Big Data Applications Symposium - Summer 2017

Project Name:

proactive rebalance citibike station with predictive machine learning model

Team: Chuan-Heng, Lin
Hao-Yuan, Chen

Abstract: We aim to develop a predictive algorithm in order to help user to find the optimal bike station without paying extra fee. Therefore the expected outcome will be the citibike station with empty dock under the constraint that user only travel within 45 mins.

Motivation

Who are the users of this application? CitiBike user

Who will benefit from this application? CitiBike user

Why is this application important?

Users can avoid wasting their time from riding to the stations without an available dock.

What actuation(s) or remediation actions are performed by the application?

Our application can provide reachable stations with our algorithm

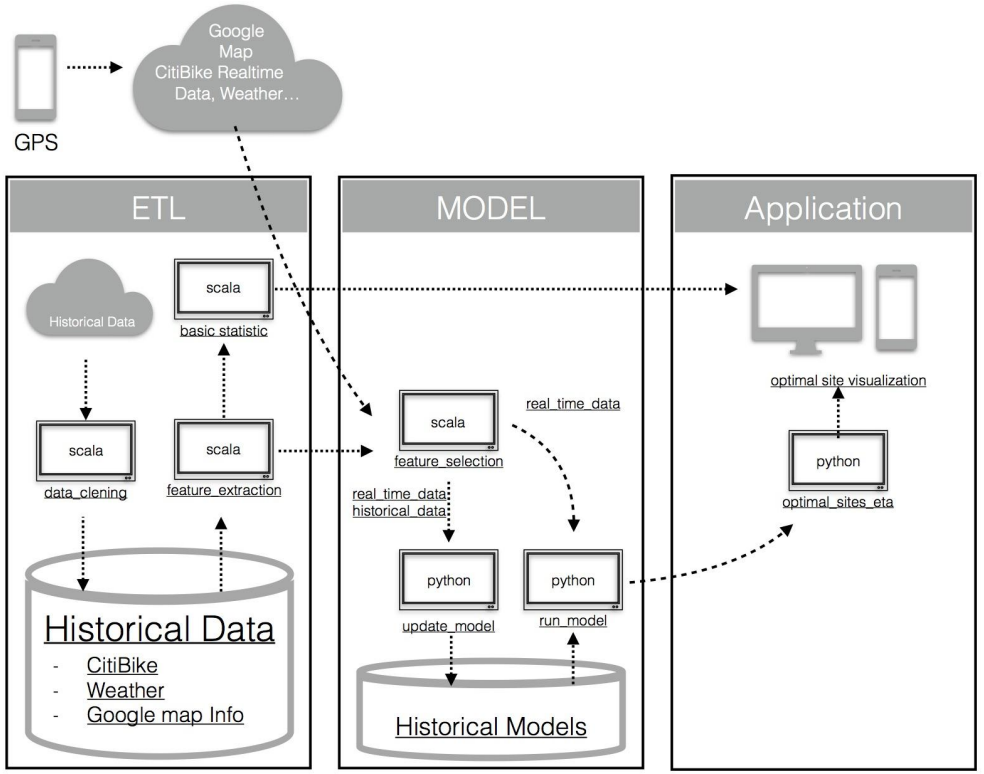
Data Sources

Name: CitiBike Historical Trip Data
Description: station_start_time, station_end_time, ...
Size of data: 2GB

Name: CitiBike Real-Time Station Status
Description: num_bikes_available, num_docks_available, ...
Size of data: 30 kb

Name: CitiBike Real-Time Station Information
Description: lat, lon, ...
Size of data: 30 kb

Design Diagram



Platform(s) on which the analytic ran:

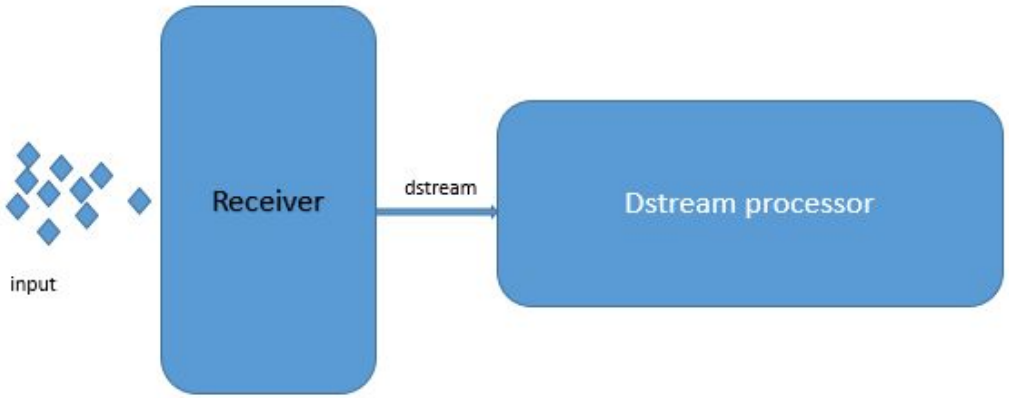
Local machine

Experiments - Spark Streaming

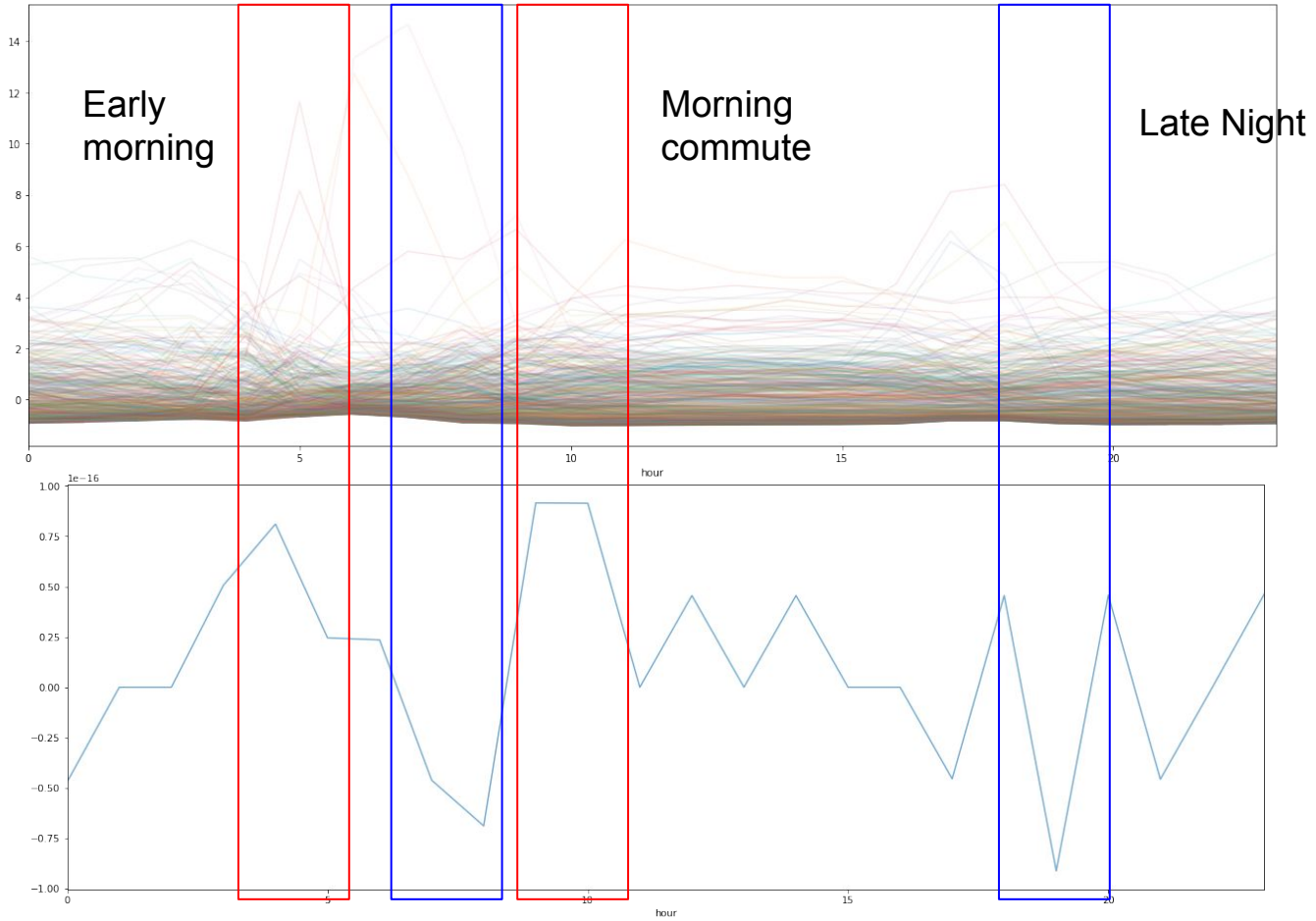
Implement Custom Receiver

- onStart(): Things to do to start receiving data
- onStop(): Things to do to stop receiving data.

