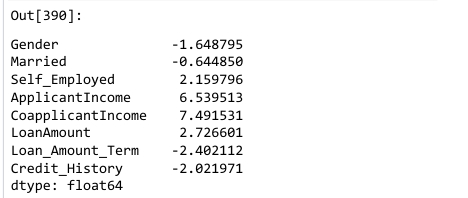
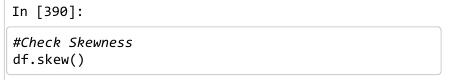
|  |
| --- |
| Problem definition  Many people apply loan in banks for fulfilling their financial needs. The responsibility of analyzing weather a person repays that loan or not lies on the bank official and then to approve the loan. A person pays or not may depend on multiple factors like gender, Income, Education, Property area, Loan Amount, married or not, dependents and so on. In this project we will use the historical data and clean the data, analyze the data and build the machine learning model to predict if the bank official will approve or not based on the factors applicable for that person who has applied for the loan.  Data Analysis  Gender:  According to the dataset given 71.57% males have got their loan application approved and 49.1 females have got their application approved. With this we may say that males have greater chance of loan approval than females. Let’s move to the next factor    Marital status:  Out of 614 people who have applied for loan 398 are married 213 are unmarried and 3 more marital status is not documented, among married 290 people loan application was approved that amounts to 76.8 % approval chance similarly out of 213 people unmarried people who have applied for loan 140 applications were approved that amounts to 65.72 % approval chance    Dependents:  Out of 614 loan applications 345 people had no (0) dependents, 102 people have 1 dependent, 101 people have 2 dependents and 51 people have 3+ dependents. 68.98 % of applications were approved who had 0 or no dependents, 64.7 % of applications were approved who have 1 dependent, 75.24 % of application were approved who have 2 dependents, 67.7 % of applications were approved who have 3+ dependents    Education:  Out of 614 loan applications 480 were graduated and 134 were not graduated and 340 were approved of graduates an 82 were approved for non-graduates so if. So 70.83 % application were approved for graduates and 61.19 % applications were approved for non-graduates    Self Employed:  Out of 614 applications 82 were self employed and 500 were not self-employed. Out of 82 applications who are self employed 56 were approve. Out of 500 application who are not self employed 343 applications were approved so that says 68.29 % of applications who were self employed have been approved and 68.6 % of applications who were not self employed have been approved.    Property Area:  Out of 614 applicants 233 applicants are from semi urban area, 202 are from urban area and 179 are from rural area, 76.82 % applications have been accepted from semi urban, 65.84 % applications have been accepted from urban area and 61.45 % applications have been accepted from rural area |
| EDA (Exploratory data analysis) |
| Exploratory Data Analysis (EDA) is an approach of analyzing data sets to summarize their main characteristics, often with visual methods, a statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modelling or hypothesis testing task. we can say that EDA is statisticians’ way of storytelling where you explore data, find patterns and tell insights. EDA is a phenomenon under data analysis used for gaining a better understanding of data aspects like: - main features of data variables and relationships that hold between them identifying which variables are important for our problem  **1. variable identification:**  Looking at the input data, what will be the output variable  **2. checking and clearing null values:**  It is possible that there may be null values in the dataset. missing values usually occurs when an entry is left empty. It can be filled with the mean, median or mode of the column through fillna () function. Or sometimes SimpleImputer () function is used to replace the null values along with label encoding them in this dataset there are Null values in gender, Married, Dependents, Self\_Employed, LoanAmount, Loan\_Amount\_Term, Credit\_History. Let’s clear Gender, Self\_Employed, Loan\_Amount\_Term, Credit\_History and Dependents will mode and then lets clear rest of the null values with mean  **3. Handling Categorical variables**  Most of the statistical models cannot take Objects / Strings as input they only takes numbers as inputs, with LabelEncoder () it is possible to categorize the string into Numbers as 1,2,3 and so on like, Gender, Married, Dependents, education, Self\_Employed and Property\_Area have string input they all are converted through LabelEncoder().fit\_transform function  **4. Descriptive Statistics** –  It is a way of giving a brief overview of the dataset we are dealing with, including some measures and features of the sample. For this we can use the describe() function the describe() function tells the mean(), min(), max(), 25th percentile, median 50th percentile and 75th percentile of the data.     1. **Correlation**   Correlation is the statistical metric for measuring to what extent Different variables are interdependent, like if one variable changes how it affects the change in other variables. corr () function is used to see the correlation among the dependent variable and independent variable you can see correlation in the following figure |

1. **Skewness**

The data can be right skewed or left skewed if the median or mean is high and data is highly spread it can be observed through the skew() method, if the skew score is negative and greater than 5 it means data is negatively skewed on left side and if the data is more than +5 it means the data is skewed on right side.



1. **Pre-processing Pipeline**

Pipelines are the special way to simplify the code, Pipeline is generally used if we have to perform the code repeatedly usually when there is different train and Test data

Here Our EDA process is completed now moving towards next step

**Building Machine Learning Models**

1. **Separating Features and Target column** –

It is necessary to separate the independent/Features column into a variable (x) and target column into a variable (y). here we have to separate all columns in x Data Frame (variable) and income variable in y Data Frame (variable)

1. **Splitting the Data for Training and Testing**

In ML the separated data is split into 4 parts for Training and Testing of features (x) and for Training and Testing of Target (y) like x\_train, x\_test, y\_train, y\_test. It is possible through a inbuilt library of sklearn’s train\_test\_model

1. **Training the Models**

To find the best model it is necessary to train 3-4 models, In the same way I have trained LogisticRegressionModel, KNeighborsClassifier, SVC, DecisionTreeClassifier, RandomForestClassifier, AdaBoost, GaussianNB, BaggingClassifier and GradientBoosting to check which model is giving the best max score

i) LogisticRegression Model is giving the max score of 0.85 and the random state of 63

ii) KNeighborsClassifier Model is giving the max score of 0.84 and the random state of 73

iii) SVC Model is giving the max score of 0.85 and the random state of 63

iv) DecisionTree Model is giving the max score of 0.75 and the random state of 72

v) RandomForest Model is giving the max score of 0.84 and the random state of 73

vi) AdaBoost Model is giving the max score of 0.84 and the random state of 77

vii) GaussianNB Model is giving the max score of 0.85 and the random state of 104

viii) BaggingClassifier Model is giving the max score of 0.82 and the random state of 69

ix) GradientBoostingClassifier Model is giving the max score of 0.82 and the random state of 64

Here the above models are giving quite good max scores, but you can see that Logistic Regression is the best so let’s choose Logistic Regression as our best algorithm for this data set

7) **Exporting the Model**:

Let’s export our model using joblib library in the form of obj or .pkl

