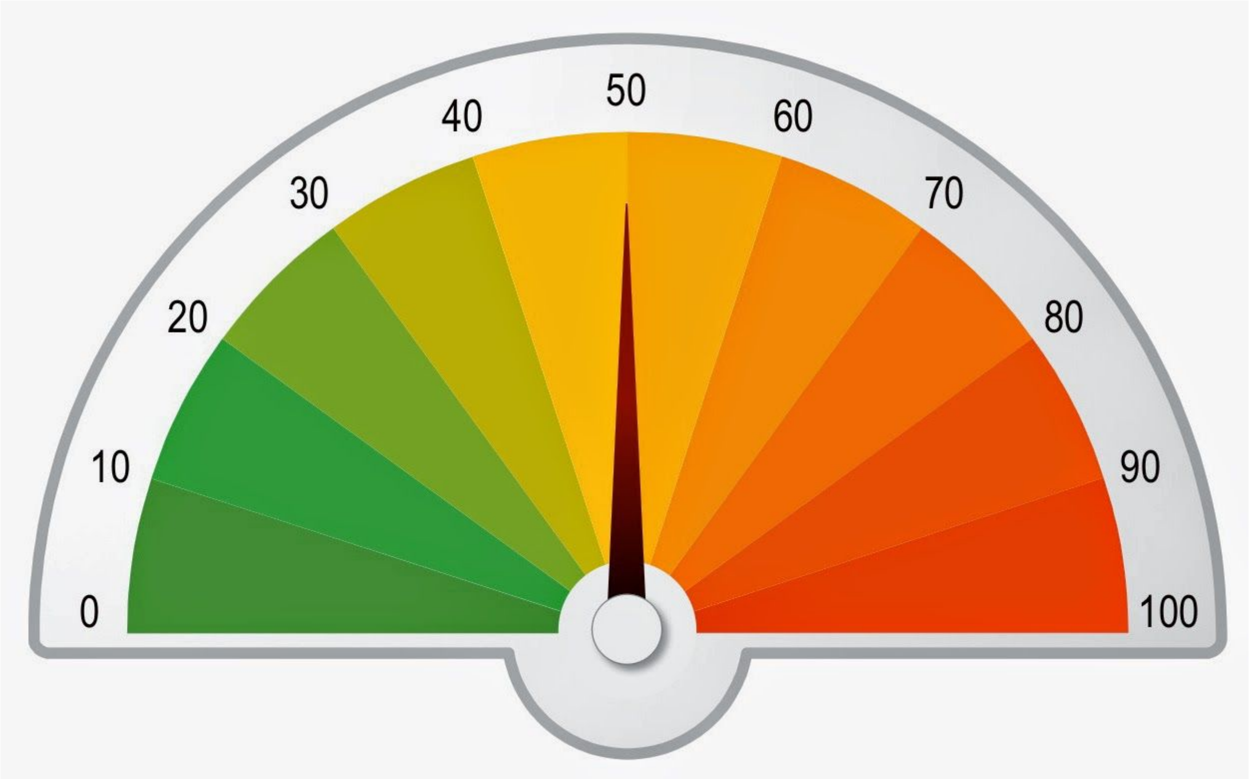
Study Report

Give Me Some Credit | Kaggle

Prabhat R.

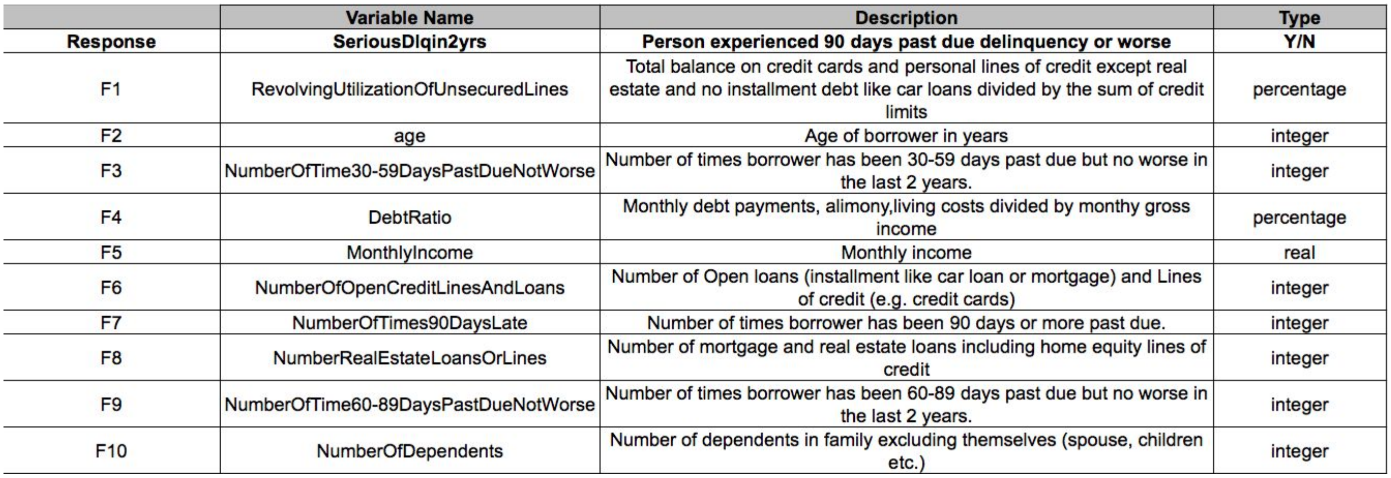


Introduction:

The availability of vast amounts of data and the advent of machine learning allows us to make accurate predictions and gain insights into customer behavior. This helps in mitigating the risk involved in lending money, cutting the cost in case of defaulting customers and potentially generating significant revenue for these financial institutions. For this report we will be using the data set available in the kaggle competition *Give Me Some Credit*, to predict the probability that a customer will experience financial distress in the next two years based information such as age, debt ratio, monthly income etc. This report presents results of prediction models like logistic regression, random forest and gradient boosted machines, XGBoost evaluated with AUROC metrics.

Data:

The data consists of information for the following variables for each customer, along with a variable indicating their status of experiencing financial distress in the next two years.



	Variable Name	Description	Type
Response	**SeriousDlqin2yrs**	**Person experienced 90 days past due delinquency or worse**	**Y/N**
F1	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
F2	age	Age of borrower in years	integer
F3	NumberOfTime30-59DaysPastDueNotWorse	Number of times borrower has been 30-59 days past due but no worse in the last 2 years.	integer
F4	DebtRatio	Monthly debt payments, alimony,living costs divided by monthy gross income	percentage
F5	MonthlyIncome	Monthly income	real
F6	NumberOfOpenCreditLinesAndLoans	Number of Open loans (installment like car loan or mortgage) and Lines of credit (e.g. credit cards)	integer
F7	NumberOfTimes90DaysLate	Number of times borrower has been 90 days or more past due.	integer
F8	NumberRealEstateLoansOrLines	Number of mortgage and real estate loans including home equity lines of credit	integer
F9	NumberOfTime60-89DaysPastDueNotWorse	Number of times borrower has been 60-89 days past due but no worse in the last 2 years.	integer
F10	NumberOfDependents	Number of dependents in family excluding themselves (spouse, children etc.)	integer

Data distribution and Imputations:

The following map shows the missing data in the given data set. The missing data is in each row is indicated by yellow horizontal lines.

