Loan Approval Prediction based on Machine Learning Approach

ABITH.K.SUNIL and DIVYA MARY BIJI

1 Christ Deemed to be University, Pune  
mail@christuniversity.in

**Abstract. *:*** With the enhancement in the banking sector lots of people are applying for bank loans but the bank has limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted will be a safer option for the bank is a typical process. So in this paper, we try to reduce this risk factor behind selecting a safe person to save lots of bank efforts and assets. This is done by mining the Big Data of the previous records of the people to whom the loan was granted before and based on records/experiences the machine was trained using the machine learning model which gives the most accurate result. The main objective of this paper is to predict whether assigning a loan to a particular person will be safe or not. This paper is divided into four sections (i)Data Collection (ii) Comparison of machine learning models on collected data (iii) Training of system on most promising model (iv) Testing

**Keywords:** Loan, Machine Learning, Training, Testing, Prediction.

1. INTRODUCTION

A Prediction Model uses data mining, statistics, and probability to forecast an outcome. Every model has some variables known as predictors that are likely to influence future results. The data was collected from various resources then a statistical model is made. It can use a simple linear equation or a sophisticated neural network mapped using complex software. As more data becomes available the model becomes more refined and the error decreases meaning then it’ll be able to predict with the least risk and consume as less time as it can. The Prediction Model helps the banks by minimizing the risk associated with the loan approval system and helps the applicant by decreasing the time taken in the process.

The main objective of the Project is to compare the Loan Prediction Models made implemented using various algorithms and choose the best one out of them that can shorten the loan approval time and decrease the risk associated with it. It is done by predicting if the loan can be given to that person based on various parameters like credit score, income, age, marital status, gender, etc. The prediction model not only helps the applicant but also helps the bank by minimizing the risk and reducing the number of defaulters.

In the present scenario, a loan needs to be approved manually by a representative of the bank which means that person will be responsible for whether the person is eligible for the loan or not and also calculating the risk associated with it. As it is done by a human it is a time-consuming process and is susceptible to errors. If the loan is not repaid, then it accounts for a loss to the bank and banks earn most of their profits from the interest paid to them. If the banks lose too much money, then it will result in a banking crisis. This banking crisis affects the economy of the country. So the loan must be approved with the least amount of error in risk calculation while taking up the least time possible. So a loan prediction model is required that can predict quickly whether the loan can be passed or not with the least amount of risk possible.

2. Literature Review

The author, Vaidya, Ashlesha [1] uses logistic regression as a machine learning tool in the paper and shows how predictive approaches can be used in real-world loan approval problems. His paper uses a statistical model (Logistic Regression) to predict whether the loan should be approved or not for a set of records of an applicant. Logistic regression can even work with power terms and nonlinear effects. Some limitations of this model are that it requires independent variables for estimation and that a large sample is required for parameter estimation.

A work by Amin, Rafik Khairul, and Yuliant Sibaroni [2] was referenced which used a decision tree algorithm called C4.5 to implement a predictive model. This algorithm creates a decision tree that generally gives high accuracy in decision-making problems DA data f 1000 cases is sed which 70% is approved and rather is rejected. This paper shows the wwasC4.5 algorithm performance n recognizing the eligibility of th applicant to repay his/her loan. From the conducted tests it is found that the highest precision value is 78.08% which was found using a data partition of 90:10. The rgreatestccallableis 96.4% and was reached with a data tition 80:20. Partition 80:20 is considered to be best since it has a high re ghesthighestracy. The research and work done by Arora, Nisha, and Pankaj Deep Kaur [3] aimed at forecasting whether an applicant can be a loan defaulter or not. It uses Bolasso to select the relevant attributes based on their robustness and then applied it to classification algorithms like Random Forest, SVM, Naive Bayes, and KNearest Neighbours (KNN) to test how accurately they can predict the results. It is concluded that Bolasso enabled Random Forest algorithm (BS-RF) provides the best results in credit risk evaluation and gives better accuracy by using optimized feature selection methods.

In a paper authored by Yang, Baoan, et al. [4], the use of artificial neural networks in an early warning system for predicting loan risk is discussed wherein it covers the early warning signals for deteriorating financial situations. The ability of an applicant to repay the loan is determined to be the most relevant aspecofin financial analysis. The early warning system in this paper uses an artisan facial neural network that is utilizing the traditional early warning concepts. This system based on ANN proves to be a very effective decision tool and early warning system for banks and other commercial lending organizations.

