

# Customer Loan Approval Prediction Using Logistic Regression

Mahankali Gopinath<sup>a,1</sup> K. Srinivas Shankar Maheep<sup>b</sup> and R. Sethuraman<sup>c</sup>

<sup>a,b</sup>UG Student, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India

<sup>c</sup> Assistant Professor, Dept of CSE, Sathyabama Institute of Science and Technology, Chennai, India

**Abstract:** Banking Sector contains loan where it is a process of lending or borrowing a sum of money by one or more individuals, organizations, etc. from Banks. The Person who lends that money from respective financier incurs a debt, and he is responsible to pay back the money with the Interest decided by Bank within a certain period. Generally what Bank's look into before applying for a loan is Credit History, Credit loss and Income of Applicant. So basically, loans play a major role regarding Income for Bank. Due to rapid urban development people who are applying for loans got increased rapidly. Therefore, finding the applicant to whom loan can be approved become a complexed process. In this paper, we want to predict the loan eligibility based on details of the customer. Fields that required are Matrimonial Status, Income, Education, Loan Amount, Credit History and other income sources of Applicant dependants. To predict the status, we will use Logistic Regression to spot the eligible applicants so bank will engage with them for granting loans to those people who can payback in a given time.

**Keywords:** Loan, Exploratory Data Analysis, Prediction, Logistic Regression

## 1. Introduction

The banking industry plays a significant role in present mostly in developing countries where money is usually required for all of them, so they will increase their market to capital value by gaining profits. Banks allows their customers to save lots of money in individual accounts. So, then Banks allows to lend money to business people or others who can utilize it for his or her capital growth and meet their Business requirements, and payback to the bank within a specific period of your time including Interest amount. So, interest is the profit gained by the banks by giving loans to folks that are in need. But Banks are worried about whether the person whose loan got granted will be ready to payback loan amount or not. so as to predict it, they basically inspect things associated with applicant like Credit score, Applicant Income.

---

<sup>1</sup> Mahankali Gopinath. UG Scholar, Dept of CSE, Sathyabama Institute of Science and Technology, India.  
Email: mahankaligopinath382@gmail.com

**Credit Score.** Here the credit score plays the key role in information given by the customer. In most scenarios Credit score is required for Loan sanction. If the applicant didn't payback his loan amount, then eventually his credit score will automatically get decreased. Giving Loans to people is one among the most business strategies for pretty much every bank. Banks will get most of the profits from the loan within the types of Interest. The most goal of Bank authorities is to grant loans to trustworthy people, so they'll pay back on the deadline comes. In recent times, banks are approving loans at their customers after a step-by-step procedure, but there's still no guarantee that the applicant's loan was granted or not. To approve loan, banks will undergo to estimate risk involved within the application, which is important for them as they cannot lend money to those that cannot afford to pay back in time which affects their economic status during this huge competitive market. So, we've got collected Dataset from the Kaggle contains loan applicant's details contains various fields like Gender, Applicant Income, Credit history etc.

After doing an Exploratory Data analysis on these data sets, we have discovered that probabilities of getting Loan granted are higher for applicants who have credit history equals to '1' with greater applicant income, with Education level mostly graduated and eventually who lives urban areas with properties. So, we'd like a model supported Fields Credit history, Education and Gender of Applicant. This model developed using Logistic Regression. we'll use this model to check with another data set and results obtained are stored in other file with Predicted Approval status as 'Yes' or 'No'. We have chosen Logistic Regression because it gives an Accuracy Rate of 80.945% approximately.

