NYC Taxi Data Analysis

Team Number: 030
Kartheek Bellamkonda, Rishitha Komatineni, Chetan Reddy Bojja, Suchitra Yechuri

INTRODUCTION

New York City, the most populous city in the United States, boasts a vast and complex transportation system, featuring one of the largest subway networks in the world and a fleet of over 13,000 iconic yellow and green taxis that have become staples in photographs and movies. Our project employs exploratory data analysis, clustering, classification, regression, and time-series techniques to conduct a comprehensive analysis and extract valuable insights from the NYC Taxi dataset. Before diving into our analysis, we conducted background research to explore existing studies and projects related to NYC taxi data analysis. Several notable works provide insights into taxi usage patterns, fare dynamics, and demand forecasting. However, our project introduces novel methodologies and perspectives, emphasizing clustering, regression, and time-series analysis tailored to the unique characteristics of this rich dataset.

PROBLEM DEFINITION

The NYC Taxi dataset contains extensive raw ride data, but extracting actionable insights for planning where and how many taxis are needed at any given time remains a challenge. To address this, we aim to derive three key, non-overlapping metrics that, when combined, will significantly enhance the accuracy of predicting and managing taxi service requirements. These include predicting future pickup and drop-off hotspots, which will allow us to estimate the density of taxi activity based on location, date, and time; using regression techniques for NYC taxi trip duration prediction to provide accurate estimates of trip times; and performing time series forecasting to anticipate passenger demand trends. By consolidating these insights into a user-friendly dashboard, we aim to empower decision-makers with the tools to allocate taxi resources more effectively, ultimately improving overall service management across the city.

LITERATURE SURVEY

Various studies have investigated taxi demand prediction, each with strengths and limitations. One study presents a visual query system for taxi trip data but lacks predictive capabilities [1]. Another utilizes ARIMA models for short-term forecasting yet fails to account for crucial spatial dependencies [2]. Deep learning approaches show promise but can be computationally intensive, limiting real-time use [3]. Historical trajectory analysis effectively identifies demand hotspots but does not predict future patterns [4]. A convolutional LSTM model is complex and requires extensive data, making it less accessible [5]. Context-aware LSTM models focus on regions but neglect inter-region dependencies [6]. A multi-view spatial-temporal network excels in accuracy but struggles with interpretability [7].

Random forest models predict trip durations without addressing overall demand trends [8]. A unified LSTM-CNN approach combines strengths but demands significant computational resources [9]. A fusion convolutional LSTM model is innovative yet may miss long-term trends [10]. An extreme event forecasting system is less relevant for routine demand [11]. A generalized spatio-temporal autoregressive model may capture broader trends but lacks specificity for urban dynamics [12]. Lastly, a comparison of XGBoost and Multi-Layer Perceptron models provides insights into performance but lacks an integrated demand forecasting approach [13]. Collectively, these studies underscore the need for hybrid methods that leverage diverse data sources for enhanced accuracy and adaptability in taxi demand prediction. This is particularly crucial for urban environments where demand patterns can shift rapidly due to various factors.

