Taxi Demand Analytics And Prediction

COMP4952 Progress Report

Presenter: Yao Yuan

z5092195

Methodology

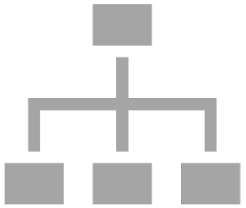
Data

Models

Evaluation

Data

Data filtering



Clustering

Models for Forecasting

LSTM Neural Network model

XGBoost model

Naïve statistical average model

Recap – what I have done in thesis B

LSTM model construct

```
def split_sequence(sequence, n_steps):
```

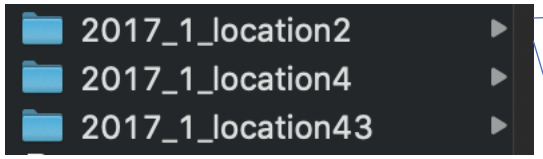
- Input data

NOTES:
The reason I did not put the location into the input data is that the data from location A is useless memory in LSTM to location B which will decrease the accuracy in location B.

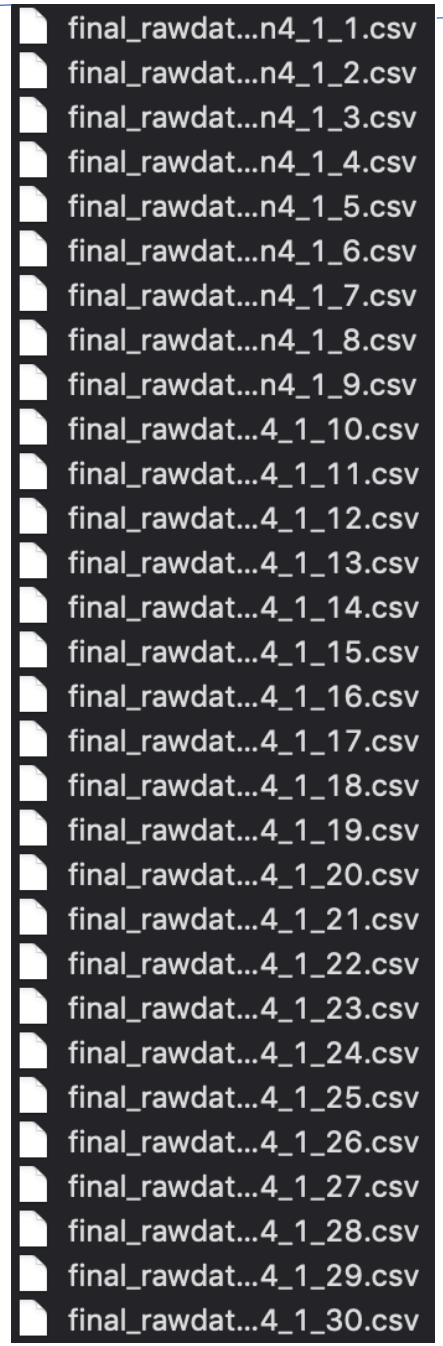
Hours	count
0	136
1	145
2	190
3	189
4	128
5	54
6	36
7	28
8	36
9	34
10	48
11	38
12	31
13	43
14	38
15	33
16	29
17	28
18	34
19	20
20	30
21	32
22	22
23	35

x1	x2	x3	y
136	145	190	189
145	190	189	128
190	189	128	54
189	128	54	36
128	54	36	28
54	36	28	36

n_steps = 3



Not Efficient for Many locations



Hours	count
0	136
1	145
2	190
3	189
4	128
5	54
6	36
7	28
8	36
9	34
10	48
11	38
12	31
13	43
14	38
15	33
16	29
17	28
18	34
19	20
20	30
21	32
22	22
23	35

XGBoost data extraction

dayofweek	PULocationID	Hours	count
6	1	3	1
6	1	4	1
6	1	5	2
6	1	6	2
6	1	7	2
6	1	10	3
6	1	11	2
6	1	12	3
6	1	13	3
6	1	14	3
6	1	15	3
6	1	16	6
6	1	17	4
6	1	18	3
6	1	23	5
0	1	3	2
0	1	4	1
0	1	5	1
0	1	6	2
0	1	7	4
0	1	10	1
0	1	11	1
0	1	15	5
0	1	16	5

