# LOAN DEFAULT PROBLEM

PROJECT BY

SUMITH GANNARAPU

Function Name in R code: userfunction which produces evaluations and metrics for a binary classifier

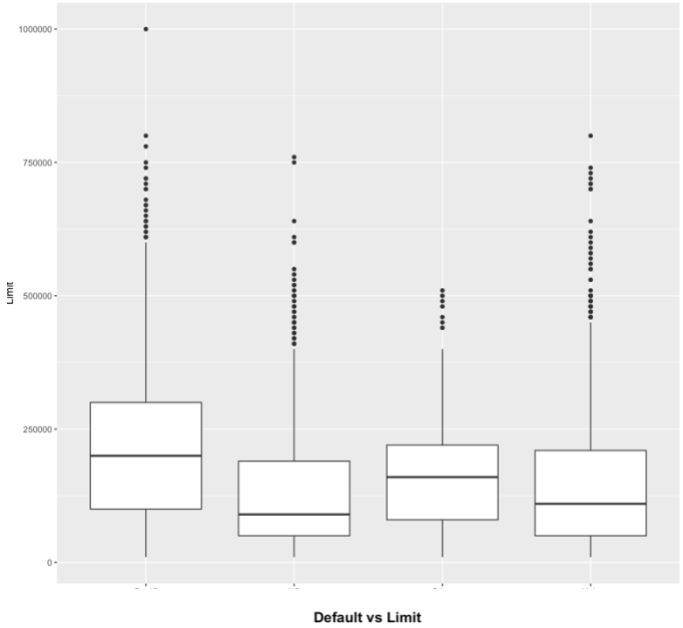
Input Parameters:

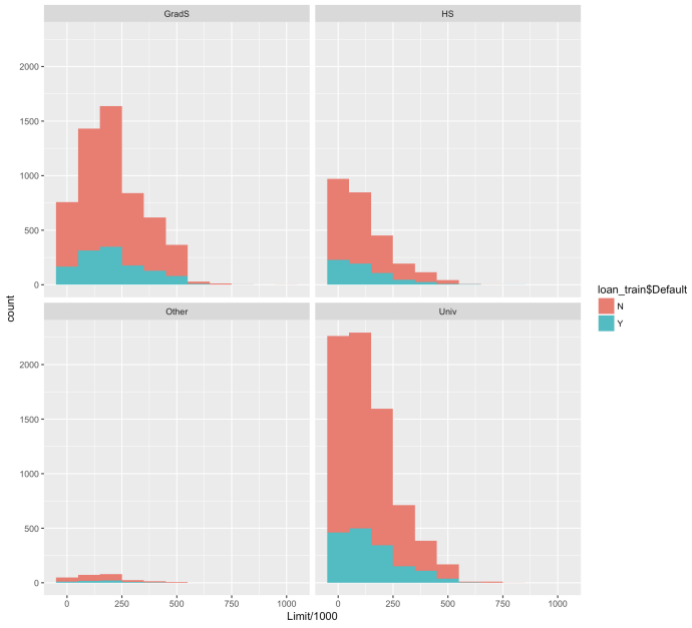
- True Values (1's and 0's) where 1 indicates the customer has defaulted the payment and 0 indicates otherwise.
- Predicted probabilities of the default status of payment for corresponding month.

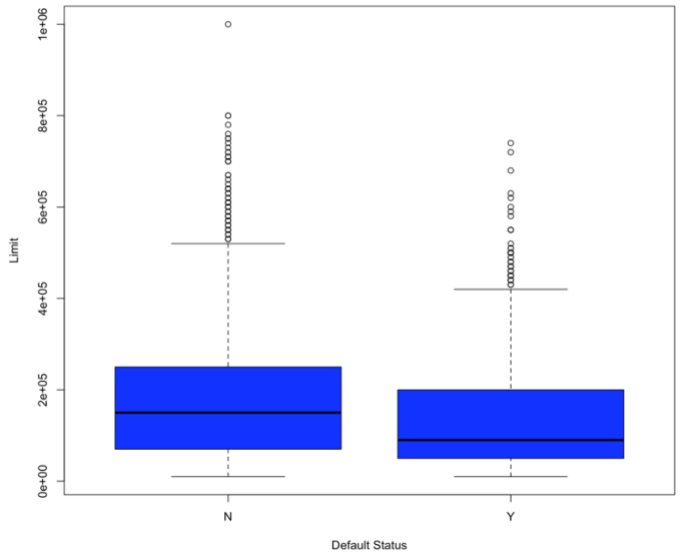
Output Values:

