Capstone Project

Machine Learning Foundations

Rashini J Widanaarachchi

Nov 21st, 2021

Loan Eligibility Prediction Machine Learning Model

Project Overview

We may not always have the money we require to do certain things or to buy certain things. In such situations, individuals and businesses/firms/institutions go for the option of borrowing money from lenders. loans are the credit extended to us by lenders on fulfilling certain key parameters.

Customer first applies for Loan in a Financial Institution such as a bank and after that company validates the customer eligibility for loan. To check the eligibility banks evaluate your credit history and credit worthiness such as income, history, etc. The higher your score of criteria the better are the chances of your loan application getting approved.

Factors determine the loan eligibility

Credit score- the lender analyses the repayment history of the borrower and concludes whether the borrower can repay on time or will he default on payments

Income and Employment History- income and income stability in the form of consistent and stable work history

Debt-to-income ratio- you must have a low debt-to-income ratio so the lenders can think that you have enough cash at hand every month

Adding a co-signer with a higher credit score and income can boost approval chances.

In this project, a Loan Eligibility Prediction Machine Learning Model has been developed which could check the possibility of Loan Approval. Once the customer data is provided this model could predict whether it is safe to provide a loan to the relevant customer. This model output will be provided through an API and it can be integrated to any application to get results as soon as the customer data is entered.

Problem Statement

The goal is to create a Loan Eligibility machine learning model to identify the customers segments that are eligible for loan amount so that they can specifically target these customers.

Metrics

Precision is a common metric for binary classifiers; it considers both true positives and false positives. It is also known as Positive Predictive Value (PPV).

$$\textbf{Precision} = \frac{\text{true positives}}{\text{true positives + false positives}}$$

According to the problem, out of the predicted credit worthy individuals who has the possibility of getting a loan how many have received a loan will be measured here.

False positive values in the model output will represent the customers as credit worthy but in real world those customers don't have an eligibility to pay back the loan. It is critical to keep this value minimal as giving a loan to an individual who doesn't have ability to pay back can be a greater lost to the company.Therefore, it is important to identify customers who can pay back the loan. Developing a model with high Precision/ high true positives compared to the total positives predicted from the model will help to achieve this.

Analysis

The model has been developed based on the data obtained from the Kaggle Platform, which is an online community of data scientists and machine learning practitioners. The link to the data repository is given below.
https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset/code

The data contained 614 rows data with 13 variables including customer details which have been provided in requesting a loan. These data were extracted to a csv format and accessed through python for the model development.

The customer details provided are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History, and others. Data description is as follows.

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)
Applicant Income	Applicant income
Co applicant Income	Co applicant income
Loan Amount	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)

The results of the exploratory data analysis for the above data provided with the below insights on the data.

