***Loan Application Status Prediction***

Submitted by:

SNEHA SANTRA

**ABSTRACT**

Banks are making major part of profits through loans. Though lot of people are applying for loans. It’s hard to select the genuine applicant, who will repay the loan. While doing the process manually, lot of misconception may happen to select the genuine applicant. Therefore we are developing loan prediction system using machine learning, so the system automatically selects the eligible candidates. This is helpful to both bank staff and applicant. The time period for the sanction of loan will be drastically reduced.Here we use machine learning model to that can predict whether the loan of the applicant will be approved or not on the basis of the details provided in the dataset.

**INTRODUCTION**

* Business Problem Framing

A loan is the core business part of banks. The main portion the bank’s profit is directly come from the profit earned from the loans. Though bank approves loan after a regress process of verification and testimonial but still there's no surety whether the chosen hopeful is the right hopeful or not. This process takes fresh time while doing it manually. We can prophesy whether that particular hopeful is safe or not and the whole process of testimonial is automated by machine literacy style. Loan Prognostic is really helpful for retainer of banks as well as for the hopeful also.

* Conceptual Background of the Domain Problem

The project is sub-divided following section. These are:

1. Loading necessary libraries

2. Loading Dataset from a CSV file

3. Summarization of Data to understand Dataset (Descriptive Statistics)

4. Visualization of Data to understand Dataset (Plots, Graphs etc.)

5.Processing the data for modeling

6.skewness and outliers detection for better accuracy

7.Build the model and select the right model and save it

* Review of Literature

There are 13 columns including Loan Id,Gender,Loan ammount,Married status,Applicant Income,Education,Loan ammount term,creditb history,loan status.

And 5 columns are directly corelated with target variable those are Applicant Income,Coapplicant income,loan ammount,credit history,loan-ammount-term.



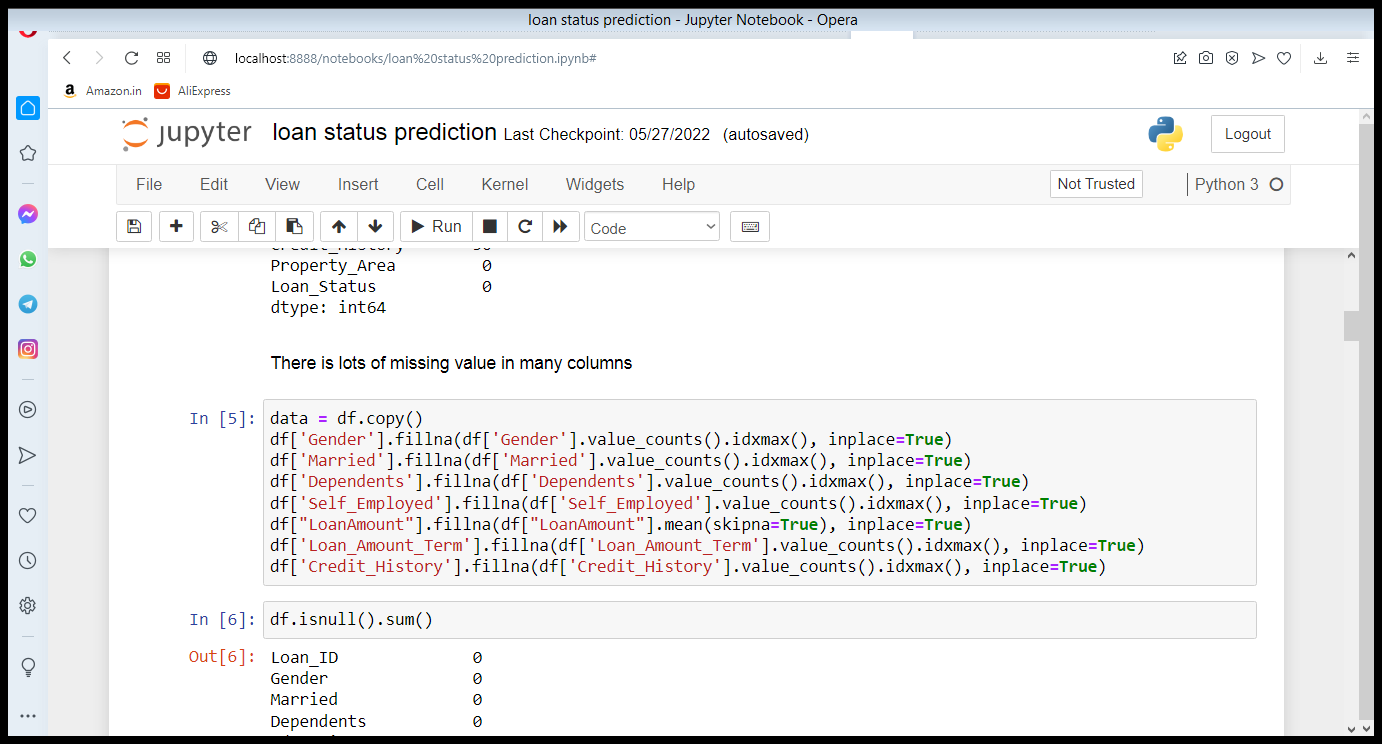
This is the heatmap of the dataset and it is declared the loan status either it is yes or no

So we use the Logistic Regression model to choose it .

