HOUSING LOAN FOREBODING:
A FINE SYSTEM THAT SORTS OUT GERMANE CLIENTS

Shivam Singh Rawat (PES2201800095), Skandan K A (PES2201800064), Sumer Singla (PES2201800073), Waris K R (PES2201800315)

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

For the Academic year 2020-21

PES University EC Campus 1 kM before Electronic City, Hosur Road,

Bangalore-100A, INDIA

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 Gender, Loan Status and Education etc. Loan_status is our Target Variable. Data cleaning includes dealing with missing values using Mean (without outliers) and Mode (with Outliers) Replacement Method. Also Correlation is visualized using Heatmap.

In the Upcoming Phases, the plan is to progress further by performing feature engineering, Model Building using regression etc..

INTRODUCTION

1.1 PROBLEM STATEMENT

A finance company deals with home loans and has it's 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 criterias 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 the process of loan eligibility checking to get automated. Eligibility is based on customer details provided while filling an application form.

1.2 NEED OF A SOLUTION

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.

1.3 HYPOTHESIS

First, listing out most important factors which may have an affect 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 chances of 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.

STOCK TAKING

2.1 VARIABLE DESCRIPTION

Given below is the description for each variable.

VARIABLE	DESCRIPTION
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (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

2.2 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

3.1 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, 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.

3.2 BIVARIATE

In Bivariate Analysis, after individually analysing every variable in univariate analysis, previously mentioned hypotheses can be tested analysing 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 EDA.

DATA CLEANING

4.1 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 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.

4.2 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.

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 Applicant Income and Co-Applicant Income 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 train dataset is used for validating the model

6.1 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.

6.2 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.

6.3 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.

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.