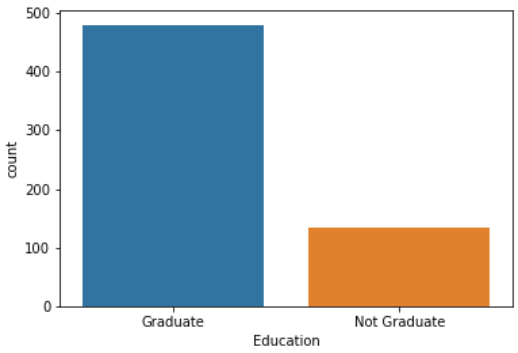

Figure 1: Education wise Analysis

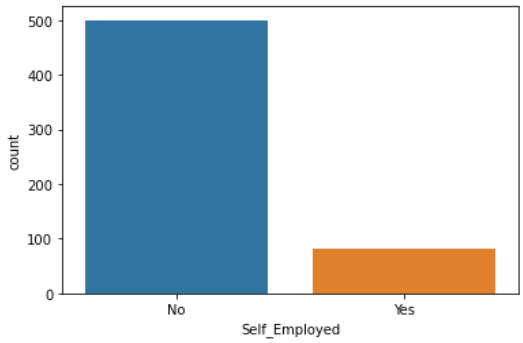

Figure 2: Self-employed/not

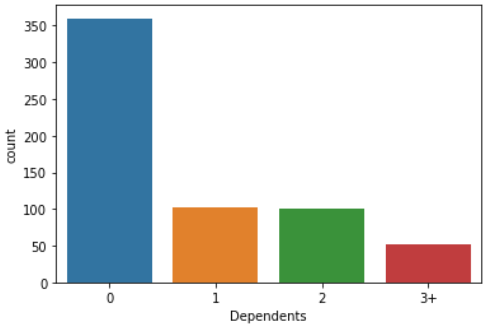

Figure 3: Dependent's availability

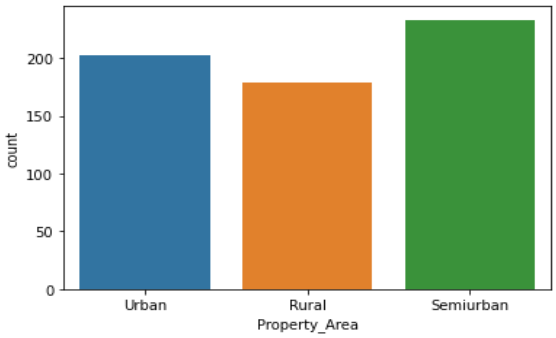

Figure 4: Property area wise analysis

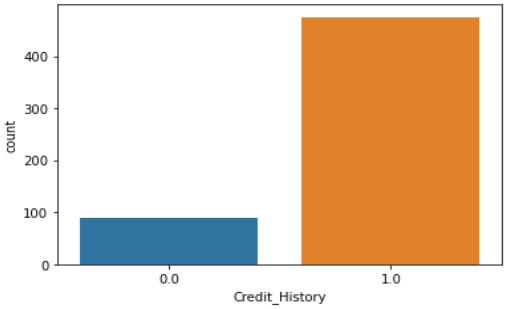

Figure 5: Credit History availability

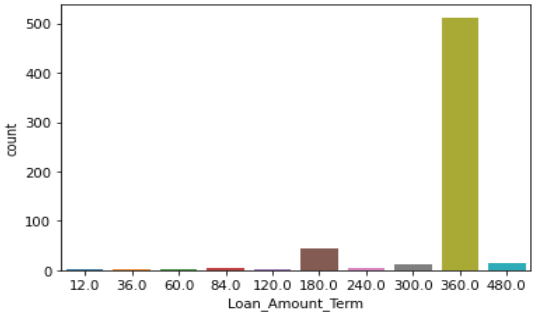

Figure 6: Loan Amount Termination period (in days)

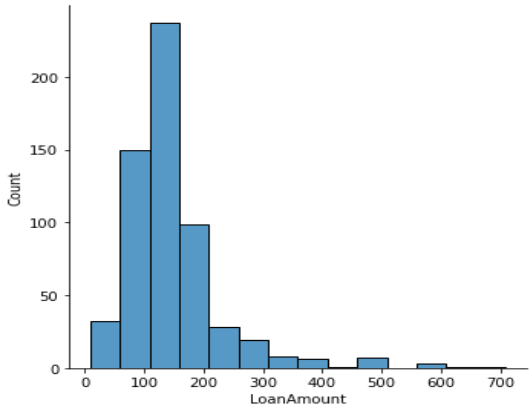

Figure 7: Laon Amount

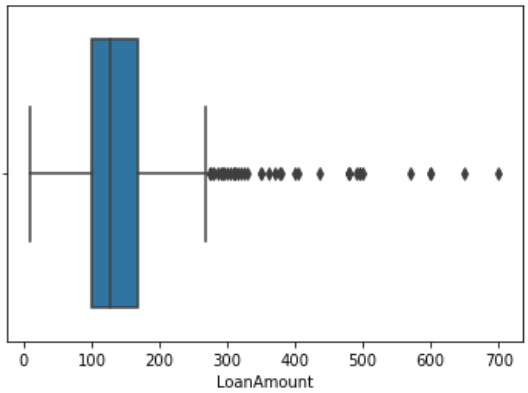

Figure 8: Loan Amount data spread

The Data is inclusive of customers who are majority is male and married. There are no dependents for most of the customers. For those who are with one or two dependents are majorly married people (figure2). 78% of the total customers are graduates (figure 1). Most of the graduates have applied for loans and out of them majority has received the loan as well. There are both self-employed and salaried customers who have applied for loans and majority are salaried customers (figure 2). All these customers are available with property area in either of Urban, Rural or Semiurban areas (figure 4). Out of them majority has been in Urban or Semiurban.

The Mean loan amount of the customers have been $149K. The range of the loan amount requested is $(700-9) K = $691k. The loan amount consists of a high range value. According to figure7, the loan amount data has been right skewed and there are many outliers visible (figure 8). Majority of these loans have been with a loan amount term of 30 years (360 months).

