Report

Taxi & Ride Demand Prediction Using Historical and Regional Data

Goutam Agarwal

1. Introduction

Ride-hailing services like Uber have become an essential part of urban transportation systems. Efficient allocation of drivers and pricing decisions depend on accurately forecasting ride demand across regions and time. Without accurate demand predictions, companies face challenges such as:

- Driver shortage → longer passenger wait times.
- Oversupply → idle drivers and wasted resources.
- Poor pricing strategies → loss of revenue opportunities.

This project aims to forecast Uber ride demand in New York City (NYC) using historical and regional data. The study combines EDA, clustering, time series analysis, and machine learning models to provide accurate and interpretable demand forecasts.

2. Data Preparation

The dataset consists of Uber ride request data from NYC, including timestamps, location information, and ride counts.

Data Preprocessing & Cleaning:

- Removed outliers (extreme spikes due to abnormal events).
- Segmented NYC into regions/zones for localized demand forecasting.
- Applied smoothing techniques to reduce noise in the demand signal.

Feature Engineering:

- Extracted time-based features (hour, day of week, weekend/weekday).
- Created historical lag features (previous demand values).
- Computed rolling averages and moving windows to capture trends.

- Incorporated cluster labels (using K-Means) to identify demand behavior groups across different regions.

3. Exploratory Data Analysis (EDA)

EDA was conducted to uncover spatio-temporal demand patterns:
- Hourly and daily trends → demand peaks during morning and evening rush hours.
- Weekday vs. weekend differences → weekdays showed consistent commuter demand; weekends had irregular late-night peaks.
- Regional variations → Manhattan recorded the highest ride requests; outer boroughs showed scattered demand.
- Clustering (K-Means) → grouped regions with similar demand patterns, improving model interpretability.

Visualizations included:
- Heatmaps of demand across NYC regions.
- Time series plots for demand cycles.
- Cluster maps to highlight regional groupings.

4. Methodology

Baseline Model:
- Developed a baseline Linear Regression model to forecast ride demand.
- This established a reference point for comparison with advanced models.

Time Series Modeling:
- Applied smoothing techniques (moving averages, exponential smoothing).
- Tested time series models (ARIMA, SARIMA, Prophet) to capture temporal dependencies.

Machine Learning Models:
- Trained multiple ML models including:
  - Linear Regression (LR)
  - Random Forest (RF)
  - Gradient Boosting Regressor (GBR)
  - XGBoost Regressor (XGBR)

Hyperparameter Tuning:
- Applied grid search and randomized search to optimize model parameters.
- Evaluated models using cross-validation to prevent overfitting.

Evaluation Metrics:
- Root Mean Square Error (RMSE)

- Mean Absolute Error (MAE)
- Mean Absolute Percentage Error (MAPE)

5. Results

- Baseline Model (LR) captured general trends but failed to adapt to non-linear demand changes.
- Random Forest & GBR improved accuracy by capturing complex temporal and regional interactions.
- XGBoost Regressor achieved the best performance with lowest RMSE and MAE after hyperparameter tuning.
- Clustering + Regional Segmentation improved interpretability and allowed localized predictions.

Key Insight: Demand forecasting improves significantly when historical lags, regional segmentation, and clustering are combined with advanced ML models.

6. Conclusion & Future Work

This project demonstrated an end-to-end pipeline for Taxi & Ride Demand Prediction Using Historical and Regional Data.

Key Contributions:

✅ Performed EDA to uncover spatio-temporal patterns in Uber ride demand.
✅ Developed a baseline regression model for demand forecasting.
✅ Applied smoothing techniques and time series models to capture temporal behavior.
✅ Used K-Means clustering to group regions with similar demand trends.
✅ Engineered time-based features and historical lags for improved model input.
✅ Trained and tuned multiple ML models (LR, RF, GBR, XGBR), with XGBoost performing best.

Future Directions:

- Incorporate weather, traffic, and event data for richer prediction.
- Explore deep learning approaches (LSTMs, Transformers) for sequential demand forecasting.
- Deploy the model as a real-time demand forecasting dashboard for operational use.

Final Note: This project highlights how combining EDA, clustering, time series analysis, and ML models can significantly enhance ride-hailing demand forecasts, enabling better driver allocation, pricing, and customer experience.