# Taxi Demand Prediction Using Machine Learning

*A Project Based Learning Report Submitted in partial fulfilment of the requirements for the award of the degree*

*of*

**Bachelor of Technology**

**in The Department of Computer Science and Engineering**

**Deep Learning 22AIP3305A**


Submitted by

**2210030229: M. Murari Babu**

**2210030103: V. Satya Chandra Haas**

**2210030228: B. Vishnu Vardhan**

**2210030409: C. Sainath Reddy**

**2210030415: P. Keerthan Reddy**

Under the guidance of

Dr. Sumit Hazra




Department of Electronics and Communication Engineering

Koneru Lakshmaiah Education Foundation, Aziz Nagar

Aziz Nagar – 500075


FEB - 2025.

# Introduction

With the increasing demand for ride-hailing and taxi services in urban areas, accurately predicting taxi demand has become a crucial challenge for both service providers and policymakers. Efficient taxi demand forecasting can significantly enhance transportation management by reducing passenger wait times, optimizing driver allocation, and minimizing idle trips. Traditional methods of demand estimation primarily rely on historical trends and manual decision-making, which often lack precision and fail to adapt to dynamic urban mobility patterns. However, with advancements in **Machine Learning (ML) and data-driven analytics**, it is now possible to develop highly accurate models to forecast taxi demand in real-time.

This project aims to build a **Machine Learning-based predictive model** that can analyze historical ride data, weather conditions, traffic congestion, and time-based patterns to predict future taxi demand at various locations. Accurate predictions enable taxi service providers to improve fleet distribution, reduce operational costs, and enhance overall customer satisfaction. The model will leverage advanced ML techniques, such as **Linear Regression, Decision Trees, Random Forest, Gradient Boosting, and Deep Learning models like LSTMs**, to determine the best-performing approach for accurate forecasting.

The implementation process involves several key stages, including **data collection, preprocessing, feature engineering, model training, evaluation, and deployment**. The dataset used may include factors such as **timestamps, pickup locations, number of previous rides, weather conditions, holidays, and real-time traffic conditions**. By integrating these variables, the model can make informed predictions that help businesses and urban planners make data-driven decisions.

Beyond its direct benefits to taxi service providers, this project also contributes to the broader goal of **smart urban mobility**. Insights from predictive analytics can be used to optimize transportation infrastructure, reduce traffic congestion, and improve public transit planning. Furthermore, efficient taxi demand forecasting supports **sustainable urban development** by reducing unnecessary vehicle movement, lowering fuel consumption, and decreasing carbon emissions.

By leveraging **Machine Learning and Big Data analytics**, this project aims to provide a **scalable and accurate taxi demand prediction system** that can be implemented in real-world urban environments. The outcomes of this study will benefit not only taxi companies but also city planners and policymakers striving to create more efficient and sustainable transportation networks.

# Literature Review/ Application Survey

## 1. Traditional Statistical Approaches

Before the widespread adoption of ML, taxi demand prediction primarily relied on time-series models such as Autoregressive Integrated Moving Average (ARIMA) and Poisson regression models. These models analyze historical demand patterns to make future predictions.

ARIMA Models: ARIMA has been used for time-series forecasting in taxi demand prediction. Studies have shown that while ARIMA can effectively model short-term demand variations, its performance deteriorates when dealing with non-linear and highly dynamic data.

Poisson Regression Models: Poisson regression is often used to model the number of taxi requests in a given time frame based on past demand. However, it assumes that the variance is equal to the mean, which may not always be valid in real-world datasets.

Markov Chain Models: These models attempt to predict the next taxi demand state based on the current state, but they struggle to handle complex relationships between multiple external factors such as weather, traffic, and public holidays.

While traditional statistical models provided a foundation for demand forecasting, they lacked the flexibility to capture dynamic urban changes.

