# LOAN PREDICTION USING LOGISTIC REGRESSION

## INTRODUCTION:

Banks and financial institutions play a very important role in stabilizing our nation's economy.They facilitate lending borrowing mechanism to ensure financial stability.Since, banks play a big role in ensuring financial stability they must ensure that they lend/supply deposits with utmost care.Here, we implement a machine learning technique to predict the basis on which customers can be sanctioned a loan.Several machine learning techniques can be leveraged to predict the loan status of a customer.Most popular techniques are Linear Regression, Logistic Regression, Support Vector Machine etc.

## POTENTIAL CUSTOMERS/CLIENTS:

Banks and financial institutions.

## PROBLEM STATEMENT:

Predict if a customer can be sanctioned loan based on his personal information.

## DATA SET:

| Variable Name | Description | Type | Example |
|---|---|---|---|
| Loan ID | Gives unique identification number | Character | AX1234 |
| Gender | Male or Female | Character | Male |
| Dependents | Number of person(s) dependent on the client/customer | Character | 3+ |
| Education | Whether the customer is literate/illiterate | Character | Yes |
| Self Employed | Whether is employed or self Employed | Character | Yes |
| Applicant's Income | Income/Remuneration of the client | Integer | 342242(Rs) |
| Co Applicant's Income | Income/Remuneration of client's family member | Integer | 235211(Rs) |
| Loan Amount | Loan Amount to be sanctioned | Integer | 232421(Rs) |
| Loan Amount Term | Duration(Time) sanctioned | Integer | 360 |
| Credit History | Guidelines met earlier | Integer | 1 |
| Property Area | Type of human settlement area | Character | Urban |
| Loan Status | Approval/Rejection of loan(Yes/No) | Character | Y |

## METHODLOGY:

Basically any machine learning technique follows 5 basic steps

**DATA COLLECTION**

- collect data from known source

**DATA PREPROCESSING**

-Feature Scaling:- Normalize data

-Feature Engineering- Add/Delete columns

-Handling Missing Values/Complex data.

-Splitting Train/Test data

**MODEL BUILDING**

-Identifying correct model for the given case.

SVM/Linear/Logistic/Naive Bayes/KNN

**MODEL PREDICTION**

-Predict the given model given the explanatory variables

**SUMMARY AND EVALUATION**

- Summarize data using confusion matrix to check for accuracy

## ASSUMPTIONS:

1. Only these 12 variables are considered for the prediction.

2. Logistic model is perfect for the given scenario.

## DATA COLLECTION:

Here, for sample lets collect data for 12 samples and consider first 9 columns

| Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount |
|---|---|---|---|---|---|---|---|---|
| LP001002 | Male | No | 0 | Graduate | No | 5849 | 0 | NA |
| LP001003 | Male | Yes | 1 | Graduate | No | 4583 | 1508 | 128 |
| LP001005 | Male | Yes | 0 | Graduate | Yes | 3000 | 0 | 66 |
| LP001006 | Male | Yes | 0 | Not Graduate | No | 2583 | 2358 | 120 |
| LP001008 | Male | No | 0 | Graduate | No | 6000 | 0 | 141 |
| LP001011 | Male | Yes | 2 | Graduate | Yes | 5417 | 4196 | 267 |
| LP001013 | Male | Yes | 0 | Not Graduate | No | 2333 | 1516 | 95 |
| LP001014 | Male | Yes | 3+ | Graduate | No | 3036 | 2504 | 158 |
| LP001018 | Male | Yes | 2 | Graduate | No | 4006 | 1526 | 168 |
| LP001020 | Male | Yes | 1 | Graduate | No | 12841 | 10968 | 349 |
| LP001024 | Male | Yes | 2 | Graduate | No | 3200 | 700 | 70 |
| LP001027 | Male | Yes | 2 | Graduate | | 2500 | 1840 | 109 |

## DATA PREPROCESSING:

Here, we have data with Missing and NA values.We preprocess the data to make changes.This is done to ensure uniformity in data.Similarly, we have complex data's like 3+ in the dependents column.Such data's need to be handled with care.Further, we split data into train and test(80:20)

## WAYS TO HANDLE MISSING VALUES AND COMPLEX DATA:

1.Replace data with either Average or Mode values depending on the circumstance.Here we have modified ID LP001014 sample.Dependents which contain 3+ are being replaced as 4.