**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem

Data Processing: There are some null values in this dataset so we have to fill this for data procesing

****

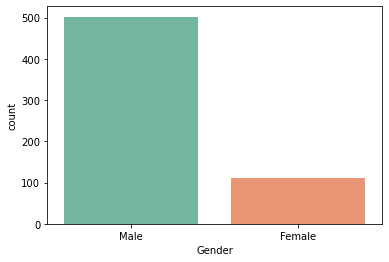
After that there is no null value present in this dataset

Then we have to do our EDA part to analysis the dataset properly

**EDA:**

****

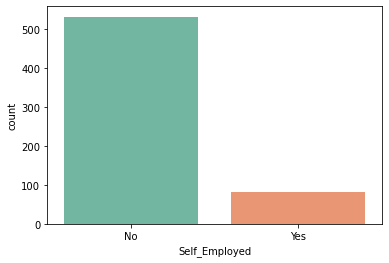
We do a countplot of married couple and we saw that applicant couple are mostly married couple

****

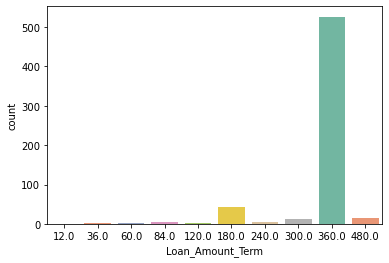
Here we saw most of all male candidate are apply for the loan



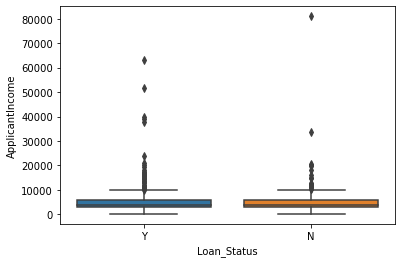
As we can see that most of the loan status is yes for the applicant most of the loan status is in positive



Here we see most of the applicant are not employed



Here we see in countplot that most of the loan ammount is 360.0



Here is a boxplot to see Mean ApplicantIncome of 0 and 1 are almost the same (o: no,1: Yes)

* Data Sources and their formats

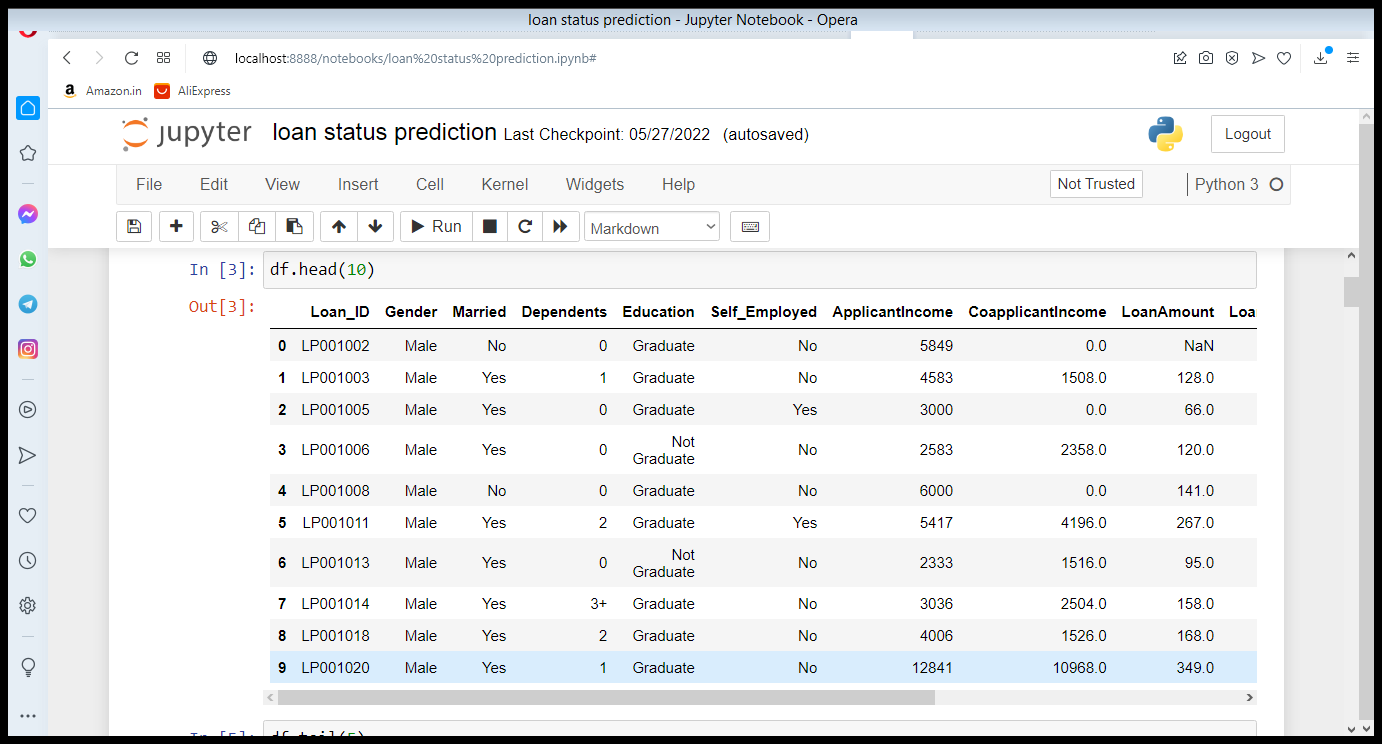
The source of this dataset is Github . The dataset consists

