Fare Price Prediction

Vishal Nadar(G01276965)

Sumanth Pola(G01330039)

December 12, 2020

Abstract

Uber is the world’s first and largest ride-sharing company that helps to connect riders and local drivers with customers. UberX and Lyft claim to charge 30% less than taxis – a good way to get the customers' attention. Nowadays, we see Machine Learning and Artificial Intelligence in almost all fields, so we try to use the same to predict the cab ride price. In this project, we did experimentation on the real-life dataset and explored how machine learning algorithms could be utilized to find different patterns in data. We primarily discussed the price prediction of different Uber and Lyft cabs and how various attributes like weather, temperature, wind, and other factors impact its pricing. This problem can be classified as a regression supervised learning category. We used different machine learning algorithms, such as Linear Regression, Decision Tree, Random Forest Regressor, XGBoost, and Gradient Boosting Regressor. Finally, we choose the one that proves best for the price prediction of cab rides. We decided the algorithm which increases the accuracy and decreases overfitting to get the best results. We found that XGBoost gave the best accuracy with three parameters tuned for price prediction out of all the algorithms we used. We also created a function to predict the price of a cab ride when the user provides the corresponding attributes of the cab ride. While preparing the Uber and Lyft Dataset of Boston for 2018, we gained a lot of experience. It was also fascinating to learn how many factors influence Uber and Lyft cab pricing.

1. INTRODUCTION

Uber Technologies, Inc., or Uber is an American mobility service company established in San Francisco that operates in over 900 cities worldwide. Its services include ride-hailing, food delivery (Uber Eats and Postmates), package delivery, couriers, freight transportation, electric bicycle and motorized scooter rental through a Lime partnership, and ferry transportation through local operators. Uber does not own any vehicles. Instead, it earns a 25% commission on each booking. Fares are quoted in advance to the consumer, but they change depending on the local supply and demand at the time of the booking. So, for our project, we are considering factors such as temperature, weather, visibility, wind, and many more to predict the price of the cab ride.

We have a training set and a test set in supervised learning. The supervised learning algorithm aims to infer a function that maps the input vector to the output vector with a low error rate from the training and test set. In the Uber and Lyft Dataset of Boston, we used machine learning techniques to estimate the price of the cab ride. From a total of 55 columns, several features will be chosen. All NULL values will be replaced, and different machine learning techniques will be applied to get the machine learning algorithm with the best accuracy. This machine-learning algorithm will predict the price of the cab ride when the user enters the attributes.

2. PROBLEM STATEMENT

There is heavy competition in every field in today's environment, including offering cab rides at the lowest price. Here we build a model that can predict the lower price for cab rides by including factors other than distance, i.e., temperature, weather, visibility, wind, time, and many other factors. Predicting a meager price will be troublesome for the company's profits, and predicting a very high price might cause the user to consider other options. Hence, indicating the price with such factors, the error rate should be as low as possible so that the user gets a fair price and keeps the company's profit in mind.

The additional problem with considering other factors is that computing the price of the cab ride may take extra time. So, calculating a fair price in less computing time is required so that the user doesn't have to wait a long time to get the cost of the cab ride. To solve this issue, we will be selecting only the essential features which give the highest prediction accuracy and in the most efficient time.

3. LITERATURE REVIEW

For the literature review, we looked at various websites and papers. Because this dataset was collected from Kaggle, only a few competitions on the same data set prior. The data were visualized and analyzed by the majority of the participants. Other research has revealed the following factors:

Abel Brodeurand and Kerry Nield (2018) investigate the impact of rain on Uber trips in New York City. Since the launch of Uber rides in May 2011, passengers and fares in all other rides, such as taxi rides, have decreased. In addition, dynamic pricing encourages Uber drivers to compete for rides when demand spikes unexpectedly, such as during rainy hours. With the rise of rain, Uber journeys have increased by 22 percent, while the number of taxi rides per hour has only increased by 5%. Since the introduction of Uber, taxis have not responded differently to increased demand during rainy hours than during non-rainy hours.

