Credit Risk Prediction

Payal Motwani
Ravi Patel
Arun Thiravianathan

FA 541 Applied Statistics with Applications in Finance

Spring 2022

Introduction

Credit risk is the risk a lender takes while extending loans to a borrower, which is the possibility of the borrower defaulting on payment of Principal amount and or interest payment. In this assignment, we will attempt to build a model to predict the credit risk of an individual with maximum accuracy. We will do this using a classification model. We will first test multiple models to find one that will work best with our data. Then we will tune that best type of model to create a final model that has the best performance. The project can be called a success if our final model has high accuracy as that means that our model can predict if a loan will default with high proficiency.

Data

Dataset

We are going to work with a dataset from Kaggle which includes account-level details of more than 32000 customers who have taken debt from a bank.
You can find the dataset here: https://www.kaggle.com/datasets/laotse/credit-risk-dataset

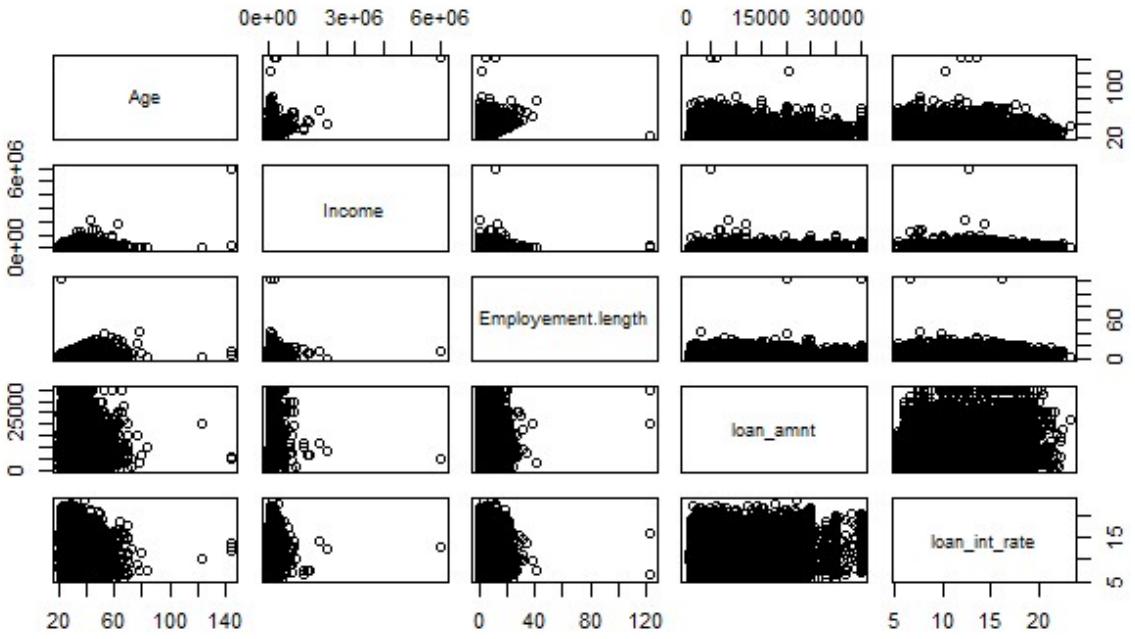
Our dataset has a total of 32,581 data points and 11 variables:

	Variable	**Metrics**	**Type**	**Description**
1	Age	years	Numerical	Captures the age of borrower at the time default is being measured
2	Income	Dollars $	Numerical	Captures the annual income of borrower
3	Home Ownership	Rent/ Mortgage/ Own	Categorical	Captures the status of home ownership of the borrower
4	Employment length	Years	Numerical	Describes the period borrower has been employed
5	Loan intent	Education/ medical/ Venture/ Home improvement/ personal/ debt consolidation	Categorical	Describes the reason for which amount was borrowed
6	Loan amount	Dollars $	Numerical	Captures the amount of borrowing
7	Loan grade	A to G	Categorical	Indicates the riskiness of a borrower. A indicates low risk borrower while G indicates riskiest borrower.
8	Loan Interest rate	Percentage	Numerical	Indicates the rate at which loan was taken.
9	Loan to Income ratio	0-1	Numerical	Indicates the ratio between borrower's existing liabilities and Income. Here, 0 means borrower has no existing liabilities, 1 means entire income of borrower goes to debt payments.
10	Historical default	Y/N	Categorical	Indicates whether customer has ever defaulted on any loan before
11	Loan status	0/1	Categorical	Indicates the current status of the loan; 0 indicates No default, 1 indicates a default

Data Preprocessing

Initial screening of data indicated existence of certain NA values, there were total of 895 in employment length and 3116 data points in Interest rate that were missing. Since these data points are only a small percentage of total data, we decided to drop them rather than extrapolating.

After removing NA values, second step involved looking for outliers. To find outliers, we used scatterplot matrix.



From the above scatterplot matrix we can see clear outliers in Age, Income and Employment length.

We see few values in person's age of more than 100 years. It represents a very realistic case, 144 years seems too unrealistic for age, hence we decided to drop the values above 100 years, keeping the view that it was mis recorded. We also observed outliers in employment length, it's impossible for a person of 22 years of age to be employed for 123 years. We dropped these two data points since they clearly are mis recorded.

