IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

Loan Prediction using Machine Learning

Mr. Vikas Gaikwad[1], Mr. Shreeyash Dhotre[2], Mr. Yash Raundal[3], Mr. Umesh Sangule[4], Prof. Sharad M Rokade[5]

Students BE, Department of Computer Engineering[1,2,3,4]
Head, Department of Computer Engineering[5]
Sir Visvesvaraya Institute of Technology, Nashik, Maharashtra, India
Savitribai Phule Pune University, Pune, India

Abstract: *With the advancement in technology, there are so many enhancements in the banking sector also. The number of applications is increasing every day for loan approval. There are some bank policies that they have to consider while selecting an applicant for loan approval. Based on some parameters, the bank has to decide which one is best for approval. It is tough and risky to check out manually every person and then recommended for loan approval. In this work, we use a machine learning technique that will predict the person who is reliable for a loan, based on the previous record of the person whom the loan amount is accredited before. This work's primary objective is to predict whether the loan approval to a specific individual is safe or not.*

Keywords: Loan Dataset, Logistic Regression, Random Forest, Flask

Problem Statement: *A Company wants to automate the loan eligibility process (realtime) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers. Here they have provided a partial dataset.*

I. INTRODUCTION

As the data are increasing daily due to digitization in the banking sector, people want to apply for loans through the internet. Artificial intelligence (AI), as a typical method for information investigation, has gotten more consideration increasingly. Individuals of various businesses are utilizing AI calculations to take care of the issues dependent on their industry information. Banks are facing a significant problem in the approval of the loan. Daily there are so many applications that are challenging to manage by the bank employees, and also the chances of some mistakes are high. Most banks earn profit from the loan, but it is risky to choose deserving customers from the number of applications. One mistake can make a massive loss to a bank. Loan distribution is the primary business of almost every bank. This project aims to provide a loan [1, 8] to a deserving applicant out of all applicants. An efficient and non-biased system that reduces the bank's time employs checking every applicant on a priority basis. The bank authorities complete all other customer's other formalities on time, which positively impacts the customers. The best part is that it is efficient for both banks and applicants.

This system allows jumping on particular applications that deserve to be approved on a priority basis. There are some features for the prediction like- 'Gender', 'Married', 'Dependents', 'Education', 'Self_ Employed', 'ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term', 'Credit_History', 'Property_Area', 'Loan_Status', 'Cibil_Score'.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

II. LITERATURE SURVEY

A prediction is a statement about what someone thinks will happen in the future. People make predictions all the time. Some are very serious and are based on scientific calculations, but many are just guesses. Prediction helps us in many things to guess what will happen after some time or after a year or after ten years. Predictive analytics is a branch of advanced analytics that uses many techniques from data mining, statistics, modeling, machine learning, and artificial intelligence to analyze current data to make predictions. "Adyan Nur Alfiyatin, Hilman Taufiq" [2] and their friends work on the house price prediction. They use regression analysis and Particle Swarm Optimization (PSO) to predict house price". One other similar work on the Mohamed El Mohadab, Belaid Bouikhalene [3] and Said Safi to predict the rank for scientific research paper using supervised learning. Kumar Arun, Garg Ishan and Kaur Sanmeet [1] work on bank loan prediction on how to bank approve a loan. They proposed a model with the help of SVM and Neural networks like machine learning algorithms. This literature review helps us carry out our work and propose a reliable bank loan prediction model.

III. PROPOSED MODEL

Prediction of granting the loan to the customers by the bank is the proposed model. Classification is the target for developing the model and hence using Logistic Regression with sigmoid function is used for developing the model. Preprocessing is the major area of the model where it consumes more time and then Exploratory Data Analysis which is followed by Feature Engineering and then Model Selection. Feeding the two separate datasets to the model, and then preceding the model. Logistic regression is a type of statistical machine learning technique/algorithm which is used to classify the data by considering outcome variables on extreme ends and tries to make a logarithmic line that distinguishes between them. By this way prediction can be made through Logistic Regression.

