International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 198-201, June-2022

https://doi.org/10.47001/IRJIET/2022.606027

Loan Prediction Using Logistic Regression

[1]Simonraj E, [2]Mrs. Shivaleela S

[1]PG Student of MCA, Dr. Ambedkar Institute of Technology, Bangalore, India

[2]Assistant Professor, Department of MCA, Dr. Ambedkar Institute of Technology, Bangalore, India

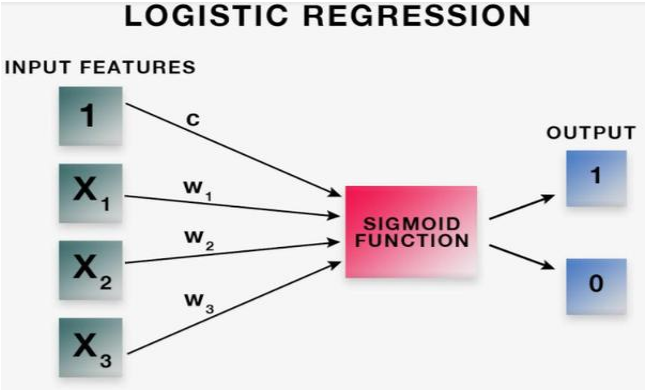
***Abstract* - For a variety of reasons, the banking industry continues to call for a more exacting predictive modeling framework. For the banking industry, it is challenging to predict credit defaulters. One of the criteria used to evaluate a loan's quality is its status, which comes after the loan application stage. Everything is not immediately visible. The loan status serves as the basis for the credit scoring model. Discovering defaulters is necessary and, ultimately, legitimate clients, credit data is reviewed with credibility using the credit score model. A model for credit rating credit data is what this study aims to produce. Many machine learning approaches are used in the development of the financial credit score model. In this research, we propose an analytical model for credit data based on machine learning classifiers. We put Min-Max normalization and linear regression together. The Jupyter notebook software suite is used to achieve the objective. The reason this model is recommended is that it delivers the most precise critical information. To forecast commercial banks loan status, and deploy a machine learning classifier.**

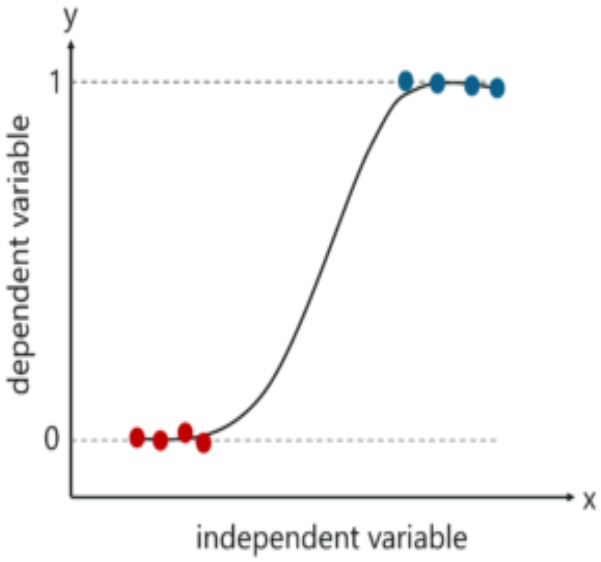
Keywords: Loan, Machine Learning, Statistical Analysis.

I. INTRODUCTION

Logistic Regression with Classification techniques is necessary for machine learning and data mining applications. Around 70% of Data Science problems are classified as classification problems. There are many classification problems that can be resolved, but the logistic regression is a well-liked and practical method for dealing with the binary classification problem. In essence, the logistic regression model is a member of the supervised classification algorithm family. Logistic regression examines the link between dependent and independent variables by estimating the probability with the use of a logistic function. When referring to dependent and independent variables in this context, the dependent variable is the target class variable we intend to predict, while the independent variables are the features we intend to use to do so. Estimating probabilities in logistic regression refers to determining the likelihood that an event will occur.