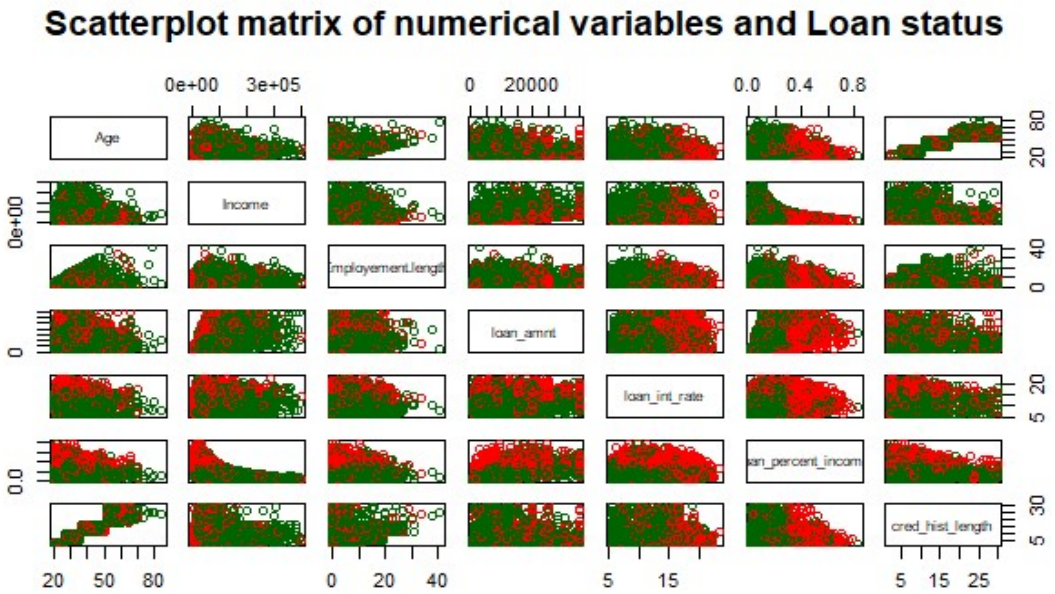
We also observed an outlier in Income, since it been a very large value could affect our analysis, we decided to drop the value.

Numerical Variables

Summary Statistics of Numerical Variables after preprocessing

Variable	Min	Pctl. 25	Median	Pctl. 75	Max	Mean	SD
Age	20	23	26	30	84	27.69	6.1481
Income	4,000	39,222	55,500	80,000	397,800	64,914	39,237.21
Employement length	0	2	4	7	41	4.774	4.0313
Loan amount	500	5,000	8,000	12,500	35,000	9,640	6309.388
Loan interest rate	5.42	7.90	10.99	13.48	23.22	11.04	3.2298
Loan Percent to income	0.01	0.09	0.15	0.23	0.83	0.1699	0.1063
Credit hist length	2	3	4	8	30	5.78	4.0230

Exploratory data Analysis- Numerical



To visualize the relationship between Loan status and Numerical variables, we drew up scatterplot matrix.
Green points represent accounts that didn't default and red points represent accounts that defaulted.
From the scatterplot, we get a basic idea that higher loan amount, higher loan rates, lower income, higher loan to income percentage tend to default more. Age, employment length, and credit history length may not have a large impact on the chance of default.

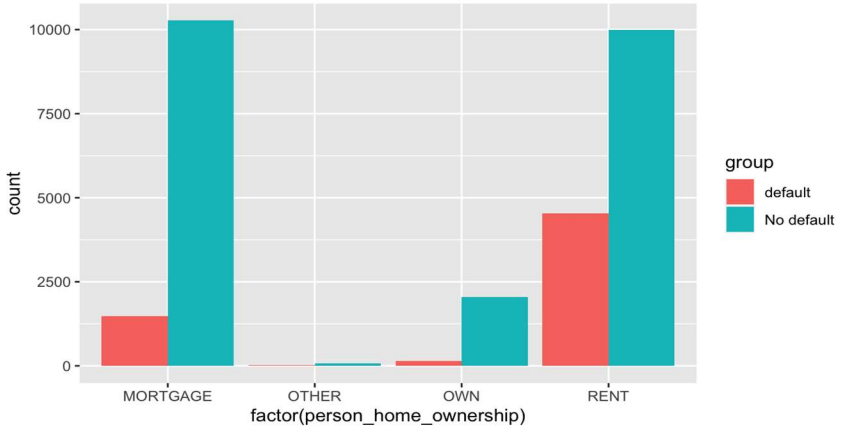
Categorical Variables

Variable	Counts
loan_status	0: 22,363 1: 6,195
person_home_ownership	Mortgage: 11,748 Own: 2,187 Rent: 14,531 Other: 92
loan_intent	Debt Consolidation: 4,548 Education: 5,692 Home Improvement: 3,188 Medical: 5,279 Personal: 4,860 Venture: 4,991
loan_grade	A: 9,379 B: 9,125 C: 5,681 D: 3,239 E: 867 F: 208 G: 59
cb_person_default_on_file	N: 23,473 Y: 5085

Exploratory data Analysis- Categorical

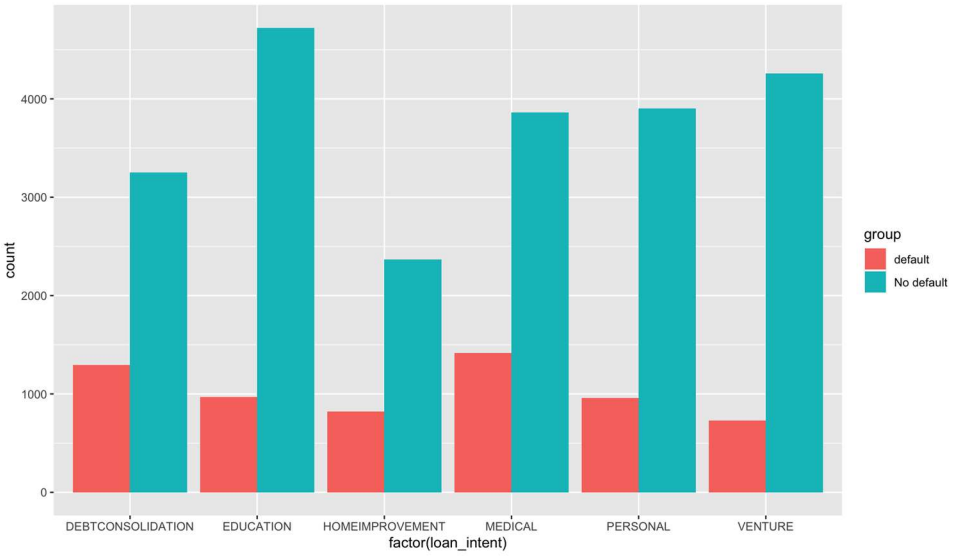
We drew bar charts of categorical variables with loan status as category to visualize the relationship. Green bar represent no default status and red bar represent default status

