Report

Oleksandra Kharchenko, Serik Tashanov, Quoc Tuan Nguyen

December 9, 2022

1 Introduction

New York City Bike share system (hereinafter BSS) is the largest bike program in the USA that is working 24/7, every day of the year. In Manhattan, Brooklyn, Queens, the Bronx, Jersey City, and Hoboken, it has more than 25,000 bikes and more than 1,500 stations (New York Citi Bike). The bike-sharing program has grown steadily since its beginning in 2013 and became one of the crucial elements of the overall mobility system throughout New York City.

As the rapid spread of the coronavirus disease (COVID-19) has become a global pandemic that disrupted urban mobility due to government lockdowns, BSS has attracted significant attention by acting as an alternative mode of transportation for city residents (Pase, 2020). New York City commuters increased their use of bike sharing systems in preference to the subway during the global pandemic (Lopes, 2020). The demand for bikes increased dramatically which led to the problem for the City Bike system to keep up with such extension. People were complaining about the lack of free bikes or locks on the stations (New York Times, 2021). The company's capacity was not able to properly serve high customer demand during the period. BSS is in need of predicting future demand after the pandemic periods.

Despite the growing research interest in influential factors of BSS usage, we still lack a comprehensive understanding of how an exogenous shock such as the global pandemic may have altered how such factors can predict bike usage. Most of the previous work was focused on predicting trips by using calendar along with weather features. Rogozhin (2020) in his project created a model to predict bike-share usage and compare his prediction to actual ridership for 2020. However, he did not use other features that can influence the number of trips. Divya Singhvi and Eoin O' Mahony (2015) included population and taxi usage to the weather conditions and predict bike usage during morning rush hours. Morency (2022) used short-term prediction, in a 15 min horizon of bike-sharing to predict the demand during the COVID. The majority of the study on COVID-19's effects on BSS focused on the variables that affect whether bike-sharing ridership rises or falls (e.g Jeffrey Jobe (2021); Lopes (2020)), or changes in the demand for BSS during the global pandemic (Chibwe, 2021).

Assessing different studies, we have found that there is no efficient model that takes into account different features related to traffic, environment, economics and COVID-19 for making predictions after the pandemic. These features can increase the accuracy of the model which may lead to a more realistic prediction. Thus, the goal of this report is to find the optimal combination of features that could describe the bike demand during the global pandemic and the machine learning model that can give the highest accuracy, so that this project tends to give a comprehensive picture of the factors affecting the bike-sharing demand in New York after the pandemic.

2 Methods

The data observed was taken from January 2014 to December 2021. The bike usage data was collected from the BSS webpage (see `Data_Downloading.ipynb`) (New York Citi Bike). Each downloaded file contains trip duration, start/stop time, start/end station id, start/end station longitude and latitude, bikeid, usertype, birth year and gender. For our project, we need to calculate the total number of trips each day for each year. Therefore, we combined all years together in one dataset and calculated the total number of trips each day.(see `Data_Downloading.ipynb`) Figure 1 shows the changes in the number of trips throughout the years.

General features

Taking into account the specific properties of using bikes as means of transportation and considering the analysis of previous works (Jiang, 2021), we have solid groundings to include features such as traffic level and environmental factors measurements as one of the driving factors in building successful models.

Traffic indicators. The hourly traffic on Metropolitan Transportation Authority (MTA) Bridges and Tunnels data was obtained from DATA.NY.GOV webpage (New York State). The file contains the plaza id, where the observations were taken on an hourly basis, the direction of traffic: I = Inbound; O = Outbound, number of vehicles passing through each bridge or tunnel's toll plaza with and without using E-ZPass (a toll collection system). For further analysis, we calculated the total number of cars that went in and out of the New-York each day as a traffic indicator.

Environmental factors. For pollution analysis, we used three pollutant indicators: the level of SO2, CO and Ozone. The monthly data for New York-Newark-Jersey City was downloaded from epa.gov (United States Environmental Protection Agency). We combined three factors together and got daily pollutant indicators for a given period in one dataset. Daily data on weather conditions in NYC was obtained from the National Climatic Data Center (National center for environmental information). Factors such as maximum, and minimum temperature, average wind speed, precipitation, snow, and snow depth were further used in our model as weather measurements.

