International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7 Issue-4s, November 2018

An Exploratory Data Analysis for Loan Prediction Based on Nature of the Clients

X.Francis Jency, V.P.Sumathi, Janani Shiva Sri

Abstract -In India, the number of people applying for the loans gets increased for various reasons in recent years. The bank employees are not able to analyse or predict whether the customer can payback the amount or not (good customer or bad customer) for the given interest rate. The aim of this paper is to find the nature of the client applying for the personal loan. An exploratory data analysis technique is used to deal with this problem. The result of the analysis shows that short term loans are preferred by majority of the clients and the clients majorly apply loans for debt consolidation. The results are shown in graphs that helps the bankers to understand the client's behaviour.

Keywords - Loan analysis, exploratory data analysis technique, client's analysis, financial categories analysis

I.INTRODUCTION

The term banking can be defined as receiving and protecting money that is deposited by the individual or the entities. This also includes lending money to the people which will be repaid within the given time. Banking sector is regulated in most of the countries as it is the important factor in determining the financial stability of the country. The provision of banking regulation act allows public to obtain loans.Loans are good sum of money borrowed for a period and expected to be paid back at given interest rate. The purpose of the loan can be anything based on the customer requirements. Loans are broadly divided as open-ended and close-ended loans. Open-ended loans are the loans for which the client has approval for a specific amount. Examples of open-end loans are credit cards and a home equity line of credit (HELOC). Close-ended loans decreases with each payment. In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, instalment loan and student loans are the most common examples of close-ended loans. Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan.

Unsecured loans are also known as personal or signature loans. Here the lender believes that the borrower can repay the loan based on financial resources possessed by the borrower. Liquidity risk is the risk that arises from the lackof marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss.

Manuscript published on 30 November 2018.
*****Correspondence Author(s)
Ms.X.Francis Jency, CSE Department, Kumaraguru College of Technology, Coimbatore, India
Ms.V.P.Sumathi, CSE Department, Kumaraguru College of Technology, Coimbatore, India
Janani Shiva Sri,C S Department, Kumaraguru College of Technology, Coimbatore, India

The interest rate risk is the risk in which the interest rates priced on loans will be too low to earn the bank money. The primary objective of the bank is to provide their wealth in the safer hands. In recent times, banks approve loan after verifying and validating the documents provided by the customer. Yet there is no guarantee whether the applicant is deserving or not. This paper classifies the customers based on certain criteria. The classification is done using Exploratory Data Analysis. Exploratory Data Analysis (EDA) is an approach to analyse the datasets that summarizes the main characteristics with visual methods. The purpose of using EDA is to uncover the underlying structure of a relatively larger set of variables using visualizing techniques.

II.LITERATURE SURVEY

In [1] the researchers analyse the data set using data mining technique. Data mining procedure provides a great vision in loan prediction systems, since this will promptly distinguish the customers who are able to repay the loan amount within a period. Algorithms like "J48 algorithm", "Bayes net", Naive Bayes" are used. On applying these algorithms to the datasets, it was shown that "J48 algorithm" has high accuracy (correct percent) of 78.3784% which provides the banker to decide whether the loan can be given to the costumer or not.

In paper [2], "loan prediction using Ensemble technique", used "Tree model", "Random forest", "svm model" and combined the above three models as Ensemble model. A prototype has been discussed in paper [2] so that the banking sectors can agree/reject the loan request from their customers. The main method used is real coded genetic algorithms. The combined algorithms from the ensemble model, loan prediction can be done in an easier way. It is found that tree algorithm provides high accuracy of 81.25%.

In paper [3], using R-language, an improved risk prediction clustering algorithm is used to find the bad loan customers since probability of default (PD) is the critical step for the customers who comes for a bank loan.

So, a frame work for finding PD in the data set is provided by data mining technique. R- Language has the technique called as KNN (K-nearest neighbour) algorithm and it is used for performing multiple imputation calculation when there are missing values seen in the data set.