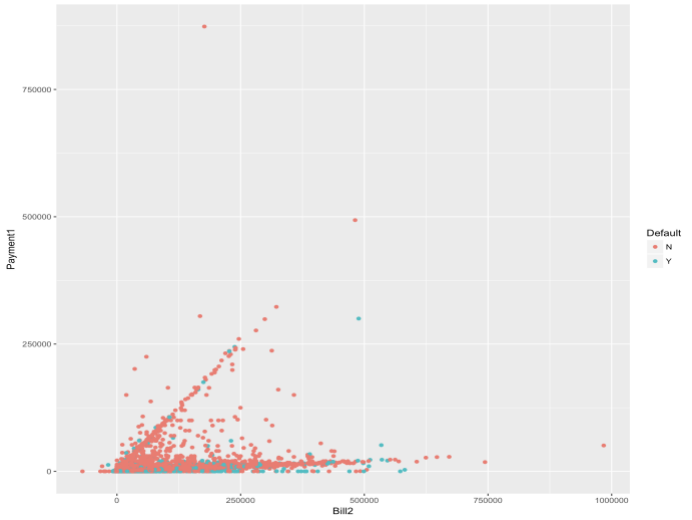
1. Confusion Matrix & Statistics
    - confusion matrix of TP, FP, TN, and FN classes
    - Accuracy, Kappa, Sensitivity, Specificity
2. K-S Chart & Statistics
    - Measures the degree of separation
    - Between the positive (y=1) and negative (y=0) distributions
3. ROC (Receiver Operating Characteristic) Curve and AUC (Area under the curve)
    - ROC Curves is used to evaluate the tradeoff between true-positive and false-positive rates of classification algorithms.
    - AUC Measure for evaluating the performance of a classifier and it's the area under the ROC Curve.
4. Distribution of predicted probabilities values for true positives and true negatives
    - Plot of predicted output Vs Predicted probability of true positive & true negative observations.
5. Concordant Pairs
    - It gives percentage of concordant and discordant pairs for a given model.
    - proportion of pairs for which scores are tied.
    - total possible combinations of 'Good-Bad' pairs
6. D statistic
    - Difference between the mean predicted value for the positive cases and mean predictive value for the negative cases
7. Cumulative gains chart
    - gain chart shows Predicted Positive Rate vs True Positive Rate
8. Log Loss
    - Log Loss quantifies the accuracy of a classifier by penalizing false classifications.