Economic, industry and COVID-19 specific data

As we know from history, huge economic and social crises are hard to predict because these events bring new learning points and imply that existing predicting technologies have weaknesses. So the COVID-19 pandemic has revealed insights into how global and industry trends are going to be shaped in response to social reactions. Thus, starting from the 22nd of March 2020 (Wikipedia, 2021), mobility within New York was limited by governmental authorities. In order to bring this effect into our model, we selected specific indicators on the quantity of people affected by COVID -19 as well as industry trends on transportation and global trends.

Tests on COVID-19. The daily number of COVID-19 cases was taken from the NYC OpenData website (New York Open Data). Originally, the table had a count of COVID-19 patients who were hospitalized, the count of deaths occurring among confirmed COVID-19, total case count. For further analysis, we took into account the number of people that get positive COVID tests.

Overall and industry trends. Oil prices were taken daily from macrotrends website (Macrotrends). We took this indicator because the price of oil during COVID-19 period decreased sharply as a reaction to artificially limited transportation within the world. Figure 3 shows the average oil price for the period

Downloaded from finance.yahoo, the SP 500 indicator can reveal the current situation in the market (Yahoo). We assume that overall economic factors may have an impact on BSS.

The employment rate was received from FRED Economic Research (FRED Economic Research). However, we got the employment rate for each month instead of the everyday rate. Therefore, we computed the approximate level of employment for each day using the formula daily $n = n_t + z*(n_{t+1} - n_t)$, where z is the number of a day, n_t the employment rate in month t and n_{t+1} the employment rate in month t+1. Figure 2 shows the changes in employment during the period.

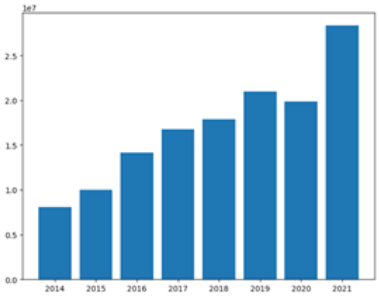

Figure 1: The total number of trips each year (x10^7)

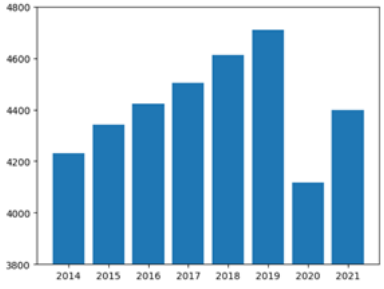

Figure 2: Employment rate for each year (x10^3)

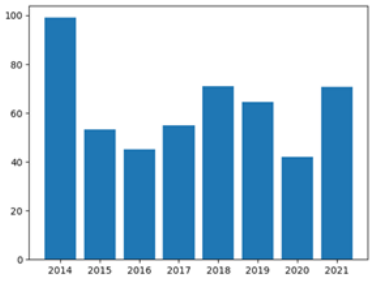

Figure 3: Average oil price for each year

Gathering all features together, we received one dataset with 79 features that potentially can describe the bike demand (see `Data_preparation_and_exploration.ipynb`). Further, they will be

used for feature engineering.

3 Results

3.1 Data exploration (see `Data_preparation_and_exploration.ipynb`)

Calendar features

By analyzing the data related to the calendar, we tried to visualize yearly, seasonal and weekday trends. We can observe the fact of existence the above-mentioned trends (Figures 4 and 5). Yet, 2020-year looks like an outlier in the yearly trend from 2014 to 2021. On a monthly graph, we can explore the pretty shaped seasonality effect on daily trips. Thus, bikes are more popular during the summer period. In addition, people tend to use bikes more during working days than at weekends.