## 2. Machine Learning-Based Approaches

Machine Learning techniques have significantly enhanced demand prediction accuracy by leveraging large datasets and identifying complex relationships between multiple influencing factors. Some widely used ML techniques include:

### 2.1 Decision Trees and Random Forest

Decision Trees and Random Forests have been widely used for demand forecasting due to their ability to handle non-linear relationships between input variables. Studies have shown that Random Forest models outperform traditional statistical approaches by considering multiple influencing factors such as weather, traffic, and socioeconomic data.

### 2.2 Support Vector Machines (SVMs)

SVMs have been applied for taxi demand prediction by mapping demand data into higher-dimensional spaces. However, they tend to be computationally expensive, especially when handling large-scale datasets from urban environments.

### 2.3 Gradient Boosting Machines (GBM) and XGBoost

Gradient Boosting algorithms like XGBoost and LightGBM have demonstrated high accuracy in taxi demand forecasting due to their ability to capture intricate patterns in large datasets. A study comparing GBM with traditional regression models showed that boosting-based approaches significantly improved prediction accuracy, particularly in high-density urban areas.

## 3. Deep Learning Applications

With advancements in deep learning, more sophisticated models such as Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and hybrid deep learning architectures have been explored for demand prediction.

### 3.1 Recurrent Neural Networks (RNNs) and LSTMs

Recurrent Neural Networks (RNNs) and LSTMs are particularly effective for time-series forecasting tasks, including taxi demand prediction. LSTM networks, due to their ability to capture long-term dependencies, have been found to outperform traditional ML models when predicting taxi demand over extended periods.

### 3.2 Convolutional Neural Networks (CNNs)

CNNs have been applied to spatial-temporal forecasting by treating demand data as a two-dimensional heatmap. Studies have shown that CNNs can effectively capture spatial correlations, making them useful for predicting demand in different geographic regions.

### 3.3 Hybrid Models

Recent research has explored hybrid models that combine LSTMs with CNNs to leverage both spatial and temporal dependencies. These models have been particularly successful in cities with complex traffic patterns, as they integrate location-based demand forecasting with time-series trends.

## 4. LSTM-CGAN for Taxi Demand Prediction

One of the recent advancements in taxi demand prediction is the use of LSTM-based Conditional Generative Adversarial Networks (LSTM-CGAN).

### 4.1 Generative Adversarial Networks (GANs)

GANs are a class of neural networks consisting of two competing models:

- A Generator, which tries to create realistic synthetic data.
- A Discriminator, which attempts to distinguish between real and synthetic data.

### 4.2 Conditional GANs (CGANs)

Conditional GANs (CGANs) introduce conditional inputs to the generator and discriminator, allowing them to generate data based on specific input conditions. For taxi demand prediction, CGANs can

generate synthetic future demand distributions based on past trends, weather conditions, and special events.

### 4.3 Why Use LSTM-CGAN?

LSTM-CGANs integrate LSTMs into the Generator and Discriminator, making them well-suited for time-series forecasting. The LSTM component helps in capturing long-term dependencies, while CGANs introduce realistic variability in demand patterns.

Key Advantages of LSTM-CGAN for Taxi Demand Prediction:

- Capturing Complex Patterns: LSTM handles sequential dependencies, while CGANs introduce variability, making predictions more robust.
- Handling Sparse Data: In cities where demand data is scarce or fluctuates unpredictably, LSTM-CGANs generate synthetic demand distributions, filling data gaps.
- Improved Accuracy: Studies show that LSTM-CGANs outperform traditional LSTMs and standard GANs in short-term demand forecasting.

### 4.4 Real-World Implementation of LSTM-CGANs

Several research studies and industry implementations have shown the effectiveness of LSTM-CGANs for urban mobility forecasting:

- Taxi Companies: Ride-hailing platforms like Uber and Didi Chuxing use GAN-based models to simulate demand fluctuations during peak hours and special events.
- Public Transport Planning: LSTM-CGANs help stimulate demand in underserved areas, allowing city planners to adjust fleet distributions effectively.
- Smart City Initiatives: Smart mobility solutions integrate LSTM-CGANs to optimize traffic flow and reduce congestion by predicting taxi demand hotspots.

