Trip Duration Prediction in Bicycle Sharing Systems

Meghana Meghana
mmeghana@vt.edu
PID: mmeghana

Elham Mohammadrezaei
elliemh@vt.edu
PID: elliemh

Uditi Goyal
ud2607@vt.edu
PID: ud2607

Riya Dani
riyadn99@vt.edu
PID: riyadn99

ABSTRACT

In this work, we attempt to analyze bike sharing systems in one of the major cities of the United States - New York City. A specific problem that we are trying to analyze and model is that of trip duration prediction. To this end, we employ semi fine-grained features representing station information by transforming fine grained station-level features into more coarse grained buckets. Stations are assigned to their respective buckets based on a two-fold categorization involving geographical clustering and classification based on each station's purpose that we call Purpose-aware Categorization (PaC). We posit that this would not only simplify the problem, but also give more insights to the learning process. We compare the performance of models employing the above mentioned features with the more often employed fine-grained station level features and analyze the results. We also try to understand the importance of having destination station information in trying to predict the trip duration. We find that employing our semi-fine grained features indeed gives better results than using station-level features. Further, we also come to understand that having destination information would significantly improve the results.

KEYWORDS

trip prediction, bike sharing, NYC citibike, point of interest, data analysis, machine learning, clustering

1 INTRODUCTION

Bike sharing systems have established themselves as a well renowned means of transportation around various parts of the world, especially in big cities. It helps tourists easily travel around a city and explore new places and city locals commute to their place of work or study. In fact, it has transformed public transportation in big cities and allowed people to travel easily when visiting new places. New York is one of the most populated and busiest cities in the world, which accordingly increases the demand on such convenient means of transport. Hence, understanding the demand of such systems at various points in the city is of the essence. To this end, in this project, we analyze the bike sharing system in New York, and more so try to model the duration of a bicycle trip using an array of features.

Predicting how long a trip could take is useful in many scenarios. We assess the importance of trip duration prediction from the bike sharing service's perspective. Being able to determine how long it might take would be useful in demand analysis - it could help them decide how and where to restock in order to satisfy the public's needs. For instance, if multiple cycles are picked up from a seemingly busy station, and trip duration for most of it is predicted to be very long, owing to factors such as it being a leisure ride, one can predict that those bikes might be unavailable for a considerably long period of time, and this could give us insights into restocking the bicycles in various stations. Furthermore, trip duration prediction can also help in analysing the usage patterns of the users and even give the users more personalized service based on the predicted duration, an example being smart recommendations for cool down spots if the duration is assumed to be long. Such downstream possibilities would incentivize the public to adopt using bicycles, thereby contributing to reduction of carbon footprint in urban context.

In the following section, we look at various previous works utilizing bike sharing data, and how it is used in various applications.

2 RELATED RESEARCH

Bike sharing systems (BSS) are gaining popularity in cities throughout the world. They are recognized as integral parts of public transportation systems, fully anticipating societal trends such as the sharing economy and healthy and sustainable urban lifestyles. While BSSs are well established in large metropolises such as Paris, London, and New York City, large and mid-sized cities have only recently begun to experiment with them or are in the process of implementing new systems[10].

Chiariotti et al. state that Bike-sharing programs are booming in Smart Cities all around the world. They are a low-cost, environmentally friendly mode of transportation that helps to alleviate traffic congestion. However, these new services are still in the works, and various obstacles must be overcome. The management of rebalancing trucks, which ensures that bikes and stalls in docking stations are always available when needed, despite fluctuations in service demand, is a critical issue. They offer a dynamic rebalancing technique in this paper that uses previous data to predict network circumstances and act quickly if necessary. They employ Birth-Death Processes to model station occupancy and determine when to redistribute bikes, as well as graph theory to choose the rebalancing path and stations. The suggested paradigm is validated using data from New York City's bike-sharing system. The numerical simulations reveal that a dynamic technique that can adapt to the network's fluctuating nature outperforms static rebalancing schemes[5].

2.1 Patterns, Predictions, Planning and Visualization

Ghosh et al. in their recent research on "Dynamic Repositioning to Reduce Lost Demand in Bike Sharing Systems" say that Bike Sharing Systems (BSSs) are commonly used in major cities across the world due to concerns about increased carbon emissions, traffic congestion, and the use of nonrenewable resources associated with substantial private vehicle use. Base stations are strategically located around a city in a BSS, and each station is filled with a certain number of bikes at the start of the day. Customers can rent bicycles from one station and return them to another. Due to the unexpected movements of consumers renting bikes, base stations are either congested (more than required) or depleted (fewer than required) of bikes. According to existing statistics, congestion/starvation is a regular occurrence that results in a big number of dissatisfied customers and a significant drop in customer demand. To address this issue, they present an optimisation formulation for repositioning bikes utilizing automobiles while taking into account vehicle routes and future predicted demand. In addition, they provide two ways that use decomposability in the problem (bike repositioning and vehicle routing) as well as aggregation of base stations to greatly minimize computation time. Finally, they compare our technique to two benchmark approaches on two real-world data sets from bike sharing systems to show how useful it is. These methods are tested using a simulation in which client movements are created from real-world data sets[7].

