Probability of Default Model

Problem Statement- To develop logistic and decision tree model for predicting whether the customer will default or not. Aim of the project will be to keep sensitivity high as accepting a defaulter cost the bank more than rejecting a non-defaulter.

Data – Data is present in R data format and consist of following variables:-

Observations: 29,091

Variables: 8

$ loan_status : int 0 0 0 0 0 0 1 0 1 0 ...
$ loan_amnt : int 5000 2400 10000 5000 3000 12000 9000 3000 10000 1000 ...
$ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 1 5 2 3 2 2 4 ...
$ home_ownership: Factor w/ 4 levels "MORTGAGE","OTHER",..: 4 4 4 4 4 3 4 4 4 4 ...
$ annual_inc : num 24000 12252 49200 36000 48000 ...
$ age : int 33 31 24 39 24 28 22 22 28 22 ...
$ emp_cat : Factor w/ 5 levels "0-15","15-30",..: 1 2 1 1 1 1 1 1 1 1 ...
$ ir_cat : Factor w/ 5 levels "0-8","11-13.5",..: 4 5 2 5 5 2 2 4 4 3 ...

Out of these loan status is the outcome or dependent variable and rest are independent variable. Loan status equal to 0 means that the customer will not default and loan status equal to 1 means that the customer will default.

Categorical Variables – grade, home-ownership, emp_cat & ir_cat.

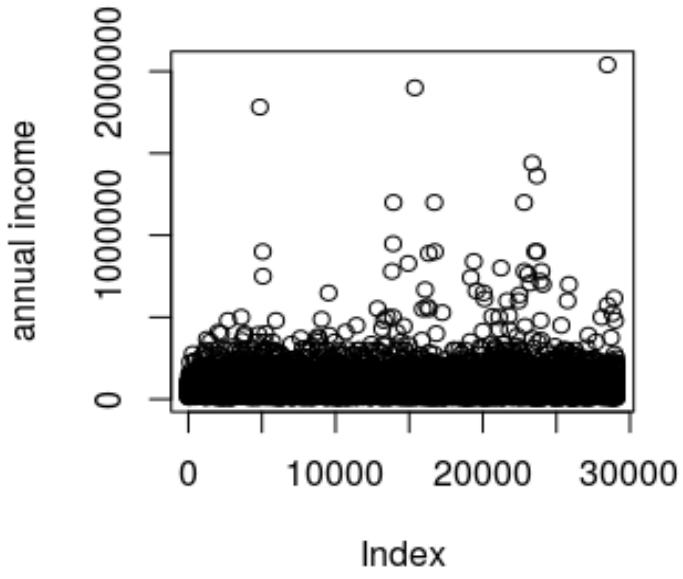
Numerical Variables – Loan amount (Loan_amnt), annual income (annual_inc) & age.

Default Rate – Percentage of defaulters out of total observations. The default rate in the given data came out to be 11.254% (generally the industry practice is that it must be greater than 5%). But, still the dataset is unbalanced and the model may underestimate the defaulters.

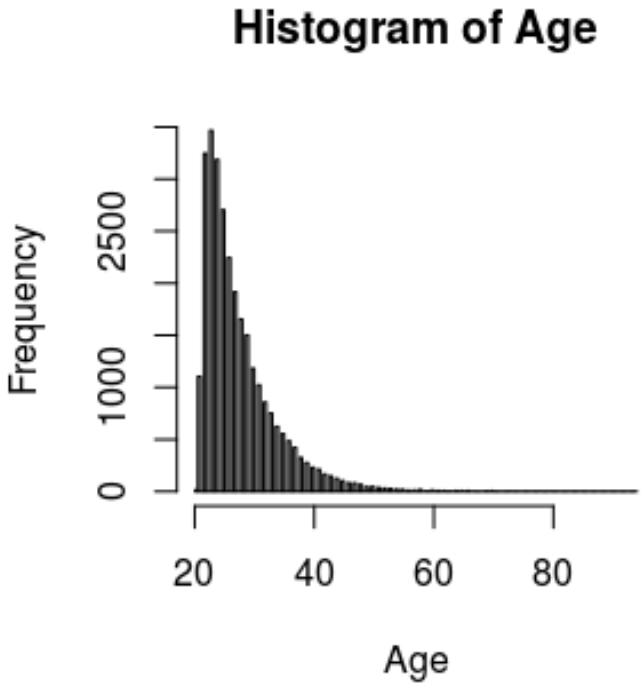
Missing Value Treatment – The data is already treated for missing values. Missing values are found in the ir_cat variable. 9.54% of values are found missing in the ir_cat variable, as this is greater than 5% we cannot impute these directly. But Dataset is already treated for missing values.