When there is a demand-supply imbalance, Uber uses an algorithmic technique called surge pricing. It occurs when both the rider's demand and the driver's availability reduces. During such a time of the rise in demand for rides, fares tend to be high usually. Surge pricing is necessary since it aids in matching the driver's efforts with consumer demand. 2018 (Junfeng Jiao) Uber's surge multiplier was investigated in Austin, Texas. It was discovered that during periods of high usage, Uber would increase their prices to reflect this demand via a surge multiplier. According to (Uber 2015) communications, this pricing is intended to attract more drivers into service at specific hours while simultaneously reducing demand from users. (Sheldon & Chen, 2016) While there is conflicting evidence, surge pricing appears to moderate supply and demand while maintaining wait times under 5 minutes.

Some papers compare the iconic yellow taxi and its modern competitor, Uber. Vsevolod Salnikov, Renaud Lambiotte, Anastasios Noulas, and Cecilia Mascolo, 2014 discovered instances where UberX, the lowest version of the Uber taxi service, is more expensive for the same journey than yellow cabs. According to our findings, depending on the length of the route, tourists may find it more cost-effective to choose Yellow Cabs or Uber. It is, however, the same journey they are willing to undertake that is important.

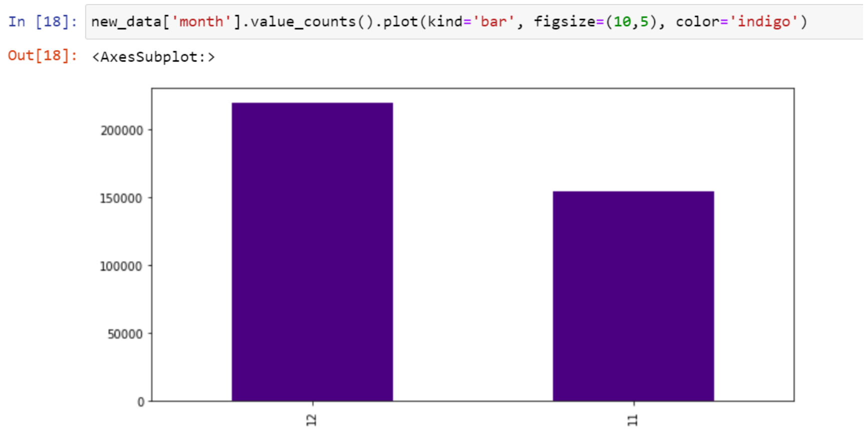
4. METHODS AND TECHNIQUES

4.1 Cleaning and Pre-processing:

Using pd.read_csv, we first opened and saved the data into a data frame. Then we looked to see any duplicate rows and found none. As the data was too large, we only considered the data of certain cab types such as Lux, Lyft, Shared, UberX, Taxi, UberPool, and black. We also checked for null values in our data, and we found that there were 55095 null values in the price column, and we replaced them by the mean of the price column, which is 10.5.

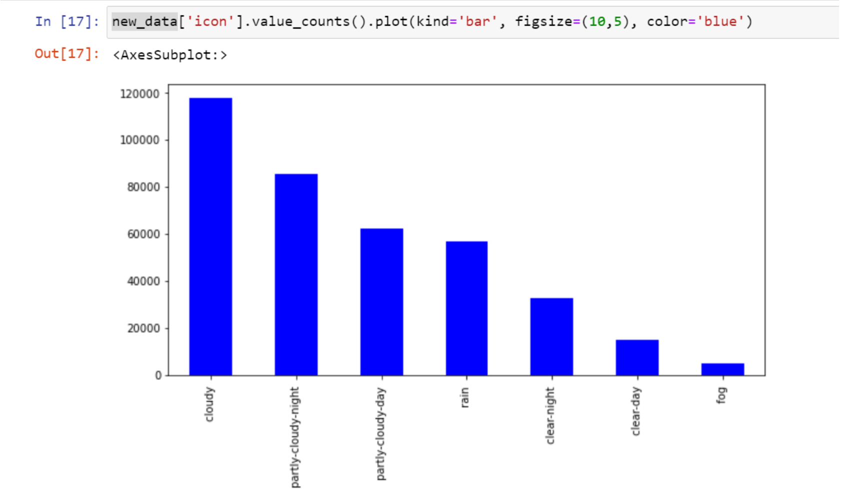
4.2 Data Visualization:

Data Visualization is a technique from which we can derive conclusions from the data. It can be in the form of charts, graphs, maps, and other data visualization tools. We can find patterns, outliers and also understand trends from this technique.



```
In [18]: new_data['month'].value_counts().plot(kind='bar', figsize=(10,5), color='indigo')
Out[18]: <AxesSubplot:>
```

Using this technique, we found that our data is of November and December months. Also, we discovered that cloudy weather has the most data due to which we can conclude that people opted for more cab rides in cloudy weather.



```
In [17]: new_data['icon'].value_counts().plot(kind='bar', figsize=(10,5), color='blue')
Out[17]: <AxesSubplot:>
```

4.3 Feature Engineering:

Our dataset consisted of many data types: an object, float64, and int64. Thus, we used a label encoder to deal with this problem by converting the object data into a numeric value(int). The changed attributes are DateTime, id, timezone, destination, product_id, short_summary, long_summary, name, source, icon, and price. Recursive Feature Elimination (RFE) is used to eliminate features to increase the accuracy and reduce the run-time of the model. The accuracy was calculated using the XGBoost machine learning algorithm, and the best accuracy was found at lower than ten attributes. To reduce the computing time of RFE, we manually removed four more features. We got the best accuracy at six attributes: the source, destination, product_id, name, distance, and surge multiplier.

4.4 Model Building:

Using the sklearn library, we have split the data using the train_test_split function into the ratio of 0.7 to 0.3 of train data and test data. We have implemented various learning algorithms on this split train and test data such as LinearRegression, DecisionTreeRegressor, RandomForestRegressor, GradientBoostingRegressor, and XGBRegressor.

a) **Linear Regression**: Linear Regression is a linear approach for modeling the relationship between a scalar response and one or more explanatory variables. Linear Regression was the first type of regression analysis studied rigorously and used extensively in practical applications. It is a statistical approach that models the relationship between input features and output.

b) **Decision Tree Regressor:** A Decision Tree is a decision-making tool that uses a flowchart-like tree structure or is a model of decisions and all their possible results, including outcomes, input costs, and utility. Decision-Tree algorithm falls under the category of supervised learning algorithms. This model is very good at handling tabular data with numerical or categorical features.

c) **Random Forest Regressor:** A Random Forest is an ensemble technique capable of performing regression and classification tasks using multiple decision trees and a method called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine numerous decision trees to determine the final output rather than rely on individual decision trees.

