Maximizing Profit of Citi Bike with Data

Manufacturing Data Science Term Project

Po-Yen Chu	Jie-Xiang Xu	Sen-Yun Gu	Leonard Tsai	Zheng-Liang Wu
B10704031	B09704079	B09704061	B10705010	B10704027
@ntu.edu.tw	@ntu.edu.tw	@ntu.edu.tw	@ntu.edu.tw	@ntu.edu.tw

1 Research Motivation

Shared bicycles not only provide convenient transportation options but also help reduce urban traffic congestion and environmental pollution. However, shared bicycle operators need to effectively manage bicycle stations to ensure the efficient use of vehicles and parking spaces while increasing operating profits. To achieve this goal, this project aims to optimize the operation of shared bicycle stations to increase borrowing volume, membership, and reduce costs.

2 Research Background

Shared bicycles have become a popular mode of transportation in modern cities. While the largest shared bicycle system in Taiwan is YouBike, detailed data is difficult to collect. Therefore, this group has opted to analyze the borrowing data of the Citi Bike system in New York City, which is also the largest shared bicycle system in the United States. This analysis aims to understand the static and dynamic factors affecting Citi Bike and establish predictions for future borrowing volume to plan staffing and resource allocation.

3 Problem Definition

The target unfolds as shown in **Figure 1**. At the first level, profit is unfolded into revenue enhancement and cost reduction. Firstly, regarding revenue, we aim to balance the supply and demand of stations to ensure that each area has bicycles available for borrowing. To achieve this, our proposed solution is to predict the number of permanent parking pillars and the number of dispatch personnel needed for each time slot in the next quarter. On the cost reduction side, we further divide it into reducing labor scheduling costs and reducing member feedback bonuses.

In terms of reducing labor scheduling costs, we analyze the impact of time-related

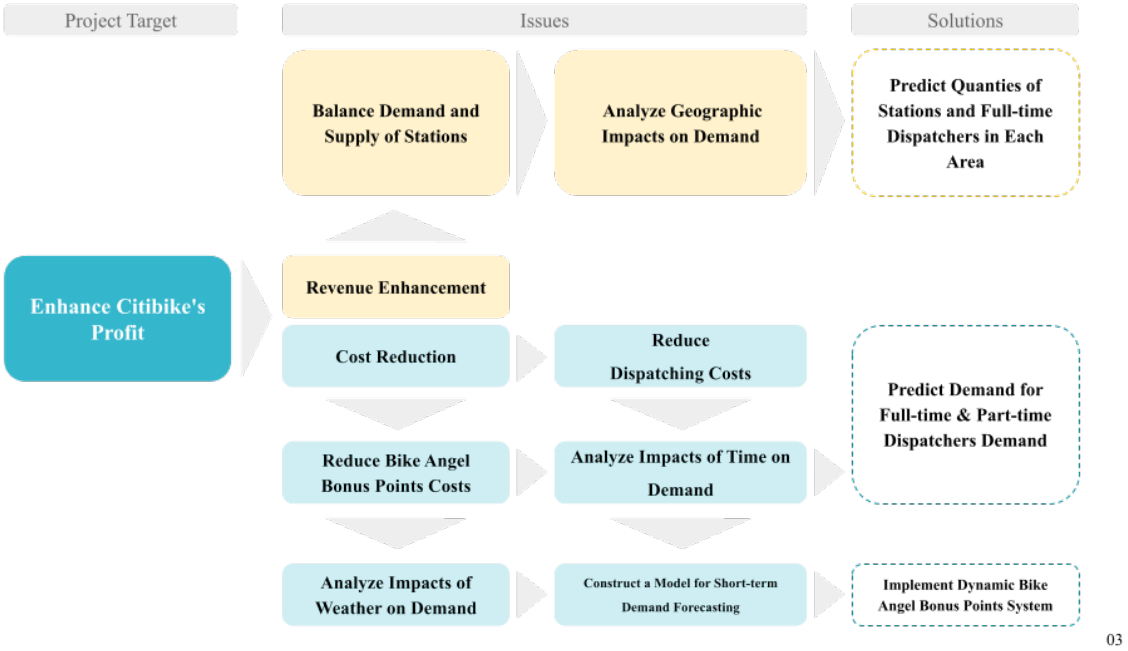

Figure 1: Issue Tree

dynamic data on demand volume, and use this to forecast the allocation of permanent and temporary personnel for each time slot in the next quarter, establishing a demand prediction model for small intervals. Regarding reducing member feedback bonuses, since Citi Bike provides bonuses to members for helping to dispatch vehicles to popular areas, we aim to control the distribution of member feedback bonuses through such interval prediction models.

4 Data

4.1 Data Exploration

The primary dataset for this research is the borrowing data of Citi Bike in May 2023 obtained from Kaggle. To accurately grasp the condition of Citi Bike users borrowing bicycles, we analyze the borrowing conditions during weekdays and holidays, as shown in **Figure 2**. It can be observed that the peak borrowing periods on weekdays are from 8:00 to 9:00 and from 16:00 to 19:00, while on holidays, the peak period is from 14:00 to 17:00. This can be attributed to the commuting needs of working individuals on weekdays, leading to higher demand for bike borrowing, while on weekend afternoons, the public tends to use Citi Bike for leisure and recreational purposes.

