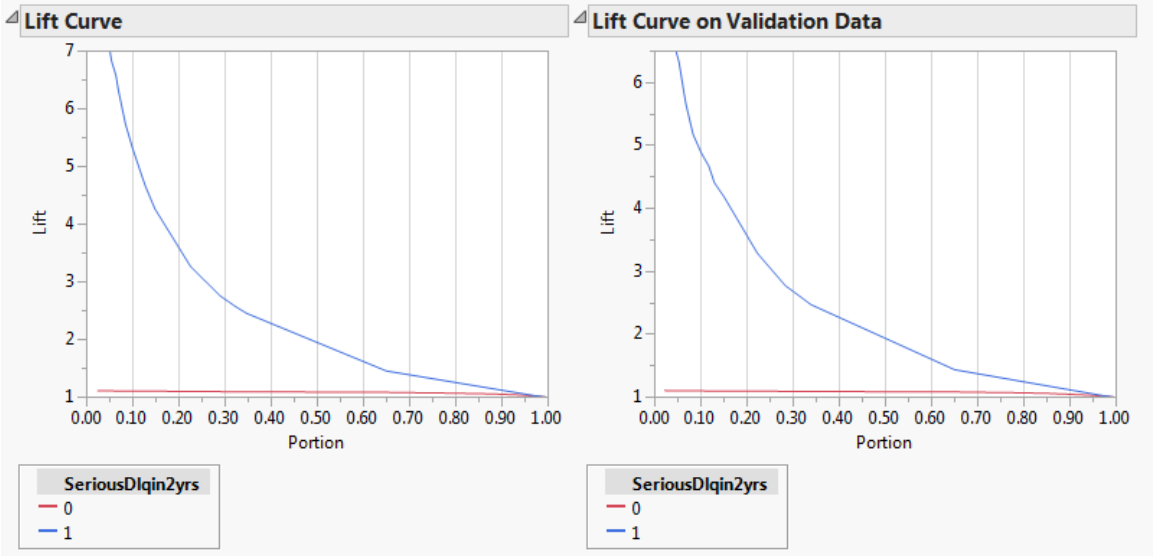


SeriousDlqin2yrs	Area
0	0.8614
1	0.8614

Receiver Operating Characteristic on Validation Data



SeriousDlqin2yrs	Area
0	0.8492
1	0.8492



Profit:

Confusion Matrix

Training

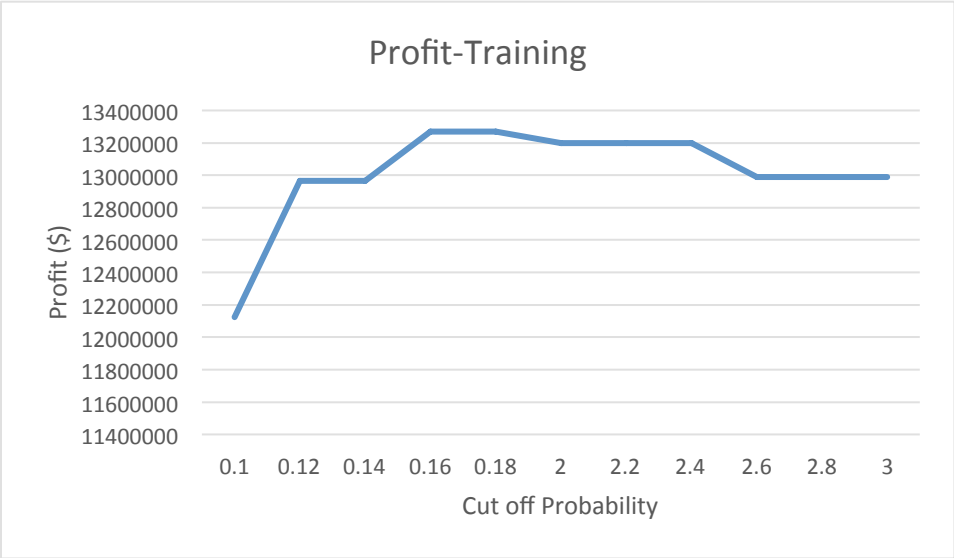
Cut Off	0.16		Actual/predict	0	1	
Profit	13268200		0	8142	824	8966
			1	383	651	1034
				8525	1475	10000

