Uber Cab-fare Prediction

A COMPREHENSIVE REPORT ON UBER DATA ANALYSIS AND TAXI CAB-FARE PREDECTION USING MACHINE LEARNING

PROJECT SUBMITTED TO

The Indian Institute of Social Welfare & Business Management

IISWBM

In the Partial Fulfilment of the Requirements for the Degree of

MASTER OF BUSINESS ADMINISTRATION

PROJECT SUBMITTED BY

ANSUMAN PAL

University Roll No.: 107/MBA/221011

Department: MBA (Day)

Specialization: Business Analytics & Systems Management (BASM)

UNDER THE GUIDANCE AND SUPERVISION OF

Jayanta Das

CERTIFICATE FROM THE SUPERVISOR

This is to certify that Mr. Ansuman Pal, a student of Indian Institute of Social Welfare and Business Management pursuing the course of Master of Business Administration (Business Analytics and Systems Management) under the University of Calcutta has worked under my supervision and guidance for his Dissertation and prepared a research report with the title named as "Uber Data Analysis and Taxi Cab-Fare prediction using Machine Learning" which he is submitting.

Place: KOLKATA
Date:

Signature: _____
Name: Jayanta Das
Designation: Dissertation Guide & SupervisorName of the Institute: IISWBM, KOLKATA

DECLARATION

I, Ansuman Pal, solemnly declare that the project titled " Uber Data Analysis and Taxi Cab-Fare prediction using Machine Learning " submitted to the Indian Institute of Social Welfare and Business Management (IISWBM) as part of my pursuit of the Master of Business Administration (MBA) degree, is an authentic representation of my original work. I completed this project under the supervision and guidance of my mentor, Prof. Jayanta Das, during my tenure as a student at IISWBM. I also confirm that this project has not been previously submitted or used for the fulfilment of any other academic qualification, including but not limited to degrees, diplomas, fellowships, associate ships, or similar titles, at any other educational institution.

Signature: _____

Name: Ansuman Pal

Roll No.: 107/MBA/221011

MBA(Day)

BA&SM Batch:2022-2024

INDEX

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to Prof. Jayanta Das for providing me with invaluable guidance, unwavering support, and insightful supervision, whichwere essential in the successful completion of this project. Her expertise, support, andconstructive criticism have been crucial throughout this journey.

I express my gratitude to the Indian Institute of Social Welfare and Business Management (IISWBM) for providing me with the necessary resources and creating a favourable environment for carrying out this research.

I am deeply appreciative of the professors, whose insights and feedback have enhanced the quality of this research. Their expertise and dedication to academia have been inspiring.

Furthermore, I would like to thank the Head of Department, Prof. Dr. Gairik Das, for his constantsupport and encouragement. His leadership has created a conducive environment for academic excellence at IISWBM.

In addition, I would like to convey my appreciation to everyone who contributed to this endeavour in any way, whether it was directly or indirectly. I would like to express my gratitudeto my family, friends, and colleagues, for their continuous support, understanding andassistance.

Thanking You.

Ansuman Pal
Roll No: 107/MBA/221011
MBA(Day), BA&SM
Batch: 2022-2024

ABSTRACT

Uber, a pioneering force in the transportation industry, has fundamentally transformed the way people commute and interact with urban mobility. Established in 2009, Uber swiftly disrupted traditional taxi services by introducing a seamless and convenient ride-sharing platform accessible via a mobile application.
At the heart of Uber's business model lies its innovative technology, which connects riders with a network of independent drivers, offering on-demand transportation at competitive prices. This decentralized approach not only empowers drivers to work flexibly but also provides users with a reliable and efficient alternative to traditional transportation options.

Central to Uber's success is its commitment to enhancing user experience through continuous innovation, which includes features such as real-time tracking, cashless payments, and personalized promotions. Despite facing regulatory challenges and public scrutiny, Uber has expanded its offerings beyond ride-sharing, diversifying into food delivery, freight logistics, and autonomous vehicles. As Uber continues to navigate an ever-evolving landscape, its relentless pursuit of innovation and commitment to redefining urban mobility solidify its position as a transformative force in the transportation ecosystem.

Uber's business strategy revolves around leveraging technology to optimize operational efficiency and improve customer experiences. Through data-driven decision-making, Uber dynamically adjusts pricing based on supply and demand, maximizing revenue while ensuring competitive fares for users. The company's extensive driver-partner network spans across cities worldwide, providing economic opportunities for individuals seeking flexible employment. Additionally, Uber invests heavily in research and development, exploring emerging technologies such as artificial intelligence and self-driving vehicles to shape the future of transportation. Despite facing regulatory challenges and controversies related to safety and labor practices, Uber remains committed to fostering innovation and reshaping urban mobility. As the company continues to evolve and diversify its services, it seeks to address societal needs, reduce congestion, and enhance accessibility to transportation for individuals around the globe.

EXECUTIVE SUMMARY

This executive summary encapsulates the key findings and implications of a comprehensive analysis conducted on Uber data and taxi cab-fare prediction using machine learning techniques.
The analysis focused on harnessing large-scale datasets from Uber rides to gain insights into user behavior, demand patterns, and route preferences. Through advanced data analysis and feature engineering, significant factors influencing taxi cab fares, such as distance, time of day, and traffic conditions, were identified. Subsequently, machine learning algorithms, including linear regression and decision trees, were deployed to develop predictive models for estimating taxi cab fares accurately.

The performance of these models was evaluated using metrics such as Root Mean Squared Error (RMSE) and R-squared, demonstrating their efficacy in predicting fares with high accuracy. The implications of this analysis extend to optimizing pricing strategies, enhancing user experiences, and improving operational efficiency within ride-sharing industry.

By leveraging machine learning for data analysis and fare prediction, Uber can further solidify its position as a leader in urban transportation, driving innovation and delivering value to users and stakeholders alike.

Advanced Data Analytics Insights: The analysis of Uber data revealed intricate insights into user preferences, peak demand hours, and popular routes. Through sophisticated data analysis techniques, patterns emerged regarding user behavior and service utilization, empowering targeted marketing strategies and resource allocation for Uber.

Enhanced Service Optimization: By harnessing machine learning algorithms, Uber can optimize its service by strategically deploying vehicles and drivers in high-demand areas during peak hours, effectively reducing wait times and improving customer satisfaction. Predictive analytics can forecast future demand patterns, allowing proactive adjustments to operations and maintaining service quality.

Accurate Fare Prediction: Machine learning algorithms, particularly linear regression and decision trees, demonstrated their efficacy in accurately predicting taxi cab fares based on various factors like distance, time, and traffic conditions. These predictive models enable Uber to offer transparent pricing to users, improving cost estimation accuracy and overall user experience.

Strategic Business Implications**:** The insights derived from machine learning-based data analytics empower Uber to make data-driven decisions, optimize operational efficiency, and enhance revenue streams. Accurate fare prediction fosters trust and loyalty among users, driving customer retention and bolstering brand reputation.

Future Directions: Continued investment in machine learning capabilities will enable Uber to extract deeper insights, refine service offerings, and maintain a competitive edge in the ride-sharing market. Embracing emerging technologies like artificial intelligence and predictive modeling will further enhance Uber's ability to anticipate user needs and deliver personalized experiences, ensuring sustained growth and innovation.

Chapter 1

Introduction

Now a day's cab rental services are expanding with the multiplier rate. The ease of using the services and flexibility gives their customer a great experience with competitive prices.

1.1 Problem Statement

You are a cab rental start-up company. You have successfully run the pilot project and now want to launch your cab service across the country. You have collected the historical data from your pilot project and now have a requirement to apply analytics for fare prediction. You need to design a system that predicts the fare amount for a cab ride in the city.

1.2 Data

Understanding of data is the very first and important step in the process of finding solution of any business problem. Here in this case our company has provided a data set with following features,it is needed to go through each and every variable of it to understand and for better functioning.
Size of Dataset Provided: - 16067 rows, 7 Columns (including dependent variable)
Missing Values: Yes
Outliers Presented: Yes
Below mentioned is a list of all the variable names with their meanings:

Variables	Description
fare_amount	Fare amount
pickup_datetime	Cab pickup date with time
pickup_longitude	Pickup location longitude
pickup_latitude	Pickup location latitude
dropoff_longitude	Drop location longitude
dropoff_latitude	Drop location latitude
passenger_count	Number of passengers sitting in the cab

Chapter 2

Methodology

➢ Pre-Processing

It was required to build a predictive model, to look and manipulate the data before modelling is started which includes multiple preprocessing steps such as exploring the data, cleaning the dataas well as visualizing the data through graph and plots, all these steps is combined under one shed whichis **Exploratory Data Analysis**, which includes following steps:

- Data exploration and Cleaning
- Missing values treament
- Outlier Analysis
- Feature Selection
- Features Scaling
    - Skewness and Log transformation
- Visualization

➢ Modelling

Once all the Pre-Processing steps has been done on our data set, then ,will now further move to our next step which is modelling. Modelling plays an important role to find out the good inferences from the data. Choice of models depends upon the problem statement and data set. As per our problem statement and dataset, some models on our preprocessed data are prepared and post-comparing the output results best suitable model for our problem is selected. As per our data set following models need to be tested:

- Linear regression
- Decision Tree
- Random forest,
- Gradient Boosting

➢ Model Selection

The final step of our methodology will be the selection of the model based on the different output and results shown by different models. Then, multiple parameters which will be study further in our report to test whether the model is suitable for our problem statement or not.