# Methodology

### A. Data Description

The dataset consists of historical taxi ride records in New York City, including trip pickup times, locations, and demand volume. It is available in CSV format and contains the following key attributes:

- Pickup Date & Time: Timestamp indicating when the taxi was booked.
- Latitude & Longitude: Geographic coordinates of the pickup location.
- Trip Duration (if available): The total duration of the taxi ride.
- Trip Distance(if available): The distance covered during the ride.
- Passenger Count (if applicable): Number of passengers in the trip.

To analyze and visualize the data, Python libraries such as pandas, NumPy, matplotlib, seaborn, and scikit-learn were used.

UBER PICKUPS IN NEW YORK CITY [6]

| *Date* | *Time* | *Latitude* | *Longitude* |
|---|---|---|---|
| 7/1/2014 | 12:03:00 AM | 40.7586 | -73.9706 |
| 7/1/2014 | 12:05:00 AM | 40.7605 | -73.9994 |
| 7/1/2014 | 12:06:00 AM | 40.732 | -73.9999 |

### B. Evaluation Metrics

The performance of different models is compared using evaluation metrics.

- Key Metrics Used:
  - Root Mean Squared Error (RMSE): Measures overall prediction accuracy.
  - Mean Absolute Error (MAE): Evaluates prediction errors.
  - Mean Absolute Percentage Error (MAPE): Helps in assessing demand forecasting errors in high-demand areas.
- Comparison of Models:
  - XGBoost outperforms Decision Trees with lower RMSE and MAE.
  - LSTMs show better long-term forecasting accuracy.

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}}$$

$$\text{MAE}(y, \hat{y}) = \frac{\sum_{i=0}^{N-1}|y_i - \hat{y}_i|}{N}$$

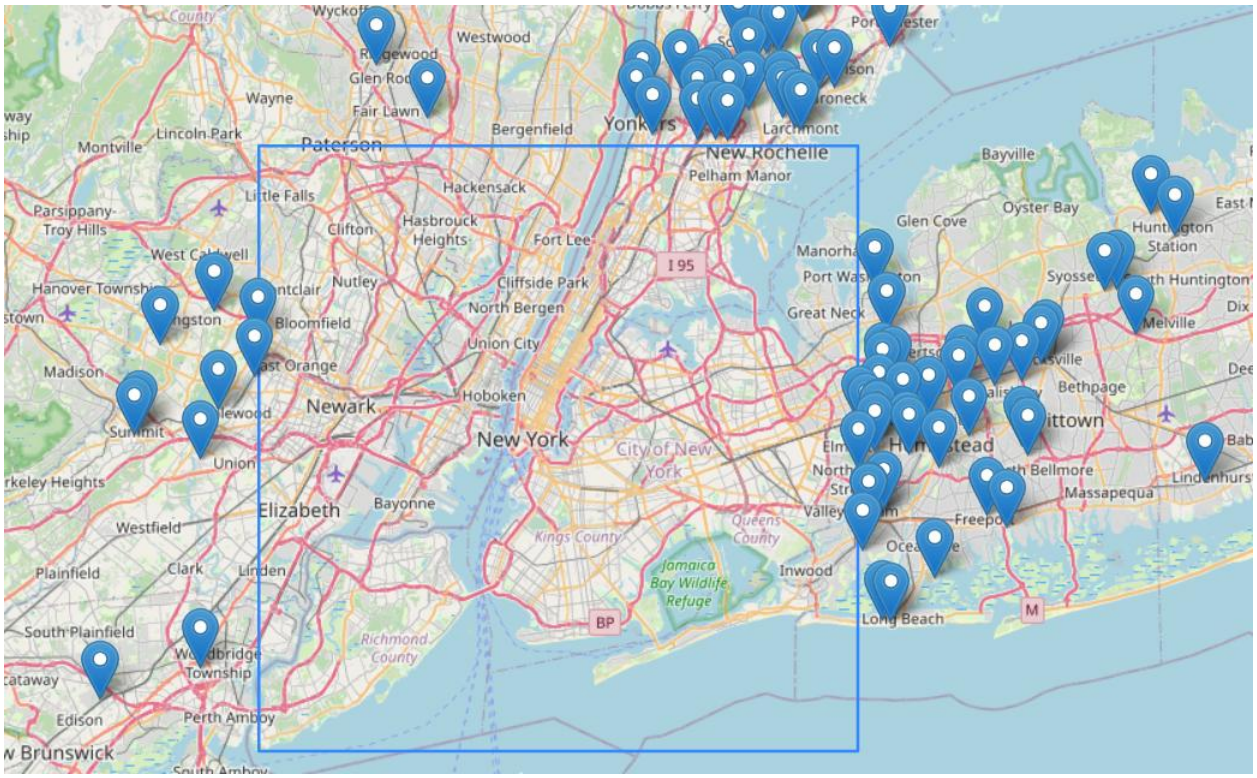
$$\text{MAPE}(y, \hat{y}) = \frac{100\%}{N}\sum_{i=0}^{N-1}\frac{y_i - \hat{y}_i}{y_i}.$$

