

COMP47750

Machine Learning with Python

Assignment 2

Feature Engineering

Objective

The objective of this assignment is to assess the impact of feature engineering on a traffic volume prediction task. The data is available in the file `metro_traffic_15_19.csv`. This file is an extract from the UCI dataset here <[link](#)>.

The idea is to generate effective features for predicting traffic volume from data available a day in advance. You can assume that reasonable forecasts for the weather features are available a day in advance.

Task 1

1. Load the dataset into a dataframe that can be used for predicting `traffic_volume` a day in advance.
2. Produce some plots at different time-scales to see if there is periodicity in the traffic volume.

Task 2

1. Extract hour, day and month features from the time-stamps.
2. Divide the data into train and test sets keeping one third of the data for testing.
3. Build two different regression models and test the accuracy. Try Linear Regression and one other regression model from `scikit learn`.

Task 3

1. Given that the linear numeric encoding of the hour, day and month features may miss cyclical signals, investigate and test a cyclical strategy for encoding these features. Does this strategy improve accuracy for the models tested in Task 2?

Task 4

1. Identify effective subsets of the features for this prediction task. These can be the same subset for all models or model-specific subsets.

Submission: This is an individual (not group) project. Submission is through the Brightspace page. Your submission should comprise your notebook. You should use markdown cells in the notebook to report your findings and conclusions.