Group 283: BANK LOAN DEFAULTER PREDICTION

First Name	Last Name	Email (hawk.iit.edu)	Student ID
Venkata Narsimha Sri Dattu	Manapragada	vmanapragada@hawk.iit.edu	A20453958
Shreeyesh	Chauhan	schauhan3@hawk.iit.edu	A20449780
Manasi	Shah	mshah114@hawk.iit.edu	A20454824

Table of Contents

1. Introduction

Main idea of the project is to analyze the historical data of the bank loans and predict the defaulters which would help the banks to take appropriate decisions. Here, the profile of “potentially bad customers” and “good customers” is analyzed. We will be considering factors such as the annual income, work experience, purpose of taking the loan and credit score. In this way, bank can detect default behavior in the earlier stage rather than being too late for bank.

2. Data

Application/Domain: Banking and Finance
Source of the dataset: https://www.kaggle.com/zaurbegiev/my-dataset
Size of dataset: 17.9MB
Number of instances: 1,00,514
Number of attributes: 19

Description of attributes:
Loan: Loan Id, Loan status, Current loan amount, Monthly debt, Month since last delinquent, Term ,Current credit balance
Credit : Credit score, Year of credit history, Number of credit problems, Maximum open credit,
Customer: Customer Id, Annual Income, Number of open accounts, Year of current job,Home ownership, Tax Lieus, Purpose.

3. Problems to be Solved

- To Predict whether loan should be sanctioned to a person on the basis of his history.
- To validate whether credit score of people who has paid the loan (Fully Paid) is equal to credit score of defaulters (Changed Off).

4. Solutions

- Build classification model
- Two sample hypothesis testing

  Plan:

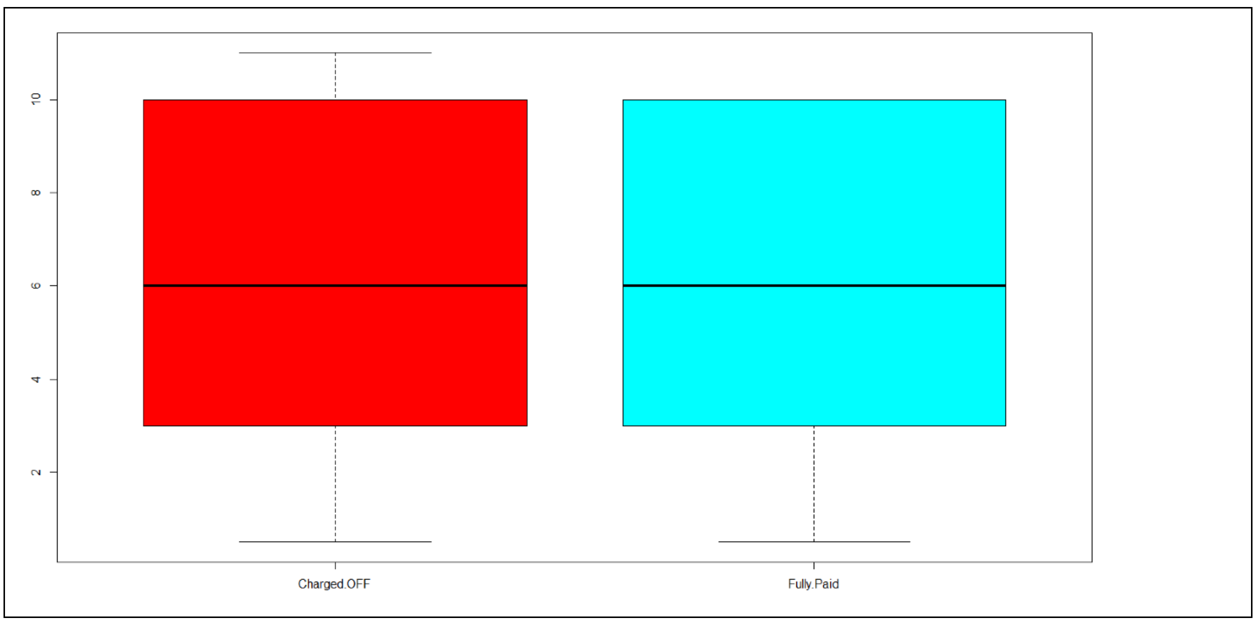
  1. Data Cleansing:
     - Missing value
     - Attribute selection
     - Noisy value
     - Conversion of data types

