MACHINE LEARNING APPROACH TO PREDICT SURGE PRICES BASED ON WEATHER DATA

By

VAISHNAVI M WADAWADAGI

INTRODUCTION

Surge pricing is a pricing Strategy used by companies to increase the cost of a product or service during the time when there is high demand. This strategy is predominantly used by companies which are in the field of ride hailing. Car companies like Uber and Lyft have been extensively using this strategy for better beneficiary. But there can be issues that the prices might be overcharged.

This is where Machine Learning (ML) plays a vital role. ML has emerged as a powerful tool for companies to optimize their surge pricing strategies. As it is known that Machine learning has the ability to understand behavioral patterns with regards to customers, it will be helpful in predicting when the demand for a product increases so that pricing can be adjusted accordingly.

In our case where we are trying to predict the surge prices based on the weather data. Surge pricing based on weather data is a specific type of surge pricing that takes into account the impact of weather on demand for certain goods or services. ML has the ability to dynamically adjust prices on a real time basis, with eventual change in market conditions. This is one of the key advantages in using machine learning.

Looking into other advantages, with ML for surge pricing we can personalize prices based on customer behavior, as machine learning algorithms can result in a pricing system based on the needs and preferences of the customer, while ensuring that the overall demand remains stable.

It is known that there is always a flip side to the coin. As there are several advantages in using machine learning for surge pricing, there are potential risks associated with using machine learning for surge pricing. For instance, Algorithms might end up giving results for a certain group of customers, such as those who are less price-sensitive or those who live in the area with less demand.

Overall, the use of machine learning for surge pricing has an immense effect on both companies and customers. It has the potential to benefit companies and customers by optimizing the pricing strategies, increasing the efficiency and improving the customer experience.

PROBLEM STATEMENT

The problem statement of this project is to develop a predictive model that uses car and weather data to predict surge multipliers for various cab services, and suggest the most cost-effective cab service based on the predicted surge multiplier. The model will take into account weather conditions, distance, time of day, and other relevant factors that impact demand for cab services.

To achieve this, the project will need to collect and integrate car data and weather data from various sources. The car data will include information such as the type of cab service, the number of drivers available, and the current location of the drivers. The weather data will include temperature, precipitation, and other relevant weather conditions.

The model will then use machine learning algorithms to analyze the data and predict the surge multiplier for various cab services based on the current weather conditions, time of day, and other factors. The model will also suggest the most cost- effective cab service based on the predicted surge multiplier, enabling users to make informed decisions about which service to use.

AIM AND OBJECTIVE

Surge multiplier is a key factor that affects the price of a ride during the period of peak hours, special events and bad weather etc. This is an important factor that should be considered while estimating the cost of a taxi ride. We aim to build a machine learning model where we predict the surge prices based on different weather conditions,that eventually optimizes the company’s pricing strategies and also satisfies the customer by providing a predictable pricing model. In order to achieve a win-win situation for both customers and companies by providing a more accurate and transparent pricing mechanism, over here we build a machine learning model.

Objective 1: Our primary objective emphasizes on reducing the loss that is minimizing the prediction error of the model: as we predict the surge prices of Uber and Lyft based on the weather data, by which we end up getting better and more accurate results.

Objective 2: We've performed certain statistical tests like t-test and ANOVA in order to know which cab provides a better pricing. As we know that t-test is used to compare the means between two groups, in this case Uber and Lyft, we can understand which company provides better pricing for the customers. Then we also perform ANOVA (Analysis of Variance) which helps us to compare the variances across means of both Uber and Lyft. We take up the price and distance features into consideration in order to know which cab provides better prices.

RELATED WORK

1. Matthew Battifarano, Zhen (Sean) Qian proposed a general framework in predicting the short term evolution of surge multipliers in real time using log linear model with L1 regularization, coupled with pattern clustering by exploring the spatio-temporal correlations between the urban environment, traffic flow characteristics and surge multipliers in their study in predicting the real-time surge pricing of ride-sourcing companies.

2. Qi Luo, Romesh Saigal work provided a macroscopic perspective in handling the complicated spatiotemporal pricing problem in ride sharing and similar matching markets where they investigated the dynamic pricing problem in on demand ride-sharing using a continuous time and continuous space approach.

3. The Effects of Uber's Surge Pricing: A Case Study by Jonathan Hall, Cory Kendrick and Chris Nosko analyzed the impact of Uber's surge pricing as how it affects the demand for rides, the number of available drivers, and the earnings of drivers during peak hours which leads us to know the supply and demand of its services, as well as on driver earnings and customer welfare.