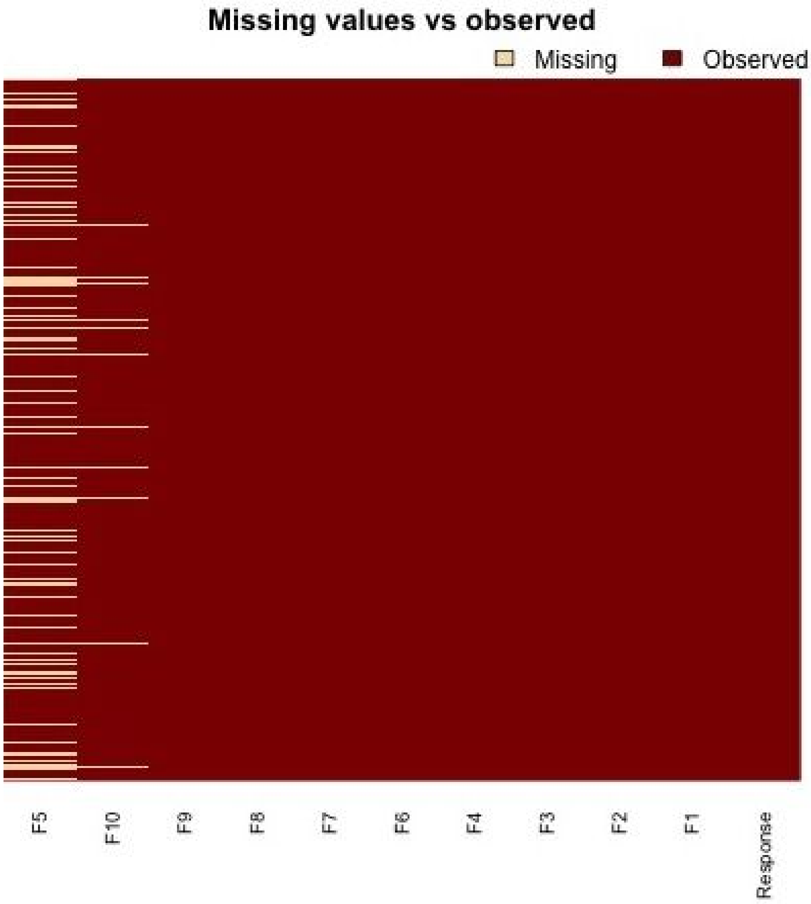


Fig.1. Missing Data Map

As can be seen from the missmap above, there are missing values in the predictor variables F5 and F10 i.e. Monthly Income and Number of Dependents. Number of records with missing monthly income are **29731** and the records with missing Number of Dependents fields are **3924**.

Now let's looks at the distribution among the predictor variables, identify any outliers or missing values and take appropriate actions.

1|age

The predictor variable has no missing values, and the distribution is fairly normal as can be seen in the histogram and QQ plot in fig.3. Below, which indicates sampling from a normally distributed population.

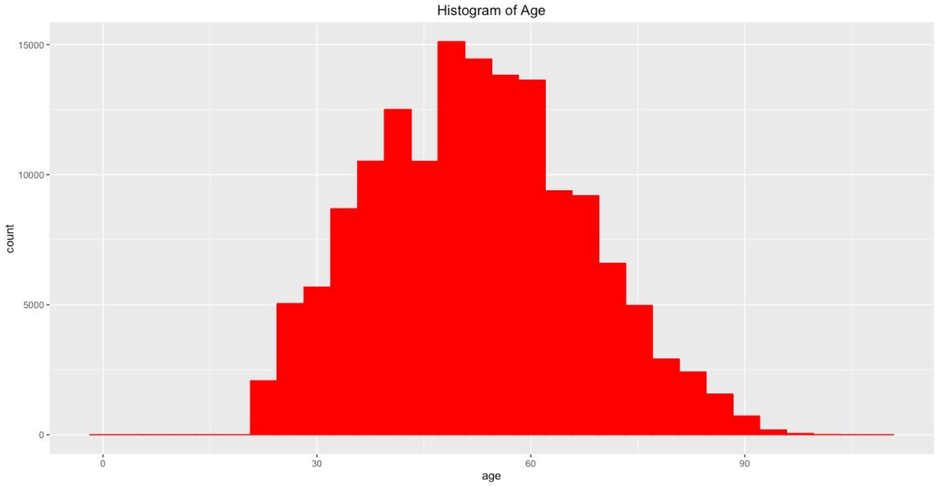


Fig.2.a. Hist of Age

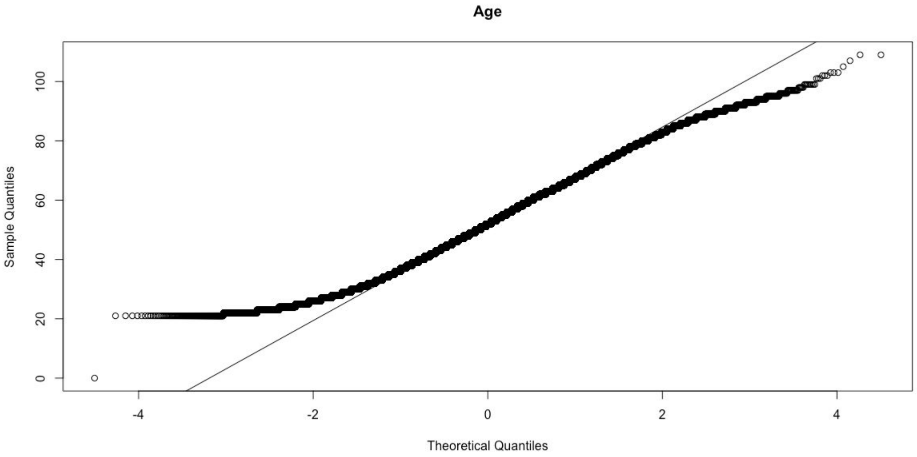


Fig.2.b.QQplot for age

2| Revolving Utilization Of Unsecured Lines

MIN	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	0.03	0.15	6.05	0.56	50710

The revolving utilization of unsecured lines is a ratio and should have a value between 0 and 1, the number of abnormal values (i.e. >1) in this predictor are 3321, and no missing values. The mean is skewed because of the high values so it makes sense to impute the median values for all the abnormal entries. The distribution after adjusting the abnormal values looks pretty reasonable. The histogram of the distribution can be seen below.