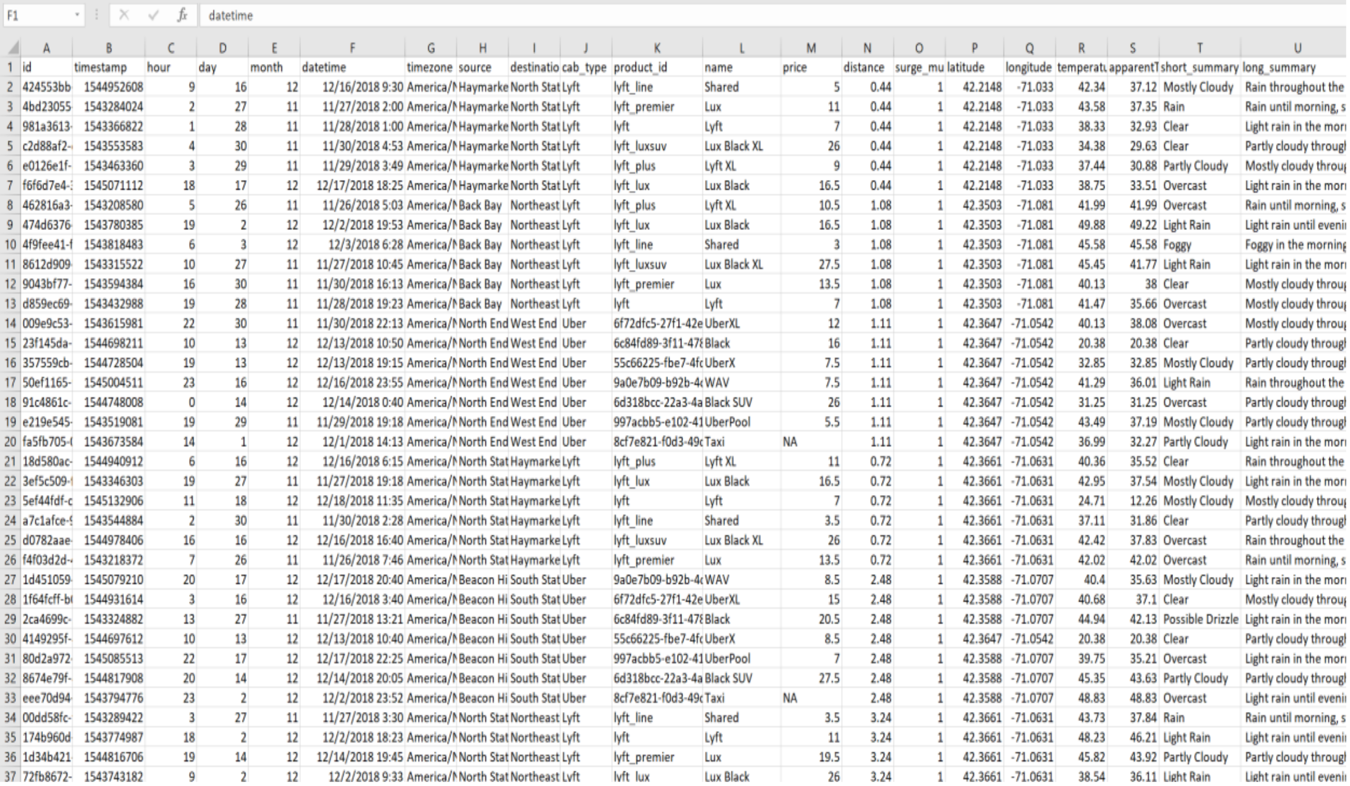
d) **Gradient Boosting Regressor:** This model combines the predictions from multiple decision trees to generate the final predictions. In gradient boosting, each predictor corrects its predecessor's error. In contrast to Adaboost, the weights of the training instances are not tweaked. Instead, each predictor is trained using the residual errors of predecessor as labels. Gradient Boosting trains many models gradually, additive, and sequential.

e) **XGB Regressor:** This algorithm creates decision trees in sequential form. Weights play an essential role in XGBoost. Weights are assigned to all the independent variables, then fed into the decision tree, predicting results. The weight of variables predicted wrong by the tree increases, and the variables are then provided to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model.

5. DISCUSSION AND RESULTS

5.1 Datasets

- The dataset is taken from www.Kaggle.com and is in .csv form.
- The dataset consisted of Uber data and Lyft data. We have considered data of lux, shared, Lyft, taxi, UberX, Uberpool, and black.
- Uber data has 56 features/columns and 322844 entries. It is of shape (322844,56)
- Data has three types of data types which were as follows: - integer, float, and object. Object data was converted to int data type.
- The dataset is not complete, which means we also have null values in a column named price. So, we filled it with the mean price value, which is 10.5**.**



