2022 Fall Deep Learning Final-Project
NYC Street Bike Sharing Demand Class Prediction

Yichen Guo,[1] **Tao Liang,** [2] **Zhaorui Ma,** [3]

[1] Center for Urban Science Progress. New York University Tandon School of Engineering, United States
[2] Department of Civil Engineering. New York University Tandon School of Engineering, United States
[3] Department of Electrical and Computer Engineering. New York University Tandon School of Engineering, United States
GitHub link: https://github.com/liangtao1216/DLFinalProject.git

Abstract

Bike-sharing has become a popular and environmentally friendly mode of transportation in recent years, offering people additional travel options and convenience while reducing pressure on traditional transportation systems. However, the increasing popularity of shared bikes has highlighted the need for more bike lanes and supporting facilities in many cities. In this project, we aim to provide urban planners with a tool to identify which segments of a city's streets are most in need of additional bike lanes and supportive measures, using Citi Bike in New York City as a case study. To achieve this, we employed three machine learning methods - Random Forest, MLP, and Graph Neural Network (GNN) - to predict the demand classes for shared bike usage on different street segments. Transportation planners and bike companies can use these predictions to prioritize investments in bike lanes and bike stops. Our final GNN model reached a classification accuracy of 63%, which outperformed our other baseline models.

Introduction

Background

The convenience and environmental benefits of bike-sharing have led to its increasing popularity as a mode of transportation. By addressing the "last mile" problem, bike-sharing allows people to ride shared bikes from the starting point to a public transportation hub (such as a subway station or bus stop) and then from the hub to their final destination, instead of walking or using a personal vehicle. However, the lack of adequate bike lanes and supporting infrastructure can limit the adoption of bike-sharing. Factors such as the location and density of bike-sharing stations and their impact on existing transportation plans and sidewalks must also be considered in the efforts of governments and legislators to address the popularity of bike-sharing and provide necessary bike lanes and facilities. It is not fully understood which factors are most influential in determining the demand for bike lanes and amenities in a given area, or which machine learning model is best suited to predict these needs. In this project, we aim to explore the characteristics of areas with a higher demand for bike lanes and supporting facilities, and to use machine learning and deep learning techniques to predict the demand classes of each segment in New York City. By predicting segment bike flow, we aim to optimize the operation of the bike-sharing system and prioritize bike lane investments based on their potential impact on bike-sharing trips, rather than traditional factors such as the number of car-bicycle collisions or the value of a particular roadway to the network. We will compare the performance of different models in order to identify the most effective approach.

Related works

There have been a number of research papers published on the topic of predicting link flow in the Citibike system using various machine learning and data analysis techniques. Here are a few examples:

"Predicting Bike Rentals in New York City: A Graph-Based Approach": This paper proposes a graph-based approach for predicting bike rentals in the Citibike system, using Graph Convolutional Networks (GCN) to learn patterns and relationships in the data. In this paper, the authors propose a graph-based approach for predicting bike rentals in the Citibike system using Graph Convolutional Networks (GCN) (Chen et al. 2017).

The authors first collected and pre-processed data on bike rentals in the Citibike system, including information on the time and location of each rental, as well as weather and holiday data. They then represented the data as a graph, with nodes representing bike rental stations and edges representing the connections between the stations. The graph was constructed such that the node features represented the characteristics of the rental stations, and the edge features represent the connections between the stations. The authors then splited the data into training, validation, and test sets and pre-processed the data for use with a GCN model. They then trained a GCN model and evaluated its performance on the validation and test sets. The GCN model was able to learn patterns and relationships in the data and make accurate predictions of bike rentals in the Citibike system and outperformed several baseline models, including a linear regression model and a random forest model. They found that the GCN model outperformed the baseline models in terms of prediction accuracy.

"Traffic Flow Prediction in Bike Sharing Systems: A Deep Learning Approach" (Zhou et al. 2018): This paper presents a deep learning approach for predicting traffic flow in bike sharing systems, including the Citibike system. The authors proposed a hybrid deep learning model that com-

bined a long short-term memory (LSTM) network with a convolutional neural network (CNN) to predict the traffic flow of bike sharing systems. The model was trained on a large dataset of bike rental records from a real bike sharing system in China. The authors evaluated the performance of their proposed model using various evaluation metrics and compared it with several baseline models. They found that their hybrid model outperformed the baselines and demonstrated good prediction accuracy and generalization ability.