4. Kun-Huang Huarng , Tiffany Hui-Kuang Yu studied the impact of surge pricing on customer retention where they used fuzzy set/Qualitative Comparative Analysis to generate relations and then qualitative analysis with structural associations to propagate the values and refine these relations.

5. A comparative Analysis of User Satisfaction between Uber and Lyft in the US by David C. Yen, published in the International Journal of Business and Economics Research (2018). This study conducted a survey of 215 Uber and Lyft users in the US to compare their satisfaction levels with the two-ride sharing services. The results showed that overall, users had similar satisfaction

levels with both services, but there were some differences in areas such as cost, availability and safety.

6. A Comparative Analysis of Surge Pricing Strategies in the Ride-Sharing Industry: Uber vs. Lyft by Zhihao Zheng and Yujie Zhang, published in the International Journal of Production Research(2019). This paper compared the surge pricing strategies of Uber and Lyft in the US, using data from the two companies' pricing policies. The authors found that both companies use similar surge pricing algorithms, but Lyft tends to have lower surge multipliers than Uber.

7. Factors Affecting Surge Pricing in Ridesharing Platforms: A study Of Uber and Lyft by Kaur et al.(2020). This paper analyzed the factors influencing surge pricing in Uber and Lyft. The study found that high demand, low supply and time of day were the most significant factors based on location, with surge prices being higher in urban areas with more traffic.

8. Integrated Reward Scheme and Surge Pricing in a Ridesourcing Market by Hongzhi Cheng, Weiwei Chen, Jingyi Ma and Jianwei Huang was published in IEEE Transactions on Intelligent Transportation Systems in 2020. This paper proposes a new pricing scheme that integrates a reward scheme with surge pricing in a ridesourcing market, aiming to improve the quality of service while balancing the interests of both drivers and passengers. The reward scheme encourages drivers to provide better service, while the surge pricing helps to match supply and demand during peak hours.

9. The Welfare Effects of Surge Pricing in Taxi Services by Peter Cohen and Robert Hahn, published in the Journal of Political Economy in 2016. This paper examines the welfare effects of surge pricing in taxi services where they developed a model to estimate the impact of surge pricing on various stakeholders, including passengers, drivers and taxi companies

10. The impact of ride-hail surge factors on taxi bookings was published in Transportation Research Part C Emerging Technologies by Sumit Agarwall, Ben Charoenwong, Shih-Fen Cheng and Jussi Keppo studied the impact of ride-hailing surge factors on allocative efficiency of taxis in Singapore. They combined a reduced form estimation with structural analyses using machine-learning based demand predictions to study how location-time-specific surge factors affect taxi bookings, while accounting for various confounding variables.

PRELIMINARIES

In this project we are trying to predict the surge prices based on the weather data. Regression algorithms are well suited for prediction problems. There are several regression algorithms that are well suited in prediction problems. In this case we will be using Random Forest, a supervised machine learning technique which is based on the concept of ensemble learning. To give a note of ensemble learning, it is a meta estimator that fits a number of decision trees classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

The ensemble technique that we are using is Random Forest. Random forest is one such ensemble learning based algorithm which is applicable for both classification and regression problems. As the name suggests it is a classifier that contains a number of decision trees. Decision tree is a tree-like model where each internal node represents a test on a specific feature, each branch represents the outcome of the test and each leaf node represents a class label or a numerical value. There are possibilities that a single decision tree might suffer the problem of overfitting, meaning it may be too complex and unable to generalize well on new data.

This is where Random Forest comes for rescue. Random Forest combines multiple decision trees to form a forest of trees. Each tree is constructed on various subsets of the given dataset. By doing this so, the trees in the forest are decorrelated, meaning they are less likely to make the same mistakes, and more likely to capture different aspects of the data. The final prediction of the Random Forest is then determined by aggregating the predictions of all the trees in the forest. This eventually leads to the concept of ensemble learning resulting in reducing overfitting and helping to improve the generalization performance of the model. This is an effective machine learning model that also results in providing better accuracy.

DATA COLLECTION

Context and intent of choosing a secondary dataset from Kaggle- Ride costs of Uber and Lyft are not constant like those on public transportation. They are significantly impacted by the availability and demand for transportation at any particular moment. What then precisely fuels this demand? The time of day would be the initial assumption; because most people commute to work or home between 9 am and

5 pm, these hours experience the largest surges. The weather could also play a factor, as more people should ride rides when it's raining or snowing. The real-time data using Uber and Lyft api requests and corresponding weather conditions in the absence of any publicly available information about rides/prices given by any company.

