ITCS 6100 - Project Deliverable 3

New York City Taxi & Limousine Commission (TLC) Trip Record Data

Team Members

1. Harsh Raval (801257980)
2. Pranjali Mehta (801255574)
3. Shruti Nagpure (801256223)
4. Urvashi Murari (801205124)
5. Zubin Ladwa (801256671)

Q1. What was unique about the data? Did you have to deal with imbalance? What data cleaning did you do? Outlier treatment? Imputation?

Answer : This dataset contains 55 million entries. The unique things about data is it has no missing values, no obvious data entry error,1 to 6 passengers, Latitudes lie between 40 and 42, Longitudes lie between -75 and -72.

Imbalance:

- The data set that we used had few imbalances.
- It had many null and negative values.
- Missing values were dropped as we had enough data to build models on.
- We performed Random Sampling in Data Selection while splitting the data into training and testing sets to avoid and treat the presence of biases.
- Various algorithms and strategies were implemented to train and optimize model performance.

We found below outliers:

1.) Some trip durations are over 100000 seconds which were clear outliers and should be removed. We used **data[data.trip_duration > 86400]** check to remove those. We removed the records with passenger count > 7, 8 or 9 as they are extreme values and look very odd to be occupied in a taxi.

2.) We used **data = data[data.passenger_count <= 6]** check to eliminate passenger counts more than 6**.**

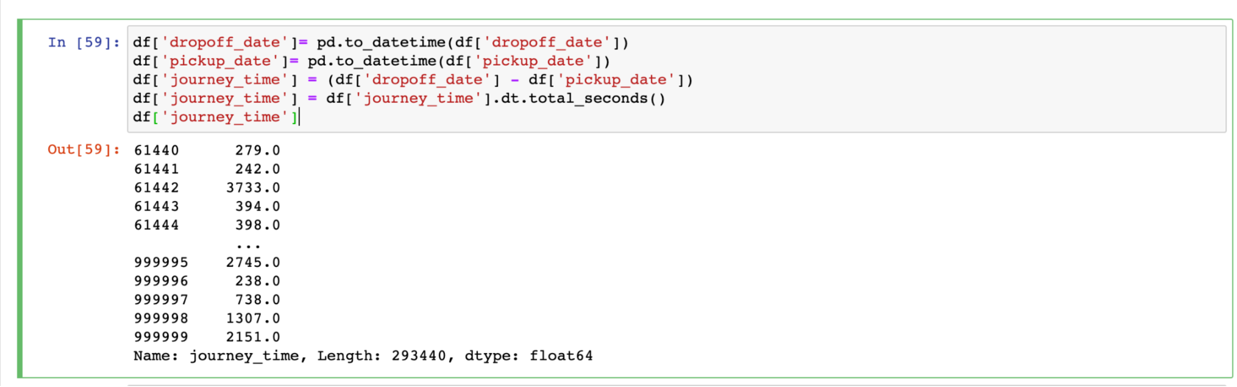
3.) Trips over 30 km/h are being considered as outliers but we cannot ignore them because they are well under the highest speed limit of 104 km/h on state controlled highways.There were few trips which covered huge distance of approx 200 kms within very less time frame, which is unlikely and should be treated as outliers.

4.) We removed those trips which covered 0 km distance but clocked more than 1 minute to make our data more consistent for predictive models. Because if the trip was canceled after booking, then that should not have taken more than a minute. This is our assumption.

Q2. Did you create any new additional features / variables?

Answer. Yes we did.