"Bike Sharing Demand Prediction: A Comparative Study"(Zhang et al. 2018): This paper compares the performance of several machine learning models for predicting bike sharing demand, including the Citibike system. The authors considered several factors that may affect bike sharing demand, including weather, time, and location. They used a dataset of bike rental records from a real bike sharing system in Beijing, China to train and evaluated the performance of different prediction models.

The authors compared several machine learning models, including linear regression, decision tree, random forest, gradient boosting, and support vector machine (SVM). They also compared the performance of these models using various evaluation metrics, including mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

Overall, there has been a significant amount of research on the topic of predicting the Citibike demand, and a variety of machine learning and data analysis techniques have been proposed and evaluated. These techniques include graph-based approaches, deep learning models, and traditional machine learning models, among others.

Data

Citibike Dataset

The Citibike dataset is a publicly available dataset containing bike rental records from the Citibike bike-sharing system in New York City. It includes information on the start and end stations, start and end times, and duration for bike rentals. We chose data from September 2021 for this study because it represents the warmer season and has the highest volume of rentals.

Since our goal is to find out which street needed to install the bike lane or needed to install new bike stations, we need trip data on each segment. However, the Citi-bike data shared with the public is origin-destination data, meaning you know the stations where a trip starts and ends but not the route the cyclist took in between. In order to derive the flow on street, we used OSMnx, which query data from OpenStreetMap, to estimate the trajectories between start stops and end stops, then assign these trips to the corresponding segment.

To analyze the data, we used the Citibike median daily flow data as the target variable, which we then divided into five classes to more easily understand the demand ratings for citibike on different street segments. These classes were based on the distribution of bike flow percentages, with quantile ranges allowing for five bins (0%-20%, 20%-40%, 40%-60%, 60%-80%, 80%-100%) and ensuring balanced class representation.

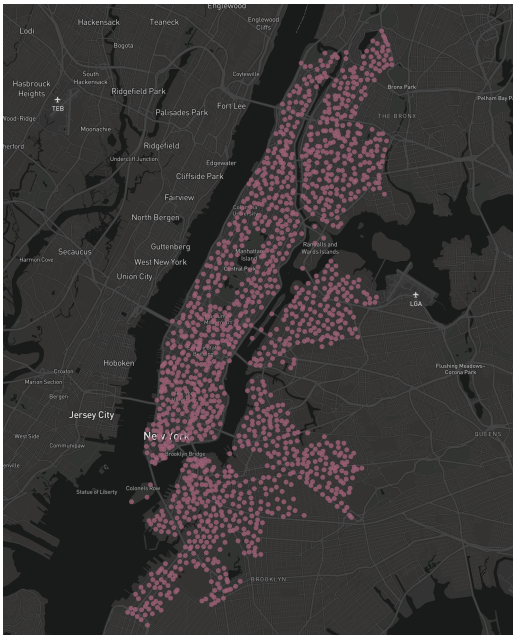


Figure 1: Citibike station distribution

We also plotted the Citibike station distribution in Figure 1. Each point represents a Citibike station.

Features

We assigned our data sources into two categories: segment related data and demographic data. The segment related features are extracted from Single Line Street Base Map (LION).The segment data included features like:

- street width
- number of travel lanes
- number of parking lanes
- with bike lane or not
- protected bike lane or not
- unprotected bike lane or not
- truck route or not
- borough dummy variables
- distance to nearest subway entrance

The demographic attributes were sourced from the 5-Year American Community Service in 2020 (ACS5), the units is census tract. The demographic attributes includes:

- population
- median age
- median income
- percent with bachelor's degree
- unemployment rate
- percent without vehicle
- mode split: public transportation user

For demographic data, we spatial joined the census tract boundaries with the segments to assign the information to each segment.

Data Limitation

It is important to note that the data comes with certain limitations. The shortest routes generate using OSMnx may not be the optimal method, Google API and HERE API have better routing service, but the problem is too costly.

