Student Name: Betty Wong
Student Number: 500-802-428
Createed on: 2017-10-31 11:42:31 PM

Forecasting Bike Sharing Demand Replication

Abstract

Bike Sharing Systems consist a fleet of bikes placing in blocks of docking stations. These bikes are rented and returned to any block of docking stations. People can rent a bike via membership as a registered frequent-user, or pay-as-you-go casual user. Predicting bike-usage-demand would help allocate adequate bikes in each block of docking stations for serving the community effectively.

The original model for bike demand prediction was conducted by Jayant Malani et at (1) in 2013 for Kaggle competition. Based on their research, I will resue their models with different parameters, and different dataset to gain a better insight of usage demand and draw a deeper conclusion.

1. INTRODUCTION

Bicycle sharing is increasingly becoming a promising initiative to encourage public cycling, which benefits users and community as a whole. In the last few years, bike-sharing has taken a major role in the transportation network of big cities as an alternative way of getting around the city. Bike share is growing at an astounding rate across 55 cities in US, with over 88 million trips made on a bike share since 2010. In 2016 alone, riders took over 28 million trips, on par with the annual ridership of entire Amtrak system, and higher than the number of people visiting Walt Disney World each year. 25% more trips were taken in 2016 than the previous year. Therefore, a good prediction of Bicycle sharing would be necessary.

The original study by Jayant Malani and his team used various predictive models such as Liner Regression, Bagging Regression, Ensemble of Regression, and Time Series with data from 2011 to 2012 to perform bike rental demand prediction. The results indicated XGBoost Regressor outperformed all other regressors as it has the lowest RMLSE (root mean square log error). I will reuse some of the testing with different parameters, adding few more predictive models, and different dataset to refining the predictions.

2. PUBLISHED RESERCHES

In the past, many researches have been done on Bike Sharing demand Prediction. Patrick Vogel et at (2) used data mining technique to analyze operation data on bike sharing in Vienna. With their study, they are able to prove their hypothesize statement "bike activity and demanding customers depend on stations locations" was correct.

Jon Froehlich et at (3) studied the Bicing – biking sharing system in Barcelona in 2009. They used three elements date, time, station's geo-location, number of available bicycles, and number of vacant parking slots. They scrape the Bike website every two minutes and collect data for 13 weeks for analysis. Stations were clustered according to number of available bikes and activity score in the curse of day.

Student Name: Betty Wong
Student Number: 500-802-428

"Visualization of the stations showed spatial dependencies – For example, uphill stations tend to empty and less active stations were located at the edge of the bike-sharing network".

Andreas Kaltenbrunner et at (4) explored and predicted trends in a bicycle-based public transport system. They analyzed human mobility data in an urban area using the amount of available bikes in the stations of the community bicycle program Bicing in Barcelona. Based on data sampled from operator's website, it was possible to detect temporal and geographic mobility patterns within the city. These patterns were applied to predict the number of available bikes or any stations some minutes/hours ahead. The predictions could be used to improve the bicycle program and the information given to the users via Bicing website.

Divya et at (5) predicted the bike usage pattern of New York City's Citi Bike System during morning rush hours of 7:00 am to 11:00 am. They used taxi usage in addition to emporal, demographic and weather factors as covariates in predicting "pairwise trips". They observed that analyzing "pairwise trips at the neighborhood level" instead of looking at individual stations in bike sharing systems can substantially improve the predictions. This model can assist planner to predicting bike demand at macroscopic level, between pairs of neighborhood.

2.1. Replication papers reviewed

The origianl research paper was created for Kaggle Competition – Bike Sharing Demand. The participants are asked to forecast bike rental demand of Bike sharing program in Washington, D.C. based on historical usage patterns in relation with weather, time and other data.

2.2. Description of original study

In past, various studies have analyzed several factors that affect bike usage and flow. In the original study, the authors tried to predict bike rental demand in each neighbourhood on hourly basis. They have combined the demand for all the stations that would fall within one census tract area. Various regression models with covariates like weather, population and time were being used in their study.

Their observation included:

- 97% of the rides were of short durations which were less than 20 min, and very few rides that were greater than 100 minutes
- The bike rentals for each station is different, and stations located in central Washington DC had the maximum bike rentals
- Weather which has the maximum bike rentals is clear weather, followed by light mist, and light rain.
- Registered users and casual users follow a different trend of bike rental demands

Student Name: Betty Wong
Student Number: 500-802-428

2.3. Information about the original study

Research questions

Create a model for predicting hourly demand of bike in each neighbourhood, defined as a census tract, of Washington DC area

Participants

Jayant Malani, Neha Sinha, Nivedita Prasad, Vikas Lokesh

Predictive Models

Three Baselines:

- The mean hourly demand for bikes using the response values in the train set, and the RMSLE on the test data is 0.60296
- The average demand for each hour in train set, and the RMSLE on the test data is 0.57589
- The average demand for each area for each hour in training set, and the RMSLE on the test data is 0.50283

