CitiBike Ridership as a Function of Weather

Data 621: Final Project - Group 4

Michael Yampol Ann Liu-Ferrara Sachid Deshmukh
Vishal Arora

12/20/2019

Introduction

- Since 2013 CitiBike has become increasingly popular in New York City for alternative transportation.
- We examine the relationship between number of daily trips across user types (Subscriber vs. Customer) and genders influenced by weather.
- We create predictions of future ridership given weather and other details, learning from past historic ride and weather data.

Literature Review

- We analyze more than a dozen articles examining bikeshare systems in various cities around the world, as far away as Shanghai, Vancouver, as well as the local NYC systems.
- According to these articles, Weather has varying influence on ridership, as other factors are also examined.
- Rider demographic is strongly significant, as riders tend to be overwhelming male and favoring certain age ranges.
- One key determinant of ridership is whether the user is an annual subscriber (and presumably local) vs a casual customer purchasing daily system access (presumably a tourist).

Data - CitiBike

- We obtained trip-level CitiBike ridership details from system inception in 2013 through the present (November 2019).
- This included a total of 77 monthly data files, where each row contains details on a single ride.
- There were a total of 90 million records spread across the files, which took up 17GB in disk space (unpacked).
- Because of the amount of individual trip data available, we subset the data using a deterministic algorithm which enables us to work with 1/1000 of the full dataset.

Data - Weather

- We also obtained daily weather observations recorded at Central Park for each day from 2013-2019
- This was freely obtained from the website of the National Climatic Data Center (NCDC), a US government agency.
- This data included measurements of daily max and min temperature, wind speed, precipitation/snowfall amounts, and other indicators such as fog, thunder, etc.

Data Exploration and Cleaning

- We examined the distribution of the major variables, such as user age, gender, membership type, and trip-related details such as total time the bike was used and estimated the distance traveled between start and end stations.
- We examined the relationship between weather-related variables, such as temperature and foul-weather indicators, to assess impact on trips taken.
- Certain data elements were missing or had outlier values which seemed unreasonable. We evaluated whether to truncate, censor or impute such values. The decision taken was truncation.

Data Preparation

- We prepared daily aggregated data on the CitiBike trip dataset, where the data was further grouped by gender and user type (Annual Subscriber or daily Customer).
- Within such grouping, we aggregated values including trip count, average user age, mean and median trip duration, total estimated distance traveled, and summation of user surcharges for lengthy trips.
- We removed variables (such as total estimated distance traveled) which exhibited extremely high correlation (0.97) with the target variable (trip count).
- We separated the data into TRAIN and TEST datasets by date: TRAIN included data from 2013 through 2018, and TEST included the subsequent data from the first 11 months of 2019.

Variable Selection: BORUTA

- BORUTA is a random-forest based algorithm which designates each candidate variable as important, unimportant, or tentative.
- The algorithm assessed 20 variables as “important” in prediction of the number of trips.
- We used these variables as the basis for the data frames supplied to various models.

Building Models

Standard Linear and Log-Linear models

- We build 3 set of linear models – 3 standard multiple linear regressions, and for each a corresponding log-linear regression on the target variable (trips taken) as this variable cannot be negative.
- This included models built with and without the DATE element, and a model omitting all variables which were assessed as not significant.

Generalized Linear Models (GLM)

- We created a pair of Poisson GLM models and a pair of Gaussian GLM models.
- Initial Poisson and Gaussian models were created from all the "important" variables in the dataframe.
- in each case, a subsequent reduced model was created by omitting those variables deemed not significant in the full model regression.

Results from the Models

Model quality was assessed using RMSE (lower is better):

Model_Name	RMSE
Linear Model 1: no dates	10.19
Log-Linear Model 1a: no dates	10.2
Linear Model 2: add dates	9.14
Log-Linear Model 2a: add dates	6.59
Linear Model 3: reduced to significant vars	9.13
Log-Linear Model 3a: reduced to significant vars	6.53
GLM model 1: Full Poisson Model	17.12
GLM model 2: Poisson reduced to significant vars	17.15
GLM model 3: Full Gaussian Model	9.05
GLM model 4: Gaussian reduced to significant vars	9.13

Explain and Compare Model Results

- Three Standard Linear models were compared; best result is **Linear Model 3** with RMSE=9.13
- Three Log-Linear models were compared; best result is **Log-Linear model 3a** with RMSE=6.53
- Four GLM models were compared; **GLM model 3: Full Gaussian model** has the lowest RMSE=9.05, but there are fewer variables in **GLM Model 4: Reduced Gaussian** , where the RMSE is almost the same (9.13), As this model is more parsimonious, it becomes preferred.
- Overall, the preferred model is **Log-Linear Model 3a** , with RMSE=6.53 .

Conclusion

From this analysis we illustrate the relationship of the preferred model and its strong impact on explaining the relationship between the variables in the model and the test results.

- The following variables have **positive** association with number of trips: DATE, gender=MALE, user_type=Subscriber, age, DailyMaximumTemperature
- The following variables have **negative** association with number of trips: gender=FEMALE(or, Unknown), user_type=Customer (or, Unknown), Precipitation, Snowfall, Depth of Snow on the ground, Fog(Indicator)

Discussion

Limitations in dataset collection and manipulations, and possibilities for further development:

- We have only daily weather observations. In the future, if we obtain hourly data, we could perform more granular and accurate predictive analysis.
- We used 1/1000 sample of the CitiBike data due to its very large size. Because the format of the CitiBike dataset is inconsistent from month to month, upload to AWS Athena was not successful.
- We used numeric value of DATE as one of the predictive variable. The positive coefficient on DATA reflects the trend in increased ridership over time.
- A future enhancement would be to model seasonality, which could be achieved using monthly dummies or a “Winter” vs. “Summer” indicator.

References

- CitiBike website: https://www.citibikenyc.com/
- CitiBike data description: https://www.citibikenyc.com/system-data
- CitiBike trip data repository: https://s3.amazonaws.com/tripdata/index.html
- National Climatic Data Center (NCDC) weather data : https://www.ncdc.noaa.gov/cdo-web/search