Yexin Li et al. in their study on "Traffic Prediction in a Bike-Sharing System" say that In many large cities, bike-sharing systems are widely used, providing a convenient method of transportation for individuals' commutes. The bikes in a system must be rebalanced often since the rents/returns of bikes at different stations during different periods are imbalanced. Real-time monitoring is ineffective in addressing this issue because reallocating the bikes when an imbalance occurs takes too long. They present a hierarchical prediction model in this research to anticipate the amount of bikes that will be rented from/returned to each station cluster in the future so that reallocation may be done ahead of time. To cluster bike stations into groups, they first propose a bipartite clustering technique, which results in a two-level hierarchy of stations. A Gradient Boosting Regression Tree predicts the total number of bikes that will be rented in a city (GBRT). Then, to anticipate the rent proportion across clusters and the inter-cluster transition, a multi-similarity-based inference model is presented, based on which the number of bikes rented from/returned to each cluster can be simply inferred. They test their model on two bike-sharing systems in New York City (NYC) and Washington, D.C. (D.C.), finding that it outperforms baseline techniques (0.03 decrease in error rate), particularly during anomalous times (0.18/0.23 reduction in error rate)[9].

N. Oliveira et al. offer an interactive visualization method to examine the dynamics of public bike-sharing systems by profiling its historical record in order to better understand how they are used. Their design facilitates the discovery of multiple patterns in temporal and spatial domains by coordinating a pixel-oriented timeline with a map and introducing a technique of partial reordering of time series. They use Citi Bike, New York City's bike-sharing program, as a use case and build a prototype to display changes in the system over a ten-month period, ranking stations based on several features, and using any time interval in daily and monthly timelines. Various analyses are presented to validate the visualization system as a useful operational tool that can assist the staff of large-city bike-sharing programs in exploring such large datasets in order to better understand commuting dynamics, overcome management problems, and provide better service to commuters[11].

Vogel et al. in their paper on "Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns" use comprehensive operational data from bike-sharing systems to derive bike activity patterns in this paper. In bike-sharing systems, inequities in bike allocation are a regular problem. To get insight into the complicated bike activity patterns at stations, They apply Data Mining. Bike distribution imbalances are revealed by activity patterns, which leads to a better knowledge of the system structure. For the design and management of bike-sharing systems, a structured Data Mining approach aids planning and operational decisions[12]. Kultenbrunner et al. in their recent research on Urban cycles and mobility patterns have used the amount of available bikes in the stations of the community bicycle program Bicing in Barcelona to analyze human mobility data in an urban region. It is feasible to detect temporal and geographic mobility trends inside the city using data taken from the operator's website. These patterns are used to forecast the quantity of available bikes at any station a few minutes or hours in advance. The forecasts could be utilized to improve the bicycle program and the information provided to users through the Bicing website[8].

Gast et al. have done an interesting study on Probabilistic Forecasts of Bike-Sharing Systems for Journey Planning. They investigate the topic of predicting the future availability of bicycles in bike-sharing system stations (BSS). This is important when making recommendations to ensure that the likelihood of a user being able to complete a journey is high enough. They do this by using probabilistic predictions from a BSS queuing theoretical time-inhomogeneous model. Using historical data from the Vélib' BSS of the City of Paris, the model was parameterized and successfully validated. Because it does not account for the stochasticity inherent in the real system, They present a critique of the standard root-mean-square-error (RMSE), which is often used in bike-sharing studies as an indication of prediction accuracy. Instead, They develop a new scoring-rule-based metric. They compare our model's average score to traditional predictors utilized in the literature. For prediction horizons of up to a few hours, They show that our model outperforms them. We also examine how, in general, counting the quantity of available bikes is only useful for forecasting horizons of a few hours or less[6].

2.2 Trip Duration Prediction

As for the trip duration prediction, Wang et al. have done an interesting job. In their research on "Travel time estimation of a path using sparse trajectories", they propose a citywide and real-time model for estimating the travel time of any path (represented as a sequence of connected road segments) in real time in a city, based on GPS trajectories of vehicles received in current time slots and