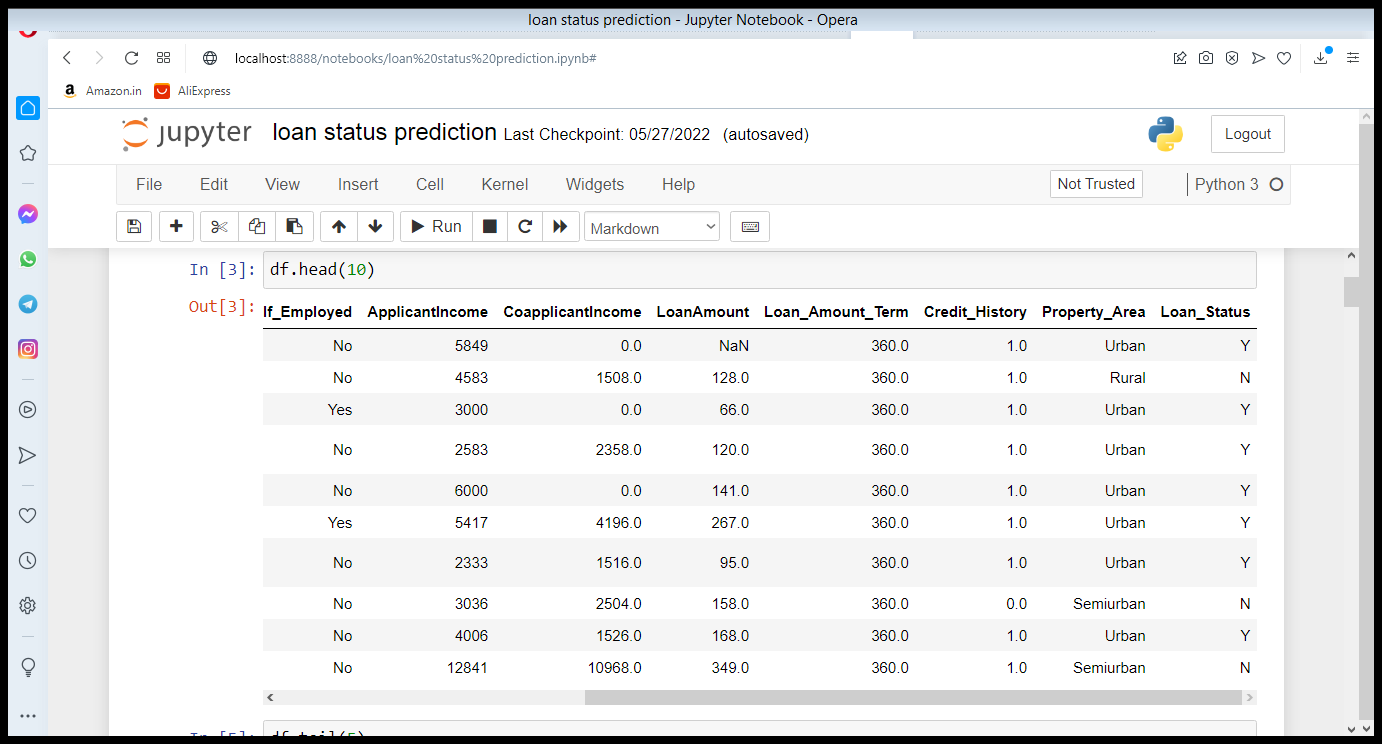
of text and metadata which is scrapped

The dataset has 13 columns and 614 rows in the

dataset where each row corresponds to a new article. For the classification problem under

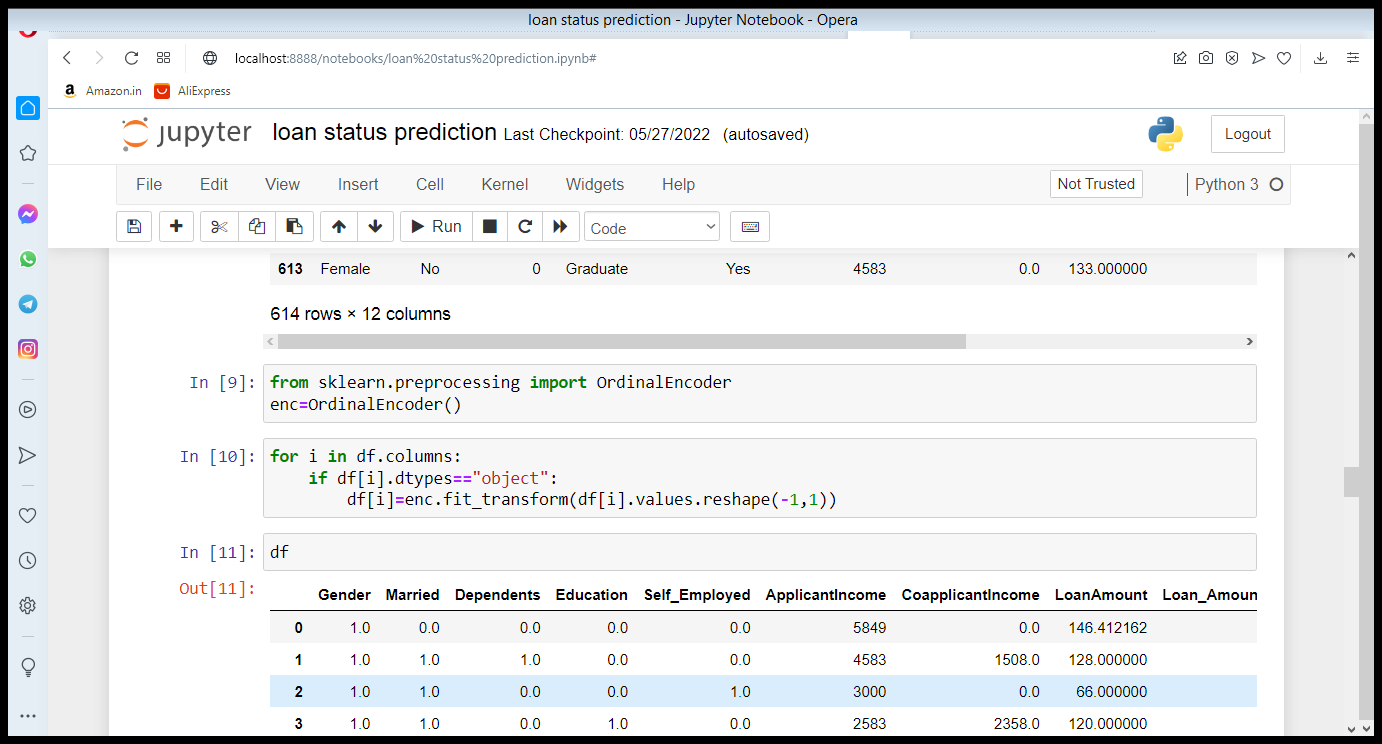
consideration we have used some object and some float columns for input variable and our output column is **loan status** that is also object type.





