BANK LOAN PREDICTION

PRESENTED BY :

TEAM 1

MENTOR :

RAJASEKHAR

DATE : 15/03/2020

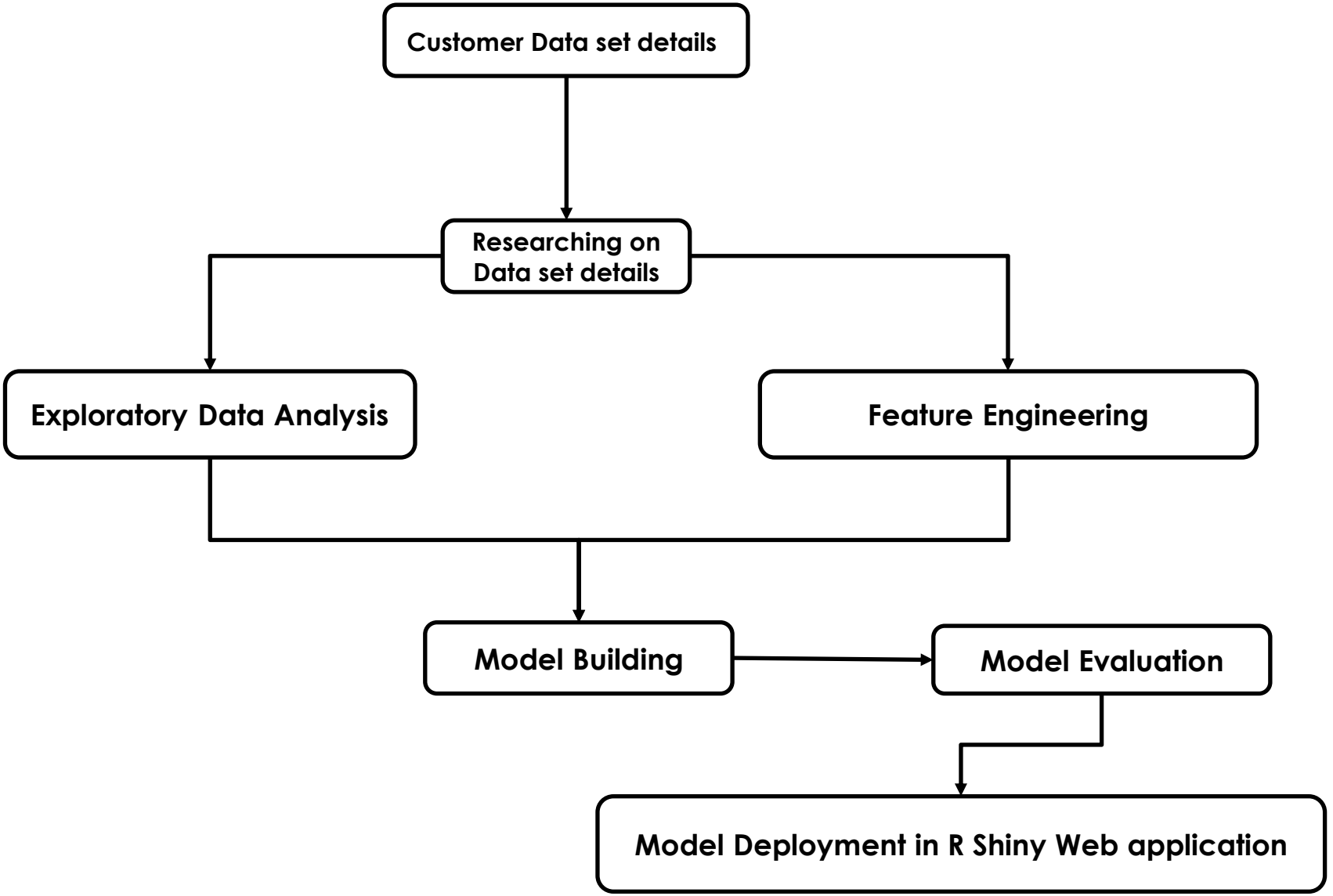
Business Problem:

The customers have taken loans from the banks to start up business if there will clear the loan amount to the banks or not.

Objective:

To predict the Whether the customer will fall under default or not.

Project Architecture / Project Flow



EXCELR
Raising Excellence

Exploratory Data Analysis (EDA) and Feature Engineering

Data set details

The given dataset is a "Bank_final.csv" .The dataset is having 26 input variables where the dependent variable is to predict customer is a default or not. This dataset is having 150000 observations.

Date ranges :

Data set Details	1 Week
EDA	1 ½ week
Model Building	1 Week
Model Evaluation	1 ½ week
Feedback	
Deployment	10 days
Final presentation	1 day

