PREDICTING PICKUP DENSITY OF TAXI CABS IN NYC

MATT SHUMWAY, ELI SAMPSON, COLE EDGREN

Abstract. In this paper we examine a large dataset containing information on NYC taxi cabs. With rigorous data augmentation and thorough parameter search, we found that it is possible to predict the number of pickups for a given neighborhood at a given time of day. These findings will allow for taxi services to better optimize their routes.

1. Research question and overview of the data

The main question this project seeks to answer is: Can taxicab pick-up density at a given location and time be accurately predicted? Having a successful predictive model for this problem would provide information on taxicab activity and demand throughout the city. Thus, it is of interest to several groups: city planners, policymakers, taxi operators, as well as other ride-share companies.

There is good reason to believe such a model could be built, as commuting patterns are often habitual. For example, many commute to Manhattan in the morning and home in the evening, with a significant portion regularly using taxis. Similarly, thousands of trips are made to JFK International and LaGuardia airports (LaGuardia notably lacks subway access). These consistent patterns suggest a machine learning model could effectively learn and predict them.

There are many projects similar to ours that have been conducted. Daultan, Raman, and Kindt [Dau23] also created a model for predicting pickup density on locations in New York City. Their project utilized data from the years 2009-2015 which contained exact latitude and longitude coordinates (the taxi organization later replaced exact coordinate data with taxi zones for privacy); they created geographically hashed bins to segment their data, merged their dataset with relevant weather data, and used cyclic encoding to develop a random forest regression model that achieved a 0.95 R^2 value [Dau23]. Breeman [bre23] developed a model that predicted taxi fare, but also performed an in-depth analysis of similar NYC taxi data and created elegant visualizations. In this project, we seek to achieve similar results but without the usage of exact coordinates, a decision we explain in greater detail shortly and also in section 2.

The dataset we used was pulled from the NYC Taxi and Limousine Commission (TLC) [TC24]. In particular, we pulled data on all yellow taxi cab

Date: December 22, 2024.

trips from the beginning of the year 2023 until present (August 2024). Each trip contained the following information (not including blatantly irrelevant features): time of the pickup and drop-off, the passenger count, the distance of the trip, the location ID of the pickup and drop-off, payment type, congestion surcharge, and airport fees. We hypothesized that the most relevant features for this model are simple: (1) the location and (2) the date/time.

One potential weakness of our dataset is that it contains pre-binned location IDs instead of exact coordinates. Each location ID was created roughly based on the NYC Department of City Planning's Neighborhood Tabulation Areas (NTAs), designed to approximate neighborhoods [NYC24]. While this segmentation allows for interpretable results, the varying area of each taxi zone potentially introduces bias in the analysis. This is the only available source of data on NYC taxi cab trips we could identify—every other relevant dataset is derived from the TLC data.

2. Data Cleaning / Feature Engineering

Our dataset initially contained 61,719,218 data points. Since we hypothesized that location and time would be the most important features, we retained only the pickup datetime and pickup location ID. Next, we dropped the NaN values, which only accounted for 5.78% of this filtered dataset. As the percentage of NaN values was relatively small, we felt that this was justified for such a large dataset. Additionally, some unique locations were associated with multiple location IDs. To address this, we replaced redundant IDs with a single representative ID for each location, ensuring consistency in the data.

The target variable, pickup density, is not directly provided in the dataset, so we created it by aggregating data by pickup location and time intervals. This aggregation reduced the dataset from over 58 million points to approximately 38,000, making it manageable for modeling.

To prepare for regression, we converted non-numerical data into numerical representations. Pickup location IDs, treated as categorical data, were replaced with the latitude and longitude of each zone's centroid. This approach leverages spatial continuity while avoiding the high dimensionality of one-hot encoding for 250+ zones. Time features were encoded using cyclical transformations to capture periodic relationships, applying sine and cosine transformations to day and hour labels.

Pickup density exhibited high variance, with downtown Manhattan zones exceeding 44,000 pickups per interval, while less populated areas had fewer than 10. To reduce variance while preserving relative relationships, we applied a log transform to the raw pickup counts.

After feature engineering, the dataset includes six features: latitude, longitude, time sine, time cosine, day sine, and day cosine. These features retain the original data's essential information, enabling effective regression modeling while avoiding high dimensionality.

3. Data Visualization and Basic Analysis

To build a model that predicts pickup density for a given day and hour, it is important to visualize the data to better understand its characteristics and identify any observable patterns.

