# KDD Cup 2017

M10515031 黃佳郁
M10515036 謝奇元
M10515104 羅煜賢

指導教授:李漢銘 教授

# Problem Description

## TASK 1:

## Travel Time Prediction

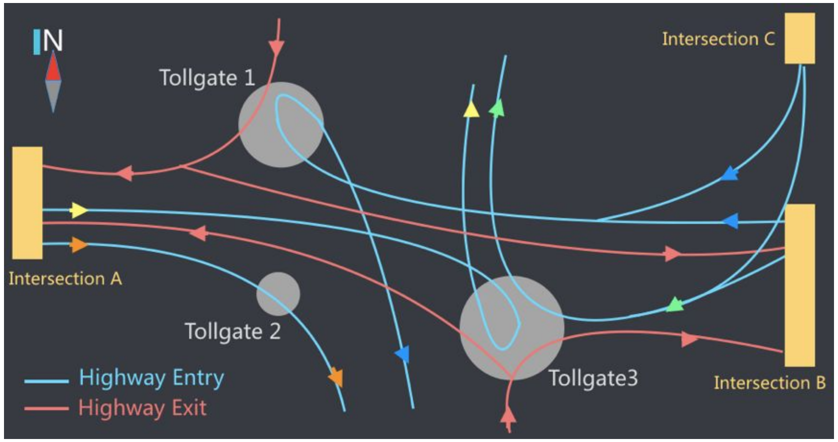
For every 20-minute time window, estimate the average travel time of each route.

A. Intersection A - Tollgates 2 & 3
B. Intersection B - Tollgates 1 & 3
C. Intersection C - Tollgates 1 & 3

## TASK 2:

## Traffic Volume Prediction

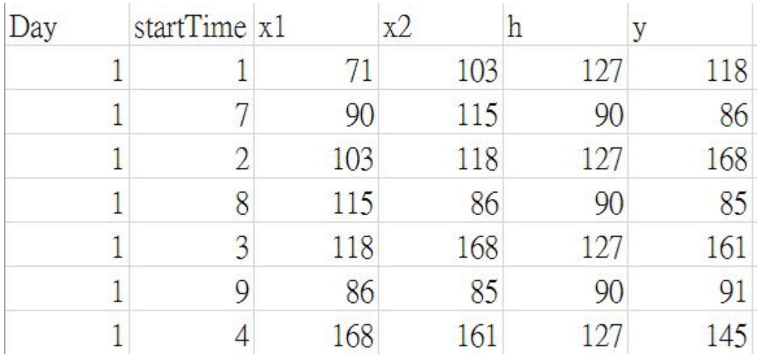
For every 20-minute time window, predict the entry and exit traffic volumes at tollgates 1, 2 and 3 from Oct. 25th - Oct. 31st.



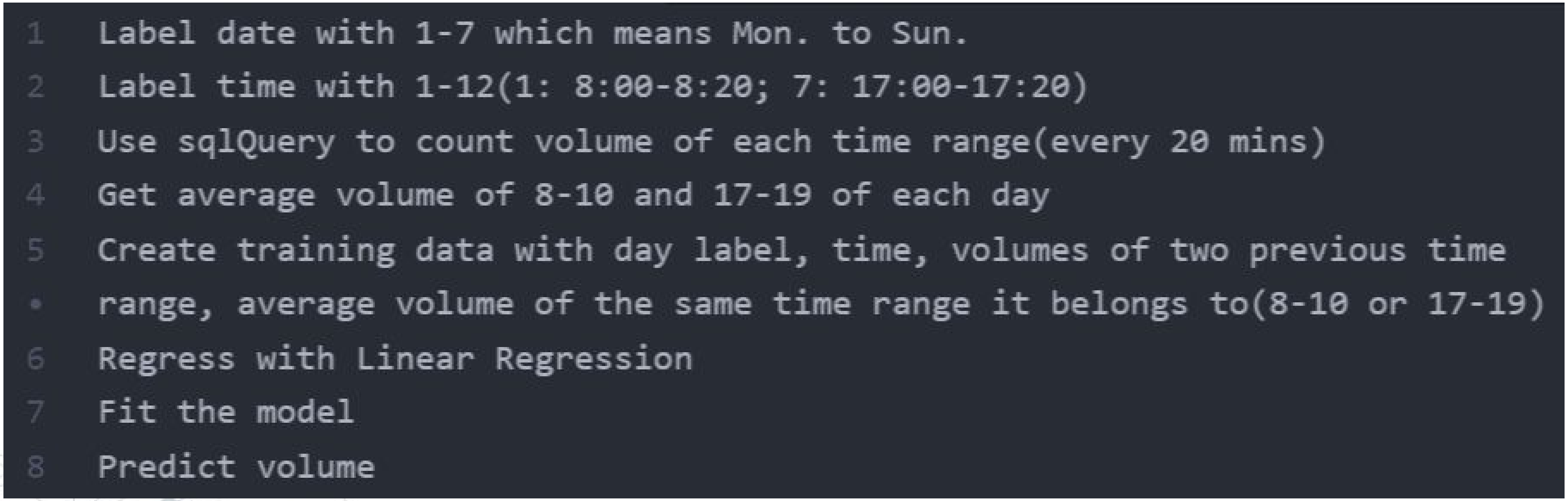
**Competition website : KDD Cup 2017**

# How To Do?

# TASK 2



| Day | startTime | x1 | x2 | h | y |
|---|---|---|---|---|---|
| 1 | 1 | 71 | 103 | 127 | 118 |
| 1 | 7 | 90 | 115 | 90 | 86 |
| 1 | 2 | 103 | 118 | 127 | 168 |
| 1 | 8 | 115 | 86 | 90 | 85 |
| 1 | 3 | 118 | 168 | 127 | 161 |
| 1 | 9 | 86 | 85 | 90 | 91 |
| 1 | 4 | 168 | 161 | 127 | 145 |



```
Label date with 1-7 which means Mon. to Sun.
Label time with 1-12(1: 8:00-8:20; 7: 17:00-17:20)
Use sqlQuery to count volume of each time range(every 20 mins)
Get average volume of 8-10 and 17-19 of each day
Create training data with day label, time, volumes of two previous time
range, average volume of the same time range it belongs to(8-10 or 17-19)
Regress with Linear Regression
Fit the model
Predict volume
```

# Previous Result

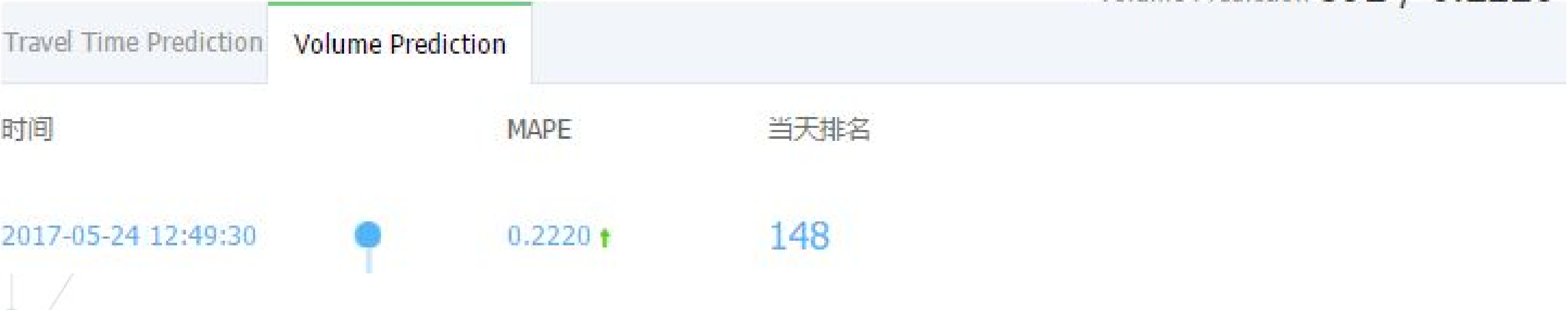
**Linear Regression** **0.2539**

**Feature:**

- What day?
- Time range
  - One unit / 20 min
  - 8-10, 17-19
- Volume of two previous time range
  - 8:00-8:20 => 7:20-7:40, 7:40-8:00
- Average volume of the same time range(8-10 or 17-19) of that day

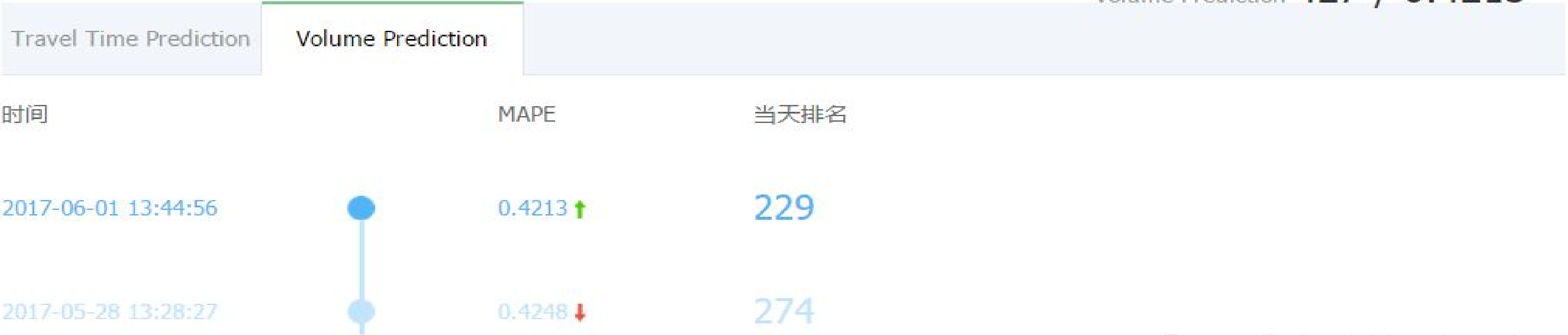
# Grade

Volume Prediction 592 / 0.2220



| 时间 | MAPE | 当天排名 |
|---|---|---|
| 2017-05-24 12:49:30 | 0.2220 ↑ | 148 |

Volume Prediction 427 / 0.4213



| 时间 | MAPE | 当天排名 |
|---|---|---|
| 2017-06-01 13:44:56 | 0.4213 ↑ | 229 |
| 2017-05-28 13:28:27 | 0.4248 ↓ | 274 |

# Using Xgboost

**n_estimators=10**

- 0.2457

**n_estimators=20**

- 0.2401

**n_estimators=100**

- 0.2451

Using xgboost method surely improve the result. First, we try different n_estimators.