Exploratory Data Analysis

Univariate Analysis

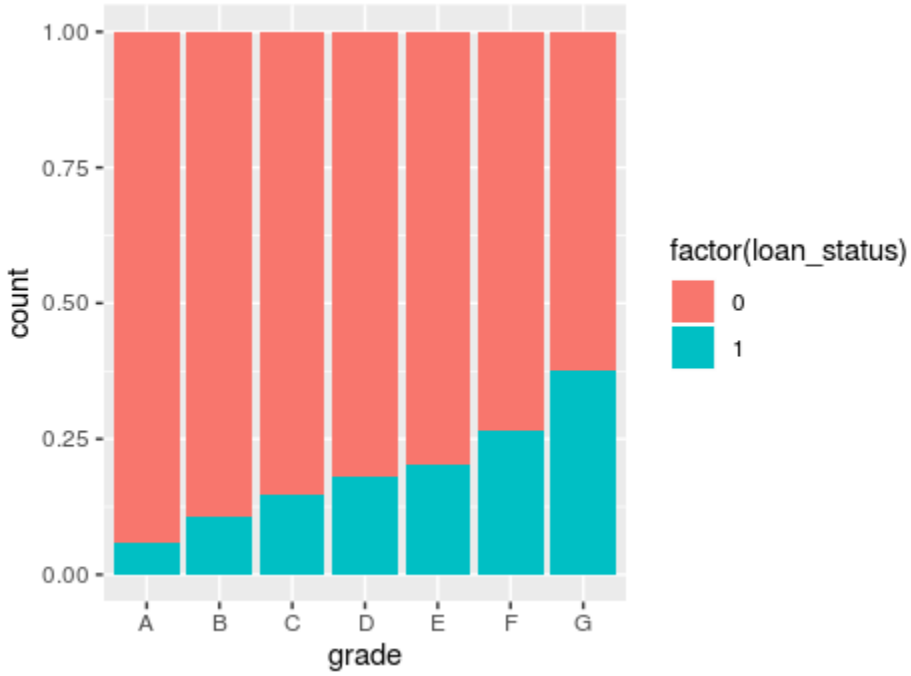


We can conclude that there are very less loan takers above annual income 1000000$.

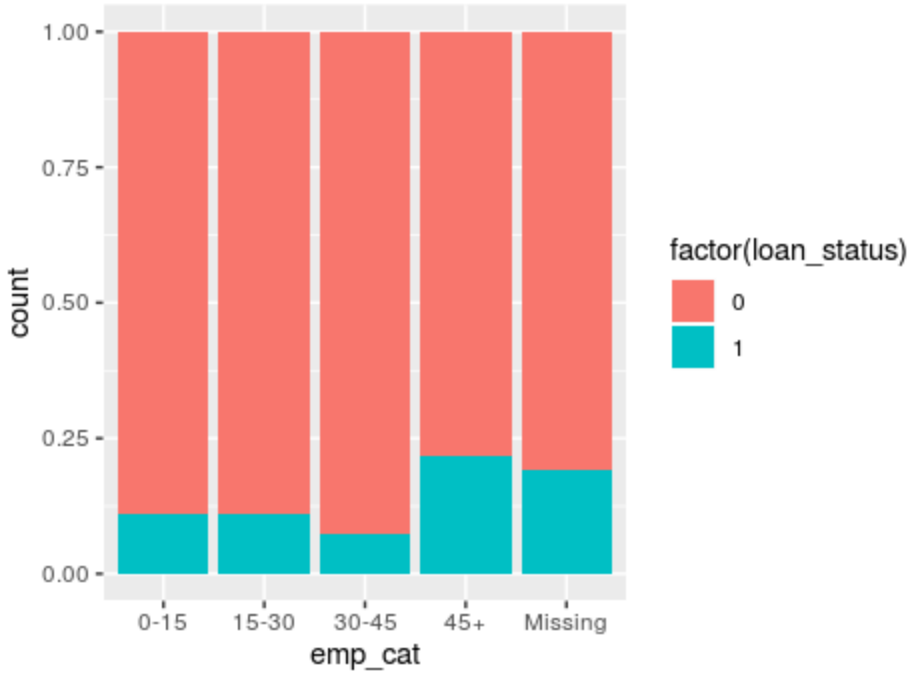


We can see that as the age is increasing the number of loan takers are decreasing.

