REPORT

Introduction

Nowadays, there are numerous risks related to loans both for the banks and the borrowers, who get the loans. The risk analysis about loans needs understanding about the risk and the risk level. Banks need to analyse their customers for loan eligibility so that they can specifically target those customers.

As the number of transactions in banking sector is rapidly growing and huge data volumes are available, the customers behaviour can be easily analysed and the risks around loan can be reduced. So, it is very important to predict the loan type and loan amount based on the data.

Loan Prediction is very helpful for employee of banks as well as for the applicant also. The aim of this project is to provide quick, immediate and easy way to choose the deserving applicants. The Company deals in all loans. They have presence across all urban, semi urban and rural areas. Customer first apply for loan after that company or bank validates the customer eligibility for loan. Company or bank wants to automate the loan eligibility process (real time) based on customer details provided while filling application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and other. This project has taken the data of previous customers of various banks to whom on a set of parameters loan were approved. So the machine learning model is trained on that record to get accurate results. Our main objective of this project is to predict the safety of loan. To predict loan safety, the Logistic Regression algorithm is used.

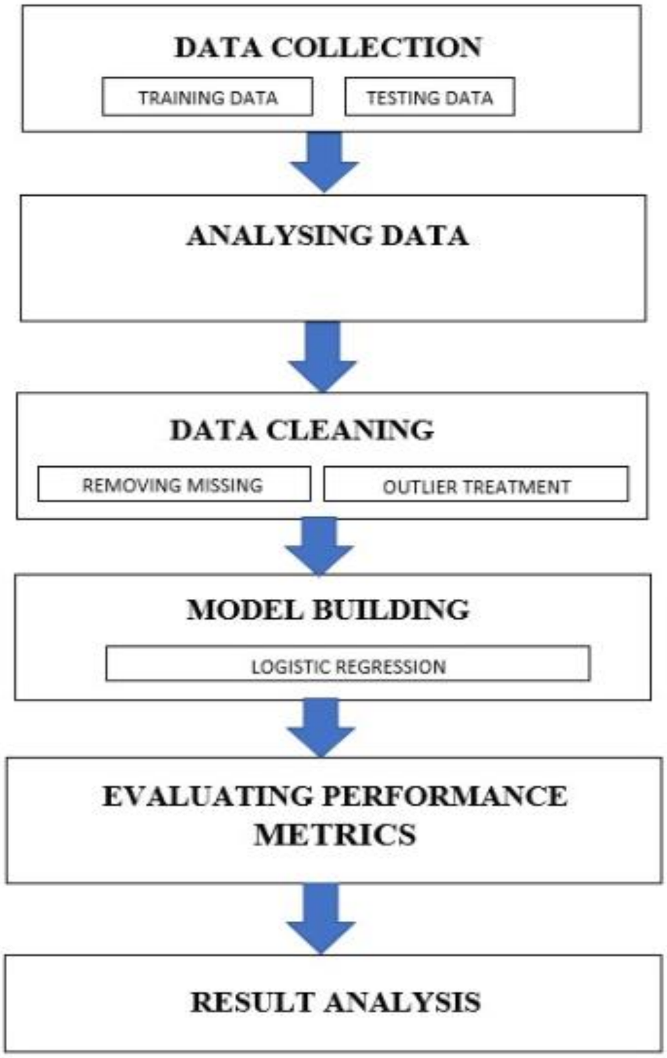
Data Description

The data set contains information about customers who have fully paid off their loans, current on their payments, in grace period, late, and in collection. This data set consists of 13 columns and around 600 to 620 rows. The data set contains numeric and categorical variables; some values are missing too. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan Status in train dataset and 12 independent variables in test dataset.

- **Loan_ID :** Unique Loan ID
- **Gender:** Male/ Female
- **Married:** Applicant married (Yes/No)
- **Dependents:** Number of dependents on the applicant
- **Education:** Applicant Education (Graduate/ Under Graduate)
- **Self Employed:** Self employed (Yes/No)
- **ApplicantIncome:** Applicant income
- **CoapplicantIncome:** Co-applicant 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:** Loan approved (Yes/No)

Approach

The purpose of this project is to build a logistic regression model based on statistical analysis to predict loans and find a reasonable classification between loans and profit for the bank.