## Visualization:



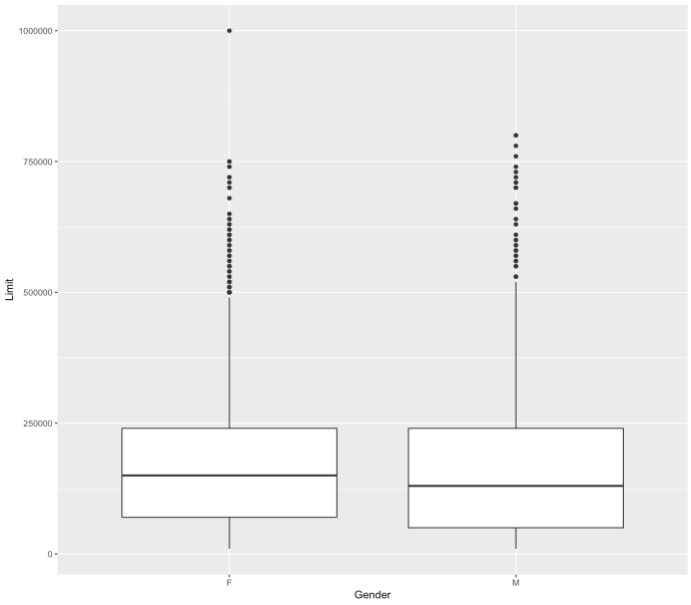



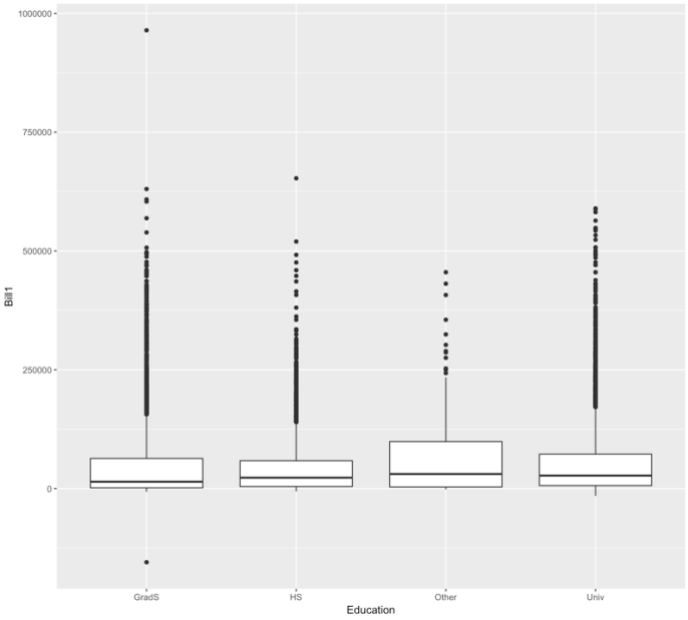




- (Plot 1) From the box plot we can clearly say that Graduate degree and University Degree students have more credit limit than High School and other students.
- (Plot 2) Balance limits and count of defaulted clients are almost same for University and Graduate Level. Additionally, the ratio of defaulted clients at high school level seems almost the same as the university and graduate levels.
- (plot 3) People with higher limit tend to default less than those with low limit. which can be clearly understood because banks tend to increase limit for those who make timely payments.
- (Plot 4) Bill Generated in this month needs to be paid in next month. according to the plot, the ratio of the payers are more than the Defaulters