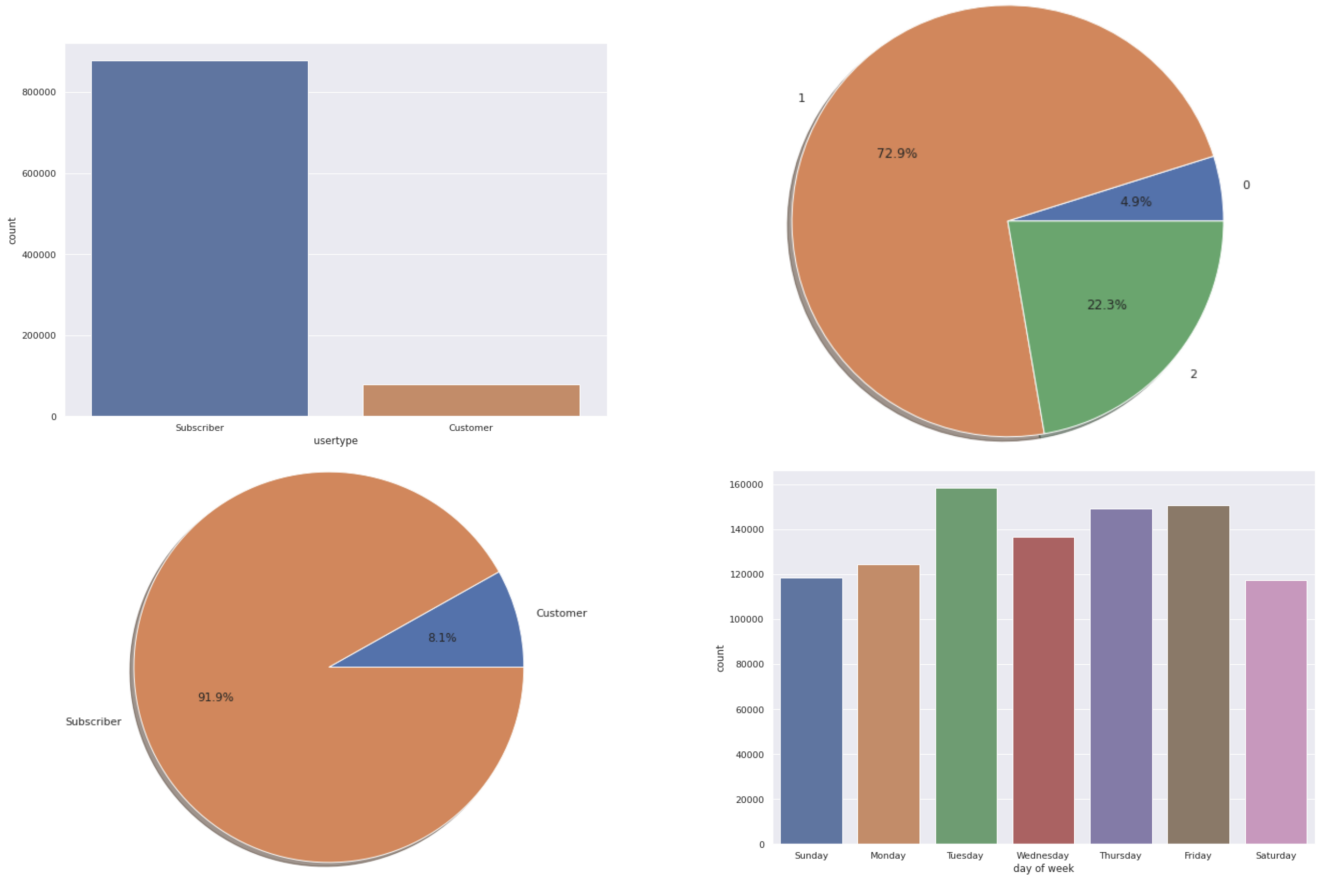

Figure 1: (top-left) Depicts the deviation between user types of the bike sharing service (top-right) Depicts the percentage of male (1), female (2), unknown genders(0) that use the bike sharing service (bottom-left) Depicts the percentage of customers vs subscribers that use the bike sharing service (bottom-right) Visualization of the popular days of the week to use the bike sharing service

over a period of history, as well as map data sources, in this paper. Despite the fact that this is a strategically vital duty in many traffic monitoring and routing systems, the problem has yet to be solved due to the three challenges listed below. The first is the data sparsity issue, which means that many road segments may not be driven by any GPS-equipped vehicles at this moment. In the vast majority of cases, we are unable to find a trajectory that traverses a query path perfectly. Second, there are numerous techniques to estimate the corresponding trip time for a portion of a path with trajectories utilizing (or combining) the trajectories. Finding the best combination is a difficult task, as there is a tradeoff between the length of a path and the number of trajectories that pass through it (i.e., support). Third, we must be able to respond to consumers' requests in real time, regardless of where they are in the city. This necessitates the development of a cost-effective, scalable, and effective solution that can provide citywide and real-time trip time estimation. To overcome these issues, we use a three-dimensional tensor to describe different drivers' trip times on different road segments in different time intervals. They use a context-aware tensor decomposition strategy to fill in the tensor's missing values using geospatial, temporal, and historical contexts learnt from trajectories and map data. Then, using a dynamic programming solution, they create and prove an object function to describe the aforementioned tradeoff, and identify the most optimal concatenation of trajectories for an estimate. They also propose employing frequent trajectory patterns (mined from past trajectories) to narrow down concatenation candidates and a suffix-tree-based index to handle trajectories received in the current time slot. Approximately the course of two months, they tested our strategy using GPS trajectories provided by over 32,000 cabs. The results show that our strategy outperforms baseline approaches in terms of efficacy, efficiency, and scalability[14].