The data was collected in custom application Scala and was saved in DynamoDB. The data was collected for every cab ride with a 5 minutes gap and weather data every one hour.

The Cab ride data includes the price for various Uber and Lyft cab types in the specified location. Additionally, you can see if there was a price increase at that time. The data contains parameters like distance, the type of the cab, time stamp, surge multiplier, ID of the ride, destination and source, product ID and the name of the cab.

Weather data includes weather characteristics for all areas taken into account, such as temperature, rain, clouds, humidity and wind.

Description of features- The dataset contains these independent variables which might contribute to the surge multiplier. The distance variable provides distance between the source and destination. The price gives us the price estimate in USD. The default value of the surge multiplier is set to 1. Visible type of the cab is given by the variable's name.

In the weather data, the temperature is given in Fahrenheit and 12 unique location values are recorded. The pressure is given in millibar and the rain is given in inches for the last hour. The wind speed is given in miles per hour. The time stamp is the epoch time when the data was collected.

PROPOSED WORK

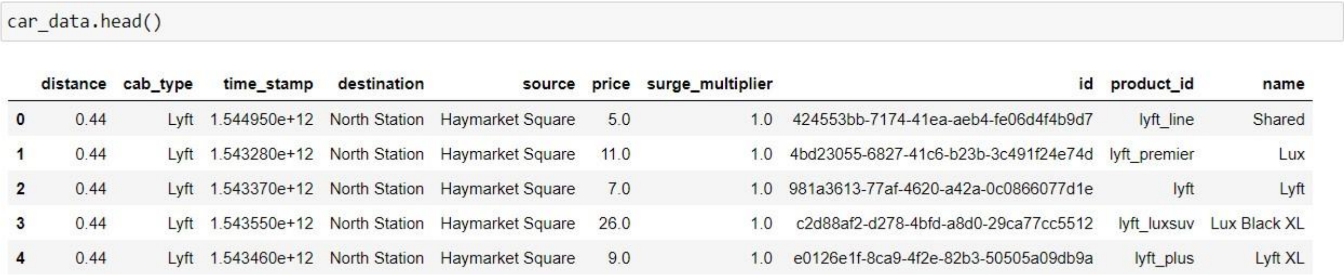
As mentioned above in this project we aim to build a machine learning model that can predict the surge multiplier based on the weather data.

The model will also suggest the most cost- effective cab service based on the price using statistical techniques, enabling users to make informed decisions about which service to use. So the process involved in building the model is discussed below.

DATA PREPROCESSING

Data preprocessing involves preparing the data for machine learning algorithms to improve their accuracy and for better deployment of the model. This involves cleaning the data where we impute or remove the missing values, transforming the data into standard forms, and taking up relevant features into consideration.

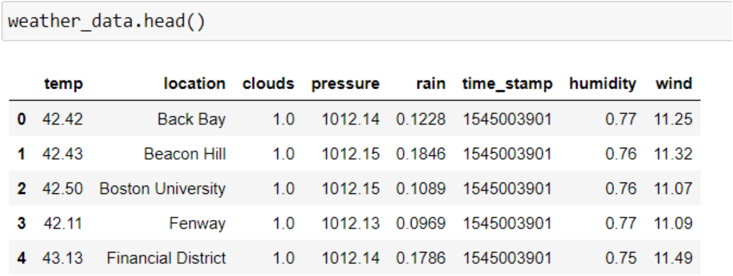
We are provided with two datasets. We have cab_rides.csv and weather.csv. We've 6,93,071 records with 12 features in our cab_rides dataset, while in weather there are 6276 records with 10 features. The head of the both datasets is displayed below.



```
car_data.head()
```

	distance	cab_type	time_stamp	destination	source	price	surge_multiplier	id	product_id	name
0	0.44	Lyft	1.544950e+12	North Station	Haymarket Square	5.0	1.0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	lyft_line	Shared
1	0.44	Lyft	1.543280e+12	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux
2	0.44	Lyft	1.543370e+12	North Station	Haymarket Square	7.0	1.0	981a3613-77af-4620-a42a-0c0866077d1e	lyft	Lyft
3	0.44	Lyft	1.543550e+12	North Station	Haymarket Square	26.0	1.0	c2d88af2-d278-4bfd-a8d0-29ca77cc5512	lyft_luxsuv	Lux Black XL
4	0.44	Lyft	1.543460e+12	North Station	Haymarket Square	9.0	1.0	e0126e1f-8ca9-4f2e-82b3-50505a09db9a	lyft_plus	Lyft XL

