Taxi demand prediction using linear regression

Case study: Wall street, NYC

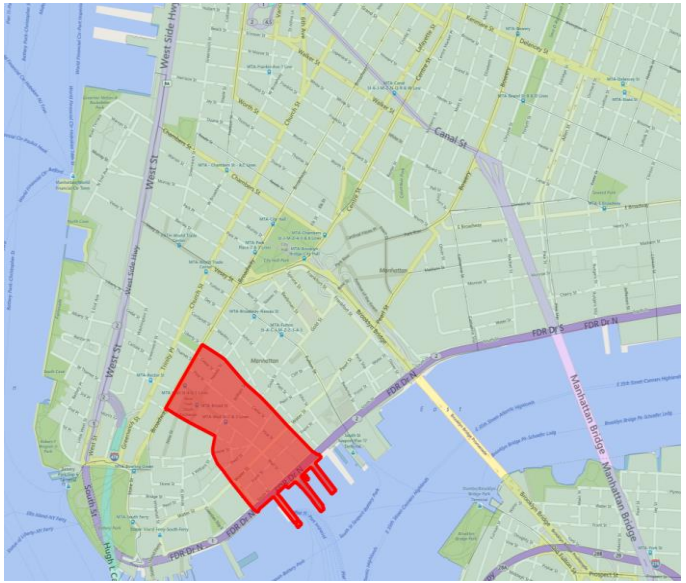
Aliasghar Mehdizadeh Dastjerdi

Abstract

This study developed a linear regression model to predict taxi demand in Wall street neighborhood, NYC. Two dataset of taxi pick up demand and weather information were used to address the objectives of the study. This study found that, (i) linear regression is an appropriate method for such predictive analysis , and (ii) including a lagged variable in the predictive model improves its performance.

Motivation

Some big cities, e.g. New York City, suffer from a lack of balance between supply and demand in the taxi service. This causes a number of issues such as increasing passenger waiting time, increasing total length and time of empty taxi roaming and accordingly, rising fuel consumption and air pollutants. Therefore, it is of importance to understand taxi supply and demand in order to enhance the efficiency of the city's taxi system. In this context, taxi demand prediction could provide valuable insights to both city planners and taxi dispatchers by addressing relevant questions including, (i) how to position cabs where they are most needed, (ii) how many taxis to dispatch, and (iii) how taxi demand varies over time. This study aims at predicting the number of taxi pickups given a one-hour time window within Wall Street, NYC.



Dataset(s)

Two main dataset were merged in this study for data analysis and modelling resulted in 65712 observations (rows) and 15 variables (columns).

- Yellow/green Cab (Taxi pick up demand from 2009 till 2016): http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
- Weather information : NOAA climate data website: http://www.ncdc.noaa.gov/cdo-web/

Variables are categorized in three classes:

1. Time stamp of taxi pickups in one hour window
2. Number of pickups
3. Weather condition (exp., min & max temp, wind speed, visibility, pressure, precipitation etc.)

Data Preparation and Cleaning

In this study data preparation and cleaning can be summarized as follow:

1. Extracting/isolating observations within wall street neighborhood using geographic coordinate system
2. Merging two dataset using a common key between them, i.e., “wban” : station id
3. Grouping taxi pickup demand from 15 mins into 1 hour interval.
4. Integrating “time of day” (0,1, … 23) and “day of week” (Monday , … , Saturday) as dummy variables into model
5. Adding a new column called “lagged_pickups” for modeling with the following linear equation. It means that taxi pickups in time *t* also depends on taxi pickups in time *t -1.*

