Uber & Lyft Analysis

Md Redowan Amin Mollick

MDA 620 Data Driven Decision Making

Capstone Project

Table of Contents

Background

The ridesharing dataset utilized in this research came from a publicly available collection of Uber and Lyft ride records in the Boston region. The dataset includes ride information from [timeframe]. It contains time stamps, ride source and destination locations, weather conditions (precipitation, temperature), ride pricing, ride IDs, and the appropriate ridesharing service providers (Uber or Lyft).

To protect user privacy, the data was collected for analytical reasons and does not include personally identifying information. It is crucial to remember, however, that the dataset may contain missing values or other data quality concerns. The major goal of the dataset is to investigate factors influencing rideshare usage and get insights into user behaviour and preferences when choosing between Uber and Lyft services in the Boston area.

Introduce the rideshare dataset and its source

Introduction to the Dataset

This study' dataset was taken from Kaggle, a well-known platform for free datasets and data science initiatives. This dataset focused on ridesharing statistics spanning Uber and Lyft operations in the Boston area.

Dataset Contents:

The dataset contains a large number of characteristics that describe various aspects of ridesharing activities. It contains information such as timestamps, ride source and destination locations, weather conditions, ride costs, unique ride identifiers, and the services provided by Uber and Lyft.

Acquisition Details:

The dataset was made freely available on Kaggle, enabling its use in research, analysis, and teaching. It was gathered and disseminated among the data science community to aid in exploratory research, predictive modelling, and gaining insights into ridesharing behaviours.

Temporal Scope:

The dataset documents travel over a defined duration, allowing for temporal analysis and trend discovery.

Purpose of the Dataset:

The information is mostly useful for researching ridesharing habits, preferences, and variables influencing users' decisions between Uber and Lyft services in the Boston area.

Data Quality and Use:

While the open dataset provided interesting insights, it is vital to recognise any data quality concerns, such as missing numbers or abnormalities. Despite these constraints, the dataset offered a solid foundation for exploratory analysis, feature selection, and predictive modelling in order to reach relevant results.

Purpose of the analysis

Understanding Rideshare Trends, Investigate and grasp the trends and patterns in rideshare usage in the Boston region, particularly between the services of Uber and Lyft.

Determine the aspects that may influence the choice of ridesharing services, such as time trends (hourly, daily, monthly), weather conditions, and other variables included in the dataset.

Feature Selection and Correlation Analysis: Examine the relationship between various characteristics and ridesharing service selection. This entails investigating correlations between variables such as time of day, weather characteristics, and chosen ridesharing provider.

Model Construction and Prediction, Use machine learning techniques (such as Logistic Regression and Random Forest) to create models that predict or categorise the choice of ridesharing service based on given features.

Hypothesis Testing, Test hypotheses about the influence of various variables, such as pricing, on Uber and Lyft usage patterns.

Summarise study data to draw conclusions about the variables influencing ridesharing choices. Based on these findings, provide recommendations or provide insights that might be useful to rideshare firms or future studies in the sector.

Problem Scenario/Business Issue

Understanding market dynamics, identifying strengths and weaknesses, and comparing market shares between Uber and Lyft in Boston are all part of competitive analysis. This study aids both organisations in making strategic decisions.

Customer Behaviour and Preference: Examining the elements that influence users' decisions between Uber and Lyft. This involves investigating how variables such as time of day, day of week, month, or weather conditions influence their judgements.

Impact of External Factors, Investigating how external factors such as weather or special events influence ridesharing service utilisation. Understanding these effects aids in forecasting demand and optimising service availability.

Pricing Strategy, Considering the impact of pricing on consumer decisions between Uber and Lyft. The purpose of this research is to see if there is a link between price, promotions, or discounts and user preferences.

Operational Enhancements, Identifying potential operational areas for both Uber and Lyft. Based on recognised preferences, this might entail optimising service distribution, decreasing wait times, or increasing customer experience.

