New York City Traffic Violation Data Analysis and Prediction

310712009 楊家碩 GMBA12
109550162 郭子頡 資工13
310551001 林奕宏 資科12
0716302 黃靖雅 資工11

Table of Contents

目標問題 / Target Problem

Motivation

鑒於台灣道路交通普遍混亂與違規情形(e.g.路邊違停)氾濫，期望藉由分析公開且齊全的紐約交通違規資料與相應的天氣資料，來評估預測未來的違規情形的可行性，以提供有關政府/民間單位進一步的應用參考(e.g.警力人力分配、即時導航的指引參考/提醒、……等)

Problem statement

分析各城區時間性(每小時)的違規資料與當日天氣資料的統計數據，以預測未來的違規數量/罰款金額

Target performance metrics

預測結果達到77%R squared

Datasets

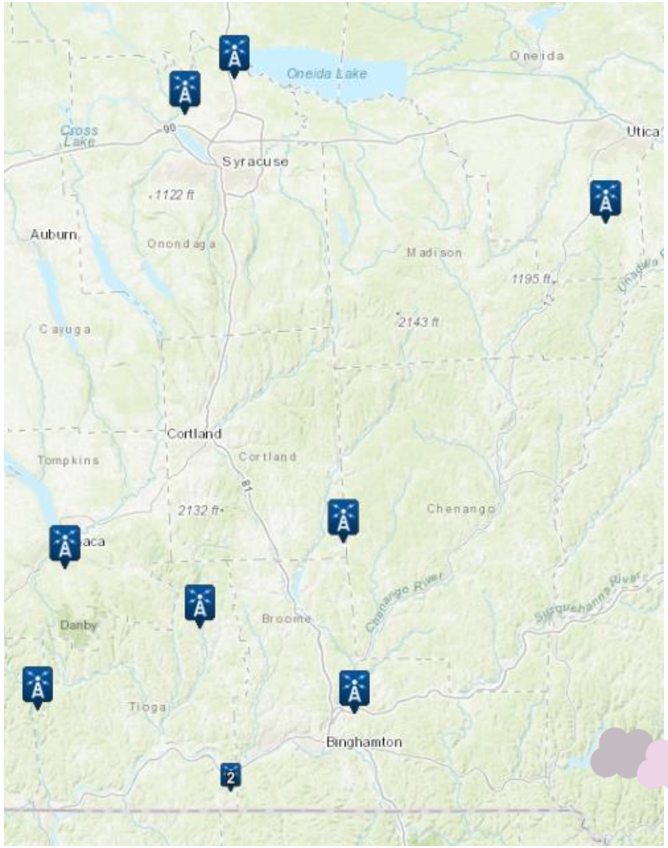
- Data sources and characteristics
- 5Vs (Volume, Variety, Velocity, Value, Veracity) of Big Data?

天氣資料的來源與蒐集

- **National Oceanic and Atmospheric Administration**
    - National Centers for Environmental Information

- **Collected Data**
    - TAVG(Avg temp)
    - AWND(Avg wind speed)
    - SNWD(Snow depth)
    - WT01 (Fog)

- **Link**
    - https://www.ncdc.noaa.gov/cdo-web/



天氣-資料特性 Data characteristics / 4V

Velocity

- Update the data daily

Variety

- Data: Row 960k Rows and 72 Columns (67 features)
- 3 種不同種類的資料
- 資料型態包含地理資訊, 數值型態, 文字等資料

天氣-資料特性 Data characteristics / 4V

Value

- 這份資料可以用來判斷當日大致的天氣類型 (e.g.是否有降雨/下雪、霧氣影響能見度...)

Veracity

- Our data is from NCDC (government's administration). It can be defined as the accuracy or truthfulness of a data set.

紐約交通違規資料的來源與蒐集

- **NYC Open Data**
  - Department of Finance (DOF)
  - Open Parking and Camera Violations

- **Collected Data**
  - Issue Date
  - Violation Type
  - Fine Amount
  - County
  - ….(15 other features)
- **Link**
  - [https://data.cityofnewyork.us/…](https://data.cityofnewyork.us/)

NYC OpenData

交通-資料特性 Data characteristics / 5V

Volume

- Time Period: 2019/1/1 ~ Now 2022/4/22
- Total Data Volume : around 20 GB

Velocity

- Update the data daily

Variety

- Data: Row 77.3Million and Columns 19
- 3 種不同種類的資料
- 資料型態包含地理資訊, 數值型態, 文字等資料

交通-資料特性 Data characteristics / 5V

Value

- 這份資料可以用來判斷交通罰單

Veracity

- Our data is from NYC Open Data. It can be defined as the accuracy or truthfulness of a data set.

Project Objectives

- Input/Output
- Group Division

專案目標
f(X) = Y

Output

Y1
Case Count

- 以每日每小時為單位計算違規案件總數
- Real-life application: 警力分派

Y2
Average fine amount

- 以每日每小時為單位計算平均的罰款金額
- Real-life application: 了解哪個時間段容易發生重大交通違規

專案目標

f(X) = Y

Input

X (V1)
Without Weather Data

- County, Month, Day, Weekday, Hour

X (V2)
With Weather Data

- X (V1) + Weather Data (Average temperature, average wind speed, snow depth, fog)

Experimental Group vs. Control Group

Control Group

Data 1
- Traffic Violations Data
  - One hot encoding for different counties
- No Weather Data

Experimental Group

Approach 1 → Data 2
- Traffic Violations Data
  - One hot encoding for different counties
- Add NOAA Weather Data

Approach 2 → Data 3
- Traffic Violations Data
  - Use one county at the time
- Add NOAA Weather Data

Basic Work Plan

- *Platform, Tools, and Analysis Workflow*
- *Schedule*

分析流程

Data Access → **Data Preprocessing** → **Modeling** → **Application & Result**

Data Access

- 取得歷年交通違規資料
- 取得歷年天氣資料

Data Preprocessing

- 處理缺失值(drop)
- One hot encoding on counties
- 資料轉換
- 合併違規、天氣資料

Modeling

- Feature Score
- 機器學習
  1. 模型建立
  2. 模型訓練及調整
  3. Metric: R^2 score

Application & Result

- 呈現出預測及分析結果
- Graphs, tables

R^2 score

R^2 score = 1 - u/v

$$u = \sum_{i=1}^{n} (y_{true_i} - y_{pred_i})^2$$

$$y_{mean} = \frac{1}{n}\sum_{i=1}^{n} y_{true_i}$$

$$v = \sum_{i=1}^{n} (y_{true_i} - y_{mean})^2$$

使用平台及工具

開發環境

Python程式語言、Google Colab

大數據平台

Pyspark

分析工具

Pandas、scikit-learn、PySpark MLlib
、Matplotlib

Data Preprocessing

- ***Challenges in PreProcessing - Weather***
- ***Challenges in Preprocessing - Violations***

Challenges in PreProcessing - Weather

- **Station attribute mapping**
  - Enormous amount of features
  - Enormous amount of missing value

- **Longitude/Latitude >> County**
  - Not exactly 1-1
  - Station may be near the boundary of multiple counties
  - Some county may not have any station at all

- **Q: How to handle missing data**
  - May result in different models
  - Leads to different results and accuracy
  - Be discussed later

Challenges in Preprocessing - Violations

- **“County” Column is very arbitrary**
  - e.g. “Manhattan” / “Brooklyn” should actually be "New York”
  - e.g. “Qns” / “Q” >> Queens
  - e.g. “B” >> “Bronx” or “Broom”?
  - e.g. “14” / “16” makes little sense

Manhattan、Brooklyn、NY	New York
Qns、Q	Queens
B、Bronx	Broom
KINGS、K	Kings
RICHM、Rich	Richmond

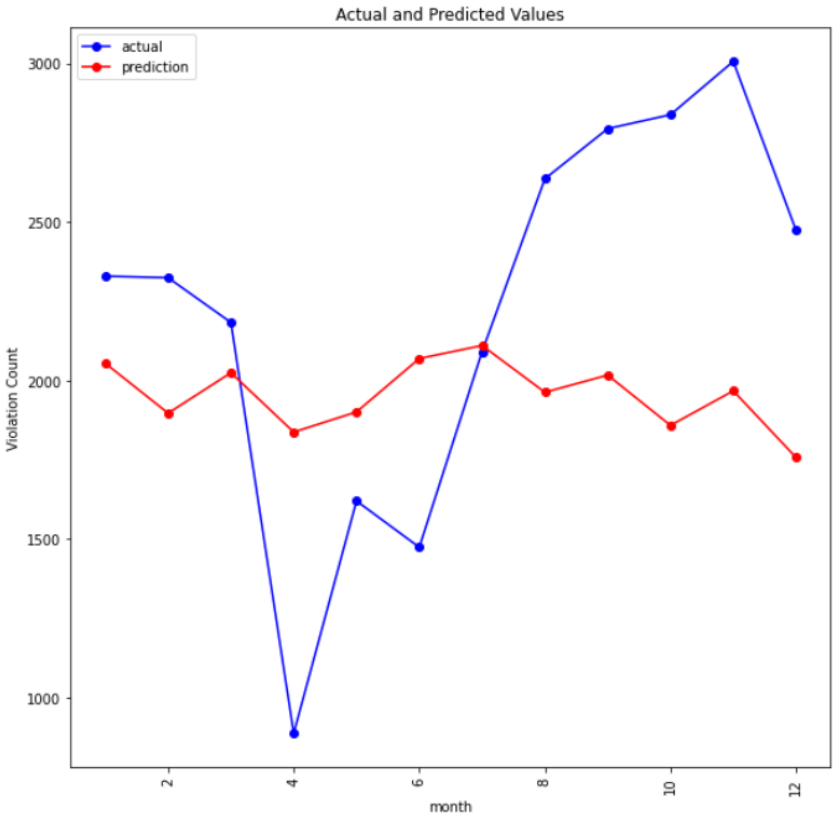
- **Solution**
  - Referencing related authorities’ County Code list
  - Matching manually