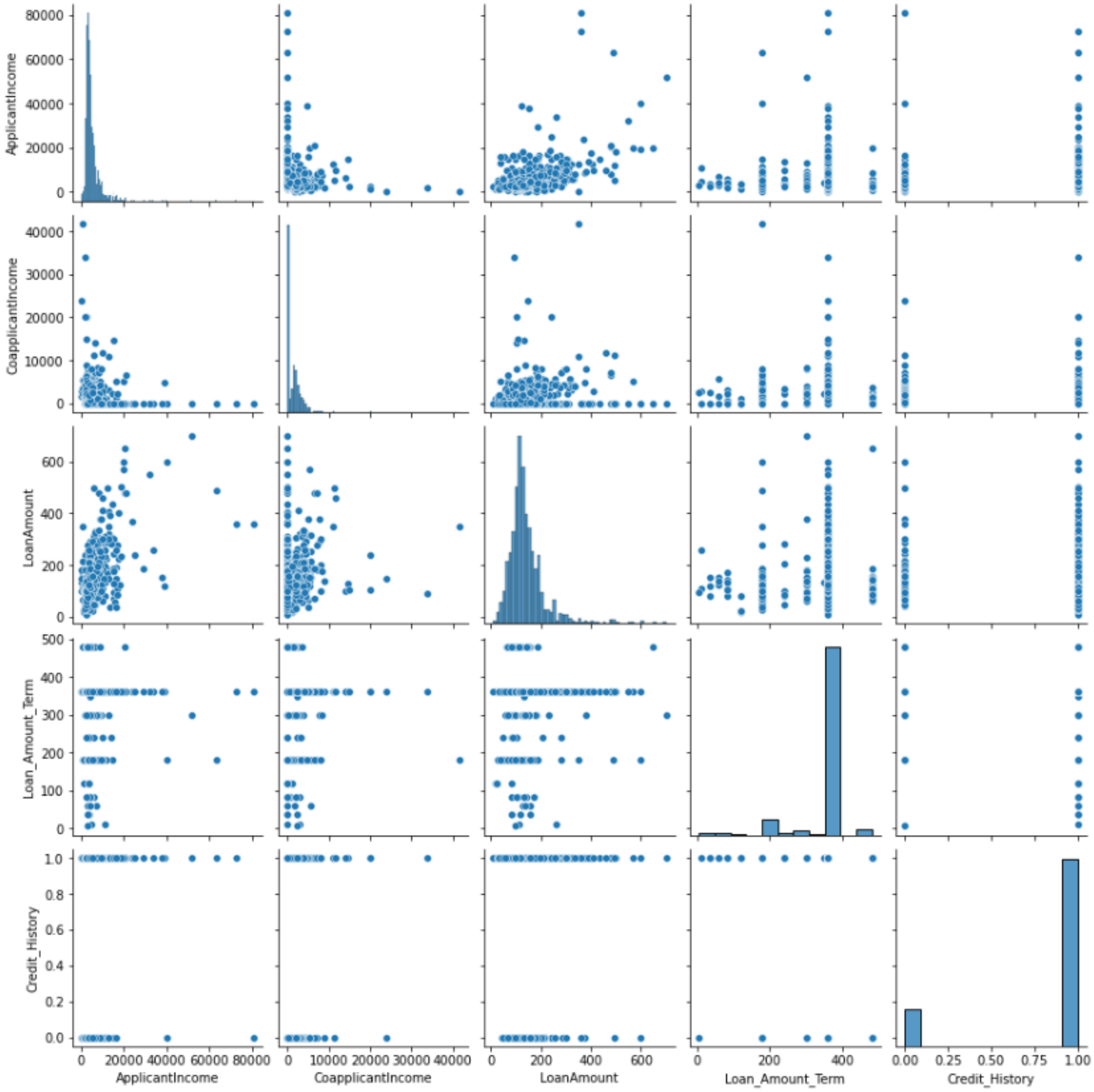
As shown in the above flow chart we have followed the same procedure. Initially we imported the data. The next step we analyse the data and check if there are any missing values, and if the data set contains missing values we discard them. After cleaning the data set we performed Exploratory Data Analysis and noted the observations. We performed the Logistic Regression for model building. The next step is to evaluate performance of the metrics models. Finally we note the result analysis.

We have built a model that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset

Visualization

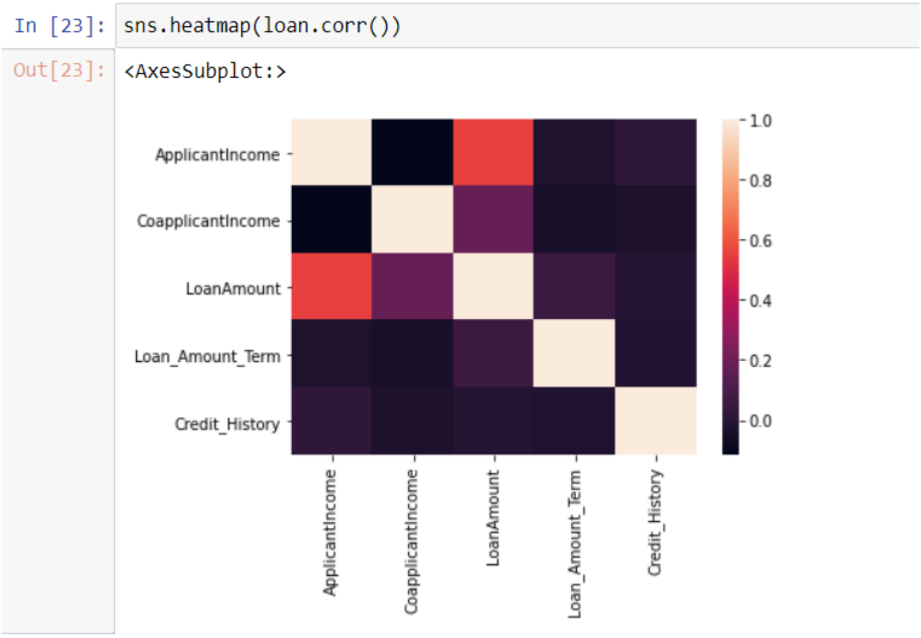
Pairplot

```
Out[22]: <seaborn.axisgrid.PairGrid at 0x23144f962b0>
```



INSIGHT: To plot multiple pairwise bivariate distributions in a dataset, you can use the pair plot() function.

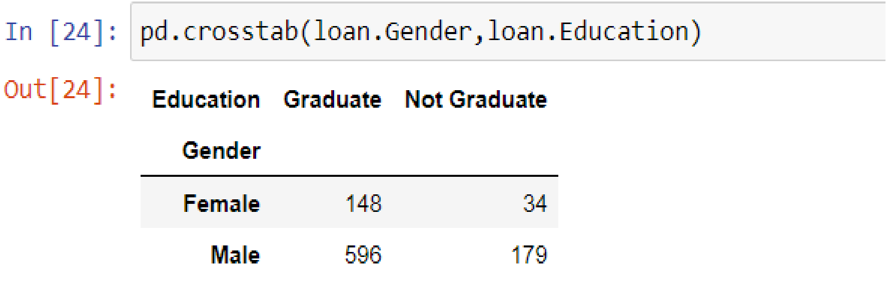
Heatmap



```
In [23]: sns.heatmap(loan.corr())
Out[23]: <AxesSubplot:>
```

INSIGHT: Each square in above figure shows the correlation between the variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables.

Crosstab->gender &education

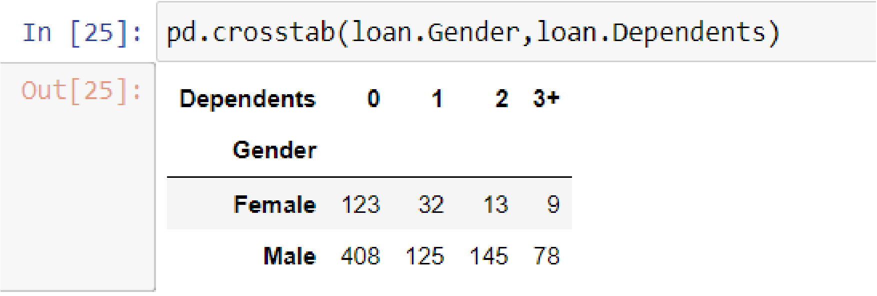


```
In [24]: pd.crosstab(loan.Gender,loan.Education)
Out[24]:
```

Education	Graduate	Not Graduate
Gender		
Female	148	34
Male	596	179

INSIGHT: In above figure we get the frequency of graduates and non-graduates males and females. The number of male graduates is more than female graduates.

Crosstab->Gender and Dependence



```
In [25]: pd.crosstab(loan.Gender,loan.Dependents)
```

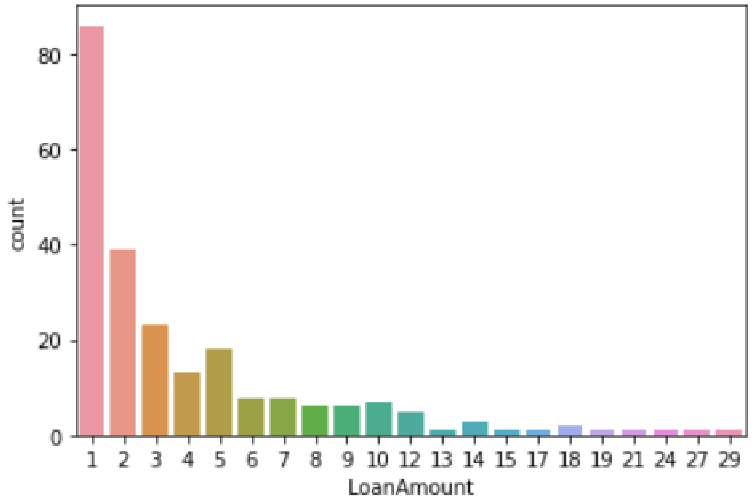
Out[25]:

Dependents	0	1	2	3+
Gender				
Female	123	32	13	9
Male	408	125	145	78

INSIGHT: From the above figure we get the frequency count of males and females who are dependents. The number of male dependents is more than female dependents.

