Citi Bike Rentals – Analytics and Forecasting

1. Abstract

Our project aims to develop a scalable solution for time series forecasting on Citi Bike dataset using Apache Spark and Facebook Prophet. Time series forecasting is crucial in various domains such as finance, retail, and energy management. However, traditional forecasting methods often struggle to handle large volumes of data efficiently. By leveraging the distributed computing capabilities of Apache Spark and the time series forecasting capabilities of Facebook Prophet, we aim to create a system that can produce scalable analytics and accurate predictions for Citi Bike rental systems. We will address key inquiries posed by Citi Bike such as Where do Citi Bikers ride? When do they ride? How far do they go? Which stations are most popular? What days of the week are most rides taken on?

2. Introduction

Time series forecasting plays a vital role in decision-making processes across industries. The increasing popularity of bike-sharing programs in urban areas has led to a need for efficient management of resources within these systems. Understanding usage patterns and accurately predicting demand can aid in optimizing fleet distribution, station placement, and infrastructure planning. However, existing forecasting methods may not scale well with increasing data volumes, leading to longer processing times and decreased accuracy. Our project addresses this challenge by combining the strengths of Facebook Prophet, a powerful forecasting library, with Apache Spark, a distributed computing framework. By distributing the computational workload across a cluster of machines, we aim to provide a robust forecasting model that provides accurate predictions for Citi Bike rentals.

3. Background

Bike-sharing demand forecasting involves analyzing historical usage data to predict future demand patterns. Accurate forecasting enables operators to make informed decisions regarding resource allocation, maintenance scheduling, and system optimization. Time-series prediction models play a crucial role in this process, capturing temporal trends, seasonality, and external factors influencing demand. Our solution employs a high-level approach that encompasses data preprocessing, model training, and evaluation. We integrate external factors such as holidays, special events, and weather conditions to enhance forecasting accuracy. Through rigorous analysis and evaluation, we aim to develop a comprehensive forecasting solution for bike-sharing demand.

4. Data used

We utilized two primary datasets for our project. The first dataset is the Citi Bike Trip Data sourced from the New York City Bike Share (NYCBS) program, specifically the Citi Bike system. This dataset provides comprehensive information on Citi Bike trips, including Ride ID, Rideable type, Start and End timestamps, Start and End station names, as well as Start and End coordinates (latitude and longitude). Our motivation for selecting this dataset stems from its ability to offer rich insights into rider behavior and usage patterns. By analyzing factors such as popular stations, ride durations, and time-of-day trends, we aimed to gain a deeper understanding of how riders interact with the bike-sharing system. We targeted the period from January 1, 2016, to February 29, 2024, to capture a substantial timeframe of data. However,

due to the monthly availability of data, we faced challenges during data acquisition. We had to download separate files for each month and merge them using Apache Spark. This merging process was complicated by changes in data format over time, necessitating manual inspection and adjustments to ensure data consistency and reliability.

The second dataset comprises Weather Data obtained from the National Centers for Environmental Information (NCEI). This dataset includes various weather factors such as precipitation, snow, and wind speed. We integrated weather data into our analysis to explore its impact on forecasting accuracy, particularly when incorporated into predictive models like the Prophet model. Weather plays a significant role in bike-sharing behavior, influencing ridership levels and ride durations. While ideally, we would have obtained weather data specific to each Citi Bike station location, we were unable to access such granular data. Therefore, we gathered general weather data for New York City. Despite this limitation, integrating weather data allowed us to investigate correlations between weather conditions and bike-sharing patterns, providing valuable insights for optimizing bike-sharing operations and enhancing service efficiency.

5. Motivation

The motivation behind our solution lies in optimizing bike-sharing operations through data-driven forecasting. By accurately predicting demand, operators can allocate resources efficiently, enhance service reliability, and ultimately, elevate the user experience. Leveraging advanced machine learning techniques, we aim to achieve these objectives by providing bike-sharing system operators with a robust decision-making tool. Our approach combines the forecasting expertise of Facebook Prophet with the distributed computing capabilities of Apache Spark. By harnessing the power of machine learning and distributed computing, we strive to empower operators with actionable insights derived from comprehensive data analysis, facilitating informed decision-making and driving continuous improvement in bike-sharing system operations.