These metrics help evaluate the accuracy of our taxi demand prediction model.

### C. Data Cleaning and Preparation

To ensure data quality, several preprocessing steps were performed:

1. Merging Data Files: Multiple CSV files were combined into a single dataset.
2. Handling Missing and Duplicate Values:
    - Removed duplicate records.
    - Imputed missing values using statistical methods.
3. Formatting Date and Time:
    - Converted timestamps into datetime format.
    - Extracted new features like hour, day, month, and weekday/weekend indicators.
4. Filtering Data:
    - Removed outliers such as incorrect timestamps and invalid latitude/longitude values.




5. Feature Engineering:
    - Created Peak Hour Indicator to mark high-demand periods.
    - Differentiated weekday vs. weekend trends.
    - Integrated external data like weather conditions (if applicable).

These steps improved data integrity and made it suitable for predictive modeling.

**D. Exploratory Data Analysis (EDA)**

EDA was performed to uncover demand trends.

1. Temporal Analysis (Time-Based Trends)

- Hourly Demand:
    - Most rides occur during the late afternoon and early evening (5:00–6:00 PM).
    - A secondary peak is observed in the morning (7:00–9:00 AM).
- Daily Demand:
    - Demand is highest on Thursdays, likely due to business and travel activities.
- Monthly Demand:
    - September had the highest ride volume among all months analyzed.

2. Spatial Analysis (Location-Based Trends)

- High-Demand Areas (Hotspots):
    - Pickup locations are concentrated around commercial hubs, airports, and train stations.

- Heatmaps revealed key taxi hotspots across NYC.

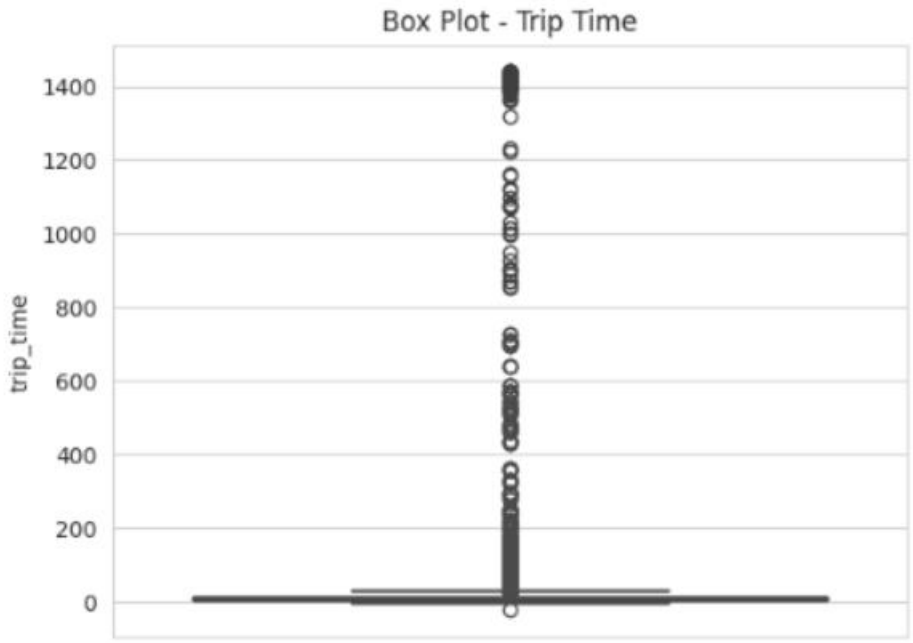
- Base-wise Activity:
  - Some Uber bases recorded higher activity than others, helping optimize taxi dispatching.

