Loan Eligibility Prediction for Mortgage Companies

Xinyue Ma

Background

As the scale of the credit business continues to expand, the requirements for the accuracy of risk management have gradually increased. The current mortgage application process still involves lots of bankers and advisors and it takes a long time. Customers need to fill in lots of information during the application, even though some of them are not necessary. We also want to pick out the most useful items for customers to fill out, in order to increase the efficiency of mortgage application.

Introduction

Mortgage companies deal in all kinds of home loans. They have a presence across all urban, semi urban, and rural areas. Customers apply for home loans and after that, the company validates the customer eligibility for the mortgage. The Company wants to automate the loan eligibility process based on customer details provided while filling online application forms. These details include Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, past application and approval data are available. We are going to build a model to identify the customer segments that are eligible for the mortgage. We see it as a binary classification problem. This project can help the mortgage company to develop market strategies specifically targeting these customers.

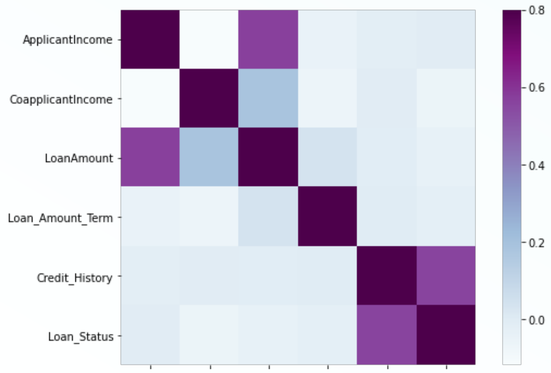
Data Description

The data is from Analytics Vidhya Loan Prediction Competition. We don't have labels in the test dataset, but we can submit the prediction result online and the website will check out the solution score.

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)

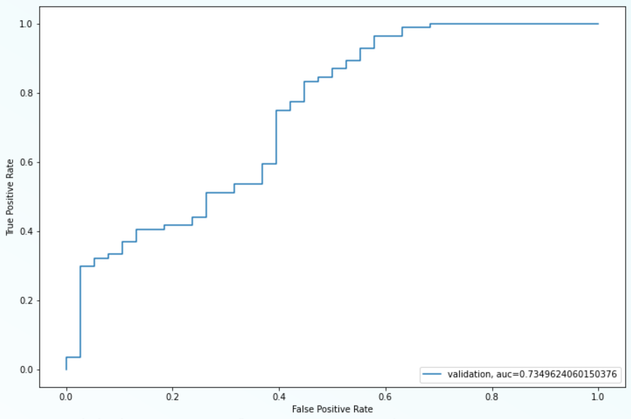
Exploratory Data Analysis

In the train dataset, the loan of 422 (around 69%) people out of 614 was approved. Most of the applicants don't have any dependents. Around 80% of the applicants are Graduate. Most of the applicants are from Semiurban area. The applicant income and co-applicant income have outliers. Using the bivariable analysis, it can be referred that the proportion of male and female applicants is same for both approved and unapproved loans. Proportion of married applicants is higher for the approved loans. Distribution of applicants with 1 or 3+ dependents is similar across both the categories of Loan_Status. People with credit history as 1 are more likely to get their loans approved. Proportion of loans getting approved in semiurban area is higher as compared to that in rural or urban areas. The proportion of loans getting approved for applicant having Total_Income is very less as compared to that of the applicants with Average, High, and Very High income. The proportion of approved loans is higher for Low and Average Loan Amount as compared to that of High Loan Amount. The most correlated variables are (ApplicantIncome - LoanAmount) and (Credit_History - Loan_Status). LoanAmount is also correlated with CoapplicantIncome.



Data Modeling

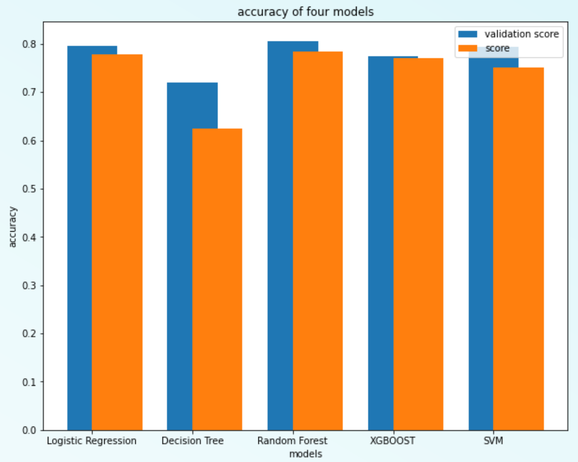
Missing data and outliers can have adverse effect on the model performance. For numerical variables, we will impute use mean or median. For categorical variables, we will impute use mode. Take the log transformation to loan_Amount variable.
The model performance will be evaluated on the basis of prediction of loan status for the test data and we use Accuracy as the evaluation method. First, we start with Logistic Regression which is used for predicting binary outcome, and the validation accuracy is 0.8013 . Then we conduct Stratified k-fold Cross Validation, which is a better approach when reducing both variance and bias. The validation accuracy is 0.801. We can plot the ROC curve and get an AUC value of 0.735.



We add three new features. Total Income: we combine the Applicant Income and Coapplicant Income. If the total income is high, chances of loan approval might also be high. EMI: EMI is the monthly amount to be paid by the applicant to repay the loan. People who have high EMI's might find it difficult to pay back the loan. Balance Income: The income left after the EMI has been paid.

Results

We use the logistic Regression model as baseline and build more complex models like RandomForest and XGBoost, then compare the performance of these models. After trying and testing 5 different algorithms, the best accuracy is achieved by Random Forest (0.7847).



Conclusion

We can see that Credit_History is the most important feature followed by Balance Income, Total Income, EMI. We recommend to have credit histories and high income to increase the chances of loan approval.