a) Home Ownership



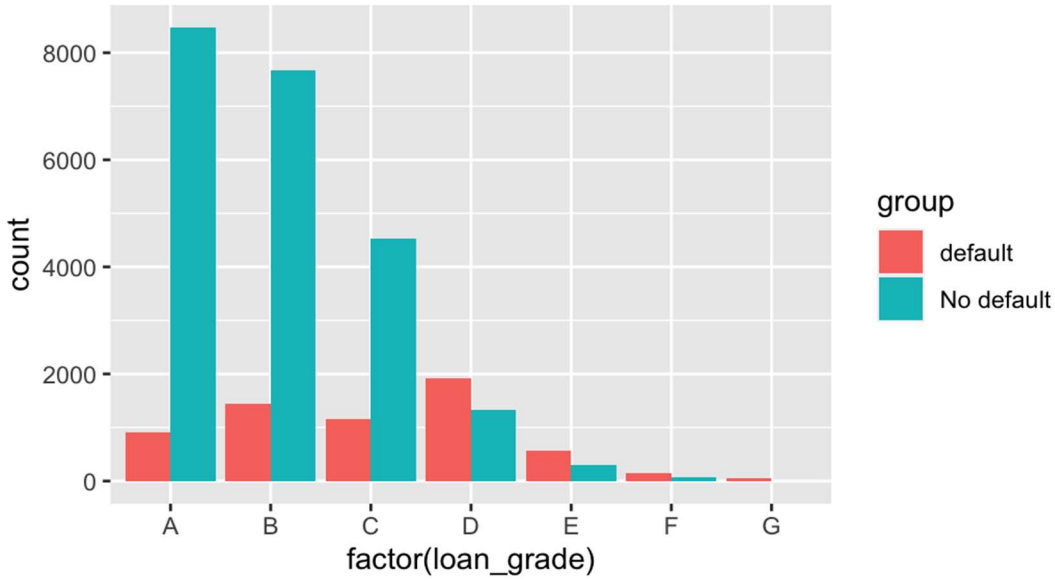
From the plot above, we can clearly see that renters tend to default on loan more, followed by mortgagers and hardly any default by home owners

b) Loan intent:



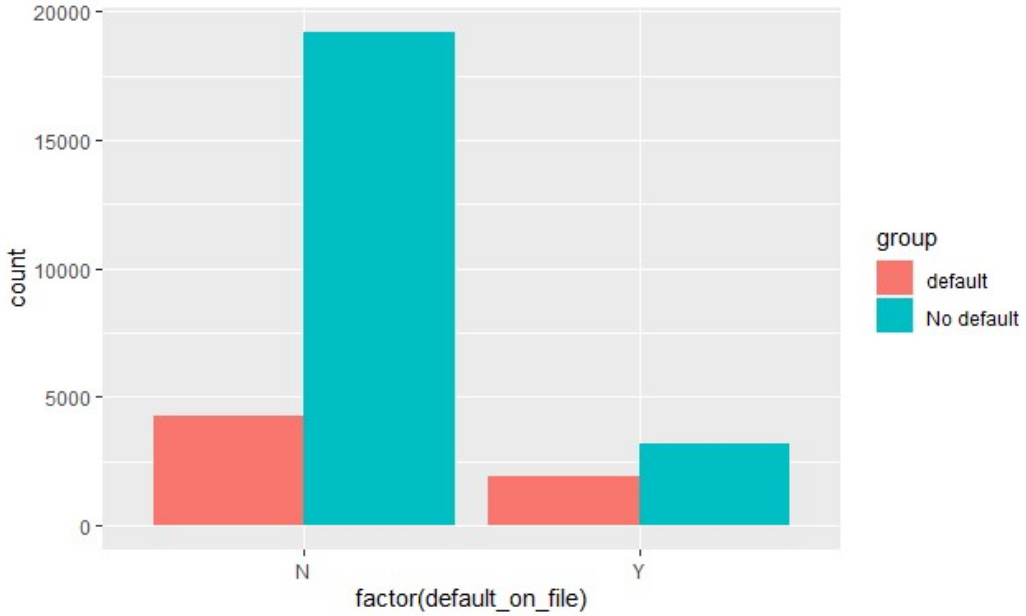
From the plot above, we can see people who took loans for Medical and Debt consolidation purpose defaulted highest amongst all.

d) Loan Grade:



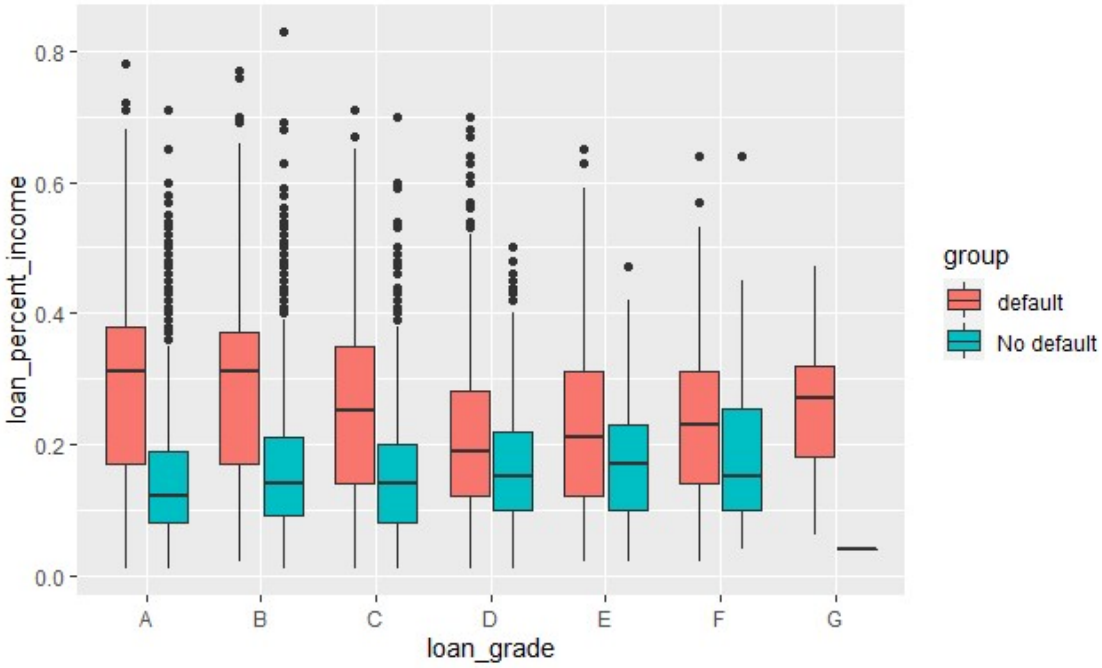
Loan Grade represent the credit quality of buyer, we can see that Grade D was highest defaulter in the category. Also none of the people in Grade G were able to pay back loans

e) Previous Default on File



People who did not have previous default on file defaulted more than previous defaulters.

f) Loan percentage to Income and Loan Grade



From the boxplot of loan Grade and Loan to percentage Income, it is very evident that mean of people with low loan to income percent have been non defaulters. Mean loan to income percentage of defaulters in grade A and B was higher than other grades. Second point to take away from the plot is that none of borrowers in Grade G were able to repay the loan

Splitting data into training and test set

We split the data into two sets, namely training and test set, to train the model and measure accuracy on untrained set.

Training data consisted of 75% of all data and test set had remaining 25%.

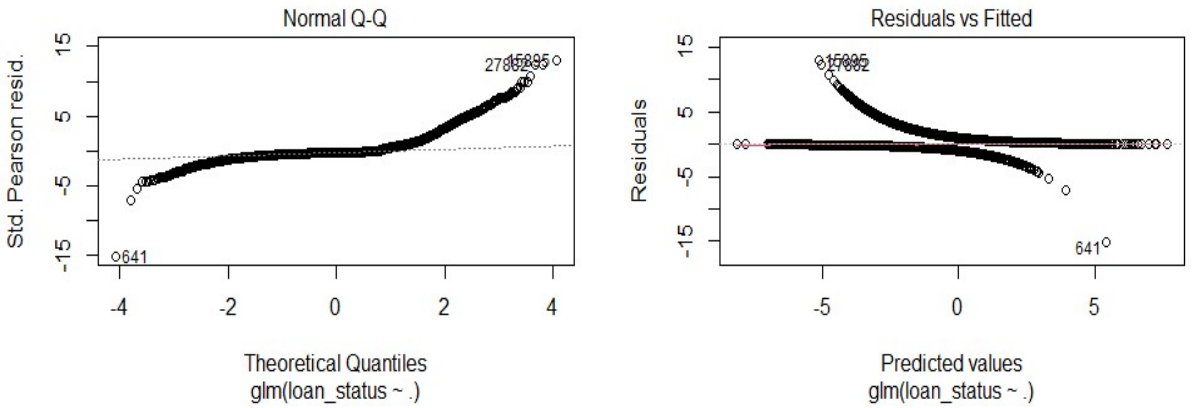
Model Training and Evaluation

We initially tried three different classification models in our project. We tried logistic regression, linear discriminant analysis, and a classification tree. Our purpose in testing these three models was to find which model has the best performance using all variables with the base models to decide on the best model to move forward with. With the best model we plan to move forward with it and further tune the model to get the best possible performance with it.

Model 1 - Logistic Regression (All Variables)

Our first model was running logistic regression on test data with all the variables as predictors and Loan status as our dependent variable.
We trained our model using Logistic regression to be able to evaluate its ability to predict our probability of default.

To assess model's performance, we will be fitting the model on Test data to calculate the accuracy of prediction