The scope of using Genetic Algorithms in building prediction models was also discussed in the paper by Metawa Noura, M. Kabir Hassan, and Mohamed Elhoseny [5]. This paper discusses a prediction model made using the Ga genetic Algorithm which can facilitate banks in making lending decisions in case of a decrease in lending supply. The main focus of the GA model is two-foldfolmaximizinging profit and minimizing errors in loan approval in case of dynamic lending decisions. Several factors like the type of loan, rating of the cartoon, too, and expected loan loss are integrated into GA chromosomes and then validation is done. The result shows that GAMCC increases the profits of the bank by 3.9% to 8.1%.

Yet another approach was used by Hassan, Amira Kamil Ibrahim, and Ajith Abraham[6] wherein they used the German dataset and built a prediction model working basically on backpropagation and implemented with three different back propagationation algorithms. They also used two different methods for two filtering functions for the attributes which resulted in DS2 giving the highest accuracy using the PLsFi filtering function.

3. Research Methodology

# Data Set



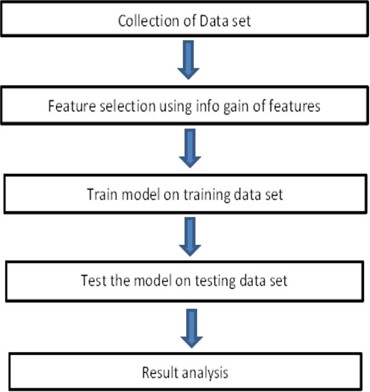
The training data set is now supplied to the machine learning model, and based on this data set the model is trained. Every new applicant details filled in at the time of application form acts as a test data set. After the operation of testing, the model predicts whether the new applicant is a fit case for approval of the loan or not based upon the inference it concludes based on the training data sets.

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| Loan\_ID | Unique Loan ID | Integer |
| Gender | Male/ Female | Character |
| Maritial\_Status | Applicant married (Y/N) | Character |



|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Description** | **Type** |
| Dependents | Number of dependents | Integer |
| Education\_Qualification | Graduate/Undergraduate | String |
| Self\_Imployed | Self-Employed (Y/N) | Character |
| Applicant\_Income | Applicant income | Integer |
| Co\_Applicant\_Income | Co-applicant income | Integer |
| Loan\_Amount | Loan amounts in thousands | Integer |
| Loan\_Amount\_Term | Term of the loan in months | Integer |
| Credit\_History | credit history meets guidelines | Integer |
| Property\_Area | Urban/ Semi Urban/ Rural | String |
| Loan\_Status | Loan Approved(Y/N) | Character |

## Loan Prediction Methodology

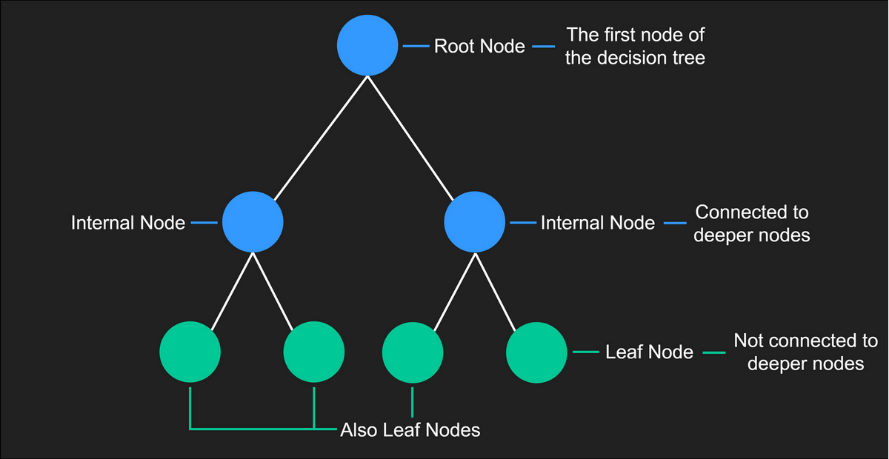


* + 1. **METHOD DESCRIPTION**

Six machine-learning classification models have been used for the prediction of android applications. The models are available in R open-source software. R is licensed under GNU GPL. The brief details of each model are described below.