* Data Preprocessing

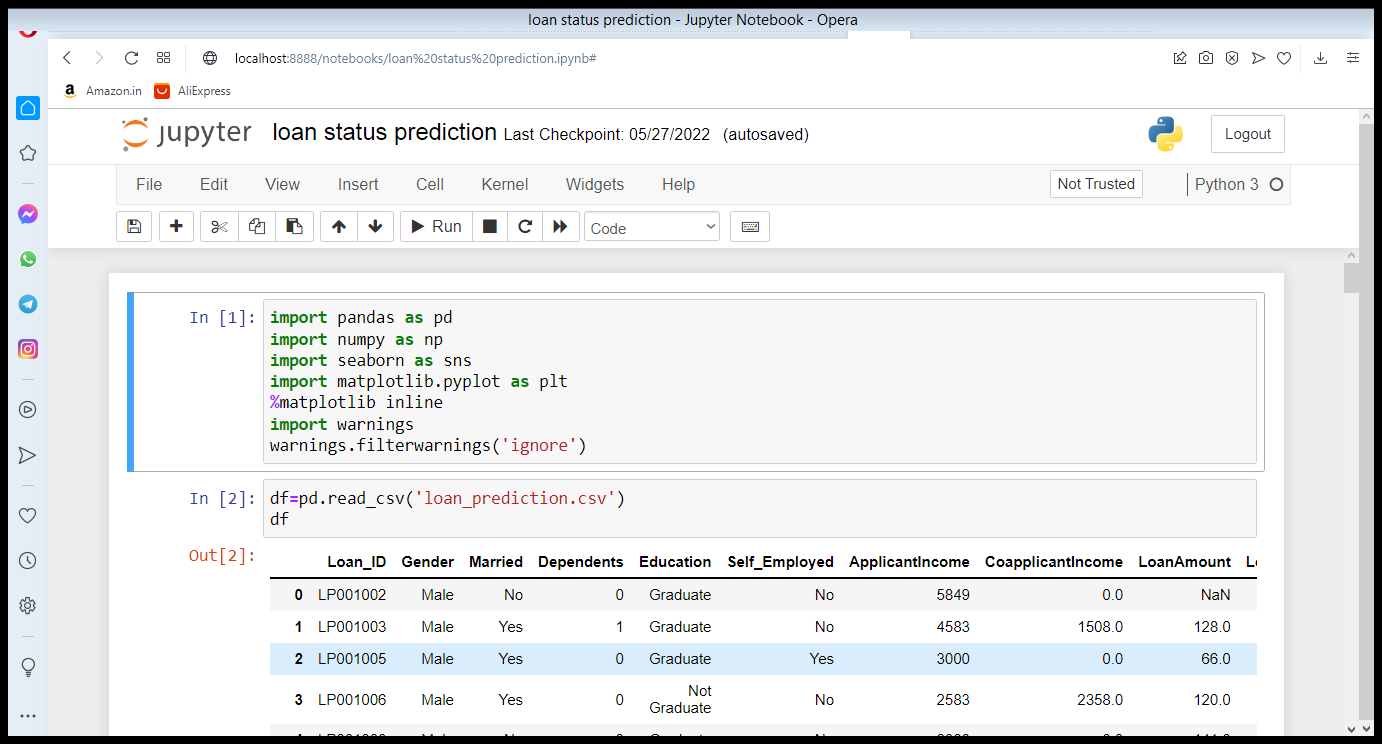
First of all fill the missing values of the dataset and drop the **Loan-ID** column because it is not needed for analysis and than endoded the object value into float value by encoding



And the dataset is ready for model building process

* Hardware and Software Requirements and Tools Used

Here we use lots of liaberies like pandas,numpy,matplot,seaborn, and we use python language for the coding purpose and import some other metrics liaberies also for model building like sklearn metrics ,classification report,accuracy report etc.