The shop owner, for instance, would like to foretell whether the customer who entered the store will purchase the play station (as an example) or not. The shopkeeper would look at a variety of customer characteristics, such as gender, age, and so forth, in order to forecast the likelihood that the customer would purchase a play station or not. The sigmoid curve is used to build the logistic function with different parameters.





Logistic Regression Hypotheses

- A binary dependent variable is required for binary logistic regression.
- In a binary regression, the dependent variable's factor level 1 should represent the desired result.
- The only variables that matter should be used.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 198-201, June-2022

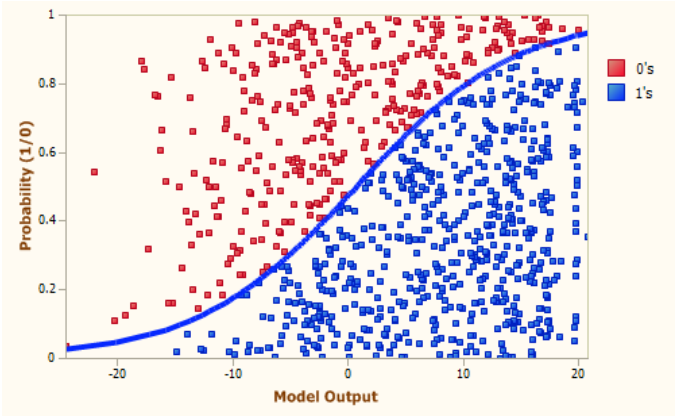
https://doi.org/10.47001/IRJIET/2022.606027

- The independent variables must be separate from one another. In other words, there should be little to no multicollinearity in the model.
- The log odds and the independent variables are linearly related.
- The sample sizes needed for logistic regression are quite large.

II. METHODOLOGY

2.1 Data Gathering

A statistical method for binary class forecasting is logistic regression. The outcome or target variable is binary in nature. There are only two possible classes when something is dichotomous. To classify a piece of mail as spam or not, a tumor as malignant or benign, or a transaction as genuine or fraudulent would be examples of classification in real life. These problems fall under the category of two class classification problems because all of the answers are categorical, i.e., Yes or No.



2.2 Logistic Fitting

Training and Predicting

```
from sklearn.linear_model import LogisticRegression
logmodel = LogisticRegression()
logmodel.fit(X_train,y_train)
predictions = logmodel.predict(X_test)
```

Numerous applications can be made for logistic regression. Here are a few examples of real-world applications.

2.3 Uses

Marketing: A marketing expert wants to predict if the subsidiary of his company will turn a profit, a loss, or just break even based on the features of the subsidiary activities.

Human Resources: Based on each employee's unique characteristics, A company's HR manager wants to predict their absenteeism trends.

Finance: Based on past transactions and customer history, a bank wants to determine whether its clients would default.

III. RESULTS AND DISCUSSION

```
Python 3.7.6 (tags/v3.7.6:43364a7ae0, Dec 19 2019, 00:42:30) [MSC v.1916 64 bit
(AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
= RESTART: C:\Users\hp\Desktop\MCA\5th sem\Internship Short\Loan Predict\Loan
Prediction.py
Libraries updated
pandas version  1.3.3
Number of columns in the dataframe: 13
Number of rows in the dataframe: 614

['Loan_ID', 'Gender', 'Married', 'Dependents', 'Education', 'Self_Employed',
'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term',
'Credit_History', 'Property_Area', 'Loan_Status']
Y   422
N   192
Name: Loan_Status, dtype: int64

   Loan_ID Gender Married  ... Credit_History Property_Area Loan_Status
0 LP001002   Male      No  ...           1.00         Urban           Y
1 LP001003   Male     Yes  ...           1.00         Rural           N
2 LP001005   Male     Yes  ...           1.00         Urban           Y
3 LP001006   Male     Yes  ...           1.00         Urban           Y
4 LP001008   Male      No  ...           1.00         Urban           Y

[5 rows x 13 columns]

       Dependents  ApplicantIncome  ...  Loan_Amount_Term  Credit_History
count      599.00           614.00  ...            600.00          564.00
mean         0.76          5403.46  ...            342.00            0.84
std          1.02          6109.04  ...             65.12            0.36
min          0.00           150.00  ...             12.00            0.00
25%          0.00          2877.50  ...            360.00            1.00
50%          0.00          3812.50  ...            360.00            1.00
75%          2.00          5795.00  ...            360.00            1.00
max          3.00         81000.00  ...            480.00            1.00

[8 rows x 6 columns]

Loan_ID           0
Gender           13
```

