**LOAN PREDICTION USING PYTHON**

## Predict Loan Eligibility for Dream Housing Finance company

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

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 provided a dataset to identify the customers segments that are eligible for loan amount so that they can specifically target these customers.

## Data Dictionary

****Train file:****CSVcontaining the customers for whom loan eligibility is known as 'Loan\_Status'

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Coapplicant 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 | (Target) Loan approved (Y/N) |

****Test file:**** CSVcontaining the customer information for whom loan eligibility is to be predicted

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Loan\_ID | Unique Loan ID |
| Gender | Male/ Female |
| Married | Applicant married (Y/N) |
| Dependents | Number of dependents |
| Education | Applicant Education (Graduate/ Under Graduate) |
| Self\_Employed | Self employed (Y/N) |
| ApplicantIncome | Applicant income |
| CoapplicantIncome | Coapplicant 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 |

**Lets look at the steps that we will follow:**

1. Problem statement
2. Hypothesis Generation
3. Getting the system ready and loading the data
4. Understanding the data
5. Exploratory data analysis
   1. Univariate analysis
   2. Byvariate analysis
6. Missing value and Outlier treatment
7. Evaluation metrics for classification problem
8. Model Building- part 1
9. Logistic Regression using stratified k fold cross validation
10. Feature Engineering
11. Model Builing part-2
    1. Logistic Regression
    2. Decision Tree
    3. Random Forest
    4. XGBoost

**Problem Statement:**

It is a classification problem where we have to predict whether a loan would be approved or not.In a classification problem we have to predict discrete values based on a given set of independent variables.

In this problem it is binary classification- we have to predict either of the two given classes. Loan prediction is a very common real life problem that each retail bank faces at least once in its life time. If done correctly, it can save lots of man hours at the end of the retail bank.

**Hypothesis Generation:**

Below are some of the factors which I think can affect the loan approval.

1. **Salary**: Applicants wih high income shoul have more chances of Loan approval.
2. **Previous History**: Applicants who have repayed their previous debts should have higher chances of loan approval.
3. **Loan Amount**: If the loan amount is less the chances of loan approval should be high.
4. **Loan Term**: Loan for less time period and less amount should have higher chances of approval.
5. **EMI**: Lesser the amount to be paid monthly to repay the loan, higher the chances of loan approval.

**Getting the system ready and loading the data:**

We will be using Python for this problem along with the listed libraies.

**Speifications:**

Python: 3.7

Pandas= 0.20.3

Seaborn=1.0.0

Sklearn=0.19.1

**Import the required modules.**

**Data:**

We have give two CSV files for this problem. Train and test.

**Train**: Will be used for training the model,i.e our model will learn from this file. It contains all the independent and the target variables.

**Test**: Contains all the independent variables but not the target variables. We will apply the model to predict the target variable for the test data.

**Univariate analysis:**

Simplest form of analyzing data where we analyse each variable individually. For categorical variables we used frequency table or bar plots which will calculate the number of each category in a particular variable. For numerical features probability density plots has been used to look at the distribution of each variable.

**Inferences:**

1. 80% applicants in the dataset are male.
2. Around 65% of the applicants in the dataset are married.
3. Around 15% of the applicants in the dataset are self employed.
4. Around 85% of the applicants has repayed their debts.
5. Most of the applicants don't have any dependents.
6. Around 80% of the applicants are graduate.
7. Most of the applicants are from semi urban area.
8. It can be inferred that most of the data in the distribution of applicant income is distributed towards the left which means it is not normally distributed. The box plot confirms the presence of lot of outliers/extreme values. This can be attirbuted to the income disparity in the society.Part of this can be driven by the fact that we are looking at people with different education levels. Let us seggregate them by education:
9. There are high number of graduates with very high inccome which appears to be the outliers.
10. We see similar distribution as that of the applicants income.Majority of the income ranges from 0 to 5000. We also see lot of outliers in the coapplicants income and it is also not normally distributed.