Exploratory Data Analysis

Before we went further with our model training, we explore the correlations between features. Figure 2 shows the correlation between data. Strong correlations (with a p-value less than 0.001) were found between the median daily Citibike flow and other variables of interest.

Methods

Data Processing

Steps below shows our data processing flow:

1. Pre-processed the data: we pre-processed the data to handle missing values, scaled numerical features and encoded categorical features.
2. Transformed the data into a suitable format: Depending on the specific model we were using. For example, for GNN model, we converted the data into a graph structure, with each intersection represented as a node and the segment represented as edges, social factors as node features, road condition as edge features, and length of the segment as edge weight.
3. Splitted the data into 80% training and 20% testing sets.
4. Fed the data into our models.

Models

Model Selection We believed that the GNN model would have better performance than traditional machine learning models under urban contexts. Although traditional ML algorithms can use independent variables to predict dependent variables, the interconnections between different entities were not taken into account. Other factors such as social, economic, and mobility flows play important roles in the formation of cities and neighborhoods. In order to take these connections into account when making predictions, GNN was used to include more information for classification tasks. Other deep learning models like neural networks are designed for Euclidean graphs and may be an optimal choice for our graphs with a spatial character.

Graph Neural Networks (GNNs) are a class of deep learning methods designed to perform inference on data described by graphs. GNNs are neural networks that can be directly applied to graphs, and provide an easy way to do node-level, edge-level, and graph-level prediction tasks.

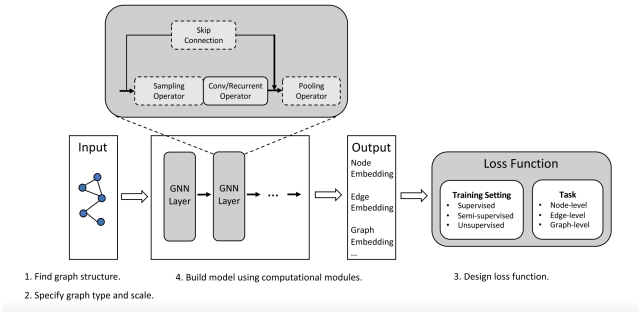


Figure 3: GNN Mechanism (Merritt 2022)

The Layer propagation rule of the GNN is:

$$H^{(l+1)} = \sigma(H^{(l)}W_1^{(l)} + \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}H^{(l)}W_2^{(l)})$$

where H is the l'th neural network layer, A is the adjaceny matrix, D is the diagonal node degree matrix, $W1$, $W2$ are learnable weight matrices initialised as $W1 = 1$, $W2 = 0$ and σ is the activation function double relu.

For our model, we defined:

- 2 GNN layers
- 16 hidden layers
- Relu Activation Function
- Adam Optimizer with decay learning rate start from 0.001

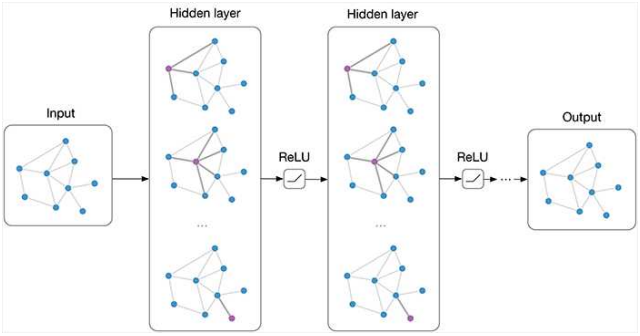


Figure 4: GNN Architecture(Karagiannakos 2020)

Baseline models We trained other models such as Decision tree, Random forest, AdaBoost, SVM and MLP to be our baseline models. Our final models were evaluated compared to the baseline models.

Loss function

We selected cross entropy loss as our loss function.

