Loan Eligibility Prediction using Machine Learning

Nishanth S.M
CSE Department, PES University
Bangalore, India
nishanthsm01@gmail.com

Rajath R Maragiri
CSE Department, PES University
Bangalore, India
rajathmaragiri8@gmail.com

Sahil Elton Lobo
CSE Department, PES University
Bangalore, India
elton.lobo21@gmail.com

Sinchan Samajdar
CSE Department, PES University
Bangalore, India
sinchan.samajdar@gmail.com

***Abstract*—A bank's profit or loss depends to a large extent on loans i.e. whether the customers are paying back the loan or defaulting. This project deals with deciding if a person, or organisation, applying for a loan, is eligible for that loan, i.e., if they are likely to be able to repay the bank. This can be useful in reducing the time and manpower required to approve loans and filter out the perfect candidates for providing loans.**

***Keywords*—SVM, Machine Learning, Loan Prediction**

I. INTRODUCTION

A monetary loan by definition, is when one or more persons, organizations, or other entities lend money to other people, organizations, or entities, in our case, a possible loan candidate, borrowing money from the bank. The recipient incurs a debt for which he or she is generally responsible for paying interest until the loan is repaid along with the principal amount borrowed. To figure out if a certain candidate is eligible for a loan, i.e., if he or she is capable of repaying the amount, is a tedious task as there can be many factors involved in making this decision. Furthermore, the acceptance or rejection of any loan application has a direct impact on the bank revenue and the profitability in quarterly issued financial statements. Loan approvals are critical processes, and for the majority of the times, it is not a straightforward procedure, and can sometimes return inaccurate results. Our project seeks to provide a solution to this overdue problem, by considering the various factors that can affect a candidate's ability to repay a loan and thereby checks, or predicts, if they will be eligible to take on a certain monetary loan from the bank. This task of checking the background and factors of each candidate, if done manually, is extremely time consuming and counter productive. Our project automates this task, and attempts to find an accurate, optimised solution to this existing problem.

II. PREVIOUS WORK

Before zeroing in on models that would best suit this problem statement we read up on various already existing research papers that were written on this particular subject.

These models help us understand the already existing models and their shortcomings in various situations, and possible ways to counter these issues.

Stated below are some research papers and their different approaches to solve similar issues:

[I]In paper [1] written by MrunalSurve, PoojaThitme, PriyaShinde, Swati,Sonawane, Sandip Pand They've explained "Data Mining Techniques To Analyze Risk Giving Loan(Bank)" in detail.The typical loan system requires a significant amount of manual labour, which is a challenging undertaking that can lead to errors and dangers.

The program's main goal is to calculate a credit rating by analysing the risk ratio.Overall, this paper presents a novel method for assessing risk while providing a loan .

[II]In paper [2] written by AboobydaJafar Hamid and Tarig Mohammed Ahmed discuss "Developing Prediction Model Of Loan Risk In Banks Using Data Mining". This model was developed utilising a data banking system to foreclose loans.The suggested model is built using three different methods: j48, Bayes Net, and naive Bayes.The sample is processed and evaluated using the Weka programme.

After the deployment of the data mining techniques protocol J48, Bayes Net, and naive Bayes, the J48 algorithm is the best algorithm of the credit class.The outcome is based on precision.

[III]In Paper [3]"A Study of Classification Based Credit Risk Analysis Algorithm," by Ketaki Chopde, Pratik Gosar, Paras Kapadia, Niharika Maheshwari, and Pramila M. Chawan.

They discuss credit score modelling to categorise loan applicants into two classes: "Good Credit" class liability to lend financial obligation and "Bad Credit" class liability to lend financial obligation.

Financial organisations can enhance the amount of credit they give while lowering possible losses by accurately judging applicants' credit qualifications.

Then, talk about the various decision tree implementation approaches for credit risk analysis.

This dataset has been selected with the assumption that the information provided regarding the users interested in taking a loan is accurate and verified.The scope of the problem we are trying to solve is for all types of banks that offer credit lines to help them automate the process.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

III. PROPOSED SOLUTION

The Model we've built is a model that evaluates the eligibility of a person for a loan. The steps in the model are to find out whether or not a loan can be granted. Loan Officers look for three major things while considering a loan, Credit history, Cash flow and the validity of collateral offered by the loanee. Our framework takes these factors into consideration.