Next, we analyzed the membership borrowing ratio of Citi Bike from Monday to Sunday and different time periods within a day to understand the borrowing patterns of members and non-members. The analysis results are shown in **Figure 3**.

It can be observed that the membership borrowing ratio on weekdays (exceeding 80%) is higher than on holidays (70%). The highest membership ratio occurs in the morning (exceeding 90%), while the lowest occurs in the afternoon and early morning

Figure 2: Average Borrowing Volume of Citi Bike

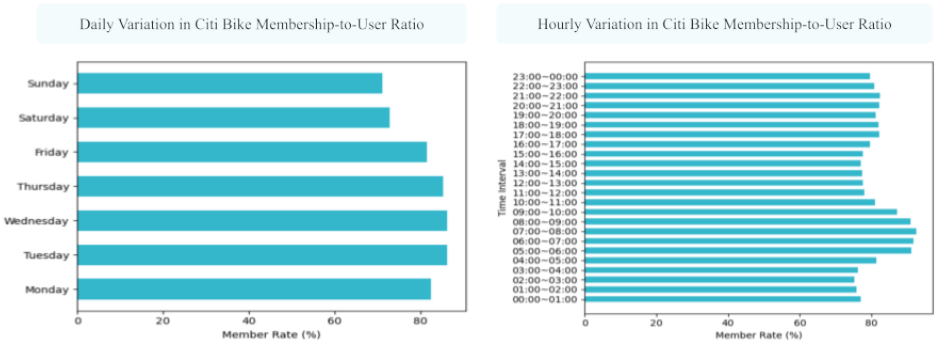

Figure 3: Citi Bike Membership-to-User Ratio

(75%). This suggests that working individuals have a higher membership ratio, while non-members are more likely to use shared bikes for leisure during weekends. Therefore, it is suggested to deploy staff at popular stations during weekend afternoons to understand the demand for station expansion and bike allocation before promoting membership, aiming to increase customer lifetime value and company revenue.

4.2 Data Sources

4.2.1 Citi Bike Borrowing Data

The borrowing data for Citi Bike in May 2023, as well as from November 2021 to October 2023, were obtained from the official website of Citi Bike NYC. The former is primarily used for subsequent station borrowing volume prediction and adjustments of bike racks and staff, while the latter is used in conjunction with periodic variables (such as weather) for LSTM and real-time prediction models. (Due to the large size of the dataset, it has not been uploaded to the submission area.)

4.2.2 Weather Data

The weather data used in this study were sourced from timeanddate.com. Historical observation data were preprocessed into numeric and dummy variables.

Column	Temp	Wind	Humidity	Barometer	Visibility	Weather Description (Dummy)
Unit	℃	km/h	%	mbar	km	None
Instance	20	9	72	1010	16	Heavy snow, Sunny, Rainy… (30 columns in total)

Table 1: Weather Data Summary

4.2.3 Traffic Volume Data

The traffic volume data for May 2023 were sourced from NYC OpenData. Using the filtering function on the website's database, the average traffic volume data for three Neighborhood Areas in Manhattan's city center were selected by weekday and time period. This dataset was integrated with the main dataset (borrowing data for May 2023) for analysis. The "WktGeom" field was converted into longitude and latitude format for visualization purposes, and "Longitude" and "Latitude" fields were created.

Column	weekday	hour	Longitude	Latitude	average_vol	neighbourhood
Instance	Friday	00:00 ∼01:00	-73.989...	40.772...	22.25	Harlem

Table 2: Traffic Volume Data Summary

4.2.4 Demographic, Telecommunications, and Public Infrastructure Data

The demographic, telecommunications, and public infrastructure data used in this study were sourced from NYC OpenData. Data for nearly 30 Neighborhood Areas in Manhattan were selected through the database filtering function. Variables such as population density, number of public schools, and public infrastructure penetration rate are relevant to borrowing volume.

To simplify the subsequent data integration process, the central coordinates of the Neighborhood Areas were chosen as their latitude and longitude coordinates, and two additional columns were added to the dataset. Furthermore, the NTA (Neighborhood Areas identification number) column in the dataset was used to match corresponding Neighborhood Areas and a new column was added for this purpose, facilitating mutual integration with other datasets. Below is a summary table of several representative fields:

Column	Total Population	Population Density (per Sq. Mi.)	Low-Income Housing
Instance	52,529	82,076.56	2,270

Number of Public Schools	Street Furniture: Bike Shelters	Empire City Subway Coverage (Percentage)
20	3	0.43

Table 3: Demographic, Telecommunications, and Public Infrastructure Data Summary

4.3 Data Preprocessing

Due to the data from timeanddate.com not being consistently observed, although the data are mostly concentrated at 51 minutes of observation, there are still instances where an hour has multiple observations or consecutive hours without observation. Therefore, to obtain continuous hourly data, the missing values are filled as follows:

- Multiple data within an hour: For continuous variables, the average value is taken. For dummy variables, the maximum value is taken (i.e., all weather conditions observed during that hour are considered).
- No data for that hour: For continuous variables, the linear average of the previous and next data is taken. For dummy variables, the middle point between the previous and next data is taken as the split point.