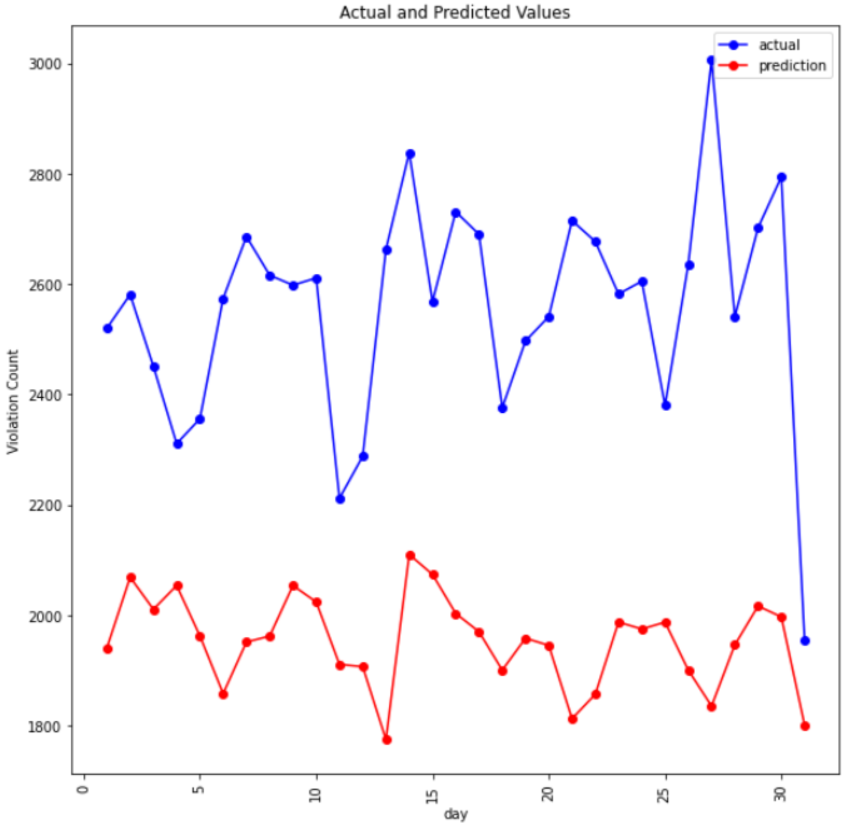
Models & Results

- Data Visualization
- Feature Score
- Result

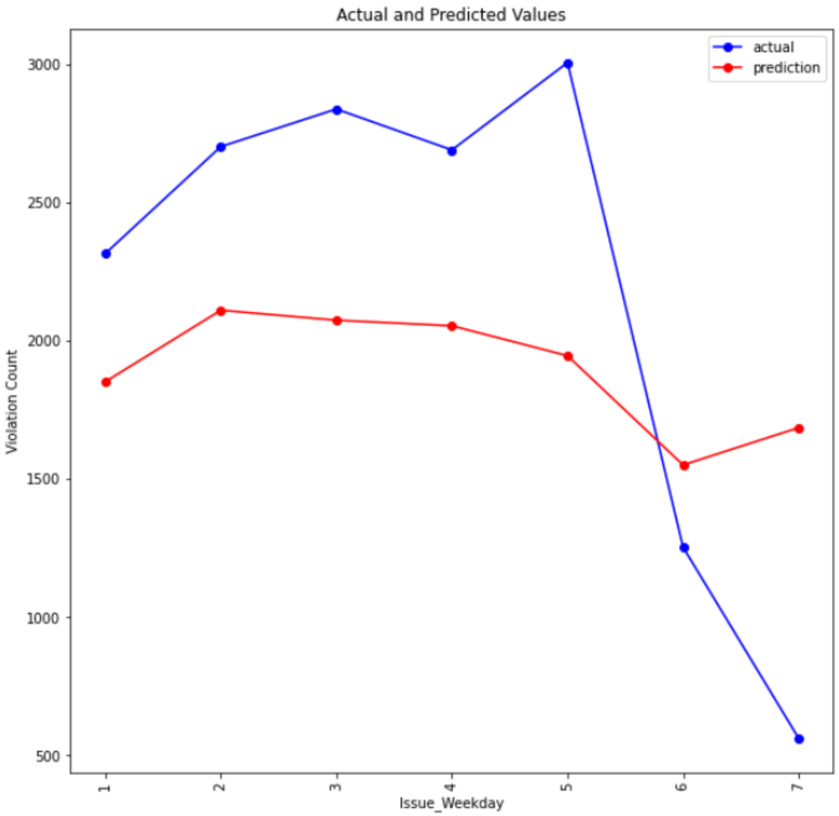
Data Visualization Prediction: Y1 (Count)

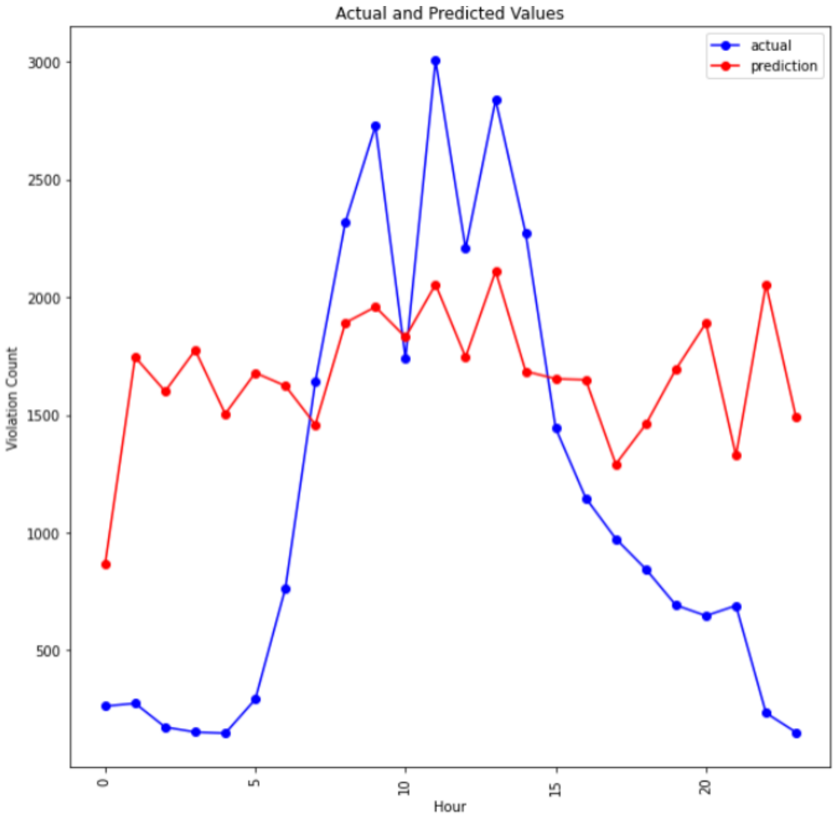
Input: Data 1 (No weather)
Prediction: Y1 (Count)