## **2. Literature Review**

In Somayyeh. Z et al[1] a model was proposed for predicting and identifying the right applicants who have applied for loan. So here Decision Tree technique is implemented to estimate the traits where accuracy rate is not much appreciable. In Sudhamathy G et al [2], This paper aims to build a model. It was based on decision tree where classification is used. It also uses the functions in the R-Package. Before preparing the model, the data is synchronized, made it ready to deliver effective results. The model prepared is used for predict using the dataset and the results shows the accuracy. In Dileep. B et al [3], Data analysis was done by figuring out techniques like Bayes classification, Decision Tree, Logistic Regression, K means algorithm, Neural Network Techniques, Perception model are combined in this model. The results in this work show the overall effective performance is very good. In Allen et al [4], The authors get to know that the economic effects of small business with credit scoring with both too high mean prices and also with more risk levels for small business. Also they find that a) banking specific and industry-wide learning curves are crucial b) These effects differ for banks that follow "rules" and c) These effects differ to people with slightly larger credits. In Altman, E. I.[5], This paper provides some empirical results of a study considering financial ratios as predictors of Japanese corporate failure. In contrast, the model proposed in this is independent of industry zone and size. This study shows that the model can predict bankruptcy with approximately 86.14% accuracy of industry and size. In J.H. Aboobyda et al [6], In this paper, a model was developed for categorize loan risk in the banking industry. It has been implemented by using data mining algorithms. The results make a comparison

between these three algorithms was conducted. J48 was best among the three based on accurate results. In A.B. Hussain et al [7], Here Two separate data mining models were created calculating the credit score that can be useful in making decisions of granting loans for the banks in Jordan residents. By the accuracy rate, the regression based model is found to perform more than the other function model. In T. Harris [8], This work tries to see the probability of default as a tool to live credit risk in an exceedingly Tunisian bank. A score calculable model was built using logistic regression, and computer science techniques. Thus, within the Tunisia context, this model is worth implementing in banking institutions so as to boost their credit risk management measures to watch and control credit. In Charles Kwofie et al [9], This study shows the good performance of logistic regression in predicting the probability of data provided by a microfinance company. The variability within the response variable within the logistic regression is not good enough.

3. Proposed Work

Python has could be a good area for data analytical which helps us in analysing the data with better models in data science. The libraries in python make the prediction for loan data and results with multiple terms considering all properties of the customer in terms of prediction. Logistic Regression is deployed to create the model and used to get the output by predicting results accurately. Credit history could be one best criteria that helps the banks and loan approvers to make its process for credit granting decisions For predicting results, we have collected Data sets in Kaggle. In order to build this model, we will import some python packages which will be used to analyse the data sets. They are NumPy, Pandas, Matplotlib and scikit-learn Libraries. All these libraries are available in Python.

3.1 Dataset Description

There are two Datasets which are required to build this model. 1) Train Dataset and 2) Test Dataset. The Test dataset contains list of customers applied for loan. By using Train Dataset, we can train the model and use it to predict loan status for Test dataset. The dataset is in CSV format. In python, Pandas is used to read the dataset. Refer the below table for field variables of the dataset.

Table 1. Data Description

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant Married(Y/N)
Dependents	Number of dependents
Education	Graduate/Under Graduate
Self-Employed	Self Employed (Y/N)
Applicant_Income	Applicant Income
Coapplicant_Income	Co Applicant Income
Loan_Amount	Loan Amount in thousands
Loan_Amount_Term	Term in Months
Credit_History	Credit History meets guidelines
Property_Area	Urban/ Rural
Loan_Status	Loan Approved (Y/N)

### 3.2 System Requirements

#### Software Requirements

- Windows XP, Windows 7, Windows 10
- Python 3.5 Mozilla Firefox(or any browser)
- Jupyter Notebook IDLE.

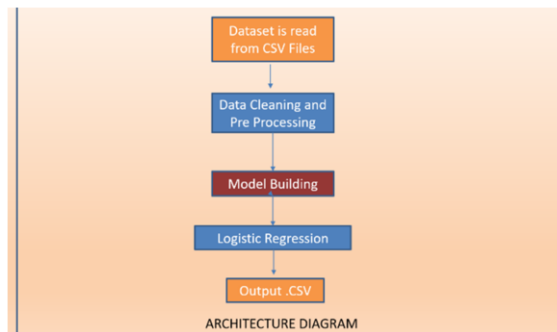
#### Hardware Requirements

- Minimum HDD: 20 GB
- Random Access Memory: 512 MB
- I3 processor-based computer or higher.

## 4. Software Methodology

After reading the dataset, we will implement Exploratory Data Analysis to understand and find out the outliers in the dataset.

### 4.1 Block Diagram



**Figure 1.** Block Diagram

In the above diagram, It shows what are the steps involved in building this model.

