Predictions on the Eligibility of Loans for Borrowers

Sai Charan Vedula

15/06/2022

Abstract

A variety of products are available for the banking system to provide, but their main source of revenue is the credit lines. Banks always profit through the interest they earn on these loans. The profit or loss of any given bank is affected by the banks by either the borrower's default or repaying any loans that they have borrowed. As a result, the Non-Performing Assets in any given bank through forecasting loan defaulters reduces. Therefore, there need to be further investigations into this occurrence since there is benefit maximization with reliable and effective forecasts. The logistic regression model, a predictive analytics tool, provides the benefit of detecting loan defaulters most reliably and conveniently. Therefore, the dataset from the Kaggle.com website is obtained for a successful forecast. Various performance indicators were calculated by the utilization of the Logistic Regression model. The sensitivity and specificity performance metrics are used to compare the models.

Additionally, in the case of checking the account details, the logistic model is of importance as it involves the demographic variables like age, loan amount, loan status, and purpose that are essential when determining the probability of loan default or how risky a borrower is when acquiring any given loan. With the utilization of the model, it is possible to easily identify the most suitable clients the bank can target and issue loans by evaluating their loan default's plausibility. From the model, it is clear that there needs to be an assessment of other attributes associated with the creditor as they play a significant role in making decisions on credit lines and predicting loan defaulters apart from giving loans only to wealthy borrowers.

Introduction

A banking system can sell various products, but one of its main sources of income is the provision of loans to potential borrowers. By lending out money to their potential borrowers, they can make profits through the interest earned from these loans. However, their profit or loss is affected by whether the borrowers repay or default on their loans. To minimize their Non-Performing Assets, the banking system can perform predictions or forecasts for loan defaulters. Therefore, there needs to be a system that makes it easy for the bank to predict the defaulters and non-defaulters borrowers. Therefore, this project implements a predictive analytics tool, Logistic Regression models, that will be essential in calculating various performance measures important when giving out loans. With the model, it will be possible to target the right customers for loans. It will be possible to ensure that banks not only give loans to wealthy borrowers but there should be a consideration of various characteristics of any potential borrowers that will help in making predictions on loan defaulters and credit decisions. ## Literature Review

In a study by Sheikh et al. (2020), the authors utilized data from various banks for customers who had their loans approved through a consideration of set criteria. A machine learning model was trained to ensure that they obtained accurate results. Their main purpose in this study was to forecast the safety of loans that the bank had given to their customers. Therefore, they utilized a logistic regression model to predict loan safety. In their study, it was identified by the authors that the various parameters that needed to be taken into account when crediting a loan are Customer Assets, Business Value, and credit history. In another study by Vaidya (2017, the author's main focus is the logistic regression technique and how this can be represented mathematically. In actualizing the probabilistic and predictive techniques to a given loan approval prediction problem, the author employs the machine learning method, logistic regression, in the study. In their study, Zhang et al. (2019) presented a logistic regression analysis to determine several classification thresholds on imbalanced datasets. Arun et al. (2016) have illustrated how banks can reduce the risk factor in picking a safe individual for crediting loans to ensure they save money and time. The authors used the technique of mining Big Data from previous records of customers who have been credited with loans from which a machine learning model is taught how to obtain accurate results.

Theory

Data

The dataset that have been used for this project is retrieved from <https://www.kaggle.com/datasets/vipin20/loan-application-data>. This data includes:

## Rows: 500

## Columns: 15

## $ X <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15…

## $ Loan\_ID <fct> LP001002, LP001003, LP001005, LP001006, LP001008, LP…

## $ Gender <fct> Male, Male, Male, Male, Male, Male, Male, Male, Male…

## $ Married <fct> No, Yes, Yes, Yes, No, Yes, Yes, Yes, Yes, Yes, Yes,…

## $ Dependents <fct> 0, 1, 0, 0, 0, 2, 0, 3+, 2, 1, 2, 2, 2, 0, 2, 0, 1, …

## $ Education <fct> Graduate, Graduate, Graduate, Not Graduate, Graduate…

## $ Self\_Employed <fct> No, No, Yes, No, No, Yes, No, No, No, No, No, , No, …

## $ ApplicantIncome <int> 5849, 4583, 3000, 2583, 6000, 5417, 2333, 3036, 4006…

## $ CoapplicantIncome <dbl> 0, 1508, 0, 2358, 0, 4196, 1516, 2504, 1526, 10968, …

## $ LoanAmount <dbl> NA, 128, 66, 120, 141, 267, 95, 158, 168, 349, 70, 1…

## $ Loan\_Amount\_Term <dbl> 360, 360, 360, 360, 360, 360, 360, 360, 360, 360, 36…

## $ Credit\_History <dbl> 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, NA, …

## $ Property\_Area <fct> Urban, Rural, Urban, Urban, Urban, Urban, Urban, Sem…

## $ Loan\_Status <fct> Y, N, Y, Y, Y, Y, Y, N, Y, N, Y, Y, Y, N, Y, Y, Y, N…

## $ Total\_Income <fct> $5849.0, $6091.0, $3000.0, $4941.0, $6000.0, $9613.0…

The Credit\_History variable has been changed to factor from numerical.

loans\_data$Credit\_History <- as.factor(loans\_data$Credit\_History)

There are unnecessary columns that need to be removed. This include the Loan\_ID and X columns.

loans\_data <- loans\_data[-c(1:2,15)]

From the dataset, it is depicted that the Total\_Income variable has the $ sign and this makes it not to be considered as numerical value. Therefore, we delete this column and create a new one by adding together the values for CoapplicantIncome and ApplicantIncome variables to convert it to numerical values.

loans\_data$Total\_Income <- rowSums(loans\_data[,c("ApplicantIncome", "CoapplicantIncome")])

To avoid redundancy, the CoapplicantIncome and ApplicantIncome variables have been removed.

loans\_data <- loans\_data[-c(6:7)]

## 'data.frame': 500 obs. of 11 variables:

## $ Gender : Factor w/ 3 levels "","Female","Male": 3 3 3 3 3 3 3 3 3 3 ...

## $ Married : Factor w/ 3 levels "","No","Yes": 2 3 3 3 2 3 3 3 3 3 ...

## $ Dependents : Factor w/ 5 levels "","0","1","2",..: 2 3 2 2 2 4 2 5 4 3 ...

## $ Education : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 1 2 1 1 2 1 1 1 ...

## $ Self\_Employed : Factor w/ 3 levels "","No","Yes": 2 2 3 2 2 3 2 2 2 2 ...

## $ LoanAmount : num NA 128 66 120 141 267 95 158 168 349 ...

## $ Loan\_Amount\_Term: num 360 360 360 360 360 360 360 360 360 360 ...

## $ Credit\_History : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 1 2 2 ...

## $ Property\_Area : Factor w/ 3 levels "Rural","Semiurban",..: 3 1 3 3 3 3 3 2 3 2 ...

## $ Loan\_Status : Factor w/ 2 levels "N","Y": 2 1 2 2 2 2 2 1 2 1 ...

## $ Total\_Income : num 5849 6091 3000 4941 6000 ...

The next step involves checking if there are any missing values in the dataset.

anyNA(loans\_data)

## [1] TRUE

There are missing values in the dataset, and therefore needs to check in each column for the missing values.

colSums(is.na(loans\_data))

## Gender Married Dependents Education

## 0 0 0 0

## Self\_Employed LoanAmount Loan\_Amount\_Term Credit\_History

## 0 18 14 41

## Property\_Area Loan\_Status Total\_Income

## 0 0 0

The rows with missing values are excluded.

loans\_data\_clean <- na.omit(loans\_data)

colSums(is.na(loans\_data\_clean))

