SDAIA
الهيئة السعودية للبيانات والذكاء الاصطناعي
Saudi Data & AI Authority

Bootcamp T5

TAXI TRIP REGRESSION

Nasser Alqahtani |Mukhtar Al bin Hamad | 4/12/2021

Abstract

The second project for Data science Bootcamp T5 on Ordinary Least Squares regression (OLS) or Regression. Through the project by building a machine learning regression model. The main purpose of this project is to provide predictions for the price of trips in NYC by using a yellow taxi. Using python libraries such as Pandas, seaborn, and other useful libraries. The first phase of the project was divided into a dataframe for training, verification, and testing on the ratios 98%, 1%, and 1% respectively. The second phase was to clean, prepare, and handle data, check for null and duplicates, find anomalies and features that don't need it, and drop them. The third phase is to build the models and select the best score of R^2 and least errors (RMSE, MAE).

Design

By applying the dataset on machine learning models such as linear regression, polynomial regression, ridge regression, lasso regression, ElasticNet, and Knn. to predict the prices of the trips.

Data

Dataset for yellow taxi trip in NYC in July 2021 This data dictionary describes yellow taxi trip data. The data base about 2,821,515 trips with 18 features like (trip_distance, RatecodeID, payment_type, etc) which main this project is Multivariate regression. The target in this project is Total amount.

Algorithms

Preparing the data, Feature Engineering, and selection:

1. Exploration the data and visualization.
2. Feature Selection by calculate the features correlation.
3. Engineering by converting categorical values to dummy.

Methods:

Linear regression, polynomial regression, ridge regression, lasso regression**,** ElasticNet, and Knn. have been used to predict the prices of the trips. By splitting the dataset to train set, validation set, and test set to measure each model scores, the best model R^2 score shows in polynomial regression.

Tools

- Python and Jupyter Notebook.
- Numpy and Pandas for data manipulation.
- Matplotlib and Seaborn for plotting visualization.
- Sklearn for ML algorithms.