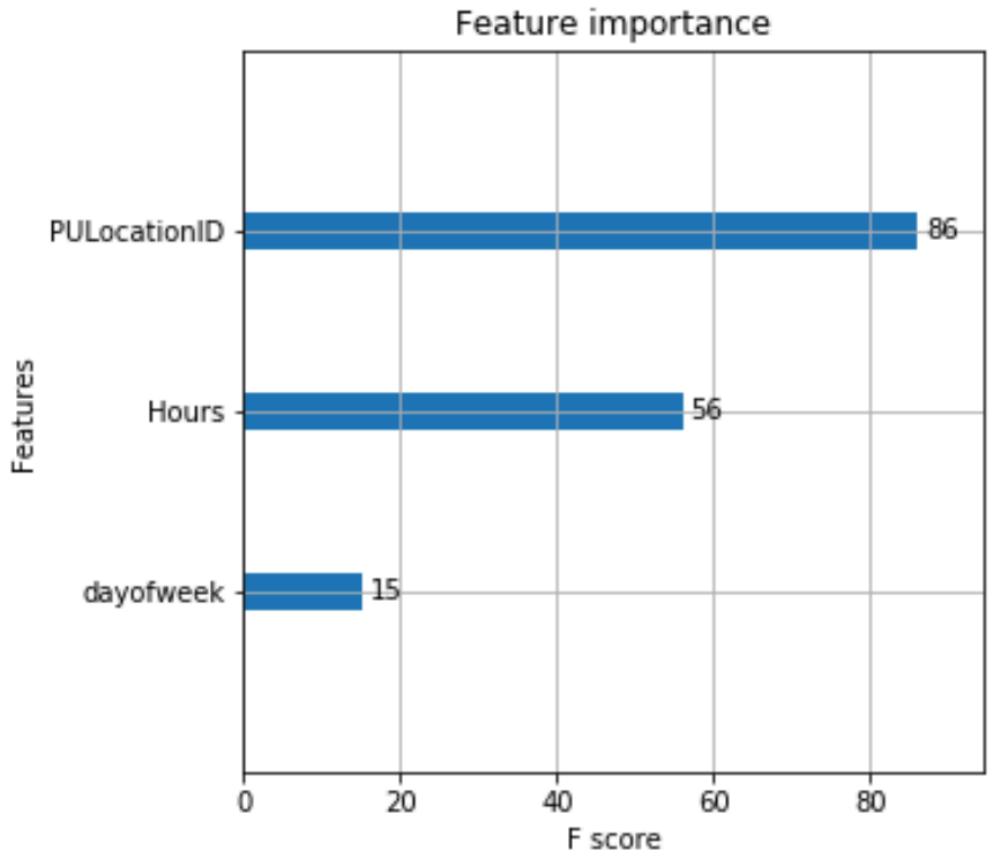
XGBoost model is a gradient boosted decision tree model which is a popular machine learning algorithm.

- (x1)Dayofweek: encode Sunday to Saturday -> 0 to 6
- (x2)PULoactionID: pick up location id
- (x3)Hours: 0 to 23 hour slot
- (y)Count: pick up count

