Applying Machine Learning Concepts on Bike-Sharing Systems

Abhinav Kapula akapula1@student.gsu.edu

Saivivek Therala stherala1@student.gsu.edu

Thirupathi Gandla tgandla1@student.gsu.edu

Introduction

Bike sharing system is a method for leasing bikes where the business tasks include enrolment of users, bike-rental, and bike return which is automated by means of a system of network areas all through a city. Utilizing this system, individuals can lease a bike from one area and return it to a better place as required. Currently, there are more than 500 bike sharing projects all over the world. Using Machine learning concepts, we are going to predict patterns on data by using the previous data. The effect of factors such as weather, geographic location, time of day, day of week etc. on the number of bikes at a bike-share station are investigated. This helps business in drawing some insights from the existing data.

History

Public bike-share systems were first conceptualized in 1965 in Amsterdam under the Witte Fietsen initiative. The entrepreneur behind the plan, Luud Schimmelpennink came up with the idea of leaving 2000 free white bikes in Amsterdam that would be free for everyone to use. Users of the system could pick up any bike, ride it to their destination and leave it there for the next user. There were no locks or bike stations in this system. This resulted in an unreliable system, as there was no way to predict where users could find free bikes. The program was also compromised by theft and vandalism, as there was no user accountability. Nearly 26 years later, the first large-scale 2nd generation of bike share systems was introduced in Copenhagen, Denmark in 1991. In this system, bikes were designed to be picked up and returned at specific locations which resulted in a more reliable system.

Bike sharing systems today

Modern third generation bike-share systems require customers to authenticate themselves through identification in order to increase user accountability. Some bike-share systems today now require users to pay with a credit card so that the user is charged the price of the bike in case of theft. In addition, most bike-share systems use proprietary parts in their bikes to discourage disassembly and resale of parts. Bike-sharing today has gone through technological improvements such as real-time status maps of bike-share stations, smartcards and electronic-locking racks and on-board communication systems for location tracking.

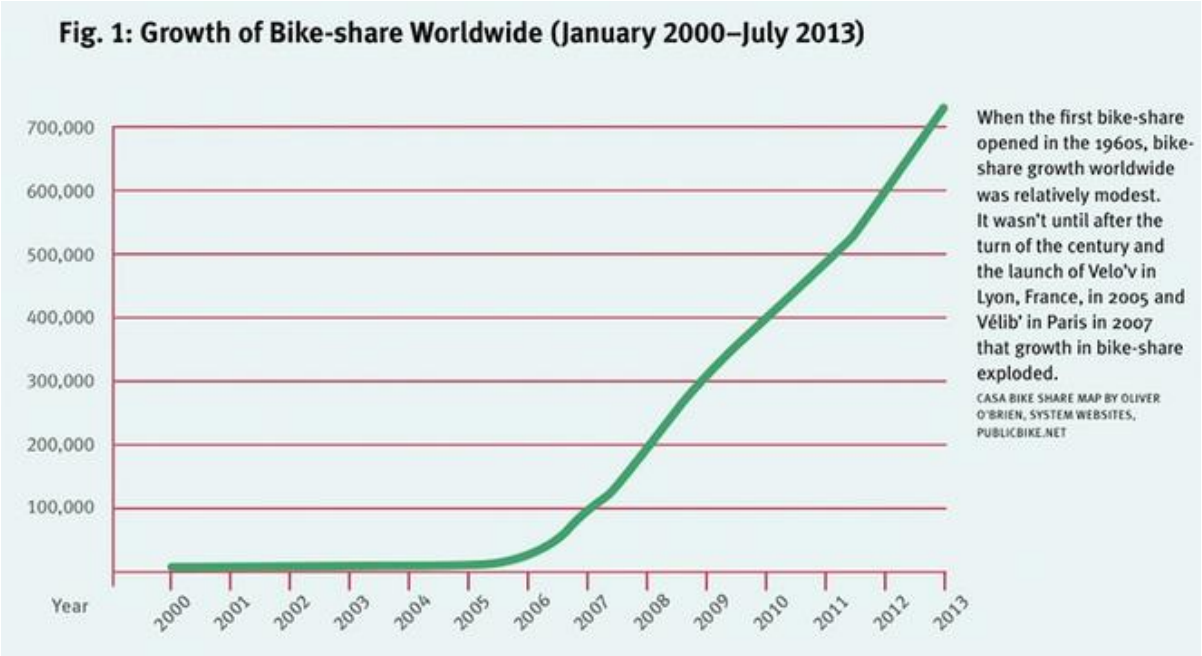


Fig. 1: Growth of Bike-share Worldwide (January 2000–July 2013)