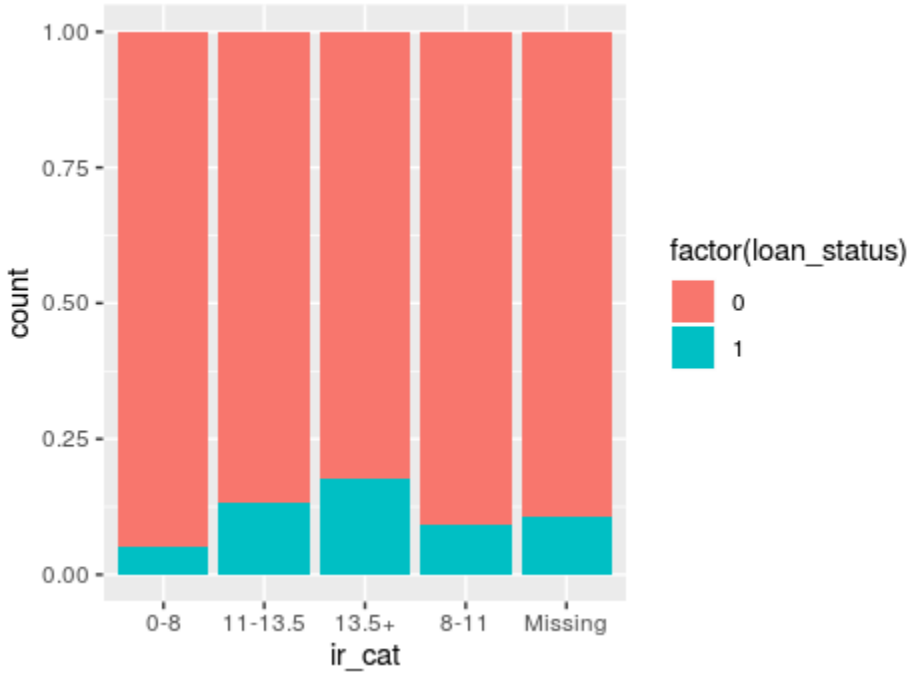
borrowers having homes on rent and mortgages are more probable to default.



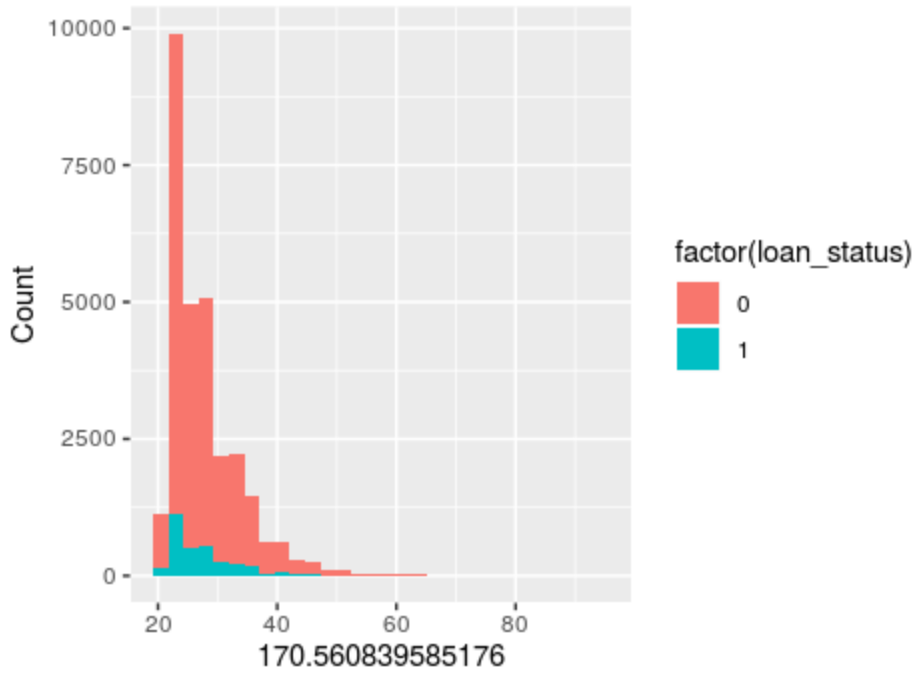
proportion of defaults are increasing as the grade is decreasing from A to G.



proportion of defaulters from emp_cat 45+ is maximum.



proportion of default for interest rate category 13.5+ is maximum.



We can see that as the annual income is increasing the number of loan takers and defaulters are decreasing.

We can see here that as age is increasing annual income is decreasing. Also the defaults are concentrated for lower income groups having lesser age.