2. Classification:
    - Logistic regression
    - K-Nearest Neighbors
3. Holdout Method:
    - Training Set
    - Test Set
4. Data Visualization:
    - Box plot
    - Bar Chart

5. Experiments and Results

Preprocessing:

For the x-variable "years in current job" we have grouped the data into 3 categories as we can observe from the below boxplot that for charged off and fully paid are equally distributed.



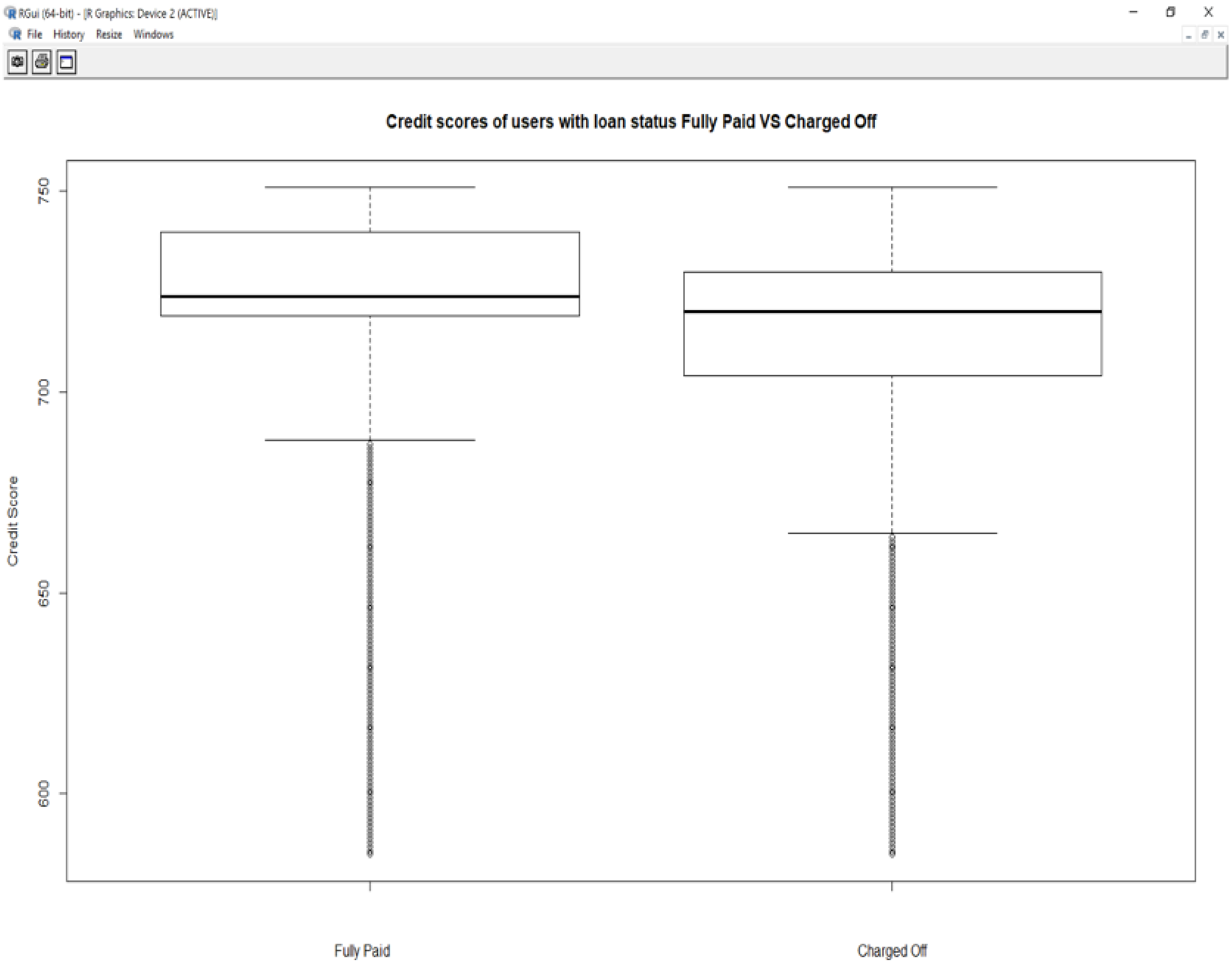
```
> sort(unique(MM3$years_in_current_job))
 [1] < 1 year  1 year    10+ years 2 years   3 years   4 years   5 years   6 years   7 years   8 years
[11] 9 years   n/a

> MM3$current_job_year <- ifelse((MM3$years_in_current_job == ('< 1 year')
+                                | MM3$years_in_current_job == ('1 year')
+                                | MM3$years_in_current_job ==('2 years')
+                                | MM3$years_in_current_job == ('3 years')
+                                | MM3$years_in_current_job == ('4 years')),'0-4 ',
+                               ifelse((MM3$years_in_current_job == ('5 years')
+                                       | MM3$years_in_current_job ==('6 years')
+                                       | MM3$years_in_current_job == ('7 years')
+                                       | MM3$years_in_current_job ==('8 years')
+                                       | MM3$years_in_current_job == ('9 years')
+                                       | MM3$years_in_current_job == ('n/a')),'5-9',
+                                      '>=10'))
>
> head(MM3$current_job_year)
[1] "5-9"  ">=10" "5-9"  "0-4 " "5-9"  ">=10"
```