Missing values(NA's) : 11,0951 (2.73 % Missing values in a Percentage in the whole data frame)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	Name	City	State	Zip	Bank	BankStat	CCSC	ApprovalDat	Approval	Term	NoEmp	NewExis	CreateJo	Retained	Franchis	UrbanRu	RevLineC	LowDoc	ChgOffDate	DisbursementDa	DisbursementGro	BalanceGro	MIS_Statu	ChgOffPrinG	GrAppv	SBA_Appv
2	ABC HOBBYCRAFT	EVANSVILLE	IN	47711	FIFTH THIRD BANK	OH	451120	28-Feb-97	1997	84	4	2	0	0	1	0	N	Y		28-Feb-99	$60,000.00	$0.00	P I F	$0.00	$60,000.00	$48,000.00
3	LANDMARK BAR & GRILLE (THE)	NEW PARIS	IN	46526	1ST SOURCE BANK	IN	722410	28-Feb-97	1997	60	2	2	0	0	1	0	N	Y		31-May-97	$40,000.00	$0.00	P I F	$0.00	$40,000.00	$32,000.00
4	WHITLOCK DDS, TODD M.	BLOOMINGTON	IN	47401	GRANT COUNTY STATE BANK	IN	621210	28-Feb-97	1997	180	7	1	0	0	1	0	N	N		31-Dec-97	$287,000.00	$0.00	P I F	$0.00	$287,000.00	$215,250.00
5	BIG BUCKS PAWN & JEWELRY, LLC	BROKEN ARROW	OK	74012	1ST NATL BK & TR CO OF BROKE	OK	0	28-Feb-97	1997	60	2	1	0	0	1	0	N	Y		30-Jun-97	$35,000.00	$0.00	P I F	$0.00	$35,000.00	$28,000.00
6	ANASTASIA CONFECTIONS, INC.	ORLANDO	FL	32801	FLORIDA BUS. DEVEL CORP	FL	0	28-Feb-97	1997	240	14	1	7	7	1	0	N	N		14-May-97	$229,000.00	$0.00	P I F	$0.00	$229,000.00	$229,000.00
7	B&T SCREW MACHINE COMPANY, I	PLAINVILLE	CT	6062	TD BANK, NATIONAL ASSOCIATI	DE	332721	28-Feb-97	1997	120	19	1	0	0	1	0	N	N		30-Jun-97	$517,000.00	$0.00	P I F	$0.00	$517,000.00	$387,750.00
8	MIDDLE ATLANTIC SPORTS CO INC	UNION	NJ	7083	WELLS FARGO BANK NATL ASS	SD	0	02-Jun-80	1980	45	45	2	0	0	0	0	N	N	24-Jun-91	22-Jul-80	$600,000.00	$0.00	CHGOFF	$208,959.00	$600,000.00	$499,998.00
9	WEAVER PRODUCTS	SUMMERFIELD	FL	34491	REGIONS BANK	AL	811118	28-Feb-97	1997	84	1	2	0	0	1	0	N	Y		30-Jun-98	$45,000.00	$0.00	P I F	$0.00	$45,000.00	$36,000.00
10	TURTLE BEACH INN	PORT SAINT JOE	FL	32456	CENTENNIAL BANK	FL	721310	28-Feb-97	1997	297	2	2	0	0	1	0	N	N		31-Jul-97	$305,000.00	$0.00	P I F	$0.00	$305,000.00	$228,750.00

Exploratory Data Analysis (EDA)

EXCELR
Raising Excellence

- *MIS_Status variable as the dependent variable. Where CHGOFF is default on loan i.e. Charged off and PIF is paid in full i.e. no default and CHGOFF having 26% and PIF having 74% in the whole dataset.*

- *The maximum amount of customers loan applied from the Bank Of America NATL ASSOC, state CA (California) and the city Los Angeles.*

- *Basic summary of the variables given below.*

```
                   ï..Name                City           State           Zip                                 Bank          BankState
SUBWAY                 :    78   LOS ANGELES:  1902   CA      :19314   Min.   :    0   BANK OF AMERICA NATL ASSOC     :28023   NC      :26847
SCHLOTZSKY'S DELI      :    39   NEW YORK   :  1583   NY      :12236   1st Qu.:20854   CITIZENS BANK NATL ASSOC       :13052   RI      :14731
QUIZNO'S CLASSIC SUBS:    37   MIAMI      :  1515   TX      : 9822   Median :48053   CAPITAL ONE NATL ASSOC         :10611   IL      :12672
DOMINO'S PIZZA         :    36   CHICAGO    :  1401   FL      : 9269   Mean   :49849   JPMORGAN CHASE BANK NATL ASSOC:10381   CA      :12449
DAIRY QUEEN            :    32   HOUSTON    :  1351   OH      : 7560   3rd Qu.:80003   WELLS FARGO BANK NATL ASSOC    : 6373   VA      :11564
(Other)                :149773   (Other)    :142246   (Other):91796   Max.   :99999   (Other)                        :81412   (Other):71588
NA's                   :     4   NA's       :     1   NA's   :     2                   NA's                           :  147   NA's   :  148
      CCSC            ApprovalDate       ApprovalFY          Term              NoEmp              NewExist        CreateJob           RetainedJob
Min.   :     0   30-Sep-97:   466   Min.   :1962   Min.   :  0.00   Min.   :   0.000   Min.   :0.00   Min.   :   0.000   Min.   :   0.000
1st Qu.:236118   17-Mar-06:   425   1st Qu.:1998   1st Qu.: 57.00   1st Qu.:   2.000   1st Qu.:1.00   1st Qu.:   0.000   1st Qu.:   0.000
Median :447110   24-Mar-06:   422   Median :2005   Median : 84.00   Median :   4.000   Median :1.00   Median :   0.000   Median :   1.000
Mean   :401568   01-Apr-97:   418   Mean   :2002   Mean   : 93.01   Mean   :   9.314   Mean   :1.32   Mean   :   1.278   Mean   :   3.686
3rd Qu.:561612   31-Mar-06:   412   3rd Qu.:2006   3rd Qu.: 84.00   3rd Qu.:   8.000   3rd Qu.:2.00   3rd Qu.:   0.000   3rd Qu.:   4.000
Max.   :928120   18-Apr-05:   402   Max.   :2007   Max.   :480.00   Max.   :9999.000   Max.   :2.00   Max.   :3000.000   Max.   :9500.000
                 (Other)  :147454
```