6. Design:

Our solution encompasses a multi-step approach designed to deliver accurate and reliable forecasts of bike-sharing demand. The key components of our solution include:

1. **Library Installation:** Necessary libraries such as Prophet, Matplotlib, and Scikit-learn were installed to fulfill the requirements of our project.

2. **Data Preprocessing:** We start by preprocessing the Citi Bike trip data to handle missing values and inconsistencies. In this step, we identify and handle null values within the dataset by dropping corresponding records. Additionally, records with ride durations less than 60 seconds were filtered out to maintain data integrity. Aggregation functions were applied to count the total number of trips on each day. The dataset was then transformed to retain only essential columns such as date and trips for forecasting purposes.

3. **Exploratory data analysis:**
   EDA was conducted to address inquiries from Citi Bike. We started with smaller dataset first and gradually increased the dataset to see the impact of increased dataset in our forecasting results.

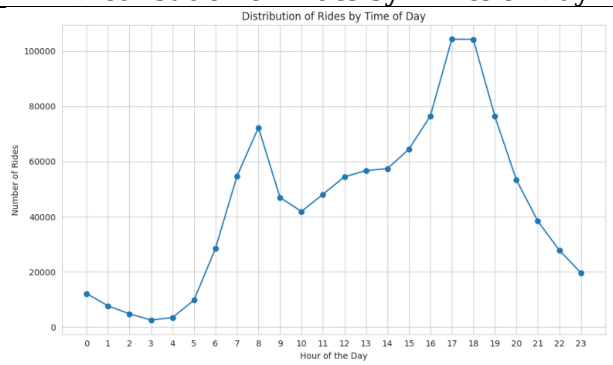
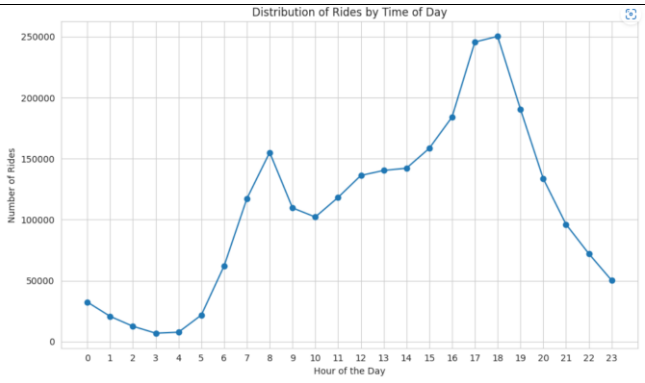
    1. **Where do Citi Bikers ride?**

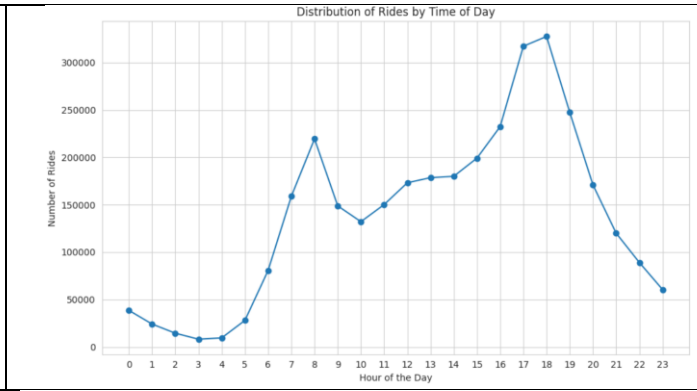
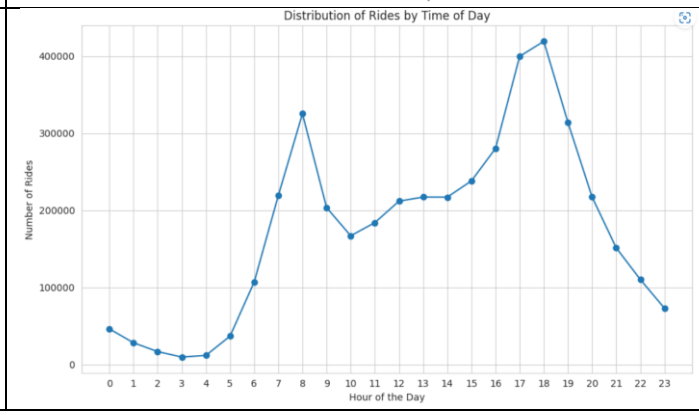
Year	Station Name
2023-2024	Grove St PATH Hoboken Terminal .. South Waterfront .. Hoboken Terminal .. City Hall - Washi.. Newport PATH Hamilton Park Newport Pkwy Bergen Ave & Sip Ave 11 St & Washingto..
2021-2024	Grove St PATH Hoboken Terminal ... South Waterfront ... Hoboken Terminal ... Newport Pkwy Newport PATH Hamilton Park City Hall - Washi... Marin Light Rail Hoboken Ave at Mo...
2019-2024	Grove St PATH Hamilton Park Newport Pkwy Hoboken Terminal ... Newport PATH South Waterfront ... Hoboken Terminal ... Marin Light Rail Harborside Liberty Light Rail