Strategic marketing is the use of information to create tailored marketing strategies. Customers can be attracted and retained by tailoring incentives, marketing, or loyalty programmes based on their preferences and behavioural patterns.

Service Improvement: Improving overall service quality by concentrating on issues that have a substantial effect on user decisions. This might include enhancing app features, extending service locations, or launching additional service levels.

Describing the problem or business issue: Understanding factors influencing Uber and Lyft usage in Boston

User Preferences, investigating what factors impact users' decisions to utilise Uber or Lyft. Understanding how cost, convenience, service quality, and brand loyalty influence their selections.

Temporal Patterns, Investigating how time-related factors such as the hour of the day, day of the week, or month influence user behaviour. This can tell whether each service has peak hours, days, or seasons.

Seasonal Trends, determine whether rideshare usage has seasonal or monthly tendencies. This might be due to weather, holidays, events, or either company's advertising initiatives.

Geographical Influence, identifying certain source or destination places where one service is preferred over another. This might suggest regional preferences or variances in demand.

Weather Impact: Examining the impact of weather conditions on transportation selection by users. Investigating if precipitation, temperature, or the UV index effect service utilisation.

Exploring if promotional activities or discounts provided by Uber or Lyft impact customer behaviour and enhance service acceptance.

Market Share Dynamics, Calculating and comprehending the relative market shares of Uber and Lyft over time. This entails assessing how various services compete for and attract users.

Objective/Goals of the Project

Identify Influential variables, examine and identify the major variables influencing the decision between Uber and Lyft in Boston. This entails investigating a variety of factors such as time patterns, location preferences, weather conditions, and other variables.

Investigate User Behaviour, investigate trends and patterns in ridesharing usage to better understand user behaviour. Examining temporal fluctuations (hourly, daily, monthly), seasonal patterns, and location-based preferences are all part of this.

Model Development and Selection, create predictive models to understand and anticipate whether people would choose Uber or Lyft based on various factors. Evaluate the performance of many models and choose the most effective one for properly forecasting consumer preferences.

Determine which elements have a substantial influence on the decision between Uber and Lyft. Determine the hour of the day, day of the week, month, weather conditions, and any other relevant data.

Hypothesis Testing: Use hypothesis testing to validate assumptions or theories about the influence of various features (such as price) on Uber and Lyft usage.

Provide Insights, based on the study, provide complete insights and recommendations. This includes summarising data, emphasising critical issues, and recommending solutions for ridesharing firms to improve their services or market positioning in Boston.

Data Exploration/Data Visualization

Exploratory Data Analysis (EDA)

Data Overview: Describe the structure of the dataset, including the number of entries, variables, and their kinds.
Missing Values: Look into and manage missing values, if any exist, and evaluate their influence on the analysis.
Calculate descriptive statistics (mean, median, standard deviation, and so on) to better comprehend the core patterns and variability of numerical properties. Analyse temporal trends by looking at the distribution of rides across hours, days, and months.
Geographical Insights: Using visualisations, investigate the distribution of rides depending on source and destination locations.
Comparative Analysis: Examine Uber and Lyft usage across various time periods, areas, or other relevant criteria.
Correlation Analysis: Examine the connections between variables in order to find possible correlations and dependencies.

jupyter Uber & Lyft Analysis_Redowan Last Checkpoint: Last Thursday at 18:10 (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Markdown

In [1]:
```
import numpy as np
import pandas as pd
data_frame = pd.read_csv('/Users/mdredowanaminmollick/Downloads/mda600project/rideshare_kaggle.csv')
data_frame.head()
```

