Uber Trip Analysis Report

1. Introduction

The Uber Trip Analysis project aims to analyze trip requests received by Uber in New York City to identify patterns, cancellations, and supply-demand gaps. This analysis helps in understanding the underlying causes of inefficiencies in Uber's operations and suggesting data-driven strategies to improve their services.

2. Project Objective

- Data Exploration and Visualization: Conduct an in-depth exploration of the 2014 Uber trip data, identifying patterns, trends, and seasonality through various visual and statistical techniques.
- Time Series Decomposition: Apply decomposition techniques to break down the data into trend, seasonality, and residual components to better understand temporal dynamics.
- Feature Engineering: Implement window-based (lagged) features to enhance the model’s ability to capture temporal dependencies and trends.
- Model Training: Develop and train three machine learning models—XGBoost, Random Forest, and Gradient Boosted Tree Regressor (GBTR)—using the engineered features.
- Model Evaluation: Evaluate the models' performance using Mean Absolute Percentage Error (MAPE) as the primary metric for accuracy and reliability.
- Ensemble Techniques: Design and apply an ensemble model that integrates predictions from the individual models to improve overall forecasting performance.
- Comparative Analysis: Compare the individual models and ensemble approach to highlight their strengths, limitations, and predictive capabilities.
- Insights and Recommendations: Summarize key findings and provide actionable recommendations based on model performance and data insights.

.

3. Dataset Overview

For this project, I have utilized a publicly available Uber trip dataset collected from the New York City area. The dataset comprises detailed information about Uber trips recorded over multiple months in the year 2014. It serves as a rich resource for analyzing ride patterns, passenger demand, and geographical trends in urban transportation.

The dataset used for this analysis contains the following key attributes:

- **Request id:** Unique identification number of the request.
- **Pickup point:** The location from where the customer requested the cab (City or Airport).

- **Driver id:** Identification number of the driver.
- **Status:** Current status of the trip - Completed, Cancelled, No Cars Available.
- **Request timestamp:** Date and time when the customer made the trip request.
- **Drop timestamp:** Date and time when the trip ended.

3. Approach

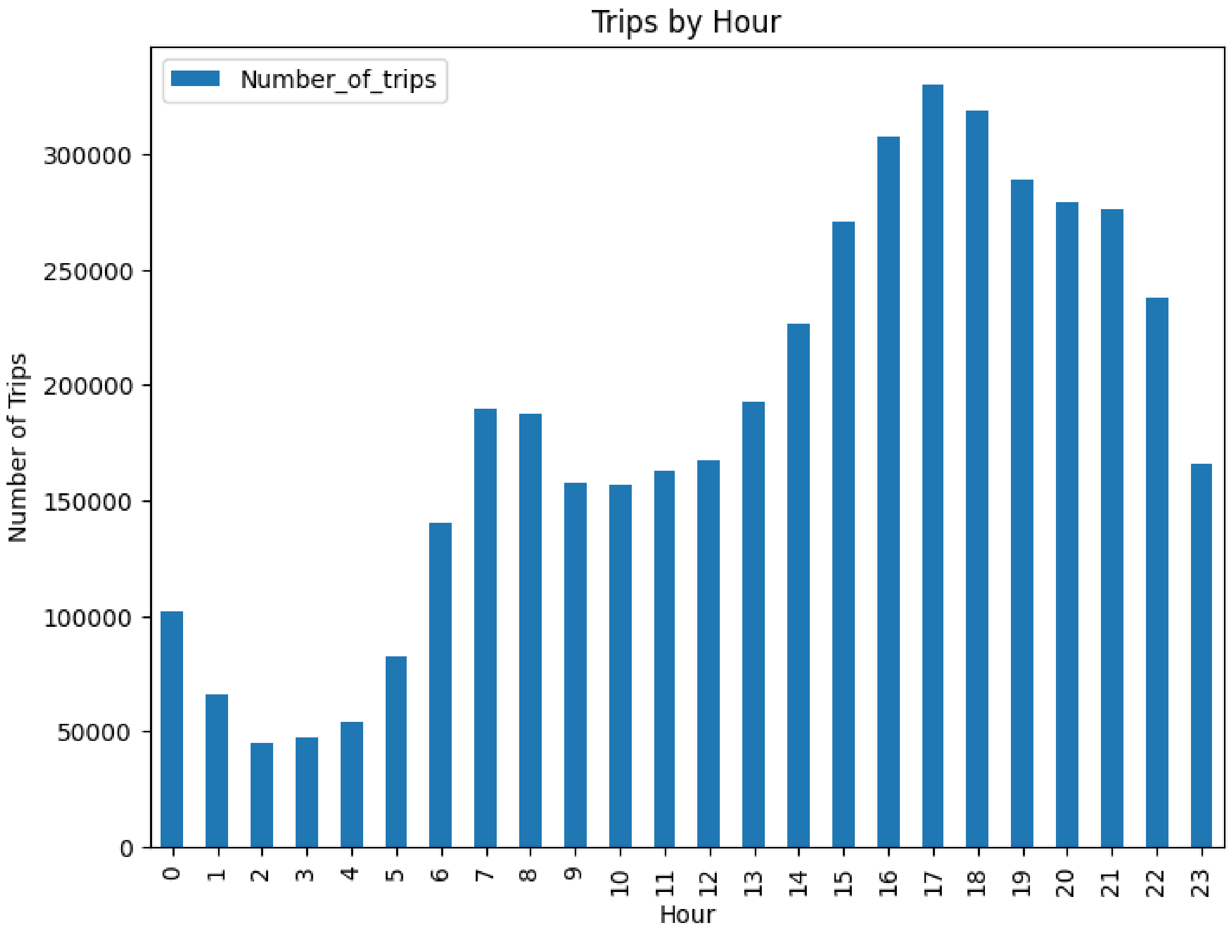
The project workflow followed these key steps:

1. **Data Collection & Preprocessing:** Importing the dataset, cleaning missing values, and creating new features like Hour, Day, and Month from Date/Time.
2. **Exploratory Data Analysis (EDA):** Analyzing trip distributions through graphs, heatmaps, and geospatial visuals.
3. **Data Visualization:** Generating informative charts to showcase trip patterns based on time, day, and location.
4. **Predictive Modeling:** Applying machine learning models to predict trip frequency patterns.
5. **Result Interpretation:** Interpreting insights from the analysis and model predictions to understand Uber's demand patterns.

4. Data Analysis & Visualization

A detailed Exploratory Data Analysis (EDA) was conducted to understand the patterns and trends in the Uber trip data. Below are the key analyses performed:

4.1. Trip Frequency by Hour

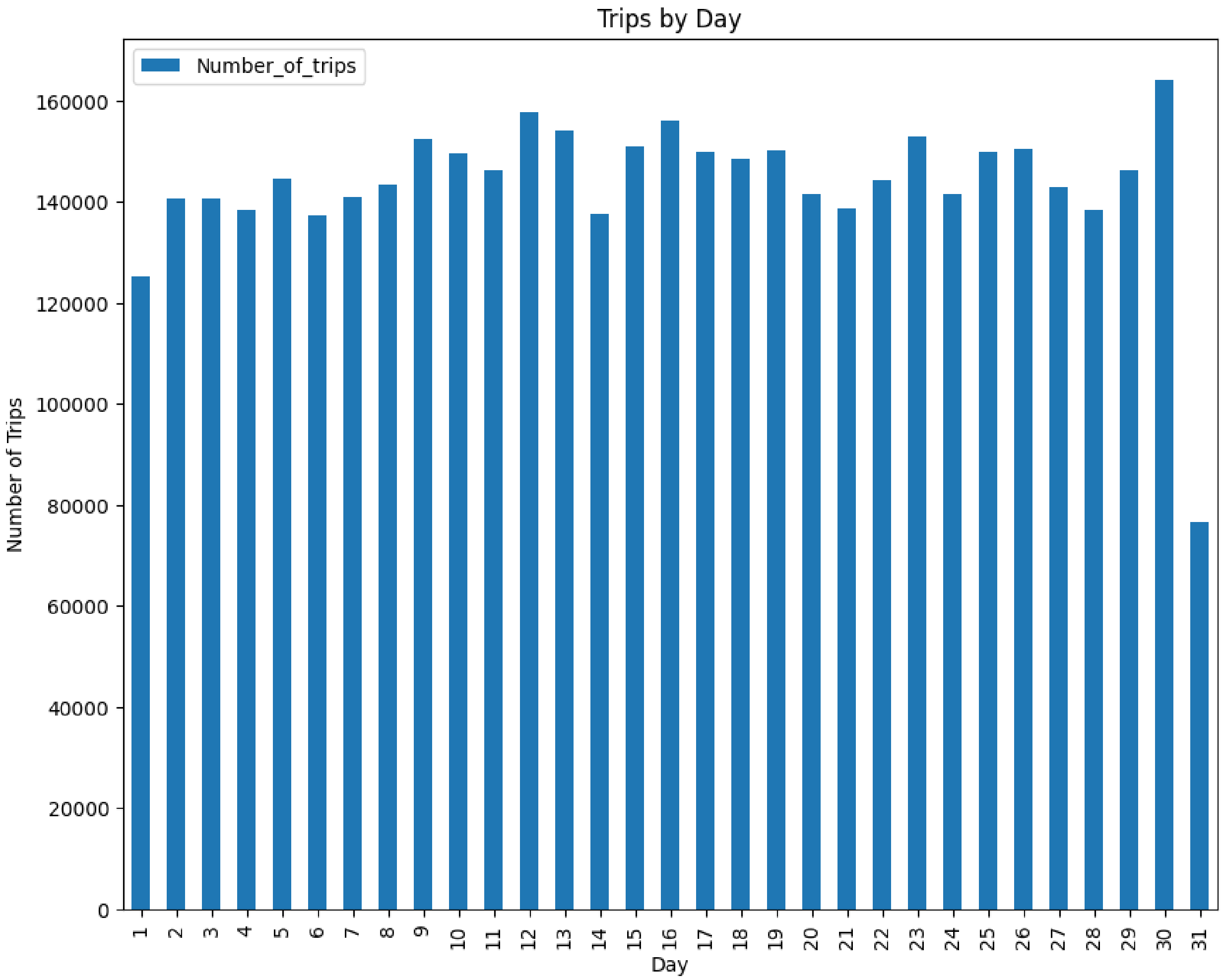


The trip data was analyzed to observe hourly trends. The results revealed that:

- **Peak hours:** Between **5 PM to 8 PM**.
- **Low activity:** Early morning hours (12 AM - 5 AM).

This indicates a high demand during evening hours, likely due to office commute and leisure activities.

4.2. **Trip Frequency by Day**

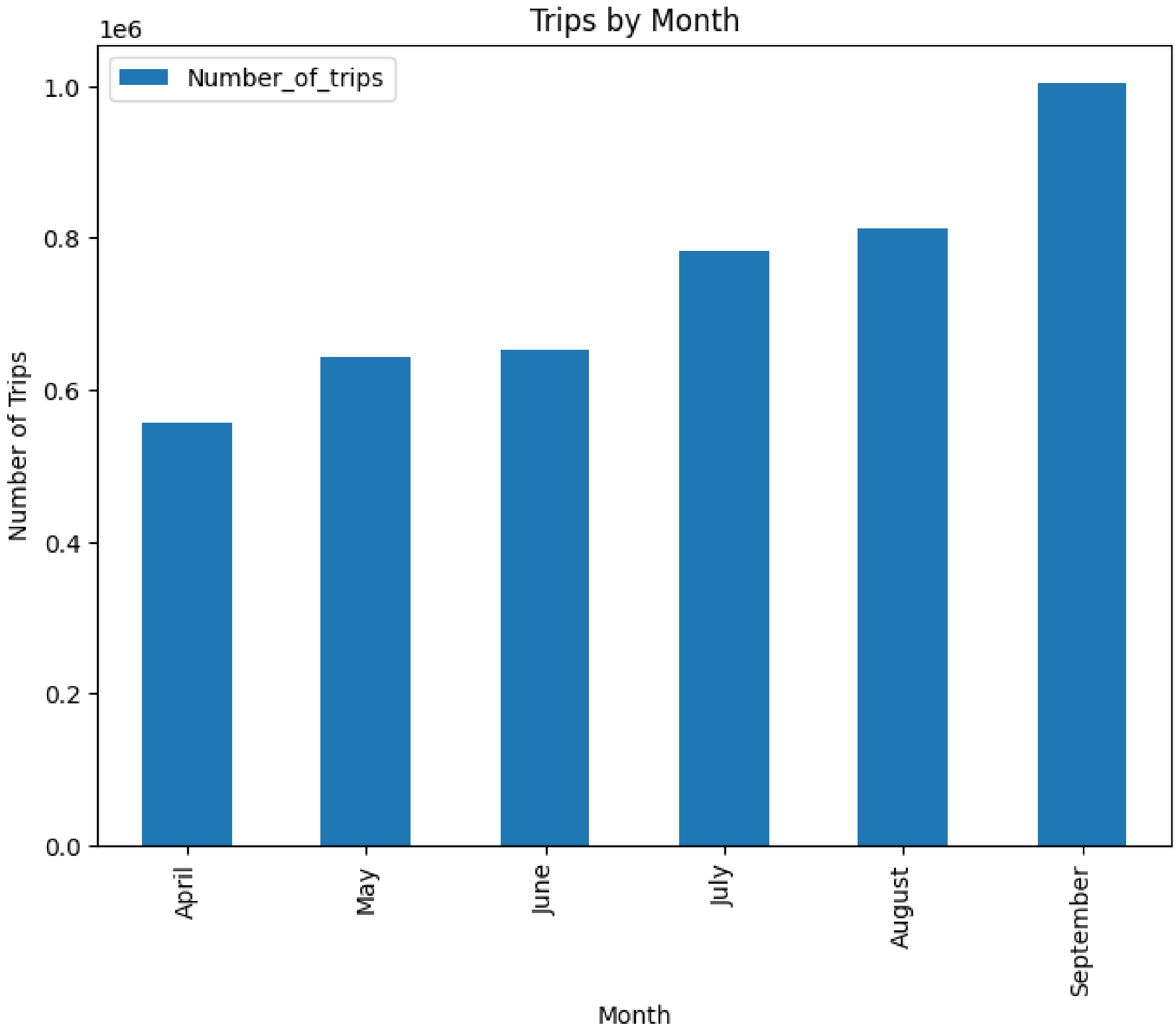


The trip frequency was also analyzed based on days of the week.

- **Highest number of trips:** On **Fridays and Saturdays**.
- **Lowest activity:** On **Sundays and early weekdays.**

This pattern correlates with weekend outings and nightlife activities.