### 4.2 Modules

The Customer Loan Approval Prediction Model has mainly 3 modules. They are

- Reading and Cleaning Dataset.
- Model Building.
- Testing Dataset with Model

#### 4.2.1 Reading and Cleaning Dataset

We need to import Pandas, NumPy, and scikit-learn libraries and use them to process the information. Reading both training and testing dataset using Pandas. By Using head () function, we are going to be ready to see the first 10 rows of the dataset so that we'll have a clear picture of what fields does dataset contains in it. After then we will store the length of rows and columns within the dataset. we'd like to grasp the varied features and columns of the dataset Understanding Distribution of Numerical Variables like Applicant Income and Loan The amount may be done by using Boxplot to know and

finding the outliers of the Dataset fields. But there are more graduates with getting high incomes, which makes them outliers in this situation. Now it's time to understanding distribution of Categorical Variables. By making a Cross table with both Credit History and Loan Amount Fields we will see that Loans Approved within the Train Data set are more with applicants having credit history equals 1. Then we are going to write a function to search out the proportion of applicants whose loans are approved with credit history adequate to 1 and it shows more than 79% of individuals have gotten loans with a Credit history of 1.

In Boxplot of Loan Amount by Gender. Outliers are present in Male Gender than compared to female. Now we are going to move forward and understand outliers in an exceedingly better way within the next module to create the model. Outliers are slightly having extreme values when compared tonormal ones in the dataset we

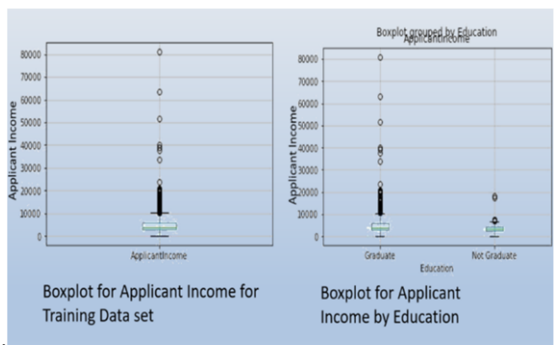


Figure2. Box Plot for Applicant Income by Education

Box plot for Applicant Income by Education shows that there are a greater number of graduates getting higher incomes can be Outliers.

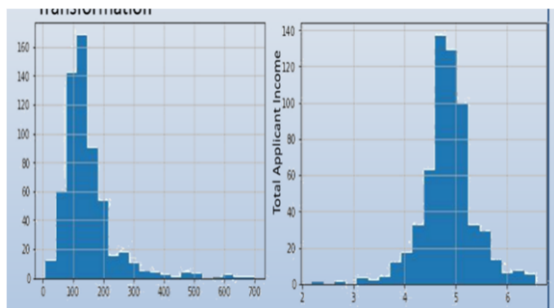


Figure3. Boxplot for Applicant Income after Log Transformation

The extreme values getting here are can happen in real, i.e., few people might apply for loans with high figures due to needs. So instead of looking them as outliers, we tried a log transformation to nullify those outliers effect. Also, we combine both Applicant and Co Applicant Income for better results. Also, we need to fill missing values for Data preparation for Model Building in next modules. We generally fill the missing values using pandas by Null values.

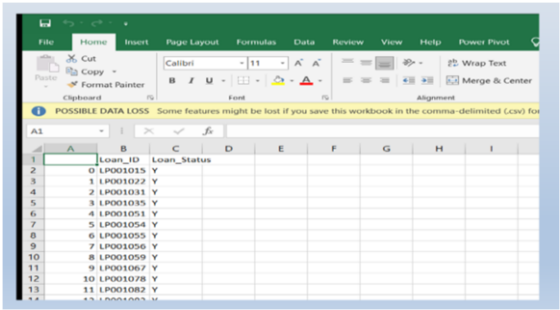
4.2.2 Model Building

After converting all our fields which contain categorical variables into numeric values as scikit-learn all inputs remain to be numeric. Here we can see all the variables in the dataset are in numeric Data type. Now we will build Generic Classification Function by importing scikit learn module and accessing performance. By Perform k-fold cross-validation with 5 folds. Training the algorithm using target and predictors.Fit the model again so that it can be referred outside the function by adding Accuracy rate and cross validation score.Look at the summary of missing values of both datasets and fill them with Null values for both categorical and numeric variables.Now all the variables are converted to numeric for processing in model. Now we create a new column named as Total Income by adding both Applicant Income and Co Applicant Income and create label encoders for them.We have used Logistic Regression in building the mode Logistic Regression is one of the popular algorithms that belongs to Supervised Learning Techniques. In this algorithm, the result must lie between 0 and 1 which means Yes or No (True/False). So, it was mainly used for solving classification-based problems. It predicts only two values, so it makes our model job very accurate and give good results as we expected.

