IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

HOME LOAN DATA ANALYSIS, VISUALIZATION AND PREDICTION

Rohan Salvi[1], Rohan Ghule[2], Talib Sanadi[3], Swapnil Waghe[4], Mrinali Bhajibhakare[5]

Computer Engineering Department, Savitribai Phule Pune University

Abstract: Loan prediction is a very common real-life problem that each retail bank faces at least once in its lifetime. If done correctly, it can save a lot of man hours at the end of a retail bank. Customers first apply for a home loan after that company validates the customer eligibility for the loan. The loan eligibility process (real time) can be automated based on customer details provided through for example filling an online application form. These details can be Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and many more. Hence the goal is to identify the customer segments that are eligible and in fact germane for loan amounts so that they can be specifically targeted. In this first Phase, some preliminary operations were carried out to understand all the features clearly. Visualization of all the variables for instance Married, to determine whether the married tend to repay the loans back or the unmarried was done. Similarly Gender, Self Employed, Applicant Income, Credit History, and on Independent Variables like Education etc.. Plotting Box and Bar Plots. All the variables were normalized before Plotting and cross checked if it is Normal using graphs. Some of the Bivariate Analysis Include Loan Status and Education etc. Loan status is our Target Variable. Data cleaning includes dealing with missing values Mean (without outliers) and Mode (with Outliers) Replacement Method. Also Correlation is visualized using Heat map. In the Upcoming Phases, the plan is to progress further by performing feature engineering, Model Building using regression etc..

Keywords **- Data Analysis, Visualization, Machine Learning, Loan Prediction, Retail Banking, Validation, Eligibility.**

I. INTRODUCTION

A Finance company deals with home loans and has its reach spread across all urban, semi-urban and rural areas. But giving a loan is a tedious process for a company as various screening and eligibility criteria have to be cross checked before a loan can be granted. This process is necessary to avoid frauds and reduce company losses. The company wants to process the loan eligibility checking to get automated. Eligibility is based on customer details provided while filling an application form,

II. LITERATURE SURVEY

In the research paper named An Approach for Prediction of Loan Approval using Machine Learning Algorithm the authors M. A. Sheikh, A. K. Goel and T. Kumar state that Logistic Regression models have been performed and the different measures of performances are computed. By using a logistic regression approach, the right customers to be targeted for granting loan can be easily detected by evaluating their likelihood of default on loan. The model concludes that a bank should not only target the rich customers for granting loan but it should assess the other attributes of a customer as well which play a very important part in credit granting decisions and predicting the loan defaulters.

In the research paper named Loan Default Prediction with Machine Learning Techniques the author L. Lai states that Loan business is one of the major income sources for banks. However, loan default problems are a major issue for loan business. With the rise of the big data era and the development of machine learning techniques, nowadays we have more options for classifying and predicting loan default, other than manual processing. Models including XGBoost, random forest, k nearest neighbors, and multilayer perceptions. Our result shows the promising application of machine learning techniques in the financial industry.

In the research paper named Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval the author A. Vaidya proposes that Decision taking is attained by probabilistic and predictive approaches developed by various machine learning algorithms. This paper discusses logistic regression and its mathematical representation. This paper adheres to logistic regression as a machine learning tool in order to actualize the predictive and probabilistic approaches to a given problem of loan approval prediction. Using logistic regression as a tool, this paper specifically delineates about whether or not loan for a set of records of an applicant will be approved. Furthermore, it also discusses other real-world applications of this machine learning mode.

Sr no.	Title Of Paper	Technique used	Efficiency	Future Scope
1	Approach for Prediction of Loan Approval	Logistic Regression	81 % (best case)	Can improve using better data processing, and ML algorithms.
2	Loan Default Prediction with Machine Learning Techniques	XGBoost	NA	Hybrid Model can be used to improve accuracy
3	Predictive and Probabilistic approach using logistic regression: Application to prediction of loan approval	Decision Tree, Probabilistic Approach	0.77	Can improve using better dataset, data cleaning, and ML algorithms.

