Uber Fares Prediction Documentation

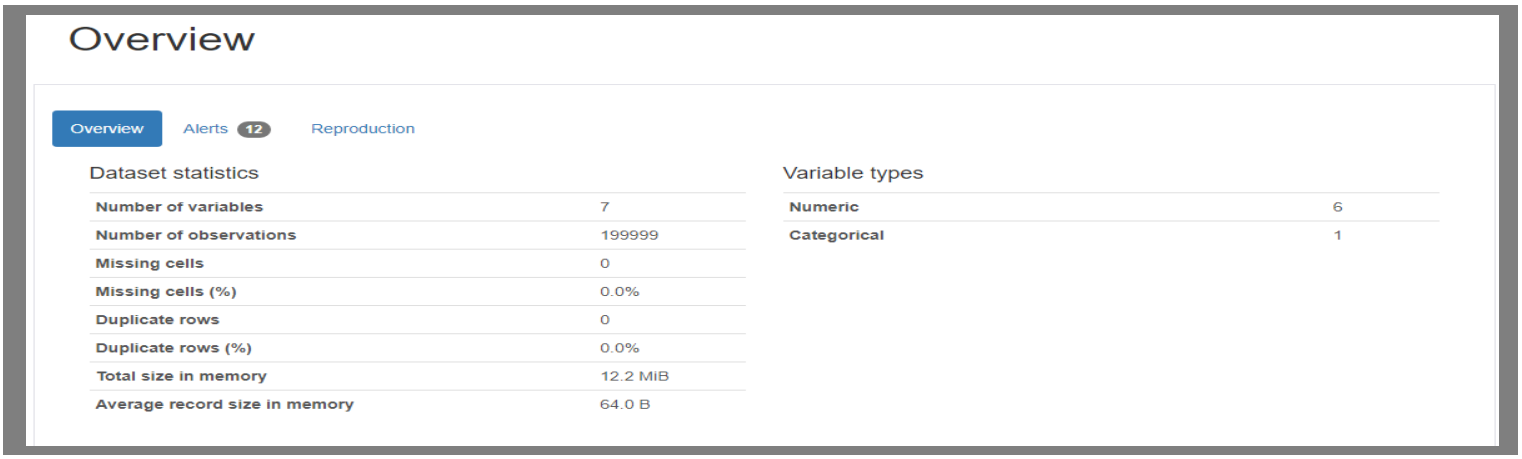
1. **Basic Info of Data:-**

   After importing the data using necessary libraries few things I always check on are :

   - What is the size of the data?
   - How does the data look like ?
   - What is the data type of the column?
   - Are there any missing values?
   - How does the data look mathematically?
   - Are there any duplicate values in the data

2. **Creating a report using Pandas Profiling:**

   - Next most important thing that helps me in deep understanding the data is profiling report
   - Visualization becomes do easy to understand with different types of plot and there are other features shown below



3. Pre-processing of Data:-

- There were few unnecessary columns such as “key”,”Unnamed :0” which I dropped using pandas drop function
- In the data there was a column named as “pickup_datetime ” which was in object data type which needed to formatted accordingly
- “pickup_datetime” was then formatted to “datetime ” formate using pandas function “pd.to_datetime()”
- Then there were lots of information hidden in the formatted column which were extracted such as year , month, day, day_is_weekend , quarter ,hour , minutes , seconds
- After then distance column was formatted using latitutdes and longitudes Using “haversine formula” which basically works considering that earth’s shape is spherical rather than the usual python library geopy which considers earth as spheroid

4. Exploratory Data Analysis :-

- After then few graphs were plotted for univariate , bi-variate and multi-variate Data
- Usually I plot 3 kinds of graphs i.e Histogram (for distribution of data), Distplot(for finding out KDE), and Boxplot (for finding the outliers)
- Graphs for Fare_amount, Passenger_count, Distance_travelled were plotted accordingly

5. Outliers Detection :-

- After exploring the data it was time for removing outliers which I have done IQR method because data wasn't normally distributed
- Further I have used capping to make the outliers same as the upper limit or whiskers rather than trimming them as according to me if we can save data then why not ?
- By plotting heatmap I have confirmed that the relation between distance and fares is much better than the other features
- And at the end I save the processed data using pandas function to_csv

```
plt.figure(figsize=(20,10))
sns.heatmap(df_wo.corr(),annot=True)
```

```
<AxesSubplot:>
```

6. Model Building using Default Values :-

- After the processed data I have tried to build the model with keep them the default setting so that I could get the idea exactly how my data is performing
- I have tried to build 10 Model they are:
    1. Linear Regression
    2. Elastic Regression
    3. SGD Regression
    4. Bayesian Ridge Regression
    5. XGBoost Regression
    6. LGBM Regression
    7. CatBoost Regression
    8. Decision Tree Regression
    9. Random Forest Regression
    10. GB Regression

7. Comparing R2Scores of very model:-

- After building all the models I have compared there scores for better understanding in which I have found most of the models have preformed pretty well where as model Bayesian Ridge Regression hasn't that good

8.Hyper Parameter Tuning of the models:-

- After creating the necessary models it was time for tune is according to my requirements
- Accordingly every models was tuned with it's Hyper parameters such as Tree based models have been tuned on Max_depth as one the important parameter other than this bootstrap ,min_samples_split , Max_samples and max_features was also taken into consideration
- Then final model i.e Decision Tree Regressor was decided as it evaluated to 80% of providing correct prediction

9.Model Deployment using StreamLit:-

- Once everything was finalized then was the main task of deploying my model using StreamLit and creating WebAPP
- For I have coded my Web page using basics of HTML by importing few of the python libraries as joblib and PyYmal
- And then I was done by creating my ***FIRST MACHINE LEARNING PROJECT***

DEMONSTRATION OF WEBAPP:-

Please provide your inputs

Enter No. of Passengers Travelling

2.00 − +

Enter your year of Journey

2015.00 − +

Enter the month of Journey

5.00 − +

Enter the day of Journey

21.00 − +

Enter your hour of journey

15.00 − +

Enter minutes roughly

0.00 − +

Enter your distance of jounery(in kilometers)

9.98 − +

Does your journey is on weekend?(0=NO,1=Yes)

0.00 − +

Quater

1

Based on your selection of Passenger counts: 2.0, Year: 2015.0, Month: 5.0 , Day: 21.0,Hour : 15.0,Minutes :0.0 , Distance Travelled : 9.98,Quater :1 and Weekend : 0.0

Fares in $ for your ride would be

10.18