IJARSCT

ISSN (Online) 2581-9429

Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 1, May 2022

Loan Prediction Using Machine Learning

Mr. Sangule Umesh[1], Mr. Dhotre Shreeyash[2], Mr.Raundal Yash[3],
Mr. Gaikwad Vikas[4], Prof. Sharad M Rokade[5]
BE Students, Department of Computer Engineering[1,2,3,4]
HoD, Department of Computer Engineering[5]
Sir Visvesvaraya Institute of Technology, Nashik, Maharashtra, India

Abstract: *In our banking system, banks have many products to sell but main source of income of any banks is on its credit line. So they can earn from interest of those loans which they credits. A bank's profit or a loss depends to a large extent on loans, whether the customers are paying back the loan or defaulting. By predicting the loan defaulters, the bank can reduce its Non- Performing Assets. This makes the study of this phenomenon very important. Previous research in this era has shown that there are so many methods to study the problem of controlling loan default. But as the right predictions are very important for the maximization of profits, it is essential to study the nature of the different methods and their comparison.*

Keywords: Loan, Outlier, Prediction, Component, Overfitting, Customer Loan, Preprocessing, Classification Models

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

Circulation of the loans is that the core business a part of as good as each and every bank. The principle parcel the bank's resources are straightforwardly came from the benefit acquire from the advances distributed by the banks. The main goal in banking system is to invest their resources in safe hands wherever it's. Now a day's several banks/financial agencies approves loan after a relapse method of verification and validation however still there's no surety whether or not the chosen candidate is the worthy right candidate out of all candidates. Through this method we are able to predict whether that particular candidate is safe or not and the whole method of validation of attribute is automated by machine learning technique [8][6]. The disadvantage of this model is that it emphasizes completely different weights to every issue however in reality sometime loan can be approved on the premise of single strong part only, that isn't possible through this method. Loan Prediction is useful for member of staff of banks as well as for the candidate. The aim of this Paper is to apply quick, immediate and easy way to choose the worthy person [6]. It will give special gain to the bank. The Loan Prediction method can automatically compute the heaviness of each attribute taking part in loan processing and on new test data information same issues. This paper has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this research is to predict the safety of loan [1][3]. To predict loan safety, the logistic regression algorithm is used. First the data is cleaned so as to avoid the missing values in the data set

II. Literature Survey

Logistic Regression is a popular and very useful algorithm of machine learning for classification problems. The advantage of logistic regression is that it is a predictive analysis. It is used for description of data and use to explain relationship between a single binary variable and single or multiple nominal, ordinal and ration level variables which are independent in nature.

DOI: 10.48175/IJARSCT-3496

Bhoomi Patel, Harshal Patil, Jovita Hembram, Shree Jaswal are used data mining methodology to predict the likely default from a dataset that contains information about home loan applications, thereby helping the banks for making better decisions in the future [3].

Xin Li, Xianzhong Long, Guozi Sun, Geng Yang, and Huakang Li This paper mainly introduces the main application of LSTM-SVM model in user loan risk prediction, and elaborates the current economic background, traditional risk forecasting method. On this basis, the prediction methodology based on LSTM method and SVM method is proposed, and the prediction results are compared with the traditional algorithm, and the feasibility of the model is confirm. However, the LSTM-SVM method proposed in this paper actually has few limits and needs to be improved in future research [7]. Aakanksha, Tamara Denning, Vivek Srikumar, Sneha Kumar Kesera[8] this paper is mainly used for voting classifier (combination of logistic regression, naïve bayes, SVM).

III. Proposed Model

In Machine Learning, we are using semi-automated extraction of knowledge of data for identifying whether a loan would be approved or not [6][8]. Classification could be a supervised learning within which the response is categorical that's its values area unit in finite unordered set. To easily the matter of classification, scikit learn are used. The praim primacy of this system is company need not has to maintain a ground team to validate and verify the customer records. They can easily check whether the loan has to be approved or not by this prediction model. In this paper we try to develop user interface flexibly graphics concepts in mind, associated through a browser interface. Our goal is to implement machine learning model so as to classify, to the best potential degree of accuracy, master card fraud from a dataset gathered from Kaggle. once initial knowledge exploration, we have a tendency to knew we might implement a random forest model for best accuracy reports. Random forest, as it was a good candidate for binary classification. Python sklearn library was used to implement the project, We used Kaggle datasets for Credit card fraud detection, using pandas to data frame for class ==0 for no fraud and class==1 for fraud, matplotlib for plotting the fraud and non fraud data, train_test_split for data extraction (Split arrays or matrices into random train and test subsets) and used Logistic Regression machine learning algorithm for fraud detection and print predicting score according to the algorithm. Finally Confusion matrix was plotted on true and predicted.

3.1 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.

3.2 Pre Processing

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.

3.3 Feature Engineering

In feature engineering a proper input dataset which is compatible as permachine learning algorithm requirements is prepared. In our model Pandas and Numpy library has been imported to run. So the performance of machine learning model improves. import pandas as pd import numpy as np

DOI: 10.48175/IJARSCT-3496

3.4 List of Techniques

1. **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)
2. **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.

