Prediction and Estimation Model for the Taxi' driver in the Airport area

TEAM MEMBERS:
NGHI LU
MINGYU LYU
ZIYI ZHONG
HONG TANG
YA CHEN TSAI

ECE 143: TEAM 6

References

- https://www.flightera.net/en/airport/Shanghai/ZSPD/departure/2018-04-01%20%2008_00
- https://www.google.com/maps
- http://www.google.com
- https://machinelearningmastery.com/how-to-configure-k-fold-cross-validation/

Purpose of Making This Project

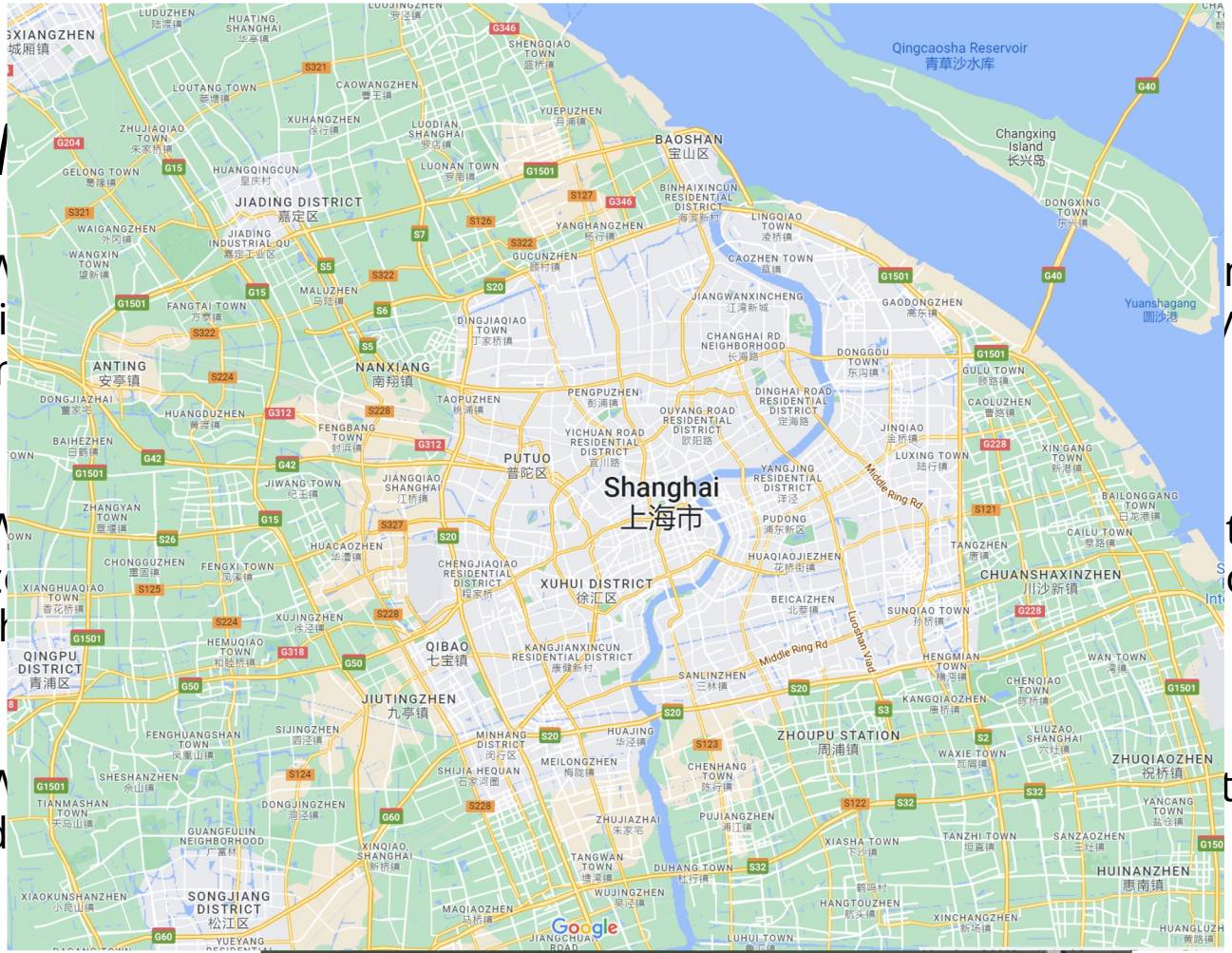
The purpose of this project is to help out taxi' drivers (in this case we will use the Data set from Shanghai government in 2018 and flights data manually imported from flight tracking website) to decide whether they should stay at the airport waiting for the passengers or go to Downtown to wait for other passengers based on the projected wait time and the cost of a operating consumption.

Overview

- Process the data that we have collected and written functions that could work with the ideas that we came up with
- Built a prediction model for taxi drivers based on the waiting time data that we have collected.
- Built the recommender model to decide whether the taxi’s drivers should wait or leave the Airport.
- Data visualization.

Hov

- W
  ti
  ir
- W
  c
  th
- W
  d



	0	1
0	0004	84
1	0009	88
2	0014	86
3	0019	92
4	0024	95
...	...	...
283	2339	24
284	2344	24
285	2349	21
286	2354	6
287	2359	0

288 rows × 2 columns

Clear

execu

Wait time of each cab

	0	1	2	3
12708	0	0	2.0	9.0
14915	0	0	0.0	43.0
14520	0	0	1.0	12.0
26760	20	56	NaN	NaN
18785	19	24	21.0	47.0
...	...	...	...	...
11853	23	26	23.0	26.0
20709	23	28	23.0	28.0
21486	23	50	23.0	51.0
17912	23	53	23.0	53.0
20166	23	54	NaN	NaN

1599 rows × 4 columns

Columns: ID of each taxi cab

0: entry hour

1: entry minute(s)

2: exit hour

3: exit minute(s)

NaN needed to be filtered out

Final wait time data frame

After filtering out the unwanted data values, we converted the values into integer to easy to work with.