- Total number of record: 89514

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 123)
```

XGBoost evaluation



Total RMSE: 54.84

Model	R^2	RMSE
Multiple linear regression	0.76	125.56
Ridge	0.75	125.56
Randomforest	0.97	40.06
XGBoost	0.98	35.01
Ensemble	0.97	42.95

They have two more features: Rainfall and temperature

https://nycdatascience.com/blog/student-works/predict-new-york-city-taxi-demand/

XGBoost evaluation – test case

dayofweek	PULocationID	Hours	count
1	4	16	20
1	4	16	11
1	4	16	14
1	4	16	19
1	4	16	15

Location 4;
All Mondays;
At 4pm

Average value of these data: 15.8

XGBoost model

Predicted value is 4.751063

XGBoost is worse than simply averaging the values

The reasons caused this:

- Data not enough (only one month feeding)
- Features not enough

Main Tasks to do in the last stage

LSTM:

- Efficient prediction at all locations

XGBoost:

- More features needed

dayofweek	PULocationID	Hours	time1	time2	time3	count
6	1	3	0	0	0	1
6	1	4	0	0	1	1
6	1	5	0	1	1	2
6	1	6	1	1	2	2
6	1	7	1	2	2	2
6	1	10	2	0	0	3
6	1	11	0	0	3	2
6	1	12	0	3	2	3
6	1	13	3	2	3	3
6	1	14	2	3	3	3
6	1	15	3	3	3	3
6	1	16	3	3	3	6
6	1	17	3	3	6	4
6	1	18	3	6	4	3
6	1	23	0	0	0	5
0	1	3	0	0	0	2
0	1	4	0	0	2	1
0	1	5	0	2	1	1
0	1	6	2	1	1	2
0	1	7	1	1	2	4
0	1	10	4	0	0	1
0	1	11	0	0	1	1

dayofweek	Hours	time1	time2	time3	PULocationID	count
1.0	0.005671077504725900	0.0	0.0	0.0	0.0037735849056603800	0.00046533271288971600
1.0	0.007561436672967860	0.00046533271288971600	0.0	0.00046533271288971600	0.0037735849056603800	0.00046533271288971600
1.0	0.00945179584120983	0.00046533271288971600	0.00046533271288971600	0.00046533271288971600	0.0037735849056603800	0.00093066542577943200
1.0	0.011342155009451800	0.00093066542577943200	0.00046533271288971600	0.00093066542577943200	0.0037735849056603800	0.00093066542577943200
1.0	0.013232514177693800	0.00093066542577943200	0.00093066542577943200	0.00093066542577943200	0.0037735849056603800	0.00093066542577943200
1.0	0.01890359168241970	0.0	0.0	0.0	0.0037735849056603800	0.00139599813866915000
1.0	0.020793950850661600	0.00139599813866915000	0.0	0.00139599813866915000	0.0037735849056603800	0.00093066542577943200
1.0	0.02268431001890360	0.00093066542577943200	0.00139599813866915000	0.00093066542577943200	0.0037735849056603800	0.00139599813866915000
1.0	0.024574669187145600	0.00139599813866915000	0.00093066542577943200	0.00139599813866915000	0.0037735849056603800	0.00139599813866915000
1.0	0.026465028355387500	0.00139599813866915000	0.00139599813866915000	0.00139599813866915000	0.0037735849056603800	0.00139599813866915000
1.0	0.02855538752362950	0.00139599813866915000	0.00139599813866915000	0.00139599813866915000	0.0037735849056603800	0.00139599813866915000
1.0	0.030245746691871500	0.00139599813866915000	0.00139599813866915000	0.00139599813866915000	0.0037735849056603800	0.00279199627738300
1.0	0.032136105860113400	0.00279199627738300	0.00139599813866915000	0.00279199627738300	0.0037735849056603800	0.00186133085155886000
1.0	0.034026465028355400	0.00186133085155886000	0.00279199627738300	0.00186133085155886000	0.0037735849056603800	0.00139599813866915000
1.0	0.043478260869565200	0.0	0.0	0.0	0.0037735849056603800	0.00232666356444858000
0.0	0.005671077504725900	0.0	0.0	0.0	0.0037735849056603800	0.00093066542577943200
0.0	0.007561436672967860	0.00093066542577943200	0.0	0.00093066542577943200	0.0037735849056603800	0.00046533271288971600
0.0	0.00945179584120983	0.00046533271288971600	0.00093066542577943200	0.00046533271288971600	0.0037735849056603800	0.00046533271288971600
0.0	0.011342155009451800	0.00046533271288971600	0.00046533271288971600	0.00046533271288971600	0.0037735849056603800	0.00093066542577943200
0.0	0.013232514177693800	0.00093066542577943200	0.00046533271288971600	0.00093066542577943200	0.0037735849056603800	0.00186133085155886000
0.0	0.01890359168241970	0.0	0.0	0.0	0.0037735849056603800	0.00046533271288971600
0.0	0.020793950850661600	0.00046533271288971600	0.0	0.00046533271288971600	0.0037735849056603800	0.00046533271288971600

To solve this problem

- Input features:
  - Dayofweek: encoded into 0 to 6
  - PULocationID
  - Hours: the corresponding hour slot
  - Time1, Time2, Time3: the previous hours of the current hour
  - Count: the number of pickups in that hour slot

- Number of data:
  - 89541

Results evaluation

- RMSE