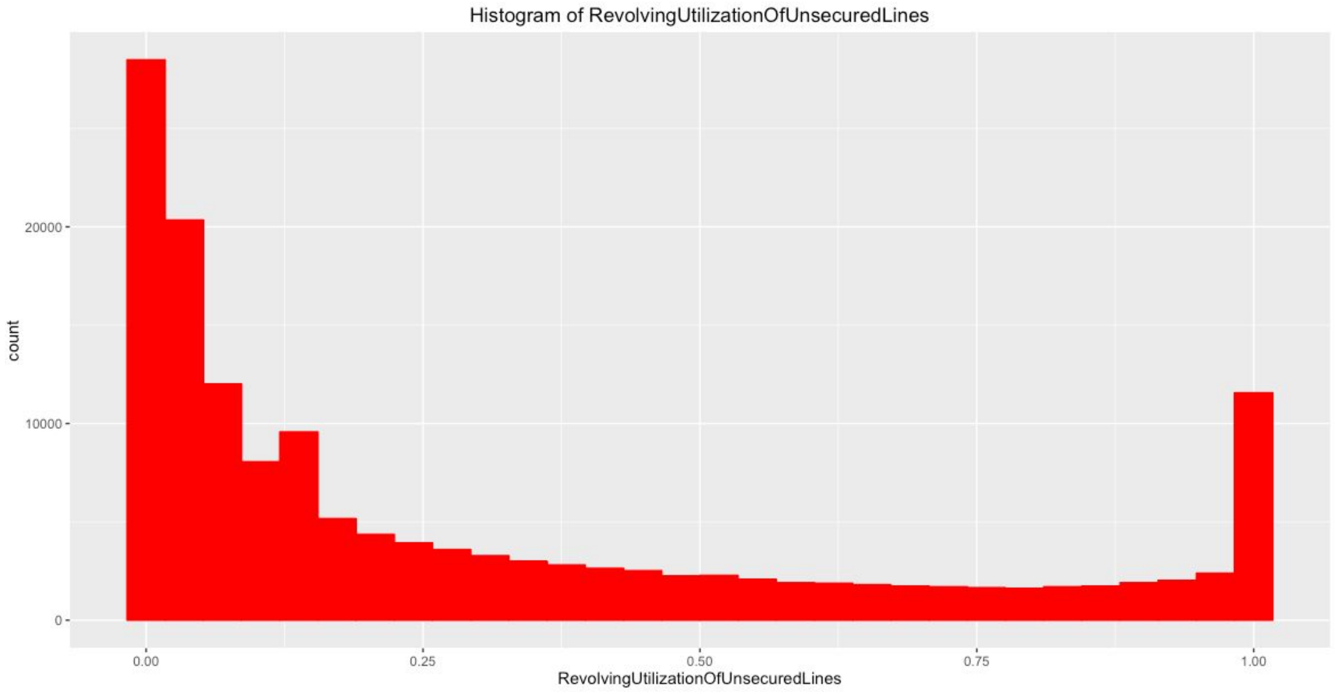


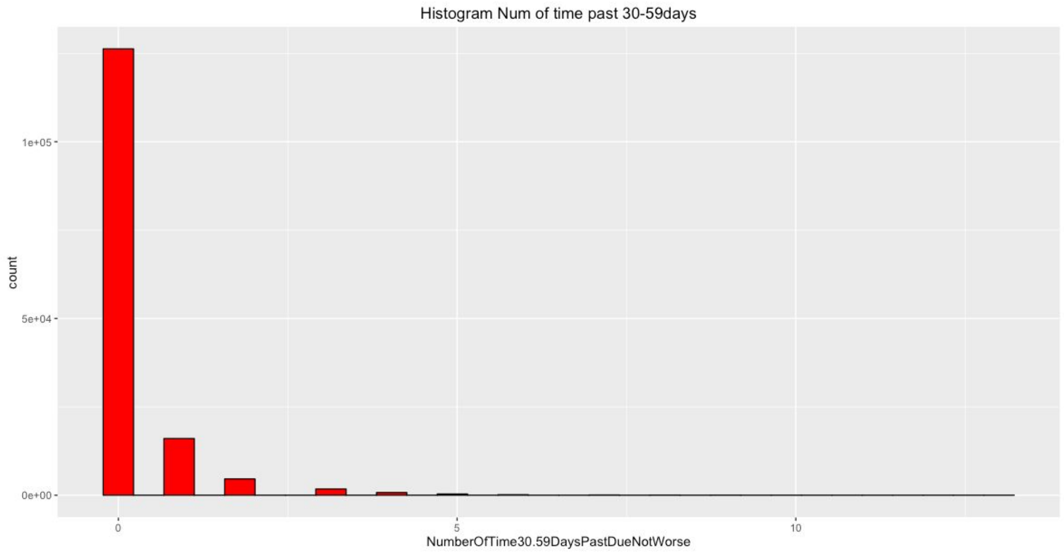
Fig.3. Histogram of RevolvingUtilizationOfUnsecuredLines

3| NumberOfTime30.59DaysPastDueNotWorse

The following table show the number of records per each count. It seems highly unlikely that anyone would have gone 96 or 98 times past 30-59 days mark. These must be missing values which were coded as 96 or 98. Lets just impute the records with the count 96 or 98 with 0.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	96	98
126018	**16033**	**4598**	**1754**	**747**	**342**	**140**	**54**	**25**	**12**	**4**	**1**	**2**	**1**	**5**	**264**

The distribution can be seen in the figure below, it looks like a decreasing power function.

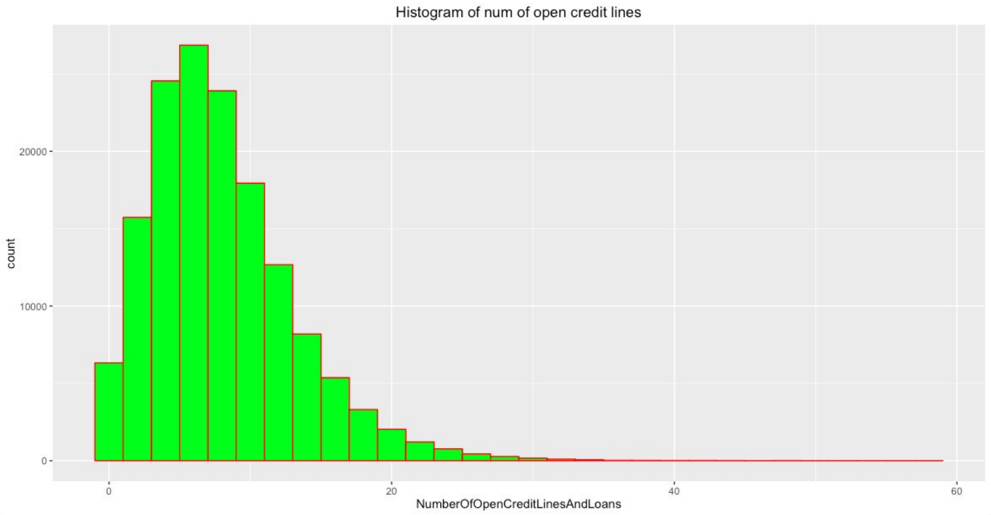


4| MonthlyIncome

There are 29,723 missing values in the monthly income predictor. Its sensible to impute them with the median of the monthly income, because the mean is skewed by people who earn on the higher side of the distribution. Additionally, the wealthy people specifically people with monthly income over $300,000 are dropped from the dataset.

5| NumberOfOpenCreditLinesAndLoans

There are no missing values in the predictor. The distribution shown below seems perfectly reasonable.



6| NumberOfTimes90DaysLate

This predictor has issues similar to NumberOfTime30.59DaysPastDueNotWorse, where the counts are 96 and 98, so we do the same imputations to correct these abnormal values by replacing them with 0.

7| NumberRealEstateLoansOrLines

The number of real estate loans or lines distribution seems reasonable, except one of the records has 54, this could pose as a leverage point later, so lets just drop the record.

8| NumberOfTime60.89DaysPastDueNotWorse

This predictor has issues similar to NumberOfTime30.59DaysPastDueNotWorse, where the counts are 96 and 98, so we do the same imputations to correct these abnormal values by replacing them with 0.

9| NumberOfDependents

There are a few (3922) missing values in this predictor, we impute the obvious choice that is 0.

10| DebtRatio

It seems weird that anyone can have a debt ratio of greater than 100,000, so lets drop those records.

Data Imbalance:

The 0 class constitutes **93.31%** of the response variable SeriousDlqin2yrs, while the class 1 constitutes **6.685%.** This is a clear case of data imbalance. To address this imbalance the class 0 has been downsampled and a new train and test data have been constructed where both the classes are of equal or almost equal proportions.

The distribution of each predictor across the two classes in the response variable SeriousDlqin2yrs is shown in the figure below.

