Taxi Trajectories: A Predictive Approach to Forecasting Fare, Duration, and Demand Dynamics in New York City

DATA240 Team 7: Divya Neelamegam, Keerthana Raskatla, Poojitha Venkatram, and Sreenidhi Polineni

1. Motivation

The motivation for our project "Taxi Trajectories" stems from the potential to significantly enhance the operational efficiency and service quality of taxi systems in New York City through predictive analytics. The NYC taxi system facilitates over 240 million trips yearly and is integral to the urban economy by improving passenger experiences. "Taxi Trajectories" leverages predictive analytics to comprehensive 2023 taxi data, to boost operational efficiency and enrich passenger experiences. Urban planners can use our insights for better zoning and transport strategies, while taxi drivers benefit from predictive data on passenger hotspots and optimal timings, enhancing earnings and reducing wait times. This approach not only improves current smartphone-based taxi services by providing deeper behavioural insights but also lays the groundwork for future transportation innovations, offering significant benefits for the community and the environment.

2. Background

The project leverages the NYC Taxi Dataset to address the unique transportation challenges of New York City, a densely populated urban environment where taxis are integral to daily mobility. This dataset includes detailed records of taxi trips, such as pickup and drop-off locations, times, distances, and fares, offering a rich foundation for advanced data analysis. Utilizing big data and machine learning technologies, the project focuses on time series analysis and predictive modelling to forecast taxi fares, trip durations, and demand patterns. These insights are vital for taxi operators optimizing fleet distribution, city planners engaged in traffic and infrastructure management, and businesses dependent on reliable transportation data. Set against the backdrop of enhancing urban transportation's sustainability and efficiency, this project aims to mitigate traffic congestion and reduce emissions, contributing to the city's environmental and operational goals.

3. Literature Review

The New York City Taxi Dataset has been extensively utilized across various research fields, demonstrating its significant value in enhancing urban planning, transportation forecasting, and economic analysis. This dataset serves as a foundational tool in studies like "Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation," where sophisticated hybrid modelling techniques are employed. In this study, authors combine Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) with Mixture Density Networks (MDN) to accurately forecast taxi demand, capturing complex nonlinear relationships and temporal dependencies. Moreover, they utilize an ensemble of machine-learning algorithms for precise fare prediction, optimizing city transportation efficiency and enhancing user satisfaction. Similarly, Rathore, Bhawana, et al. (2024) in "Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models," apply a novel clustering mechanism for fare prediction, adjusting pricing strategies dynamically in response to urban transit demands. Poongodi, M., et al. (2021) explore the use of Multi-Layer Perceptrons (MLP) and XGBoost in "New York City taxi trip duration prediction," effectively managing taxi dispatch and route planning to avoid congestion. Hou, Emily, and Kimberly Hsueh (2023) in "Back to the Past: Predicting Expected Ride Demand

from Previous Dropoffs and Beyond," utilize spatial and temporal data to predict future demand, enhancing taxi deployment efficiency. Additionally, Deri, Joya A., and José MF Moura (2016) in "Taxi data in New York City: A network perspective," employ graph Fourier transform techniques for spectral analysis of taxi trajectories, identifying critical traffic patterns and co-behavioral dynamics within Manhattan. Collectively, these studies showcase the potential of data-driven techniques to substantially improve forecasting and operational efficiencies in urban transportation systems.

4. Methodology

4.1 Data Collection

NYC Yellow Taxi Trip Records: The data for this study was obtained from the NYC Taxi & Limousine Commission's official website, covering all of 2023's Yellow Taxi Trip Records as detailed in *Figure 1*. This dataset contains over 38 million trips, averaging more than 100,000 daily rides in New York City. It includes essential details such as timestamps, locations for pickups and drop-offs, distances, fares, and additional charges, enabling a comprehensive analysis of taxi service trends and spatial-temporal demand patterns in NYC.