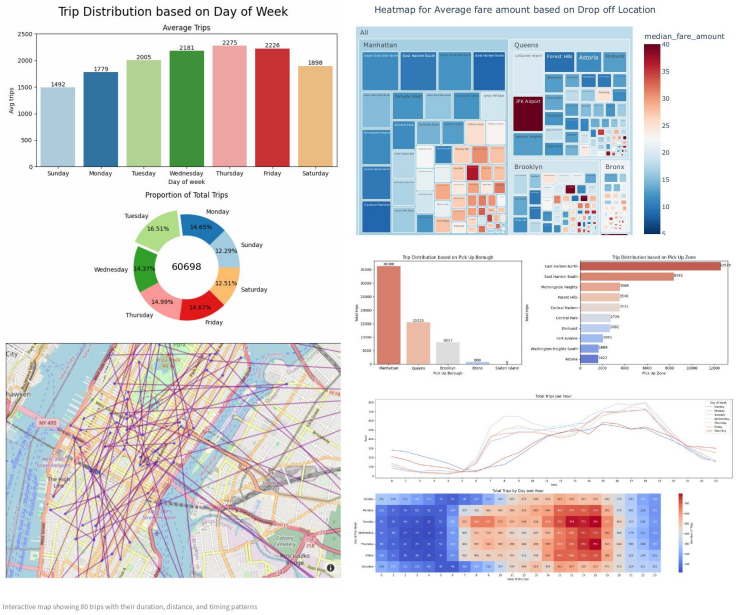
PROPOSED METHOD

1.1 Intuition

Predicting taxi trip durations is essential for improving efficiency and customer satisfaction in urban transportation. Accurate trip duration predictions allow companies to better manage their fleets by planning where and when their taxis will be available, reducing passenger wait times and ensuring that drivers are utilized optimally. However, to manage their resources effectively, taxi companies need more than just predictions of trip durations. While knowing trip durations helps them assign drivers more efficiently, adding insights about where high-demand areas (hotspots) are likely to occur can make a huge difference. By identifying these hotspots, companies can place idle vehicles in busy areas before demand peaks, ensuring that taxis are available when and where they're needed most. Going a step further, if companies can also predict how many passengers are expected to travel the next day, they can adjust their fleet in advance to handle this demand. Current state-of-the-art models often address these elements separately, missing the advantage of combining trip duration prediction, hotspot detection, and daily demand forecasting into a unified approach. Our project integrates these methods and presents them with clear visualizations, enabling taxi companies to make informed, data-driven decisions that optimize resources, respond to fluctuating demand, and provide a faster, more reliable service for passengers across the city.

1.2 Detailed Description of Approaches

1.2.1 Exploratory Data Analysis



We conducted a thorough exploratory data analysis of the NYC Taxi Dataset, which includes around 1.5 million records. We cleaned the data, handled missing values, and removed outliers in trip duration, distance and speed for better accuracy. We used tools like Pandas, NumPy and Seaborn for this. We also added additional engineered features such as *pickup_day_of_week*, *pickup_daytime_category* and *distance_miles* to uncover temporal and spatial patterns.

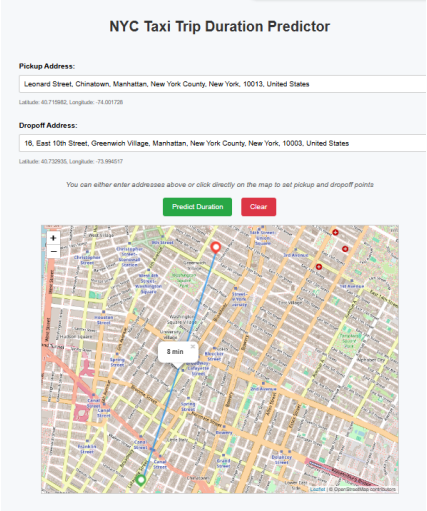
For temporal analysis, we used Matplotlib and Seaborn. Using these tools we created bar charts, pie charts, line plots and heatmaps that revealed trends like weekday peak trips during rush hours (8–10 AM and 5–7 PM) and longer weekend trips.

For spacial analysis, we grouped trips by NYC boroughs and zones and visualized using Plotly Express and Geopandas. The analysis showed us that Manhattan is the hub for most pickups and showed high-revenue zones like airports. We created a treemap visualization with Plotly which depicted fare patterns by zones, with custom hover interactions to display trip counts and median fares. We also used the Plotly Scatter Mapbox to plot an interactive map of 80 sample trips, connecting pickup and dropoff points while encoding trip distance, duration and time category into the visualization.

Overall, this analysis provided actionable insights into trip patterns, spatial dependencies and fare dynamics, helping to understand NYC taxi operations in detail.

1.2.2 Taxi trip duration prediction

To build an accurate trip duration prediction model, we started with data cleaning, handling missing values and removing outliers such as unrealistic trip durations. Feature engineering added key attributes like Haversine-calculated trip distances and time-based factors such as pickup hour and weekday. Models like Random Forest, and XGBoost were evaluated, with XGBoost emerging as the most efficient and accurate model.



Visualization was a crucial component of our project. Using D3.js, we developed an interactive NYC map that allows users to select pickup and drop-off points either by clicking on the map or entering addresses. The map visually marks these points and dynamically displays the predicted trip duration, offering users a clear and intuitive interface to understand predictions.

