Citi Bike Machine Learning Analysis

Brianna Ta, Hailey Weingord, Songwen Zhao, Yi Ju, and Zekai Su

Columbia University

1 Introduction

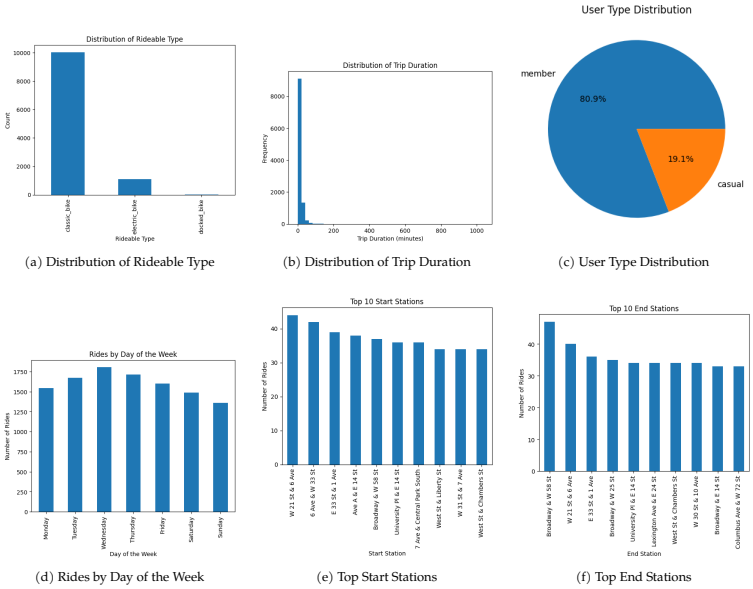
Citi Bike, New York City's largest bike-sharing system, is a cornerstone of sustainable urban transportation. Understanding trends in its usage can help optimize infrastructure, enhance user experience, and support accessibility. Analyzing historical trip data enables us to uncover key patterns, such as peak usage times, popular stations, and trip durations, providing actionable insights to improve the system.

2 Dataset

The 2023 Citi Bike dataset includes detailed trip records, capturing information such as start and end times, trip duration, station locations, bike types, and user categories. This dataset allows for the exploration of usage patterns, seasonal trends, and travel distances, offering valuable data to refine bike-sharing operations and promote sustainable mobility in New York City.

3 Data Exploration and Cleaning

Through initial data exploration, we discovered that classic bikes are by far the most frequently used type, with electric bikes being significantly less common and docked bikes almost negligible in usage. The vast majority of the riders are members, at 81.5%, and weekday usage is generally higher than weekend usage, with peaks from Tuesday to Thursday. Lastly, the trip duration distribution shows that the vast majority of rides are short (median = 9.3 minutes), clustered at lower durations, with very few trips extending beyond 200 minutes. These findings, and more, are illustrated below.



(a) Distribution of Rideable Type (b) Distribution of Trip Duration (c) User Type Distribution

(d) Rides by Day of the Week (e) Top Start Stations (f) Top End Stations

Figure 1: Exploration of Data and Variable Distributions

To clean the dataset, we began by identifying columns with missing values. For columns with a manageable amount of missing data, we imputed the missing values using the mean or mode, depending on the data type. Columns with excessive missing values were dropped to maintain data quality. Additionally, we reformatted the date and time column for easier analysis. Using the datetime method in pandas, we extracted the date and time components and created separate columns for each.

4 Machine Learning Models

4.1 Baseline Models

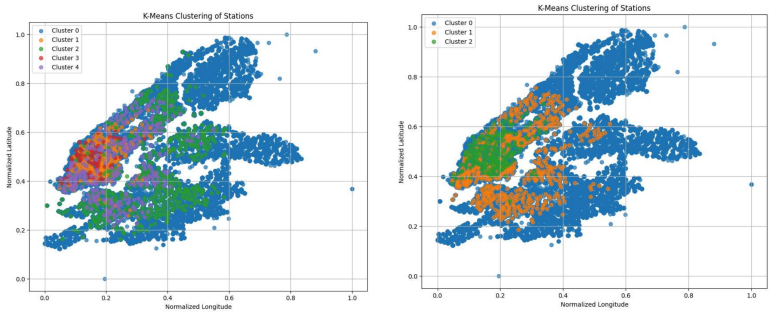
To establish baseline performance metrics, we implemented two simple models. First, we used a linear regression algorithm to predict trip duration based on features such as start and end stations, bike type, date, start and end time, and latitude and longitude. Second, we applied a logistic regression model to classify whether a Citi Bike user was a casual rider or a member, using the same input variables. The linear regression model achieved an R^2 score of 0.0148 on the test set, while the logistic regression model attained an overall accuracy of 81%. Although these baseline models lacked hyperparameter tuning and performed poorly in capturing complex patterns, they provided a starting point for further refinement.