5.1. Methods and Process

- Credit score of people who has paid the loan (Fully Paid) is same as credit score of defaulters (Changed Off)
- Null hypothesis(H0): Average of credit scores of fully paid is same as that of charged off.
- Alternative hypothesis (Ha): Average of credit scores of fully paid is not same as that of charged off.

Below box-blot depicts the basic overview of the claim made above.




```
> z.test(paid,charged_off, alternative = "two.sided", mu = 0, sigma.x=sd(paid),sigma.y=sd(charged_off), conf.level = 0.95, paired=F)

        Two Sample z-test

data:  paid and charged_off
z = 50.192, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  9.934093 10.741457
sample estimates:
mean of x mean of y
 723.0327  712.6949
```

As we observe above box plot there is no much difference in the median values and there is no necessity to perform z-test. So we have performed the chi-square test to calculate the probability value to predict the hypothesis testing.

```
> chisq.test(MM2$loan_status, MM2$loan_status)

	Pearson's Chi-squared test with Yates' continuity correction

data:  MM2$loan_status and MM2$loan_status
X-squared = 81994, df = 1, p-value < 2.2e-16
```

By taking 95% confidence level and the calculated p-value is less than the alpha value. So, we are rejecting the Null-Hypothesis. Hence, we go with the alternative hypothesis which says that both the group means of the Charged off and Fully paid people are different.

Average of credit scores of fully paid is more than that of charged off.

5.2. Evaluations and Results

Logistic Regression

Forward model:

```
Call:
glm(formula = loan_status ~ ., family = binomial(), data = train.data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-8.4904  -1.1194   0.6595   0.7815   1.9985  

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)    
(Intercept)                   -1.077e+01  3.664e-01 -29.392  < 2e-16 ***
`termLong Term`               -4.257e-01  2.366e-02 -17.990  < 2e-16 ***
annual_income                  4.772e-07  1.903e-08  25.069  < 2e-16 ***
`home_ownershipHome Mortgage` -8.362e-02  2.279e-01  -0.367 0.713707    
`home_ownershipOwn Home`      -1.869e-01  2.293e-01  -0.815 0.415043    
home_ownershipRent            -3.373e-01  2.278e-01  -1.481 0.138638    
monthly_debt                  -1.170e-05  1.044e-06 -11.204  < 2e-16 ***
years_of_credit_history        2.073e-03  1.391e-03   1.490 0.136153    
number_of_open_accounts       -2.937e-03  2.004e-03  -1.465 0.142808    
number_of_credit_problems      2.643e-03  2.373e-02   0.111 0.911329    
tax_liens                     -1.572e-01  4.267e-02  -3.685 0.000229 ***
newcs                          1.647e-02  3.991e-04  41.279  < 2e-16 ***
CLA                           -3.399e-07  6.623e-08  -5.132 2.86e-07 ***
`current_job_year0-4 `        -3.766e-03  2.327e-02  -0.162 0.871442    
`current_job_year5-9`         -1.138e-01  2.346e-02  -4.851 1.23e-06 ***
Loan_PurposeBusiness          -5.222e-02  3.310e-02  -1.578 0.114678    
Loan_PurposeHome              -3.366e-02  4.890e-02  -0.688 0.491221    
Loan_PurposePersonal           8.568e-03  6.204e-02   0.138 0.890148    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77287  on 65598  degrees of freedom
Residual deviance: 72187  on 65581  degrees of freedom
AIC: 72223

Number of Fisher Scoring iterations: 5
```