Figure 1: Execution of Dependencies

```
Married            3
Dependents        15
Education          0
Self_Employed     32
ApplicantIncome    0
CoapplicantIncome  0
LoanAmount        22
Loan_Amount_Term  14
Credit_History    50
Property_Area      0
Loan_Status        0
dtype: int64

Loan_ID            0
Gender            13
Married            3
Dependents        15
Education          0
Self_Employed     32
ApplicantIncome    0
CoapplicantIncome  0
LoanAmount         0
Loan_Amount_Term  14
Credit_History     0
Property_Area      0
Loan_Status        0
dtype: int64

Number of columns in the dataframe: 13
Number of rows in the dataframe: 614

Loan_ID            0
Gender             0
Married            0
Dependents         0
Education          0
Self_Employed      0
ApplicantIncome    0
CoapplicantIncome  0
LoanAmount         0
Loan_Amount_Term   0
```

Figure 2: Data Count and Description

```
Credit_History     0
Property_Area      0
Loan_Status        0
dtype: int64

Number of columns in the dataframe: 13
Number of rows in the dataframe: 542

Preprocessing Stage, Visualization Level 1 Done

Warning (from warnings module):
  File "C:\Python37\lib\site-packages\seaborn\_decorators.py", line 43
    FutureWarning
FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the
only valid positional argument will be `data`, and passing other arguments without an
explicit keyword will result in an error or misinterpretation.
Loan_Status   N   Y
Gender
Female       33  65
Male        133 311
```

Figure 3: Predicted value of loan prediction

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 6, Issue 6, pp 198-201, June-2022