Once the data is received, that data can be stored by calling ***store(data)***



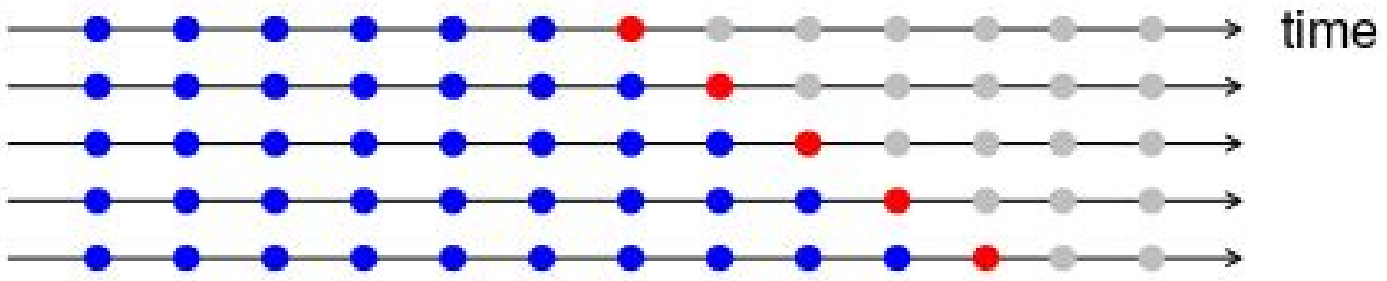
Experiments - temporal profile



Experiments/Results

Temporal K-Fold Cross Validation

- Training set increases (blue), test set is always the next subset (red)



Experiments/Results

1. Spatial distribution
    a. Red -> undockable
    b. Green -> dockable

Results

Random Forest Parameter setting:

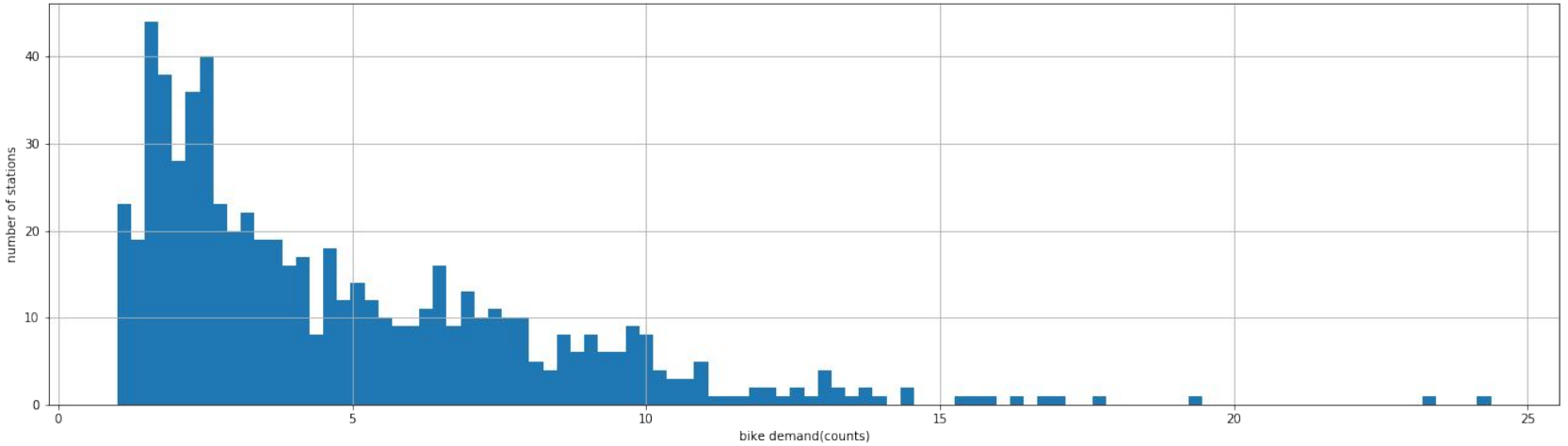
```
RandomForestRegressor(n_estimators= 10,
                      max_depth=50,
                      min_samples_leaf = 400)
```

Best Setting

MAPE 40%

Obstacles

1. Sparse and imbalanced data. Most of stations has a little demand, only certain stations has large demand.



2. The quota of Google Map API and Google Distance Matrix API is limited to thousands of request per day.

Summary

The result we get is limited since we only train the model using historical Citibike trip data. It may improve the result by combining more data source like NYC weather data. For the future work, we can keep training our model using the real-time Citibike data to make the model better and better.

Acknowledgements

CitiBike General Bikeshare Feed Specification

References

[1] Liu, Junming, et al. "Station site optimization in bike sharing systems." Data Mining (ICDM), 2015 IEEE International Conference on. IEEE, 2015.

[2] Singhvi, Divya, et al. "Predicting Bike Usage for New York City's Bike Sharing System." AAAI Workshop: Computational Sustainability. 2015.

[3] Spark Apache Org, 'Spark Streaming Custom Receivers". [Online]. Available:https://spark.apache.org/docs/1.6.1/streaming-custom-receivers.html#spark-streaming-custom-receivers. [Accessed: 10- Jul- 2017].

proactive rebalance citibike station with predictive machine learning model

Demo Time!