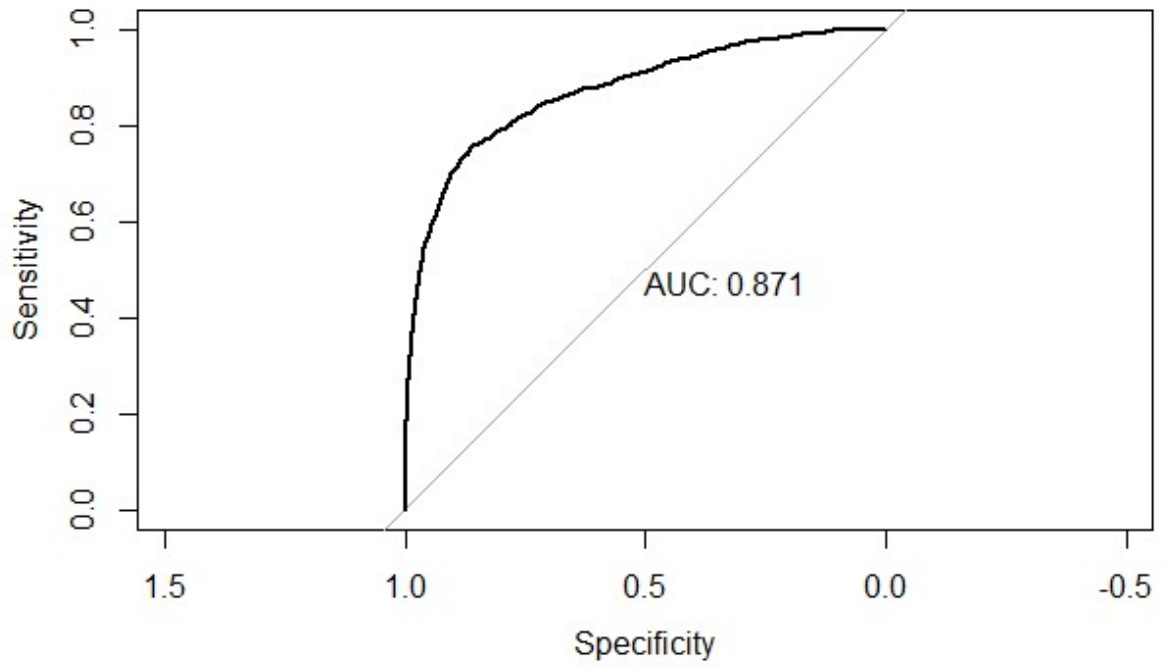


Results:

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2916  -0.5147  -0.2981  -0.1200   3.3465
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22461  on 21417  degrees of freedom
## Residual deviance: 14308  on 21395  degrees of freedom
## AIC: 14354
##
## Number of Fisher Scoring iterations: 6
```

Accuracy – Model 1

To test the accuracy of model, we fitted above model to Test data, if the predicted probability was more than 50% it was assigned to value 1 and less than 50% was assigned to zero.



Tabular result of predicted probability Model 1 as below

Probability	**Did not default**	**Defaulted**
Predicted no default	5323	683
Predicted default	259	875

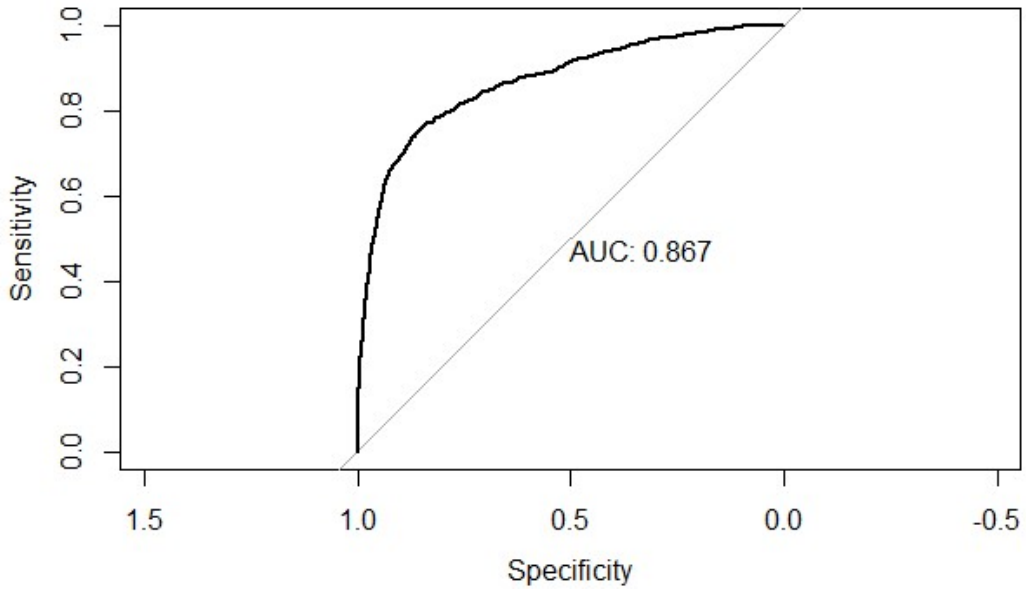
Accuracy of Model 1 was reported at 87.1%

Model 2 - Linear Discriminant Analysis

In our second Model, we ran Linear Discriminant Analysis on the entire dataset with all the predictors.