Chapter 3

Pre-Processing

3.1 Data exploration and Cleaning (Missing Values and Outliers)

The very first step which comes with any data science project is data exploration and cleaning which includes following points as per this project:

a. Separate the combined variables.
b. There are some negative values in fare amount so those values must be removed.
c. Passenger count would be max 6 if it is a SUV vehicle not more than that.Remove the rows having passengers counts more than 6 and less than 1.
d. There are some outlier figures in the fare (like top 3 values) so remove those.
e. Latitudes range from -90 to 90. Longitudes range from -180 to 180. Remove the rows if any latitude and longitude lies beyond the ranges.

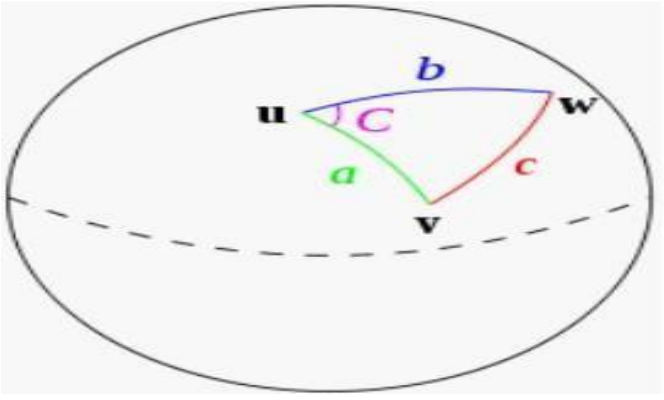
3.2 Creating some new variables from the given variables.

Here in our data set our variable name pickup_datetime contains date and time for pickup. So it is required to extract some important variables from pickup_datetime:

- Year
- Month
- Date
- Day of Week
- Hour
- Minute

The distance is calculated using the haversine formula which says:

The **haversine formula** determines the great-circle distance between two points on a sphere given their longitudes and latitudes. Important in navigation, it is a special case of a more general formula in spherical trigonometry, the law of haversines, that relates the sides and angles of spherical triangles.



So our new extracted variables are:

- fare_amount
- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- passenger_count
- year
- Month
- Date
- Day of Week
- Hour
- Minute
- Distance

3.3 Selection of variables

Now all the above variables are of no use so it is better to drop the redundant variables:

- pickup_datetime
- pickup_longitude
- pickup_latitude
- dropoff_longitude
- dropoff_latitude
- Minute

Now only these following variables will be used for further steps:

	fare_amount	passenger_count	year	Month	Date	Day of Week	Hour	distance
0	4.5	1.0	2009.0	6.0	15.0	0.0	17.0	1.030764
1	16.9	1.0	2010.0	1.0	5.0	1.0	16.0	8.450134
2	5.7	2.0	2011.0	8.0	18.0	3.0	0.0	1.389525
3	7.7	1.0	2012.0	4.0	21.0	5.0	4.0	2.799270
4	5.3	1.0	2010.0	3.0	9.0	1.0	7.0	1.999157
5	12.1	1.0	2011.0	1.0	6.0	3.0	9.0	3.787239
6	7.5	1.0	2012.0	11.0	20.0	1.0	20.0	1.555807
8	8.9	2.0	2009.0	9.0	2.0	2.0	1.0	2.849627
9	5.3	1.0	2012.0	4.0	8.0	6.0	7.0	1.374577
10	5.5	3.0	2012.0	12.0	24.0	0.0	11.0	0.000000

VariableNames	Variable DataTypes
fare_amount	float64
passenger_count	object
year	object
Month	object
Date	object
Day of Week	object
Hour	object
distance	float64

3.4 Some more data exploration

In this report it is required to predict the fare prices of a cab rental company. So here our dataset consists of 16067 observations with 8 variables including one dependent variable.

3.4.1

Below are the names of Independent variables:
passenger_count, year, Month, Date, Day of Week, Hour, distance

Our Dependent variable is: **fare_amount**

3.4.2 Uniqueness in Variable

By looking at the unique number in the variables which help us to decide whether the variable is categorical or numeric. By using python script 'nunique' to find out the unique values in each variable the following variables is received.The table is below:

Variable Name	Unique Counts
fare_amount	450
passenger_count	7
year	7
Month	12
Date	31
Day of Week	7
Hour	24
distance	15424

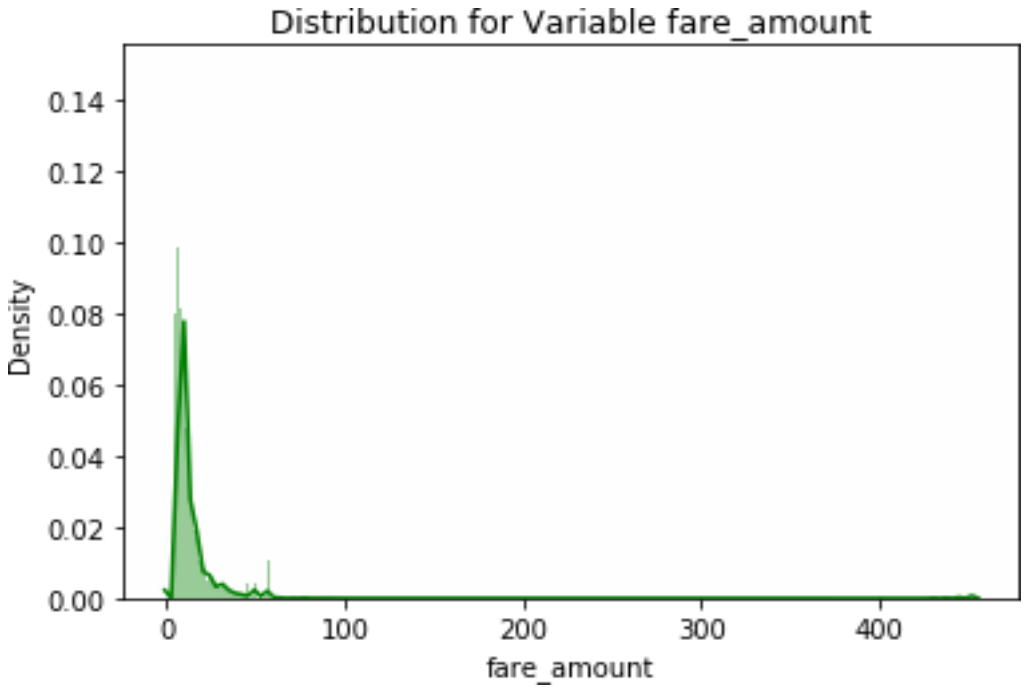
3.4.3 Dividing the variables into two categories basis their data types:

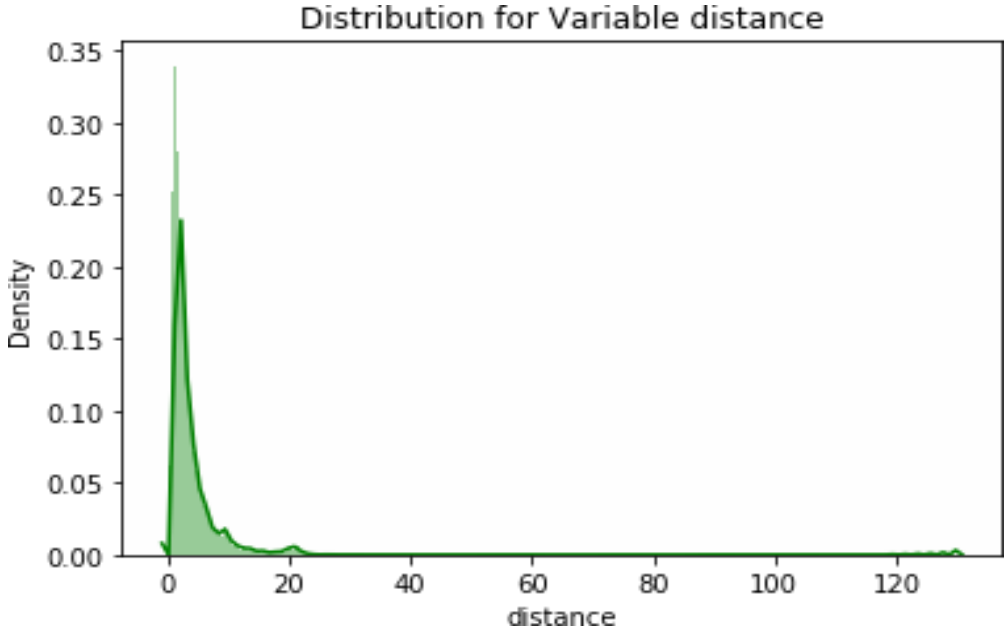
Continuous variables - 'fare_amount', 'distance'.
Categorical Variables - 'year', 'Month', 'Date', 'Day of Week', 'Hour', 'passenger_count'

