Transportation Optimization

Shaikh Ibrahim Noman
ibrahim.noman788@gmail.com
University of Texas at Austin
Austin, Texas, USA

Viren Halaharivi
vhalaharivi@gmail.com
University of Texas at Austin
Austin, Texas, USA

Akash Rajeev
akash.rajeev@utexas.edu
University of Texas at Austin
Austin, Texas, USA

Robbie Howell
howellrobbie01@gmail.com
University of Texas at Austin
Austin, Texas, USA

William Guo
wdg485@utexas.edu
University of Texas at Austin
Austin, Texas, USA

Abstract

Analyzing micro mobility data is crucial to address the growing demand for e-vehicles (scooters, bikes, mopeds). Using historical data from the City of Austin Open Data Portal [1], this study aims to predict total trip count distributed by vehicle type and hour for council districts in Austin, TX. The trip count feature was created by aggregating: Date, Day of Week, Hour, Council District and Vehicle Type. Log transformations were applied to trip distance and trip count to address their highly-skewed nature. Furthermore, sample weights were used for vehicle types to account for class imbalances created by scooters. Each having their own benefits and constraints, Linear Regression, Random Forest and XGBoost were the three models used for this regression prediction problem. This study aims to evaluate the performance of these models through a time series split, cross validation, hyper parameter tuning, evaluation metrics such as Root Mean Squared Error (RMSE) and R-Squared, and visualizations such as feature importance, distributed bar plots and residual plots. The authors also note that more complex models and higher computational power is needed to address limitations such as heteroscedasticity and the data range. The findings highlight the the variation of trip count across different timestamps, leading to a more accurate understanding of vehicle allocation, and meeting user demand.

CCS Concepts

• **Computing methodologies** → **Machine learning algorithms**; • **Applied computing** → *Transportation*; • **Information systems** → *Data analytics*.

Keywords

Transportation, E-Vehicles, Demand Forecasting, Transportation Analytics, Urban Mobility, Machine Learning, Data Science

ACM Reference Format:
Shaikh Ibrahim Noman, Viren Halaharivi, Akash Rajeev, Robbie Howell, and William Guo. 2024. Transportation Optimization. In *Proceedings of Class Project for SDS 336: Practical Machine Learning (Class Project '24).* ACM, New York, NY, USA, 7 pages. https://doi.org/N/A

Class Project '24, Fall 2024, University of Texas, Austin
2024. ACM ISBN N/A
https://doi.org/N/A

1 Introduction

Over the past couple of years, micro mobility vehicles such as scooters, bicycles and mopeds have revolutionized urban transportation by offering flexible and sustainable travel options. As demand for them continues to grow, understanding and predicting trip demand has become important for effective vehicle allocation and infrastructure planning, particularly in populated cities such as Austin, TX. Accurately modeling trip counts is essential for optimizing vehicle allocation, improving user satisfaction, and reducing operational inefficiencies.

There were various studies used in literature comparison. Although, one study has explored methods to analyze and predict micro mobility usage patterns in Austin TX, [2], it aims to forecast demand for dock less scooters more broadly. It does this by focuses on spatiotemporal modeling such as considering environmental factors, and not considering factors such as council districts. One important thing to note is that the previous study did notice an increase in accuracy when accounting for weather patterns. However, the STMGT model (the most accurate deep learning model), was difficult to interpret. This is the real trade off in this study (high accuracy or better interpretability). Our study aims to see more specific patterns by understanding trip count variation over time and across districts to improve vehicle allocation and meet user demand. For this study, more focus is on interpretability because its goal is to assist policymakers and mobility operators. Furthermore, this helps us narrow down on specific feature importance, instead of the black box nature of the STMGT model.

The analysis begins with data preprocessing steps such as dropping NA values and removing outliers through the inter-quartile range (IQR) method. Next, is the creation of the trip count feature using date, day of week, hour, council district, and vehicle type. To mitigate data skewness, log transformations are also applied, and scooter class imbalances were handled by sample weights. Lastly, three regression models (Linear Regression, Random Forest and XGBoost) were implemented to predict trip count. These models were evaluated using time series splits, cross-validation, hyper parameter tuning, visualization plots (feature importance, bar plots) and evaluation metrics such as root-mean squared error (RMSE) and R-squared.