```
## lda(loan_status ~ ., data = train_data)
##
## Prior probabilities of groups:
##         0         1
## 0.7834999 0.2165001
## ## Coefficients of linear discriminants:
##                                     LD1
## Age                          -3.375008e-03
## Income                        5.249868e-06
## Home.ownershipOTHER           2.691791e-01
## Home.ownershipOWN            -6.493479e-01
## Home.ownershipRENT            4.640299e-01
## Employement.length           -3.880504e-03
## loan_intentEDUCATION         -4.956691e-01
## loan_intentHOMEIMPROVEMENT    5.587246e-03
## loan_intentMEDICAL           -9.568160e-02
## loan_intentPERSONAL          -3.841240e-01
## loan_intentVENTURE           -5.793368e-01
## loan_gradeB                   1.911503e-02
## loan_gradeC                   7.568074e-02
## loan_gradeD                   1.925352e+00
## loan_gradeE                   2.129090e+00
## loan_gradeF                   2.214729e+00
## loan_gradeG                   3.620486e+00
## loan_amnt                    -9.046821e-05
## loan_int_rate                 4.006161e-02
## loan_percent_income           1.014699e+01
## default_on_fileY              3.150429e-02
## cred_hist_length             -2.251887e-03
```

Accuracy – Model 2



Tabular result of predicted probability Model 2 as below

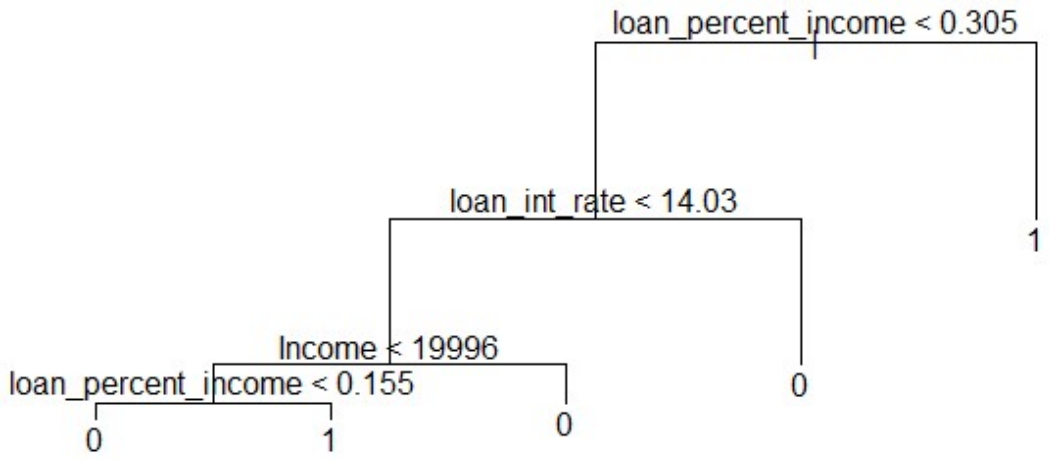
Probability	Did not default	Defaulted
Predicted no default	5255	633
Predicted default	327	925

Accuracy of Model 2 was reported at 86.7%

Model 3 – Decision Trees (Classification)

A classification tree is **a structural mapping of binary decisions that lead to a decision about the class (interpretation) of an object** (such as a pixel).

We trained our model on classification tree, tree as obtained was



Summary of Model

```
summary(tree2)

##
## Classification tree:
## tree(formula = loan_status ~ ., data = train_data)
## Variables actually used in tree construction:
## [1] "loan_percent_income" "loan_int_rate"       "Income"
## Number of terminal nodes:  5
## Residual mean deviance:  0.7259 = 15540 / 21410
## Misclassification error rate: 0.1594 = 3415 / 21418
```

Accuracy – Model 3

Probability	Did not default	Defaulted
Predicted no default	5334	891
Predicted default	248	667

Accuracy of Model 3 was reported at 84.04%

Model Pruning

From all the tree models used above, highest accuracy was obtained for Model 1 which was based on Logistic regression. We decided to try and Prune Model 1 to be able to improve accuracy. We Pruned the model basis Stepwise subset selection and Anova Chi square test. Since all the models above gave almost equal accuracy, we, will be using AIC to determine the best model. AIC for all predictor logistic regression was reported at 14354

Model 4 – Stepwise Pruning of Model 1

We used both forward and backward stepwise regression to create the best model with this method. It gave a model that removed no further variables. It has the same accuracy and AIC as the base model with all significant variables because it is the same model.

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2151  -0.5194  -0.2997  -0.1225   3.3455
##
##
##     Null deviance: 22379  on 21417  degrees of freedom
## Residual deviance: 14367  on 21398  degrees of freedom
## AIC: 14407
##
## Number of Fisher Scoring iterations: 6
```

Accuracy- Model 4

Probability	Did not default	Defaulted
Predicted no default	5321	683
Predicted default	261	875

Accuracy of Model 4 was reported at 86.77%

AIC of Model 4 was reported at 14407

Stepwise Pruning of the Model reported that Age and employment length were less significant to the predictability of loan status. This Model reduced our AIC significantly but it also reduced the accuracy of Model.

Model 5 – ANOVA on Model 1

In Model 5, we ran Anova test on Model 1, logistic regression with all predictors to evaluate their goodness of fit to model

Summary of test

```
                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                      21417      22563
person_age                 1     14.3     21416      22549 0.0001581 ***
person_income              1   1289.4     21415      21260 < 2.2e-16 ***
person_home_ownership      3    868.0     21412      20392 < 2.2e-16 ***
person_emp_length          1      1.6     21411      20390 0.2076430
loan_intent                5    337.1     21406      20053 < 2.2e-16 ***
loan_grade                 6   3427.2     21400      16626 < 2.2e-16 ***
loan_amnt                  1   1006.7     21399      15619 < 2.2e-16 ***
loan_int_rate              1     11.5     21398      15607 0.0006864 ***
loan_percent_income        1   1062.3     21397      14545 < 2.2e-16 ***
cb_person_default_on_file  1      0.1     21396      14545 0.7122529
cb_person_cred_hist_length 1      0.4     21395      14545 0.5061547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Based on the test above, we dropped employment length, previous default on file and credit history length as predictors and ran a logistic regression model on with remaining predictors.

Results of logistic regression were:

```
##
## Call:
## glm(formula = loan_status ~ Age + Income + Home.ownership + loan_intent +
##     loan_grade + loan_amnt + loan_int_rate + loan_percent_income,
##     data = train_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24554  -0.18814  -0.06242   0.08404   1.24718
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1089765)
```

```
##
##     Null deviance: 3633.1  on 21417  degrees of freedom
## Residual deviance: 2331.9  on 21398  degrees of freedom
## AIC: 13328
##
## Number of Fisher Scoring iterations: 2
```

Accuracy – Model 5

Probability	Did not default	Defaulted
Predicted no default	5333	740
Predicted default	249	818

Accuracy of Model 3 was reported at 86.14%

AIC of Model 3 was reported at 13328

Model 6 – Pruning Model 5

We then generated another ANOVA table off this model again and got the ANOVA table below.