The paper [4] had used tree model. It helps to find whether the banking sector people will be able to overcome the loan problem with their customers. It provides a high accuracy of 80.87%.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

The paper [5] uses decision tree induction algorithm and found that the algorithm finds a best way to evaluate the credit risk. To avoid the credit risk, bankers holds the

technique called as "credit score", where it helps the lenders to keep note on who are the applicants who will able to repay the amount or probability of going into the default risks. The input given for credit evaluation was customer data, WEKA software, cibil score. The methodology used in prediction system was problem and data understanding, data filtering, system modelling and finally system evaluation. This was done on the banks existing dataset containing 1140 records and 24 attributes. At last the system was tested and helps the bankers to make a correct decision on whether to accept or reject the loan approval.

The paper [6] used predictive model technique and descriptive model technique to predict the loan approval in banks. In predictive model technique, classification and regression were used and in descriptive model technique clustering and association were used. Classifiers also implement several algorithms like naive Bayes, kNN algorithms of R language and regressors implements several algorithms like decision trees, neural networks, etc., To undergo this prediction analysis, out of all these algorithms, naive Bayes produces a most accurate classifier and the algorithms like decision tree, neural network, K-NN algorithms will be more accurate regressors. The main goal of the paper is to predict the loan classification based on the type of loan, loan applicant and the assets (property) that loan applicant holds. It was found that the decision tree algorithm gave an improved accuracy of almost 85% on doing the analysis.

III.LOAN APPLICANT DATA ANALYSIS

Whenever the bank makes decision to give loan to any customers then it automatically exposes itself to several financial risks. It is necessary for the bank to be aware of the clients applying for the loan. This problem motivates to do an EDA on the given dataset and thus analysing the nature of the customer. The dataset that uses EDA undergoes the process of normalisation, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualising the data in the graphical format. Python is used for easy and efficient processing of data. This paper used the pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualisation of the results and for better understanding. For obtaining the graph Matplot library is used.

A. Annual Income Vs Purpose Of Loan

In this Figure 1, the X axis represents the purpose of loan i.e. the purpose for which the loan is applied. Debt consolidation, home improvement is some of the purposes. High, moderate and low represents the annual income of people who fall in the range as below.

Low represents the annual income of people between the range of minimum to 10 lakhs and Moderate represents the annual income of people between 10 lakhs and 25 lakhs and High represents the annual income of people above 25 lakhs. By these criteria, a new column called Category is derived.

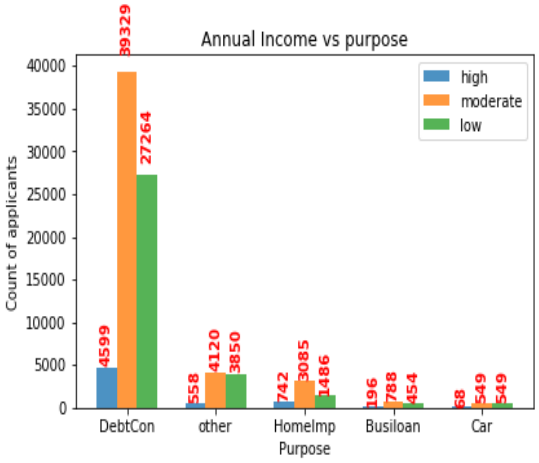

Fig 1. Annual Income vs purpose

Thus, grouping the Category that is high, moderate and low.

Inference from the Figure 1 is as follows:

- People in moderate category seek loan in the higher numbers.
- The field debt consolidation shows the highest distribution.
- Low and moderately categorized applicants try for other purpose and car loans equally.