3.5 Feature Scaling

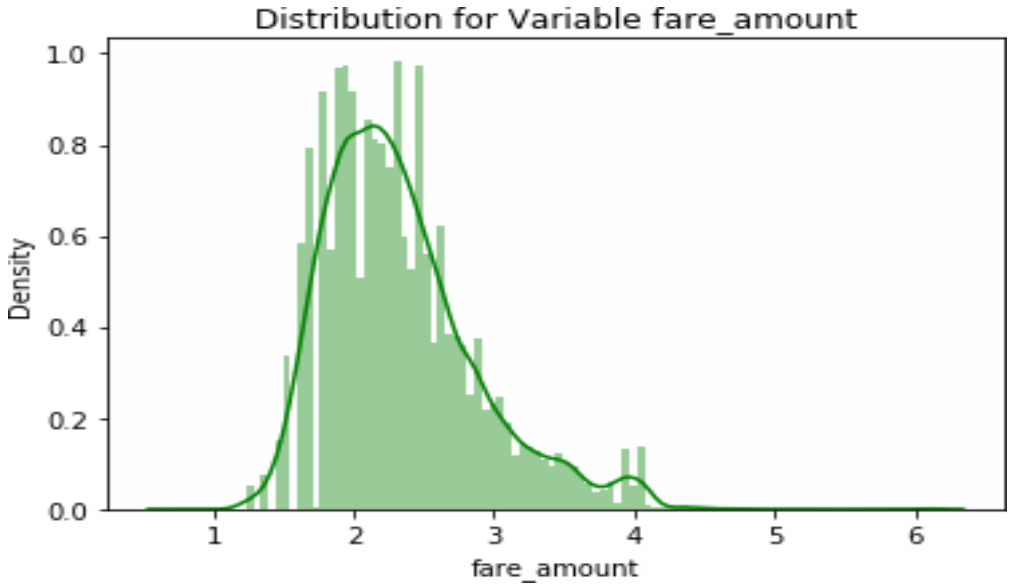
Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right. Skewness can be quantified to define the extent to which a distribution differs from a normal distribution. There is skewness in our variables and our target variable 'fare_amount' is having one-side left-skewed so by using **log transform** technique skewness is removed.

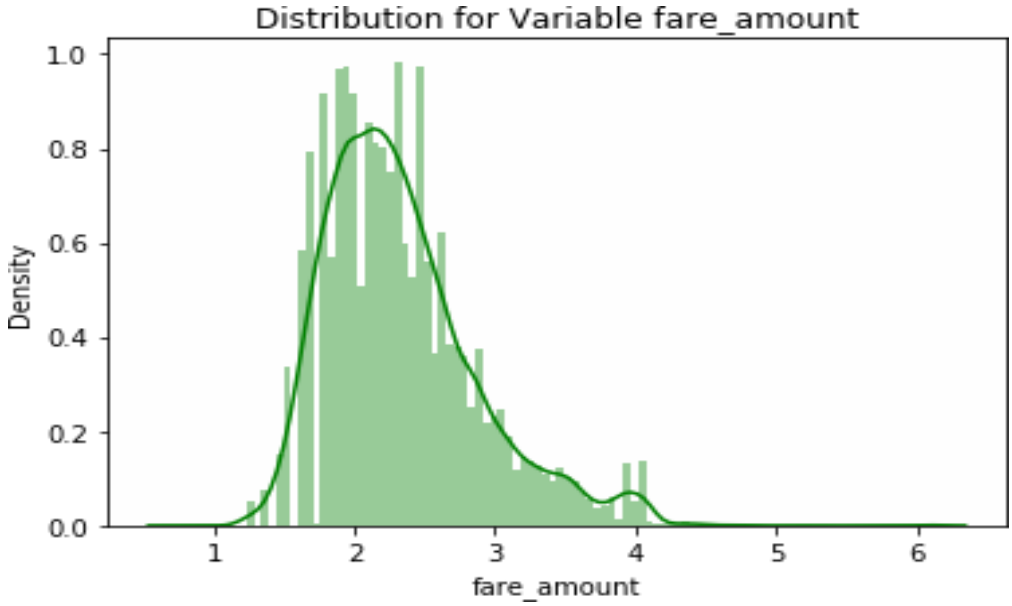
Below mentioned graphs shows the probability distribution plot to check distribution before log transformation:





Below mentioned graphs shows the probability distribution plot to check distribution after log transformation:





Chapter 4

Modelling

After a thorough preprocessing, regression models are used on our pre-processed data to predict the target variable. Following are the models which have been tested with data –

- Linear Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Before running any model, split our data into two parts which is train and test data. In our case 80% of the data is taken as our train data. Below is the snipped image of the split of train test.

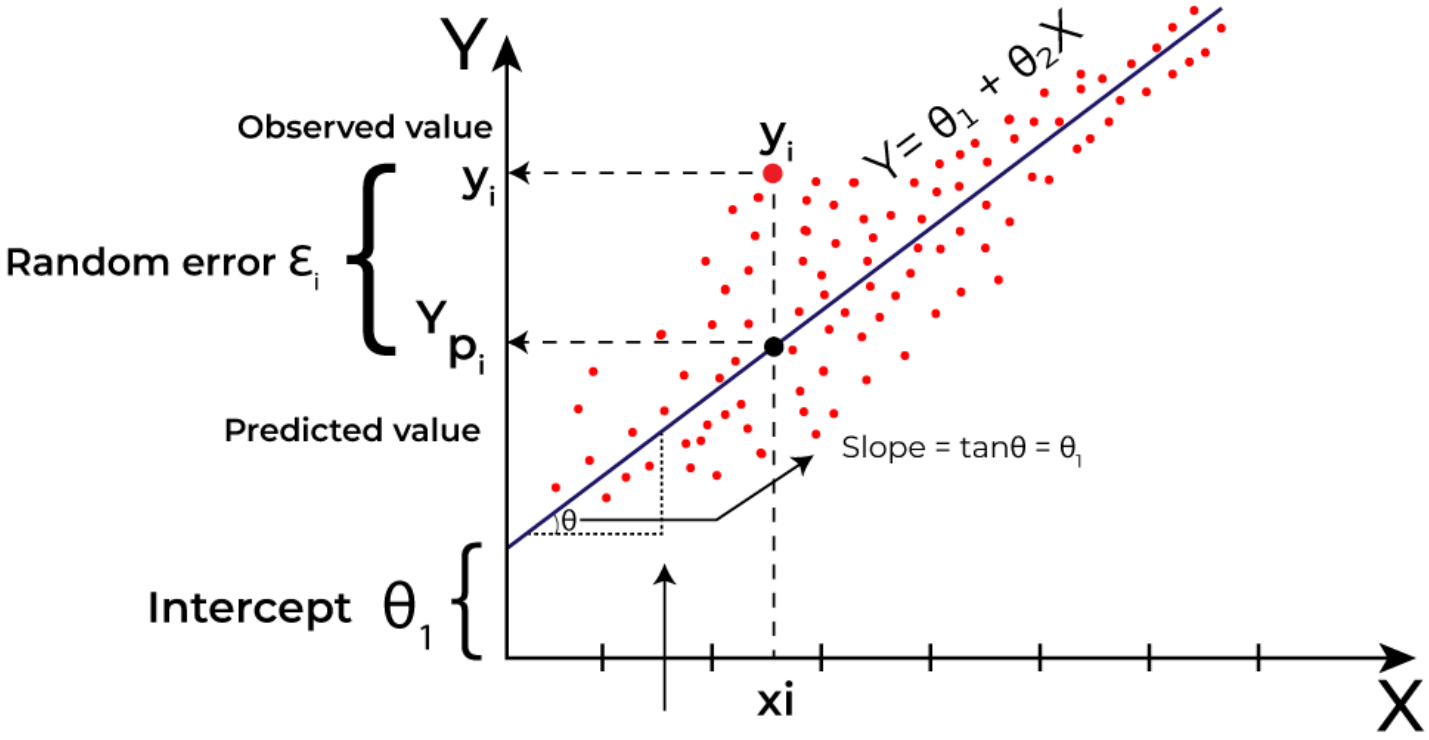
We need to split our train data into two parts

```
In [56]: from sklearn.tree import DecisionTreeRegressor
         from sklearn.metrics import mean_squared_error

In [60]: ##train test split for further modelling
         X_train, X_test, y_train, y_test = train_test_split( train_df.iloc[:, train_df.columns != 'fare_amount'],
                                 train_df.iloc[:, 0], test_size = 0.20, random_state = 1)

In [61]: print(X_train.shape)
         print(X_test.shape)

         (12339, 7)
         (3085, 7)
```

4.1 Multiple Linear Regression:-

Multiple linear regression is the most common form of linear regression analysis. Multiple regression is an extension of simple linear regression. It is used as a predictive analysis, when it is required to predict the value of a variable based on the value of two or more other variables. The variable to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).