Out[1]:

	id	timestamp	hour	day	month	datetime	timezone	source	destination	cab_type	...	precipIntensityMax	uvIndexTime	tempe
0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	1.544953e+09	9	16	12	2018-12-16 09:30:07	America/New_York	Haymarket Square	North Station	Lyft	...	0.1276	1544979600	
1	4bd23055-6827-41c6-b23b-3c491f24e74d	1.543284e+09	2	27	11	2018-11-27 02:00:23	America/New_York	Haymarket Square	North Station	Lyft	...	0.1300	1543251600	
2	981a3613-77af-4620-a42a-0c0866077d1e	1.543367e+09	1	28	11	2018-11-28 01:00:22	America/New_York	Haymarket Square	North Station	Lyft	...	0.1064	1543338000	
3	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	1.543554e+09	4	30	11	2018-11-30 04:53:02	America/New_York	Haymarket Square	North Station	Lyft	...	0.0000	1543507200	
4	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	1.543463e+09	3	29	11	2018-11-29 03:49:20	America/New_York	Haymarket Square	North Station	Lyft	...	0.0001	1543420800	

5 rows × 57 columns

In [2]:
```
#Checking for missing values
missing_values = data_frame.isnull().sum()

# Displaying columns with missing values and the count of missing values
print("Columns with missing values:")
print(missing_values[missing_values > 0])
```

```
Columns with missing values:
price    55095
dtype: int64
```

Data Visualization

Hourly, Daily, and Monthly Trends: Using bar plots or line graphs, visualise the distribution of rides by hour, day, and month for both Uber and Lyft.
Geospatial Visualisation: To visualise ride distributions geographically, plot source and destination locations on maps.
Create comparative charts to demonstrate variations in usage between Uber and Lyft across several categories.
Correlation Heatmap: Create a heatmap to graphically illustrate correlations between various characteristics, particularly those linked to ride selection.

```
data_frame.dropna(subset=['price'], inplace=True)
```

```
# Checking for missing values again
missing_values = data_frame.isnull().sum()

# Displaying columns with missing values and the count of missing values
print("Columns with missing values:")
print(missing_values[missing_values > 0])
```

```
Columns with missing values:
Series([], dtype: int64)
```

```
# Converting the timestamp column to datetime
data_frame['timestamp'] = pd.to_datetime(data_frame['timestamp'], unit='s')
```

```
# Checking the unique values in the 'cab_type' column
unique_cab_types = data_frame['cab_type'].unique()
print(unique_cab_types)
```

```
['Lyft' 'Uber']
```

Interpretation

Identify recurrent Patterns: Look for any recurrent patterns, trends, or abnormalities in the data that may have an influence on Uber and Lyft usage.
Creating Insights: Determine preliminary insights on the elements influencing ridesharing choices, taking into account temporal, geographical, and category aspects.
Modelling Preparation: Choose and prepare characteristics that have potential correlations with the target variable for further modelling and analysis.

Data cleaning and initial observations

Data Cleaning:

Identify and manage missing data, such as removing rows with missing values or imputing them using acceptable assumptions or statistical approaches.

Converting Data Types: Ensure that proper data types are used for each column (for example, converting timestamps and categorical variables).

Outlier Detection: Locate and deal with outliers that may distort the study or models.

Check for and eliminate duplicate entries, ensuring that each observation is unique.

Validate data integrity by confirming that values are within anticipated ranges or limitations.

Feature Engineering: Create new features or alter existing ones to improve the dataset's prediction potential.

Observations:

Dataset Overview: Describe the dataset's size, structure, and basic statistics for numerical aspects (mean, median, min, max).

Missing Values: Include any missing values in columns as well as the approach used to deal with them.