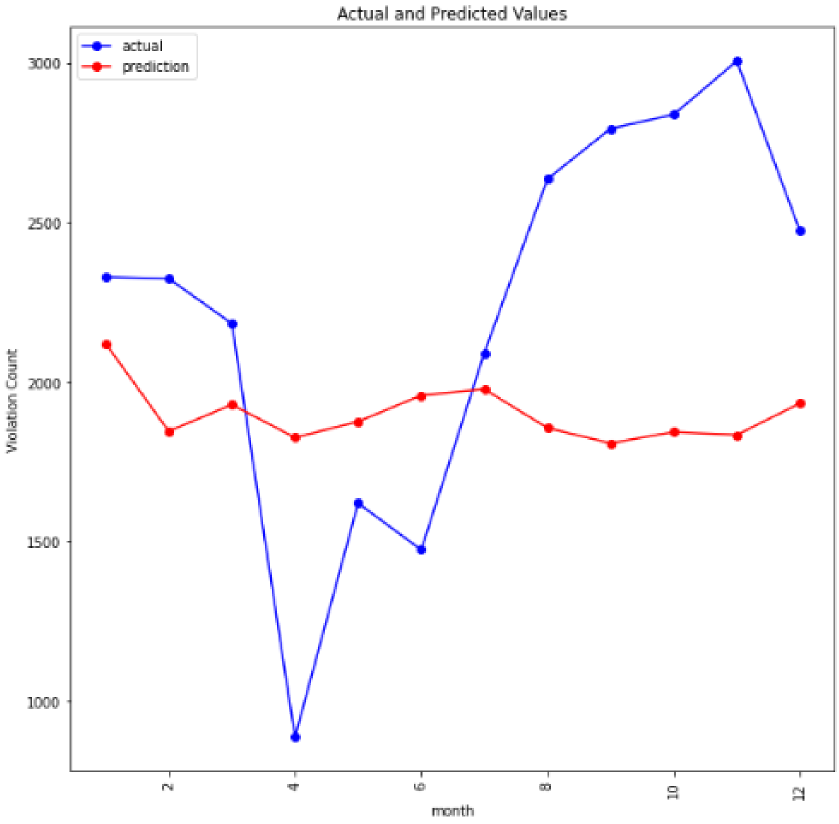


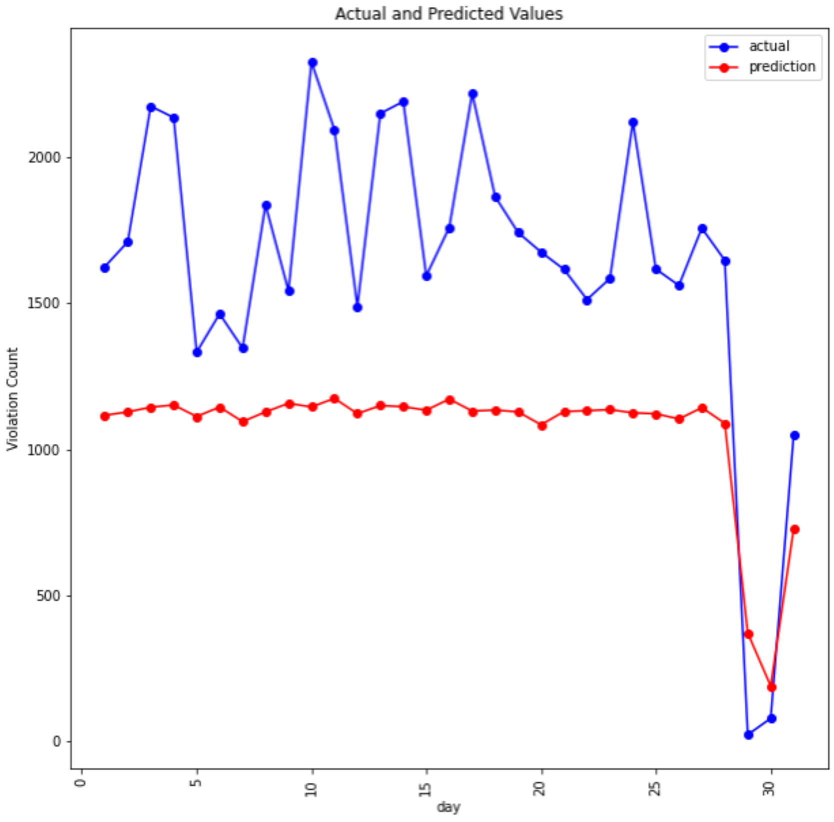
Input: Data 1 (No weather)
Prediction: Y1 (Count)



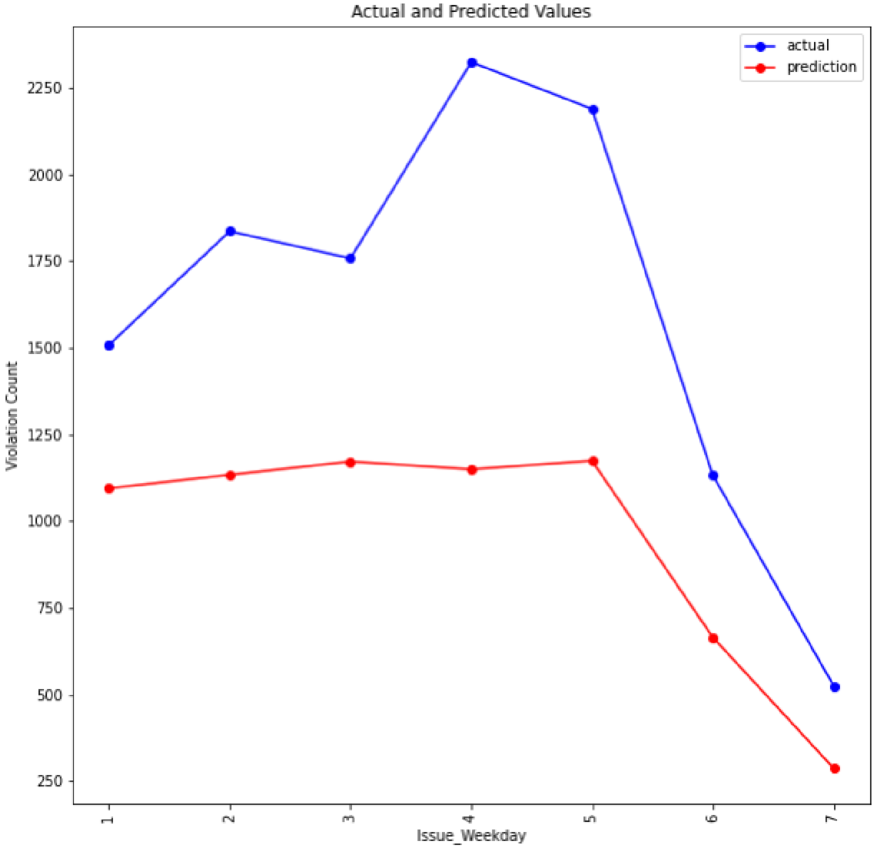


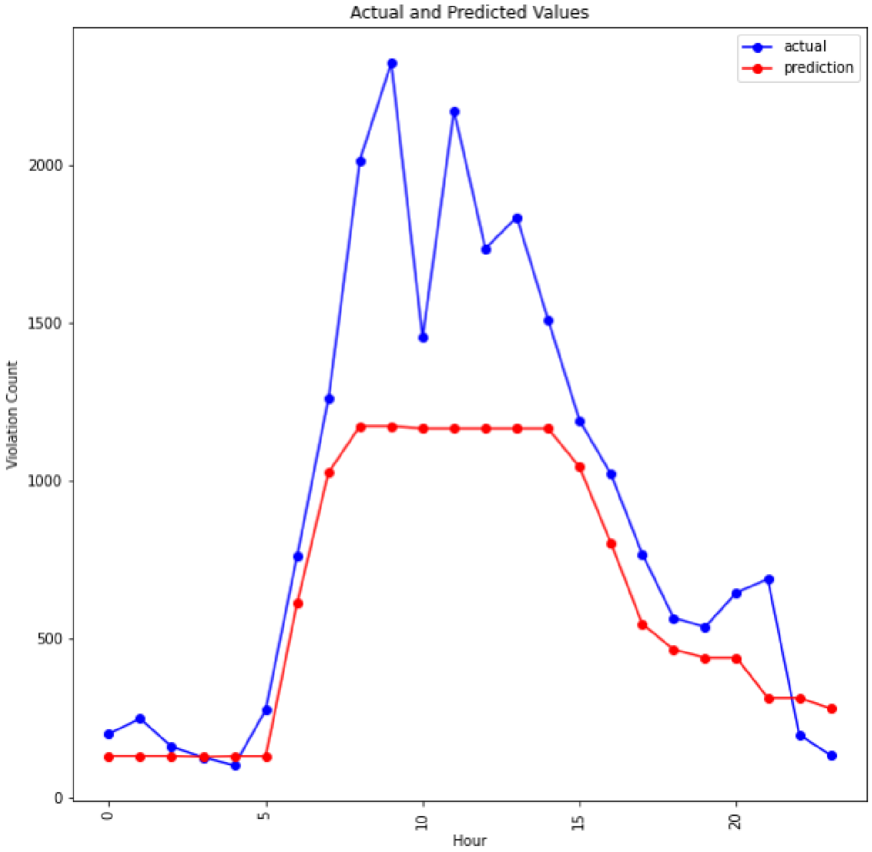
Input: Data 2 (Add NOAA Weather Data)
Prediction: Y1 (Count)



