Capstone Project

Machine Learning Engineer Nanodegree

Shih-Wen, Wu

April 2, 2021

Definition

Project Overview

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].

Taipei MRT Map

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.

Evaluation Metrics

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

Analysis

Data Exploration

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: https://data.gov.tw/dataset/128506)

The fields includes:

- date
- time
- entrance
- exit
- people
- code_entrance
- code_exit
- weekday
- datetime

Weather data

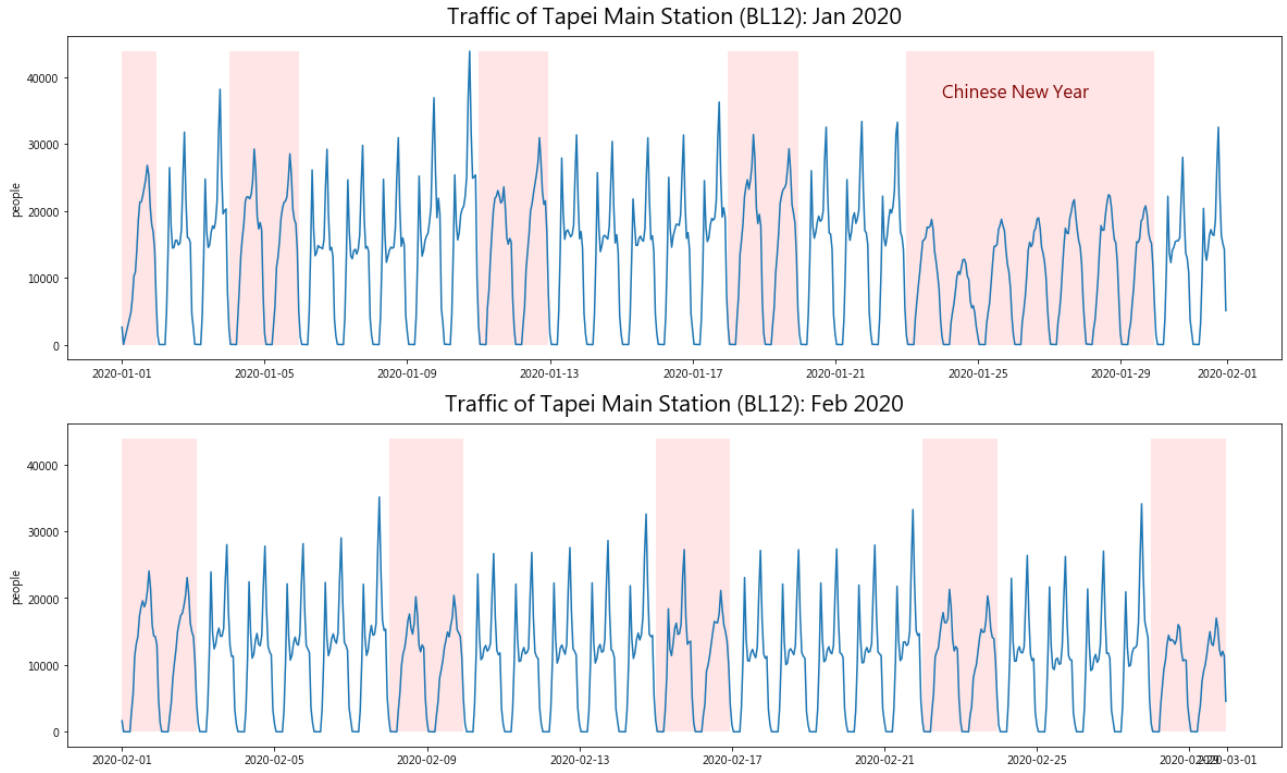
Released by Central Weather Bureau, the dataset include temperature, air pressure, humidity, wind speed and precipitation. It is also collected hourly. In correspondence to traffic data, the time range will also between January 2020 and December 2020. (data source link: https://e-service.cwb.gov.tw/HistoryDataQuery/)

The fields includes:

- air_presr
- temp_°C
- humidity
- wind_spd_m_s
- precipitation__mm_t

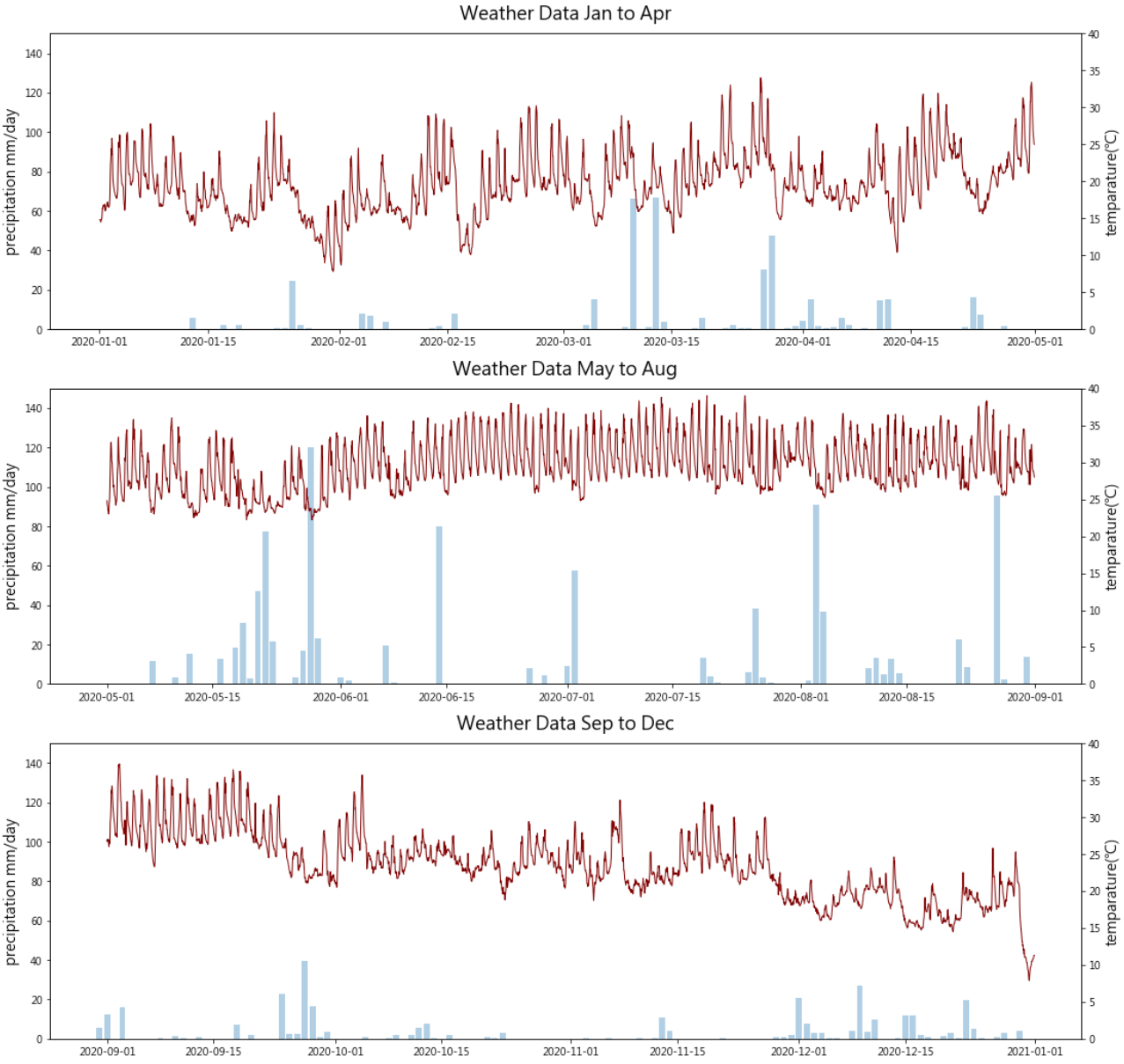
Exploratory Visualization

The plot below illustrates the traffic of Taipei Main Station (BL12) on January 2020 and February 2020. The areas with white background are weekday traffic while the red areas are weekends or holidays, in this case, Chinese New Year. It is clear to see that the traffic patterns are different between weekdays and weekends.



Traffic Data of Taipei Main Station(BL12)

The plot below shows the temperature in line red line, and precipitation in bars from Taipei observation-station in 2020.



Weather Data of Taipei observation-station

Algorithms and Techniques

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.

In general, SVR is similar to support vector machine (SVM) with some notable difference:

- ε (epsilon): The value of epsilon determines the width of the tube around the estimated function. Points fall inside this area are considered as correct predictions and are not penalized by the algorithm.
- The support vectors are the points that fall outside the tube rather than just the ones at the margin as seen in SVM.

- Slack (ξ) measures the distance to points outside the tube, by setting the parameter C, you can decide how much you care about them.

Support vector regression algorithm is regarded as a huge improvement over simple linear regression. It allows users to build non-linear models and gives them control over the flexibility vs. robustness of your models. (https://towardsdatascience.com/support-vector-regression-svr-one-of-the-most-flexible-yet-robust-prediction-algorithms-4d25fbdaca60)

Benchmark

Since time series forecast are widely used in this domain, time series model DeepAR will be used as a benchmark model in this project. By running DeepAR for traffic of Taipei Main Station, the RMSE is 1423.35, and Taipei City Hall is 1391.21.