III. PROPOSED SYSTEM

This is a standard supervised classification task. This kind of problem requires prediction of whether a loan would be approved or not. Discrete values based on a given set of independent variables are to be predicted. Loan prediction is a real-life problem that every retail bank faces at least once in its lifetime. If done correctly, it can save a lot of man hours.

First, listing out most important factors which may have an effect on Loan Approval:

Salary: High income means more chances of loan approval.

Previous history: Applicants who have repaid their previous debts would be considered a better client.

Loan amount: Lower the loan amount, higher are the chances of loan approval.

Loan term: Loan for less time period and less amount shall have higher the chances of loan approval.

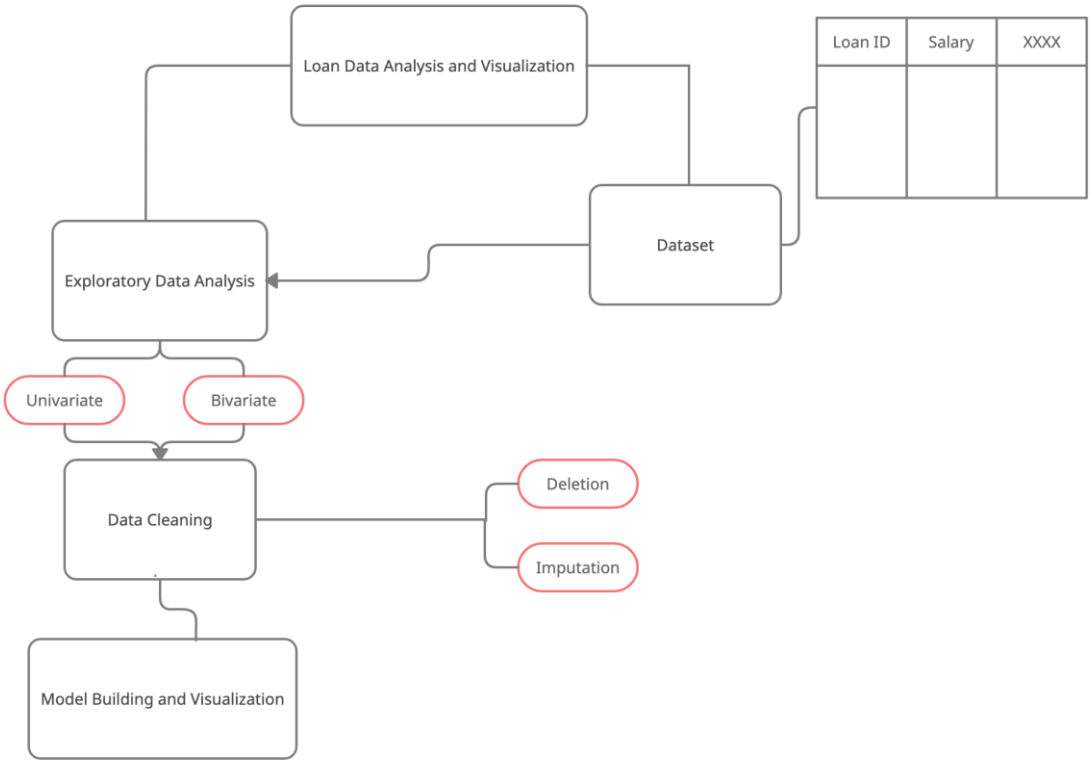
EMI: Lesser is the EMI amount, higher the chances of loan approval.

These are some of the factors which can affect the loan approval or the Target variable, there might be many more factors.

Hardware and Software Requirements:

1. Python 3
2. Jupyter Notebook
3. Numpy
4. Pandas
5. Seaborn
6. Matplotlib
7. Windows / Mac OS / Linux
8. 2 Ghz Intel/AMD processor with 2 cores and 4 GB RAM (Minimum)
9. Microsoft Excel

Architecture:



Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan.