2016-2024	Grove St PATH Hamilton Park Newport PATH Newport Pkwy Marin Light Rail Sip Ave Hoboken Terminal ... Liberty Light Rail South Waterfront ... City Hall

Insights: The most popular Citi Bike stations where citi bike rider ride are Grove St PATH, Hoboken Terminal - River St & Hudson Pl, South Waterfront Walkway - Sinatra Dr & 1 St, Hoboken Terminal - Hudson St & Hudson Pl, City Hall - Washington St & 1 St, Newport PATH, Hamilton Park, Bergen Ave & Sip Ave, Newport Pkwy, and 11 St & Washington St. These stations likely experience high levels of activity due to their strategic locations, such as being near transportation hubs, residential areas, or popular destinations like IT hub, etc. Understanding the popularity of these stations is valuable for Citi Bike operators as it helps them optimize bike distribution, station maintenance, and infrastructure planning to better serve riders in these areas.

2. When do they ride

Year	Distribution of Rides by Times of Day
2023-2024	
2021-2024	

2019-2024	
2016-2024	

Insights: The graph reveals that the busiest time for Citi Bike rentals is between 5 and 6 PM. This insight is crucial for managing the Citi Bike rental system effectively. Operators need to ensure there are enough bikes available and docking stations open during this peak period to meet the high demand. Understanding when demand is highest enables operators to plan maintenance and allocate resources more efficiently. It also allows them to encourage bike usage during less busy times, which can help balance demand throughout the day and improve overall service quality for Citi Bike users.

3. How far do they go?

Year	Average distance travelled per ride
2023-2024	1.27 km
2021-2024	1.20 km
2019-2024	1.40 km
2016-2024	0.78km

Insights: The above data reveals interesting insights into the average distance traveled per ride over different time periods. From 2016 to 2024, there's a notable increase in the average distance traveled per ride, rising from 0.78 km to 1.40 km. This suggests a trend of riders traveling longer distances over the years, indicating potential shifts in rider usage patterns. However, there's a slight decrease in the average distance traveled per ride from 2021 to 2024, dropping from 1.20 km to 1.27 km. This fluctuation might indicate variations in external factors influencing rider behavior or changes in the bike-sharing system itself. Overall, analyzing these trends provides valuable insights for understanding evolving patterns of bike usage and needs of riders.

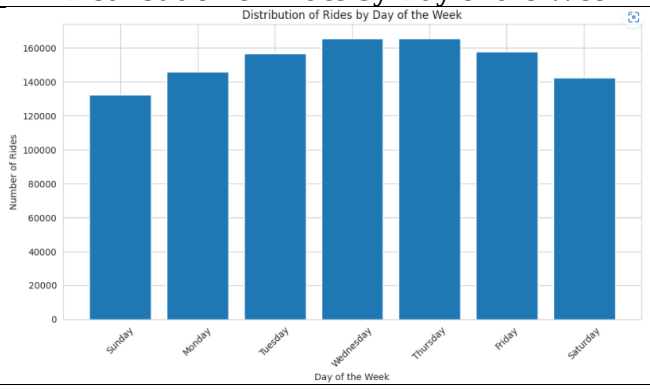
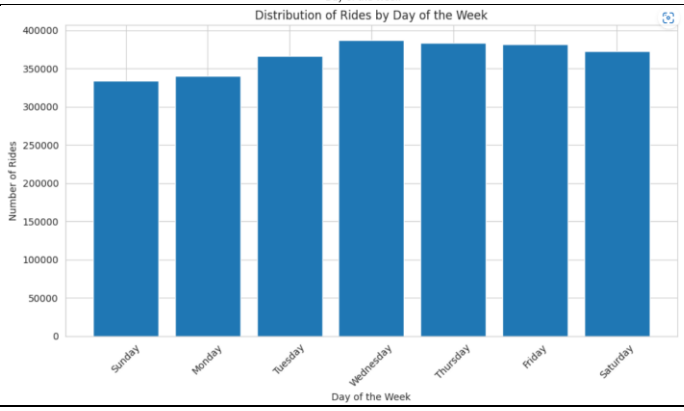
4. Which stations are most popular?

