Predicting Citi Bike Demand in NYC: A Machine Learning Approach

Artem Pankin, Hunter College, CUNY

Abstract

Demand forecasting in bike-sharing systems (BSS) is crucial for enhancing urban mobility planning, resource allocation, and service reliability. This study proposes a machine learning (ML) approach to predict the daily number of Citi Bike trips in New York City, utilizing historical trip-level data (2014-2025), weather observations, and calendar features. The dataset was processed and aggregated using DuckDB and Python, incorporating lagged and rolling average features to account for temporal autocorrelation in BSS usage. Two models were developed and compared: Lasso Regression and XGBoost. Evaluation metrics showed that XGBoost significantly outperformed the linear baseline, achieving a MAPE of 13.64%. The most influential predictors were temporal aggregates and weather characteristics such as temperature and precipitation. While the model provides reliable citywide forecasts, it does not account for spatial heterogeneity or intraday variation. These limitations point to future research using hourly station-level data and incorporating additional contextual variables. The proposed model provides a valuable tool for policymakers and BSS operators to anticipate daily demand and optimize system performance based on accessible weather forecast data.

Keywords: Transportation, Bike Sharing System, Citi Bike, Machine Learning, XGBoost

1. Introduction

The field of mobility scholarship has historically focused on understanding the behavioral variables that influence the transportation mode choices of the population (Meixell & Norbis, 2008; Singhal et al., 2014a). Specifically, bicycle-focused scholarship has devoted considerable space to analyzing how weather conditions affect bicycle ridership (de Kruijf et al., 2021; Singhal et al., 2014; Wessel, 2020; Zhao et al., 2018). The consensus in the field is that as weather conditions worsen (e.g., lower temperatures, precipitation), fewer people tend to choose biking as a mode of transportation, while good weather conditions are typically associated with an increase in bike ridership.

While data availability makes it difficult to study mobility choices related to private vehicles, researchers have begun to focus on rental services, including bike sharing systems (BSS), such as CitiBike in New York City. This mode of transport has become particularly relevant as it

has been integrated into urban policy as both a climate and mobility solution, due to its convenience for users and low carbon emissions compared to cars or public transport. However, its unpredictable demand complicates fleet deployment, revenue planning, multimodal integration, and policy evaluation. To fill this gap, I develop a machine learning (ML) forecasting tool that predicts daily trip counts across the CitiBike network using readily available weather data and minimal preprocessing. I compare two popular algorithms, Lasso Regression and XGBoost, while accounting for multicollinearity among weather variables to identify the most accurate and robust model.

Specifically, the goals of the project are to: (1) collect and clean daily CitiBike ridership and meteorological data for New York City, (2) develop ridership prediction models, and (3) evaluate the quality of different prediction models to determine the one that provides the most accurate result. Achieving these goals will allow us to create a model that can both contribute to the science of BSS ridership forecasting and be useful for urban planning, policy, and development purposes.

This paper is structured as follows: Section 2 reviews BSS demand modeling scholarship, including role of weather conditions for prediction; Section 3 details data and preprocessing and describes the ML methods and multicollinearity treatment; Section 4 presents performance results; and Section 5 discusses implications and future work with a particular focus on New York and CitiBike data.

2. Literature review

Research in the field of BSS demand forecasting has evolved rapidly, particularly in recognizing the crucial role of weather conditions as key determinants of user behavior. There is a well-established consensus that weather variables are among the important variables that influence transportation choice behavior (Bean et al., 2021; Singhal et al., 2014; Zhao et al., 2018), including the decision to use BSS. These variables have been identified as critical for accurately predicting the number of trips in BSS at both aggregate and disaggregated levels (An et al., 2019; Ashqar et al., 2019; Bean et al., 2021; de Kruijf et al., 2021; Namgung, 2020; Singhal et al., 2014; Wessel, 2020; Zhao et al., 2018). Integrating weather variables with historical data on BSS trip history consistently improved model performance compared to models relying solely on past trip counts or broader environmental data.