A. Abbreviations used

NO-Nominal Data
NU-Numeric Data
SVM- Support Vector Machine
RF- Random Forest
KNN- k Nearest Neighbours

B. Dataset

The Dataset we have used as reference for the model is taken from the loan industry. The format of our dataset is a Standard CSV(comma separated values) file.
After applying the standard preprocessing and normalization methods, we get the following data:

	Attribute	Description	Data type
1	Gender	Gender of the loanee	NO
2	Married	Marital Status of the loanee	NO
3	Dependents	Number of dependents on the loanee.	NU
4	Education	Educational qualifications of the loanee	NO
5	Self_Employeed	Employment status of the loanee	NO
6	Property_Area	Locality of the loanee's property	NO
7	Loan_Status	Loan approval status	NO

C. Detailed Solution

1. All the columns with long names were renamed to shorter, more meaningful names.
2. The Null (Select) values were treated properly. Also, less significant counts were grouped together
3. After performing all the above exploratory data analysis, ensured that there were no missing values in any of the columns.
4. A total of 13 attributes.
5. As part of outlier treatment for continuous variables, the outliers were replaced by median =values of the column.

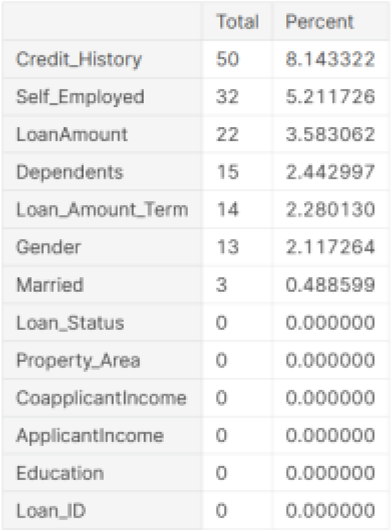
Preprocessing :

A. Data Cleaning:

● Dataset comprises 614 observations and 13 characteristics.

● Out of which one is dependent variable and rest 12 are independent variables

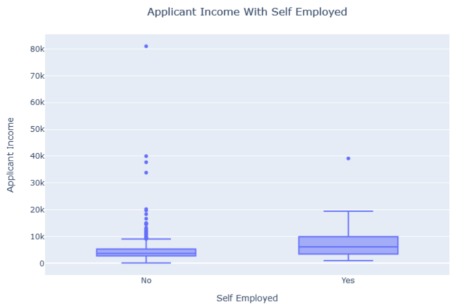
Missing Data

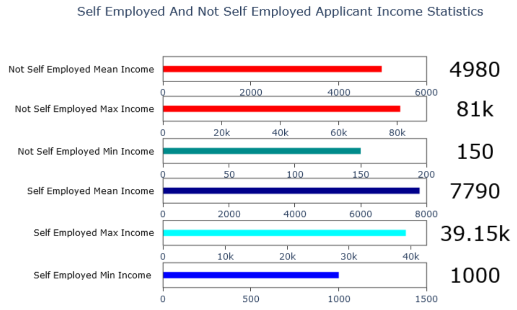


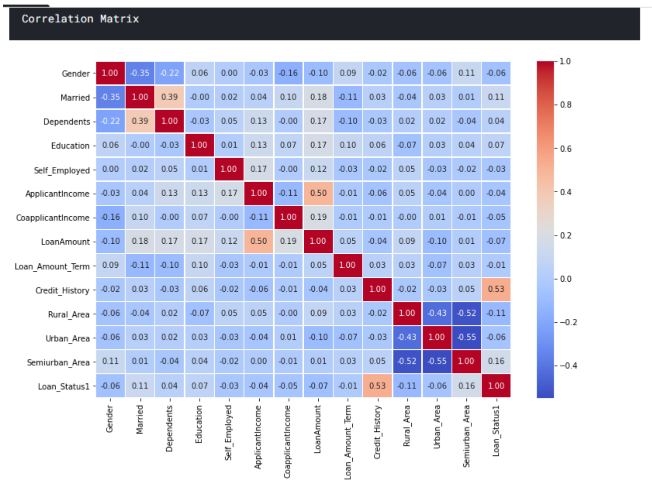
	Total	Percent
Credit_History	50	8.143322
Self_Employed	32	5.211726
LoanAmount	22	3.583062
Dependents	15	2.442997
Loan_Amount_Term	14	2.280130
Gender	13	2.117264
Married	3	0.488599
Loan_Status	0	0.000000
Property_Area	0	0.000000
CoapplicantIncome	0	0.000000
ApplicantIncome	0	0.000000
Education	0	0.000000
Loan_ID	0	0.000000