**3.1.1.1 Decision Trees (C5.0):**

The basic algorithm of the decision tree [7] requires all attributes or features should be discretized. Feature selection is based on the greatest information gain of features. The knowledge depicted in the decision tree can be represented in the form of IF-THEN rules. This model is an extension of the C4.5 classification



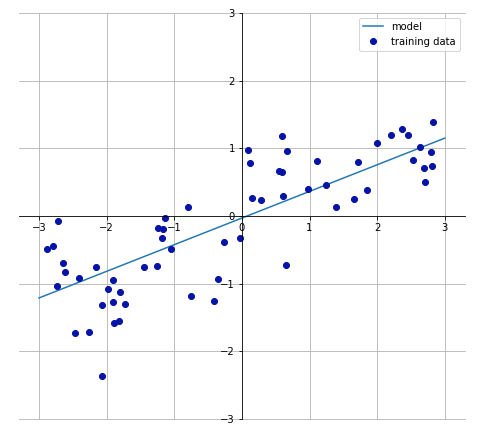
**3.1.1.2 Linear Models (LM):**

The Linear Model [10] is numerically indistinguishable from various regression analyses yet burdens its suitability for both different qualitative and numerous quantitative variables.

## Parameter setting for machine learning models

|  |  |
| --- | --- |
| **Model** | **Parameter Setting** |
| Decision Trees | Min Split = 20, Max Depth = |
|  | 30, Min Bucket = 7 |
|  |  |
|  |  |
| Support Vector Machine | Kernel Radial Basis |
| Linear Model | Multinomial |
| Logistic regression(LR) | Binary value |
|  |  |
|  |  |

