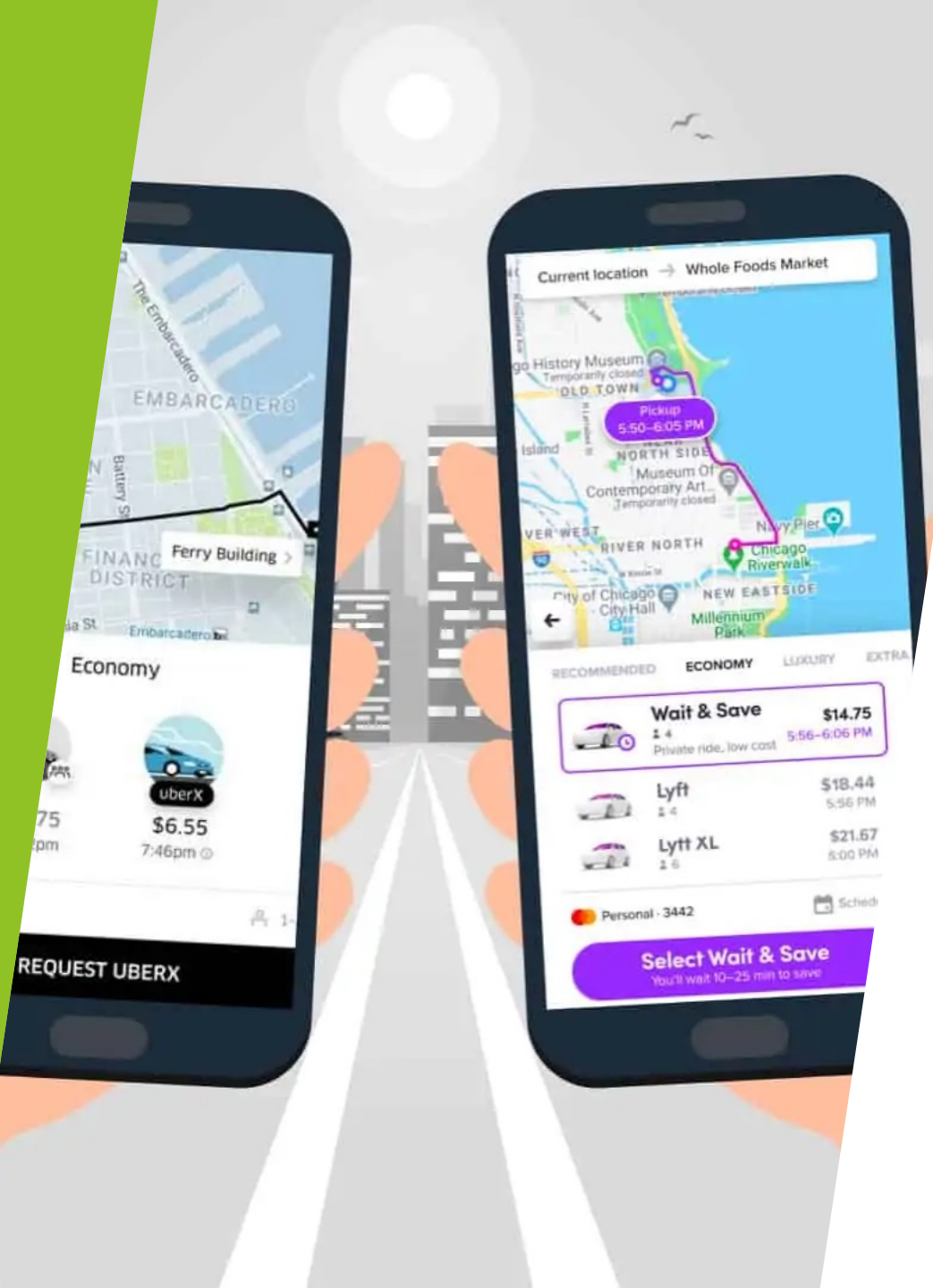
Uber Vs Lyft Price Prediction

- Saaketh Chaganty
- Tanuj Kodali
- Shreya Malapaka
- Harish Rao T



Introduction

Uber and Lyft, two ride-hailing companies, vie for the same pool of customers and drivers. Unlike fixed prices seen in public transportation, ride fares on both platforms fluctuate based on factors like distance, weather elements, as well as the starting and ending points of the journey.

- Our objective is to uncover the factors that impact ride demand and how expenses vary according to distance, geographic position, and weather conditions.
- Furthermore, we aim to gain insights into the elements that affect taxi fares and forecast trip costs by analyzing these variables.