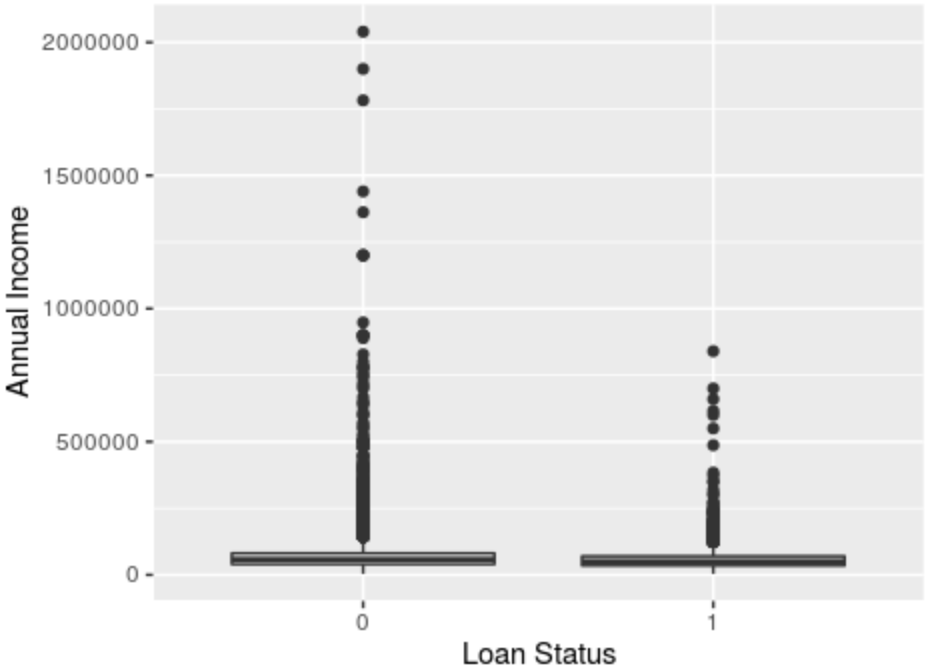
We can conclude that defaults are concentrated for lower income and higher loan amount group. Also, we can see as the annual income is increasing the loan amount is also increasing. Also, there are few extreme cases where annual income is high and the loan amount is less.

Outlier Analysis

Based on EDA we are building our model on the following logic & those data points that contradict this logic will be treated as outlier: -

- Probability of default is increasing as grade decreases from A to G.
- As the age is increasing, annual income is increasing and therefore probability of default is decreasing.
- As the annual income is increasing the Loan amount is also increasing. And probability of default is decreasing.

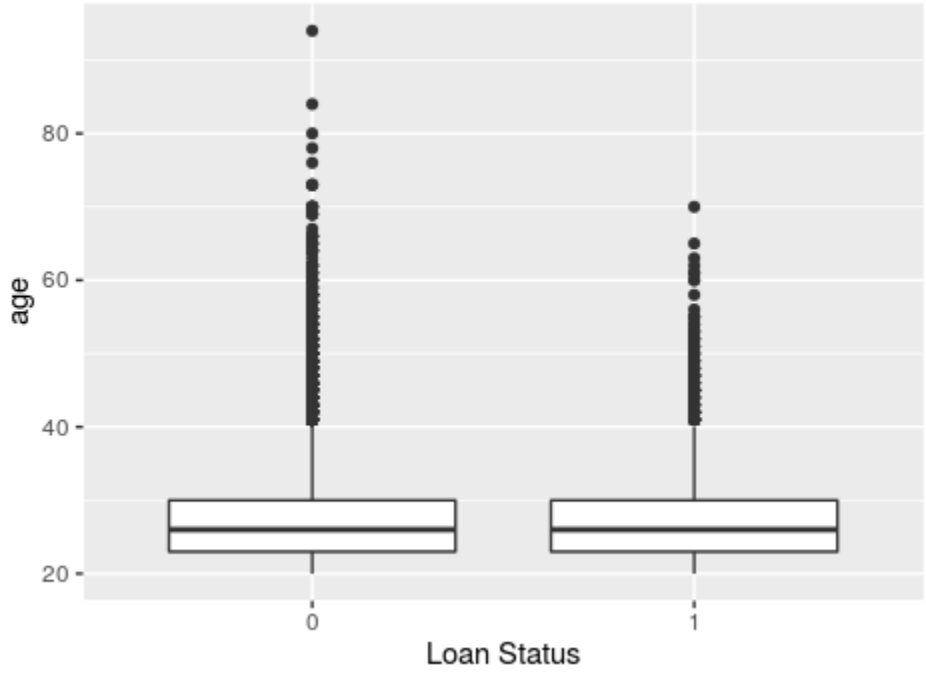
Visualizing distribution through box plot



We can see that whiskers are not present on one side but we cannot say that how many of these data points do not conform to our model therefore using IQR concept.