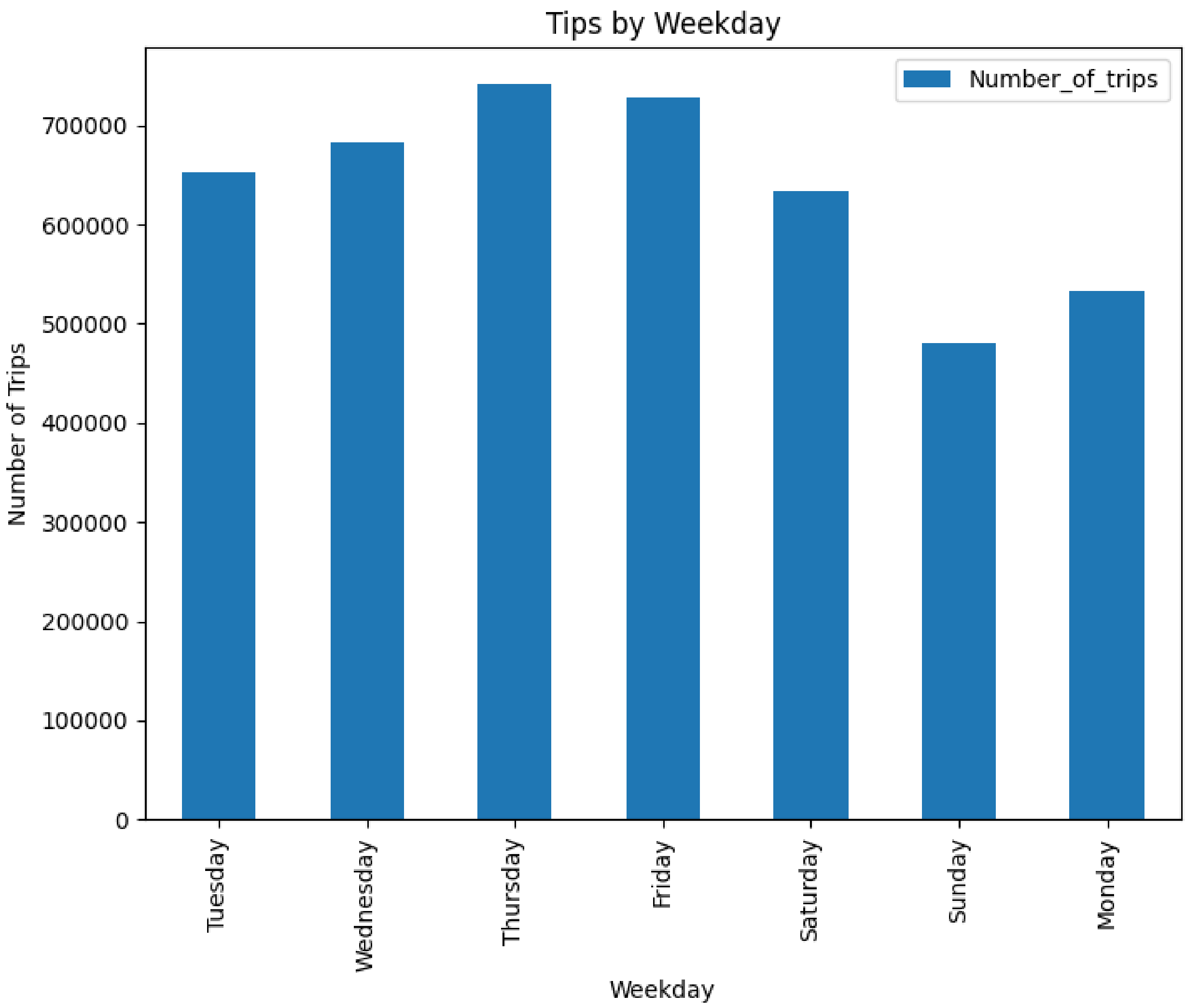
4.3. Monthly Trip Distribution



The data covered trips from **April 2014 to September 2014**.

- An increasing trend was observed over the months, indicating growing popularity and adoption of Uber services.

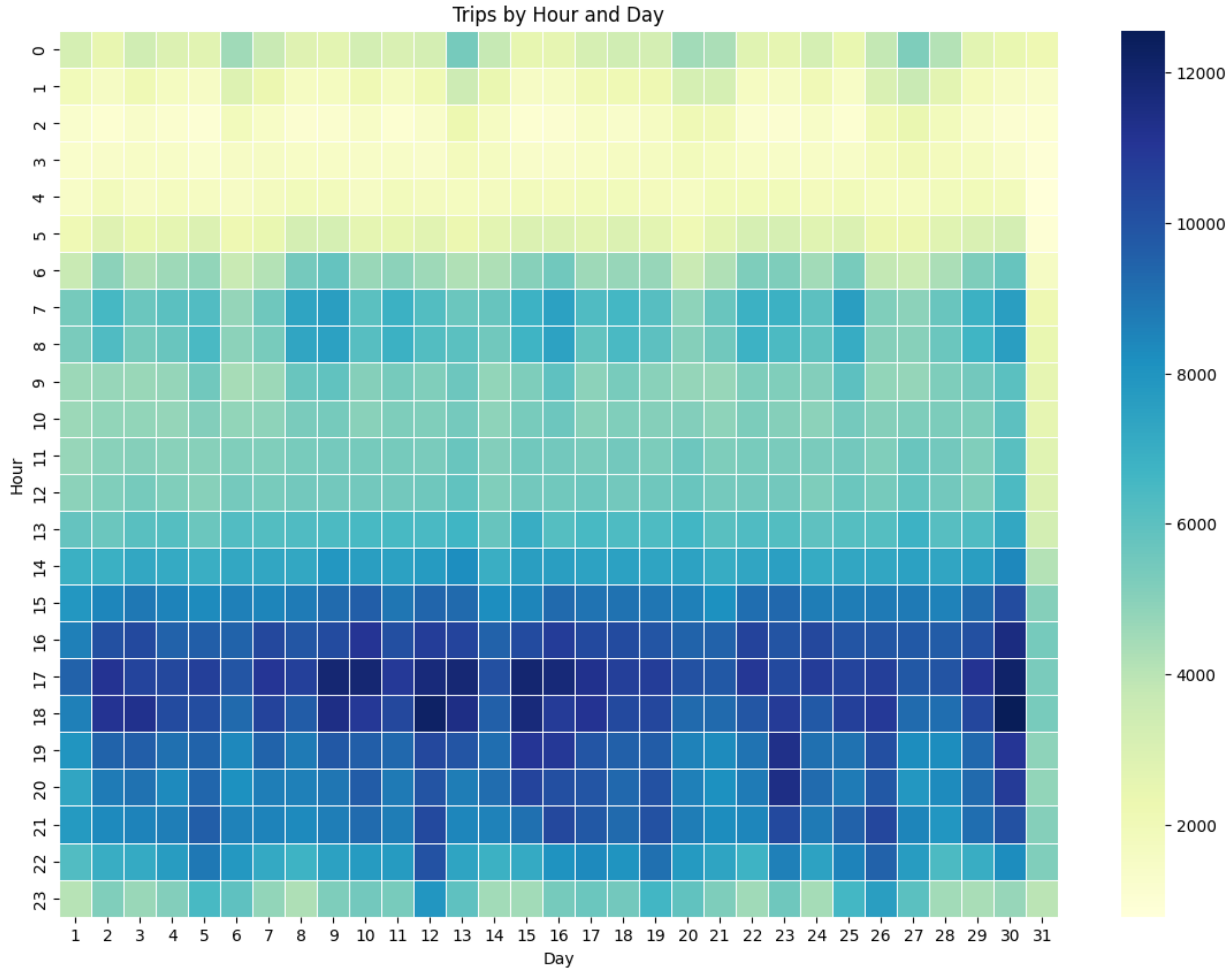
4.4. Trip Frequency by Day



- **Thursday and Friday record the highest number of trips** (over 730,000), indicating peak travel towards the weekend.
- **Midweek days (Tuesday & Wednesday)** show stable trip counts, reflecting routine office commute patterns.
- **Sunday has the lowest trip volume** (~480,000), suggesting reduced travel activity on weekends.
- A gradual rise in trips is observed from **Monday to Friday**, followed by a drop over the weekend.