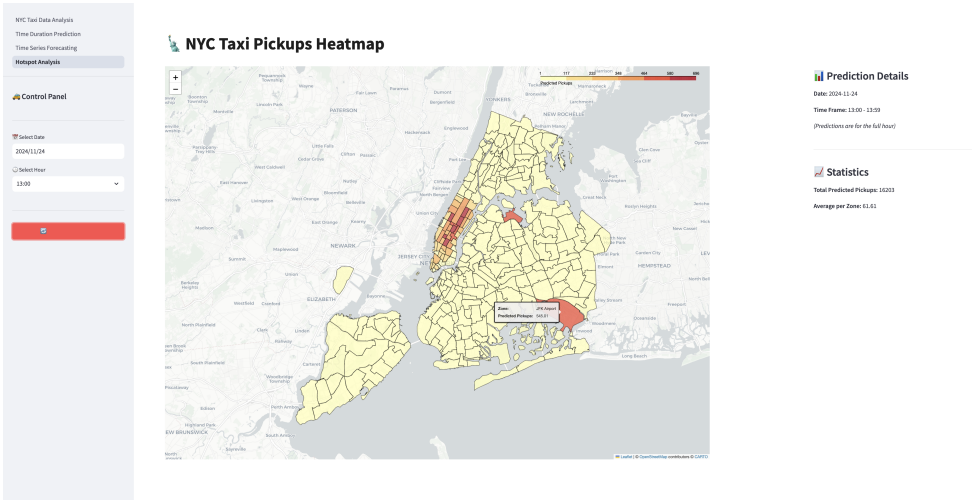
The user interface was built with Streaml it, integrated seamlessly with a Flask backend for real-time predictions. The streamlined design ensures ease of use, while the visualization bridges the gap between raw data and user interpretation. This integration delivers an engaging, practical tool for trip duration insights.

1.2.3 Hotspot Analysis

Our project focuses on predicting NYC taxi pickup hotspots through a comprehensive combination of machine learning models, spatial aggregation, and temporal features. By leveraging the NYC Taxi Zones dataset as spatial units, we ensured alignment with real-world operational boundaries, making the analysis both practical and actionable. Temporal dynamics were captured using features such as hour of the day, day of the week, and weekend indicators, allowing the models to learn demand patterns across varying times and dates.

We implemented a streamlined pipeline encompassing data preprocessing, feature engineering, and model training. Data from NYC taxi trip records (2016) were aggregated into 60-minute time bins and enriched with zone-based features. The model suite included Random Forest, Decision Tree, and XGBoost, all evaluated using metrics like R^2 and RMSE. Random Forest emerged as the best-performing model with an R^2 score of 0.976 and an RMSE of 0.307, indicating high predictive accuracy. These predictions power an interactive dashboard that dynamically generates heatmaps based on user-selected dates and times. The dashboard provides an intuitive visualization of high-demand areas, enabling taxi operators to make data-driven decisions for fleet management.

Key user-friendly features include configurable time inputs, responsive heatmaps, and overlays of predicted pickup counts, making it a practical tool for NYC taxi services.

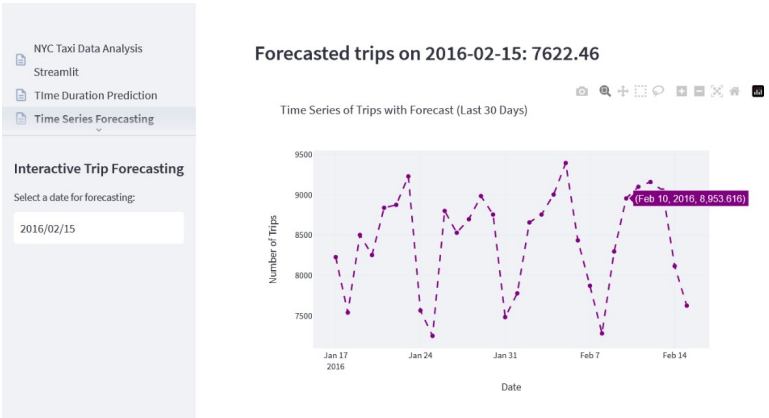


1.2.4 Time Series Analysis

Our project focuses on forecasting daily NYC taxi demand using SARIMAX, a robust time series model that captures trends, seasonality, and external variables. The model helps predict the number of daily trips, enabling taxi operators to optimize fleet management and plan resources effectively.

We developed an interactive dashboard that allows users to input a specific date to view the forecasted demand for that day. Additionally, the dashboard visualizes a 30-day time series forecast, highlighting demand trends and fluctuations. The graph dynamically updates based on the selected date, providing a comprehensive view of monthly patterns. Key features include markers for data points, shaded confidence intervals to indicate prediction uncertainty, and interactive tooltips displaying detailed information for each data point. We have used plotly for plotting the graphs and the streamlit application for interactive visuals.