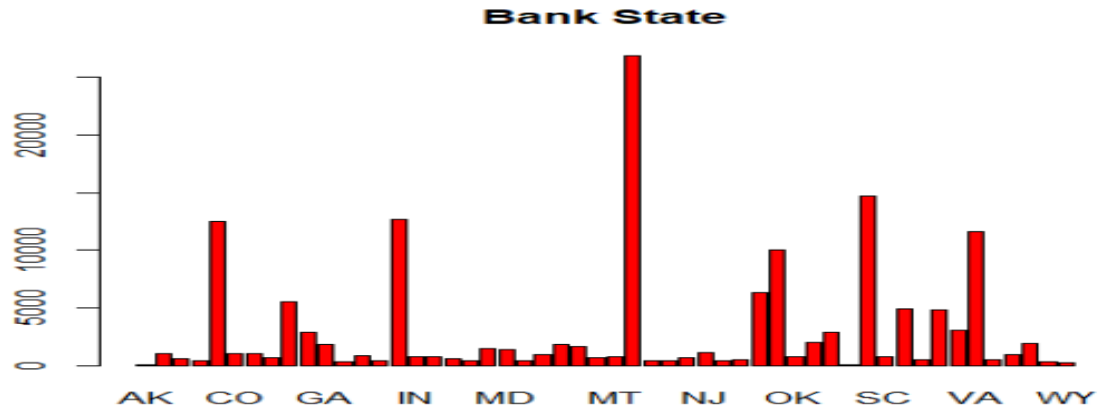
FranchiseCode	UrbanRural	RevLineCr	LowDoc	ChgOffDate	DisbursementDate	DisbursementGross
Min. : 0	Min. :0.0000	N :71611	1: 1	13-Mar-10: 201	31-May-06: 6220	$50,000.00 : 10214
1st Qu.: 0	1st Qu.:0.0000	Y :49881	C: 83	30-Jan-10: 183	31-Mar-06: 6039	$25,000.00 : 6992
Median : 1	Median :1.0000	0 :23659	N:137871	20-Feb-10: 175	30-Apr-06: 5848	$100,000.00 : 6978
Mean : 1656	Mean :0.7679	T : 4819	Y: 12044	06-Feb-10: 158	30-Jun-06: 5263	$10,000.00 : 5612
3rd Qu.: 1	3rd Qu.:1.0000	1 : 3		06-Mar-10: 148	28-Feb-06: 4902	$35,000.00 : 3092
Max. :91999	Max. :2.0000	(Other): 3		(Other) : 39601	(Other) :121502	$20,000.00 : 2748
		NA's : 23		NA's :109533	NA's : 225	(Other) :114363

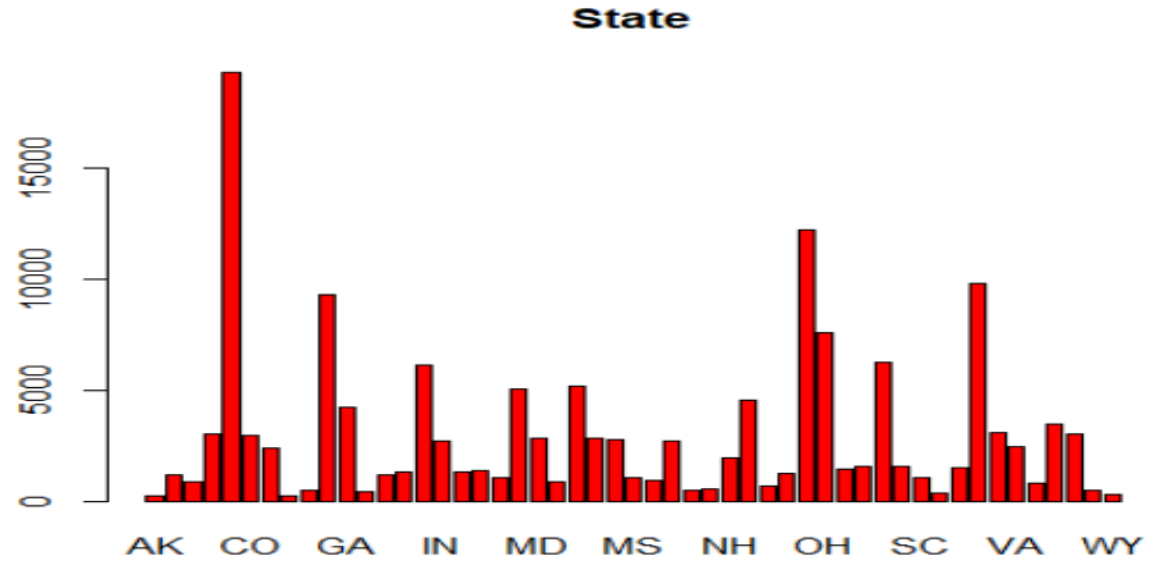
BalanceGross	MIS_Status	ChgOffPrinGr	GrAppv	SBA_Appv
$0.00 :149997	CHGOFF: 39008	$0.00 :109702	$50,000.00 :17500	$25,000.00 :15063
$12,750.00 : 1	P I F :110123	$10,000.00 : 956	$25,000.00 :13268	$5,000.00 :11900
$827,875.00 : 1	NA's : 868	$50,000.00 : 795	$10,000.00 :12588	$12,500.00 :10836
		$25,000.00 : 437	$100,000.00 :11123	$50,000.00 : 7764
		$100,000.00 : 400	$20,000.00 : 5451	$10,000.00 : 4666
		$35,000.00 : 266	$35,000.00 : 5067	$17,500.00 : 4507
		(Other) : 37443	(Other) :85002	(Other) :95263