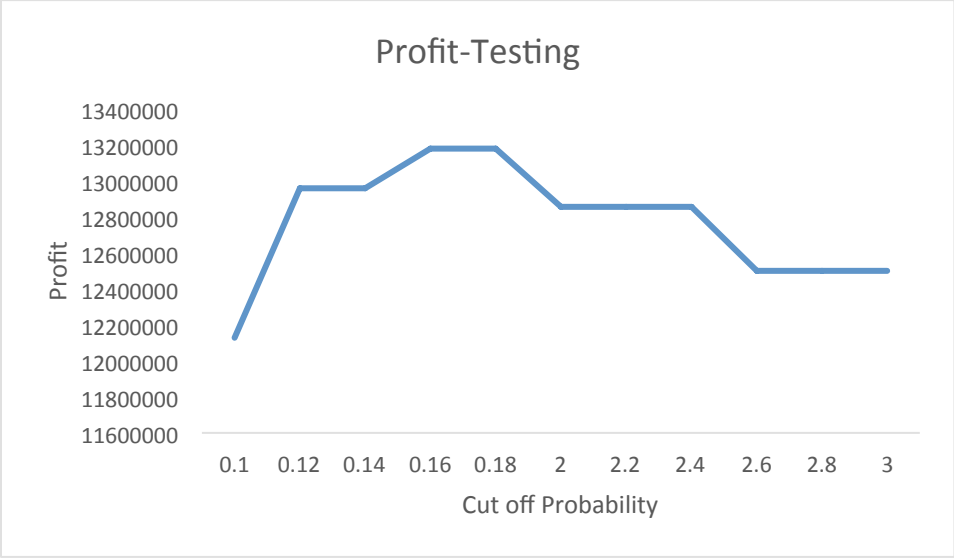
Cut Off	0.18		Actual/predict	0	1	
Profit	13268200		0	8142	824	8966
			1	383	651	1034
				8525	1475	10000

Testing

Cut Off	0.16		Actual/predict	0	1	
Profit	13179900		0	8119	845	8964
			1	387	649	1036
				8506	1494	10000

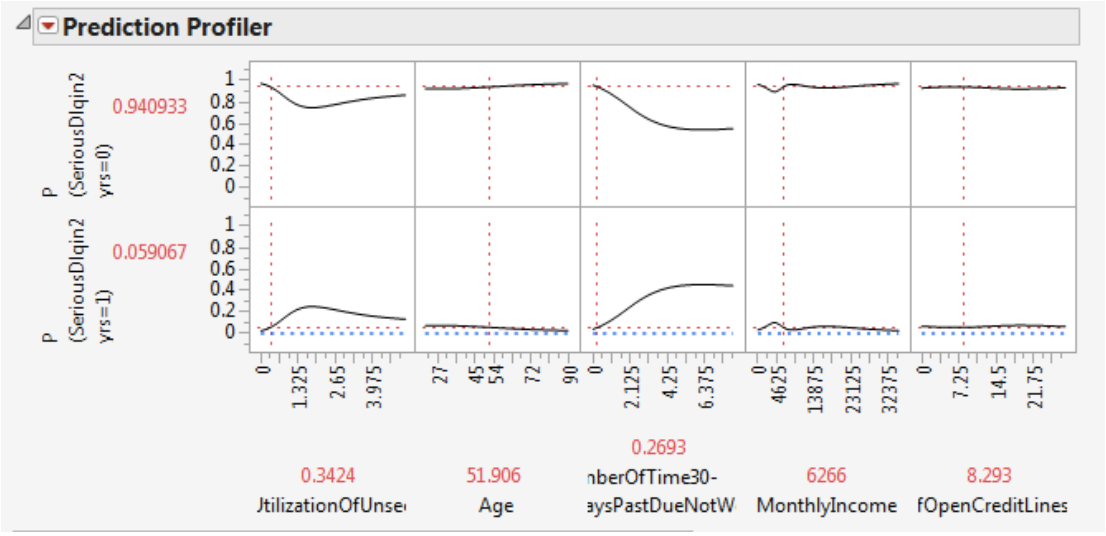
Cut Off	0.18		Actual/predict	0	1	
Profit	13179900		0	8119	845	8964
			1	387	649	1036
				8506	1494	10000





Second Best Model: Neural Network

We applied Neural Network to build a model to predict those people who are going to experience 90 days past due or worse in the next 2 years. Variables used to construct the model include Revolving Utilization Of Unsecured Lines, Age, Number Of Time 30-59 Days Past Due Not Worse, Monthly Income, and Number Of Open Credit Lines And Loans.