F1 fx datetime

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	timestamp	hour	day	month	datetime	timezone	source	destinatio	cab_type	product_id	name	price	distance	surge_mu	latitude	longitude	temperatu	apparentT	short_summary	long_summary
2	424553bb	1544952608	9	16	12	12/16/2018 9:30	America/N	Haymarke	North Stat	Lyft	lyft_line	Shared	5	0.44	1	42.2148	-71.033	42.34	37.12	Mostly Cloudy	Rain throughout the
3	4bd23055	1543284024	2	27	11	11/27/2018 2:00	America/N	Haymarke	North Stat	Lyft	lyft_premier	Lux	11	0.44	1	42.2148	-71.033	43.58	37.35	Rain	Rain until morning, s
4	981a3613-	1543366822	1	28	11	11/28/2018 1:00	America/N	Haymarke	North Stat	Lyft	lyft	Lyft	7	0.44	1	42.2148	-71.033	38.33	32.93	Clear	Light rain in the mor
5	c2d88af2-	1543553583	4	30	11	11/30/2018 4:53	America/N	Haymarke	North Stat	Lyft	lyft_luxsuv	Lux Black XL	26	0.44	1	42.2148	-71.033	34.38	29.63	Clear	Partly cloudy throug
6	e0126e1f-	1543463360	3	29	11	11/29/2018 3:49	America/N	Haymarke	North Stat	Lyft	lyft_plus	Lyft XL	9	0.44	1	42.2148	-71.033	37.44	30.88	Partly Cloudy	Mostly cloudy throu
7	f6f6d7e4-	1545071112	18	17	12	12/17/2018 18:25	America/N	Haymarke	North Stat	Lyft	lyft_lux	Lux Black	16.5	0.44	1	42.2148	-71.033	38.75	33.51	Overcast	Light rain in the mor
8	462816a3-	1543208580	5	26	11	11/26/2018 5:03	America/N	Back Bay	Northeast	Lyft	lyft_plus	Lyft XL	10.5	1.08	1	42.3503	-71.081	41.99	41.99	Overcast	Rain until morning, s
9	474d6376	1543780385	19	2	12	12/2/2018 19:53	America/N	Back Bay	Northeast	Lyft	lyft_lux	Lux Black	16.5	1.08	1	42.3503	-71.081	49.88	49.22	Light Rain	Light rain until eveni
10	4f9fee41-f	1543818483	6	3	12	12/3/2018 6:28	America/N	Back Bay	Northeast	Lyft	lyft_line	Shared	3	1.08	1	42.3503	-71.081	45.58	45.58	Foggy	Foggy in the morning
11	8612d909	1543315522	10	27	11	11/27/2018 10:45	America/N	Back Bay	Northeast	Lyft	lyft_luxsuv	Lux Black XL	27.5	1.08	1	42.3503	-71.081	45.45	41.77	Light Rain	Light rain in the mor
12	9043bf77-	1543594384	16	30	11	11/30/2018 16:13	America/N	Back Bay	Northeast	Lyft	lyft_premier	Lux	13.5	1.08	1	42.3503	-71.081	40.13	38	Clear	Mostly cloudy throu
13	d859ec69-	1543432988	19	28	11	11/28/2018 19:23	America/N	Back Bay	Northeast	Lyft	lyft	Lyft	7	1.08	1	42.3503	-71.081	41.47	35.66	Overcast	Mostly cloudy throu
14	009e9c53-	1543615981	22	30	11	11/30/2018 22:13	America/N	North End	West End	Uber	6f72dfc5-27f1-42e	UberXL	12	1.11	1	42.3647	-71.0542	40.13	38.08	Overcast	Mostly cloudy throu
15	23f145da-	1544698211	10	13	12	12/13/2018 10:50	America/N	North End	West End	Uber	6c84fd89-3f11-478	Black	16	1.11	1	42.3647	-71.0542	20.38	20.38	Clear	Partly cloudy throug
16	357559cb-	1544728504	19	13	12	12/13/2018 19:15	America/N	North End	West End	Uber	55c66225-fbe7-4fc	UberX	7.5	1.11	1	42.3647	-71.0542	32.85	32.85	Mostly Cloudy	Partly cloudy throug
17	50ef1165-	1545004511	23	16	12	12/16/2018 23:55	America/N	North End	West End	Uber	9a0e7b09-b92b-4c	WAV	7.5	1.11	1	42.3647	-71.0542	41.29	36.01	Light Rain	Rain throughout the
18	91c4861c-	1544748008	0	14	12	12/14/2018 0:40	America/N	North End	West End	Uber	6d318bcc-22a3-4a	Black SUV	26	1.11	1	42.3647	-71.0542	31.25	31.25	Overcast	Partly cloudy throug
19	e219e545-	1543519081	19	29	11	11/29/2018 19:18	America/N	North End	West End	Uber	997acbb5-e102-41	UberPool	5.5	1.11	1	42.3647	-71.0542	43.49	37.19	Mostly Cloudy	Partly cloudy throug