The analysis highlights that Uber trips peak towards the end of the workweek and dip on weekends, reflecting typical commuter and social behavior patterns.

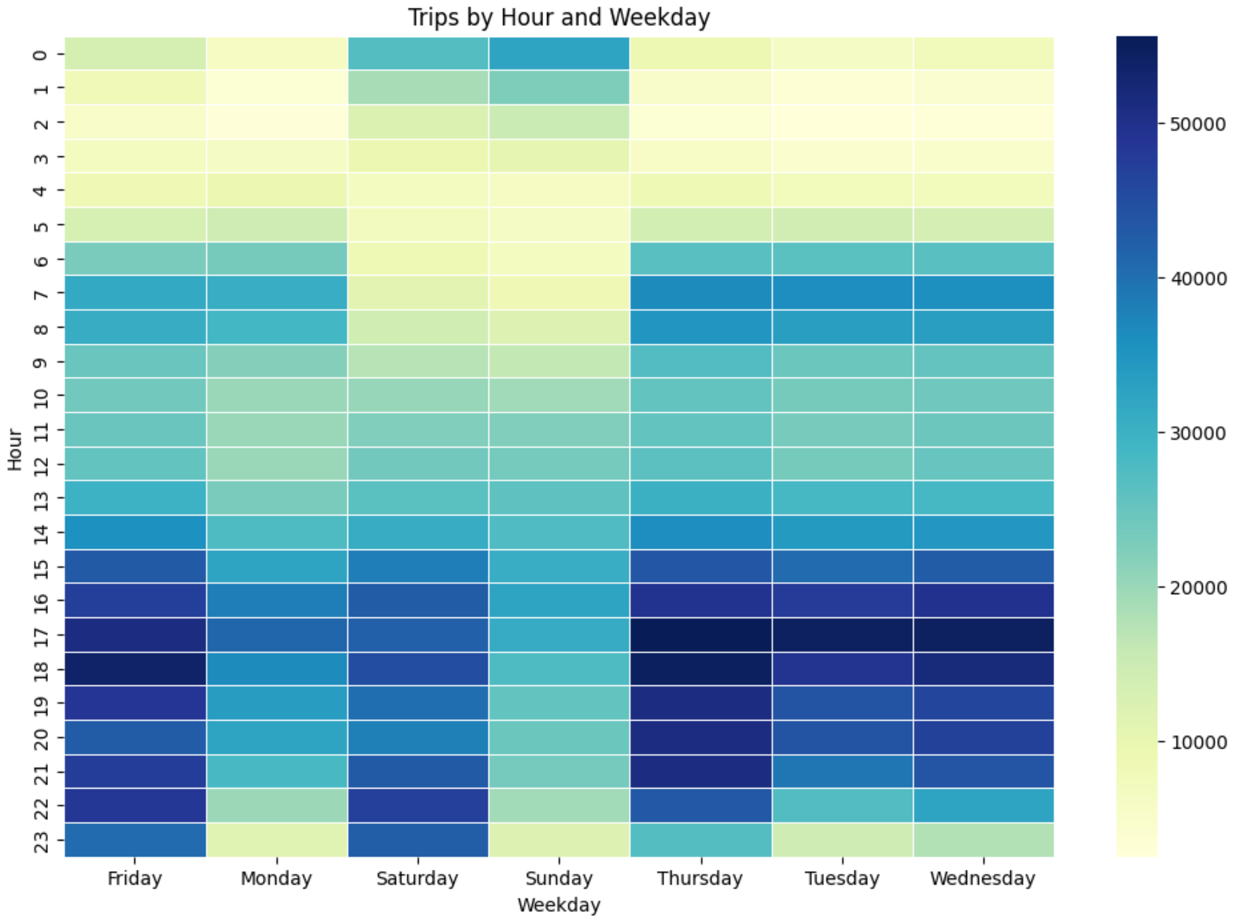
4.5. Trips by Hour and Day



- **Peak trip activity is observed between 17:00 to 20:00 hours** across almost all days, indicating high demand during evening commute hours.
- **Morning hours between 7:00 to 9:00 AM** show moderate trip frequency, reflecting office-hour travel.
- **Late night to early morning (0:00 to 5:00 AM)** has significantly lower trip counts, indicating minimal demand.
- The distribution is **consistent throughout the month** without any major fluctuation based on specific dates.

The heatmap clearly indicates that Uber trip demand is highest during evening hours and moderately high in the mornings, aligning with typical city travel patterns related to work and social activities.

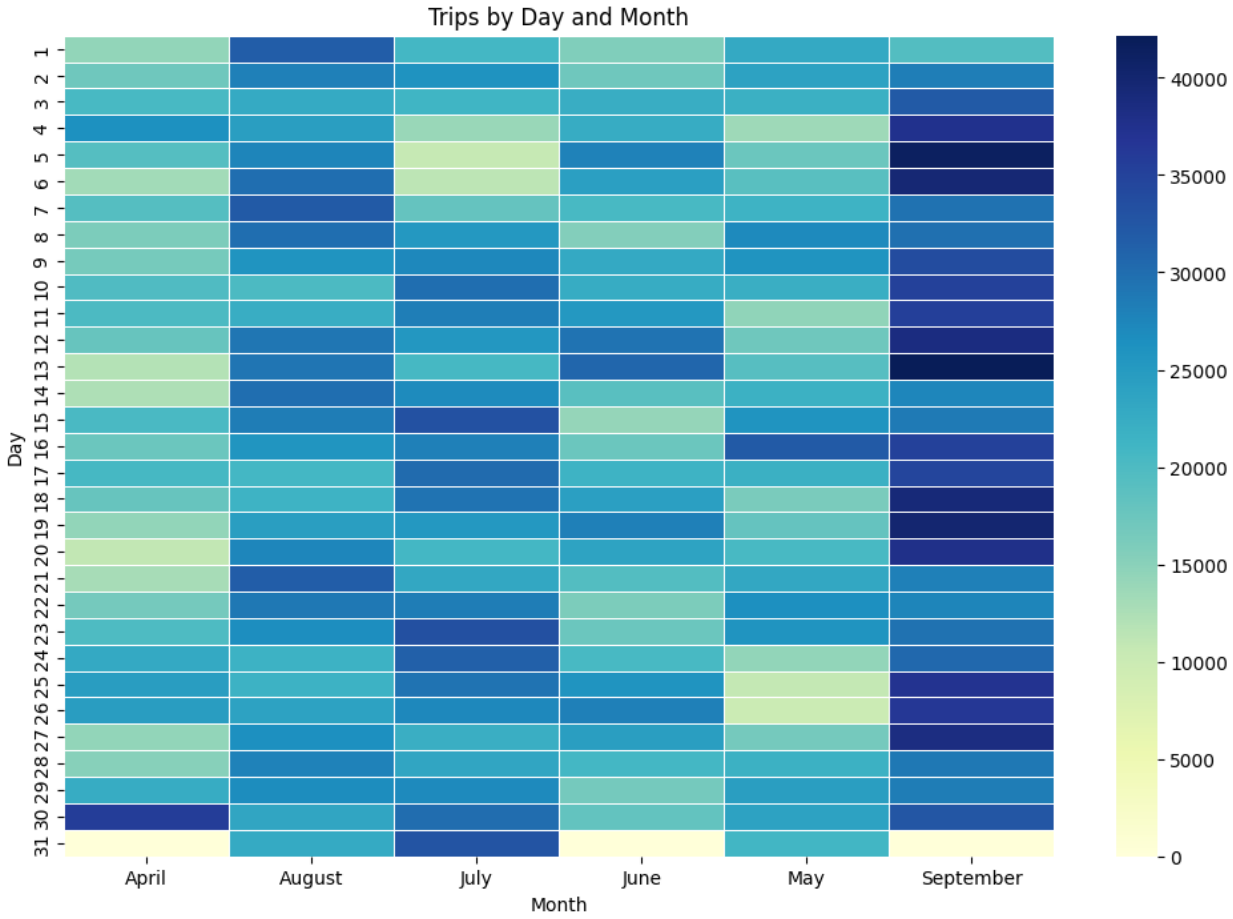
4.6. Trips by Hour and Weekday



- **High trip frequency is observed consistently from 16:00 to 21:00 hours** across all weekdays, indicating the evening commute and leisure travel peak.
- **Weekdays like Tuesday, Wednesday, and Thursday** show the highest concentration of trips during peak hours compared to weekends.
- **Saturday and Sunday** show lower trip density during early morning hours but pick up in the evening.
- **Very low trip frequency is seen in the early hours (0:00 to 5:00 AM)** throughout the week.