Fig 1. Cab rides data

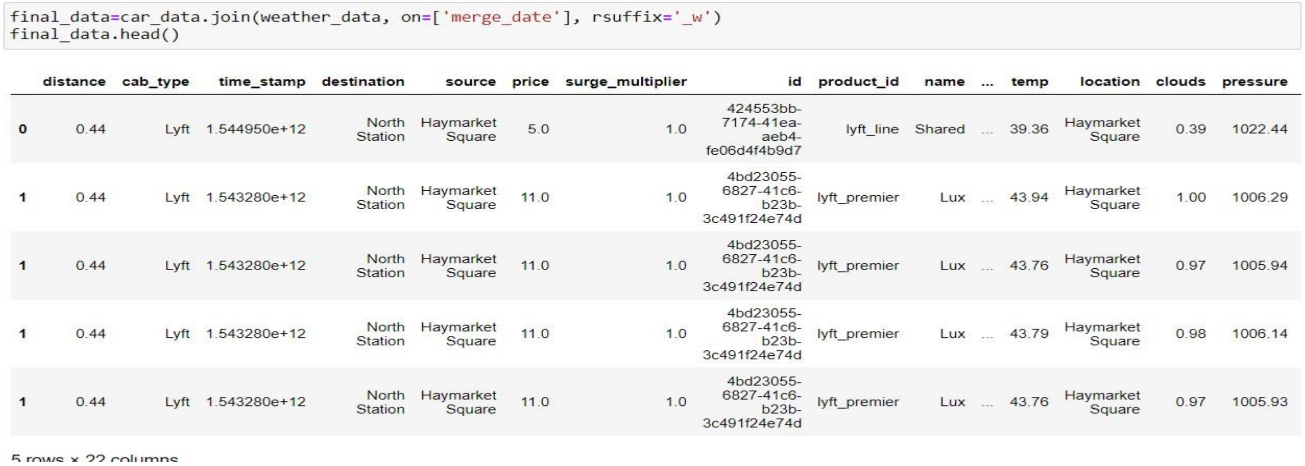


```
weather_data.head()
```

	temp	location	clouds	pressure	rain	time_stamp	humidity	wind
0	42.42	Back Bay	1.0	1012.14	0.1228	1545003901	0.77	11.25
1	42.43	Beacon Hill	1.0	1012.15	0.1846	1545003901	0.76	11.32
2	42.50	Boston University	1.0	1012.15	0.1089	1545003901	0.76	11.07
3	42.11	Fenway	1.0	1012.13	0.0969	1545003901	0.77	11.09
4	43.13	Financial District	1.0	1012.14	0.1786	1545003901	0.75	11.49

Fig 2. Weather data

We've to merge these data to predict the surge prices based on weather. We've time_stamp data as a common feature in both the datasets. Based on this we merge the data and it can be seen that the time_stamp values are in Unix time format so we convert the observation into readable and desired format. After merging both the datasets we have 12,69,926 records with 22 features and is displayed below



```
final_data=car_data.join(weather_data, on=['merge_date'], rsuffix='_w')
final_data.head()
```

	distance	cab_type	time_stamp	destination	source	price	surge_multiplier	id	product_id	name	...	temp	location	clouds	pressure
0	0.44	Lyft	1.544950e+12	North Station	Haymarket Square	5.0	1.0	424553bb-7174-41ea-aeb4-fe06d4f4b9d7	lyft_line	Shared	...	39.36	Haymarket Square	0.39	1022.44
1	0.44	Lyft	1.543280e+12	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux	...	43.94	Haymarket Square	1.00	1006.29
1	0.44	Lyft	1.543280e+12	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux	...	43.76	Haymarket Square	0.97	1005.94
1	0.44	Lyft	1.543280e+12	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux	...	43.79	Haymarket Square	0.98	1006.14
1	0.44	Lyft	1.543280e+12	North Station	Haymarket Square	11.0	1.0	4bd23055-6827-41c6-b23b-3c491f24e74d	lyft_premier	Lux	...	43.76	Haymarket Square	0.97	1005.93

5 rows × 22 columns

Fig 3. Merged data

Now we proceed with the next step which is handling the missing values. There are plenty of missing values within our merged data. As we are predicting the surge multiplier based on weather conditions, the data that are missing related to weather must be dropped. This is because if we impute those missing values with any of the traditional methods our model might not perform well and we also ignore the surge values that are more than 3 because the samples are very less for surge multipliers greater than three.

