MIT School of Computing
Department of Information Technology
Machine Learning Lab

Assignment no: -4

AIM: Perform an EDA and also Improve the mean error in the test set:

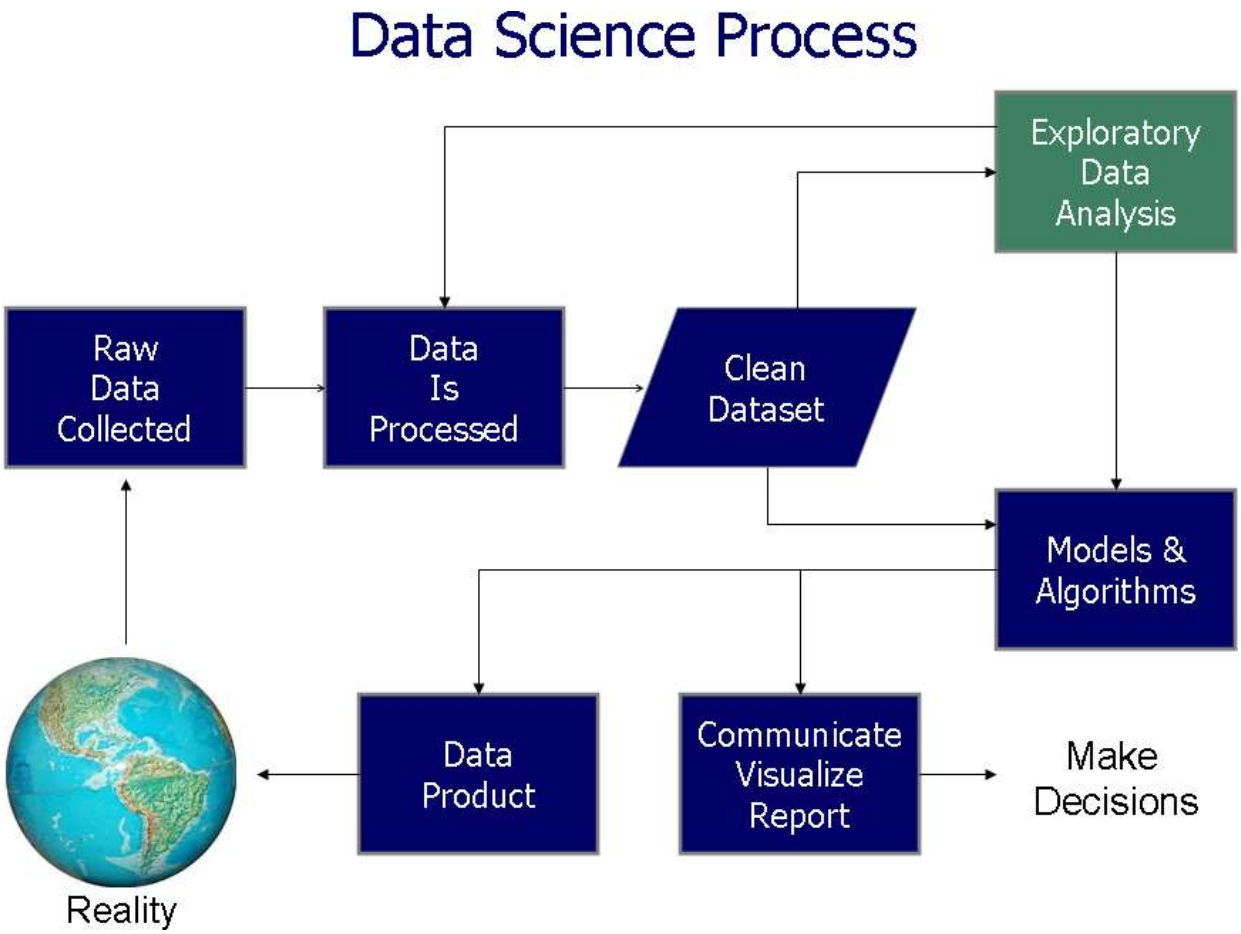
INDEX TERMS: Data Preprocessing, EDA, Mean error Test.

PROBLEM DEFINITION: Using feature engineering or insight from EDA or analyzing the output of regression model, iterate to improve mean error in the test set.

Theory:

Exploratory data analysis:

Exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.



Here are the main reasons we use EDA:

- Detection of mistakes • checking of assumptions
- Preliminary selection of appropriate models
- Determining relationships among the explanatory variables,
- Assessing the direction and rough size of relationships between explanatory and outcome variables.

The four types of EDA:

1. **Univariate non-graphical**: The data that come from making a particular measurement on all of the subjects in a sample represent our observations for a single characteristic such as age, gender, the speed at a task, or response to a stimulus.

2. **Multivariate non-graphical**: The multivariate non-graphical type of EDA generally depicts the relationship between multiple variables of data through cross-tabulation or statistics.

3. **Univariate graphical**: Unlike the non-graphical method, the graphical method provides the full picture of the data. The three main methods of analysis under this type are histogram, stem and leaf plot, and box plots. The histogram represents the total count of cases for a range of values. Along with the data values, the stem and leaf plot shows the shape of the distribution. The box plots graphically depict a summary of minimum, first quartile median, third quartile, and maximum.

4. **Multivariate graphical:** This type of EDA displays the relationship between two or more set of data. A

Algorithm/steps:

Step 1: Import your data set and have a good look at the data.

Step 2: Now let's try to classify these columns as Categorical, Ordinal or Numerical/Continuous.

Step 3: Now we are all set to perform Univariate Analysis.

Step 4: Now let's find some relationship between two variables, particularly between the target variable "Loan_Status" and a predictor variable from the dataset. Formally, this is known as bivariate analysis.

Step 5 : Let's move on to analyzing more than two variables now. Yay !! You guessed it right, we call it "Multivariate analysis". You should first create an hypothesis like in step 3 and act in that direction.

Step 6: calculate the mean absolute error and Coefficient of Regression

Write your comment in conclusion