The findings of this study aim to improve understanding of trip demand across different vehicle types and timestamps. Aligning vehicle allocation with predicted trip patterns allows micro mobility services to better meet user demands while reducing inefficiencies. the remainder of this report includes the following sections: Data, Methodology, Results, Discussion, Conclusion, Acknowledgment and References.

2 Data

This section goes over the data we used for our project and the steps we took to make sure it was fit for predictive modeling. The dataset we chose to analyze for this project is the Shared Micro mobility Vehicle Trips dataset, which comes from the publicly available City of Austin Open Data Portal. This dataset consists of 15,048,635 rows, each of which represents a trip taken using a public micro mobility vehicle, such as a scooter or bicycle, from April 3, 2018 to April 4, 2022. The data were reported by vehicle providers like Lime and Bird as part of the requirements for operating shared mobility services.

The dataset contains 18 variable columns, each containing information relevant to a micro mobility trip, such as trip duration and distance. Some of the columns, like trip ID and device ID, are not useful for this project as we are looking at overall trends of vehicle usage. The variables that are of the most interest to us are the ones that provide location or time data, such as the council district start and end, as well as the month, day of the week, hour, and vehicle type used on the trip. These variables are the most helpful for predicting when and where trip demand is the highest, as we can aggregate these to see how many trips occur at each location or time.

We discovered early on in our project that using all fifteen million rows was simply infeasible, as we did not have the computing power necessary to run our models on a dataset of that size. In order to fix this issue, we decided to drop all the observations from before 2020, which left us with roughly six million rows. We chose this cutoff since it still left us with a significant amount of data, and allowed us to analyze the most recent trends in the data.

As we continued our early data exploration, we discovered that many variables were extremely skewed by outliers. We saw trip durations of over a million meters, and there were some trips that had a start date in 1970, far before these vehicles were even available. These observations were clearly not representative of actual trips taken, so we removed them by dropping all data outside of the inter-quartile range for each variable. With these outliers removed, we saw immediate improvements with smoother visualizations and data distributions.

Even after removing outliers, we had issues with scooters coming up as the only predicted result from our early models since there is a significantly higher number of scooters than bikes or mopeds in our dataset. In order to ensure that vehicles were adequately represented, we added sample weights to each vehicle type. Rerunning our models with these weights resulted in much more accurate predictions and better representation of each vehicle.

The most important step we took in setting up our data was creating the trip count column, which represented the number of trips taken from a certain location at a certain time over the time period that our data was collected. We achieved this by aggregating all trips that shared a vehicle type, day of the week, hour, and starting council district. However, even after creating this column, we saw that it was massively skewed and log transformed it to make it more interpretable. With trip count created and our data cleaned up, preprocessing was complete and we started testing different models to find the best predictor for our data.

3 Methodology

We tested three different types of models to try and determine which would most accurately predict our data: linear regression, random forest model, and XGBoost. The following evaluation strategy was used:

- Time-Series Split: Since the data isn't i.i.d, and the goal is to predict future values, the data was trained on 2020-2021 data, and tested on the 2022 data.
- Cross Validation: Applied to handle over fitting, and gives a better understanding of how the model will perform on unseen data. Demand patterns might vary by time and location, and cross-validation ensures there isn't a bias toward a specific subset of data
- Hyper parameter tuning (Random Forest and XGBoost): By accounting for complex relationships, optimal hyper parameter models can lead to more accurate predictions and reduced variances. Specified sample weights to fix vehicle class imbalances
- Evaluation Metrics (RMSE and R-squared): RMSE: gives us an understanding of how far off our predictions are from actual values on average, and R-squared was used to determine the proportion of variance explained by each model
- Visualizations: Feature importance plots provided insights to what features contributed the highest to model predictions. Bar plots showed busy times for trip demand, and also helped see demand per vehicle.

3.1 Linear Regression

The first model fitted was a linear regression model. Due to this model's relative simplicity (easier interpretation), we decided to keep this model as our baseline. Since we also have large amounts of data (5 million rows), the linear regression would be less resource intensive. In addition, there was no hyper parameter tuning applied. The assumptions of this model are: linear relationship between predictor and features, homoscedasticity and Gaussian normal distribution of the data.