	0	1	2	3
12708	0	0	2	9
14915	0	0	0	43
14520	0	0	1	12
18785	19	24	21	47
27981	18	19	20	44
...	...	...	...	...
14931	20	18	23	58
16881	20	23	20	31
20650	21	27	21	40
22063	21	31	21	40
23333	23	11	23	34

1095 rows × 4 columns

How it works?

After creating a dataframe, we created the following variables:

x1: each car’s entry time to the airport parking lot.

x2: total cabs in the parking lot at a specific time slot

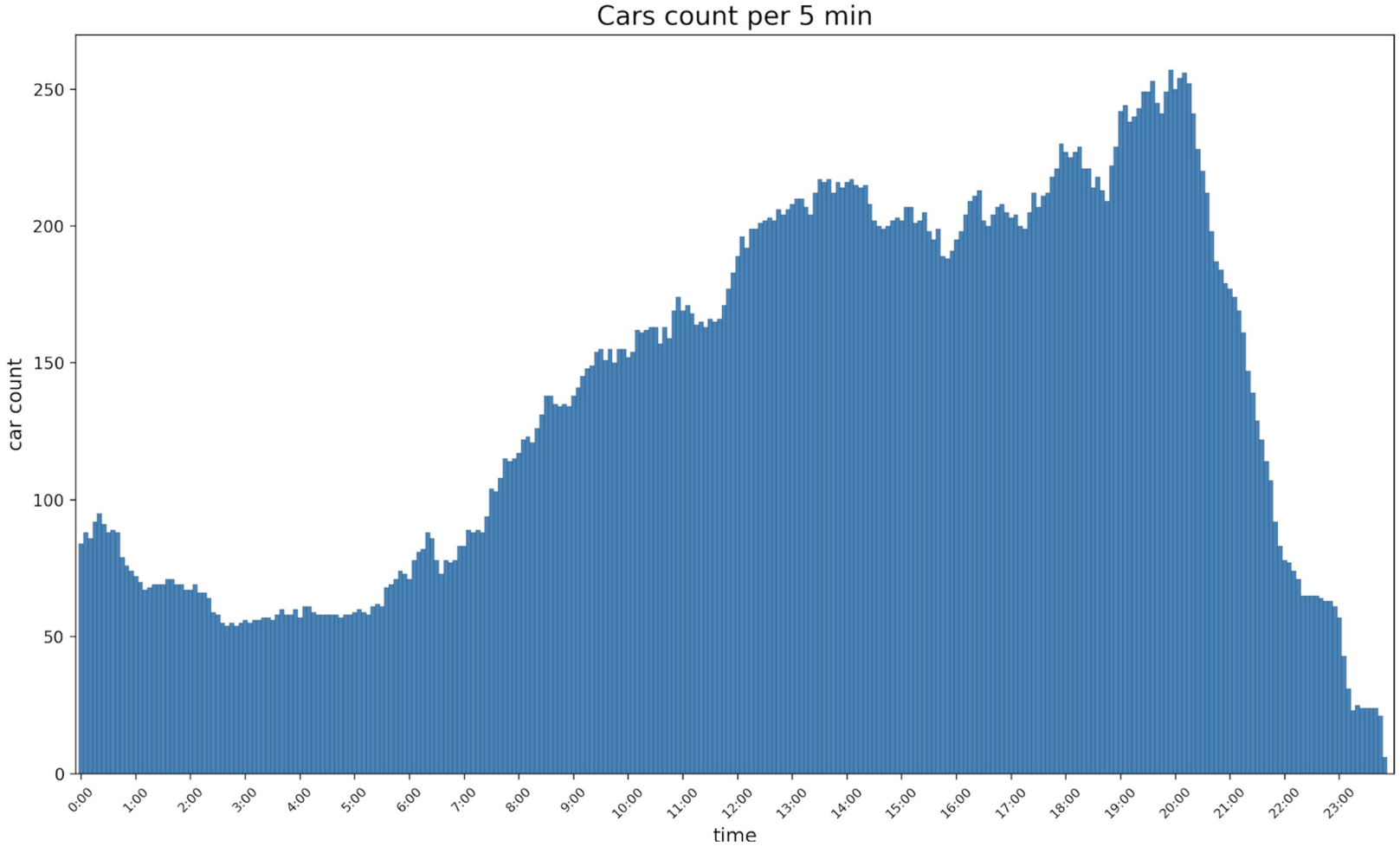
x3: total flights in the specific time slot