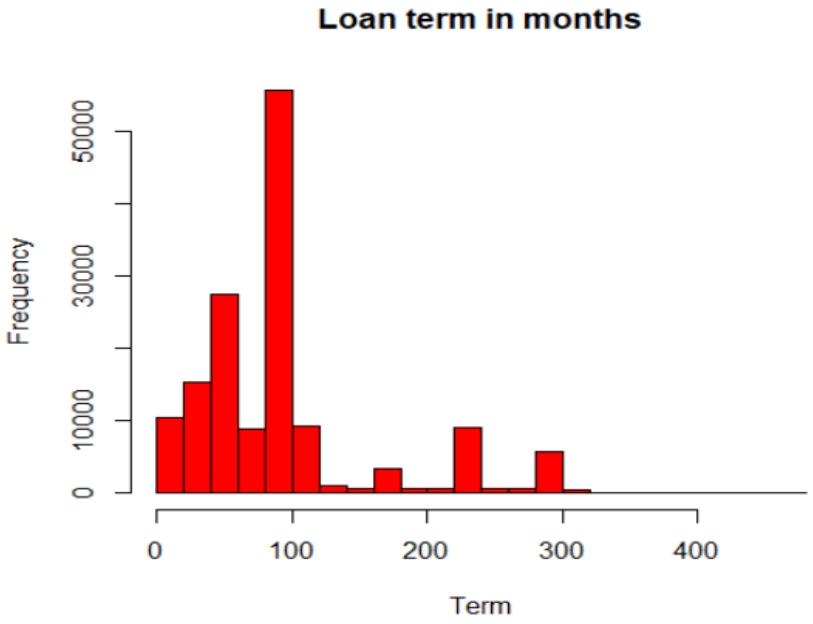
EDA

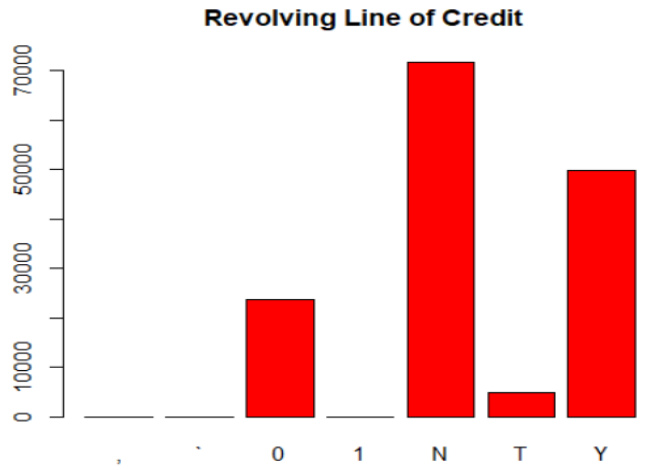
EXCELR
Raising Excellence

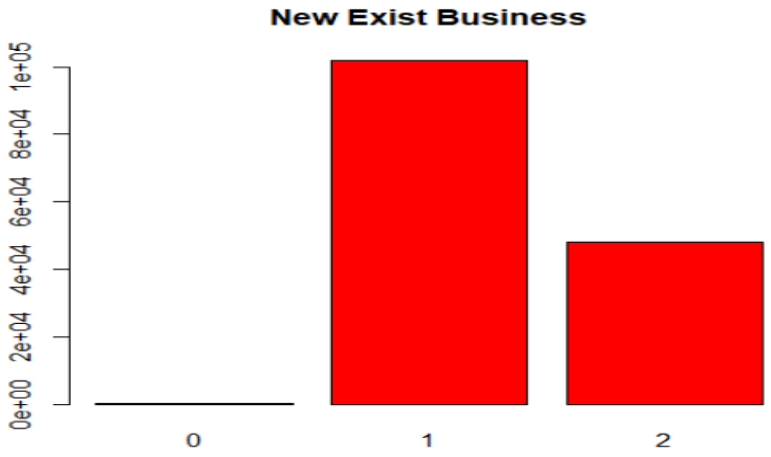
Basic Plots :

