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.

5. Challenges and Future Research Directions

Despite significant advancements, taxi demand prediction using LSTM-CGANs still faces several challenges:

- Computational Complexity: Training GANs combined with LSTMs requires high computational power and large datasets. Efficient cloud-based implementations are needed.
- Data Scarcity: Real-world taxi demand datasets are often incomplete or biased. Future research should focus on data augmentation techniques.
- Model Stability: GAN training can be unstable. Techniques like Wasserstein GAN (WGAN) and Spectral Normalization should be explored for stable training.
- Integration with External Data: Incorporating real-time weather, road closures, and event data into LSTM-CGAN models will enhance prediction accuracy.

- Fairness and Ethics: Demand forecasting should ensure equitable taxi distribution, preventing biases that may lead to unfair pricing strategies.

6. Conclusion

The use of Machine Learning, Deep Learning, and Generative Adversarial Networks (GANs) has significantly improved taxi demand forecasting. While traditional ARIMA models and basic ML techniques offer some accuracy, LSTM-CGANs provide a more advanced solution by capturing both sequential dependencies and data variability. Future research should focus on improving GAN stability, integrating external real-time data, and developing scalable cloud-based solutions. The successful implementation of LSTM-CGANs in urban mobility systems will revolutionize smart transportation planning, optimize taxi fleet management, and enhance customer satisfaction.

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.