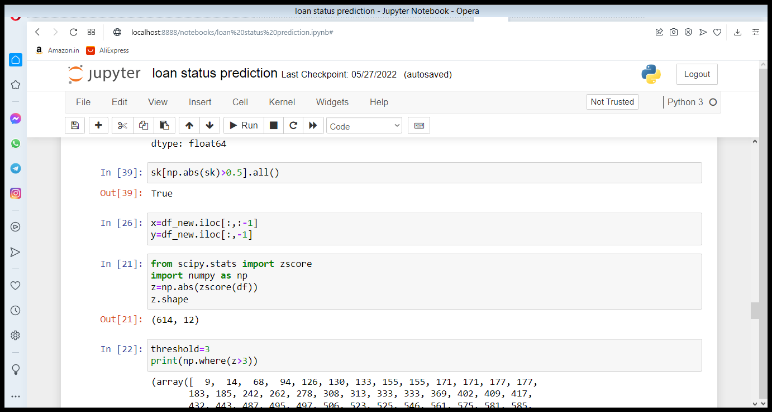
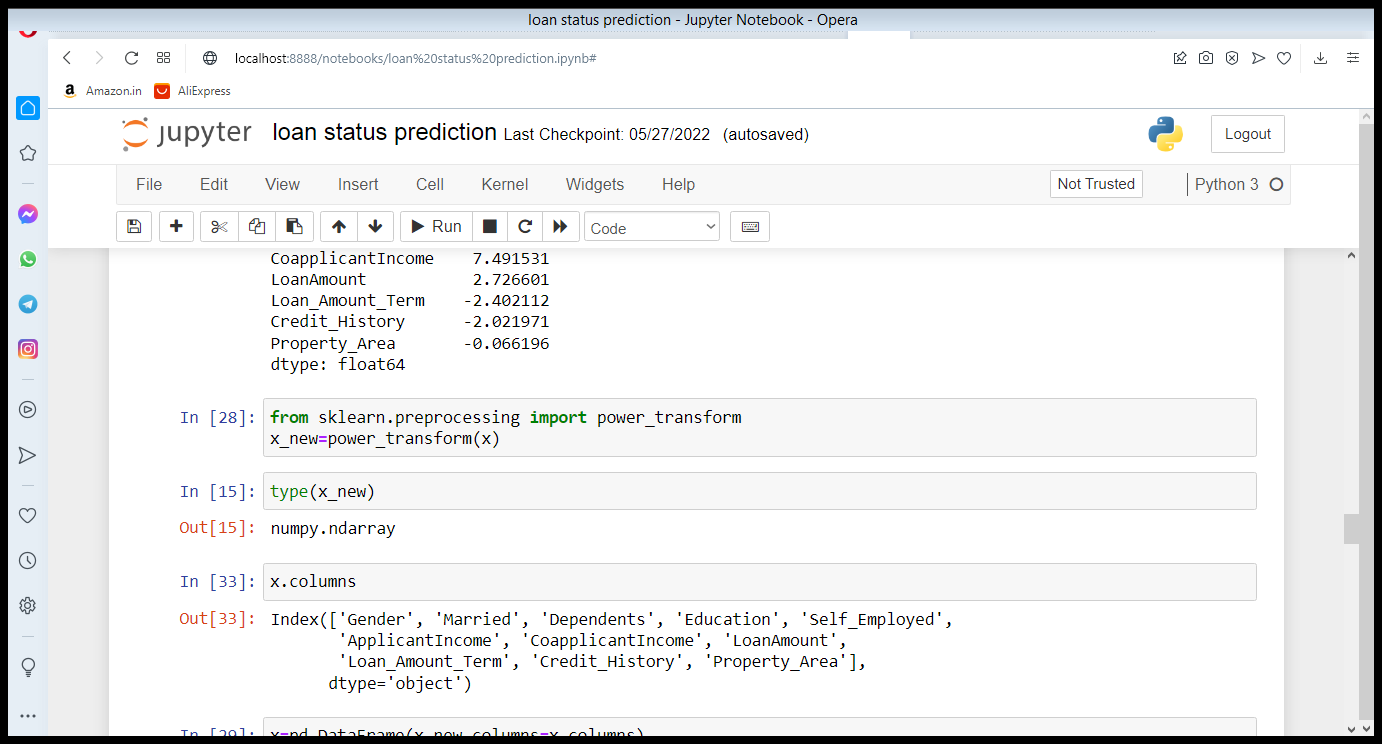


**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

First of all detect the skewness and outliers

Detect the skewness and solve it by power transfrom method and detect the outliers and solve it by z-score method

****

Then it is set for model building process and we drop the target variable for the further process

* Testing of Identified Approaches (Algorithms)