5 Forecasting Usage for Manhattan Areas

5.1 Data Scope

This part uses the dataset combined from "Citi Bike borrowing volume data for May 2023" and "Population characteristics, telecommunications, and public infrastructure penetration rate data" from various neighborhoods. Additionally, we calculate the membership composition ratio for each area from the "Citi Bike borrowing volume data for May 2023" and add this as a new field. The total number of variables in the original dataset is 125.

5.2 Feature Selection

1. Initial variable selection: 44 key factors affecting usage volume are selected from 125 area data.
2. Data transformation: Population characteristic data for each area mostly have the same values in different records, so these values are grouped by the NTA field and averaged. The data on the types of buildings have finer dimensions compared to the Neighborhood Areas, so these field data are also grouped by the NTA field and summed.
3. Removal of highly correlated variables: Due to high correlation among many variables, removing all variables with a correlation above 0.7 would significantly reduce the number of variables in the dataset. Therefore, after adjusting the threshold for removing variables multiple times, variables that are correlated above 0.7 with three or more other variables are ultimately removed to maintain most of the information in the dataset without overfitting or multicollinearity issues.

4. Selection of important variables: After conducting VIF multicollinearity tests, it was found that some variables had multicollinearity issues. Therefore, Lasso Regression was used to select important variables to help reduce multicollinearity problems in the dataset. In the end, only 37 variables were included in the model.

5.3 Linear Regression Analysis Results

The Linear Regression results are presented below, utilizing forward selection with a p-value threshold of 0.05.

OLS Regression Results

Dep. Variable:	group_count	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.965
Method:	Least Squares	F-statistic:	124.3
Date:	Sat, 09 Dec 2023	Prob (F-statistic):	2.69e-15
Time:	21:34:20	Log-Likelihood:	-301.95
No. Observations:	28	AIC:	617.9
Df Residuals:	21	BIC:	627.2
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8516.5871	7836.662	-1.087	0.289	-2.48e+04	7780.643
bldgclass_sum	0.0271	0.002	12.173	0.000	0.022	0.032
Street Furniture: Parking Pay Stations	137.5436	26.332	5.223	0.000	82.783	192.304
bldgclass_W	-643.7581	182.306	-3.531	0.002	-1022.883	-264.633
Empire City Subway Coverage (Percentage)	4.065e+04	1.22e+04	3.328	0.003	1.52e+04	6.61e+04
Low-Income Housing (NYCHA)	1.3834	0.480	2.881	0.009	0.385	2.382
bldgclass_Q	-313.8403	124.751	-2.516	0.020	-573.274	-54.407

Omnibus:	1.793	Durbin-Watson:	2.253
Prob(Omnibus):	0.408	Jarque-Bera (JB):	0.829
Skew:	0.389	Prob(JB):	0.661
Kurtosis:	3.322	Cond. No.	1.33e+07

Figure 4: Linear Regression Analysis Summary

The results of the linear regression analysis presented above allow us to categorize the regression coefficients of each variable into two groups: positive coefficients for the total number of buildings, number of paid parking lots, subway coverage rate, and number of low-income households; negative coefficients for the number of educational institutions and outdoor recreational facilities. Below, we interpret the coefficients of these variables individually:

- The positive coefficient for the total number of buildings: It is speculated that areas with a higher total number of buildings tend to have a larger population, thus resulting in a significant increase in the demand for Citi Bike usage.
- The positive coefficient for the number of paid parking lots: It is speculated that areas with a higher number of paid parking lots indicate more vigorous commercial activities, attracting a large crowd to the area and consequently increasing the demand for Citi Bike borrowing.
- The positive coefficient for subway coverage rate: It is speculated that commuters often use Citi Bike to travel from home to subway stations and then from subway

stations at their destinations to their workplaces, leading to a significant increase in Citi Bike usage demand.

- The positive coefficient for the number of low-income households: It is speculated that Citi Bike, being relatively low-cost compared to other public transportation options, is preferred as a means of transportation by members of low-income households.

- The negative coefficient for the number of educational institutions: It is speculated that there are fewer suitable routes for Citi Bike usage within the vicinity of school campuses, with students or visitors mostly moving around within the campus by walking. Additionally, students who regularly use bicycles to travel between campuses are more likely to purchase bicycles rather than rent them, thus significantly reducing the demand for Citi Bike borrowing.

- The negative coefficient for the number of outdoor recreational facilities: It is speculated that as Citi Bike usage is also considered a form of recreational activity, an increase in the number of outdoor recreational facilities serves as a substitute variable, thereby significantly reducing the demand for Citi Bike borrowing.