Figure 1: Attributes of dataset

```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'RatecodeID', 'store_and_fwd_flag',
       'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra',
       'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge',
       'total_amount', 'congestion_surcharge', 'airport_fee'],
      dtype='object')
```

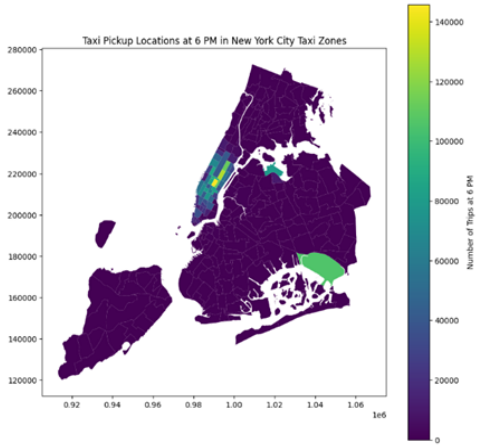
Taxi Zone Shapefile: The study also utilizes the Taxi Zone Shapefile from the TLC to define NYC taxi zones, each marked with unique identifiers like zone name and borough. This shapefile is crucial for precise spatial analysis, allowing the study to map taxi trip data to specific locations, thereby enhancing insights into citywide traffic patterns and demand dynamics. Combining trip data with the zone shapefile offers a detailed view of fare, duration, and traffic trends across New York City throughout the year.

4.2 Data Preprocessing and Exploration

The data preprocessing for the Taxi Trajectories project transforms raw NYC taxi data and geographical information into a structured format for analysis. The `NYCDataProcessor` class standardizes data by converting column names to lowercase and renaming inconsistent elements, such as changing 'airport_surcharge' to 'airport_fee'. Unnecessary columns and rows with missing critical values are removed to refine the dataset. Spatial data integration is crucial, as it merges trip data with NYC taxi zones' shapefile, enhancing each trip with the geographical coordinates of pickup and dropoff locations. This process not only improves data quality but also supports advanced spatial analyses and visualizations, such as the distribution of taxi pickups depicted in *Figure 2*.

Temporal feature engineering enhances the Taxi Trajectories project by extracting and refining time-based attributes from trip timestamps, such as the day of the week, hour, and month for pickups and drop-offs.

Figure 2: The spatial distribution of taxi pickups at 6 PM across New York City taxi zones



These features are further categorized into time segments (Morning, Afternoon, Evening, and Late Night) to analyze variations in taxi usage. These insights are crucial for modelling time-based fare strategies and predicting demand fluctuations, as depicted in *Figure 3* and

Figure 4, which show taxi usage variations and hourly trip fluctuations, respectively.

Figure 3: Day of the week analysis

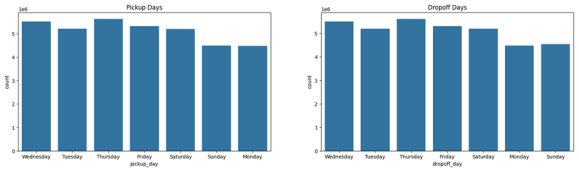
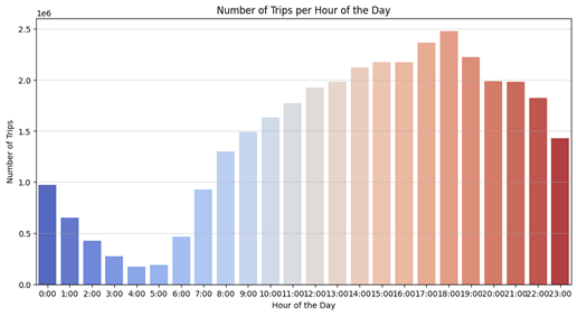
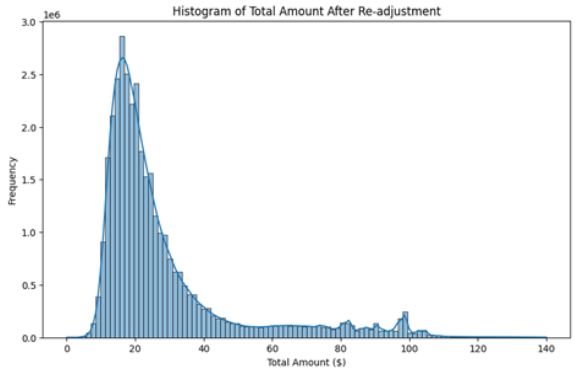


Figure 4: Hourly Distribution of Taxi Trips



In the preprocessing phase for the Taxi Trajectories project, trips with over six passengers were excluded to maintain a focus on typical taxi usage. Additionally, trips originating or ending outside a predefined New York City bounding box were removed for geographical accuracy. The exploratory data analysis produced visuals like *Figure 5*, a histogram showing fare distributions and identifying outliers. This phase also involved robust outlier handling methods, such as z-scores and interquartile range (IQR), for metrics like trip duration, distance, and fare, ensuring the data's integrity and relevance for further analysis.