*Loan Prediction*

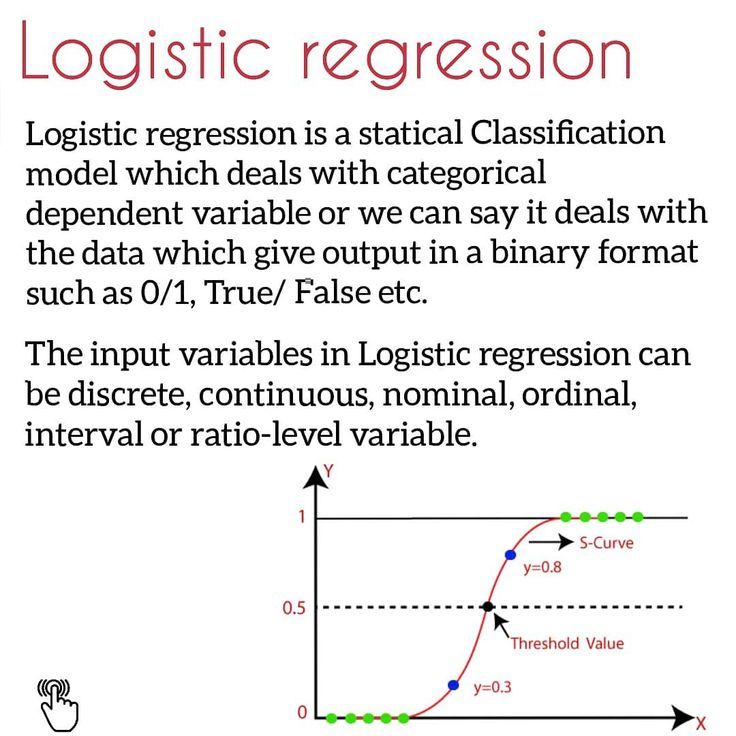


**3.1.1.3 Logistic regression(LR)**

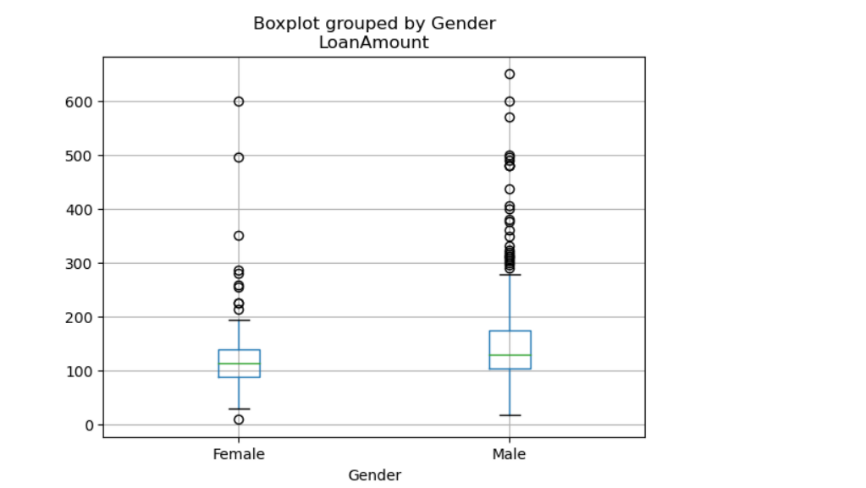
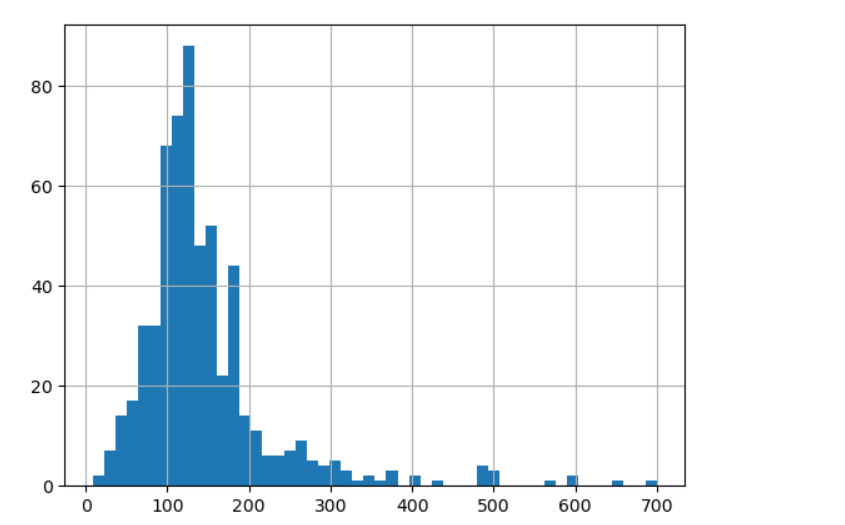
Logistic regression uses an equation as the representation, very much like linear regression. Input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta) to predict an output value (y). A key difference from linear regression is that the output value being modeled is a binary value (0 or 1) rather than a numeric value. Below is an example logistic regression equation:

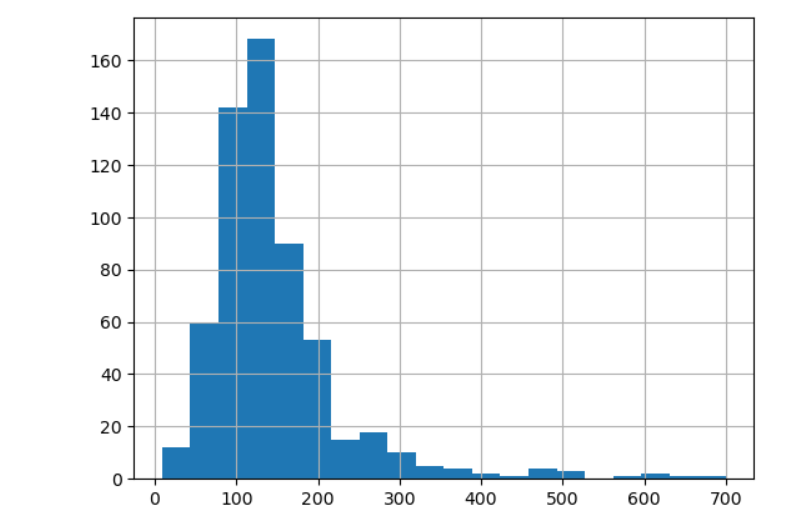
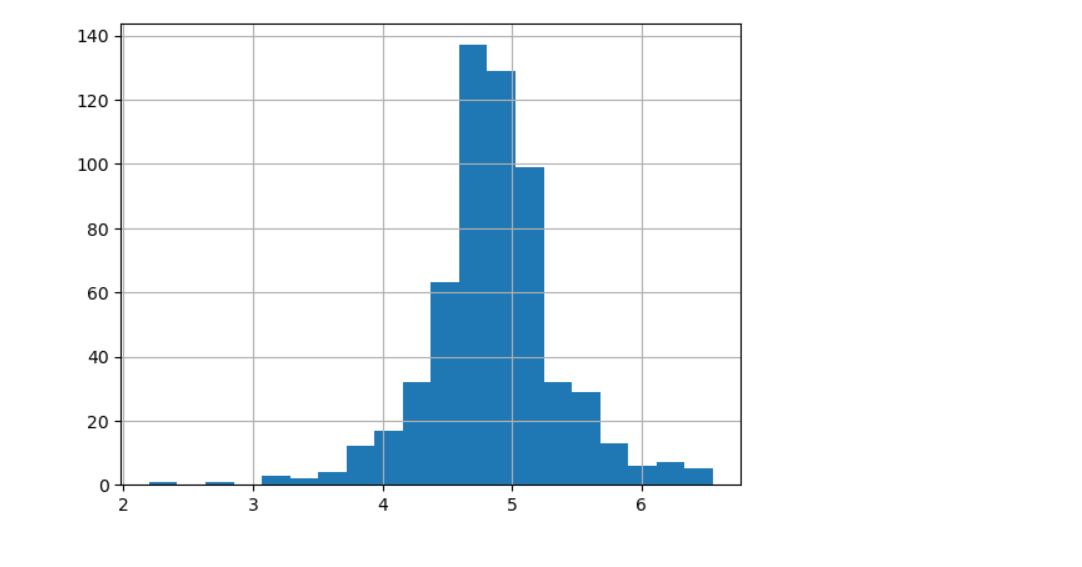
y = e^(b0 + b1\*x) / (1 + e^(b0 + b1\*x))

