RESEARCH ARTICLE OPEN ACCESS

Loan Prediction by using Machine Learning Models

Pidikiti Supriya[1], Myneedi Pavani [2], Nagarapu Saisushma[3]
Namburi Vimala Kumari[4] , K Vikas[5]

[1, 2, 3,4] B-Tech, Dept of CSE, VVIT, Guntur, AP.
[5]Assoc Professor, Dept of CSE, VVIT, Guntur, Ap.

ABSTRACT

With the enhancement in the banking sector lots of people are applying for bank loans but the bank has its limited assets which it has to grant to limited people only, so finding out to whom the loan can be granted which will be a safer option for the bank is a typical process. So in this project we try to reduce this risk factor behind selecting the safe person so as 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 on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this project is to predict whether assigning the loan to 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. In this paper we are predict the loan data by using some machine learning algorithms they are classification, logic regression, Decision Tree and gradient boosting.

Keywords: Machine learning, Decision Tree, prediction, Python.

I. INTRODUCTION

This Problem is done by mining the Big Data of the previous records of the people to whom the loan was granted before and on the basis of these records/experiences the machine was trained using the machine learning model which give the most accurate result. The main objective of this paper is to predict whether assigning the loan to a particular person will be safe or not. We have implemented this loan prediction problem using Decision tree algorithm and data cleaning in Python as there are missing values in the dataset. We use map function for the missing values. The aim of this paper is to apply machine learning technique on dataset which has 1000 cases and 7 numerical and 6 categorical attributes. The creditability of a customer for sanctioning loan depend on several parameters, such as credit history, Installment etc. [2].

2 LITERATURE REVIEW

Data mining is the process of analyzing data from different perspectives and extracting useful knowledge from it[3]. It is the core of knowledge discovery process. The various steps involved in extracting knowledge from raw data as depicted in figure-1. Different data mining techniques include classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression etc. Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to classification technique. This approach frequently employs Decision tree based classification Algorithm. In classification, a training set is used to build the model as the classifier which can classify the data items into its appropriate classes. A test set is used to validate the model.

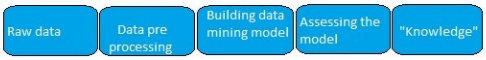


Fig.1: Steps in knowledge extraction

2.1 Data Mining in Banking

Due to tremendous growth in data the banking industry deals with, analysis and transformation of the data into useful knowledge has become a task beyond human ability. Data mining techniques can be adopted in solving business problems by finding patterns, associations and correlations which are hidden in the business information stored in the data bases. By using data mining techniques to analyze patterns and trends, bank executives can predict, with increased accuracy, how customers will react to adjustments in interest rates, which customers are likely to accept new product offers, which customers will be at a higher risk for defaulting on a loan, and how to make customer relationships more profitable[5]. Globalization and the stiff competition had led the banks focus towards customer retention and fraud prevention. To help them for the same, data mining is used. By analyzing the past data, data mining can help banks to predict credible customers. Thus they can prevent frauds, they can also plan for launching different special offers to retain those customers who are credible. Certain areas that effectively utilize data mining in banking industry are marketing, risk management and customer relationship management.

Marketing: It is one of the most widely used areas of data mining in the banking industry. The consumer behavior with reference to product, price and distribution channel can be analyzed by the marketing department. The reaction of the customers to the existing and new products can also be known. This information can be used by the banks to promote the products, improve quality of products and services, and gain competitive advantages. Bank analysts can also analyze the past trends, determine the present demands and forecast the customer behavior of various products and services, in order to grab more business opportunities

Risk Management: It is widely used for managing risks in the banking industry. Bank executives need to know the credibility of customers they are dealing with. Offering new customers credit cards, extending existing customers' lines of credit, and approving loans can be risky decisions for banks, if they do not know anything about their customers. Banks provide loans to their customers by verifying the various details relating to the loan, such as amount of loan, lending rate, repayment period etc. Even though, banks are cautious while providing loan, there are chances of loan repaying defaults by customers. Data mining technique helps to distinguish borrowers who repay loans promptly from those who default.

Customer Relationship Management: Data mining can be useful in all the three phases of a customer relationship cycle such as customer acquisition, increasing value of the customer and customer retention. Customer acquisition and retention are very important concerns of any industry, especially the banking industry. Banks have to cater the needs of the customers by providing the services they prefer. This will ultimately lead to customer loyalty and customer retention. Data mining techniques help to analyze the customers who are loyal from those who shift to other banks for better services. If the customer is shifting from his bank to another, reasons for such shifting and the last transaction performed before shifting can be known, and this will help the banks to perform better and retain their customers.

3. Proposed Model

3.1 Machine learning: Decision Tree

Decision tree algorithm in machine learning methods which efficiently performs both classification and regression tasks[2]. It creates decision trees. Decision trees are widely used in the banking

industry due to their high accuracy and ability to formulate a statistical model in plain language. In Decision tree each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value).

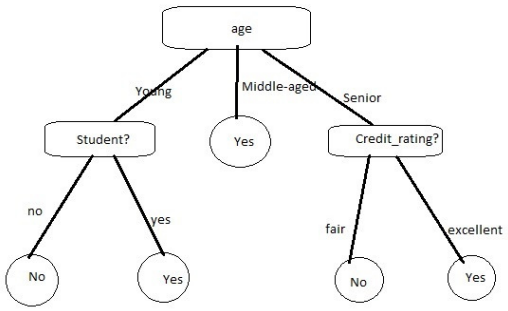


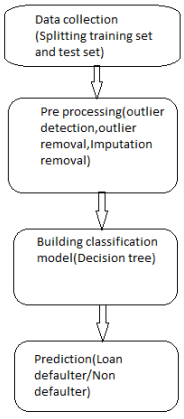
Fig. 2: Decision tree

4. Architecture of proposed model

4.1 Methodology

The methodology adopted for predicting loan Defaulters using Decision tree Technique is derived using a flow diagram.