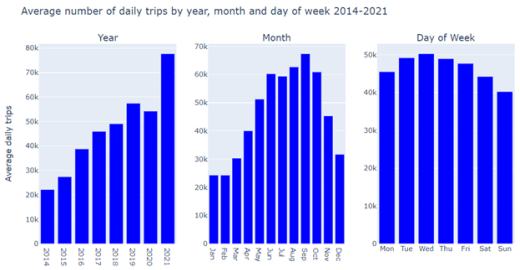

Figure 4: Daily, monthly, yearly number of trips

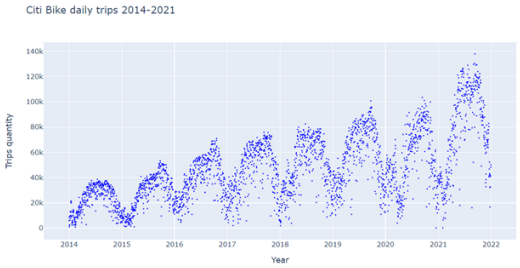

Figure 5: BSS usage throughout the period

Weather effect

We could assume that the weather is one of the driving forces that can shape the seasonality effect. As we can observe from Figure 6 below, the shape of the "daily trips" curve can imitate the trend of average temperature. Thus, we believe that along with the calendar, weather features can become one of the key inputs for the high-performing model.

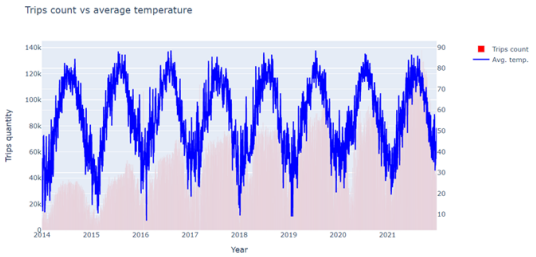

Figure 6: Trips count vs average temperature

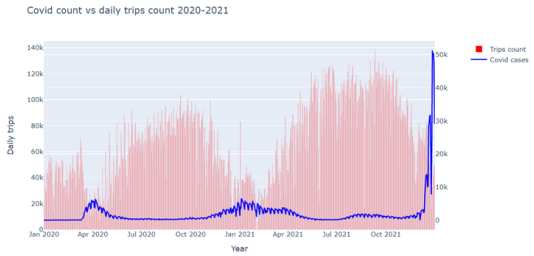

Figure 7: COVID-19 count vs daily trips

In addition, we made the following encodings:

1. Extracted year, weekday, month and day from date (along with sine and cosine transformation) to capture the yearly, monthly and weekends cycles and trends.
2. Added US holidays as binary (0/1) to capture the effect of holidays.
3. We took logs of precipitation, snowfall and snow depth in order to better capture weather conditions, that is, at least some precipitation or snow will yield better results on our model.

COVID-19 related factors

Despite the fact that the seasonality effect usually starts to increase the number of ridership in March, the number of ridership in April 2020 started to dramatically decrease. According to Figure 7, we can observe how the number of COVID-19 cases can affect the number of daily trips. With a high number of COVID cases-19 in April 2020, the tendency of ridership started to decrease. As a result, COVID-19 data is included in the model as a feature to take its impact on ridership in 2020 into consideration.

3.2 Model development (see `Models.ipynb`)

Train and testing split

The main difficulty with machine learning is that the algorithm must work well on new, untrained inputs as well as those that our model was trained on (Courville, 2016). The train-test split approach is used to gauge how well machine learning algorithms perform when used to predict outcomes from data that was not used to train the model. Its objective is to test the robustness of the model.

It could not make sense to randomly divide the data into groups and attempt to use future data to forecast previous data because the data is time series (meaning it is indexed in the order of time). Data should not shuffle at random to start, but a test set should be placed aside (Rafferty, 2021). Thus, the observations were divided into a training set, a validation set, and a test set: data from 2014 to 2020 is for the training set, while for 2021, we will use 75 percent of data (before week 40 of 2021) for the validation set, and the last 25 percent will be used for the test set. For feature selection and prediction, the training set was utilized. The validation set will be used for hyperparameter tuning. For the evaluation and forecasting of bike rentals, we used the test set.