Unfortunately, the linear regression model didn't perform well. Due to its failure in capturing non-linear relationships, the performance metrics were relatively low. In addition, the homoscedasticity assumption wasn't addressed even after applying log transformations. Even though the result wasn't what was expected, the linear regression model helped in discovering that there are nonlinear relationships in the data, and each observation is highly correlated.

3.2 Random Forest

To address the issues in the linear regression model, a random forest regression model was fitted. Due to its ensemble nature, not only does this model account for non-linear relationships, but each tree contributes to the model independently to the final model (reduces variance). For the evaluation strategy, a 5-fold cross-validation was applied after using Randomized Search CV. The following are the hyper parameters used: n estimators: 50, min samples split: 5, min samples leaf 4, max features: None, max depth: None.

The model performed relatively well, but there a couple of issues. The model was over fitting the data by capturing random noise. This was evident due to the high RMSE. In addition, the computational time was still relatively high even after Randomized Search CV. Overall, the results were close to what expected, but there were issues that needed to be resolved.

3.3 XGBoost

Finally, a XGBoost model was trained. This type of model tends to be better than random forest models due to in-built features to avoid over fitting to the data and iteratively correcting errors in successive decision trees as opposed to aggregating the results of different decision trees like in the random forest model. In addition, XGBoost has the advantage of being more computationally efficient than random forest models, though it comes at the cost of interpretability. Using Grid Search CV, a 5-fold cross validation was applied. The optimal hyper parameters are: colsample by tree: 1, eta: 0.3, max depth: 7, subsample: 1

The results went as expected. The model provided a good trade off between bias and variance due to its lower R-squared (compared to random forest), and a lower RMSE. The model also did relatively well in capturing information unique to each council district. The XGBoost model was determined as the best performing model (in relation to other models), and was used to make final predictions.

4 Results

We used RMSE (Root Mean Squared Error) and R-squared values to measure the performance of our model. RMSE measures the average magnitude of prediction errors in the trip count, with a lower RMSE indicating better model accuracy. In the context of this study, the RMSE provides insight into how well the model captures temporal and spatial variations in trip counts across districts. Policymakers and mobility operators can use this information to identify discrepancies between predicted and actual trip counts, which is critical for improving vehicle allocation. The values for each model are listed below:

Model	R-squared	RMSE
XGBoost	0.97	0.31
Random Forest	0.93	0.98
Linear Regression	0.74	0.60

Table 1: Performance metrics of different models

Since XGBoost performed the best, it was used to determine final results. The RMSE value was .31 indicating that the average prediction error for trip counts is relatively low. This suggests the model's predictions are highly accurate and closely aligned with the actual observed values. Our R-squared value was .97 indicating that 97% of the variance in trip count is explained by the model. This value shows the model's capability to capture temporal and spatial patterns.

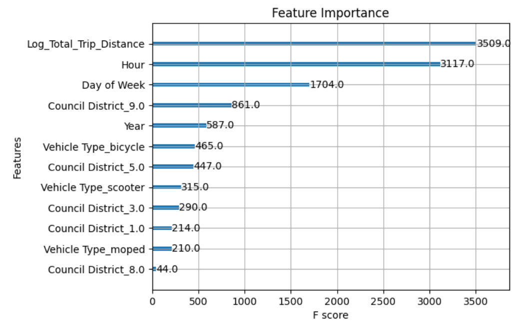

Figure 1: Plotting Features

The feature importance plot shows the significance of predictors in the XGBoost model for explaining trip count variations. The importance is measured in F-score, which represents how frequently a feature is used in splitting data across the decision trees in the model. A higher F-score indicates greater importance. The Log_Total_Trip_Distance had the highest f-score of 3509 emphasizing the strong relationship between the total trip distance (log-transformed) and trip counts. It suggests that trip distance is a primary determinant of variations in demand.