4.2 Random Forest Regression

We implemented a Random Forest Regression model to predict trip duration for Citi Bike users. To enhance model performance, we employed random search for hyperparameter tuning, optimizing key parameters such as the number of estimators, maximum tree depth, minimum samples required for splits, minimum leaf size, and the maximum number of features. After training the tuned model, we observed a significant reduction in the Root Mean Squared Error (RMSE) from 7.32 on the baseline model to 4.67, indicating that hyperparameter tuning substantially improved the model's predictive accuracy. Furthermore, the R^2 value increased to 0.59, demonstrating that the tuned model explains 59% of the variance in trip duration, a notable improvement over the baseline.

4.3 Clustering

We performed clustering to analyze patterns among Citi Bike stations, focusing on geographic groupings. Using K-means clustering, we first explored the data with 5 clusters. At this level, we observed significant overlap between clusters 1 and 3, as well as between clusters 2 and 4. Additionally, cluster 0 emerged as the dominant cluster, overlapping extensively with stations from all other clusters. To further investigate this phenomenon, we reduced the number of clusters to n=3. While this reduction simplified the clustering structure, one cluster continued to dominate. By leveraging the longitude and latitude columns from the dataset, we identified the location of this dominant cluster: midtown east, a bustling area of the city with high activity levels.



(a) K-Means Clustering, n=5 (b) K-Means Clustering, n=3

Figure 2: Comparison of K-Means Clustering Results

4.4 Forecasting

We used time-series models, including ARIMA and SARIMA, to analyze Citi Bike's daily ride counts and uncover seasonal patterns. Usage peaked in spring and summer, dipped in winter, and showed distinct weekly trends, with commuter rides dominating weekdays and leisure rides on weekends. The models also highlighted the impact of external factors, such as a sharp drop in rides on Christmas Day due to holidays. SARIMA outperformed ARIMA by capturing recurring seasonal patterns, as

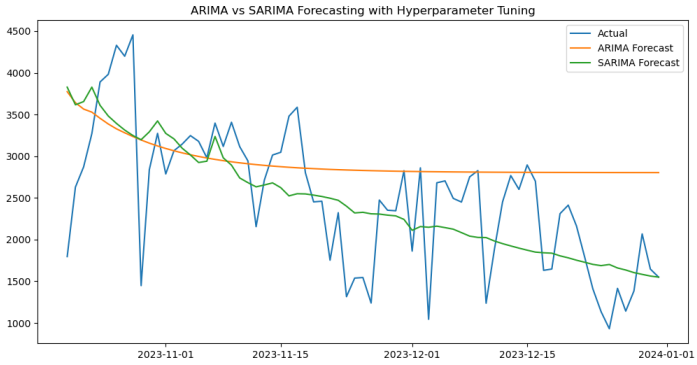


Figure 3: Comparison between tuned ARIMA and SARIMA models

Model	RSME	MAE
ARIMA	876.051	670.523
SARIMA	650.277	516.328

Table 1: Tuned ARIMA uring the levs SARIMA Model Evaluation

reflected in improved RMSE and MAE scores. These insights enable Citi Bike to optimize bike and dock availability during peak demand and schedule maintenance during off-peak times, fostering efficiency, user satisfaction, and sustainable transportation.

4.5 Neural Network

The last model we implemented was a neural network to predict trip duration. After developing and training the model, the neural network resulted in the following loss curve for both the training and test datasets: We can see that the training loss curve decreases steadily over the epochs, indicating that

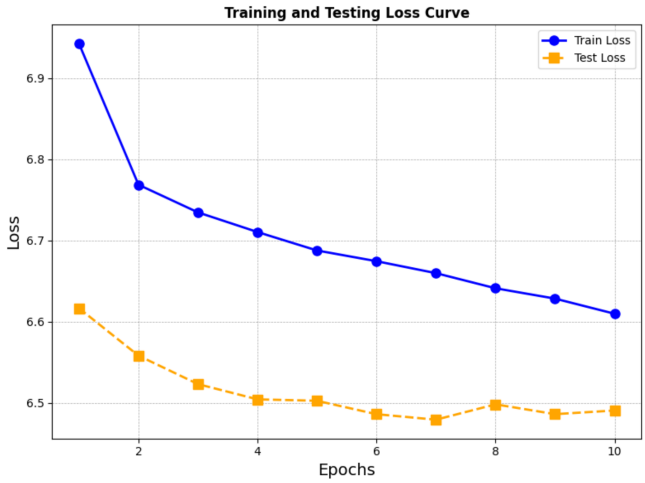


Figure 4: Training and Testing Loss Curve

it is successfully learning patterns in the training data. The test loss curve decreases initially but starts to stabilize and fluctuate around epoch 6. This indicates that the model's generalization to unseen data may be limited. Interestingly, the loss for the training set is consistently higher than the test set, which is unusual. Lastly, the R^2 score on the test set was only 6.1%, indicating that the model did not fit our dataset very well.

5 Conclusion

In conclusion, we developed and optimized four models to analyze the Citi Bike dataset. Random Forest and Neural Network models were implemented to predict trip duration, with Random Forest demonstrating superior performance. Clustering revealed that midtown east is the dominant region for Citi Bike activity in NYC, while time-series forecasting highlighted seasonal trends and the impact of holidays and weather on ride patterns. Future work could explore incorporating real-time data, such as live weather updates or traffic conditions, to further enhance prediction accuracy and operational insights.