Model selection

Previously developed models (see `Data_preparation_and_exploration.ipynb`) As it was previously stated in the introduction, unexpected global issues and challenges have revealed weaknesses in previous knowledge and brought new learning points. Models developed Rogozhin (2020) expectedly were not able to perform well enough after COVID-19 period. These models, which are based only on weather and calendar features, have scored 51%, 38% and 40% (r-square) on Linear, Polynomial and Ridge regressions for "after covid" period (Figure 8).

Models	Train < 2019, Test = 2019	Train < 2021, Test = 2021
Linear Regression	('R2= 0.77', 'MAE = 8756.0', 'RMSE = 10615.0')	('R2= 0.51', 'MAE = 20027.0', 'RMSE = 23342.0')
Ridge Regression	('R2= 0.89', 'MAE = 5500.0', 'RMSE = 7326.0')	('R2= 0.4', 'MAE = 22365.0', 'RMSE = 25706.0')

Figure 8: Results for Post-COVID period

Model selection using AutoML (see `Auto_ML.ipynb`)

Because of the change, we decided to select models again. A cloud-based tool called Azure AutoML may be used to automatically create machine learning pipelines for activities like categorization, regression analysis, and forecasting. Its objectives include choosing which model to use and how to pre-process the input dataset in addition to tuning the hyper-parameters of a particular model (Microsoft). This one could be used to give the general idea for model choosing and how to process the input data.

Firstly, we performed many regression algorithms on all features to predict bike rentals in New York. All of these models were performed at Azure Auto Machine Learning to find some of the best models for the datasets. The number of models tested is 12 (Microsoft). Initially, we put all features into Auto ML without doing feature selection. The accuracy for some of them is as below (Table 1).

Table 1: The AutoML results

Model name	R^2 **for training data**	R^2 **for training data**
1 MaxAbsScaler, LightGBM	0.95	0.5
2 MaxAbsScaler, XGBoostRegressor	0.94	0.51
3 StandardScalerWrapper, XGBoostRegressor	0.92	0.36
4 SparseNormalizer, XGBoostRegressor	0.91	0.04
5 StandardScalerWrapper, ExtremeRandomTrees	0.89	0.3
6 MaxAbsScaler, ElasticNet	0.87	-0.68

From that results, we chose to test the data on the two models: lightGBM, and XGBoostRegressor. LightGBM is a gradient-boosting framework that uses tree-based learning algorithms. It is designed to be distributed and efficient with the following advantages: faster-training speed, higher efficiency, lower memory usage, and capability of handling large-scale data (LightGBM). This model, however, might be delicate to the tiny amount of data. XGBoost is also an implementation of the gradient tree boosting algorithm that is widely recognized for its efficiency and predictive accuracy (XGBoost). While XGBoost uses a pre-sorted method and a histogram-based approach to get the optimal split, LightGBM employs a revolutionary technique called Gradient-based One-Side Sampling (GOSS) to filter out the data instances for determining a split value.

The results from Table 1 suggested the effective scaling method is MaxAbsScaler, which scales each feature by its maximum absolute value. Though the hyperparameter tuning is automatically done by Azure AutoML, through the experiment, the results could be better if we do it manually.

While the tree-based models outperformed, it was decided to test Ridge, Lasso and Linear regressions because they showed good results for data before the pandemic. Lasso is one of the types of linear regression that uses the shrinkage method. The process of shrinking data values toward a central tendency, such as the mean, is referred to as shrinkage. Ridge regression is similar to Lasso but shrinks the value towards squares of the mean.

In summary, we selected the following machine learning models for further consideration: LightGBM, XGBoostRegressor, Lasso, Ridge and linear regression.

Feature Selection

XGBoost and LGBM

Though we found a lot of features that could affect the number of trips, it is important to select features carefully to increase the model accuracy and avoid overfitting. We proceeded to use Recursive feature elimination with cross-validation (RFECV) to select features. The importance of each feature is first determined using any particular attribute (such as `coef_, feature importances_`), after which the estimator is trained on the initial set of features. The least significant features are then removed from the existing set of features. The subset of candidate inputs, which resulted in the best prediction performance, was selected for each model development (Scikit-learn).