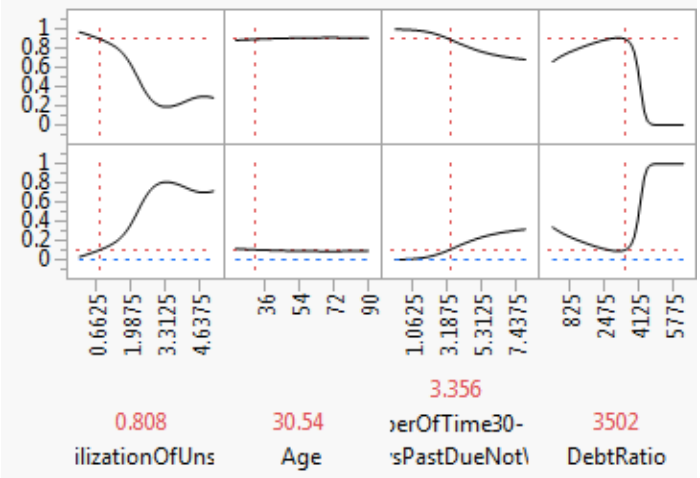
Those five variables above are chosen from Logistic Regression stepwise with the lowest p-value. Except for those five, we also applied three more variables: NumberOfTimes90DaysLate, Debt Ratio, and Number Of Dependents. The reason why we apply those three variables are as follows:

a) **NumberOfTimes90DaysLate**

Originally, we have chosen NumberOfTime30-59DaysPastDueNotWorse as one of the variables to build our model. We further select NumberOfTimes90DaysLate to include into our model because they are highly co-related. From the business insight, those who experienced past due before, even for a short period of time, may have chance to go through a longer period past due in the future. Moreover, those who had 90 days late payback before are having very high probability failing to pay back again.

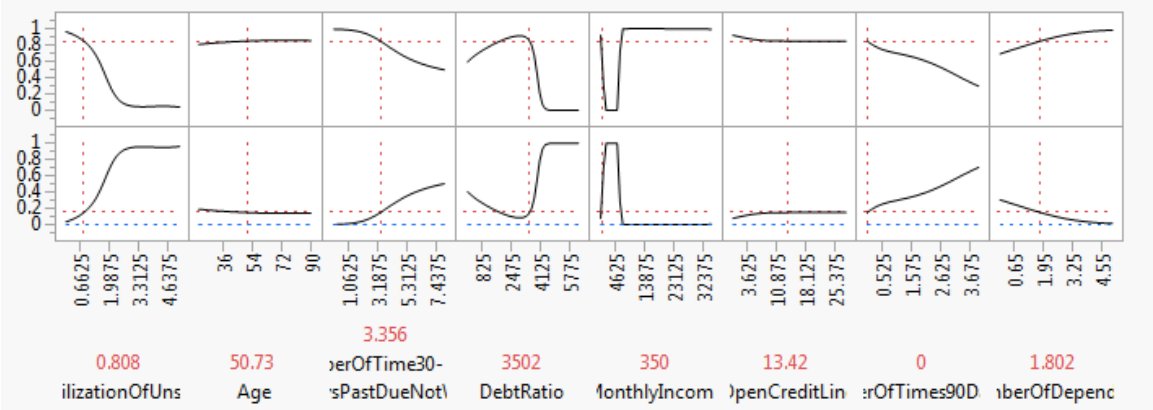
b) **Debt Ratio**

Since debt ratio is the payments, alimony and living cost divided by monthly income, we can assume that for those who have high debt ratio would have higher chance to experience seriously delinquency in the next 2 years. From the profiler, we can find that for those with age around 30, when their debt ratio has passed a certain level such as 3502 in the graph below, the possibility of 90 days past due turns very high.



c) <u>Number of Dependents</u>:

From our basic instinct, we might consider that those who have more dependents are having bigger chance to experience 90 days past due delinquency. However, it doesn't seem to be like that according to the profiler. The better business insight for explaining this is that people with more dependents are usually more responsible. They know how their capability is so they can raise so many children. It is possible simply because they depend on social welfares to raise their children. On the other hand, those who have less or no dependent are usually single. Since they don't have the pressure to raise a family, they might tend to loan money to enjoy their lives and turn out to have no way to pay the money back.