Where y is the predicted output, b0 is the bias or intercept term and b1 is the coefficient for the single input value (x). Each column in your input data has an associated b coefficient (a constant real value) that must be learned from your training data. The actual representation of the model that you would store in memory or a file is the coefficients in the equation (the beta value or b’s).



* + 1. **Exploratory data analysis**

****

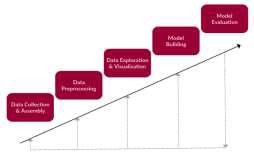


**3.1.3 METHOD DESCRIPTION**

The paper will be comparing different prediction models and deduce their limitations as well as advantages. Since all the research papers used different sets of data to infer the accuracy and for cross-validation of data, the authors have used the same data for all the models which will give a clearer view of their performance and lead to a better comparison of the same. Based on the results, a modified prediction model will be created to ensure maximum accuracy and performance.

# Implementation

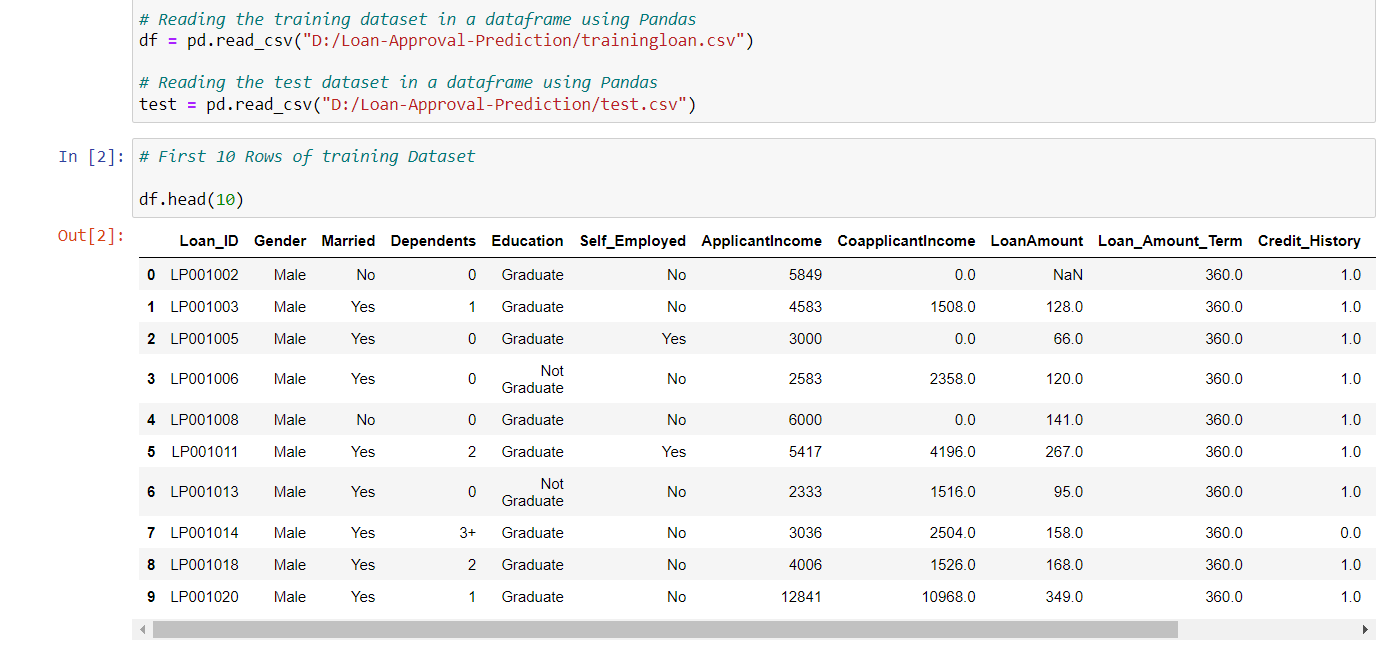
**Flow Chart**



**Figure 1.** Flow chart.

# Data collection and importing

The training set was imported in CSV format and a simple function is applied to check whether it was working or not