Methodology

Data Preporcessing

Taipei MRT Hourly Data

The raw data is an oriented-destination traffic data includes all MRT station during operation hours (from 06:00 to 24:00). Each single file contains data for a month that consist of more than 7 million rows. The first few rows as shown below:

日期	時段	進站	出站	人次
2020-01-01	00	松山機場	松山機場	1
2020-01-01	00	松山機場	中山國中	0
2020-01-01	00	松山機場	南京復興	0
2020-01-01	00	松山機場	忠孝復興	0
2020-01-01	00	松山機場	大安	0
2020-01-01	00	松山機場	科技大樓	0
2020-01-01	00	松山機場	六張犁	0
2020-01-01	00	松山機場	麟光	0
2020-01-01	00	松山機場	辛亥	0
2020-01-01	00	松山機場	萬芳醫院	0
2020-01-01	00	松山機場	萬芳社區	0

Raw dataset of MRT station

The first 2 columns are date and hour, 3rd and 4th are entrance and exit station recorded in Chinese, and last column is number of people. The preprocessing for this dataset includes:

1. Clean up the redundant content such as the dash line between headers and data.
2. Extract the station needed in the project which are Taipei Main Station and Taipei City Hall.
3. The purpose of this project is to predict the total traffic of each station, so we need to sum of people both entering and exiting the study station in the data.
4. Keep date and hour column which are keys to join with weather data.

The result are shown below:

	date	time	people
1	2020-01-01	0	2607
2	2020-01-01	1	17
3	2020-01-01	6	4799
4	2020-01-01	7	6957
5	2020-01-01	8	10221
6	2020-01-01	9	10837
7	2020-01-01	10	14249
8	2020-01-01	11	18653
9	2020-01-01	12	21292
10	2020-01-01	13	21334

Processed dataset of MRT station

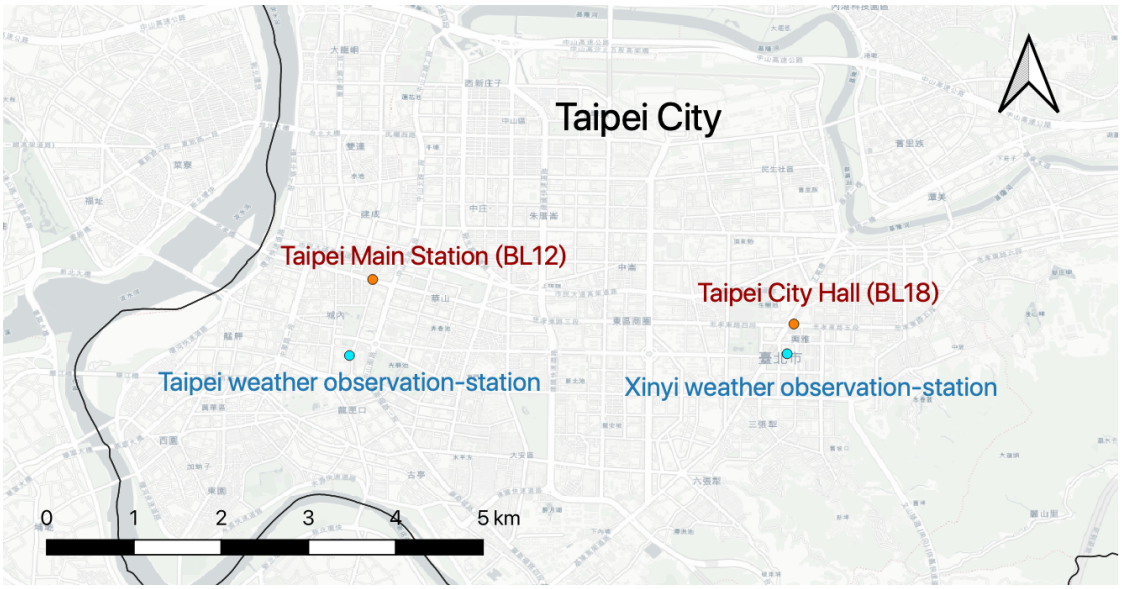
Weather data

The dataset is from Central Weather Bureau, the original dataset includes 20 columns. Depends on the type of weather observation-stations, some of them have data for all 20 columns, while some of them are not. In this project, 5 type of weather data are extracted, namely air pressure(hPa), temperature(Celcius), humidity, wind speed(meters per second), precipitation(mm per hour). Below illustrates the dataset after processing:

		air_presr	temp_°C	humidity	wind_spd_m_s	precipitation__mm_t
datetime	**station_ID**					
2020-01-02 01:00:00	**C0AC70**	0.749296	0.285266	0.802632	0.229885	0.0
2020-01-02 02:00:00	**C0AC70**	0.726761	0.285266	0.789474	0.195402	0.0
2020-01-02 03:00:00	**C0AC70**	0.704225	0.282132	0.789474	0.241379	0.0
2020-01-02 04:00:00	**C0AC70**	0.695775	0.272727	0.789474	0.195402	0.0
2020-01-02 05:00:00	**C0AC70**	0.712676	0.272727	0.802632	0.218391	0.0

