NYC CitiBike Demand Prediction and Optimization

This project leverages geospatial data, chronological patterns, and machine learning techniques to predict demand and optimize bike allocation across CitiBike stations in New York City. The goal is to create a framework for real-time demand forecasting, station-level performance insights, and an optimized bike reallocation strategy that improves operational efficiency, revenue and user experience.

1. Data Exploration and Preprocessing

All of the data for this project was provided from the CitiBikenyc.com/system-data. The data set includes ride details such as start and end coordinates, timestamps, and membership type. These features were first preprocessed to calculate the distance between stations using the Haversine formula, and key time-series variables such as the day of the week and hour of the day were extracted. This enables us to analyze the ride patterns based on geographical and time factors.

A major finding from the exploration was that the dataset contains 2,164 unique start stations spread across four boroughs: Brooklyn (700 stations), Manhattan (694), Queens (451), and the Bronx (319).

Here are some quick facts on the data processed from 2023:

1. Brooklyn had the longest average ride distances **(1,993 meters)**, while the Bronx had the shortest **(1,501 meters)**.
2. The stations with the longest average ride distances were concentrated in the Bronx and Queens, such as the station at 46 Rd & 11 St in Queens with an average ride distance of **11,245 meters**.
3. The analysis also revealed that all boroughs have peak ridership around 4-5 PM, with Manhattan exhibiting the highest peak
4. Over 80% of all the rides were from members and not casual riders.
5. Only 10% of members rode electric bikes compared to classic bike types

These findings suggest that certain areas experience longer trips, likely due to their location in relation to popular destinations or transportation hubs.

2. Station-Level Insights

The next step involved aggregating the data by station to compute key metrics such as total rides, average distance, casual versus member ride counts, and weekend activity. This station-level analysis helps identify trends in station performance and informs the clustering process for demand prediction.

Clustering stations based on geographic location using K-means helped group similar stations, which provides insights into station performance based on factors like proximity to high-demand areas and ridership behavior.

3. Demand Prediction with XGBoost

To predict future demand at each station, an XGBoost regression model was trained on the processed dataset. Given the size of the CitiBike dataset, which includes several station-level attributes and ride details, using an efficient algorithm like XGBoost ensures that the model can be trained quickly without compromising performance. The model utilized variables like average ride distance, membership type, and temporal features (hour of the day) to predict the total number of rides at each station. A random 80/20 split of the data was used for training and testing, ensuring that the model was robust and unbiased.

The model achieved a Root Mean Square Error (RMSE) of 14.93, which was a significant improvement over the baseline model RMSE of 563.11. This indicates that the XGBoost model effectively captures demand patterns, with a substantial reduction in prediction error compared to a simple mean-based prediction.

4. Bike Reallocation Strategy

An essential outcome of this project was the formulation of a bike reallocation strategy. By analyzing stations with excess or insufficient demand, the model can recommend a redistribution of bikes to ensure optimal availability. For instance, stations with underperforming demand could receive bikes from high-demand stations, maintaining a balanced bike supply across the network. This strategy is crucial for CitiBike operators to ensure bikes are available where they are most needed, especially during peak hours.

5. Conclusions and recommendations

Utilizing the predicted bike usage per station, I was able to deduce the top 5 stations that are projected to have the highest monthly revenue. Not surprisingly, all of the stations were located within Manhattan. However, the interesting part is all of the stations were located at or near Central Park. The station with the highest projected revenue was Central Park South and 6th Avenue with a projection little over $2,700 dollars which is the highest out of all the stations looked at in the data.

Looking at a borough level, Manhattan as a whole has the highest average projected monthly revenue. Brooklyn is actually in second which would make a good argument to incorporate more stations within Manhattan. Instead of incorporating more stations in Brooklyn, you could reallocate electric bikes to the outer boroughs as they have higher average distances.

Lastly, I would recommend adding more electric bikes to the Central Park Stations. On average in Manhattan, casual riders use an electric bike 70% of the time, at Central Park South and 6th Ave casual riders use electric bikes only 61% of the time. This leads me to believe there are not enough electric bikes at that station and a possible loss of revenue as electric bikes generate more revenue.

This project provides a data-driven approach to enhancing CitiBike station performance through demand prediction and resource optimization. The insights gained from the station-level analysis and demand forecasting can be used to inform real-time operational decisions, improving the user experience and operational efficiency. The bike reallocation strategy, based on predicted demand, is a key step towards achieving a more responsive and efficient bike-sharing system.

Below is the full code used for this project as well as visuals to help explain the data.