

Predicting Loan Approval for a Customer

By Avi Hritwik

1. Introduction

1.1 Background

In present days a lot of people apply for loans as it is a convenient way to fulfil their necessary requirements. And most of the bank have a lengthy process to approve their customer's loan, so this is advantageous to predict approval of loan for a customer in a short time. So that the customer have a idea about their request.

1.2 Problem

Data that might contribute to determine loan approval for a customer might include his credit history, Income, Gender (as there are many schemes provided for different gender), Education and income of co applicant. This project aims to predict whether a loan will be approved based on these data.

1.3 Interest

The loan company will be interested in accurate prediction of loan approval as it will help them provide a better and confident service to their customer. And even customer will be able to know about their status.

2. Data Acquisition and cleaning

2.1 Data Sources

The data I have for different customer are available here in csv format <https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/download/test-file> .

2.2 Data Cleaning

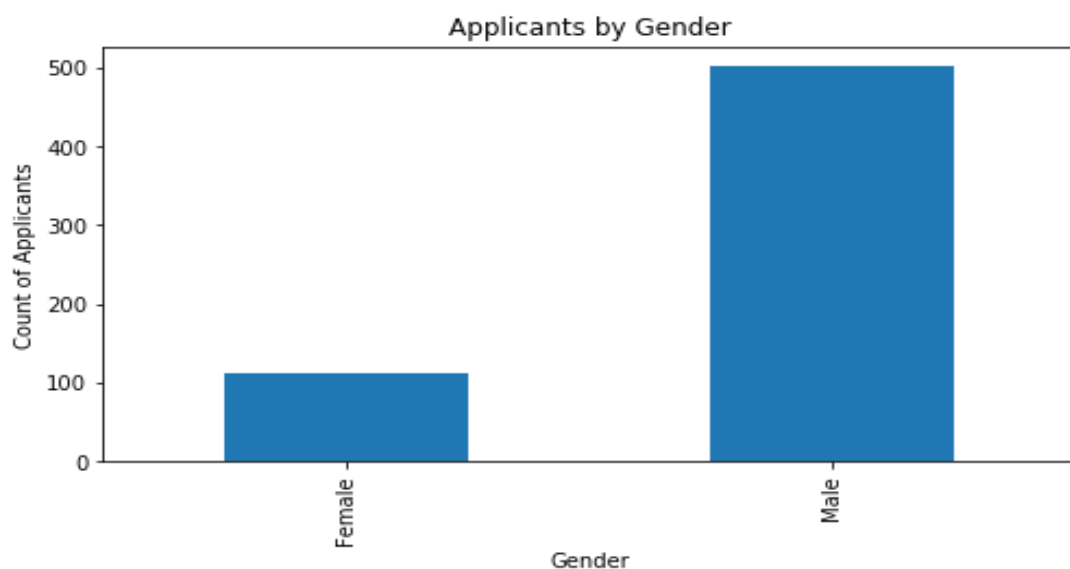
As data which I scrapped from the source was not processed, so it contained a lot missing value. And a lot loan were having duplicate values, So I had to clean that up ad make a more organised and usable data set.