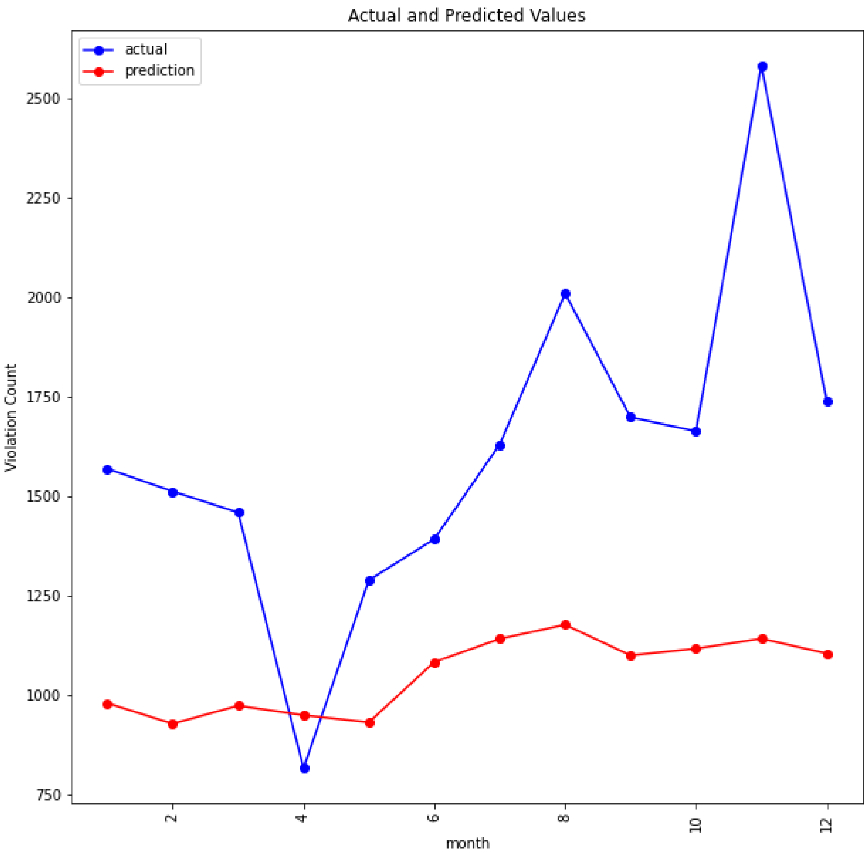


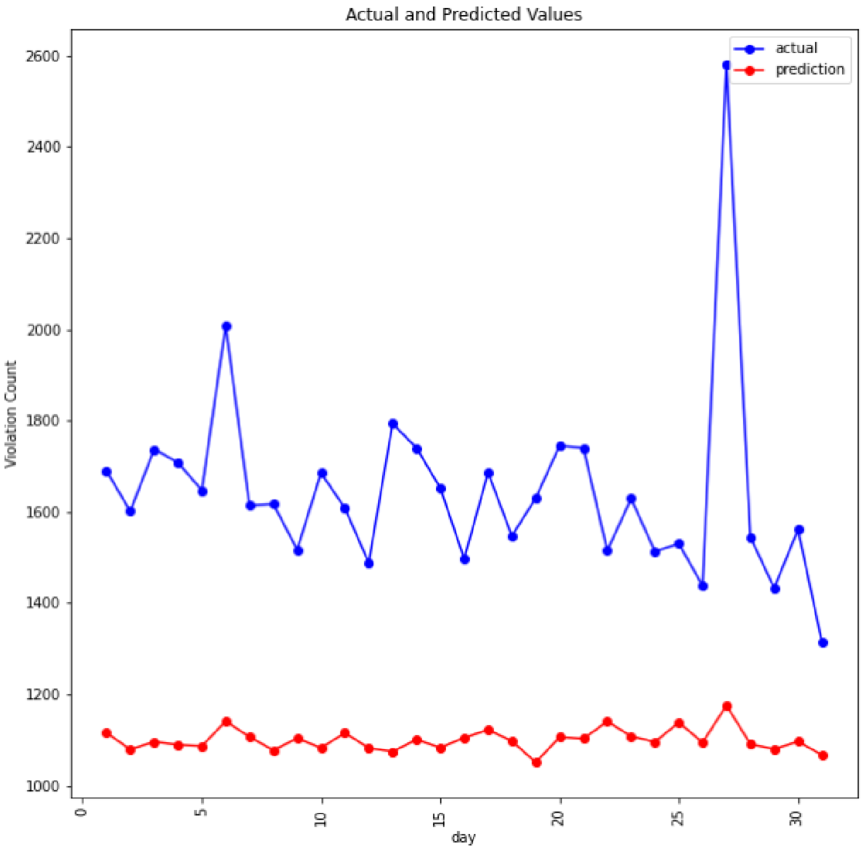
Input: Data 2 (Add NOAA Weather Data)
Prediction: Y1 (Count)



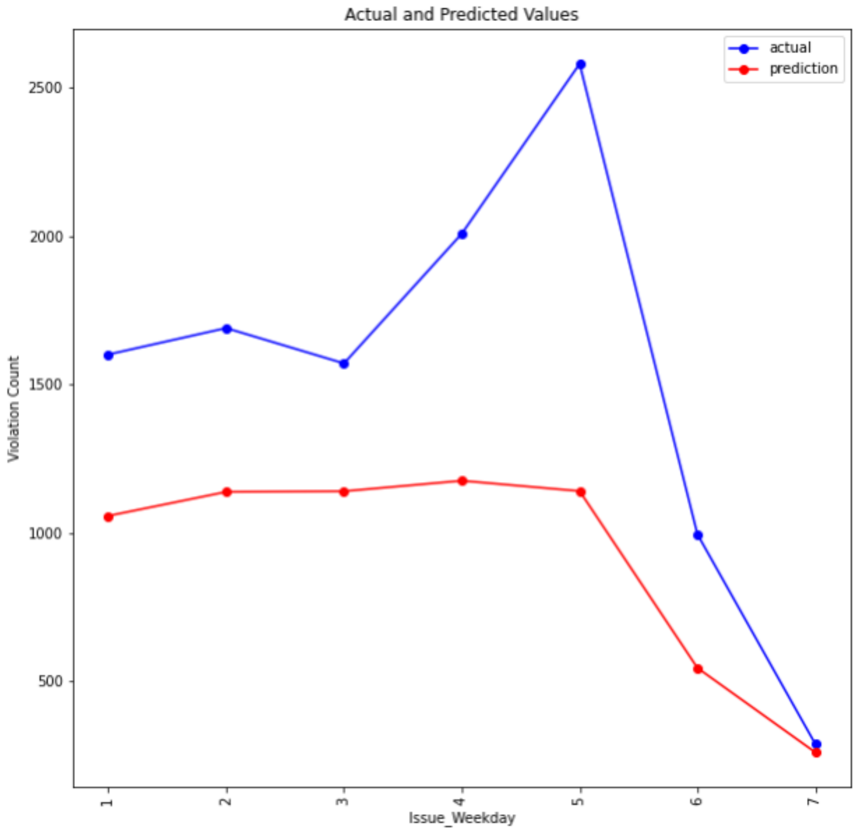


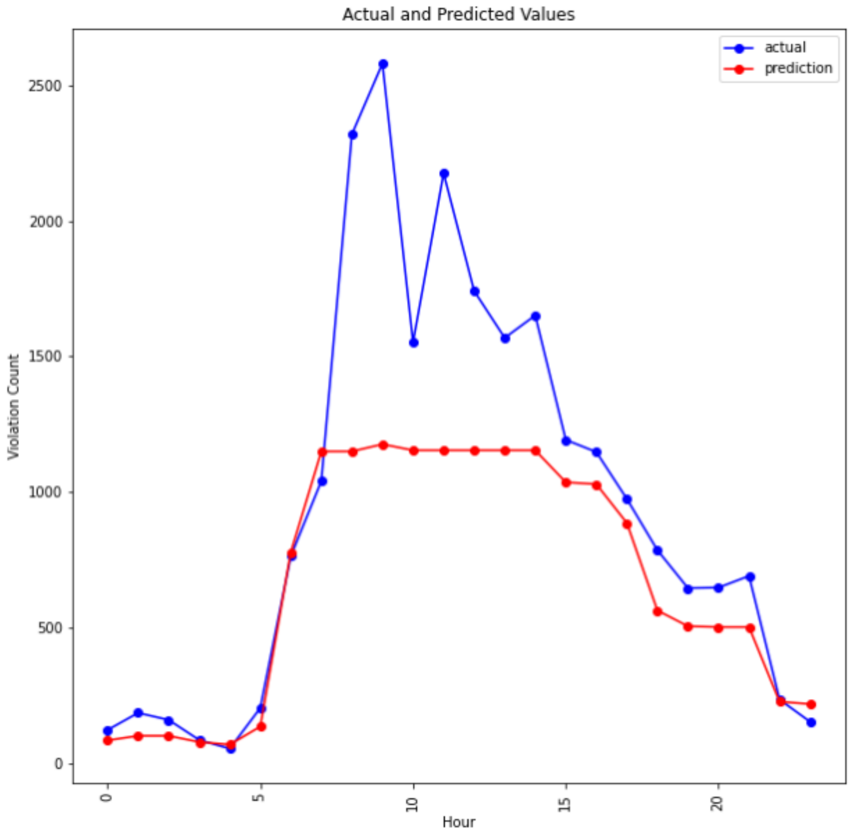
Input: Data 3 (Add NOAA Weather Data)
Prediction: Y1 (Count)



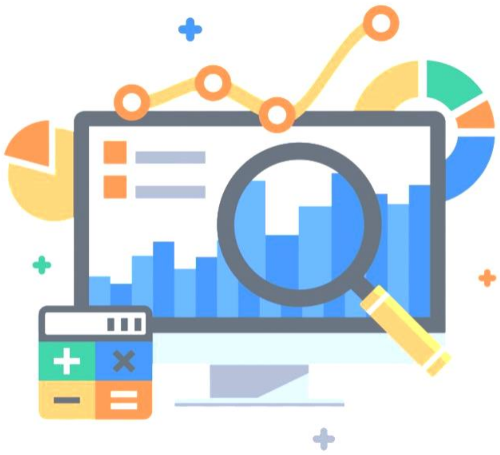


Input: Data 3 (Add NOAA Weather Data)
Prediction: Y1 (Count)