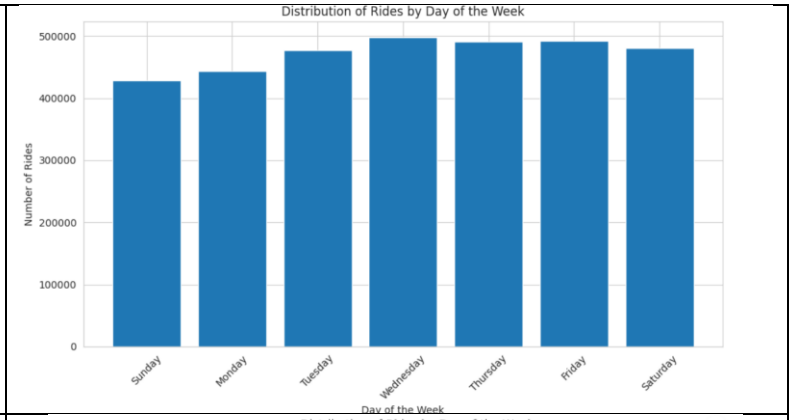
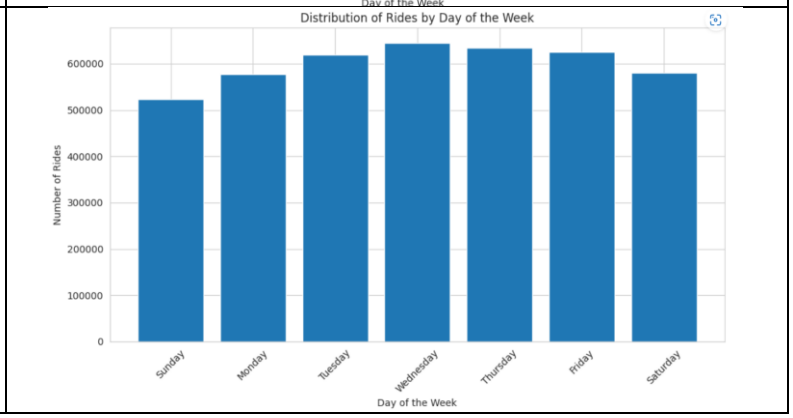
Year	Station Name
2023-2024	Grove St PATI Hoboken Terminal .. South Waterfront .. Hoboken Terminal .. City Hall - Washi.. Newport PATI Hamilton Parl Newport Pkwy Bergen Ave & Sip Av 11 St & Washingto..
2021-2024	Grove St PATH Hoboken Terminal ... South Waterfront ... Hoboken Terminal ... Newport Pkwy Newport PATH Hamilton Park City Hall - Washi... Marin Light Rail Hoboken Ave at Mo...
2019-2024	Grove St PATH Hamilton Park Newport Pkwy Hoboken Terminal ... Newport PATH South Waterfront ... Hoboken Terminal ... Marin Light Rail Harborside Liberty Light Rail

2016-2024	Grove St PATH Hamilton Park Newport PATH Newport Pkwy Marin Light Rail Sip Ave Hoboken Terminal ... Liberty Light Rail South Waterfront ... City Hall

Insights: The most popular Citi Bike stations are Grove St PATH, Hoboken Terminal - River St & Hudson Pl, South Waterfront Walkway - Sinatra Dr & 1 St, Hoboken Terminal - Hudson St & Hudson Pl, City Hall - Washington St & 1 St, Newport PATH, Hamilton Park, Bergen Ave & Sip Ave, Newport Pkwy, and 11 St & Washington St. These stations likely experience high levels of activity due to their strategic locations, such as being near transportation hubs, residential areas, or popular destinations like IT hub, etc. Understanding the popularity of these stations is valuable for Citi Bike operators as it helps them optimize bike distribution, station maintenance, and infrastructure planning to better serve riders in these areas.