Outlier cutoff is coming to be 140000 by using IQR concept but from EDA we found that above 1000000$ annual income the loan amount was decreasing (which goes against the logic on which model is based) also there were no defaults in this range so we can safely remove these points (Only 0.02% of the data is removed).

Checking for Outliers in Age Column



Here also the whiskers are present only on one side of plot therefore, checking the IQR cutoff. Cutoff by IQR is coming out to be 40.5 but we cannot remove data points above 40.5 years age as although probability of default is decreasing with age but if we remove these points, we will create too much bias in the model.

Therefore, removing only values greater than 80 years of age. (This removes only two observations)

Standardizing the data

Both logistic regression and decision tree works better with standardized data.

Test & Training Set

Training set has been made by randomly selecting two third values from whole data and remaining values are taken for test set.

Model Development

Logistic Model – First of all the model is developed on all the variables.

```
Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.959203   0.070848 -41.768  < 2e-16 ***
loan_amnt            0.017426   0.025762   0.676 0.498765
annual_inc          -0.266530   0.036264  -7.350 1.99e-13 ***
age                  0.003743   0.023429   0.160 0.873053
gradeB               0.366545   0.110278   3.324 0.000888 ***
gradeC               0.586500   0.124966   4.693 2.69e-06 ***
gradeD               0.850300   0.141454   6.011 1.84e-09 ***
gradeE               1.079657   0.168508   6.407 1.48e-10 ***
gradeF               1.510455   0.238834   6.324 2.54e-10 ***
gradeG               1.975495   0.378583   5.218 1.81e-07 ***
home_ownershipOTHER  0.469890   0.316311   1.486 0.137403
home_ownershipOWN   -0.005191   0.091166  -0.057 0.954590
home_ownershipRENT   0.006743   0.052319   0.129 0.897451
emp_cat15-30         0.205539   0.084518   2.432 0.015020 *
emp_cat30-45        -0.102781   0.274155  -0.375 0.707735
emp_cat45+           0.388000   0.630024   0.616 0.537994
emp_catMissing       0.761766   0.117989   6.456 1.07e-10 ***
ir_cat11-13.5        0.583812   0.135441   4.310 1.63e-05 ***
ir_cat13.5+          0.534536   0.150190   3.559 0.000372 ***
ir_cat8-11           0.305940   0.121824   2.511 0.012028 *
ir_catMissing        0.341423   0.133086   2.565 0.010304 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13608  on 19386  degrees of freedom
Residual deviance: 13022  on 19366  degrees of freedom
AIC: 13064
```

We can see that only grade, annual income & interest rate category variable have higher significance.

Model Results: -

best_accuracy 0.891376108933361"

best_auc 0.607329750291599"

best_sensitivity 0.737891737891738"

best accuracy cutoff 0.39"

best auc cutoff 0.11"

best sensitivity cutoff 0.1"

Developing model only on significant variables (grade, annual income & interest rate category)

Model Results: -

best_accuracy 0.891479265525067"
best_auc 0.608456525807913"
best_sensitivity 0.730294396961064"
best accuracy cutoff 0.41"
best auc cutoff 0.11"
best sensitivity cutoff 0.1"

We can see that AUC has improved for this model.

Now, Checking VIF for this model.

```
               GVIF Df GVIF^(1/(2*Df))
grade      9.618631  6        1.207608
annual_inc 1.007988  1        1.003986
ir_cat     9.553267  4        1.325925
```

We can see that VIF for grade and interest rate category is significantly high therefore they cannot be used together in model.

Removing grade and building the model on rest two

Model Results: -

best_accuracy 0.890757169383124"
best_auc 0.601166472304072"
best_sensitivity 0.743589743589744"
best accuracy cutoff 0.22"
best auc cutoff 0.1"
best sensitivity cutoff 0.1"

Since auc decreased therefore including grade and removing interest rate category.

Developing model using only grade & annual Income.

Model Results: -

best_accuracy 0.891582422116773"
best_auc 0.612465549683464"
best_sensitivity 0.760683760683761"
best accuracy cutoff 0.39"
best auc cutoff 0.12"
best sensitivity cutoff 0.1"

AUC has improved in the model by including only grade and annual income

In our hypothesis we concluded that age could be one of default determining factor Therefore, including age although it came out to be insignificant in initial model.

"best_accuracy 0.891582422116773"
"best_auc 0.61366876239769"
"best_sensitivity 0.760683760683761"
"best accuracy cutoff 0.39"
"best auc cutoff 0.12"
"best sensitivity cutoff 0.1"

We can see that AUC has improved. Therefore, this is our final logistic regression model.