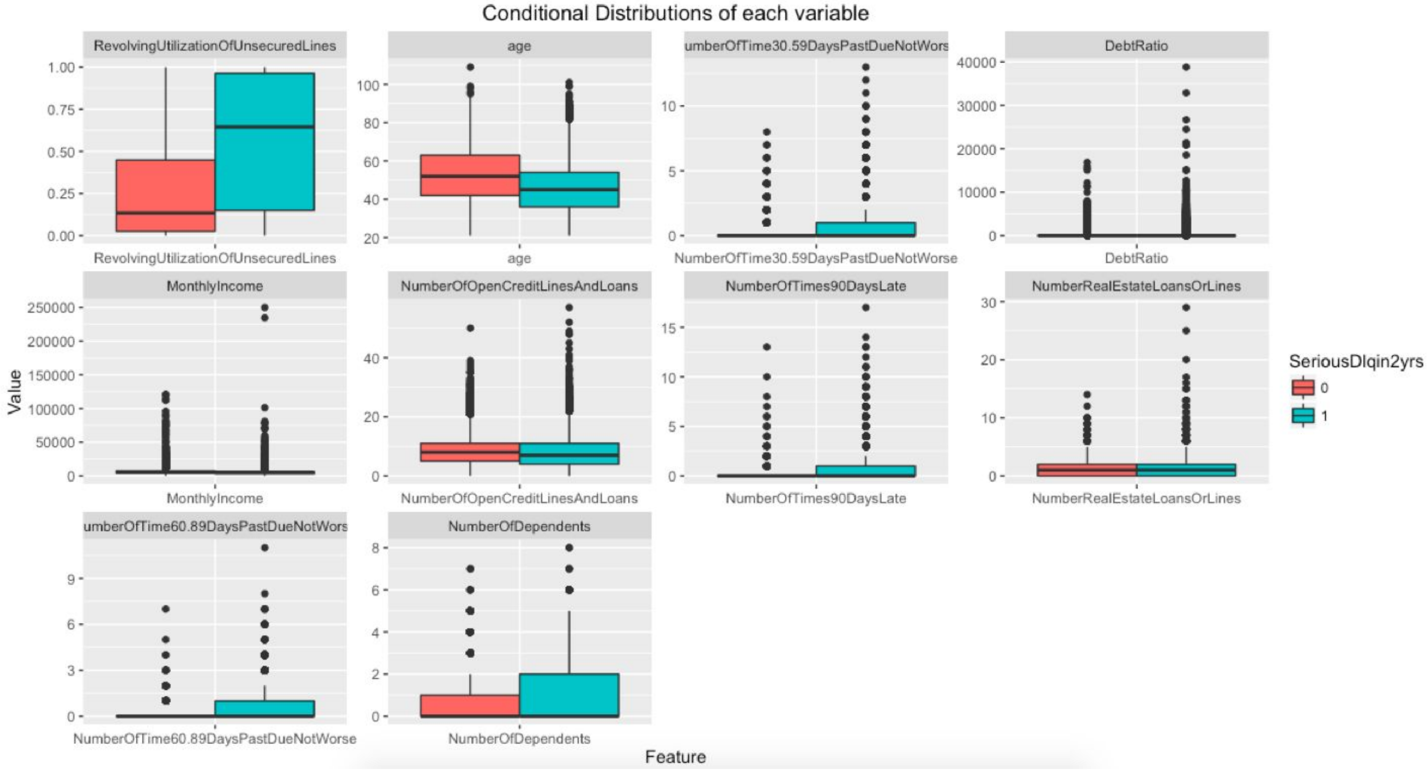


Fig.6. Feature distribution across the classes in response variable

The figure 6 suggest that the predictor **Revolving Utilization Of Unsecured Lines** is quite discriminatory in predicting the response variable with the distributions quite well separated across the two classes, **age** too seems quite powerful in predicting the response class.

Just out of curiosity the data was plotted in the first two principal components, note: the variance explained by the first two principal components is only 36%. Its just something to see how separable the data is. Well, its not that separable. The data was plotted in the first three principal components too, first three components explain 48.42% of the variance, still not really that separable. This can be seen in Fig.8.

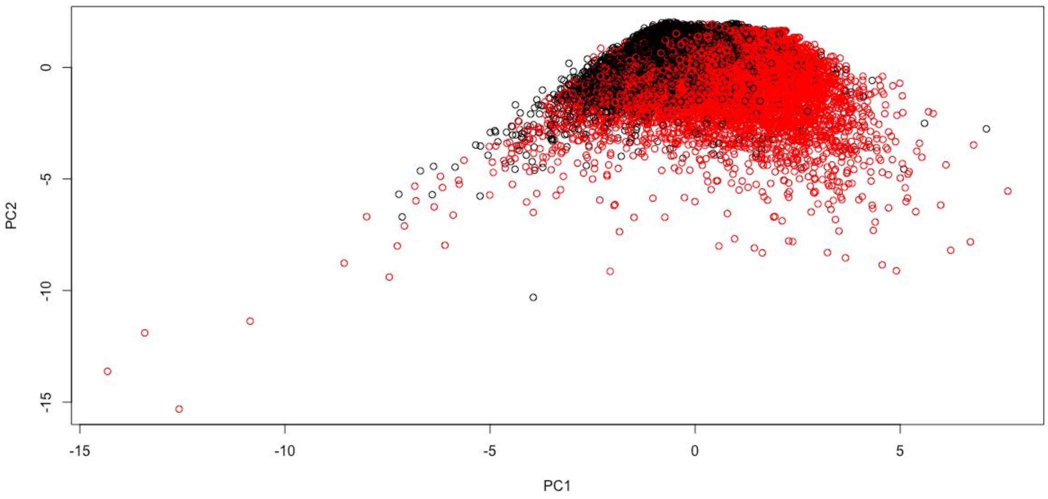


Fig.7. The data is plotted in the first two components, the red points belong to class 0 and the black points belong to the class 1.