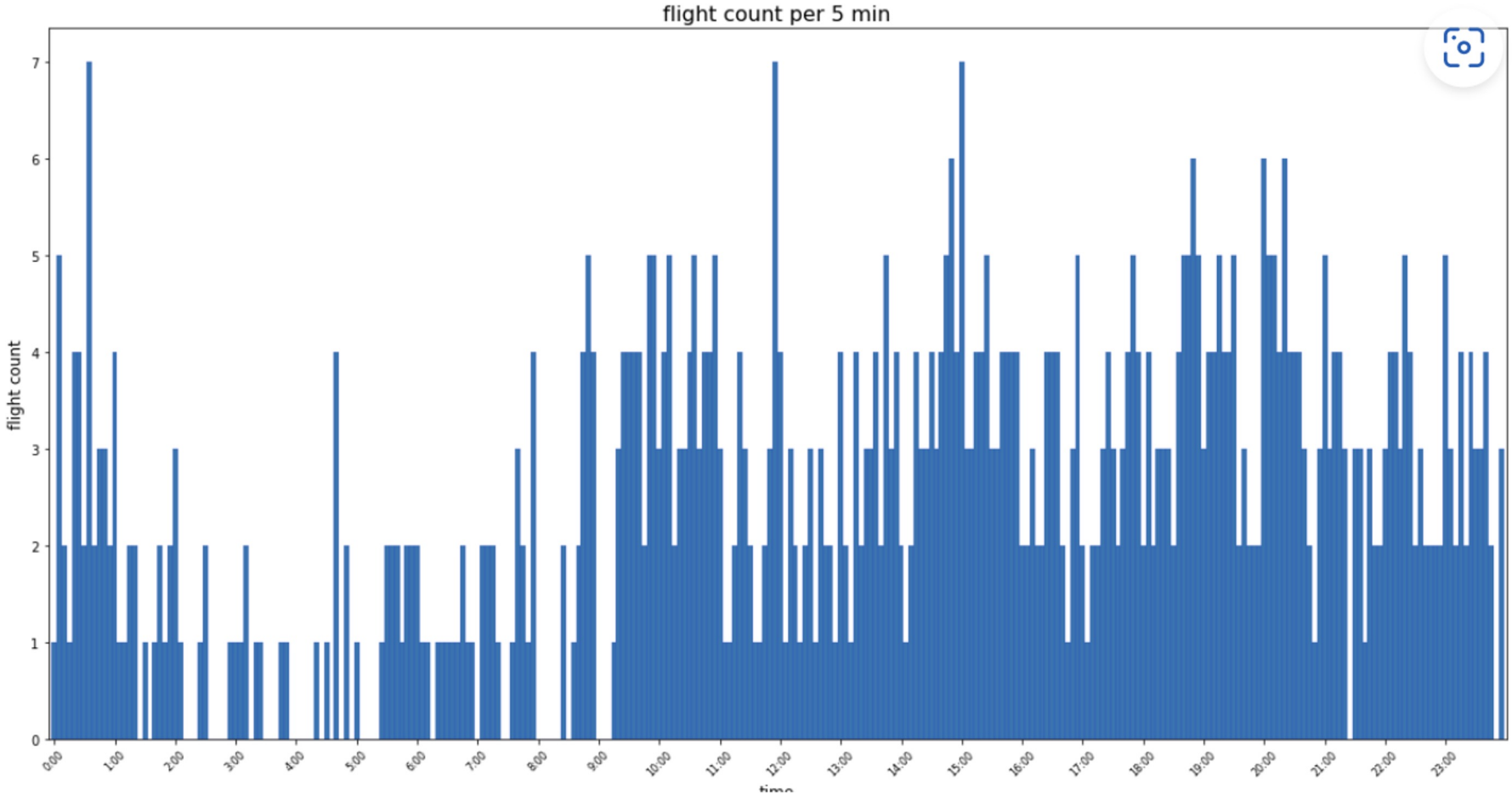
y: the average wait time in that time slot

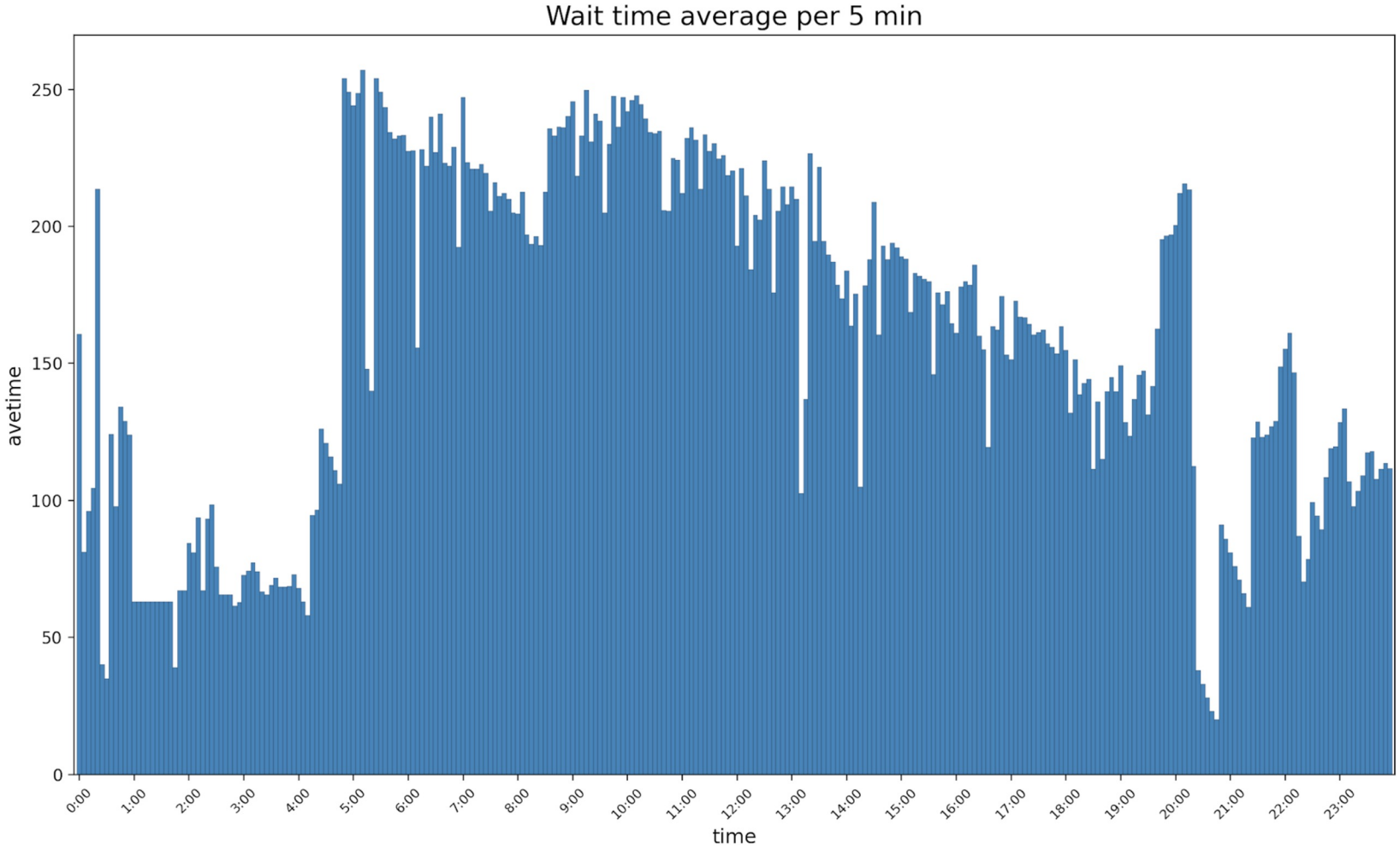
Using all the x(s) and y(s) values we need to train our model.

```
[[0, 84, 22], [0, 84, 22], [0, 84, 22], [232, 243, 34], [219, 229, 27], [243, 252, 40], [0, 84, 22], [125, 163, 33], [145, 196, 25], [0, 84, 22]]
```

```
[160.75, 160.75, 160.75, 145.7273, 138.6364, 213.375, 160.75, 290.0, 221.1429, 160.75]
```