Processed dataset of weather station

Since there are many weather observation-station around Taipei City, 2 observation-stations are selected accordingly for the 2 MRT Stations in this project, which are Taipei weather observation-station for MRT Taipei Main Station, and Xinyi weather observation-station. The map below shows the locations of each stations:



Locations of MRT and weather stations

Implementation

The implementation process are shown below:

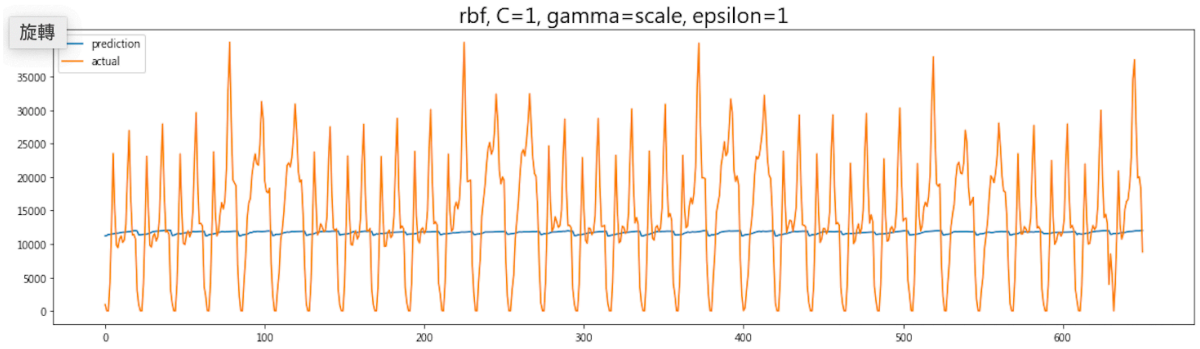
1. Join prepared weather data and traffic data
2. Train data will be the data with date between 2020-01-01 and 2020-11-30, and the rest will be test data (2020-12-01 to 2020-12-31).
3. Input parameters for the model include 5 from weather data, namely air pressure(hPa), temperature(Celcius), humidity, wind speed(meters per second), precipitation(mm per hour), and 1 from traffic data, time (hour).

datetime	air_presr	temp_°C	humidity	wind_spd_m_s	precipitation__mm_t	time	people
2020-01-01 01:00:00	1024.9	14.8	82	3.6	0.0	1	17
2020-01-01 06:00:00	1024.4	14.9	87	4.1	0.0	6	4799
2020-01-01 07:00:00	1025.2	14.9	90	3.1	0.0	7	6957
2020-01-01 08:00:00	1025.6	15.4	89	3.5	0.0	8	10221
2020-01-01 09:00:00	1026.0	16.4	82	4.5	0.0	9	10837

