Sharing-Bicycle System Prediction

Horace Wang

Mar.26 2019

Introduction

Sharing-bicycle has become a new popular business around the big cities in the world. Normally, a company will produce a large number of bicycles and locate them in different stations around the cities, and the customers, who have registered before, could use their information to unlock the bicycle and use as long as they want. If the user decides to return the bicycle, he only needs to ride the bicycle to the nearest location and lock its bicycle there. And the next customer might come, uses the bicycle and starts its new cycle again. There are two major advantages of this model: first, the customer no longer needs to purchase the bicycle to enjoy its convenience; second, without worrying about his own bicycle, the customer is free to choose to ride a bicycle or not, at anytime and anywhere.

Problem

However, all these convenient features are based on one foundation: the customer could always get a bicycle whenever he wants. During the rush hour, it’s normal to have far more people in the business district than the other areas, so the company needs to move some bicycles to the business district from other stations. Without a useful tool, it’s hard for the company to predict when and how many should they relocate these bicycles. The purpose of this report is to use the knowledge of data science to help these companies have a better prediction of the future bicycle situation of each station.

Data

Citibike is one of the largest companies which provide bicycle-sharing service in New York. To better help its customer to see if a station has a bicycle available or not, Citibike keeps releasing the most up-to-date station availability JSON file on its website (https://feeds.citibikenyc.com/stations/stations.json). The file has the following characteristics for each row of data:

executionTime
stationBeanList
stationName
availableDocks
totalDocks
latitude
longitude
statusValue
statusKey
availableBikes
stAddress1
stAddress2
city
postalCode
location
altitude
testStation
lastCommunicationTime
landmark

Methodology

The data gathered from the official Citibike website will be cleaned and prepared first. And the data from Foursquare data will be included to generate the surroundings of each station. Afterward, different types of analyzing tool will be used, including Linear Regression, SVR, Decision Tree, and Natural Network. Based on the result and score of each method, the best one will be selected and used to produce the best prediction for the future.

Discussion

The data will be separated randomly into two groups, one for training and one for testing. Because there is only one set of data available, each method will use identical training and testing groups, to eliminate the noise from different data. With the steps mentioned above, different methods provide a different prediction about future station availability. The analysis uses the data from the previous 2 hours to predict the result in the 3rd hour. Listed below are the results of the same 4 station with different prediction methods:

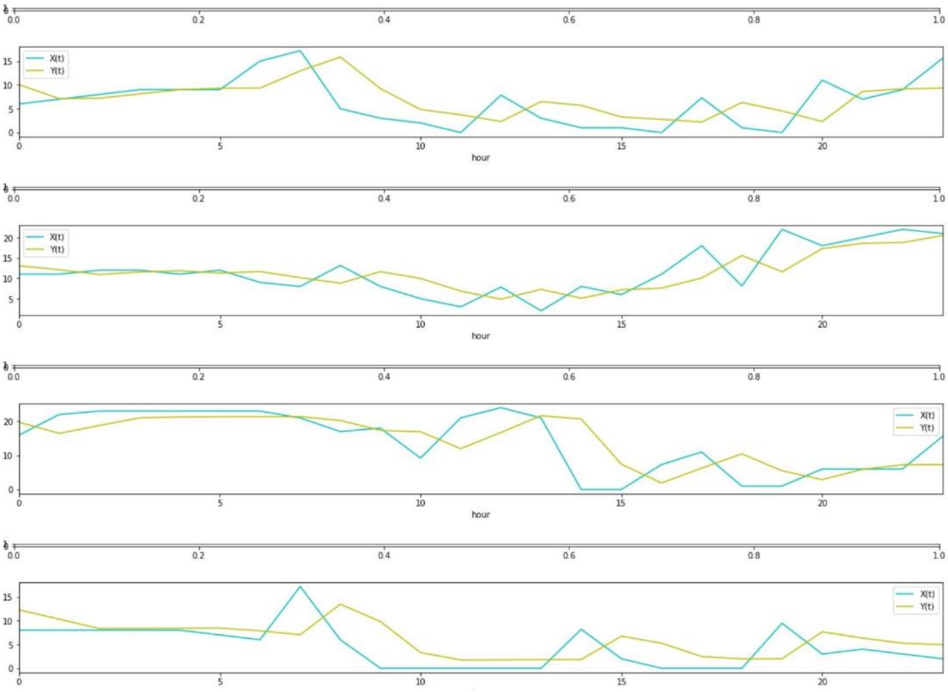

Figure 1: Linear Regression

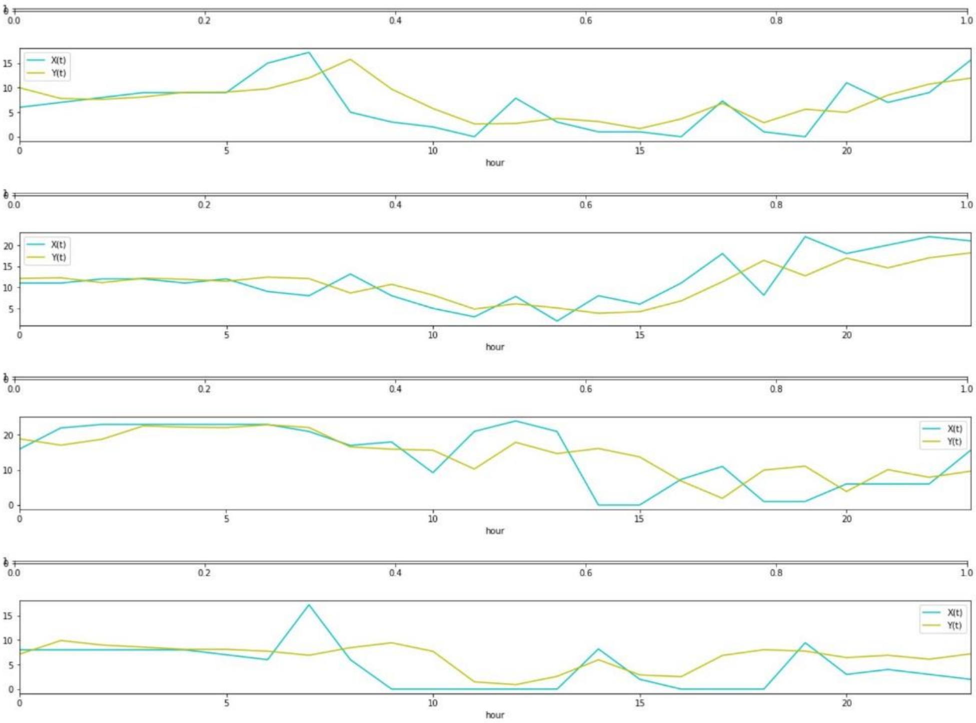

Figure 2: SVR

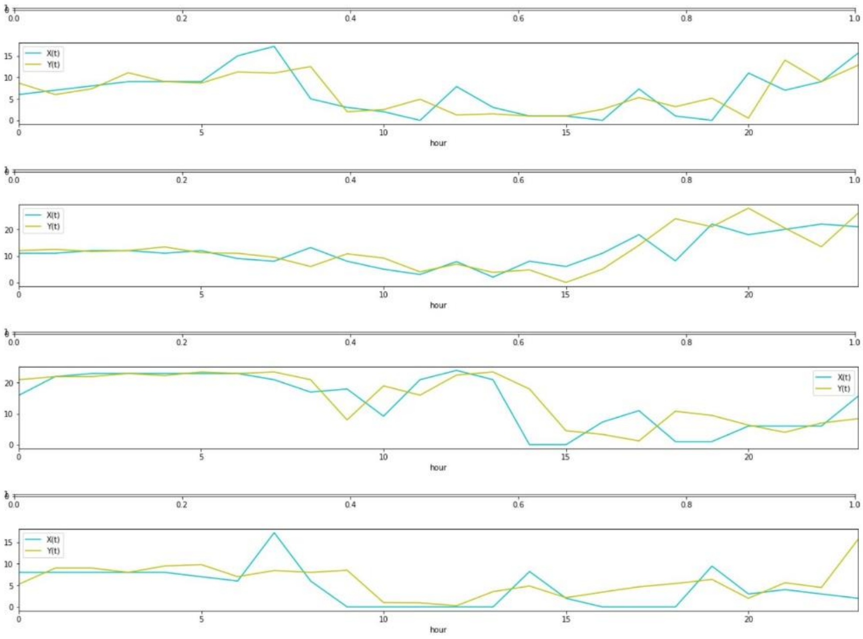

Figure 3: Decision Tree

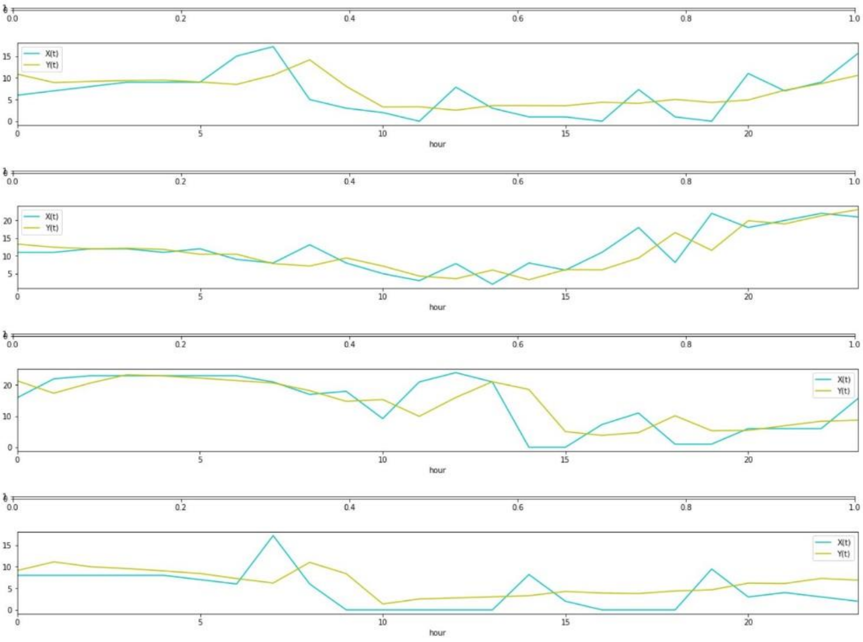

Figure 4: Natural Network

In each plot, the blue line is the predicted result, and the yellow line is the actual value. Although it only shows 4 stations of each method here in the plot, there are actually more than 700 stations in the data. Thus, some method may seem to have a better prediction in the plot, it doesn't mean it is the best method, because the best method needs to have the best numeric value result. The R-square and means squared error of each method are also calculated, and this is the result:

	Linear Regression	SVR	Decision Tree	Neural Network
R-square	0.547	0.556	0.485	0.622
MSR	18.453	18.063	20.960	15.387

The higher the R-square is, the more accurate the result is, and the lower the MSR is, the smaller the distance between the prediction value and the actual value is. From the above two factors, we could see clearly that the Neural Network is the best model to predict the bicycle sharing situation issue in New York.

Conclusion

Although we conclude that Neural Network is the best model to describe the bicycle-sharing issue among these four, it doesn't mean there is no other better model. Neural Network has an R-square value of 0.622, which means there is still a lot possibility to improve this model. For example, if we consider the relationship between each station, (eg. Negative relationship between the stations located in the center of a city and the edge of the city), the result may be more accurate. Or, if we could obtain more data from the customer side, the result may be improved. With more data available, we could have a much better direction for analyzing this situation in the future.