Several papers have established the weather variables as a critical predictor of BSS demand using linear and spatial econometric methods, such as, for instance, Ridge or Negative binomial regression (Ashqar et al., 2019; Giot & Cherrier, 2014; He et al., 2019; Wessel, 2020). These studies included variables such as temperature, wind speed, and precipitation, and incorporated calendar events (e.g., holidays and weekends) to further improve prediction accuracy. Other important features that researchers suggest including are trend and lag variables (Singhal et al., 2014; Zhao et al., 2018). Since both cycling and weather are autocorrelated phenomena, incorporating variables that reflect this tends to improve model performance.

Advances in machine learning have introduced more sophisticated approaches to BSS demand forecasting and confirmed the previous approach of utilizing weather and calendar data. Deep Learning architectures, including Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNN), have been widely used to explore spatiotemporal patterns of BSS and specifically capture station-based demand (Pan et al., 2019; Sardinha et al., 2021; Yang et al., 2018).

Among various models, ensemble tree-based methods, especially Random Forest (RF) regressors, have emerged as one of the most effective models in forecasting the number of trips (Li et al., 2015; Namgung, 2020; Warmayana et al., 2025; Zhou et al., 2019). Notably, these models sometimes outperform more complicated DL models in capturing the non-linear relationship of BSS demand. Based on that finding, the current study adopts the XGBoost model, which uses a tree ensemble while providing better accuracy than the conventional Random Forest approach.

Despite these advances in BSS forecasting, there is still no well-established consensus on which models perform best on both aggregated and non-aggregated datasets. As Albuquerque et al. (2021) state that “main research gap [in BSS forecasting] is related to the selection of machine learning techniques that are best fitted and have better performance to solve BSSs at a multilevel scope”, therefore requiring more case studies using various modelling strategies. The current paper aims to contribute to the outlined research gap by utilizing infrequently used algorithms based on the historical BSS and weather data.

3. Data and Methodology

3.1 Data Processing

The study employed two publicly available datasets: Citi Bike System Data and Open-meteo Historical Weather data.

For the Citi Bike System data, raw data was collected in the form of CSV files for each month from January 2014 to April 2025. Each record corresponds to an individual trip made by the BSS user. This temporal range covers all of the available data for the Citi Bike BSS since its inception in 2013 (system-wide data is available since 2014) and includes more than 200 million trips made by the users in the provided timeframe. Using the DuckDB data management system and the corresponding Python package, the data is processed into the database, ensuring processing efficiency, taking into account the computational limitations of the personal computer. Subsequently, the records corresponding to the maintenance trips are filtered out. Following this, the records are summarized by the date, meaning that the number of trips made during a particular date from 2014 to 2025 is calculated.

The historical weather data is collected using the Open-meteo Historical Weather data Application Programming Interface (API). The API facilitates sending requests for detailed timeframes, specified by geographical locations expressed in coordinates, and weather variables. The weather variables include, but are not limited to, mean temperature, precipitation, windspeed and direction, daylight, cloud cover, and humidity. The variables are chosen based on how relevant they might be in affecting transportation behavior, and, thus, determining a daily number of BSS trips. This dataset is subsequently aggregated and joined with the bike trip dataset using the date field as a join field.

Given the inability of ML models to process the date feature, a series of distinct features were developed to reflect the temporal aspect of the data that might affect the number of trips made within BSS. These variables include the year, season, day of the week, and whether this day is a weekend. Additionally, using US Public Holidays data, a feature has been developed that determines if the specified date falls on a holiday. Furthermore, because daily bike-share usage shows short-term seasonality, we supplement our feature set with two temporal aggregates. First, we include one, seven, and thirty-day lags of total trips, so that trends, demand shocks, as well as the autocorrelation effect of cycling and the lagging effect of weather, can be reflected in the forecasting. Second, we calculate the 7 and 30-day moving averages of ridership to smooth out

high-frequency noise associated with weekends and weather fluctuations. Prior scholarship on BSS modelling has shown that combining lagged and rolling-mean predictors yields more accurate and stable forecasts (Singhal et al., 2014; Zhao et al., 2018). The final dataset used for model creation looks as follows, with 21 columns and 4,099 records.