B. Trust Customer Classification

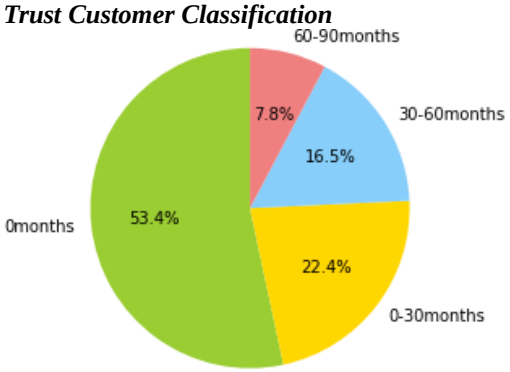

Fig 2. Trust Customers

From the Figure 2 it is inferred as follows:

- There are many customers who does not have delinquents, has applied for the loan which intimates or indirectly conveys that the applicant has some chances to get approval of the loan as the applicant have no delinquents. The result is about 53.3% of applicants.
- And it is also inferred that the number of people applied for loan gradually decreases with the increase in the delinquent months. This shows that, the applicant has minimum chances to get the approval of loans.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

C. *Loan Term Vs Delinquent Months*

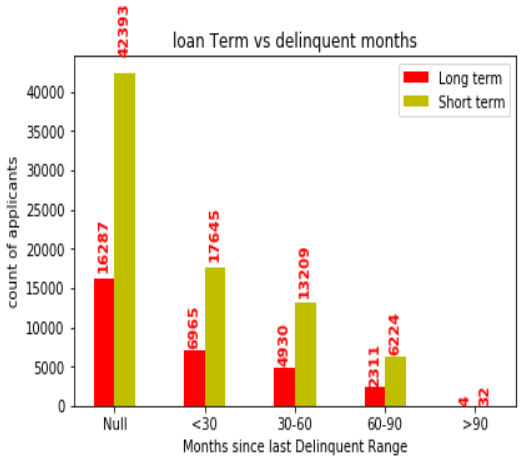

Fig 3. loan term vs delinquent months

- Figure 3 deals with the customers who can pay the loan within term period against customers who cannot repay their monthly depts within the particular term.

From the Figure 3 it can be concluded that:

- This analysis can find a higher number of customers who are able to repay without deliquiates and for short term.
- Almost all applicants who are even delinquent more than 90 months prefers only short term.
- Applicants who delinquent more than 90 months are less in number and it indirectly conveys that their loan will never be sanctioned and if its yes, the applicant will not be able to pay it back

D. Loan Term Vs Credit Category

From the categories poor, fair, good, very good and undefined, the value the credit category is found, as applicants without credit score fall into the category undefined, people have credit score between 300-850 between these there are some categories such as credit score between 300 and 579 falls in poor, credit score between 580 and 669 falls in category fair and the credit score between 670 and 739 is good and above this is considered to be very good.

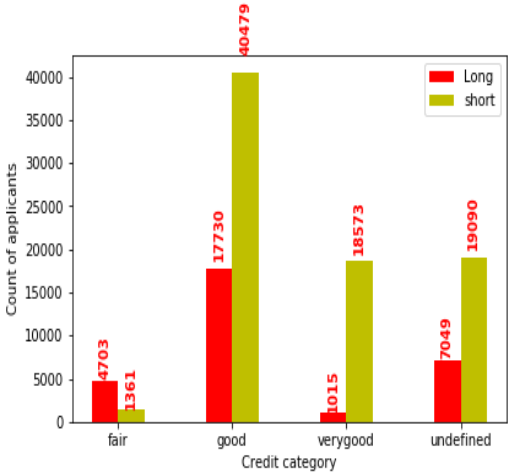

Fig 4. Loan term vs Credit category

Figure 4 spectacles the repayment period of the loan versus credit score under various categories by grouping the derived column credit category and loan term.Figure 4 have deduced the following:

- Customers with good and very good credit score prefer for short term payback period in contradiction to customers with fair credit score.
- People applying loan for first time prefer short term loans because lender doesn't do a credit check so that the applicant avail loans easily.