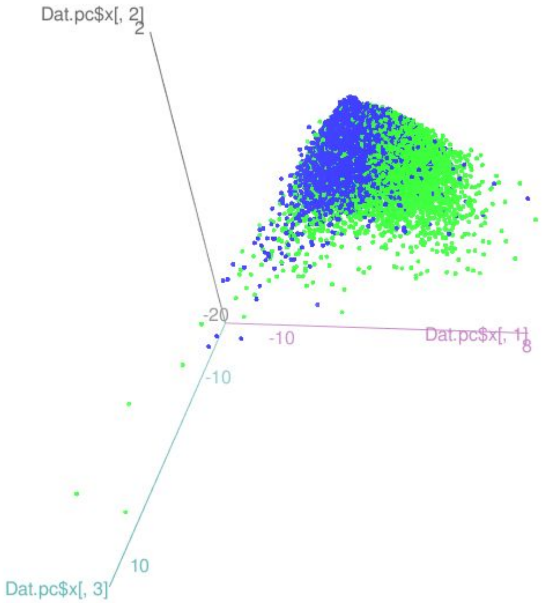
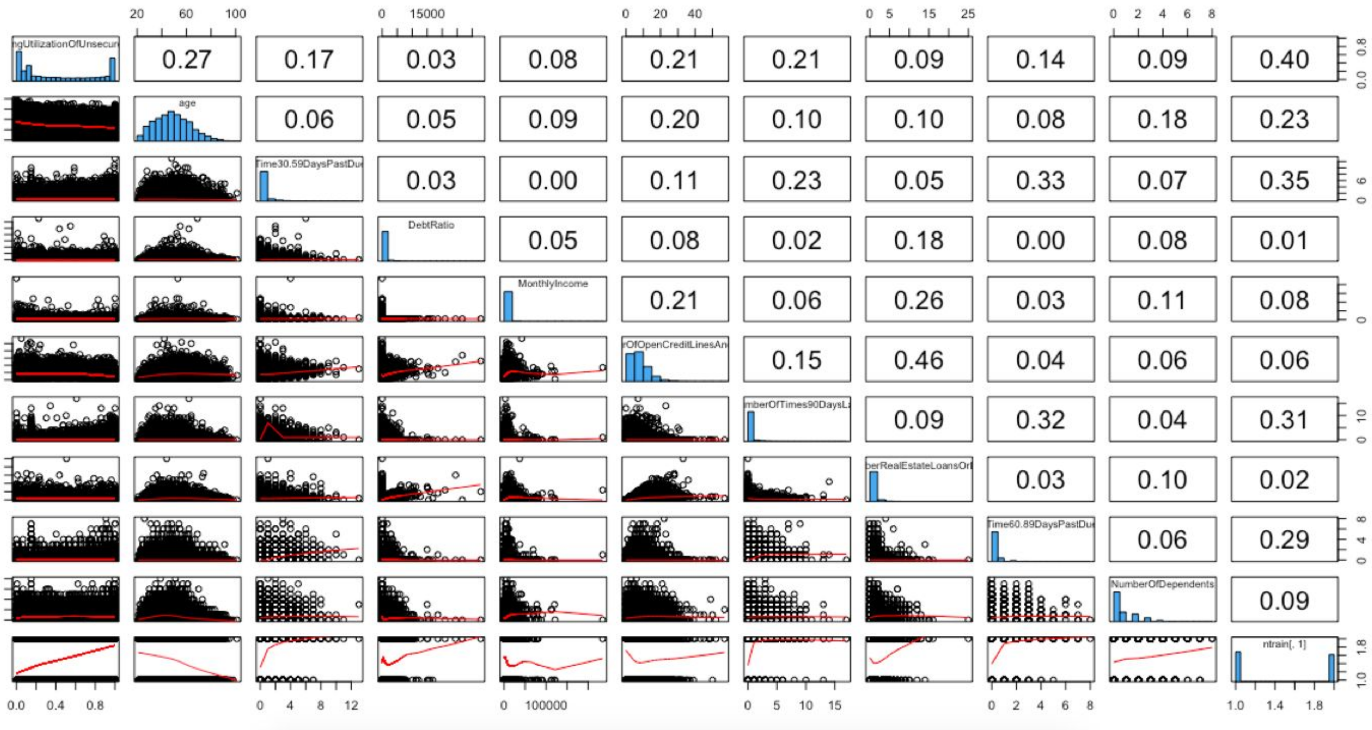


Fig.8. The data is plotted in the first three principal components, the color signifies the class the record belongs to.

The following figure shows the scatterplot matrix for all the predictors and the response.



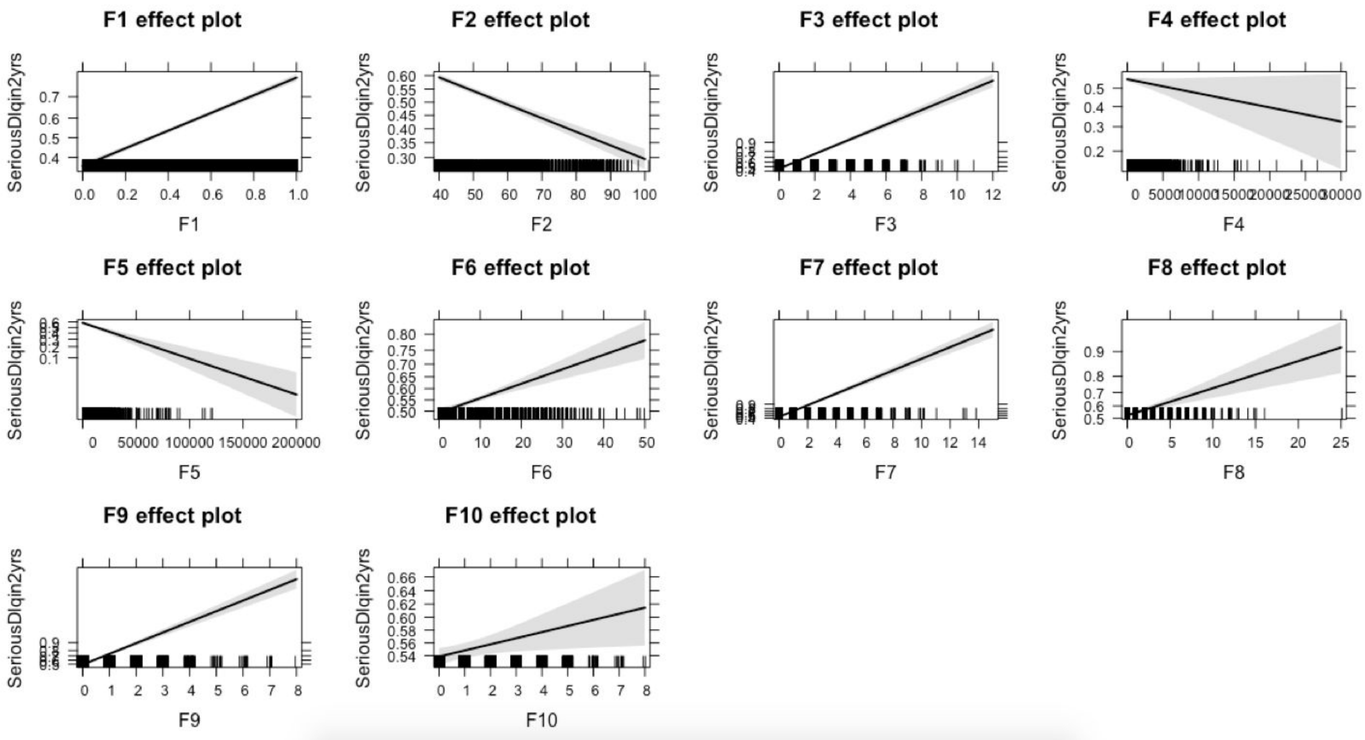
The scatterplot matrix shows that **Revolving Utilization Of Unsecured Lines** , **age, NumberOfTime30.59DaysPastDueNotWorse** and **NumberOfTime60.89DaysPastDueNotWorse** have high correlation with the response variable. The predictors of type number of days past due are correlated with each other indicating collinearity/interaction.

Prediction Models:

The following models were build on the Give Me Some Credit Data.

Logistic Regression:

A main effects logistic regression model was fit first. Each effects plots below in fig.10. shows the effect of each predictor when the other main effects are held constant. This allows us to get an idea of how each predictor affects the predictions.



<u>**Fig.10.** Effects of each predictor are shown in this fig. The probability of SeriousDlqin2yrs is shown on the y-axis while the value of each predictor is plotted in the x-axis.</u>

Next we plot the complete main effects + two way interaction model, and use the anova table with chisq test to find that the complete model is better than just the main effects model with a p-value of $2.2 * 10^{-16}$. So, a choice was made to use the stepwise selection using the AIC measure, to identify the best model. The stepwise model selected is shown below:

SeriousDlqin2yrs ~F1 + F2 + F3 + F4 + F5 + F6 + F7 + F8 + F9 + F10 + F1:F2 + F1:F3 + F1:F6 + F1:F7 + F1:F8 + F1:F9 + F2:F4 + F2:F5 + F2:F6 + F2:F7 + F2:F9 + F2:F10 + F3:F6 + F3:F7 + F3:F8 + F3:F9 + F4:F6 + F4:F8 + F4:F10 + F5:F7 + F5:F8 + F5:F9 + F6:F8 + F6:F10 + F7:F9 + F7:F10 + F9:F10

The chisq test also seconds the significance of the step wise model over the complete model.

