42577 Introduction to Business Analytics

1st Anton Bang Lejbølle
Mail: s254345@student.dtu.dk
StudentID: s254345

2nd Jonas Hansen
Mail: s254354@student.dtu.dk
StudentID: s254354

3rd Ludvig Lindmark
Mail: s254344@student.dtu.dk
StudentID: s254344

4th Mathilde Brinch Sørensen
Mail: s254124@student.dtu.dk
StudentID: s254124

***Report Contribution*—All members contributed equally to the project. Ludvig led the section on data visualisation and analysis and the clustering the stations; Mathilde focused on building a prediction model for station clusters; Anton led the section on aggregate destination trends; and Jonas took the lead on individual ride destination prediction.**

I. Introduction

Citi Bike has become essential to New York City's urban mobility, but operational success requires bikes to be available where demand occurs. Without accurate demand forecasts, the company cannot efficiently reposition its fleet overnight, resulting in either stock-outs or wasteful over-allocation. This paper compares multiple forecasting models to predict hourly demand across station clusters, comparing multiple approaches and translating predictions into specific bike allocation requirements, providing operators with actionable insights for optimal fleet management. Additionally, the paper explores aggregate and individual destination trends to enhance redistribution strategies and to anticipate station-level supply and demand imbalances.

II. Data analysis and visualization

The data used in the analysis is the Citi Bike New York City data for the year 2018. The dataset contains detailed records of individual trips, including start and end time, along with corresponding start and end location, the trip duration and it also includes some limited user characteristics.

To ensure the best predictions possible, the data is initially cleaned. Firstly, the rows where no start nor end station existed were dropped. Secondly, outliers outside the range 16-90 years old were removed. Thirdly, trips with unrealistic durations were excluded. Therefore, the durations were limited to 24 hours. Furthermore, the duration of the trips were calculated to make sure there wasn't a mix-match between the actual trip duration and the duration reported.

In the notebook Data analysis and visualization.ipynb we see the average number of trips. The trips are concentrated around the day time hours, with peaks around the classical rush hours around 7 to 9 am and again at 4 to seven pm. Which tells us that a lot of the users use the bikes for the commute back and forth to work. Furthermore, notebook Data analysis and visualization.ipynb shows that the trips mostly occur on the weekdays rather than weekends. Delving deeper into this, the graph in figure 1 reveals a much clearer pattern.

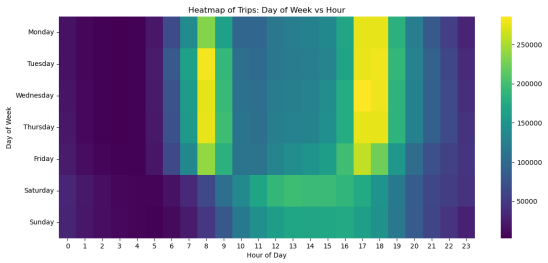

Fig. 1. Heatmap of the hourly pattern across the weekdays

During the workweek there is not only a generally higher demand, but the commuting rush hours clearly stands out. Indicating, that a large part of the demand comes from people commuting to and from work. During the weekends, most of the trips are during the day which could mean that the demand primarily comes from casual users going around the city.

Looking at the monthly demand in the notebook Data analysis and visualization.ipynb, there is evidently a much higher demand during the summer months, with a rise in spring and decline in the fall. With further investigation, the heatmap in figure 2

These temporal patterns with strong weekday peaks and significant seasonality emphasize the importance of incorporating hour-of-day, weekday and monthly effects in our prediction model.

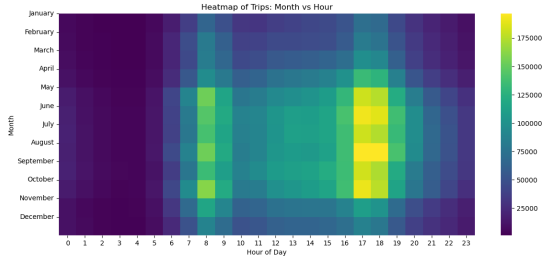

Fig. 2. Heatmap of the hourly pattern across the months

When predicting the hourly demand, the data is split into three separate datasets. The training set, containing the months January to September, the validation set only being October and lastly the test set containing data for November and December. Per the assignment, the data was not allowed to be shuffled. From a time series perspective, the fact that only one year of data was provided to us limits the models exposure to seasonal patterns occurring in these months. The supervised learning models aren't able to train on this pattern which might cause it to struggle when predicting the demand in November and December, potentially yielding lower accuracy and performance of the model.