For model building process we consider 5 methods for the best accuracy

1**.Logistic Regression**



2.**DecisionTree Classifier**



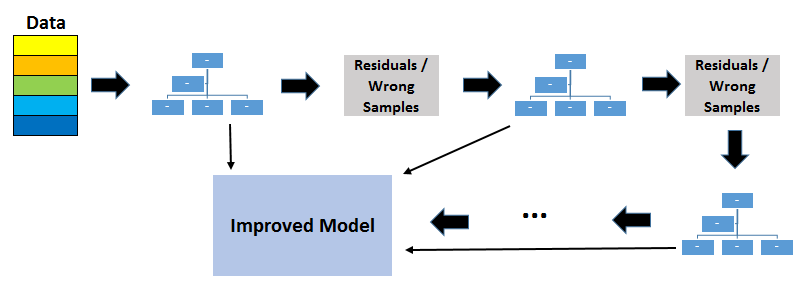
3**.RandomForest Classifier**



4.**SVC model**



5.**GradientBoosting Classifier**

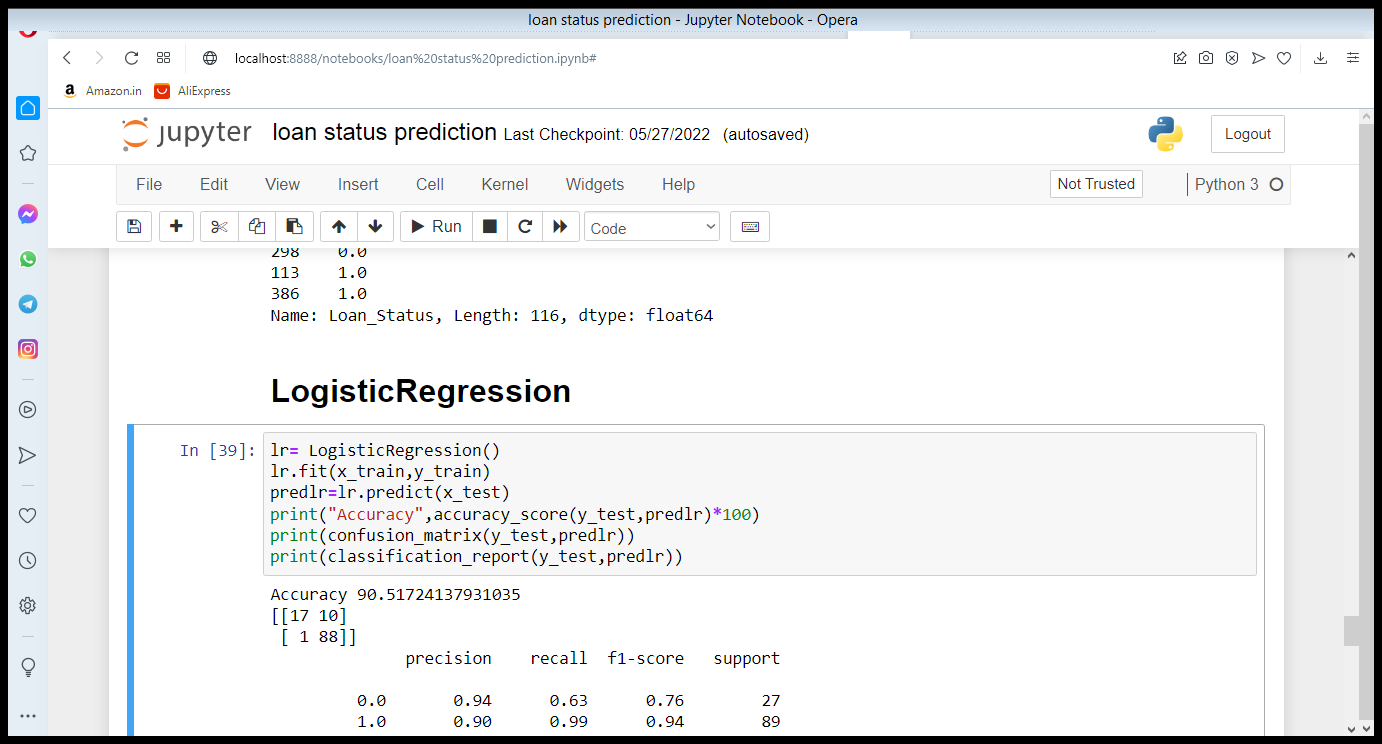


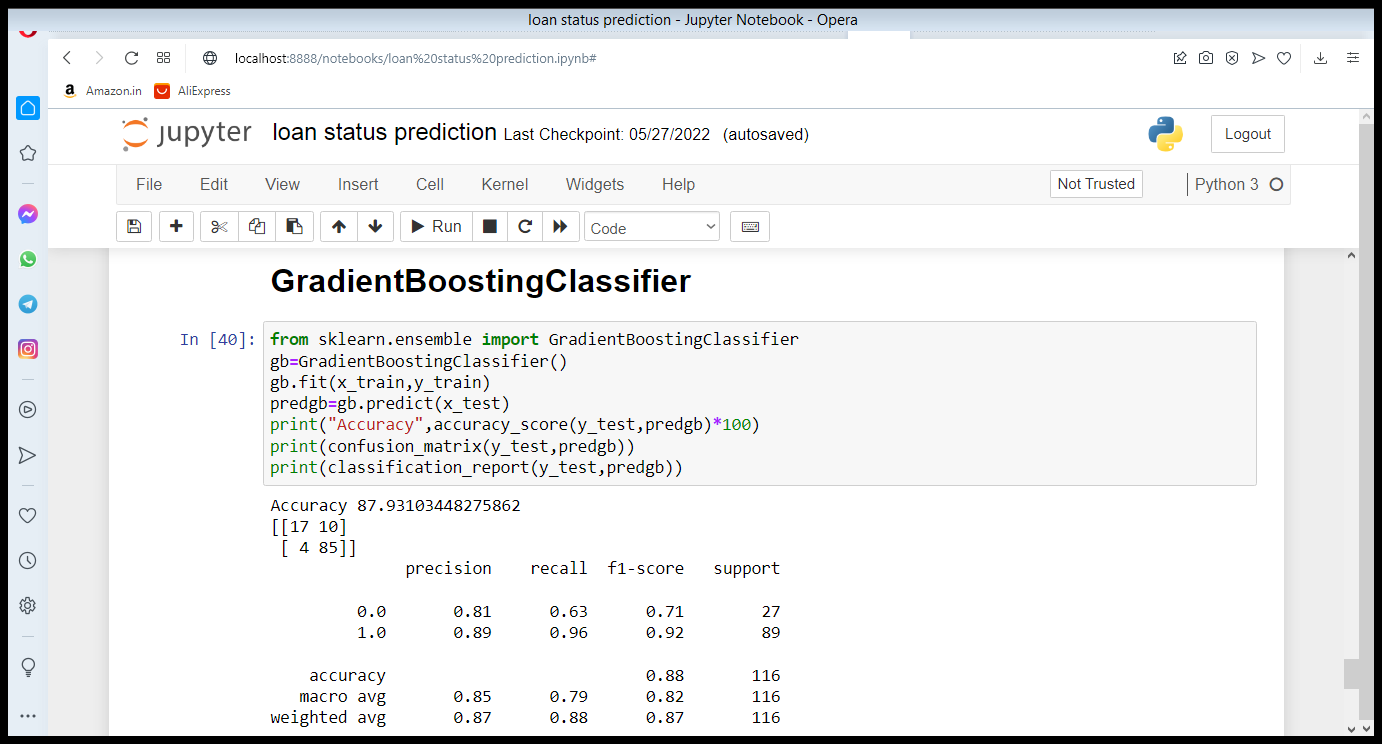
And then we use cross validation method for identify the best model accuracy score and save it