3. Exploratory Data Analysis

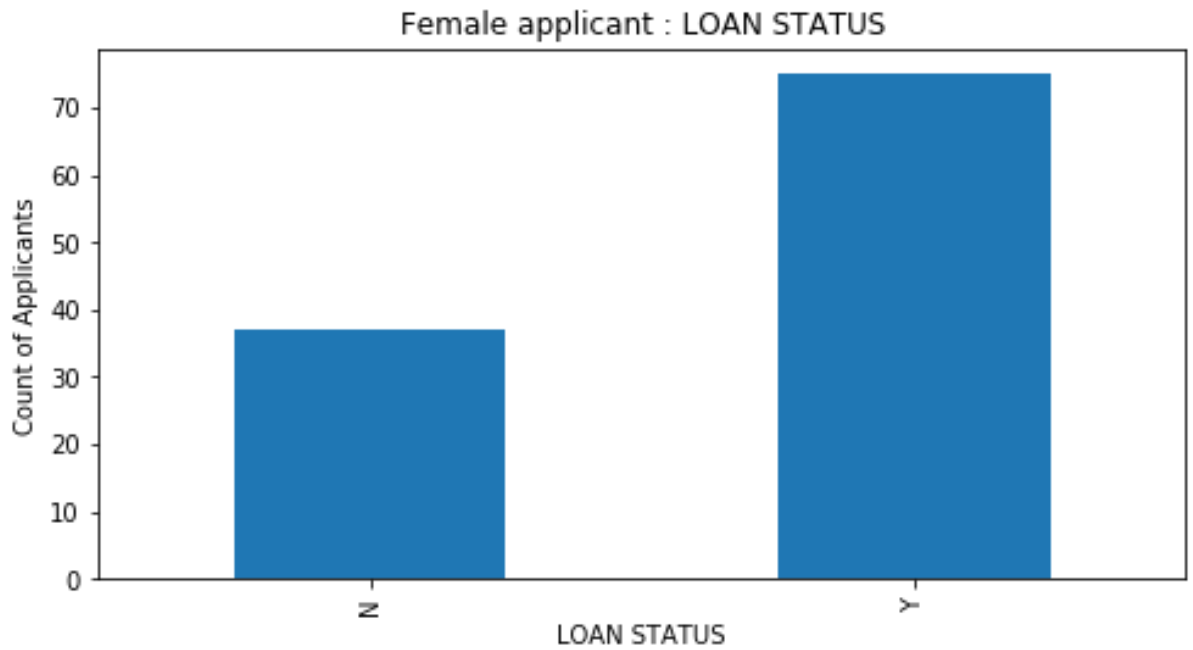
3.1 Target Variable

As the model had to predict loan approval the target variable should be the final status of the loan. So, in the data set the variable Loan_Status contains the outcome of the loan application.

3.2 Relation between Gender and Loan applicants



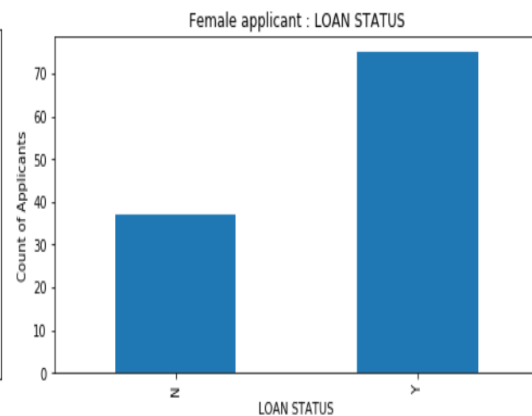
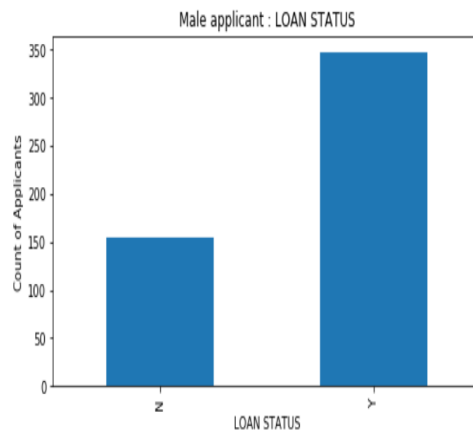
As we can see the loan applications by male are more than female.



More than 70% of loan applied by female are likely to be approved.