## Gender Married Dependents Education

## 0 0 0 0

## Self\_Employed LoanAmount Loan\_Amount\_Term Credit\_History

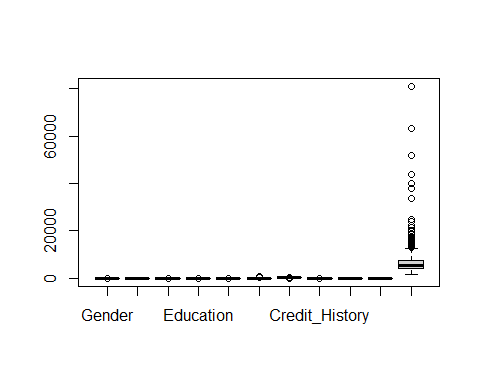
## 0 0 0 0

## Property\_Area Loan\_Status Total\_Income

## 0 0 0

Next, the dataset is checked for any outliers that may lead to bias results.

boxplot(loans\_data\_clean)



From the boxplot above, it is clear that there are outliers. Therefore, the data is cleaned up to avoid biased analysis.

loans\_data\_cleaned1 <- loans\_data\_clean[loans\_data\_clean$Total\_Income < 8000,]

summary(loans\_data\_cleaned1)

## Gender Married Dependents Education Self\_Employed

## : 4 : 2 : 8 Graduate :247 : 19

## Female: 67 No :121 0 :197 Not Graduate: 84 No :276

## Male :260 Yes:208 1 : 49 Yes: 36

## 2 : 54

## 3+: 23

##

## LoanAmount Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status

## Min. : 17.0 Min. : 36.0 0: 50 Rural : 96 N:102

## 1st Qu.: 96.0 1st Qu.:360.0 1:281 Semiurban:127 Y:229

## Median :118.0 Median :360.0 Urban :108

## Mean :117.7 Mean :345.3

## 3rd Qu.:137.5 3rd Qu.:360.0

## Max. :275.0 Max. :480.0

## Total\_Income

## Min. :1442

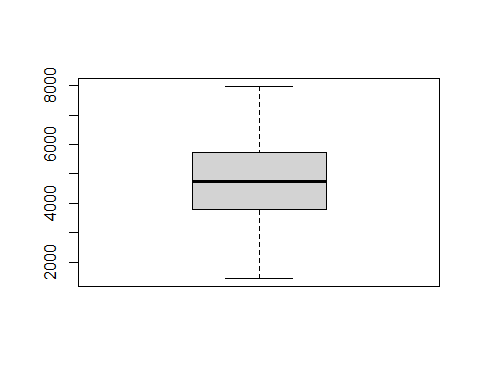
## 1st Qu.:3791

## Median :4727

## Mean :4819

## 3rd Qu.:5733

## Max. :7978

The dataset is now free from outliers, as illustrated in the boxplot below. Since our target variable is Loan\_Status, we check its proportion.

table(loans\_data\_cleaned1$Loan\_Status)

##

## N Y

## 102 229

The target variable has a ratio of 1:2, indicating that it is balanced; therefore, the dataset can be split to train and test set. The dataset has been split into 80% training set and 20% test set.

RNGkind(sample.kind = "Rounding")

## Warning in RNGkind(sample.kind = "Rounding"): non-uniform 'Rounding' sampler

## used

set.seed(123)

sets <- sample(x = nrow(loans\_data\_cleaned1),

size = nrow(loans\_data\_cleaned1)\*0.8)

#splitting

train\_loan <- loans\_data\_cleaned1[sets,]

test\_loan <- loans\_data\_cleaned1[-sets,]

Methodology

Having cleaned and pre-processed the dataset, the next setion involved the building of the model. The model include the logistic regression model which is created using the glm() function.

model\_1 <- glm(formula = Loan\_Status ~ 1, data = train\_loan, family = binomial)

model\_2 <- glm(formula = Loan\_Status~., data = train\_loan, family = binomial)

summary(model\_2)

##

## Call:

## glm(formula = Loan\_Status ~ ., family = binomial, data = train\_loan)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.2305 -0.3414 0.3899 0.6598 2.3865

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) 1.221e+01 1.018e+03 0.012 0.99044