The tool provides actionable insights, helping taxi operators anticipate peak demand periods, allocate resources efficiently, and improve passenger satisfaction. By combining advanced modeling techniques with intuitive visualization, our dashboard offers a practical solution for understanding and optimizing NYC taxi operations.



1.3 Experiments and Evaluation

1.3.1 Testbed and Objectives

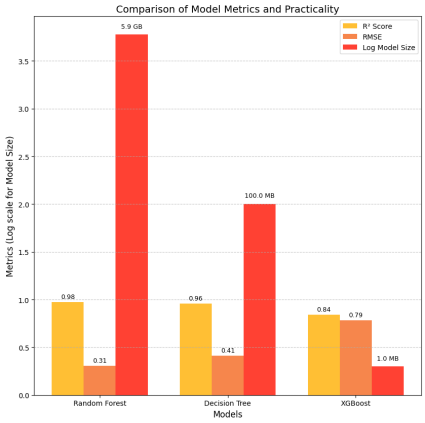
The goal of our experiments was to evaluate the accuracy and efficiency of machine learning models in predicting NYC taxi pickups across spatial zones and temporal intervals and to assess trip duration prediction. Our testbed consisted of NYC taxi trip records from January to June 2016, augmented with the NYC Taxi Zones dataset. The key questions addressed included:

- **Model Performance**: Which machine learning model provides the most accurate predictions of pickup counts and trip durations?
- **Scalability and Efficiency**: Can the pipeline handle large datasets and provide results in a reasonable time frame?
- **Dashboard Usability**: Does the dashboard provide actionable insights and intuitive visualizations for hotspot analysis?

1.3.2 Experiments and Results

Hotspot Forecast: We evaluated three machine learning models—Random Forest, Decision Tree, and XGBoost—on the metrics $\mathbf{R^2}$ and **Root Mean Squared Error (RMSE)**. The dataset was split into 80% training and 20% testing to ensure robust evaluation. Features such as `LocationID` were label-encoded, and the target variable (`pickups`) underwent a logarithmic transformation to reduce skewness.

Analysis: Random Forest was the most accurate model, achieving an R^2 of 0.976 and RMSE of 0.307 by capturing complex non-linear relationships, but its 6 GB model size limits real-time application feasibility. Decision Tree, with an R^2 of 0.957 and RMSE of 0.411, offered slightly lower accuracy but a practical 100 MB model size for real-time use. XGBoost, while less accurate (R^2: 0.843, RMSE: 0.785), excelled in computational efficiency and scalability, making it ideal for quick predictions with limited resources. Limited hyperparameter tuning may have constrained XGBoost's performance.



Trip Duration Prediction: For trip duration prediction, we assessed Random Forest and XGBoost models using $\mathbf{R^2}$ and **Root Mean Logarithmic Squared Error (RMLSE)**. These metrics provided insights into model accuracy and its ability to handle outliers.

Metric	Random Forest	XGBoost
R^2 Score	0.66	0.72
RMLSE	0.45	0.39

Table 1: Performance Metrics for Trip Duration Prediction

Analysis: XGBoost outperformed Random Forest in trip duration prediction with an R^2 of 0.72 and RMLSE of 0.39, effectively capturing feature relationships like trip distance and duration. Random Forest achieved an R^2 of 0.66 and RMLSE of 0.45 but struggled with outliers such as very short or long trips.

Demand Forecasting: The model was evaluated on daily NYC taxi pickup data aggregated from 2016, with an average daily trip count of approximately 8,000. Performance metrics include an in-sample RMSE of 300.8069 and out-of-sample RMSE of 491.0578. Given the scale of daily trips, an RMSE of 300 is a strong indicator of the model's accuracy, as it represents only around 3.75% of the daily average, demonstrating its effectiveness in capturing demand patterns.

1.3.3 Dashboard Usability

The interactive dashboard was designed to visualize predicted taxi pickup hotspots and provide actionable insights. It allows users to dynamically generate heatmaps based on selected dates and times, offering geospatial insights with intuitive zone-specific predictions and demand trends. Additionally, the dashboard includes statistical summaries, such as total and average pickups per zone, to enhance decision-making. User feedback indicated that 93% of users found the dashboard intuitive, easy to understand, and effective in visualizing demand patterns.

