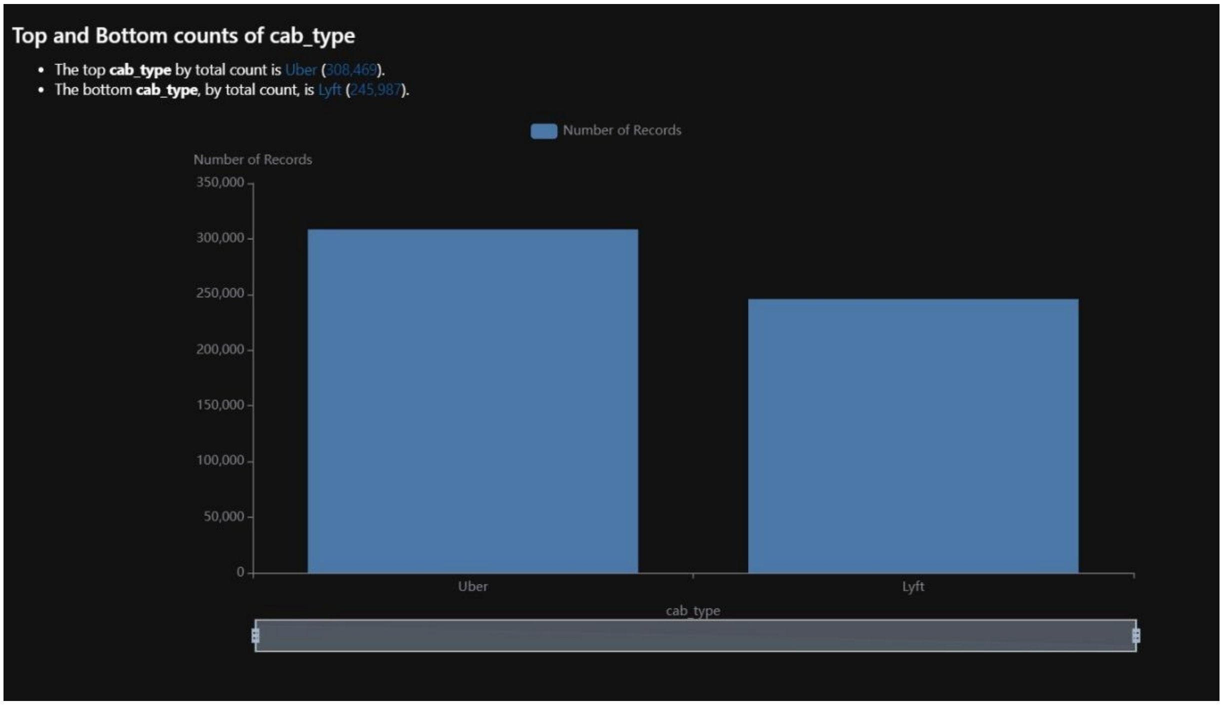
TAXI PRICE PRDICTION

Machine Learning Project

MAY 28, 2022

Phase 1

- **Pre-processing:**
  - **<u>“taxi_rides”:</u>**



Top and Bottom counts of cab_type

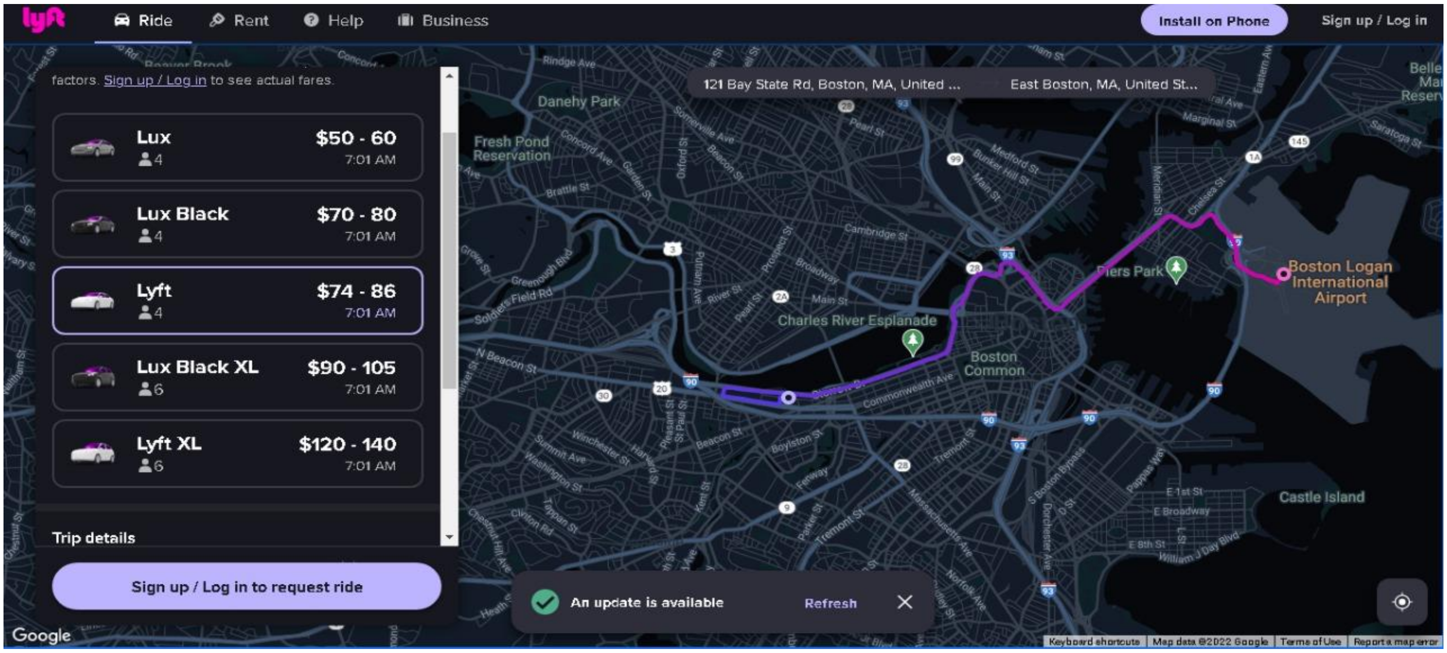
- The top **cab_type** by total count is Uber (308,469).
- The bottom **cab_type**, by total count, is Lyft (245,987).

  - “cab_type” is object, it’s either “uber” or “lyft”, so one hot
  - “time_stamp” is float and it indicates time so it’s converted to datetime.
  - “price” contains null values in taxi, so it’s calculated by getting the price of the trips that are done by uberX which has the same source and destination.
  - “time_stamp” was replaced by “date” and “hour” and was dropped later.
  - “id” values are unique, so it doesn’t help in prediction of the “price”. It will be dropped later after merge step.
  - “lyft” has 6 product IDs, uber doesn’t have any. Count of “name” is identical to count of “product_id”, so a map was made to map “product_id” to “name.
  - “product_id” was dropped as “name” can replace it.

- For “lyft” cabs, names were compared to see how they affect the price, we got the description from official company website:https://help.lyft.com/hc/ru/articles/115012927427-Lyft-ride-modes-overview

  and the order is:
  - Shared: Share a car with riders headed in the same direction at a discounted price.
  - Lyft: Standard Lyft car for up to 3* riders
  - Lyft xl: SUV for up to 5* riders
  - Lux: Luxury car for up to 3* riders
  - Lux black: Premium black car service with leather seats for up to 3* riders
  - Lux black xl: Premium black SUV with leather seats for up to 5* riders

- Ordinal encoding was done based on this order on “lyft_types”



121 Bay State Rd, Boston, MA, United ...

East Boston, MA, United St...

factors. Sign up / Log in to see actual fares.