Methodology

Data Preprocessing

Based on the above insights the null values of the variables have been addressed in the model. The categorical variable null values have been filled with the modes of the relevant fields while the loan amount null values have been filled up with the median value. This is due to the availability of outliers in the variable and when filling the null, it is important to select a measure which does not affect from the outliers.

The binary variables have been renamed with 1's and 0's. The 'loan amount term' variable values have a higher frequency for the months 360,180,480 and 300 while the rest of the months show a lower frequency. By grouping the field, the years with lower frequencies can be grouped to one. Therefore, the loan amount term variable has been grouped in 10-year groups

(E.g.- ['<120', '120-240', '240-360', '>360']) and converted to a simple categorical variable termed as 'Loan_Amount_Term_grp'.

For the ease of model prediction, the categorical variables have been broken down into dummy variables and compared with the loan status which is the depended variable of the model. e.g. - Loan_Amount_Term_grp, Property_Area and Dependents

The relationship between the dependent variable and the independent variable has been measured using the correlation matrix. The correlation matrix represents how strong the relationship between each variable is and how strong the relationship is between each variable and dependent variable.

	y_act
y_act	1.000000
ApplicantIncome	-0.004710
CoapplicantIncome	-0.059187
LoanAmount	-0.033214
is_male	0.017987
is_Married	0.091478
is_Self_Employed	-0.003700
Loan_Amount_Term	-0.022549
Credit_History_availability	0.540556
is_graduate	0.085884
Loan_Term_<120	0.019096
Loan_Term_120-240	-0.012959
Loan_Term_240-360	0.048785
Loan_Term_>360	-0.098067
Property_Area_Rural	-0.100694
Property_Area_Semiurban	0.136540
Property_Area_Urban	-0.043621
Dependents_0	-0.003044
Dependents_1	-0.038740
Dependents_2	0.062384
Dependents_3+	-0.026123

Figure 9: correlation_matrix values of loan status

The above figure shows the relationship the loan status variable has with the independent variable. As an example, according to the above matrix the 'Credit_History_availability' shows a high relationship to the loan status comparatively. This can be a variable which will have an impact in predicting the loan eligibility. Likewise based on the above values the X variables suitable for the model has been selected.

Implementation and Refinement

During the first stage, the classifier was trained on the preprocessed training data. The data was broken down into train and test split. Which includes 30% of data to test and 70% of the data to train the model.

Train sample size = 429
Test sample size = 185

This is a classification Problem. Therefore, the python built in Machine Learning (ML) models Logistic Regression, Decision Tree Classifier and Random Forest Classifier can be applied to

identify the most suitable ML model for the problem. To identify the best model for the problem the test data has been run on every model. Model precision is the metric used to compare and select the most suited model for the problem. For each of these models different hyperparameters have also been tested and a suitable hyperparameter for the model has also being finalized.

	model_name	model	accuracy	precision	f1_score	roc_auc
0	lgr1	LogisticRegression(n_jobs=3, verbose=1)	0.783784	0.756410	0.756482	0.760513
1	dt1	DecisionTreeClassifier(max_depth=10, min_sampl...	0.713514	0.727891	0.690450	0.673462
2	rf1	(DecisionTreeClassifier(max_features='auto', r...	0.756757	0.769784	0.744861	0.803654
3	rf2	(DecisionTreeClassifier(max_features='auto', r...	0.772973	0.774648	0.759382	0.789359
4	rf3	(DecisionTreeClassifier(max_depth=10, max_feat...	0.794595	0.773333	0.775195	0.795513
5	rf4	(DecisionTreeClassifier(max_depth=20, max_feat...	0.762162	0.767606	0.747923	0.790705

Figure 10: Different models used for testing

Initially a set of X Variables was defined and run for different models and the precision value is compared. If the precision value is not up to expected value, the process is repeated carrying out feature engineering or hyperparameter changes. Figure 10 shows few of the model types tested during the process for the problem. After carrying out model training with changing parameters and comparing the precision values a model is being finalized.

Result

Based on the above metric values rf3 model has been considered as the best model for the problem. This has a comparatively a high precision/ high true positive values. Also, it has a high accuracy which represent a proportion of true negative values as well.

The finalized Model

model_name	model	accuracy	precision	f1_score	roc_auc
rf3	(DecisionTreeClassifier(max_depth=10, max_feat...	0.794595	0.773333	0.775195	0.795513

Figure 11:Finalized model

Model= RandomForestClassifier(max_depth=10, n_estimators=500, n_jobs=3, verbose=1)

Conclusion

The finalized model is saved. This model will be used in an API which will give the predicted output when the customer details are provided.