When the first bike-share opened in the 1960s, bike-share growth worldwide was relatively modest. It wasn't until after the turn of the century and the launch of Velo'v in Lyon, France, in 2005 and Vélib' in Paris in 2007 that growth in bike-share exploded.

CASA BIKE SHARE MAP BY OLIVER O'BRIEN, SYSTEM WEBSITES, PUBLICBIKE.NET

Bike sharing system growth worldwide

Challenges of modern bike-share systems

Let us define the term overflow as a situation where a bike-share station is in danger of being too full (i.e. customers can't park their bikes there). This leads to customers being forced to use another bike-share station or park the bike privately overnight. It would therefore be useful for customers to have a prediction system that tells them if a bike-share station will be full when they arrive. In addition, when commuters have a wide range of choices in regard to where they park their bike, it would be helpful to know which station would be the least likely to be full.

Commute patterns will place imbalances in bike-share systems. In addition to the overflows mentioned above, let us define shortage as a condition where a bike-share station is in danger of running out of available bikes. Shortages and overflows occur as part of the daily commute pattern.

When shortages arise, it is important that redistribution trucks restore balance to the bike-share distribution. It is therefore just as important to understand when bike-share stations run empty, and this is in fact just the inverse problem (predicting overflows and the likelihood of overflows).

Mathematical description of the problem

For regression purposes, *bs* is defined as the number of bikes available at the given bike-share station. Let *w* be the weather recorded at time *t* , where air temperature *T* is measured in °C, cloud cover *CC* is measured in okta and precipitation (last 24h) *RR* is measured in mm. Let *sta* be the station described by the latitude *lat* , longitude *lng* , altitude *al t* , and the station_id *id*. And finally, let *t* be the time described by the variables: time of the day *h*, day of the week *wday*, day of the month *d*, month *m*.
$w = (T,CC,RR)$
$t = (h,wday,d,m)$

st = (*lat* , *lng* ,*al t* , *id*)

The contribution of our work is to answer the following questions:
1. Given a station *sta*, the weather conditions *w* and the time *t* , can we predict the bike-share station status *bs* using a machine learning algorithm?
2. What are the best performing estimators for this particular task?
3. How do factors such as time of day, weather etc. affect bike share traffic?

Goals of our work are as follows

Enable the bike-share operators to be proactive. If they can receive a prediction about the net loss in number of bikes *X* at a given station in a timeframe *T*, they can then place *X* +*d* Number of bikes at that station, where *d* is a small buffer that ensures that the station will not be empty. Similarly, if they are able to predict an influx, they can do the opposite. In order to achieve these goals, the system must be able to predict the future based on historical data. It must also be flexible enough to take into account factors that cause variance like wind, rain, temperature etc., while generalizing enough to understand concepts like weekdays and weekends, morning and afternoon rush hours etc. that produce periodic patterns in the data.

We also found that the standard deviation of the number of bikes available at each station during the day was high. This meant that they could not simply average the recordings for any given time during the day and use that to predict the number of bikes. In order to explain the high variance in the data, they took into account factors like:

Data features

The feature values chosen for the data were:

- Working day
- Weather
- Temperature
- Humidity
- Wind speed

Data format

Bike-share data is stored as CSV (Comma separated values) files.

Weather data

Additionally, in order to determine the importance of weather on bike-share stations, weather data collection was required as well.

Algorithms

Classification:

In classification, the target variables are class labels, i.e. all observations are of class A, but the problem is to determine which sub-class of class A they fall into.