For the XGBoost model, the optimal number of features is 21 (Figure 9), and they are: Year, Week, Day of Week, Day of Year, traffic volume, covid count, Employment, S&P price, S&P volume, SO2, PRCP, SNWD, TMAX, TMIN, sin doy, cos doy, sin dow, cos dow, dow 4, month 7, holiday. For the LightGBM model, the RFECV choose 28 features (Figure 10), and they are: Year, Month, Week, Day, Day of Week, Day of Year, traffic volume, covid count, employment, oil prices, S&P price, S&P volume, CO, SO2, Ozone, AWND, PRCP, SNWD, TMAX, TMIN, sin doy, cos doy, sin dow, cos dow, month 4, day 18, day 30, holiday.

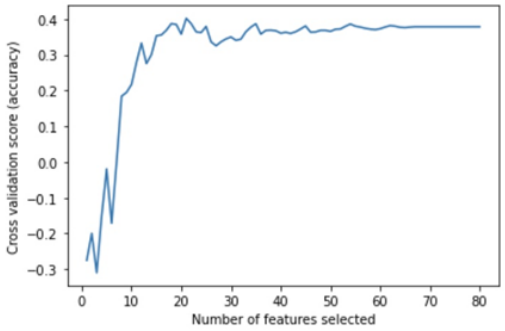

Figure 9: RFECV for XGBoost

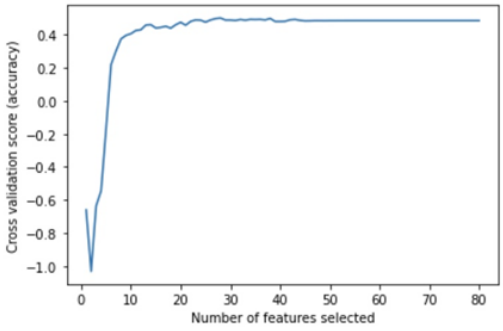

Figure 10: RFECV for Light GBM

Linear Regressions

Firstly, for linear models, we try to use RFECV and feature importance. However, such methods yield low results. By trial and error, we decide to choose 15 features that can logically affect the number of trips such as employment, oil prices, S&P price, SNWD, PRCP, TMAX, TMIN, sin doy, sin dow, SO2, CO, S&P volume, traffic volume, day of week.

Optimization of model hyperparameters

The performance of the models can be impacted by hyperparameters, which are critical to the

outcome of machine learning algorithms (Courville, 2016). There are multiple hyperparameter tuning techniques that can be used in various situations, including manual tuning, random search, and grid search. However, we chose to manually tune the model for finding the numbers of the combination of hyperparameters. For hyperparameter tuning, we use the validation set.

Evaluation method

An R-squared score, which measures how far the expected results deviate from the observed data, can be provided by the machine learning model when it makes a prediction (GeeksforGeeks, 2022). No single R-squared score can be used to judge whether a model is "excellent" or not. Instead, the R^2 score's effectiveness changes depending on the model's objectives. We also use the Root mean square error (RMSE), and the mean absolute error (MAE) to evaluate the performance of each model. The best predictive capacity is thought to be demonstrated by the model with the highest R-squared, lowest RMSE, and MAE score. The process of our pipeline could be described below. By trial and error, we tested many combination of features as well as hyperparameters to find the optimal results so far.

Model performance and outcome

The prediction accuracy of the ML models is shown in Figure 11. As seen in Figure 12, there is some variation between the anticipated and real number of rides per day, but the forecast is pretty close to the actual number of rides. Initially, the R^2 value of the prediction was calculated, as was previously described, to ascertain how accurate the prediction is. The XGBoost seems to be more robust than other models.