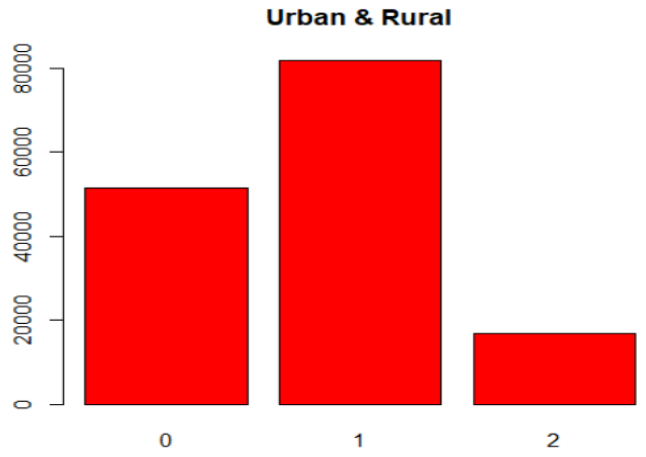


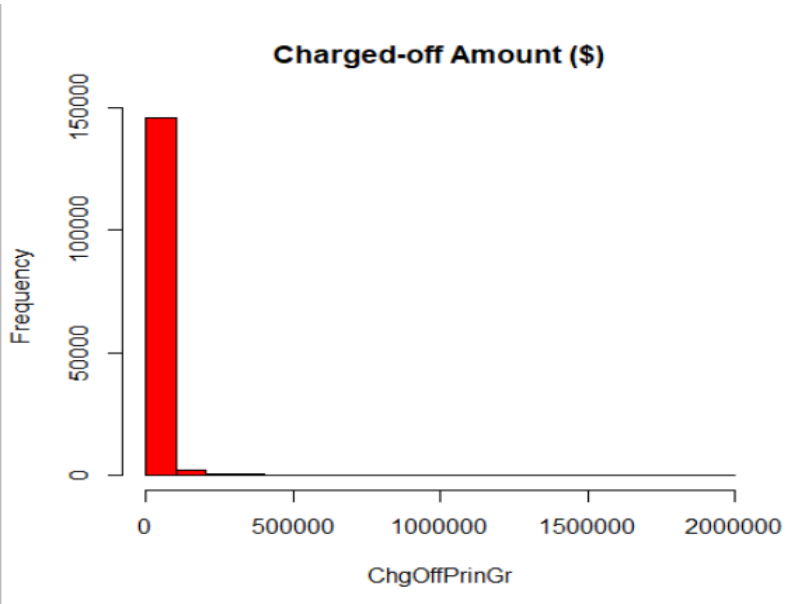


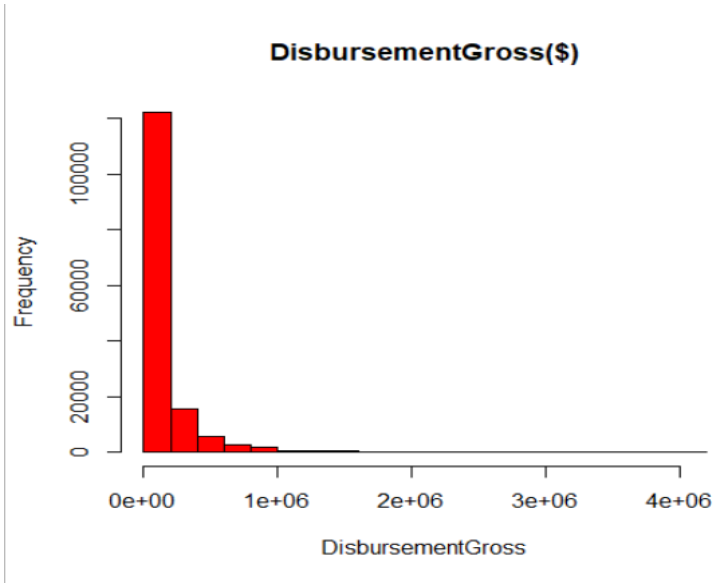

Loan term in months
Frequency
0
10000
30000
50000
0
100
200
300
400
Term

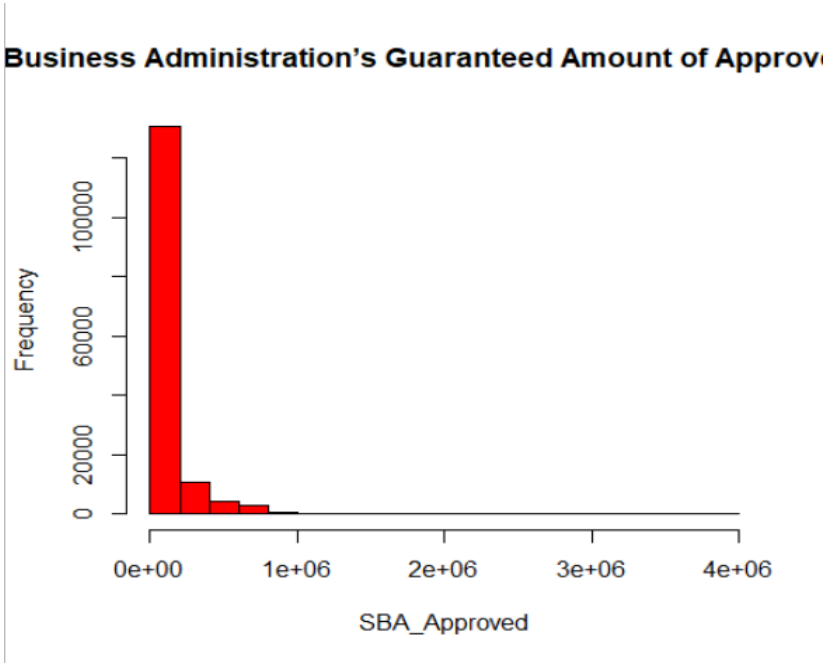

Revolving Line of Credit
0
10000
30000
50000
70000
,
`
0
1
N
T
Y

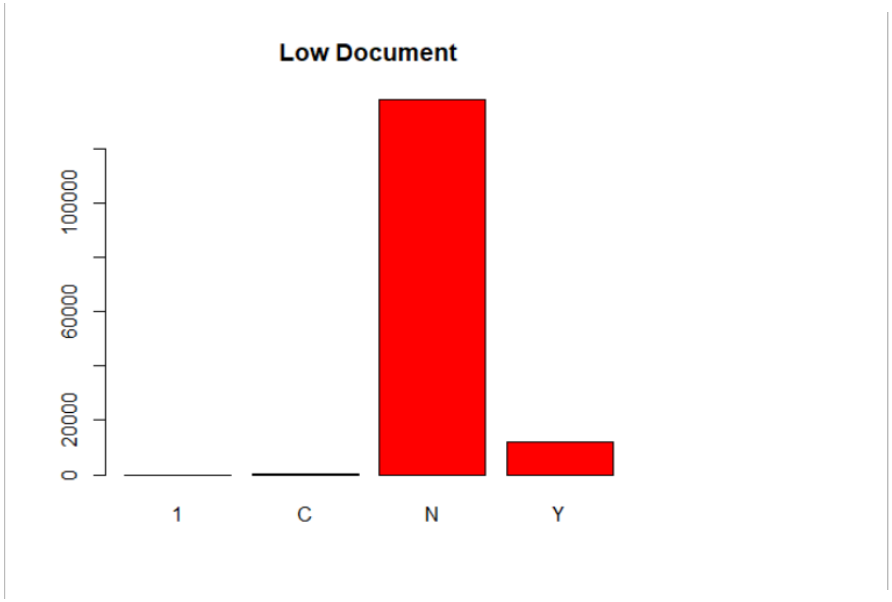

New Exist Business
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
0
1
2

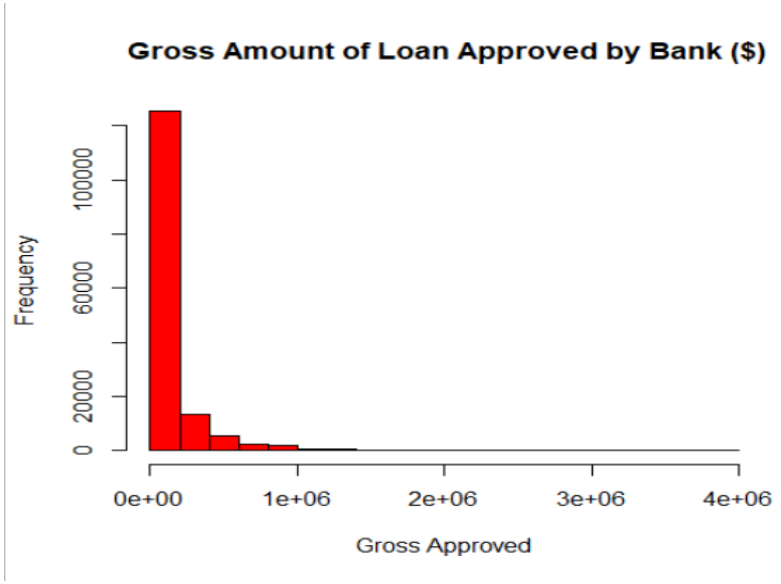

Urban & Rural
0
20000
40000
60000
80000
0
1
2

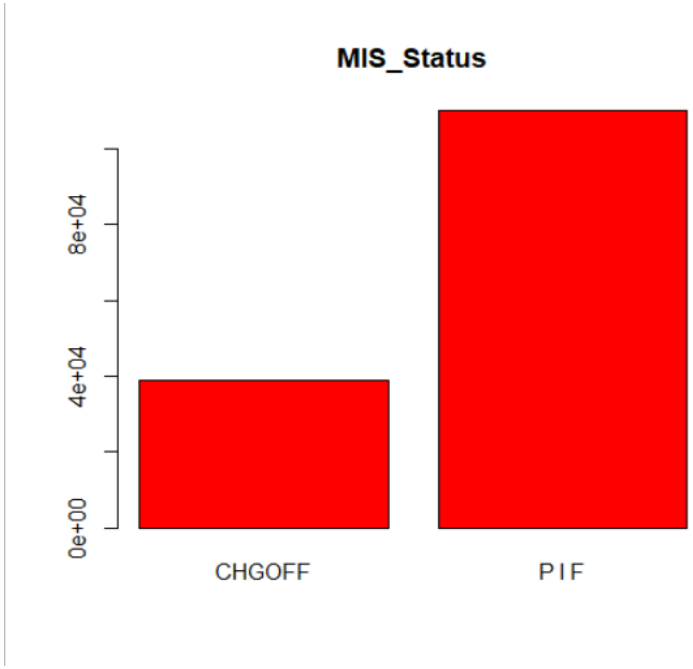

Charged-off Amount ($)
Frequency
0
50000
100000
150000
0
500000
1000000
1500000
2000000
ChgOffPrinGr


DisbursementGross($)
Frequency
0
20000
60000
100000
0e+00
1e+06
2e+06
3e+06
4e+06
DisbursementGross


Business Administration's Guaranteed Amount of Approv
Frequency
0
20000
60000
100000
0e+00
1e+06
2e+06
3e+06
4e+06
SBA_Approved


Low Document
0
20000
60000
100000
1
C
N
Y


Gross Amount of Loan Approved by Bank ($)
Frequency
0
20000
60000
100000
0e+00
1e+06
2e+06
3e+06
4e+06
Gross Approved


MIS_Status
0e+00
4e+04
8e+04
CHGOFF
PIF

Feature Engineering

EXCELR
Raising Excellence

- Few variables such as – Name, City, State, Zip, Bank, Bank State, CCSC, Approval Date, Approval FY, ChgOff Date, Disbursement Date, Balance Gross were eliminated due to considering Weight of Evidence and Information Value they variables were irrelevant for prediction.
- Variables such as – Gross Approval, SBA Approval, Charge off amount, Disbursement Gross were changed from currency format to numeric format.
- If there having any Franchise considered as (0) and No Franchise code as (1).