# Using Xgboost

**linear booster; n_estimator=50**

- 0.2500

**tweedie**

- 0.2550

**gamma**

- 0.2665

According to the result, n_estimators=20 when using linear booster got the best result.

# Using Xgboost

**learning_rate=0.55**

◎ 0.2248

**learning_rate=0.6**

◎ 0.2220

**learning_rate=0.65**

◎ 0.2295

According to the result, when reducing learning_rate, the result might be better.

# Briefly conclusion for Task 2

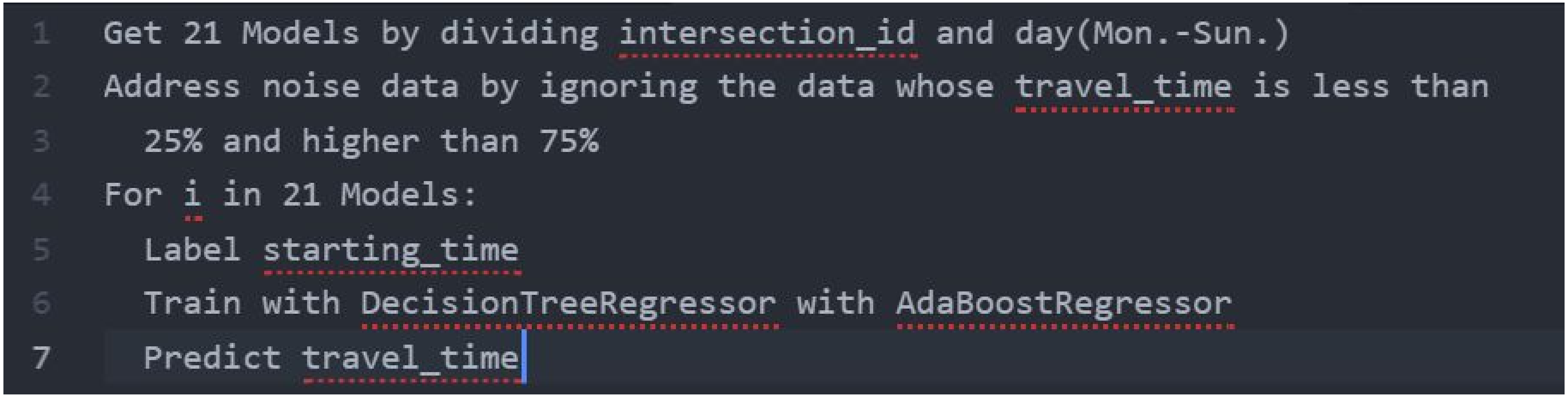
- We can infer that the model fall into overfitting/underfitting problem  from our experiments. (Overfitting, estimator >= 50; Underfitting estimator = 10.) Therefore, we chose 20.

- Second, we chose differents algorithms, including Linear Regression, Gamma, Tweedie. Linear Regression is better than others, according to results.

- Next, we modified learning rate for our model, get 0.6 is better than 5.5 and 6.5.

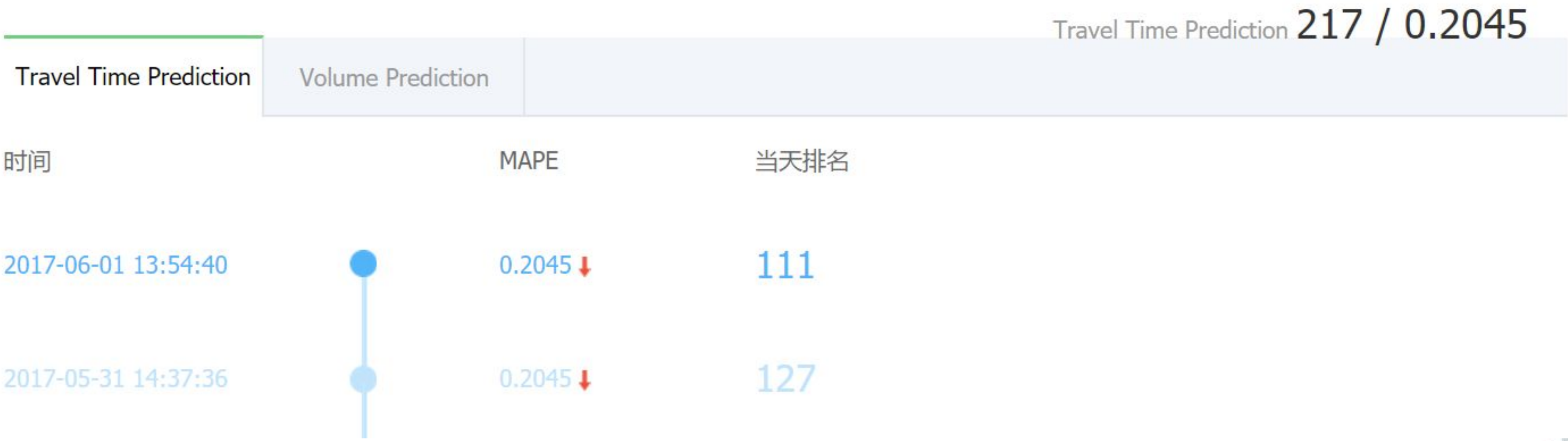
- Finally, we get 0.2220 grade by using Linear Regression with estimators 20 and 0.6 learning rate in our xgboost parameters

# TASK 1



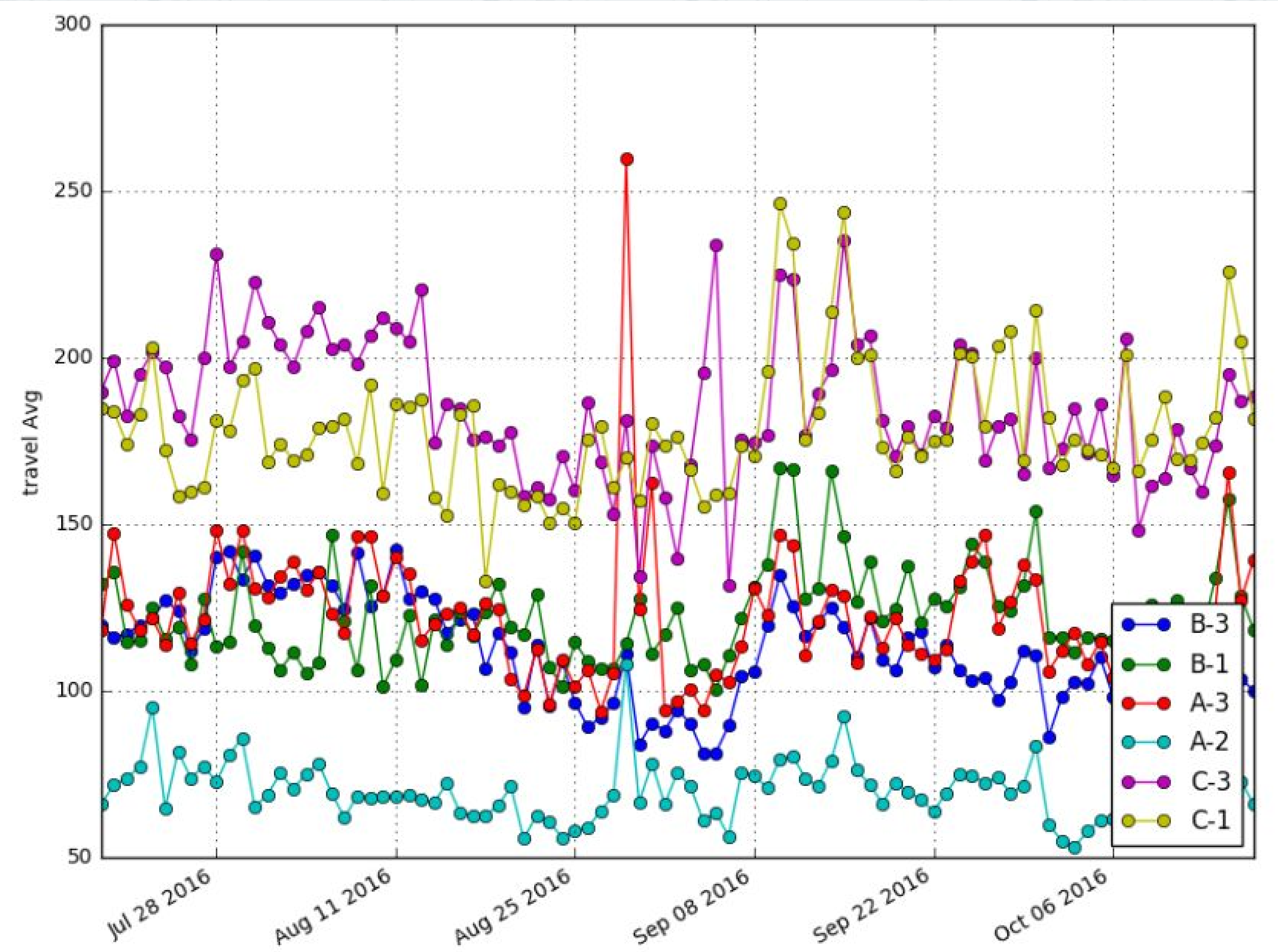
```
Get 21 Models by dividing intersection_id and day(Mon.-Sun.)
Address noise data by ignoring the data whose travel_time is less than
  25% and higher than 75%
For i in 21 Models:
  Label starting_time
  Train with DecisionTreeRegressor with AdaBoostRegressor
  Predict travel_time
```