- Credit history.
  - Applicantshaving High Applicant income.
  - Applicants with High academic degree.
  - Applicants having Properties in urban territory.
- So, we made our model with CreditHistory, Education & Gender.

4.2.3 Testing Dataset with Model

For this Model, the Accuracy Rate is 80.945% with Cross Validation Score of 80.946% which is best value when compared to recent models based on other Machine Learning Models.Output CSV file is stored in local storage with fields Unique Loan ID with predicted Loan Approval Status as Yes or No. By using this results Banks can go ahead and complete the loan approval process for right applicants.



	A	B	C	D	E	F	G	H	I
1		Loan_ID	Loan_Status						
2		LP001019	Y						
3		LP001022	Y						
4		LP001033	Y						
5		LP001035	Y						
6		LP001051	Y						
7		LP001054	Y						
8		LP001055	Y						
9		LP001056	Y						
10		LP001059	Y						
11		LP001062	Y						
12		LP001078	Y						
13		LP001082	Y						

Figure 4: Output CSV File Screenshot

## 5. Conclusion and Future Scope

Finally, by using logistic regression model we can predict whether the loan can be approved or not. so as to implement this various input variables were accustomed to get the output. Whenever a program takes the computer file it returns the output within the type of binary i.e., either 0 or 1. If the output is 0 then '0' is displayed and it indicates that the loan is not approved. If the output is 1 then '1' is displayed and it indicates that the loan is approved. This model may help the banking system to take right decision in approving or rejecting Loan applicants in less time.

## References

- [1] Z. Somayyeh, and M. Abdolkarim, Natural Customer Ranking of Banks in Terms of Credit Risk by Using Data Mining A Case Study: Branches of Mellat Bank of Iran, *Jurnal UMP Social Sciences and Technology Management*, vol. 3, no. 2, pp. 307–316, 2015.
- [2] Sudhamathy G and Jothi Venkateswara, *Analytics Using R for Predicting Credit Defaulters*, IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.
- [3] Dileep B. Desai, Dr. R.V.Kulkarni, A Review: Application of Data Mining Tools in CRM for Selected Banks, (IJCSIT) *International Journal of Computer Science and Information Technologies*, Vol. 4 (2), 2013, 199 – 201.
- [4] Allen, N., Berger, W., Scott, F., & Nathan, H. M. (2002). Credit Scoring and the Availability, Price, and Risk of Small Business Credit. FRB of Atlanta Working Paper No. 2002-6, FEDS Working Paper No. 2002-26.
- [5] Altman, E. I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *The Journal of Finance*, 23(4), 589-609.
- [6] J.H. Aboobyda, and M.A. Tarig, ,Developing Prediction Model Of Loan Risk In Banks Using Data Mining, *Machine Learning and Applications: An International Journal (MLAIJ)*, vol. 3, no.1, pp. 1–9, 2016.
- [7] Hussain.A.B, and Shorouq F.K.E., Credit risk assessment model for Jordanian commercial banks: Neural scoring approach, *Review of Development Finance*, Elsevier, vol. 4, pp. 20–28, 2014.
- [8] T.Harris, Quantitative credit risk assessment using support vector machines: Broadversus Narrow default definitions, *Expert Systems with Applications*, vol. 40, pp.4404– 4413, 2013.
- [9] Kwofie, Charles & Owusu-Ansah, Caleb &Boadi, Caleb. (2015). Predicting the Probability of Loan-Default: An Application of Binary Logistic Regression. *Research Journal of Mathematics and Statistics*. 7. 46-52. 10.19026/rjms.7.2206.