4.1 Visualizations

The following visualizations aim to provide insights into the predicted trip counts for micro-mobility vehicles in Austin, TX, with a focus on temporal and spatial demand patterns. These visualizations were made using the predicted trip counts from our model. Note: To address the highly skewed nature of the data, log transformations were applied to trip counts, allowing for clearer insights into the variations in demand. By analyzing the distribution of trip counts across hours of the day, days of the week, and council districts, these figures illustrate the key features influencing demand. As stated it's important to note that these visualizations were compiled using log transformed data. We see the same trend of data when analyzing hourly demand as seen in Figure 1, Figure 2, and Figure 8 (The dataset had a disproportionate amount of scooter data compared to bikes and mopeds).

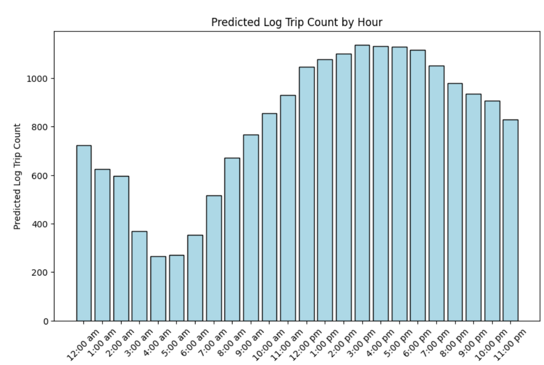

Figure 2: Hourly Demand

There were very low predicted trip counts from 3:00 - 6:00 AM. Most of the usage occurred from mid-day to the late evening.

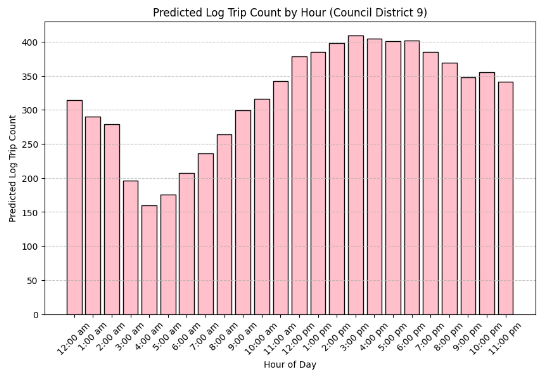

Figure 3: Busiest Council District

The Log Trip Count by Hour Council District 9 shows peak demand between 3:00 pm and 7:00 pm and minimal demand from 3:00 am to 6:00 am. Demand begins increasing around 6:00 am, aligning with the morning commute. These trends highlight key demand patterns for efficient vehicle allocation

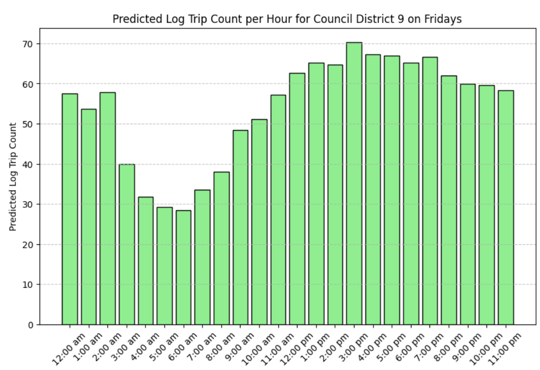

Figure 4: Busiest Council District on Busiest Day

On Fridays we have the most predicted trip counts, and you can see it on the figure above. You see a huge increase in demand in the early morning period compared to the overall data.

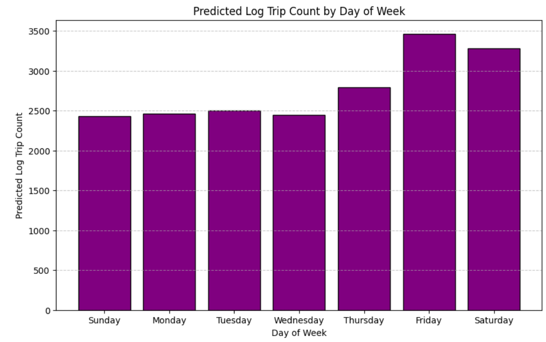

Figure 5: Weekly Demand

Here is predicted trip count by day. As you can see Thursday, Friday, and Saturday are our days with the most trips.