The Company wants to automate the loan eligibility process (real time) 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.

It's a classification problem, given information about the application we have to predict whether they'll be to pay the loan or not. We'll start by exploratory data analysis, then pre-processing, and finally we'll be testing different models such as Logistic regression and decision trees. We'll be using seaborne for visualization and pandas for data manipulation. We'll import the necessary libraries and load the data. We can look at few top rows using the head function. We can see that there's some missing data, we can further explore this using the pandas describe function. For numerical values a good solution is to fill missing values with the mean, for categorical we can fill them with the mode (the value with the highest frequency).

Next we have to handle the outliers, one solution is just to remove them but we can also log transform them to nullify their effect which is the approach that we went for here. Some people might have a low income but strong CoappliantIncome so a good idea is to combine them in a Total Income column. We're going to use sklearn for our models. To try out different models we'll create a function that takes in a model, fits it and measures the accuracy which means using the model on the train set and measuring the error on the same set.

IV. Analysis And Design

Given below is the description for each variable.

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)

Loan_ID: Unique Loan ID

Gender: Male/ Female

Married: Applicant married (Y/ N)

Dependents: Number of dependents

Education: Applicant Education (Graduate/ Undergraduate)

Self_Employed: Self-employed (Y/ N)

ApplicantIncome: Applicant income

CoapplicantIncome: Co Applicant income

LoanAmount: Loan amount in thousands

Exploring the dataset:

This project works with two datasets - training dataset and a test dataset. Firstly, observing what all columns are there in the datasets and what are their data types.

Object: Object type denotes that variables are categorical. Categorical attributes in our dataset are - Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status.

Int64: Denotes integer variables. ApplicantIncome is the only attribute of type int64.

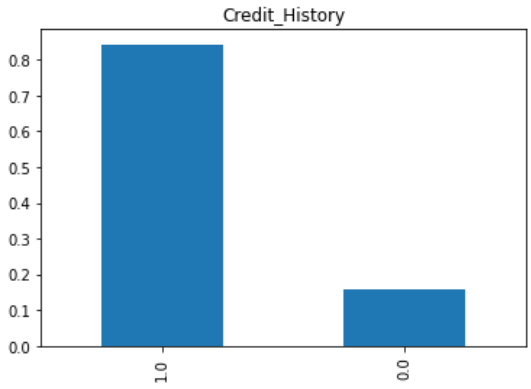
Float64: Denotes decimal valued or numeric variables. Floating attributes in our dataset are - CoapplicantIncome, LoanAmount, Loan_Amount_Term and Credit_History.

Next, observing the shape of the datasets. Training dataset has 614 rows and 13 columns on the other hand Test dataset has 367 rows and 12 columns.

Exploratory Data Analysis:

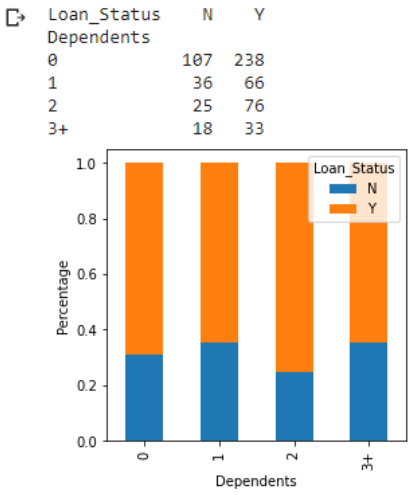
1. **Univariate:**

Univariate analysis is the simplest form of analyzing data where each variable is to be examined individually including the target variable 'loan_status'. Among 614 cases, Loan_Status value came out to be accepted for 422 cases i.e. approved for around 69% and rejected for 192(31.27%). We converted categorical variables into quantitative ones for the visualization purpose. Univariate analysis gave us information like male to female count of loan applicants etc. We have 489 male, 112 female, 398 married, 213 unmarried, 500 self-employed, 82 not self-employed, 475 repaid their debts, 89 have unpaid debts. Independent variables in categorical features include Dependents, Education, and Property_Area. 480 people graduated, 134 did not graduate. 233 live in semi-urban areas, 202 from urban and 179 from rural. 345 have no dependents, 102 have one dependent, 101 have two and 51 have three or more. ApplicantIncome, CoapplicantIncome, LoanAmount, Loan_Amount_Term are independent numerical variables. Credit_History along with ApplicantIncome has the most impact on prediction of Loan Application Approval/Rejection. The image below shows the visualization of the credit history variable in the form of Bar Plot. First we took the count of the "Credit_History" variable of our training dataset and then the value counts, by which we came to know that there are 475 repaid debts and 89 not-repaid debts, after that we normalized the values by converting them to percentages, finally we used these values to plot a bar plot which compares 1 (repaid debts) and 0 (not repaid debts)

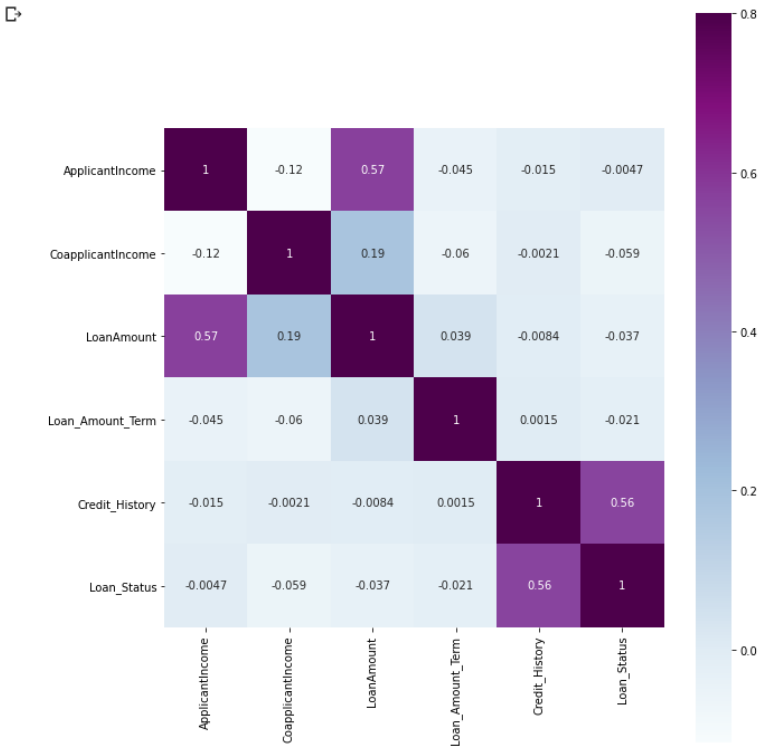


2. Bivariate:

In Bivariate analysis, after individually analyzing every variable in univariate analysis, previously mentioned hypotheses can be tested analyzing every variable again, this time with respect to the target variable. Between variables like, Loan_status vs married, it was observable how marriage influences Loan payment. In Loan_status vs Loan amount it was observable if the amount of money influences loan payment. It was observed that Loan_Status was highly influenced by Credit_History i.e. people who repaid debts were more likely to get approved and otherwise in case of people with pending debts. Further, proportions of loans getting approved for people having low Total Income is very less as compared to other groups. Similarly, higher proportions of loans were approved if loan amount was low or average as compared to high loan amounts. As mentioned above we have cleaned and chosen proper replacements of data for all the operations as a part of Exploratory Data Analysis. In Bivariate Analysis we created a crosstab for creating a table representing data in the Categorical Independent Variable vs Target Variable format, by using this format we represented Gender vs Loan_status, Married vs Loan_status, Education vs Loan_status, Self_Employed vs Loan_status, Credit_history vs Loan_status, Property_Area vs Loan_status, and Dependents vs Loan_status which is shown in the image below.



