D. Y. Patil College of Engineering, Akurdi, Pune,411044
Department of Artificial Intelligence & Data Science

Machine Learning Mannual

Assignment No. 3

Regression Analysis

1.1 Title: To apply Linear Regression for Uber price prediction

1.2 Aim: Apply Linear Regression Algorithm on Uber dataset and predict price.

1.3 Objective: To develop uber price prediction system and implement in Python.

1.4 Software Requirements:

Anaconda with Python 3.7

1.5 Hardware Requirement:

PIV, 2GB RAM, 500 GB HDD

1.6 Learning Objectives:

Learn Linear, Ridge & Lasso Regression algorithm on given dataset.

1.7 Problem Statement:

Predict the price of the Uber ride from a given pickup point to the agreed drop-off location. Perform following tasks:

1. Pre-process the dataset.
2. Identify outliers.
3. Check the correlation.
4. Implement linear regression and Ridge, Lasso regression models.
5. Evaluate the models and compare their respective scores like R2, RMSE, etc.

1.8 Theory Concepts:

The Assignment is about on world's largest taxi company Uber inc. In this assignment, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

1.8.1 Data Preprocessing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning

models. Data preprocessing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model. It involves below steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

Outliers - Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations. To ensure that the trained model generalizes well to the valid range of test inputs, it's important to detect and remove outliers.

Correlation - Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable. Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

Positive Correlation: Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.

Negative Correlation: Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.

No Correlation: Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable(s) doesn't increase or decreases. Linear regression - Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis.Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent (y) variables, hence called as linear regression. Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable

is changing according to the value of the independent variable. Mathematically, we can represent a linear regression as:

$y= a0+a1x+ \varepsilon$

Here, y= Dependent Variable (Target Variable)

x= Independent Variable (predictor Variable)

a0= intercept of the line (Gives an additional degree of freedom)

a1 = Linear regression coefficient (scale factor to each input value).

ε = random error

The values for x and y variables are training datasets for Linear Regression model representation.

1.8.2 Ridge & Lasso Regression

1.9 Algorithm

1. Import the Required Packages
2. Read Given Dataset
3. Import the Linear Regression
4. Data Preprocessing
5. Define input & output
6. Initialize the model
7. Fit the dataset
8. Evaluate the model

1.10 Evaluation Parameters after Implementation of Code

Parameters	Linear Regression	Ridge Regression	Lasso Regression
MSE			
RMSE			
MAE			
R2_score			
Adj_R2			

1.11 Conclusion

Thus we learn how to implement Linear Regression model with Ridge & Lasso Regression models on uber dataset.