```
RMSE = math.sqrt(mean_squared_error(y_test[:], results[:,0]))
```

- MAE

```
MAE = statistics.mean(abs(y_test[:] - results[:,0]))
```

- MPE

```
MPE = np.sum(abs(y_test[:] - results[:,0]))/np.sum(y_test[:])
```

More data

model	# of data	RMSE	MAE	MPE
LSTM	89541	38.93	18.85	0.18
	170661	32.36	16.47	0.15
XGBoost	89541	35.75	16.48	0.15
	170661	34.80	16.00	0.14

From these results, we can see that

- More data generally have better results
- LSTM model is better than XGBoost model when having more data

Feature Importance

LSTM model	RMSE	MAE	MPE
All features	37.02	18.28	0.17
Without Hours	43.00	20.32	0.19
Without Locations	38.69	18.90	0.18
Without weekdays	38.89	21.61	0.20
Without Time1	37.44	19.57	0.18
Without Time1 & 2	37.31	17.33	0.16
Without All Times	146.54	96.70	0.90

From these results, we can see that

- Delete all times has most influences on the prediction of LSTM model (Since time connections are destroyed)
- Within Time1, Time2, Time3, the Time3 has larger effects on the results
- Hours, Locations, weekdays also affect the results
- Hours > weekdays > Locations

Feature Importance

XGBoost model	RMSE	MAE	MPE
All features	48.24	24.75	0.23
Without Hours	51.18	26.43	0.25
Without Locations	47.86	24.02	0.22
Without weekdays	48.09	24.77	0.23
Without Time1	48.12	23.89	0.22
Without Time1 & 2	46.28	23.94	0.22
Without All Times	87.57	53.46	0.50

From these results, we can see that

- Delete all times has most influences on the prediction of XGBoost model
- When without Time1 and Time2, the results get better. <- XGBoost is a gradient boosting regression model. Time1 and Time2 might affect the true prediction.
- Location and weekdays do not have large effects on the results
- Hours do have large influence on the predictions

Different parameter for LSTM cells

# LSTM cells	# parameters	RMSE	MAE	MPE	Time
25	2810	39.99	19.05	0.18	Around 20mins
50	10610	38.96	18.85	0.18	Around 45mins
100	41210	36.68	18.84	0.18	Around 60mins

From these results, we can see that

- The more LSTM cells you have, the better RMSE you will get
- The more LSTM cells you have, the longer the training process time will be taken. (for the same number of data records)

Different parameter for LSTM Train-Test ratio

Training:Test	RMSE	MAE	MPE
2:8	42.62	20.96	0.19
3:7	41.73	19.34	0.18
4:6	39.49	18.66	0.17
5:5	40.54	18.90	0.17
7:3	38.14	17.85	0.16
8:2	38.96	18.85	0.18

From these results we can see that

- The smaller this training-testing ratio is, the larger error predictions will have
- When ratio is 7:3, it has the best results

Different parameter for LSTM epochs

Epochs	RMSE	MSE	MPE	Time
100	38.93	17.99	0.17	Around 30mins
150	38.96	18.85	0.18	Around 45mins
200	37.02	18.28	0.17	Around 60mins

From these results we can see that

- The more epochs you have, the better results you will get
- However, the more epochs, the longer the training will take

Different parameter for XGBoost Train-Test ratio

Training:Test	RMSE	MAE	MPE
2:8	57.60	31.34	0.29
3:7	55.06	29.15	0.27
4:6	53.78	27.65	0.25
5:5	50.86	26.59	0.25
7:3	49.23	25.00	0.23
8:2	48.23	24.75	0.23

From these results we can see that

- The smaller this training-testing ratio is, the larger error predictions will have
- The effects are larger than LSTM model

Different parameter for XGBoost Learning-rate

Learning rate [0,1]	RMSE	MAE	MPE
0.01	49.37	29.89	0.28
0.1	48.23	24.75	0.23
0.3	47.60	23.14	0.22
0.5	48.71	24.45	0.23

From these results we can see that

- The learning rate at 0.3 has the lowest values of error

Different parameter for XGBoost Objectives

Objective	RMSE	MAE	MPE
reg:squarederror	47.60	23.14	0.22
reg:logistic	35.75	16.48	0.15
count:poisson	37.06	17.06	0.16

From these results we can see that

- When Objectives are reg:logistic, it has the best result

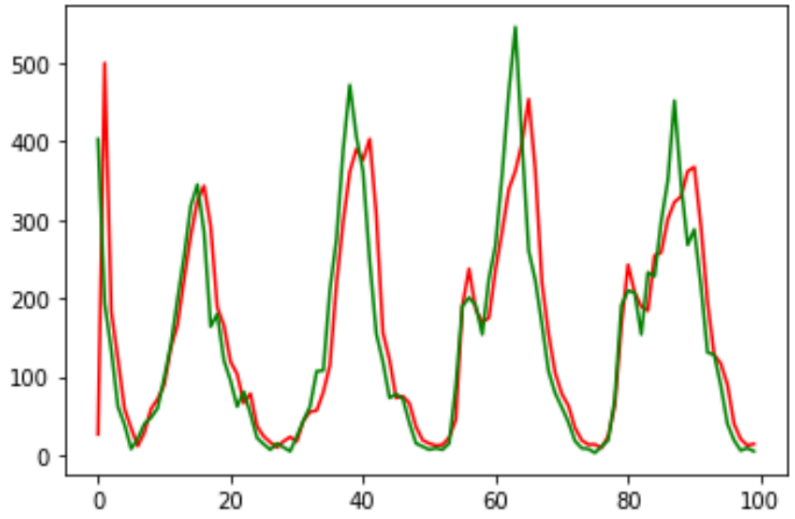
- reg:squarederror:
  - regression with squared loss
- reg:logistic
  - logistic regression
- count:poisson
  - –poisson regression for count data, output mean of poisson distribution

Different locations

- Location 43: Central park

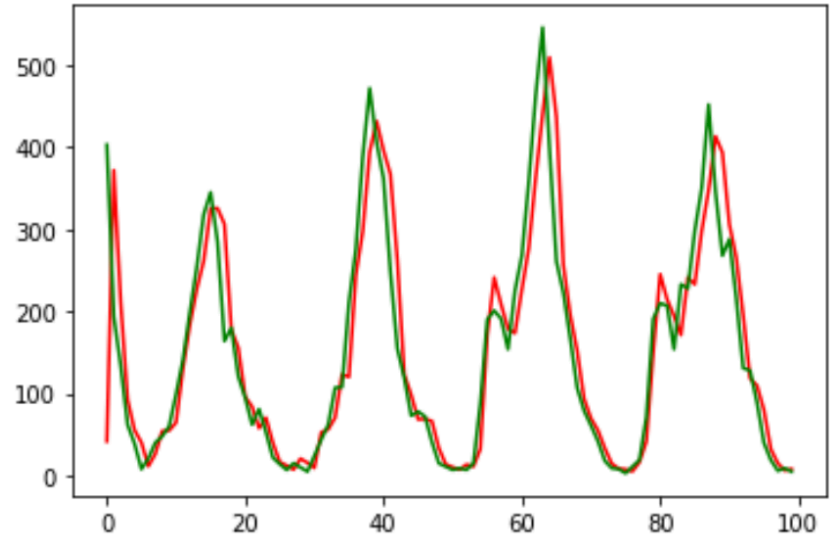
Central Park is an urban park in Manhattan, New York City, located between the Upper West Side and the Upper East Side. Central Park is the most visited urban park in the United States, with an estimated 37.5–38 million visitors annually, and one of the most filmed locations in the world