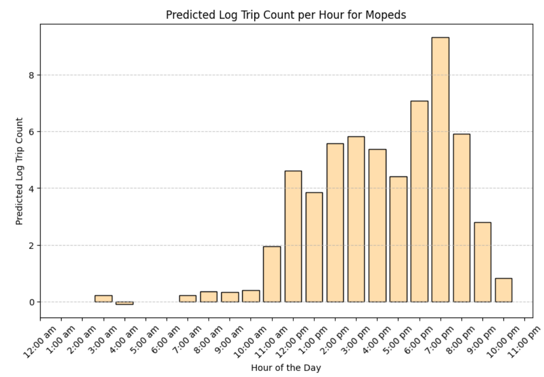

Figure 6: Moped Demand

Here is predicted moped trip count over a span of a day. Note: Mopeds were discontinued after 2022. Comparatively very few observations of mopeds in our dataset compared to bikes and scooters

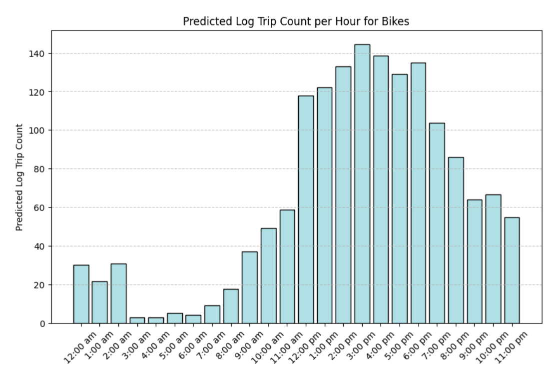

Figure 7: Bike Demand

Here is predicted bike trip count over a span of a day. Trip count is low from 12:00 AM to 7:00 AM and after 7:00 PM. Trip count ramps up significantly from 8:00 AM to 7:00 PM. Note that bike usage is minimal, as indicated by the low y-axis values of predicted log counts.

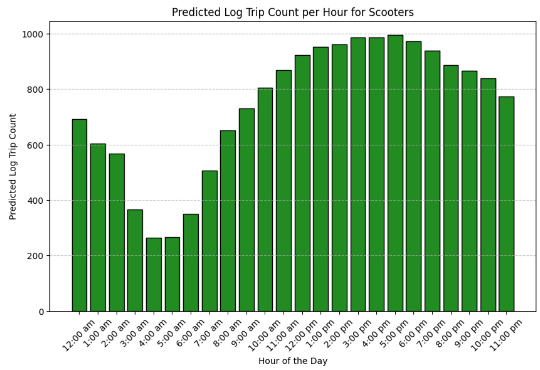

Figure 8: Scooter Demand

Here is predicted trip count over a span of a day. Trip counts are low between 3:00 AM and 5:00 AM and after 9:00 PM. Moderate trip counts occur during morning commutes (6:00 AM–9:00 AM) and evening commutes (4:00 PM–8:00 PM). Higher trip counts are observed during midday hours.

5 Discussion

5.1 Introduction

In interpreting our analysis, it is important to revisit some key guiding questions: "What is the most accurate model for predicting trip demand across different areas and times?" and "How should platform providers allocate resources given fluctuating demand?" These questions provide a framework for understanding our results and ensuring that our findings are meaningful and actionable.

As outlined in the results section, the primary model employed in this analysis was XGBoost, a gradient-boosting mechanism. This model was chosen due to its ability to achieve the lowest Root Mean Square Error (RMSE) and the highest R-squared value, indicating strong predictive performance. However, it is essential to note that, while this model performed well within the scope of our analysis, it may not represent the most optimal solution for this prediction problem. In a more comprehensive study, researchers might consider experimenting with a variety of other regression models to refine and improve the results.

5.2 Interpretation

The model's performance is reflected in an RMSE of 0.31 and an R-squared value of 0.97. The high R-squared value suggests that the model explains 97% of the variance in the data, which is promising. However, the RMSE of 0.31 indicates that there is still room for improvement in the model's accuracy. Given the dynamic and fluctuating nature of trip demand in urban environments, capturing finer details in demand is crucial for effective resource allocation and improving platform performance.