2.3 Miscellaneous

"Gender Gap" is one of the topics that has been taken to attention in recent researches in this area and many valuable works has been found. Wang and Akar in their research on Gender Gap in Bike Share Ridership explore the environmental determinants of bike sharing utilization for males and females using data from New York City's Citi Bike Share system. They also simulate the effects of bicycle infrastructure, land use considerations, and public

transportation services on the proportion of female trip arrivals. The findings imply that the environmental factors of bike share utilization are basically similar for males and females. The projected magnitudes, however, suggest that our factors of interest may have distinct effects on males and females. Installing more bicycle racks, for example, is linked to increased bike share usage for both men and women. They also discovered that this element had a greater impact on women than on males. A 1% increase in the number of bicycle racks is associated with a 1.18 percent rise in the proportion of women who arrive on excursions. The results can be used to evaluate the efficacy of future infrastructure expenditures aimed at closing the gender gap in bike share usage. The findings also provide useful information on how to increase total bike share ridership and so promote local bicycle culture[13]. Abasahl et al. in another research on "Gender gap generators for bicycle mode choice in Baltimore college campuses" investigate gender equity in bicycle mode choice and barriers inhibiting women from bicycling. To identify gender difference generators, socioeconomic data, travel preferences, mode accessibility, and individual aspects of the examined population are used. They use a bivariate statistical analysis and a two-level nested logit model to study gender equity. Females are around 30% less likely to bike from home to college, according to the bivariate statistical study, and are much more sensitive to environmental and infrastructure factors. The results of the two-level nested logit model, which are complementary to the bivariate statistical analysis, show that females are discouraged from riding because of great distances, longer journey times, lack of access to a bicycle, and a dangerous environment. They also discovered that undergraduate ladies are less likely than other groups of students to ride to campus. Their findings show that integrating bicycle and transit services, upgrading infrastructure to segregate bicycle and motorized traffic, improving bicycle facility safety, and increasing public awareness of local bicycle routes all encourage female biking[4].

3 DATA

3.1 NYC CitiBike Data

The dataset that we analyzed displays the traffic prediction in a bike-sharing system in NYC from the most recent NYC CitiBike dataset in 2019. It displays a variety of valuable information, such as trip duration, start time, stop time, start station id, start station name, end station name, user type, birth year, gender, and day of week. [9]

We create a variety of graphical representations to view the relationship between different attributes in the bike sharing data as seen in Fig 1. One representation is a count plot of the type of users. There are approximately 900000 subscribers while there are approximately 15000 customers. This demonstrates that more people that use the bike sharing service are subscribed to it so they can repeatedly use it. Another graphical representation we have created is the count of the number of bike rides started on each day of the week. The graph shows that Tuesday is the most popular day of the week for using the bike sharing service, as it has a total of approximately 160000 rides. Saturday is the least popular day of the week to ride the bike, as it has only about 120000 rides in total. This representation helps visualize the popular days of the week for the bike sharing service. The pie chart representation was created to easily visualize the gender ratio of those using the bike sharing service. 72.9 percent of people using it are men, while 22.3 percent are women. 4.9 percent have not specified their gender.

3.2 NYC POI Data

This dataset [2] contains a total of 14229 Points of Interest (POI) split into 14 domains such as Residential, Commercial and Health Care.

4 METHODOLOGY/APPROACH

Prediction of trip duration is treated as a problem of regression. Our experiment is molded by two factors in the context of the bike sharing systems. Firstly, we examine the presence of redundancy in network activity and a possibility of simplifying the station network. Secondly, we analyse the importance of destination or drop off station of a trip in determining trip duration. The following subsections elaborate on the rationale for pursuing each of these considerations:

4.1 Problem Simplification

An exploratory analysis of the bike sharing network shows us that it is very dense - we see that the average distance between a station and its nearest neighbor is $\sim$ 230 meters. Hence, the usage patterns and trends would be very similar for multiple stations. This calls for a possibility of problem simplification, where multiple nodes in the network could come under the same bucket. A simple and intuitive means of doing this would be to geographically cluster the stations. This would identify possible neighborhoods and then categorize stations based on which neighborhood they fall into. While this is a great way to simplify the network, we wondered if there was some other, more meaningful way to group the stations. Hence, we employ two kinds of problem simplification strategies - geographic positional clustering and Purpose-aware Categorization.

4.1.1 Geographical Clustering. Geographical clustering is done using k-means clustering. The distance metrics used are solely based on the Euclidean distance between the latitude and longitude. Geographical clustering has been employed in various previous works for problem reduction or simplification.