```
Test RMSE: 47.895844
Test MAE: 33.32
Test MPE: 0.20
```



LSTM

```
Test RMSE: 46.876737
Test MAE: 32.71
Test PE: 0.20
```



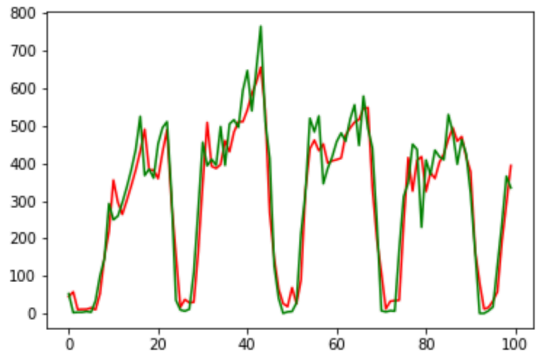
XGBoost

Different locations

- Location 132: JFK Airport
- Location 138: LaGuadia Airport

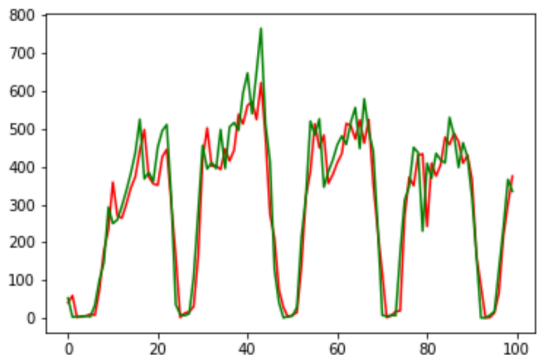
LSTM

```
Test RMSE: 77.286099
Test MAE: 55.68
Test MPE: 0.19
```



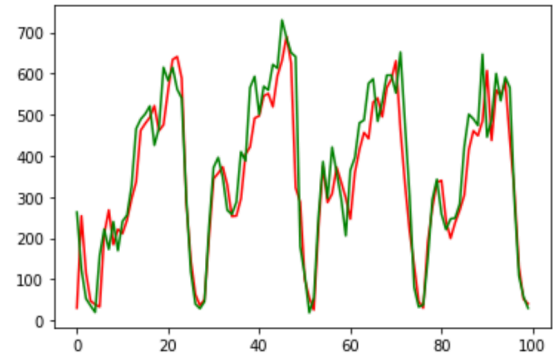
XGBoost

```
Test RMSE: 80.912750
Test MAE: 58.11
Test PE: 0.19
```

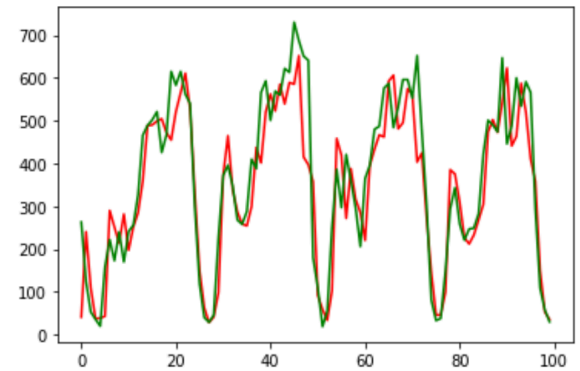


Location 132