20	fa5fb705-	1543673584	14	1	12	12/1/2018 14:13	America/N	North End	West End	Uber	8cf7e821-f0d3-49c	Taxi	NA	1.11	1	42.3647	-71.0542	36.99	32.27	Partly Cloudy	Light rain in the mor
21	18d580ac-	1544940912	6	16	12	12/16/2018 6:15	America/N	North Stat	Haymarke	Lyft	lyft_plus	Lyft XL	11	0.72	1	42.3661	-71.0631	40.36	35.52	Clear	Rain throughout the
22	3ef5c509-	1543346303	19	27	11	11/27/2018 19:18	America/N	North Stat	Haymarke	Lyft	lyft_lux	Lux Black	16.5	0.72	1	42.3661	-71.0631	42.95	37.54	Mostly Cloudy	Light rain in the mor
23	5ef44fdf-c	1545132906	11	18	12	12/18/2018 11:35	America/N	North Stat	Haymarke	Lyft	lyft	Lyft	7	0.72	1	42.3661	-71.0631	24.71	12.26	Mostly Cloudy	Mostly cloudy throu
24	a7c1afce-	1543544884	2	30	11	11/30/2018 2:28	America/N	North Stat	Haymarke	Lyft	lyft_line	Shared	3.5	0.72	1	42.3661	-71.0631	37.11	31.86	Clear	Partly cloudy throug
25	d0782aae-	1544978406	16	16	12	12/16/2018 16:40	America/N	North Stat	Haymarke	Lyft	lyft_luxsuv	Lux Black XL	26	0.72	1	42.3661	-71.0631	42.42	37.83	Overcast	Rain throughout the
26	f4f03d2d-	1543218372	7	26	11	11/26/2018 7:46	America/N	North Stat	Haymarke	Lyft	lyft_premier	Lux	13.5	0.72	1	42.3661	-71.0631	42.02	42.02	Overcast	Rain until morning, s
27	1d451059	1545079210	20	17	12	12/17/2018 20:40	America/N	Beacon Hi	South Stat	Uber	9a0e7b09-b92b-4c	WAV	8.5	2.48	1	42.3588	-71.0707	40.4	35.63	Mostly Cloudy	Light rain in the mor
28	1f64fcff-b	1544931614	3	16	12	12/16/2018 3:40	America/N	Beacon Hi	South Stat	Uber	6f72dfc5-27f1-42e	UberXL	15	2.48	1	42.3588	-71.0707	40.68	37.1	Clear	Mostly cloudy throu
29	2ca4699c-	1543324882	13	27	11	11/27/2018 13:21	America/N	Beacon Hi	South Stat	Uber	6c84fd89-3f11-478	Black	20.5	2.48	1	42.3588	-71.0707	44.94	42.13	Possible Drizzle	Light rain in the mor
30	4149295f-	1544697612	10	13	12	12/13/2018 10:40	America/N	Beacon Hi	South Stat	Uber	55c66225-fbe7-4fc	UberX	8.5	2.48	1	42.3647	-71.0542	20.38	20.38	Clear	Partly cloudy throug
31	80d2a972	1545085513	22	17	12	12/17/2018 22:25	America/N	Beacon Hi	South Stat	Uber	997acbb5-e102-41	UberPool	7	2.48	1	42.3588	-71.0707	39.75	35.21	Overcast	Light rain in the mor
32	8674e79f-	1544817908	20	14	12	12/14/2018 20:05	America/N	Beacon Hi	South Stat	Uber	6d318bcc-22a3-4a	Black SUV	27.5	2.48	1	42.3588	-71.0707	45.35	43.63	Partly Cloudy	Partly cloudy throug
33	eee70d94	1543794776	23	2	12	12/2/2018 23:52	America/N	Beacon Hi	South Stat	Uber	8cf7e821-f0d3-49c	Taxi	NA	2.48	1	42.3588	-71.0707	48.83	48.83	Overcast	Light rain until eveni
34	00dd58fc-	1543289422	3	27	11	11/27/2018 3:30	America/N	North Stat	Northeast	Lyft	lyft_line	Shared	3.5	3.24	1	42.3661	-71.0631	43.73	37.84	Rain	Rain until morning, s
35	174b960d	1543774987	18	2	12	12/2/2018 18:23	America/N	North Stat	Northeast	Lyft	lyft	Lyft	11	3.24	1	42.3661	-71.0631	48.23	46.21	Light Rain	Light rain until eveni
36	1d34b421	1544816706	19	14	12	12/14/2018 19:45	America/N	North Stat	Northeast	Lyft	lyft_premier	Lux	19.5	3.24	1	42.3661	-71.0631	45.82	43.92	Partly Cloudy	Partly cloudy throug
37	72fb8672-	1543743182	9	2	12	12/2/2018 9:33	America/N	North Stat	Northeast	Lyft	lyft_lux	Lux Black	26	3.24	1	42.3661	-71.0631	38.54	36.11	Light Rain	Light rain until eveni