## GenderFemale 8.635e-01 1.466e+00 0.589 0.55579

## GenderMale 1.219e+00 1.405e+00 0.868 0.38556

## MarriedNo -1.319e+01 1.018e+03 -0.013 0.98967

## MarriedYes -1.313e+01 1.018e+03 -0.013 0.98971

## Dependents0 -1.097e+00 1.743e+00 -0.630 0.52898

## Dependents1 -1.908e+00 1.791e+00 -1.065 0.28680

## Dependents2 -1.019e+00 1.788e+00 -0.570 0.56863

## Dependents3+ -1.046e+00 1.831e+00 -0.571 0.56770

## EducationNot Graduate -5.373e-01 3.894e-01 -1.380 0.16761

## Self\_EmployedNo -2.984e-02 7.558e-01 -0.039 0.96851

## Self\_EmployedYes -2.702e-01 9.074e-01 -0.298 0.76587

## LoanAmount -5.070e-03 5.537e-03 -0.916 0.35983

## Loan\_Amount\_Term -6.502e-03 3.798e-03 -1.712 0.08688 .

## Credit\_History1 4.045e+00 6.107e-01 6.624 3.49e-11 \*\*\*

## Property\_AreaSemiurban 1.326e+00 4.579e-01 2.897 0.00377 \*\*

## Property\_AreaUrban -3.592e-02 4.433e-01 -0.081 0.93542

## Total\_Income 2.243e-04 1.632e-04 1.375 0.16920

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 325.53 on 263 degrees of freedom

## Residual deviance: 216.97 on 246 degrees of freedom

## AIC: 252.97

##

## Number of Fisher Scoring iterations: 14

Next step involve performing the feature selection of the created model

model\_stepwise <- step(object = model\_1,

direction = "both",

scope = list(upper = model\_2),

trace = FALSE)

summary(model\_stepwise)

##

## Call:

## glm(formula = Loan\_Status ~ Credit\_History + Property\_Area +

## Loan\_Amount\_Term, family = binomial, data = train\_loan)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.2974 -0.3309 0.4516 0.7913 2.4455

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -0.940051 1.272418 -0.739 0.46003

## Credit\_History1 3.878285 0.580768 6.678 2.42e-11 \*\*\*

## Property\_AreaSemiurban 1.292314 0.427928 3.020 0.00253 \*\*

## Property\_AreaUrban 0.061087 0.394078 0.155 0.87681

## Loan\_Amount\_Term -0.005552 0.003510 -1.582 0.11375

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 325.53 on 263 degrees of freedom

## Residual deviance: 227.52 on 259 degrees of freedom

## AIC: 237.52

##

## Number of Fisher Scoring iterations: 5

Results

Having created the logistic regression model, there were various insights identified. First, the model identified that the LoanAmount and Total\_Income are the determinant factors for loan approval.

optimum\_model <- glm(formula = Loan\_Status ~ Credit\_History + Property\_Area + Loan\_Amount\_Term + LoanAmount + Total\_Income, family = binomial, data = train\_loan)

summary(optimum\_model)

##

## Call:

## glm(formula = Loan\_Status ~ Credit\_History + Property\_Area +

## Loan\_Amount\_Term + LoanAmount + Total\_Income, family = binomial,

## data = train\_loan)

##

## Deviance Residuals:

## Min 1Q Median 3Q Max

## -2.2945 -0.3506 0.4517 0.7518 2.4694

##

## Coefficients:

## Estimate Std. Error z value Pr(>|z|)

## (Intercept) -1.6853291 1.4628950 -1.152 0.249

## Credit\_History1 3.9014176 0.5808989 6.716 1.87e-11 \*\*\*

## Property\_AreaSemiurban 1.3438571 0.4349054 3.090 0.002 \*\*

## Property\_AreaUrban 0.0783288 0.4055971 0.193 0.847

## Loan\_Amount\_Term -0.0049702 0.0035436 -1.403 0.161

## LoanAmount -0.0047831 0.0052063 -0.919 0.358

## Total\_Income 0.0002229 0.0001500 1.486 0.137

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## (Dispersion parameter for binomial family taken to be 1)