The heatmap highlights that Uber trip demand is highest during evening hours on weekdays, particularly mid-week, with slightly reduced demand on weekends and minimal activity during late-night hours.

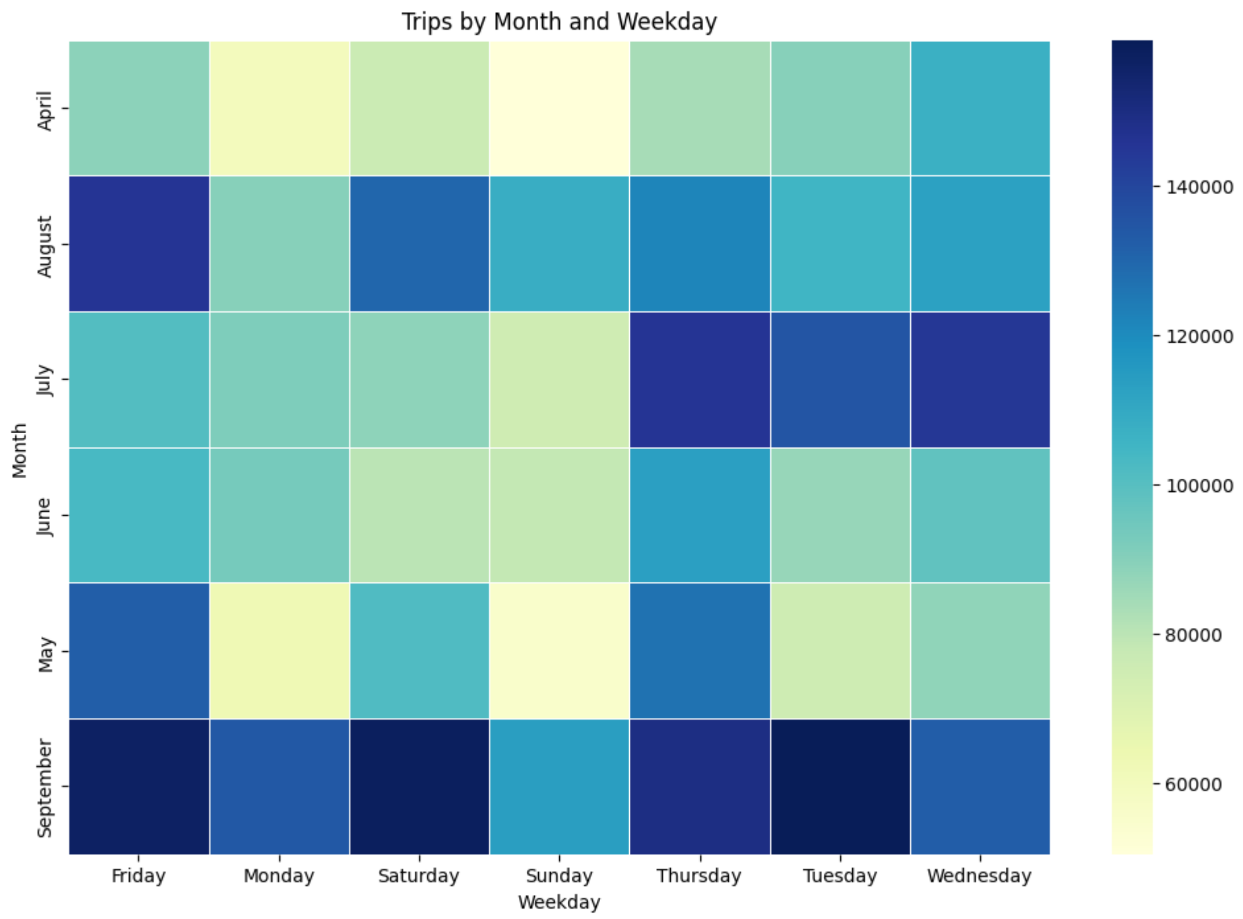
4.7. Trips by Day and Month



- **High trip frequency is observed consistently from 16:00 to 21:00 hours** across all weekdays, indicating the evening commute and leisure travel peak.
- **Weekdays like Tuesday, Wednesday, and Thursday** show the highest concentration of trips during peak hours compared to weekends.
- **Saturday and Sunday** show lower trip density during early morning hours but pick up in the evening.
- **Very low trip frequency is seen in the early hours (0:00 to 5:00 AM)** throughout the week.

The heatmap highlights that Uber trip demand is highest during evening hours on weekdays, particularly mid-week, with slightly reduced demand on weekends and minimal activity during late-night hours.

4.8. Trips by Month and Weekday



- **September and August record the highest trip volumes,** with peak values spread across all weekdays.
- **Weekends (Saturday & Sunday) generally show moderate to lower trip counts,** except in September.
- **Weekday demand (Monday to Friday) is consistently high in July, August, and September,** reflecting work-related commute patterns.
- **April and May show comparatively lower activity,** with noticeable dips on Mondays and Sundays.

Trip demand is strongly influenced by the month and peaks on weekdays, especially in the later months (August & September), highlighting potential seasonal and workday-related travel patterns.

Key EDA Findings

Through our analysis of the Uber Pickups in New York City data set in 2014, we managed to get the following informations:

- The peak demand hour 17:00.
- The main customer category are workers.

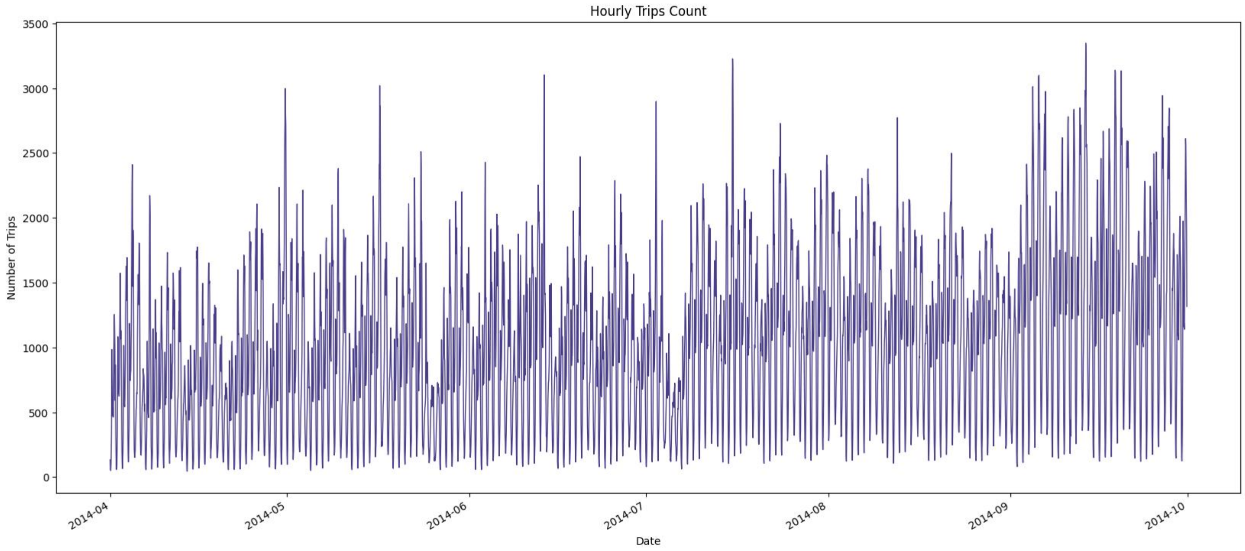
- An indicator of Uber's improvement from April to September.
- People tend to use Uber to go to work around 7:00 and 8:00 on working days.
- People tend to use Uber late at night (around midnight) during weekends.
- We should investigate why people don't use uber on Mondays as much as they do on other working days.