Below is a screenshot of the linear regression model build and its output:



1. Linear Regeression Model

In [62]:
```
# Importing libraries for Linear Regression
from sklearn.linear_model import LinearRegression
```

In [63]:
```
# Building model on top of training dataset
fit_LR = LinearRegression().fit(X_train , y_train)
```

In [64]:
```
#prediction on train data
pred_train_LR = fit_LR.predict(X_train)
```

In [65]:
```
#prediction on test data
pred_test_LR = fit_LR.predict(X_test)
```

In [66]:
```
##calculating RMSE for test data
RMSE_test_LR = np.sqrt(mean_squared_error(y_test, pred_test_LR))
```

In [67]:
```
print("Root Mean Squared Error For Test data = "+str(RMSE_test_LR))
```

```
Root Mean Squared Error For Test data = 0.2503511796785927
```

In [68]:
```
from sklearn.metrics import r2_score
#calculate R^2 for train data
r2_score(y_train, pred_train_LR)
```

Out[68]:
```
0.746855951097612
```

In [69]:
```
r2_score(y_test, pred_test_LR)
```

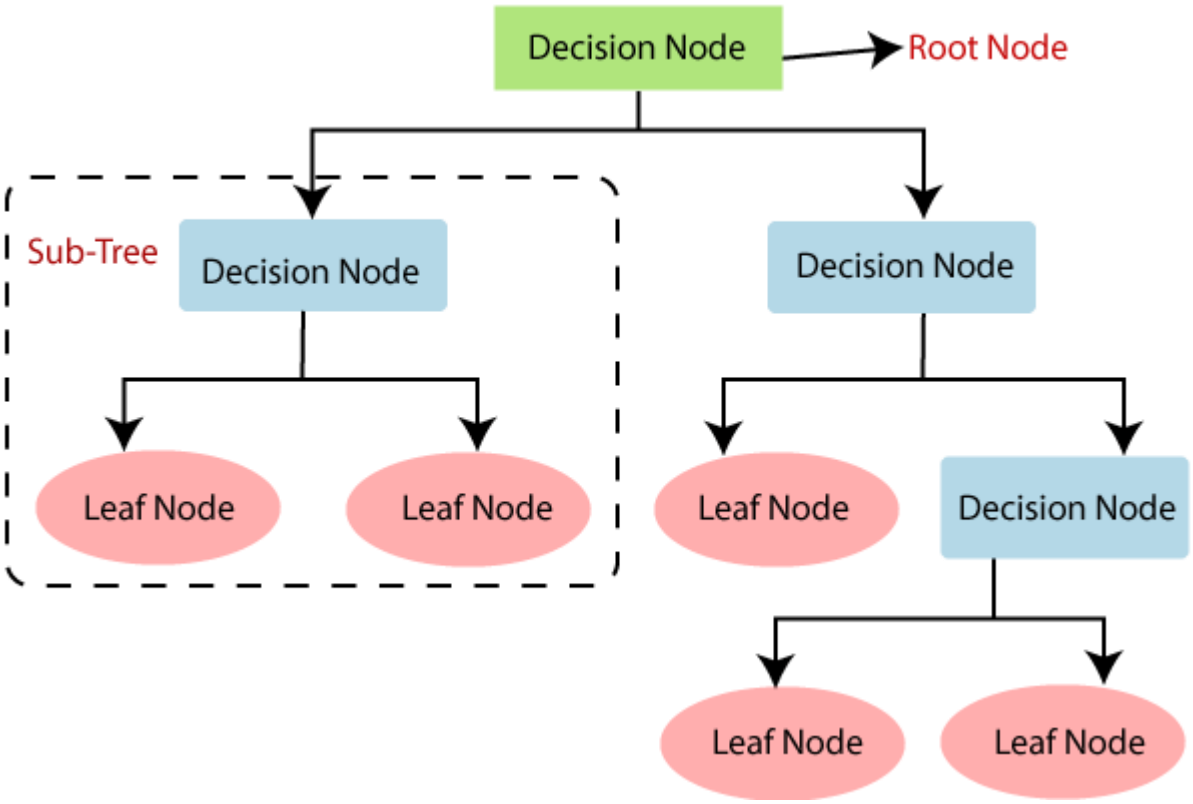
Out[69]:
```
0.7778537029821875
```

4.2 Decision Tree

A tree has many analogies in real life, and turns out that it has influenced a wide area of machine learning, covering both classification and regression. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. As the name goes, it uses a tree-like model of decisions.

Below is the screenshot of the query and the result shown, and the results are compared of each model in a combined table later on.

Decision Tree Algorithm



2. Decision Tree Model

In [71]:
```
fit_DT = DecisionTreeRegressor(max_depth = 2).fit(X_train,y_train)
```

In [72]:
```
#prediction on train data
pred_train_DT = fit_DT.predict(X_train)

#prediction on test data
pred_test_DT = fit_DT.predict(X_test)
```

In [73]:
```
##calculating RMSE for train data
RMSE_train_DT = np.sqrt(mean_squared_error(y_train, pred_train_DT))

##calculating RMSE for test data
RMSE_test_DT = np.sqrt(mean_squared_error(y_test, pred_test_DT))
```

In [74]:
```
print("Root Mean Squared Error For Training data = "+str(RMSE_train_DT))
print("Root Mean Squared Error For Test data = "+str(RMSE_test_DT))
```

```
Root Mean Squared Error For Training data = 0.30120638747129796
Root Mean Squared Error For Test data = 0.28969521517125973
```

In [75]:
```
## R^2 calculation for train data
r2_score(y_train, pred_train_DT)
```

