Machine Learning Project
Team - PB0208

Lovejeet Singh Parihar (IMT2019048)
Manuj Malik (IMT2019052)

15 December 2021

1 Project Definition

The primary aim of this machine learning project is to predict the number of visitors for a certain restaurant on a given date.

We are provided with the historical visit and reservation data for the restaurants, along with the store info which includes the location and genre for the restaurant.

2 Exploratory Data Analysis (EDA)

2.1 Overview of Data

- Chwiggy Reservation Data :
  1. chw_store_id : air format, restaurant's id
  2. visit_date_time : time when people arrives, time of reservation
  3. reserve_date_time : time when reservation was made
  4. reserve_visitors : number of people for that reservation
- Yomato Reservation data :
  1. yom_store_id : hpg format, restaurant's id
  2. visit_date_time : time when people arrives, time of reservation
  3. reserve_date_time : time when reservation was made
  4. reserve_visitors : number of people for that reservation
- Calendar info :
  1. Date
  2. Day of Week
  3. Holiday Flag
- Relation between store id's :
  1. chw_store_id
  2. yom_store_id
- Number of restaurants :
  1. Chwiggy : 829
  2. Yomato : 4690

2.2 Conclusions from EDA

1. Left skewness was observed in the restaurant data, where the graph was plotted for average visitors per day.

2. Outliers were detected for visitors data of chwiggy.

3. Outliers were seen for both chwiggy and yomato reserve average visitors and were removed for training purposes.

4. The average visitors per day is high on Saturday, Sunday and Friday and low on Monday and Tuesday.

5. People preferred to visit restaurants, there seems an effect of holidays on the number of people visiting

6. Average number of people visiting the restaurants is about the same for both the years.

7. The amount of visitors visiting is low in 2017 is either due to less data for the year or more collection of data from 2016.

8. No specific trend was seen in the graph of date wise analysis of average visitors per day, and mean appears to remain constant.

9. Sudden spike of visitors visiting the restaurant is noticed around the july of 2016, and sudden decrease around the january of 2017.

10. After plotting the graph of Date wise analysis of total reserve visitors in chwiggy restaurants, it was observed that more restaurants for chwiggy in the month of november 2016, or more collection of data.

11. No apparent pattern was seen in the graph of Date wise analysis of total reserve visitors in yomato restaurants.

12. In the month wise analysis of average visitors, pattern was observed to be same for the first four months of 2016 and 2017.

13. The total visitors abruptly increase in the seventh month of 2016 probably due to more restaurant data for these months.

14. For the bar plot graph of holiday/non-holiday, it was observed that average number of visitors are more on holidays than non-holidays.

