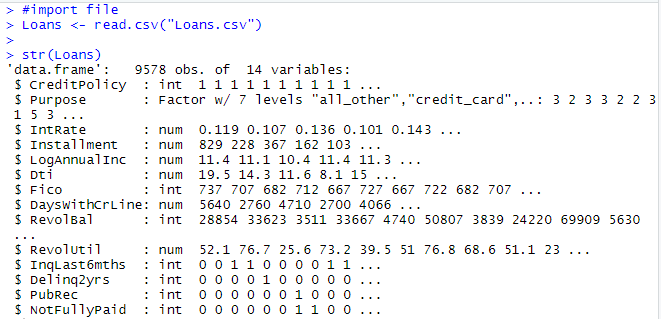
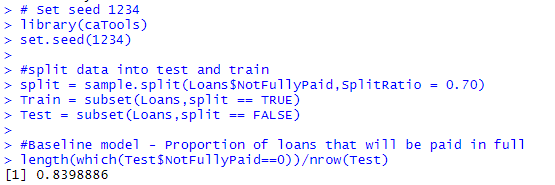
**Lending Club**

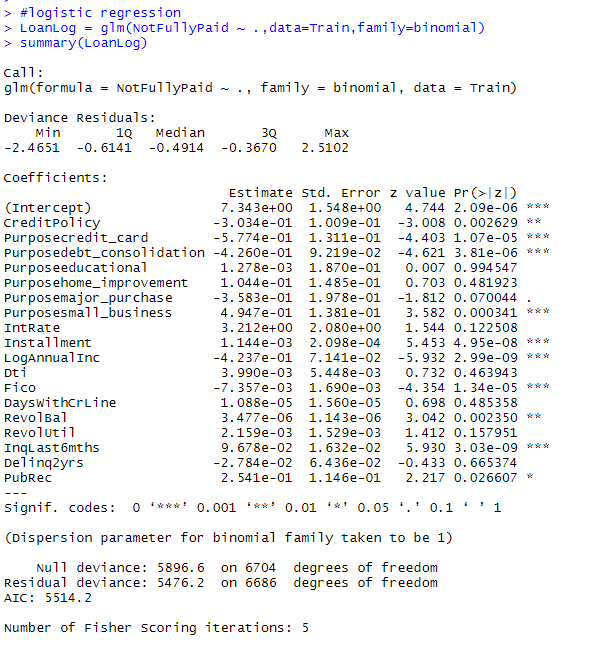
Set seed and split data into train and test.





Accuracy of baseline model that all loans will be paid back in full on the test set is 83.9888%

Logistic regression model of NotFullyPaid on all independent variables -

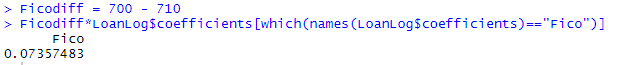


CreditPolicy, Purposecredit\_card, Purposedebt\_consolidation, Purposesmall\_business, Installment, LogAnnualInc, Fico, RevolBal, InqLast6mths and PubRec

are the statistically significant variables. Among these, CreditPolicy, Purposecredit\_card, Purposedebt\_consolidation, LogAnnualInc and Fico reduce the odds of loan not fully paid while increase in other variables increases the odds of loan not fully paid.

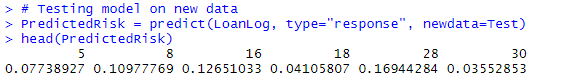
Logit(A) – Logit(B)

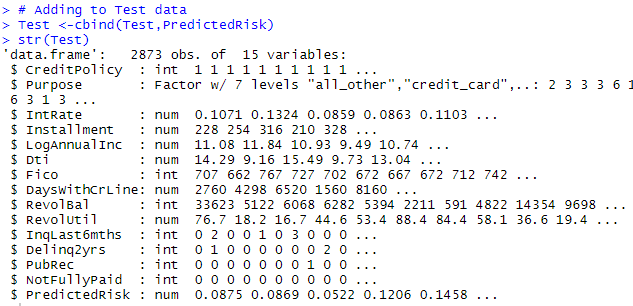
Logit(A) – Logit(B) = βFICO (700) - βFICO (710) as the loans are identical and all other independent variables cancel out in the logit equation.

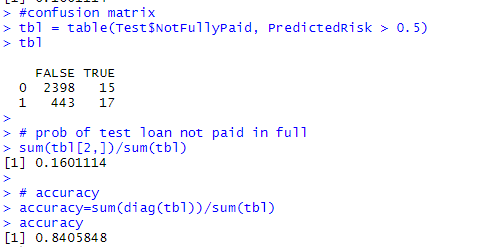


Logit(A) – Logit(B) = 0.0736

Predicted probabilities in Test -



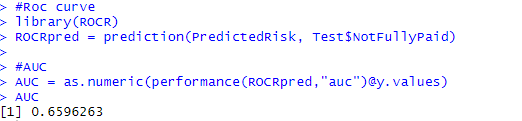




Probability of test set loans not paid back in full = 0.1601.