5.4 Model Extension and Application

By incorporating the predicted values of various variables for the next quarter, such as the total number of buildings, the number of paid parking lots, subway coverage rate, number of low-income households, number of educational institutions, and number of outdoor recreational facilities into a linear regression model, we can calculate the borrowing demand for each region in the next quarter. Meanwhile, we also incorporate some simple computational logic (refer to **Figure 5**) to determine the demand for parking docks in each region for full-time dispatchers (refer to **Figure 6**).

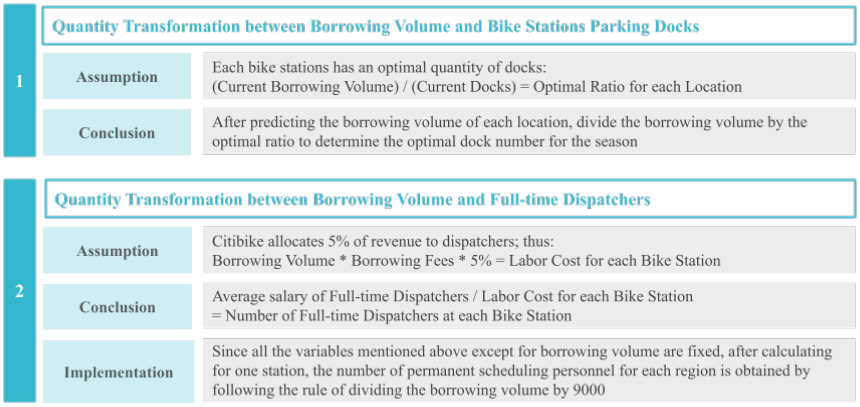

Figure 5: Quantity Transformation between Borrowing Volume, Bike Stations Parking Docks and Full-time Dispatchers

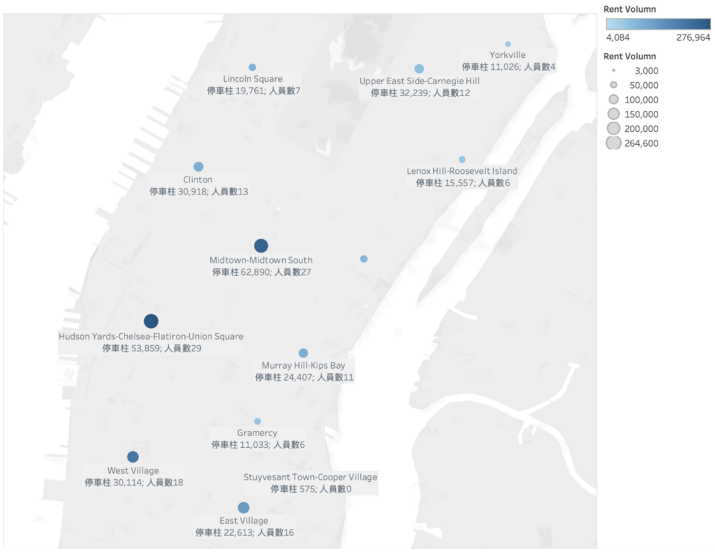

Figure 6: Demand for Parking Docks in each Region for Full-time Dispatchers

Furthermore, for the convenience of decision-making on the enterprise side, we subtract the predicted borrowing demand, the demand for parking docks for full-time dispatchers, and the data from the previous quarter to obtain the increase (decrease) in borrowing demand, the demand for parking docks for full-time dispatchers in the next quarter (refer to **Figure 7**).

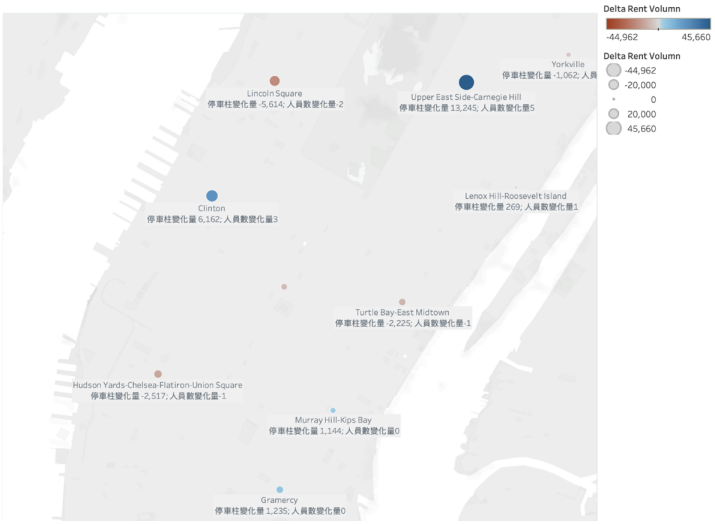

Figure 7: Demand for Parking Docks in each Region for Full-time Dispatchers in the Next Quarter

6 Optimizing the Schedule for Full-time Dispatchers

With the analysis of geographic impacts completed, the focus now shifts to examining weather and time factors to refine the schedule for full-time dispatchers. This optimization aims to minimize labor costs while maximizing operational efficiency.