2. In the Loan Amount column we replace the NA values by Average value.This is done because average can give a central representation of the loan amount.

3.Another way to handle to handle missing values is to delete them.

## MODIFIED DATA:

| | Loan_ID | Gender | Married | Dependents | Education | Self_Employed | ApplicantIncome | CoapplicantIncome | LoanAmount |
|---|---|---|---|---|---|---|---|---|---|
| 1 | LP001002 | 1 | 0 | 0 | 1 | 0 | 5849 | 0 | 146.4 |
| 2 | LP001003 | 1 | 1 | 1 | 1 | 0 | 4583 | 1508 | 128.0 |
| 4 | LP001006 | 1 | 1 | 0 | 0 | 0 | 2583 | 2358 | 120.0 |
| 7 | LP001013 | 1 | 1 | 0 | 0 | 0 | 2333 | 1516 | 95.0 |
| 8 | LP001014 | 1 | 1 | 4 | 1 | 0 | 3036 | 2504 | 158.0 |
| 9 | LP001018 | 1 | 1 | 2 | 1 | 0 | 4006 | 1526 | 168.0 |
| 12 | LP001027 | 1 | 1 | 2 | 1 | 1 | 2500 | 1840 | 109.0 |
| 13 | LP001028 | 1 | 1 | 2 | 1 | 0 | 3073 | 8106 | 200.0 |
| 14 | LP001029 | 1 | 0 | 0 | 1 | 0 | 1853 | 2840 | 114.0 |
| 16 | LP001032 | 1 | 0 | 0 | 1 | 0 | 4950 | 0 | 125.0 |
| 17 | LP001034 | 1 | 0 | 1 | 0 | 0 | 3596 | 0 | 100.0 |
| 18 | LP001036 | 1 | 0 | 0 | 1 | 0 | 3510 | 0 | 76.0 |

Here we have made the data relatively cleaner because the dependents have been changed to 4 instead of 3+ which is obscure to handle.

## SAMPLE CODE:

Here, we replace NA values in Loan Amount by Average value of the column

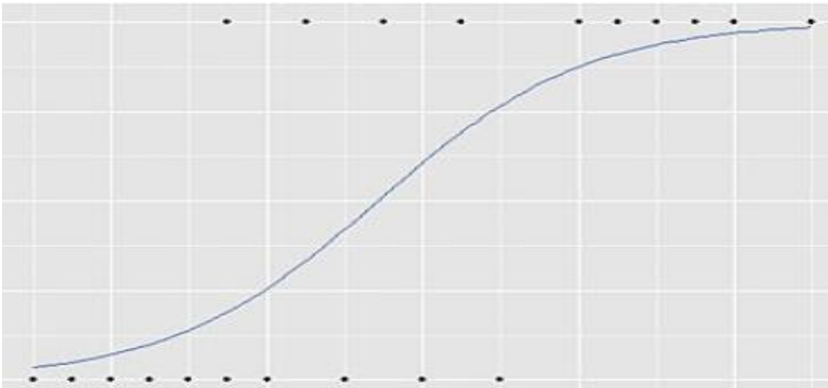
```
#2 Clean NA values in  loan Amount,Loan Amount term, credit history train
trainloan$LoanAmount = ifelse(is.na(trainloan$LoanAmount),
                              ave(trainloan$LoanAmount , FUN= function(v) mean(v, na.rm=TRUE)),
                              trainloan$LoanAmount)
```

## MODEL BUILDING:

### Logistic Regression:

Logistic Regression is a model used to predict a certain class of objects which linear regression fails to achieve.For example in rain prediction, there are only 2 classes- Yes and No.Hence, logistic regression is used in scenarios where classifying the data becomes the sole objective.

Basically logistic regression plots a Sigmoid/logit function curve.It determines the probability in which a certain class of objects lie.



$$\ell = \log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\frac{p}{1-p} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

where $\ell$ is the log-odds, b is the base of the logarithm, and βs are parameters of the model determined from the train data.

These equations are used to predict the probability of loan prediction.Consider example below for test data , where

Applicant Income= Rs. 3036

Co Applicant Income = Rs. 2504

Credit History = 0

Dependents= 4

Property Area = 1

Substituting in the above equation we get,

$\frac{p}{1-p}$ = 0.124. Hence p = 0.11.Assuming threshold of 0.5.We reject the loan sanction if p<0.5.Here the concerned person cannot be sanctioned loan because p = 0.11.This result is verified below.

