ABSTRACT
In the realm of transportation network companies (TNCs), the utilization of demand forecasting models and algorithms has revolutionized essential services, such as guiding idle vehicles, efficient carpooling, rate fluctuations, and price surging. In this work, we investigate the effect of varying learning models and input features on the prediction of future demand. Specifically, we extract simple yet key spatial and temporal features of pickups and dropoffs from NYC taxi data, including day of the week, time within the day, and location, and run them through multiple learning models with varying attributes and complexities, such as regression and Long Short-Term Memory (LSTM) models. Our results highlight the capabilities of harnessing accessible and simple features to enhance ride-demand prediction models, paving the way for more inclusive and transparent advancements in urban transportation services.

CCS CONCEPTS
• Information systems → Spatial-temporal systems; • Computing methodologies → Machine learning;

KEYWORDS
pickup predictions, spatial-temporal, learning models, NYC taxi

1 INTRODUCTION
TNCs like Uber have access to different facets of data that exceed what is open-sourced on the internet, and this dependence on inaccessible features renders many applications and algorithms ineffective for the public. Existing research predominantly concentrates on more complex sets of features, including past pickup data, weather conditions, traffic patterns, and passenger behavior, to predict future pickups. This creates a reliance on vast amounts of intricate data that may not be publicly available. The limited availability of ride data has catalyzed our investigation into finding alternative, accessible data sources and APIs.

For our first phase of investigation, we leveraged the NYC Taxi and Limousine Commission dataset, which offers a substantial trove of historical yellow taxi data from 2009 to 2023 [2]. Focusing on the objective of efficient prediction of future pickup patterns in urban areas, we started with the identification of the minimal set of critical features necessary for such predictions, as well as an investigation of spatial and temporal variability of ride patterns within the different zones in NYC.

In addition to simple regression models, we analyzed the impact of various features and their combinations in LTSM models. Our primary objective persists: dissecting readily available ride data to forecast future demand. Rather than using complex inputs, our research assesses whether simple features alone are sufficient in creating machine learning models for predicting future pickups, as we examine the correlation between historical dropoffs and expected ride demand.

2 LEARNING MODELS
A visual analysis of the data across various zones indicated patterns in pickups and dropoffs, with peaks during morning and evening rush hours on weekdays and a peak around midday on weekends. Accordingly, we hypothesized that rides on weekends form a linear relationship and weekdays a polynomial one. We investigated this hypothesis with quadratic and cubic polynomial regressions, finding that the cubic polynomial regression models tended to overestimate the predictions of pickups, whereas the quadratic models approximated well 1.

Since there are non-linear correlations between dropoffs and pickups in weekdays, we chose a deep learning model that specializes in time series and sequence data processing: Long Short-Term Memory (LSTM). The architecture of our LSTM model comprises two consecutive LSTM layers, of 100 LSTM cells each, which are subsequently followed by two fully connected Dense layers, as illustrated in Figure 1.

To determine the optimal hyperparameter configuration, we adopted a grid search approach; we fine-tuned specific hyperparameters, incorporated MinMaxScaling for data normalization, experimented with different batch sizes, and set the number of epochs for training. To enhance the model’s generalization ability, we integrated a dropout function after each LSTM layer to mitigate the impact of outliers. Our dropout rate of 0.20 removes the outliers that account for 20 percent of the predictions.

1 For evaluation, we used multiple error metrics, but we only display R² scores due to space limitation.
We then systematically examined four distinct combinations of whereas weekends displayed similar outcomes with both linear and the deep learning nature of LSTM, we wanted to assess if employ-
ing simpler combinations of features would yield consistent and the training set, with the other 20% designated for testing. Given
for our LSTM model, we arbitrarily allocated 80% of the data to be
3.2 LSTM Models
3.1 Regression Models
We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.

3.3 Discussion

We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.

3.3 Discussion

We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.

3.3 Discussion

We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.

3.3 Discussion

We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.

3.3 Discussion

We note that although LSTM allows for better prediction of future demand, it does not consistently perform well across different zones. Areas that are more central, with a higher and more consistent volume of traffic, were more predictable with our models. The quality of prediction does not only depend on the learning model and the volume of training data available, but it also depends on the spatial and temporal patterns within that data. This emphasizes the importance of studying varying urban dynamics when designing predictive models for spatial data, as regions with different traffic flows can exhibit distinct behavioral characteristics.

In summary, this study allowed us to investigate essential feature and model selection for accessible ride data. In future work, we plan to improve our predictions by considering the spatio-temporal variation in the data and combining learning models to extract relevant model features that would enhance the predictions.