Finally, due to the fact that K-means relies on Euclidean distance the geographic coordinates were converted to meters before clustering. Transforming the coordinates at approximately 40.7N (New York City's latitude), we used the conversion factors: 1 latitude $\approx$ 111,000 meters and 1 longitude $\approx$ 85,000 meters. This ensures the clustering algorithm measures actual physical distance between stations rather than angular separation in degrees.

III. Prediction Challenge

This task required developing a time-series prediction model to forecast hourly demand (pickups and dropoffs) for 24-hour forecast horizons at the cluster level. We formulated this as a supervised machine learning regression problem, treating hourly demand aggregates as the target variable and systematically engineering temporal features to capture the underlying patterns inherent in bike-sharing demand dynamics.

Part 1

To cluster the stations spatially we initially used the Elbow method to determine how many clusters to use. The graph can be seen in the notebook prediction_challenge.ipynb but unfortunately the results from the method becomes superfluous as there is a requirement in the project description to use a minimum of 20 clusters which is far more than what the Elbow method would recommend. Therefore we choose to split it into K=20 clusters as can be seen in figure 3.

Part 2

We ranked the 20 clusters generated in part 1 by total pickups and selected the one with highest demand and the one with lowest.

2.1 Features

We developed a comprehensive feature set to capture temporal, cyclical, and autoregressive components of demand:

- Temporal Features: Hour of day, day of week, month, binary indicators for weekends, and rush-hour periods (07:00–09:00 and 17:00–19:00 on weekdays).
- Autoregressive Features: Lagged demand at 1-hour, 24-hour, and 168-hour intervals to capture short-term persistence, diurnal cycles, and weekly seasonality patterns.
- Rolling Statistics: 3-hour and 24-hour rolling averages and 24-hour rolling maximum values to provide contextual information about recent and typical demand levels.
- Cyclical Statistics: Mean demand computed by hour of day and day of week, estimated exclusively from training data to prevent information leakage.

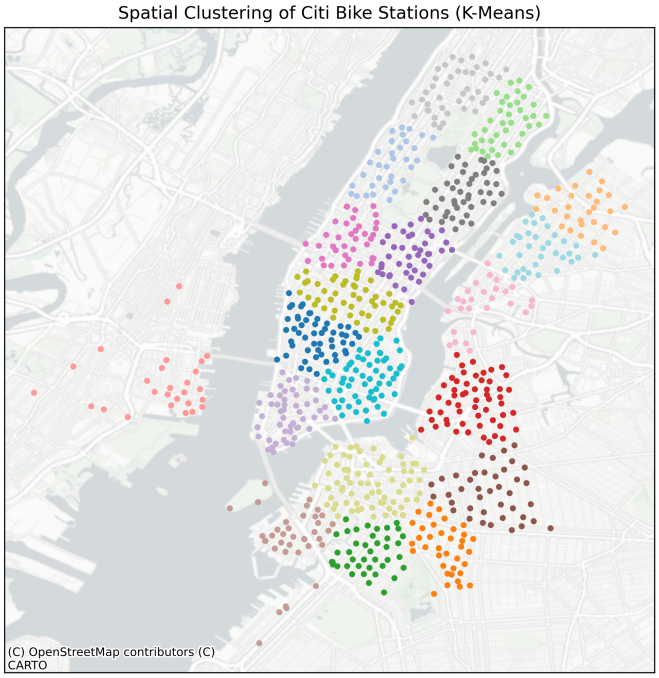

Fig. 3. Spatial Clustering of Citi Bike Stations

2.2 Model Training

Four models were evaluated: a baseline (last week's demand), Ridge Regression, Gradient Boosting (GB) and Hierarchical GB (HGB). Rather than pursuing hyperparameter tuning, fixed, conventionally-established parameter values were used. This approach was justified by the characteristics of the forecasting problem; demand forecasting on clustered trip data is a mature problem with well-understood solutions [1]. Linear models and tree-based ensembles with default configurations typically generalize well when applied to similar problems [2]. Accordingly, Ridge Regression was applied with $\alpha = 1.0$ and Gradient Boosting with $n\_estimators = 150$, $\max\_depth = 5$, and $learning\_rate = 0.1$. These values reflect conventional practice in machine learning and require no tuning to achieve strong performance [3].