Data Collection & Cleaning

- Gathering data constitutes a pivotal aspect of conducting Exploratory Data Analysis.
- Null values were addressed by substituting them with the mean of the corresponding feature. Additionally, outliers were excluded from the dataset, and irrelevant columns and rows were removed to streamline the data.
- Our data originates from Kaggle. https://www.kaggle.com/datasets/brllrb/uber-and-lyft-dataset-boston-ma

Machine Learning Algorithms

Supervised Machine Learning involves training models using labeled data to make predictions. In the realm of regression algorithms, three methods are employed:

- a. Multivariable Linear Regression: This technique explores relationships between multiple input variables and the target variable, enabling the prediction of continuous outcomes.
- b. Support Vector Regression: By using support vectors, this method constructs a predictive model that aims to minimize prediction errors, particularly suited for scenarios with complex data patterns.
- c. Random Forest Regressor: Utilizing an ensemble of decision trees, this algorithm excels at capturing intricate relationships within data, making it proficient in predicting continuous values.

Multivariable Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

$$\frac{1}{2m}\sum_{1}^{m}(h(x^{(i)}) - y^{(i)})^2$$

- Multiple linear regression is a statistical technique employed to assess the connection between a single dependent variable and multiple independent variables. In our scenario, we possess 21 independent variables alongside a solitary target variable.
- In the subsequent formula, y denotes the sought-after predicted target price, while x1, x2, and so forth signify the various independent variables.

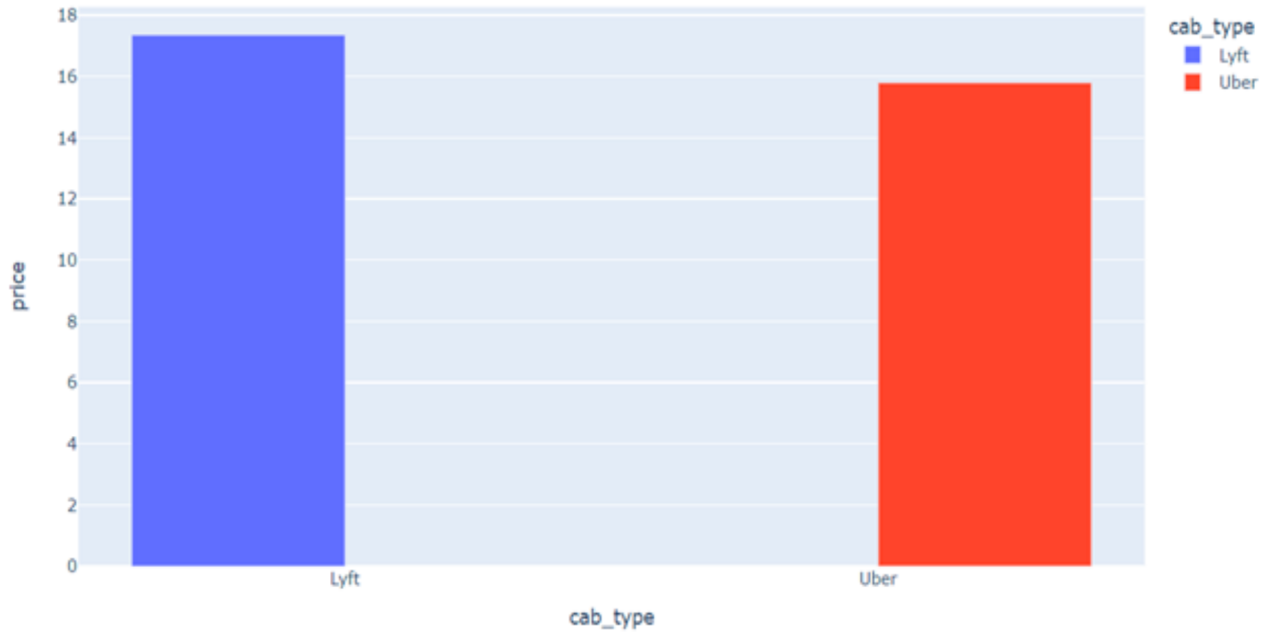
Support Vector Regression

- Differing from the Linear Regression model's approach of closely aligning a line with most data points or minimizing prediction errors, SVR endeavors to fit a hyperplane that encompasses the majority of data points.
- Due to SVR's higher time complexity for fitting, only a subset comprising 10% of the dataset was used for training.
- GridSearchCV was employed to systematically search for optimal parameters.
- In our context, the most suitable parameters are as follows:
  - 'C': 10
  - 'epsilon': 0.5
  - 'gamma': 1e-07
  - 'kernel': 'linear'