Decision Tree Model

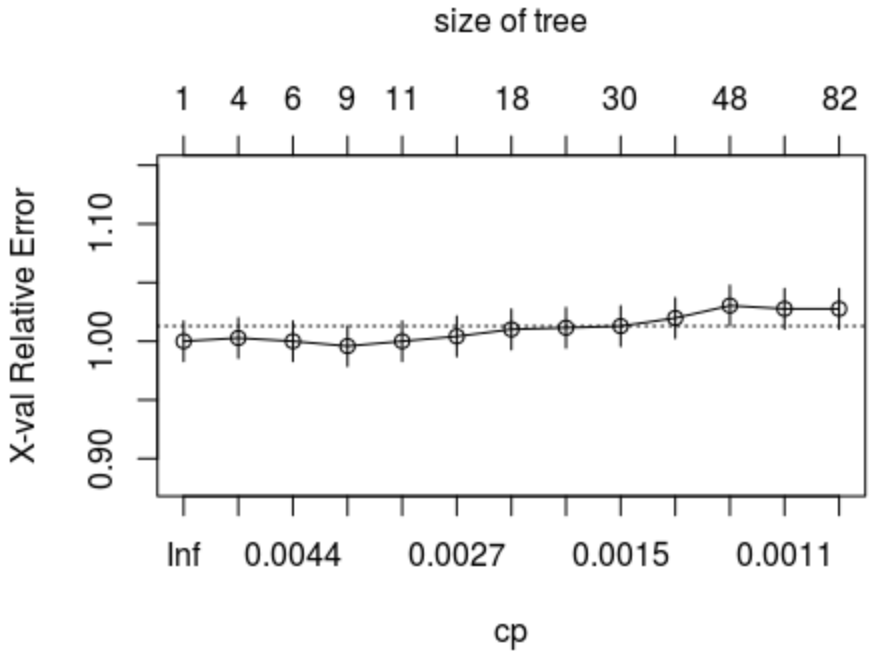
Initial model was developed using all variables(dec_mod1) and setting no limit on complexity parameter. The tree could not branch properly. After this the complexity parameter was set to 0.001 and then the model was developed only on grade, annual income & age. Still the model is not branching well.

Decision tree model performs well for small balanced data, Therefore, through under sampling the data is made balanced to include such that one third values are defaults and randomly selecting two third non-default values.

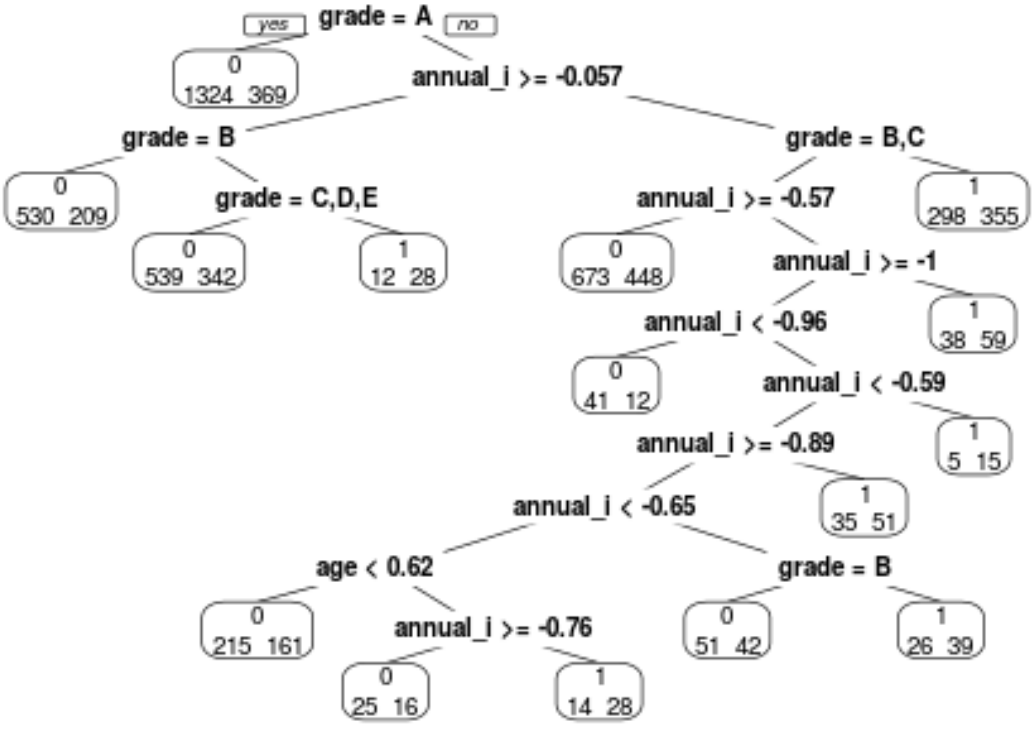
After this model was developed using grade, annual income & age.