##

## Null deviance: 325.53 on 263 degrees of freedom

## Residual deviance: 225.25 on 257 degrees of freedom

## AIC: 239.25

##

## Number of Fisher Scoring iterations: 5

Having the optimum\_model the predictions on the test data for loan eligibility was possible and the results of the model are illustrated.

test\_loan$Eligible <- predict(optimum\_model, type = "response", newdata =test\_loan)

head(test\_loan)

## Gender Married Dependents Education Self\_Employed LoanAmount

## 4 Male Yes 0 Not Graduate No 120

## 5 Male No 0 Graduate No 141

## 7 Male Yes 0 Not Graduate No 95

## 9 Male Yes 2 Graduate No 168

## 12 Male Yes 2 Graduate 109

## 14 Male No 0 Graduate No 114

## Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status Total\_Income

## 4 360 1 Urban Y 4941

## 5 360 1 Urban Y 6000

## 7 360 1 Urban Y 3849

## 9 360 1 Urban Y 5532

## 12 360 1 Urban Y 4340

## 14 360 1 Rural N 4693

## Eligible

## 4 0.7374255

## 5 0.7628301

## 7 0.7127500

## 9 0.7180456

## 12 0.7213703

## 14 0.7166116

Atthreshold for this model was set to 0.5, where if the eligibility prediction is greater than 0.5 the borrower is eligible and if the eligibility probability is less than 0.5 the the customer is not eligible for loan.

test\_loan$eligibility\_label <- ifelse(test = test\_loan$Eligible > 0.5,

yes = "Y",

no = "N")