15. Asian cuisine was preferred by average number of visitors visiting the restaurants.

16. Area wise analysis of visitors didn't yield great results.

17. Most reservations are made for the dinner time, yomato restaurants.

18. Similar trend was observed for the chwiggy restaurants as well.

2.3 Graphs and Plots

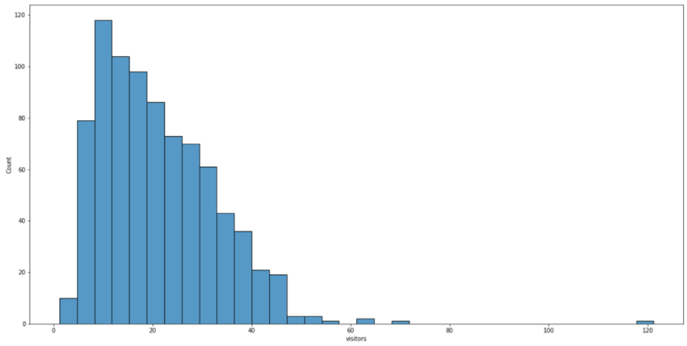

Figure 1: Restaurant Wise analysis of data

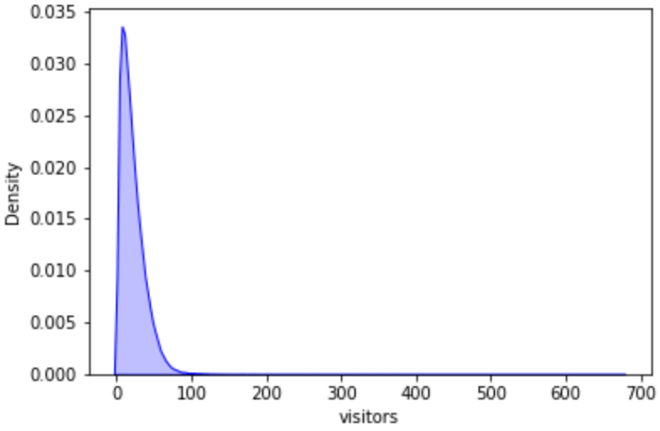

Figure 2: Density Plot of Visitors

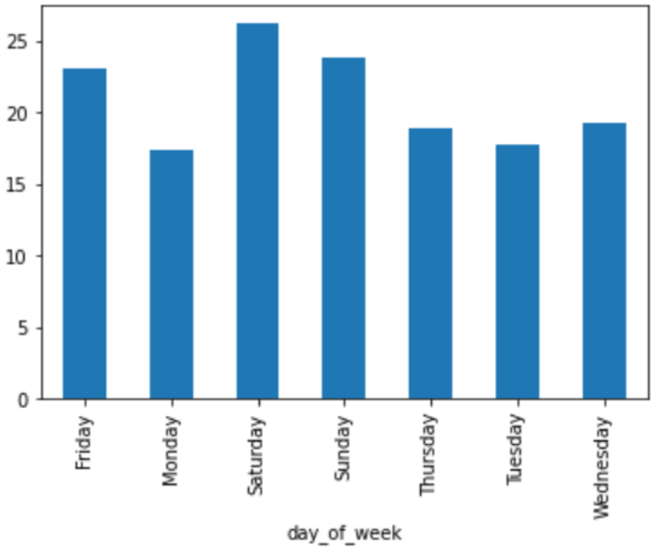

Figure 3: Bar plot for day wise analysis of average visitors per day

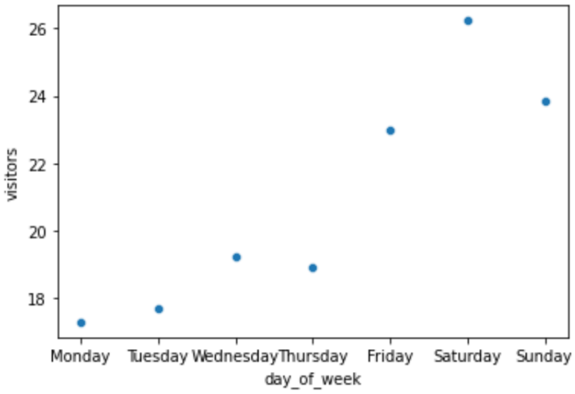

Figure 4: Average number of visitors v/s weekdays plot

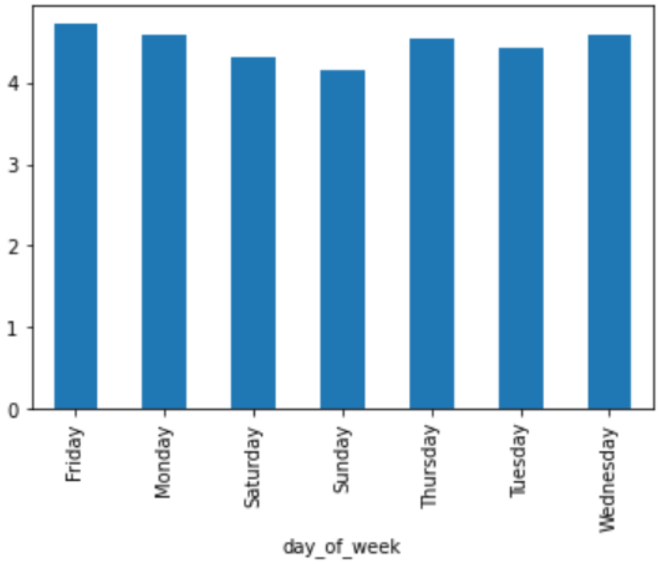

Figure 5: Day wise analysis of average reservations per day in chwiggy restaurant