The Table 2 represents the determined R^2 value

Table 2: The results for tree-based models

Model name	R^2 **for training data**	**Results for testing data**
1 XGBoost	0.73	R^2 = 0.73; RMSE = 12540; MAE = 11011
2 LightGBM	0.65	R^2 = 0.53; RMSE = 16521.24; MAE = 14799.18

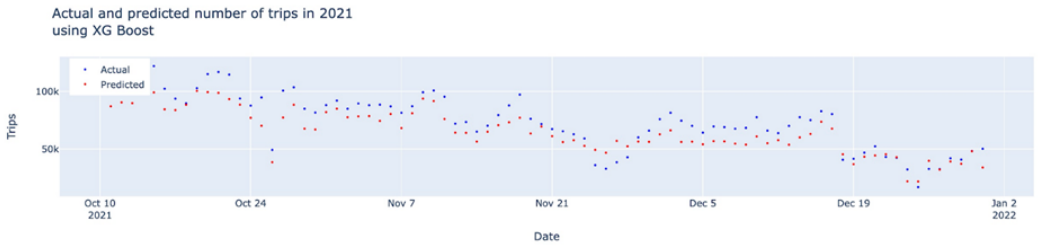

Figure 11: Actual and predicted number of trips in 2021 using XG Boost

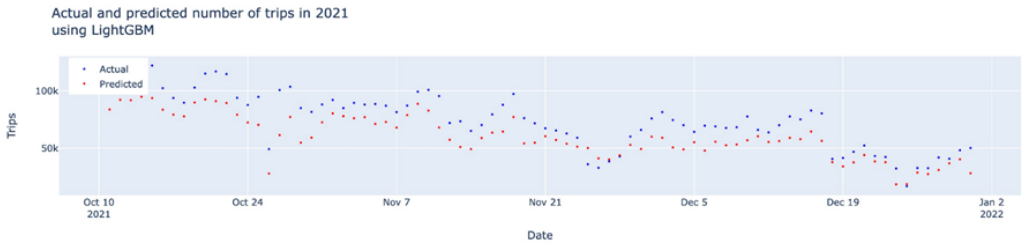

Figure 12: Actual and predicted number of trips in 2021 using LightGBM

For linear models, we received Table 3. It showed almost the same low performance for these models.

Table 3: The results for linear models

Model name	R^2 **for training data**	**MSE**
1 Linear Regression	0.27236	16787.56
2 Lasso	0.27233	16787.95
3 Ridge	0.27221	16789.55

4 Conclusion

In this paper, we pursued to find the optimal model and its appropriate features to predict the demand for bike sharing system in New York after COVID-19. In the process of features analysis, we found that during such a severe period, global economic and local indicators enrich the model with valuable metrics to perform better. As a result, S&P 500, oil prices, employment rate, traffic volume and pollution level became a good complement to weather and calendar features that were extensively used by previous 'predictors'.

The bike usage dataset was collected from New York Citi Bike, and divided into training, testing and validation. The RFECV algorithm was applied to select the most strongly representative features for the tree-based models. Yet, for linear models, we used the trial and error method to select the most relevant features. Selected features were fed into many algorithms: XGBoost, LightGBM, Linear, Ridge and Lasso regression. The parameters of the tree-based models were tuned to perform the optimal regression so that algorithms reached promising results. The XGBoost algorithm outperformed all other algorithms, achieving an R^2 and MAE of 0.73 and 11,011 respectively.

During this study, we confront several limitations. Firstly, it is a lack of publicly available data on some aspects that could have a crucial impact on our models. We were not able to find metrics related to mobility such as taxi and subway usage, and daily population growth. Secondly, we do not have access to precise business metrics related to operations. Exact information on the company's capacity, logistics and limitations could have revealed insightful information for a better model and feature engineering.

The analysis could reveal that COVID-19 has made both past experiences and established models less applicable in current realities. Thus, the usage of past models cannot guarantee an effective prediction of bicycle usage. As a solution, many stakeholders and shareholders of bike-sharing systems over the world can use our model to better understand how bike usage can be predicted in a highly volatile and unstable period of history.