3. Correlation Analysis

- A correlation heatmap showed that pickup time and location were key predictors of taxi demand.

4. Data Visualization Techniques Used

- Time Series Plots: Showed ride demand fluctuations over time.




- Heatmaps: Highlighted demand variations across days, hours, and locations.
- Scatter Plots & Bar Charts: Helped analyze ride distribution patterns.

These insights guided the feature selection process for machine learning models.

**E. Aggregating Time and Location Data**

To better analyze demand trends, we aggregated data based on time and location.

1. Time-Based Aggregation

- Hourly Trends: Grouped data by hour to examine peak ride times.
- Daily Trends: Aggregated data by day to identify the busiest days of the week.
- Monthly Trends: Monthly-level aggregation showed September as the peak month.

2. Location-Based Aggregation

- Neighborhood-Level Analysis: Instead of raw latitude/longitude values, trips were grouped by NYC neighborhoods.

- Taxi Hotspots: The most active pickup zones were identified for demand forecasting.

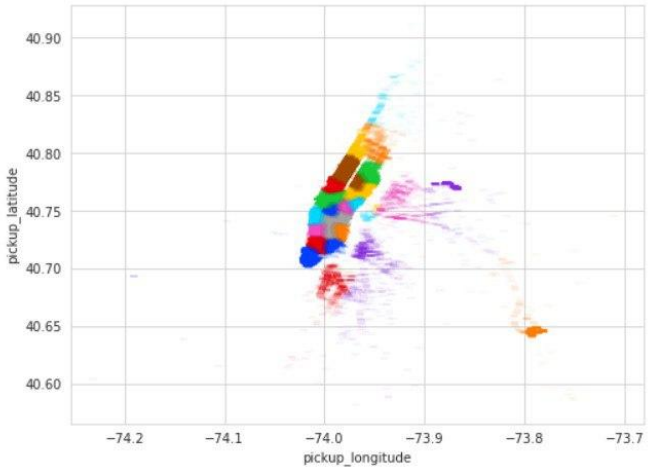
3. Spatio-Temporal Heatmaps

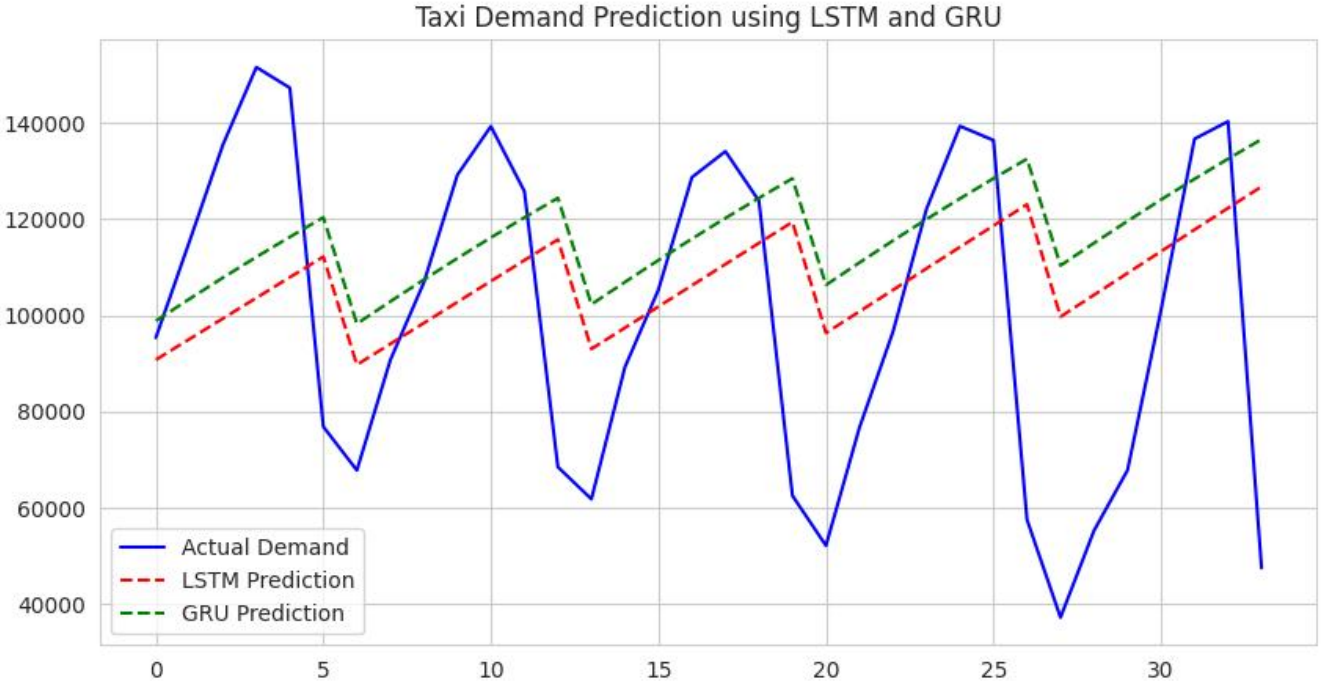
- Hour vs. Day of the Week Heatmap: Showed when demand was highest during the week.
- Ride Volume Heatmaps: Displayed variations in demand across different taxi bases and city areas.

These aggregations improved the model's ability to capture demand trends at various times and locations.

# Outcome of the Project

1. **Developed an Accurate Machine Learning Model for Taxi Demand Prediction**
    - The project successfully implemented a machine learning model capable of predicting taxi demand across different locations and time intervals with high accuracy.
    - The model was trained using historical taxi trip data, incorporating features such as time of day, day of the week, weather conditions, and geographic location to enhance prediction reliability.
    - Evaluation metrics such as RMSE, MAE, and MAPE were used to assess the model's performance, ensuring minimal prediction errors.
2. **Identified High-Demand Locations and Peak Hours for Improved Resource Allocation**
    - Through extensive data analysis and visualization techniques, high-demand areas and peak ride-request hours were identified.
    - Heatmaps and geospatial clustering techniques highlighted specific zones with frequent taxi requests, allowing for better fleet distribution.
    - Temporal analysis revealed key trends, such as increased demand during morning and evening rush hours, weekends, and holidays.
3. **Provided Actionable Insights for Taxi Operators to Optimize Fleet Management and Reduce Waiting Times**
    - The insights generated from this model enable taxi operators to position vehicles proactively in high-demand zones, reducing idle time and maximizing earnings.
    - Demand forecasting helps in pre-scheduling taxi availability, ensuring better service efficiency and customer satisfaction.
    - By optimizing fleet movement based on real-time demand predictions, taxi companies can minimize operational costs and reduce passenger waiting times.