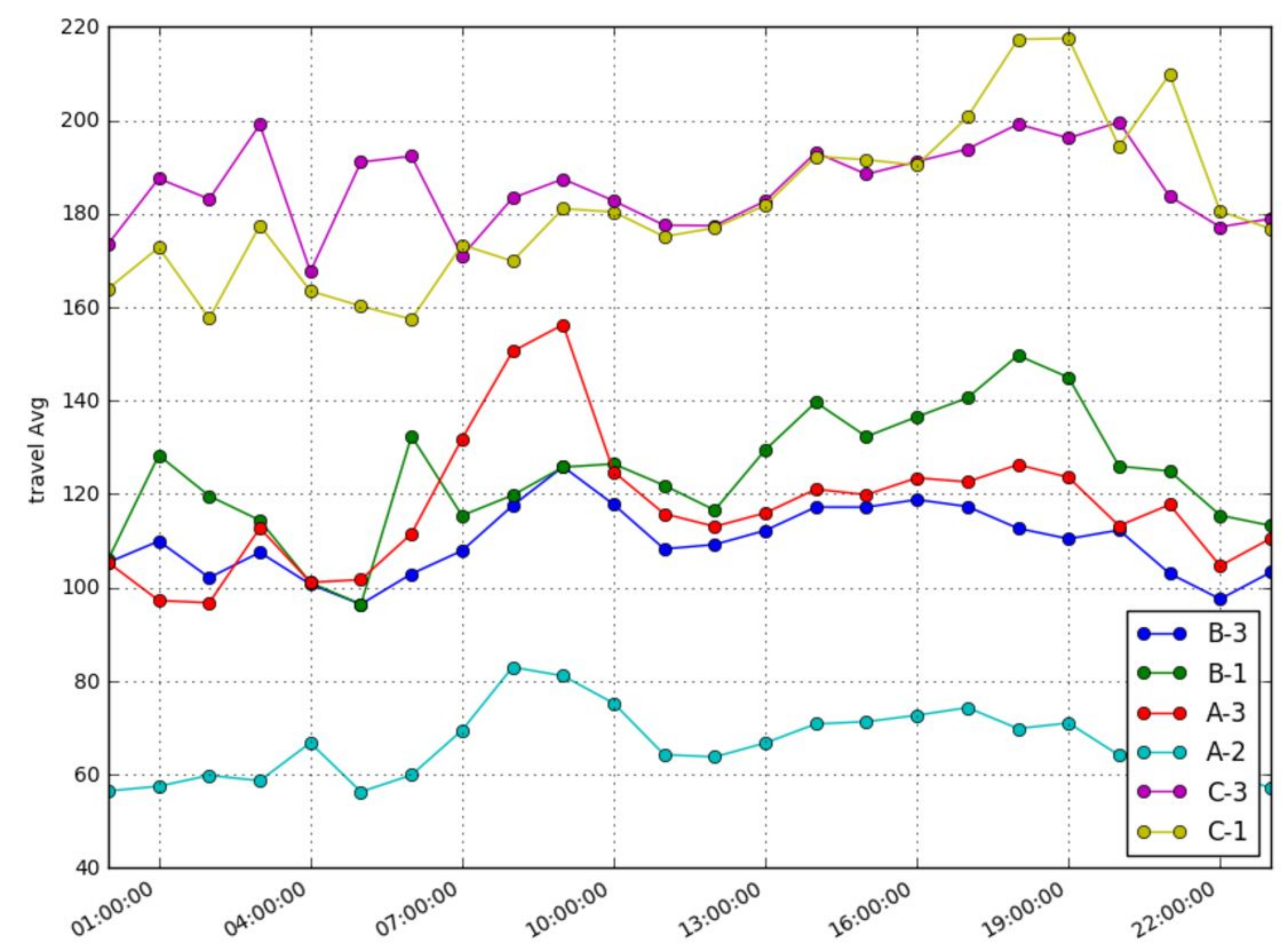
# Grade



Travel Time Prediction 217 / 0.2045

| 时间 | MAPE | 当天排名 |
|---|---|---|
| 2017-06-01 13:54:40 | 0.2045 ↓ | 111 |
| 2017-05-31 14:37:36 | 0.2045 ↓ | 127 |


300
250
200
150
100
50
travel Avg
Jul 28 2016
Aug 11 2016
Aug 25 2016
Sep 08 2016
Sep 22 2016
Oct 06 2016
B-3
B-1
A-3
A-2
C-3
C-1


220
200
180
160
140
120
100
80
60
40
travel Avg
01:00:00
04:00:00
07:00:00
10:00:00
13:00:00
16:00:00
19:00:00
22:00:00
B-3
B-1
A-3
A-2
C-3
C-1

# Data Preprocessing

- Divide data into 21 sets
  - with intersection id (A, B, C)
  - what day? (Mon. - Sun.)
- Label Mon.- Sun. with 0-7

Training models => 21 models ( 3 * 7 )

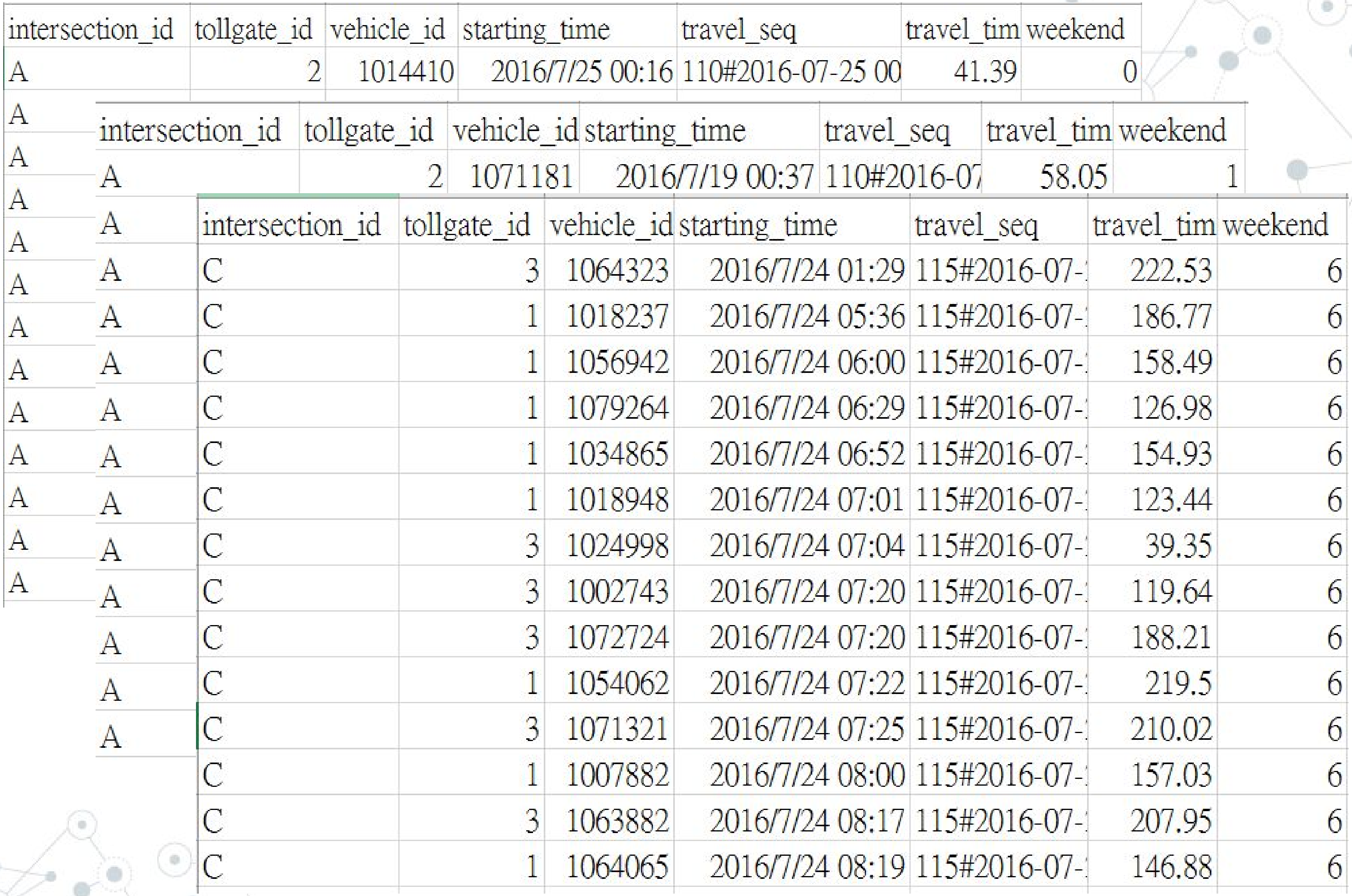
intersection A - Mon.

intersection A - Tue.

intersection A - Wed.