Out[75]: 0.70012186420221

In [76]:
```
## R^2 calculation for test data
r2_score(y_test, pred_test_DT)
```

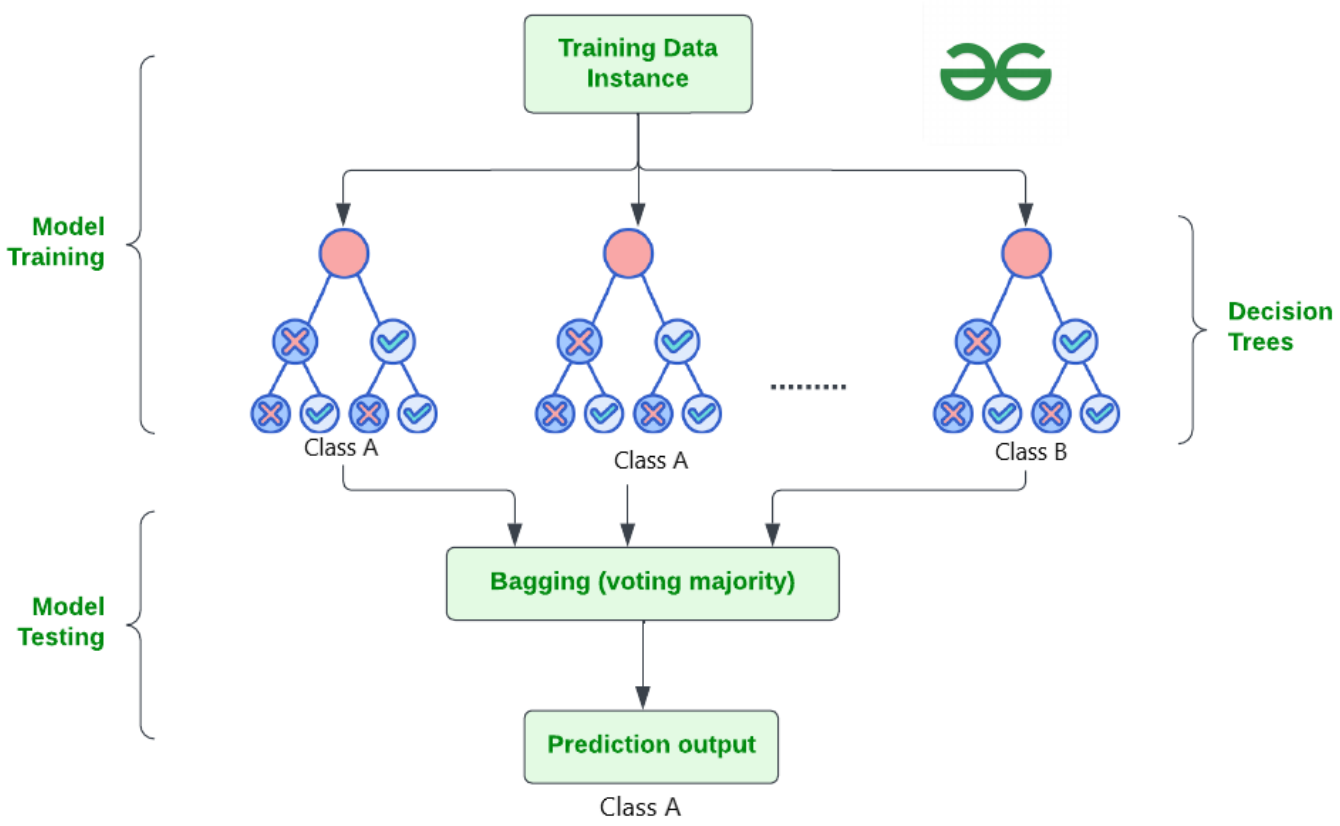
Out[76]: 0.7025442022580384

4.3 Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other task, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

To say it in simple words: Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Below is a screenshot of the model and its output:



3. Random Forest Model

In [77]:
```
# Importing libraries for Random Forest
from sklearn.ensemble import RandomForestRegressor
```

In [78]:
```
fit_RF = RandomForestRegressor(n_estimators = 200).fit(X_train,y_train)
```

In [79]:
```
#prediction on train data
pred_train_RF = fit_RF.predict(X_train)
#prediction on test data
pred_test_RF = fit_RF.predict(X_test)
```

In [80]:
```
##calculating RMSE for train data
RMSE_train_RF = np.sqrt(mean_squared_error(y_train, pred_train_RF))
##calculating RMSE for test data
RMSE_test_RF = np.sqrt(mean_squared_error(y_test, pred_test_RF))
```

In [81]:
```
print("Root Mean Squared Error For Training data = "+str(RMSE_train_RF))
print("Root Mean Squared Error For Test data = "+str(RMSE_test_RF))
```

```
Root Mean Squared Error For Training data = 0.09594886483675674
Root Mean Squared Error For Test data = 0.23918653965477282
```

In [82]:
```
## calculate R^2 for train data

r2_score(y_train, pred_train_RF)
```

Out[82]:
```
0.9695704079865043
```

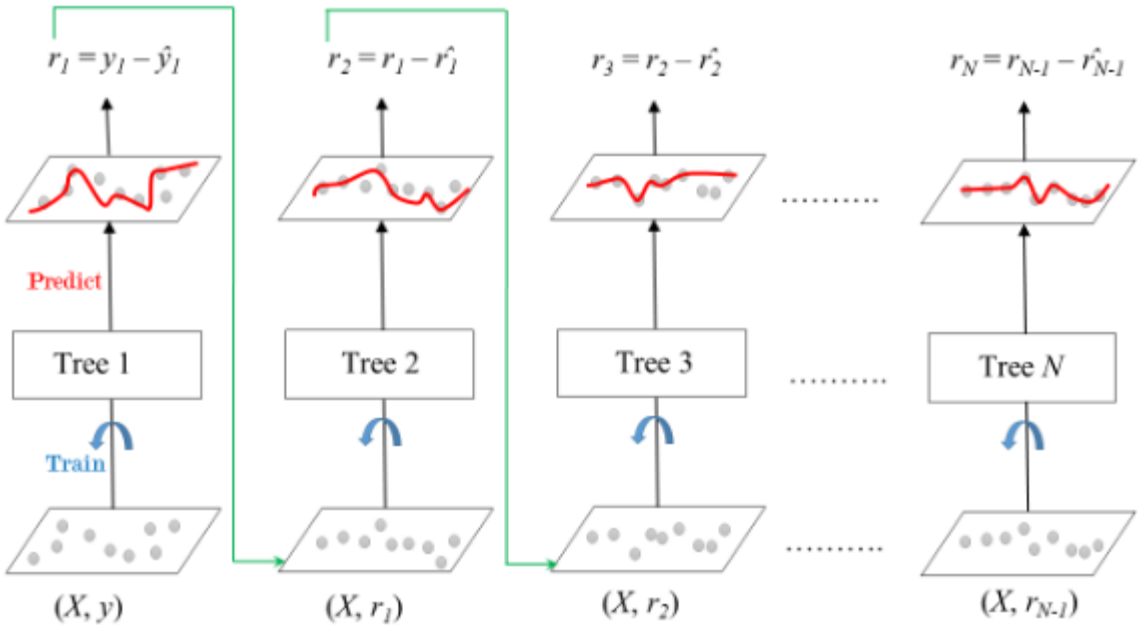
In [83]:
```
#calculate R^2 for test data
r2_score(y_test, pred_test_RF)
```

Out[83]:
```
0.7972255343157659
```

4.4 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.

Below is a screenshot of the model and its output:



4. Gradient Boosting

In [86]:
```
# Importing library for GradientBoosting
from sklearn.ensemble import GradientBoostingRegressor
```

In [87]:
```
# Building model on top of training dataset
fit_GB = GradientBoostingRegressor().fit(X_train, y_train)
```

In [88]:
```
#prediction on train data
pred_train_GB = fit_GB.predict(X_train)

#prediction on test data
pred_test_GB = fit_GB.predict(X_test)
```

In [89]:
```
##calculating RMSE for train data
RMSE_train_GB = np.sqrt(mean_squared_error(y_train, pred_train_GB))
##calculating RMSE for test data
RMSE_test_GB = np.sqrt(mean_squared_error(y_test, pred_test_GB))
```

In [90]:
```
print("Root Mean Squared Error For Training data = "+str(RMSE_train_GB))
print("Root Mean Squared Error For Test data = "+str(RMSE_test_GB))
```

```
Root Mean Squared Error For Training data = 0.22921680482502263
Root Mean Squared Error For Test data = 0.22939164285908767
```

In [92]:
```
#calculate R^2 for test data
r2_score(y_test, pred_test_GB)
```

Out[92]:
```
0.813493068270751
```

In [93]:
```
#calculate R^2 for train data
r2_score(y_train, pred_train_GB)
```

Out[93]:
```
0.8263361773771449
```

Chapter 5

Conclusion

5.1 Model Evaluation

The main concept of looking at what is called residuals or difference between our predictions f(x[I,]) and actual outcomes y[i].

In general, most data scientists use two methods to evaluate the performance of the model:

I. **RMSE** (Root Mean Square Error): is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{mo\ del,i})^2}{n}}$$

II. **R Squared(R^2):** is a statistical measure of how close the data are to the fitted regression line. It is also called 'goodness of fit'. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression. In other words, it explains as to how much of the variance of the target variable is explained.

III. Both train and test data results are shown, the main reason behind showing both results is to check whether our data is overfitted or not.

Below table shows the model results before applying hyper tuning:

Model Name	**RMSE**		**R Squared**	
	Train	**Test**	**Train**	**Test**
`Linear Regression	0.27	0.25	0.74	0.77
Decision Tree	0.30	0.28	0.70	0.70
Random Forest model	0.09	0.23	0.96	0.79
Gradient Boosting	0.22	0.22	0.82	0.81

Through analysis, it is discovered that-

1. The Linear Regression model performs normally in training data, but its accuracy improves in Testing data.
2. The Decision Tree model accuracy drops from training data to testing data.
3. The Random Forest model performs excellently in training data, but as soon as test dataset is given, its accuracy drops to quite low.
4. The Gradient Boosting model accuracy remains same in both training and testing dataset.

Therefore, it is needed to select that model whose accuracy increases from training data to testing data.

Therefore, Linear Regression model and Decision Tree models are selected as the best predictive models for predicting the Uber- cab fare.

For better model performance, **Root Mean Squared Error (RMSE**) must be **low** and **r2_score** must be **high** i.e **closer to 1**

We find that by applying 4 models, **Linear Regression and Decision Tree** turns out to be the **best model for our Uber taxi cab fare prediction**

5.2 Model Selection

On the basis RMSE(Root Mean Squared Error) and R-squared results a good model should have least RMSE and max R-Squared value. So, from above tables it can be seen that:

1. The performance of linear regression and decision tree models on both training and testing datasets.
2. Upon examination, it is observed that on the training data, both models exhibited relatively low RMSE values and high R-squared values, indicating a good fit to the training data.
3. However, the true test of a model's effectiveness lies in its performance on unseen data.
4. Remarkably, when applied to the testing data, both linear regression and decision tree models showcased a further reduction in RMSE and an increase in R-squared values compared to their performance on the training data.
5. This outcome suggests that not only did these models generalize well to unseen data, but their accuracy actually improved when confronted with new observations.
6. **Consequently, based on their ability to minimize error and maximize explanatory power on both training and testing datasets, it is asserted that both linear regression and decision tree models emerge as the most robust and effective choices for our predictive analysis**.