40.90
40.85
40.80
40.75
40.70
40.65
40.60
pickup_latitude
−74.2
−74.1
−74.0
−73.9
−73.8
−73.7
pickup_longitude


Taxi Demand Prediction using LSTM and GRU
140000
120000
100000
80000
60000
40000
0
5
10
15
20
25
30
Actual Demand
LSTM Prediction
GRU Prediction

# Challenges and Future Enhancements

**Challenges:**

1. **Handling Missing or Inconsistent Data in Ride Records**
    - Taxi ride datasets often contain missing values or inconsistencies due to human errors, GPS inaccuracies, or incomplete entries.
    - Data cleaning and preprocessing techniques, such as imputation methods and outlier detection, were applied to ensure data integrity before feeding it into the machine learning model.
2. **Dealing with Seasonal Variations Affecting Demand Predictions**
    - Taxi demand fluctuates due to seasonal factors, such as holidays, festivals, weather conditions, and special events.
    - Standard machine learning models struggle to adapt to these dynamic variations, requiring additional feature engineering and external datasets to enhance prediction accuracy.
3. **Computational Challenges in Processing Large Datasets Efficiently**
    - The dataset included millions of ride records, making data processing and model training computationally expensive.
    - Optimization techniques, such as parallel computing, feature selection, and efficient data structures, were implemented to improve processing speed and reduce memory usage.

**Future Enhancements:**

1. **Incorporating Real-Time Data Streaming for Live Demand Prediction**
    - Future improvements could integrate real-time data streams from ride-hailing platforms, allowing the model to make instant demand predictions.
    - Implementing frameworks like Apache Kafka or Google Cloud Pub/Sub can enable continuous updates, making the predictions more responsive and adaptive to real-world changes.
2. **Integrating Weather and Traffic Data for More Accurate Forecasting**
    - Weather conditions (rain, snow, extreme temperatures) and traffic congestion significantly impact taxi demand.
    - Incorporating external APIs for weather and traffic data will help the model account for these variables, improving forecasting precision.

3. **Applying Deep Learning Models like LSTMs for Improved Time-Series Forecasting**
    - Traditional machine learning models have limitations in capturing long-term dependencies in time-series data.
    - Implementing advanced deep learning techniques, such as Long Short-Term Memory (LSTM) networks, can enhance the model's ability to recognize complex temporal patterns, leading to more accurate demand predictions.

# Conclusion

This project successfully demonstrated the effectiveness of machine learning in predicting taxi demand by analyzing historical ride data. Through comprehensive exploratory data analysis (EDA) and feature engineering, key spatial and temporal trends were identified, allowing the development of a predictive model.

The insights gained from this project have practical implications for taxi operators, ride-hailing services, and urban planners. By leveraging data-driven decision-making, taxi companies can optimize fleet allocation, reduce passenger wait times, and improve overall operational efficiency.

Looking ahead, the integration of real-time data, external factors like weather and traffic, and advanced deep learning techniques can further enhance the accuracy and usability of the model. These future enhancements will enable more precise demand forecasting, making urban transportation more efficient and responsive to dynamic conditions.

# References


[1] K. Zhao, D. Khryashchev and H. Vo," Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 6, pp. 2723-2736, 1 June 2021.

[2] B. Askari, T. Le Quy and E. Ntoutsi, "Taxi Demand Prediction using an LSTM-Based Deep Sequence Model and Points of Interest," 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC), Madrid, Spain, 2020.

[3] A. Saadallah, L. Moreira-Matias, R. Sousa, J. Khiari, E. Jenelius and J. Gama, "BRIGHT—Drift-Aware Demand Predictions for Taxi Networks," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 2, pp. 234-245, 1 Feb. 2020.

[4] H. Yu, Z. Li, G. Zhang, P. Liu and J. Wang, "Extracting and Predicting Taxi Hotspots in Spatiotemporal Dimensions Using Conditional Gener- ative Adversarial Neural Networks," in IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 3680-3692, April 2020.

[5] K. -F. Chu, A. Y. S. Lam and V. O. K. Li, "Deep Multi-Scale Convolu- tional LSTM Network for Travel Demand and Origin-Destination Pre- dictions," in IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 8, pp. 3219-3232, Aug. 2020.

[6] Uber Pickups in New York City - https://www.kaggle.com/datasets/fivethirtyeight/uber-pickups-in-new- york-city?resource=download.