Lux	4	$50 - 60 7:01 AM
Lux Black	4	$70 - 80 7:01 AM
Lyft	4	$74 - 86 7:01 AM
Lux Black XL	6	$90 - 105 7:01 AM
Lyft XL	6	$120 - 140 7:01 AM

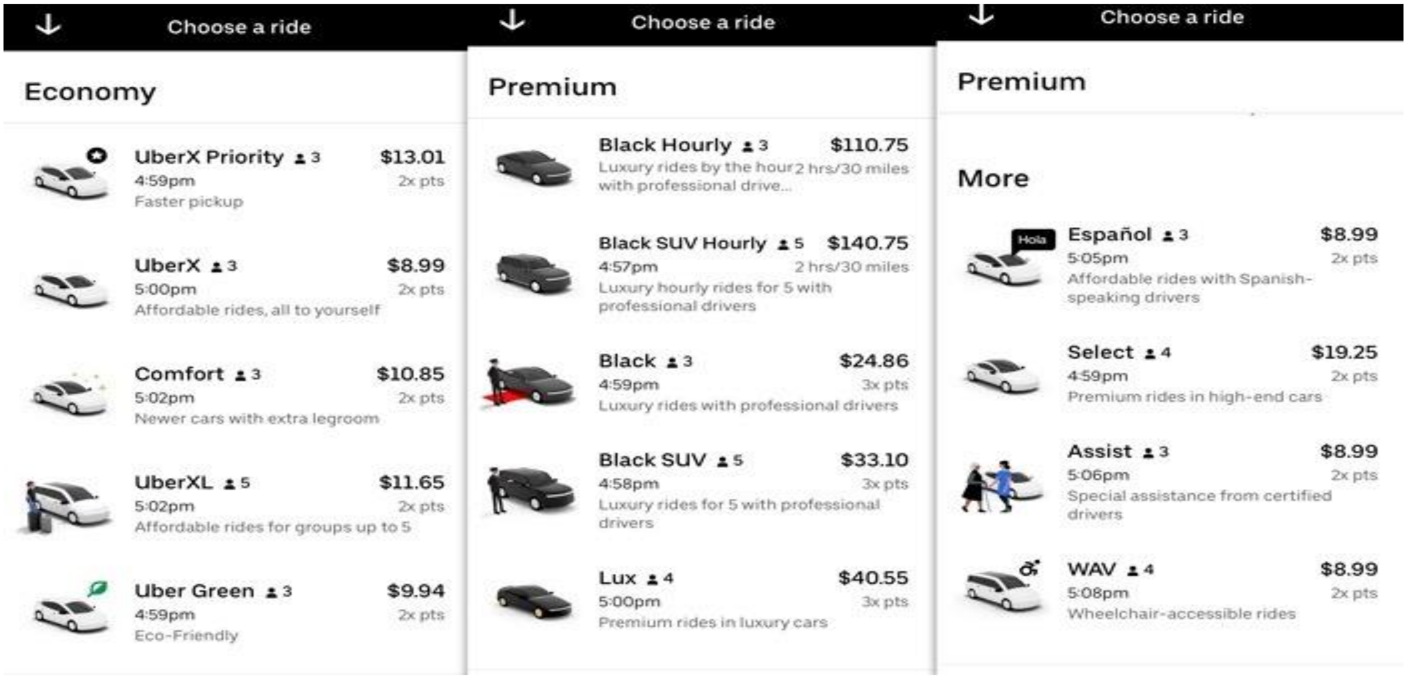
- Price differences between the Uber ride types:
  The cost of your Uber ride is largely determined by the Uber service that you select.

  The costs of the different services, from least expensive to most expensive: Uber Pool, Uber X, Uber Comfort, Uber XL, Uber Select, Uber Black, Uber SUV.

  An example Uber from The Grove to the Century City Mall in Los Angeles
  Note: This ride is 4.6 miles and 18 minutes. The price may change due to traffic, time of day, or discounts

Uber Ride Type	Est. Ride Cost (4.6 miles, 18 minutes)
Pool	$9-$11
X	$9-$12
Comfort	$12-$16
XL	$15-$20
Select	$24-$30
Black	$30-$40
Black SUV	$42-$52

- For “uber” cabs, names were compared to see how they affect the price, the order is:



Choose a ride

Economy

UberX Priority 👤 3 $13.01
4:59pm 2x pts
Faster pickup

UberX 👤 3 $8.99
5:00pm 2x pts
Affordable rides, all to yourself

Comfort 👤 3 $10.85
5:02pm 2x pts
Newer cars with extra legroom

UberXL 👤 5 $11.65
5:02pm 2x pts
Affordable rides for groups up to 5

Uber Green 👤 3 $9.94
4:59pm 2x pts
Eco-Friendly

Choose a ride

Premium

Black Hourly 👤 3 $110.75
Luxury rides by the hour 2 hrs/30 miles
with professional drive...

Black SUV Hourly 👤 5 $140.75
4:57pm 2 hrs/30 miles
Luxury hourly rides for 5 with professional drivers

Black 👤 3 $24.86
4:59pm 3x pts
Luxury rides with professional drivers

Black SUV 👤 5 $33.10
4:58pm 3x pts
Luxury rides for 5 with professional drivers

Lux 👤 4 $40.55
5:00pm 3x pts
Premium rides in luxury cars

Choose a ride

Premium

More

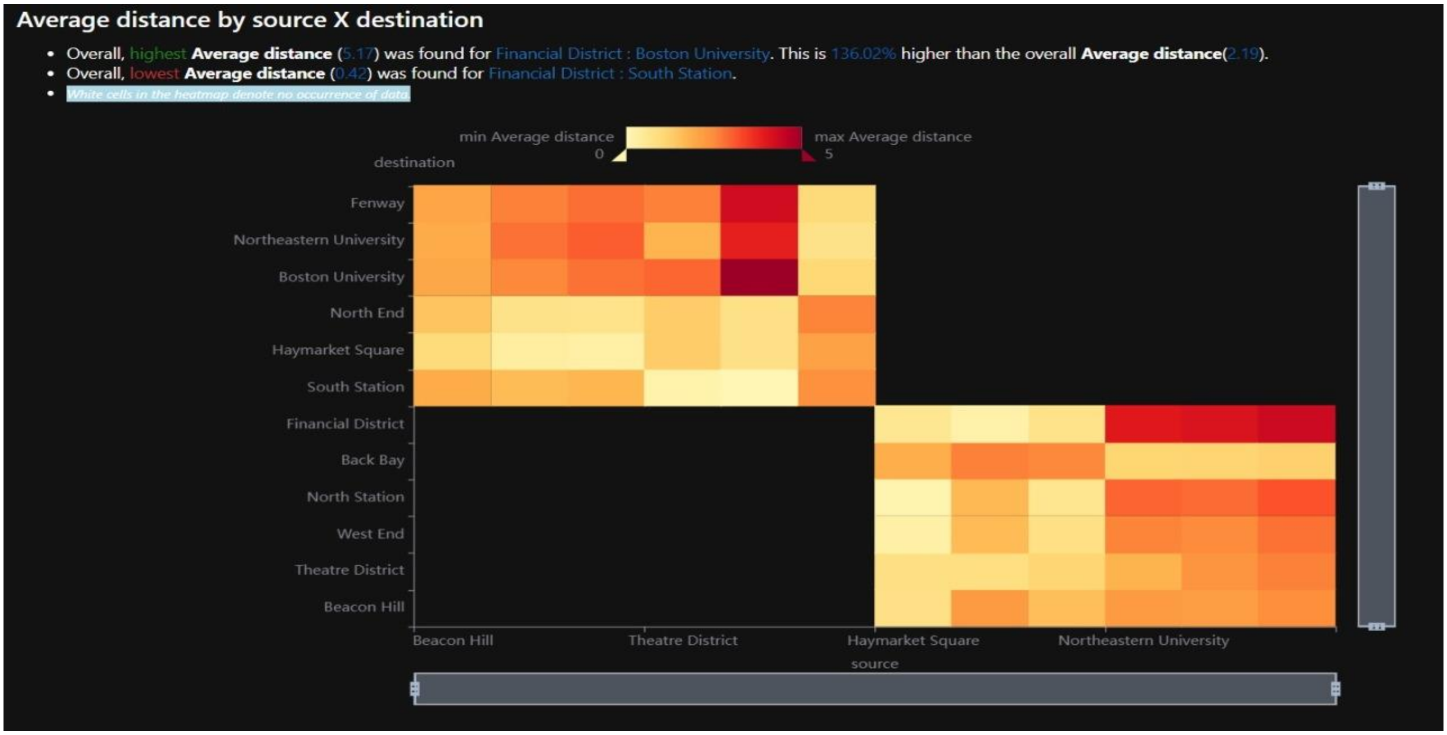
Español 👤 3 $8.99
5:05pm 2x pts
Affordable rides with Spanish-speaking drivers