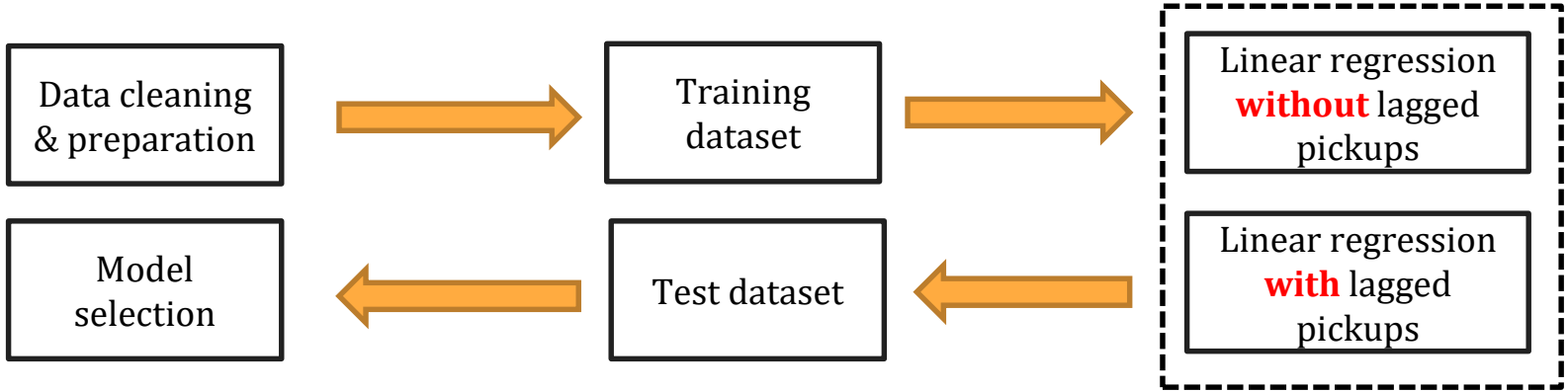
$$pickups_t = \beta_0 + \beta_1 * pickups_{t-1} + \sum_j \beta_j * Weather\_variable_j + \sum_k \beta_k * day\_of\_week_k + \sum_m \beta_m * time\_of\_day_m$$

Research Question(s)

This study investigates the following research questions:

- What is the number of taxi pickups in an hour interval in Wall street neighborhood?
- Does including the lagged_variable improve the model accuracy and its predictive power?

To address the research questions, we will take the following steps:

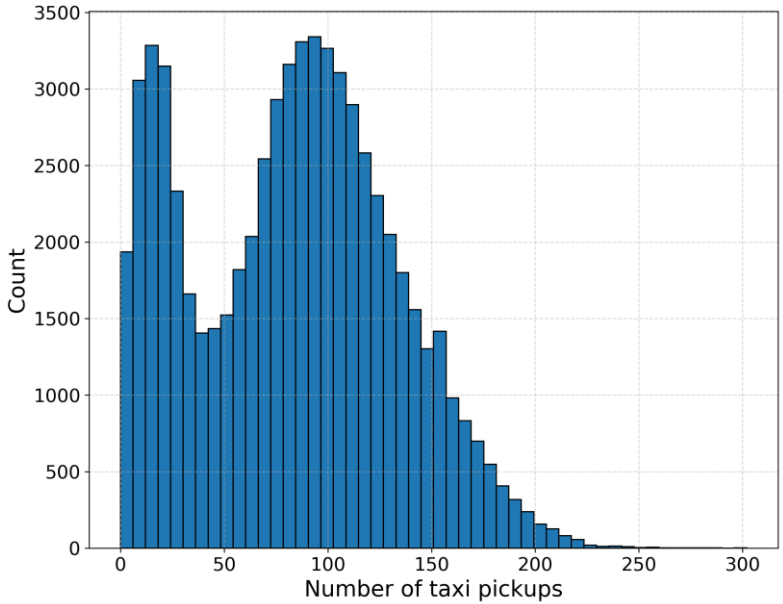


Methods

This study employs linear regression model as a supervised learning method to predict the number of taxi pickups. This is an appropriate method since taxi pickups as the target variable is known and continuous in nature.