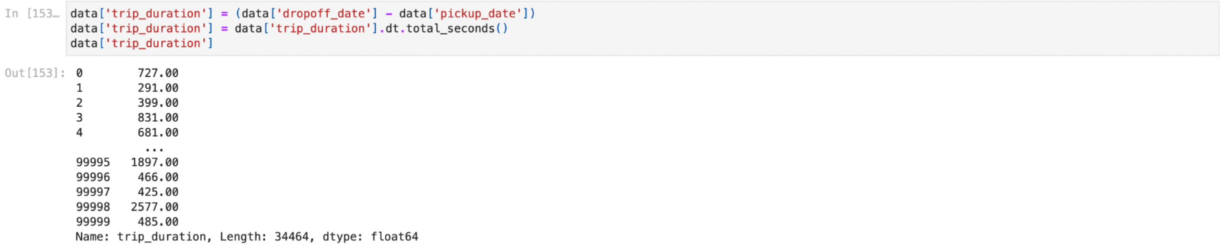
a. **Journey_time** : The two features "pickup datetime" and "dropoff datetime" in the dataset, which show the beginning and conclusion of the ride, respectively, are visible. As part of feature engineering, we will develop a feature that will use these features to compute ride duration because we are aware that the duration of the drive has a significant impact on the cost of a taxi journey.



```
In [59]: df['dropoff_date']= pd.to_datetime(df['dropoff_date'])
         df['pickup_date']= pd.to_datetime(df['pickup_date'])
         df['journey_time'] = (df['dropoff_date'] - df['pickup_date'])
         df['journey_time'] = df['journey_time'].dt.total_seconds()
         df['journey_time']

Out[59]: 61440      279.0
         61441      242.0
         61442     3733.0
         61443      394.0
         61444      398.0
                    ...
         999995    2745.0
         999996     238.0
         999997     738.0
         999998    1307.0
         999999    2151.0
         Name: journey_time, Length: 293440, dtype: float64
```

b. trip_duration : Difference of pickup_time and dropoff_time



```
In [153… data['trip_duration'] = (data['dropoff_date'] - data['pickup_date'])
         data['trip_duration'] = data['trip_duration'].dt.total_seconds()
         data['trip_duration']

Out[153]: 0        727.00
          1        291.00
          2        399.00
          3        831.00
          4        681.00
                    ...
          99995   1897.00
          99996    466.00
          99997    425.00
          99998   2577.00
          99999    485.00
          Name: trip_duration, Length: 34464, dtype: float64
```

c. Calculated and assigned new columns to the data frame such as weekday, month and pickup_hour which helped us gain more insights from the data.

Q 3. What was the process you used for evaluation? What was the best result?

Answer: Our dependent variable contains continuous values so we used a regression technique to predict our output. We also did not perform any scaling of the features because the linear regression model takes care of that inherently. This is a plus point to use the Linear regression model. It is quite fast to train even on very large datasets.

We used the following algorithms to train and test our models:

1. **Linear Regression :** Linear regression model is **extremely fast** to train on the high dimensional datasets consisting of even **millions** of records.
2. **Random Forest :** A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
3. **XGBoost :** It is an optimized distributed gradient boosting library. It uses a gradient boosting (GBM) framework at core.

After training our data on the 3 models we came to the following conclusions:

●XGBoost proved to be much more efficient in predicting the output. But it takes much more time to train it over the large dataset with more complexity as compared to the RF and Linear regression model but less time then the SVR.

●It didn't helped us much to generalize the model by tuning hyper parameters for the RF model as there is not much difference in the RMSE scores of the default model and the tuned model of the feature selection group in fact both vary on every iteration and sometimes the tuned model gives poor results than the default model.

●Contrast to the Random Forest regressor, XGBoost regressor prediction results were consistent on every iteration.

●Feature extraction didn't help in any way to improve the RMSE score with any of the regression models.

Q4. Is there Bias in your work? What were the problems you faced? How did you solve them?

Answer: The dataset that we used was year 2016 data from NYC-TLC dataset. The data set that we used had few biases. Also it had many null and negative values for e.g.: The data that had Null value for fare amount hence it was dropped and trip distances that were negative were also dropped. Biases were removed by dropping the null values and having all values as float data. For training the model we used a linear regression model, there were no assumptions that were made while selecting the model. In addition, we added an extra column called “Journey Time. and “trip_durtion.”

Q5. What future work would you like to do?

Answer:

- It would be wonderful to supply a dataset to a GPU instance to make the training task faster.
- We can also trigger Sagemaker Training jobs on larger processing instances.
- We can work on better Outliers detection in the future.
- We can utilize Clusters in AWS and make a cluster for "normal" data and another for “non-normal” data.