4.1.2 Purpose-aware Categorization (PaC). While Geographical clustering serves as a good way of bringing together close by stations and preventing redundancy in the features, we believe we could do more in understanding and simplifying the problem in terms of the usage patterns. This brings us to purpose-aware categorization of stations. In this study, each bike station is assumed to serve a specific purpose based on its surroundings, or technically speaking, its context. This makes sense because a station‘s usage is heavily dependent on what is around it. The kind of incoming and outgoing users and their trips are heavily governed by its location. A user would likely check in to (drop off the bicycle at) a station which is closest to their destination. Given the dense spread of the bike sharing stations, it is very unlikely for a user to stop at one station and walk miles to get to their purported destination when there is a bike station much closer to their destination. Hence, the main purpose of a station can be determined by what is very close

Feature Type	Feature Name	Values	Used in Setting M1/M2/M3?	Description
User	gender	0=unspecified 1=male 2=female	M1, M2, M3	Gender of user
	usertype	0=Customer 1=Subscriber	M1, M2, M3	Type of user
Time of Trip	hour	0-23	M1, M2, M3	Hour component of start time
	minute	0-59	M1, M2, M3	Minute component of start time
	is_weekend	1 if Sat/Sun 0 otherwise	M1, M2, M3	Whether that day was a weekday or not
Station Details	cluster_start	0-24	M1, M2	Encoded geographic cluster of pick up station
	category_start	0-4	M1, M2	Encoded purpose category of pick up station
	cluster_end	0-24	M2	Encoded geographic cluster of drop off station
	category_end	0-4	M2	Encoded purpose category of drop off station
	encoded id	0-881	M3	Encoded station ID of pick up station

Table 1: Features used in trip prediction modeling for Settings M1, M2, M3

to it, and we resolve to a context-aware clustering or categorizing of each of the stations based on the surroundings. In order to get some insight on the surrounding areas, we look at Points of Interest (POI(s)) in the areas. The rationale behind doing so is that the activity involved in an area is directly related to the POIs around it. POI dataset [2] is used for this, which gives us a comprehensive list of POIs diligently categorized into several categories. The POIs are superimposed with the stations network, and each station is categorized based on what it is surrounded by. "Radial Nearest Neighbors" [3] is employed to capture the POIs that surround a given station within a predetermined radius θ. Finally, a maximum vote is taken to categorize the main purpose of the station. We call this Purpose-aware Clustering (PaC). If we are looking at a station s_k and $n^{\theta}_{t_j}$ is the count of POIs of category t_j within radius θ from s_k, the "purpose" of station, P_{s_k} as we describe in PaC is:

$$P_{s_k} = \underset{t_j \in T}{argmax} \left(\left\{ n^{\theta}_{t_1}, n^{\theta}_{t_1} ... n^{\theta}_{t_{n_t}} \right\} \right)$$

Doing so would simplify the station network into a more coarse grained, but more insightful distribution of the stations which would help us understand what each station would mainly be utilized for. Eg: If a station is categorized as "residential", we can assume that bikes in this station are checked out and checked in most often during peak hours to leave to and return from work respectively. Similarly, a station classified as "commercial" could have high demand patterns in the weekend due to increased shopping crowds over the weekends.

4.2 Prediction using pick-up station alone vs. both pick-up and drop-off stations

Pragmatically speaking, prediction of trip duration in the context of bike sharing systems would be slightly different from other urban transportation networks such as taxi transport systems. With such systems, we most certainly have information about the destination right from the beginning of the trip, which becomes an important and an instrumental feature in the prediction. With bicycle sharing systems, we may not be able to depend on obtaining the destination station at the beginning of the journey. To this end, our study first attempts at predicting trip duration with only the pick up station information given, followed by modeling trip duration prediction with both pick up and destination station information given and analyse the importance of destination information.

4.3 Representation

Each of the n_s stations would be represented as s_k where $k \in [1, n_s]$

$$S = \{s_k | k \in [1, n_s]\}$$

Each of the n_c geographic clusters would be represented as c_i where $i \in [n_c]$

$$C = \{c_i | i \in [1, n_c]\}$$

Each of the n_t purpose categories would be represented as t_j where $j \in [1, n_t]$

$$T = \{t_j | k \in [1, n_t]\}$$

Each station s_k belonging to a cluster c_i and purpose category t_j can be annotated as $s_k(c_i, t_j)$.

4.4 Predictive Modeling

As mentioned in the beginning of this section, this prediction problem calls for a regression model. As for model architectures, we model Random Forest Regression, Linear Regression, XGBoost Regression and Artificial Neural Networks for the same. To expand on the two considerations mentioned in sections 4.1 and 4.2 , we would be analyzing the performances of these models in three settings, each of which vary in terms of how the station information is presented as features. Note that c_i is geographic cluster i, t_j is purpose category j, s_k is station k

- M1: Station information simplified to tuple (c_i, t_j) for pick up station only
- M2: Station information simplified to tuple (c_i, t_j) for pick up and destination stations
- M3: Station information for pick up station only with no simplification; Each station would just be indicated using its ID, s_k

Comparing performances for M1 and M2 would help us in examining whether having source station information alone can help

in making good predictions, while comparing M1 and M3 would help us determine if the proposed problem simplification is a valid abstraction step to take.