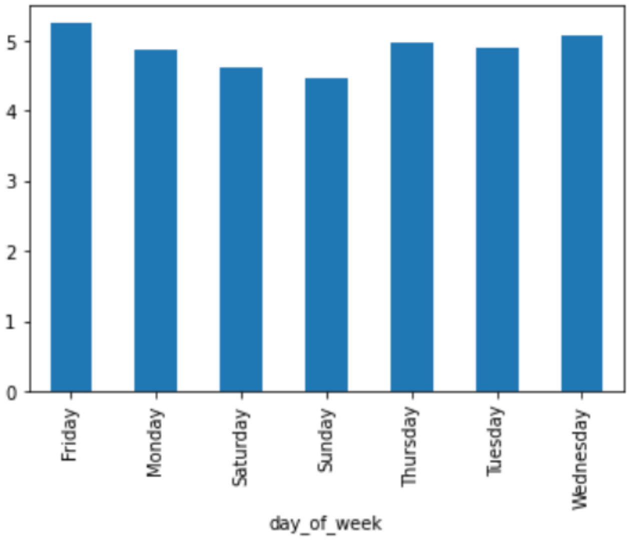

Figure 6: Day wise analysis of average reservations per day in yomato restaurant

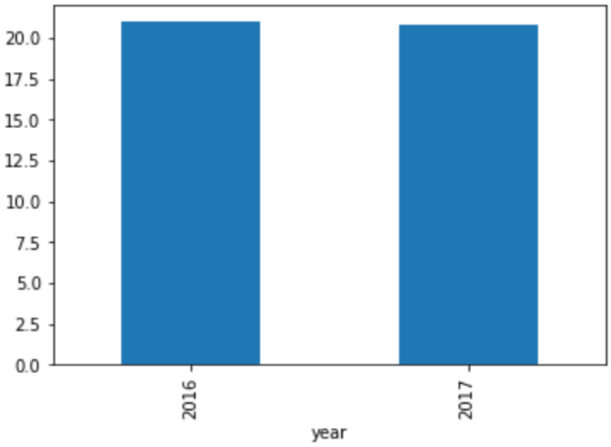

Figure 7: Year wise analysis of average visitors per day

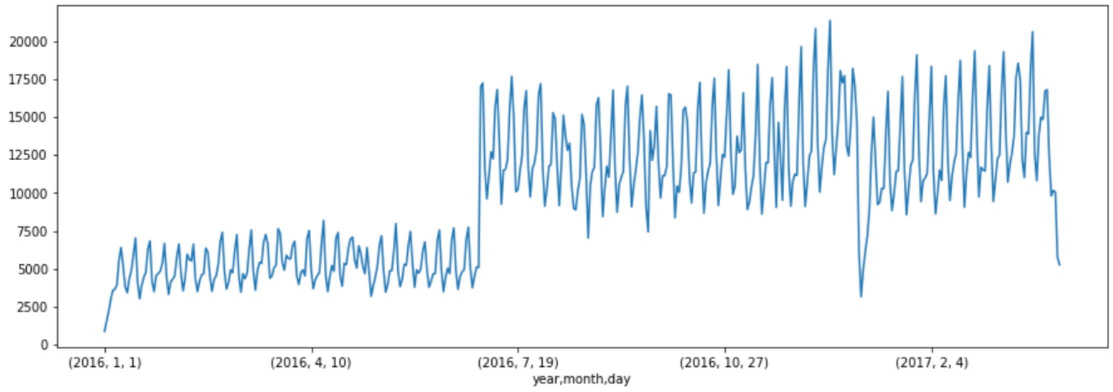

Figure 8: Date wise analysis of total visitors

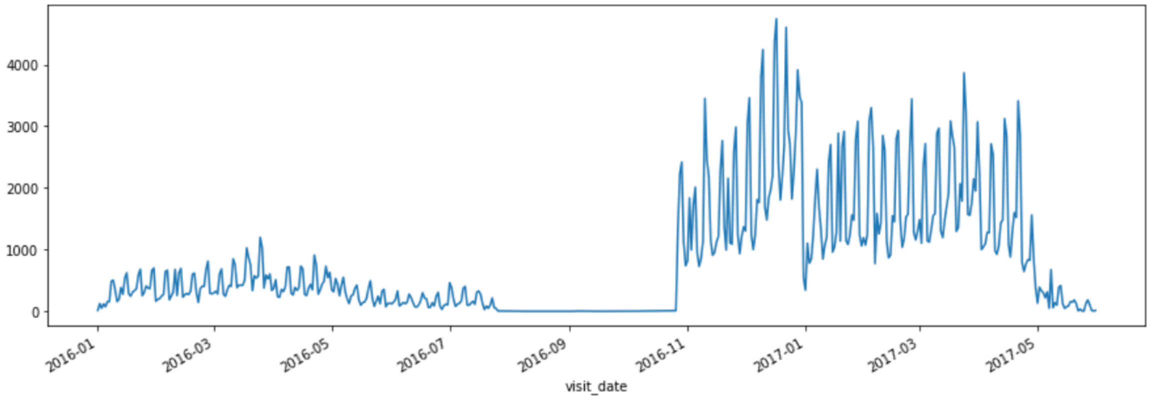

Figure 9: Date wise analysis of total reserve visitors in chwiggy restaurants

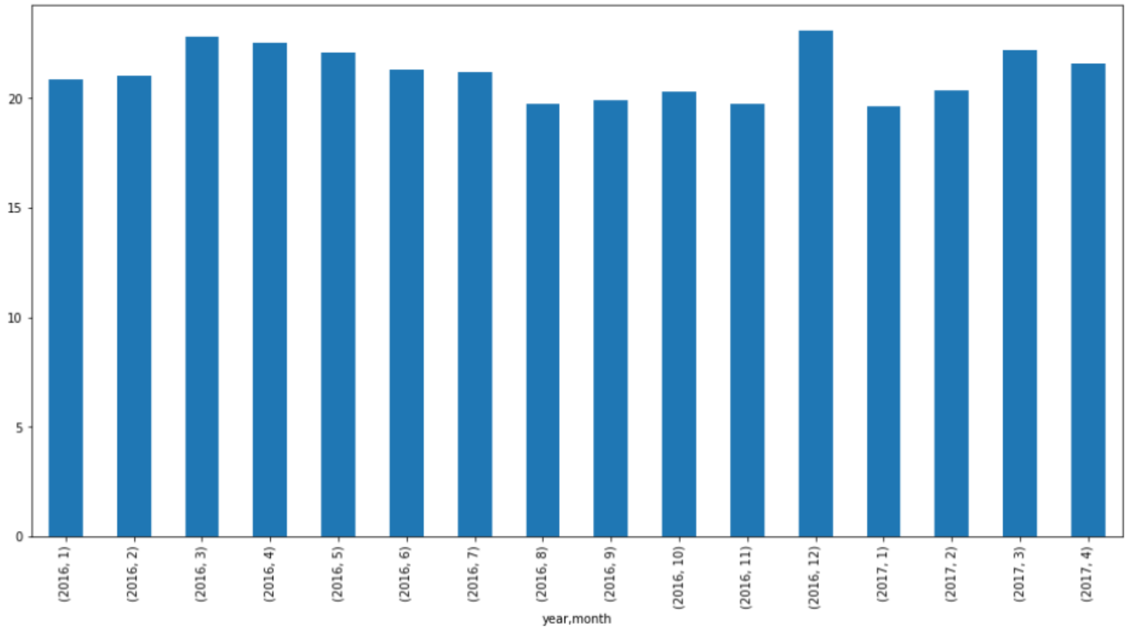

Figure 10: Month wise Analysis of average visitors per day

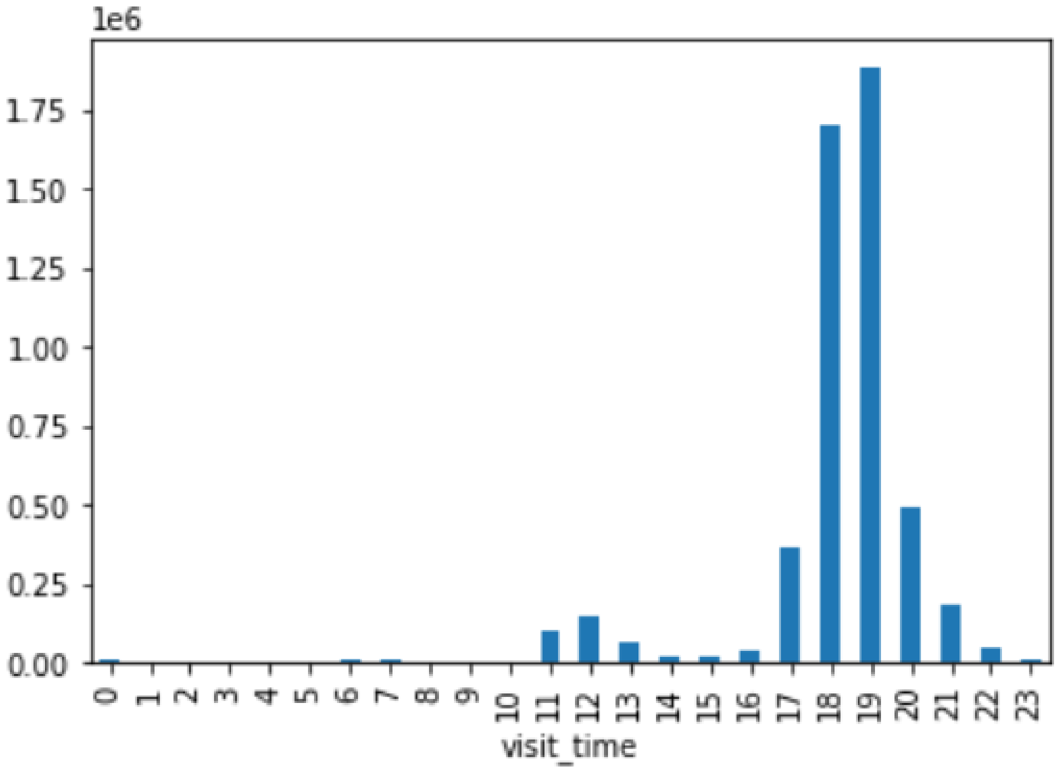

Figure 11: Time wise analysis of total reserve visitors in yomato restaurant

3 Feature Selection and Engineering

3.1 Preprocessing

1. No null values were present in any of the relations. The nan values generated after merging were filled with zeros.
2. Date format of the file date_info was different from other files, so it was modified and made consistent with others.
3. Date information was merged with the train data.
4. There were some outliers for number visitors in the train dataset. So we removed the outliers and the range of +/- 2.5 × (Inter Quartile Range) worked best.
5. Separate columns for month, day and year were made.

3.2 Generating New features

Feature Set 1 - Features generated with only the visitors data from train :-