Figure 5: Histogram of total amount



Trip distances are calculated using the Haversine formula, which provides a measure of the great-circle distance between two points, an essential factor in fare calculations and route optimization. Furthermore, the direction of travel is computed and added as a feature, shedding light on the directional flow of traffic and potentially informing traffic management strategies. *Figure 6* shows the features after Preprocess. These elaborate preprocessing steps are meticulously implemented to ensure that the data is primed for in-depth analysis, setting a strong foundation for accurate and insightful analytical outcomes.

Figure 6: Attributes after Preprocessing

```
full_df.columns
```

```
Index(['VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime',
       'passenger_count', 'trip_distance', 'store_and_fwd_flag',
       'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra',
       'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge',
       'total_amount', 'airport_fee', 'pickup_latitude', 'pickup_longitude',
       'dropoff_latitude', 'dropoff_longitude', 'direction', 'pickup_datetime',
       'dropoff_datetime', 'pickup_day', 'dropoff_day', 'pickup_day_no',
       'dropoff_day_no', 'pickup_hour', 'dropoff_hour', 'pickup_month',
       'dropoff_month', 'pickup_timeofday', 'dropoff_timeofday',
       'distance_miles', 'trip_duration'],
      dtype='object')
```

4.3 Model

i) Duration Prediction in Taxi Trajectories

Accurate duration prediction is essential for improving fleet management and passenger satisfaction in urban transport. By employing machine learning models, we aim to enhance our understanding of trip dynamics and provide more accurate forecasts, aiding operational and strategic decisions in taxi services.

Linear Regression Model: We chose Linear Regression for its simplicity and interpretability, assuming a linear relationship between trip duration and various features. Implemented using the `LinearRegression` class from `sklearn.linear_model`, this model is straightforward, focusing on core parameters without complex hyperparameters, ensuring quick computations and easy interpretation.

XGBoost Regression Model: The XGBoost regression model was used to handle more complex datasets effectively, capturing non-linear relationships through an ensemble of decision trees. Configured to minimize squared errors with `reg: squared

error` and optimize RMSE, this model includes tuning options like learning rate and tree depth, allowing for precise adjustments to enhance prediction accuracy.

ii) **Time Series Analysis**

Utilizing the ARIMA (AutoRegressive Integrated Moving Average) model for time-series forecasting on this dataset offers significant advantages. ARIMA excels in identifying and leveraging seasonal patterns and cyclic behaviours, which are prevalent in taxi demand due to changes in tourism, business activities, and local events. Implementing this model allows for accurate predictions of demand fluctuations, facilitating superior fleet management and operational effectiveness. Moreover, the flexibility of the ARIMA model to integrate past anomalies ensures robust adaptability to the distinctive dynamics of urban transportation systems. This leads to optimized resource deployment and enhanced service dependability for both taxi drivers and passengers.

iii) **Demand Forecasting**

The NYC taxi demand forecasting effort began with intensive data preprocessing and feature engineering to develop robust forecasting models. Datetime and geographic data from pickup and dropoff locations were retrieved from the dataset. DateTime features to capture temporal dynamics, autoregressive features to include past taxi pickups (PU_count), and rolling statistics to smooth oscillations and find longer-term patterns improved the dataset's prediction power. Correlation analysis was crucial for handling multicollinearity. Features like 'total_amount', 'ewma_24h', and 'lag_2d' were eliminated due to large intercorrelations found in correlation plots of the newly developed features. Ensuring that the remaining attributes provided independent and relevant information for more accurate and dependable predictions enhanced the model. Linear Regression was chosen for its simplicity and interpretability, revealing how predictor factors affect taxi demand. This computationally efficient model is ideal for large datasets and rapid iterations and serves as a baseline for performance comparison. However, Random Forest was used to capture complicated, non-linear correlations and interactions between factors like time and location, which are crucial to demand forecasting. This model's robustness to overfitting and its capability to provide intrinsic evaluations of feature importance made it an invaluable part of the methodology.