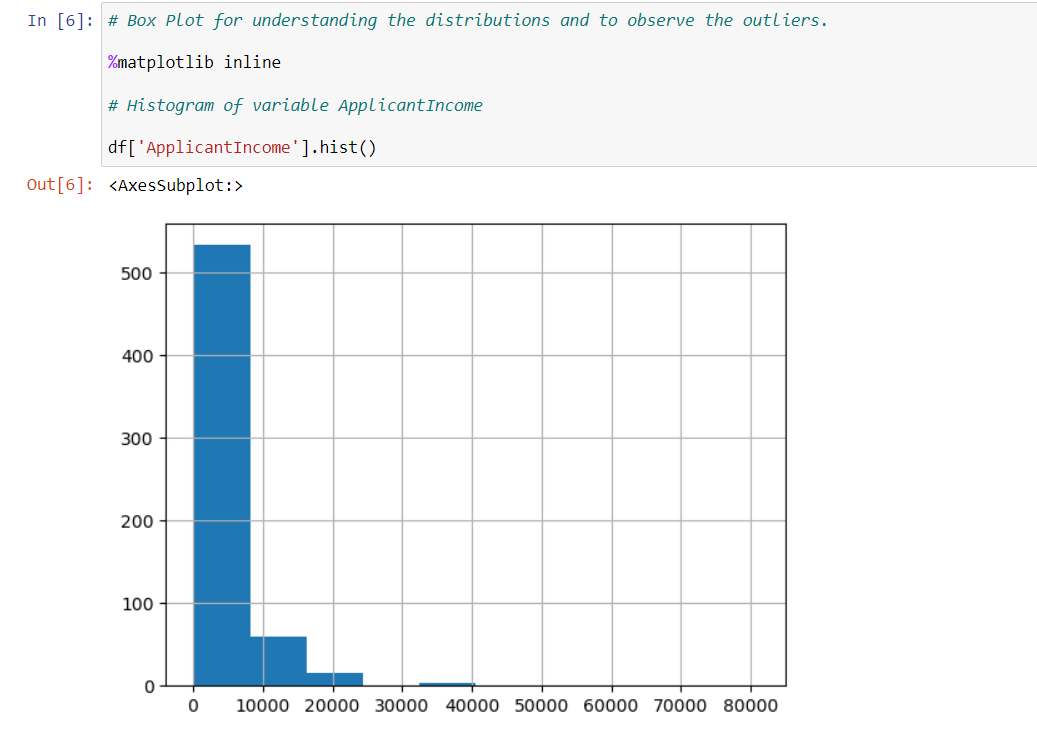


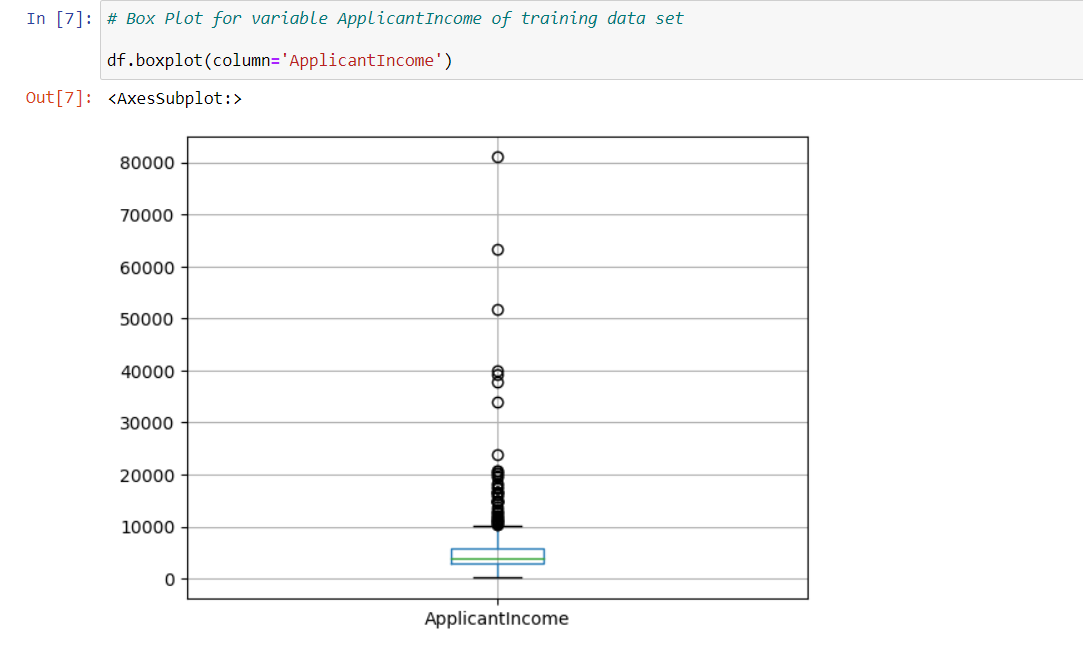
**Figure 2.** Importing and checking training data.