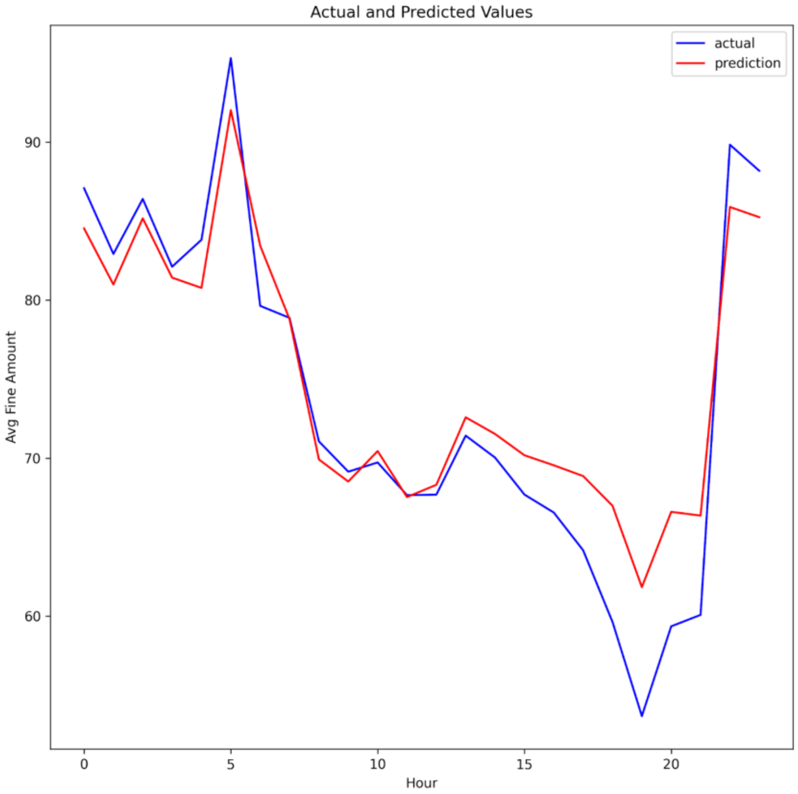


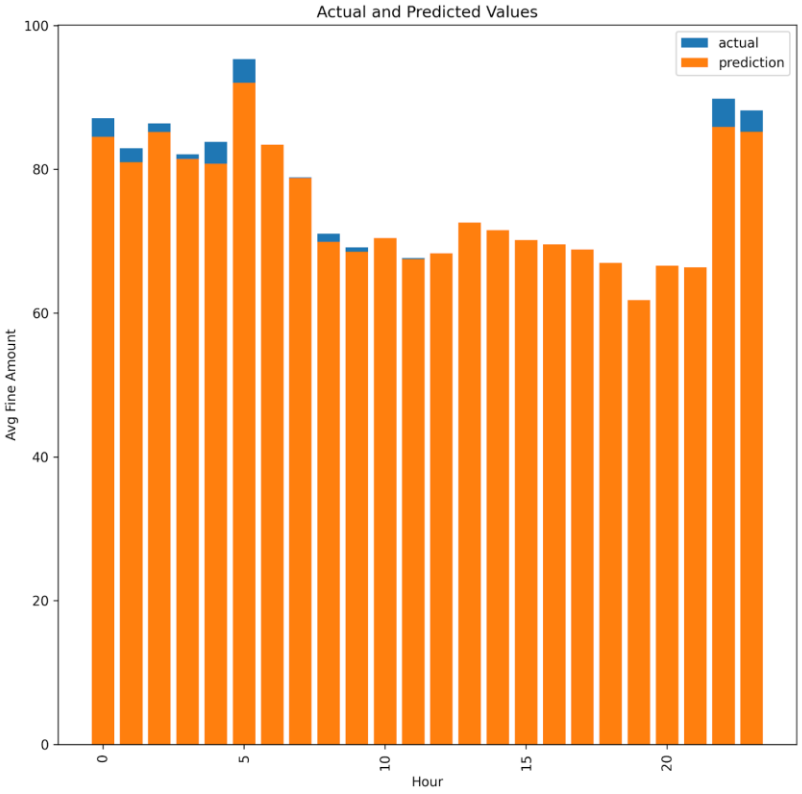


Data Visualization
Prediction: Y2 (Average Fine Amount)

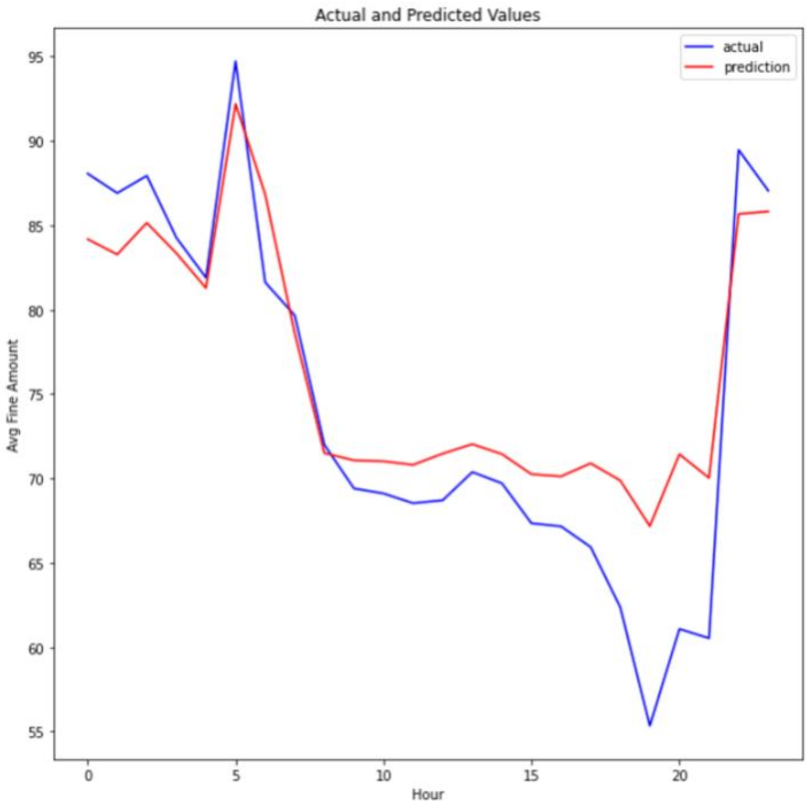


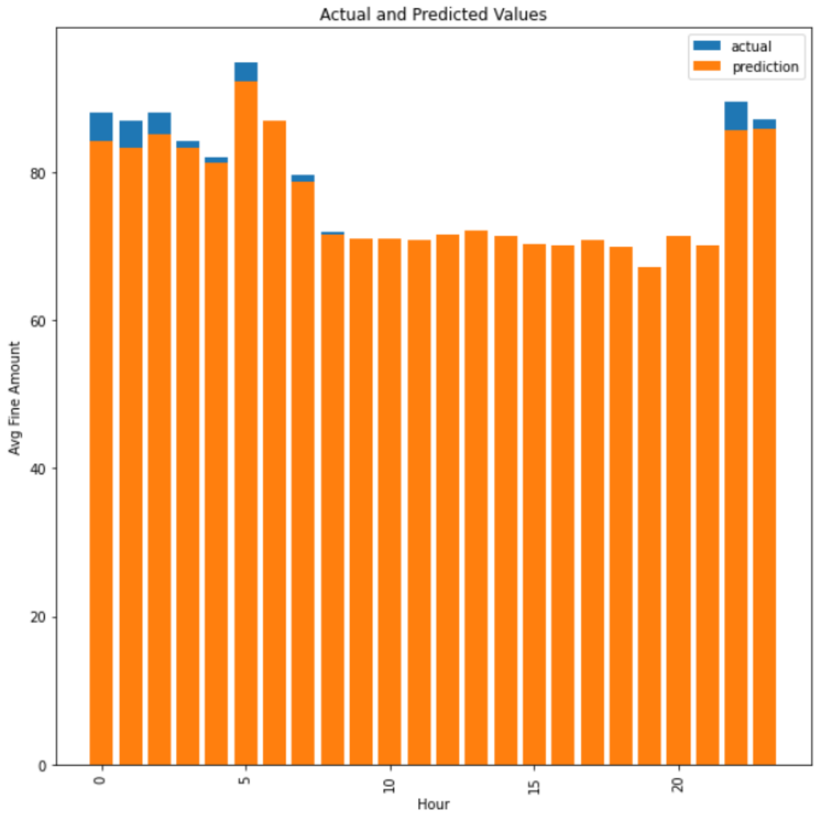
Input: Data 1 (No weather)
Prediction: Y2 (Average Fine Amount)