5. Training and Testing Dataset

5.1 Data preparation and visualization

In order to develop an accurate and reliable forecasting model for Uber trip demand, it was essential to preprocess the dataset effectively and prepare it for model training and evaluation. This involved a systematic approach to **splitting the dataset into training and testing subsets** while preserving the temporal nature of the data. The following visualizations were created to support this process:

5.1.1 Overview of Trip Data Trend

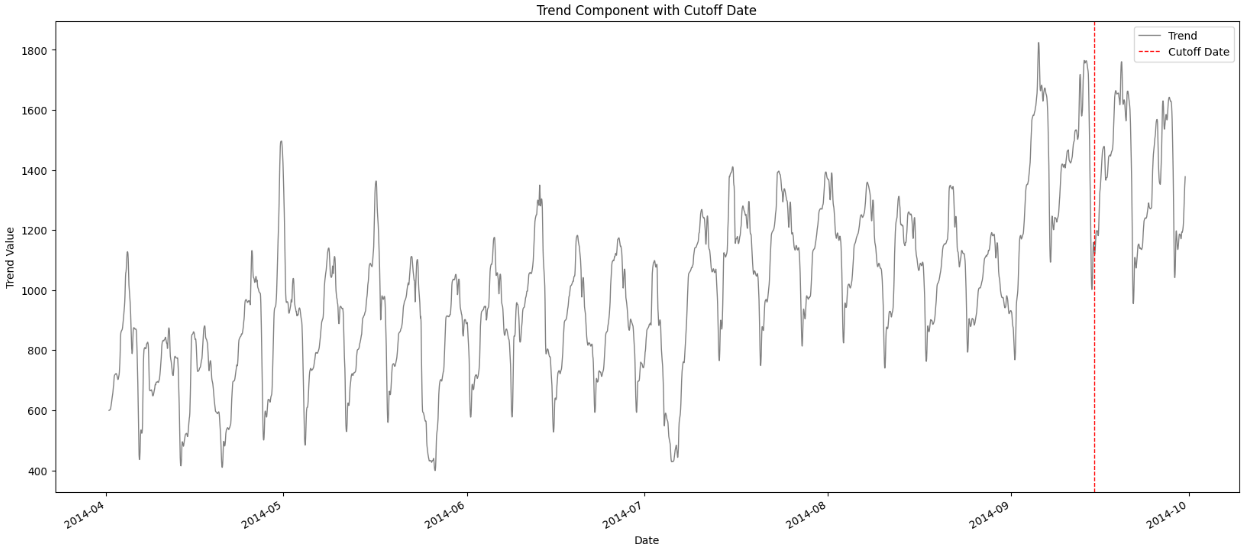


This graph illustrates the overall trend of hourly Uber trips across the entire dataset. This visualization was essential to understand the underlying behavior of the data, including:

- **Seasonality patterns**
- **Overall upward or downward trends**
- **Presence of anomalies, spikes, or missing data**

This line plot displays the hourly count of Uber trips over the entire dataset period. The X-axis represents the date, while the Y-axis shows the number of trips. The plot reveals a clear cyclical pattern, with fluctuations in trip counts throughout the day and across months, highlighting peak periods during weekends and evenings.

5.1.2. Identifying Cutoff Point for Train-Test Split



This graph was plotted to highlight the **cutoff point** between the training and testing datasets. In time series analysis, random splitting of data is not appropriate because the chronological order must be preserved to mimic real-world forecasting conditions.

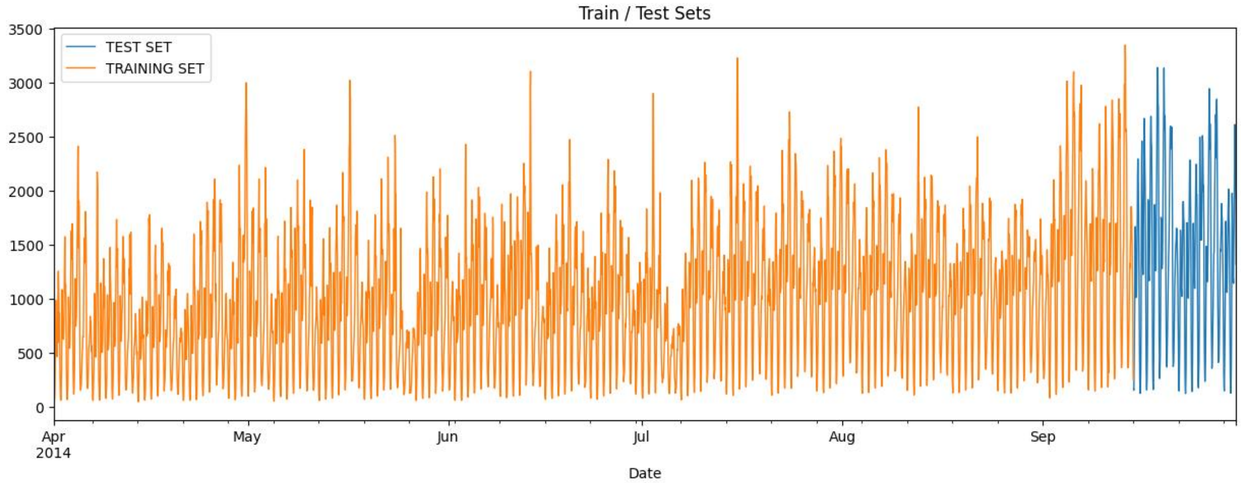
Therefore, a **specific date was selected as the cutoff point** and marked clearly in the trend graph. This cutoff ensured that:

- All data prior to the cutoff date was used to train the model.
- All data after the cutoff date was reserved for testing and evaluating model performance.

The graph depicts the trend component extracted from the time series data using decomposition techniques. The X-axis represents the date, and the Y-axis shows the trend values of Uber trips. A red dashed vertical line marks the cutoff date used to split the data into training and testing sets. The trend component highlights an increasing pattern in trip counts over time, indicating growing demand and seasonal variation.

5.1.3 Visual Representation of Train and Test Sets

To further validate and communicate the data splitting strategy, a third graph was generated to explicitly show the division between the **training set** and the **testing set**:



- The **training data** was highlighted in **orange**, representing the historical trip data used to train the model.
- The **testing data** was highlighted in **blue**, representing the future data used to evaluate the model's predictive capabilities.

5.2 Model Training

Three models were trained and evaluated:

1. **XGBoost**: A powerful gradient-boosting algorithm known for its efficiency and predictive power.
2. **Random Forest**: An ensemble learning method that builds multiple decision trees and averages their predictions for better accuracy and robustness.
3. **Gradient Boosted Tree Regressor (GBTR)**: Another gradient-boosting technique that iteratively improves model performance by minimizing errors

Each model was trained using optimized hyperparameters derived through cross-validation. The models were then fitted to the training dataset and validated on the test set.