# Study of distribution of attribute

Box plots and histograms are used for the study of distribution factors. In the snapshot below one such factor (applicant income) has been used as an example.

**Figure 3.** Study of distribution of data.



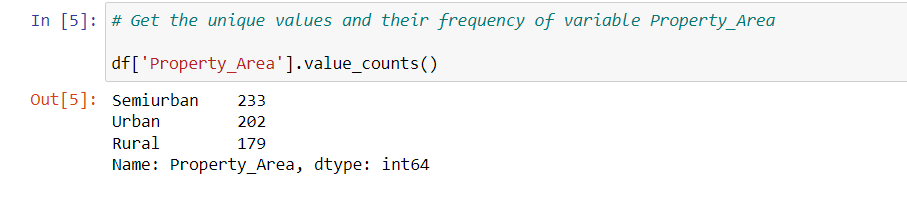


There are many extreme values due to the income gap and differences in education levels.

# 

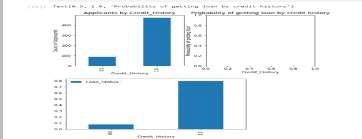
# Categorical Variable analysis

The following snapshot shows the method to calculate the chances of getting a loan based on credit history.



**Figure 4.** Checking the importance of data.

# Plotting of graph to infer more results

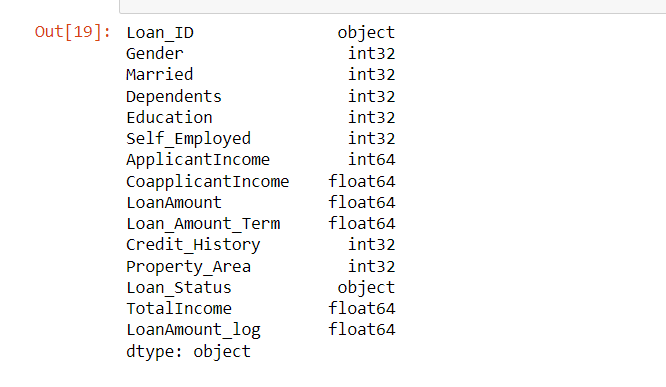


**Figure 5.** Plotting the relationship between factor and result to recognize patterns.

The graph depicts that it’s eight times easier to get a loan if an applicant has a valid credit history.

# Checking missing values

Data is being processed so that it can be determined how many values are missing from each column. The count of missing values present in the non-numerical attributes are processed and computed using the statistics.