**Bivariate Analysis:**

1. It can be inferred that proportion of male and female applicants is more or less same for approved and unapproved loans.
2. Proportion of married applicants is higher for the approved loans.
3. Distribution of applicants with 1 or 3+ dependents is similar across both the categories of the loan status.
4. There is nothing significant we can infer from self employed vs loan status plot.
5. It seems people with credit history as 1 are more likely to get their loans approved.
6. Proportion of loan getting approved in semi urban area is higher as compared to that in rural and urban areas.
7. We dont see any change in the mean income so we create bins for the applicants income variable based on the values in it and analyze the corresponding loan status for each bin.
8. It can be inferred that applicants income does not affect the chances of loan approval which is in contradicts to our hypothesis in which we assume higher the applicants income higher the chances of loan approval.
9. It shows that if the coapplicants income is less the chances of loan approval is high. But this does not look right. The possible reason behind this is most of the applicants do not have co applicants where the coapplicants income become 0 and hence the loan approval is not dependent on it. So we make a new variable in which we will combine the applicants and the co applicants income to visualize the combine effect of the income on loan approval.
10. We can see that the proportion of loan getting approved for applicants having low total income is very less as compared to that of applicants with average, high and very high income.
11. We can see that proportion of approved loan is higher for low and average loan amount as compared to that of high loan amount which supports our hypothesis in which we considered the chances of loan approval will be high when the loan amount is less.

Dropping the bins created for the exploration part.

We convert 3+ in dependent variable to 3 to make it numerical variable. And Y and N in loan status to 1 and 0 respectively so that we can find its corelation with numerical variables and also few models like logistic regression will take only numeric values as input.

Lets look at the correlation between all the numeric variables using heatmap.

We can see that the most correlated variables are (Applicant Income- Loan Amount) and ( credit history - loan status). Loan amount is also correlated with coapplicants income.

# **Missing value and Outlier Treatment:**

There are missing values in Credit history, Loan amount term, Loan amount, self employed, dependents,married, gender.

There are very less missing values in gender, married, dependents, credit history, self employed featues so we can fill them with mode of the features.

To remove the outliers we use log transformation.

Now the distribution looks much closer to normal and effect of extreme values has been significantly subsided.

# **Model Building-part 1(Logistic Regression)**

Make dummy variables for the categorical variables. Dummy variables turn categorical variables into values of 0 and 1, making them lot easier to quantify and compare.

Using the train\_test\_split function from sklearn to divide our train dataset as train and validation dataset.

Fit the logistic regression model.

### **Predicting the loan status for validation set and calculating its accuracy**

Our predictions are around 80% accurate. we predicted 80% of the loan status correctly.

### **Prediction for the test data set**

# **Logistic Regression using strtified k fold cross validation**

Let us visualize the ROC curve

# **Feature Engineering**

We will create the following three new features:

1. Total income - If the total income is high the chances for loan approval might also be high.
2. EMI- People who have high emi might find it difficult to repay the loan.
3. Balance Income- This is income left after emi is paid. If this value is high there is high chances for the person to repay the loan which in turn increase the chances of approval.

Dropping the variables used to create the new variables. Because the correlation between the old features and the new features will be very high and logistic regression assumes that the variables are not highly correlated. And also to remove the noise from the dataset.

# **Model Building- part 2**

Will build the following models:

1. Logistic regression- From the submission we got an accuracy of 0.7847. Feature engineering has not improved our model.
2. Decision tree- Accuracy 0.63 which is less than Logistic regression model.
3. Random Forest- Trying to increase the accuracy by tuning the hyperparameters of the model. Using grid search to get the optimized values of the hyper parameters. Also will tune the max\_depth and n\_estimators parameters.  
   Optimized value for max\_depth =3, n\_estimators =41.

Accuracy =0.7738

### **Feature importance:**

We can see that the credit history is the important feature followed by the balance income, EMI. Feature engineering helped us in predicting our target variable.

1. XGBoost- Accuracy= 0.7361

After trying and testing 4 different algorithms, the best accuracy on the public leaderboard is achieved by Loistic Regression(0.7847), followed by random forest(0.7638)