Table 1. Summary of model variables

Variable	Data Type	Description
temperature_2m_mean	Numeric	Mean Air temperature at 2 meters above ground (°C)
apparent_temperature_mean	Numeric	Mean apparent temperature at 2 meters above ground (°C)
precipitation_sum	Numeric	Sum of daily precipitation (rain, showers and snowfall) (mm)
rain_sum	Numeric	Sum of daily rain (mm)
snowfall_sum	Numeric	Sum of daily snowfall (mm)
precipitation_hours	Numeric	The number of hours with rain
wind_speed_10m_max	Numeric	Maximum wind speed on a day at 10 meters above ground (m/s)
wind_gusts_10m_max	Numeric	Maximum wind gusts on a day at 10 meters above ground (m/s)
cloud_cover_mean	Numeric	Mean cloud cover as an area fraction (%)
relative_humidity_2m	Numeric	Relative humidity at 2 meters above ground (%)
daylight_duration	Numeric	Number of seconds of daylight per day
weekday	Categorical	Day of week, encoded 0 = Monday … 6 = Sunday
year	Categorical	Year offset from 2014 (0 = 2014, 1 = 2015, …, 11 = 2025)
season	Categorical	Season code (1 = winter, 2 = spring, 3 = summer, 4 = autumn)
is_holiday	Categorical	Holiday flag (1 = public holiday, 0 = non-holiday)
lag_1	Numeric	Number of trips on the previous day (1-day lag)
lag_7	Numeric	Number of trips 7 days earlier (weekly lag)
lag_30	Numeric	Number of trips 30 days earlier (monthly lag)
rolling_7_mean	Numeric	7-day moving average of daily trip counts
rolling_30_mean	Numeric	30-day moving average of daily trip counts
number_of_trips	Numeric	Total daily CitiBike trip count (target variable)

3.2 ML Modelling

Regression models have demonstrated their efficacy in producing accurate predictions within the domain of BSS forecasting scholarship and, therefore, are used as a baseline for quality assessment in this study (Ashqar et al., 2019; Wessel, 2020). In the context of the weather data, researchers frequently employ Lasso and Ridge Regularization regression models to address the challenge of multicollinearity. As weather variables tend to be correlated, it is plausible that