Select 👤 4 $19.25
4:59pm 2x pts
Premium rides in high-end cars

Assist 👤 3 $8.99
5:06pm 2x pts
Special assistance from certified drivers

WAV 👤 4 $8.99
5:08pm 2x pts
Wheelchair-accessible rides

        - UberPool
        - Taxi
        - UberX
        - UberXl
        - WAV
        - Black
        - Black SUV
    - Ordinal encoding was done based on this order on “uber_types”
    - “cab_type” for both “lyft” and “uber” were dropped as their values were encoded in “lyft_type” and “uber_type”
    - In “distance” there were outliers, so they were removed it .



Average distance by source X destination

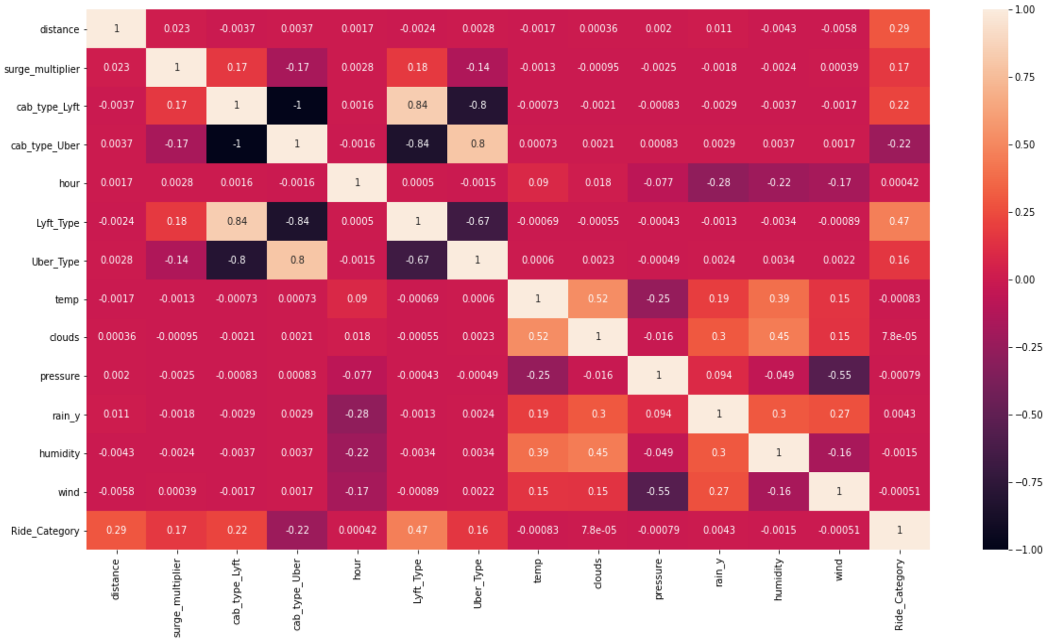
- Overall, highest **Average distance** (5.17) was found for Financial District : Boston University. This is 136.02% higher than the overall **Average distance**(2.19).
- Overall, lowest **Average distance** (0.42) was found for Financial District : South Station.
- *White cells in the heatmap denote no occurrence of data.*

    - **<u>In “distance“: we got all street name from {data/Boston/street_name website } using web scraping then we got the latitude and longitude of every street name using google.geoglocation api , so we can calculate distance easily between any source and destination.</u>**

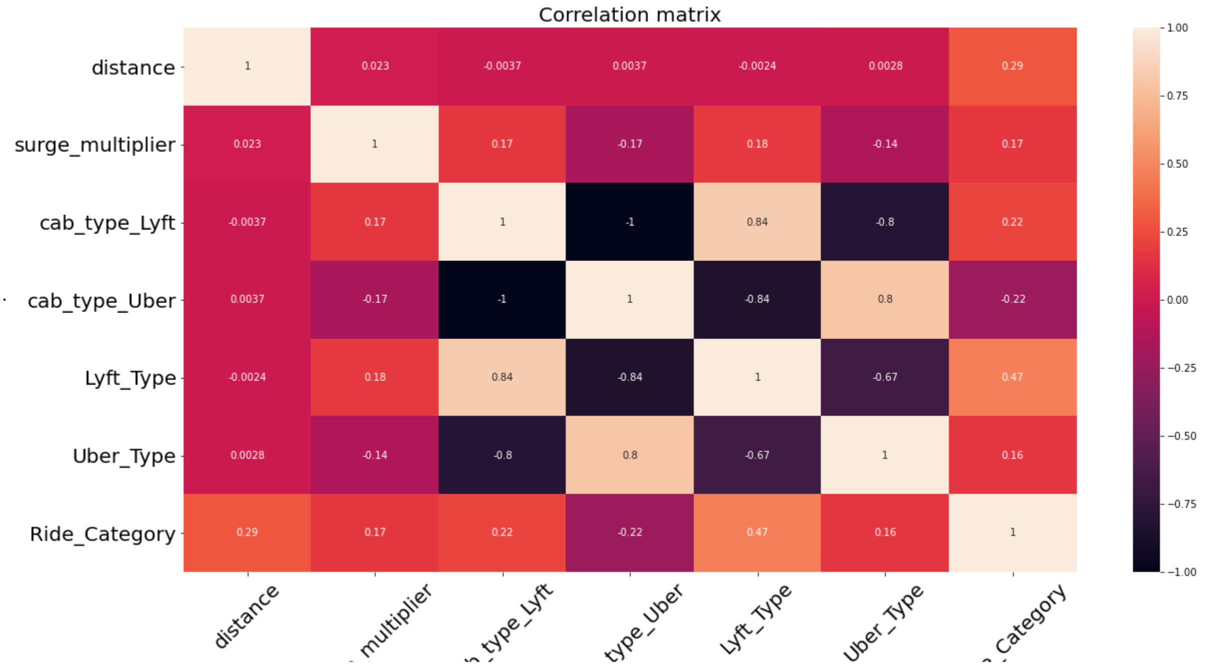
- **“weather”:**
- “time_stamp” is float, so it’s converted to datetime.
- For “rain” 85% of the data is null, K-Nearest Neighbors imputation method was used. Normalizing data was applied in order not to generate biased replacement for the missing values.
- Regression techniques:
    - Polynomial regression
    - Multiple regression
- For Multiple regression:
    - Mean Square Error 16.33271939415199
    - r2_score: 75.16022240596294 %
    - Training time: 0.10199832916259766 seconds
    - For Polynomial regression: Degree : 6
    - Mean Square Error 3.203520780009517
    - r2_score: 95.94640692535074 %
    - Training time 0.12099742889404297 seconds
- Features used for regression are:
    - Distance
    - surge_multiplier
    - Lyft_Type
    - Uber_Type
    - cab_type_Uber