Avoiding hyperparameter tuning eliminated the need for a validation set, allowing more data to be retained for training. The time-series was split at October 31, 2018: pre-cutoff observations formed the training set, and post-cutoff observations the test set. This temporal split provides an unbiased estimate of general performance, since the models were fully specified before encountering test data.

Features were standardized before Ridge Regression training to ensure consistent regularization across variables. Gra-

dient Boosting was applied to unscaled features as tree-based models are invariant to feature scaling.

Finally the GB model was augmented with a hierarchical layer to leverage cross-cluster patterns. In the first stage, independent GB models are trained per cluster. In the second stage, GB predictions serve as input to a global neural network and cluster-specific residual models. The global model captures pooled demand patterns, while residual models learn cluster-unique deviations. Final predictions combine both components. The hierarchical layer employed neural networks with two hidden layers (32, 16 units), ReLU activation, early stopping, and L2 regularization ($\alpha = 0.001$).

2.3 Results

Four modeling approaches were evaluated across clusters 16 and 3 using thre performance metrics: coefficient of determination (R^2), mean absolute error (MAE) and Normalized MAE. Results are presented in Table I.

TABLE I
PREDICTION PERFORMANCE ACROSS CLUSTERS

Cluster	Model	R^2	MAE	Norm. MAE (%)
		Pickups		
16	Baseline	0.488	92.41	38.91
16	Ridge	0.815	64.71	27.24
16	GB	0.946	31.48	13.25
16	**HGB**	**0.948**	**30.52**	**12.85**
3	Baseline	0.218	5.77	52.82
3	Ridge	0.627	4.17	38.16
3	GB	0.703	3.81	34.90
3	**HGB**	**0.704**	**3.80**	**34.82**
		Dropoffs		
16	Baseline	0.511	93.01	38.81
16	Ridge	0.818	66.22	27.63
16	GB	0.957	30.22	12.61
16	**HGB**	**0.957**	**29.62**	**12.36**
3	Baseline	0.194	5.68	52.52
3	Ridge	0.655	3.91	36.20
3	GB	0.688	3.73	34.50
3	**HGB**	**0.702**	**3.63**	**33.64**

2.4 Discussion

Gradient Boosting substantially outperforms baseline and Ridge Regression approaches across both clusters for both pickups and dropoffs. On Cluster 16, GB achieves normalized MAE of 13.25% (pickups) and 12.61% (dropoffs) compared to 38.91% and 38.81% for baseline, respectively. Ridge Regression achieves intermediate performance (27.24% and 27.63%), demonstrating that linear models are insufficient for capturing demand patterns in bike-sharing systems.

Hierarchical Gradient Boosting was evaluated to leverage cross-cluster patterns. HGB achieves modest improvements over GB on large clusters, while improvements on smaller clusters are marginal ($< 1\%$, Cluster 3). The residual-learning mechanism offers limited benefit when cluster demand is inherently noisy, suggesting that information sharing across clusters is most effective for well-sampled, stable clusters. The modest 2–3% improvement from HGB raises the question of whether its added architectural complexity is justified.

Comparing the two clusters, cluster 3 shows significantly higher percentage error (34.90%) but a smaller absolute error (3.81 bicycles). This is expected since the demand naturally varies more in small clusters. GB still captures 70% of the demand patterns ($R^2 \approx 0.70$).

2.5 Conclusion

Hierarchical Gradient Boosting achieved the best performance. The divergent performance between clusters reflects fundamental differences in demand characteristics. Cluster 16, representing high-traffic areas, exhibits stable, regular usage patterns conducive to precise forecasting. Cluster 3, representing lower-density areas, exhibits sporadic demand with higher variance relative to mean volume. This heterogeneity means predictions for Cluster 3 will have proportionally larger percentage errors, but the absolute bicycle-level errors remain practical for deployment.