IV. DATASET DESCRIPTION AND PRE-PROCESSING

4.1 Data Collection

Data has been collected from the Kaggle one of the most data source providers for the learning purpose and hence the data is collected from the Kaggle, which had two data sets one for the training and another testing[12]. The training dataset is used to train the model in which datasets is further divided into two parts such as 80:20 or 70:30 the major datasets is used for the train the model and the minor dataset is used for the test the model and hence the accuracy of our developed model is calculated.

4.2 Data Prepocessing

Data mining technique has been used in Pre-Processing for transforming raw data which is collect using online form into useful and efficient formats. There is a need to convert it in useful format because it may have some irrelevant, missing information and noisy data. To deal with this problem data cleaning technique has been used.

Before data mining the data reduction techniques is used to deal with huge volume of data. So data analysis will become easier and it intends to get accurate results. So data storage capacity increase and cost to analysis of data reduces.

The size of data can be reduced by encoding mechanisms. So it may be lossy or lossless. If the original data is obtained after reconstruction from compressed data, such reductions are called lossless reduction else it is called lossy reduction. Wavelet transforms and PCA (Principal Component Analysis) methods are effective for reduction.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

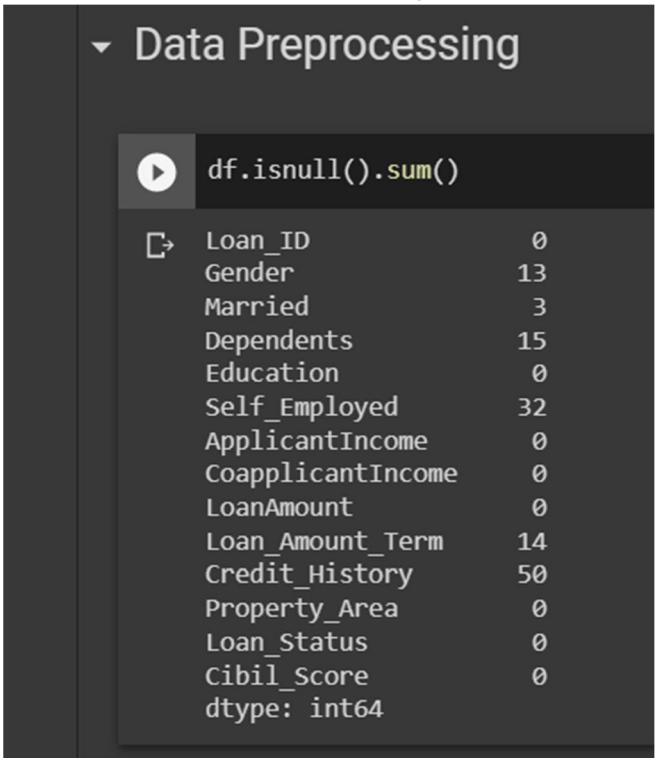

Fig: Data Prepocessing

4.3 Feature Engineering

In feature engineering a proper input dataset which is compatible as per machine learning algorithm requirements is prepared. In our model Pandas, Numpy and Matplotlib library has been imported to run. So the performance of machine learning model improves.

```
import pandas as pd
import numpy as np
import matplotlib,pyplot as plt
```

4.4 List of Techniques

Imputation

There is one more measure problem i.e. missing values when data is prepared for our machine learning model. There may be many reason of missing values like human errors, interruptions in flow of data, security concerns, and so on. The performance of machine learning model severely affected by missing values.

```
train['Gender'].fillna(train['Gender'].mode()[0],inplace=True)
train['Married'].fillna(train['Married'].mode()[0],inplace=True)
train['Dependents'].fillna(train['Dependents'].mode()[0],in place=True)
```