Pruning the decision tree model for better performance.

```
CP nsplit rel error  xerror     xstd
1  0.0087397      0   1.00000 1.00000 0.017126
2  0.0048298      3   0.97378 1.00276 0.017137
3  0.0039865      5   0.96412 1.00000 0.017126
4  0.0036799      8   0.95216 0.99586 0.017111
5  0.0031049     10   0.94480 1.00000 0.017126
6  0.0022999     14   0.93238 1.00414 0.017142
7  0.0018399     17   0.92548 1.01012 0.017163
8  0.0017249     20   0.91996 1.01150 0.017168
9  0.0013799     29   0.90156 1.01288 0.017173
10 0.0011960     36   0.89190 1.01978 0.017198
11 0.0011500     47   0.87626 1.03036 0.017234
12 0.0010733     61   0.86017 1.02760 0.017224
13 0.0010000     81   0.83349 1.02760 0.017224
```



We can see that minimum xerror is achieved for cp = 0.003104876. After this the decision tree is pruned for this complexity parameter.



Strategy Curve –

It has been asked that we have we have to keep the acceptance rate between 50% to 75%. Therefore checking the acceptance rate, bad rate, cutoff, accuracy, sensitivity & AUC. We have to find the cutoff such that the rest of parameters remain optimized.

Decision Tree Model

	accept_rate	cutoff	bad_rate	accuracy	sensitivity	AUC	
[1,]		0.75	0.3996	0.0898	0.7595	0.3438	0.577
[2,]		0.74	0.3996	0.0898	0.7595	0.3438	0.577
[3,]		0.73	0.3996	0.0898	0.7595	0.3438	0.577
[4,]		0.72	0.3996	0.0898	0.7595	0.3438	0.577
[5,]		0.71	0.3996	0.0898	0.7595	0.3438	0.577
[6,]		0.70	0.3996	0.0898	0.7595	0.3438	0.577
[7,]		0.69	0.3996	0.0898	0.7595	0.3438	0.577
[8,]		0.68	0.3996	0.0898	0.7595	0.3438	0.577
[9,]		0.67	0.3996	0.0898	0.7595	0.3438	0.577
[10,]		0.66	0.3996	0.0898	0.7595	0.3438	0.577
[11,]		0.65	0.3996	0.0898	0.7595	0.3438	0.577
[12,]		0.64	0.3996	0.0898	0.7595	0.3438	0.577
[13,]		0.63	0.3996	0.0898	0.7595	0.3438	0.577
[14,]		0.62	0.3996	0.0898	0.7595	0.3438	0.577
[15,]		0.61	0.3488	0.0830	0.6195	0.5318	0.581
[16,]		0.60	0.3488	0.0830	0.6195	0.5318	0.581
[17,]		0.59	0.3488	0.0830	0.6195	0.5318	0.581
[18,]		0.58	0.3488	0.0830	0.6195	0.5318	0.581
[19,]		0.57	0.3488	0.0830	0.6195	0.5318	0.581
[20,]		0.56	0.3488	0.0830	0.6195	0.5318	0.581
[21,]		0.55	0.3488	0.0830	0.6195	0.5318	0.581
[22,]		0.54	0.3488	0.0830	0.6195	0.5318	0.581
[23,]		0.53	0.3488	0.0830	0.6195	0.5318	0.581
[24,]		0.52	0.3488	0.0830	0.6195	0.5318	0.581
[25,]		0.51	0.3488	0.0830	0.6195	0.5318	0.581
[26,]		0.50	0.3488	0.0830	0.6195	0.5318	0.581

We can see that decision tree model is not giving better results as compared to logistic Model.