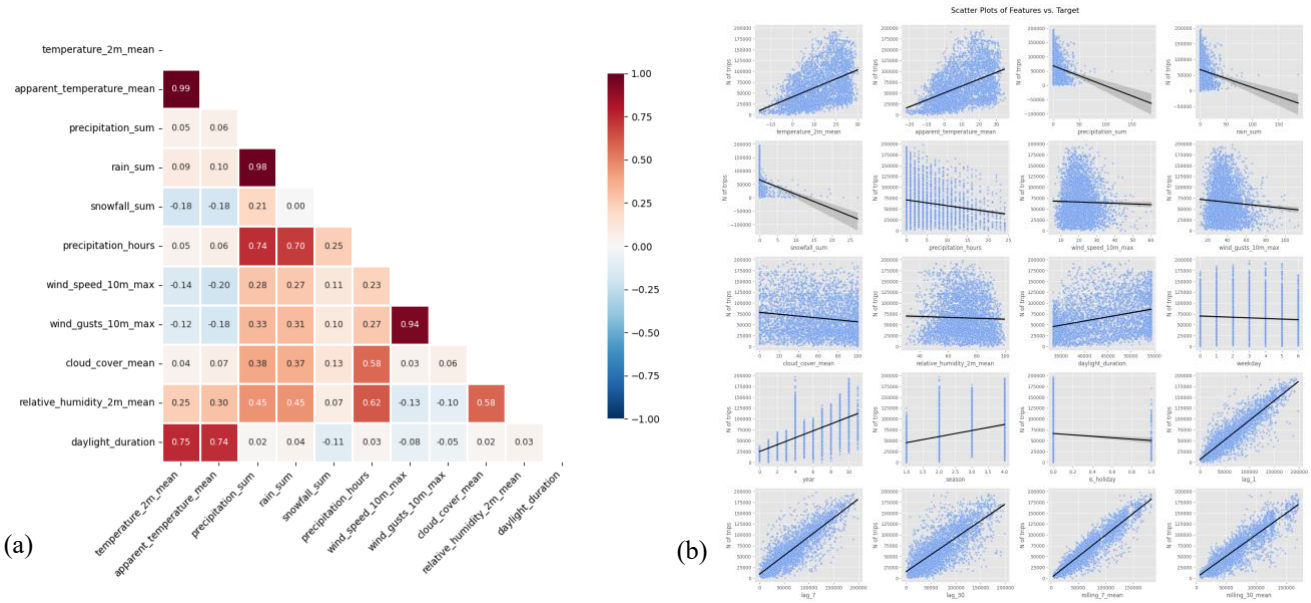

Figure 1. Exploratory correlations.
(a) Correlation matrix of weather variables.
(b) Scatter plots of features vs. number of trips.

multicollinearity may emerge as a concern in our data as well. This assumption is confirmed by the features correlation matrix (Fig. 1a), which shows high correlation coefficients for features such as temperature and apparent temperature, rain and precipitation, etc. Additionally, an exploration of correlations between established variables and the target variable (daily number of trips) is conducted to establish the necessity for a regression model based on the linear relationship of features and the target variable (Fig. 1b).

As linear methods struggle to deal with multicollinearity, the Lasso Regression model is employed, as it adds penalty to the loss function, helping to shrink coefficients of less important predictors to zero and thus essentially performing feature selection and reducing overfitting of the model.

Prior to training the model, the categorical features with multiple categories are encoded using one-hot encoding to ensure compatibility with a linear model framework. Numerical features are scaled to standardize their range, thereby enhancing the interpretability of the resulting coefficients. Subsequently, the model is trained using a cost function that combines Mean Squared Error (MSE) with L1 regularization (Lasso), and hyperparameter tuning of the regularization strength parameter λ using cross-validation is performed. The cost function minimized during training is defined as:

$$J(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where y_i denotes the true value, $\hat{y}_i$ is the predicted value computed as $\hat{y}_i = \mathbf{x}_i^T\beta$, where $\mathbf{x}_i$ is the vector of input features for observation i. n is the number of observations, p is the number of features, β is the vector of coefficients, and λ controls the strength of the L1 penalty.

As highlighted in the previous section, one of the most used models for BSS forecasting are tree-based models, such as decision trees and Random Forest. The primary model of this project, XGBoost regression, combines advantages of the tree-based approaches while offering higher precision, which is essential in the case of BSS daily forecasting. Prior to training the model, the categorical features with multiple categories, like weekday and season, are encoded using one-hot encoding. The year feature is kept as is to preserve timeseries character in the data, as recent XGBoost updates allow to train models on categorical without encoding it. I train the model on the MSE cost function as defined above, and perform hyperparameter tuning using the automated hyperparameter optimization library HyperOpt (Bergstra et al., 2013) using cross-validation.

Lastly, to ensure that the model does not produce unrealistic negative predictions, given that the target variable, cannot be negative, I apply a non-negativity constraint to the output by replacing all negative values with zero after prediction. Formally, this post-processing step is defined as:

$$\hat{y}_i = \max(\hat{y}_i, 0)$$

This correction is particularly important when using regression models that do not inherently restrict output values, such as linear models or tree-based ensembles.

4. Results

I estimate the performance of the base model (Lasso Linear Regression) and the XGBoost model using three evaluation metrics: Mean Absolute Error (MAE), root Mean Squared Error (rMSE), and Mean Absolute Percentage Error (MAPE), which are defined as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$rMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Table 2. Performance comparison of Lasso and XGBoost models using MAE, rMSE, and MAPE

Model	MAE	rMSE	MAPE (%)
Lasso Regression	9931	13392	14.48
XGBoost Regression	5891	8510	13.64

where notations are the same as in the previous section. MAE measures the average magnitude of errors in absolute terms, rMSE penalizes larger errors more strongly due to squaring, and MAPE expresses the error as a percentage, offering an interpretable metric relative to the actual values.

As demonstrated in the Table 2 and Figure 3, Lasso Regression shows good result with MAE of around 10,000 trips, indicating that model's prediction is on average off by that number of trips, considering the maximum of trip made in one day in the dataset is 200,000, while the mean is approximately 66,000 trips. However, as demonstrated in Table 2, the XGBoost model outperformed the Lasso regression model, achieving lower MAE (5891 vs 9931), lower rMSE (8510 vs. 13392), and lower MAPE (13.64% vs. 14.48%). Essentially, XGBoost's predictions are off by 13.64% on average compared to the actual values, which is considered good forecasting quality in the machine learning industry.