Joined dataset

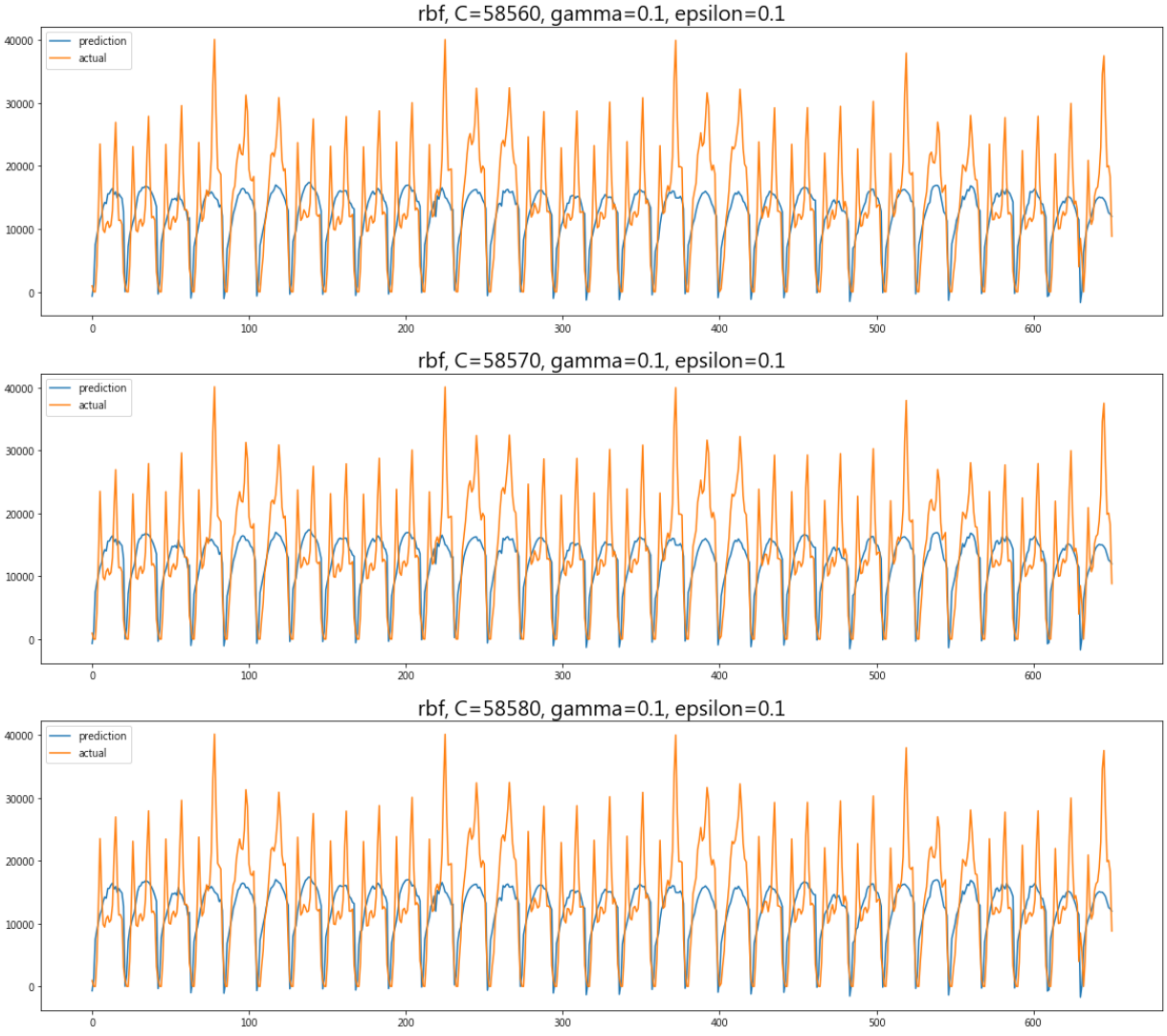
Refinement

First run the model with dataset of Taipei Main Station, with default parameter: C=1, gamma='scale' and epsilon=0.1. The result RMSE is 8591.97 as shown below.