Next interactions between different main effects were plotted for the purpose of analysis and get a better understanding of the data. Some of these interaction plots can be seen below:

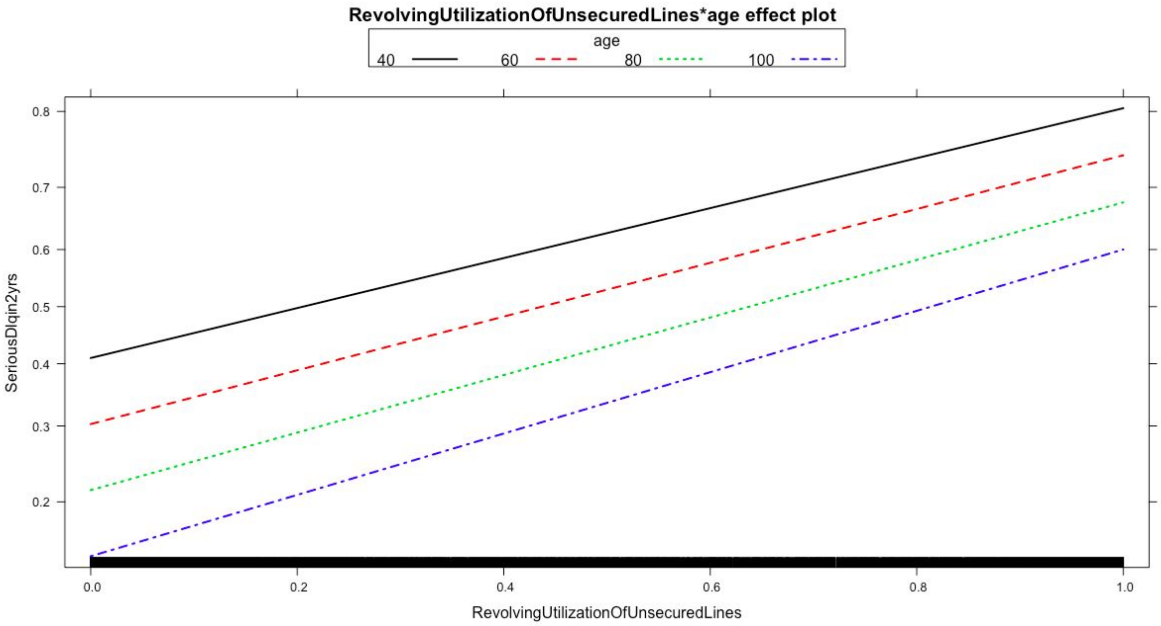


Fig.11. The above figure shows RevolvingUtilization vs probability of SeriousDlqin2yrs, each line represents a different value for age, while all the other predictors are held constant, the lines are quite parallel indicating no significant interactions.

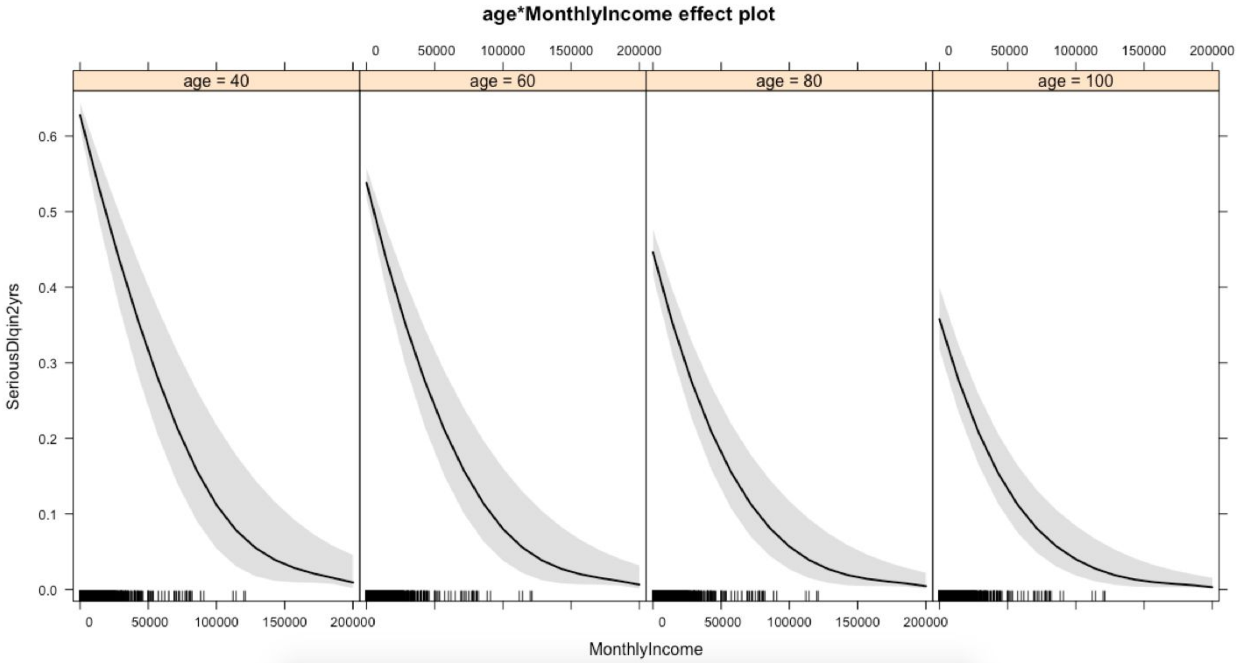


Fig.12. The probability of seriousdlqin2yrs seems to decrease for the same monthly income with increasing age, an expected observation.

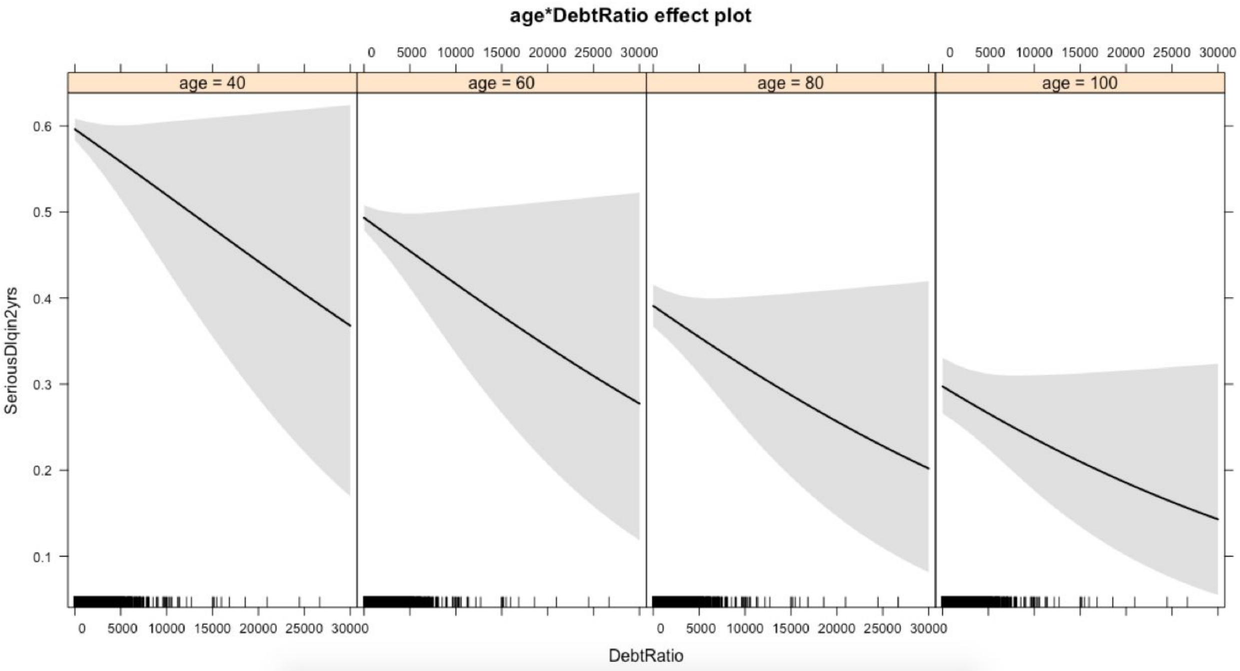


Fig.13. The probability of DeriousDlqin2yrs vs debtratio for various values of age can be seen in the above plots. Seems that older people are less likely to be experience serious delinquency for the same monthly income.

