Report of HW1 - Dengue Case Prediction

Student ID: 109062211 Name: 張惇媛

I. Regression Equation in Basic Part

For CityA, $CaseNumber = (0.4230755630621288) * Temperature + (0.6054281404463512) * CaseofPrevWeek + (0.11606595999479018)$

For CityB, $CaseNumber = (0.7686972068305014) * Temperature + (0.21647129531893558) * CaseofPrevweek + (0.24849235608027498)$

For CityC, $CaseNumber = (0.6105556184517776) * Temperature + (0.514931820647193) * CaseofPrevweek + (0.6892163871029104)$

II. Variables in Advanced Part

1. Average Case Number of Past 4 Weeks

- Difference between Basic Part and Advanced Part:
  I only used the average case number of past four weeks as input of regression model in Advanced Part because of the reasons below:
  - The correlation between Temperature and Case number is low in the scatter plot of Temperature and Case number. As a result, I drop the columns of Temperature which used in Basic Part.
  - The case number is time series data. I should make use of more information in the past to forecast the future outcome more precisely. Nevertheless, I only used the data of previous week in Basic Part. Therefore, I change the input to the average case number of past four weeks in Advanced Part.

III. The Difficulty I Encountered

1. Not very familiar with Python
   Since I am not very familiar with Python and the usage of some packages, like Scikit-Learn, NumPy, pandas and csv, I spend time on programming and handling errors.
2. The way of preprocessing the data
   Dealing with missing data and outlier is important because the noise may interfere with the optimization of regression model. But, how to determine whether to pick the re-filled values or to drop the data is difficult.
3. The selection for parameters
   The detection of the outlier is important because the noise interfere with the regression model. Also, poor learning rate may let the cost function go far from the minimum and become infinity number. As a result, the selection is important. Testing for the ideal values is required.

IV. Summary: The Solution of the Hardness and Reflections

From starting do the assignment from scratch, I faced lots of challenges, which sometimes make me felt frustrated. However, referring to the guidelines of functions and the example code of regression, I make use of the resource on the Internet and do some changes to meet my needs. In addition, I do several times testing and rolling adjustment, which try to find better selection for parameters. Also, I plot the picture and evaluate MAPE to verify my regression model.

Finally, I train the regression model in practice. Although the process is hard, I get a sense of achievement when the assignment is done. Besides, it is easier to write code in Advanced Part than in Basic Part because we can use the existing packages. So, we may spend time understanding some common function methods and fully make use of this to implement in the future works of this machine learning class.