iv) **Fare Prediction in Taxi Trajectories**

In the realm of NYC taxi services, precise fare prediction is pivotal for operational efficiency and passenger satisfaction. Our approach to fare prediction harnesses the power of machine learning, employing three distinct models: Linear Regression, XGBoost Regression, and Random Forest Regression. Linear Regression forms the foundation of our model framework, valued for its simplicity and interpretability. By assuming a linear relationship between fare amount and key features such as distance travelled, tolls, and time of day, we leverage the straightforward parameter estimation of the `LinearRegression` class from `sklearn.linear_model`, enabling efficient computations and easy interpretation. Complementing Linear Regression, XGBoost Regression is introduced to capture non-linear relationships inherent in NYC taxi data. By minimizing squared errors with the `reg: squared error` objective and optimizing for Root Mean Squared Error (RMSE), the XGBoost model offers tunable parameters like learning rate and tree depth, enhancing prediction accuracy for fare amounts. Further enhancing prediction accuracy, Random Forest Regression is incorporated into our model ensemble. Leveraging the `RandomForestRegressor` class from `sklearn.ensemble`, this technique constructs multiple decision trees during training and aggregates predictions, effectively capturing complex relationships among features while mitigating overfitting. Through the synthesis of these models, we aim to provide accurate fare estimates, facilitating informed decision-making and

optimizing operational efficiency in NYC taxi services.

5. Experiment and Result

i) Duration Prediction in Taxi Trajectories

In the evaluation of our models for predicting taxi trip durations, we employed Linear Regression as the baseline and XGBoost for advanced predictions, using Root Mean Square Error (RMSE) for assessment. The Linear Regression model achieved an RMSE of 4.312 on the validation set and 4.045 on the test set, indicating reasonable generalization but potential overfitting issues. These results are detailed in *Table 1*, which summarizes the RMSE scores across both the validation and test datasets for each model.

Table 1: Attributes after Preprocess

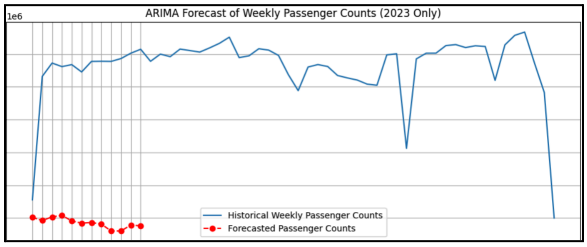
Model	Performance Evaluation
Linear Regression	Validation Root Mean Square Error (RMSE) : 4.312 Test Root Mean Square Error (RMSE) : 4.045
XGBoost	Validation Root Mean Square Error (RMSE) : 23.778 Test Root Mean Square Error (RMSE) : 18.67

In contrast, the XGBoost model demonstrated superior performance, with an RMSE of 23.778 in validation and 18.67 in testing, nearly halving the error compared to Linear Regression. This highlights Linear Regression model's ability to effectively capture complex data interactions, making it a robust choice for enhancing operational efficiency in taxi service management.

ii) Time Series Analysis

The SARIMAX(4, 1, 2)x(2, 0, [], 52) model applied to the 2023 weekly passenger counts in New York City displays robust predictive capabilities, with significant autoregressive and moving average parameters indicating the model's strong ability to leverage historical data patterns. The significant seasonal component at L52 confirms the model's effectiveness in capturing annual trends, crucial for forecasting in urban transport contexts. While the model's high log-likelihood suggests an effective fit to the observed data, the sizeable values of AIC and BIC may prompt further refinement to streamline the model without compromising on predictive accuracy. Diagnostic plots show that the residuals are mostly stable, although some deviations from normality are noted, which are common in real-world data scenarios and do not drastically diminish the model's utility. The forecast plot as shown in *Figure 7* successfully captures the overall passenger trends with some minor deviations, suggesting that with slight adjustments for outlier management and enhanced seasonal modeling, the SARIMAX model could offer even more reliable forecasts.

