Bike Share Challenge

Nguyen Tien Huy – 15 August 2019

Data Exploration

The task is to build a predict model for outputting the net rate of bike renting for a given station (net rate is defined as trips ended minus trips started at the station for a given hour).
Three kinds of information are provided:

- Station information (id, name, longitude, latitude, dock count, city): 76 stations.
- Trip data (id, duration, started station, started time, ended station, ended time, Subscription Type): 354152 trip records from September 2014 – August 2015
- Daily weather measurements (Date, temperature, Dew point…): 1825 records from September 2014 – August 2015 for 5 cities San Francisco, Rewood City, Palo Alto, Mountain View and San Jose.

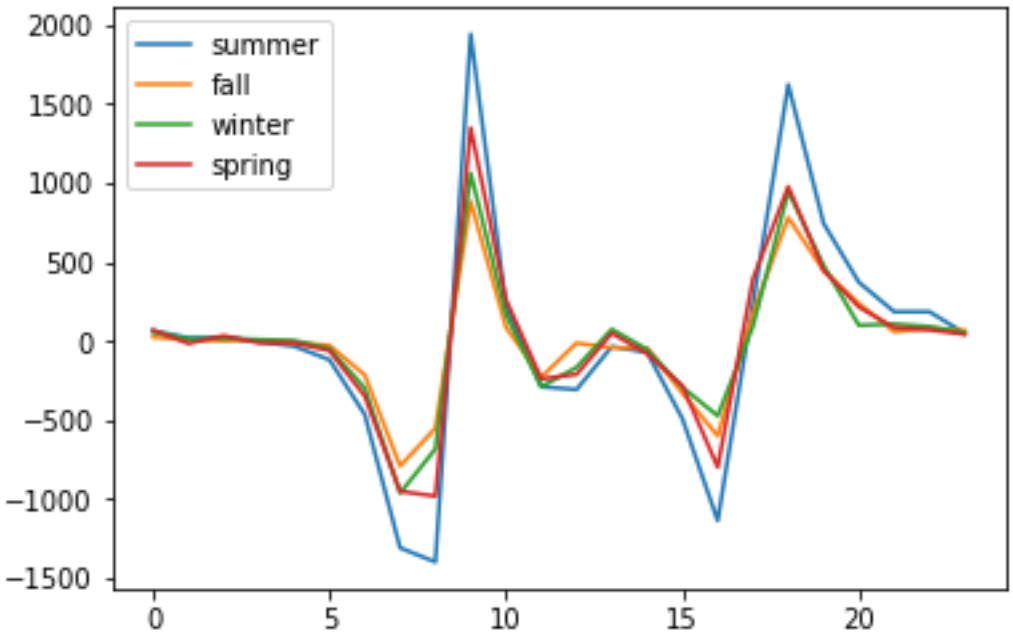


Figure 1 Net ratio over 24 hours

I also did some analysis to consider whether seasons and weekdays affect to the net ratio. As we can observe from Figure 1, the seasonal net ratio's patterns are similar and the summer one has the most ratio value while the fall one has the least ratio value. That shows that seasons affect to the net ratio.