6.1 Feature Selection & Engineering

6.1.1 Weather Variables Organization

For binary weather variables, 15 types from May 2023 are one-hot encoded. Additionally, since there can only be a maximum of two weather variables per hour, feature fusion is performed to avoid multicollinearity issues caused by the majority of weather variables being 0. Finally, it is organized into four major weather variables. The classification is as follows:

- Sunny: Clear, Sunny, Partly sunny, Passing clouds, Scattered clouds
- Cloudy: Partly cloudy, More clouds than sun, Mostly cloudy, Broken clouds, Cloudy, Overcast, Low clouds
- Rainy: Light rain, Rain
- Foggy: Fog

6.1.2 Clustering and Data Transformation

Based on the 2.19 million borrowing samples in Manhattan for May 2023, clustering is performed for each hour of each day to calculate the borrowing volume and the proportion of member borrowing per hour. Humidity is transformed into a percentage, and traffic flow is averaged within each time period.

6.1.3 Collinearity Check

For the multicollinearity problem of continuous variables, VIF inspection is conducted, and variables with results exceeding 10 are deleted, including air pressure and visibility.

6.2 Linear Regression Analysis Results

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2387.4228	989.353	-2.413	0.016	-4329.717	-445.129
average_vol	14.2424	0.397	35.895	0.000	13.463	15.021
Sunny	9.4882	868.171	0.011	0.991	-1694.903	1713.879
Cloudy	244.8884	859.900	0.285	0.776	-1443.264	1933.041
Rainy	-1317.0857	269.146	-4.894	0.000	-1845.471	-788.700
Foggy	-316.1605	862.813	-0.366	0.714	-2010.031	1377.710
Temp	132.7179	12.669	10.476	0.000	107.847	157.589
Wind	-5.8137	9.207	-0.631	0.528	-23.890	12.262
Humidity	-3.1878	3.695	-0.863	0.389	-10.442	4.067

(a) Borrowing Volume per hour

	coef	std err	t	P>|t|	[0.025	0.975]
const	85.1151	4.324	19.684	0.000	76.626	93.604
average_vol	0.0278	0.002	16.017	0.000	0.024	0.031
Sunny	0.8338	3.794	0.220	0.826	-6.615	8.283
Cloudy	0.3316	3.758	0.088	0.930	-7.046	7.710
Rainy	-0.2181	1.176	-0.185	0.853	-2.527	2.091
Foggy	1.8067	3.771	0.479	0.632	-5.596	9.210
Temp	-0.7897	0.055	-14.263	0.000	-0.898	-0.681
Wind	-0.0009	0.040	-0.022	0.983	-0.080	0.078
Humidity	0.0546	0.016	3.379	0.001	0.023	0.086

(b) Membership-to-user Ratio

Figure 8: Relationship between Weather Variables, Traffic Flow and Two Target Variables

6.2.1 Relationship between borrowing volume per hour, weather variables and traffic flow

Through OLS regression model, with a significance level set at 0.05, it can be inferred that as traffic flow increases, borrowing volume also increases; borrowing volume decreases by 1317 times per hour on rainy days; for each degree increase in temperature, borrowing volume increases by 132 times.

When dividing a day into four periods for OLS regression analysis and examining its relationship with traffic flow and weather variables, the results obtained with a significance level set at 0.05 are as follows:

1. Overall: Both temperature and traffic flow increase, borrowing volume increases.
2. Early Morning Period (1-6 am): Other variables are not significant, indicating that users in the early morning period are the least affected by weather factors.
3. Morning Period (7-12 am): Borrowing volume increases by 570 times on sunny days.
4. Afternoon Period (1-6 pm): Borrowing volume increases by 1173 times on sunny days, 967 times on cloudy days, and decreases by 1661 times on rainy days. It is inferred that users in the afternoon period are most sensitive to weather changes.
5. Evening Period (7-12 pm): Borrowing volume decreases by 1760 times on rainy days, indicating that users in the evening are more concerned about poor visibility on rainy days and choose not to ride shared bicycles.

6.2.2 Relationship between membership-to-user ratio per hour, Weather variables and traffic flow

Through the OLS regression model, with a significance level set at 0.05, it can be inferred that as traffic flow increases and humidity rises, the membership-to-user ratio also increases, while the membership-to-user ratio decreases with higher temperatures. Therefore, for members, even in cold temperatures and high humidity, they choose to ride shared bicycles, indicating a lower sensitivity to weather, whereas non-members are more sensitive to weather changes.

6.3 Model Extension and Application

Based on the required number of full-time employees in different areas, after considering the borrowing demand at different times and weather conditions, suitable employee schedules are planned as the results of prescriptive analysis (refer to **Figure 9**).

6.3.1 Employee Allocation Principles

Each employee works 40 hours a week, not exceeding 8 hours a day, with more employees on weekdays than weekends, and more daytime employees than nighttime employees. Additionally, the number of employees increases in high temperatures and decreases on rainy days.

