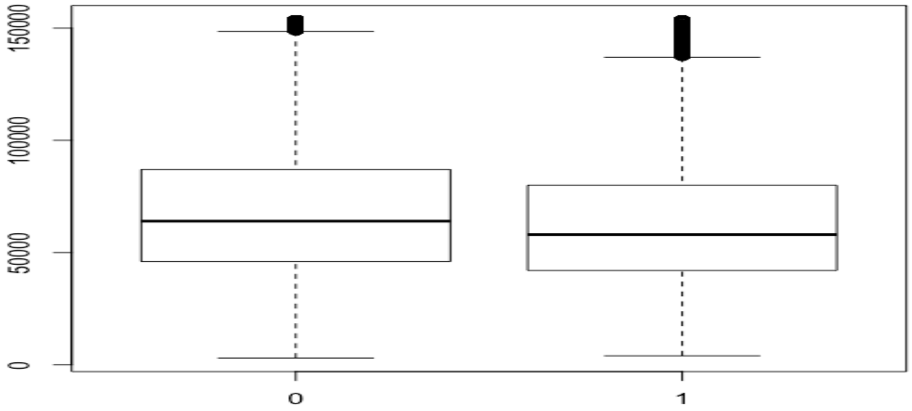
The variable selection process:

1. Transform all the columns into numeric or factor. Otherwise we cannot do any regression analysis on the data.
2. See the NAs in each column. If a large part of the column is NAs that means lots of information is missing in this variable, so I delete those that have more than 50000,or 5% of total, NA records.
3. Next is EDA. As logistic regression does not require the independent variables to be normal, I focus on solving the problem of "complete separation", because this means our model is useless, at least for Inference. So I build boxplots by Loan Status.



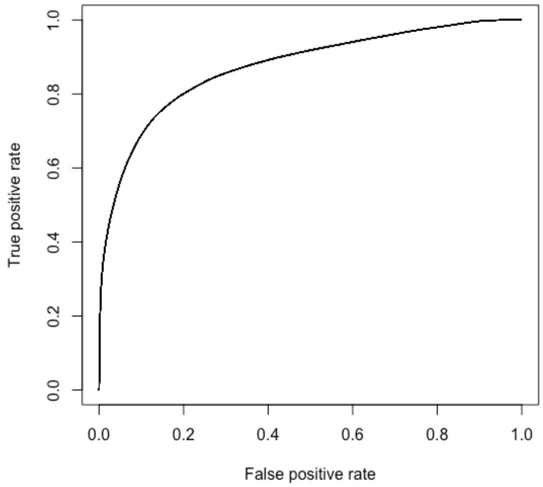
By inspecting the plots, we can know some of the variables will properly cause complete separation. I also draw histograms for all variables.

In the process, I found that annual income and dti have extreme values; I think this may come from those who are un-honest. This will harm the model because it is not real data. So I take 95% of both variables. It's hard to tell whether other variables are abnormal, so I did not do the same thing to them.

4. Next is about correlation. Some of them are highly correlated, thus by inspecting correlation matrix, I only leave one of the highly correlated columns. It contains most information.
5. Then we come to model building part. In logistic regression, AIC is a better criteria when comparing models and selecting variables. So I use stepAIC(Forward) to choose variables..
6. After we get the model, we will look at summary() and anova(). Some

coefficients are not significant, so we can discard them.

7. we need to do analysis.
   A. VIF(). If the vif value is very high(>8), then we need to discard that variable, since it's collinear with others.
   B. Next is outlierTest. We need to detect the outliers, and remove them from the data set. An then we should do regression again.
   C. Mcfadden test.
   D. hoslem.test
   E. durbinWatsonTest() this is not that useful because this test requires that the variables are normally distributed.
   F. ROC plot: (I got 0.87 from the final model)



"As a rule of thumb, a model with good predictive ability should have an AUC closer to 1 (1 is ideal) than to 0.5."