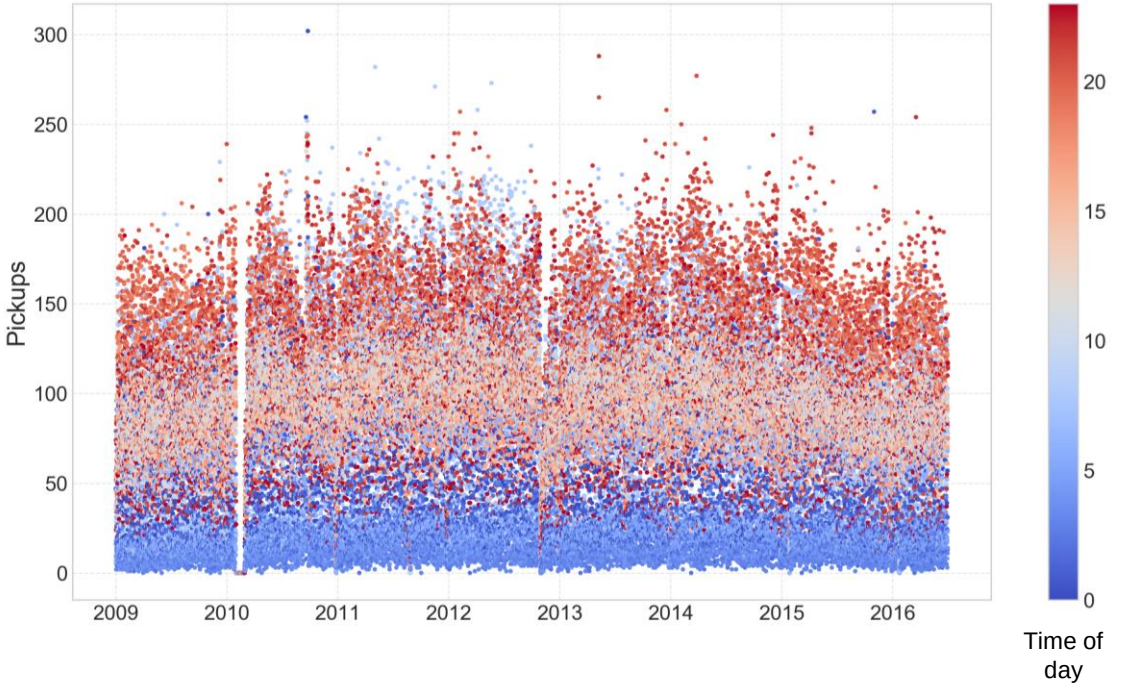
Findings

- This figure shows the histogram for the number of taxi pickups.
- The histogram presents two peeks, possibly because the data comes from two different distributions.
- It might related to the difference between the number of pickups during weekdays and weekends or peek and off-peek hours.
- It needs to be investigated.



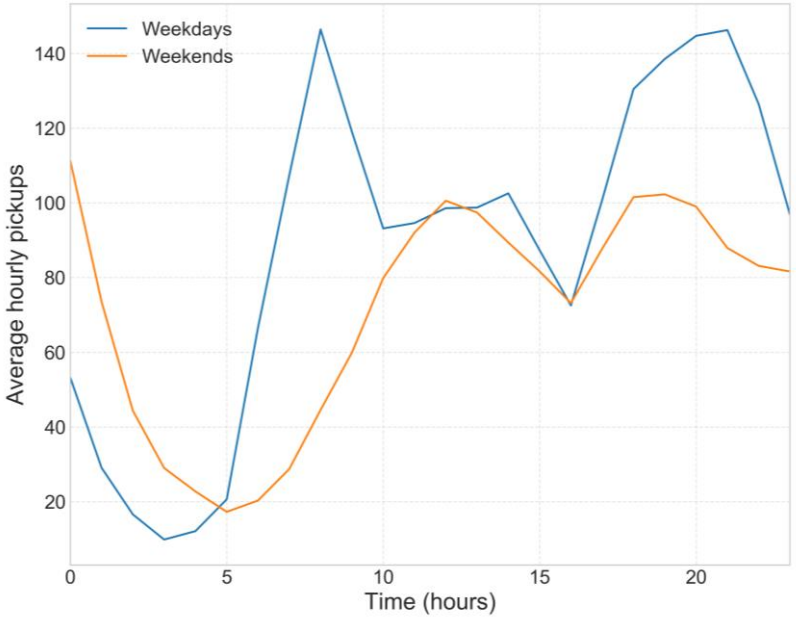
Findings

- The scatter plot shows the relationship between taxi pickups and time of day (12 AM, 1 AM,… , 23 PM) from 2009 till mid 2016.
- There is a clear trend showing, from early morning to late evening the demand for taxi service increases i.e. more reddish points on the top .
- It indicates that time of day is an important predictor of demand for taxi service in the zone.