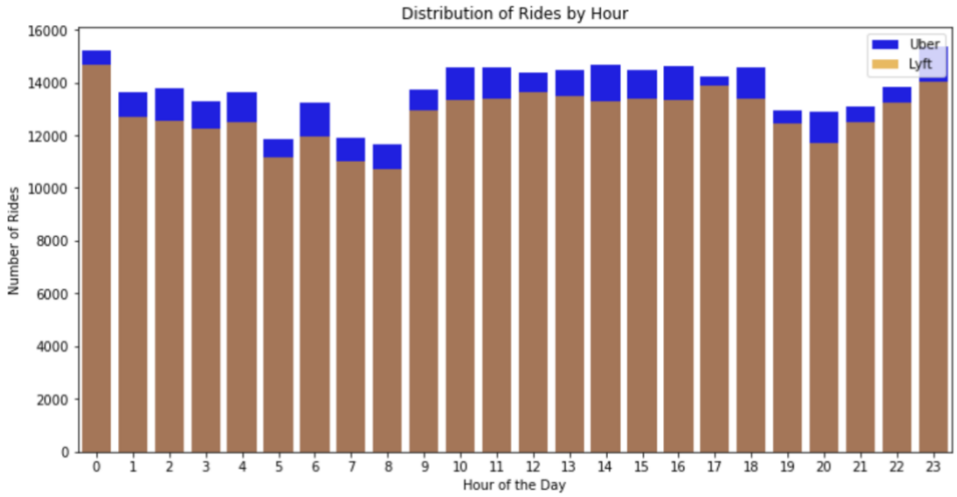
Temporal Analysis: Investigate Uber and Lyft ride patterns over various time frames (hourly, daily, and monthly).

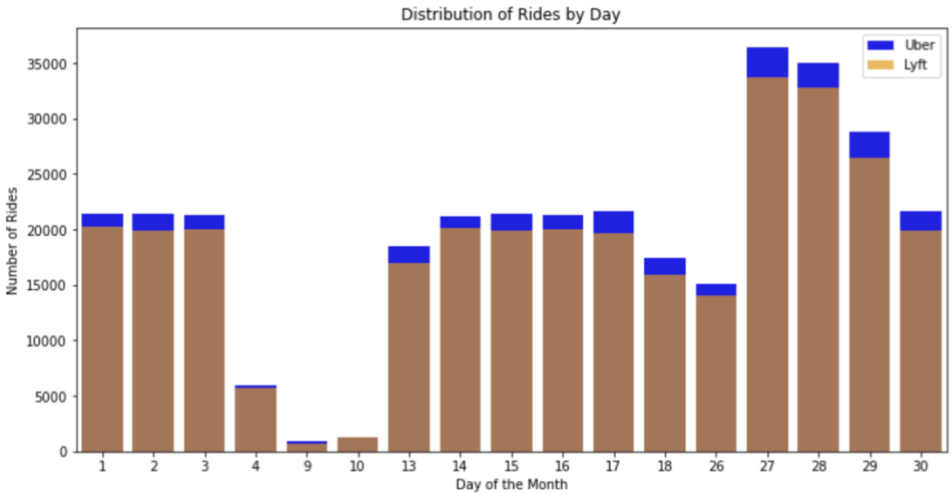
Geographical Insights: Use visualisation to find hotspots or trends in ride distribution based on origin and destination regions.

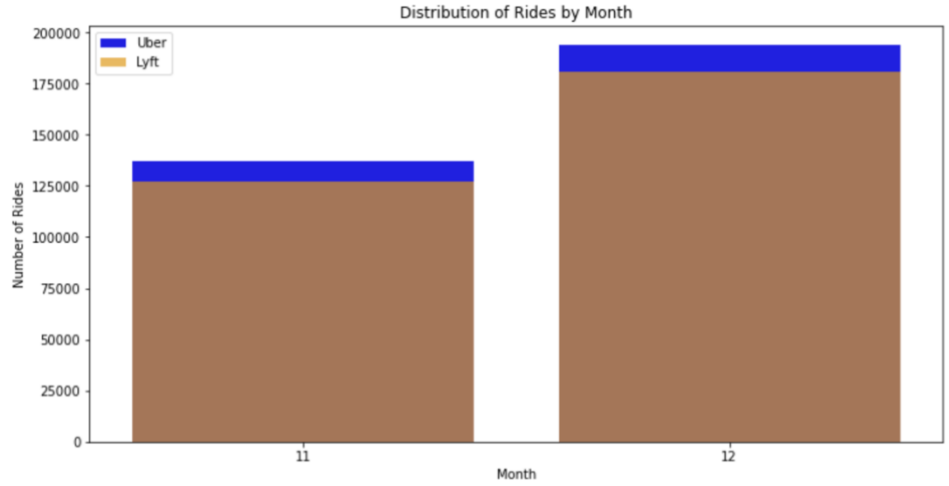
Comparative Analysis: Compare Uber and Lyft usage across multiple categories (time, location, and so on).