We plot the correlation matrix in order to know how variables are related to each other. This correlation matrix also helps us in checking multicollinearity which occurs when two variables are highly correlated which can lead to inaccurate parameter estimates.. In our case we don’t have the problem of multicollinearity while our variables are both positively and negatively correlated

Exploratory Data Analysis

Now we will explore the data by performing some basic Exploratory Data Analysis in order to get a better understanding of the data. We plot the histogram of all numerical variables to know the distribution of the dataset and also to know the frequency of values within the specified range where we can create a separate histogram for each column in the Data frame. Snippets of histogram are displayed below.

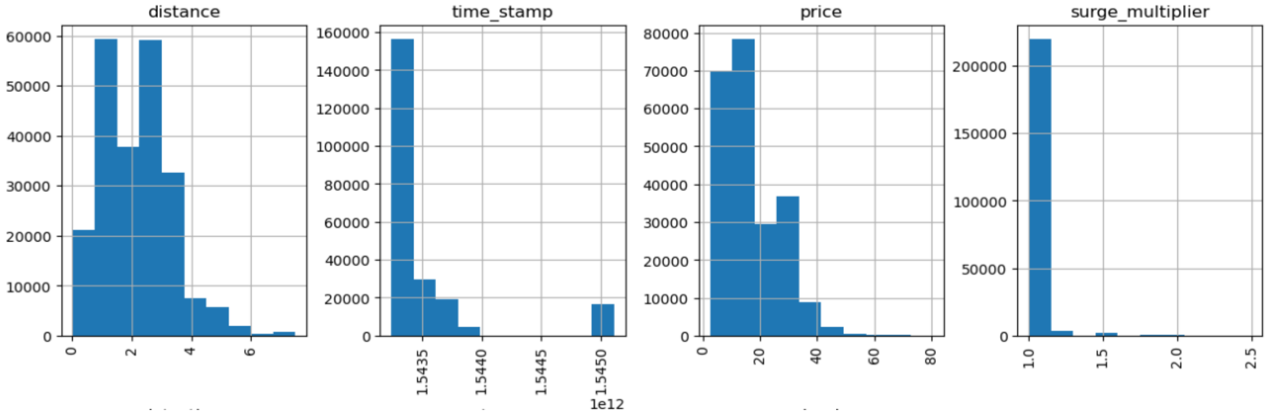


Fig 3. Histogram of our merged data

As we have an idea on how well the numerical variables are distributed, we now focus on exploring the categorical variables. We plot a count plot on the types of cab, sources and destinations of the cabs and results are displayed below