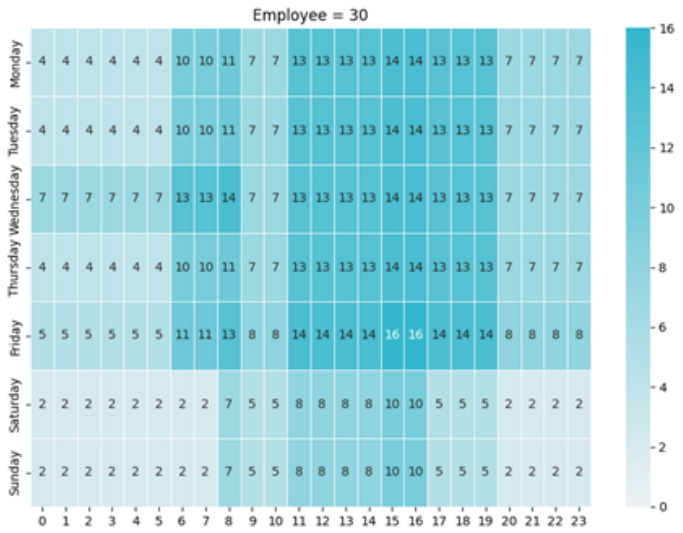

Figure 9: The number of employees in each period when the number of employees reaches 30

1. Weekdays: Employees are split for work, satisfying both the scheduling needs for going to work and returning home simultaneously. On Wednesdays and Fridays, the number of nighttime scheduling employees is increased to meet the borrowing demand for vehicles at various stations during the day, and the highest number of employees is scheduled on Fridays to prepare for the weekend by scheduling vehicles in advance.

2. Weekends: The highest number of employees work in the afternoon on weekends to meet the public's usage habits. Employees are also scheduled to promote membership plans to non-members on sunny days with high temperatures and low humidity.

7 Utilizing Weather Data to Predict Short-Term Hourly Borrowing Volume

7.1 Data Overview

- X: Weather data and isHoliday (whether it is a national holiday)
- Y: ['Count'] (borrowing volume)
- Training Scope: November 2021 to October 2023
- Testing Scope: November 2023

7.2 Time Series LSTM Model

7.2.1 Constructing Phase

The design of LSTM is such that inputting the preceding 720 hours of X will yield the Y of the last hour in these 720 hours. Therefore, the timestep of the sequence is 720.

Unlike typical LSTMs that predict the future based on the past, since the predicted X is expected data (weather data can mix historical weather data with weather forecasts, and whether it is a national holiday is predetermined), this model places the predicted hour's X into input.

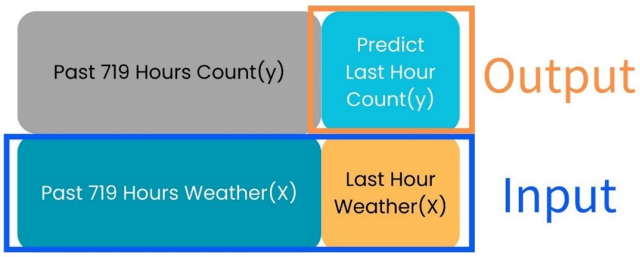

Figure 10: The Data Scope of Input & Output

7.2.2 Tuning Phase

This model uses Adam optimizer, and several loss functions were tested and selected (see **Figure 11**). The x-axis of the figure represents epochs, the y-axis represents batch, and the z-axis represents the MSE of the testing data (the scales of the z-axis of the three graphs are different). It can be observed that using MAE as the loss function and a batch size of 300 results in the lowest and relatively stable MSE, thus choosing these hyperparameters to adjust our model.

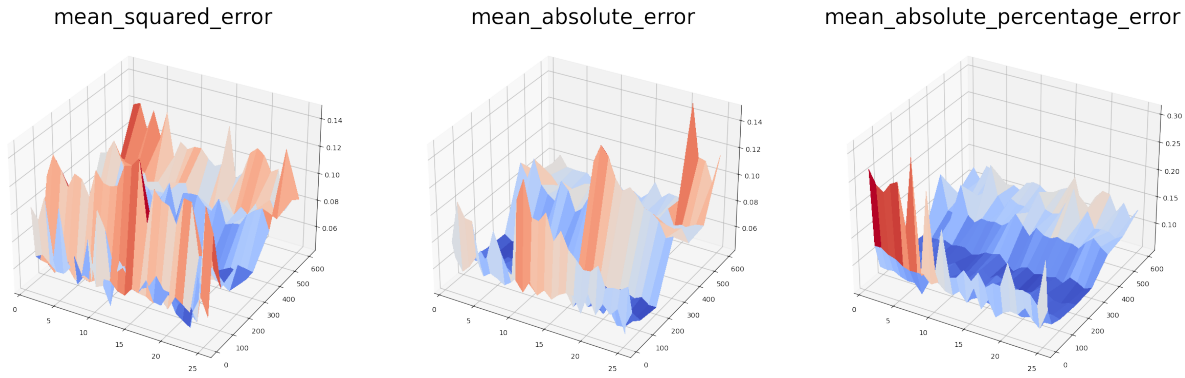

Figure 11: Model Evaluation utilizing Three Different Loss Functions

The performance of using this model on Testing Data (November 2023) is as follows (MSE_before_inverse_transform = 0.050742034) (refer to **Figure 12**).