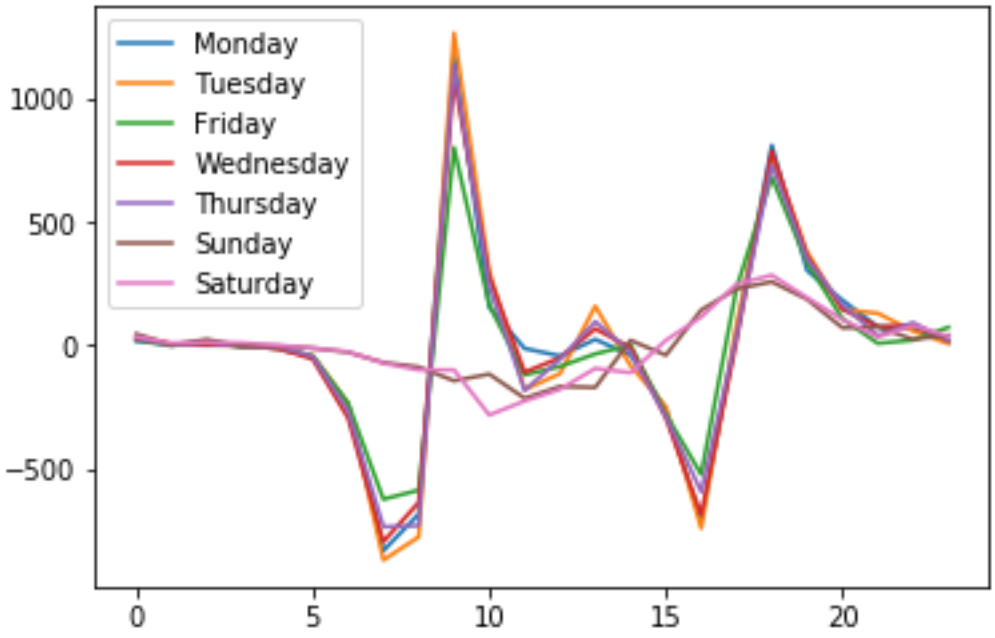


Figure 2 Net ratio over 24 hours

In Figure 2, we observe an interesting information that there is a different between workweek and weekend. The rush hours in both of the two Figures are around 8-9am and 4-6pm. That is matched to work commute time.

From this analysis, we can see that information of season and weekday affects to the net ratio.

Model approach

From 198156 records of (date, station id, net ratio), I randomly split it into the train/test/dev sets are as follows:

- Train (60%) 126184 records
- Test (20%) 36107 records
- Dev (20%) 35865 records

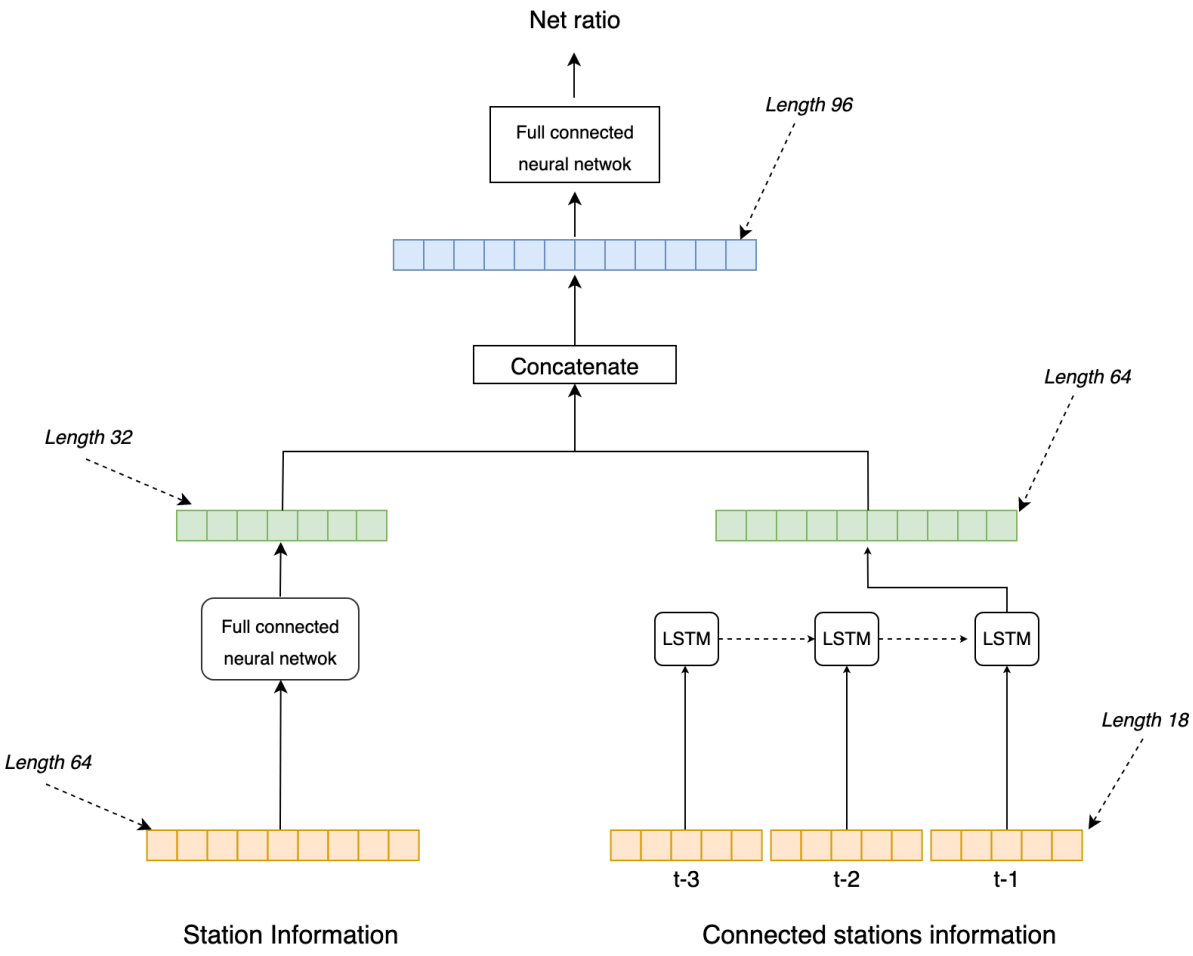
Each record contains the following information:

- Station information - a vector with length 64 (float and int features are normalized in range (0,1))
    - Latitude (float)
    - Longitude (float)
    - Dock_count (int)
    - Weekday (one-hot vector with length 7)
    - Season (one-hot vector with length 4)
    - Weather (21 float features and one-hot vector with length 5 for weather event)
    - Hour_category (one-hot vector with length 24)
- Connected stations information – a matrix 3*18: contains information (started trips count, ended trips count, net count) of the most 5 connected stations and itself in previous 3 hours.

According to my analysis, most of trips (99%) are ended in less than 3 hours. Note: two station are connected if there are trips started in one station and ended in the other. These values are normalized in range (0,1)).

- Output target: net_ratio (float) (this value is normalized in range (-15,15))

I propose a neural LSTM which is described in Figure 3. The hyper-parameters are tuning via a grid search. Because the connected stations information is a sequence data, LSTM is suitable to it.



The neural LSTM regression for net prediction

Figure 3 The proposed model

Performance analysis

I use the mean squared error (MSE) and root mean squared error (RMSE) to evaluate my model and compared with a baseline (only using station information without connected stations information)

Table 1 MSE and RMSE results

Method	MSE on test set	RMSE on test set
Baseline	7.3	2.7
Proposed	5.2	2.2

The results from the table mean that my model has an average root mean squared error of 2.2 and the baseline's one is 2.7.

Discussion

I also inspect the cases (around 20 cases) with high error (error > 20). An interesting observation is that these cases are related to one of three stations (station id: 69, 70, and 88) and around 4-6pm on workdays. Both of the two models (proposed model and baseline) also have high error for these cases. It means the connected stations information is not the cause.
Because the weather information is for a whole day not hourly, it is also not the cause. When I inspect whether anything special took place in those date but I collected nothing.
Therefore, I assume that there are two reasons:

1. There is special events in these date but the information of these events are not available online.
2. There is a mistake in collecting data, which creates noisy records.

Conclusion and Improvement

In this approach, I propose a neural LSTM model with the station information and the connected station information. This model achieves the RMSE of 2.2. I think this error is acceptable for the logistics team. As I discussed in the above section, the external information may affect to the net ratio (e.g., train schedules change, holiday, and disaster). Therefore, these kinds of information should be included for predicting the net ratio.

Implementation Notes

I implement the model on Keras with Tensorflow backend.
Each epoch takes 8s to complete and the total number of epochs is 100.
Github: https://github.com/ntienhuy/Bike-sharing-challenge