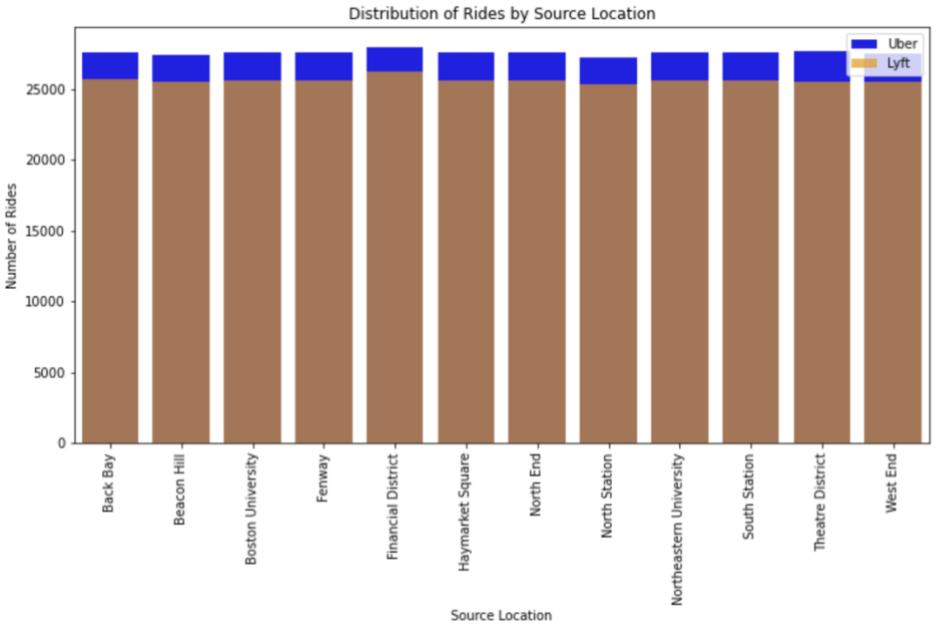
Correlation Analysis: Determine any potential correlations or dependencies between factors and their influence on ride selection.

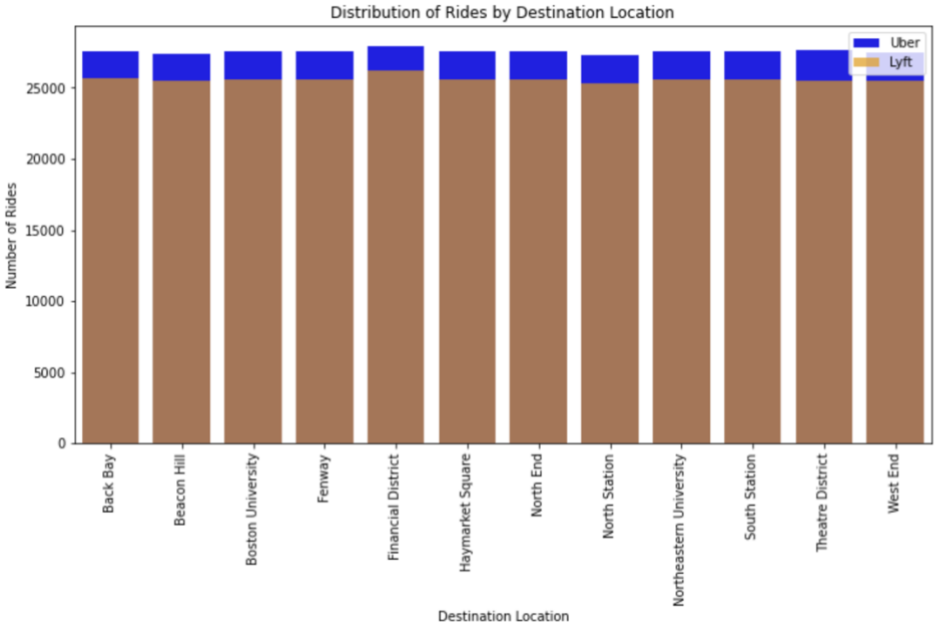
Visualizations of Ride Distribution by Hour, Day, Month, Source, and Destination