Training:

	0.16	actual/pre	0	1	
profit	13572800	0	7968	998	8966
		1	316	718	1034
			8284	1716	10000

Testing:

	0.16	actual/pre	0	1	
profit	13226400	0	7984	980	8964
		1	354	682	1036
			8338	1662	10000

Training

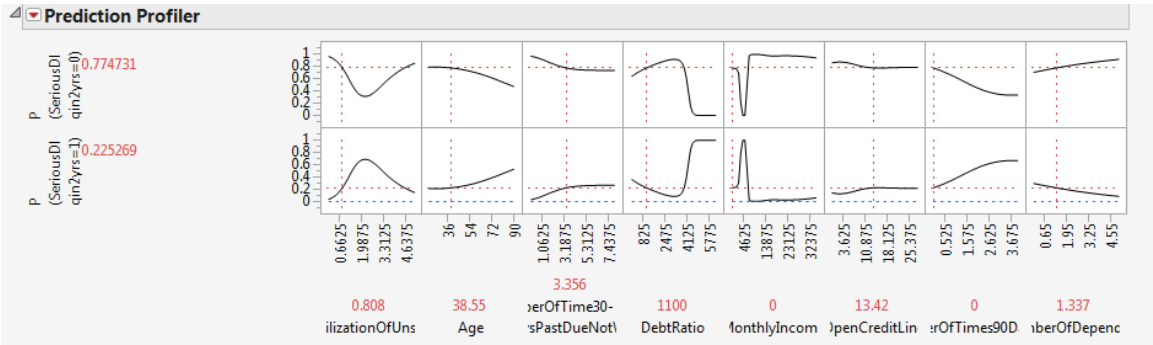
SeriousDlqin2yrs

Measures	Value
Generalized RSquare	0.4510289
Entropy RSquare	0.3718493
RMSE	0.2444472
Mean Abs Dev	0.1207636
Misclassification Rate	0.0785
-LogLikelihood	2088.5379
Sum Freq	10000

Validation

SeriousDlqin2yrs

Measures	Value
Generalized RSquare	0.4053429
Entropy RSquare	0.3296228
RMSE	0.2534032
Mean Abs Dev	0.1245426
Misclassification Rate	0.082
-LogLikelihood	2231.8313
Sum Freq	10000



