

Loan Credibility Prediction System using Data Mining Techniques

Anuja Kadam¹, Pragati Namde², Sonal Shirke³, Siddhesh Nandgaonkar⁴, Dr.D.R. Ingle⁵

^{1,2,3,4}Student, Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

⁵Head of Department of Computer Engineering, Bharati Vidyapeeth College of Engineering, Navi Mumbai, Maharashtra, India

Abstract - As we know that now-a-days there is a rapid growth in banking sector, resulting lots of people are applying for bank loans. Finding out the applicant to whom the loan will be approved is a difficult process. Data mining techniques are becoming very popular nowadays because of the wide availability of huge quantity of data and the need for transforming such data into knowledge. Techniques of data mining are implemented in various domains such as retail industry, telecommunication industry, biological data analysis, etc. In this paper, we proposed a model which predicts loan approval/rejection of an applicant using data mining techniques. This can be done by training the model with the data of the previous records of the people applied for loan.

Key Words: Logistic Regression, Data Mining, Classification, Credibility, Loan, Prediction.

1. INTRODUCTION

The prime goal in banking domain is to invest their assets in safe hands. Lending money to unsuitable loan applicants results in the credit risk. Today many banks approve loans after a long procedure of verification, yet there is no guarantee whether the picked candidate is the right candidate or not. Estimating the risk, which is involved in a loan application, is one of the most significant concerns of the banks in order to survive in the highly competitive market.

Data mining algorithms are used to study the loan-approved data and exact patterns, which would help in predicting the reasonable defaulters, thereby helping the banks for making better choices in the future. Loan Prediction is extremely useful for employee of banks and for the applicant also. The main aim of this model is to provide a speedy, immediate and simple approach to pick the deserving applicants.

In [1] the author acquaints a structure to successfully recognize the Probability of Default of a Loan applicant. The metrics got from the predictions reveal the high accuracy of the built model. In [2] an effective model was proposed for predicting the right customers who have applied for loan. Decision Tree is applied to foresee the traits significant for believability. The model proposed in [3] has been built using data from banks to predict the status of loans. This model uses three classification algorithms namely j48, bayes Net

and naive Bayes. The model was implemented using Weka. In [4] a decision tree model was utilized as a classifier and for feature selection genetic algorithm is utilized. The model was tried utilizing Weka. In [5] two data mining models were created for credit scoring that helps in decision making of giving loans for the banks in Jordan. With the consideration of accuracy rate, the regression model is found to perform better than radial function model. The work in [6] analyses support vector machine-based models for credit-scoring created using the different default definitions. The work inferred that the expansive definition models are better than the narrow definition models in their performance. In [7] financial data analysis was done by figuring out techniques like Decision Tree, Random forest, Boosting, Bayes classification, Bagging algorithm etc. Techniques like Support Vector Machine, Decision Tree, Logistic Regression, Neural Network, Perception model are combined in this model. The accuracy rate of each of these techniques is studied. The analysis results show the performance is extraordinary based on accuracy.

2. LITERATURE SURVEY

A. Data Mining

Data Mining is the process of examining underlying and potentially useful patterns in big chunks of source data. Similar to precious-stone mining, in statistics analysts extract fragments of potentially useful information from the deep recesses of database systems. Here a goal is set to discover connections between the informational streams that weren't perceived previously. Data mining has other names: knowledge discovery, information harvesting, etc.

Data Mining employs machine learning (ML), artificial intelligence (AI), statistical information, and database technological insights.

The purpose of data mining is twofold: the creation of predictive power using the current information for predicting future values, finding descriptive power for a better description of patterns in the present data.

B. Data Mining in Banking Sector

There are numerous areas in which data mining can be used in the banking industry, which include customer segmentation and profitability, credit scoring and approval, predicting payment default, marketing, detecting fraudulent transactions, cash management and forecasting operations,

optimizing stock portfolios, and ranking investments. In addition, banks may use data mining to identify their most profitable credit card customers or high-risk loan applicants. To help bank to retain credit card customers, data mining is used. By analysing the past data, data mining can help banks to predict customers that are likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers. Credit card spending by customer groups can be identified by using data mining.