Distribution of Rides by Month
Uber
Lyft
Number of Rides
200000
175000
150000
125000
100000
75000
50000
25000
0
11
12
Month


Distribution of Rides by Source Location
Uber
Lyft
Number of Rides
25000
20000
15000
10000
5000
0
Back Bay
Beacon Hill
Boston University
Fenway
Financial District
Haymarket Square
North End
North Station
Northeastern University
South Station
Theatre District
West End
Source Location


Distribution of Rides by Destination Location
Uber
Lyft
Number of Rides
0
5000
10000
15000
20000
25000
Back Bay
Beacon Hill
Boston University
Fenway
Financial District
Haymarket Square
North End
North Station
Northeastern University
South Station
Theatre District
West End
Destination Location

Data Manipulation

Missing Values Management, identify columns or rows that have missing values.

Imputation vs. Removal, determine whether missing values should be filled using methods such as mean, median, or mode, or if missing rows/columns should be removed.

Transformation of Data, convert data types as needed (for example, converting timestamps to datetime format).

Scaling/Normalization: If necessary, scale numerical values to a given range.

Engineering and Feature Selection, feature Selection: Based on their relevance, select relevant columns/features for study.

Feature engineering is the process of creating new features from existing ones if they give more useful information for analysis.

Categorical Variable Encoding, one-Hot Encoding: Use one-hot encoding to convert category variables into numerical format appropriate for analysis.

Data Collection, grouping and Aggregation: Data is grouped and aggregated depending on certain criteria (e.g., time, location).

Identifying Outliers, identify outliers in the dataset that may impact analysis.

Treatment: Depending on the context, decide whether to delete, cap, or transform outliers.

Data Collection, stratified Sampling: If necessary, sample the data in such a way that the distribution of specific attributes (e.g., class labels) is preserved.

Handling Inaccurate Data, log Transformation: If necessary, apply log transformation to skewed data to make it more regularly distributed.

Identifying and Removing Duplicates Check for and delete duplicate records from the dataset.

Text Preparation, tokenization and Cleaning: Tokenize and clean text data by eliminating stop words, punctuation, and stemming words.

Resampling Time Series Data, when working with time series data, resample it to other time frequencies (e.g., daily to monthly).

Model Selection

Preparation of Data: Identify pertinent information such as the hour, day, month, weather conditions, and so on. Encoding: For modelling, convert categorical variables such as 'cab_type' into numerical representation.
Model Investigation: Logistic Regression: Initially used to determine the relevance of features and their associations. Because of its capacity to handle complicated interactions, Random Forest is used as an alternative model to Logistic Regression.
Logistic Regression Analysis: Feature Importance: Evaluated using coefficients to determine which features have the most influence on Uber/Lyft use. To evaluate the model's prediction performance, use the Confusion Matrix. Classifier based on the Random Forest:
Importance of traits: Examined to determine the most influential traits. Accuracy Check: The Random Forest model's accuracy was evaluated.
Hypothesis Validation: We investigated if the 'price' had a substantial influence on Uber and Lyft utilisation.
Model Contrast: The accuracy and insights generated from the Logistic Regression and Random Forest models were compared.