We were most interested in creating plots/animations visualizing the spatial and temporal relationships of taxi cab pickups. Below we present two figures, one displaying how pickup density changes between days of the week 1a and another displaying how pickup density changes throughout the hours of the day 1b.

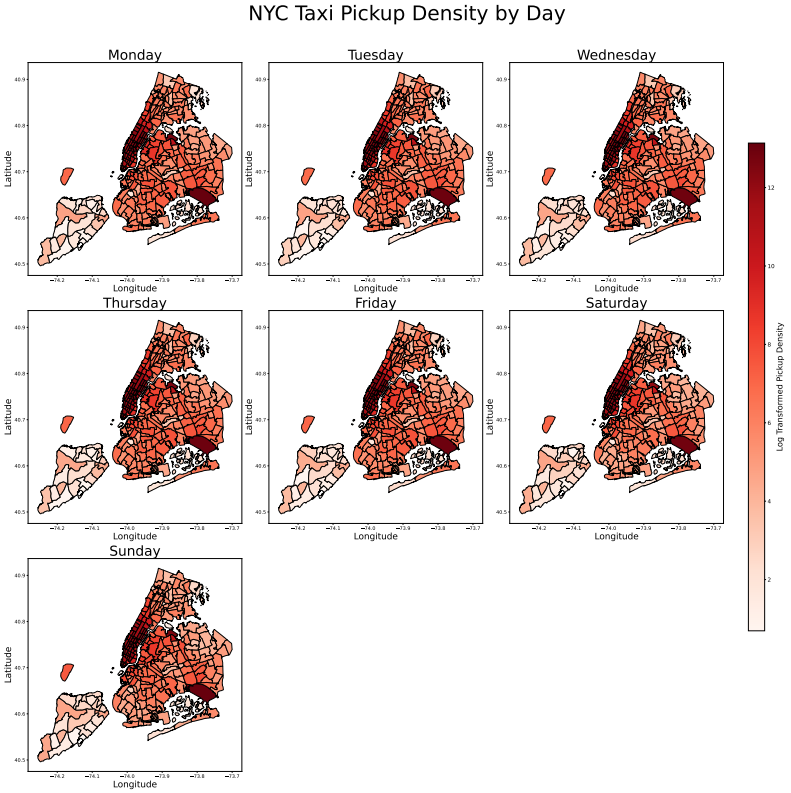
The key findings from such visualizations and analyses are the following. Although there is not significant variance in daily pickup densities in taxi zones across different days of the week, there is significant variance between the hours of the day. Regardless of that, however, JFK airport always seems to be busy - no matter the day or the time. Manhattan is clearly the busiest borough for taxis - the most rudimentary predictive model dictating where to place your fleet would always suggest Manhattan and the JFK airport. Taxi pickups are the most desired between the hours of 6:00 AM and 6:00 PM.

4. Learning Algorithms and In-depth Analysis

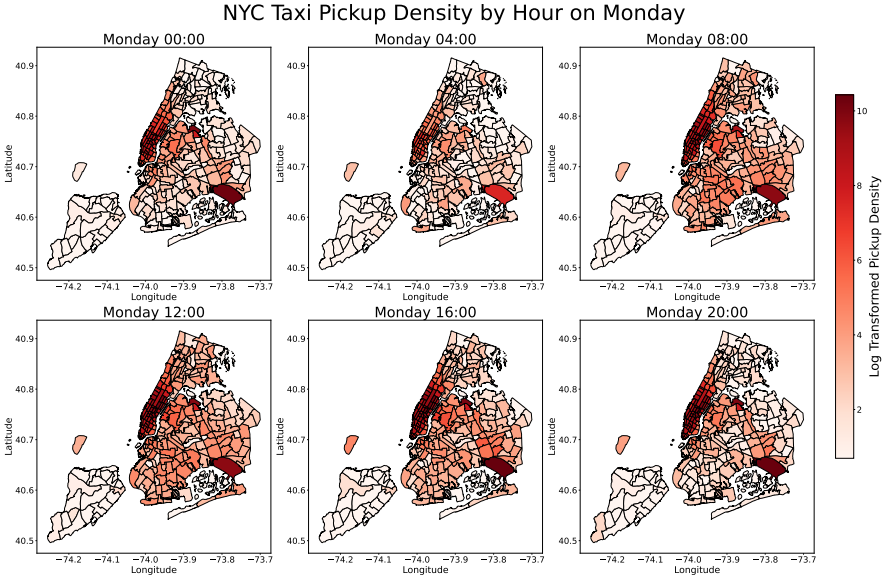
As we are predicting a continuous output variable, we did not run any classification models such as logistic regression or näive bayes. We ran a variety of regression algorithms and tuned our models with the Optuna framework instead of relying on a traditional grid search or random search. Optuna uses Bayesian optimization to iteratively build a probabilistic model of the objective function and selects hyperparameters based on an acquisition function [ASY+19]. This allowed for a more intelligent exploration of the hyperparameter space. We ensured an 80-20 train-test split of our data and used 5-fold cross validation in the search of optimal hyperparameters. For each model, we measured both the R^2 and MSE on the holdout test set: a summary of performance is listed in figure 2.