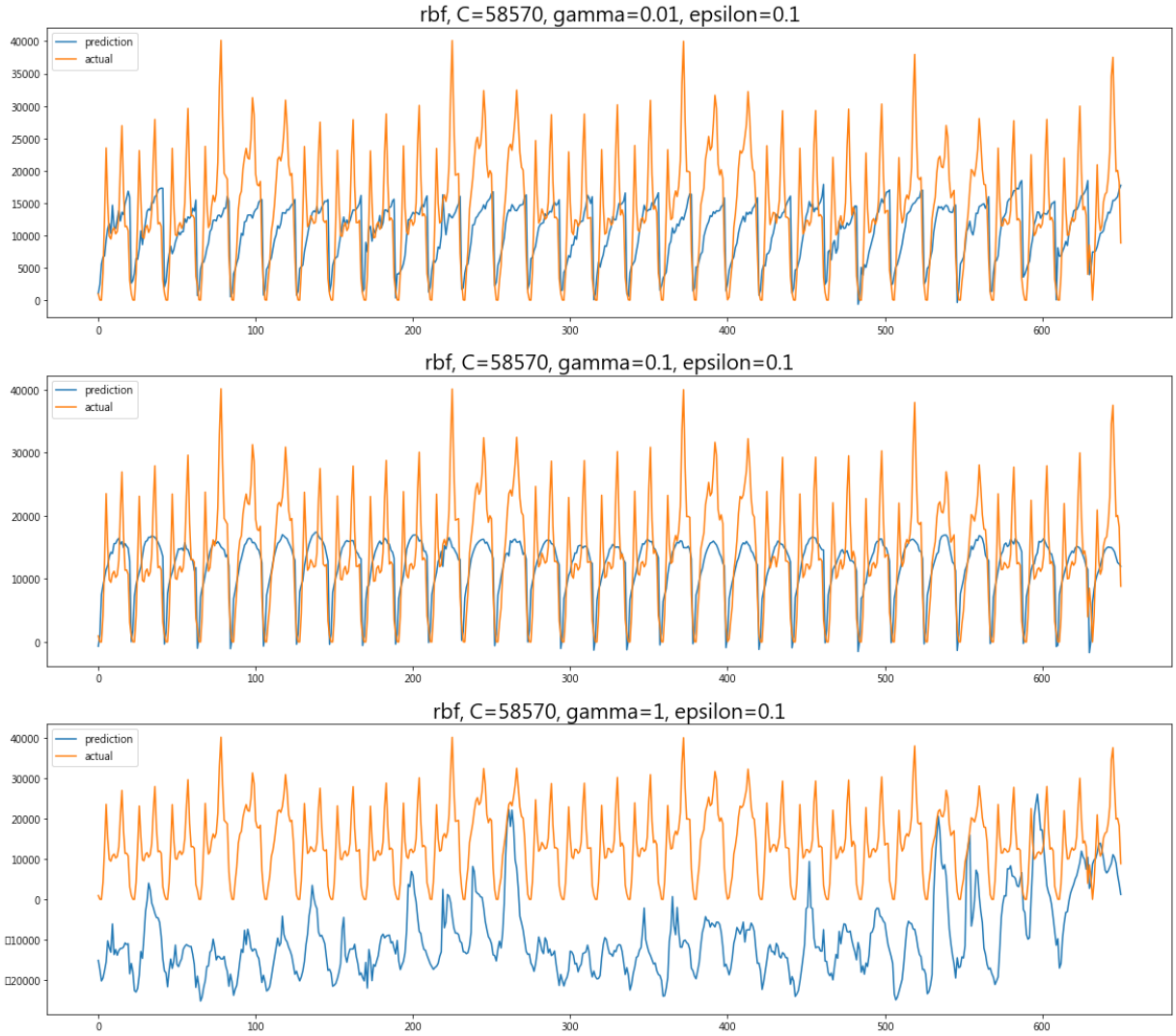


In order to improve the model, 3 parameters for SVR in sk-learn are adjusted, which are C for regularization parameters, epsilon and gamma. The kernel used is rbf (Radial basis function kernel).

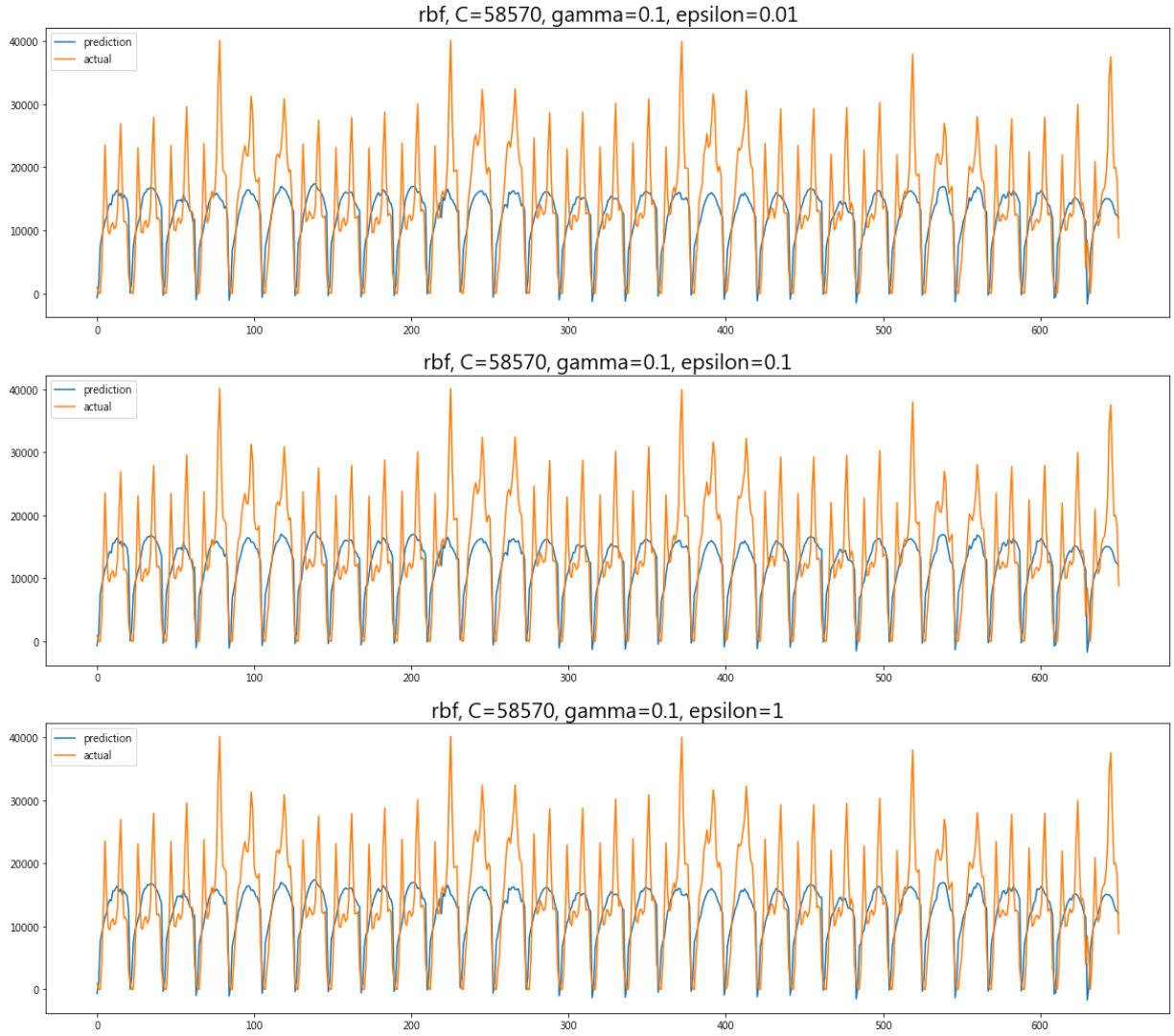
The following process is adjusting C to 58,560, 58,570 or 58,580 while keeping other parameters at 0.1. The RMSEs respectively are 6,437.081, 6,437.073 and 6,437.077, where RMSE is the lowest while C is equal to 58,570.