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

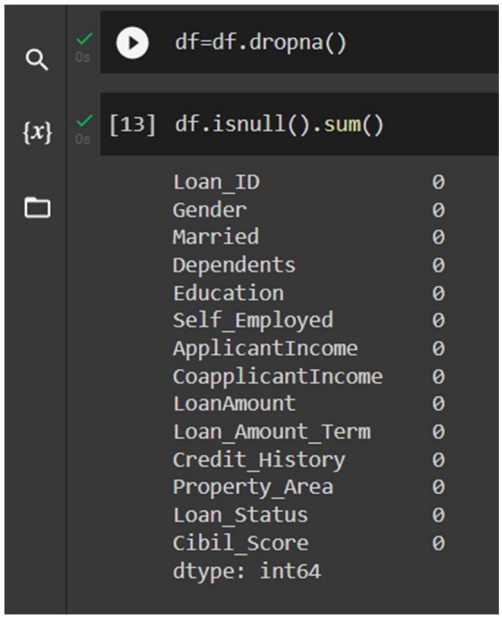

Fig : Impuatiotion

Handling Outliers

To detect the outliers the data is demonstrated visually and afterwards handled the outliers. When the ouliers decisions visulaized are of high precision and accurate. Percentiles is another mathematical method to detect outliers. In this method,it assumes a certain percentage of value from top or taken it from bottom as an outlier. The key point is here to set the percentage value once again, and this depends on the distribution of your data as mentioned earlier.

Binning

The key point between performance and overfitting is binning. In my opinion, for numerical values columns, except very few overfitting cases, binning might be redundant for some kind of algorithms, due to its effect on the performance of model. However, for categorical columns, the labels which have low frequencies might affected from the robustness of statistical models in negative manner. After assigning a comman category to all these less frequent values helps to keep the model robust.

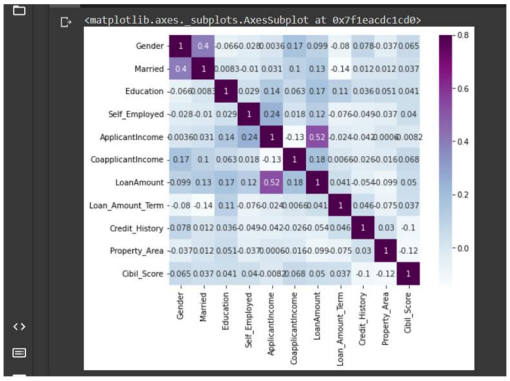

Fig : Heap Map

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

V. MODEL SELECTION

The process of selecting a final machine learning model from among a group of candidate machine learning models for a particular training dataset of Loan customer is called model selection. There are different types of model like logistic regression, SVM, KNN, etc. All these models have some merits and demerits for example predictive error gives the statistical noise in the data, the incompleteness of the sample data, and the limitations of each different model type. The chosen model meets the requirements and constraints of the stakeholders (Bank and Customers) project stakeholders. A model should have parameters like

- Skillful as compared to naive models.
- Skillful relative to other tested models.
- Skillful relative to the state-of-the-art.

Thus, Prediction of loan approval is a type of a classification problem and hence this model is used.

From sklearn.linear_model import LogisticRegression model = LogisticRegression() model.fit(x_train, y_train)

VI. MODEL EVALUATE

Model evaluation is technique which is used for the evaluating the performance of the model based on some constraints it should be kept in mind while evaluating the model that it can't underfoot or overfit the model. Various methods are present to evaluate the performance of the model such as Confusion metrics, Accuracy, Precision, Recall, F1 score etc.

