Citibike: Bikeshare usage demand forecasting

https://www.citibikenyc.com/

A social experiment for predicting future demand of Citibike shared bicycle system in New York city

Sachid Deshmukh – M.S. in Data Science Student CUNY SPS

Ann Liu-Ferrara - M.S. in Data Science Student CUNY SPS

Ahmed Sajjad - M.S. in Data Science Student CUNY SPS

Highlights

- Paper proposes Machine Learning and Timeseries based forecasting methodologies for predicting future usage demand of Citibike bikeshare system in New York city
- Paper focuses on analyzing Citibike data and identifying important patterns to support business viability and overall profitability of Citybike bikeshare system
- Paper proposes various feature engineering techniques for predicting future usage demand of Citibike bikeshare system in New York city
- Paper evaluates feature importance and proposes key features which contributes towards predicting future usage demand of Citibike bikeshare system in New York city

Article Info

Keywords
Bikeshare,
Weather, Cycling,

Abstract

Bicycling is an activity which yields many benefits: Riders improve their health through exercise, while traffic congestion is reduced if

CitiBike, New York City

Article History
Created:
02/16/2020

riders move out of cars, with a corresponding reduction in pollution from carbon emissions. In recent years, Bike Sharing has become popular in a growing list of cities around the world. The NYC “CitiBike” bicycle sharing scheme went live (in midtown and downtown Manhattan) in 2013, and has been expanding ever since, both as measured by daily ridership as well as the expanding geographic footprint incorporating a growing number of “docking stations” as the system welcomes riders in Brooklyn, Queens, and northern parts of Manhattan which were not previously served. One problem that many bikeshare systems face is money. An increase in the number of riders who want to use the system necessitates that more bikes be purchased and put into service in order to accommodate them. Heavy ridership induces wear on the bikes, requiring for more frequent repairs. However, an increase in the number of trips does not necessarily translate to an increase in revenue because riders who are clever can avoid paying surcharges by keeping the length of each trip below a specified limit (either 30 or 45 minutes, depending on user category.) We seek to examine CitiBike ridership data, joined with daily NYC weather data, to study the impact of weather on shared bike usage and generate a predictive model which can estimate the number of trips that would be taken on each day. The goal is to estimate future demand which would enable the system operator to make expansion plans. Our finding is that ridership exhibits strong seasonality, with correlation to weather-related variables such as daily temperature and precipitation. Additionally, ridership is segmented by user type (annual subscribers use the system much more heavily than casual users), gender (there are many more male users than female) and age (a large number of users are clustered in their late 30s).

1. Introduction

Since 2013 a shared bicycle system known as CitiBike has been available in New York City. The benefits to having such a system include reducing New Yorkers’ dependence on automobiles and encouraging public health through the exercise attained by cycling. Additionally, users who would otherwise spend money on public transit may find bicycling more economical – so long as they are aware of CitiBike’s pricing constraints. There are currently about 12,000 shared bikes which users can rent from about 750 docking stations located in Manhattan and in western portions of Brooklyn and Queens. A rider can pick up a bike at one station and return it at a different station. The system has been expanding each year, with increases in the number of bicycles available and expansion of the geographic footprint of docking stations. For planning purposes, the system operator needs to project future ridership in order to make good investments. The available usage data provides a wealth of information which can be mined to seek trends in usage. With such intelligence, the company would be better positioned to determine what actions might optimize its revenue stream.

- Because of weather, ridership is expected to be lower during the winter months, and on foul-weather days during the rest of the year, than on a warm and sunny summer day. Using the weather data, we can seek to model the relationship between bicycle ridership and fair/foul or hot/cold weather.
- What are the differences in rental patterns between annual members (presumably, local residents) vs. casual users (presumably, tourists?)
- Is there any significant relationship between the age and/or gender of the bicycle renter vs. the rental patterns?

2. Data Collection

We are leveraging two major sources of dataset for this scientific experiment

1. CitiBike trip dataset

CitiBike makes a vast amount of data available regarding system usage as well as sales of memberships and short-term passes. For each month since the system's inception, there is a file containing details of (almost) every trip. (Certain "trips" are omitted from the dataset. For example, if a user checks out a bike from a dock but then returns it within one minute, the system drops such a "trip" from the listing, as such "trips" are not interesting.) There are currently 77 monthly data files for the New York City bikeshare system, spanning July 2013 through November 2019. Each file contains a line for every trip. The number of trips per month varies from as few as 200,000 during winter months in the system's early days to more than 2 million trips this past summer. The total number of entries was more than 90 million, resulting in 17GB of data. Because of the computational limitations which this presented, we created samples of 1/1000 and 1/100 of the data. The samples were created deterministically, by subsetting the files on each 1000th (or, 100th) row.

