Capstone Project Proposal

Machine Learning Engineer Nanodegree

Shih-Wen, Wu

February 12, 2021

Domain Background

Taipei Mass Rapid Transit (MRT), branded as Taipei Metro, is a metro system in Taipei City, Taiwan. In 2020, there are 131 stations in service, serving area includes Taipei City and New Taipei City. The daily trip is around 2.16 million in 2019, and 2.01 million in 2020[1].

In order to provide a decent service and maintain the system with scale as such, predicting the hourly traffic amount correctly will be a key to success. It is also important for urban planning and policy making if there is a reliable traffic forecast generated on a regular basis.

Since commuters are normally accessing stations by walking, adverse weather condition is generally considered as a drawback for public transits[2]. (Syeed Anta Kashfi et al.). Xiaoyuan Wang et al. also suggest that urban residents tend to change traveling choice according to weather condition[3].

In this project, Taipei MRT hourly traffic data and weather data will be used to create a forecast model.

Problem Statement

The goal is to predict the hourly traffic by given weather conditions such as temperature, precipitation etc.

Datasets and Inputs

There are two dataset in this project:

Taipei MRT Hourly Data

Released by Taipei City Council on monthly basis in csv format. The dataset include 5 columns, namely date, time, orient station, destination station and the number of traffic. The dataset used will begin at January 2020 and end at December 2020. (data source link: http://163.29.157.32:8080/fi/dataset/98d67c29-464a-4003-9f78-b1cbb89bff59)

Weather data

Released by Central Weather Bureau, the dataset include temperature, air pressure, humidity, wind speed and precipitation. It is also collected hourly. In correspondence to traffic data, the time range will also between January 2020 and December 2020. (data source link: https://e-service.cwb.gov.tw/HistoryDataQuery/)

Solution Statement

In this project, support vector regression (SVR) will be applied to predict the MRT traffic. It is a supervised learning algorithm that is used to predict discrete values. Compared to simple regression which try to minimize error, SVR allows us to fit the error within a certain threshold.

Benchmark Model

Since time series forecast are widely used in this domain, time series model DeepAR will be used as a benchmark model in this project.

Evaluation Metrics

Root mean square error will be applied to evaluate the model.

Project Design

To increase the accuracy, the forecast model will be generated on station basis. Stations with high traffic will be selected, such as Taipei Main Station and Taipei City Council. Next, the weather data of the nearest weather station of the selected MRT station will be collected.

The training and test set will be split based on weather condition. The test set will randomly select 3 days on each month, total 36 days. (3 days* 12 months), and the rest will be used as training set. The result will be compared with benchmark model using root mean square error.
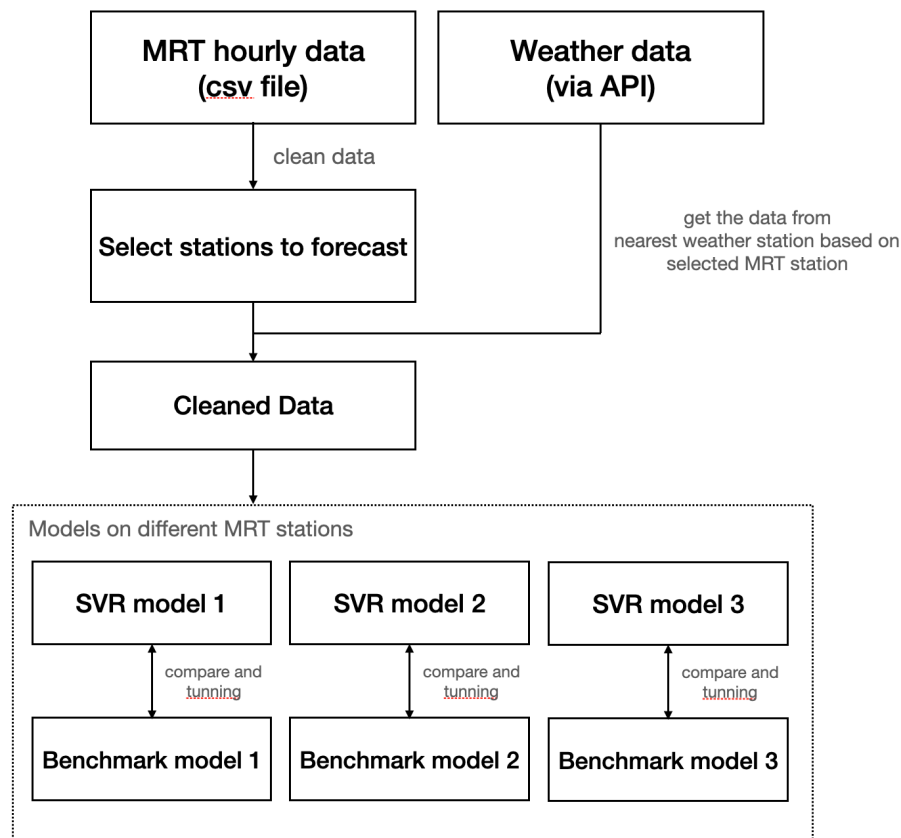
Project structure

Reference:

[1] Central News Agency 2020, <https://www.cna.com.tw/project/20200416-metro/page1.html>

[2] Syeed Anta Kashfi , Jonathan M. Bunker, Tan Yigitcanlar "Modelling and analysing effects of complex seasonality and weather on an area's daily transit ridership rate" *Journal of Transport Geography* 54 (2016)

[3] Xiaoyuan Wang et al."The Effects of Weather on Passenger Flow of Urban Rail Transit" *Civil Engineering Journal* Vol.6 No.1 (2020)