5. What days of the week are most rides taken on?

Year	Distribution of Rides by Day of the Week
2023-2024	
2021-2024	

2019-2024	
2016-2024	

Insights: The distribution of rides by day of the week showcases interesting trends, with an increase observed from Monday through Wednesday, followed by a peak on Thursdays and Fridays, and then a decline from Friday to Saturday and Sunday. This pattern suggests that ridership steadily builds up from the beginning of the week, reaching its peak towards the end, particularly on Thursdays and Fridays. The drop in rides during the weekend, especially from Friday to Sunday, indicates a decrease in usage as the week transitions into the weekend. Understanding these fluctuations in ride distribution across different days of the week is valuable for Citi Bike operators, as it helps them anticipate and manage demand variations effectively.

4. **Model Selection:** We employ the Prophet model, a state-of-the-art time-series forecasting algorithm developed by Facebook. Prophet is well-suited for analyzing data with strong seasonal patterns and is capable of handling holidays and special events, making it ideal for our bike-sharing demand prediction task.

5. **Training Model:** The dataset was split into training and testing sets, with records from 2024 utilized for testing and other for training. Columns were renamed to adhere to the format required by the Prophet model (date as 'ds' and trips as 'y'). Our approach involved systematically training and evaluating the Prophet model on various subsets of the data, starting with smaller datasets and progressively expanding to assess the impact of increased data volume on forecasting accuracy. Throughout the project, we developed 12 distinct models by combining three different model configurations (baseline, incorporating holidays and special events, and considering weather effects) with four distinct dataset periods (2016-2024, 2019-2024, 2021-2024, and 2023-2024). The Prophet model was trained on the training dataset and performance was evaluated on testing dataset. Interactive visualizations were developed to facilitate analysis of forecasting results.

6. **Incorporating External Factors:** In addition to historical bike-sharing data, we incorporate external factors such as weather conditions, holidays, and special events into the forecasting model. This enables us to capture the influence of these factors on bike-sharing demand and improve forecasting accuracy.

7. **Evaluation:** We assess the performance of our forecasting model using standard metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics provide insights into the accuracy and reliability of our model's predictions, allowing us to quantify its performance against ground truth values. In addition to evaluating prediction accuracy, we also compare the runtime performance of our model. We conduct experiments by running the code on both single and multiple nodes, leveraging Apache Spark's distributed computing capabilities. By measuring runtime performance on different computing architectures, we aim to assess the scalability and efficiency of our solution.

7. Evaluation

In our evaluation, we assess the performance of our forecasting solution across three key scenarios: baseline, holidays and special event incorporation, and weather integration. Performance metrics includes Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) were computed.

1. **Baseline Comparison:**
   We establish a baseline by evaluating the performance of our initial forecasting model without considering external factors such as holidays, special events, or weather conditions. This baseline model serves as a reference point for assessing the impact of incorporating additional features into our forecasting approach. By comparing the baseline model's performance to the other scenarios, we can quantify the improvements achieved through feature incorporation.

Year	RMSE	MAE	MAPE
2023-2024	584.35	503.49	0.32
2021-2024	523.65	460.66	0.28
2019-2024	480.48	357.04	0.29
2016-2024	606.27	462.43	0.38

   Across the years examined, there are fluctuations in the performance of the model. For instance, from 2016 to 2024, there's a noticeable increase in RMSE, indicating a higher level of prediction error over time. Conversely, MAE and MAPE show variations, with some periods demonstrating improved accuracy compared to others. Notably, the period from 2019 to 2024 exhibits the lowest MAE and MAPE values, suggesting better overall model performance during this time frame.

2. **Holidays and Special Event Incorporation:**
   We trained a second model that takes into account holidays and even the impact of the Covid-19 pandemic. To do this, we made a special list of US Federal Holidays using a pandas library called pandas.tseries.holiday. We also considered covid-19 pandemic effect starting from March 15, 2020 to March 1, 2023.