Backward Model:

```
> backwardmodel = step (full,scope=list(upper = full, lower=~1),direction = "backward", trace = FALSE)
There were 50 or more warnings (use warnings() to see the first 50)
> summary(backwardmodel)

Call:
glm(formula = loan_status ~ `termLong Term` + annual_income +
    `home_ownershipOwn Home` + home_ownershipRent + monthly_debt +
    tax_liens + newcs + CLA + `current_job_year5-9` + Loan_PurposeBusiness,
    family = binomial(), data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -1.1185   0.6598   0.7813   2.0001

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -1.088e+01  2.885e-01 -37.696  < 2e-16 ***
`termLong Term`         -4.257e-01  2.365e-02 -18.004  < 2e-16 ***
annual_income            4.801e-07  1.897e-08  25.317  < 2e-16 ***
`home_ownershipOwn Home` -1.025e-01  3.317e-02  -3.091  0.00199 **
home_ownershipRent      -2.551e-01  1.996e-02 -12.782  < 2e-16 ***
monthly_debt            -1.206e-05  9.723e-07 -12.409  < 2e-16 ***
tax_liens               -1.539e-01  3.409e-02  -4.514 6.35e-06 ***
newcs                    1.650e-02  3.960e-04  41.681  < 2e-16 ***
CLA                     -3.407e-07  6.563e-08  -5.191 2.09e-07 ***
`current_job_year5-9`   -1.117e-01  1.954e-02  -5.713 1.11e-08 ***
Loan_PurposeBusiness    -4.494e-02  2.440e-02  -1.842  0.06551 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77287  on 65598  degrees of freedom
Residual deviance: 72191  on 65588  degrees of freedom
AIC: 72213

Number of Fisher Scoring iterations: 5
```

Best Subset Model

```
> bestsubsetmodel=step(base,scope=list(upper = full, lower=~1),direction = "both", trace = FALSE)
There were 50 or more warnings (use warnings() to see the first 50)
> summary(bestsubsetmodel)

Call:
glm(formula = loan_status ~ newcs + annual_income + `termLong Term` +
    monthly_debt + home_ownershipRent + CLA + `current_job_year5-9` +
    tax_liens + `home_ownershipOwn Home` + Loan_PurposeBusiness,
    family = binomial(), data = train.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -1.1185   0.6598   0.7813   2.0001

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)              -1.088e+01  2.885e-01 -37.696  < 2e-16 ***
newcs                     1.650e-02  3.960e-04  41.681  < 2e-16 ***
annual_income             4.801e-07  1.897e-08  25.317  < 2e-16 ***
`termLong Term`          -4.257e-01  2.365e-02 -18.004  < 2e-16 ***
monthly_debt             -1.206e-05  9.723e-07 -12.409  < 2e-16 ***
home_ownershipRent       -2.551e-01  1.996e-02 -12.782  < 2e-16 ***
CLA                      -3.407e-07  6.563e-08  -5.191 2.09e-07 ***
`current_job_year5-9`    -1.117e-01  1.954e-02  -5.713 1.11e-08 ***
tax_liens                -1.539e-01  3.409e-02  -4.514 6.35e-06 ***
`home_ownershipOwn Home` -1.025e-01  3.317e-02  -3.091  0.00199 **
Loan_PurposeBusiness     -4.494e-02  2.440e-02  -1.842  0.06551 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 77287  on 65598  degrees of freedom
Residual deviance: 72191  on 65588  degrees of freedom
AIC: 72213

Number of Fisher Scoring iterations: 5
```