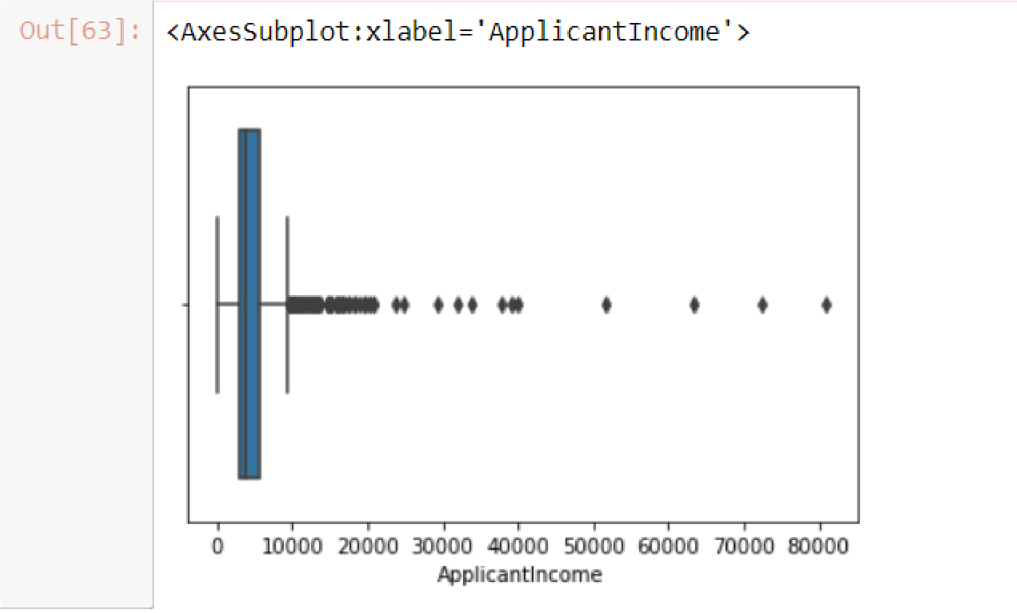
Countplot ->loan amount

```
Out[46]: <AxesSubplot:xlabel='LoanAmount', ylabel='count'>
```



INSIGHT: A count plot basically counts the categories and returns a count of their occurrences. The loan amount decreases gradually as we move ahead.

Boxplot->Applicant Income



```
Out[63]: <AxesSubplot:xlabel='ApplicantIncome'>
```

INSIGHT: The box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.

Algorithm

For this House Loan Amount Prediction, we have used Logistic Regression model. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes i.e Logistics Regression is an algorithm that is used to predict the target variable which consists of 2 possible cases only.

Logistic Regression is easier to implement, interpret, and very efficient to train. It gives a good accuracy for many simple data sets. Also, this algorithm is less inclined to over-fitting. To evaluate the performance of the algorithm, we made use of metrics such as Accuracy, Recall, F1 score, ROC-AUC.

Result

From the Exploratory Data Analysis, we could generate insight from the data. How each of the features relates to the target.

Accuracy: 82%
F1: 88%
Precision: 80%
Recall: 98%
ROC AUC: 72%

Conclusion

In this project, we have implemented customer loan prediction using supervised learning technique for loan candidate as a valid or fail to pay customer. Logistic Regression Algorithm was implemented to predict customer loan. Optimum results were obtained using this algorithm. This algorithm gives high accuracy. From a correct analysis of positive points and constraints on the part, it can be safely ended that the merchandise could be an extremely efficient part. This application is functioning properly and meeting to all or any Banker necessities. This part is often simply obstructed in several different systems. There are numbers cases of computer glitches, errors in content and most significant weight of option is mounted in machine-driven prediction system, therefore within the close of future the therefore called software system might be created more secure, reliable and dynamic weight adjustment. In close to future this module of prediction can be integrated with the module of machine-driven processing system.

Future scope

The system is trained on old training dataset in future software can be made such that new testing data should also take part in training data after some fix time.

References

https://bhaktithaker.medium.com/loan-application-status-prediction-using-logistic-regression-fa19dbfb2f55

https://www.kaggle.com/ninzaami/loan-predication

https://courses.analyticsvidhya.com/courses/loan-prediction-practice-problem-using-python

https://github.com/novitaprahastha/Final_Project