https://doi.org/10.47001/IRJIET/2022.606027

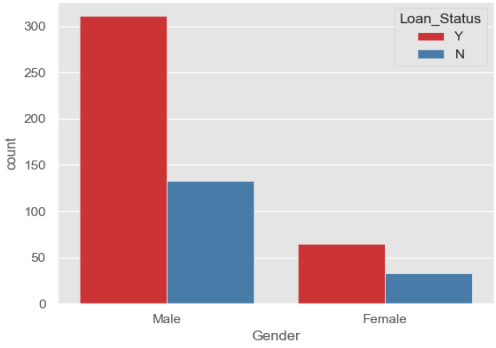


Figure 4: Gender vs Loan Status

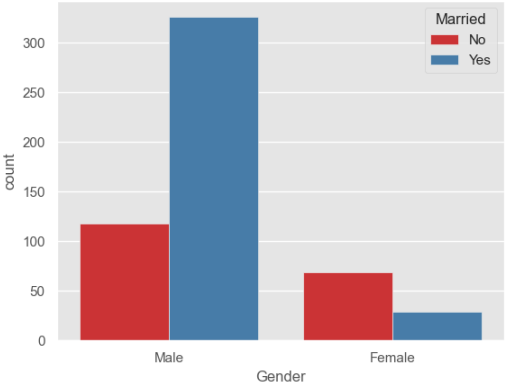


Figure 5: Gender vs Married

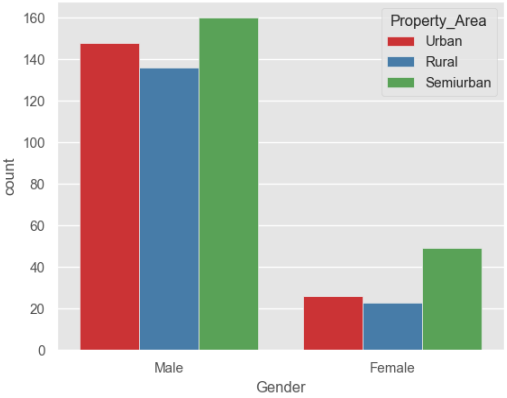


Figure 6: Gender vs Property Area

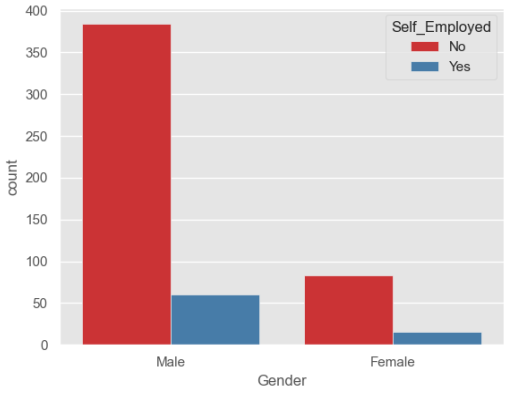


Figure 7: Gender vs Self Employed

Benefits:

It is simpler to use, analyze, and train logistic regression. It makes no presumptions regarding the distribution of classes in feature space. It is simple to extend it using multinomial regression and a natural probabilistic perspective of class predictions. It provides an evaluation of the adequacy of a predictor (coefficient size), as well as the direction of relationship (positive or negative). It swiftly classifies unclassified records.

Disadvantages:

If there are less data than features, logistic regression should not be utilized because this could lead to over-fitting. It establishes linear boundaries. The primary flaw in logistic regression is the assumption that the relationship between the dependent variable and the independent factors is linear.

It is only effective when predicting discrete functions. The discrete number set is thereby linked to the logistic regression's dependent variable. Since the decision surface of logistic regression is linear, it cannot resolve non-linear issues. Rarely do real-world circumstances involve linearly separable data

IV. CONCLUSION

An efficient technique for simulating the interaction between a binary response variable and one or more explanatory variables —where the latter may be continuous or categorical is logistic regression. A variety of techniques can be used to evaluate how well the resulting model fits.

ACKNOWLEDGEMENT

We would like to thank everyone who contributed to the success of the study and who helped us reach a final conclusion, whether directly or indirectly.

REFERENCES

[1] Abdelmoula, A.K. Bank credit risk analysis with k-nearest-neighbor classifier: Case of Tunisian banks. Accounting and Management Information Systems, 14(1), 79-106.

[2] Attig, A., & Perner, P. The Problem of Normalization and a Normalized Similarity Measure by Online Data. Tran. CBR, 4(1), 3-17.

[3] Bach, M.P., Zoroja, J., Jaković, B., & Šarlija, N. (2017). Selection of variables for credit risk data mining models: preliminary research. In 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 1367-1372.

ISSN (online): 2581-3048
Volume 6, Issue 6, pp 198-201, June-2022
https://doi.org/10.47001/IRJIET/2022.606027

[4] Byanjankar, A., Heikkilä, M., & Mezei, J. (2015). Predicting credit risk in peer-to-peer lending: A neural network approach. In IEEE Symposium Series on Computational Intelligence, 719-725.

[5] Devi, C.D., & Chezian, R.M. (2016). A relative evaluation of the performance of ensemble learning in credit scoring.

Citation of this Article:

Simonraj E, Mrs. Shivaleela S, "Loan Prediction Using Logistic Regression" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 6, Issue 6, pp 198-201, June 2022. Article DOI https://doi.org/10.47001/IRJIET/2022.606027

****