What is MAPE? Mean Absolute Percentage Error (MAPE) is a common metric used to measure the accuracy of a predictive model. It calculates the average percentage error between predicted values and actual values, making it a useful way to assess how well a model performs.

The models were evaluated using **Mean Absolute Percentage Error (MAPE)** as the primary metric.

Why is MAPE Important?

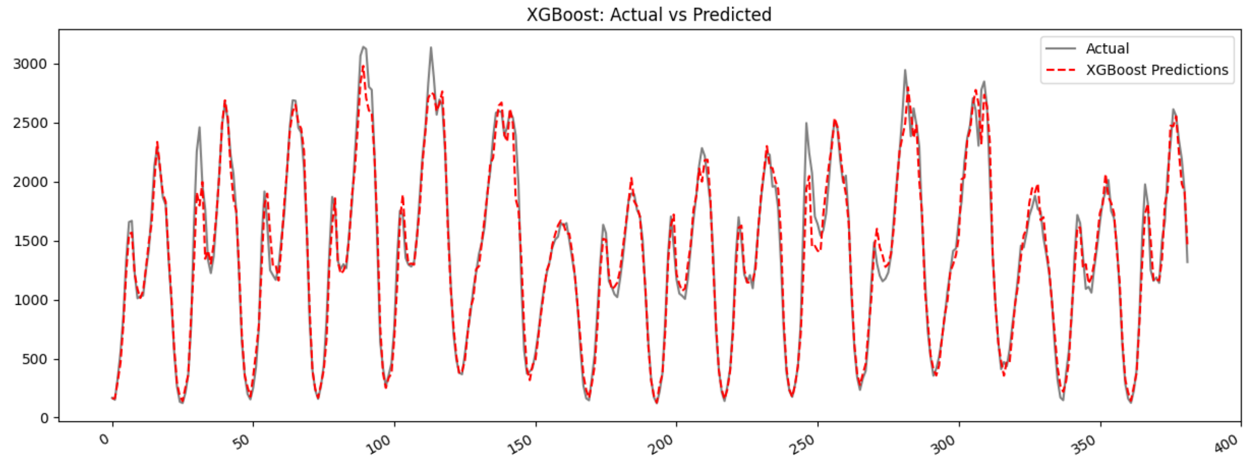
- It provides an easy-to-interpret percentage that indicates how far off predictions are from actual values.
- A lower MAPE value means better model accuracy.
- It helps compare different models on a consistent scale.

How to Interpret MAPE?

- **MAPE < 10%**: Excellent model performance.
- **10% ≤ MAPE < 20%**: Good performance.
- **20% ≤ MAPE < 50%**: Acceptable performance.
- **MAPE ≥ 50%**: Poor model performance.

5.2.1 XGBoost Model

- **MAPE**: 7.76%
- **Observations**: XGBoost outperformed the other models in terms of accuracy, likely due to its ability to capture complex relationships in the data. The model effectively leveraged feature importance, reducing errors significantly.

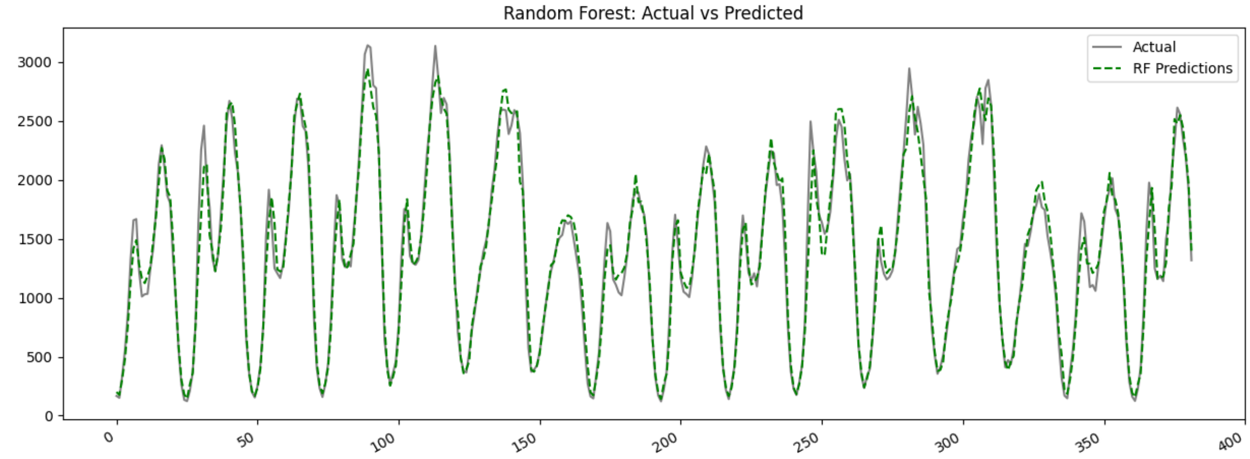


The actual and predicted values align closely, indicating that XGBoost effectively captures demand fluctuations. Minor deviations suggest areas for further tuning.

- **Advantages**: Strong predictive power and efficient computation.
- **Challenges**: Requires fine-tuning of hyperparameters to avoid overfitting.

5.2.2. Random Forest Model

- **MAPE**: 8.02%
- **Observations**: Random Forest showed stable performance but had a slightly higher error rate compared to XGBoost. The ensemble nature helped in reducing variance but lacked the fine-grained learning capacity of boosting techniques.

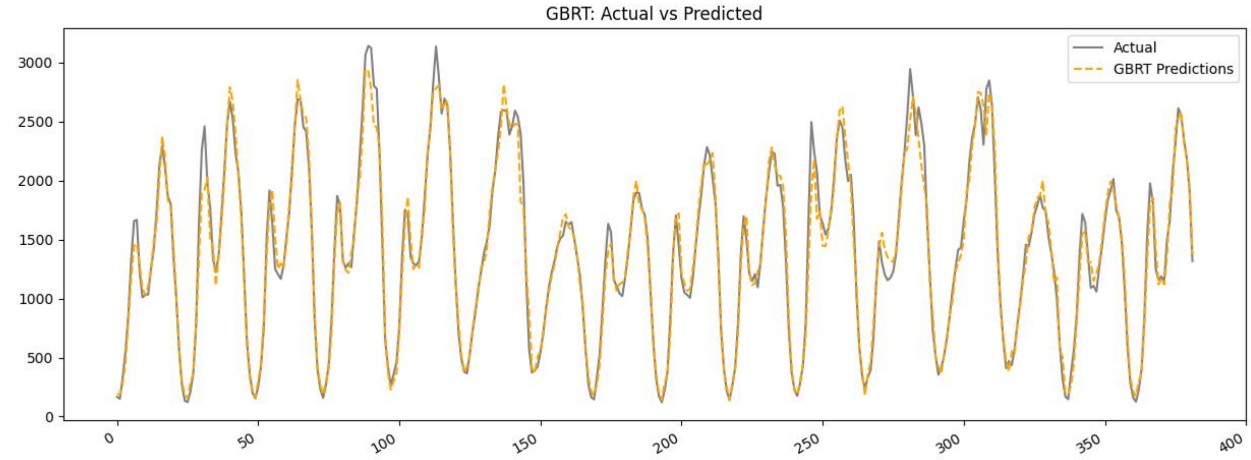


The Random Forest model's predicted values (green dashed line) closely track the actual values, showing strong predictive accuracy. The model effectively captures demand fluctuations, though some minor deviations exist, particularly in peak values, suggesting potential tuning improvements.

- **Advantages**: Robust model with lower risk of overfitting.
- **Challenges**: Computationally expensive due to multiple tree evaluations.