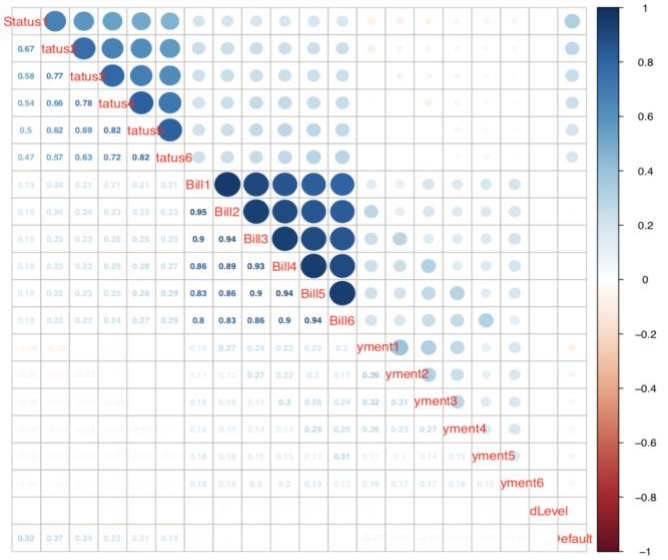
**Outlier Analysis:**

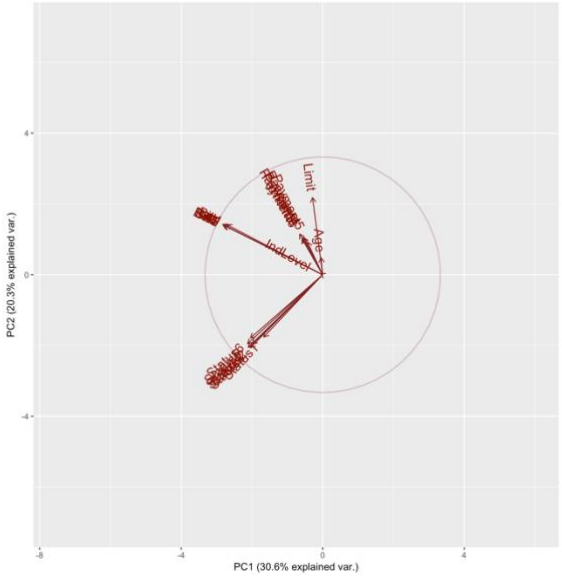
- From Plot 1, Male Students and Female Students seems to have same number of outliers but one of the Female student have one extreme outlier.
- From Plot2, Usage of Graduate Students and University students credit card bill for the last month is more whereas usage of High School and Other are less and also more outliers are observed in University Students and Grad Students.




## Correlation & Principal Component Analysis:




Correlation Plot: This shows that Bill Amounts have high correlation with each other and Status variables also have high correlation.

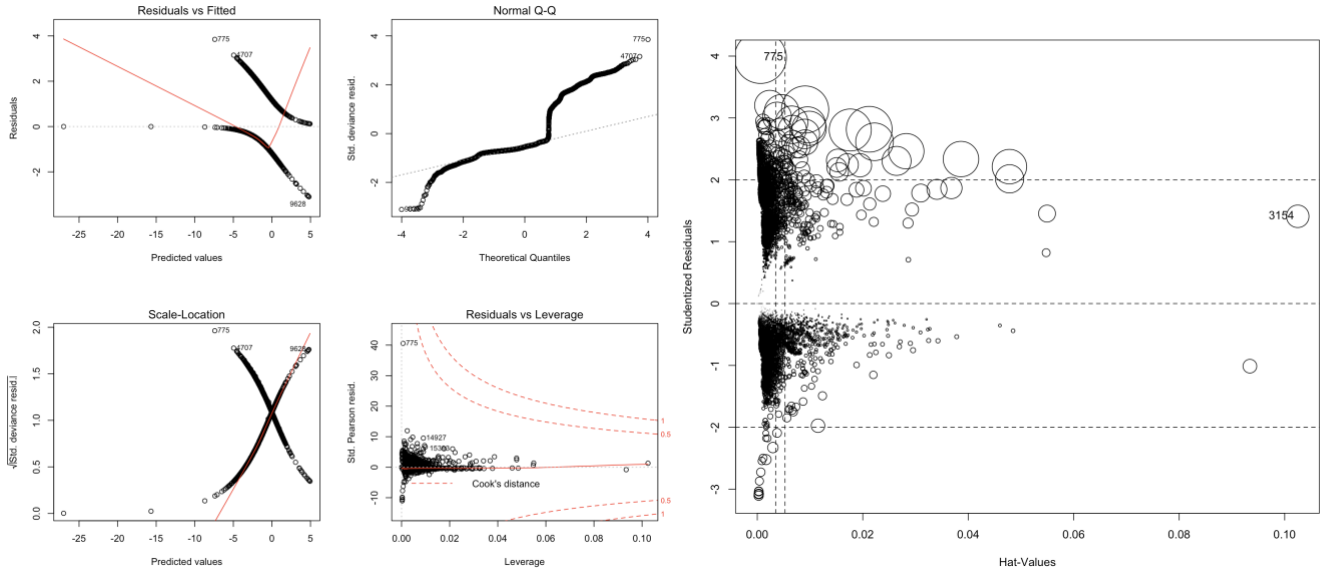
PCA Plot: First two components (PC1 and PC2) itself explains 51% of the variance in the whole matrix. Status 1 to 6, Age, Limit more towards PC2, Bill 1 to 6 more towards PC1and Payment 1 to 6 participating in PC1 & PC2 components.