- There having any existing business considered as (1) and New business as (2).
- In MIS Status charge off considered as (0) and paid in full as (1).
- In low document process considered as LowDoc (1) and No lowDoc as (0).
- In there Revolving Line of Credit considered as yes (1) and No as (0).
- Out of dictionary values of few variables were changed accordingly.
- To avoid overfitting ChgOffPrinGr variable was eliminated.

Model Building

Template for Model results presentation

EXCELR
Raising Excellence

Model – Logistic Regression

Data set details

After eliminating some variable in the dataset there having 14 input variables 150000 observations.

Data Partition details

Randomly dataset divided into 70% training dataset and 30% testing dataset.

Algorithms

Logistic Regression is a classification **algorithm**. It is used to predict a binary outcome (1 / 0, Yes / No, True / False) given a set of independent variables.

Algorithm details and configuration

```
model <- glm(train$MIS_Status~.,data=train,family =
"binomial“)
```

glm = function for logistic regression.
Y = MIS_Status variable.
Data = Training dataset.
Family = Binomial object are used.

Confusion Matrix Details :

Accuracy : : 0.8314 (83.14%)
95% CI : (0.8279, 0.8348)

Kappa : 0.5288
Sensitivity : 0.9229
Specificity : 0.5699
Pos Pred Value : 0.8598
Neg Pred Value : 0.7211
Prevalence : 0.7408
Detection Rate : 0.6837

Detection Prevalence : 0.7951
Balanced Accuracy : 0.7464

Model Predictions

EXCELR
Raising Excellence

- The model builds with train data and after that predicted with test data to know the performance of the model.

- Other models were tried but their results were not better.