Cross entropy loss For a multi-class classification task with K classes and true label y and predicted probability distribution p:

$$L = -\sum_{i=1}^{K} y_i \cdot \log(p_i)$$

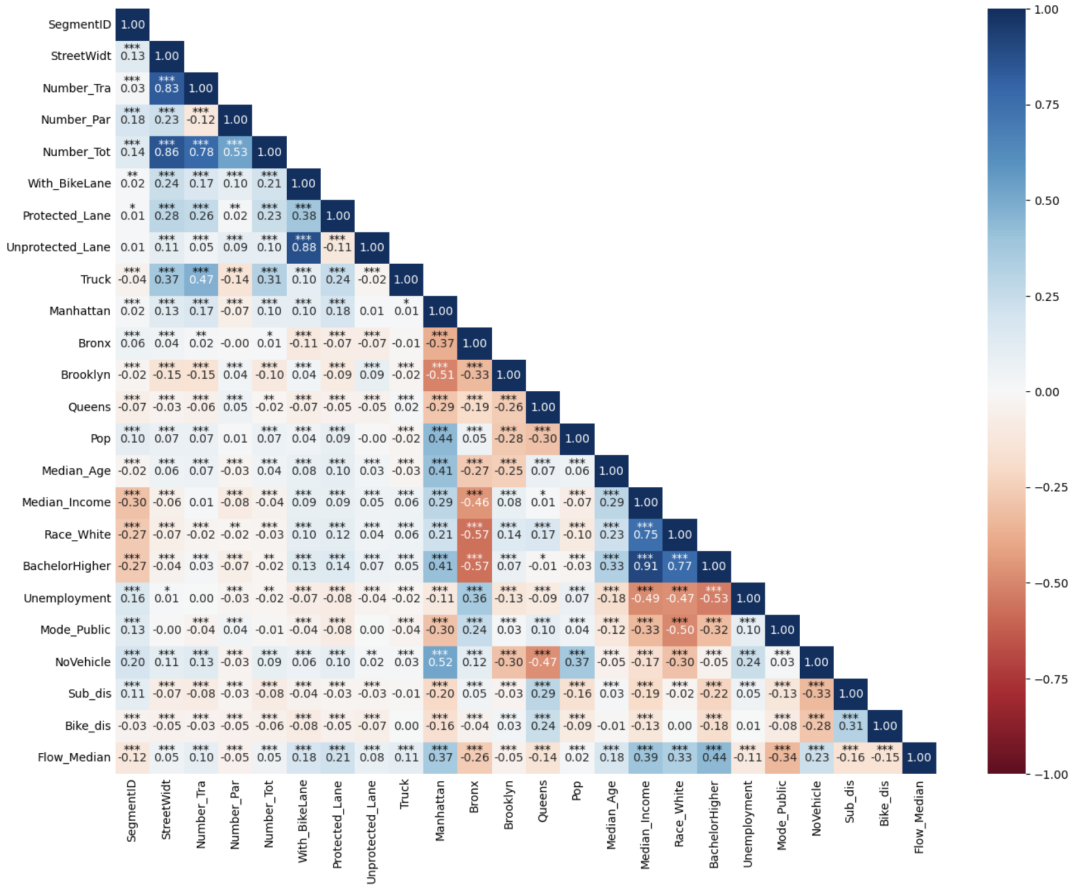


Figure 2: Feature correlation.(*: p-value ≤ 0.05, **:p-value ≤ 0.01, ***; p-value ≤ 0.001)

where y_i is the binary indicator for class i (equal to 1 if the sample belongs to class i and 0 otherwise) and p_i is the predicted probability of the sample belonging to class i.

Result

Table 1 includes performances of our baseline models and our GNN models. While the baseline models achieved validation accuracy around 40%, our GNN model reached a validation accuracy of 63%.