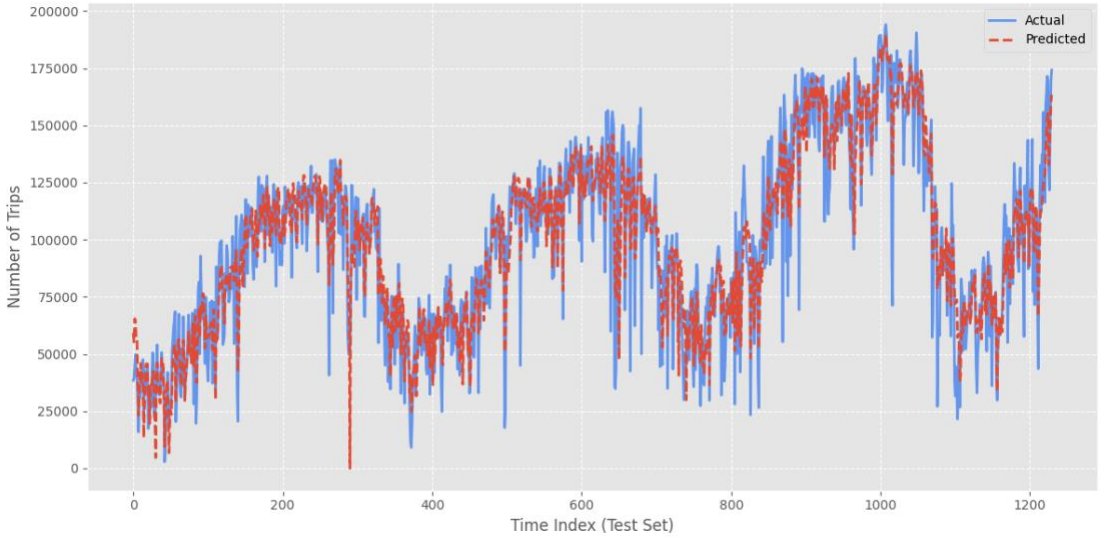

Figure 2. Line Plot of Actual vs Predict values for Lasso Regression.

