Project 4: Metro Interstate Traffic Volume Prediction

Data Science and Machine Learning for Engineering Applications

Politecnico di Torino

Introduction

We want to train a regression model that predicts the metro traffic volume given the weather, holiday, and date variables. The model can be useful for increasing the number of trains in the metro on days and time slots when the highest volume of people will be estimated, also based on weather forecasts and holidays. Similarly, the number of trains can be reduced when less traffic volume is estimated, so as to save resources and reduce pollution.

Objective

Estimate the metro traffic volume given the weather and date features. Optional: identify the attributes most correlated with traffic volume.

Dataset and Task

Task

The task is a regression problem. The dataset consists of two files in CSV format: a *Metro_Interstate_Traffic_Volume_train.csv* file that contains the data to be used for training (and possibly validation) and a file *Metro_Interstate_Traffic_Volume_test.csv* that should be used for testing. Remember that the test data should not be used in the training phase under any circumstances. Once the model has been trained, the final evaluation should be done on the test data. The test data has been selected based on the data (last year).

Dataset Attributes

The following attributes are present in the dataset:

- holiday: Categorical US National holidays plus regional holiday, Minnesota State Fair.
- temp: Numeric Average temperature in kelvin.
- rain_1h: Numeric Amount in mm of rain that occurred in the hour.
- snow_1h: Numeric Amount in mm of snow that occurred in the hour.
- clouds_all: Numeric Percentage of cloud cover.
- weather_main: Categorical Short textual description of the current weather.
- weather_description: Categorical Longer textual description of the current weather
- date_time: DateTime Hour of the data collected in local CST time.
- year: Year as integer extracted from the date.
- month: Month as integer extracted from the date.

- day: Day as integer extracted from the date.
- traffic_volume: Numeric Hourly I-94 ATR 301 reported westbound traffic volume (target variable).

Reference

The original dataset can be found in the following URL: https://archive.ics.uci.edu/ml/datasets/Metro+Interstate+Traffic+Volume.

Note

The dataset used in the project is a re-sampling of the original dataset (i.e., it is a modified version of the original dataset). The data collection is the result of manual collection, so it is necessary to pay attention to possible errors, missing values, etc. For the regression task, we are interested in predicting the **traffic_volume** column.

Important: The dataset has an in-built order. However, the model must be able to estimate the traffic volume for the years following the training data. Therefore, pay attention to the input features you use. If you want, you can also tackle the problem as a time series.