Cars count per 5 min
car count
0
50
100
150
200
250
time
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00


flight count per 5 min
flight count
0
1
2
3
4
5
6
7
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00


Wait time average per 5 min
avetime
0
50
100
150
200
250
time
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00

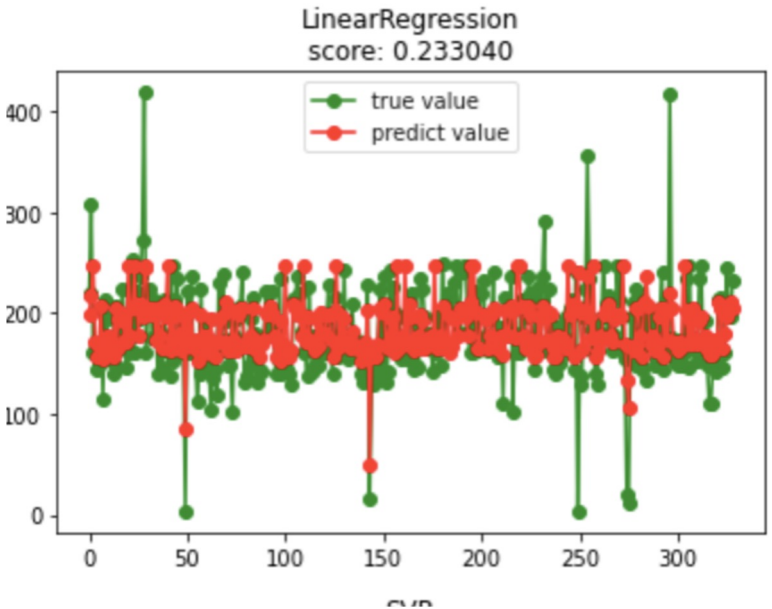
Models training' graph

- Using the third party module from Python (sklearn, mlt), we plotted a couple graphs with different results.

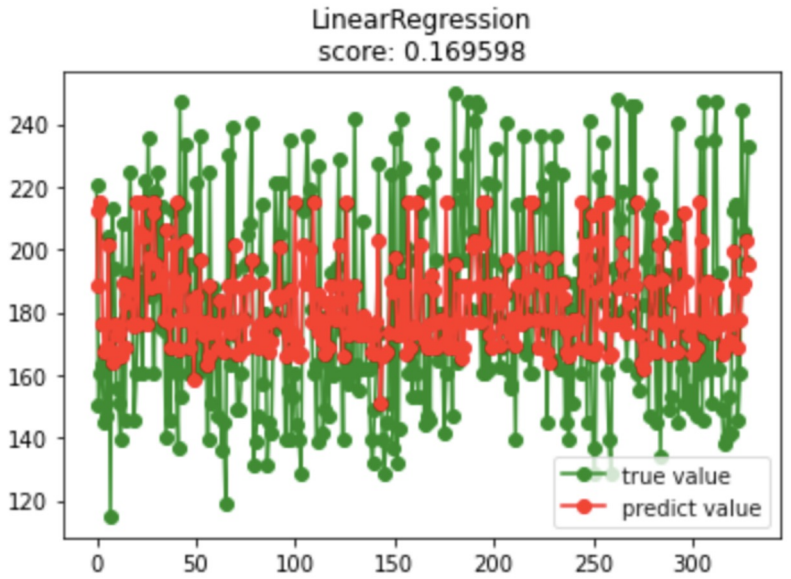
- After removing the outliers, we could see that the results were more accurate than before

- By experimenting, we compared and selected the suitable models.

Before



After

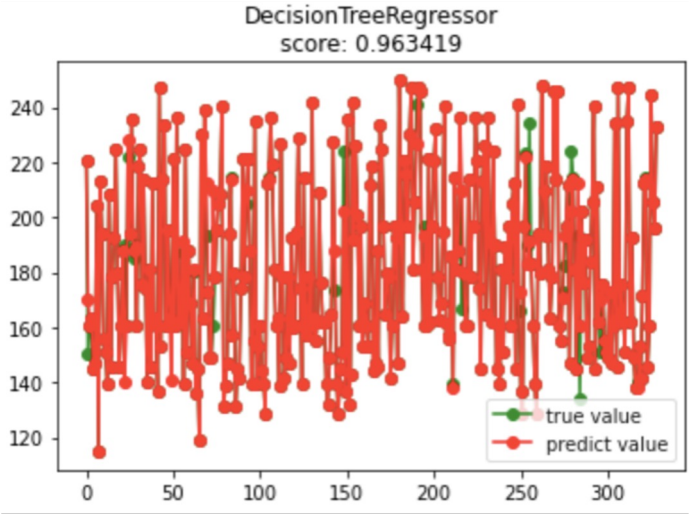


Diagrams showing training result before we process the data(get rid of all the outliers)

From the result from Tree model and Linear Regression, wse can conclude that our model is non-linear