5.3 Data-Visualization on Uber Data:-

1. Number of passengers and fare

In the below graph it can be seen that single passengers are the most frequent travellers, and the highest fares seems to come from cabs which carry just 1 passenger.

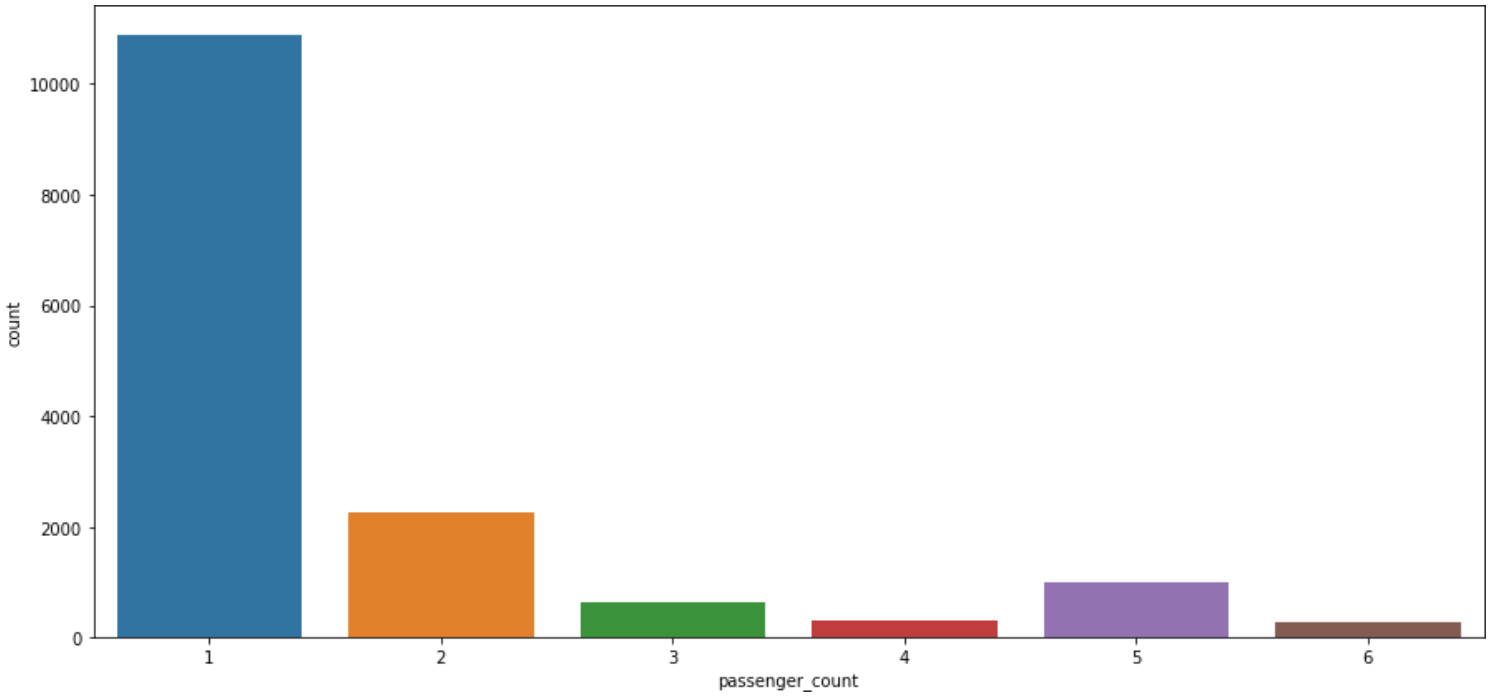


Figure-5.1:Passengers vs fare

According to analysis, rides with 1 passengers have the maximum number of Uber cab bookings. This shows that for Uber single passengers have the biggest market.

Hence , they can give these passengers more offers and discounts and introduce passes to retain these customers and frequently give offers, passes and discounts to encourage group-bookings .
Increase marketing spending and promotional offers.

2. **Date of month and fares**

The fares throughout the month mostly seem uniform.

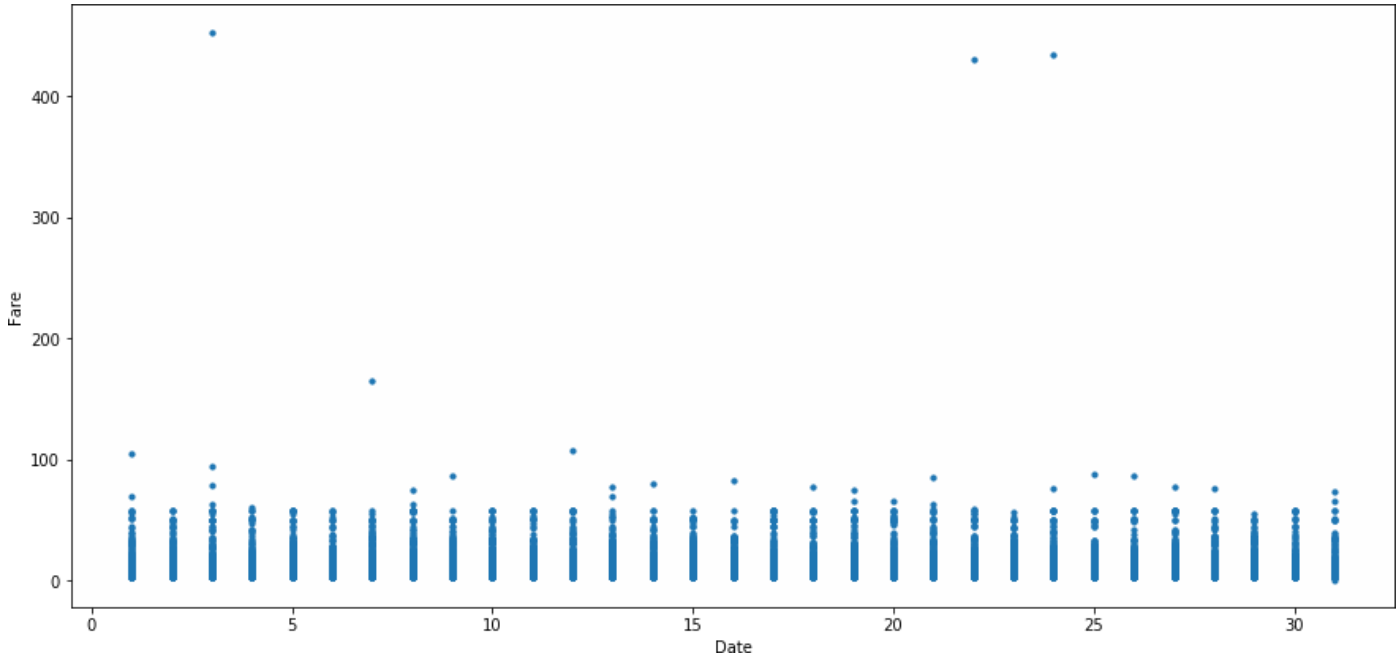


Figure 5.2:Date of month vs number of cabs booked

Date of the month has no impact on Uber bookings.