Example: All emails fall into the class *Email*, but spam classifiers strive to determine which ones fall into the *spam* and *relevant_email* subclasses respectively.

Regression:

In regression, the target variables are real/continuous values, i.e. the problem is to determine how big or small the target variable will be.

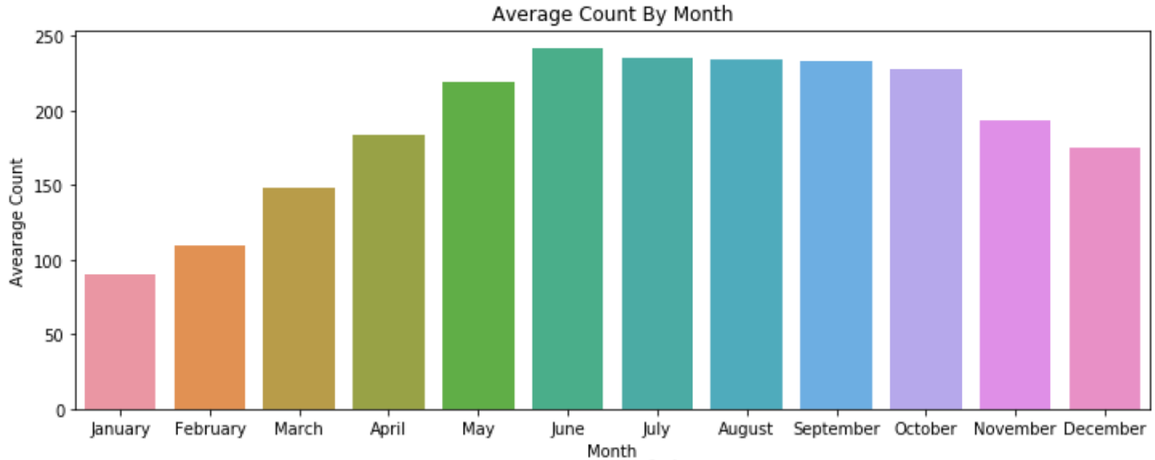
Example: While browsing Netflix, Alice gave a 5/5 rating to some comedy movies, 2/5 to a few thriller movies and 3/5 to some mixed-genre movies. She then comes across a new movie Which is in the thriller/horror category. Based on her previous experiences, Netflix can try to use a regressor to predict how well she will like the movie on a scale from 1-5.

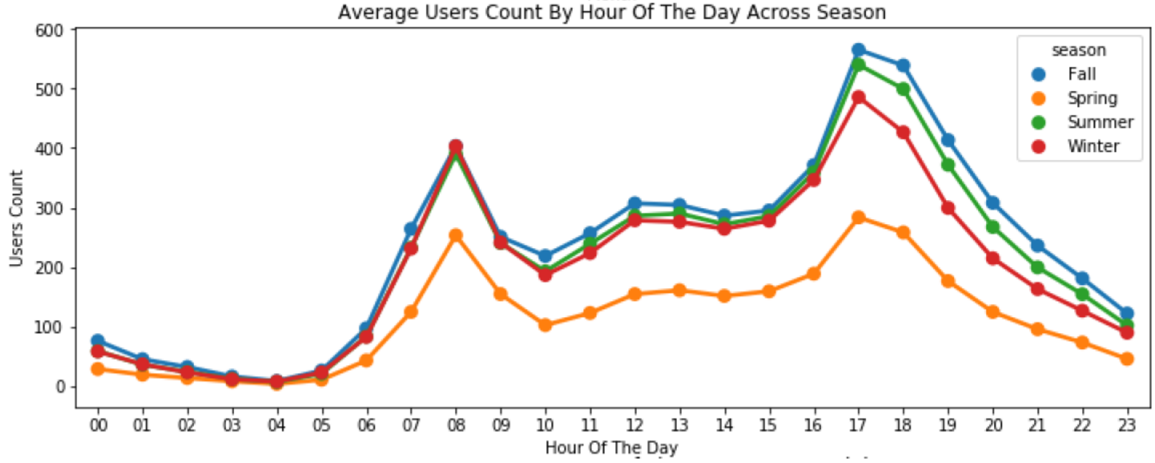
In the bike-share domain, this would mean that while classification algorithms might output whether or not a station will be in an overflow, balanced or shortage state, regression algorithms will output the expected number of bikes at the station. The training set will be slightly different for regression algorithms, as will the output for the test set.

