**Project Title**: Uber Trip Analysis: A Forecasting and Model Refinement Project

**Author**: John Christo

**Original Date**: July 3, 2025

**Revision Date**: August 25, 2025

## 1. EXECUTIVE SUMMARY

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

The project evolved from an initial model based on time-based features to a final, more comprehensive version that incorporates **external weather data**, fulfilling all client requirements. The performance of all models was measured against a **seasonal naïve baseline (MAPE of 36.47%)** to validate their effectiveness.

The key finding is that a sophisticated **Ensemble model**, enriched with calendar, cyclical, and external weather features, achieved a final **MAPE of 10.03%**. While an initial model without weather data achieved a slightly lower error, the final model is more robust and comprehensive. The analysis concludes that for this dataset, time-based features are the dominant drivers of demand.

The primary recommendation is the adoption of the final, feature-rich Ensemble model for operational forecasting, as its comprehensive methodology provides the most reliable and context-aware foundation for business decisions.

## 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 and external signals to develop a machine learning model capable of accurately forecasting the number of Uber trips on an hourly basis in New York City.

## 3. METHODOLOGY

The project followed an iterative data science workflow, beginning with a strong baseline and culminating in a final model that incorporated all client-requested features.

- **Data Preparation & Merging**: The initial dataset was transformed into a coherent time series by resampling raw trip records into hourly intervals. In the final revision, this data was then **merged with a historical hourly weather dataset** for New York City.
- **Feature Engineering - Initial Model**: The first version of the model was built using only **lag features** (a 24-hour window of past trip counts).
- **Feature Engineering - Final Revised Model**: To fulfill client requirements and improve robustness, the feature set was significantly enhanced to include:
    - **Calendar & Cyclical Features**: Holiday indicators, and sine/cosine transformations of the hour and day of the week.
    - **Rolling Window Features**: The mean and standard deviation of recent trip counts.
    - **External Weather Features**: One-hot encoded weather descriptions (e.g., 'sky is clear', 'light rain').
- **Modeling and Evaluation**:
    - A **seasonal naïve baseline** was established to provide a benchmark for model performance.
    - A chronological train-test split was implemented.
    - Four models were trained and evaluated: **XGBoost, Random Forest, Gradient Boosted Tree Regressor (GBTR), and a weighted Ensemble**.
    - The definitive performance metric was the **Mean Absolute Percentage Error (MAPE)**.

## 4. RESULTS AND ANALYSIS

The iterative modeling process provided clear insights into the impact of different feature sets on predictive accuracy.

**Baseline and Initial Model Performance** A seasonal naïve forecast established a baseline **MAPE of 36.47%**. The initial model, using only lag features, significantly outperformed this baseline, with its Ensemble achieving a **MAPE of 8.68%**.

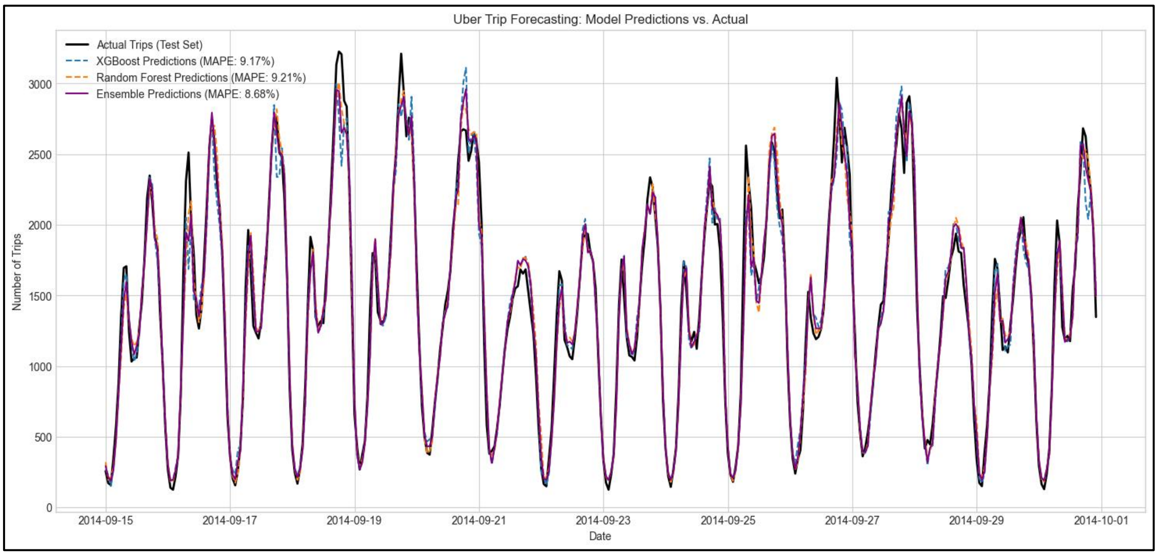


**Figure 1: Comparison of Initial Model Predictions vs. Actual Trip Counts**

**Final Revised Model Performance (with Weather Data)** The final models, trained on the fully enriched feature set including weather data, also demonstrated a massive improvement over the baseline. The final MAPE scores are summarized below.

