Project Report

Uber Trip Analysis
Machine Learning Project

By: Sujeet Kumar

Submitted To

Unified Mentor
Gurugram , Haryana , 122002

Content

1. Introduction

The rise of ride-hailing platforms like Uber has led to a significant increase in the volume of trip data collected. This dataset offers a unique opportunity to analyze ride patterns, understand customer behavior, and predict future demand. This project focuses on analyzing Uber trip data collected in New York City during 2014 and early 2015. The aim is to uncover trends, build predictive models, and evaluate model performance using machine learning techniques.

2. Project Statement

The main objectives of this project are:

- To identify temporal and spatial trends in Uber ride patterns.
- To forecast future ride demand using machine learning models.
- To evaluate the performance of different regression models (XGBoost, Random Forest, GBRT).
- To build an ensemble model for better prediction accuracy.

Key questions addressed:

- What are the peak hours and busiest days for Uber rides in NYC?

- How well can machine learning models forecast Uber demand based on historical data?
- Does combining multiple models improve forecasting performance?

3. Methodology

Dataset Description

- **Source**: NYC TLC (Taxi & Limousine Commission) Uber FOIL dataset (April–September 2014, Jan–Feb 2015).
- **Features**:
  - `Date/Time`: Timestamp of pickup.
  - `Lat`, `Lon`: Pickup coordinates.
  - `Base`: Dispatching company.
  - `Trips`, `Active Vehicles` (in aggregated dataset).

Data Preprocessing

- Combined monthly CSVs for Uber data.
- Converted `Date/Time` to datetime format.
- Extracted temporal features: `Hour`, `Day`, `DayOfWeek`, `Month`.

- Grouped data by hourly intervals to create a time series.
- Applied seasonal decomposition to visualize trends and seasonality.
- Applied a **window-based approach** to create lagged features for time series forecasting.

Modeling Approaches

Three main regression models were applied:

- **XGBoost Regressor**:
  - Tuned with GridSearchCV (243 combinations).
  - Captured trends and non-linear relationships effectively.
- **Random Forest Regressor**:
  - Used for its robustness and ability to handle high-dimensional data.
- **Gradient Boosting Regressor (GBRT)**:
  - Applied with extensive hyperparameter tuning.

Model Evaluation

- Evaluation metric: **MAPE (Mean Absolute Percentage Error)**.

- Cross-validation with **TimeSeriesSplit (5 folds)** to maintain temporal structure.
- Comparison across models.
- **Ensemble Model**: Combined predictions of the three models using reciprocal MAPE-based weights.

4. Results

Model	MAPE
XGBoost	**8.37%**
Random Forest	**9.61%**
Gradient Boosted Tree	**10.02 %**
Ensemble	**8.60%**

XGBoost outperformed the others individually.

Ensemble model offered a balance between performance and stability.

Visual plots showed all models tracked the test data trend closely.

Time series decomposition revealed daily trends and seasonal components (especially for weekdays vs weekends).

5. Conclusion

This Uber trip analysis project provided valuable insights into temporal ride patterns and demonstrated the strength of machine learning in time series forecasting.

Key Findings:

- Peak hours are during commute times and weekends.
- XGBoost delivered the most accurate predictions.
- The ensemble model, though slightly less accurate than XGBoost, offered robust performance and generalized better across the test period.
- Using window-based lagged features helped capture temporal dependencies effectively.

Practical Implications:

- Ride-hailing companies can leverage such predictive models for fleet optimization, dynamic pricing, and resource allocation.
- Time-based feature engineering and ensemble learning play a crucial role in real-time forecasting systems.