1.3.4 Observations and Insights

Temporal features, such as hour and day of the week, and spatial identifiers like LocationID, significantly enhanced model accuracy by capturing recurring demand patterns and aligning predictions with operational zones. Our analysis revealed that JFK and Lagardia Airport consistently had the highest number of pickups, showing a reliable and frequent demand. Additionally, there is a clear weekly trend in passenger numbers, with certain days of the week like Tuesday and Thursday consistently experiencing higher demand.

The interactive dashboard proved highly effective, with heatmaps providing actionable insights for resource allocation and statistical summaries offering additional depth for decision-making. However, our experiments highlighted a trade-off: the best-performing model, Random Forest, was less scalable due to its large size and longer runtime, whereas Decision Tree provided a balance between accuracy and real-time usability, making it better suited for deployment.

CONCLUSIONS

Our project successfully combined predictive modeling with interactive visualizations to optimize NYC taxi operations. Random Forest demonstrated the highest accuracy for predicting pickups, with an R^2 of 0.976, but its large model size (6 GB) rendered it impractical for real-time applications. Decision Tree emerged as a more suitable alternative, balancing accuracy (R^2: 0.957) with a manageable model size (100 MB).

XGBoost provided accurate predictions for trip durations, while the SARIMAX model effectively forecasted demand trends. The dashboard offered actionable insights through intuitive heatmaps and graphs, significantly enhancing decision-making. Future work could involve integrating external data, such as weather or event schedules, for enhanced accuracy. Additionally, exploring lightweight ensemble models or hybrid approaches may help achieve a balance between precision and scalability for real-time applications.

EFFORT DISTRIBUTION

All team members have contributed a similar amount of effort.

REFERENCES

[1] Ferreira, N., Poco, J., Vo, H. T., Freire, J., & Silva, C. T. (2013). Visual exploration of big spatio-temporal urban data: A study of New York City cab trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2149-2158.

[2] Yazici, M. A., Kamga, C., & Singhal, A. (2013). A big data driven model for taxi drivers' airport pick-up decisions in New York City. In *2013 IEEE International Conference on Big Data* (pp. 37-44).

[3] Zhao, K., Khryashchev, D., Freire, J., Silva, C., & Vo, H. (2016). Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In *2016 IEEE International Conference on Big Data (Big Data)* (pp. 833-842).

[4] Tang, J., Liu, F., Wang, Y., & Wang, H. (2015). Uncovering urban human mobility from large scale taxi GPS data. *Physica A: Statistical Mechanics and its Applications*, 438, 140-153.

[5] Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., & Damas, L. (2013). Predicting taxi–passenger demand using streaming data. *IEEE Transactions on Intelligent Transportation Systems*, 14(3), 1393-1402.

[6] Zhang, J., Zheng, Y., & Qi, D. (2017). Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 31, No. 1).

[7] Xu, J., Rahmatizadeh, R., Bölöni, L., & Turgut, D. (2017). Real-time prediction of taxi demand using recurrent neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 19(8), 2572-2581.

[8] Davis, N., Raina, G., & Jagannathan, K. (2016). A multi-level clustering approach for forecasting taxi travel demand. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 223-228).

[9] Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., ... & Lv, W. (2017). The simpler the better: a unified approach to predicting original taxi demands based on large-scale online platforms. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1653-1662).

[10] Poongodi, M., Hamdi, M., Sharma, A., Ma, M., & Singh, P. K. (2021). DDoS detection mechanism using trust-based evaluation system in VANET. *IEEE Access*, 9, 15478-15489.

[11] Laptev, N., Yosinski, J., Li, L. E., & Smyl, S. (2017). Time-series extreme event forecasting with neural networks at Uber. In *International Conference on Machine Learning* (Vol. 70, pp. 1-5).

[12] Safikhani, A., Kamga, C., Mudigonda, S., Faghih, S. S., & Moghimi, B. (2017). Spatio-temporal modeling of yellow taxi demands in New York City using generalized STAR models. arXiv preprint arXiv:1711.10090.

[13] Yao, H., Wu, F., Ke, J., Tang, X., Jia, Y., Lu, S., ... & Li, Z. (2018). Deep multi-view spatial-temporal network for taxi demand prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 32, No. 1).