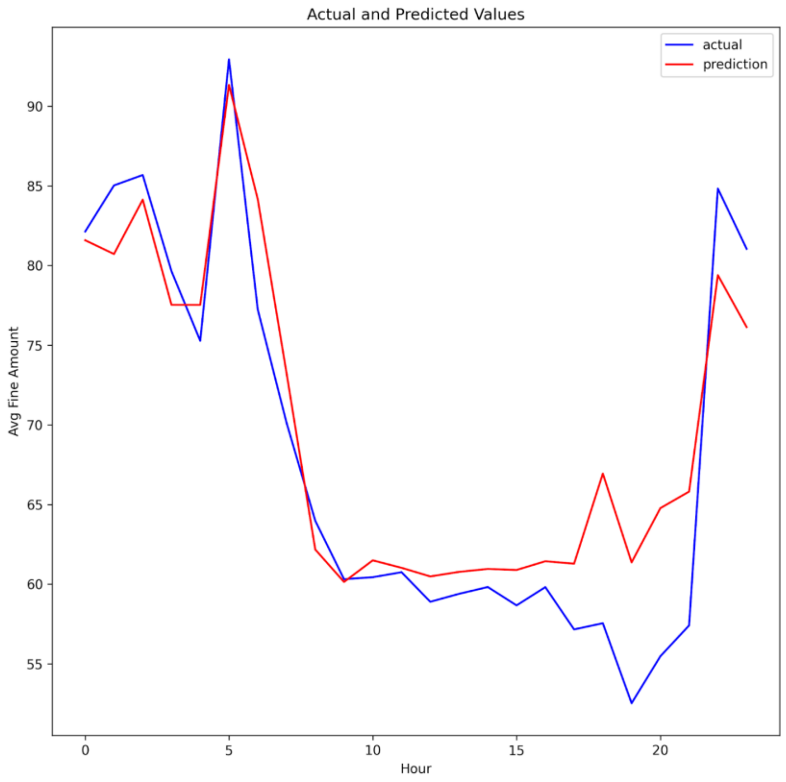


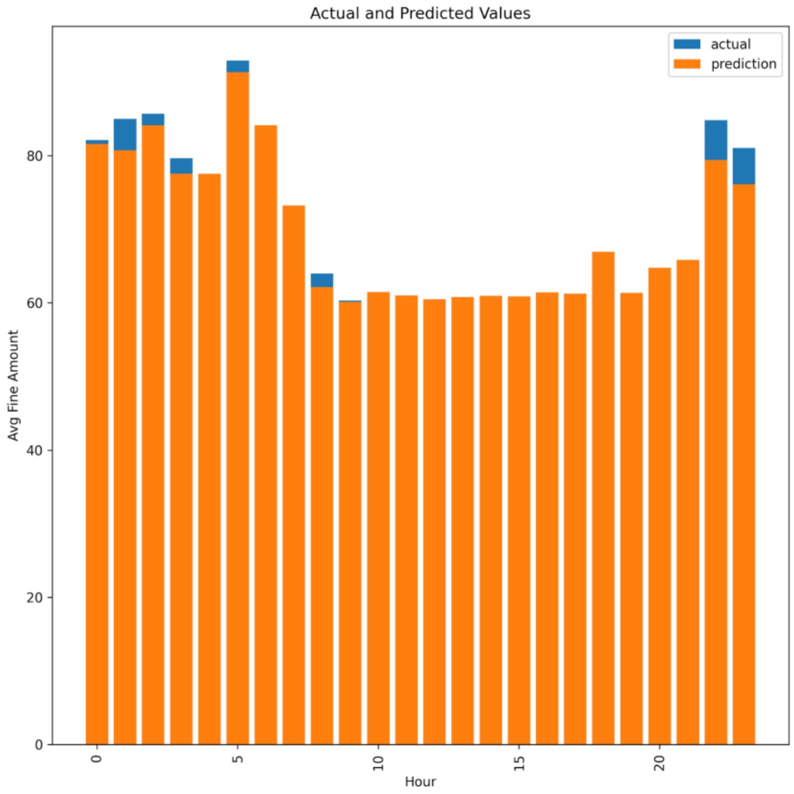
Input: Data 2 (Add NOAA Weather Data)
Prediction: Y2 (Average Fine Amount)





Input: Data 3 (Add NOAA Weather Data)
Prediction: Y2 (Average Fine Amount)

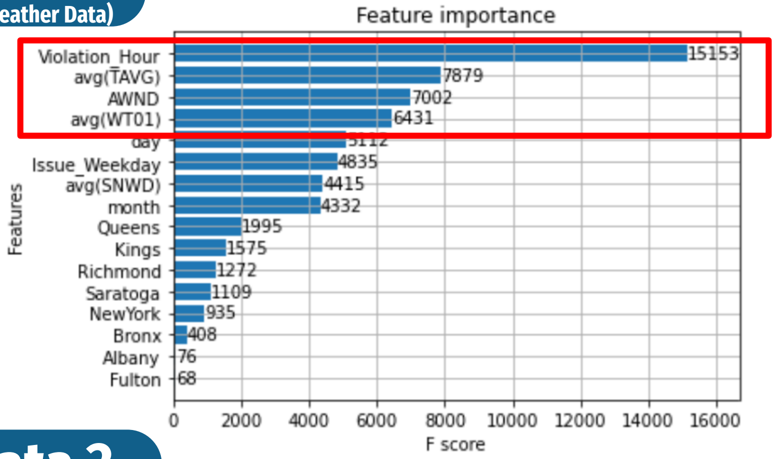




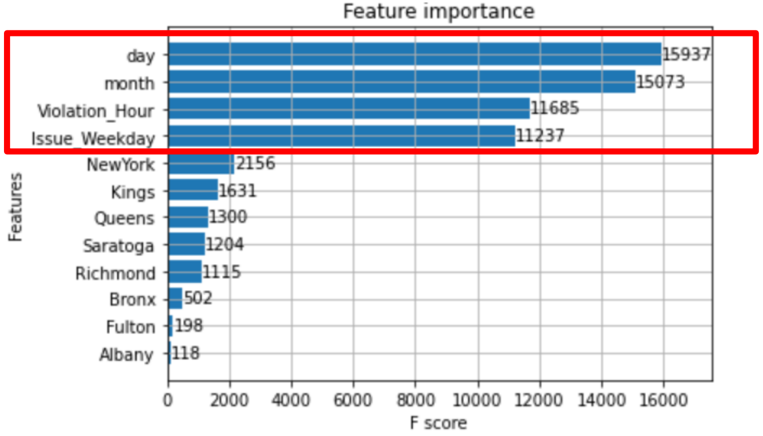
Feature Importance Analysis

Feature Importance
Y1 (Count)

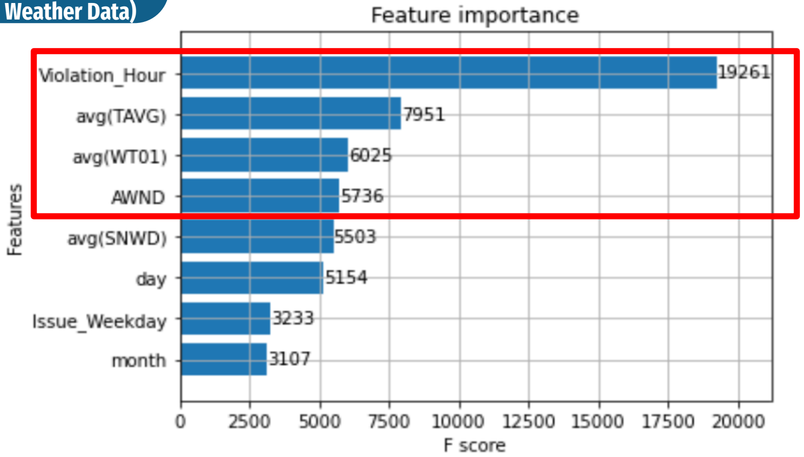
Data 2
(With Weather Data)



Data 1
(No weather)

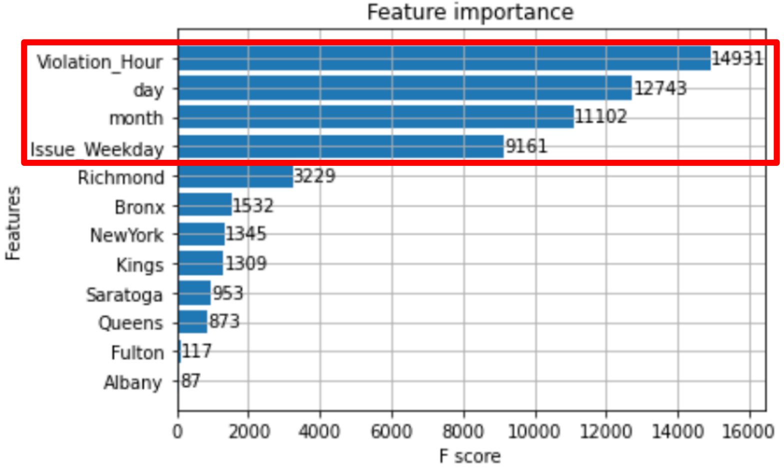


Data 3
(With Weather Data)

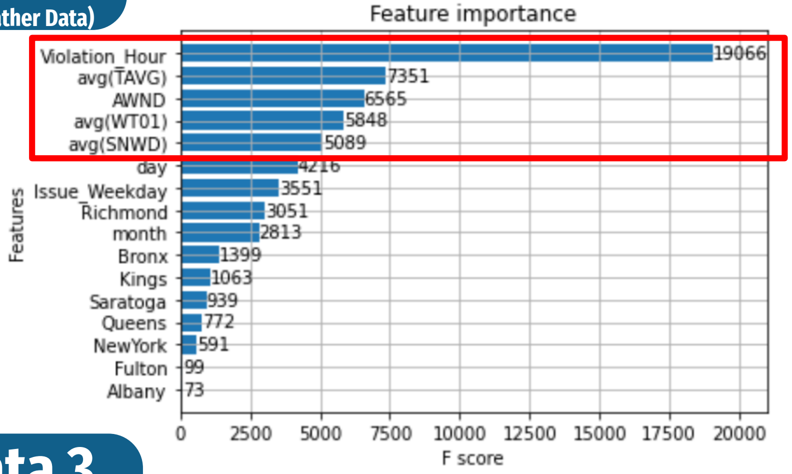


Feature Importance
Y2 (Average Fine Amount)