3. Hours and Fares

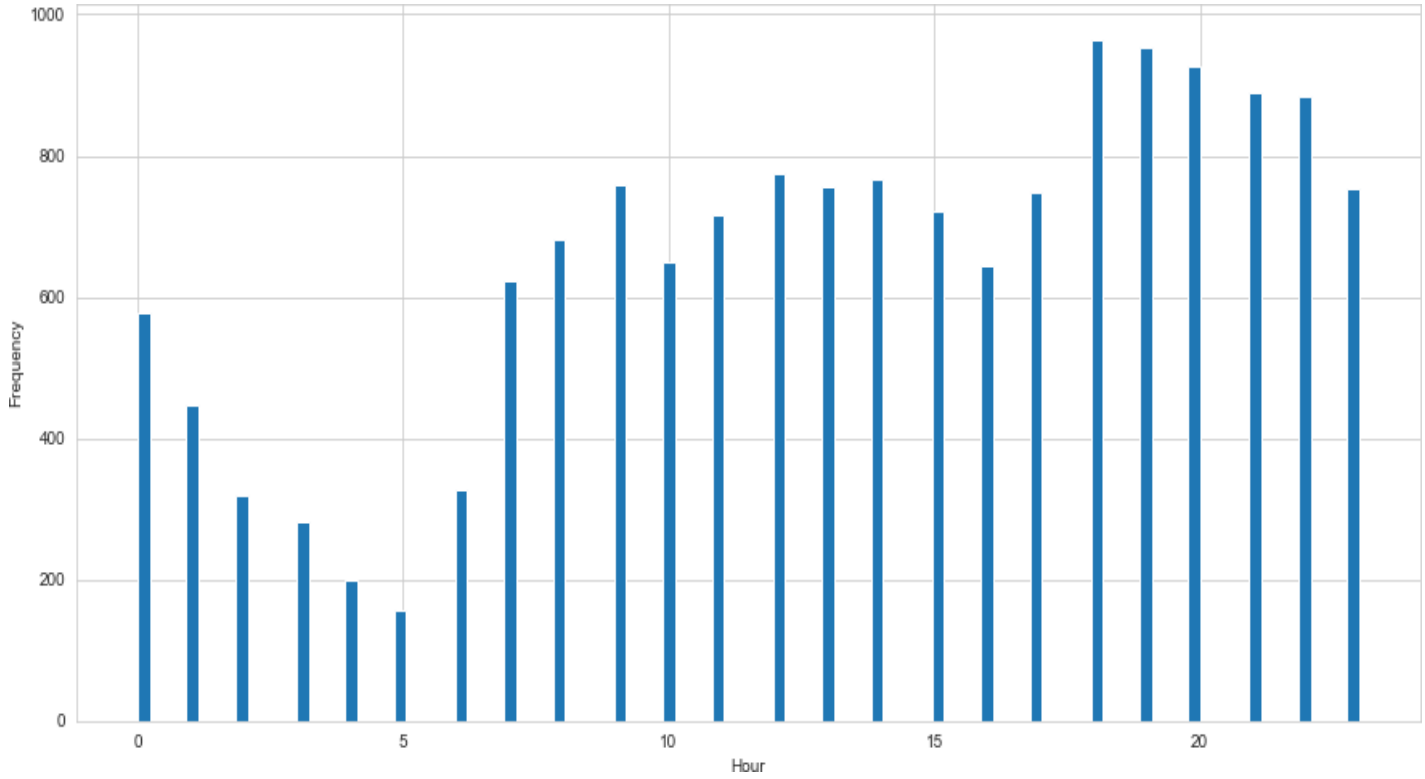


Figure 5.3: Time of day vs Frequency of cabs booked

Analysis tells us that, there are some pockets of peak rush-hour in the day.They are as follows-

Time 1- 00:00 – 5:00 hrs is the lowest traffic time, so Uber bookings are the lowest in the day.
Time 2- 5:00 – 9:00 hrs is the second busiest time for cab bookings.
Time 3- 10:00 – 18:00 hrs is normal rush hour for Uber bookings.
Time 4- **18:00 – 23:00 hrs** is the peak most-booked rush-hour for Uber rides.

4. **Impact of Day and fare**

- Cab fare is high on Friday, Saturday and Monday, may be during weekend and first day of the working day they charge high fares because of high demands of cabs.

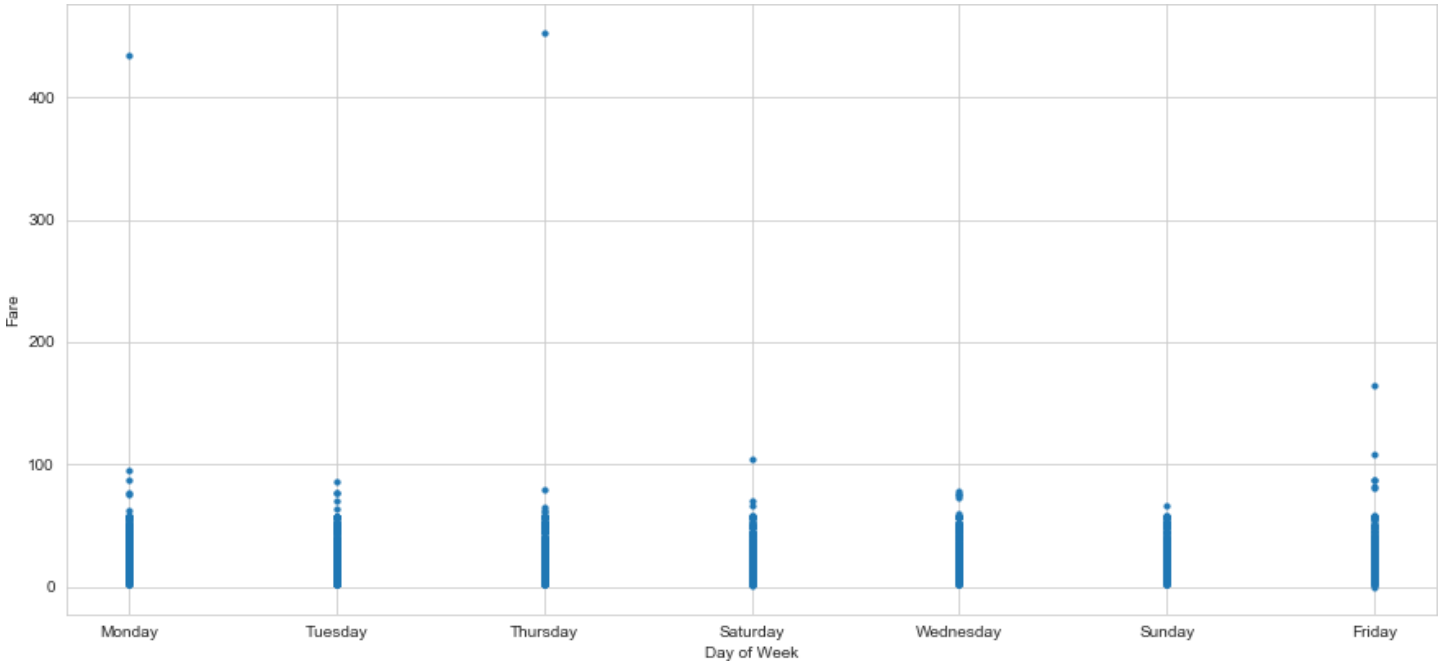


Figure 5.4: Weekday vs fare of rides

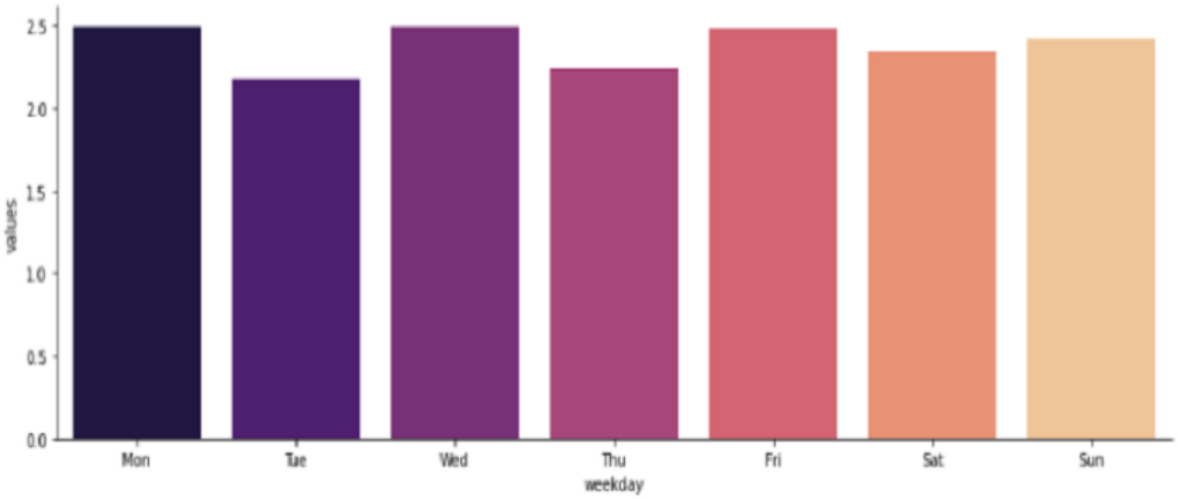


Figure 5.5: Day of week vs number of cabs booked

Observation : The day **MONDAY** seem to have most influence on the number of cabs ride.

5.4 Conclusion and Recommendation:

Conclusion:

In conclusion, the Uber cab fare prediction analysis has yielded valuable insights and implications for optimizing service efficiency and enhancing user experience.

Through comprehensive data analytics, we have gained a deeper understanding of user behavior, peak hours, and popular routes, enabling Uber to strategically allocate resources and tailor its services to meet customer demands effectively.

The development and evaluation of predictive models, particularly linear regression and decision tree algorithms, have demonstrated their efficacy in accurately forecasting taxi cab fares based on various factors such as distance, time of day, and traffic conditions.