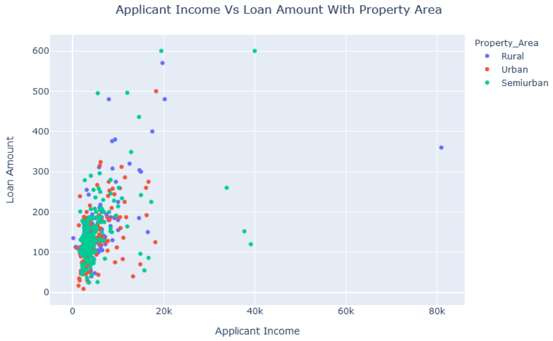
We have dropped all null values.

B. Exploratory data analysis.







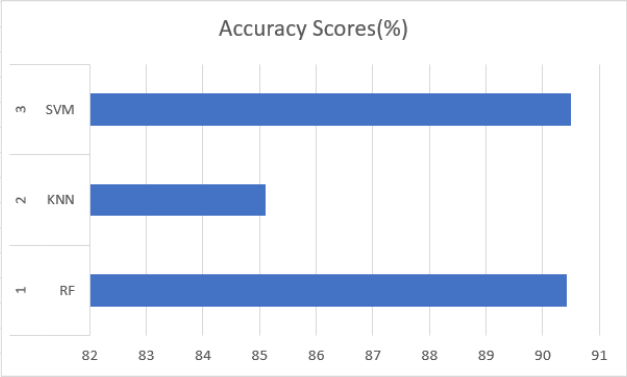


6. Data was then split into a train and test set with the 70-30 ratio.
7. For the train set, Min max scaler was applied to standardize the values for all numerical variables
8. After this we ran the dataset through a series of ML models that included SVM, RF and KNN, to extrapolate which model best suits our problems needs.
9. Based on the experimental results found below SVM was considered to be best fitting, as the accuracy was found to be 90.5%

IV. Experimental Results

1. Top variables which contribute most towards loan eligibility are:
   a. Credit History
   b. Employment
   c. Income
   d. Education
   e. Loan amount
   f. Loan term
2. Strategy used for the Model:
   -Main focus is on Credit History where the credit score should be in the Range of 630-730.
   -Income criteria is taken as a contingent with Loan Amount. These two categories go hand in hand, as the Loan amount is only relevant with respect to the Income of the Loanee.
3. Here are the results of the models tested for the dataset used:
   a. Null values were handled so that no rows were dropped.
   b. Outliers were replaced with median values.
   c. Using the models we got the following accuracy scores:

	Model	Accuracy Scores(%)
1.	RF	90.4255
2.	KNN	85.106
3.	SVM	90.5



V. Conclusion

In conclusion, it is safe to say that the SVM model is the most accurate to predict whether a person is eligible for a loan. One of the most important aspects of the Loan industry is gauging whether or not a person is eligible for a certain loan amount. Our model incorporates the various attributes available in our dataset taking into consideration the correlated attributes effectively.

In contrast to the other tested models such as RF and KNN, SVM ensures that there is a reduction in the overfitting of the dataset which results in higher accuracy prediction.

Our model can be of great use to Loan companies, as it automates the process of Loan eligibility which helps with the retrieval of credit lines and other assets which would otherwise be categorised as dead assets causing losses in banks.

References

[1] MrunalSurve, PoojaThitme, PriyaShinde, Swati Sonawane,
Sandeep Pandit, “DATA MINING TECHNIQUES TO ANALYSES RISK GIVING LOAN(BANK)”, IJARIIE-ISSN(O)-2395-4396 Vol-2 Issue-1 2016.

[2] AboobydaJafar Hamid and Tarig Mohammed Ahmed, “DEVELOPING PREDICTION MODEL OF LOAN RISK IN
BANKS USING DATA MINING”, Machine Learning and Applications: An International Journal (MLAIJ) Vol.3, No.1,
March 2016.

[3] KetakiChopde, Pratik Gosar, ParasKapadia, NiharikaMaheshwari,
Pramila M. Chawan, “A Study of Classification Based Credit
Risk Analysis Algorithm”, International Journal of Engineering
and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-
1, Issue-4, April 2012.