test\_loan

## Gender Married Dependents Education Self\_Employed LoanAmount

## 4 Male Yes 0 Not Graduate No 120

## 5 Male No 0 Graduate No 141

## 7 Male Yes 0 Not Graduate No 95

## 9 Male Yes 2 Graduate No 168

## 12 Male Yes 2 Graduate 109

## 14 Male No 0 Graduate No 114

## 15 Male Yes 2 Graduate No 17

## 16 Male No 0 Graduate No 125

## 24 Yes 2 Not Graduate No 112

## 33 Male No 1 Graduate Yes 106

## 34 Male Yes 0 Graduate No 114

## 38 Female Yes 0 Graduate No 144

## 41 Male No 0 Graduate No 80

## 70 Female No 0 Graduate No 136

## 71 Male Yes 0 Graduate No 172

## 76 Male No 0 Graduate No 113

## 109 Male Yes 2 Graduate No 216

## 137 Female Yes 0 Graduate No 84

## 142 Male No 0 Graduate No 168

## 148 Male Yes 1 Graduate No 30

## 152 Male Yes 0 Graduate Yes 152

## 154 Male Yes 2 Not Graduate No 113

## 155 Male No 0 Graduate No 50

## 169 Male No 0 Graduate No 63

## 177 Male Yes 2 Graduate No 101

## 194 Male No 0 Graduate No 76

## 209 Male No 0 Graduate No 59

## 214 Male Yes 3+ Not Graduate Yes 130

## 227 Male Yes Not Graduate Yes 138

## 231 Male Yes 1 Graduate No 104

## 239 Female No 1 Graduate No 112

## 250 Male Yes 0 Graduate No 90

## 251 Male Yes 0 Not Graduate No 201

## 262 Male No 0 Graduate No 134

## 263 Female No 1 Graduate No 155

## 266 Male No 0 Graduate No 151

## 269 Female No 0 Graduate 135

## 270 Female No 1 Graduate No 90

## 286 Male No 0 Graduate No 89

## 289 Female No 0 Graduate No 115

## 292 Male Yes 2 Graduate No 127

## 298 Female Yes 1 Graduate No 135

## 299 Female No 0 Graduate No 151

## 312 Male No 0 Not Graduate No 111

## 328 Male Yes 0 Graduate Yes 184

## 329 Female Yes 0 Graduate No 110

## 331 Male No 1 Graduate No 117

## 337 Male Yes 1 Graduate 160

## 342 Female No 0 Graduate No 46

## 346 Male Yes 0 Graduate No 160

## 372 Male Yes 2 Graduate No 155

## 374 Male No 1 Graduate No 111

## 381 Male Yes 0 Graduate 128

## 384 Male No 0 Graduate Yes 128

## 386 Male No 1 Graduate 113

## 389 Male Yes 0 Graduate No 136

## 399 Male No 0 Not Graduate No 109

## 416 Female No 0 Graduate No 60

## 417 Female No 1 Graduate No 160

## 447 Male Yes 2 Not Graduate No 110

## 458 Male Yes 0 Graduate No 173

## 463 Male Yes 0 Graduate No 153

## 473 Male Yes 3+ Graduate No 100

## 483 Male Yes 0 Graduate No 128

## 489 Male Yes 2 Graduate Yes 160

## 496 Female Yes 1 Graduate No 105

## 499 Male Yes 1 Graduate Yes 95

## Loan\_Amount\_Term Credit\_History Property\_Area Loan\_Status Total\_Income

## 4 360 1 Urban Y 4941

## 5 360 1 Urban Y 6000

## 7 360 1 Urban Y 3849

## 9 360 1 Urban Y 5532

## 12 360 1 Urban Y 4340

## 14 360 1 Rural N 4693

## 15 120 1 Urban Y 2385

## 16 360 1 Urban Y 4950

## 24 360 0 Rural N 5282

## 33 360 1 Rural N 4692

## 34 360 1 Semiurban Y 5167

## 38 360 1 Semiurban Y 5126

## 41 360 1 Urban N 3600

## 70 360 0 Semiurban N 4300

## 71 360 1 Urban Y 6274

## 76 480 1 Urban N 3750

## 109 360 0 Urban N 7400

## 137 360 1 Rural N 4583

## 142 360 1 Urban Y 5417

## 148 360 1 Urban Y 2963

## 152 360 1 Rural Y 6327

## 154 360 1 Rural N 2281

## 155 360 1 Urban Y 3254

## 169 480 0 Semiurban N 2237

## 177 360 1 Rural Y 3664

## 194 360 1 Semiurban Y 3858

## 209 360 1 Urban Y 2479

## 214 360 1 Rural Y 5703

## 227 360 1 Urban N 4735

## 231 360 1 Semiurban Y 4545

## 239 360 1 Rural Y 3812

## 250 360 1 Urban Y 3677

## 251 360 0 Semiurban N 5558

## 262 360 1 Semiurban Y 4269

## 263 36 1 Semiurban N 3481

## 266 360 1 Rural Y 7542

## 269 360 1 Rural N 3418

## 270 360 1 Urban Y 4436

## 286 360 1 Rural Y 6211

## 289 360 1 Semiurban Y 4124

## 292 360 0 Semiurban N 4400

## 298 360 1 Urban Y 4666

## 299 480 1 Rural N 7541

## 312 360 1 Semiurban Y 5332

## 328 360 1 Semiurban Y 7978

## 329 360 1 Urban N 6784

## 331 360 1 Urban Y 6177

## 337 360 1 Rural