Year	RMSE	MAE	MAPE
2023-2024	550.92	472.02	0.31

2021-2024	524.13	461.77	0.28
2019-2024	459.75	391.78	0.27
2016-2024	506.88	366.41	0.31

The evaluation table demonstrates that integrating holidays and considering the impact of the Covid-19 pandemic into the forecasting model leads to enhanced prediction accuracy across different time periods. Specifically, we observe consistent reductions in RMSE, MAE, and MAPE values compared to the baseline model. This suggests that accounting for holidays and pandemic effects allows the model to better capture the underlying patterns in Citi Bike rental demand data.

3. **Weather Integration:**
   In addition to holidays and special events, we further refine our forecasting model by integrating weather factors like snow and precipitation in our analysis. Weather conditions play a significant role in bike-sharing behavior, affecting factors such as number of riders and ride durations. By incorporating weather data into our model, we aim to capture the impact of weather on bike-sharing demand and improve forecasting accuracy.

Year	RMSE	MAE	MAPE
2023-2024	640.14	521.14	0.33
2021-2024	509.72	454.53	0.27
2019-2024	414.61	353.83	0.24
2016-2024	482.08	347.56	0.29

The results demonstrate tangible improvements in forecasting performance across different time periods, as indicated by reduced RMSE, MAE, and MAPE values compared to previous models. For example, in the period from 2019 to 2024, RMSE decreases from 459.75 to 414.61, MAE decreases from 391.78 to 353.83, and MAPE decreases from 0.27 to 0.24. This underscores the importance of considering weather factors in forecasting models to better reflect real-world conditions and improve the reliability of predictions for Citi Bike rentals.

Runtime Evaluation:

In evaluating our code's runtime performance, we conducted tests on both a single node and multiple Spark nodes. By measuring the time taken for different stages of our code, we aimed to assess how execution time varied with computational resources. Transitioning from a single node to multiple Spark nodes allowed us to gauge the scalability and efficiency of our code, as distributed computing potentially reduces processing time through parallel execution. This evaluation provided valuable insights into the performance characteristics of our code, guiding decisions on resource allocation and optimization for future implementations.

Year		Single Node	Multiple Nodes
2023 - 2024	Data Preparation	11.78s	11s
	Data Cleaning	41.26s	28.69s
	Data Analysis	51.79s	35.69s
2021 - 2024	Data Preparation	23s	20.36s
	Data Cleaning	1.28 min	1.09 min

	Data Analysis	1.54 min	1.23 min
2019 - 2024	Data Preparation	30.97s	27.83s
	Data Cleaning	2 min	1.24 min
	Data Analysis	2.35 min	1.43 min
2016 - 2024	Data Preparation	44.35s	38.17s
	Data Cleaning	2.38 min	1.52 min
	Data Analysis	3.32 min	2.13 min

The results indicate improvements in runtime efficiency when utilizing multiple nodes for data processing tasks. For instance, in the period from 2023 to 2024, data preparation tasks show a slight improvement in runtime on multiple nodes compared to a single node, with execution times decreasing from 11.78s to 11s. Similarly, data cleaning and analysis tasks demonstrate more significant reductions in runtime on multiple nodes, highlighting the scalability benefits of distributed computing. This trend is consistent across all evaluated time periods, with data cleaning and analysis tasks consistently showing decreased runtime on multiple nodes compared to a single node. Overall, the results underscore the advantages of leveraging distributed computing frameworks like Apache Spark for handling large volumes of data efficiently, leading to improved runtime performance in data processing workflows.

8. Conclusion

In conclusion, our project has demonstrated the effectiveness of leveraging advanced forecasting techniques, specifically the Prophet model, to predict Citi Bike rental demand. By systematically training and evaluating 12 distinct models across different time periods and considering various external factors such as holidays, Covid Pandemic effect and weather effects, we have gained valuable insights into the dynamics of bike-sharing usage patterns. Our approach has highlighted several strengths, including the ability to accurately capture temporal trends and seasonality, as well as the flexibility to incorporate external factors for improved forecasting accuracy.

Moving forward, integrating more granular and diverse datasets, such as real-time weather data at the station level and demographic information, could provide deeper insights into the factors influencing bike-sharing demand. Furthermore, incorporating machine learning techniques for anomaly detection and adaptive modeling could improve the model's robustness in handling unexpected events or disruptions. Overall, our project lays the groundwork for continued research in forecasting methodologies for urban mobility systems.