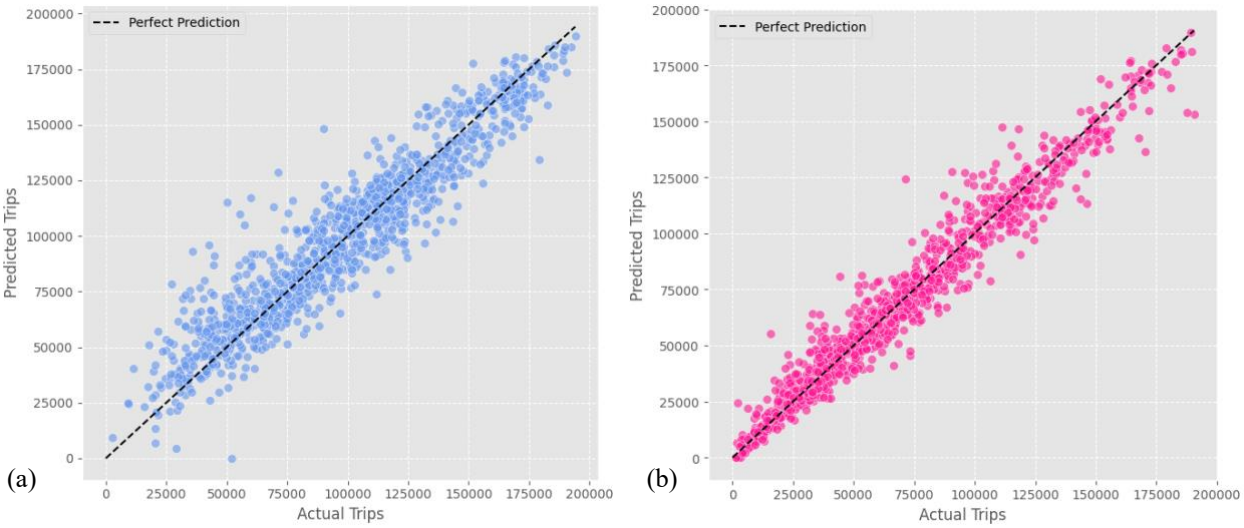

Figure 3. Predicted vs. Actual Citi Bike trip counts for (a) Lasso Linear Regression and (b) XGBoost Regression models. Each point represents one observation, with the dashed line indicating perfect prediction. Both models demonstrate a strong linear relationship, though the XGBoost model (b) shows a tighter clustering around the diagonal, suggesting improved predictive accuracy.

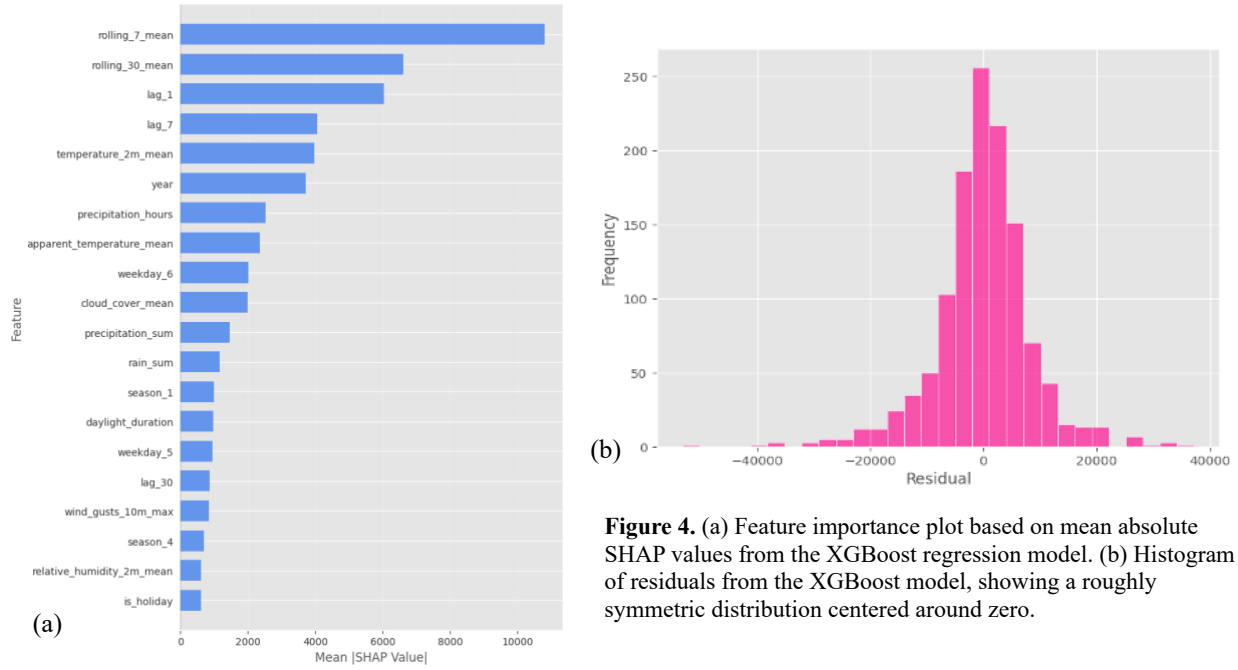

Figure 4. (a) Feature importance plot based on mean absolute SHAP values from the XGBoost regression model. (b) Histogram of residuals from the XGBoost model, showing a roughly symmetric distribution centered around zero.

Additionally, based on the rMSE score, it is also possible to conclude that the XGBoost model is much more reluctant to outliers in data, compared to the Lasso Regression model. As demonstrated in Figure 4, predictions of the XGBoost model show a strong alignment with the perfect prediction line, as scatterplot points on the chart are located closer to the perfect prediction line with far fewer outliers. The residuals histogram of the XGBoost model displays a roughly symmetrical distribution centered around zero, indicating generally unbiased prediction. There is

a slight skew towards the right, indicating that oftentimes the model underpredicts number of trips. Most residuals fall within the range of ±20,000 trips, suggesting stable model performance. As seen in Figure 4a, the most influential features for predicting daily number of trips include recent trends, particularly 7- and 30-day moving averages of trips, short-term lags, as well as temperature and precipitation variables. The least influential variables, such as some of the weekday, season, and wind variables, are excluded from the chart due to their low significance to the prediction.