```
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                21417     3633.1
## Age                  1     2.23     21416     3630.9 6.041e-06 ***
## Income               1   156.37     21415     3474.5 < 2.2e-16 ***
## Home.ownership       3   139.84     21412     3334.6 < 2.2e-16 ***
## loan_intent          5    53.84     21407     3280.8 < 2.2e-16 ***
## loan_grade           6   607.65     21401     2673.1 < 2.2e-16 ***
## loan_amnt            1    83.86     21400     2589.3 < 2.2e-16 ***
## loan_int_rate        1     0.62     21399     2588.7   0.01743 *
## loan_percent_income  1   256.80     21398     2331.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We to prune Model 5 further by removing loan interest rate and running logistic regression again to see if it had a positive impact on the model (to see if the AIC of the model decreased).

Summary of Model 6

```
##
## Call:
## glm(formula = loan_status ~ Age + Income + Home.ownership + loan_intent +
##     loan_grade + loan_amnt + loan_percent_income, data = train_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.25653  -0.18807  -0.06239   0.08245   1.24305
##     Null deviance: 3633.1  on 21417  degrees of freedom
## Residual deviance: 2333.2  on 21399  degrees of freedom
## AIC: 13339
##
## Number of Fisher Scoring iterations: 2
```

As we can see that removing loan interest rate increased AIC from Model 5

Model accuracy: 86.09%

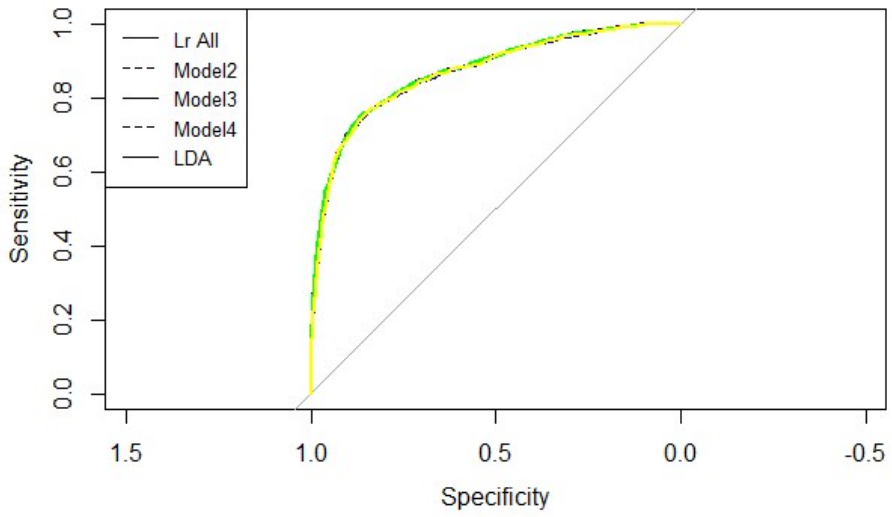
Model AIC: 13339

Removing loan interest rate did not increase accuracy or decrease AIC.

Summary of Performance of all Models

S.no	Model	AIC	Accuracy
Model 1	Logistic Regression with all Variables	14354	87.1%
Model 2	Linear Discriminant Analysis		86.7%
Model 3	Classification Tree		84.04%
Model 4	Logistic Regression with Stepwise Pruning	14407	86.77%
Model 5	Logistic Regression with ANOVA Variables 1	13328	86.14%
Model 6	Logistic Regression with ANOVA Variables 2	13339	86.09