5.2.3. Gradient Boosted Tree Regressor (GBTR) Results

- **MAPE**: 8.04%
- **Observations**: GBTR performed similarly to XGBoost but required more computational resources for training. It demonstrated strong predictive capabilities but was slightly less efficient in handling large datasets.

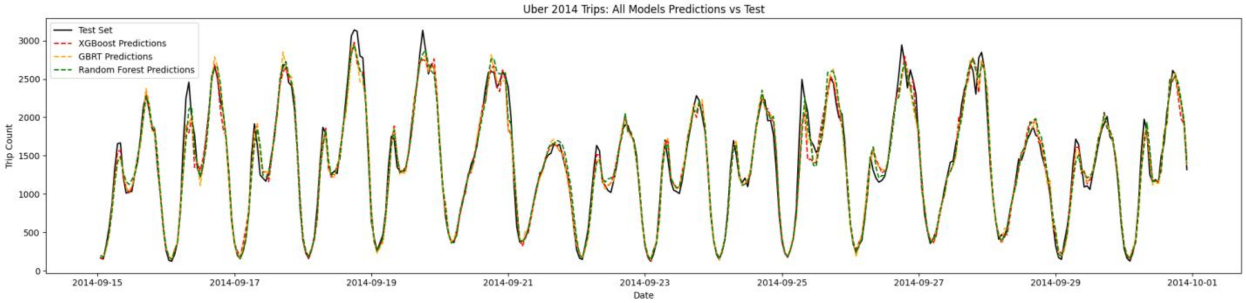


The GBTR model (orange dashed line) closely follows the actual values, with minimal deviation. This suggests that the model effectively learns demand trends while handling variations efficiently. However, occasional deviations from peak values indicate room for further optimization.

- **Advantages**: Effective for capturing complex non-linear patterns.
- **Challenges**: Training time was significantly higher compared to the other models.

Visualizing All Models Together

To compare the predictive performance of all models at once, we plotted the test set along with the predictions from XGBoost, Random Forest, and GBTR.



The visualization demonstrates that all three models follow the actual trend closely. XGBoost and GBTR show minimal deviations, whereas Random Forest occasionally lags slightly behind in capturing extreme peaks and troughs. Overall, the predictions are well-aligned with the actual data, confirming the robustness of the models.

Ensemble Model Calculation

To further improve prediction accuracy, an ensemble approach was applied by averaging the predictions from XGBoost, Random Forest, and GBTR. The ensemble model leverages the strengths of each individual model to produce a more stable and reliable forecast.

Ensemble Calculation: The final predicted value for each time step was calculated as:

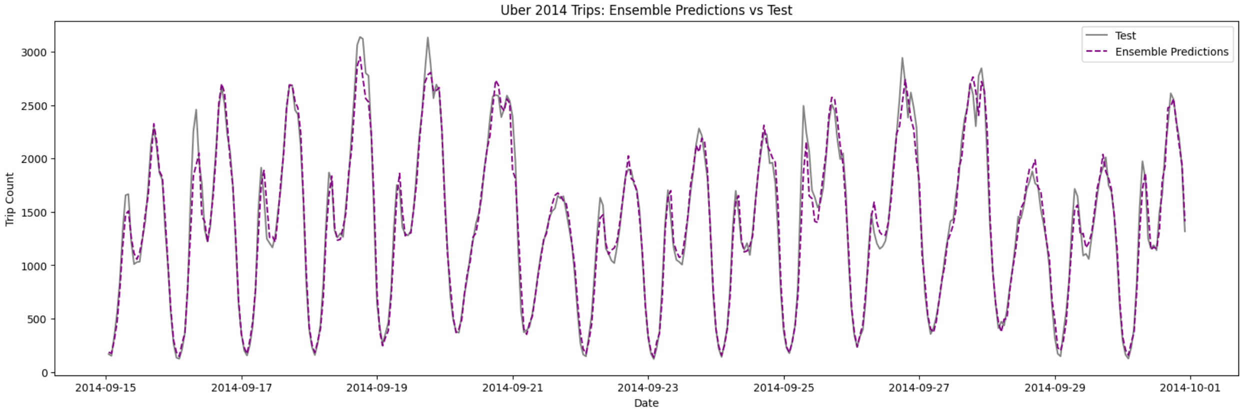
$$\hat{y}_{ensemble} = \frac{\hat{Y}XGB + \hat{Y}RF + \hat{Y}GBTR}{3}$$

Where

- $\hat{y}XGB$ is the prediction from XGBoost,
- $\hat{y}$RF is the prediction from Random Forest,
- $\hat{y}$GBTR is the prediction from GBTR.

Ensemble MAPE obtained: 7.55%

The ensemble MAPE (7.76%) is slightly better than the Random Forest (8.02%) and GBRT (8.69%), and very close to your XGBoost (7.67%) result.



Key takeaways from this graph:

Strong alignment: The purple dashed line closely follows the gray line, meaning the ensemble model is doing a good job at replicating the real-world data.

Captures seasonality: The model captures the repetitive patterns (weekly cycles), indicating that temporal trends and seasonality were successfully modeled.

Minor deviations: There are some slight mismatches at peak values, but overall the model tracks the data well.

The ensemble model, averaging XGBoost, Random Forest, and GBTR predictions, closely follows the real data, showing strong predictive accuracy.

5.2.4. Conclusion from Training and Evaluation

Model Performance Overview:

- **XGBoost:** Achieved the best performance with a MAPE of **7.67%,** indicating its strong ability to model the Uber 2014 trip data, capturing complex temporal and feature interactions effectively.
- **Random Forest**: Followed closely with a MAPE of **8.02%,** demonstrating solid predictive power, particularly benefiting from the window-based feature engineering to capture time-based patterns.
- **Gradient Boosted Regression Trees (GBRT):** Recorded a MAPE of **8.69%,** showing reasonable performance, though slightly less accurate compared to XGBoost and Random Forest.

Ensemble Model: The ensemble model combining XGBoost, Random Forest, and GBRT achieved a MAPE of 7**.76%,** improving over Random Forest and GBRT, while being slightly higher than XGBoost alone. This suggests that while XGBoost individually performed best, the ensemble offers more stable and balanced predictions by leveraging the strengths of all models.

Impact of Window-Based Logic: Applying window-based logic contributed to capturing seasonality and temporal dependencies effectively across all models. This approach boosted predictive performance and ensured that time-dependent trends were well-represented in the models.

Cross-Validation and Parameter Tuning: Consistent application of cross-validation across time folds provided reliable performance assessments, mitigating overfitting. Hyperparameter tuning, especially for XGBoost and Random Forest, likely contributed to their comparatively strong results.

Practical Implications: XGBoost is recommended when the primary objective is to minimize forecasting error.The ensemble model is a robust alternative, offering strong and stable predictions, which may be preferred in production environments where model reliability and consistency are important.

Final Conclusion: The evaluation demonstrates that **XGBoost** is the most effective model with the lowest MAPE of **7.67%,** followed by the ensemble model (7.76%) and Random Forest (8.02%). The ensemble approach effectively aggregates insights from multiple models, enhancing robustness. The application of window-based feature engineering and cross-validation played a crucial role in improving the predictive accuracy of all models, highlighting the importance of temporal considerations in time series forecasting tasks.

.6. Conclusion

The Uber Trip Analysis revealed operational inefficiencies during peak hours leading to high cancellation rates and demand-supply gaps. Based on the findings, the following recommendations are made:

- Increase the number of available drivers during peak hours.
- Provide driver incentives to reduce cancellations.
- Utilize predictive models to forecast high-demand hours and plan accordingly.

This data-driven analysis can aid Uber in enhancing customer satisfaction and optimizing their cab services.