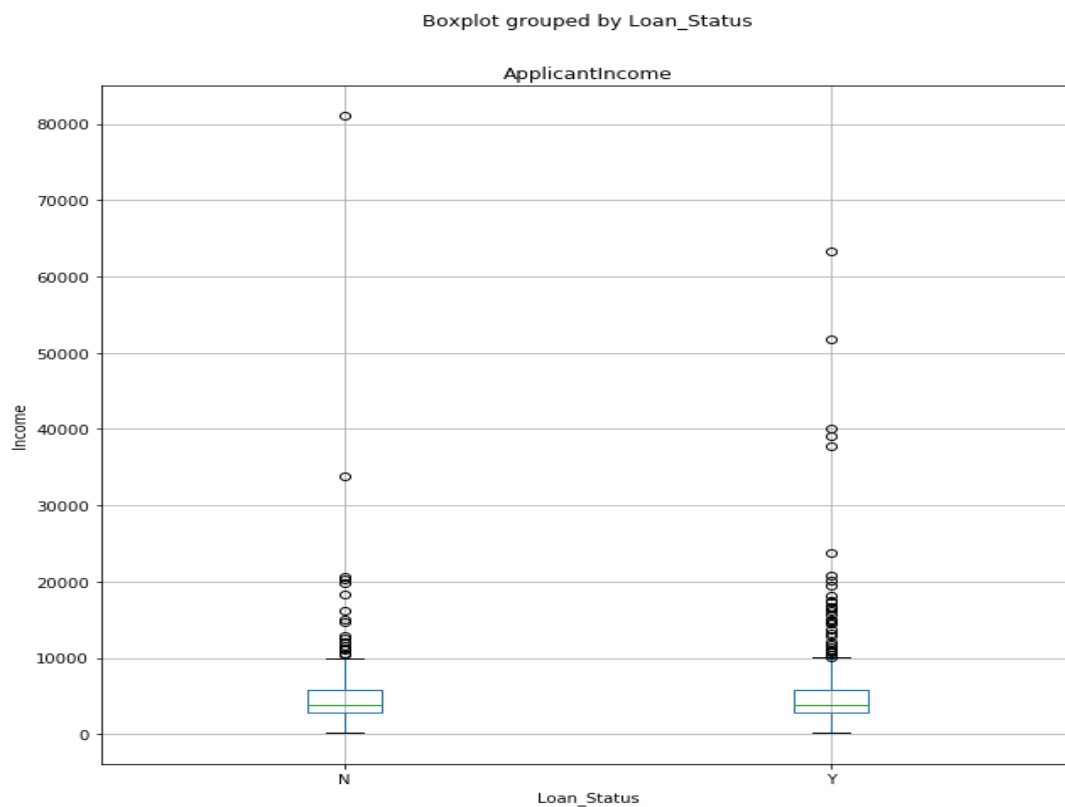
N 155
Y 347
Name: Loan_Status, dtype: int64

N 37
Y 75
Name: Loan_Status, dtype: int64



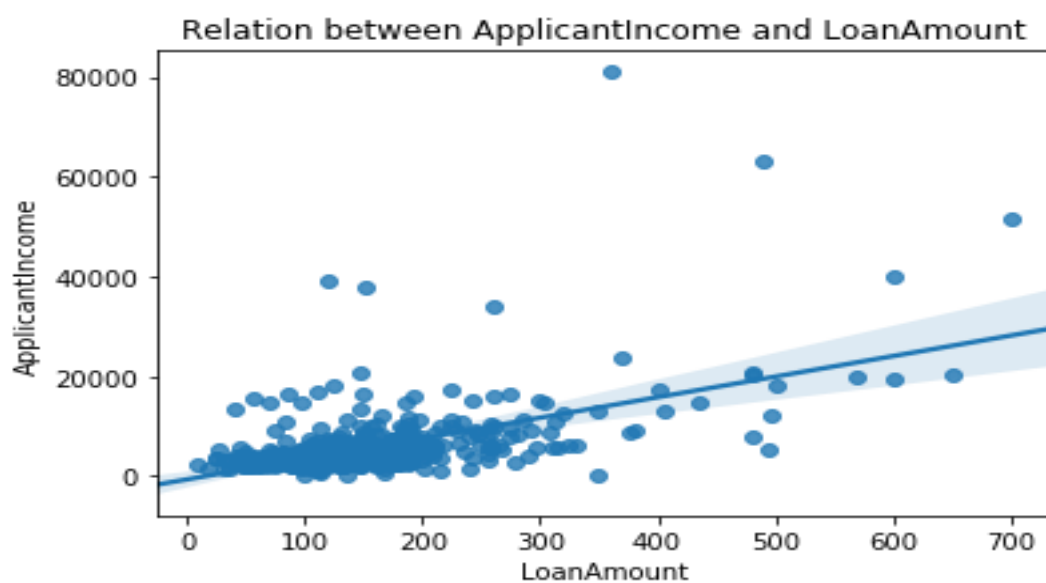
But the percentage approval for male is higher than female applicant