Random Forest

The Random Forest algorithm was introduced by Leo Breiman in 2001. It has been shown to have excellent performance as shown in Caruana & N-M , and came in a close second only to boosted decision trees, beating relatively complex algorithms such as neural networks and SVMs. Random Forests combine bagging with the concept of random feature selection introduced by Ho and Amit and Geman.

Analysis





The results of the analysis yield some interesting statistics:

- Season: If we can see the above graph plots, we can infer that the demand for bike rental is more during the summer till mid-Fall seasons, due to less rains and wind speeds.

- Day of week: As you might expect in a city of bicycle commuters, there is roughly 2.5 times the amount of traffic on weekdays as there is on weekends.

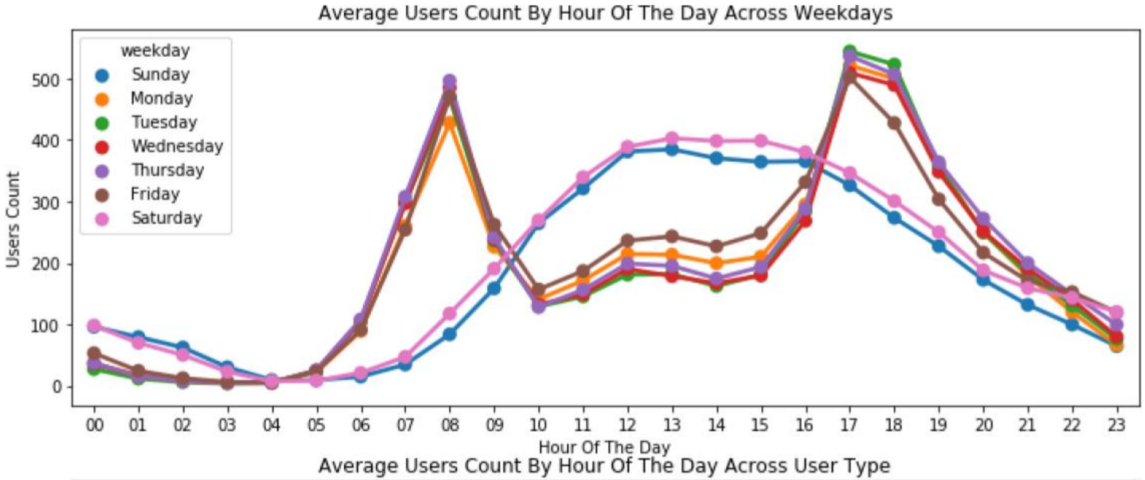
- Hour of the day: During weekdays, there is an extreme demand for bikes between 7-8 am and between 4-6 pm. Whereas there is a steady increase in demand by afternoon on weekends.