```
Test RMSE: 73.507729
Test MAE: 55.99
Test MPE: 0.19
```




```
Test RMSE: 83.599607
Test MAE: 63.88
Test PE: 0.22
```



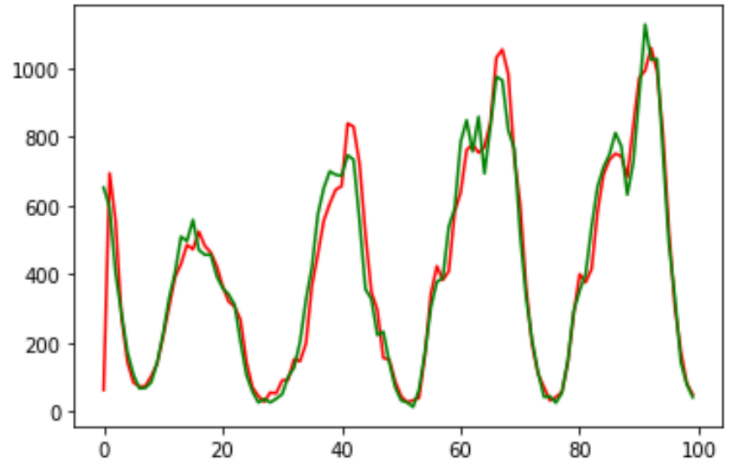
Location 138

Different locations

Midtown Manhattan is the largest central business district in the world and ranks among the most expensive pieces of real estate. **Midtown Manhattan** is the central portion of the borough of Manhattan in New York City. Midtown is home to some of the city's most iconic buildings, including the Empire State Building, Grand Central Terminal and Times Square.

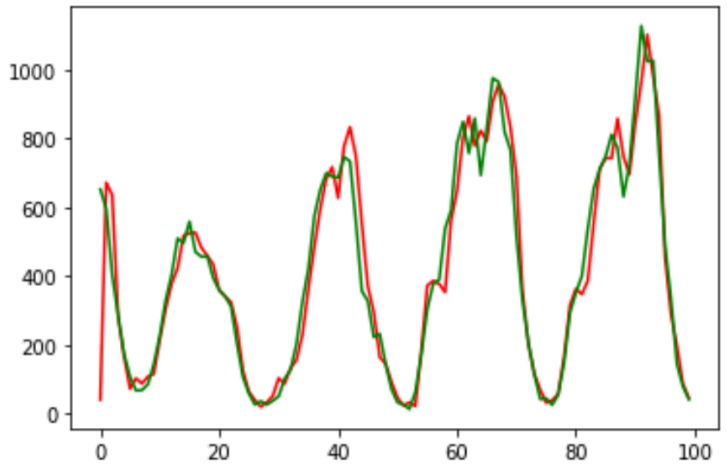
- Location 161: Midtown center