…

intersection C - Sun.



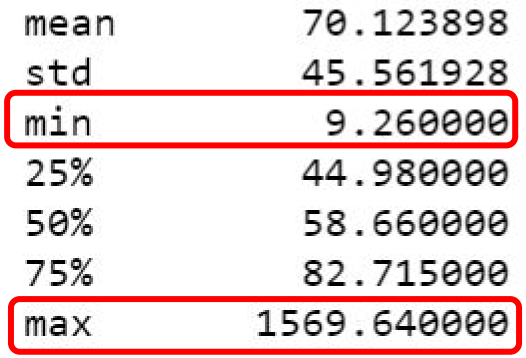
| intersection_id | tollgate_id | vehicle_id | starting_time | travel_seq | travel_tim | weekend |
|---|---|---|---|---|---|---|
| A | 2 | 1014410 | 2016/7/25 00:16 | 110#2016-07-25 00 | 41.39 | 0 |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |

| intersection_id | tollgate_id | vehicle_id | starting_time | travel_seq | travel_tim | weekend |
|---|---|---|---|---|---|---|
| A | 2 | 1071181 | 2016/7/19 00:37 | 110#2016-07 | 58.05 | 1 |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |
| A | | | | | | |

| intersection_id | tollgate_id | vehicle_id | starting_time | travel_seq | travel_tim | weekend |
|---|---|---|---|---|---|---|
| C | 3 | 1064323 | 2016/7/24 01:29 | 115#2016-07- | 222.53 | 6 |
| C | 1 | 1018237 | 2016/7/24 05:36 | 115#2016-07- | 186.77 | 6 |
| C | 1 | 1056942 | 2016/7/24 06:00 | 115#2016-07- | 158.49 | 6 |
| C | 1 | 1079264 | 2016/7/24 06:29 | 115#2016-07- | 126.98 | 6 |
| C | 1 | 1034865 | 2016/7/24 06:52 | 115#2016-07- | 154.93 | 6 |
| C | 1 | 1018948 | 2016/7/24 07:01 | 115#2016-07- | 123.44 | 6 |
| C | 3 | 1024998 | 2016/7/24 07:04 | 115#2016-07- | 39.35 | 6 |
| C | 3 | 1002743 | 2016/7/24 07:20 | 115#2016-07- | 119.64 | 6 |
| C | 3 | 1072724 | 2016/7/24 07:20 | 115#2016-07- | 188.21 | 6 |
| C | 1 | 1054062 | 2016/7/24 07:22 | 115#2016-07- | 219.5 | 6 |
| C | 3 | 1071321 | 2016/7/24 07:25 | 115#2016-07- | 210.02 | 6 |
| C | 1 | 1007882 | 2016/7/24 08:00 | 115#2016-07- | 157.03 | 6 |
| C | 3 | 1063882 | 2016/7/24 08:17 | 115#2016-07- | 207.95 | 6 |
| C | 1 | 1064065 | 2016/7/24 08:19 | 115#2016-07- | 146.88 | 6 |

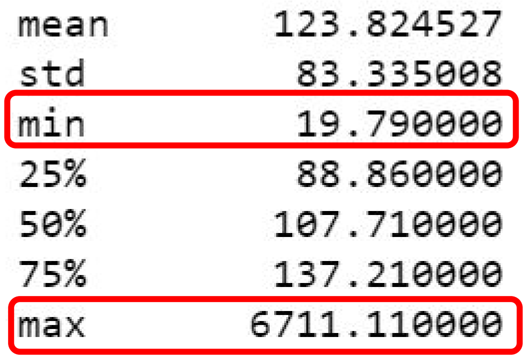
# Problem - Travel Time

## A-2



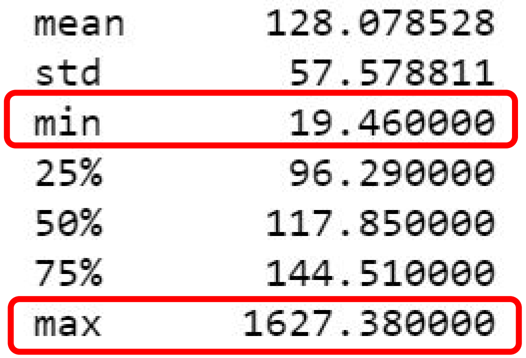
```
mean      70.123898
std       45.561928
min        9.260000
25%       44.980000
50%       58.660000
75%       82.715000
max     1569.640000
```

## A-3



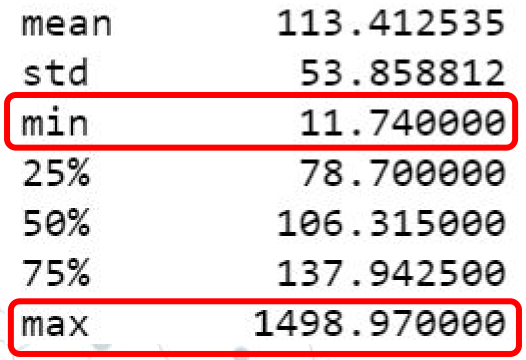
```
mean     123.824527
std       83.335008
min       19.790000
25%       88.860000
50%      107.710000
75%      137.210000
max     6711.110000
```

## B-1



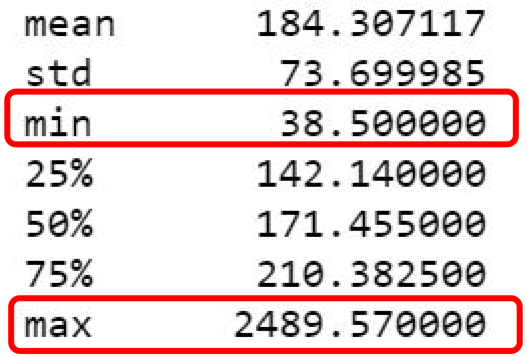
```
mean     128.078528
std       57.578811
min       19.460000
25%       96.290000
50%      117.850000
75%      144.510000
max     1627.380000
```

## B-3



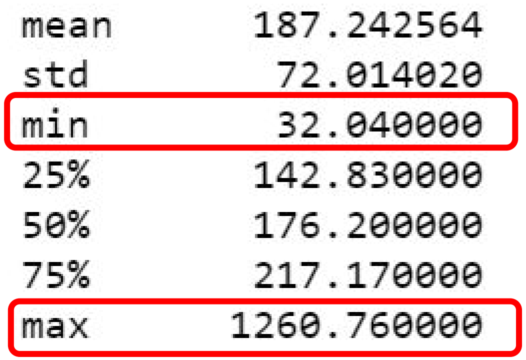
```
mean     113.412535
std       53.858812
min       11.740000
25%       78.700000
50%      106.315000
75%      137.942500
max     1498.970000
```

## C-1

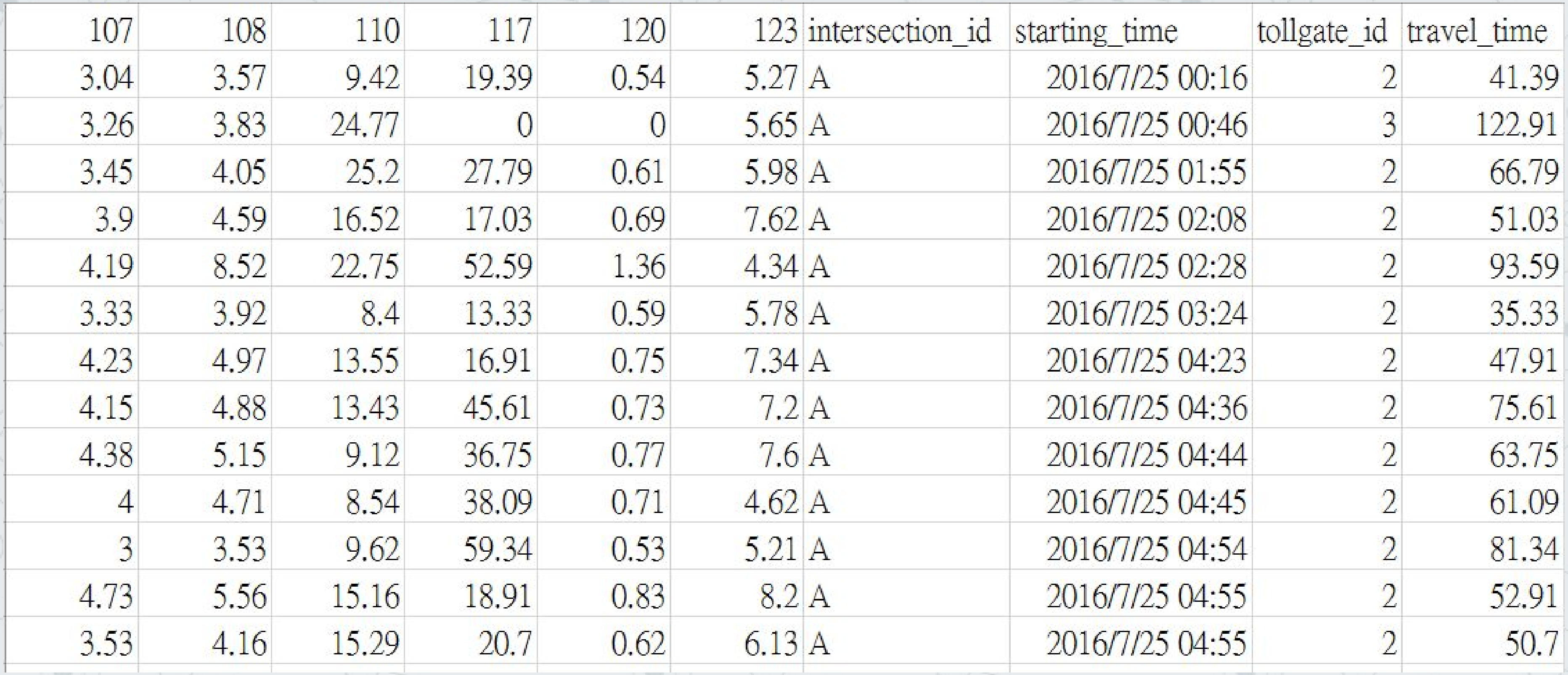


```
mean     184.307117
std       73.699985
min       38.500000
25%      142.140000
50%      171.455000
75%      210.382500
max     2489.570000
```