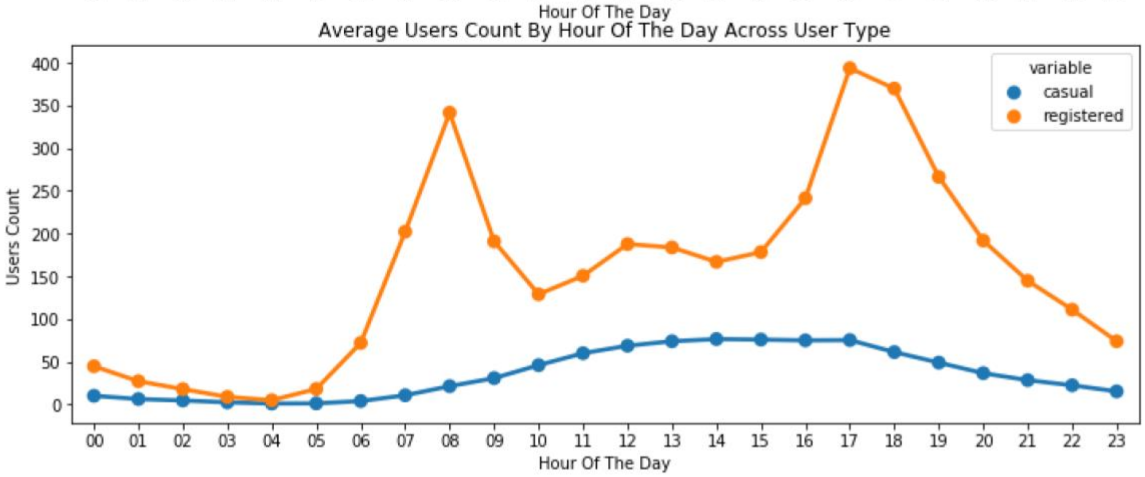
- Registered Users: Unlike guest users, the registered users follow specific pattern throughout the day.

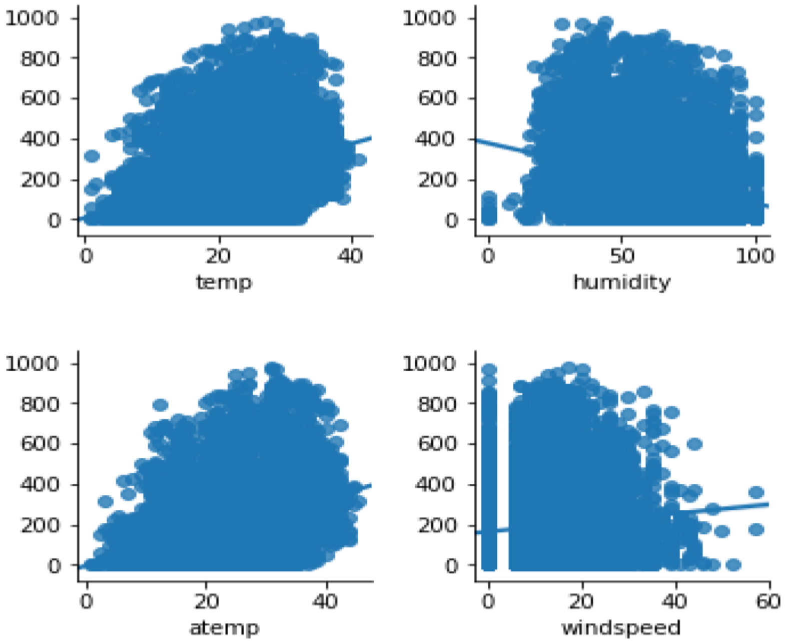
- Wind speeds: As wind speeds increase, there is a drop in bike rentals.

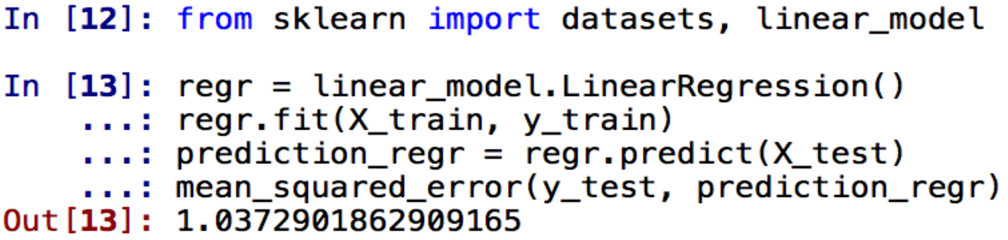
- Temperature: As temperature increases, customers tend to use more bikes for rent.

- Humidity: As humidity increases, customers would not like to commute though bikes.