Before

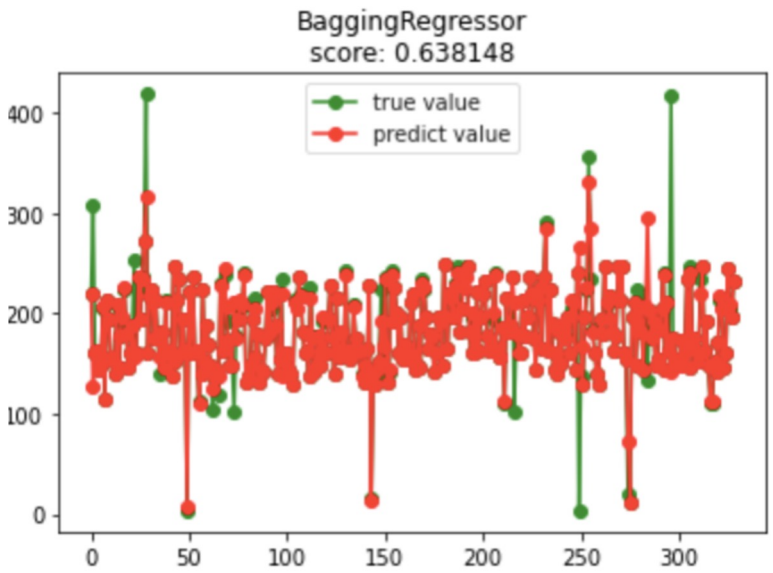
After



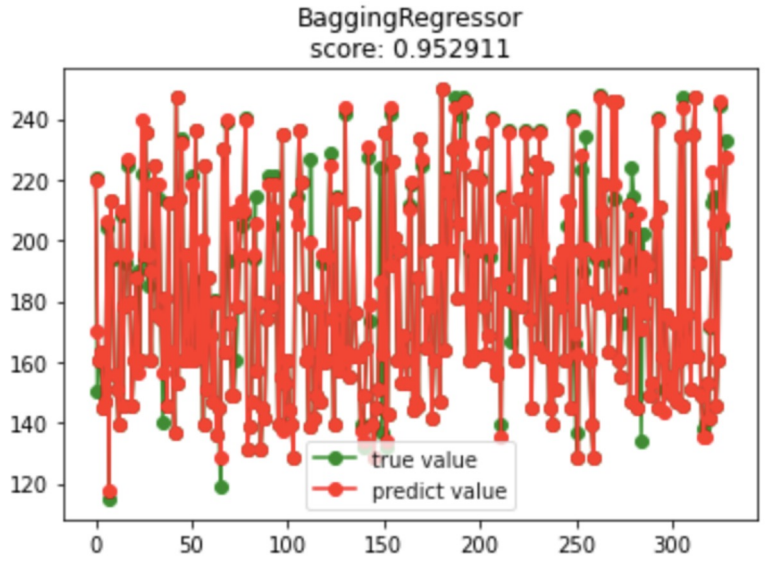
Diagrams showing training result before we process the data(get rid of all the outliers)

Ensemble method includes two averaging algorithms based on randomized decision tree, which is why Ensemble works well in non-linear data

