UNIFIED MENTOR
YOUR SKILL, SUCCESS & JOURNEY

PROJECT – 2

NAME - AAGASH.M
UNID - UMID11062542183

PROJECT TITLE - **“Uber Trip Analysis & Forecasting”**

Abstract

This report details the process of forecasting hourly Uber trip demand in New York City using machine learning. By analyzing trip data from April to September 2014, this project identifies temporal patterns and builds predictive models to estimate future ride volume. The raw trip data was transformed into an hourly time series, which was then used to train three distinct regression models: XGBoost, Random Forest, and a Gradient Boosted Tree Regressor (GBTR). The models were evaluated using Mean Absolute Percentage Error (MAPE). The XGBoost model emerged as the most accurate, achieving a MAPE of 8.37% on the test set. This report covers the full project lifecycle, from data preparation and exploratory analysis to model implementation, evaluation, and conclusion.

Table of Contents

1. Introduction

1.1. Project Background

Ride-hailing services like Uber have become an integral part of urban transportation. The vast amount of trip data generated by these services presents a significant opportunity for data analysis and predictive modeling. Accurate demand forecasting is critical for operational efficiency, enabling better resource allocation, dynamic pricing strategies, and improved customer service.

1.2. Problem Statement

The primary challenge is to accurately forecast the number of Uber trips on an hourly basis in New York City. By understanding the underlying patterns of demand, Uber can optimize the availability of drivers to meet rider needs, particularly during peak hours.

1.3. Objectives

The main objectives of this project are:

- To process and transform raw Uber trip data into a usable time series format.
- To analyze and visualize the data to uncover trends, seasonality, and other temporal patterns.
- To build and train multiple machine learning models to forecast hourly trip demand.
- To evaluate and compare the performance of these models to identify the most accurate one.

2. Data Acquisition and Preparation

2.1. Data Source

The dataset was obtained via a Freedom of Information Law (FOIL) request by FiveThirtyEight from the NYC Taxi & Limousine Commission (TLC). It contains detailed trip-level data for Uber pickups from April to September 2014.

2.2. Data Description

The raw dataset for 2014 consists of six CSV files, one for each month. Each record represents a single Uber trip and contains the following fields:

- Date/Time: The exact timestamp of the pickup.
- Lat: The latitude of the pickup location.

- Lon: The longitude of the pickup location.
- Base: The TLC base company code affiliated with the trip.

2.3. Data Cleaning and Transformation

The raw data is not structured for time series forecasting. The following transformation was performed:

1. **Concatenation:** All six monthly CSV files were combined into a single dataframe.
2. **Timestamp Conversion:** The Date/Time column was converted to a proper datetime object.
3. **Aggregation:** The trip-level data was resampled into **hourly intervals**. The total number of trips (or "counts") within each hour was calculated. This aggregation transformed the dataset into a univariate time series, with the hourly trip count as the target variable.

3. Exploratory Data Analysis (EDA)

3.1. Hourly Trip Volume Analysis

The complete time series of hourly trips from April to September 2014 was plotted.

Analysis: The plot reveals strong cyclical patterns, indicating daily and weekly seasonality. There is also a noticeable increase in overall trip volume towards the end of the dataset, particularly in September.

3.2. Trend and Seasonality Decomposition

To better understand the underlying components of the time series, a seasonal decomposition was performed.

Analysis:

- **Trend:** The trend component confirms a relatively stable period followed by a significant upward trend in September 2014.
- **Seasonality:** A consistent daily pattern of peaks and troughs is clearly visible, confirming the 24-hour cycle of ride demand.

4. Methodology

4.1. Feature Engineering: The Sliding Window

To enable the machine learning models to make predictions, a "sliding window" approach (also known as creating lagged features) was used.

- **Input:** The input for predicting the trip count for a specific hour is the sequence of actual trip counts from the **previous 24 hours**.
- **Output:** The model uses this 24-hour sequence to predict a single value: the trip count for the **next hour**.

This method allows the models to learn the relationship between recent historical data and the immediate future.

4.2. Model Selection

Three powerful, tree-based ensemble models were chosen for this regression task:

1. **XGBoost (Extreme Gradient Boosting):** Known for its high performance and efficiency.
2. **Random Forest Regressor:** An robust model that is less prone to overfitting.
3. **Gradient Boosted Tree Regressor (GBTR):** Another boosting algorithm similar in principle to XGBoost.

Hyperparameter tuning was performed using GridSearchCV with time-series cross-validation to find the optimal settings for each model.

4.3. Train-Test Split Strategy

Due to the time-dependent nature of the data and the observed trend change, a chronological split was used instead of a random one:

- **Training Set:** Data from **April 1, 2014, to September 15, 2014**.
- **Test Set:** Data from **September 15, 2014, to September 30, 2014**.

This approach ensures that the model is trained on past data to predict future data, simulating a real-world scenario.

4.4. Evaluation Metric

The primary metric used to evaluate model performance is the **Mean Absolute Percentage Error (MAPE)**.

MAPE = (1/n) * Σ(|Actual - Predicted| / |Actual|) * 100

MAPE is intuitive because it expresses the average prediction error as a percentage, making it easy to interpret the accuracy of the forecasts. A lower MAPE indicates a better model.

5. Results and Discussion

5.1. Model Performance Comparison

Each trained model was used to make predictions on the unseen test set. The resulting MAPE scores are summarized in the table below.

Model	Mean Absolute Percentage Error (MAPE)
XGBoost	**8.37%**
Random Forest Regressor	9.61%
Gradient Boosted Tree Regressor (GBTR)	10.02%
Ensemble Model (Weighted Average)	8.60%

Discussion: The XGBoost model delivered the best performance with the lowest MAPE of 8.37%. This indicates that, on average, its hourly predictions were within 8.37% of the actual trip counts. The Random Forest and GBTR models also performed well, and the ensemble

model provided a robust alternative that slightly underperformed the standalone XGBoost model.

5.2. Machine Learning Input and Output

- **Clear Input:** For each prediction, the model was given a list of 24 numbers representing the trip counts of the preceding 24 hours.
- **Clear Output:** The model produced a single number representing the forecasted trip count for the next hour.

5.3. Visualizing Model Predictions

To qualitatively assess performance, the predictions from all models were plotted against the actual trip counts from the test set.

[Insert Chart: The final comparison plot showing Test data vs. predictions for all models, from page 23 of the source PDF.]

Analysis: The chart visually confirms the quantitative results. All models successfully captured the daily cyclical patterns of peaks and troughs. The XGBoost predictions (red dashed line) appear to track the test data (gray line) most closely, consistent with its lower MAPE score.

6. Conclusion

6.1. Summary of Findings

This project successfully demonstrated the effectiveness of machine learning for forecasting Uber trip demand. By transforming raw data into an hourly time series and using a sliding window for feature engineering, we were able to train models that accurately predict future ride volume. The **XGBoost model proved to be the most effective predictor**, achieving a MAPE of **8.37%**.

DATASET LINK -
"C:\Users\USER\OneDrive\Desktop\Internship\Project 2\Uber-Jan-Feb-FOIL.csv"

DATA VISVALIZATION COLAB LINK

https://colab.research.google.com/drive/17YpVFSFcG2BE-jgpbBHRIkmsWCH-_IHC?usp=sharing