Part 3

Predictions from the Hierarchical Gradient Boosting (HGB) model developed in Task 2 were used to determine bike allocation requirements. For each cluster $i \in \mathcal{C}$ and hour $t \in \mathcal{T}$, the hourly net outflow is computed as:

$$\phi_{i,t} = d_{i,t} - a_{i,t}$$

where $d_{i,t}$ represents departures (pickups) and $a_{i,t}$ represents arrivals (dropoffs) in cluster i during hour t. The cumulative net outflow for cluster i over the operational day is:

$$\Phi_{i,t} = \sum_{h=1}^{t} \phi_{i,h}$$

This cumulative measure captures the running balance of bikes, where positive values indicate a surplus and negative values indicate a deficit.

3.2 Minimum Bike Allocation

The required initial bike allocation for cluster i is determined by the maximum cumulative deficit encountered during the 24-hour period:

$$r_i = \max_t \left(-\min_t(\Phi_{i,t}), 0 \right) \tag{1}$$

This ensures non-negative inventory throughout the operational day. If the cumulative flow never reaches a deficit (minimum cumulative flow ≥ 0), no additional bikes are required.

3.3 Results and Discussion

Based on the HGB model predictions for November 4, the bike company should allocate 32 bicycles to Cluster 3 and 82 bicycles to Cluster 16 to ensure zero shortage throughout the 24 hours.

Figure 4 illustrates how cumulative net flow evolves throughout the day. Cluster 3 accumulates surplus in early hours (00:00–10:00), then enters deficit through afternoon and

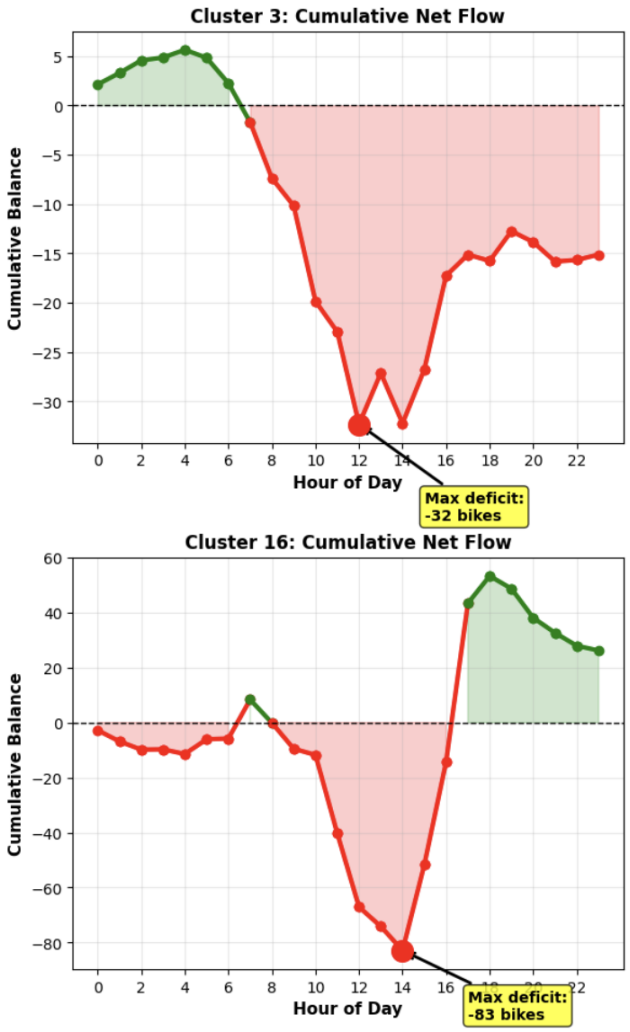

Fig. 4. Cumulative Net Flow and Minimum Bike Allocation for November 4. Green regions indicate surplus, red regions indicate deficit. Cluster 16 requires 83 bikes to prevent shortage (peak deficit at 16:00), while Cluster 3 requires 32 bikes (peak deficit at 14:00). Initial allocations are determined by maximum cumulative deficit.