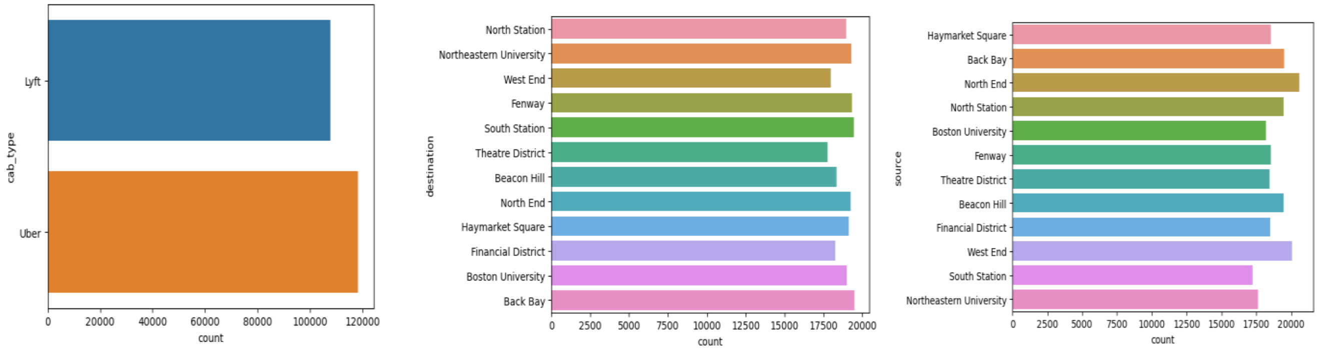


Fig 4. Countplot of categorical variables

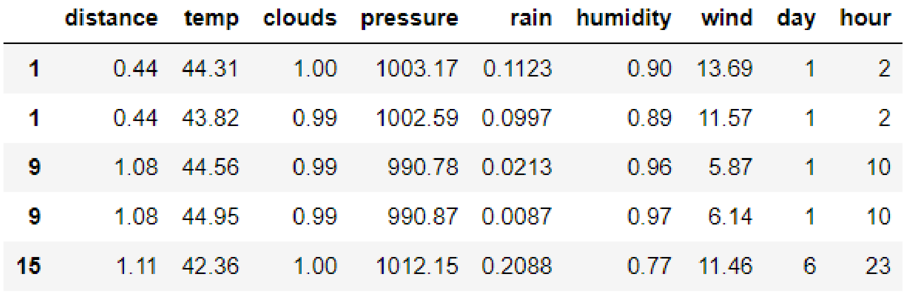
MODEL BUILDING

In this project, we are using Random Forest Classifier to classify the Surge Multiplier into 6 classes i.e., [1, 1.25, 1.5, 1.75, 2.0, 2.5] . Random Forest Classifier is a classification algorithm based on ensemble learning, specifically the Random Forest Technique. The algorithm creates multiple decision trees, each of which is trained on a random subset of the features and data points. The prediction of the random forest classifier is then the mode of the predictions of all the decision trees.

This algorithm is very popular due to its ability to handle large datasets with many features and its resistance to overfitting. It is also relatively easy to use and can handle both categorical and numerical data.

As our main objective is to predict the Surge multipliers based on the weather conditions, we need to select the appropriate features such that we can reduce the complexity and increase accuracy of our model. In this case, we are using Feature Selection Technique to select the features which have more effect on surge multiplier and storing them in x variable whereas the target variable y contains surge multipliers.

The important features which affect the surge multipliers are distance, temperature, clouds, pressure, rain, humidity, wind, day of the week and hour.



	distance	temp	clouds	pressure	rain	humidity	wind	day	hour
1	0.44	44.31	1.00	1003.17	0.1123	0.90	13.69	1	2
1	0.44	43.82	0.99	1002.59	0.0997	0.89	11.57	1	2
9	1.08	44.56	0.99	990.78	0.0213	0.96	5.87	1	10
9	1.08	44.95	0.99	990.87	0.0087	0.97	6.14	1	10
15	1.11	42.36	1.00	1012.15	0.2088	0.77	11.46	6	23

Fig 5. x data

While solving the classification problem, we need to label the surge multipliers present in the data as {1→0, 1.25→1, 1.5→2, 1.75→3, 2→4, 2.5→5} using Label Encoder.

Random Forest Classifier makes use of decision trees and decision trees are sensitive to the scale of the input features. Features with larger values may dominate the others leading to biased results. Thus, Standard Scaling is used to transform the features to have a mean of 0 and a standard deviation of 1, ensuring that all features have equal weight in the decision-making process.

Then, PCA is applied to reduce the dimensionality of the input features while retaining 90% of the original information. Using PCA before Random Forest Classifier can reduce the computational complexity of the algorithm, improve model performance and reduce the risk of overfitting.

Once that is done, we can split the data into training and testing dataset with test size as 0.3 and import the Random Forest Classifier model from sklearn library.

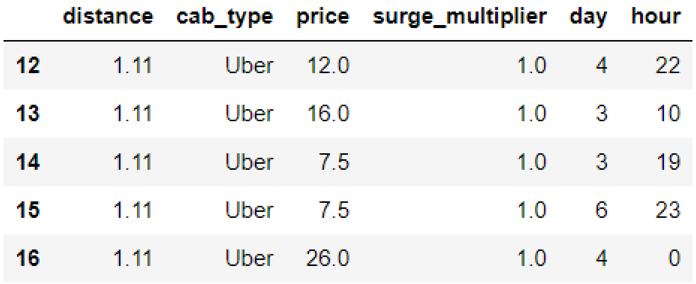
Random Forest Classifier uses the parameter n_jobs =1 which specifies the CPU to use all available cores for parallel processing during training and prediction and parameter class_weight = ‘balanced’ which specifies the algorithm to automatically adjust the weights of the different classes based on their frequencies in the training data. This is useful when the dataset has imbalanced classes, where one or more classes have a much smaller number of samples than the others.

The next part of the project comes in the form of Comparative Analysis where the purpose is to gain a deeper understanding of the two cab companies Uber and Lyft using Statistical tests.

The Statistical tests used here are t-test and ANOVA test. The t-test is a statistical hypothesis test used to determine whether there is a significant difference between the means of two groups. The t-test calculates the t-statistic which is a ratio of the difference between the sample means to the standard error of the difference between the means. The t-statistic is compared to a critical value from the t-distribution based on the degrees of freedom and the level of significance chosen for the test. If the calculated t-statistic is greater than the critical value, then we reject the null hypothesis that the means are equal and conclude that there is a significant difference between the means.