5. Discussion and Conclusion

This study presents a machine learning approach to forecasting daily BSS trips on New York City's Citi Bike. The model uses trip-level data from 2014 to 2025, enriched with historical weather and calendar context data. Raw data on individual trips are processed using the database management system DuckDB and Python, and then aggregated by date to generate daily ridership counts. Two models were developed: Lasso Regression and XGBoost Regression. Based on the evaluation metrics discussed in the previous section, the XGBoost model significantly outperformed the linear baseline, demonstrating prediction with better accuracy according to all three metrics: MAE, rMSE, and MAPE. The final MAPE score of 13.64% indicates that the developed model predicts the daily number of trips with an error of 13.64%, thus producing a reliable daily forecast with a relatively low average error.

The most influential predictors were rolling mean and lag features, reflecting the autocorrelated nature of BSS trips, as well as temperature and precipitation, confirming the findings from the scholarship (Bean et al., 2021; Sardinha et al., 2021; Zhao et al., 2018). These results suggest that the model can be useful as an analytical and decision-support tool for urban policy makers and system operators to estimate the number of daily trips made using the Citi Bike system, using available and forecasted weather data (Wessel, 2020).

However, these results should be interpreted considering several limitations. First, the current model does not incorporate any spatial parameters, although there is evidence that bicycles within BSS have unequal spatial distribution among the cities. Second, there are several environmental and other features that might affect the daily ridership, such as levels of air pollution (Namgung, 2020), large-scale public events, availability of bike infrastructure (Bustamante et al., 2022), bike-sharing system-level factors, public transportation disruptions, and street closers. Third, daily aggregations may smooth out intra-day fluctuations in demand, which is essential information for operational decisions within the BSS. Hourly forecasts can provide more

actionable insights for operations and rebalancing. Finally, although the XGBoost model performs well overall, residuals remain large for extreme trip volumes, suggesting reduced reliability and underperformance in edge cases.

These limitations led to potential avenues for future work. It should focus on developing models based on hourly data as well as station-based demand to increase the granularity of predictions and account for spatial heterogeneity in demand. Future studies should also include additional environmental and contextual variables that can potentially influence bicycle demand, especially when developing spatially oriented models. In addition, future work should focus on improving model performance, especially for the edge cases, with the goal of reducing the MAPE score to less than 10%.

Data availability. The dataset used in this study is publicly accessible from the Citi Bike Portal https://citibikenyc.com/system-data and using the Open-Meteo API https://open-meteo.com/en/docs/historical-weather-api

Code availability. The code used in this paper can be accessed from the public GitHub repository https://github.com/temapankin/XGBoost-for-CitiBike-forecasting

References

Albuquerque, V., Sales Dias, M., & Bacao, F. (2021). Machine Learning Approaches to Bike-Sharing Systems: A Systematic Literature Review. *ISPRS International Journal of Geo-Information*, *10*(2), Article 2. https://doi.org/10.3390/ijgi10020062

An, R., Zahnow, R., Pojani, D., & Corcoran, J. (2019). Weather and cycling in New York: The case of Citibike. *Journal of Transport Geography*, *77*, 97–112. https://doi.org/10.1016/j.jtrangeo.2019.04.016

Ashqar, H. I., Elhenawy, M., & Rakha, H. A. (2019). Modeling bike counts in a bike-sharing system considering the effect of weather conditions. *Case Studies on Transport Policy*, *7*(2), 261–268. https://doi.org/10.1016/j.cstp.2019.02.011

Bean, R., Pojani, D., & Corcoran, J. (2021). How does weather affect bikeshare use? A comparative analysis of forty cities across climate zones. *Journal of Transport Geography*, *95*, 103155. https://doi.org/10.1016/j.jtrangeo.2021.103155