**Figure 6.** Checking missing values.

# Finding extreme values and nullifying them

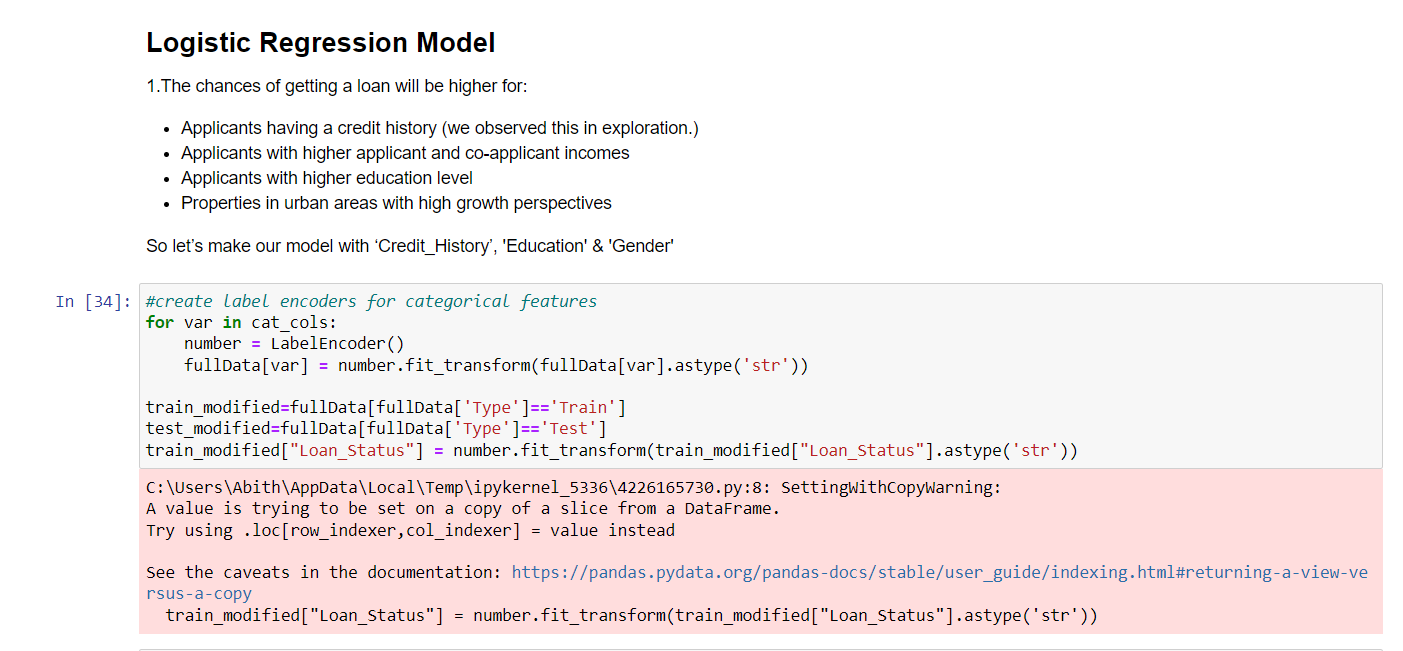
# 

**Figure 7.** Nullifying extreme values.

4. Result and Discussion

# Building a predictive model using Logistic Regression

A generic classification function is defined whose input is a model that helps determine the cross-validation scores and accuracy using the *K*-fold method.



**Figure 8.** A predictive model using Logistic Regression using Credit History.

# Building a predictive model using a Decision Tree

# 

**Figure 9.** A predictive model using a Decision Tree.

5. Conclusion

From a proper analysis of positive points and constraints on the component, it can be

safely concluded that the product is a highly efficient component. This application is

working properly and meeting all Banker requirements. This component can be easily

plugged into many other systems.

There have been several cases of computer glitches, and errors in content, and the most important weight of features are fixed in an automated prediction system, So shortly the so-called software could be made more secure, reliable, and dynamic weight adjustment. In near future, this module of prediction can be integrated with the module of the automated processing system. the system is trained on old training datasets in the future software can be made such that new testing dates should also take part in training data after some fixed time. The predictive models based on Logistic Regression, Decision Tree and, give the accuracy as 80.945%, 93.648%, and 83.388% whereas the cross-validation is found to be 80.945%, 72.213%, and 80.130% respectively. This shows that for the given dataset, the accuracy of the model based on the decision tree is highest but random forest is better at generalization even though its cross-validation is not much higher than logistic regression.

6. References

1. Vaidya and Ashlesha, Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017.
2. Amin, Rafik Khairul and Yuliant Sibaroni, Implementation of a decision tree using C4. 5 algorithms in decision making of loan application by the debtor (Case study: Bank Pasar of Yogyakarta Special Region), 2015 3rd International Conference on Information and Communication Technology (ICoICT). IEEE, 2015.
3. Arora, Nisha, and Pankaj Deep Kaur, A Bolasso-based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment, Applied Soft Computing 86 (2020), 105936.
4. Yang, Baoan, et al, An early warning system for loan risk assessment using artificial neural networks, Knowledge-Based Systems 14.5-6 (2001), 303-306.
5. Mettawatawa, Noura, M. Kabir Hassan, and Mohamed Elhoseny, Genetic algorithm based model for optimizing bank lending decisions, Expert Systems with Applications 80 (2017), 75-82.
6. Hassan, Amira Kamil Ibrahim, and Ajith Abraham. "Modeling consumer loan default prediction using ensemble neural networks, 2013 International Conference On Computing, Electrical And Electronic Engineering (IEEE). IEEE, 2013.