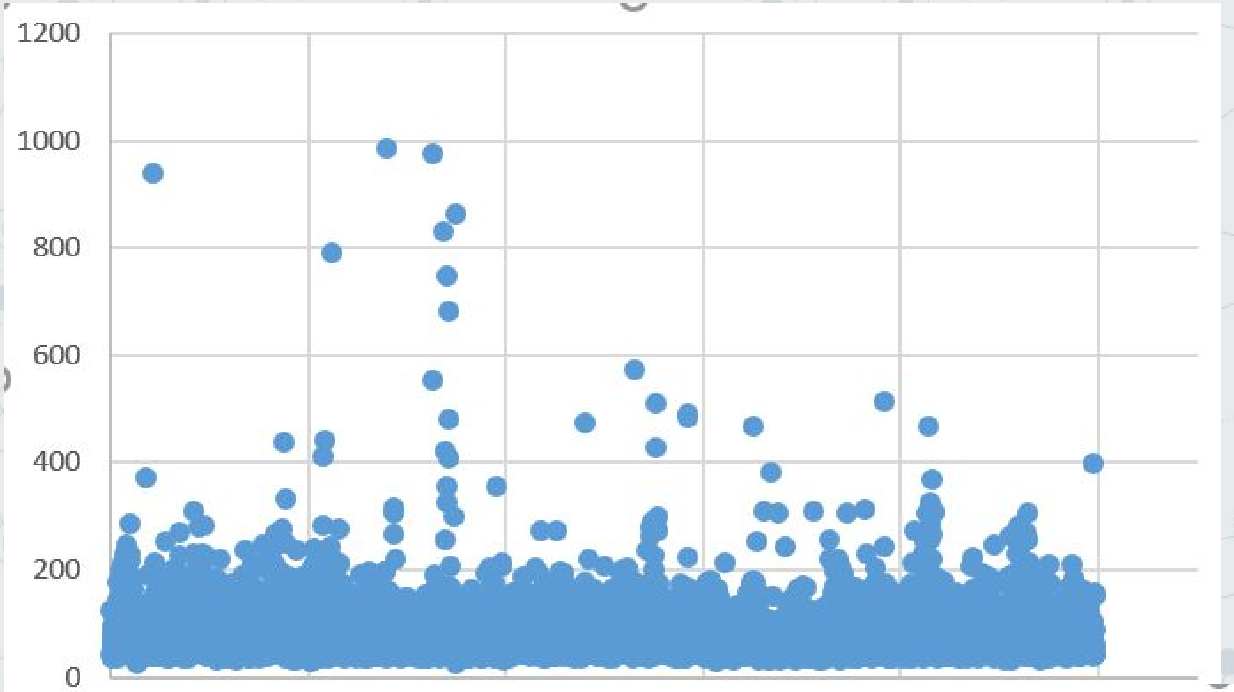
## C-3



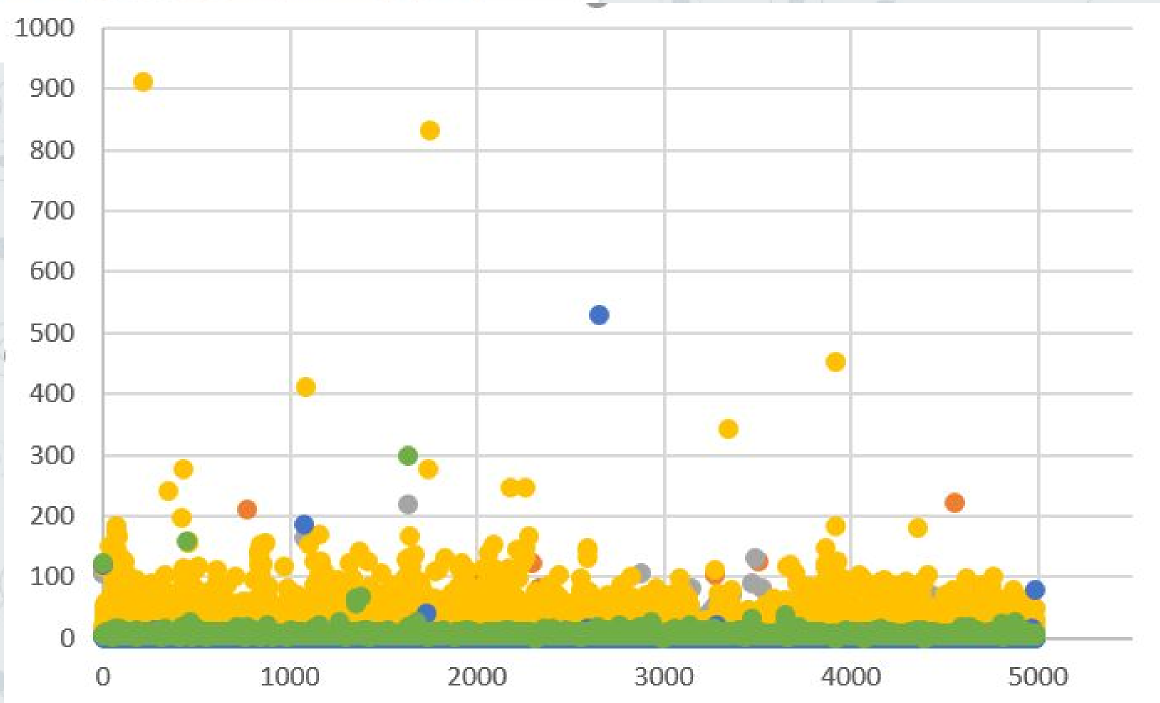
```
mean     187.242564
std       72.014020
min       32.040000
25%      142.830000
50%      176.200000
75%      217.170000
max     1260.760000
```



| 107 | 108 | 110 | 117 | 120 | 123 | intersection_id | starting_time | tollgate_id | travel_time |
|---|---|---|---|---|---|---|---|---|---|
| 3.04 | 3.57 | 9.42 | 19.39 | 0.54 | 5.27 | A | 2016/7/25 00:16 | 2 | 41.39 |
| 3.26 | 3.83 | 24.77 | 0 | 0 | 5.65 | A | 2016/7/25 00:46 | 3 | 122.91 |
| 3.45 | 4.05 | 25.2 | 27.79 | 0.61 | 5.98 | A | 2016/7/25 01:55 | 2 | 66.79 |
| 3.9 | 4.59 | 16.52 | 17.03 | 0.69 | 7.62 | A | 2016/7/25 02:08 | 2 | 51.03 |
| 4.19 | 8.52 | 22.75 | 52.59 | 1.36 | 4.34 | A | 2016/7/25 02:28 | 2 | 93.59 |
| 3.33 | 3.92 | 8.4 | 13.33 | 0.59 | 5.78 | A | 2016/7/25 03:24 | 2 | 35.33 |
| 4.23 | 4.97 | 13.55 | 16.91 | 0.75 | 7.34 | A | 2016/7/25 04:23 | 2 | 47.91 |
| 4.15 | 4.88 | 13.43 | 45.61 | 0.73 | 7.2 | A | 2016/7/25 04:36 | 2 | 75.61 |
| 4.38 | 5.15 | 9.12 | 36.75 | 0.77 | 7.6 | A | 2016/7/25 04:44 | 2 | 63.75 |
| 4 | 4.71 | 8.54 | 38.09 | 0.71 | 4.62 | A | 2016/7/25 04:45 | 2 | 61.09 |
| 3 | 3.53 | 9.62 | 59.34 | 0.53 | 5.21 | A | 2016/7/25 04:54 | 2 | 81.34 |
| 4.73 | 5.56 | 15.16 | 18.91 | 0.83 | 8.2 | A | 2016/7/25 04:55 | 2 | 52.91 |
| 3.53 | 4.16 | 15.29 | 20.7 | 0.62 | 6.13 | A | 2016/7/25 04:55 | 2 | 50.7 |



total travel time



travel time of each link

# Remove Noise

- As the table shown, there exists noise in the data (e.g. maximum time, minimum time)

- As the figure shown, the distribution position of total time is similar to that of travel time of each link

- Take the data which total travel time is in the range of 25%-75% as training data

