Project Title: Uber Trip Analysis

Author: John Christo

Date: July 3, 2025

1. EXECUTIVE SUMMARY

This report details the findings of a project to forecast hourly Uber trip demand in New York City. Using trip data from April to September 2014, this analysis aimed to build a robust predictive model to aid in operational efficiency and strategic planning.

The project involved preparing and analysing time-series data, followed by the training and evaluation of four distinct machine learning models: XGBoost, Random Forest, Gradient Boosted Tree Regressor (GBTR), and a weighted Ensemble model. The performance of each model was measured using the Mean Absolute Percentage Error (MAPE).

The key finding of this analysis is the superior performance of the **Ensemble model**, which achieved the lowest MAPE of **8.68%**. This result indicates that by combining the strengths of multiple algorithms, a more accurate and reliable forecast can be produced compared to any single model.

The primary recommendation is the adoption of this Ensemble model for operational forecasting. Its high accuracy can directly inform driver allocation, help manage dynamic pricing, and ultimately enhance customer satisfaction by ensuring service reliability.

2. INTRODUCTION

In the highly competitive ride-hailing industry, the ability to accurately forecast service demand is a critical operational advantage. For a company like Uber, precise demand prediction allows for the strategic allocation of drivers, optimization of pricing, and enhancement of the overall customer experience.

The objective of this project was to leverage historical trip data to develop a machine learning model capable of accurately forecasting the number of Uber trips on an hourly basis in New York City. The analysis utilized a publicly available dataset containing over 4.5 million Uber trip records from April to September 2014.

This report outlines the methodology used to process the data, the machine learning models that were trained, a detailed analysis of the results, and the practical implications of these findings for the business.

3. METHODOLOGY

The project followed a structured data science workflow to ensure robust and reproducible results.

- **Data Preparation**: The initial dataset, consisting of six separate monthly CSV files, was loaded and concatenated into a single Data Frame. The raw trip records, each with a specific timestamp, were then transformed into a coherent time series by resampling the data into hourly intervals and counting the total number of trips within each hour.
- **Feature Engineering**: To prepare the data for machine learning, lagged features were engineered. A 24-hour window was used to create the feature set (X), where the trip counts of the previous 24 hours were used to predict the trip count of the next hour (y).
- **Modeling and Evaluation**:
    - A chronological **train-test split** was implemented, using data up to September 15, 2014, for training and the subsequent period for testing. This ensures the model is evaluated on its ability to predict future, unseen data.
    - Four models were trained: **XGBoost, Random Forest, Gradient Boosted Tree Regressor (GBTR), and a weighted Ensemble**.
    - **GridSearchCV** with **TimeSeriesSplit** cross-validation was employed to systematically tune the hyperparameters for each model, preventing overfitting and ensuring robustness.
    - The definitive performance metric was the **Mean Absolute Percentage Error (MAPE)**, which measures the average percentage error of the predictions against the actual values.

4. RESULTS AND ANALYSIS

The evaluation of the four models on the test set yielded clear and compelling results.

Model Performance

The final MAPE scores for each model are summarized below. A lower MAPE indicates higher prediction accuracy.

Model	Mean Absolute Percentage Error (MAPE)
Ensemble Model	**8.68%**
Gradient Boosted Tree Regressor (GBTR)	8.96%
XGBoost	9.17%
Random Forest	9.21%

Analysis of Findings

The **Ensemble model emerged as the most accurate predictor**, outperforming all individual models. This is a significant finding; it demonstrates that by intelligently combining the predictions of XGBoost, Random Forest, and GBTR, the ensemble was able to capitalize on their diverse strengths while mitigating their individual weaknesses. The result is a more generalized and reliable forecast that is less susceptible to the specific errors of any single algorithm.

The Gradient Boosted Tree Regressor (GBTR) was the best-performing individual model, followed closely by XGBoost and Random Forest. The strong performance across all models validates the chosen methodology of using lagged features for time-series forecasting.