CONFUSION METRICS:

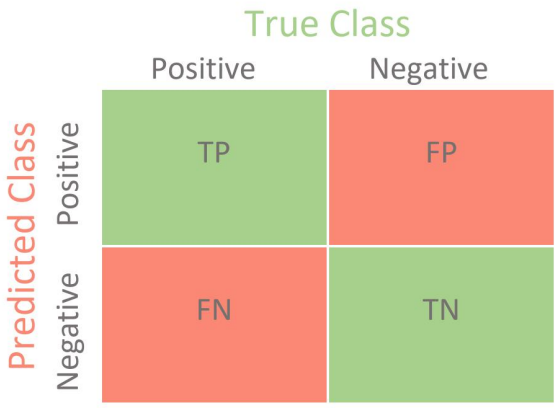

Fig : Confusion Matrix

ACCURACY

Accuracy of the model has been measured by predefined metrics. In a balance class model shows high accuracy but in the case of unbalanced class the accuracy is very less.

$$\frac{(TP + TN)}{(TP + FP + TN + FN)}$$

PRECISION

Percentage ratio of positive instances and total predicted positive instances gives precision value. In the below equation denominator represents the model positive prediction done from the whole given dataset. Precision value tells the perfectness of our model. In our data set good precision value has been obtained.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 7, May 2022
Impact Factor: 6.252

$$\frac{TP}{TP + FP}$$

RECALL

Percentage ratio of positive instances with actual total positive instances is recall value. Here denominator (TP + FN) shows the total number of positive instances which are present in whole dataset. As a result it has obtained 'how much extra right ones, the model will failed if it shows maximum right ones'

$$\frac{TP}{TP + FN}$$

F1 Score

The harmonic mean (HM) of precision and recall values is called F1 Score. Model will be best performer if it shows maximum F1 Score. Numerator shows the product of precision and recall if one goes low either precision or recall, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted (precision) having positive valueand doesn't miss out on positives and predicts them negative (recall).

$$\frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = \frac{2 * Presion * Recall}{Presion + Recall}$$

VII. CONCLUSION

The process of prediction starts from cleaning and processing of data, imputation of missing values, experimental analysis of data set and then model building to evaluation of model and testing on test data. On Data set, the best case accuracy obtained on the original data set is 0.811. The following conclusions are reached after analysis that those applicants whose credit score was worst will fail to get loan approval, due to a higher probability of not paying back the loan amount. Most of the time, those applicants who have high income and demands for lower amount of loan are more likely to get approved which makes sense, more likely to pay back their loans. Some other characteristic like gender and marital status seems not to be taken into consideration by the company.

REFERENCES

[1] Kumar Arun, Garg Ishan, Kaur Sanmeer, Loan Approval Prediction based on Machine Learning Approach.
[2] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita, Wayan Firdaus Mahmudy, 'Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization'. International Journal of Advanced Computer Science and Applications (Vol. 8, No. 10, 2017).
[3] Mohamed El Mohadab, Belaid Bouikhalene, Said Safi, 'Predicting rank for scientific research papers using supervised learning' Applied Computing and Informatics 15 (2019) 182–190.
[4] K. Hanumantha Rao, G. Srinivas, A. Damodhar, M. Vikas Krishna: Implementation of Anomaly Detection Technique Using Machine Learning Algorithms: Internatinal Journal of Computer Science and Telecommunications (Volume2, Issue3, June 2011).
[5] J. R. Quinlan. Induction of Decision Tree. Machine Learning, Vol. 1, No. 1. pp. 81-106., 1086.
[6] G. Arutjothi, C. Senthamarai: Prediction of loan status in commercial bank using machine learning classifier, International Conference on Intelligent Sustainable Systems (ICISS), 2017.
[7] J.R. Quinlan. Induction of decision trees. Machine learning Springer, 1(1):81–106, 1086.
[8] Vishnu Vardhan case study of bank loan prediction,https://medium. com/@vishnumbaprof/case-study-loan-prediction-ac035f3ec9e4.
[9] S.S. Keerthi and E.G. Gilbert. Convergence of a generalize SMO algorithm for SVM classifier design. Machine Learning, Springer, 46(1):351–360, 2002.
[10] J.M. Chambers. Computational methods for data analysis. Applied Statistics, Wiley, 1(2):1–10, 1077