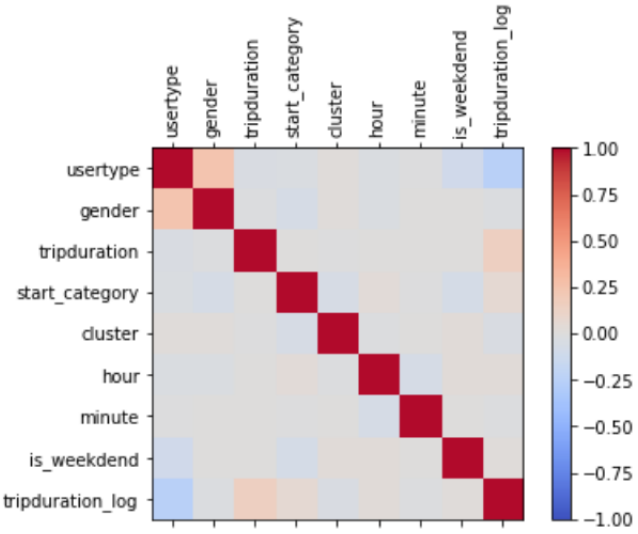

Figure 2: Correlation plot between features and candidate response variable

Furthermore, Fig 2, shows that there is barely any linear correlation between the features and the final trip duration (feature name: tripduration). To enhance the linear learning ability of the data, we perform a natural logarithmic transformation on trip duration. We now see in Fig 2 how $log(tripduration)$ (feature named tripduration_log) has more linear correlation than tripduration.

4.5 Features Used

Prediction of trip duration would need more than just station information. Knowing what time of the day it was, and whether or not it was a weekend could help a lot in understanding patterns in the nature of trips. Ex: weekend cycling pursuits would usually be for leisure and would hence probably take longer. Similarly, knowing some details about the user such as the user's gender could also help in assessing trip lengths. Table 1 lists the features employed in detail, along with the clear indication of which features were used for settings M1, M2 and M3.

5 EXPERIMENTS

5.1 Experimental Setup

5.1.1 Geographical Clustering: A k value of 25 was employed to ensure the right balance between fine and coarse-grained clustering. With this value for k, we see a good division between the stations not only across the map as a whole but also within each borough as well. Fig 4 shows the clusters formed as a result of the above mentioned clustering process. The library Folium is used to create the plots on the map.

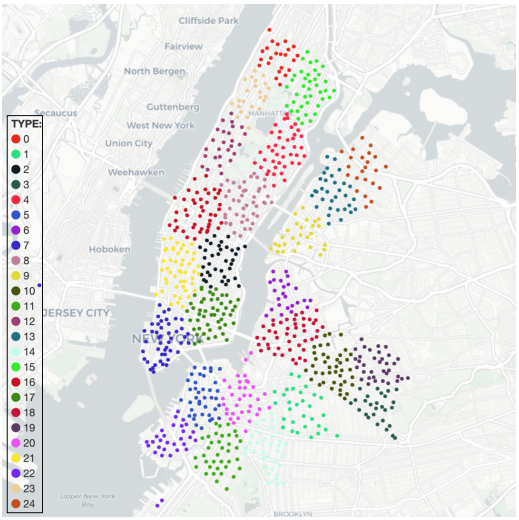

Figure 4: Bike share stations divided into clusters

5.1.2 Purpose-aware Categorization: Five categories are chosen from the POI dataset, namely "Residential", "Education Facility", "Recreational Facility", "Transportation Facility", "Commercial" and "Health Services". The radius for the categorization was chosen based on realistic measurements. A user would typically cycle to a drop off station nearest to their destination. Given the dense nature of the network, it is assumed that an average user would consider an "after cycling" commute of one block. Since the standard block length in Manhattan is 274 meters [1], we take the radius of 300 meters, which is close to one block length. Fig 3 (next page) shows two plots - to the left is of POIs considered for the categorization, and to the right is a plot of the stations color coded with the category assigned to them. The library Folium is used to create the plots on the map.

5.1.3 Modeling: Random Forest was trained with 250 participating decision trees. ANN is composed of two hidden layers with 5 nodes in the first hidden layer and 10 in the second hidden layer. Each of the layers has ReLU activation. Adam Optimizer is used with a learning rate of 0.001. Batch size used for training is 32. The linear regression model is a simple linear regression without any penalties incorporated, and XGBoost model is used with its default parameters.

5.2 Results

Table 2 lists the results of our experiment. We see that for M1, XG Boost performs the best with a mean squared error (MSE) of 0.519 and adjusted R squared score of 0.090. As for M2, random forest regression performs the best with an MSE and adjusted R squared of 0.374 and 0.343 respectively. Finally for M3, XG boost again performs the best with an MSE and adjusted R squared of 0.523 and 0.082 respectively. Now, we move on to compare the performances of settings M1 M2 and M3 by comparing the best performing models, as stated above, for each of them. It is quite evident that presence of drop off station information significantly improves the learning ability of the models and makes for a model