IV. System Architecture

4.1 Implementation Details (Modules)

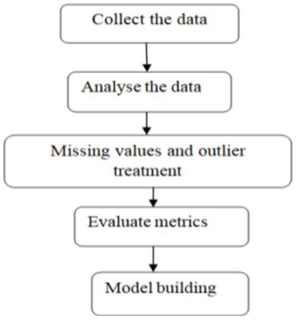
1. **Loan Dataset:** Loan Dataset is very useful in our system for prediction of more accurate result. Using the loan Dataset the system will automatically predict which costumer's loan it should approve and which to reject. System will accept loan application form as an input. Justified format of application form should be given as an input to get processed.
2. **Determine the Training and Testing Data:** Typically , Here the system separate a dataset into a training set and testing set ,most of the data use for training ,and a smaller portions of data is use for testing. after a system has been processed by using the training set, it makes the prediction against the test set.
3. **Data Cleaning and Processing:** In Data cleaning the system detect and correct corrupt or inaccurate records from database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing , modifying or detecting the dirty or coarse data. In Data processing the system convert data from a given form to a much more usable and desired form i.e. make it more meaningful and informative.

4.2 Models Used

1. **SVM:** In this approach, each data item is plotted in a n dimensional space, where n represents the number of features with each feature represented in a corresponding co- ordinates. A hyper plane is determined to distinguish the classes (possibly two) based on their features.
2. **Decision Tree:** Decision tree is a type of supervised learning algorithm(having a pre-defined target variable) that is mostly used in classification problems. In this technique, we split the population or sampleinto two or more homogeneous sets(or sub-populations) based on the most significant splitter/differentiator in input variables. Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable.
3. **Random Forest:** Random Forest is a tree-based bootstrapping algorithm wherein a certain no. of weak learners (decision trees) are combined to make a powerful prediction model. For every individual learner, a random sample of rows and a few randomly chosen variables are used to build a decision tree model. Final prediction can be a function of all the predictions made by the individual learners. In the case of a regression problem, the final prediction can be the mean of all the predictions.
4. **Logistic Regression:** This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome. The logistic function used is a logit function which is a log of odds in the favour of the event. Logit function develops a s-shaped curve with the probability estimate similar to a step function

DOI: 10.48175/IJARSCT-3496

V. Overview



VI. Mathematical Model

Consider any decision problem, where forgiven number of inputs, decision oriented solution is available so our project is NP complete but some cases like not proper input format provided or if dataset not trained proper it's NP hard.

Let s be System:

S=I, P, O

S: is a System

I=I1, I2

P= DC, DP, DV, NBA, CL

O=RD

I1: Loan Dataset

I2: Trained Dataset.

DC: Data Cleaning

D DP: Data Processing

DV: Data Verification

NBA: Naïve Bayes Algorithm

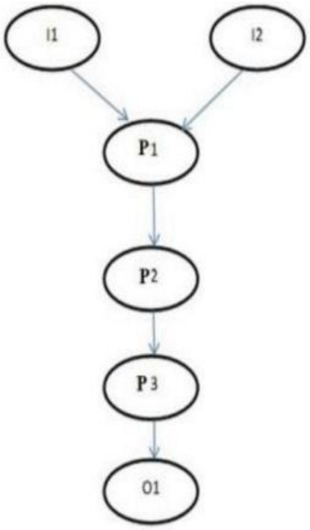
CL: Classification

RD: Report Deliver Success

Condition: Proper features trained Dataset will give proper output

Failure Condition No Trained Dataset.

6.1 Mathematical Model



6.2 Other Specification

A. Advantages

1. It can provide special advantages to the bank.
2. The Loan Prediction System can can automatically calculate the weight of each features taking part in loan processing and on new test data same features are processed with respect to their associated weight.
3. It is easy to implement and gives quick result.

B. Disadvantages

1. The disadvantage of this model is that it emphasize different weights to each factor but in real life sometime loan can be approved on the basis of single strong factor only, which is not possible through this system.

C. Application

1. In the Finance Company.
2. Banking Sectors
3. Private Loan Companies

VII. Conclusion and Future Scope

7.1 Conclusion

We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed. We did bivariate and multivariate analysis to see imapct of one another on their features using charts. We analysed each variable to check if data is cleaned and normally distributed. We cleaned the data and removed NA values We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association. We calculated correaltion between independent variables and found that applicant income and loan amount have significant relation. We created dummy variables for constructing the model. We constructed models taking different variables into account and found through odds ratio that credit credit history is creating the most impact on loan giving decision.

7.2 Future Scope

In future, this model can be used to compare various machine learning algorithm generated prediction models and the model which will give higher accuracy will be chosen as the prediction model. This paper work can be extended to higher level in future. Predictive model for loans that uses machine learning algorithms, where the results from each graph of the paper can be taken as individual criteria for the machine learning algorithm

References

[1]. Nikhil Madane, SiddharthNanda,"Loan Prediction using Decision tree", Journal of the Gujrat Research History, Volume 21 Issue 14s, December 2019.

[2]. https://www.kaggle.com/telco-churn

[3]. PhilHyo Jin Do ,Ho-Jin Choi, "Sentiment analysis of real-life situations using loca- tion, people and time as contextual features," International Conference on Big Data and Smart Computing (BIGCOMP), pp. 39–42. IEEE, 2015.

[4]. S. Vimala, K.C. Sharmili, ―Prediction of Loan Risk using NB and Support Vector Machine‖, International Conference on Advancements in Computing Technologies (ICACT 2018), vol. 4, no. 2, pp. 110-113, 2018.

[5]. K I Rahmani, M.A. Ansari, Amit Kumar Goel, "An Efficient Indexing Algorithm for CBIR,"IEEE- International Conference on Computational Intelligence & Communication Technology ,13-14 Feb 2015.

[6]. Kumar Arun, GargIshan, Kaur Sanmeet, ―Loan Approval Prediction based on Machine Learning Approach‖, IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 18, Issue 3, pp. 79-81, Ver. I (May-Jun. 2016).

DOI: 10.48175/IJARSCT-3496