Our model also identified which features contributed most significantly to predicting trip demand. As shown in Figure 1, the feature with the highest importance was Log_Total_Trip_Distance, followed by Hour, Day of Week, Council_District_9.0, and Year. This indicates that the distance and temporal factors, such as the hour of the day and day of the week, are the most important variables in predicting trip demand. Interestingly, platform type (e.g., scooter, bike, or moped) received relatively lower importance, suggesting that factors like time and distance may play a larger role than the specific type of vehicle when determining trip demand.

In Figure 3, we present the predicted log trip count by hour in Austin's Council District 9, which had the highest number of observations in our dataset. This district encompasses central Austin, including the University of Texas campus, downtown areas, and popular neighborhoods like South Congress. These areas are characterized by high population density and significant activity, with numerous places of interest such as government offices, educational institutions, and parks. This is likely why trip demand in District 9 is notably high. It is worth noting that the University of Texas and surrounding areas likely contribute a substantial portion of the demand, and further research could focus specifically on this region to explore more targeted resource allocation strategies. Given the high demand in District 9, we recommend that platform providers prioritize resource allocation here, particularly in dense areas, to maximize platform usage.

Temporal factors, such as the day, hour, and month, have also proven essential in predicting trip demand. Figures 2 and 4 depict the distribution of trip counts by hour on Fridays, showing a consistent increase in demand beginning around 7:00 AM, peaking between 12:00 PM and 7:00 PM, and then tapering off. Figure 5 illustrates trip demand across the weeks, revealing that demand peaks on Fridays and Saturdays, likely due to the increased mobility of users during the weekends. Based on these patterns, we recommend that platform providers allocate resources in areas that experience high late-afternoon and evening traffic, such as parks, shopping centers, and business districts, to ensure high platform utilization. Furthermore, providers should prepare for higher weekend demand by ensuring adequate resource availability and maintenance.

We also observed notable differences in demand patterns across different platform types. Figure 8 shows that scooter demand peaks during midday, with a secondary demand spike in the evening. This suggests that scooters are more frequently used during lunch breaks, errands, or recreational trips. Therefore, platform providers should consider focusing scooter availability near business parks, eateries, and areas that experience lunchtime foot traffic.

Bike demand, as depicted in Figure 7, follows a similar pattern but with a significant increase in demand from 8:00 AM to 7:00 PM, potentially reflecting a commuter-focused usage trend. To meet this demand, we recommend that providers allocate bikes in high-density residential areas near transportation hubs such as bus stations and train stations along CapMetro's Red Line. However, given that the dataset contained fewer bike observations, further analysis would be beneficial to confirm these trends before implementing policy changes.

Finally, mopeds, as shown in Figure 6, exhibit demand patterns similar to scooters and bikes, though conclusions are less clear due to their discontinuation in 2022 and limited dataset representation. Given these limitations, further research and more data would be necessary to develop more reliable insights regarding moped demand.

5.3 Literature Comparison

To contextualize our findings within the broader literature, we reviewed studies and reports analyzing micro mobility trends in other regions. A joint report by the League of American Bicyclists and Lime[3] examined micro mobility data in Washington, D.C., and Bloomington, Indiana, and revealed that micro mobility usage is significantly higher in areas with dedicated bike infrastructure, such as bike lanes. Although our dataset did not explicitly include information on whether trips occurred in areas with bike lanes, we observed that council districts with the highest micro mobility demand, such as District 9, are areas known to have such infrastructure. This aligns with the report's conclusions and suggests a potential link between infrastructure and demand in our analysis as well.

The National League of Cities' micro mobility Report[4] also emphasizes the role of bike infrastructure in managing micro mobility demand, while highlighting the importance of equity in resource allocation. The report stresses that micro mobility programs should extend to under-served areas to ensure equitable access to transportation options. While our analysis did not include socioeconomic factors, the importance of equitable resource distribution is evident, particularly in cities like Austin, where diverse communities have varying levels of access to transportation services.

Additionally, a Vox article[5] on Lime's resource allocation practices describes the use of rebalancing—a process where employees strategically redistribute scooters based on anticipated demand, maintain vehicles in need of repair, and manage resources to optimize platform availability. This highlights the operational challenges and human effort involved in effective resource allocation. Notably, these considerations were absent from our analysis, which did not account for the logistical complexities of employee-driven redistribution or environmental challenges like Austin's hot and variable climate. Incorporating these aspects into future studies could provide a more holistic understanding of resource allocation strategies.