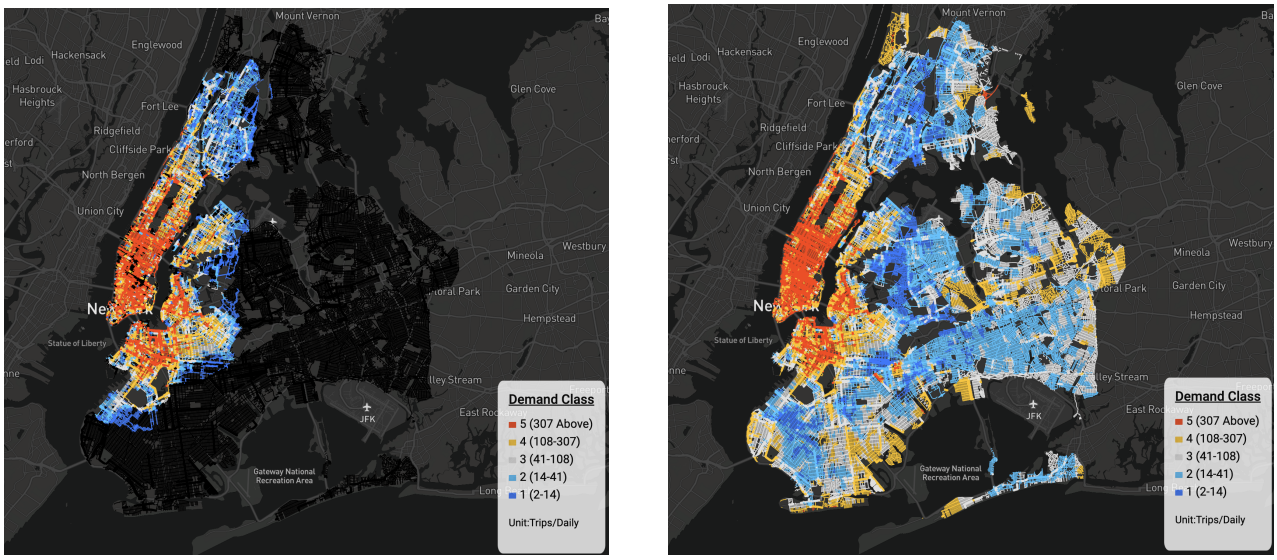
Model	Train Accuracy	Val Accuracy
Decision Tree	45.36%	42.89%
Random Forest	46.57%	42.64%
MLP	39.33%	39.16%
AdaBoost	40.64%	38.79%
SVM	39.04%	39.21%
GNN	64.06%	63.68%

Table 1: Model performance

Figures 5 show our model prediction and our true sample class. It is clear that the two maps matched pretty and both revealed a general trend of the city wise median Citibike flow. The major areas with high median daily Citibike flows are Manhattan, downtown Brooklyn and Long Island City. The higher demand areas also correlates with current Citibike station distribution. In lower demands areas such as Bronx, however, we saw a relatively dense bike station distribution.

Conclusion

This project investigates the potential of using GNNs to predict and understand the street citbike flow, compared to conventional ML and DL models. By applying the GNN model to analyze the impact of social factors between neighborhoods and road conditions on citibike demand, we found strong evidence that GNNs can outperform traditional prediction algorithms. This is because GNNs can easily represent the inherent spatial structure of cities through graphs and handle heterogeneous urban data. The findings from this project can be applied to various urban domains and may inform research and policy on proactive city government response, impact assessment, urban planning, and asset management.



(a) Actual

(b) Predicted

Figure 5: Citibike Median Daily Flow Class

Implications and future suggestions

By referencing the station distribution and our prediction of daily median Citibike flow, we could tell there existed uneven allocation of shared bike resources like Bronx. The results could be useful to suggest the potential shared bike station installation in other areas.

Due to the limitation of computing resources, our GNN model reached memory limit and crashed at our final accuracy of 63.68%. All models were trained using Google Collab Pro GPU environment, so a higher demand for computing resources could lead to better performance.

References

Chen, Y.; Zhang, Y.; Zhang, J.; Zou, J.; and Chen, H. 2017. Predicting Bike Rentals in New York City: A Graph-Based Approach. *arXiv preprint arXiv:1711.07757*.

Karagiannakos, S. 2020. Graph Neural Networks. Accessed on: Insert date accessed.

Merritt, R. 2022. What Are Graph Neural Networks? Accessed on: Insert date accessed.

Zhang, X.; Sun, Y.; Zhang, Y.; and Zhang, J. 2018. Bike Sharing Demand Prediction: A Comparative Study. *arXiv preprint arXiv:1806.04913*.

Zhou, Z.; Hu, Y.; Liu, T.; Wang, J.; Chen, Y.; and Wang, Z. 2018. Traffic Flow Prediction in Bike Sharing Systems: A Deep Learning Approach. *IEEE Access*, 6: 44127–44137.