## PREDICTED TEST RESULTS:

```
         4          8          9         16         17         21         33         35         38 
0.75021552 0.11588597 0.84928120 0.76478023 0.73996732 0.06732068 0.67343288 0.66848434 0.86453867 
        44         47         51         56         65         67         69         80        108 
0.86090480 0.73956040 0.86399637 0.92175647 0.12996167 0.06284721 0.75617352 0.85622751 0.70176091 
       110        114        120        129        132        133        134        142        160 
0.86230313 0.85390072 0.76376784 0.05217037 0.75310095 0.87046372 0.87037767 0.76469372 0.84659577 
       161        165        192        194        196        197        206        212        219 
0.86644181 0.76396933 0.86938234 0.87033132 0.84296270 0.67363434 0.87026740 0.12037186 0.21411096 
       228        234        236        245        250        252        255        258        265 
0.92015864 0.86981047 0.66356936 0.85210858 0.75353314 0.85600177 0.06676750 0.75652453 0.87017926 
       272        276        313        325        328        330        340        342        349 
```

Now we convert them only into discrete forms(binary) in 1's and 0's

## OBSERVATIONS:

```
Coefficients:
      (Intercept)    ApplicantIncome      Credit_History        Dependents0        Dependents1  
       -2.803e+00         -1.030e-06           3.805e+00          1.822e-01          4.754e-02  
      Dependents2        Dependents4       Property_Area0     Property_Area1  CoapplicantIncome  
        7.840e-01          1.372e-01          -3.209e-01          7.237e-01         -3.467e-05  
```

## MODEL CODE:

```
#Fit appropriate model
seat_glm = glm(Loan_Status ~ ApplicantIncome+Dependents+Property_Area+CoapplicantIncome, family = "binomia
summary(seat_glm)
```

Here, We have assumed Applicant income,Dependents and Property Area and Coapplicants to be the explanatory variables.

## RESULTS&SUMMARY:

```
       Predicted_output
Actual  0  1
     0  2 24
     1  0 96

               Accuracy : 0.8033          
                 95% CI : (0.7216, 0.8697)
    No Information Rate : 0.9836          
    P-Value [Acc > NIR] : 1               

                  Kappa : 0.1159          

 Mcnemar's Test P-Value : 2.668e-06       

            Sensitivity : 1.00000         
            Specificity : 0.80000         
```

## OPTIMIZATION:

We achieved an accuracy of 80.33%We can further optimize our model by applying few techniques like Parameter Tuning, Controlling Train/Testsplit and by changing the probability threshold.Here, we will try to optimize our model by selecting right set of explanatory variables.To achieve this, we need to find explanatory variables that impact our output significantly.This can be done by viewing summary.

```
                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)       -2.071e+00  6.145e-01  -3.371 0.000749 ***
ApplicantIncome    8.940e-06  2.508e-05   0.357 0.721462    
Credit_History     3.756e+00  4.559e-01   8.239  < 2e-16 ***
LoanAmount        -1.833e-03  1.817e-03  -1.009 0.313073    
CoapplicantIncome -2.704e-05  3.525e-05  -0.767 0.443077    
Married0          -5.457e-01  2.433e-01  -2.243 0.024912 *  
Self_Employed0    -3.151e-03  3.001e-01  -0.011 0.991622    
Property_Area0    -2.644e-01  2.769e-01  -0.955 0.339652    
Property_Area1     7.437e-01  2.979e-01   2.497 0.012537 *  
Education0        -2.791e-01  2.866e-01  -0.974 0.330160    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This table shows that Intercept,Credit_History are 99.9% significantly important to explain the dependent variable.Married and Property Area are 99 percent significantly important in explaining the dependent variable.Hence these explanatory variables cannot be excluded.

## OPTIMIZED MODEL CODE:

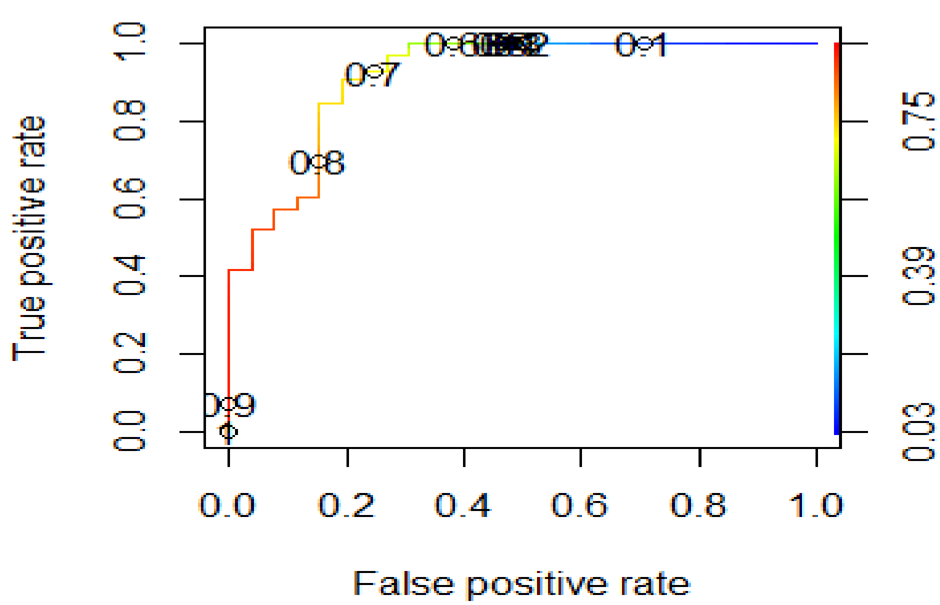
```
seat_glm = glm(Loan_Status ~ ApplicantIncome+Credit_History+LoanAmount
               +CoapplicantIncome+Married+Self_Employed+Property_Area
               +Education, family = "binomial", data = trainloan)
summary(seat_glm)