3. PROPOSED SYSTEM

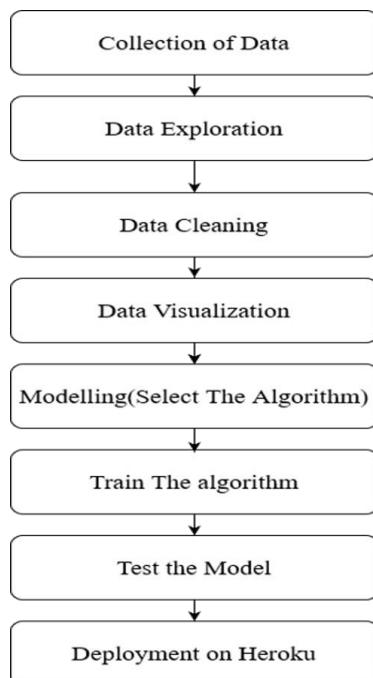


Fig-1: Implementation of Proposed Model

This problem is a supervised classification problem because we need to predict whether the "Loan Status" of customers are either "Yes" or "No".

This can be solved with any of the algorithm listed below:

- Logistic regression
- Decision tree
- Random Forest

The above-listed algorithm is a few of the algorithms that can be used to solve this problem.

1. Collection of Data

Two data sets are given, one is the training data set and the other is the testing data set. Columns in the data sets are as shown in the table.

Table -1: Columns

Features	Descriptions
Loan_ID	Unique Loan ID
Gender	Male/Female
Married	Married(Yes)/not married(No)
Dependents	Number of dependents
Education	Education(Graduate/Not graduate)
Self_Employed	Self employed(Yes/No)
ApplicantIncome	Applicant income
CoapplicantIncome	Co-applicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	Credit history meets guidelines
Property_Area	Urban/semi urban/rural
Loan_Status	Loan approval(Y/N)

2. Data Exploration

All the packages needed to explore the data were imported. Then the data was explored by checking the first few columns and rows

After which the summary of the statistics, the missing values, number of rows and columns was checked.

3. Data Cleaning

After exploring the data set, some missing values were found which had to be filled using the median for numerical data and for categorical data, mode instead of mean.

In [8]: `train.isnull().sum()`

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status	dtype: int64
Out[8]:	0	13	3	15	0	32	0	0	22	14	50	0	0	0

After filling out the missing data, the data is found to be cleaned.

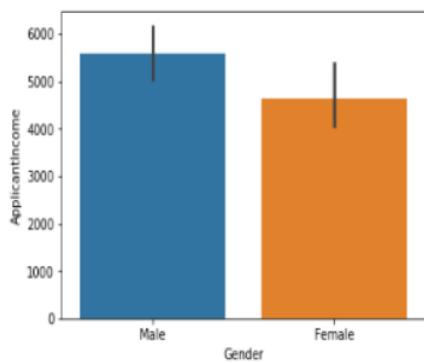
```
In [24]: train.isnull().sum()
```

```
Out[24]: 
Loan_ID      0
Gender       0
Married      0
Dependents   0
Education    0
Self_Employed 0
ApplicantIncome 0
CoapplicantIncome 0
LoanAmount    0
Loan_Amount_Term 0
Credit_History 0
Property_Area 0
Loan_Status    0
dtype: int64
```

3.Data Visualization

The value count was done and then data was visualized.

```
Out[36]: <matplotlib.axes._subplots.AxesSubplot at 0x2900e2ebbe0>
```



Even after the data analysis, there is still no unique factor to determine loan status. Categorical data was converted into numerical data.

4.Data Modelling

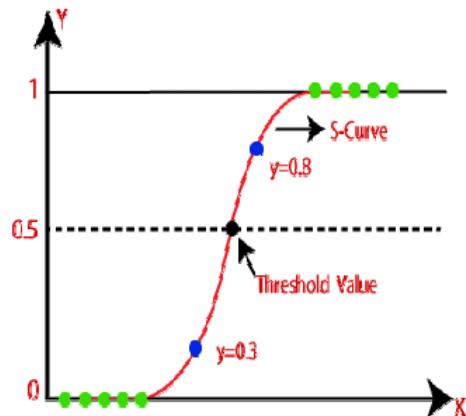
After Data Visualization, the data is modelled/trained. For this the packages of three algorithms (Logistic regression, Decision tree and Random forest) were imported. The model was then defined and the accuracy score was evaluated.

Logistic Regression was the best fit with highest accuracy score 81.12%.