3.2 Relation between Applicant income and Loan Approval



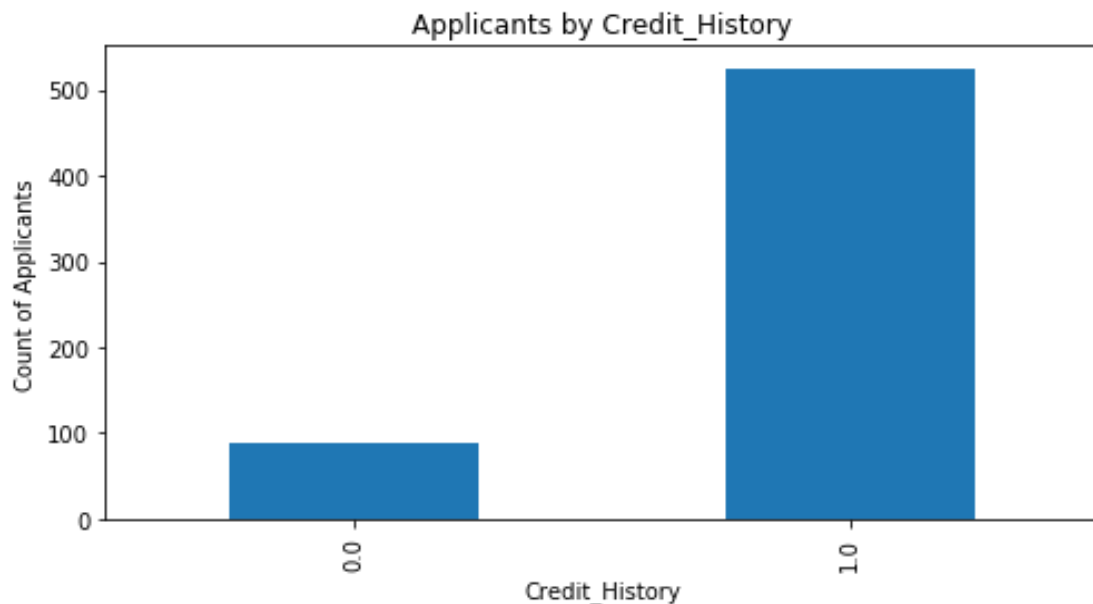
Both loan which are not approved or which are approved has a same mean income of the applicant. So, only feature income of applicant doesn't really affect the approval.

3.3 Relation between Gender and Loan applicants

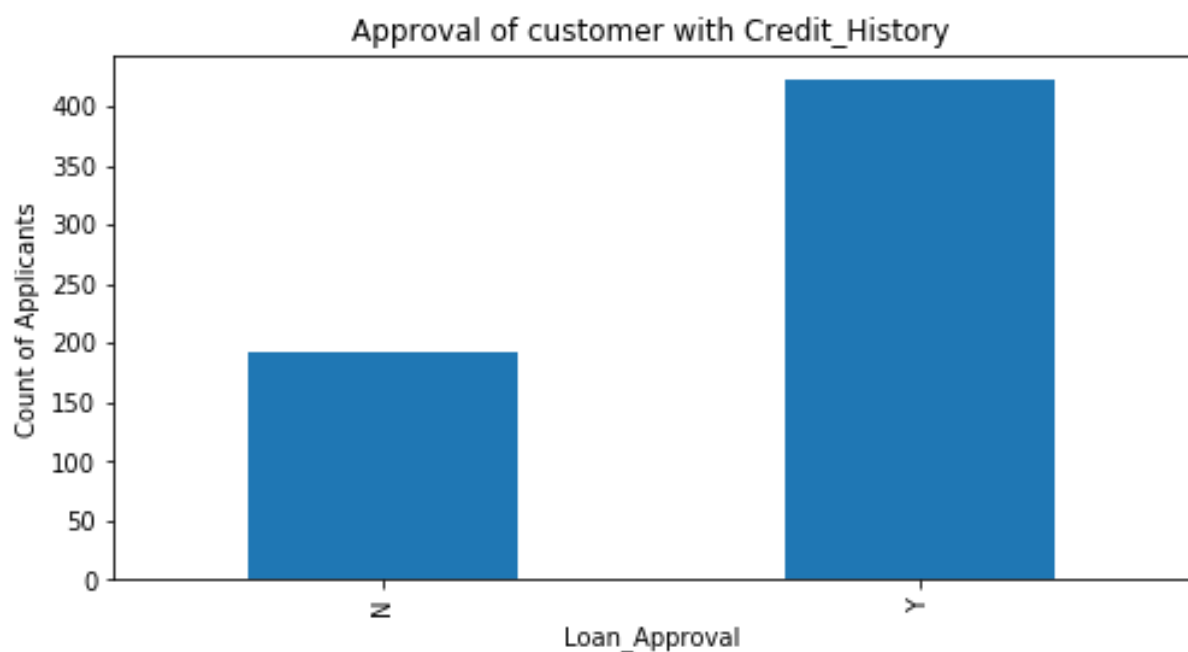


As the income of applicant increases the loan amount is also increasing. And most of the applicant belong to a narrow range of income as well as loan amount.