5 References

Chibwe, J., Heydari, Shahram, Faghih Imani, Ahmadreza and Scurtu, Aneta. (2021). An exploratory analysis of the trend in the demand for the London bike-sharing system: From London Olympics to Covid-19 pandemic. https://doi.org/https://doi.org/10.1016/j.scs.2021.102871

Courville, I. G. a. Y. B. a. A. (2016). Deep Learning. https://www.deeplearningbook.org/
Divya Singhvi, S. S., Peter I. Frazier, Shane G. Henderson,, & Eoin O' Mahony, D. B. S., Dawn B. Woodard. (2015). Predicting Bike Usage for New York City's Bike Sharing System. /lhttps://people.orie.cornell.edu/shane/pubs/AAAI15.pdf
FRED Economic Research. All Employees: Total Nonfarm in New York City. https://fred.stlouisfed.org/series/SMS369356100000000010
GeeksforGeeks. (2022). Python – Coefficient of Determination-R2 score. https://www.geeksforgeeks.org/python-coefficient-of-determination-r2-score/
Jeffrey Jobe, G. P. G. (2021). Bike share responses to COVID-19. Transportation Research Interdisciplinary Perspectives. https://doi.org/https://doi.org/10.1016/j.trip.2021.100353
Jiang, W. (2021). Bike sharing usage prediction with deep learning: a survey. https://link.springer.com/content/pdf/10.1007/s00521-022-07380-5.pdf?pdf=button
LightGBM. https://lightgbm.readthedocs.io/en/v3.3.2/
Lopes, J. F. T. a. M. (2020). The link between bike sharing and subway use during the COVID-19 pandemic: The case-study of New York's Citi Bike. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7345406/
Macrotrends. https://www.macrotrends.net/2480/brent-crude-oil-prices-10-year-daily-chart
Microsoft. Set up AutoML training with the Azure ML Python SDK v2. https://learn.microsoft.com/en-us/azure/machine-learning/how-to-configure-auto-trainsupported-algorithms
Microsoft. What is automated machine learning (AutoML)? https://learn.microsoft.com/en-us/azure/machine-learning/concept-automated-ml
Morency, A. M. D. a. C. (2022). Bike-Sharing Demand Prediction at Community Level under COVID-19 Using Deep Learning. https://doi.org/https://doi.org/10.3390/s22031060
National Centers for Environmental Information. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00094728/detail
New York Citi Bike. How it works. https://citibikenyc.com/how-it-works
New York Citi Bike.System data. https://ride.citibikenyc.com/system-data
New York Open Data. COVIC -19 Daily Counts of Cases Hospitalizations https://data.cityofnewyork.us/Health/COVID-19-Daily-Counts-of-Cases-Hospitalizations-an/rc75-m7u3
New York State. Hourly Traffic on Metropolitan Transport https://data.ny.gov/Transportation/Hourly-Traffic-on-Metropolitan-Transportation-Auth/qzve-kjga
New York Times (2021). N.Y.C.'s Bike Parking Problem: 1.6 Million Riders and Just 56,000 Spots. https://www.nytimes.com/2021/01/26/nyregion/bike-parking-nyc.html
Pase, F., Chiariotti, F., Zanella, A., Zorzi, M. (2020). Bike Sharing and Urban Mobility in a Post-Pandemic World. https://doi.org/https://doi.org/10.1109/ACCESS.2020.3030841
Rafferty, G. (2021). Forecasting Time Series Data with Facebook Prophet: Build, improve, and optimize time series forecasting models using the advanced forecasting tool.
Rogozhin, S. (2020). Analysis and prediction of Citi Bike usage in the unpredictable 2020. https://medium.com/towards-data-science/analysis-and-prediction-of-citi-bike-usage-in-the-unpredictable-2020-3401da26881b
Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.feature$_s$*election.RFE.html*
United States Environmental Protection Agency. https://www.epa.gov/outdoor-air-quality-data/download-daily-data
Wikipedia. (2021). COVID-19 pandemic in New York City. Retrieved 28/11/2022 from `https://en.wikipedia.org/wiki/COVID-19_pandemic_in_New_York_CityStay-at-home_order`
XGBoost. https://xgboost.readthedocs.io/en/stable/
Yahoo. `https://finance.yahoo.com/quote/%5EGSPC/history?period1=1388534400&period2=1640908800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true&guccounter=2`