- Training set size is 60% and testing set size is 40%

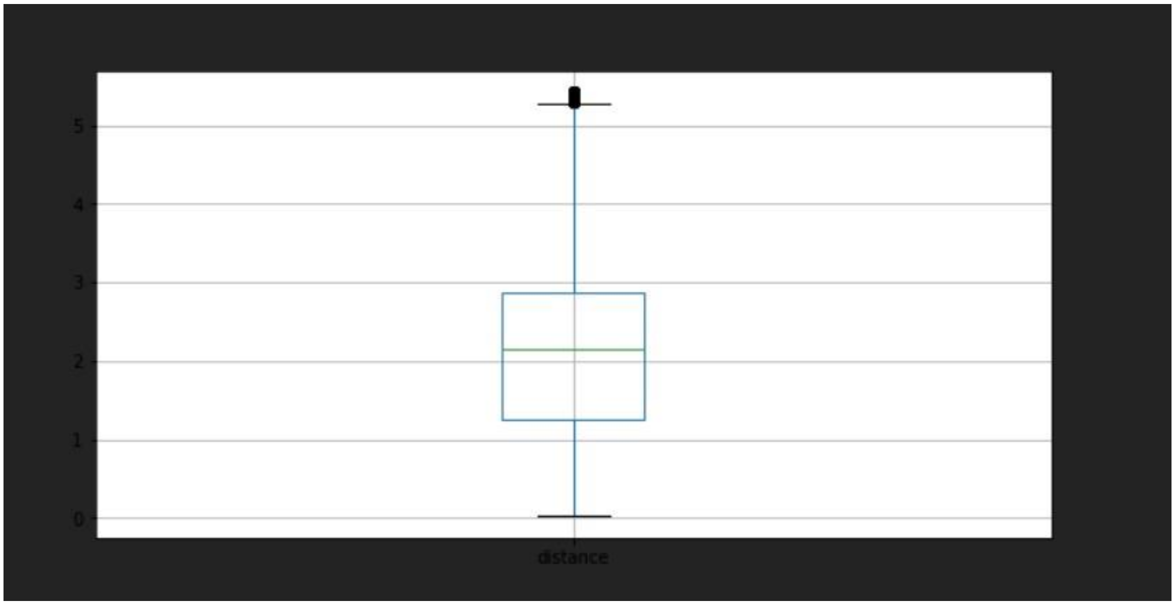
How data affect each other



- How data affect each other after feature selection



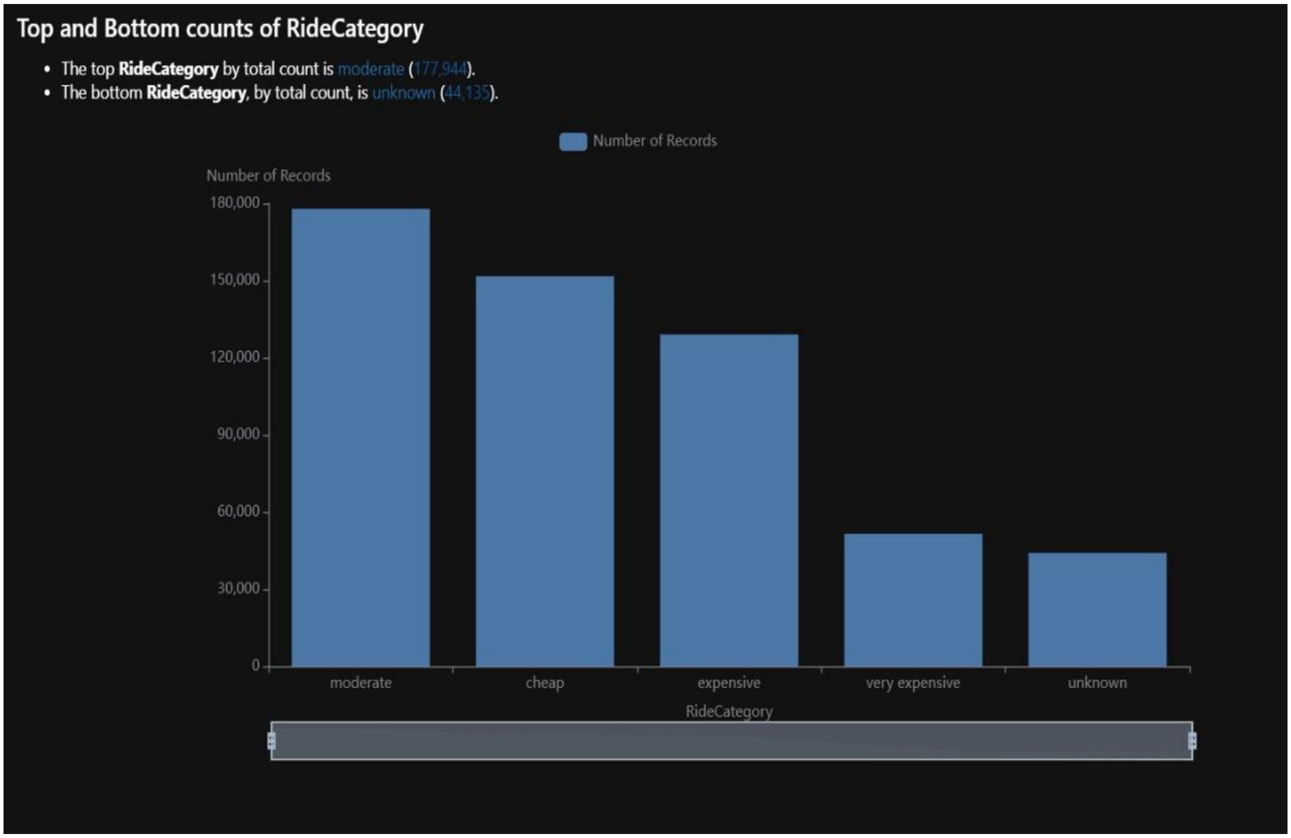
- Data visualization for “distance” outliers:



- Conclusion:
  - In this phase, we applied preprocessing for all features and feature selection and we have concluded that weather is not a good indicator for price and the product id (name) & distance was the main indicators for price. But for surge it was very effective in calculating the price for Lyft cabs but not for Uber cabs .

Phase 2

- **Pre-processing:**



Top and Bottom counts of RideCategory

- The top **RideCategory** by total count is moderate (177,944).
- The bottom **RideCategory**, by total count, is unknown (44,135).

