Gojek Assignment

APPROACH:

- Calculate Online Hours for drivers for each date by summarising Ping data
    - 1st grouping with respect to driver_id, date and hour of that date.
    - Calculating number of pings in a specific hour of a day
    - Multiplying number of pings with 15 since each ping is at an interval of 15 sec
    - dividing the sum of all seconds by 3600 to convert to hours
    - Rounding of the hours to calculate online_hours.
- Feature Engineering
    - Calculate Features for summarised Online_hours like
        - Mean
        - Median
        - 1st quantile
        - 2nd quantile
        - Max
        - Min
        - Range
        - Gender wise mean hours
    - Generate features for Date
        - Day of year
        - Day of Week
        - Day of Month
- Combining Generated features with driver details data
- Checking for driver ids in both training and test data
- Checking for missing data
- Checking for duplicate records
- Write a function to generate features for both train and test and then split back to train , test.
- Training baseline linear regression model which gives predicts the mean online hours
    - MSE-7.56
- To improve the predictions train Random Forest and Gradient Boosted trees.
- Grid search in order to find the optimal parameter for the models
    - GBM(grid search with 3 fold cross validation accuracy)
        - no of trees
        - learning rate
        - Max depth of trees
    - MSE-7.16
    - For Random forest(grid search with 3 fold cross validation accuracy)
        - Max depth of trees
        - Number of trees
    - MSE-7.2
- Training RF and GBM with best parameters on training data and predicting using both the model

- Taking mean of both the prediction
- Imputing the prediction with 0 for drivers present in test but not in train
    - MSE-7.1

CHALLENGE:

- There are lot of duplicate values present in both train and test data
- Online hours distribution is different in train and test
- More data required for training to capture the variation in weeks