TABLE II
BIKE ALLOCATION (PREDICTED VS ACTUAL) NOV 4

Cluster	Predicted	Actual	Diff.
3	32	26	-6
16	83	142	+59

evening, reaching maximum deficit of 32 bicycles at hour 14:00. Cluster 16 shows early morning deficit, recovers midday, then experiences sharp evening deficit peaking at 83 bicycles at hour 16:00.

The model demonstrates effective allocation performance. Cluster 16's larger absolute errors are expected given its substantially higher volume; both clusters show proportionally similar prediction accuracy relative to their operational scale. However, the November 4 underprediction on Cluster 16 (59 bicycles predicted vs. 142 actual) illustrates the challenge for larger clusters.

The company could implement a differentiated buffering strategy balances revenue protection against inventory holding costs, where a bigger confidence interval is introduced for larger clusters where larger absolute error is expected. The main limitation is incomplete training data. The model was trained on January to October but evaluated on November and December, months with unseen seasonal patterns. Access to a full year of data would be expected to improve forecast accuracy.

IV. EXPLORATORY COMPONENT

In the exploratory component of this study, we aimed at investigating the where users of Citi Bike in New York City are biking from and to. First we will investigate if there are any aggregate trends in the data about where users are biking to and from, and if we from this can zone the city into meaningful segments based on biking patterns. Secondly, we will attempt to predict the destination of individual bike rides based on a set of features including weather data [4], time of day, holiday status [5], gender indicators, birth year, user type and the starting location of the user.

A. Aggregate Destination Trends

In order to investigate aggregate trends in the data, we calculated the net demand for each station, defined as the number of pickups minus the number of dropoffs. Stations with a positive net demand are stations where more people are picking up bikes than dropping them off, indicating a higher demand for bikes at these locations. Conversely, stations with a negative net demand are locations where more bikes are being dropped off than picked up, indicating a surplus of bikes. Initially, we looked at the total net demand for each station over the entire year of 2018, this showed some patterns, but most stations had a net demand close to zero, indicating a balance between pickups and dropoffs.

As we found in the data analysis and visualization section, commuting rush hours are a significant driver of bike usage in New York City. Therefore, we decided to segment the net demand into two time periods: before noon (00:00-11:59) and after noon (12:00-23:59). This segmentation allowed us to capture the directional flow of bikes during the morning and afternoon/evening periods, which are likely to reflect commuting patterns. By calculating the net demand for each station during these two periods, we were able to identify stations that consistently experienced a surplus or deficit of bikes during specific times of the day. This method revealed clearer patterns in bike usage, as these time-segmented net demands generally had larger magnitudes compared to the total net demand. For both time segments, we classified stations based on the sign of their net demand, categorizing them into three groups: positive net demand, negative net demand, and zero net demand. Interestingly, we found that almost 90% of stations had the opposite sign of net demand in the two time segments, indicating two distinct clusters of

stations: those that are net providers of bikes in the morning and net receivers in the afternoon/evening, and vice versa. This finding aligns with typical commuting patterns, where individuals travel from residential areas to work locations in the morning and return in the evening. The stations outside these two main clusters, can be considered outliers.

Now that we have identified these two distinct groups of stations based on their net demand patterns, we will create a classification model to predict the group membership of each station based on its geographical location. After plotting the stations on a map, it is evident that the two groups are not linearly separable, this means that we must use a non-linear classification model. We considered the following models for this task: decision trees, random forests, bagging, gradient boosting, neural networks, and a support vector machines with a radial basis function kernel. We decided on a 80/20 train-test split of the stations, and to tune hyperparameter we used grid search with a 3-fold cross-validation on the training set. To compare the models we used accuracy as the performance metric, since we wanted to predict the class labels, both classses are equally important, and the classes are pretty balanced with a 60/40 distribution. Before training the models we converted the latitude and longitude coordinates to kilometers with zero mean to ensure that the models would not be biased by the different scales of the two features.