We do the same preprocessing as we did in phase 1 and we noticed that Ridge Category in order of:

1: unknown

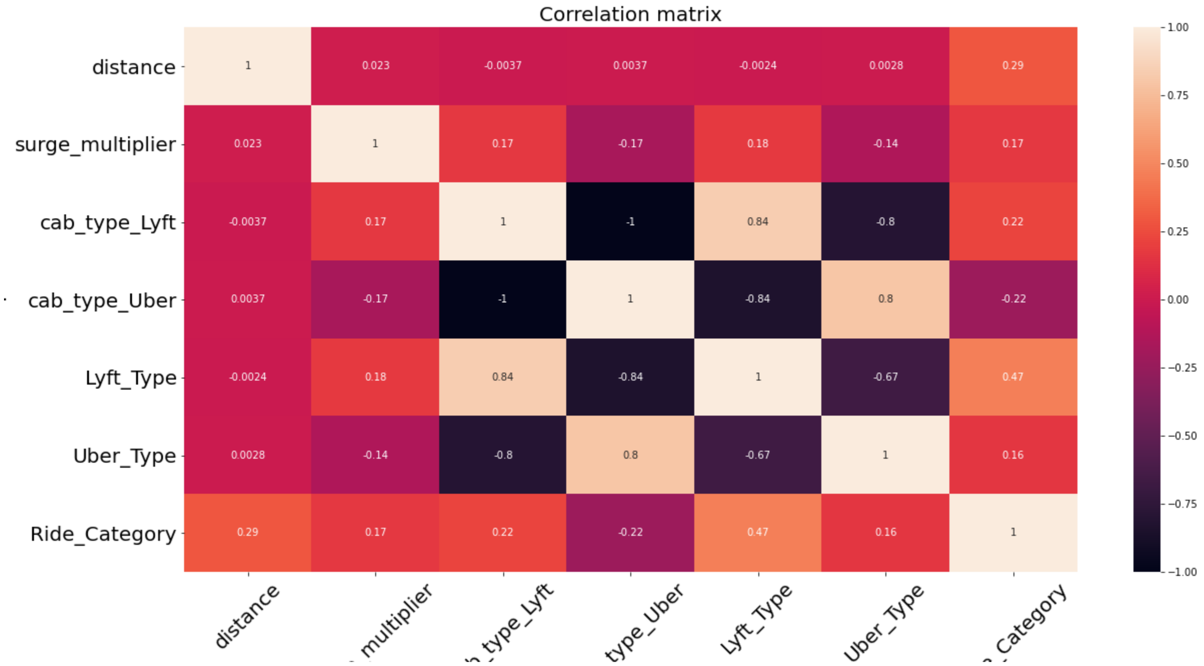
2: cheap

3: moderate

4: expensive

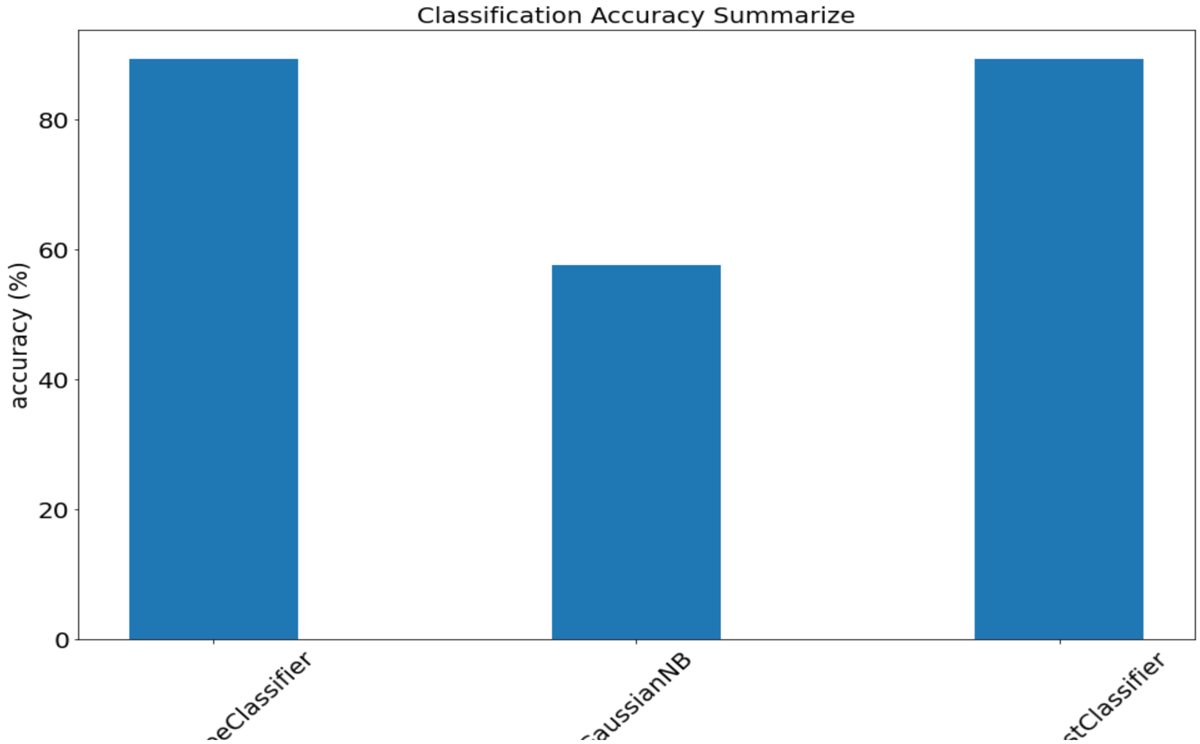
5: very expensive

Feature Selection:

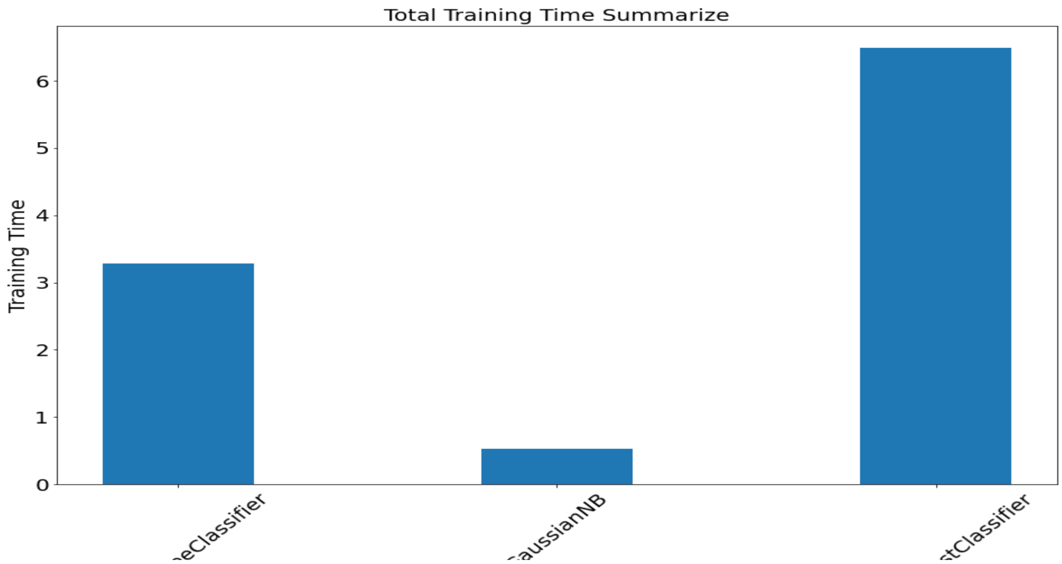


Summarization:

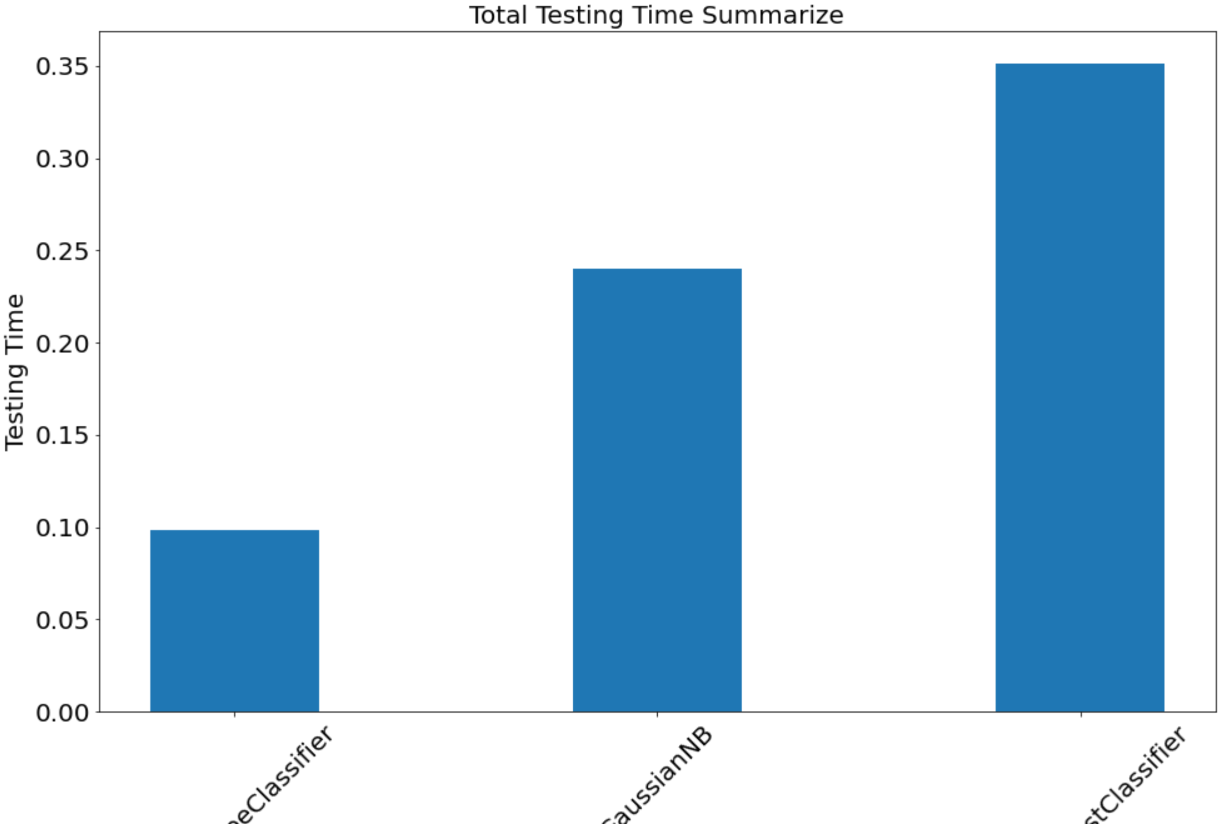
1: classification accuracy



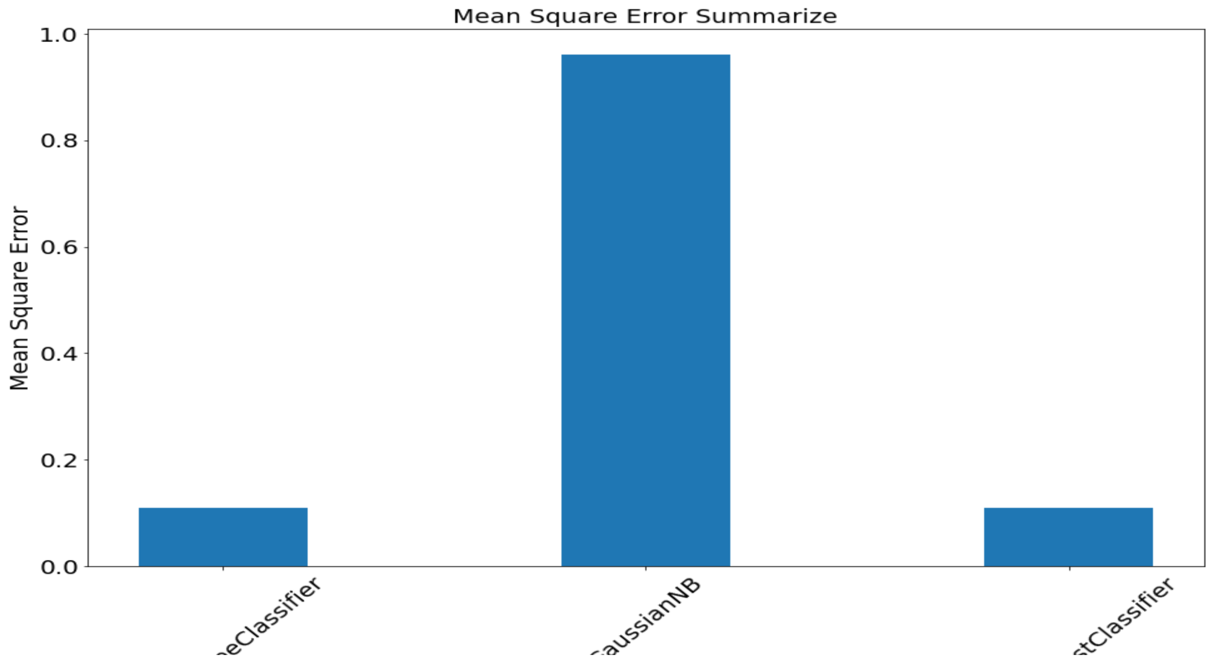
2: Total training time



3: Total test time



4: Mean Square Error

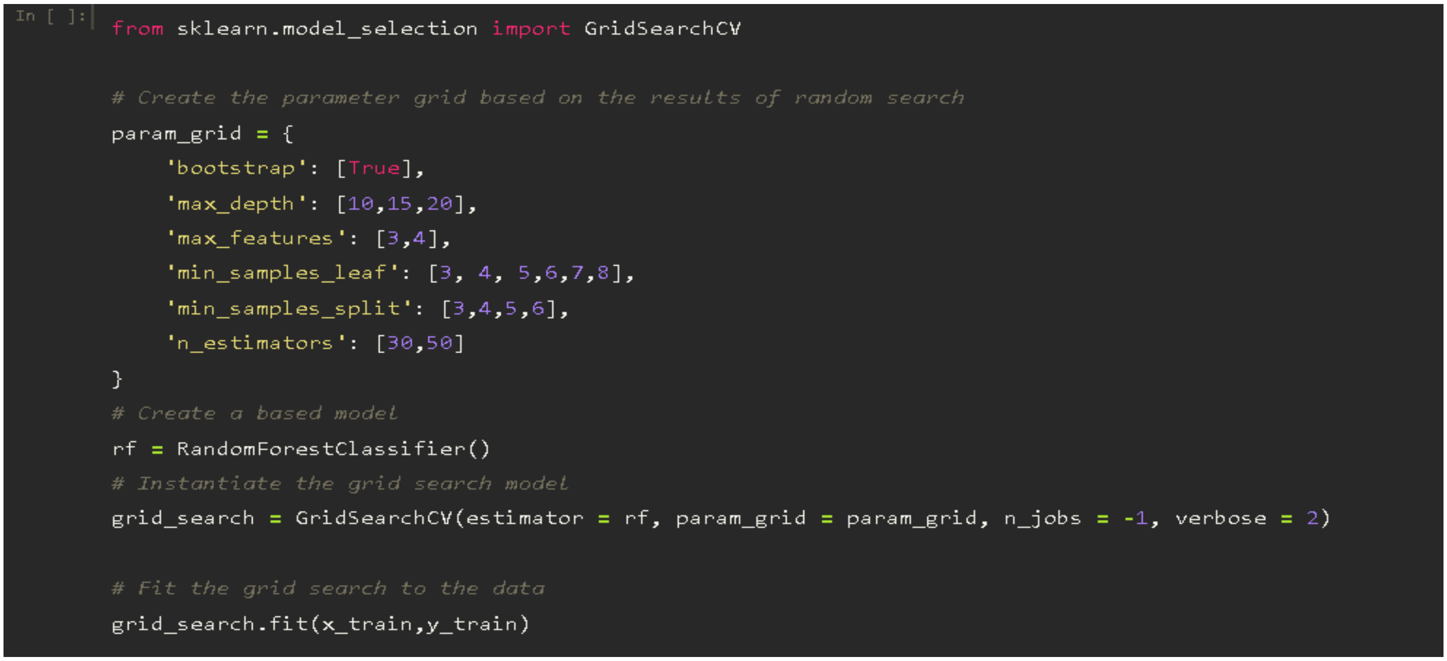


Hyperparameter tuning:

RandomForestClassifier's hyper parameters.

{'bootstrap': True, 'ccp_alpha': 0.0, 'class_weight': None, 'criterion': 'gini', 'max_depth': None, 'max_features': 4, 'max_leaf_nodes': None, 'max_samples': None, 'min_impurity_decrease': 0.0, 'min_impurity_split': None, 'min_samples_leaf': 1, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0, 'n_estimators': 3, 'n_jobs': None, 'oob_score': False, 'random_state': None, 'verbose': 0, 'warm_start': False}

Try to Explain in detail how hyperparameter tuning affected RandomForestClassifier models’performance.



```
In [ ]: from sklearn.model_selection import GridSearchCV

        # Create the parameter grid based on the results of random search
        param_grid = {
            'bootstrap': [True],
            'max_depth': [10,15,20],
            'max_features': [3,4],
            'min_samples_leaf': [3, 4, 5,6,7,8],
            'min_samples_split': [3,4,5,6],
            'n_estimators': [30,50]
        }
        # Create a based model
        rf = RandomForestClassifier()