Logistic Regression Model

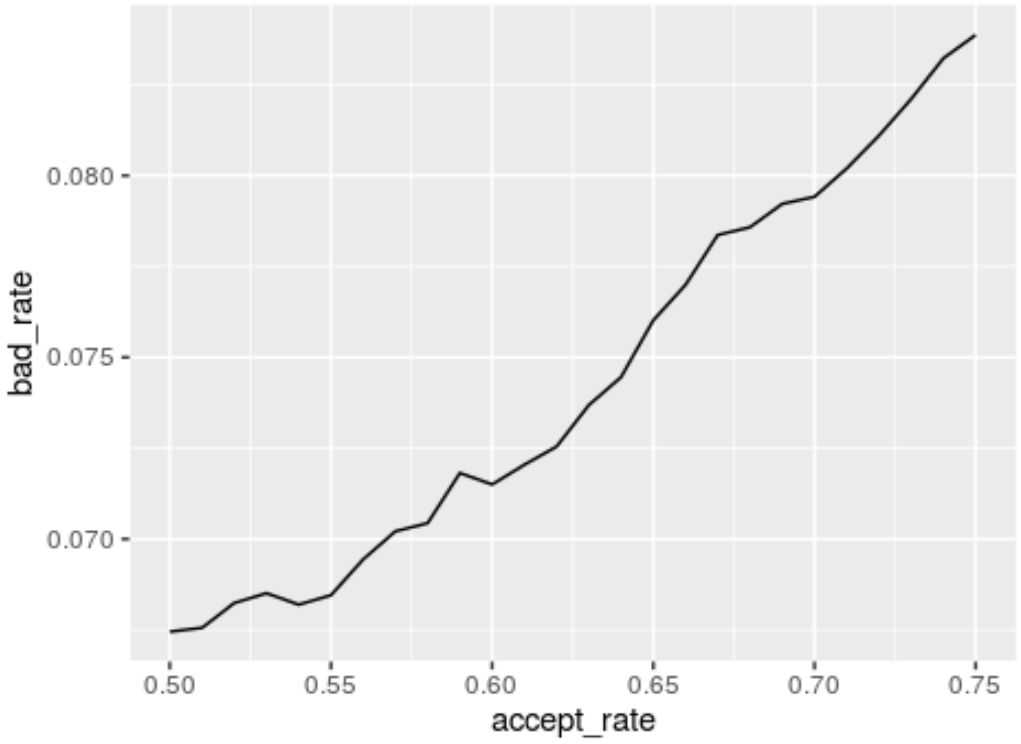
Strategy Table

	accept_rate	cutoff	bad_rate	accuracy	sensitivity	AUC
[1,]		0.75	0.1473	0.0844	0.7323	0.4169	0.5938
[2,]		0.74	0.1446	0.0831	0.7256	0.4340	0.5976
[3,]		0.73	0.1416	0.0825	0.7181	0.4454	0.5983
[4,]		0.72	0.1392	0.0812	0.7116	0.4615	0.6018
[5,]		0.71	0.1369	0.0803	0.7046	0.4748	0.6037
[6,]		0.70	0.1350	0.0796	0.6972	0.4872	0.6050
[7,]		0.69	0.1334	0.0791	0.6895	0.4976	0.6053
[8,]		0.68	0.1318	0.0792	0.6810	0.5043	0.6034
[9,]		0.67	0.1306	0.0777	0.6746	0.5204	0.6069
[10,]		0.66	0.1293	0.0772	0.6668	0.5309	0.6071

```
[11,]        0.65 0.1282   0.0765   0.6592      0.5423 0.6078
[12,]        0.64 0.1272   0.0754   0.6521      0.5556 0.6097
[13,]        0.63 0.1258   0.0729   0.6468      0.5774 0.6163
[14,]        0.62 0.1250   0.0729   0.6383      0.5840 0.6145
[15,]        0.61 0.1240   0.0720   0.6307      0.5954 0.6152
[16,]        0.60 0.1229   0.0720   0.6221      0.6021 0.6133
[17,]        0.59 0.1217   0.0720   0.6137      0.6087 0.6115
[18,]        0.58 0.1208   0.0713   0.6058      0.6192 0.6117
[19,]        0.57 0.1195   0.0702   0.5986      0.6315 0.6131
[20,]        0.56 0.1184   0.0702   0.5901      0.6382 0.6112
[21,]        0.55 0.1173   0.0690   0.5827      0.6505 0.6125
[22,]        0.54 0.1159   0.0688   0.5744      0.6581 0.6111
[23,]        0.53 0.1148   0.0687   0.5658      0.6648 0.6093
[24,]        0.52 0.1137   0.0684   0.5575      0.6724 0.6079
[25,]        0.51 0.1126   0.0683   0.5491      0.6790 0.6061
[26,]        0.50 0.1113   0.0677   0.5411      0.6885 0.6058
```

We can see that for 60% acceptance rate and 0.1229 cutoff we are getting optimized values of bad rate (7.2%), accuracy(62.21%), sensitivity(60.21%) and AUC(61.33%).

Strategy Curve



Conclusion

Best model achieved using logistic model with grade , annual income and age as variables, acceptance rate 60% and cutoff 0.1229.