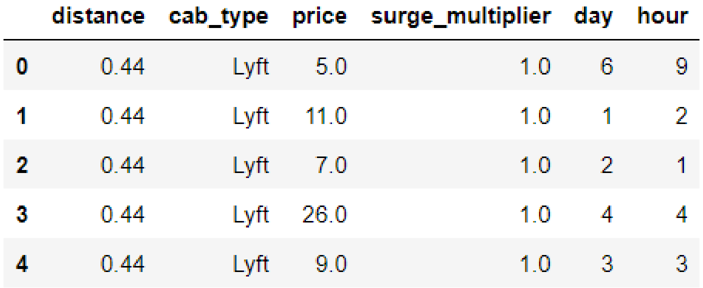
ANOVA is a statistical technique used to analyze the differences between two or more groups or treatments. It is a hypothesis testing method that determines whether the means of the groups are significantly different from each other.

Here, we split the car dataset into two datasets based on their cab type - Uber and Lyft



	distance	cab_type	price	surge_multiplier	day	hour
12	1.11	Uber	12.0	1.0	4	22
13	1.11	Uber	16.0	1.0	3	10
14	1.11	Uber	7.5	1.0	3	19
15	1.11	Uber	7.5	1.0	6	23
16	1.11	Uber	26.0	1.0	4	0

Fig 6. Uber data



	distance	cab_type	price	surge_multiplier	day	hour
0	0.44	Lyft	5.0	1.0	6	9
1	0.44	Lyft	11.0	1.0	1	2
2	0.44	Lyft	7.0	1.0	2	1
3	0.44	Lyft	26.0	1.0	4	4
4	0.44	Lyft	9.0	1.0	3	3

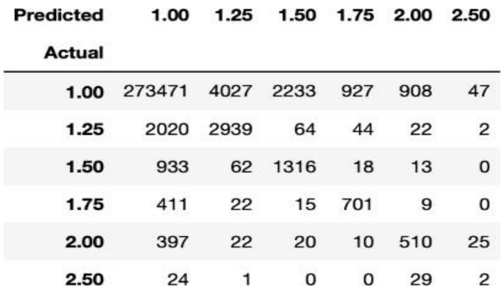
Fig 7. Lyft data

t-test is applied on both the datasets and means are compared based on the prices. Similarly for ANOVA test, a simple linear regression model is fit to the dataframe with distance as the response variable and price as the predictor variable. The OLS function comes from the statsmodel library and returns a RegressionResults object that contains information about the model fir. This model is then fitted to the data. Then, ANOVA is computed and the resulting ANOVA table provides information on the significance of each term in the model as well as the overall goodness of fit of the model.

Then, Tukey's HSD (honestly significant difference) test is performed to determine which groups have significantly different means from each other. It is considered a post-hoc test because it is used after the ANOVA has already shown that there is a significant difference between the means of at least two groups.

EVALUATION

From the model we built, we got the F1 score which is a measure of a test's accuracy that takes both precision and recall into account as 96% and accuracy of the model as 95%. So we can say that the ML model is doing quite a decent job here. The below diagram shows the complete confusion matrix



Predicted	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**	**2.50**
Actual						
1.00	273471	4027	2233	927	908	47
1.25	2020	2939	64	44	22	2
1.50	933	62	1316	18	13	0
1.75	411	22	15	701	9	0
2.00	397	22	20	10	510	25
2.50	24	1	0	0	29	2

Fig 8. Confusion matrix

RESULTS AND DISCUSSION

The accuracy of the Model built to predict the surge multipliers was found out to be 95%

We also found out the features in the dataset that have the most influence or impact on the surge multipliers. An additional advantage of Random Forest is that the importance of every feature comes as a by-product of training.