        # Instantiate the grid search model
        grid_search = GridSearchCV(estimator = rf, param_grid = param_grid, n_jobs = -1, verbose = 2)

        # Fit the grid search to the data
        grid_search.fit(x_train,y_train)
```

Android Application

Tools:

1: Heroku Cloud: we upload our model on Heroku cloud

So, we can use model anytime.

2: Flask Framework: to connect model with local server as framework.

3: Android Studio: we use it to deploy our application.

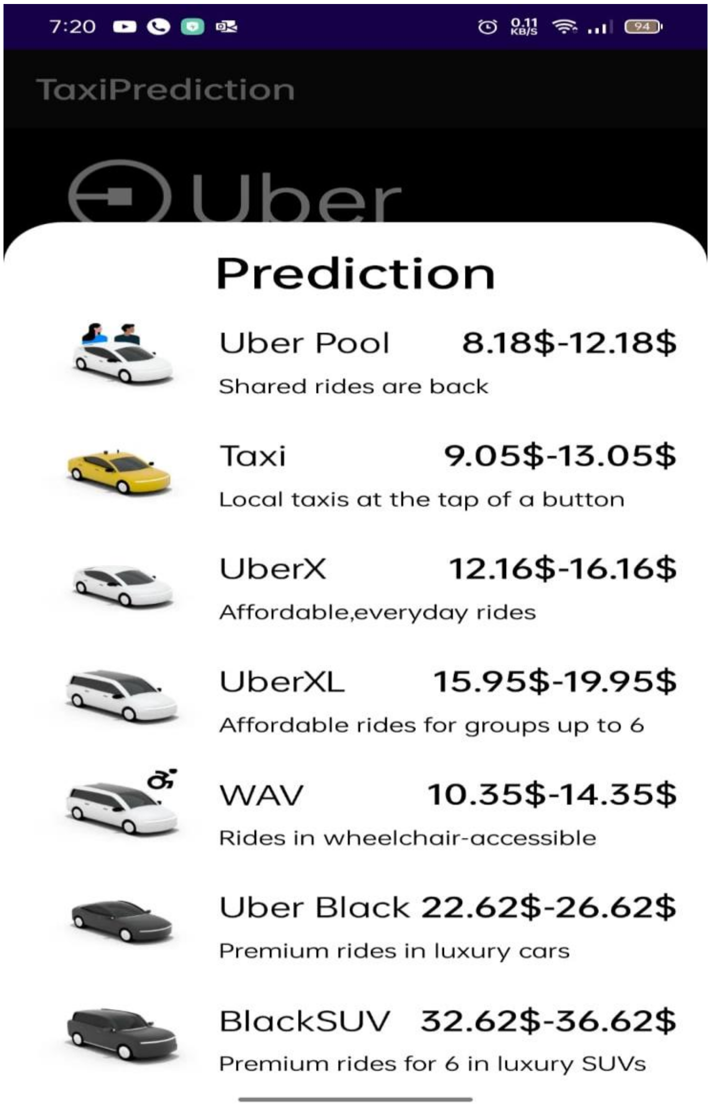
4: Google Places API: we use it to get any street name in all the world.

5: Google Map API: we use it to get maps from google

6: Google Distance API: to calculate distance between two points in map

Comparing Between our Application and the real application to predict the price between same distance

Our uber app

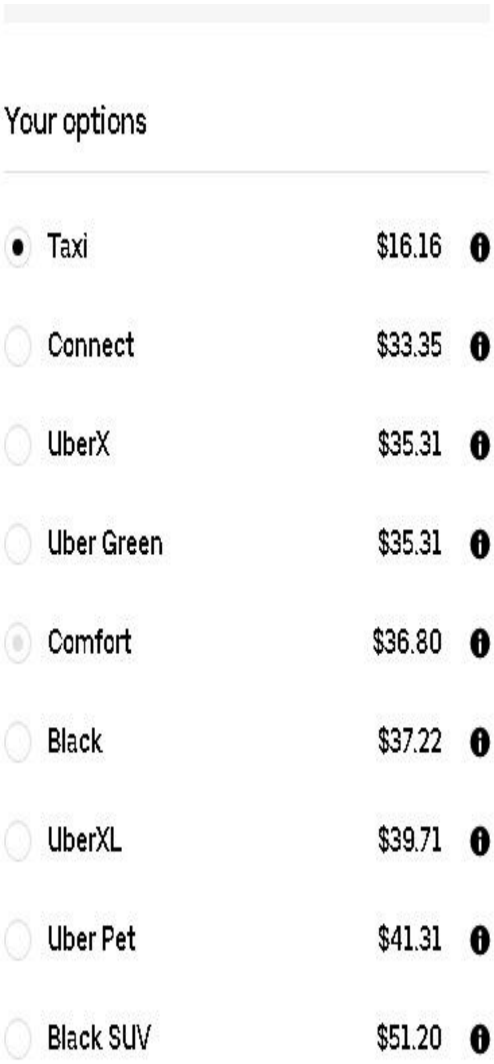


TaxiPrediction

Prediction

Uber Pool Shared rides are back	8.18-12.18
Taxi Local taxis at the tap of a button	9.05-13.05
UberX Affordable,everyday rides	12.16-16.16
UberXL Affordable rides for groups up to 6	15.95-19.95
WAV Rides in wheelchair-accessible	10.35-14.35
Uber Black Premium rides in luxury cars	22.62-26.62
BlackSUV Premium rides for 6 in luxury SUVs	32.62-36.62