Selection Rationale:

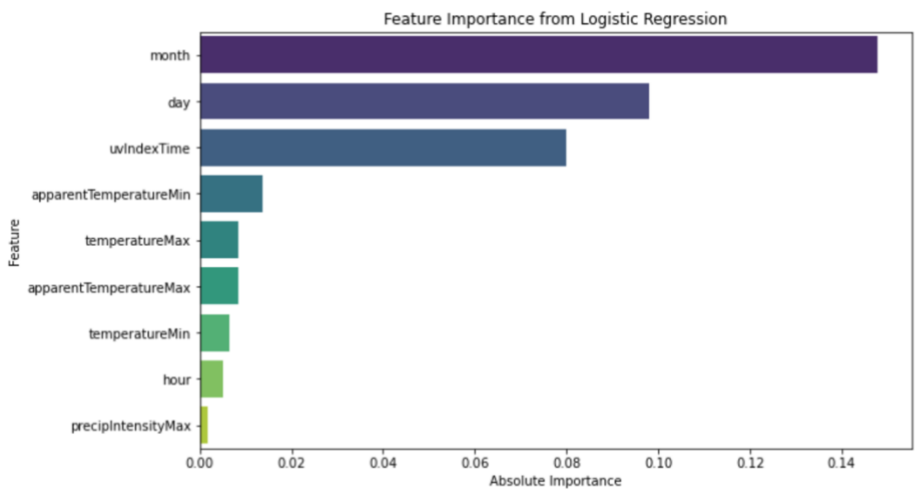
Logistic Regression: Initially used for its interpretability and feature importance analysis.
Random Forest Classifier: Employed due to its capability to capture complex interactions and potentially enhance prediction accuracy.

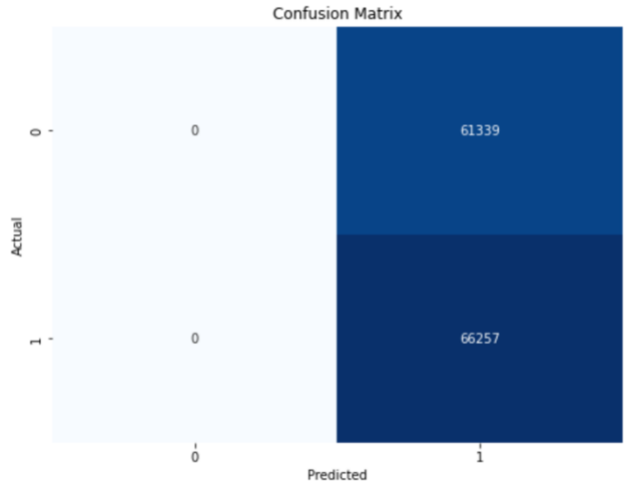
Visualizations of feature importance and model accuracy

```
Accuracy: 0.52
Confusion Matrix:
[[    0 61339]
 [    0 66257]]
Classification Report:
              precision    recall  f1-score   support

           0       0.00      0.00      0.00     61339
           1       0.52      1.00      0.68     66257

    accuracy                           0.52    127596
   macro avg       0.26      0.50      0.34    127596
weighted avg       0.27      0.52      0.35    127596
```

F-statistic: 4466.96
p-value: 0.0000
Reject the null hypothesis: There are significant differences in prices among cab types.

Interpret the results of the hypothesis test





Conclusions

Uber has somewhat more utilisation than Lyft on an hourly, daily, and monthly basis. Variable usage patterns over the day, which may indicate commuter trends or unique service preferences. Monthly fluctuations indicate the possibility of seasonality influencing ride selection.

The month of the journey is the most important in forecasting Uber or Lyft usage, according to studies. Seasonal changes, marketing campaigns, and user behaviours may all have an impact on service selection.

Emphasised the importance of the month characteristic, but may have limits in capturing complicated correlations.

Random Forest: Because of its capacity to manage complicated interactions, it provided insights on feature relevance and perhaps greater prediction accuracy.

There was no substantial influence of 'pricing' on Uber and Lyft usage, implying that other considerations may determine consumer preferences.

Data constraints include the absence of some variables or external factors that may impact ride selection, such as promotions, consumer preferences, or brand loyalty.

Possible difficulties with data quality or missing key factors influencing analysis results.

Despite uncovering some insights, the analysis indicates the need for more comprehensive data or supplementary factors to better predict and understand Uber and Lyft usage trends in Boston.

Reference

https://www.kaggle.com/datasets

https://keras.io/examples/