```
Test RMSE: 68.540150
Test MAE: 47.60
Test MPE: 0.10
```



LSTM model

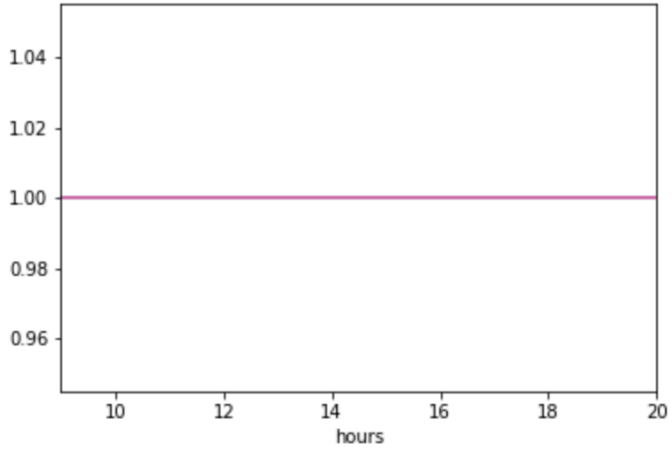
```
Test RMSE: 77.175297
Test MAE: 54.26
Test PE: 0.11
```



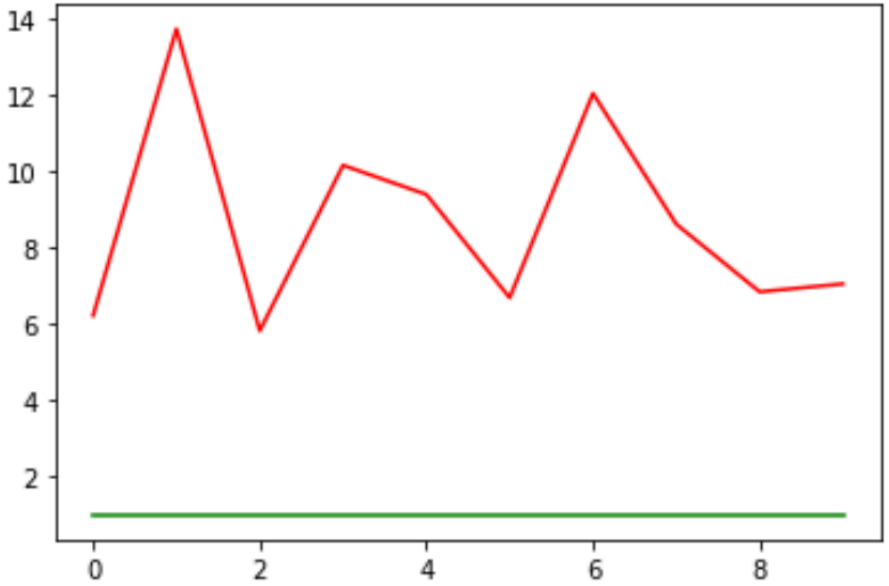
XGBoost model

Different locations

- Location 2: Queens. Jamaica Bay

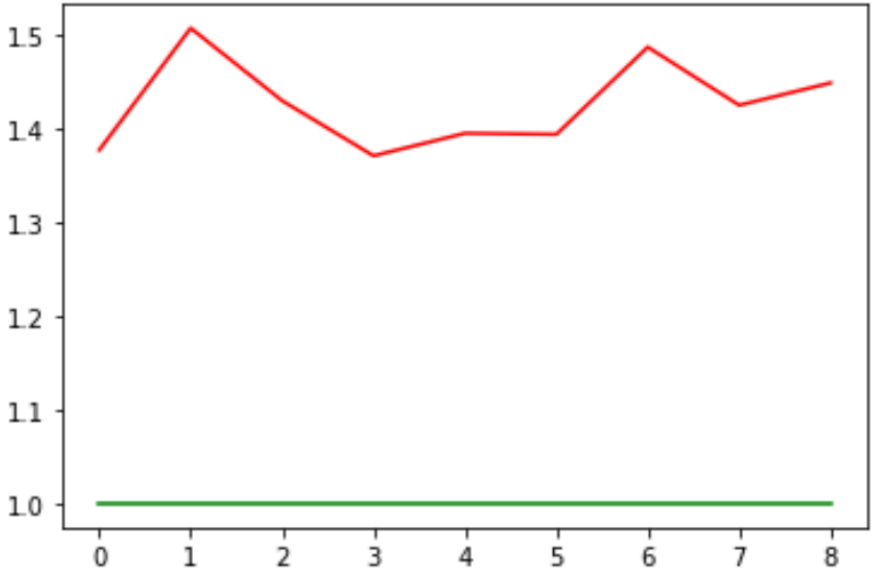