Results & Analysis

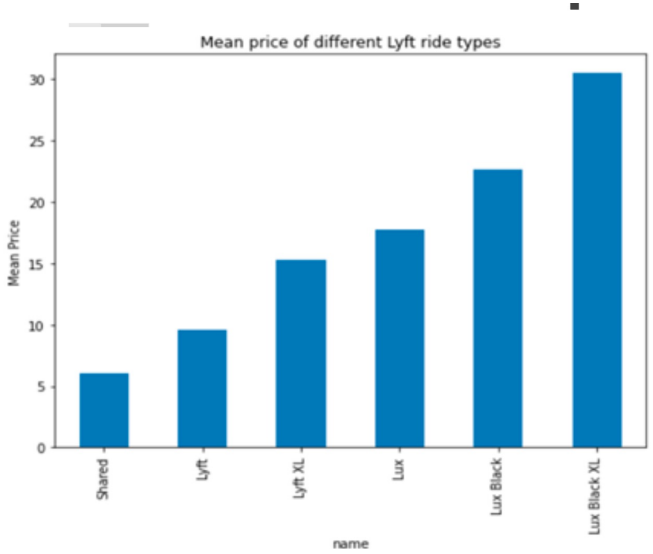
Comparing Prices based on Cab type

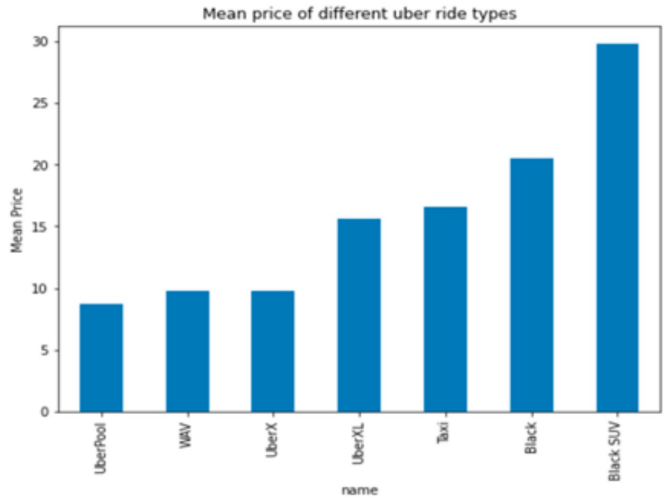
- The typical cost for a cab ride with Lyft is roughly $17, whereas Uber's average fare stands at approximately $16. This suggests that Lyft generally sets its prices a bit higher in contrast to Uber.