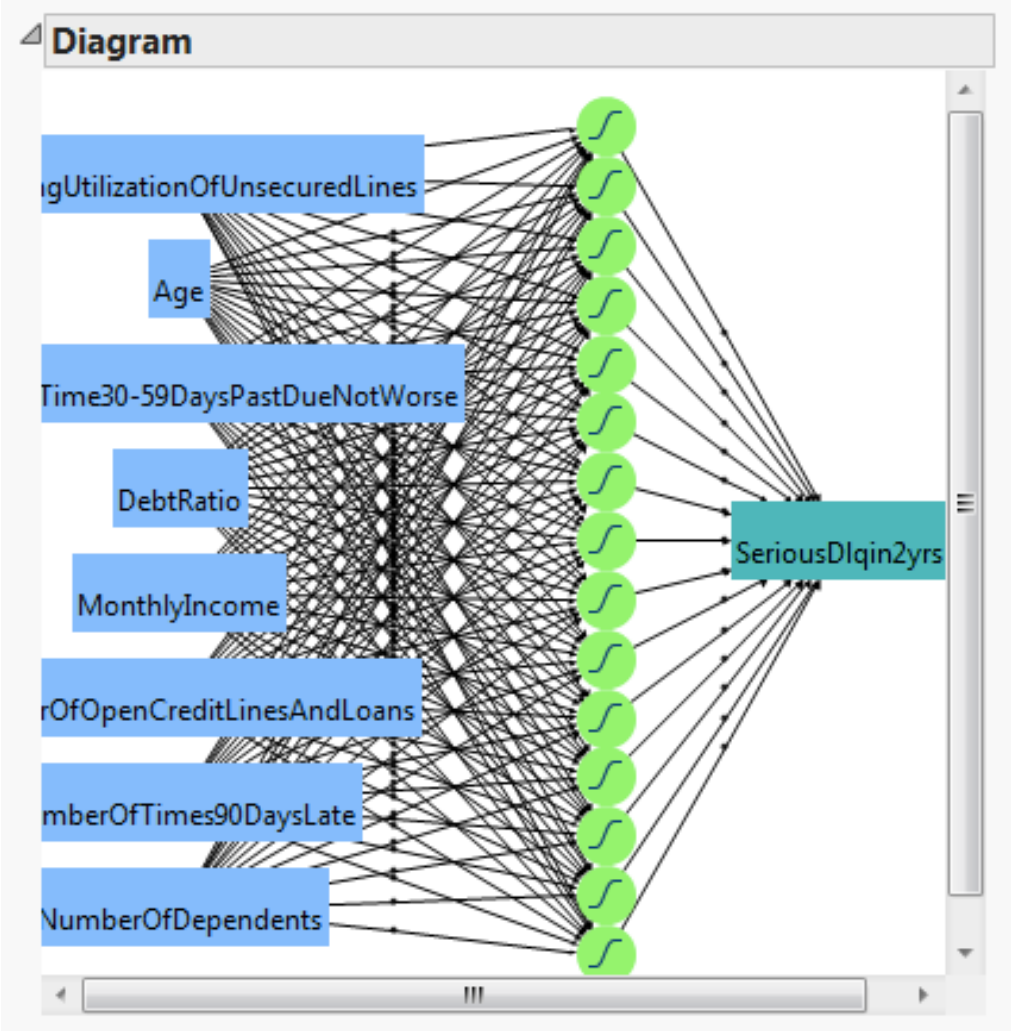
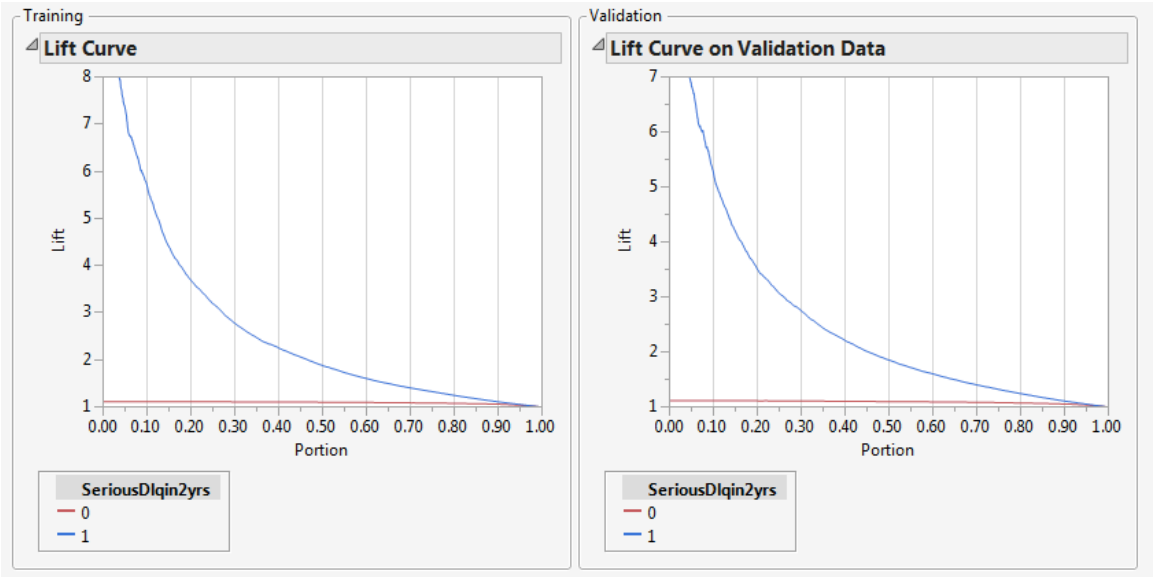


Diagram
gUtilizationOfUnsecuredLines
Age
Time30-59DaysPastDueNotWorse
DebtRatio
MonthlyIncome
rOfOpenCreditLinesAndLoans
mberOfTimes90DaysLate
NumberOfDependents
SeriousDlqin2yrs


Training
Lift Curve
Lift
Portion
SeriousDlqin2yrs
0
1
Validation
Lift Curve on Validation Data
Lift
Portion
SeriousDlqin2yrs
0
1