Findings

- The plot shows how average hourly pickups varies from 12:00 AM to 23:00 PM.
- The plot shows different pattern for the number of pickups over weekdays and weekends.
- It indicates that day of week plays an important role to predict taxi demand in Wall street neighborhood.



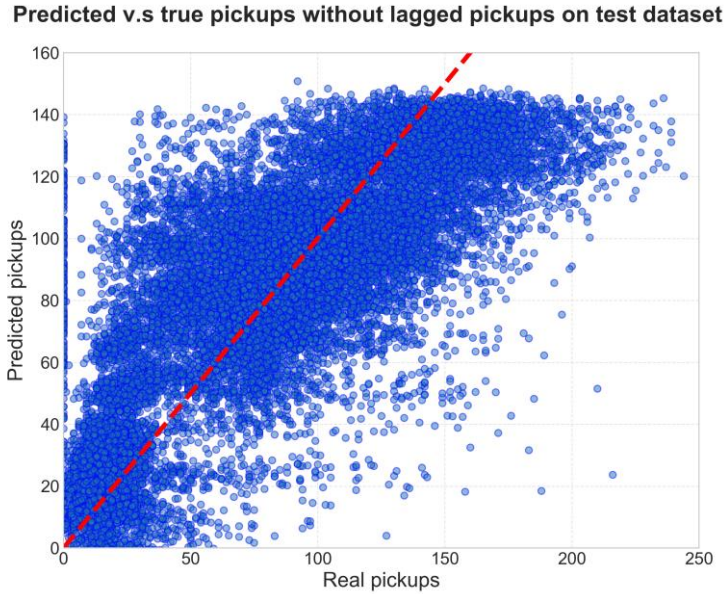
Findings

- The table compares the correlation coefficient, MAE, RMSE, and R-Squared metrics of two regression models in order to evaluate their prediction error rates and model performance.
- Model 1, was established without the lagged variable while in Model 2 , lagged pickups was included.
- Model 2 is preferred since it provides lower error rates and higher model performance.

Accuracy indices	Model 1	Model 2
Pearson correlation coefficient	0.771	0.919
Mean absolute error (MAE)	23.450	14.551
Root mean square error	31.057	19.190
R-Squared	0.595	0.844

Findings

- The plots compare the predictive power of Model 1 and Model 2.
- The distance of a point from the red line (ideal 45-degree angle line) highlights how well or how poorly the prediction performed. Therefore, in line with the results of the table:
- Incorporating the lagged variable into the model increased its performance .



Limitations

- This study assumed that taxi pickups comes from Gaussian distribution and hence, we used Gaussian linear model. It might be problematic, since pickups are count data and always positive. Future study should address this issue by implementing Poisson regression model.
- This study only included the first lagged pickups into the model. Future study should investigate to what extent including the second, third, …. lagged variables improves the model performance.

Conclusions

Overall, linear regression model for predicting taxi pickups in Wall street, NYC , performed well. Including a lagged variable in the regression model improves its accuracy and performance by achieving a value of 14.551, 19.190 and 0.844 for MAE, RMSE and R^2 respectively. The model could be useful to city planners and taxi dispatchers in studying patterns in taxi ridership.

Acknowledgements

I would like to acknowledge that no one provided me with any feedback to improve this assignment. However, there are a number of online data science projects completed on NYC taxi dataset. I was inspired by them. Two of them are referred in the next page.

References

1. Gong, Y., Fang, B., Zhang, S. & Zhang, J. (2016). Predict New York City Taxi Demand. Viewed 20 January 2020, < https://nycdatascience.com/blog/student-works/predict-new-york-city-taxi-demand/>

2. Sing, R. (2019). Taxi demand prediction in New York City. Viewed 21January 2020, <https://medium.com/@ranasinghiitkgp/taxi-demand-prediction-in-new-york-city-916cde6a3492>