Sardar Patel Institute of Technology,Mumbai

Department of Electronics and Telecommunication Engineering

T.E. Sem-V (2018-2019)

ETL54-Statistical Computational Laboratory

Lab-5A: Regression Analysis and Modeling

Name: Manish Dsilva **Roll No.15**

Objective: To carry out linear regression (including multiple regression) and build a regression model using Python Platform

Outcomes:

1. To learn how to define, fit, and use a model in Python

2. To interpret the results

System Requirements: Ubuntu OS with Anconda platform with Pandas, numpy, scipy, matplotlib, seaborn and scikit-learn ML library.

Part-A: Simple linear regression and Multiclass linear regression with data preprocessing (Handling NA values)

Use the houseprices.csv files to build the models and evaluate the models.

Refer Jupyter Notebook 51 and 52 for this part of the experiment.

General Steps:

1. Load the dataset (Use pandas)

- Here we load the Dataset from a CSV into a DataFrame

2. Data Preprocessing (Handling NA values)

- If the Values are NA we must fill them with mean or median

3. Exploratory Data Analysis (understanding the relationships between the variables with help of plot, scatter-plot, enery-plot etc) Use matplotlib

- We use a scatter plot to Visualize our Data

4. Data Partition (80% for training and 20% for testing) (Use scikit-learn)

- Important Step to Split Data into Train Data and Test Data

5. Build the model (use scikit learn)

- We build a model that is establish a relation between dependent variable and independent variable.

6. Summarize the model.

- We would determine coefficient and intercept

7. Prediction

- We predict the test values of X with the help of our model.

8. Evaluate the model

- We evaluate the model based on error our model makes.

9. Tuning the model

- With the help of these parameters we Fine tune our model

Part-B: Logistic Regression

Use University admission binary.csv file.

Follow the general steps to carry out logistic regression as mentioned in Part-A.

Calculate the performance metrics-Accuracy, Miss-classification rate, Receiver operating characteristics.

Refer Jupyter Notebook 53 for this part of the experiment.

Area Under Curve:

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. Before defining AUC, let us understand two basic terms :

True Positive Rate (Sensitivity) : True Positive Rate is defined as *TP/ (FN*+TP). True Positive Rate corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points.

$$TruePositiveRate = \frac{TruePositive}{FalseNegative + TruePositive}$$

False Positive Rate (Specificity) : False Positive Rate is defined as *FP / (*FP+TN). False Positive Rate corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points.

$$FalsePositiveRate = \frac{FalsePositive}{FalsePositive + TrueNegative}$$

False Positive Rate and True Positive Rate both have values in the range [0, 1]. FPR and TPR bot hare computed at threshold values such as (0.00, 0.02, 0.04, …., 1.00) and a graph is drawn. AUC is the area under the curve of plot False Positive Rate vs True Positive Rate at different points in [0, 1].As evident, AUC has a range of [0, 1]. The greater the value, the better is the performance of our model.

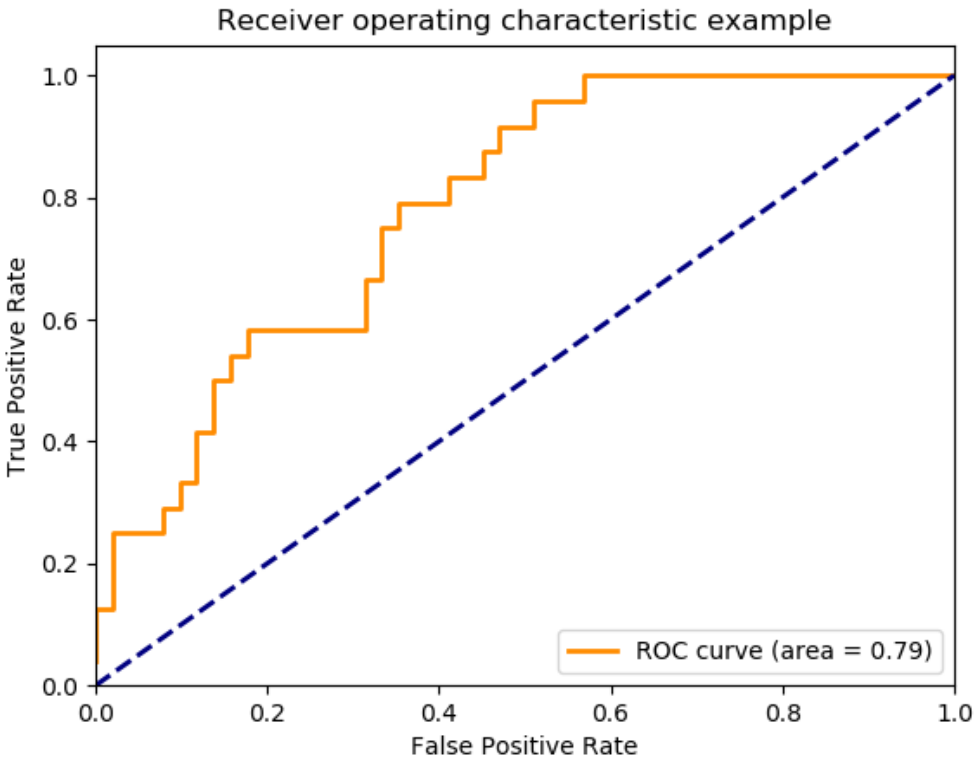
Receiver Operating Characteristics:

ROC stands for Receiver Operating Characteristic (from Signal Detection Theory) initially - for distinguishing noise from not noise so it's a way of showing the performance of Binary Classifiers only two classes - noise vs not noise it's created by plotting the fraction of True Positives vs the fraction of False Positives

- Precision and Recall are popular metrics to evaluate the quality of a classification system
- ROC Curves can be used to evaluate the tradeoff between true- and false-positive rates of classification algorithms

Properties:

- ROC Curves are insensitive to class distribution
- If the proportion of positive to negative instances changes, the ROC Curve will not change



Conclusion:

- Linear Regression is a model where we predict a numerical data(dependent variable) with respect to one or more features(independent variables)

- Logistic Regression is a binary classification model where we predict based on independent variables whether the classification is ‘0’ or ‘1’ or such discrete levels

- We have learnt about multi-variate(multiple independent variables) and single-variable regression.

- We have learnt about various parameters which tell us about the accuracy of our predicted model.(ROC,AUC)

- We have learnt how to build a model for our data,train it and then use it for prediction purposes on Python using Jupyter Notebook

This experiment has 3 attachments in the form of Jupyter Notebooks(.ipynb) 51,52 and 53 which contain the Code,Output and Graphs.