Models that being studied:

- Linear Regression
  - Ridger regressor in Sklearn library
  - Lambda = 1000 on 10 fold cross validation give the best result

- SVR (smiliar to SVM)
  - SVR in Sklearn library
  - Performance is very slow
  - Penalty parameter = 1e0.3 and gamma = 0.1 and it took a day to run SVR on researcher's laptop

- Gradient Boosting Tree Regressor
  - Gradient Boosting trees in Sklearn library
  - It took around 3 hours to train a model with one set of hyper parameters
  - Number of trees = 300, maximum depth of each trees = 30, and gamma = 0.1 give the best result

- XGBoost Regressor
  - XGBoost in python library
  - It took 30 minutes to train a model
  - Number of trees = 125, and maximum depth of each trees = 12 give the best result

- Random Forest
  - Random Forest in Sklearn library

Student Name: Betty Wong
Student Number: 500-802-428

    - Number of trees = 100, maximum depth of each trees = 25, and maximum features = all features give the best result

- Extratree Regressor
    - It took 25 minutes to train a model
    - May be due to presence of noisy features, the result is not as good as Random Forest
- Bagging Regressor
    - Bagging regressor in Sklearn library
    - Number of trees = 120, and Bootstrap = True give the best result

- Ensemble of Regressors
    - Create a new train set by averaging the predictions of the three different predictors – Random Forest, XGBoost, and Bagging regressors
    - Run linear regression with x_train as the newly created data set, y_train as the actual values of the count, and two different linear regressors to predict demand by registered and cauls users
    - The result is not as good as XGBoost

- Time Series
    - Use the bike demand of the recent past of a station to predict the current bike demand in that station
    - Using the previous day's demand does not yield good result as the demand significantly varies with the day of the week
    - Using the bike demand data at that station on the same day of the previous weeks also not as good as XGBoost

Artifacts

Bike Sharing Demand – https://www.kaggle.com/c/bike-sharing-demand

R. Alexander Rixey, 2012, Station-Level Forecasting of Bike Sharing Ridership: Station Network Effects in Three U.S. Systems, TRB 2013 Annual Meeting

Buck, D., and Buehler, R. 2012. Bike lanes and other determinants of capital bikeshare trips. In 91st Transportation Research Board Annual Meeting

Etinne, C., and Latifa, O. 2012. Model-based count series clustering for bike_sharing system usage mining, a ase study with the velib's system of paris

Context variables

Duration: 97% of the riders were of short durations, less than 20 minutes, and only few rides that were greater than 100 minutes

Weather: most number of data points has clear weather

Student Name: Betty Wong
Student Number: 500-802-428

Hour of the day: registered users and total count has similar trend – two peaks in a day, one during morning and the other on evening, and casual user has the maximum demand in the evenings

Day of the week: registered users use bikes more on the weekdays whereas casual users use bikes mostly on the weekends

Census tract id: the most important feature for demands by both registered and casual users

Summary of the results

XGBoost regressor outperformed all other regressors which have the lowest RMSLE. The census tract ID is the most important feature for both registered user and casual user. The demand for casual users is much higher on weekends and holidays wheras the deamdn by registered users varies by small value over the different working day of the week.

Model	Average RMSLE for 10 fold cross validation	RMSLE on test set
Baseline – the average demand for each area for each hour using all data in training set	0.57812	0.57028
Linear Regressor	0.54675	0.55623
Gradient Boosting Regressor	0.48213	0.48327
XGBoost Regressor	0.40350	0.40505
Random Forest	0.41891	0.42144
ExtraTree Regressor	0.42712	0.43176
Bagging Regressor	0.41627	0.41433
Ensemble of Regressors	0.41092	0.41003
Time series	0.46831	0.48124

3. REPLICATION STUDY

3.1. Information about the replication

3.2. Comparison of replication result with original study results

Student Name: Betty Wong
Student Number: 500-802-428

3.3. Drawing conclusions across studies

4. SUMMARY

Student Name: Betty Wong
Student Number: 500-802-428

5. REFERENCE

[1] Jayant Malani, Neha Sinha, Nivedita Prasad, Vikas Lokesh: Forecasting Bike Sharing Demand

[2] Patrick Vogel, Torsten Greiser, Dirk Christian Mattfeld: Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns

[3] Jon Froehlich, Joachim Neumann, Nuria Oliver: Sensing and Predicting the Pulse of the City through Shared Bicycling

[4] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, Rafael Banchs: Urban cycles and mobility patterns: Exploring and prediction trends in a bicycle-based public transport system

Bike Sharing Demand - https://www.kaggle.com/c/bike-sharing-demand

Comprehensive Guide to Data Exploration - https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

Capital Bike Sharing - https://www.capitalbikeshare.com/system-data