1. Got the mean number of visitors for each restaurant and added it as a feature. This adds a weighting factor which gives more weight to the restaurant who have a higher mean number of visitors. Similarly we add min and max number of visitors for each restaurant.
2. Got the mean number of visitors for each date and added it as a feature. It weights the dates according to the mean number of visitors on that day.
3. Got the mean number of visitors for each genre and added it as a feature. It weights the genres according to the mean number of visitors on that day.
4. Got the mean number of visitors for each day of week and added it as a feature. It weights the days of week according to the mean number of visitors on that day.
5. Got the number of restaurants for each area and added it as a feature. It gives information about the area in the training example.
6. Got the mean, maximum and minimum number of visitors for each month of year and added it as a feature. It weights the months of year according to the mean number of visitors on that day.
7. Two more binary features were added, one telling if it is a weekend or not and the other one telling if it is a month end (day ≥ 26) or not.
8. Got the mean, max and min number of visitors for each restaurant for a given month and added it as a feature.
9. Got the mean, max and min number of visitors for each restaurant for a given year and added it as a feature.
10. Got the mean number of visitors for each restaurant for a given day of week and added it as a feature.
11. Got the mean number of visitors for each restaurant for a holiday/non-Holiday and added it as a feature.

All the above features had a significant impact on the score. Infact most of the improvement in the score were due to adding of these features at different times.

Feature Set 2 - Features generated using the reserve data :-

1. Total number of reserve visitors were calculated for whatever restaurants and dates the information was available. The store relation table was used to link some chwiggy restaurants with yomato restaurants and hence get their information from the yomato reserve.
2. Got the mean, min, max number of reserve visitors for each restaurant and added it as a feature to the train dataset.

3. Got the mean number of reserve visitors for each date and added it as a feature to the train dataset
4. Got the mean number of reserve visitors for each genre and added it as a feature to the train dataset.
5. Got the mean number of reserve visitors for each day of week and added it as a feature to the train dataset.
6. Got the mean, min, max number of reserve visitors for each month of year and add it as a feature to the train dataset.
7. Got the mean number of reserve visitors for holiday/working day and add it as a feature to the train dataset.

The above set of features (from the reserve sets) had some impact on the score but the impact was not as large as Feature Set 1.