Based on the performance metrics in Figure 2, tree-based methods significantly outperformed linear regression models. LightGBM achieved the best results, with the lowest MSE and highest R^2, followed closely by our Random Forest and Decision Tree models. The strong performance of tree-based methods likely stems from their ability to naturally capture nonlinear relationships in the data without requiring additional feature engineering. In contrast, linear regression relies on manually designed transformations to model complex nonlinear patterns, which were not applied in this analysis, limiting its effectiveness.

Because of the strong performance of the tree based methods, we feel confident in concluding that our question can be answered positively: it is indeed possible to accurately predict pickup densities of taxi-cabs in NYC in a given interval of time.



(A) Pickup density across days of the week. Even with careful examination, variation in daily pickup densities is very small between days.



(B) Pickup density across hours of the day, plotted in 4-hour intervals for Monday.

FIGURE 1. Exploration of yellow taxi cab pickup densities in NYC taxi zones by day and hour.

Model	R^2 (↑)	MSE (↓)
Linear Regression	0.1711	5.5787
Ridge Regression	0.1711	5.5789
Decision Tree	0.9567	0.2912
LightGBM	**0.9785**	**0.1442**
Random Forest	<u>0.9780</u>	<u>0.1483</u>

FIGURE 2. Performance comparison of various models based on R^2 and MSE (remember MSE compares a logged difference due to our data engineering). Higher R^2 values and lower MSE values indicate better performance. LightGBM achieves the best performance, followed closely by Random Forest.

It is valuable to compare the high R^2 of our XGBoost model to the 0.95 achieved in [Dau23]; however, be very careful in this comparison as there are significant differences in modeling. We attribute this higher R^2 value not necessarily to a better model, but mainly to the fact that our segmentation of NYC zones was more coarse, so there are less zones to predict the density of.

For animations comparing our models predicted pickup densities against true data (what we feel like is the most attractive aspect of this project), please refer to the README file of the github: animation.

5. Ethical Implications and Conclusions

Our project has few ethical concerns. The dataset, collected under the TPEP/LPEP programs, ensures anonymity and includes only trip details, not passenger information. Our model's success may raise privacy concerns or fears of tracking people. However, since the dataset is fully anonymized and contains no passenger information, these concerns are unfounded. Our data is equivalent to publicly observing taxicabs in operation.

We believe our model has minimal potential for harm but recognize it could lead to resource misallocation. By predicting taxi density, it aims to optimize transportation in New York. However, if shared with drivers, many might flock to high-density areas, neglecting lower-density zones. While this could temporarily reduce overall efficiency, we trust the market would eventually rebalance driver distribution.

In order to ensure the responsible use of our model way and to prevent any malicious use, we think that it or similar models should only be used by companies or organizations dedicated the transportation of New Yorkers.

6. Code

Refer to the 'analysis' folder of our github repository for the data and modeling pipelines.

References

[ASY+19] Takuya Akiba, Shotaro Sano, Takeru Yanase, Toshihiko Ohta, and Masanori Koyama. Optuna: A hyperparameter optimization framework. https://optuna.org/, 2019. Accessed: 2024-11-26.

[bre23] breemen. Nyc taxi fare - data exploration. https://www.kaggle.com/code/breemen/nyc-taxi-fare-data-exploration, 2023. Accessed: 2024-11-25.

[Dau23] Samuel Daulton. Nyc taxi data prediction. http://sdaulton.github.io/TaxiPrediction/, 2023. Accessed: 2024-11-25.

[NYC24] NYC Taxi and Limousine Commission. Nyc taxi zones, 2024. Accessed: 2024-11-26.

[TC24] New York City Taxi and Limousine Commission. Tlc trip record data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2024. Accessed: 2024-11-25.