7.2.3 Result Interpretation

As shown in the figure, this model has a certain accuracy in predicting the borrowing volume per hour. Based on the characteristics of this model and weather forecasts, enterprises can use it for short-term predictions of the total borrowing volume in the Manhattan area. Therefore, the operational improvement suggestion proposed in this study is the Bike Angel dynamic bonus point system.

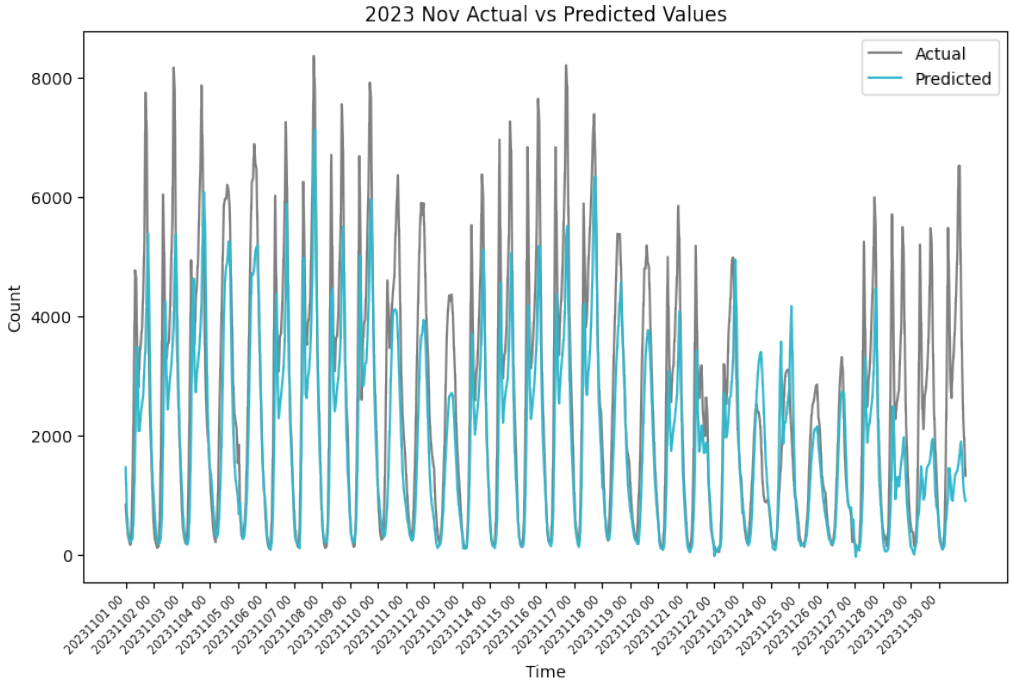

Figure 12: Nov. 2023 Borrowing Volume Prediction

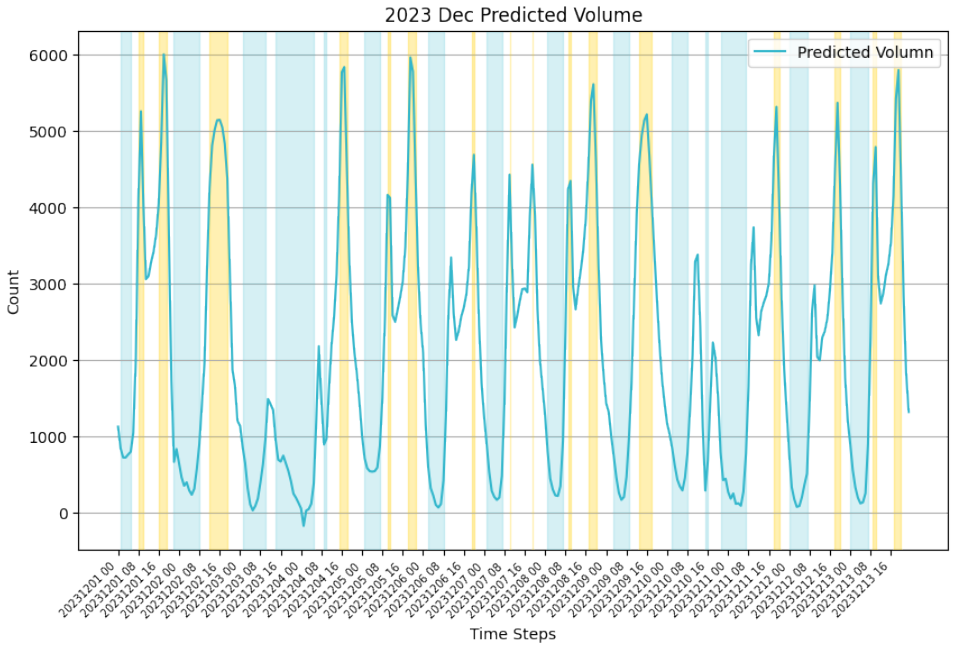

Figure 13: Dynamic Bonus Point System Implemented with December 2023 Borrowing Volume Prediction

7.3 Dynamic Bike Angel Bonus Point System

LSTM models can predict the borrowing volume per hour for the next week, allowing Citi Bike to prepare in advance as follows:

Firstly, the prediction results can be used to dynamically adjust the bonus point ratio of Bike Angels. During peak periods where larger borrowing volumes are expected, Citi Bike can increase the bonus point ratio to incentivize users to assist in vehicle scheduling through the Bike Angels system, thereby reducing additional scheduling costs. Conversely, during off-peak periods with lower predicted borrowing volumes, the point ratio can be reduced to minimize unnecessary feedback costs.