3.3 Adding Weather Data

In an attempt to use the longitude and latitude features, we tried adding the weather data for certain locations on certain dates.

1. Got the historical weather data using meteostat.
2. Found the nearest station for each restaurant, using longitude and latitude, and collected the historical weather data in the required date range.
3. Adding average temperature, minimum temperature, maximum temperature and precipitation value as features to the training set.

The weather features didn't work out much and didn't improve the score.

4 Models Used

4.1 Linear Regression

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. A linear regression line has an equation of the form Y = a + bX

4.2 Decision Tree Regressor

A type of supervised learning algorithm. The default values for the parameters controlling the size of the trees (e.g. max_depth, min_samples_leaf, etc.) lead to fully grown and unpruned trees which can potentially be very large on some data sets. To reduce memory consumption, the complexity and size of the trees should be controlled by setting those parameter values.

4.3 eXtreme Gradient Boost (XGBoost)

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible and portable.

4.4 Light Gradient Boost Machine (LightGBM)

LightGBM is a gradient boosting framework that uses tree based learning algorithms. With LightGBM you can run different types of Gradient Boosting methods. You have: GBDT, DART, and GOSS which can be specified with the boosting parameter. LightGBM threading also scales by the number of features.

4.5 Kernel Ridge Regressor

Penalizes the size (square of the magnitude) of the regression coefficients. Kernel ridge regression combines Ridge regression and classification (linear least squares with l2-norm regularization) with the kernel trick. In Kernel Ridge Regression (krr), also called Kernel Regularized Least Squares, the basis functions Φ are generated from a kernel function, which takes two vectors from the input space as input. Kernel functions are such that their output is maximal when $x = x'$ and decreases as the distance increases, which provides a locality property. KNN is a memory intensive algorithm and it is already classified as instance-based or memory-based algorithm.

4.6 Lasso Regression

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage.

4.7 Elastic Net

The elastic net algorithm uses a weighted combination of L1 and L2 regularization. Regularized regression model using the penalties of both lasso and ridge.

5 Table of Models and their scores

Sr no.	**Model Used**	**Best Score**
1	Linear Regression	0.494432 (Locally)
2	Decision Tree Regressor	Public = 0.55139 Private = 0.55495 (Incomplete features)
3	XGBoost	Public = 0.52097 Private = 0.51911
4	LightGBM	Public = 0.49463 Private = 0.49022
5	Lasso Regression	Public = 0.53149 Private = 0.52889
6	Elastic Net	Public = 0.53149 Private = 0.52889
7	Stacking of lightGBM, GBoost, Decision Tree Model	Public = 0.51081 Private = 0.51015
8	Stacking of lasso Regressor, XGBoost, Decision Tree Model	Public = 0.49605 Private = 0.49276
9	Stacking of lightGBM, ENet, Decision Tree Model	Public = 0.51630 Private = 0.51630
10	Stacking of lasso, lightGBM, ENet, XGBoost, Decision Tree Model	Public = 0.51479 Private = 0.51483

6 Individual Contributions

6.1 Lovejeet Singh Parihar (IMT2019048)

- Performed Exploratory Data Analysis.
- Did the merging and preprocessing steps mentioned above.
- Did the feature engineering and added new features.
- Added historical weather data.
- Trained linear regression, XGBoost, Decision Tree and LightGBM models and tuned the parameters by a manually written code that go over different parameters randomly.

6.2 Manuj Malik (IMT2019052)

- Performed Exploratory data analysis.
- Some preprocessing was done after the first evalutation. For example, NaN values in the whole dataset is filled with 0s, tried to fill the NaN values with mode, but it didnt work
- Did the code for Linear, Elastic Net, Kernel Ridge, Decision tree, KNeighborsRegressor, SGDRegressor, XGBoost. Grid and random search CV was tried and tested for fine tuning of hyperparamters.
- Trained and modeled different variations of models to be stacked and chaining of models was done.

7 Conclusions

- After adding the features from feature-set-1 provided in this report, notable score impact was seen. Mean number of visitors v/s id was the main reason behind the boosting of the score.
- Light Gradient Boost Machine used in the code impacted the score significantly.

8 References

https://towardsdatascience.com/ridge-lasso-and-elasticnet-regression-b1f9c00ea3a3
https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
https://neptune.ai/blog/lightgbm-parameters-guide
https://sites.google.com/view/lauraepp/parameters