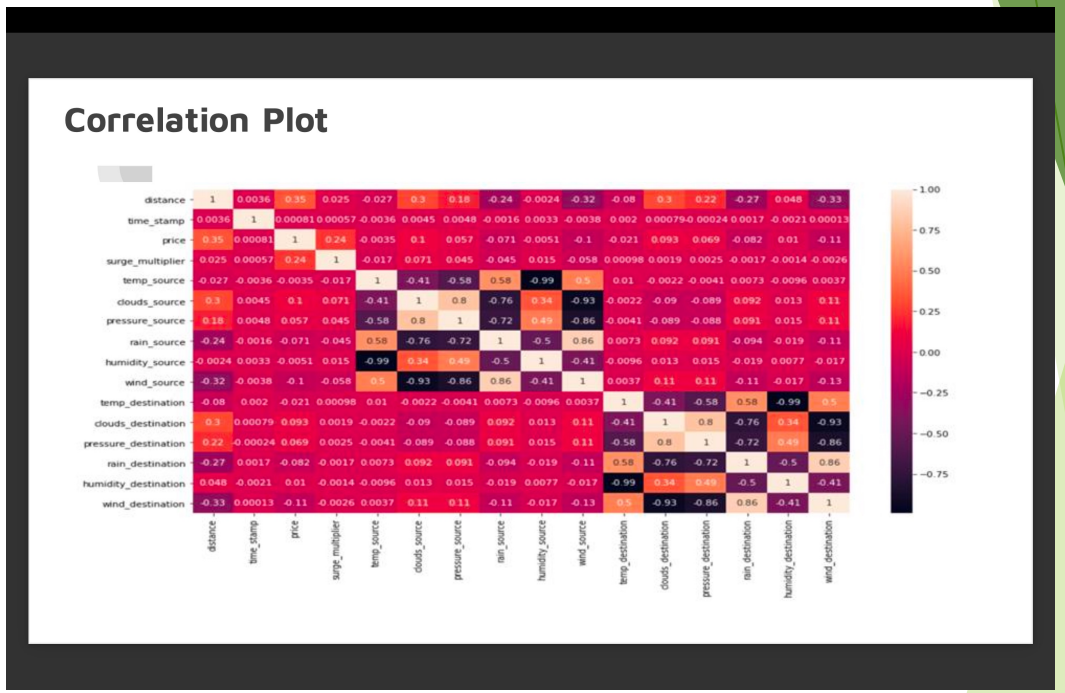


Relation between Price & Distance

- Analyzing the fare discrepancies between Uber and Lyft for their respective categories of cab services, it is evident that Lyft offers a more affordable option within the economy class, while in the luxury segment, Lyft tends to have higher pricing.



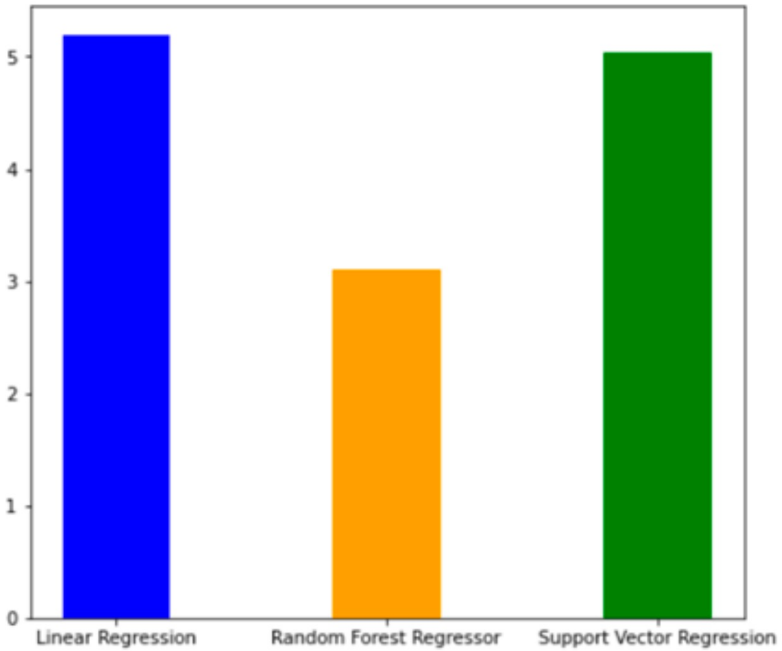


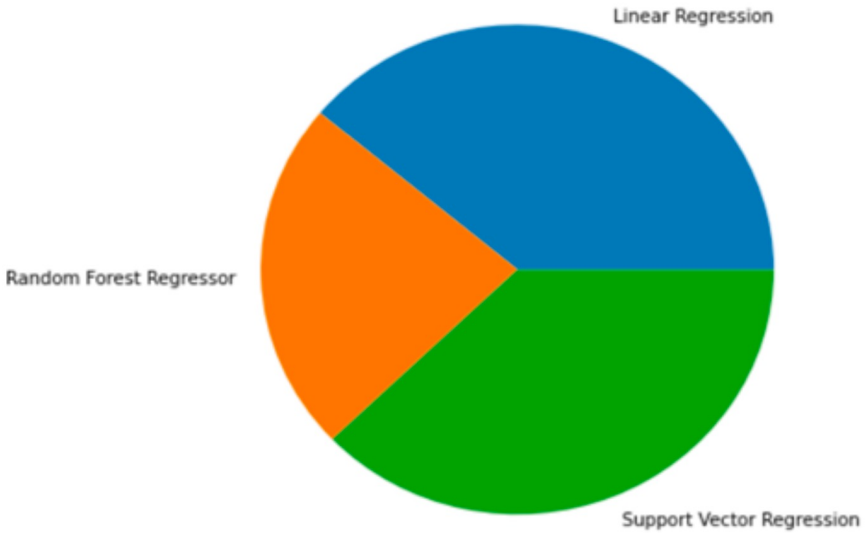


Correlation

- The distance covered, cloud cover at the source and destination, and destination pressure exhibit a positive relationship with the price.
- Conversely, humidity at the source, timestamp, surge multiplier, and precipitation at the source display a negative correlation with the price.
- While cloud cover and pressure are positively associated with the price, other weather factors may not share a direct positive correlation, but they significantly influence conditions that impact the fare.