2. Central Park daily weather data

Also we obtained historical weather information for 2013-2019 from the NCDC (National Climatic Data Center) by submitting an online request to https://www.ncdc.noaa.gov/cdo-web/search. Although the weather may vary slightly within New York City, we opted to use just the data associated with the Central Park observations as proxy for the entire city's weather. We believe that the above data provides a reasonable representation of the target population (all CitiBike rides) and the citywide weather.

3. Methodology

3.1 Descriptive Analytics:

- Analyze data distribution and data skewness
- Analyze feature correlations
- Analyze Feature Importance
- Analyze timeseries decomposition plots
- Seasonal and Trend timeseries analysis
- Analyze autocorrelation and partial autocorrelation

3.2 Data Correction

- Data Imputation
- Outlier removal

3.3 Feature Engineering

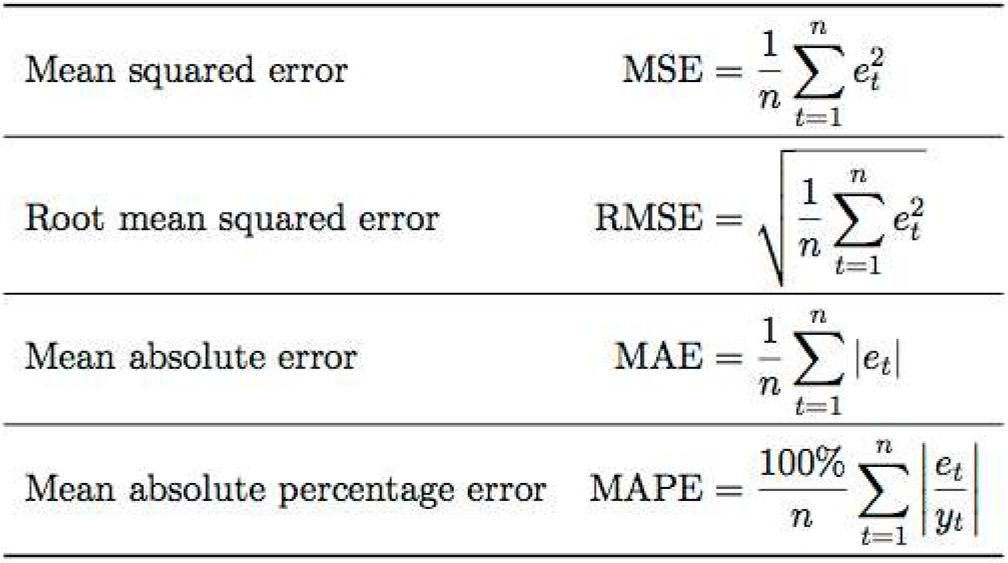
- Create important features from date
- Create Lagged Feature
- Create windowing features
- Create important features from weather data
- Apply timeseries smoothing functions
- Apply various feature transformation techniques (BoxCox, Log etc)

3.4 Model Building

- Build Multiple Regression model
- Build Regularized Regression Model (glmnet)
- Build Ensemble model Random Forest
- Build Ensemble model gradient boost
- Build nonparametric KNN Regressor
- Build nonparametric SVM Regressor
- Build Timeseries Forecasting model ARIMA
- Build Timeseries Forecasting model HoltWinters

- Build Timeseries Forecasting model ETS
- Build Ensemble of Timeseries Forecasting model

3.5 Optimization Functions



Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \left\lvert \frac{e_t}{y_t} \right\rvert$

4. Conclusion

4.1 Recommend best model for usage demand forecasting

4.2 Identify key features that influence future demand

4.3 Identify key features to enable important business decisions

4.4 Propose optimization areas for customer expansion and retention to increase business profitability

5. References

5.1 CitiBike website: https://www.citibikenyc.com/

5.2 CitiBike data description: https://www.citibikenyc.com/system-data

5.3 CitiBike trip data repository: https://s3.amazonaws.com/tripdata/index.html

5.4 National Climatic Data Center (NCDC) weather data: https://www.ncdc.noaa.gov/cdo-web/search