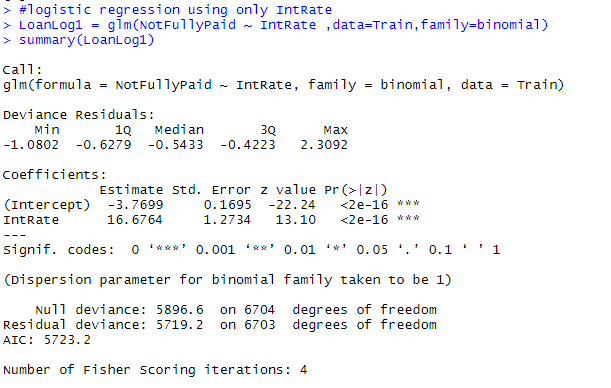
Accuracy of model on test set = 84.0584% which is marginally higher than accuracy of baseline model 83.9888%.

Test AUC



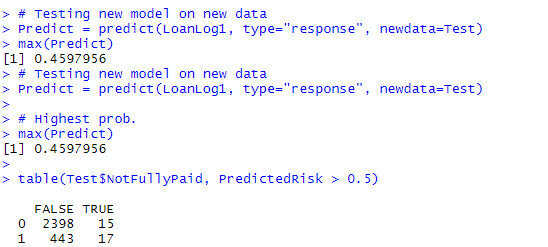
The test set AUC of the model is 65.9626%. As AUC of this model is low and the accuracy of model is only slightly higher than the baseline, this model is not useful for investors to make profitable investments.

Logistic regression using only IntRate as dependent variable -



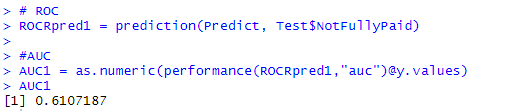
IntRate is statistically significant in this model as p value is < 0.05.However, it was not coming out significant in the previous model with all variables. This indicates a sign of multicollinearity when the presence or absence of a variable changes the significance of variables.

Predictions -



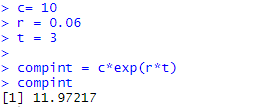
The highest predicted probability of a loan not being paid back in full on test set is 0.4598. As this probability is less than the threshold (0.5), we predict no loan can be paid back in full using this model and threshold of 0.5.

AUC



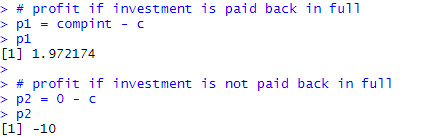
The AUC of this model = 0.6107. This AUC is lower than the AUC with all independent variables. Thus the model with all independent variables having higher AUC is stronger as it has higher discrimination power.

Continuous compounding revenue -



Thus $10 investment with annual interest rate of 6% pays back $11.9722 after 3 years.

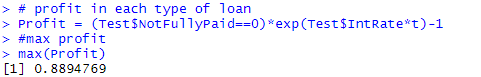
Profit to investor -



Profit to the investor if the investment is paid back in full = $1.9722

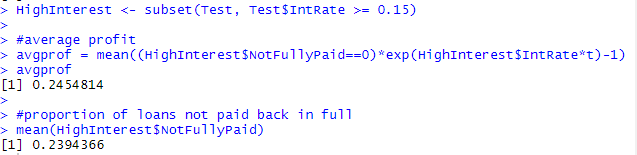
Profit to the investor if the investment is not paid back in full = -$10

Profit of $1 investment in each loan



The maximum profit of $1 investment in each loan in test set = $0.8895.

High Interest -

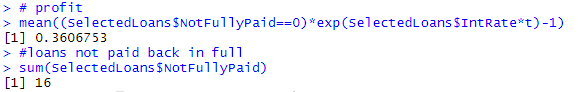


Average profit of $1 investment in high-interest loans is $0.2455.

23.9437% of the high-interest loans were not paid back in full.

Sort loans by PredictedRisk -





Profit of investor who invested $1 = $0.3607.

16 of the 100 selected loans were not paid back in full.

This profit is much higher as compared to the profit of $20.94 for $100 investment simple strategy or $0.2094 profits from $1 investment < $0.3604.

The important assumption of predictive modelling that does not hold in financial situations is that future will be same as past. Financial world changes very fast and past data is not always a good indicator of future as predictive models are not able to account for all types of financial risks. As an analyst, getting more data which is in real-time and reliable and reduces the uncertainty or risk can improve the situation.