```
Test RMSE: 10.988136
Test MAE: 9.46
Test MPE: 9.46
```



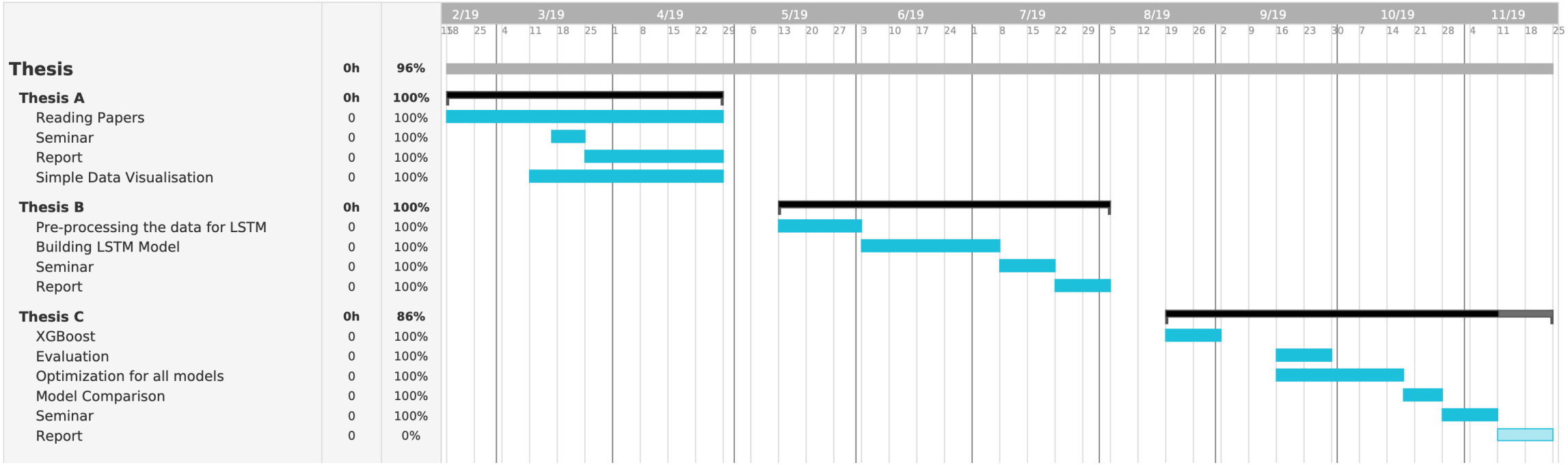
LSTM

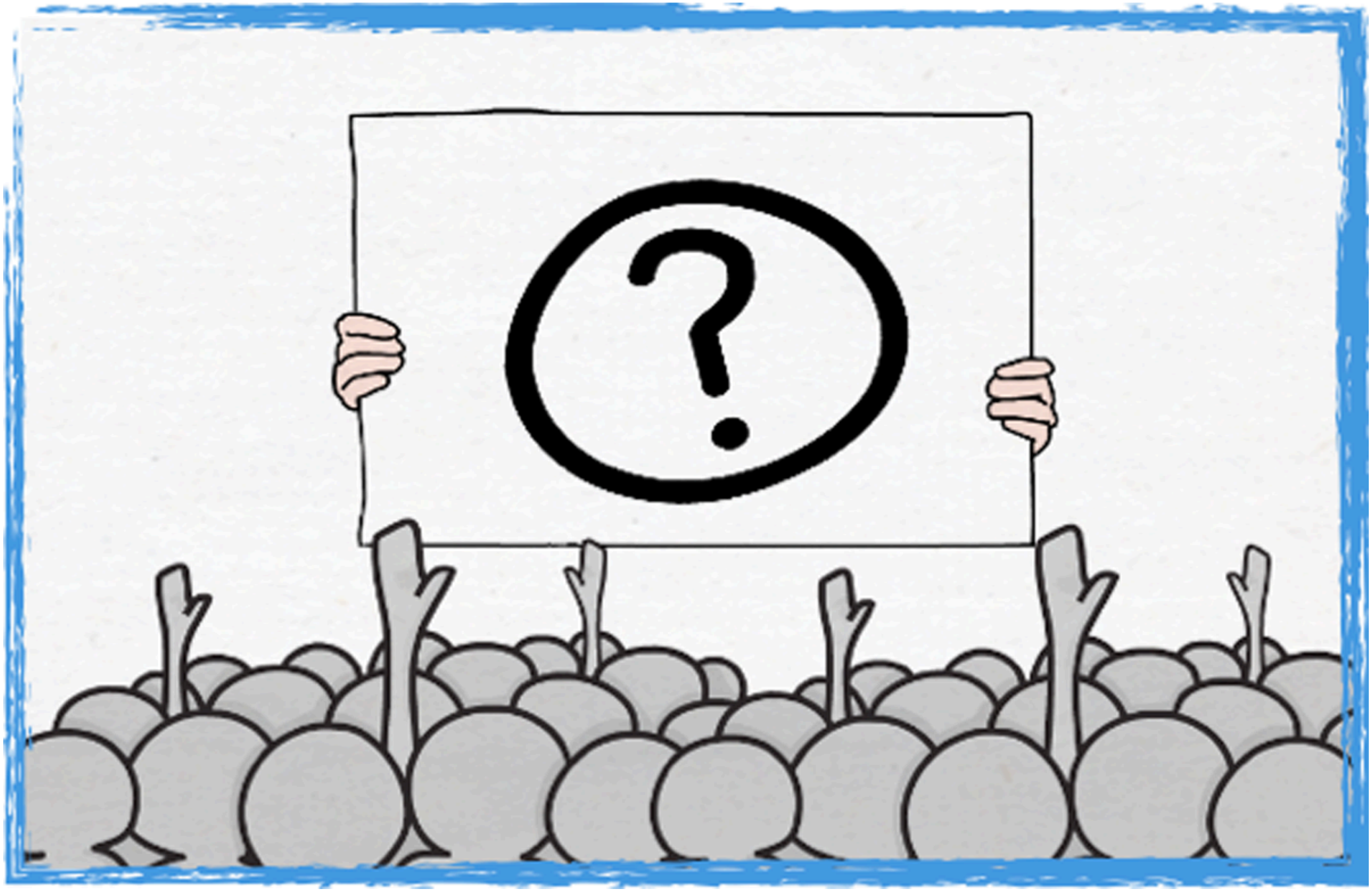
```
Test RMSE: 0.447102
Test MAE: 0.44
Test PE: 0.44
```



XGBoost

Gantt Chart Time Line






Thank You!