Next, adjusting gamma input, the graph below indicates the results while gamma are set to 0.01, 0.1 and 1 respectively, where the lowest RMSE appears at 6,437.073 while gamma is equal to 0.1:

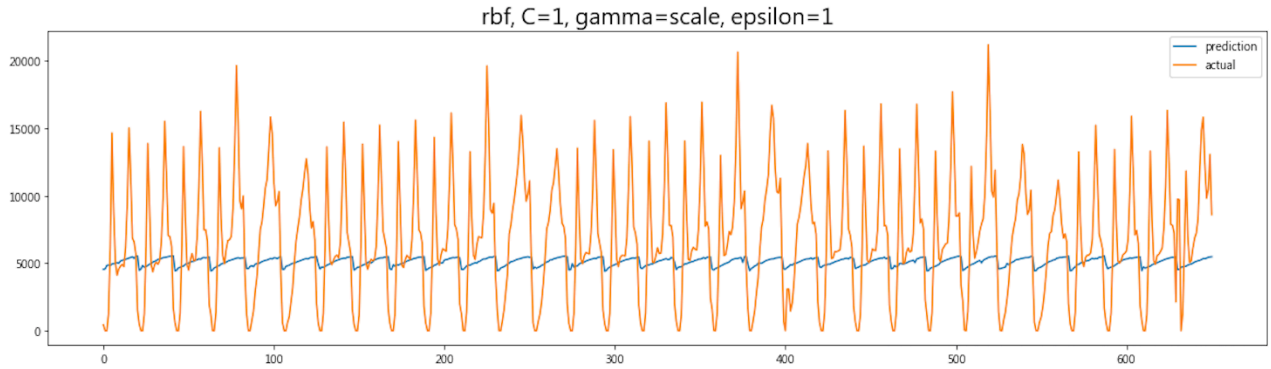


Lastly, while keeping C equals to 58,570 and gamma equals to 0.1. Setting epsilon equals to 0.01, RMSE is 6,437.11, when setting it to 0.1, RMSE is 6,437.07, when setting it to 1, RMSE is 6,437.20. The difference is insignificant as the graph shows below:

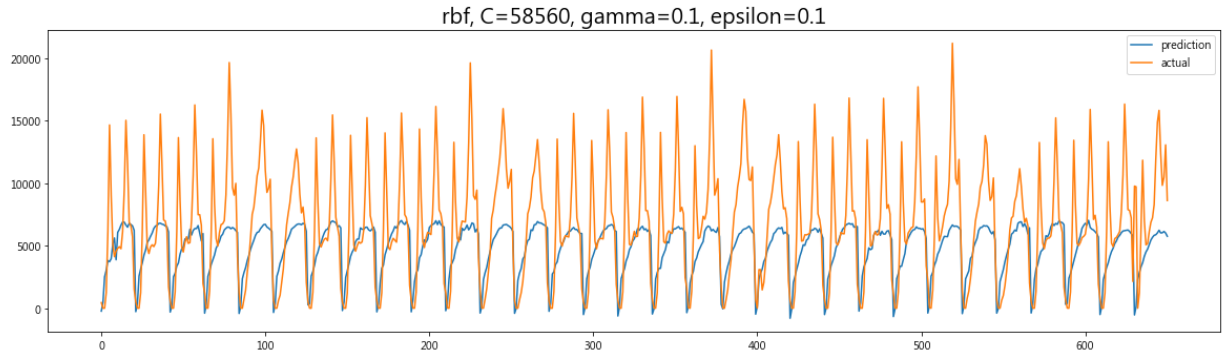


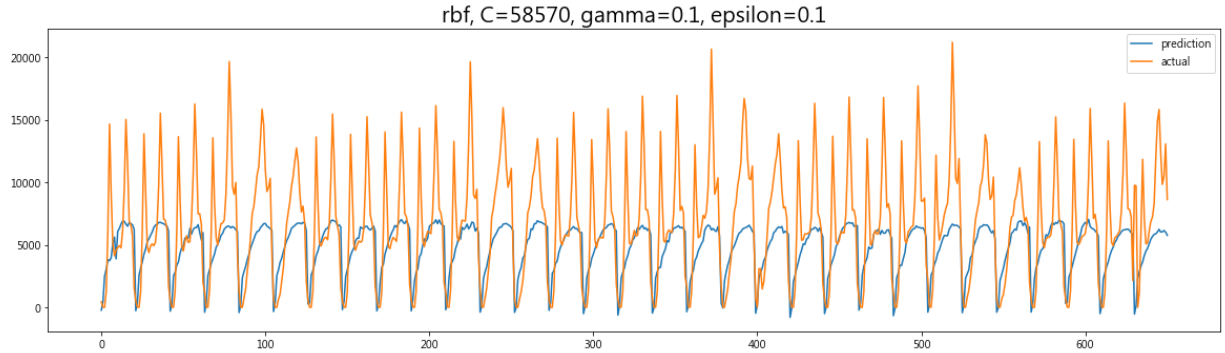
After adjusting the 3 parameters, the RMSE improves from 8,591.97 to 6,437.07, a decrease around 25%.

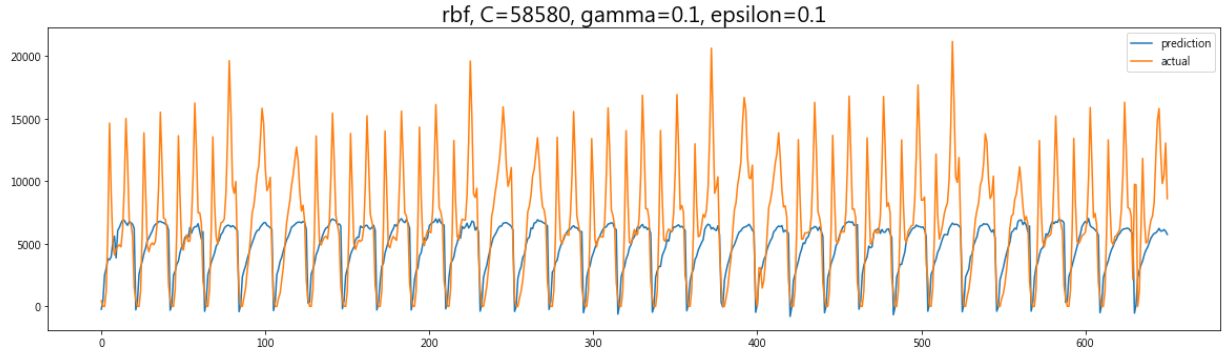
For dataset from Taipei City Hall, the RMSE from default parameters is 4,545.14:



When setting the parameters C to 58,570, gamma and epsilon to 0.1, the RMSE drops to 3,949.29:







Result

Model Evaluation and Validation

The final parameters of the model were chosen based on the RMSE, which are :

- C = 58570
- gamme = 0.1
- epsilon = 0.1

For dataset of Taipei Main Station, the RMSE drops from 8,591.97 to 6,437.07, around 25% decrease. For dataset of Taipei City Hall, the RMSE drops from 4545.14 to 3949.29.

Justification

Compared with benchmark model, which is the result generated by time series forecast model DeepAR, from AWS, there is a significant difference between their RMSEs.

For the result of Taipei Main Station, the RMSE from DeepAR is 1423.35 versus 6437.07 from SVR model, while for the result of Taipei City Hall, the RMSE from DeepAR is 1391.21 versus 3949.29 from SVR model.

Conclusion

Reflection

The process used for this project can be summarized into following steps:

1. Collecting dataset from 2 different source, namely Central Weather Bureau and Taipei MRT Station.
2. Data preprocessed
3. Create benchmark using DeepAR
4. Create SVR model with sk-learn
5. Testing the parameters

The first challenge I encountered was when collecting the weather data, the api was not available yet, so I have to write a scraper to get the data I needed. I studied a number of tutorial website teaching how to write a python scrape so that I can get the data I need.

The next challenge was to use the DeepAR with data I prepare. Thanks to the tutorial lesson from Udacity and the documentation provided, this make this step a lot easier.

The most interesting part in the project is when exploring such wonderful traffic data and bring it into a model, it is interesting to find out that the result appear a lot different than what have expected. I believe in the future the dataset can be used in more different area.

Improvement

As shown in the justification section, there is still a significant difference of RMSE between SVR model and benchmark model. By check the weather data, possible reasons are listed below:

- There only few raining days during the period of dataset (January to November, 2020), which make the precipitation feature less useful, even thought there are studies show that commuters are affected by the weather.
- For daily commuters, they have to go to work by public transit no matter how bad the weather is.
- On the other hand, casual commuters are the people who affected by the weather.

In order to improve the model, more effective features need to be explored. The performance of model with only weather data is not accurate enough

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)