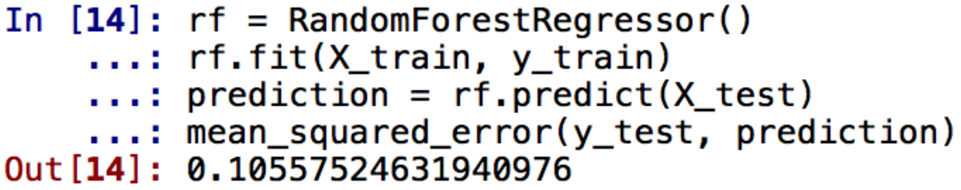








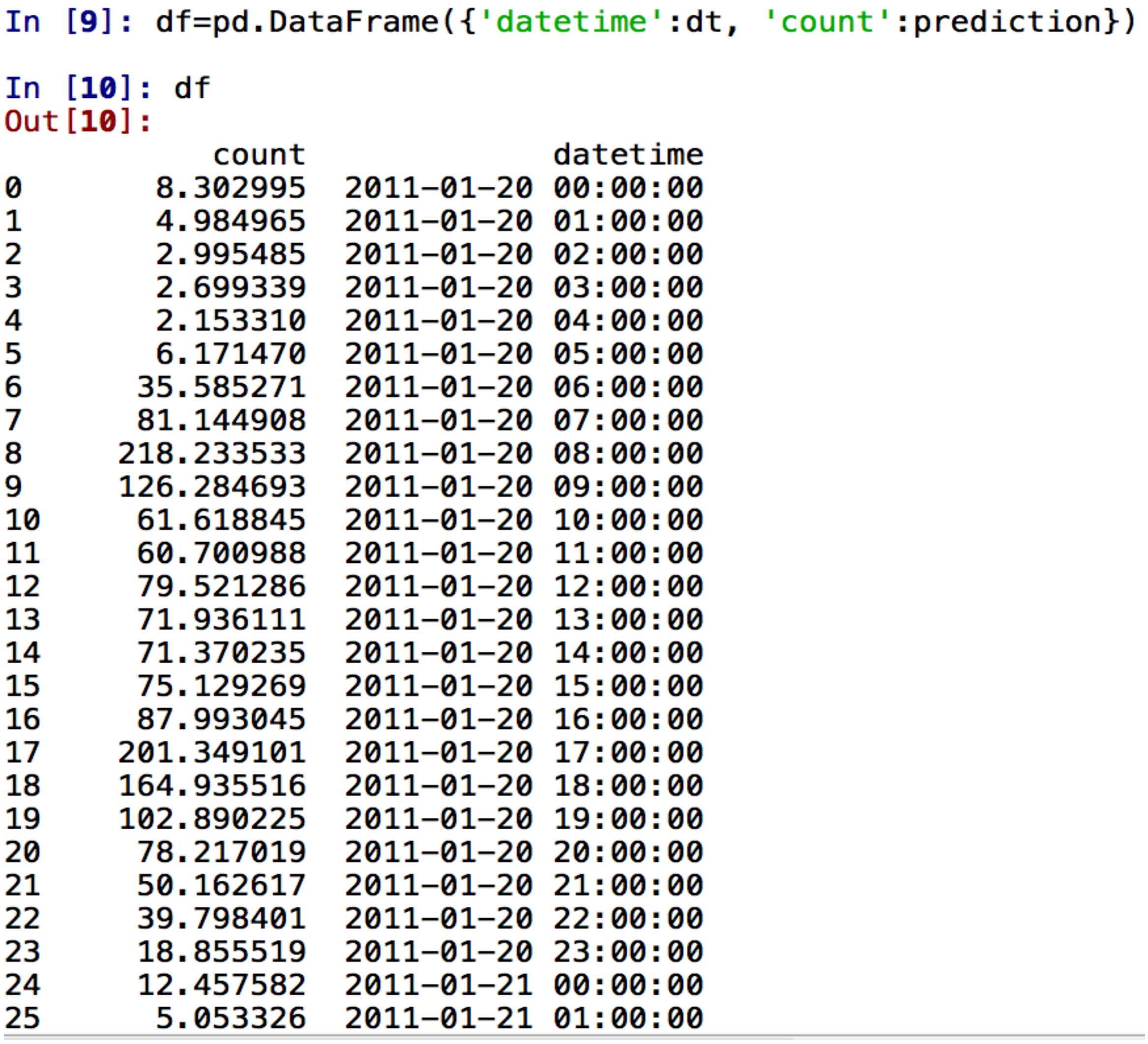
```
In [12]: from sklearn import datasets, linear_model

In [13]: regr = linear_model.LinearRegression()
    ...: regr.fit(X_train, y_train)
    ...: prediction_regr = regr.predict(X_test)
    ...: mean_squared_error(y_test, prediction_regr)
Out[13]: 1.0372901862909165
```



```
In [14]: rf = RandomForestRegressor()
    ...: rf.fit(X_train, y_train)
    ...: prediction = rf.predict(X_test)
    ...: mean_squared_error(y_test, prediction)
Out[14]: 0.10557524631940976
```