* Key Metrics for success in solving problem under consideration

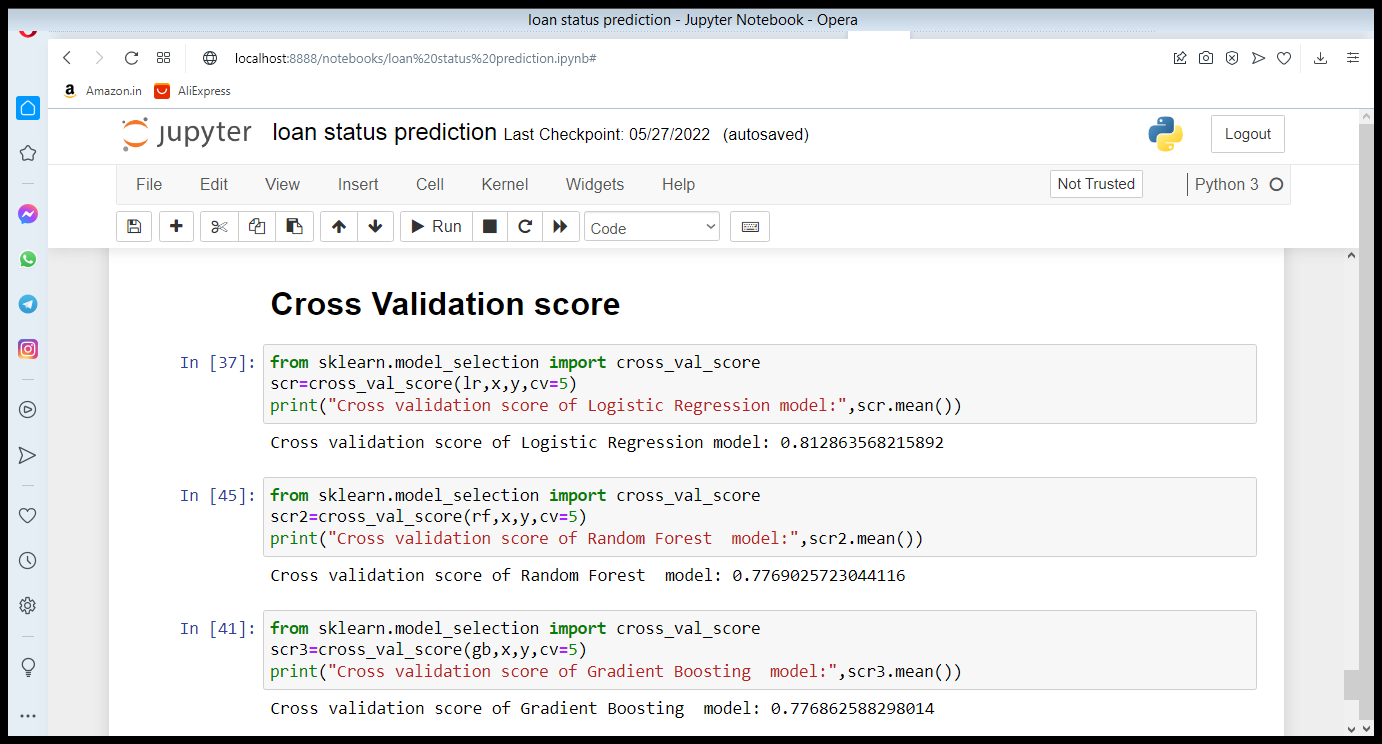
As we saw we use 5 methods but the best accuracy score is given in logistic Regression method

Best two model is logistic regression and GradientBoosting Classifier





Then the cross validation accuracy score is given best score for logistic regression method