Random Forest:

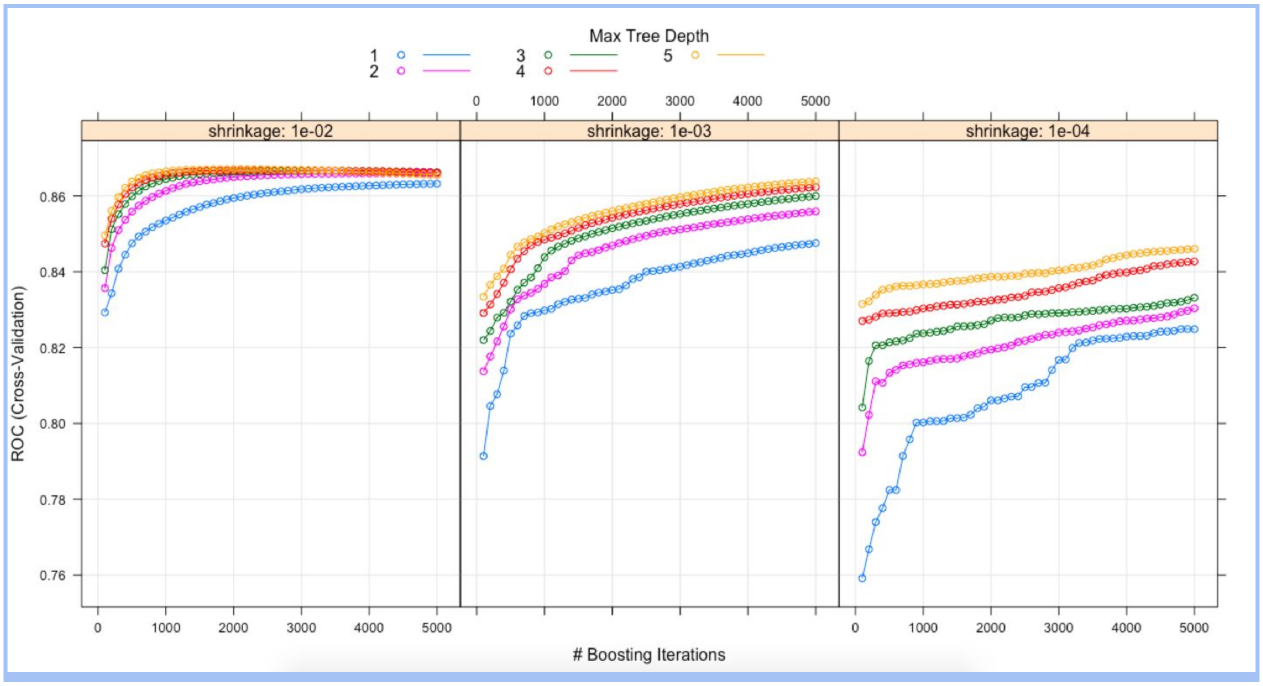
Next random forest was used to build an ensemble of trees and make prediction. The only parameters to tune here are the number of the trees and the number of predictors considered at each node. After some trial and error method the model that performed best had the following parameters:

Number of trees: 5000 ***Number of predictors considered at each node: 2***

The evaluations on the test data and performance on kaggle are shown in the evaluation section.

Gradient Boosted Machines:

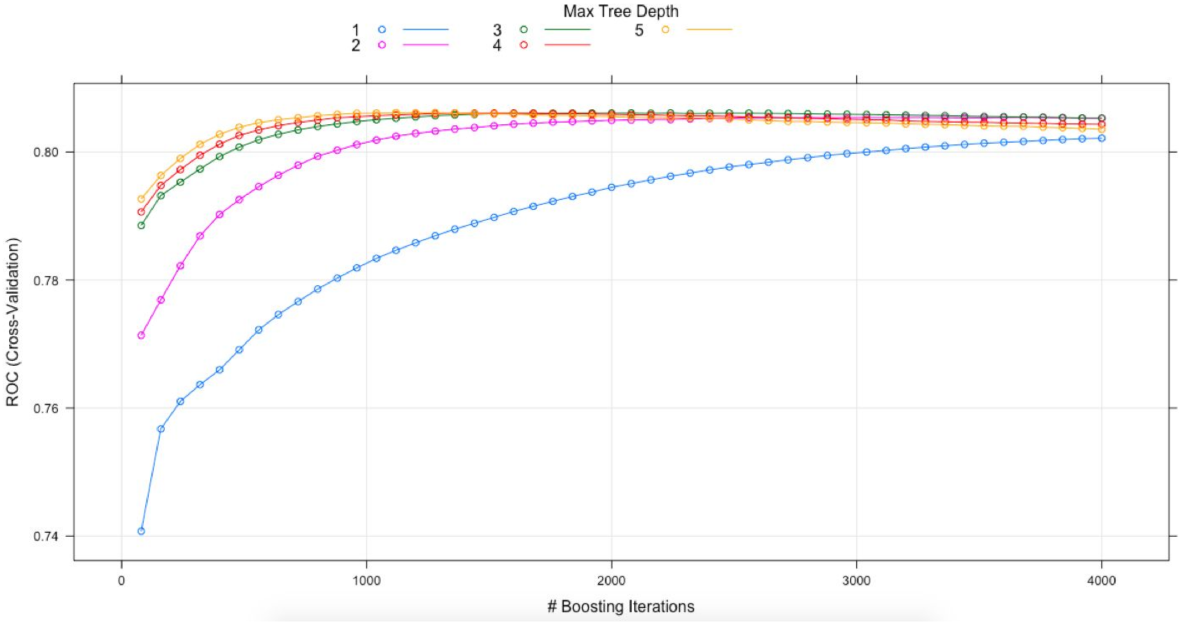
Gradient boosted models were built on the training data, using the gbm package, a grid search was setup to identify the parameters of the gbm model, namely interaction depth, number of trees, and the shrinkage factor. The metric that was evaluated in the grid search was the AUROC curve, the graph of the metric for different values of each parameter are shown in the graphs below:



The best parameters as indicated by the above graphs and the grid search are:

Num. of trees	***Interaction depth***	***shrinkage***	***num. Of min. obs in terminal leaves***
2100	**5**	**0.01**	**10**

Next a gradient boosted model with a relationship constraint b/n the response variable and a few predictors was constructed. The relationships defined were, the probability of serious delinquency in the next two years has a monotonically increasing relationship with the debt ratio, number of times 30-59 days past due not worse, number of times 60-89 days not worse and number of times 90 days or worse. The same grid search as above is applied to search for the best parameters for gbm with defined monotonic relationships. The graph of auroc for different parameters of this case is shown below:

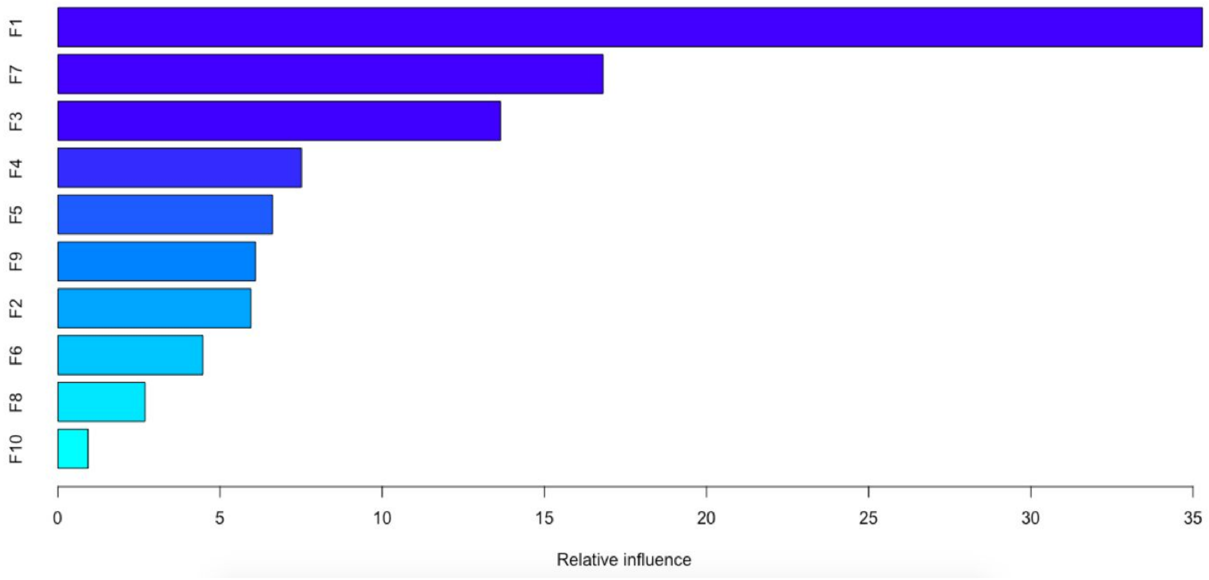


The best parameters picked by the grid are:

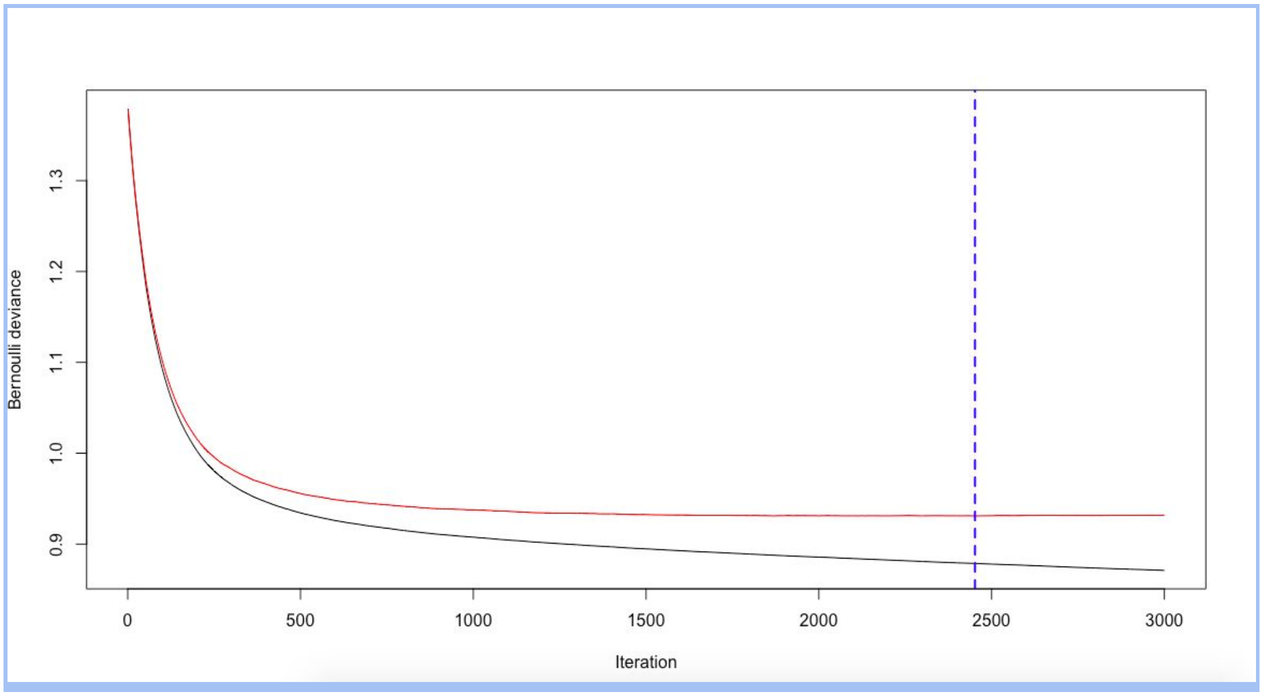
Num. of trees	*Interaction depth*	*shrinkage*	*num. Of min. obs in terminal leaves*
2080	**3**	**0.01**	**10**

The variable importance plot is shown below:

This is in keeping with our inference from above that Revolving Utilization is very discriminative of the response variable.



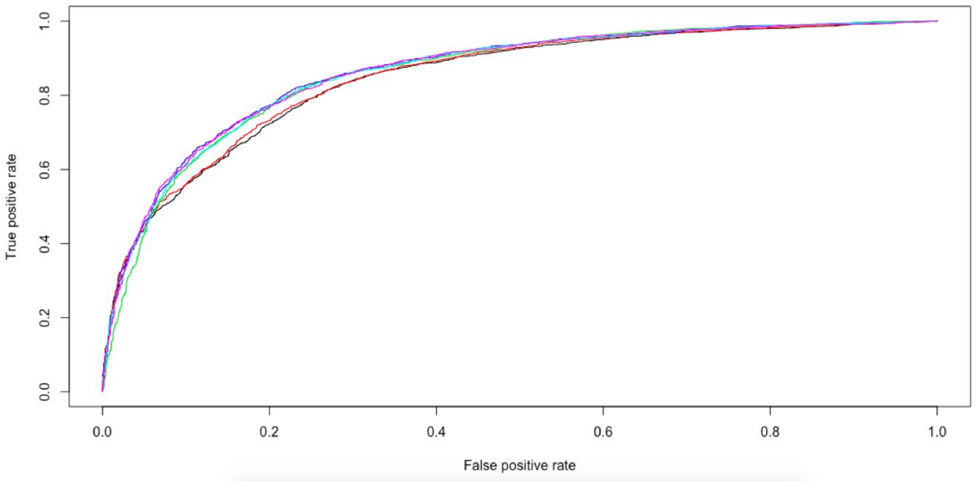
The training, and cross validation log loss for the final gbm model with the tuned parameters are shown below:



Evaluation:

Model	*AUROC on test*	*AUROC Kaggle*
Logistic Regression main effects	**0.8458615**	**0.855397**
Logistic Regression step wise	**0.8465**	**0.837890**
Random Forest	**0.8526727**	**0.859272**
GBM	**0.8583818**	**0.862852**
GBM with defined variable relationship	**0.8566018**	**0.861493**
XGBoost	**0.8571582**	**0.863536**

Below are the ROC curves for the above models:



Black-*Logistic regression Main effects* **Red- *Logistic Regression step wise***

Green- *Random Forest* **Blue-*GBM***

Cyan-*GBM with defined variable relationship* **Purple-*XGBoost***