Secondly, the prediction results can also be used to evaluate the manpower needs of temporary workers for the next week. During expected peak periods, Citi Bike can expand its temporary workforce in advance; while during off-peak periods, it can reduce the budget for temporary workers. This way, more precise control of labor costs can be achieved.

To improve the prediction accuracy of the LSTM model, Citi Bike adjusts parameters during the model training phase to find the model with the smallest error. More accurate prediction results can make the above operational decisions more reliable.

In conclusion, utilizing the LSTM model to predict changes in borrowing volume per hour can not only help Citi Bike accurately grasp manpower and resource needs but also improve vehicle redistribution efficiency through systems such as dynamic point adjustment, thereby enhancing operational efficiency and reducing costs.

8 Research Conclusion

8.1 Project Summary

This project aims to improve the efficiency of personnel and vehicle scheduling for shared bicycles using multiple linear regression models and LSTM models to increase Citi Bike's profit.

Multiple linear regression models were separately established to predict the demand for parking docks and the number of permanent personnel needed in various areas of Manhattan, as well as the shift schedules for permanent personnel across different time periods in Manhattan as a whole, reducing vehicle idle time and improving vehicle scheduling efficiency. The LSTM model implements a dynamic point feedback adjustment system and evaluates the need for temporary dispatchers on a weekly basis, accurately managing the cost of Citi Bike's Bike Angel member scheduling and the budget for temporary dispatchers.

By implementing the above solutions, Citi Bike can increase revenue and reduce costs, ultimately achieving the goal of increasing profits for Citi Bike.

8.2 Future Research Suggestions

8.2.1 Data Limitations

- Due to the lack of information on parking dock vacancies at stations where the number of parking docks varies, the relationship between borrowing volume demand and the actual scheduling needs cannot be confirmed.

- Weather data coverage: This study can only find data for Central Park, New York as

the observation station. Whether this model can be expanded to other areas of New York for application remains to be discussed, and due to the inherent limitations of weather data, slight differences in weather factors between different areas may affect the accuracy of the model.

- Data maintenance and prediction: With current weather forecasting technology, the time frame for high-accuracy forecasts is very short, affecting the timeliness of the LSTM model. Additionally, if the update frequency of the original data sets for different regions is inconsistent, it will be difficult to maintain, and predictions of factors for different regions (such as population growth rates in the region) usually require rigorous research.

8.2.2 Issues Not Covered by the Study

- Citi Bike data outside Manhattan: Due to the large size of the original data set and limitations in hardware and computing power, we only sampled data from the Manhattan area for analysis. Future research can expand the scope of study to explore more possible influencing factors.
- Analysis of network relationships between station geographical locations.
- This data set includes the starting and ending points and times of each trip, which can be used for analyzing factors such as these with data analysis tools like GNN.
- Combining the solutions proposed in this study with operational research methods to plan scheduling issues (such as minimizing scheduling costs).
- Geographical terrain data from Manhattan can also be incorporated into the data set to analyze issues such as whether terrain (such as uphill or average slope) affects riding willingness.

8.2.3 Introduction of Automation Tools

If applied in practice, enterprises can combine the models, analyses, and methods proposed in this study to create automated tools, increasing operational efficiency and improving profits.

References

[1] *Internet Master Plan: Adoption and Infrastructure Data by Neighborhood*. Accessed Nov 4, 2023, `https://data.cityofnewyork.us/City-Government/Internet-Master-Plan-Adoption-and-Infrastructure-D/fg5j-q5nk/about_data`.

[2] *Automated Traffic Volume Counts*. Accessed Nov 5, 2023, `https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/7ym2-wayt/about_data`.

[3] Citi Bike, *Citi Bike Trip Histories*. Accessed Nov 1, 2023, `https://citibikenyc.com/system-data`.

[4] Citi Bike, *How we're rebalancing the Citi Bike system*. Accessed Nov 20, 2023, https://ride.citibikenyc.com/blog/rebalancing-the-citi-bike-system.

[5] timeanddate.com, *Weather in New York, New York, USA*. Accessed Nov 6, 2023, https://www.timeanddate.com/weather/usa/new-york.

[6] *Building Classification | City of New York*. Accessed Nov 4, 2023, https://www.nyc.gov/assets/finance/jump/hlpbldgcode.html.

[7] Citi Bike powered by lyft (2023). *May 2023 Monthly Report*.

[8] *New York Population*. Accessed Nov 24, 2023, https://worldpopulationreview.com/states/new-york-population.