Before

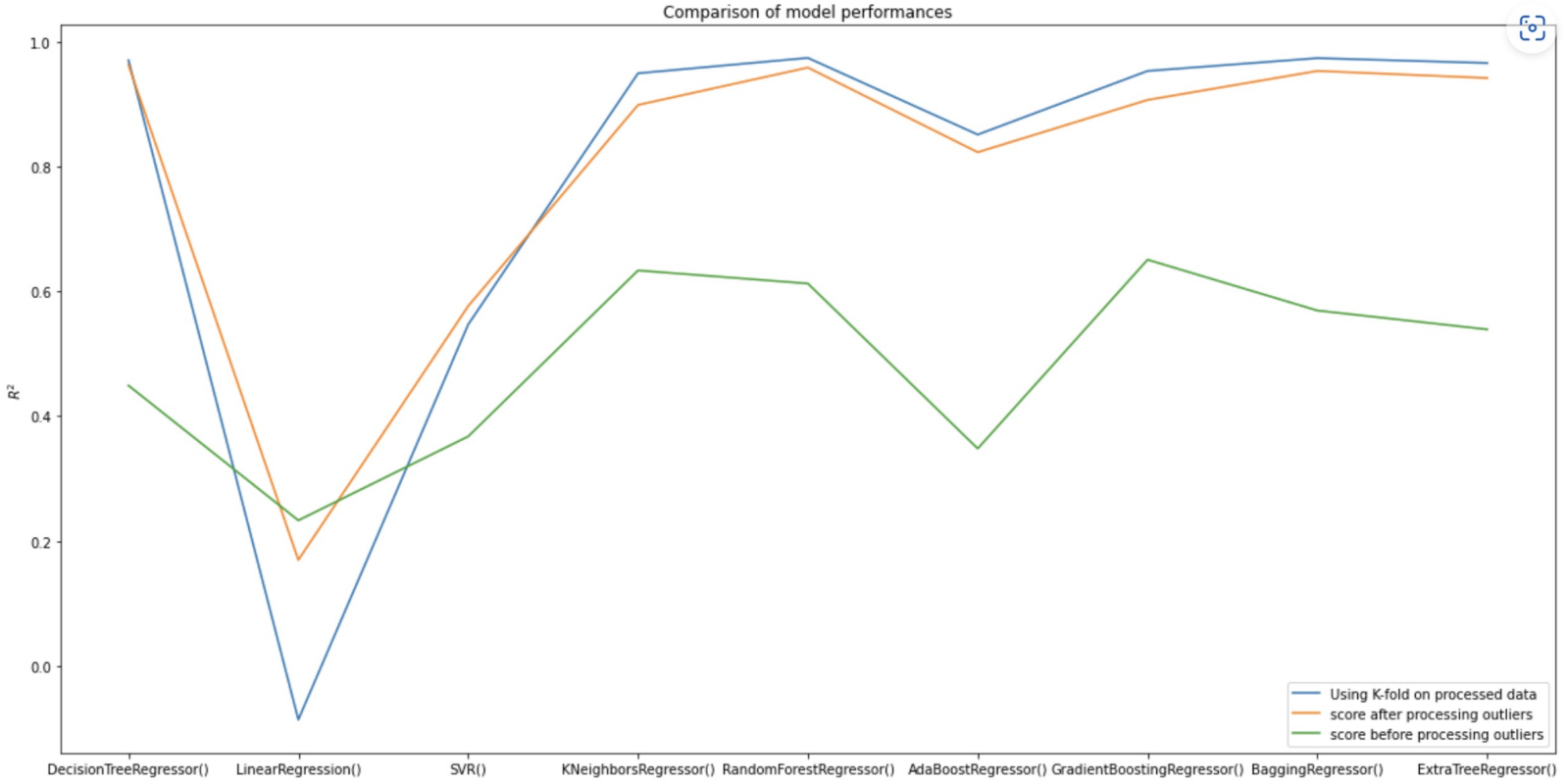


After



Diagrams showing training result before we process the data(get rid of all the outliers)

1. Processing outliers can generally improve the performance of ML models we select (except Linear Regression, probably some outliers improve data linearity)
2. K-fold can improve performances on Ensemble and Tree methods



Profit comparing

Assuming Wait time in Taxi Car pool is T_w, and the average Transit time between Airport and Downtown area is T_r. In this project, based on Map Search, we define the T_r = 60 min

We Define Operation Cost per minutes as c, average Loaded Rate per 5 min as λ, and the average Profit per hour in Shanghai as P_a.

We are able to build an Expectation Model as following:

Expected Revenue from Waiting in the airport: $P_w = T_r * P_a - c * T_r$

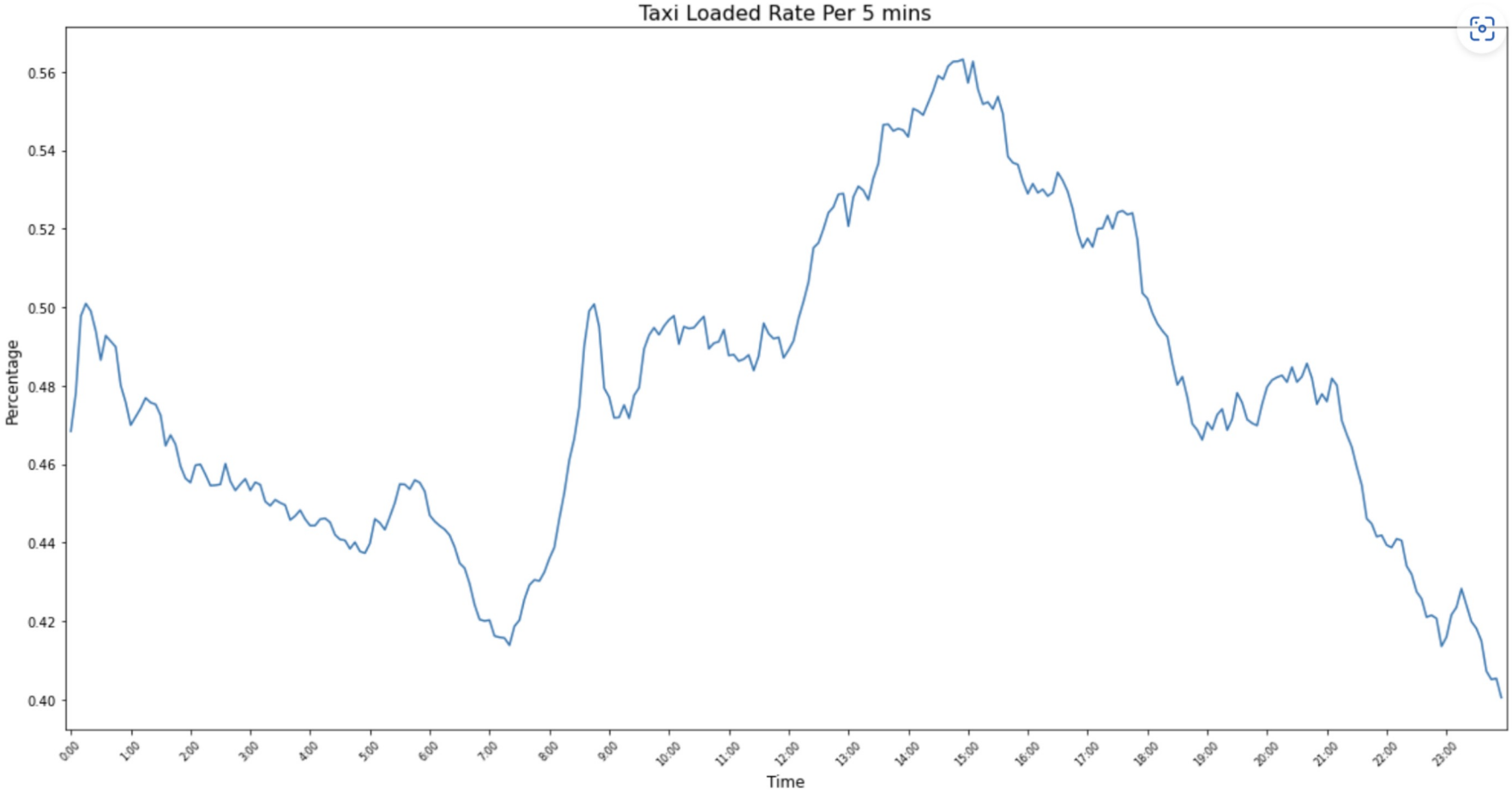
Expected Revenue from going back to Urban Area: $P_d = T_w * P_a * \lambda - c * T_w - c * T_r$