E. Loan Term Vs Years In Current Job

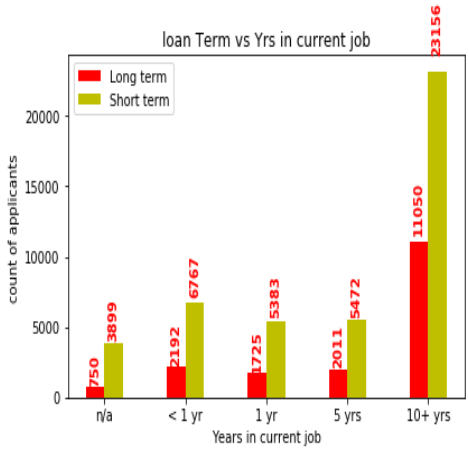

Fig 5. Loan term vs Years in current job

Figure 5 displays the count of applicants who have various years of experience in current job against the term period of repayment of loans.
From the Figure 5 it is concluded that:

- Applicants who have various years of experience in the same job claim the loan for a short period of time.
- Also, the applicants who are freshers claim the loan for a short period of time.
- This also infers that, long term loans are borrowed by people who are yet to start their own business and can repay only after this business set well and brings profit.
- Long term goals carry a greater risk to the money lenders and thus not very easily approved by the banks.

F. Loan Payment Chances Vs Home Ownership

Loan payment chances which have been classified into canpay, maypay and not payable. From Current credit balance, subtracting the monthly debt of that person, one can find whether the person would be able to repay the loan or not. By subtracting, if applicants have less than 50,000 balance, the applicant is considered in the not payable category and applicants having balance between 50,000 and 3 lakhs are considered to be may pay category and above than that the applicant are considered to be can pay category. From the Figure 6 it is concluded that,

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

- People living in rent home fall under not payable category among ownership credentials
- Applicants who have kept home in mortgage apply loan in highest numbers.

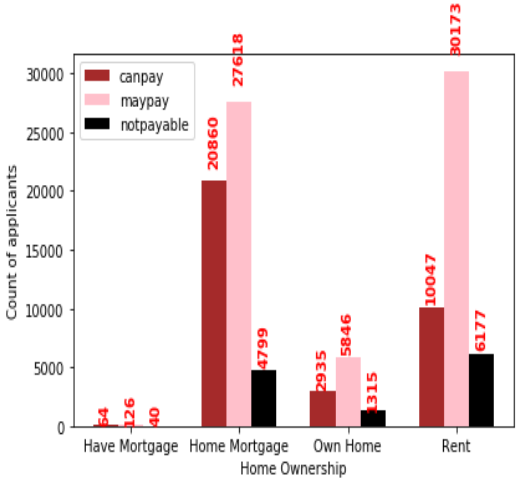

Fig 6. Loan payment chances vs Home ownership

IV.CONCLUSION AND FUTURE WORK

The main purpose of the paper is to classify and analyse the nature of the loan applicants. From a proper analysis of data set and constraints of the banking sector, seven different graphs were generated and visualized. From the graphs, many conclusions have been made and information were inferred such as short-term loan was preferred by majority of the loan applicants and the clients majorly apply loan for debt consolidation.

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. A. Goyal and R. Kaur, "A survey on Ensemble Model for Loan Prediction", International Journal of Engineering Trends and Applications (IJETA), vol. 3(1), pp. 32-37, 2016.
2. A. J. Hamid and T. M. Ahmed, "Developing Prediction Model of Loan Risk in Banks using Data Mining".
3. G. Shaath, "Credit Risk Analysis and Prediction Modelling of Bank Loans Using R".
4. A. Goyal and R. Kaur, "Accuracy Prediction for Loan Risk Using Machine Learning Models".
5. 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(3), pp. 705-718, 2016.
6. Gerritsen, R. (1999). Assessing loan risks: a data mining case study. IT professional, 1(6), 16-21.
7. Hsieh, N. C., & Hung, L. P. (2010). A data driven ensemble classifier for credit scoring analysis. Expert systems with Applications, 37(1), 534-545.
8. https://en.wikipedia.org/wiki/Exploratory_data_analysis
9. https://pandas.pydata.org/pandas-docs/stable/
10. https://www.experian.com/blogs/ask-experian/credit-education/score-basics/what-is-a-good-credit-score/

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)