Project: Metro Interstate Traffic Volume

Abstract

This aim of this project is to create a regression predictive model using different machine learning models to predict traffic volume of metro interstate located in Minneapolis US for the purpose of planning maintenance and for advertisement purpose. I worked with data provided by UCI Machine learning repository. To build the model, I used five features from the dataset (e.g. day_off , Temp , clouds_all , Hour , day_of_week) to predict the target which is traffic volume. I used three machine learning models: Linear regression, decision tree regressor, random forest regressor. Decision tree regressor and random forest regressor gave a promising result.

Design

The need of predicting of traffic volume is critical for the metro company to plan for regular maintenance. Moreover, the prediction of traffic volume is important for the sake to focus on advertisement to improve the sales in these advertisement companies. Using Machine learning models to predict the traffic volume will enable the metro company to plan accordingly for the next year to avoid and sudden stop in metro due to a lack of maintenance

Data

The dataset has 48,204 traffic volume observations with 8 features of which 4 are numerical, 3 categorical and 1 datetime. Using one hot encoding, 11 numerical values can be extracted from one of the categorical features. Moreover, datetime can be split into hour, day of the week, month, and year. Also, holiday feature is converted into Boolean values which is a numerical value.

Algorithms

- Data cleaning (e.g deleting outliers , drop null values)
- One hot encoding for weather_main column
- Split date_time into hour, day , month and year.
- Convert holiday column into Boolean valuue
- Creating day_off column following the logical function: off_day_colunm = Day + Holiday
- Plot Correlation heatmap to find the features related with target.
- Selecting the features that are most correlated with the target.
- Standardizing the selected features.

Models

Linear regression, decision tree regressor, random forest regressor.

Model Evaluation and Selection

The entire dataset of 48,204 records was split into 80/20 train vs. test. At first the datasets were trained via three models and based on the result mentioned below linear regression was eliminated due to low R- square value. Next the parameters of decision tree regressor, random forest regressor were tuned with 5-fold cross validation on the training portion only to select the best argument parameters where the result is mentioned below.

First Result for three models on test set.

Linear regression: R- square = 17.8% MAE = 1583.86

Decision Tree Regressor: R- square = 90.7% MAE = 389.46

Random Forest Regressor: R- square = 91.6% MAE = 363.98

Second Result for two models of test set

Decision Tree Regressor: R- square = 93.9% MAE = 492.23

Random Forest Regressor: R- square = 93.9% MAE = 487.17

Tools

- Numpy and Pandas for data manipulation
- Scikit-learn for modeling
- Matplotlib and Seaborn for plotting

Communication

The below figure how apart is the actual value from the predicted value for 15 samples with the test set only.