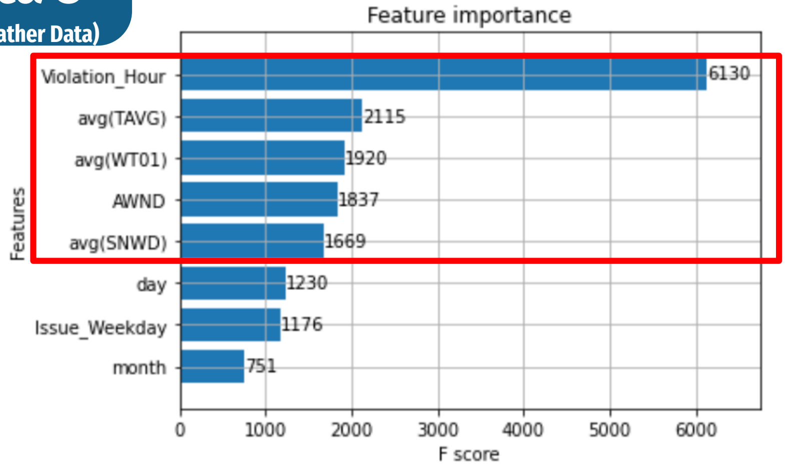
Data 1
(No weather)



Data 2
(With Weather Data)



Data 3
(With Weather Data)



Model Prediction

Prediction Model - Machine Learning using Pyspark
Y1 - Count

Y1	W/O NOAA Weather Data	W/ NOAA Weather Data	
Test data 1 (2020+2022)	Data1	Data2	Data3
Linear Regression	0.1846	0.1832	0.1086
Gradient-Boosted Trees	0.6915	0.7085	0.7509
Decision Tree	0.6851	0.6881	0.7534
Random Forest	0.6935	0.6971	0.7712
Test data 2 (2022)	Data1	Data2	Data3
Linear Regression	0.1696	0.2040	0.1091
Gradient-Boosted Trees	0.5828	0.7274	0.7300
Decision Tree	0.6180	0.7207	0.7418
Random Forest	0.6434	0.7422	0.7328

Prediction Model - Machine Learning using Pyspark
Y2 - Average Fine Amount

Y2	W/O NOAA Weather Data	W/ NOAA Weather Data	
Test data 1 (2020+2022)	Data1	Data2 (Approach 1)	Data3 (Approach 2)
Linear Regression	0.4996	0.5102	0.1791
Gradient-Boosted Trees	0.6165	0.6355	0.3864
Decision Tree	0.6089	0.6426	0.4150
Random Forest	0.6254	0.6311	0.4880
Test data 2 (2022)	Data1	Data2 (Approach 1)	Data3 (Approach 2)
Linear Regression	0.4417	0.4976	0.1762
Gradient-Boosted Trees	0.5580	0.6314	0.3672
Decision Tree	0.5637	0.6657	0.3754
Random Forest	0.5723	0.6452	0.5298

Prediction Model - Machine Learning
Y1 - Count

Y1	W/O NOAA Weather Data	W/ NOAA Weather Data	
Test data 1 (2020+2022)	Data1	Data2	Data3
XGBoost	0.6929	0.7091	0.7362
Decision Tree	0.6003	0.6362	0.6229
Random Forest	0.6776	0.7262	0.7680
Test data 2 (2022)	Data1	Data2	Data3
XGBoost	0.5744	0.7274	0.7253
Decision Tree	0.4513	0.6545	0.6355
Random Forest	0.5327	0.7491	0.7680

Prediction Model - Machine Learning
Y2 - Average Fine Amount

Y2	W/O NOAA Weather Data	W/ NOAA Weather Data	
Test data 1 (2020+2022)	Data1	Data2 (Approach 1)	Data3 (Approach 2)
XGBoost	0.6342	0.6365	0.5083
Decision Tree	0.4105	0.4095	0.1949
RandomForest	0.6058	0.6371	0.5163
Test data 2 (2022)	Data1	Data2 (Approach 1)	Data3 (Approach 2)
XGBoost	0.5724	0.7274	0.7253
Decision Tree	0.3422	0.6545	0.6355
RandomForest	0.5327	0.7491	0.7680

Prediction Model - Deep Learning (Neural Network)

Case count

Y1	w/o weather	w/ weather	
Test data 1 (2020+2022)	Data1	Data2	Data3
NN	0.7012	0.6997	-0.0705
Test data 2 (2022)	Data1	Data2	Data3
NN	0.7601	0.7327	0.2004

Overfitting?

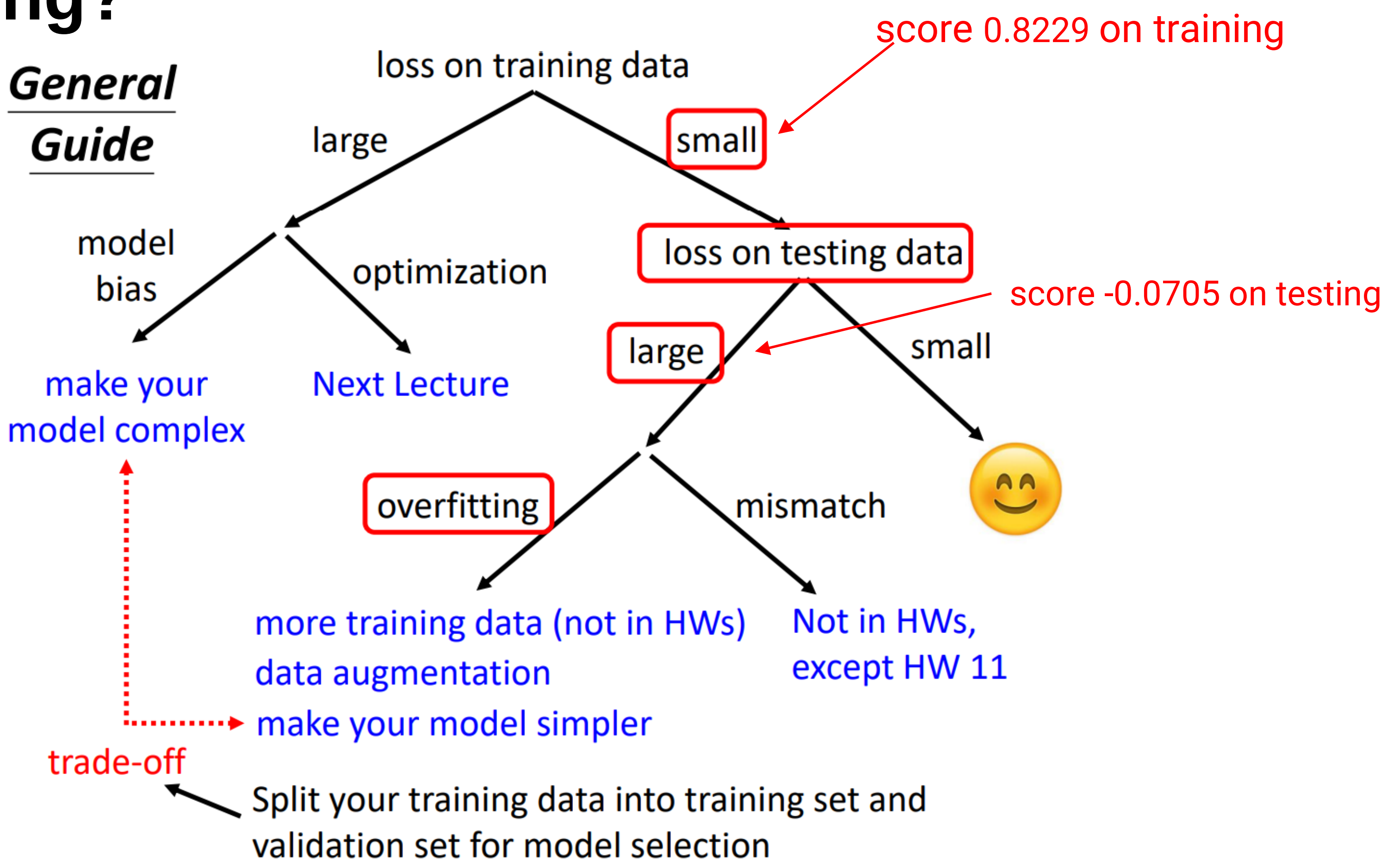
Avg fine amount

Trained on different models with several numbers of layers and choose the best scores

Y2	w/o weather	w/ weather	
Test data 1 (2020+2022)	Data1	Data2	Data3
NN	0.6255	0.6249	0.1290
Test data 2 (2022)	Data1	Data2	Data3
NN	0.6553	0.6332	0.1892