The steps involved in Building the data model is depicted below:



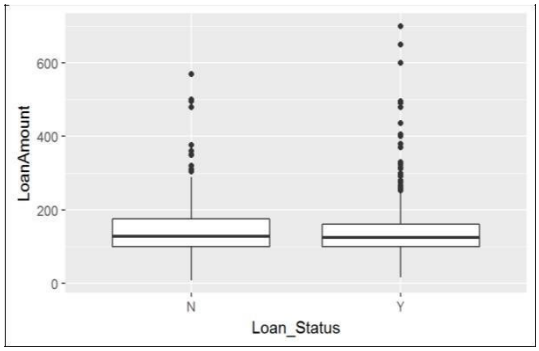
4.2 Data Collection

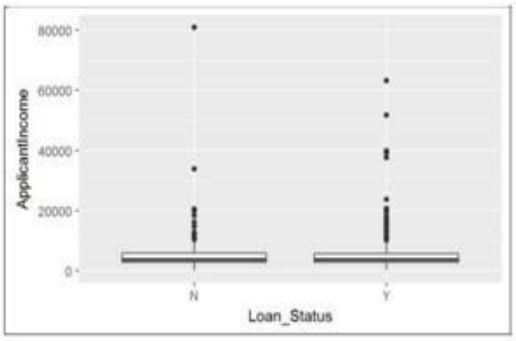
The dataset collected for predicting loan default customers is predicted into Training set and testing set. Generally 80:20 ratio is applied to split the training set and testing set. The data model which was created using Decision tree is applied on the training set and based on the test result accuracy, Test set prediction is done. Following are the attributes

Atrribute Name	Category
Loan_ID	Qualitative
Gender	Categorical
Married	Categorical
Dependents	Qualitative
Education	Categorical
Self_Employed	Categorical
ApplicantIncome	Qualitative
CoapplicantIncome	Qualitative
LoanAmount	Qualitative
Loan_Amount_Term	Qualitative
Credit_History	Qualitative
Property_Area	Categorical

4.3 Pre processing

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency o the algorithm[4]. The outliers have to be removed and also variable conversion need to be done. In order to overcoming these issues we use map function.





4.4 Correlating attributes

Based on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can include Property area, education, loan amount, and lastly credit History, which is since by intuition it is considered as important. The correlation among attributes can be identified using corplot and boxplot in Python platform[1] .

4.5 Building the classification model using Decision tree algorithm [2]

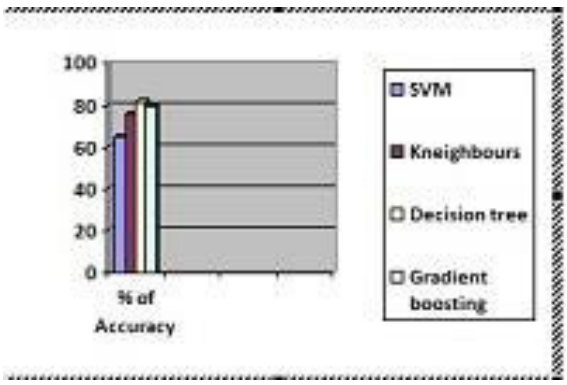
For predicting the loan defaulter's and non defaulter's problem Decision tree algorithm is used. It is effective because it provides better results in classification problem. It is extremely intuitive, easy to implement and provide interpretable predictions. It produces out of bag estimated error which was proven to be unbiased in many tests. It is relatively easy to tune with. It gives highest accuracy result for the problem.

4.6 Predicting default outcomes

```
Predictions<--
unname(predict(final_dt,newtest[]))
solution<-
data.frame(Loan_ID=test[1],Loan_Status=pr
edictions##0.811053
```

We noticed that 299 cases in the test set are predicted as "Y", which is more than 81%, whereas in the training set only about 69% had this status[6].

5. Experimental Results



6. Conclusion

The analytical process started from data cleaning and processing, Missing value imputation with micepackage, thenexploratory analysis and finally model building and evaluation. The best accuracy on public test set is 0.811. This brings some of the following insights about approval.

Applicants with Credit history not passing fails to get approved, Probably because that they have a probability of a not paying back. Most of the Time, Applicants with high income sanctioning low amount is to more likely get approved which make sense, more likely to pay back their loans. Some basic characteristic gender and marital status seems not to be taken into consideration by thecompany.

REFERENCES

[1]Cowell,R.G.,A.P.,Lauritez,S.L.,and Spiegelhalter,D.J.(1999). Graphical models and Expert Systems. Berlin: Springer. This is a good introduction to probabilistic graphical models.

[2] Kumar Arun, Garg Ishan, Kaur Sanmeet, May-Jun. 2016. Loan Approval Prediction based on Machine Learning Approach, IOSR Journal of Computer Engineering (IOSR-JCE)

[3] Wei Li, Shuai Ding, Yi Chen, and Shanlin Yang, Heterogeneous Ensemble for Default Prediction of Peer-to-Peer Lending in China, Key Laboratory of Process Optimization and Intelligent Decision-Making, Ministry of Education, Hefei University of Technology, Hefei 23009, China

[4] Clustering Loan Applicants based on Risk Percentage using K-Means Clustering Techniques, Dr. K. Kavitha, International Journal of Advanced Research in Computer Science and Software Engineering.

[5] Research on bank credit default prediction based on data mining algorithm, The International Journal of Social Sciences and Humanities Invention 5(06): 4820-4823, 2018.

[6] Short-term prediction of Mortgage default using ensembled machine learning models, Jesse C. Sealand on july 20, 2018.