Input data fed to the model

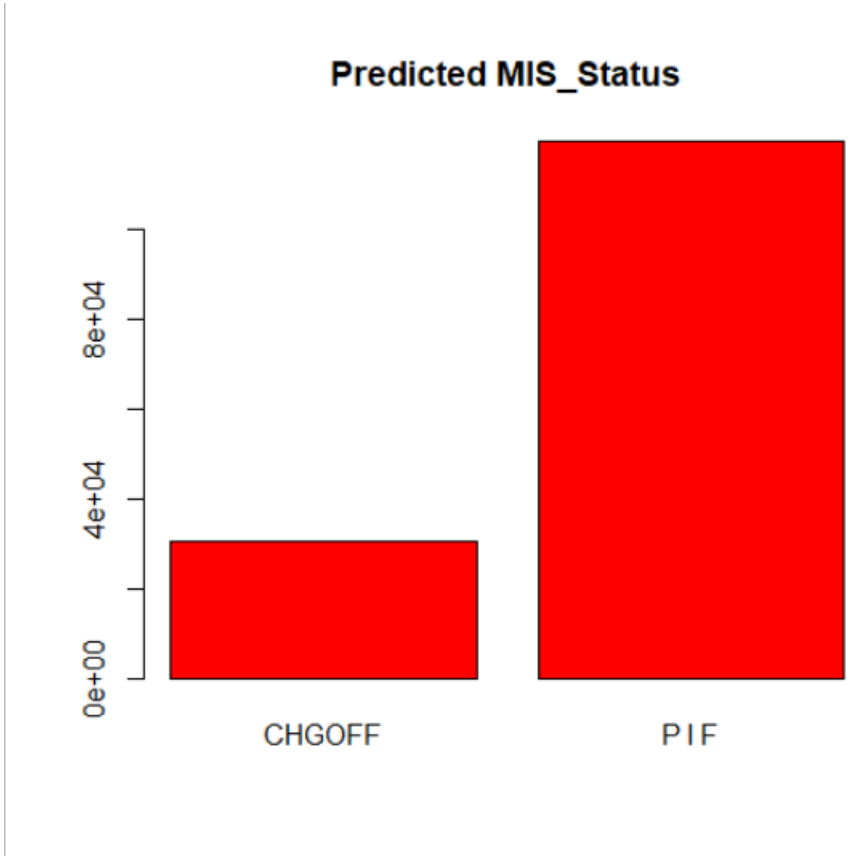
Predicted probabilities for each row

	Term	NoEmp	NewExist	CreateJob	RetainedJob	FranchiseCode	UrbanRural	RevLineCr	LowDoc	DisbursementGross	MIS_Status	GrAppv	SBA_Appv	prob	pred_values	pred_MIS_Status
2	60	2	1	0	0	0	0	0	1	35000	0	35000	28000	0.103688974	0	P I F
3	240	14	1	7	7	0	0	0	0	229000	0	229000	229000	0.000343376	0	P I F
4	84	1	2	0	0	0	0	0	1	45000	0	45000	36000	0.035451029	0	P I F
5	84	1	2	0	0	0	0	0	1	70000	0	70000	56000	0.037840385	0	P I F
6	60	5	1	0	0	0	0	0	1	70000	0	70000	56000	0.111401392	0	P I F
7	60	16	1	0	0	0	0	0	1	1.00E+05	0	1.00E+05	80000	0.11424821	0	P I F
8	300	12	2	0	0	0	0	0	0	615000	0	615000	461250	3.60E-05	0	P I F
9	144	90	1	0	0	0	0	0	0	1250000	0	1250000	937500	0.093465556	0	P I F
10	240	1	2	5	0	0	0	0	0	291000	0	291000	291000	0.000427733	0	P I F

Model Results

EXCELR
Raising Excellence

Model generated the output for the customers loan data set in which 79.51% customers are genuine(0)non default paid in full(P I F) and 20.48% Chargeoff (CHGOFF) are default (1) customers.



Model Deployment using R shiny

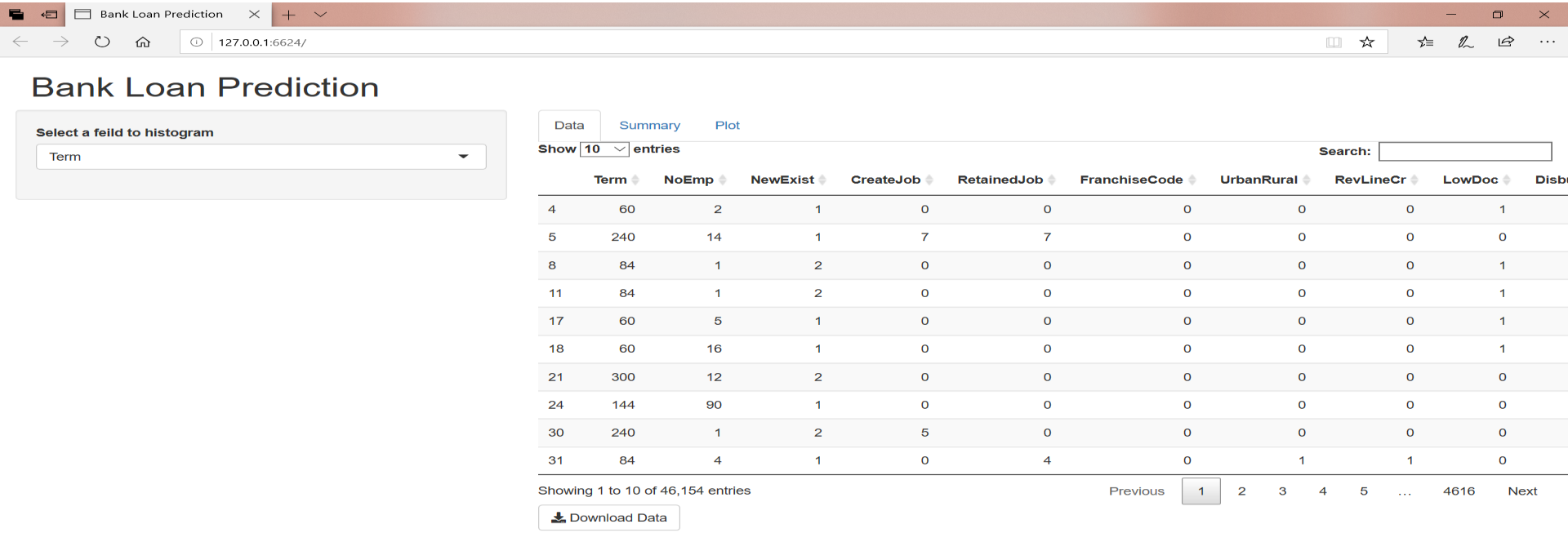
Model Deployment using R shiny

- The model deployment was accomplished through Rshiny.

- We are using three tab panel, they are –

1. Data
2. Summary
3. Plot

- The bank loan customer data were invoked into Data tab panel which shown below. Their display total dataset and at the end there having a Download Data button.



- Click on Download Data button they're downloaded only predicted MIS Status in with CVS format.