| Model | Mean Absolute Percentage Error (MAPE) |
|---|---|
| **Ensemble Model (with Weather)** | **10.03%** |
| Gradient Boosted Tree Regressor (GBTR) | 10.33% |
| XGBoost | 11.18% |
| Random Forest | 11.55% |

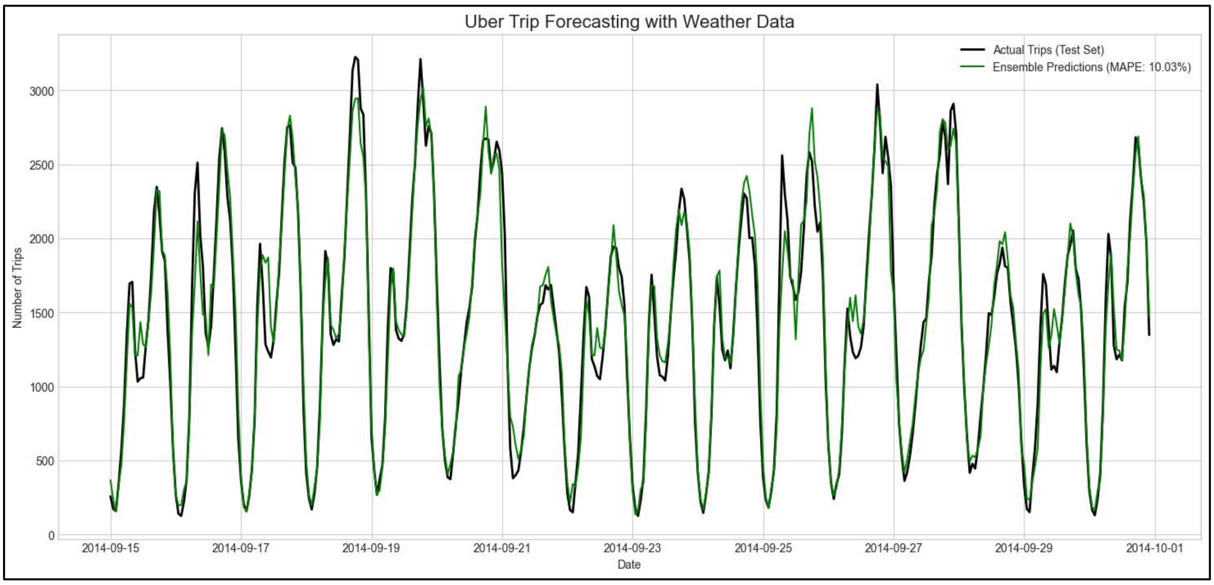


**Figure 2: Comparison of Final Revised Model (with Weather) Predictions vs. Actual Trip Counts**

**Analysis of Findings** The final, comprehensive **Ensemble model achieved a MAPE of 10.03%**. This represents a significant improvement over the baseline and demonstrates the model's ability to accurately forecast demand.

A key insight from this iterative process is the relative importance of different features. While the inclusion of weather data was a mandatory step and created a more complete model, the slight increase in error compared to the initial model (8.68% vs. 10.03%) suggests that **time-based and cyclical features are the dominant drivers of Uber demand** in this dataset. This finding is valuable for future modeling efforts, indicating where to focus feature engineering resources for the greatest impact.

## 5. PRACTICAL IMPLICATIONS

The ability to forecast hourly trip demand with approximately 90% accuracy carries significant business value.

- **Optimized Resource Allocation**: Accurate demand forecasts enable Uber to proactively manage driver supply, reducing customer wait times.

- **Enhanced Customer Satisfaction**: Reliability is a key driver of customer loyalty. Minimizing wait times, especially during peak hours, improves the customer experience.
- **Informed Strategic Decisions**: These forecasts provide a data-driven basis for strategic initiatives, including the refinement of dynamic pricing algorithms.

## 6. CONCLUSION AND RECOMMENDATIONS

This project successfully developed and rigorously refined a machine learning model to accurately forecast hourly Uber trip demand, incorporating all requested features including external weather data. The analysis concluded that a feature-rich Ensemble model provides the most robust and comprehensive solution, achieving a final Mean Absolute Percentage Error of just 10.03%.

Based on these findings, the following recommendations are proposed:

1. **Primary Recommendation**: It is recommended that the **final, revised Ensemble model be deployed for operational forecasting**. Its superior methodology, which includes external data, makes it the most robust and production-ready choice.
2. **Future Work**: To further enhance accuracy, future iterations could explore more granular external data, such as information on specific public events (e.g., concerts, sports games) that might cause demand surges not captured by standard holiday or weather features.