Importantly, these models have exhibited strong performance not only on the training data but also on the testing data, indicating their robustness and generalizability to unseen observations. By leveraging these insights and predictive capabilities,
Uber can optimize operational efficiency, improve cost estimation accuracy, and ultimately enhance customer satisfaction and loyalty. Moving forward, continued investment in data analytics and predictive modeling
will be essential for Uber to maintain its competitive edge and drive innovation in the dynamic landscape of ride-sharing services.

Recommendation:

Based on the insights gained from the Uber cab fare prediction analysis using predictive modelling, several recommendations can be proposed to enhance operational efficiency and improve customer satisfaction:

1. **Refine Pricing Strategies:** Utilize the predictive models to refine pricing strategies, especially during peak hours and in high-demand areas. Dynamic pricing mechanisms can be implemented to adjust fares in real-time based on predictive insights, optimizing revenue generation while ensuring competitive pricing for users.

2. **Enhance User Experience**: Leverage predictive modeling to provide users with accurate fare estimates before booking rides. Implement features within the Uber app that enable users to input their destination and receive an estimated fare, considering factors such as distance, time of day, and traffic conditions. This transparency can enhance user trust and satisfaction.

3. **Optimize Driver Allocation**: Utilize predictive models to optimize driver allocation by anticipating demand patterns and adjusting driver deployment accordingly. By strategically positioning drivers in areas with high predicted demand, wait times can be minimized, and service availability can be maximized, leading to improved user experiences.

4. **Predictive Maintenance:** Extend the use of predictive modeling beyond fare prediction to include predictive maintenance of vehicles. By analyzing historical data on vehicle performance and maintenance records, predictive models can forecast when vehicles are likely to require maintenance, enabling proactive servicing and reducing downtime.

5. **Personalized Promotions**: Leverage predictive analytics to offer personalized promotions and incentives to users based on their historical ride patterns and preferences. By analyzing user data, including frequency of rides, preferred routes, and timing, targeted promotions can be designed to incentivize users to book rides during off-peak hours or to specific destinations.

6. **Continuous Model Evaluation and Improvement**: Regularly evaluate and fine-tune predictive models to ensure accuracy and relevance. Incorporate feedback from users and drivers to identify areas for improvement and update the models accordingly. Additionally, stay abreast of advancements in predictive modeling techniques and integrate new methodologies as appropriate to maintain competitiveness.

7. **Invest in Data Security and Privacy**: Given the sensitivity of user data involved in predictive modeling, prioritize data security and privacy measures. Implement robust data encryption protocols, access controls, and compliance with regulatory requirements such as GDPR to safeguard user information and maintain trust.

CITATION

1) Srinivas, R., Ankayarkanni, B., & Krishna, R. S. B. (2021, May 6). Uber Related Data Analysis

   using Machine Learning. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*.

   (Srinivas et al., 2021)

   https://doi.org/10.1109/iciccs51141.2021.9432347

2) Gunawardena, T., & Jayasena, K. (2020, December 2). Real-Time Uber Data Analysis of Popular Uber Locations in Kubernetes Environment. *2020 5th International Conference on Information Technology Research (ICITR)*.

   (Gunawardena & Jayasena, 2020)

   https://doi.org/10.1109/icitr51448.2020.9310851

3) Fu, Y., & Soman, C. (2021, June 9). Real-time Data Infrastructure at Uber. *Proceedings of the 2021 International Conference on Management of Data*. https://doi.org/10.1145/3448016.3457552

   (Fu & Soman, 2021)

4) Cohen, P., Hahn, R., Hall, J., Levitt, S., & Metcalfe, R. (2016, September). *Using Big Data to Estimate Consumer Surplus: The Case of Uber*. https://doi.org/10.3386/w22627

   (Cohen et al., 2016)

5) Indulkar, Y., & Patil, A. (2020, September). Sentiment Analysis of Uber & Ola using Deep Learning. *2020 International Conference on Smart Electronics and Communication (ICOSEC)*. https://doi.org/10.1109/icosec49089.2020.9215429

   (Indulkar & Patil, 2020)

6) Shokoohyar, S., Sobhani, A., & Ramezanpour Nargesi, S. R. (2020, February 11). On the determinants of Uber accessibility and its spatial distribution: Evidence from Uber in Philadelphia. *WIREs Data Mining and Knowledge Discovery*, *10*(4). https://doi.org/10.1002/widm.1362

   (Shokoohyar et al., 2020)

7) Qiu, J. (2021, June). What does Uber bring for consumers? *Data Science and Management*, *2*, 20–27. https://doi.org/10.1016/j.dsm.2021.05.002

   (Qiu, 2021)

8) Zhu, L., & Laptev, N. (2017, November). Deep and Confident Prediction for Time Series Uber. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. https://doi.org/10.1109/icdmw.2017.19

   (Zhu & Laptev, 2017)

9) Bharathi, A., & Prakash, S. S. (2019, April). An approach to predict taxi-passenger demand using quantitative histogram on Uber data. *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. https://doi.org/10.1109/icacce46606.2019.9079980

(Bharathi & Prakash, 2019)

10) Shokoohyar, S. (2018). Ride-sharing platforms from drivers' perspective: Evidence from Uber and Lyft drivers. *International Journal of Data and Network Science*, 89–98. https://doi.org/10.5267/j.ijdns.2018.10.001

(Shokoohyar, 2018)

11) Maass, K., Sathanur, A. V., Khan, A., & Rallo, R. (2020, January). Street-level Travel-time Estimation via Aggregated Uber Data. *2020 Proceedings of the SIAM Workshop on Combinatorial Scientific Computing*,

76–84. https://doi.org/10.1137/1.9781611976229.8

(Maass et al., 2020)

12) Zhao, K., Khryashchev, D., & Vo, H. (2021, June 1). Predicting Taxi and Uber Demand in Cities: Approaching the Limit of Predictability. *IEEE Transactions on Knowledge and Data Engineering*, *33*(6), 2723–2736. https://doi.org/10.1109/tkde.2019.2955686

(Zhao et al., 2021)

13) Poulsen, L. K., Dekkers, D., Wagenaar, N., Snijders, W., Lewinsky, B., Mukkamala, R. R., & Vatrapu, R. (2016, June). Green Cabs vs. Uber in New York City. *2016 IEEE International Congress on Big Data (BigData Congress)*. https://doi.org/10.1109/bigdatacongress.2016.35

(Poulsen et al., 2016)

14) Willis, G., & Tranos, E. (2021, January). Using 'Big Data' to understand the impacts of Uber on taxis in New York City. *Travel Behaviour and Society*, *22*, 94–107. https://doi.org/10.1016/j.tbs.2020.08.003

(Willis & Tranos, 2021)

15) Dogo, V., Garg, K., & Zheng, Y. (2020, October 7). Ride-Share Analysis in Boston City. *Proceedings of the 21st Annual Conference on Information Technology Education*. https://doi.org/10.1145/3368308.3415444

(Dogo et al., 2020)

16) Shah, D., Kumaran, A., Sen, R., & Kumaraguru, P. (2019, May 13). Travel Time Estimation Accuracy in Developing Regions: An Empirical Case Study with Uber Data in Delhi-NCR✱. *Companion Proceedings of the 2019 World Wide Web Conference*. https://doi.org/10.1145/3308560.3317057

(Shah et al., 2019)

17) Sathanur, A. V., Amatya, V., Khan, A., Rallo, R., & Maass, K. (2019, September 10). Graph Analytics and Optimization Methods for Insights from the Uber Movement Data. *Proceedings of the 2nd ACM/EIGSCC Symposium on Smart Cities and Communities*. https://doi.org/10.1145/3357492.3358625

(Sathanur et al., 2019)

18) Wang, M., & Mu, L. (2018, January). Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. *Computers, Environment and Urban Systems*, *67*, 169–175. https://doi.org/10.1016/j.compenvurbsys.2017.09.003

(Wang & Mu, 2018)

19) Zhao, J., Chen, C., Huang, H., & Xiang, C. (2020, July 7). Unifying Uber and taxi data via deep models for taxi passenger demand prediction. *Personal and Ubiquitous Computing*, *27*(3), 523–535. https://doi.org/10.1007/s00779-020-01426-y

(Zhao et al., 2020)

20) Brodeur, A., & Nield, K. (2018, August). An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC. *Journal of Economic Behavior & Organization*, *152*, 1–16. https://doi.org/10.1016/j.jebo.2018.06.004

(Brodeur & Nield, 2018)

21) Punt, M. B., van Kollem, J., Hoekman, J., & Frenken, K. (2021, June 21). Your Uber is arriving now: An analysis of platform location decisions through an institutional lens. *Strategic Organization*, *21*(3), 501–536. https://doi.org/10.1177/14761270211022254

(Punt et al., 2021)