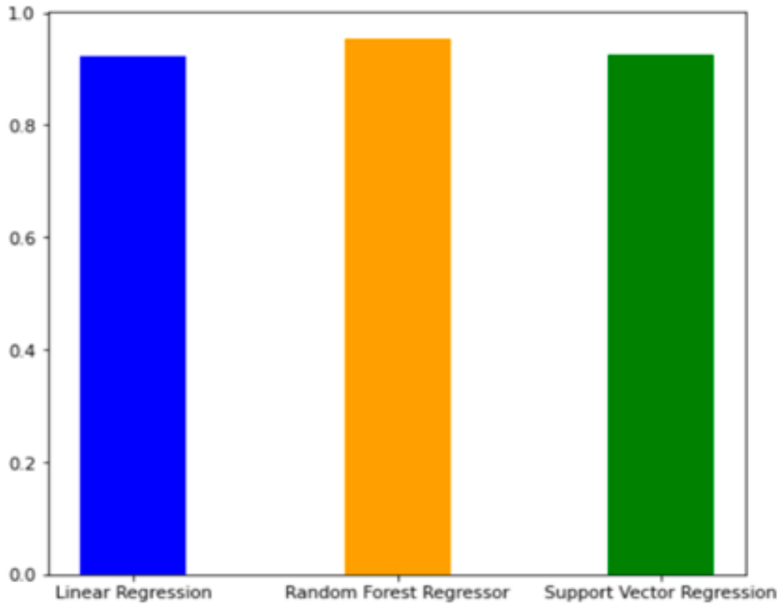
MEAN SQUARED ERROR(MSE)

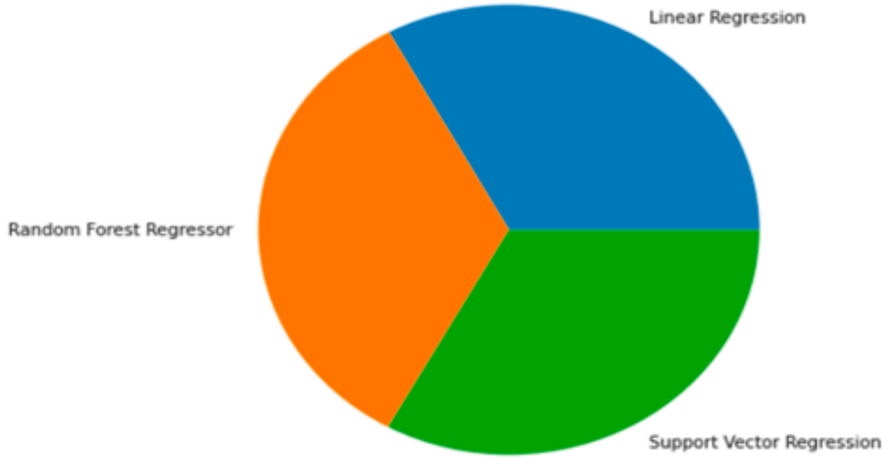




Random Forest Regressor has the lowest MSE score indicating it is closer to the data points.

R2 SCORE





Random Forest Regressor has the highest R2 score, indicating that its predicted values are closer to the real values.

Comparing Machine Learning Algorithms

Linear regression	Random Forest Regressor	Support Vector Machine
R2 SCORE:	R2 SCORE:	R2 SCORE:
0.9237	0.954	0.925
MSE:	MSE:	MSE:
5.196632	3.096856	5.045863

Conclusion

- In Boston, overall, Uber offers more cost-effective fares compared to Lyft.
- Uber holds a larger market share in the Boston ridesharing industry.
- Within the economical category, Lyft provides lower-priced options than Uber, while in the luxury segment, Uber maintains a more budget-friendly stance.
- The price of rides is significantly influenced by factors such as distance and prevailing weather conditions.
- Among the considered regression models—Support Vector Regression, Linear Regression, and Random Forest Regressor—the Random Forest Regressor yields the highest accuracy score.
- This model holds utility for both business-to-business (B2B) and business-to-consumer (B2C) applications. Ride-hailing services like Uber, Lyft, or emerging competitors can employ this model to scrutinize their rivals' pricing strategies. Additionally, consumers can utilize our model to identify the most favorable fare rates for their rides.