prob_pred = predict(seat_glm, type = 'response', newdata = testloan)
prob_pred
predicted = ifelse(prob_pred > 0.5,1, 0)
predicted
```

Here we have included Credit History,Married, Loan Amount and education in addition to the existing explanatory variables

## SUMMARY AND OBSERVATIONS:

```
      predicted
Actual   0  1
     0 14 12
     1  0 96

               Accuracy : 0.9016
                 95% CI : (0.8345, 0.9481)
    No Information Rate : 0.8852
    P-Value [Acc > NIR] : 0.346322

                  Kappa : 0.6474

 Mcnemar's Test P-Value : 0.001496

            Sensitivity : 1.0000
            Specificity : 0.8889
```

Here, we infer that our accuracy has been increased from 0.80 to 0.90 by tweaking with the explanatory variables. We can also tweak the Threshold value of 0.5 and try to optimize further.To tweak the threshold value, one needs to plot graph between  True Positive and Flase Positive.



## OPTIMIZED MODEL CODE BY TWEAKING THRESHOLD:

```
Predicted_output = ifelse(prob_pred > 0.55,1, 0)
Predicted_output
```

```
       Predicted_output
Actual  0  1
     0 15 11
     1  0 96

               Accuracy : 0.9098
                 95% CI : (0.8444, 0.9541)
    No Information Rate : 0.877
    P-Value [Acc > NIR] : 0.167537

                  Kappa : 0.6821

 Mcnemar's Test P-Value : 0.002569

            Sensitivity : 1.0000
            Specificity : 0.8972
```

Our accuracy has further increased from 90.16 to 90.98% by tweaking the threshold

## SUMMARY:

1. We have basically incorporated Logistic Regression model to predict Loan Status of a customer.

   We have taken the most important explanatory variables to explain the output variable more efficiently.
2. We have optimized our model and increased its accuracy from **80.33 to 90.98%.** This was achieved because of Optimization.Further, sensitivity of the model is kept 1.This ensures that the model correctly predicts "Yes" to the customers who truly deserve the loan.
3. Optimization of the logistic regression can be carried out in two ways.One, by tuning the parameters and other by changing the threshold value.
4. By changing the combination of explanatory variables our accuracy improved from 80.33 to 90.16%.**This shows that Credit_History is the most important explanatory variable.**
5. By altering the threshold value from 0.5 to 0.55 using ROC curve our accuracy further increased from **90.16 to 90.98.**

**REFERENCES**:


- https://www.youtube.com/results?search_query=logistic+regression+fundamentals
- https://www.researchgate.net/publication/303326261_Machine_Learning_Project
- https://www.andrewng.org/