**Basic GLM Model:**

Three coefficients of Basic GLM Model:

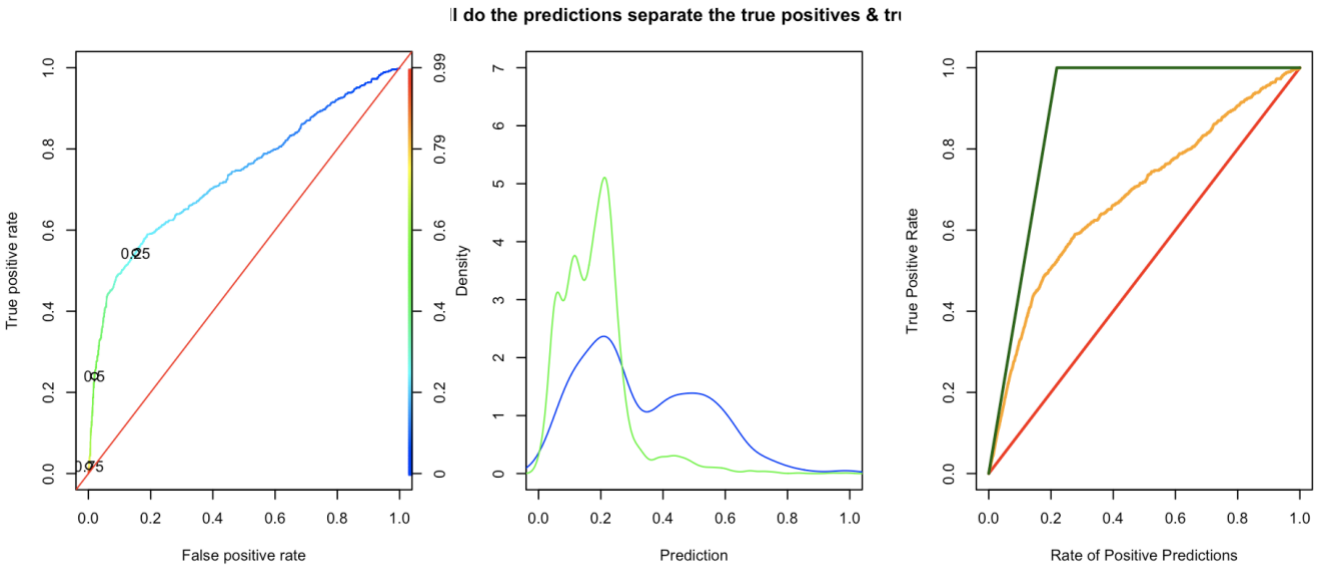
- The coefficient of Limit is significant and negative(-7.251e-07). Limit value indicates that the expected decrease in the log odds of being defaulted for a unit increase in the Limit value holding all other predictors constant is e^-7.251e-07
- The coefficient of Gender (Male = 1) is least significant and positive. GenderM value indicates that the odds of being defaulted for male (Male=1) over the odds of being defaulted for females (Male = 0) is e^8.610e-02
- The coefficient of Age is very less significant and positive. Age value indicates that the expected increase in the log odds of being defaulted for a unit increase in the Age value holding all other predictors constant is e^4.332e-03

#NULL DEVIANCE is deviance for the empty model (16813 on 15999 degrees of freedom)
#RESIUDAL DEVIANCE is a deviance based on actual basic model (14777 on 15972 degrees of freedom)




- The Residuals vs Fitted plot hard to interpret for logistic regression. In particular, the way the residuals fall on two distinct curves is an expected pattern due entirely to the fact that the response values are either zero or one, and the fitted probabilities are a non-linear but monotone function of the linear fit. But again the smooth of the residuals is essentially flat. The Normal Q-Q plot of the residuals are not normally distributed. The Scale-Location indicates that the model is heteroscedastic in nature. The Residuals vs Leverage having few outliers.
- From Influence Plot, we can see that 775, 3154 are the outliers and size of the data points related to the cook's distance.
- Variance Inflation factor is a standard error which is very high for Bill1 to Bill6 for the basic model.

**Basic GLM Model Statistics:**




| | |
|---|---|
| Accuracy | 0.8184 |
| Kappa | 0.2932 |
| Sensitivity | 0.24 |
| Specificity | 0.98 |
| AUC | 0.73467 |
| Log Loss | 0.4555 |

- From the Confusion Matrix, the accuracy of the basic model is 81% for the validation data.
- ROC curve between TP and FP and Area under curve
- The FP rate should be less and TP should be high for the ROC Curve which results in pushing curve towards the top left corner. Our results explained we have low TP rate and as a result AUC is less which is 0.73 indicates poor performance of the classifier.
- Green line in the Cumulative gain chart is ideal classifier and orange line in the gain chart is our basic logistic regression model classifier. As the percentage of sample increases, the orange line moves towards random curve instead of moving towards ideal model curve which indicates this model doesn't give good predictions.
- Log loss for the validation data is also high for this model which indicates less accuracy.
- From all above statistics, we can conclude that logistic regression model could not give good classification.

## METRIC COMPARISON

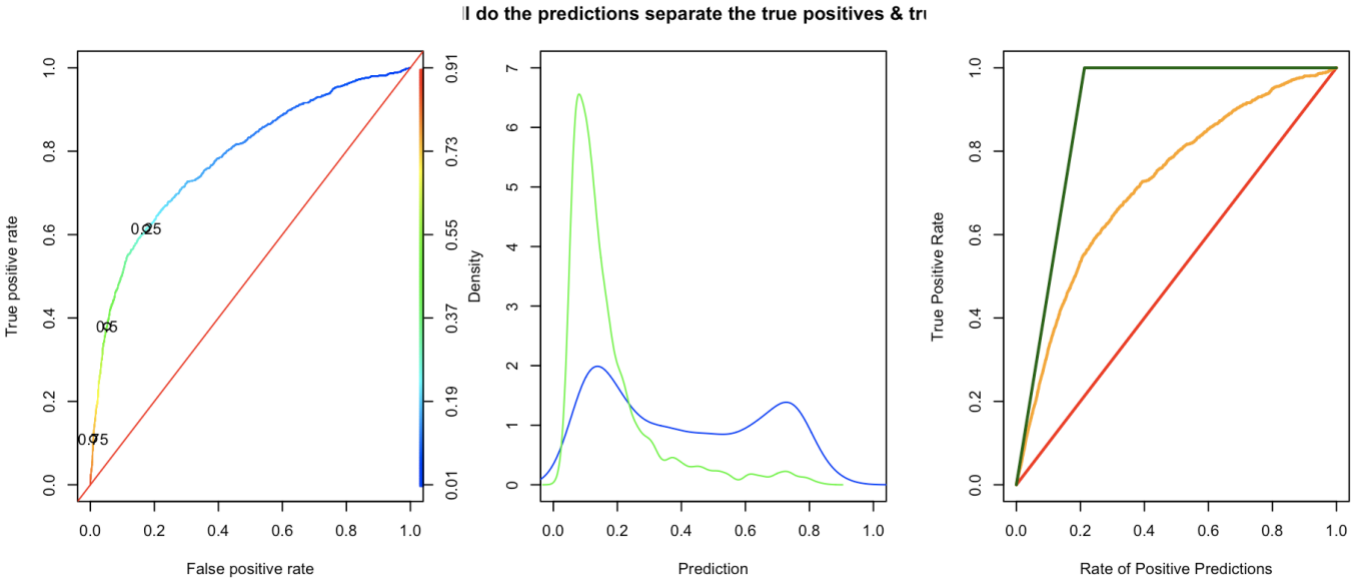
| Method | Package | CV Method | Parameter | Range of Tuning | Final Parameter |
|---|---|---|---|---|---|
| Elastic-net regularized logistic regression | Caret<br>glmnet | 5-repeat<br>10 Fold CV | alpha<br>Lambda | 0 to 1<br>0.0001 to 0.0004 | alpha = 1<br>lambda = 1e-04 |
| Random Forest | Caret<br>rf | 5-repeat<br>10 Fold CV | mtry<br>n.tree | 3 to 12<br>100 to 1600 | mtry = 10<br>n.tree = 500 |
| Decision Tree | Caret<br>rpart | 5-repeat<br>10 Fold CV | Complexity Parameter(cp) | CP at best Gini Index | cp = 0.0011 |
| Decision Tree | Caret<br>rpart | 5-repeat<br>10 Fold CV | Complexity Parameter(cp) | CP at best Information gain | cp = 0.00321 |
| AdaBoost.M1 | adabag | 5-repeat<br>10 Fold CV | mfinal<br>Coeflearn<br>maxdepth | 6:36<br>Breiman, Freund, Zhu<br>1 to 6 | 36<br>Breiman<br>1 |
| Gradient Boosting | gbm | 5-Fold CV | n.trees<br>shrinkage<br>Interaction-depth | 100 to 1000<br>0.01 to 0.03<br>3 to 5 | n.trees = 650<br>shrinkage=0.01<br>Int-depth=4 |

(b)

| | | | | CV Performance | | | |
|---|---|---|---|---|---|---|---|
| Method | Package | Parameter | Parameter | Accuracy | AUC | Kappa | Log Loss |
| Elastic-net regularized logistic regression | Caret<br>glmnet | alpha<br>Lambda | 1<br>1e-04 | 0.8196 | 0.738 | 0.302 | 0.452 |
| Random Forest | Caret<br>rf | mtry<br>n.tree | mtry = 10<br>500 | 0.839 | 0.773 | 0.40 | 0.431 |
| Decision Tree | Caret<br>rpart | Complexity Parameter(cp) | 0.0011 | 0.8371 | 0.751 | 0.373 | 0.451 |
| Decision Tree | Caret<br>rpart | Complexity Parameter(cp) | 0.00321 | 0.829 | 0.743 | 0.425 | 0.438 |
| AdaBoost,M1 | adabag | mfinal<br>Coeflearn<br>Maxdepth | 36<br>Breiman<br>1 | 0.829 | 0.775 | 0.384 | 0.426 |
| Gradient Boosting | gbm | trees<br>shrinkage<br>Interaction -depth | 650<br>0.01<br>4 | 0.826 | 0.790 | 0.390 | 0.417 |

**Best Model is Gradient Boosting:**

| | | | | CV Performance | | | |
|---|---|---|---|---|---|---|---|
| Method | Package | Parameter | Parameter | Accuracy | AUC | Kappa | Log Loss |
| Gradient Boosting | gbm | trees<br>shrinkage | 680<br>0.01 | 0.826 | 0.790 | 0.390 | 0.417 |