Best subset model is the one who is having the least AIC score. Based on the AIC scores of the above models, we found that the best subset model is best among the above.

```
> prob1=predict(bestsubsetmodel, type="response", newdata=test.data)
>
> for(i in 1:length(prob1)){
+   if(prob1[i]>0.5)
+ {
+     prob1[i]=1
+   }else{
+     prob1[i]=0
+   }
+ }
>
> accuracy(test.label,prob1)
[1] 0.7290244
```

```
> prob=predict(bestsubsetmodel, type="response", newdata=test.data)
>
>
> for(i in 1:length(prob)){
+   if(prob[i]>0.4)
+ {
+     prob[i]=1
+   }else{
+     prob[i]=0
+   }
+ }
>
>
> accuracy(test.label,prob)
[1] 0.7315854
```

```
> prob2=predict(bestsubsetmodel, type="response", newdata=test.data)
>
> for(i in 1:length(prob2)){
+   if(prob2[i]>0.6)
+ {
+     prob2[i]=1
+   }else{
+     prob2[i]=0
+   }
+ }
>
> accuracy(test.label,prob2)
[1] 0.7256707
```

```
> prob3=predict(bestsubsetmodel, type="response", newdata=test.data
>
> for(i in 1:length(prob3)){
+   if(prob3[i]>0.3)
+ {
+     prob3[i]=1
+   }else{
+     prob3[i]=0
+   }
+ }
>
> accuracy(test.label,prob3)
[1] 0.7296951
```

```
> accuracy_1<-table(test.data$loan_status,prob>0.4)
>
> accuracy_1

    FALSE  TRUE
  0   277  4170
  1   232 11721
```

Hold-out evaluation, we produce probabilities of the model and choose cut-off value as 0.4 for calculating the accuracy.

Accuracy= 73.15

K-Nearest Neighbor Model

Here, running the model which different K values like 5, 15, 20 ,21 and we found out that K value = 21 get the best accuracy.

Accuracy= 77.14%

```
> knn.5<- knn(train.data,test.data,cl=lss,k=5)
> accuracy(knn.5,ls)
[1] 0.7542683
> knn.15<- knn(train.data,test.data,cl=lss,k=15)
> accuracy(knn.15,ls)
[1] 0.770122
> knn.20<- knn(train.data,test.data,cl=lss,k=20)
> accuracy(knn.20,ls)
[1] 0.7704878
```

```
> knn.21<- knn(train.data,test.data,cl=lss,k=21)
> accuracy(knn.21,ls)
[1] 0.7714634
```

5.3. Findings

We can summarize our findings as:

- From the box plot and the hypothesis testing, we can conclude that the means of credit score of Charged Off and Fully Paid people are different.
- We built the Logistic Regression, where the AIC score of the best subset model is 72213.
- From the K-NN demonstrate with k esteem as 21, we got the most elevated exactness.

6. Conclusions and Future Work

6.1. Conclusions

- Based on the Charged off VS Fully paid box-plot it is evident that the medians are the same, but there's a distinction in fluctuation and the mean value of the Fully paid is more prominent than the Charged off individuals, which makes a difference that banks can take educated choices on future endorsing credits or loan approvals.
- By utilizing calculated relapse, we found that the subset show is best when compared with the forward and backward models.
- KNN model is more accurate than the logistic regression.

6.2. Limitations

- There are many x-variables having less co-relation with the y-variable.
- If the data would have the more x-variables like dependents and their financial data, which will increase the accuracy in the model and predictions.
- There is data redundancy in the dataset, where an x-variable has more than 50% of null values, which forced us to drop that variable.

6.3. Potential Improvements or Future Work

We can also perform Naive Bayes, Random Forest, Decision Tree which may result in higher accuracy of the Bank Loan defaulter's prediction.