After the comparison of all the different Categorical Independent variables with the Target variable we created a Heatmap to see the correlation between different variables. We found out that ApplicantIncome and Loan amount, Credit_History and Loan_Status have the highest correlation



Data Cleaning:

1. **Missing Values Treatment:**

Imputation of the missing values and treatment of the outliers is very important, because it can have adverse effects on the model performance. Once the count of the missing values present in every feature were listed, it was observed that missing values were present in the following variables: Gender, Married, Dependents, Self_Employed, LoanAmount, Loan_Amount_Term and Credit_History features. Missing values were replaced with Mode in Gender, Married, Dependents, Credit_History and Self_Employed because count was very small. For loan amount, replacement with median is used. Same approach is used for missing values in the test dataset as well.

2. **Outlier Treatment And Correlation:**

Outliers are taken care by Normalization and log Transformation. Correlation is visualized using Heat Map. It was observed that variables (ApplicantIncome - LoanAmount) and (Credit_History - Loan_Status) were Strongly Correlated.

3. **Feature Engineering:**

Based on domain knowledge, new features can be invented which might affect the target variable. Following are the three new features:

Total Income: Adding ApplicantIncome and Co-ApplicantIncome gives Total_Income which can influence the target variable.

EMI: Idea behind EMI is that applicants with high EMIs to be paid might find it hard to pay back the loan.

Balance Income: Income after the EMI has been paid is called balance income. If this value is high, it increases the chances of loan approval.

Model Building:

Here comes model building. Starting with a logistic regression model we move towards Decision tree model to a complex model like random forest. For making a Logistic Regression Model categorical variables need to be made into dummy variables i.e. into a series of 0 and 1 (as logistic regression takes only numeric values as input) making them easier to quantify and compare. Train data trains the data and test data is used for making predictions. So a separate dataset is needed to validate the predictions. In order to do so, train data is divided into two parts one for validation and other for training, 70% of train dataset is used for training the model. 30% of the training dataset is used for validating the model.

1. Logistic Regression Model

First fit the logistic regression model and then predict the Loan_Status for the validation set. Next calculating the accuracy of the predictions, following are the outcomes - Predictions using the Logistic regression model came out to be 75.67% accurate i:e, the model identified around 76 % of the loan status correctly. Now make predictions for the test dataset.

2. Decision Tree Model

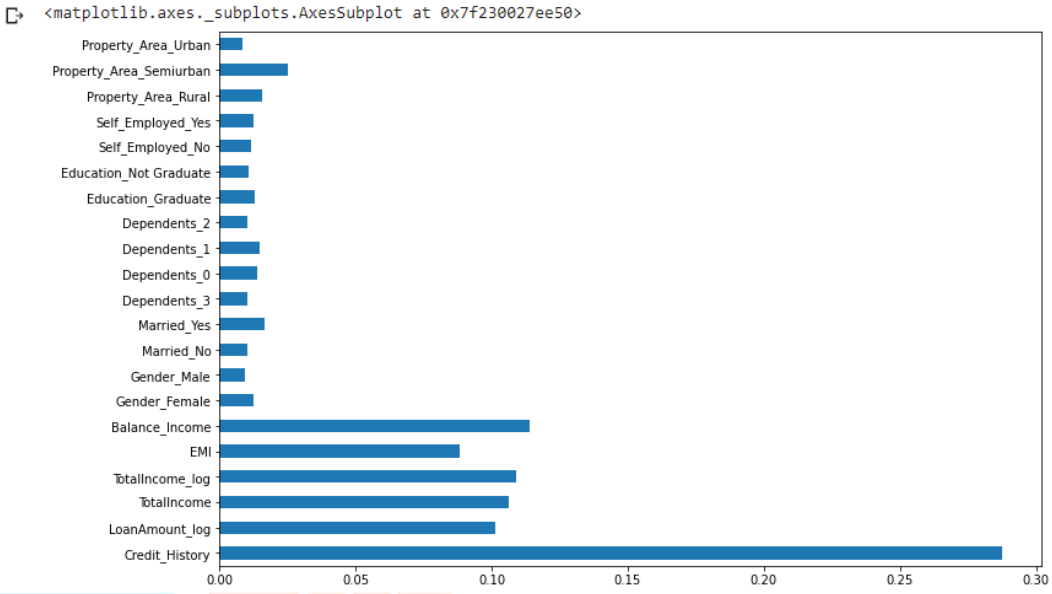
First fit the Decision Tree model and then predict the Loan_Status for the validation set. Next calculating the accuracy of the predictions, following are the outcomes - Predictions using the Decision Tree model came out to be 71.35% accurate. Now make predictions for the test dataset