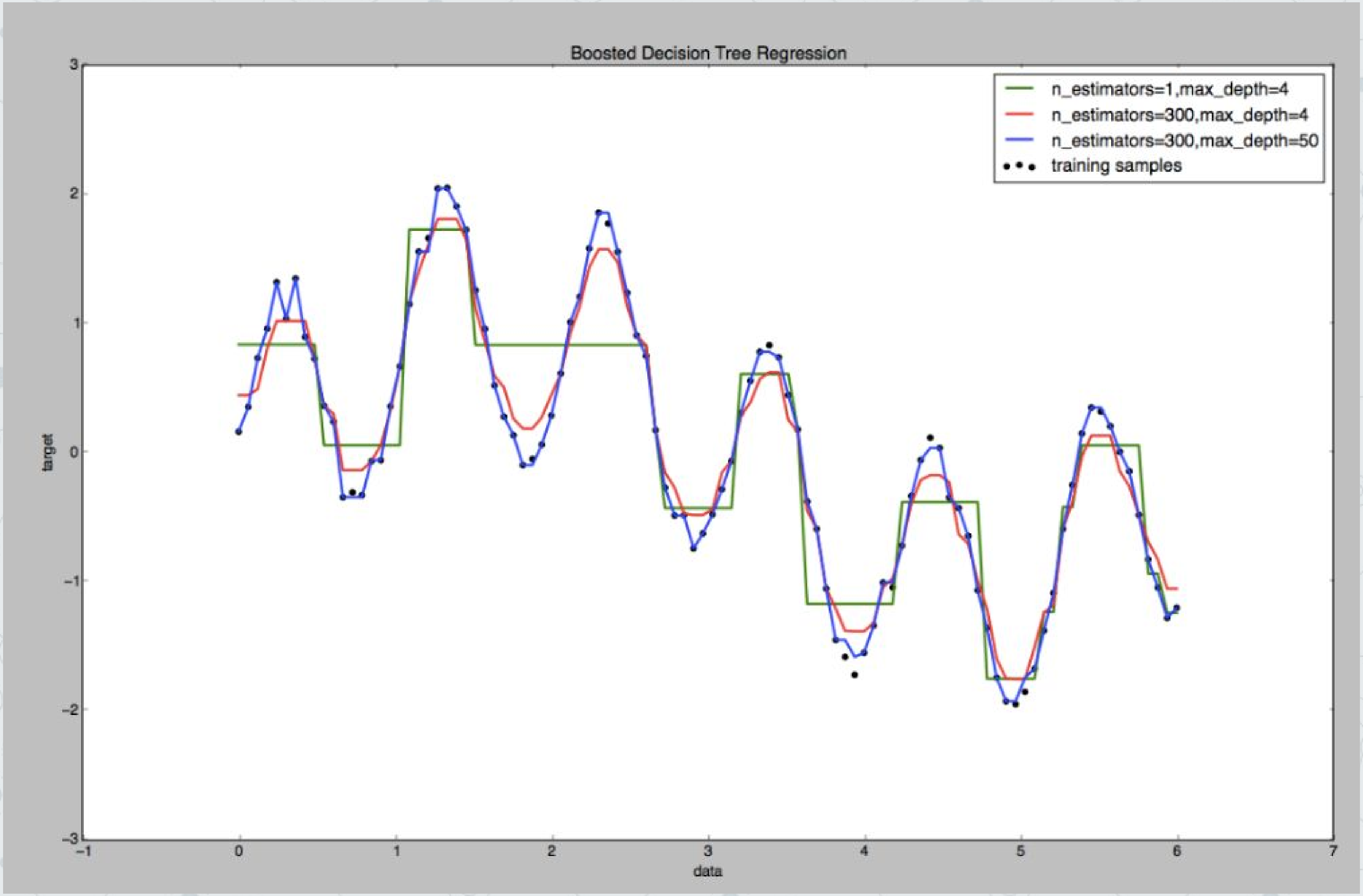
# Decision Tree with Adaboost

**Decision Tree**

- Easily overfit

**Adaboost**

- Improve accuracy
- Avoid overfitting


Boosted Decision Tree Regression
n_estimators=1,max_depth=4
n_estimators=300,max_depth=4
n_estimators=300,max_depth=50
training samples
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6
7