Figure 7: ARIMA Forecast of Weekly Passenger Counts



iii) Demand Forecasting

The performance of the two models—linear regression and random forest—was tested against a dataset characterized by significant variability in taxi pickups: mean 3,997, standard deviation 2,322, and a range from 0 to 10,114. This shows a model with commendable generalization, as a linear regression one: 41.14 and 32.39 for RMSE and MAE, respectively, on the test set, which are scores way below the standard deviation and point out that our model managed to accurately capture the central tendencies of the data without overfitting. This was proven through the model performing well on both the training and testing sets. The random forest model also showed very good training results, with an RMSE of 40.77 and an MAE of 25.10, but with bad generalization, for which the RMSE and MAE were 156.91 and 99.53, respectively, in the test set, suggesting very strong overfitting to the training data's noise and extremes. This underperformance during the test phase indicates the fact that

the model failed to generalize from the training data set to the new ones. That serves as evidence of the applicability of the linear regression model for the real-world situation, where predictions are to be stable and reliable in a situation of highly variable environments in the NYC taxi service.

iv) **Fare Prediction**

In our evaluation of models for predicting taxi trip durations, we established Linear Regression as the baseline and implemented Random Forest for more advanced predictions. To assess model performance, we utilized Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The Linear Regression model achieved a Validation MAE of 0.8328846904014493 and a Validation RMSE of 1.2074705626331743, with corresponding Test MAE and Test RMSE of 0.8338656177650092 and 1.2091696131547127, respectively. These results are summarized in *Table 2*, detailing the RMSE scores across both validation and test datasets for each model.

Table 2: RMSE and MAE for the models

Model	Validation MAE	Validation RMSE	Test MAE	Test RMSE
Linear Regression	0.8329	1.2075	0.8339	1.2092
Random Forest after PCA	0.2391	0.7798	0.2397	0.7788

The Random Forest model demonstrated significantly improved performance after Principal Component Analysis (PCA). With a Validation MAE of 0.23913172095590923 and a Validation RMSE of 0.7798225477287793, along with Test MAE and Test RMSE of 0.23969677970391715 and 0.7788123646389945 respectively. This represents a remarkable reduction in error of approximately 70% compared to Linear Regression.

This notable enhancement underscores Random Forest's capability to effectively capture intricate data interactions, positioning it as a robust choice for enhancing operational efficiency in taxi service management.

6. Discussion and Future Work

For future developments, here are some of the following enhancements that could be explored. Firstly, implementing dynamic ride-sharing could facilitate the prediction of optimal pairings between passengers who share similar destinations or travel routes, encouraging cost-sharing while reducing vehicle emissions. Secondly, managing an autonomous taxi fleet using real-time data could revolutionize how taxis are dispatched. By analyzing traffic conditions, passenger demand, and fare dynamics, autonomous taxis could be allocated more efficiently, optimizing energy use and alleviating congestion. Thirdly, developing augmented reality (AR) navigation systems for taxi drivers could significantly enhance operational efficiency. These systems would use predictive analytics to display upcoming demand hotspots, anticipated traffic congestion, and optimal routes directly from the driver’s field of view, thereby improving response times and service quality. Additionally, integrating weather forecast-based predictions is a potential area for future exploration. During the initial phase of the project, attempts to implement this feature encountered challenges due to the lack of an appropriate dataset specifically for the year 2023 that aligns with our existing taxi data. Incorporating weather data could enable the analysis of how rain or harsh weather conditions impact taxi demand and fare adjustments. This aspect of predictive analytics would provide valuable insights into demand fluctuations during adverse weather, allowing for more precise service planning and fare setting.

7. References

1. Chou, K. S., Wong, K. L., Zhang, B., Aguiari, D., Im, S. K., Lam, C. T., ... & Pau, G. (2023). Taxi Demand and Fare Prediction with Hybrid Models: Enhancing Efficiency and User Experience in City Transportation. *Applied Sciences*, *13*(18), 10192.
2. Rathore, B., Sengupta, P., Biswas, B., & Kumar, A. (2024). Predicting the price of taxicabs using Artificial Intelligence: A hybrid approach based on clustering and ordinal regression models. *Transportation Research Part E: Logistics and Transportation Review*, *185*, 103530.
3. Poongodi, M., Malviya, M., Kumar, C., Hamdi, M., Vijayakumar, V., Nebhen, J., & Alyamani, H. (2022). New York City taxi trip duration prediction using MLP and XGBoost. *International Journal of System Assurance Engineering and Management*, 1-12.
4. Hou, E., & Hsueh, K. (2023, November). Back to the Past: Predicting Expected Ride Demand from Previous Dropoffs and Beyond. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*(pp. 1-2).
5. Deri, J. A., & Moura, J. M. (2015, November). Taxi data in New York City: A network perspective. In *2015 49th asilomar conference on signals, systems and computers* (pp. 1829-1833). IEEE.