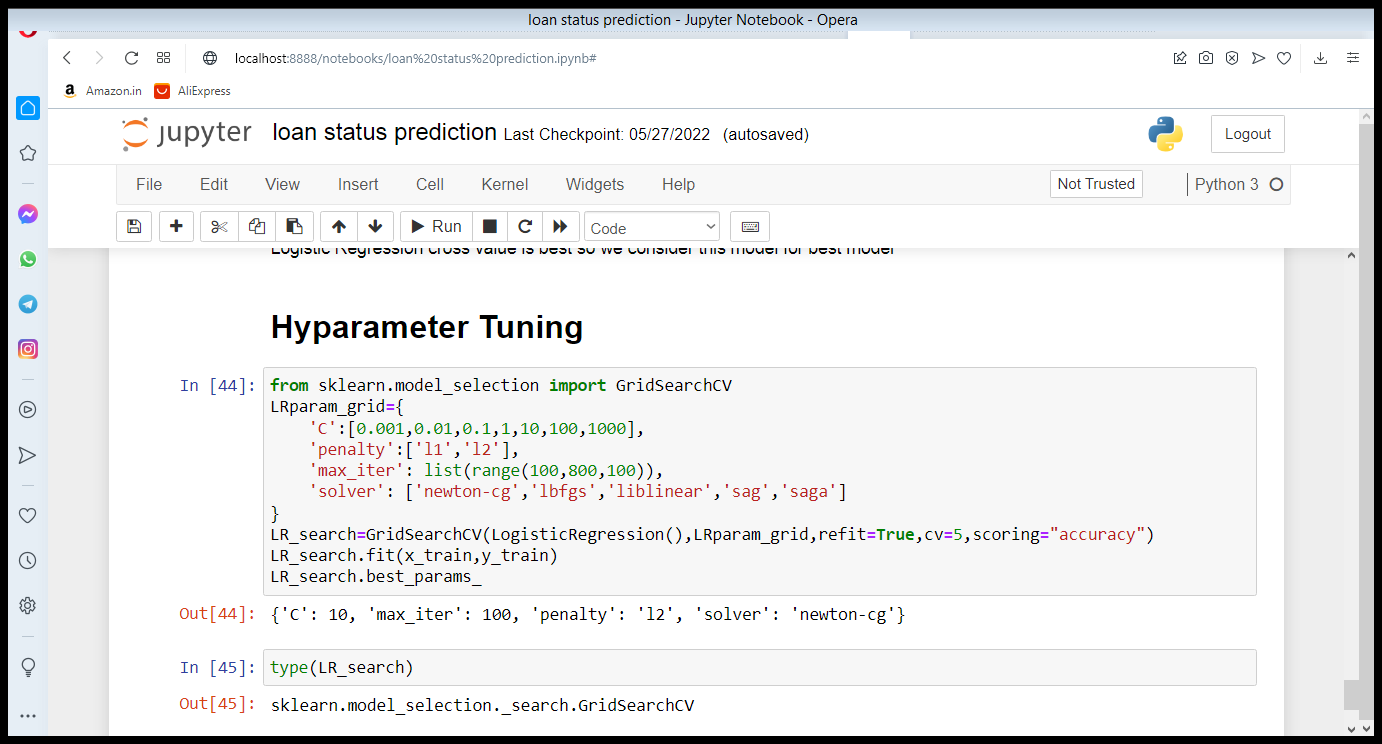


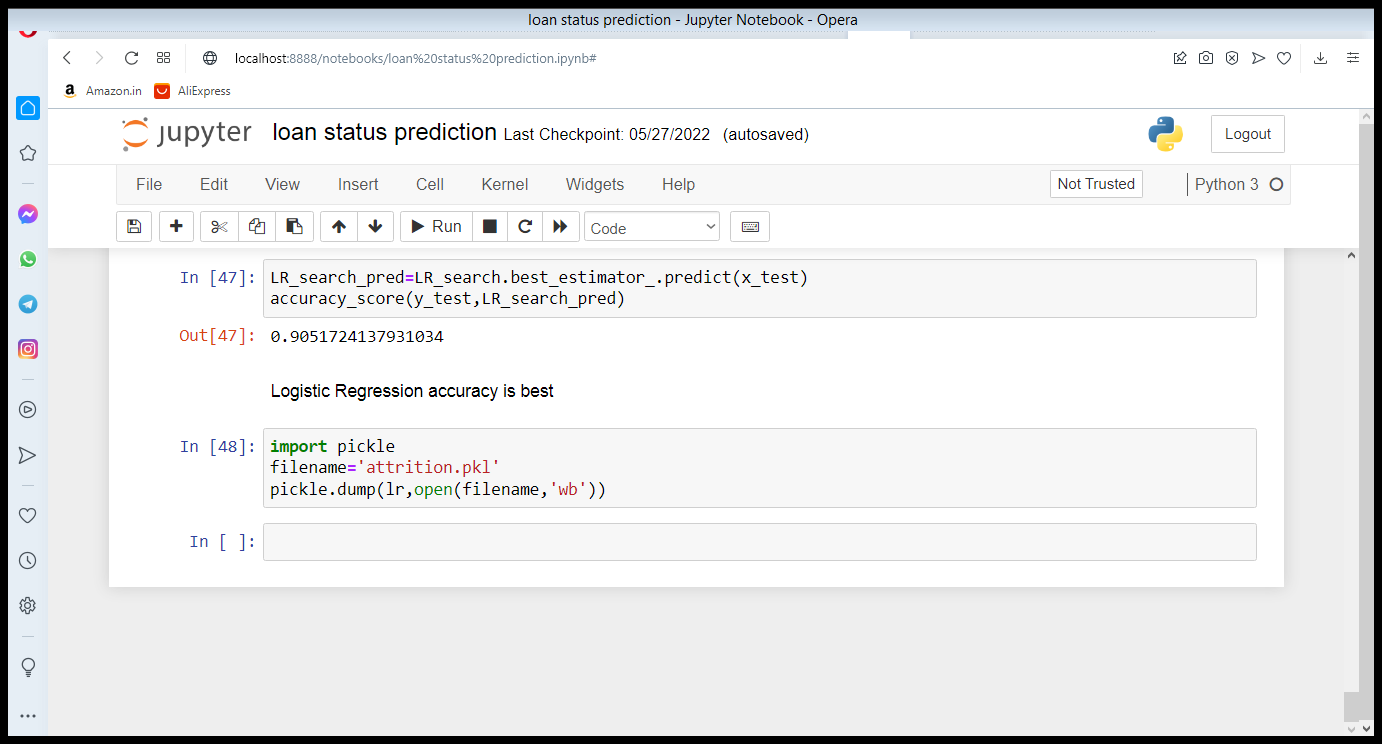
So we choose the lr method for the best model for machine learning method and save it for the process.

* Interpretation of the Results

We finally do the hyperparameter tuning for logistic regression method for the best accuracy and the accuracy score is 0.90

We import GridSearchCv for that and clculate the best grid for best result





**CONCLUSION**

* Key Findings and Conclusions of the Study

After the Final Submission of test data, my accuracy score was 90%

Feature engineering helped me increase my accuracy.

Amazingly Logistic Regression worked better than all other Ensemble models.

* Learning Outcomes of the Study in respect of Data Science

From a proper analysis of positive points and constraints on the member, it can be safely concluded that the product is a considerably productive member. This use is working duly and meeting to all Banker requisites. This member can be freely plugged in numerous other systems. There have been mathematics cases of computer glitches, violations in content and most important weight of features is fixed in automated prophecy system, so in the near future the so – called software could be made more secure, trustworthy and dynamic weight conformation. In near future this module of prophecy can be integrated with the module of automated processing system. The system is trained on old training dataset in future software can be made resembling that new testing date should also take part in training data after some fix time.

**Thank You**