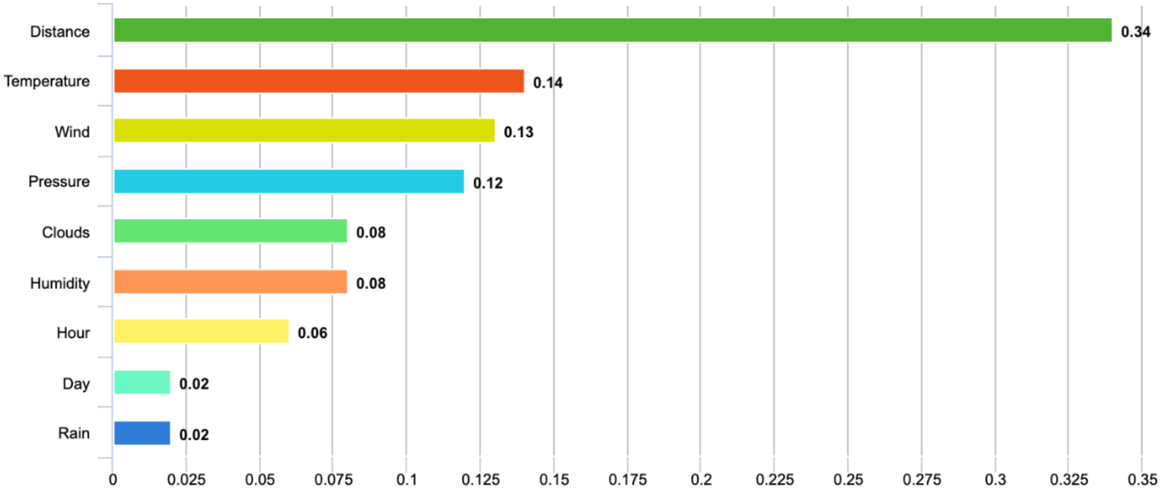
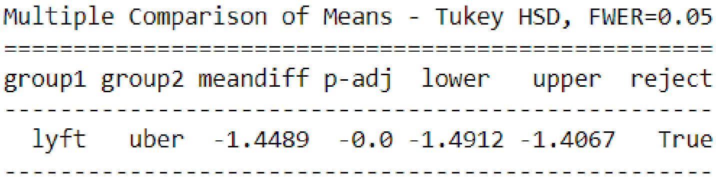


Fig 9. Feature importance

It can be seen that the model is dependent on the distance feature the most and subsequently on the other features represented in the decreasing order of importance.

After performing a t-test we found that there is a significant difference between the mean value of prices in Uber and Lyft. We also found that Lyft has a lower mean price and is considered to be better. But the discussion arises whether we can predict which cab type is better based on only prices, thus this can be continued as a future work.

After performing ANOVA test and Tukey’s HSD test, we got the following table where the reject column shows that the null hypothesis of equal means for the Lyft and Uber groups was rejected at the chosen significance level of 0.05, indicating that the mean price of one of these groups is significantly different from the other and the meandiff column shows that the mean price for the Lyft dataset is lower than the mean price for the Uber dataset, with a difference of -1.4489. The lower and upper columns show the lower and upper bounds of the 95% confidence interval for the mean difference, respectively.



```
Multiple Comparison of Means - Tukey HSD, FWER=0.05
===================================================
group1 group2 meandiff p-adj  lower   upper  reject
---------------------------------------------------
  lyft   uber  -1.4489  -0.0 -1.4912 -1.4067   True
---------------------------------------------------
```

Fig 10. Tukey HSD Table

CONCLUSION AND FUTURE WORK

This project has the potential to greatly improve the user experience of cab services by providing more accurate surge predictions and cost-effective suggestions. Additionally, cab companies could use this model to optimize their pricing strategies and improve their overall service.

Our future work focuses on implementing this model for huge volumes of data or multiple geographies where our algorithms must optimize to those locations which will be helpful for organizations when they certainly plan to expand their business across regions.

REFERENCES

[1] Kun-Huang Huarng , Tiffany Hui-Kuang Yu “Impact of Surge Pricing on Customer Retention” , November 2020, Journal of Business Research ,Volume 120

[2] Matthew Battifarano, Zhen (Sean) Qian “Predicting real-time surge pricing of ride-sourcing companies” October 2019,Transportation Research Part C: Emerging Technologies, Volume 107

[3] Sumit Agarwall, Ben Charoenwong, Shih-Fen Cheng and Jussi Keppo “ The Impact of ride-hail surge factors on taxi bookings” March 20222, Transportation Research Part C: Emerging Technologies, Volume 136

[4] Hongzhi Cheng, Weiwei Chen, Jingyi Ma and Jianwei Huang “Integrated reward Scheme and surge pricing in a ridesourcing market” April 2020, Transportation Research Part B: Methodological, Volume 134

[5] Gerard P. Cachon, Kaitlin M. Daniels and Ruben Lobel “The Role of Surge Pricing on a Service Platform with Self-Scheduling Capacity” January 2019, Sharing Economy (SSSCM), Volume6.