It is applicable for categorical dependent variable using a given set of independent variables. Thus, the outcome must be a categorical or discrete value. The output can be either Yes or No, 0 or 1, true or false, etc. But instead of giving the exact value as 0 or 1, it gives some probabilistic values which lies between 0 and 1. In Logistic regression, rather than fitting a regression line, we fit an "S" shaped logistic function, which predicts two greatest values (0 or 1). The curve from the logistic function demonstrates the probability of something, for example, regardless of whether the cells are destructive or not, a mouse is corpulent or not founded on its weight, and so on. It is a significant algorithm because it can provide probabilities and classify the use of different types of data and easily determines the most effective variables that

are used for classification. The S-structure curve is also known as the sigmoid function or the logistic function.

$$\log(1/(1-y)) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_nx_n$$



Graph-1: Logistic Function

5.Implementation using Logistic Regression

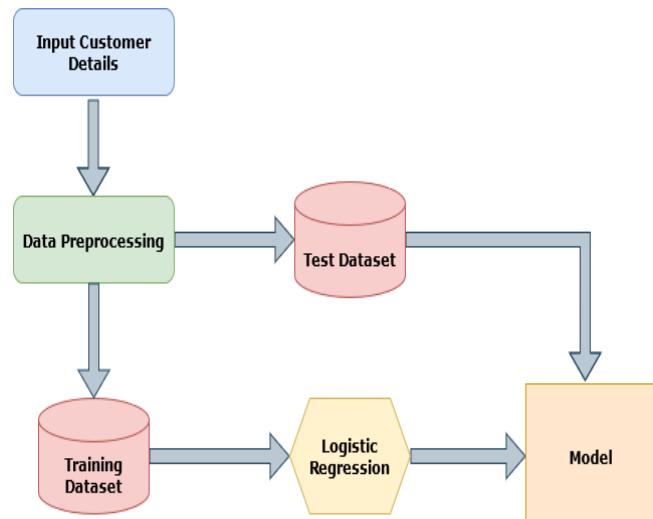


Fig -2: Implementation of Logistic Regression

The Logistic Regression was used to fit the test data set and the Prediction result was displayed Successfully.

4. EXPERIMENTAL RESULTS

Based on the data given by the loan applicant, we can predict whether the loan of particular applicant is approved or not using a User Interface. User interface contains input variables with their corresponding fields and a field to display the output. Input variables are Gender, Marital status, Dependents, Education, Applicant income, Loan Amount, Loan amount term, Credit History, Property Area. The applicant needs to give these values and based on these, the model will predict whether the loan will be approved or not.

Fig -3: User Interface

5. CONCLUSIONS

Finally, in our model by using logistic regression model we predict whether the loan is approved or not. In order to implement this various input variables were used to get the output. Whenever program takes the input data it gives the output in the form of binary i.e., either 0 or 1. If the output is 1 then '1' will be displayed and it indicates that loan is approved. If the output is 0 then '0' will be displayed and it indicates that loan is not approved.

Here, we had implemented loan credibility prediction system that helps the organizations in making the right decision to approve or reject the loan request of the customers.

In this model, Logistic Regression algorithm is used for the prediction. Incorporation of other techniques that outperform the performance of popular data mining models has to be implemented and tested for the domain.

REFERENCES

- [1]. Sudhamathy G and Jothi Venkateswaran "Analytics Using R for Predicting Credit Defaulters", IEEE international conference on advances in computer applications (ICACA), 978-1-5090-3770-4, 2016.
- [2]. M. Sudhakar, and C.V.K. Reddy, "Two Step Credit Risk Assessment Model for Retail Bank Loan Applications Using Decision Tree Data Mining Technique", International Journal of Advanced Research in Computer Engineering &Technology (IJARCET), vol. 5, no.3, pp. 705-718, 2016.
- [3]. J.H. Aboobida, and M.A. Tarig, "Developing Prediction Model of Loan Risk in Banks Using Data Mining", MachineLearning and Applications: An International Journal (MLAIJ), vol. 3, no.1, pp. 1–9, 2016.
- [4]. 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.
- [5]. A.B. Hussain, and F.K.E. Shorouq, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach", Review of Development Finance, Elsevier, vol. 4, pp. 20–28, 2014. JAC: A JOURNAL OF COMPOSITION THEORY Volume XIII, Issue V, MAY 2020 ISSN: 0731-6755 Page No: 324
- [6]. T. Harris, "Quantitative credit risk assessment using support vector machines: Broad versus Narrow default definitions", Expert Systems with Applications, vol. 40, pp. 4404–4413, 2013.
- [7]. 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.