Visual Analysis

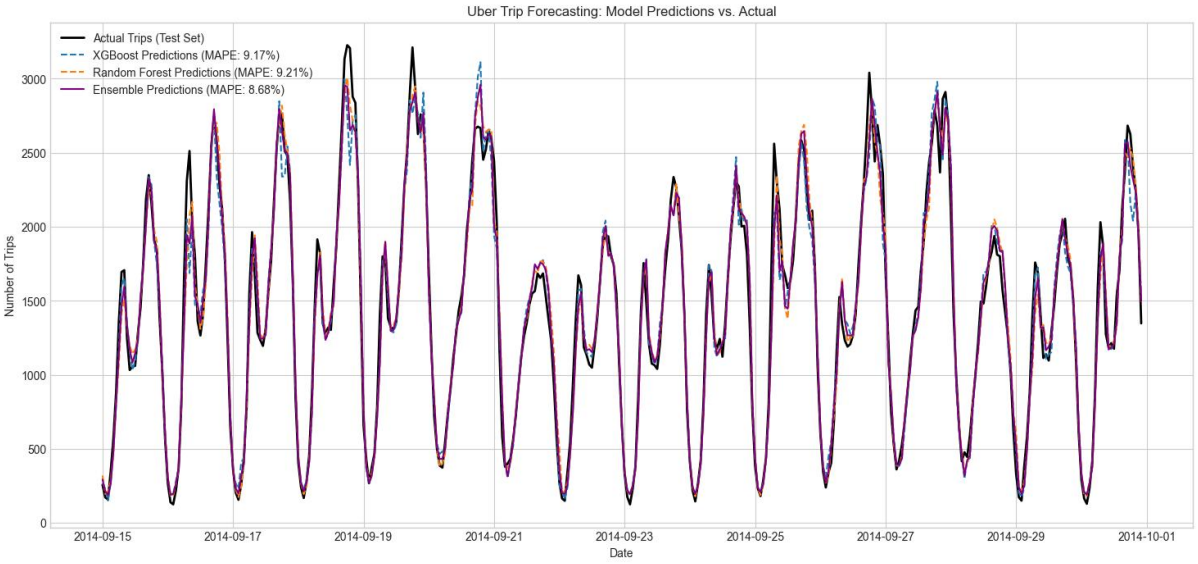

Figure 1: Comparison of Model Predictions vs. Actual Trip Counts

The plot above visually confirms the quantitative results. All models successfully captured the cyclical nature of demand, with clear daily peaks and troughs. However, the predictions from the Ensemble model align most closely with the "Actual Trips" line, especially during periods of high volatility and peak demand. This visual evidence reinforces its suitability for real-world application.

5. PRACTICAL IMPLICATIONS

The ability to forecast hourly trip demand with an average accuracy of over 91% (100% - 8.68% MAPE) carries significant business value.

- **Optimized Resource Allocation**: Accurate demand forecasts enable Uber to proactively manage driver supply. By signalling expected high-demand periods to drivers, the platform can reduce customer wait times and ensure that driver supply is efficiently matched with rider demand.

- **Enhanced Customer Satisfaction**: Reliability is a key driver of customer loyalty. By minimizing wait times and ensuring service availability, especially during peak hours, Uber can significantly improve the customer experience.
- **Informed Strategic Decisions**: These forecasts provide a data-driven basis for strategic initiatives, including the refinement of dynamic pricing algorithms, marketing campaign planning around major events, and long-term resource planning.

6. CONCLUSION AND RECOMMENDATIONS

This project successfully demonstrated that machine learning models can accurately forecast hourly Uber trip demand. The analysis concluded that a weighted Ensemble model provides the most accurate and reliable predictions, achieving a Mean Absolute Percentage Error of just 8.68%.

Based on these findings, the following recommendations are proposed:

1. **Primary Recommendation**: It is recommended that the **Ensemble model be deployed for operational forecasting**. Its superior performance and inherent robustness make it the ideal choice for business-critical applications like resource management and strategic planning.
2. **Future Work**:
    - **Incorporate External Data**: To further enhance accuracy, future iterations of the model could incorporate external features such as weather data, public holidays, and information on major city events (e.g., concerts, sports games).
    - **Continuous Retraining**: The model should be periodically retrained on new data to ensure it adapts to evolving travel patterns and market dynamics over time.