5.2 Evaluation Metrics

We used Recursive Feature Elimination (RFE) to reduce the number of features and used XGBoost to check the accuracy for each iteration.

Serial No.	No. of Feature	Accuracy (XGBoost)
1	55	0.9468444817
2	25	0.9468021057
3	40	0.9464620950
4	30	0.9466622373
5	15	0.9468355822
6	10	0.9470181364

We checked the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean squared error (MSE) of all algorithms and found that xgboost had the least values as follows:

MAE: 0.8261228071948997

MSE: 1.7183157404282017

RMAE: 1.3108454296476766

5.3 Experimental Results

Algorithm	Accuracy
LinearRegression	0.6662976321197515
DecisionTreeRegressor	0.9468883043112083
RandomForestRegressor	0.9481346720904356
GradientBoostingRegressor	0.9475575610252001
XGBRegressor	0.9482508867426854

The accuracy is calculated by executing .score(X_test, y_test) on all machine learning algorithms and we found that XGBoost gave us the best accuracy

Further Tuning of our XGBoost Model using Hyperparameters:

We also experimented with some parameters on XGBoost such as turning eta to 0.6, max_depth=8 and min_child_weight=1.5 increased the accuracy of the model to 0.9492495232473771. More parameters were also experimented with such as gamma=0.1, max_delta_step=0.1, subsample=0.9, and colsample_bytree=0.9 but all further reduced the accuracy of the model.

We have also written a sample code for the user to input the details for each attribute and then the price of a cab ride is predicted and shown.

6 CONCLUSION

- The dataset is cleaned of any duplicate values, which were none, or any null values, which were replaced by the mean of the column(price).
- Recursive Feature Elimination (RFE) reduces the number of attributes to 10 and encodes object datatypes to int using label encoder.
- We have successfully implemented different models on our dataset to get the accuracy of each algorithm, among which Random Forest and XGBoost gave us the highest accuracy of 0.9481 and 0.9482, respectively.
- We tweaked the parameters of our XGBoost classifier to improve the model and reduce overfitting.
- With the information on attributes, the user can input the value of each six attributes and predict the price of the cab ride.
- Hence, we have successfully implemented a model to predict the price of the cab ride.

6.1 Directions for Future Work

- We would like to predict the cost of cab in real time which could be used in apps to predict cost and show it to the user instantly.
- We would like to include more features and work on more larger datasets so that the cost can be further optimized.

REFERENCES

- Abel Brodeurand & Kerry Nield (2018) An empirical analysis of taxi, Lyft and Uber rides: Evidence from weather shocks in NYC
- Junfeng Jiao (2018) Investigating Uber price surges during a special event in Austin, TX
- Anastasios Noulas, Cecilia Mascolo, Renaud Lambiotte, and Vsevolod Salnikov (2014) OpenStreetCab: Exploiting Taxi Mobility Patterns in New York City to Reduce Commuter Costs
- https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston
- https://www.analyticsvidhya.com/blog/2016/03/complete-guide-parameter-tuning-xgboost-with-codes-python/
- https://gdcoder.com/decision-tree-regressor-explained-in-depth/

- https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
- https://xgboost.readthedocs.io/en/stable/parameter.html
- https://www.geeksforgeeks.org/xgboost-for-regression/
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
- https://datatofish.com/plot-dataframe-pandas/
- https://jakevdp.github.io/PythonDataScienceHandbook/04.14-visualization-with-seaborn.html
- https://thispointer.com/pandas-find-duplicate-rows-in-a-dataframe-based-on-all-or-selected-columns-using-dataframe-duplicated-in-python/
- https://blog.paperspace.com/implementing-gradient-boosting-regression-python/
- https://www.studytonight.com/post/what-is-mean-squared-error-mean-absolute-error-root-mean-squared-error-and-r-squared
- https://www.askpython.com/python/built-in-methods/unique-values-from-a-dataframe
- https://statisticsbyjim.com/regression/overfitting-regression-models/