3. Random Forest Model

Random Forest is a tree based algorithm. A certain number of weak learners aka decision trees are combined to make a powerful prediction model. Final prediction can be a function of all the predictions made by the individual learners. First fit the Random Forest model and then predict the Loan_Status for the validation set. Next calculating the accuracy of the predictions, following are the outcomes - Predictions using the Random forest model came out to be 77.83% accurate. Now make predictions for the test dataset.

V. CONCLUSION

Comparing the accuracy of all the three models, we can conclude that the Random Forest Model gives the highest prediction accuracy that is 78% thus it is the better choice for us. Lastly, find the important features using the feature_importances_attribute of sklearn for the best performing model first which is Random Forest Model. It was observed that 'Credit_History','Balance Income' features are most important. So, it is clear that feature engineering was helpful in predicting target variable accurately.



VI. ACKNOWLEDGMENT

We would like to express gratitude to our guide Prof. Mrinali Bhajibhakare for valuable suggestions and direction towards the execution of this project. We convey our heartfelt thanks to Prof. Swati Patil for her dynamic support being the project coordinator. We are very thankful to Dr. Arati Dandavate, Head of the Department, Computer Engineering, who has extended support and valuable suggestions towards achieving success in this project.

REFERENCES

[1] M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 490-494, doi: 10.1109/ICESC48915.2020.9155614.

[2] L. Lai, "Loan Default Prediction with Machine Learning Techniques," 2020 International Conference on Computer Communication and Network Security (CCNS), Xi'an, China, 2020, pp. 5-9, doi: 10.1109/CCNS50731.2020.00009.

[3] A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Delhi, 2017, pp. 1-6, doi: 10.1109/ICCCNT.2017.8203946.

[4] B. V. Srinivasan, N. Gnanasambandam, S. Zhao and R. Minhas, "Domain-Specific Adaptation of a Partial Least Squares Regression Model for Loan Defaults Prediction," 2011 IEEE 11th International Conference on Data Mining Workshops, Vancouver, BC, 2011, pp. 474-479, doi: 10.1109/ICDMW.2011.69.

[5] C. Wang and Y. Tzeng, "Prediction Model for Policy Loans of Insurance Company," The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007), Tokyo, 2007, pp. 653-658, doi: 10.1109/CEC-EEE.2007.81.

[6] Y. Chen, J. Zhang and W. W. Y. Ng, "Loan Default Prediction Using Diversified Sensitivity Undersampling," 2018 International Conference on Machine Learning and Cybernetics (ICMLC), Chengdu, 2018, pp. 240-245, doi: 10.1109/ICMLC.2018.8526936.

[7[M. V. J. Reddy and B. Kavitha, "Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis," 2010 International Conference on Signal Acquisition and Processing, Bangalore, 2010, pp. 274-277, doi: 10.1109/ICSAP.2010.10.