5.4 Next Steps

To improve upon our analysis and offer actionable recommendations for platform providers, several avenues for future work are suggested. First, incorporating data on the presence and quality of bike infrastructure, such as bike lanes and parking areas, would allow for a more direct assessment of its impact on micro mobility demand. This could help refine resource allocation strategies, particularly in districts where infrastructure investments could have the greatest effect on usage patterns.

Second, integrating socioeconomic and demographic data could provide valuable insights into equity in micro mobility access. Analyzing demand in under served areas and identifying potential barriers to access would enable platform providers to ensure fair distribution of resources, aligning with the principles emphasized in the micro mobility Report.

Third, operational data on resource rebalancing—such as employee time, travel distance, and maintenance logs—should be incorporated to evaluate the logistical challenges of resource allocation. This would help providers design strategies that minimize employee effort while maximizing service efficiency. For instance, algorithms could be developed to optimize redistribution routes, taking into account both demand and ease of access for employees.

Finally, future studies should examine the influence of environmental factors, such as weather conditions and seasonality, on micro mobility demand. Understanding how these factors affect usage patterns could help providers anticipate fluctuations and adapt their resource allocation plans accordingly.

By expanding the scope of analysis to include infrastructure, equity, operations, and environmental considerations, platform providers can develop more effective, human-centered strategies for resource allocation. These enhancements will not only improve service efficiency but also contribute to a more sustainable and inclusive micro mobility ecosystem in cities like Austin.

5.5 Limitations

We acknowledge several key limitations that impact the relevance and applicability of our micro mobility trip demand analysis.

First, our data preprocessing significantly constrained the dataset's comprehensiveness. Initially encompassing approximately 15 million rows, we were compelled to limit our analysis to observations post-2020 due to computational constraints in Google Colab. This methodological decision substantially reduced our sample size and potentially compromised the external validity of our findings. By excluding data from 2018-2020, we may have inadvertently omitted critical temporal patterns and trends that could provide more nuanced insights into trip dynamics. Future research should prioritize leveraging computational resources—such as cloud computing platforms or high-performance computing environments—to analyze the complete dataset comprehensively.

The dataset's composition presented another notable limitation in terms of vehicle type representation. Significant disparities existed across micro mobility modes, with scooter observations substantially outnumbering bike and moped records. The discontinuation of mopeds after 2022 further exacerbated this imbalance, potentially introducing sampling bias into our analysis. This uneven distribution suggests the necessity of future studies to either implement sophisticated resampling techniques or collect more balanced observational data across different micro mobility vehicle types to ensure representative insights.

Methodological challenges also emerged in our statistical approach. Our dataset exhibited characteristics of heteroscedasticity and high variance in error terms, which, while less problematic for ensemble methods like Random Forest and XGBoost, potentially compromised the reliability of our Linear Regression model's interpretations. Researchers should consider robust regression techniques or advanced statistical transformations to mitigate these variability concerns in future analyses.

Spatial granularity represented another significant constraint in our research design. While utilizing Austin's Council Districts provided a resource-efficient method of incorporating spatial dimensions, this approach inherently limits analytical precision. With

Austin comprising ten single-member council districts that span diverse geographical and zoning contexts, our spatial analysis lacks the nuanced resolution that more granular data sources could provide. Future research could enhance spatial understanding by leveraging more detailed geographic demarcations such as 2020 Census Tracts or specific neighborhood boundaries.

Additionally, our analysis acknowledged the multifaceted nature of trip demand by recognizing numerous unaccounted variables. Environmental factors like weather conditions, precipitation, heat index, and humidity can profoundly influence micro mobility usage. Similarly, social dynamics—including local events, concerts, neighborhood infrastructure, and area density—represent critical contextual elements that potentially modulate trip patterns. While our current dataset did not incorporate these variables, we recommend future research develop comprehensive models that integrate these multidimensional factors to generate more holistic and predictive trip demand analyses.