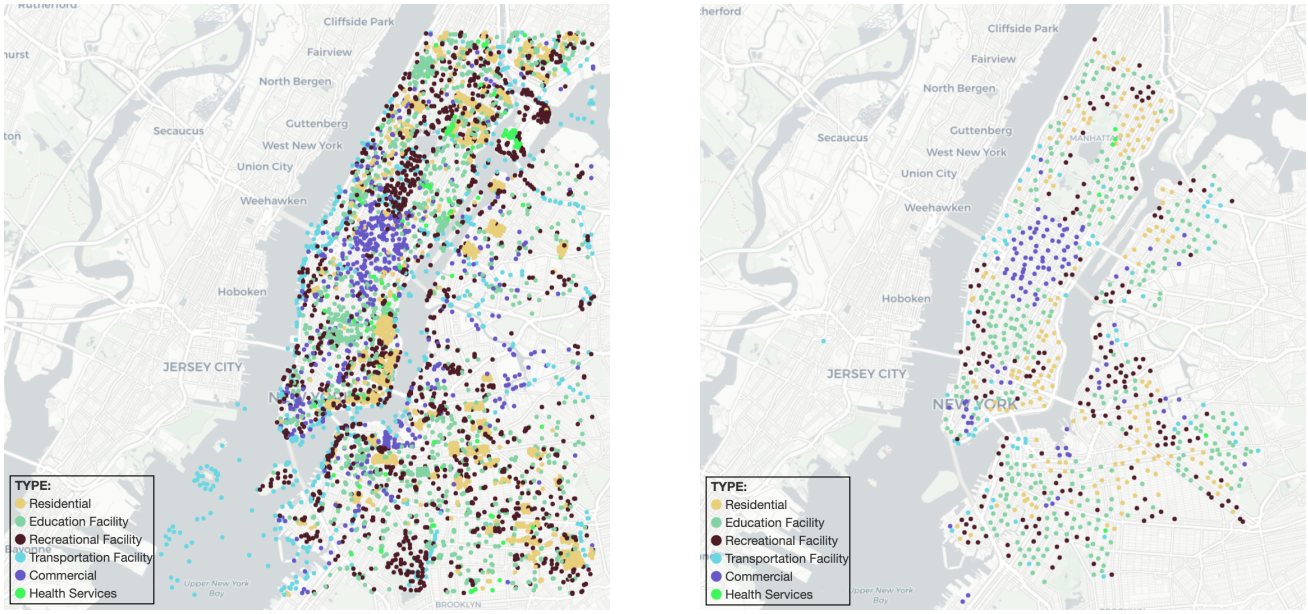

Figure 3: (left) POIs colored in terms of their type plotted on a map of New York City. (right) Bike Share stations categorized in terms of their major purpose

	M1		M2		M3	
	MSE	Adj-R2	MSE	Adj-R2	MSE	Adj-R2
Random Forest	0.626	-0.098	**0.374**	**0.343**	0.583	-0.022
ANN	0.533	0.066	0.406	0.288	0.535	0.063
XG Boost	**0.519**	**0.090**	0.433	0.241	**0.523**	**0.082**
Linear Regression	0.533	0.066	0.531	0.070	0.535	0.062

Table 2: Results showing MSE and Adjusted R-squared values for the 4 regression models across the three settings M1, M2 ad M3.

of immensely higher quality. We see that the adjusted R squared value for M2 is almost 4 times larger than that for M1 and M3. This provides enough evidence to conclude that having destination information is extremely important and can actually play a big role in increasing trip duration prediction performance. While having only source information generates models of comparable MSE, having some details about the destination could increase prediction accuracy by a lot. Furthermore, comparing M1 an M3 shows us that the performance of M1 is slightly higher than the performance of M3 in almost all the cases. The conclusions being made from this observation are two-fold. Firstly, having fine-grained station-level information is clearly not necessary since the performance with simplified features is not just as good as with station-level features but in fact better than them. Hence, we do not need to have fine-grained station-level information, when our simplified features can be confidently employed in the prediction problem. Secondly, as previously mentioned, we see that the performance of M1 is almost consistently higher than the performance of M3. This shows that having simplified features for each trip data-point is not only a requisite form of representing each station but is in fact a better form of representation for stations. It is evident from our results that our simplified features provide more insight into the usage patterns and to the task of trip duration prediction.

6 DISCUSSION

Thus far, we have been looking at trying to understand the problem of predicting the trip duration in a quantitative context. In this section, we look at it in a qualitative manner and try to understand the various attributes and patterns in the bike sharing system. Firstly, from the section 5.2, we see that using our semi fine-grained features gave better results than using fine-grained station-level features. As a matter of fact, it helped in reducing any kind of redundancy in data and creating a tighter and better representation, and also gave us better insights as evidenced by the performance. This shows that while each station could be different in various aspects, many stations are similar at a higher level, and utilizing this less complex, less fine-grained network would actually give us similar results. We also see how within each geographical cluster, we have stations of different purposes, which highlights the heterogeneity of each of these regions in New York City. To elaborate, none of these neighborhoods are strongly residential or strongly commercial, for instance, but have a good mix of many

kinds of establishments. This reflects on the cultural and social richness of New York City. We also conclude that the problem of trip prediction is very complex since there are multiple factors involved in the same. Even seemingly unrelated features like user details and the starting time of a trip can play a big role in leading to more accurate predictions. Qualitatively, this makes sense because each trip is distinct, and part of what makes a trip happen is based on the users' needs and their intent to travel, which is successfully encoded in our features using purpose aware categorization.