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. *Proceedings of the 30th International Conference on Machine Learning*, 115–123. https://proceedings.mlr.press/v28/bergstra13.html

Bustamante, X., Federo, R., & Fernández-i-Marin, X. (2022). Riding the wave: Predicting the use of the bike-sharing system in Barcelona before and during COVID-19. *Sustainable Cities and Society*, *83*, 103929. https://doi.org/10.1016/j.scs.2022.103929

de Kruijf, J., van der Waerden, P., Feng, T., Böcker, L., van Lierop, D., Ettema, D., & Dijst, M. (2021). Integrated weather effects on e-cycling in daily commuting: A longitudinal evaluation of weather effects on e-cycling in the Netherlands. *Transportation Research Part A: Policy and Practice*, *148*, 305–315. https://doi.org/10.1016/j.tra.2021.04.003

Giot, R., & Cherrier, R. (2014). Predicting bikeshare system usage up to one day ahead. *2014 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS)*, 22–29. https://doi.org/10.1109/CIVTS.2014.7009473

He, Y., Song, Z., Liu, Z., & Sze, N. N. (2019). Factors Influencing Electric Bike Share Ridership: Analysis of Park City, Utah. *Transportation Research Record*, *2673*(5), 12–22. https://doi.org/10.1177/0361198119838981

Li, Y., Zheng, Y., Zhang, H., & Chen, L. (2015). Traffic prediction in a bike-sharing system. *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 1–10. https://doi.org/10.1145/2820783.2820837

Meixell, M. J., & Norbis, M. (2008). A review of the transportation mode choice and carrier selection literature. *The International Journal of Logistics Management*, *19*(2), 183–211. https://doi.org/10.1108/09574090810895951

Namgung, M. (2020). *Performance Comparison of Public Bike Demand Predictions: The Impact of Weather and Air Pollution* [Thesis, Purdue University Graduate School]. https://doi.org/10.25394/PGS.13353482.v1

Pan, Y., Zheng, R. C., Zhang, J., & Yao, X. (2019). Predicting bike sharing demand using recurrent neural networks. *Procedia Computer Science*, *147*, 562–566. https://doi.org/10.1016/j.procs.2019.01.217

Sardinha, C., Finamore, A. C., & Henriques, R. (2021). *Context-aware demand prediction in bike sharing systems: Incorporating spatial, meteorological and calendrical context* (No. arXiv:2105.01125). arXiv. https://doi.org/10.48550/arXiv.2105.01125 arXiv:2105.01125 [cs]

Singhal, A., Kamga, C., & Yazici, A. (2014). Impact of weather on urban transit ridership. *Transportation Research Part A: Policy and Practice*, *69*, 379–391. https://doi.org/10.1016/j.tra.2014.09.008

Warmayana, I. G. A. K., Yamashita, Y., & Oka, N. (2025). Predictive Analysis for Optimizing Targeted Marketing Campaigns in Bike-Sharing Systems Using Decision Trees, Random Forests, and Neural Networks. *Journal of Digital Market and Digital Currency*, *2*(1), Article 1. https://doi.org/10.47738/jdmdc.v2i1.29

Wessel, J. (2020). Using weather forecasts to forecast whether bikes are used. *Transportation Research Part A: Policy and Practice*, *138*, 537–559. https://doi.org/10.1016/j.tra.2020.06.006

Yang, H., Xie, K., Ozbay, K., Ma, Y., & Wang, Z. (2018). Use of Deep Learning to Predict Daily Usage of Bike Sharing Systems. *Transportation Research Record*, *2672*(36), 92–102. https://doi.org/10.1177/0361198118801354

Zhao, J., Wang, J., Xing, Z., Luan, X., & Jiang, Y. (2018). Weather and cycling: Mining big data to have an in-depth understanding of the association of weather variability with cycling on an off-road trail and an on-road bike lane. *Transportation Research Part A: Policy and Practice*, *111*, 119–135. https://doi.org/10.1016/j.tra.2018.03.001

Zhou, X., Wang, M., & Li, D. (2019). Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. *Journal of Transport Geography*, *79*, 102479. https://doi.org/10.1016/j.jtrangeo.2019.102479