After training our models we around 80% accuracy on all of our models. The best model we found was bagging with 300 estimators and when testing on our test set, we get an accuracy of 84% and thereby an accuracy of 97% on the full dataset. When visualizing the classification border of this model, as seen in figure 5, it is clear that the model has successfully identified a boundary that separates the two groups of stations based on their geographical coordinates, and that the two groups are clustered in different areas of the city. This zoning of the city can be compared to existing neighborhood definitions, and when compared to New Yorks City's Zoning & Land Use map [6], we find that our zoning based on biking patterns aligns roughly with the zoning of residential and commercial areas. Stations in residential areas tend to be net providers of bikes in the morning and net receivers in the afternoon/evening, while stations in commercial and manufacturing areas tend to be net receivers in the morning and net providers in the afternoon/evening. For Citi Bike, this zoning information is highly relevant, as it makes it evident that most stations will naturally balance out trough the day. The stations outside these two categories should be given special attention, as these do not balance out naturally through the day, and therefore require a more active but simple redistribution strategy.

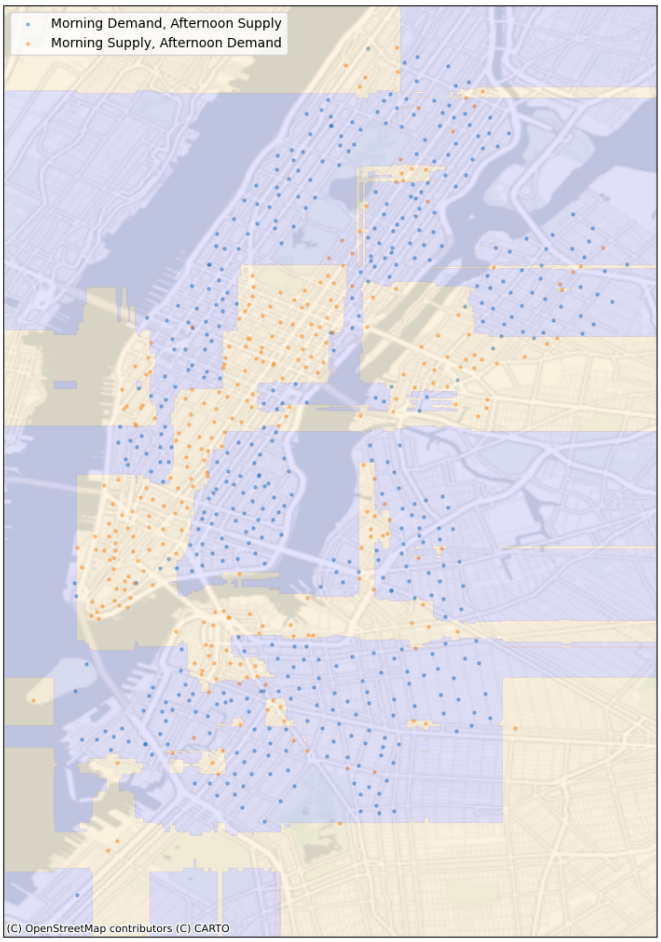

Fig. 5. New York City zoning based on bike demand

B. Individual Ride Destination Prediction

We then extended the existing dataset with publicly available weather data for New York City, segmented on an hourly basis, as well as a record of all public holidays in New York City during 2018. This external data required preprocessing to ensure compatibility with the pre-existing dataset, necessitating certain assumptions. Specifically, we merged the hourly weather data with the `trips_2018` dataset by aligning the hour and date columns, under the assumption that hourly weather granularity is sufficient for our predictive modeling purposes. While finer temporal resolution might capture more nuanced effects, such as sudden rain showers, we consider hourly data a reasonable compromise between data availability and model complexity. Public holiday information was incorporated as a binary indicator variable, taking the value of 1 if the date corresponded to a public holiday and 0 otherwise. It is acknowledged that this approach does not capture partial holidays or regional events, but provides a straightforward mechanism to encode temporal variations in demand.

With this extended dataset, we aimed to develop a neural network capable of predicting the destination of a user based on a combination of features including weather conditions, time of day, holiday status, gender indicators, birth year, user type, and the starting location of the user. Each of these features was selected based on its potential influence on user behavior. For instance, user type and demographics may reflect different commuting patterns, while start location is a key

determinant of likely destinations. Weather conditions can influence both the choice to bike and the likely route taken. Such a model is highly relevant from a business perspective, as bike-sharing companies can utilize these predictions to anticipate station-level demand. Accurate predictions allow the company to proactively redistribute bikes, thereby minimizing the risk of stockouts and ensuring consistent service availability.