3.3 Relation between Credit history and Loan applicants



Most of the applicant are those who have already taken credits in past.



More than 66% of loans of customer with credit history were approved.

4. Predictive Modelling :-

There are two types of models, i.e. Regression and Classification, that be used to predict the loan approval. And also we can use supervised and Un-Supervised method for machine learning. But our goal is to provide a prediction based on a fixed dataset. So, we should use supervised learning.

Also to choose the model between Regression and Classification. Regression is used for predicting value for a continuous variable. As our output is going to be a Loan Status which a categorical variable we will use classification.

4.1 Classification Model

For the loan prediction system we use Logistic Regression.

As we are using sci-kit library for our building predictive model. And it uses numerical data to give the output. So, first of all we convert all the variables into numerical data type. For that we are label encoding.

In label ending, if a variable have 3 different value, it assigns each value a number and then replace the values with those number correspondingly.

Loan_ID	object
Gender	int64
Married	int64
Dependents	int64
Education	int64
Self_Employed	int64
ApplicantIncome	int64
CoapplicantIncome	float64
LoanAmount	float64
Loan_Amount_Term	float64
Credit_History	float64
Property_Area	int64
Loan_Status	int64
dtype:	object

Then, we choose the columns for applying Logistic Regression and store and create a dataset.

```
X=np.asarray(df[['Gender','Married','Dependents','Education','Self_Employed','ApplicantIncome','LoanAmount','Loan_Amount_Term','Credit_History']])
y=np.asarray(df['Loan_Status'])
x[0:5]
```

```
array([[1.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
        0.00000000e+00, 5.84900000e+03, 1.46412162e+02, 3.60000000e+02,
        1.00000000e+00],
       [1.00000000e+00, 1.00000000e+00, 1.00000000e+00, 0.00000000e+00,
        0.00000000e+00, 4.58300000e+03, 1.28000000e+02, 3.60000000e+02,
        1.00000000e+00],
       [1.00000000e+00, 1.00000000e+00, 0.00000000e+00, 0.00000000e+00,
        1.00000000e+00, 3.00000000e+03, 6.60000000e+01, 3.60000000e+02,
        1.00000000e+00],
       [1.00000000e+00, 1.00000000e+00, 0.00000000e+00, 1.00000000e+00,
        0.00000000e+00, 2.58300000e+03, 1.20000000e+02, 3.60000000e+02,
        1.00000000e+00],
       [1.00000000e+00, 0.00000000e+00, 0.00000000e+00, 0.00000000e+00,
        0.00000000e+00, 6.00000000e+03, 1.41000000e+02, 3.60000000e+02,
        1.00000000e+00]])
```

Then, we split the data in train and test set, So that we can use one of the set to train our model and other one to test the performance of the model.

```
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=4)
print ('Train set:', X_train.shape, y_train.shape)
print ('Test set:', X_test.shape, y_test.shape)
```

```
Train set: (491, 9) (491,)
Test set: (123, 9) (123,)
```

Then we train the model and use that model to predict the status of test set

4.1 Error

```
jaccard_similarity_score(y_test, yhat)
```

```
0.7886178861788617
```

I am using Jacard's Index to check the error in our prediction. In Jacard's Index, if the all the prediction is same as the real status then the value will be 1. It's worst value can be 0.0 .

The result .78 represent a good predicting system.

5. Result

I have created a model which can predict the loan approval with the a Jacard's Similarity of 0.78. This can be used to predict the loan approval for different customers.

6. Conclusion

In this study, I have seen relationship between Loan amount and loan approval, Gender and Loan approval. How the credit history affect the loan approval and also the trend of loan application by different class of customer.