Finally, from the system's perspective, the presence of semi fine-grained features as opposed to fine-grained station-level features makes the prediction system more adaptable to addition of newer stations. A new station would most certainly belong to one of geographic clusters and one of the existing purpose categories. Also, new stations are more likely to be added than new POI categories or even geographic clusters. Hence, having a change-agnostic feature system like ours would go a long way in system quality and maintenance.

7 CONCLUSION

In this project, we analyzed the data of New York bike sharing system and modeled the task of predicting the trip duration by using machine learning and feature mining. This work can help analyze the amount of bikes needed for many of the cities like New York City, and help officials decide when to make the most bikes available. We conclude that having categorized buckets composed of a station's geographical cluster and major purpose is a better representation of the stations than fine-grained station level information, evidenced by better performance of the former. We also conclude that having destination information could be very beneficial in determining the trip duration.

8 CONTRIBUTIONS

- **Meghana Meghana:**
  Report: Methodology/Approach, Experiments, Abstract, Introduction, Discussion, Conclusion
  Meghana's Code: DurationPrediction.ipynb;
- **Elham Mohammadrezaei:**
  Report: Related Works
- **Uditi Goyal:**
  Report: Abstract, Introduction, Discussion, Conclusion
- **Riya Dani:**
  Report: Graph Analysis, Data;
  Riya's Code: PlotData.ipynb

Full Code Link: https://gitlab.com/meghana-holla/cs5834_tripdurationprediction/-/tree/main

Points of Interest Dataset Source: https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj

Bike Sharing Dataset Source: https://ride.citibikenyc.com/system-data

Specific 2019 Dataset Distribution: https://s3.amazonaws.com/tripdata/201912-citibike-tripdata.csv.zip

9 ACKNOWLEDGEMENTS

We would like to thank CitiBike System Data and NYC Open Data for making the bike sharing data and the Points of Interest data respectively publicly available.

REFERENCES

[1] City block. https://en.wikipedia.org/wiki/City_block. Accessed: 2021-12-12.
[2] Points of interest. https://data.cityofnewyork.us/City-Government/Points-Of-Interest/rxuy-2muj. Accessed: 2021-12-12.
[3] Fixed radius near neighbors. https://en.wikipedia.org/wiki/Fixed-radius_near_neighbors. Accessed: 2021-12-12.
[4] Farhad Abasahl, Kaveh Bakhsh Kelarestaghi, and Alireza Ermagun. Gender gap generators for bicycle mode choice in baltimore college campuses. *Travel behaviour and society*, 11:78–85, 2018.
[5] Federico Chiariotti, Chiara Pielli, Andrea Zanella, and Michele Zorzi. A dynamic approach to rebalancing bike-sharing systems. *Sensors*, 18(2):512, 2018.
[6] Nicolas Gast, Guillaume Massonnet, Daniël Reijsbergen, and Mirco Tribastone. Probabilistic forecasts of bike-sharing systems for journey planning. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, pages 703–712, 2015.
[7] Supriyo Ghosh, Pradeep Varakantham, Yossiri Adulyasak, and Patrick Jaillet. Dynamic repositioning to reduce lost demand in bike sharing systems. *Journal of Artificial Intelligence Research*, 58:387–430, 2017.
[8] Andreas Kaltenbrunner, Rodrigo Meza, Jens Grivolla, Joan Codina, and Rafael Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive and Mobile Computing*, 6(4): 455–466, 2010.
[9] Yexin Li, Yu Zheng, Huichu Zhang, and Lei Chen. Traffic prediction in a bike-sharing system. In *Proceedings of the 23rd ACM International Conference on Advances in Geographical Information Systems*. ACM SIGSPATIAL 2015, November 2015. URL https://www.microsoft.com/en-us/research/publication/traffic-prediction-in-a-bike-sharing-system/.
[10] Martin Loidl, Ursula Witzmann-Müller, and Bernhard Zagel. A spatial framework for planning station-based bike sharing systems. *European Transport Research Review*, 11(1):1–12, 2019.
[11] Guilherme N Oliveira, Jose L Sotomayor, Rafael P Torchelsen, Cláudio T Silva, and João LD Comba. Visual analysis of bike-sharing systems. *Computers & Graphics*, 60:119–129, 2016.
[12] Patrick Vogel, Torsten Greiser, and Dirk Christian Mattfeld. Understanding bike-sharing systems using data mining: Exploring activity patterns. *Procedia-Social and Behavioral Sciences*, 20:514–523, 2011.
[13] Kailai Wang and Gulsah Akar. Gender gap generators for bike share ridership: Evidence from citi bike system in new york city. *Journal of transport geography*, 76:1–9, 2019.
[14] Yilun Wang, Yu Zheng, and Yexiang Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 25–34, 2014.