The neural network architecture was designed with four layers, incorporating key regularization and normalization techniques to enhance model performance. Dropout layers were implemented to reduce the risk of overfitting by randomly omitting a subset of neurons during training, while batch normalization layers were used to stabilize and accelerate learning by re-centering and re-scaling activations. ReLU activation functions were applied to introduce non-linearity and avoid vanishing gradient issues. Input features were standardized by subtracting the mean and scaling to unit variance, computed as follows:

$$z = \frac{x - \mu}{\sigma}$$

where x denotes a feature value, μ represents the mean of the training samples, and σ is the corresponding standard deviation. Standardization ensures that features contribute equally during optimization, preventing variables with larger scales from dominating the gradient.

The network was trained using the Mean Squared Error (MSE) as the loss function, chosen for its suitability in regression tasks, while the ADAM optimizer was employed for its adaptive learning rate and robust convergence properties. To evaluate model performance in terms of spatial prediction accuracy, we converted the predicted and true geographical coordinates into a distance metric using the Haversine formula, which calculates the great-circle distance between two points on the Earth's surface:

$$\phi_1' = \frac{\pi}{180}\phi_1, \quad \lambda_1' = \frac{\pi}{180}\lambda_1, \quad \phi_2' = \frac{\pi}{180}\phi_2, \quad \lambda_2' = \frac{\pi}{180}\lambda_2,$$

$$\Delta\phi = \phi_2' - \phi_1', \quad \Delta\lambda = \lambda_2' - \lambda_1',$$

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1')\cos(\phi_2')\sin^2\left(\frac{\Delta\lambda}{2}\right),$$

$$c = 2\arctan\left(\frac{\sqrt{a}}{\sqrt{1-a}}\right),$$

$$d = R \cdot c, \quad R = 6{,}371{,}000 \text{ meters}.$$

The dataset was split into training, validation, and test sets, comprising 80%, 10%, and 10% of the total data, respectively. After training, the model achieved a mean Haversine distance of 2,347.40 meters and a median Haversine distance of 1,977.94 meters on the test set. These results indicate that the model can reasonably approximate user destinations, providing actionable insights for operational decision-making in bike-sharing systems. While the performance demonstrates promising predictive capability, limitations include the coarse temporal resolution of weather data, the simplified representation of holidays, and potential unobserved factors such as traffic conditions or transient events. Future work may focus on incorporating real-time weather updates, more granular temporal features, and additional spatial covariates to further improve destination prediction accuracy.

V. Conclusion

Accurate demand forecasting is essential for bike-sharing operations. Without reliable predictions, operators face a fundamental trade-off: insufficient allocation leads to stock-outs and lost revenue, while excessive allocation wastes resources. This paper demonstrates how machine learning can help address this challenge. Hierarchical Gradient Boosting achieves $R^2 \approx 0.95$ for large clusters and maintains low absolute errors across smaller clusters, providing reliable predictions that support better allocation decisions. The methodology is practical and immediately deployable. Access to a full year of data could further improve forecasting accuracy. By grounding fleet repositioning in data rather than guesswork, operators can optimize resource allocation and improve customer service.

The aggregate destination trends analysis revealed two distinct zones in New York City based on bike demand patterns, which align closely with residential and commercial areas. Residential zones tend to be net providers of bikes in the morning and net receivers in the afternoon/evening, while commercial zones exhibit the opposite pattern. These stations will naturally balance out through the day, simplifying redistribution efforts. Outlier stations that do not fit this pattern require special attention, as they do not balance out naturally through the day and therefore need a more active redistribution strategy.

For the individual ride destination prediction we achieved a model where we computed the mean of all of the predictions on our test-set where we achieved a mean of 2,347.40 meters. based on our set of features including weather data, time of day, holiday status, gender indicators, birth year, user type and the starting location of the user.

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[2] Pedro Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[3] Jerome H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[4] Copernicus Climate Change Service, "Nyc weather – 2016 to 2022," 2022.

[5] Office Holidays Ltd., "New York 2018 — Holidays and Observances," https://www.officeholidays.com/countries/usa/new-york/2018, 2018.

[6] New York City Department of City Planning, "Zola: Nyc's zoning & land use map," 2025, Accessed: 2025-12-01.