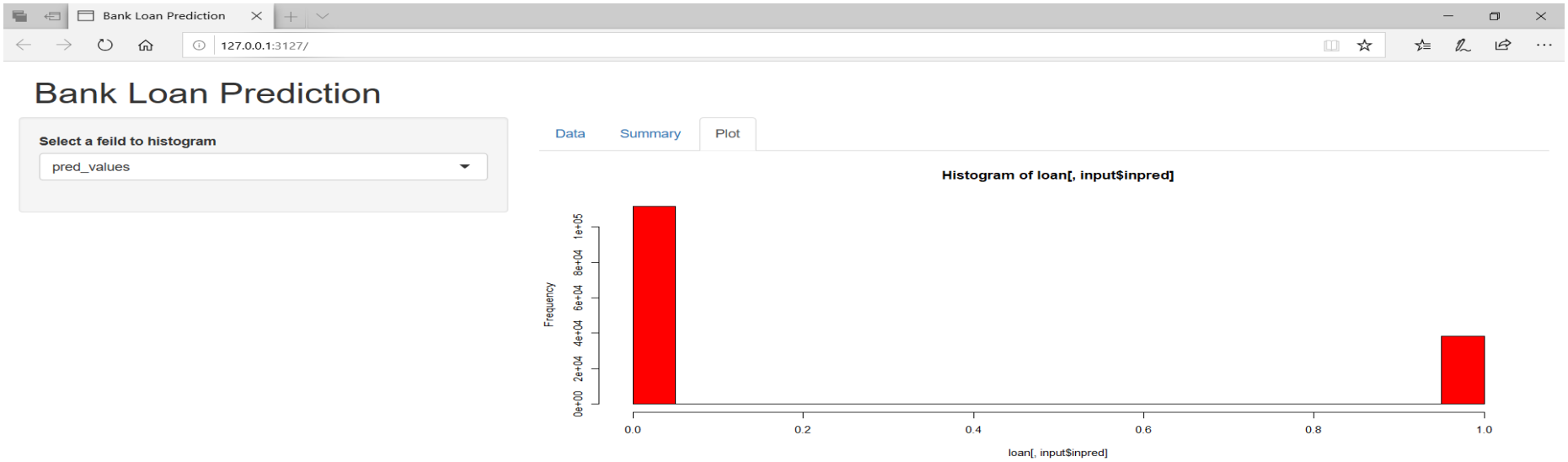
index	pred_MIS_Status
1	P I F
2	P I F
3	P I F
4	P I F
5	P I F
6	P I F
7	P I F
8	P I F
9	P I F
10	P I F
11	P I F
12	P I F

- The customer bank loan data summary were shown in second tab panel which given below.

Data | Summary | Plot

```
      Term             NoEmp           NewExist       CreateJob         RetainedJob       FranchiseCode      UrbanRural       RevLineCr
 Min.   :  0.00   Min.   :   0.000   Min.   :1.00   Min.   :   0.000   Min.   :   0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.: 57.00   1st Qu.:   2.000   1st Qu.:1.00   1st Qu.:   0.000   1st Qu.:   0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median : 84.00   Median :   4.000   Median :1.00   Median :   0.000   Median :   1.000   Median :0.0000   Median :1.0000   Median :0.0000
 Mean   : 93.01   Mean   :   9.314   Mean   :1.32   Mean   :   1.278   Mean   :   3.686   Mean   :0.4238   Mean   :0.7679   Mean   :0.3325
 3rd Qu.: 84.00   3rd Qu.:   8.000   3rd Qu.:2.00   3rd Qu.:   0.000   3rd Qu.:   4.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :480.00   Max.   :9999.000   Max.   :2.00   Max.   :3000.000   Max.   :9500.000   Max.   :1.0000   Max.   :2.0000   Max.   :1.0000
     LowDoc         DisbursementGross MIS_Status  ChgOffPrinGr          GrAppv            SBA_Appv             prob            pred_values
 Min.   :0.00000   Min.   :      0   0:110991   Min.   :      0   Min.   :    200   Min.   :    100   Min.   :0.000000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:  30000   1: 39008   1st Qu.:      0   1st Qu.:  25000   1st Qu.:  12500   1st Qu.:0.002912   1st Qu.:0.0000
 Median :0.00000   Median :  65500              Median :      0   Median :  50000   Median :  25000   Median :0.018936   Median :0.0000
 Mean   :0.08029   Mean   : 141294              Mean   :  15057   Mean   : 127692   Mean   :  92500   Mean   :0.260336   Mean   :0.2551
 3rd Qu.:0.00000   3rd Qu.: 150000              3rd Qu.:   6003   3rd Qu.: 120000   3rd Qu.:  80000   3rd Qu.:0.584585   3rd Qu.:1.0000
 Max.   :1.00000   Max.   :4029520              Max.   :1999999   Max.   :4000000   Max.   :4000000   Max.   :1.000000   Max.   :1.0000
 default_pred
 Length:149999
 Class :character
 Mode  :character
```

- Their can plot histogram with selecting each variable in the customer bank loan data and it was shown in the third tab panel.



Challenges faced?

- Understanding the dataset it takes a lot of time for each variable.
- Selecting the right algorithm with the best results and it plays a key role in the project.
- New to model deployment part with the R shiny web app is too difficult to deployment process.

How did you overcome?

- Researching on more part to understand the dataset of each variable.
- Try to build model with many algorithms, but the logistic regression model having the best performance.
- New to this part of model deployment with the R shiny web app so it's taken more time to deploy and help of mentor and online tutorial deployment part has been completed.

Conclusion

EDA process was performed on the given bank full dataset of csv format. Some variables were dropped based on the Weight of Evidence and Information Value. The dataset was split into the train and test datasets and the model builds with logistic regression with train dataset and prediction model were developed with test dataset to predict the customers loan default status. The logistic regression model was selected based on accuracy, sensitivity and specificity which has better than other models. The Model deployment was achieved through R shiny.

Thank you