ROC curve for all Models

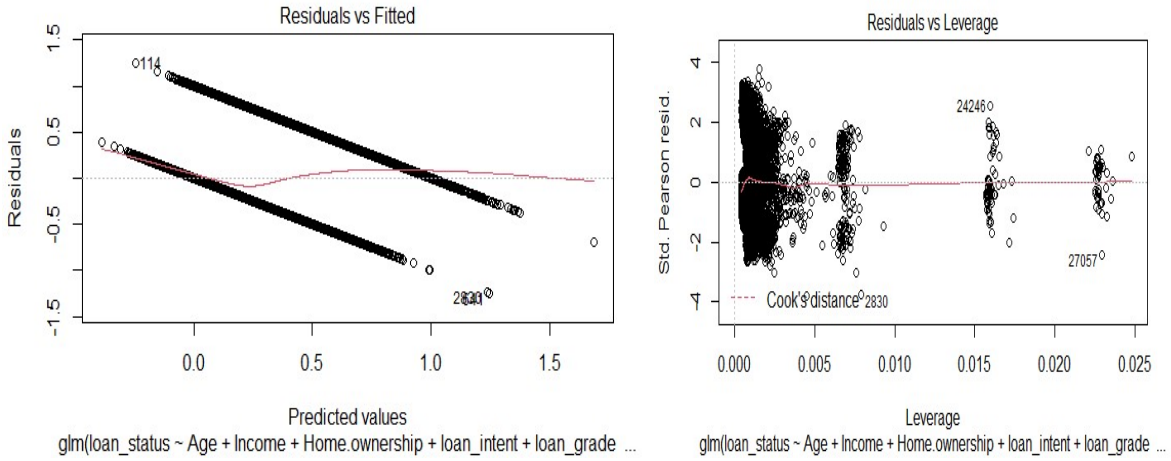


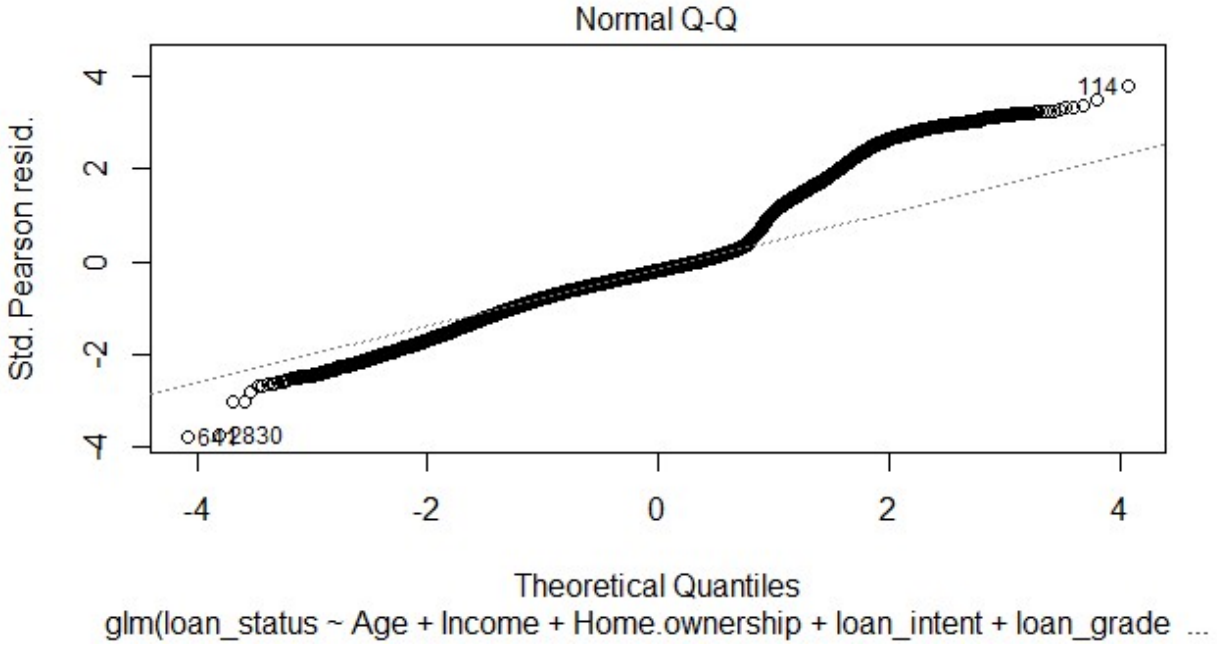
From the performance table and ROC curve above, we can see that accuracy for all the models was almost equal, curve for all the models is overlapping.
Therefore we decided to chose the model with lowest AIC, as AIC selects the best model given goodness of fit and interpretability of the Model.
Based on Model training and evaluation our final model performance summarized below:
In our final model, Age, Income, Home Ownership, Loan Intent, Loan Grade, Loan amount, Loan interest rate and Loan Percent to Income was used as predictors.

Final Model Performance

```
##
## Call:
## glm(formula = loan_status ~ Age + Income + Home.ownership + loan_intent +
##     loan_grade + loan_amnt + loan_int_rate + loan_percent_income,
##     data = train_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24554  -0.18814  -0.06242   0.08404   1.24718
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -1.255e-01  2.174e-02  -5.773 7.91e-09 ***
## Age                        -9.989e-04  3.714e-04  -2.689 0.007168 **
## Income                      1.032e-06  9.900e-08  10.427  < 2e-16 ***
## Home.ownershipOTHER         5.460e-02  4.114e-02   1.327 0.184469
## Home.ownershipOWN          -1.280e-01  9.066e-03 -14.120  < 2e-16 ***
## Home.ownershipRENT          9.286e-02  5.037e-03  18.436  < 2e-16 ***
## loan_intentEDUCATION       -9.781e-02  7.599e-03 -12.871  < 2e-16 ***
## loan_intentHOMEIMPROVEMENT  1.266e-03  8.913e-03   0.142 0.887026
## loan_intentMEDICAL         -1.893e-02  7.738e-03  -2.447 0.014413 *
## loan_intentPERSONAL        -7.582e-02  7.878e-03  -9.624  < 2e-16 ***
## loan_intentVENTURE         -1.145e-01  7.887e-03 -14.512  < 2e-16 ***
## loan_gradeB                 3.870e-03  9.924e-03   0.390 0.696560
## loan_gradeC                 1.830e-02  1.509e-02   1.213 0.225317
## loan_gradeD                 3.835e-01  1.952e-02  19.651  < 2e-16 ***
## loan_gradeE                 4.237e-01  2.546e-02  16.640  < 2e-16 ***
## loan_gradeF                 4.411e-01  3.635e-02  12.135  < 2e-16 ***
## loan_gradeG                 7.180e-01  5.730e-02  12.530  < 2e-16 ***
## loan_amnt                  -1.791e-05  7.156e-07 -25.031  < 2e-16 ***
## loan_int_rate               7.926e-03  2.236e-03   3.545 0.000394 ***
## loan_percent_income         2.005e+00  4.131e-02  48.543  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1089765)
##
##     Null deviance: 3633.1  on 21417  degrees of freedom
## Residual deviance: 2331.9  on 21398  degrees of freedom
## AIC: 13328
```





Accuracy – Final Model

Accuracy of final model reported in tabular form as below:

Probability	Did not default	Defaulted
Predicted no default	5333	740
Predicted default	249	818

Accuracy of Model 3 was reported at 86.14%

AIC of Model 3 was reported at 13328

Conclusion

In our attempt to be able to predict credit risk on an Individual, we worked on dataset from Kaggle, we first preprocessed and cleaned the data for better prediction. We trained our data on three models initially, namely, logistic regression, Linear Discriminant Analysis and Classification trees. After that Logistic regression model was pruned further to increase accuracy and decrease AIC of model.
Our final model was able to predict credit risk of an individual with 86.14% accuracy. Based on our model we can say, Age, Income, Home Ownership, Loan Intent, Loan Grade, Loan amount, Loan interest rate and Loan Percent to Income are predictors of importance.