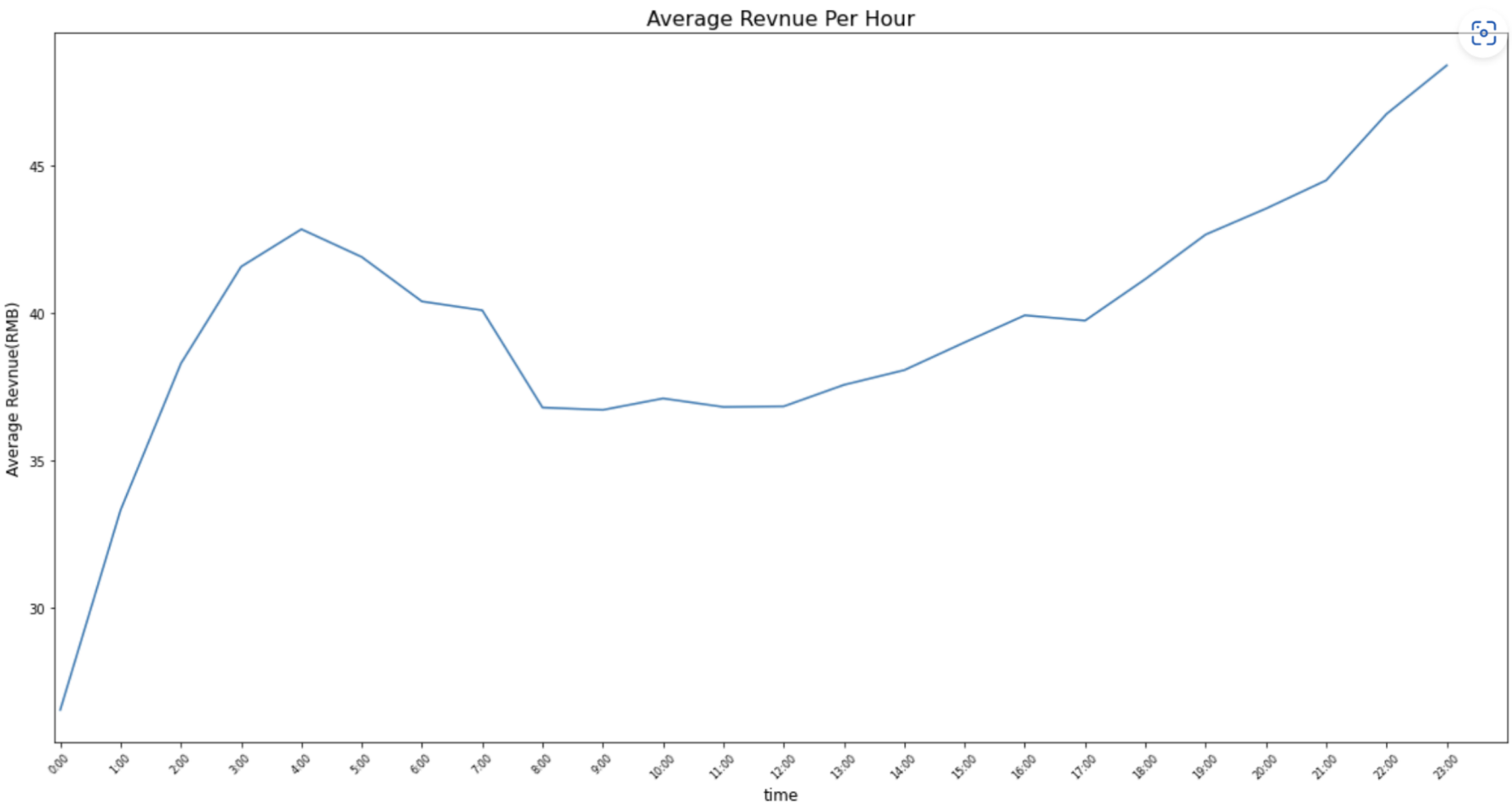
which we can simplify as the following:

Expected Revenue from Waiting in the airport: $P_w = T_r * P_a$

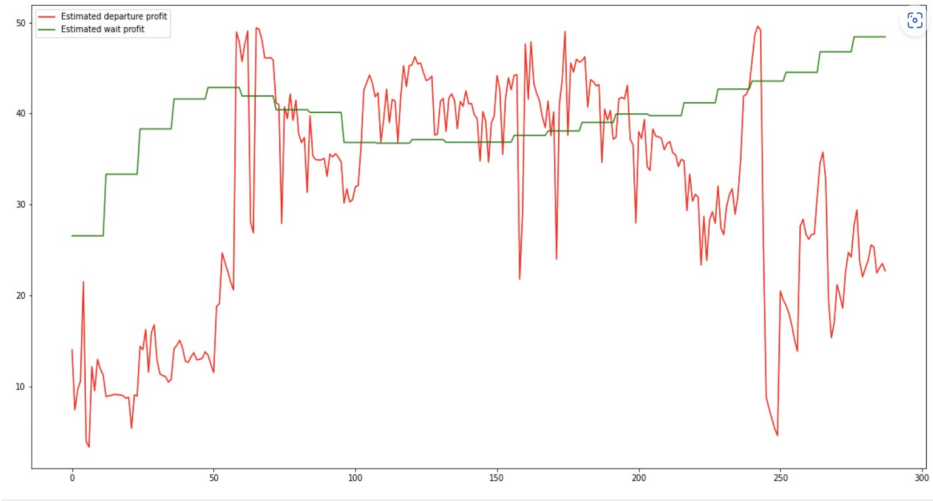
Expected Revenue from going back to Urban Area: $P_d = T_w * P_a * \lambda - c * T_w$

Assuming the operating cost is 0.12 RMB per min(electric taxi vehicle)


Taxi Loaded Rate Per 5 mins
Percentage
0.56
0.54
0.52
0.50
0.48
0.46
0.44
0.42
0.40
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Time