# Parameters Selection

### n_estimators

- 50
- 100
- 300

### loss function

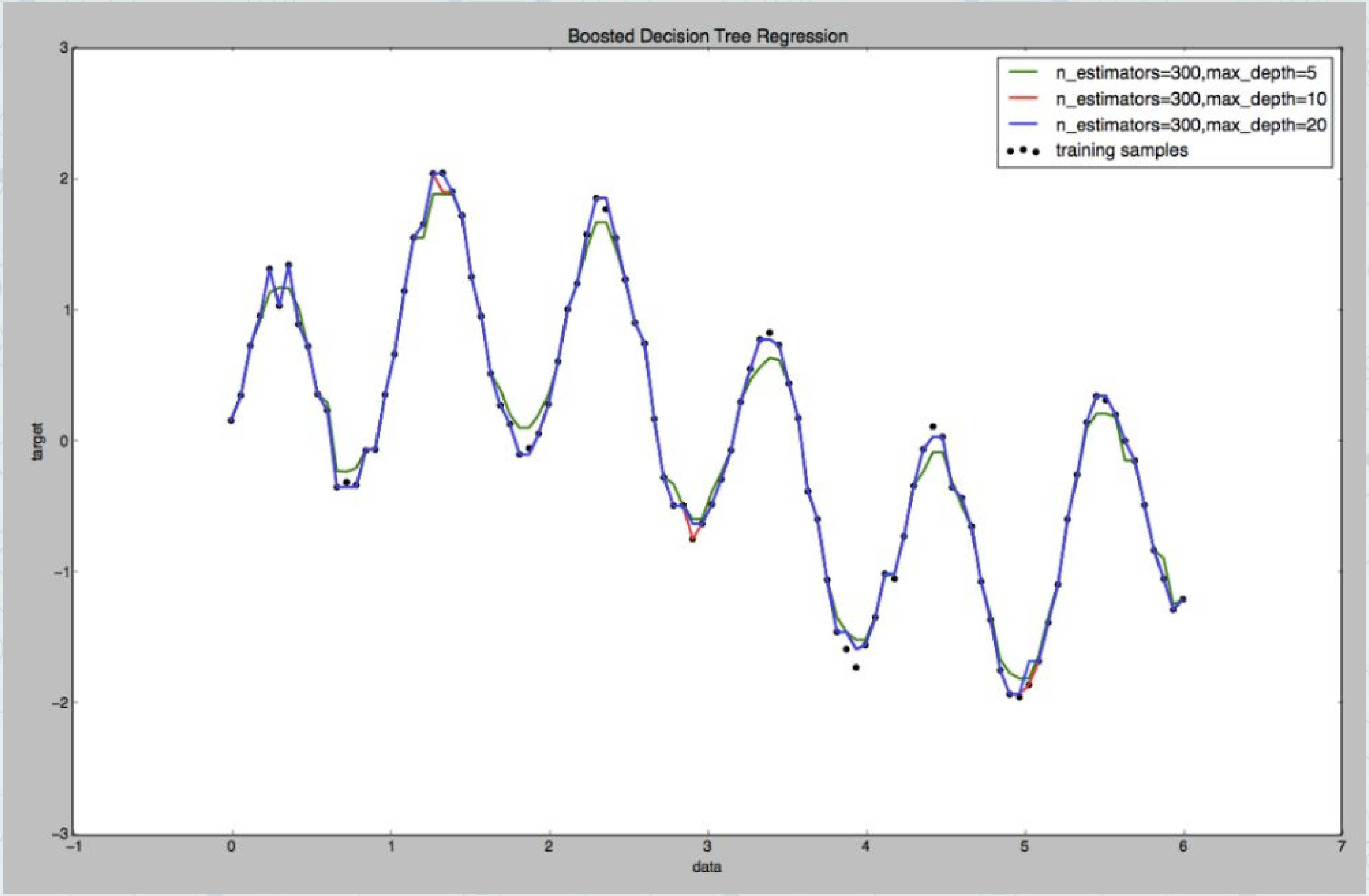
- linear
- square
- exponential

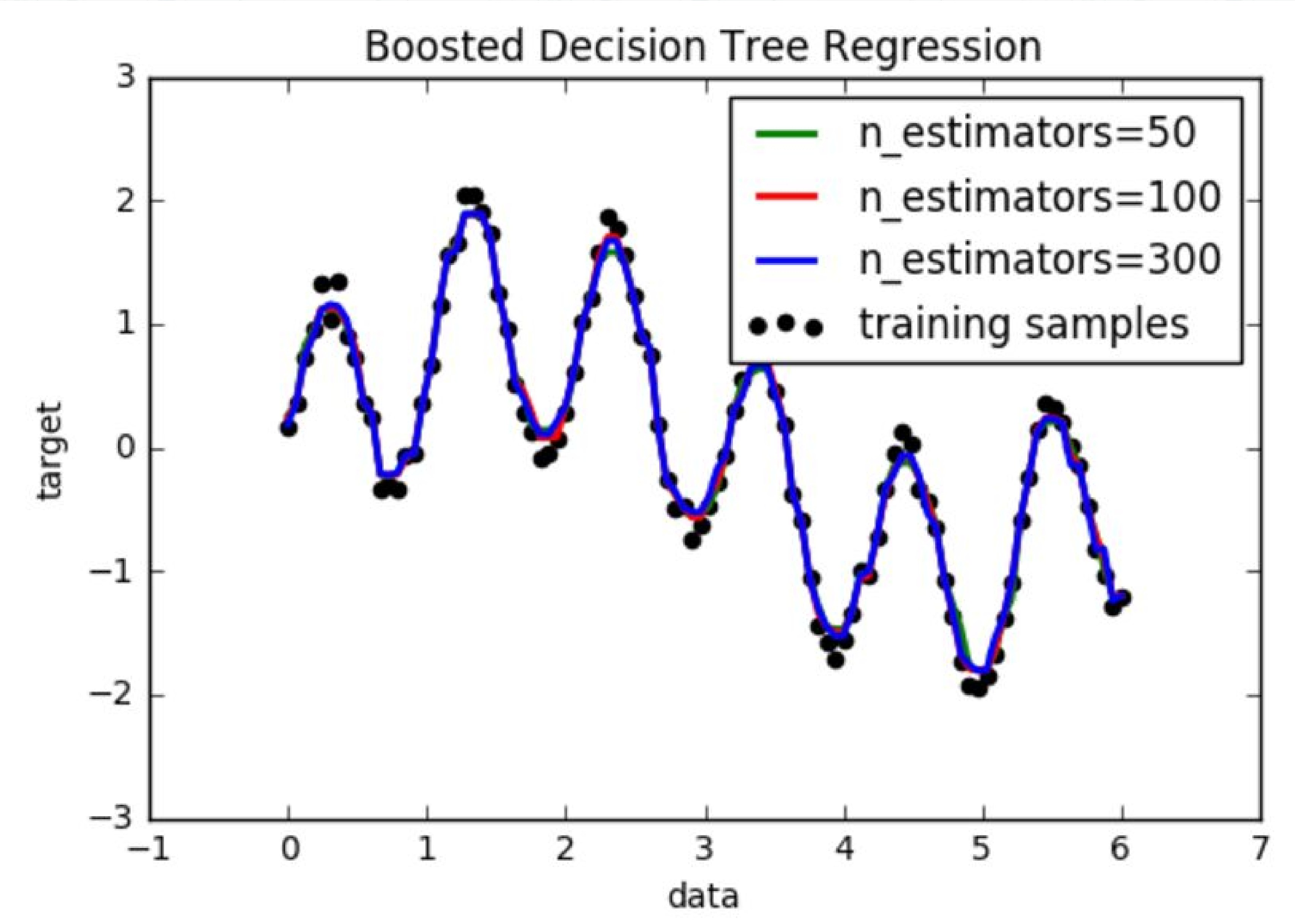
### max_depth

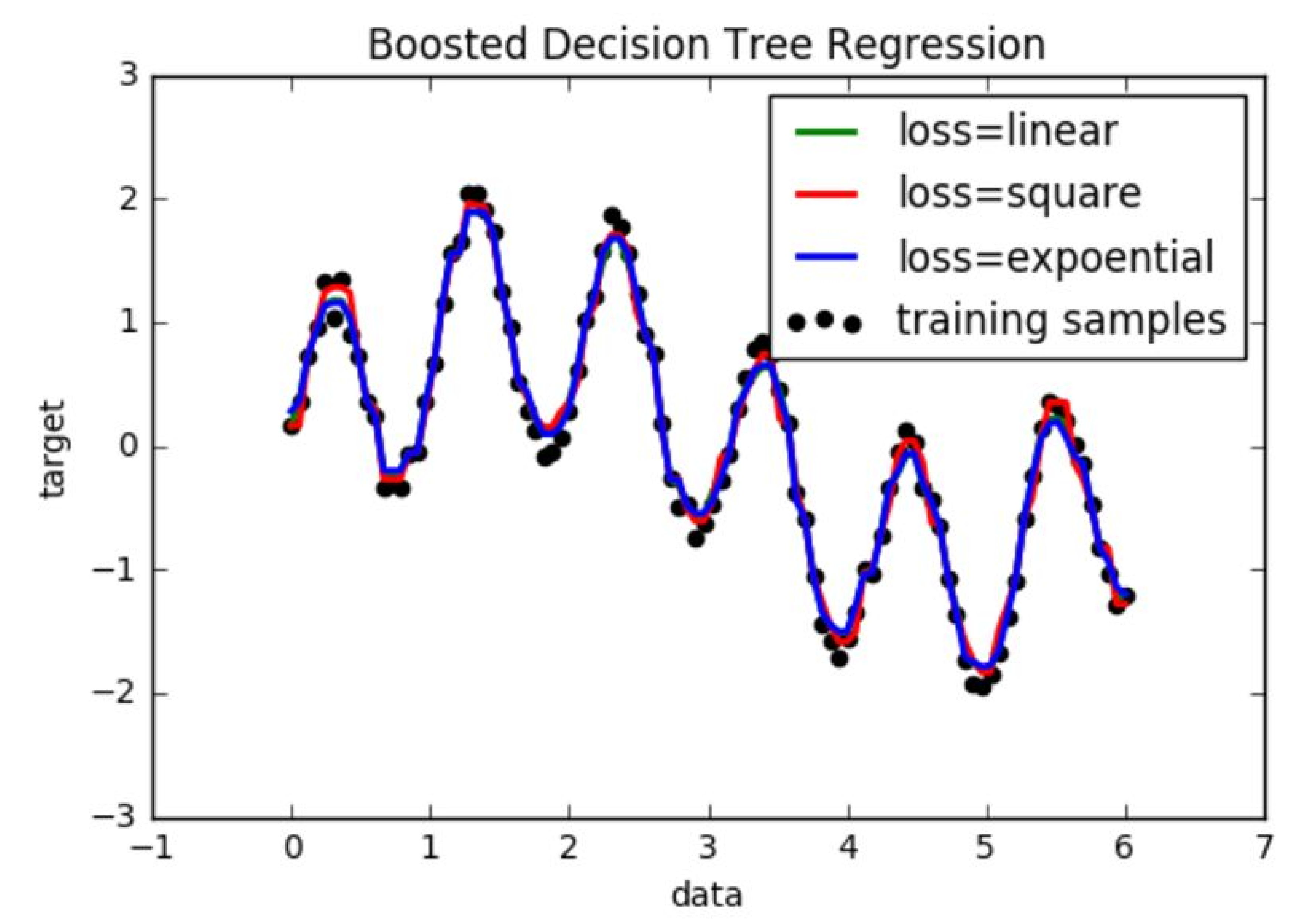
- 5
- 10
- 20

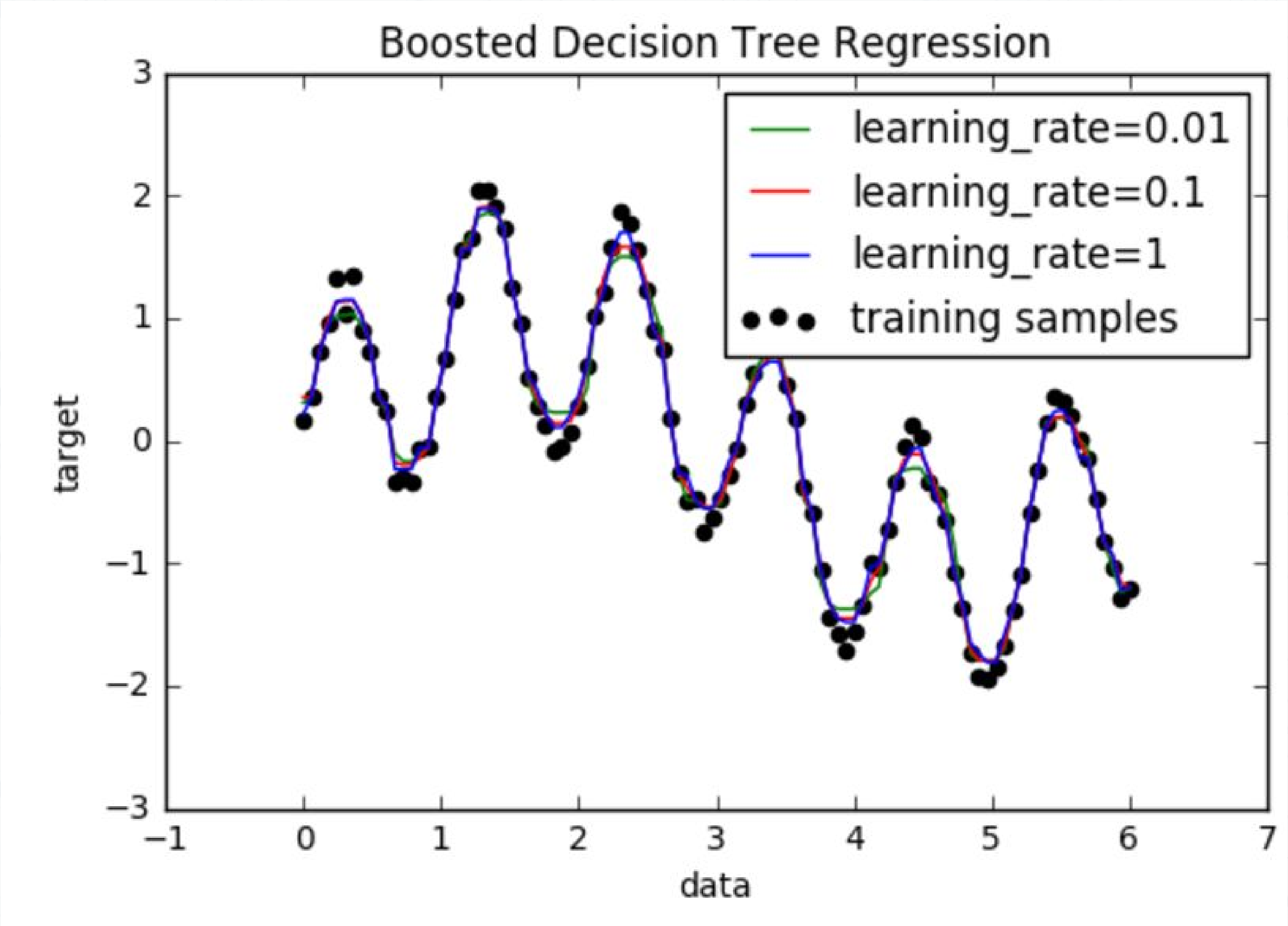
### learning_rate

- 0.01
- 0.1
- 1


Boosted Decision Tree Regression
n_estimators=300,max_depth=5
n_estimators=300,max_depth=10
n_estimators=300,max_depth=20
training samples
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6
7