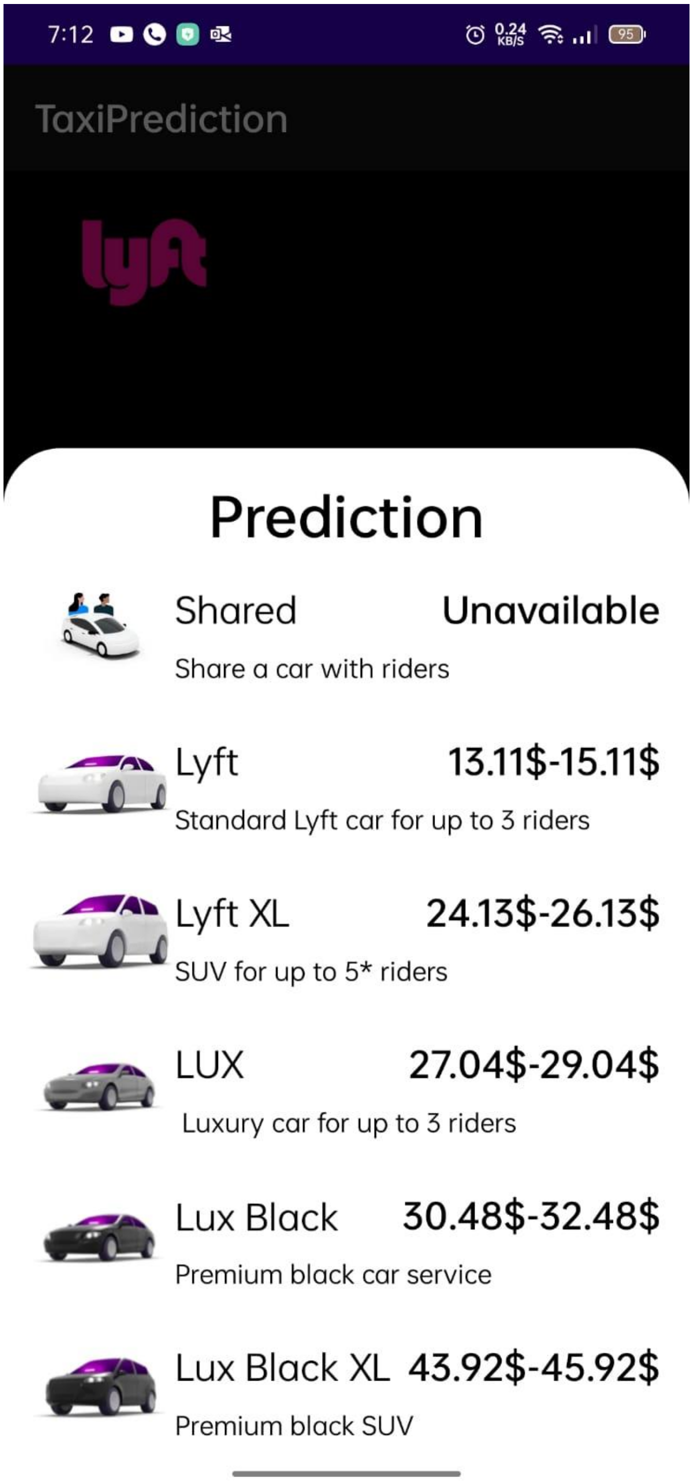
real uber app



Your options

●	Taxi	$16.16
○	Connect	$33.35
○	UberX	$35.31
○	Uber Green	$35.31
○	Comfort	$36.80
○	Black	$37.22
○	UberXL	$39.71
○	Uber Pet	$41.31
○	Black SUV	$51.20

Our Lyft app



7:12 0.24 KB/S 95

TaxiPrediction

Prediction

Shared Unavailable
Share a car with riders

Lyft 13.11-15.11
Standard Lyft car for up to 3 riders

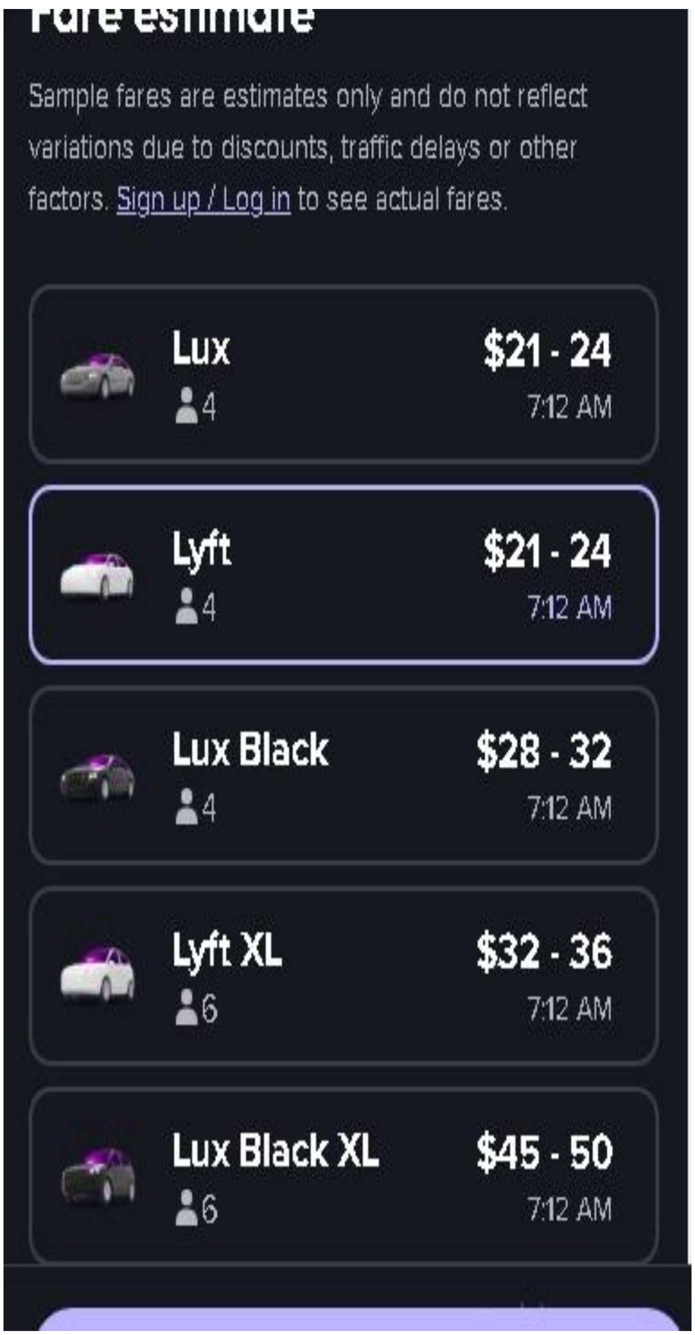
Lyft XL 24.13-26.13
SUV for up to 5* riders

LUX 27.04-29.04
Luxury car for up to 3 riders

Lux Black 30.48-32.48
Premium black car service

Lux Black XL 43.92-45.92
Premium black SUV

Real app



Fare estimate

Sample fares are estimates only and do not reflect variations due to discounts, traffic delays or other factors. Sign up / Log in to see actual fares.

Lux $21 - 24
4 7:12 AM

Lyft $21 - 24
4 7:12 AM

Lux Black $28 - 32
4 7:12 AM

Lyft XL $32 - 36
6 7:12 AM

Lux Black XL $45 - 50
6 7:12 AM