Average Revnue Per Hour
Average Revnue(RMB)
45
40
35
30
0:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
10:00
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
time

Ground truth visualization of the profit based on the whole data set



Testing data

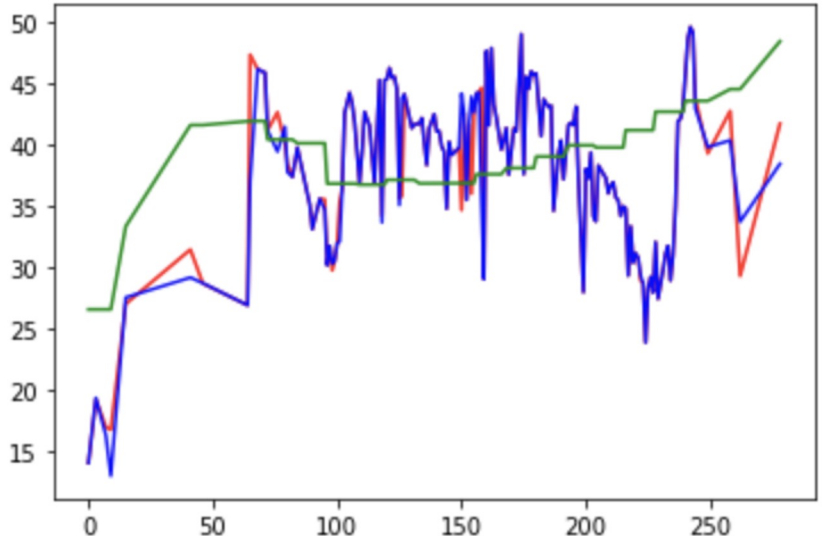
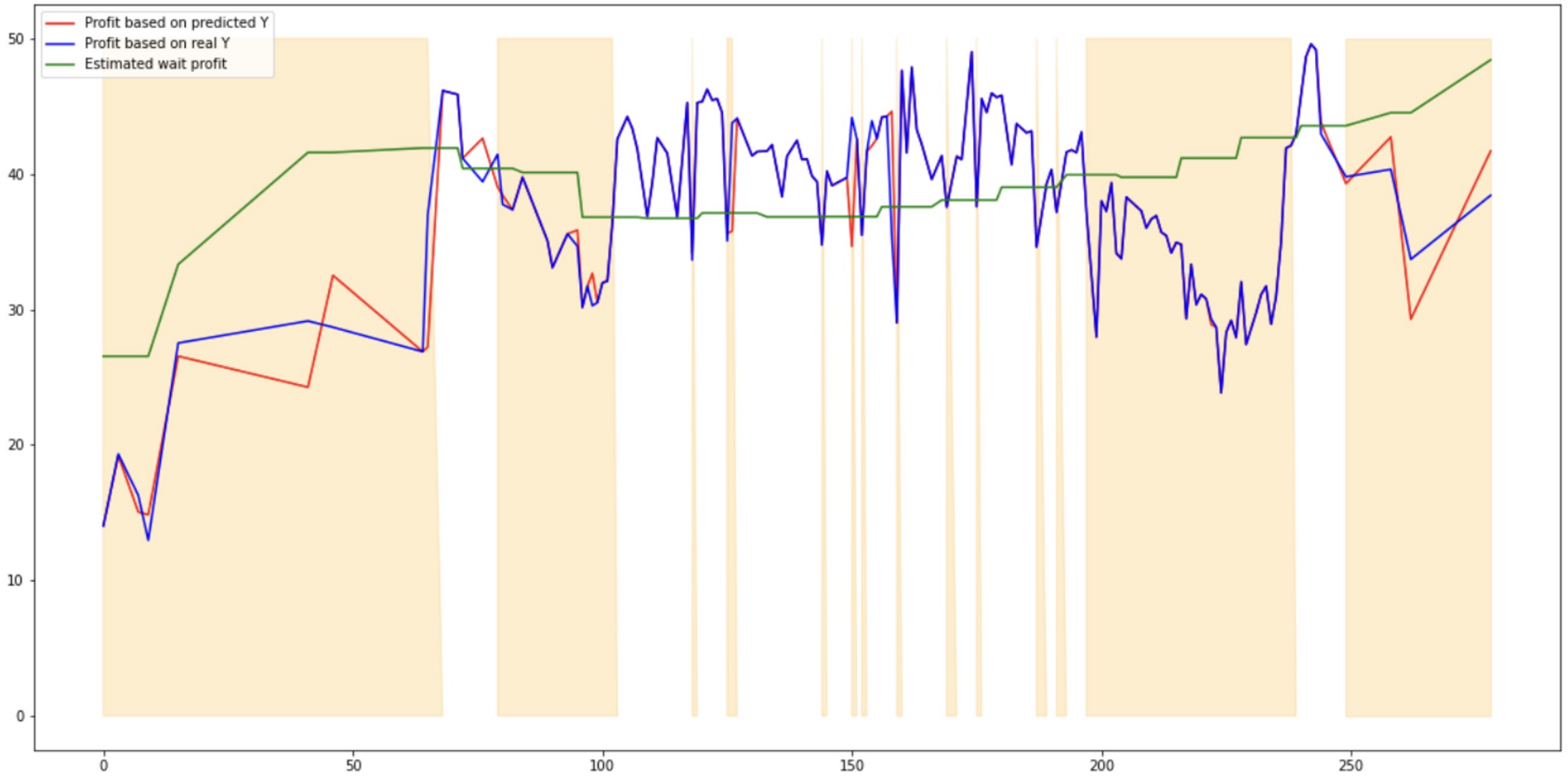


Diagram showing the profit(s) using the ExtraTreeRegressor model

Blue (depart): prediction on training model
Green (to stay): prediction on training model



Final results gave the drivers the answers to stay and pick up the passengers early in the morning and late at night

Q&A?

Thank You!