Boosted Decision Tree Regression
n_estimators=50
n_estimators=100
n_estimators=300
training samples
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6
7


Boosted Decision Tree Regression
3
2
1
0
−1
−2
−3
target
−1
0
1
2
3
4
5
6
7
data
loss=linear
loss=square
loss=expoential
training samples


Boosted Decision Tree Regression
learning_rate=0.01
learning_rate=0.1
learning_rate=1
training samples
target
data
3
2
1
0
−1
−2
−3
−1
0
1
2
3
4
5
6
7

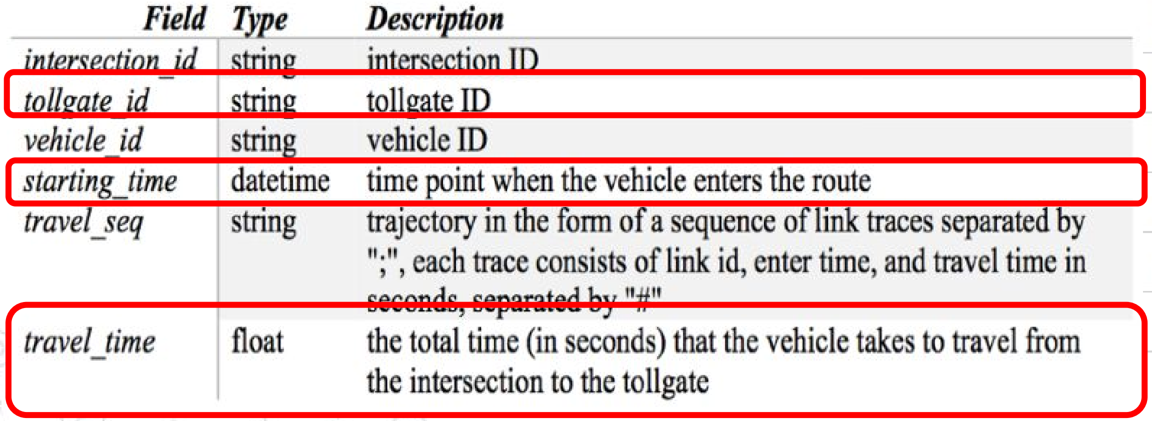
# Decision Tree Regression with Adaboost

## Feature

- Start point time
- Exit path

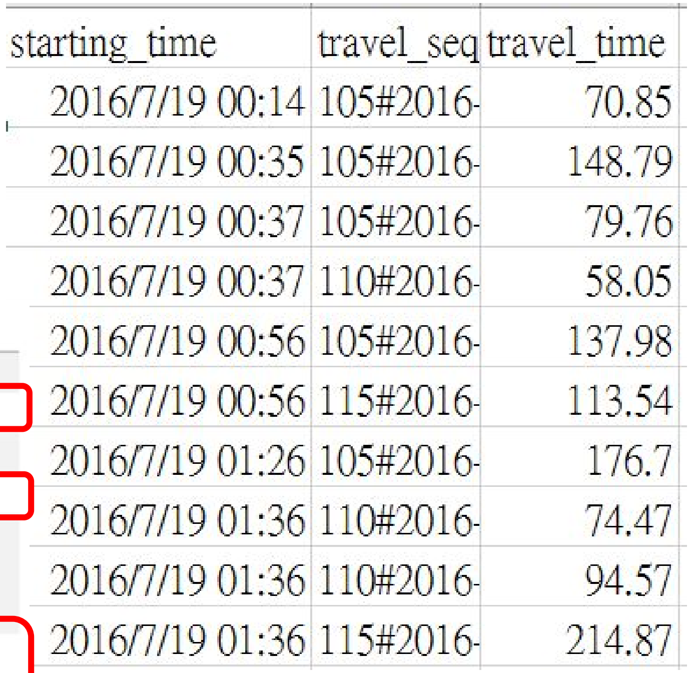
## Target

- Total travel time



| Field | Type | Description |
|---|---|---|
| *intersection_id* | string | intersection ID |
| *tollgate_id* | string | tollgate ID |
| *vehicle_id* | string | vehicle ID |
| *starting_time* | datetime | time point when the vehicle enters the route |
| *travel_seq* | string | trajectory in the form of a sequence of link traces separated by ";", each trace consists of link id, enter time, and travel time in seconds, separated by "#" |
| *travel_time* | float | the total time (in seconds) that the vehicle takes to travel from the intersection to the tollgate |

Table 5



| starting_time | travel_seq | travel_time |
|---|---|---|
| 2016/7/19 00:14 | 105#2016- | 70.85 |
| 2016/7/19 00:35 | 105#2016- | 148.79 |
| 2016/7/19 00:37 | 105#2016- | 79.76 |
| 2016/7/19 00:37 | 110#2016- | 58.05 |
| 2016/7/19 00:56 | 105#2016- | 137.98 |
| 2016/7/19 00:56 | 115#2016- | 113.54 |
| 2016/7/19 01:26 | 105#2016- | 176.7 |
| 2016/7/19 01:36 | 110#2016- | 74.47 |
| 2016/7/19 01:36 | 110#2016- | 94.57 |
| 2016/7/19 01:36 | 115#2016- | 214.87 |

# Result

**max_depth=5**

- 0.2045

**max_depth=10**

- 0.2114

**max_depth=20**

- 0.2096

Although the more depth the tree is, the better the result we got when learning, it might be overfitting.

## Get Model 2

**Use previous model to train the second model (decision tree with adaboost).**

◎ Travel time of every 30 minutes

◎ Travel time of every hour

# Decision Tree Regression with Adaboost

## Feature

- Start point time
- Exit path
- Average travel time of last 30 minutes
- Average travel time of last one hour

## Target

- Total travel time

# Result

**max_depth=5; max_depth=5**

- 0.2058

**max_depth=5; max_depth=50**

- 0.2049

**max_depth=20; max_depth=20**

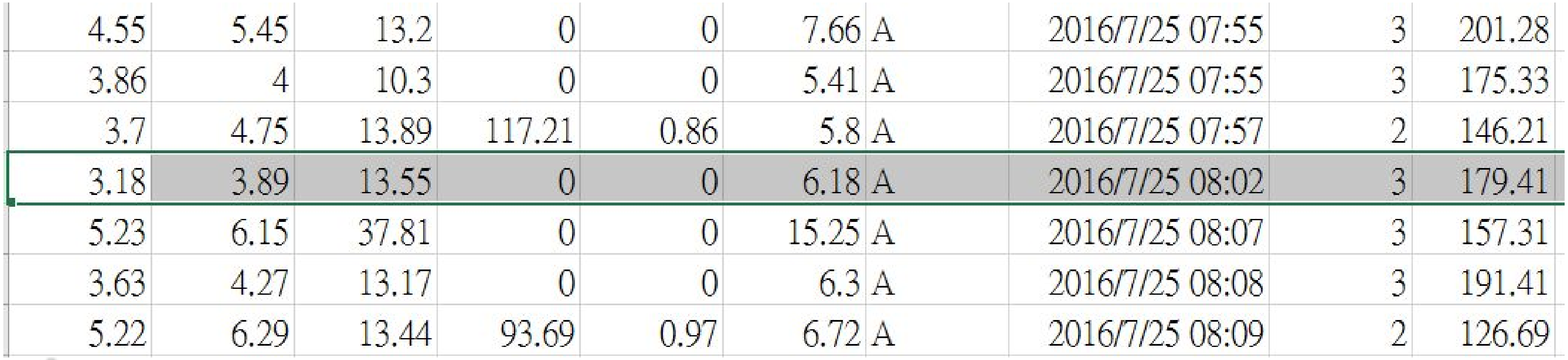
- 0.2045

According to the previous result, we thought that if using max_depth=5, the result might be the best. However, the real result does not match what we thought.

# Problem

**Noise correction is not complete.**

There exist null values, but the total travel time is inside the range of 25%-75%.



| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 4.55 | 5.45 | 13.2 | 0 | 0 | 7.66 | A | 2016/7/25 07:55 | 3 | 201.28 |
| 3.86 | 4 | 10.3 | 0 | 0 | 5.41 | A | 2016/7/25 07:55 | 3 | 175.33 |
| 3.7 | 4.75 | 13.89 | 117.21 | 0.86 | 5.8 | A | 2016/7/25 07:57 | 2 | 146.21 |
| 3.18 | 3.89 | 13.55 | 0 | 0 | 6.18 | A | 2016/7/25 08:02 | 3 | 179.41 |
| 5.23 | 6.15 | 37.81 | 0 | 0 | 15.25 | A | 2016/7/25 08:07 | 3 | 157.31 |
| 3.63 | 4.27 | 13.17 | 0 | 0 | 6.3 | A | 2016/7/25 08:08 | 3 | 191.41 |
| 5.22 | 6.29 | 13.44 | 93.69 | 0.97 | 6.72 | A | 2016/7/25 08:09 | 2 | 126.69 |

# Thanks!