Uber Trip Analysis — Predictive Modelling (Jan–Feb 2015)

Author: Tabish Deshmukh | Data Analyst / ML Developer
🔗 GitHub: github.com/taabishhhhhh
🔗 LinkedIn: linkedin.com/in/tabish-deshmukh-430279268

Abstract

This research analyzes Uber's operational data (Jan–Feb 2015) using machine learning to forecast trip demand.
The objective is to identify temporal and operational factors influencing trip volume and to develop a predictive model for accurate demand forecasting.
The Gradient Boosting Regressor achieved **R^2 = 0.9821**, **MAPE = 7.67%**, and **RMSE = 1505.93**, establishing it as the most effective model.

Dataset Overview

- **Data Period:** January–February 2015
- **Fields:** Date, Dispatching Base Number, Active Vehicles, Trips
- **Total Records:** ~350 daily entries
- **Goal:** Predict total trips per day

Methodology

1. **Data Pre-processing** — Converted dates, handled missing data, and sorted chronologically.
2. **Exploratory Data Analysis** — Identified strong correlations between active vehicles and trip volume.
3. **Feature Engineering** —
   - Added time-based attributes (day, month, weekday/weekend).
   - Created rolling means (3-day, 7-day) and lag features (1–3 days).
4. **Modelling** — Trained Random Forest, XGBoost, and Gradient Boosting Regressors.
5. **Explain ability (SHAP)** — Provided global and local feature importance analysis.

Results

Model	MAPE (%)	RMSE	R^2 Score
Random Forest	9.0485	2048.6752	0.9669
XGBoost	8.7246	1798.0989	0.9745
Gradient Boosting	**7.6726**	**1505.9369**	**0.9821**

- Trip volume depends heavily on **active vehicles**, recent trip averages, and previous day's demand (lag features).
- Demand surges on weekends, with cyclical weekly trends visible.

Explain ability Summary (SHAP)

- `active_vehicles`, `lag_1`, and `trips_rolling_mean_3` were top contributors.
- SHAP visualizations confirmed model stability and transparency.
- Force plots revealed clear relationships between operational capacity and predicted demand.

Key Insights

- Gradient Boosting model accurately forecasts daily demand.
- Vehicle availability drives trip count with lag-dependent influence.
- Strong weekday–weekend demand pattern aids scheduling optimization.

Future Work

- Integrate **external data** such as weather or events.
- Expand to **LSTM / Prophet models** for time-series forecasting.
- Build **Streamlit dashboard** for real-time trip prediction.

Contact

📧 deshmukhtabish4@gmail.com
🔗 GitHub
🔗 LinkedIn