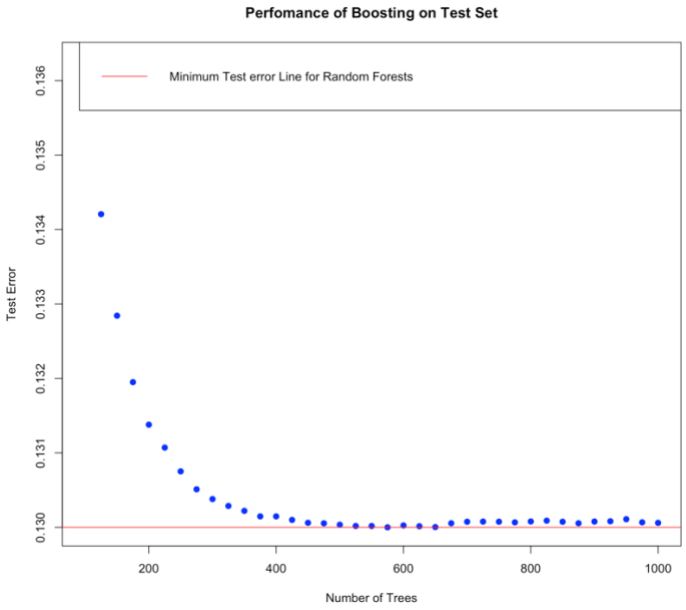


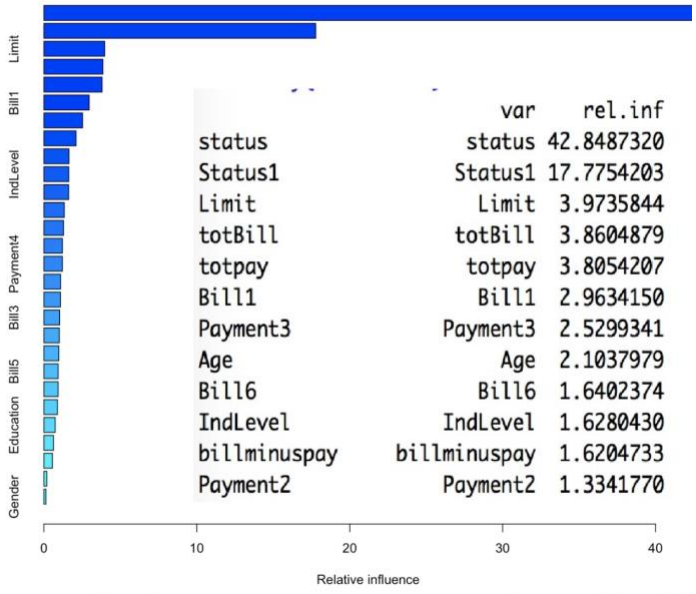

**About above plots and Statistics:**

- The above ROC plot indicates tradeoff between true-positive and false-positive rates and gbm model ROC curve mostly pushed towards the area with less false positive rate and high true positive rate among all other models we created which indicates gbm model has least false positive rate as result more accurate.
- Since ROC curve is pushed towards top left corner as a result the area under curve(AUC) is high among all other models which is 0.790.
- The gain chart, green line indicates ideal model, orange line indicates gbm model curve and red line indicates the random model. As the orange line curve is more pushed towards the ideal curve hence we got the better predictions for the model.
- Distribution of true positives & true negatives: Green line indicates in the true postive and blue line indicates true negatives

**Description for Below plots:**

- **N.tree plot** indicates the hyper parameter (Number of trees for our model). we considered the number of trees are 650 as the curve touches the minimum test error line at that point.
- **Variable Importance plot** indicates the importance of the predictors in the model. From this plot we see that the new feature called status and Status1 explains maximum variance in the data.
- Gbm model got the least test error i.e. 0.13000 and least log loss for test error 0.417 indicates that this is the most accurate model among all other models constructed. It gave a logloss of 0.43511 which is second in the private leader board in kaggle.




| | var | rel.inf |
|---|---|---|
| status | status | 42.8487320 |
| Status1 | Status1 | 17.7754203 |
| Limit | Limit | 3.9735844 |
| totBill | totBill | 3.8604879 |
| totpay | totpay | 3.8054207 |
| Bill1 | Bill1 | 2.9634150 |
| Payment3 | Payment3 | 2.5299341 |
| Age | Age | 2.1037979 |
| Bill6 | Bill6 | 1.6402374 |
| IndLevel | IndLevel | 1.6280430 |
| billminuspay | billminuspay | 1.6204733 |
| Payment2 | Payment2 | 1.3341770 |

**Conclusion:**

If boosting is done properly by selecting appropriate tuning parameters such as shrinkage parameter (lambda), the number of splits we want and the number of trees (n), then it can generalize really well and convert a weak learner to strong learner.