After comparing the data with the real value, it is indicated that the linear regression model is less accurate. The results and the accuracy of linear regression analysis are greatly improved when use of random forest model to predict the demand for bicycle rental.

Output:



```
In [9]: df=pd.DataFrame({'datetime':dt, 'count':prediction})

In [10]: df
Out[10]:
          count             datetime
0      8.302995  2011-01-20 00:00:00
1      4.984965  2011-01-20 01:00:00
2      2.995485  2011-01-20 02:00:00
3      2.699339  2011-01-20 03:00:00
4      2.153310  2011-01-20 04:00:00
5      6.171470  2011-01-20 05:00:00
6     35.585271  2011-01-20 06:00:00
7     81.144908  2011-01-20 07:00:00
8    218.233533  2011-01-20 08:00:00
9    126.284693  2011-01-20 09:00:00
10    61.618845  2011-01-20 10:00:00
11    60.700988  2011-01-20 11:00:00
12    79.521286  2011-01-20 12:00:00
13    71.936111  2011-01-20 13:00:00
14    71.370235  2011-01-20 14:00:00
15    75.129269  2011-01-20 15:00:00
16    87.993045  2011-01-20 16:00:00
17   201.349101  2011-01-20 17:00:00
18   164.935516  2011-01-20 18:00:00
19   102.890225  2011-01-20 19:00:00
20    78.217019  2011-01-20 20:00:00
21    50.162617  2011-01-20 21:00:00
22    39.798401  2011-01-20 22:00:00
23    18.855519  2011-01-20 23:00:00
24    12.457582  2011-01-21 00:00:00
25     5.053326  2011-01-21 01:00:00
```

Conclusion

This project is set out to explore the concept that bike-share traffic could be successfully analyzed and predicted with machine learning patterns. The study also sought to compare some popular estimators to determine which was best suited for a prediction system. After comparing the data with the real value, it is indicated that the linear regression model is less accurate. The results and the accuracy of linear regression analysis are greatly improved when use of random forest model to predict the demand for bicycle rental. As a final step, the study determined the effect of climatological, geographical and time-based variables on the traffic flow.

Github link: https://github.com/akapula1/bike-share

References

[1] Institute for Transportation and Development Policy (ITDP).ITDP Bikeshare Planning Guide. URL: https : / / go. itdp . org/display / live / The + Bike - Share + Planning + Guide (visited on19/06/2014).

[2] Olivier O'Brien and UCL Centre for advanced spacial analysis. Bicycle sharing systems - Global trends in size. URL: http://www.bartlett.ucl.ac.uk/casa/pdf/paper196.pdf (visited on 19/06/2014).

[3] Barclays review by customer. URL: http : / /www. tripadvisor.com / ShowUserReviews - g186338 - d2151262 - r174553920 -Barclays_Cycle_Hire-ondon_England.html#REVIEWS.

[4] Divvy: Helping Chicago's New Bike Share Find Its Balance.URL:http://dssg.io/2013/08/09/divvy-helping-chicagos-newbike-share.html (visited on 20/09/2014).
[5] Olivier O'Brien. Bike Share Map. URL: http:/ /bikes.oobrien.com/global.php.
[6] DSSG Analysis. URL: https://github.com/dssg/bikeshare/wiki/Analysis (visited on 29/10/2014).