Our research encountered significant constraints across methodological, demographic, and contextual dimensions. The analysis was limited to three predictive models—Linear Regression, Random Forest, and XGBoost—which, while providing valuable insights, represent a narrow approach to capturing the complex, non-linear relationships in urban mobility patterns. These models exhibit varying degrees of interpretability, from clear linear relationships to the more opaque "black box" nature of ensemble methods. Additionally, our dataset likely represents a subset of users with smartphone access, potentially introducing systematic biases by not comprehensively capturing diverse demographic factors such as age, income, and individual transportation preferences that influence micro mobility usage. The dynamic regulatory ecosystem further complicates analysis, as local policies regarding micro mobility operations can dramatically shift trip patterns. Future research should develop more sophisticated modeling approaches that balance predictive accuracy, interpretability, and comprehensive demographic representation.

By transparently discussing these limitations, we aim to provide a clear framework for understanding the current study's scope and identifying promising avenues for future research in micro mobility demand prediction.

6 Conclusion

Our research attempted to address the challenge of predicting micro mobility trip demand in Austin by developing predictive models across the city's council districts. Utilizing the techniques—Linear Regression, Random Forest, and XGBoost—we analyzed a the dataset to try to understand trip patterns and resource allocation strategies. XGBoost emerged as the most accurate model of the three tested, revealing significant insights into urban mobility dynamics. Key findings demonstrated peak demand periods between 12:00 PM and 7:00 PM, with bicycles showing an additional morning commuter peak at 8:00 AM, and District 9 consistently reporting the highest trip volumes. Our analysis identified strategic resource allocation opportunities, including concentrating vehicle deployment near high-traffic destinations such as restaurants in the downtown area, workplaces in the central business district, parks like Zilker Park, and transit hubs including the Austin Convention Center and major train stations, like along CapMetro's Red Line. Particularly for District 9, which encompasses areas like Campus, West Campus, parts of South Congress and parts of downtown, providers should prioritize maintaining a robust fleet during peak hours, and during weekend periods for all districts

7 Acknowledgment

7.1 Contribution Table

Name	Contribution
Ibrahim Noman	10
Viren Halahrivi	10
Robbie Howell	10
Akash Rajeev	10
William Guo	x

7.2 Use of LLM

The main LLM used in this project was ChatGPT. Below are the sections in which ChatGPT was used and the reason.

- Data Preprocessing: Used to generate code to aggregate data correctly and research methods that can reduce data skewness (Box-Cox, Log Transform).
- Modeling: Used to research what models can handle the data used in the study (large rows, high skewness), generate code for cross-validation, hyper parameter tuning and assigning sampling weights to fix vehicle class imbalance.
- Evaluation: Used to generate code for plots and explain the reason for the heteroscedasticity issue.
- Research Paper: Used to fix syntax errors, research/find other similar studies (simplify the language of other studies, so it is easier to understand), revise writing, aid in generating CCS concepts, and generate citations for the References section.

References

[1] City of Austin. *Shared Micromobility Vehicle Trips (2018-2022)*, 2022, https://data.austintexas.gov/Transportation-and-Mobility/Shared-Micromobility-Vehicle-Trips-2018-2022-/7d8e-dm7r/about_data.

[2] Xu, Yiming, et al. "Real-Time Forecasting of Dockless Scooter-Sharing Demand: A Spatio-Temporal Multi-Graph Transformer Approach." *arXiv*, 2021, https://arxiv.org/pdf/2111.01355.

[3] The League of American Bicyclists. "The League and Lime Unveil New Report on How Cities Can Use Shared Micromobility Data for Transportation Planning and Policy." *League of American Bicyclists*, 2023, https://bikeleague.org/the-league-and-lime-unveil-new-report-on-how-cities-can-use-shared-micromobility-data-for-transportation-planning-and-policy/.

[4] National League of Cities. *Micromobility in Cities: A History and Policy Overview*. Apr. 2019, https://www.nlc.org/wp-content/uploads/2019/04/CSAR_MicromobilityReport_FINAL.pdf.

[5] Griswold, Alison. "Lime, the Bike-Share Unicorn, Has Pivoted to Scooters." *Vox*, 20 May 2018, https://www.vox.com/2018/5/20/17328110/lime-limebike-dockless-scooters-san-francisco-charge-drop-off.

Received 11 December 2024