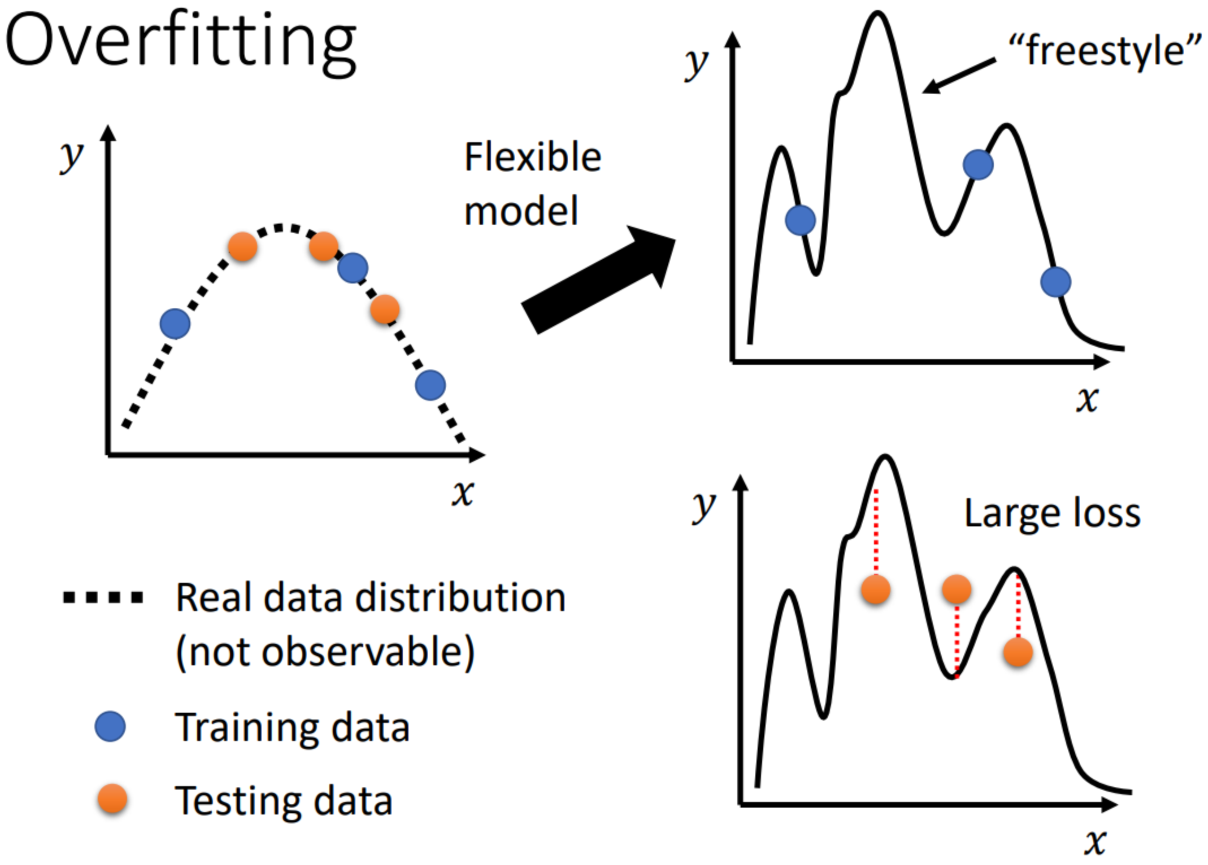
Easiness of NN to overfit on small training dataset

Overfitting?



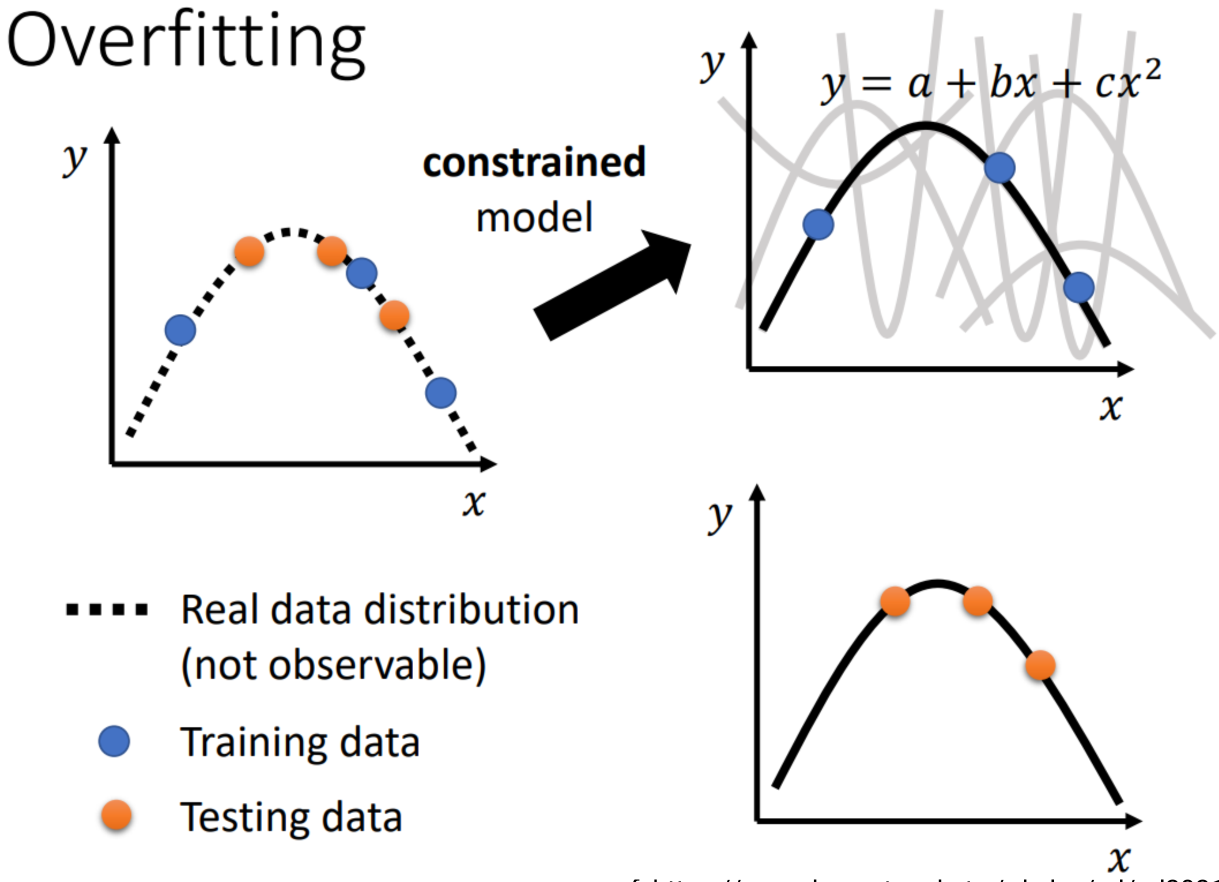
ref: https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/overfit-v6.pdf

Overfitting?



ref: https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/overfit-v6.pdf

Overfitting?



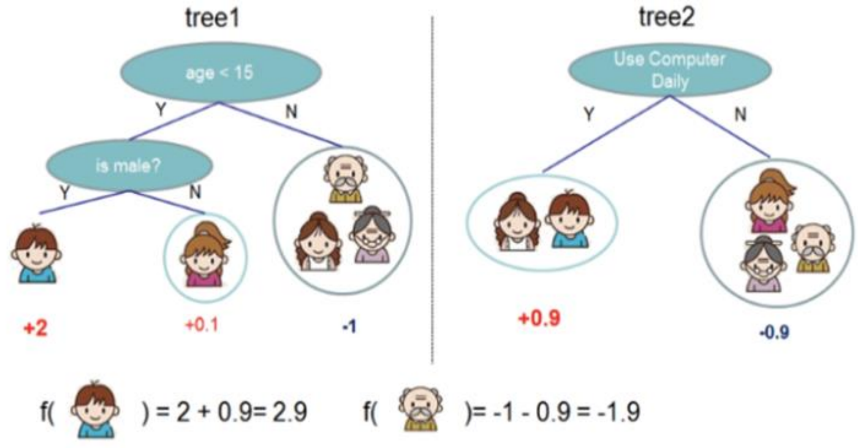
ref: https://speech.ee.ntu.edu.tw/~hylee/ml/ml2021-course-data/overfit-v6.pdf

XGboost Model

XGboost Introduction

- XGBoost (Extreme Gradient Boosting ），是一種Gradient Boosted Tree(GBDT)，將**許多弱學習器(weak learner)集合起來**變成一個比較強大的學習器(strong learner)，**每一次保留原來的模型不變**，並且**加入一個新的函數至模型中，修正上一棵樹的錯誤**，以提升整體的模型。

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$



- 左側L指的是loss function，式子後面加入一個**懲罰項Ω**，它可以**避免我們的模型overfitting**，幫助找到較好的模型。
- 可詳上圖範例圖示說明。

Challenges & Further Research

- **Possible future improvements**

Possible Future Improvements:

Implemented with ReLU or exponential

1. **Constraints on outputs**
   - (e.g. Non-negativity of case counts and fine amount)
2. **Standardize/Normalize numerical data**
   - (Reduce data fluctuation)
3. **Customized model size**
   - (With regard to the size of the dataset)
4. **Better ways to deal with missing value on weather data**
   - (e.g. Fill with geographically close counties)

Conclusions

Finding Important Features:

1. **Hour**
   - 08~09：案件數多
   - 22~05：案件數少，平均罰款高
   - 白天：平均罰款低
2. **Weather**
   - 氣溫、風速以及霧的指標對於預測有很顯著的幫助
   - 降雪量相對影響較小
3. **Week day**
   - 平日：案件數多
   - 假日：案件數少，尤其週日

Reference

- Open Parking and Camera Violations
  https://data.cityofnewyork.us/City-Government/Open-Parking-and-Camera-Violations/nc67-uf89
- Pyspark
  https://spark.apache.org/docs/latest/api/python/
- Scikit-learn
  https://scikit-learn.org/stable/
- Pandas
  https://pandas.pydata.org/
- Weather Source
  https://www.ncei.noaa.gov/support/access-support-service
  https://www.weather.gov/wrh/Climate?wfo=okx
- Nyc-parking-tickets
  https://www.kaggle.com/new-york-city/nyc-parking-tickets
- XGBoost
  https://medium.com/chung-yi/xgboost%E4%BB%8B%E7%B4%B9-b31f7ec8295e
  https://xgboost.readthedocs.io/en/stable/

THANK YOU