Y 5938

## 342 360 1 Rural N 2378

## 346 360 1 Semiurban Y 4957

## 372 360 1 Semiurban Y 5185

## 374 180 0 Urban N 5049

## 381 360 1 Semiurban Y 5833

## 384 360 1 Urban Y 7167

## 386 180 1 Urban Y 3667

## 389 360 1 Urban Y 4750

## 399 360 1 Rural Y 5568

## 416 360 1 Urban Y 2995

## 417 360 1 Urban N 2600

## 447 360 1 Rural Y 4652

## 458 360 1 Urban N 6277

## 463 360 1 Rural Y 5203

## 473 360 1 Semiurban Y 4691

## 483 360 1 Semiurban Y 5233

## 489 360 1 Semiurban Y 6666

## 496 84 1 Semiurban Y 4239

## 499 360 1 Semiurban Y 2895

## Eligible eligibility\_label

## 4 0.73742550 Y

## 5 0.76283009 Y

## 7 0.71275004 Y

## 9 0.71804559 Y

## 12 0.72137027 Y

## 14 0.71661160 Y

## 15 0.89551437 Y

## 16 0.73316123 Y

## 24 0.05557539 N

## 33 0.72427307 Y

## 34 0.91507498 Y

## 38 0.90243734 Y

## 41 0.71606329 Y

## 70 0.13911246 N

## 71 0.74669503 Y

## 76 0.55087096 Y

## 109 0.05842543 N

## 137 0.74013827 Y

## 142 0.71282655 Y

## 148 0.73539538 Y

## 152 0.75216840 Y

## 154 0.59744081 Y

## 155 0.72936033 Y

## 169 0.07379371 N

## 177 0.68146448 Y

## 194 0.90612316 Y

## 209 0.68472409 Y

## 214 0.74580003 Y

## 227 0.71107539 Y

## 231 0.90774537 Y

## 239 0.67719008 Y

## 250 0.70978727 Y

## 251 0.13550369 N

## 262 0.88908139 Y

## 263 0.96818934 Y

## 266 0.79992776 Y

## 269 0.63252073 Y

## 270 0.74336802 Y

## 286 0.79990286 Y

## 289 0.89472566 Y

## 292 0.14712365 N

## 298 0.71086327 Y

## 299 0.68765897 Y

## 312 0.91896512 Y

## 328 0.93517554 Y

## 329 0.81627403 Y

## 331 0.78959900 Y

## 337 0.72814465 Y

## 342 0.67631854 Y

## 346 0.89191001 Y

## 372 0.89890736 Y

## 374 0.12931488 N

## 381 0.92119819 Y

## 384 0.81616647 Y

## 386 0.84247296 Y

## 389 0.71372047 Y

## 399 0.75890527 Y

## 416 0.70802516 Y

## 417 0.57918970 Y

## 447 0.71863650 Y

## 458 0.74591605 Y

## 463 0.70158655 Y

## 473 0.91198291 Y

## 483 0.91092539 Y

## 489 0.92353483 Y

## 496 0.97301679 Y

## 499 0.87671028 Y

The perfomance of the logistic regression model in predicting the eligibility of loans is pretty good. IT has a prediction accuracy rate of 79.1%.

test\_loan$eligibility\_label <- as.factor(test\_loan$eligibility\_label)

model\_evaluation <- confusionMatrix(data = test\_loan$eligibility\_label,

reference = test\_loan$Loan\_Status,

positive = "Y")

model\_evaluation

## Confusion Matrix and Statistics

##

## Reference

## Prediction N Y

## N 7 0

## Y 14 46

##

## Accuracy : 0.791

## 95% CI : (0.6743, 0.8808)

## No Information Rate : 0.6866

## P-Value [Acc > NIR] : 0.039911

##

## Kappa : 0.4071

##

## Mcnemar's Test P-Value: 0.000512

##

## Sensitivity : 1.0000

## Specificity : 0.3333

## Pos Pred Value : 0.7667

## Neg Pred Value : 1.0000

## Prevalence : 0.6866

## Detection Rate : 0.6866

## Detection Prevalence : 0.8955

## Balanced Accuracy : 0.6667

##

## 'Positive' Class : Y

##

Conclusion

From the analysis and prediction of this project, it has been identified that the logistic regression model is pretty good in predicting the loan eligibility of bank customers.

References

Arun, K., Ishan, G., & Sanmeet, K. (2016). Loan approval prediction based on machine learning approach. IOSR J. Comput. Eng, 18(3), 18-21.

Sheikh, M. A., Goel, A. K., & Kumar, T. (2020, July). An approach for prediction of loan approval using machine learning algorithm. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 490-494). IEEE.

Zhang, H., Li, Z., Shahriar, H., Tao, L., Bhattacharya, P., & Qian, Y. (2019, July). Improving prediction accuracy for logistic regression on imbalanced datasets. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC) (Vol. 1, pp. 918-919). IEEE.

Vaidya, A. (2017, July). Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval. In 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE.