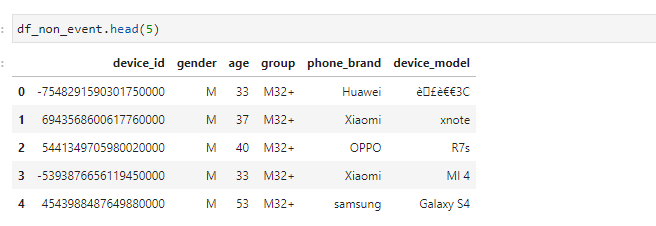
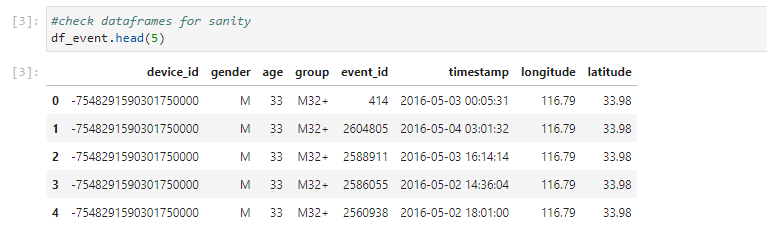
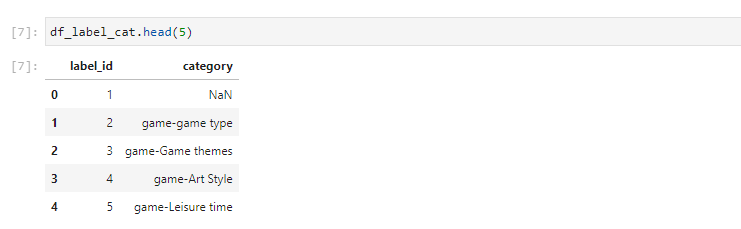
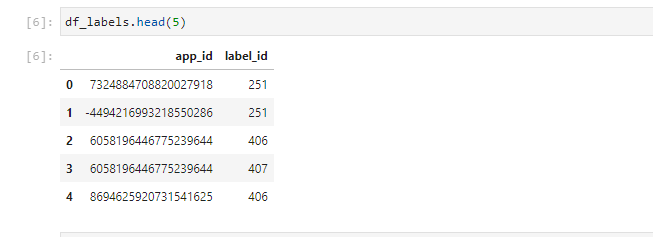
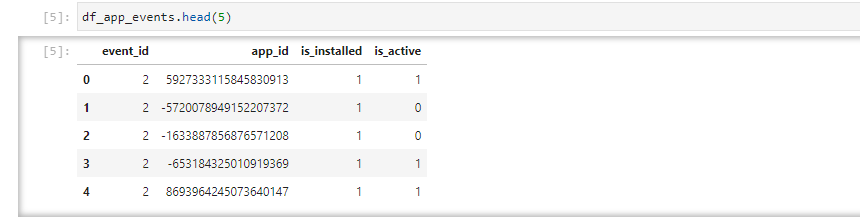
Upgrad Capstone-2022

Contribution:

Akshay : 50%

Chandrabhan Singh : 50%

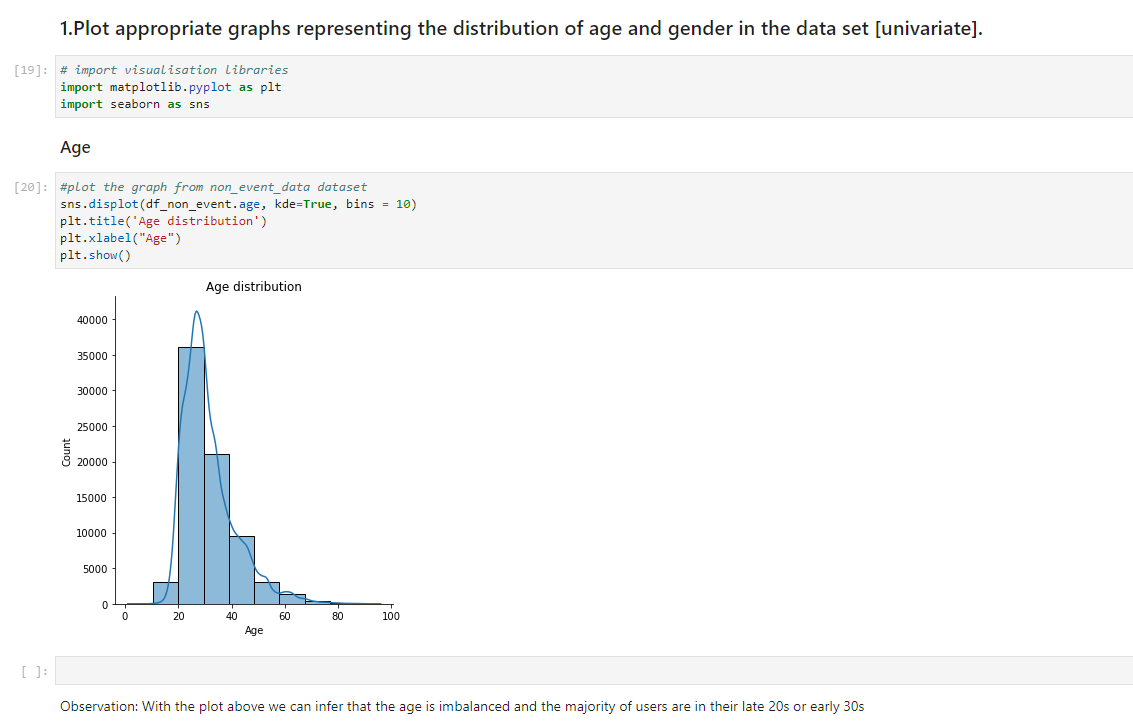
# **1) Top five rows of the data set at the beginning of the analysis**

1. Non - Event dataset
2. Event dataset   
   
3. Label Categories  
     
   
4. App Labels
5. App events

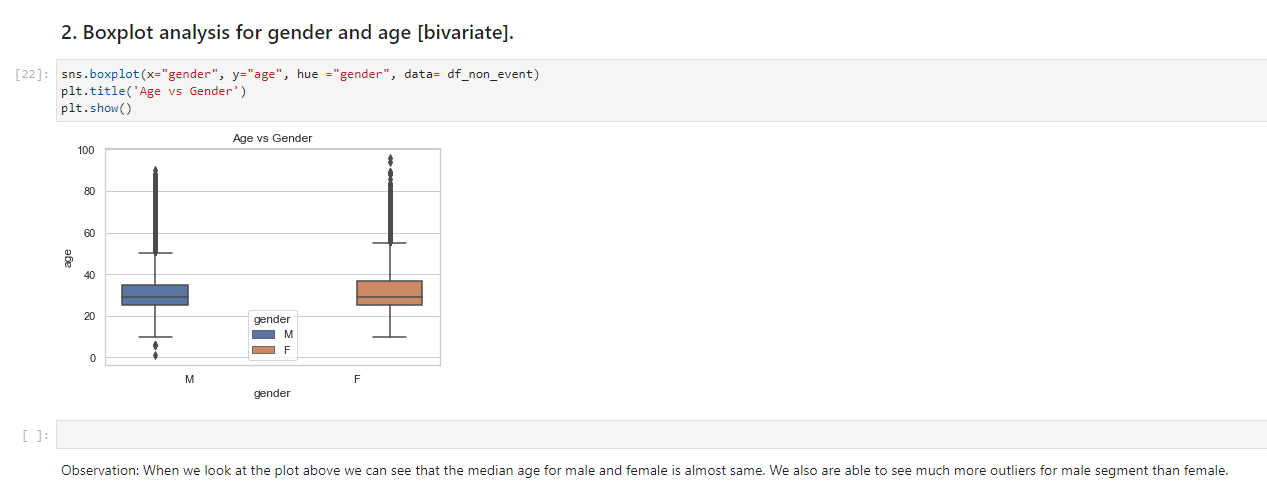
# **2) List of data cleaning techniques applied such as missing value treatment, etc.**

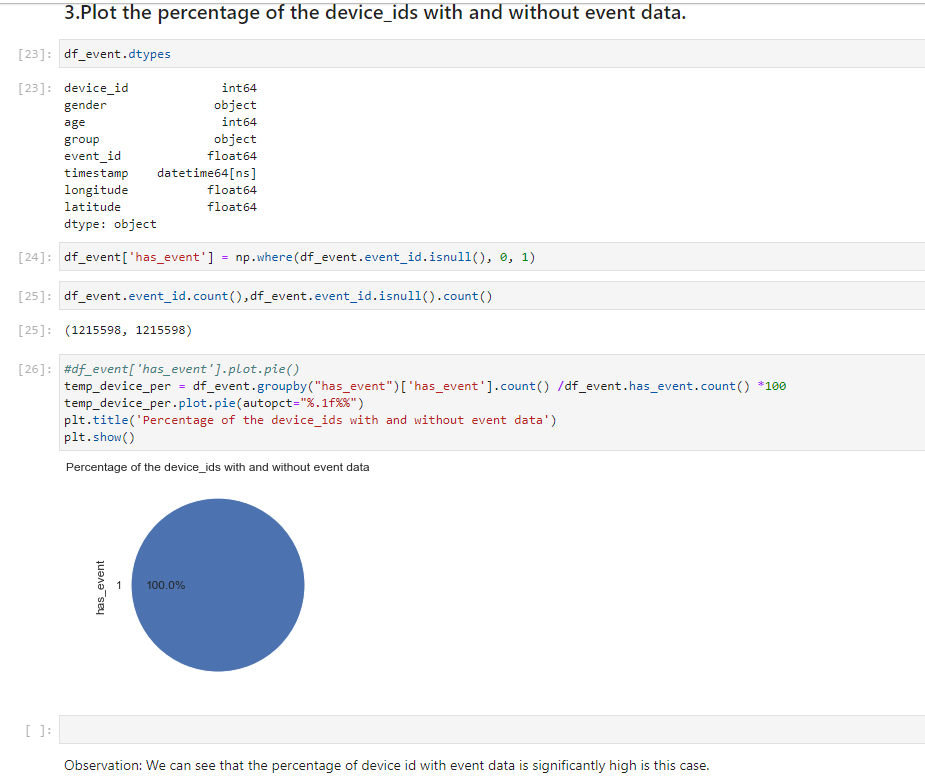
* Replacing the missing values with None
* Correction of data types like integer and datetime
* From event dataset removing data with no events available
* For non-event dataset removal of duplicate device ids

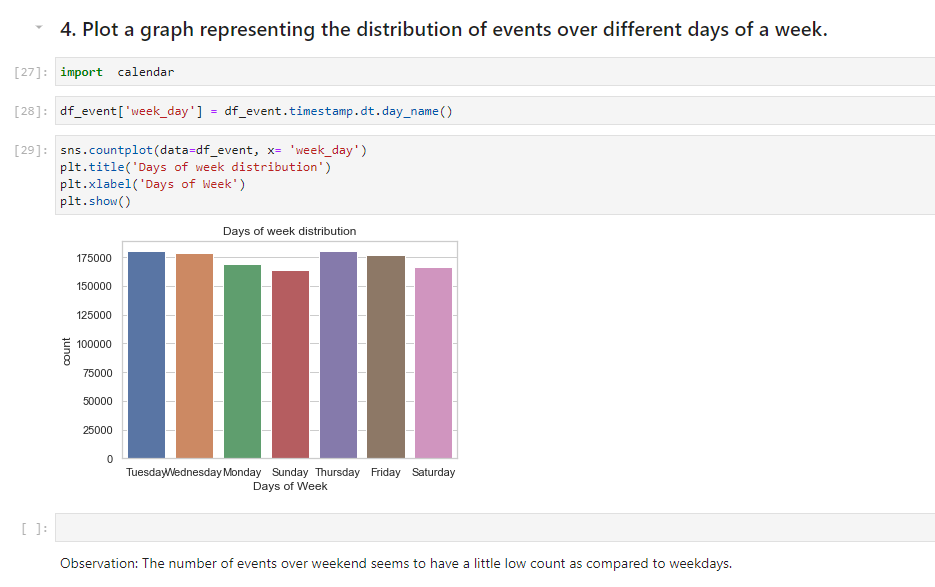
# **3) Outputs to the various EDA and Visualization codes along with the corresponding results and the insights gathered from each EDA and visualization.**

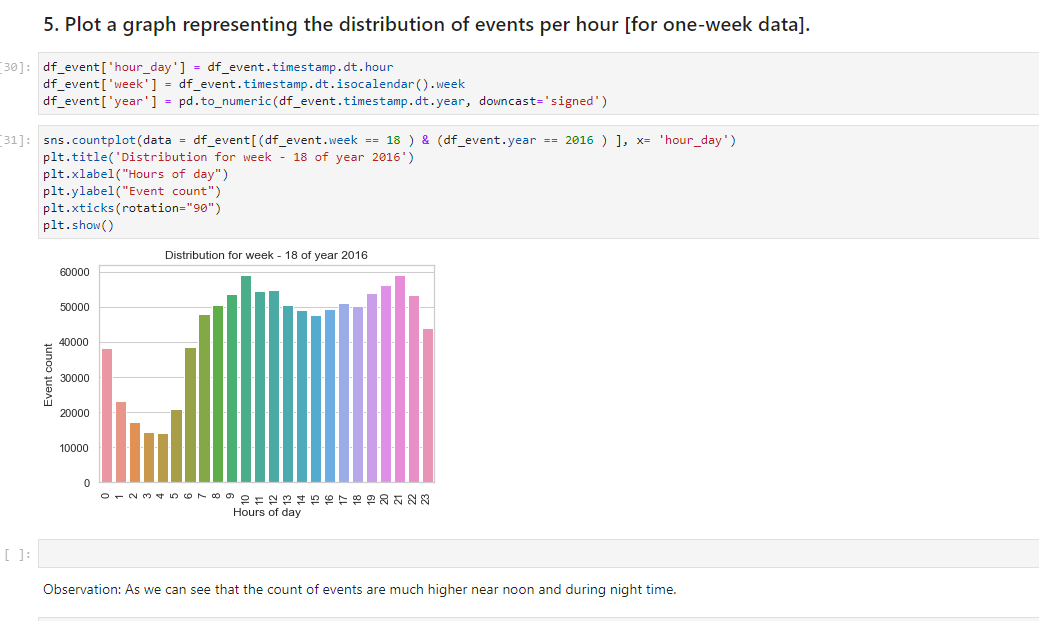
****

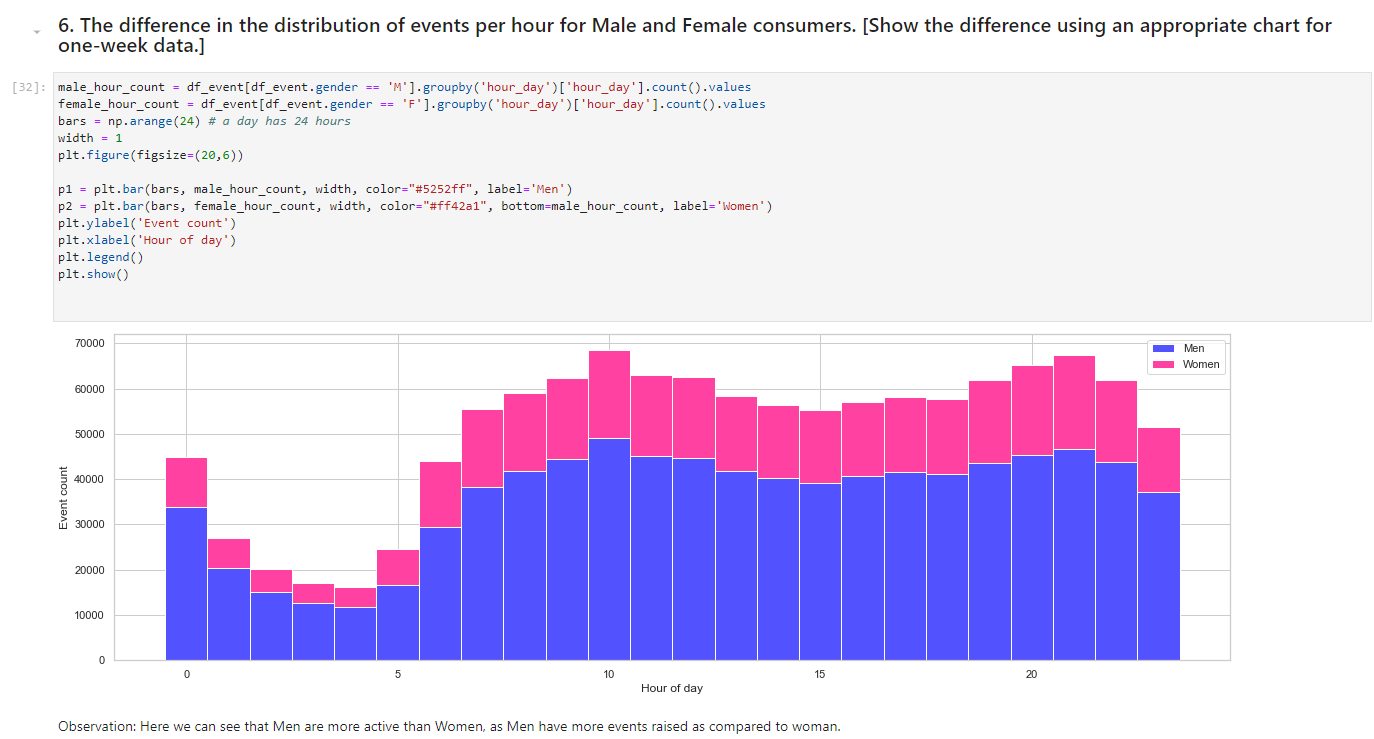
****

****

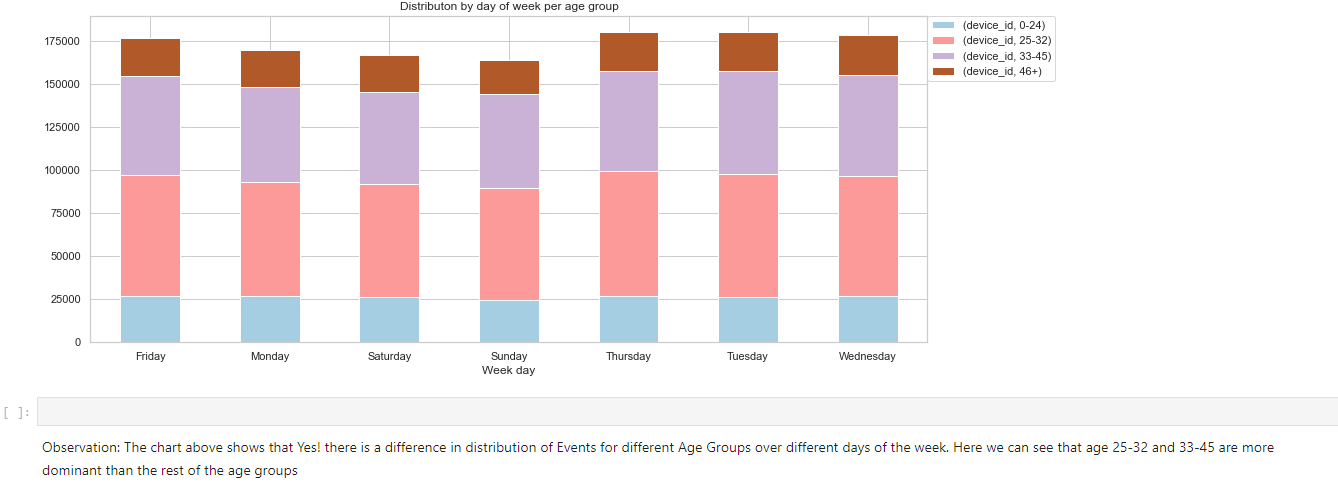
****

****

****

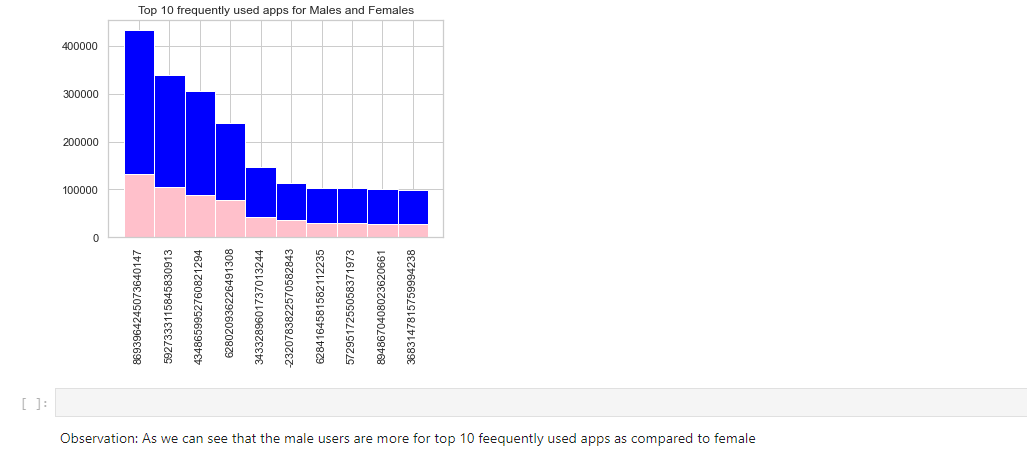
****

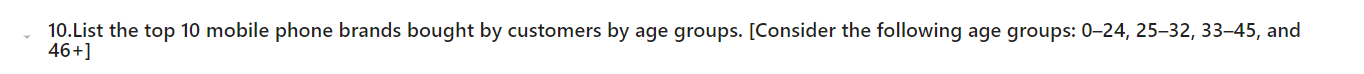
****

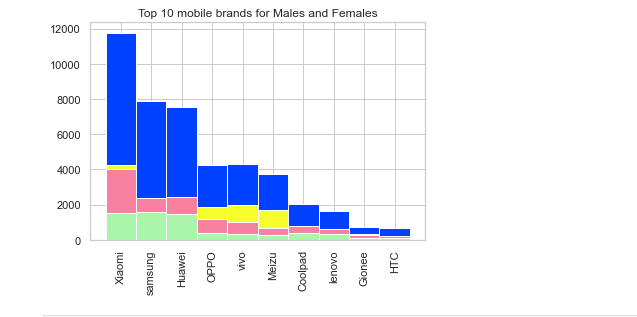
****

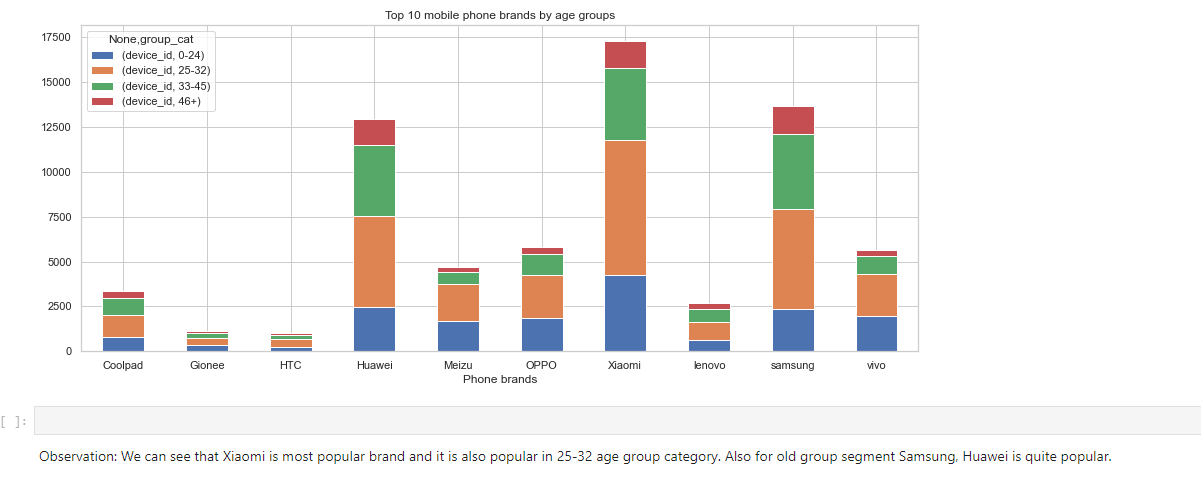
* ****

****

****

****

****

****

# **4) Geospatial visualizations along with the insights gathered from this visualization**

# 

# 

# 

# 

# 

# 

# **5) Feature engineering techniques that were used along with proper reasoning to support why the technique was used**

**Initial notebook**: In this notebook we did following feature engineering to get better of data and these features will be helpful while creating our model

* Average events per day for each device, as one gender can be using more of the mobile handsets than the other. Similarly, one age group (young) may be more used to mobile while other (old)
* Median used-time per device, as one section of users may be more active at a particular time than others.

**Advance\_Visualisation notebook:** In this notebook we did feature engineering related to latitude and longitude data.

### changes in the latitude and longitude details at different times of the day

# **5) Results interpreting the clusters formed as part of DBSCAN Clustering and how the cluster information is being used**

We used latitude and longitude for grouping the data in various cluster using DBSCAN as below:



Observation: Based on DBSCAN clustering, we have found that there are total of 3639, clusters based on longitude, latitude. We also created a special cluster for missing lat/long data.

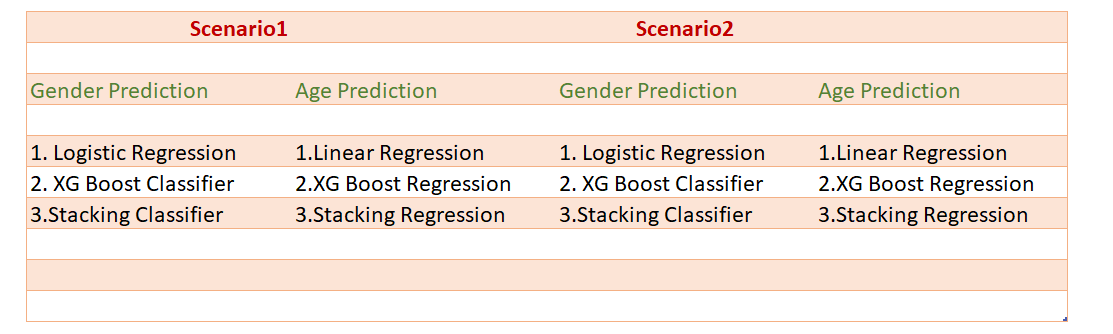
We used this clusters as a feature in our model building process.

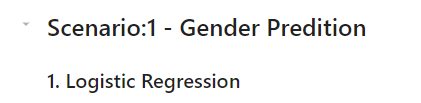
Note: There was no other subtask performed to improve the data cleaning and feature generation step.

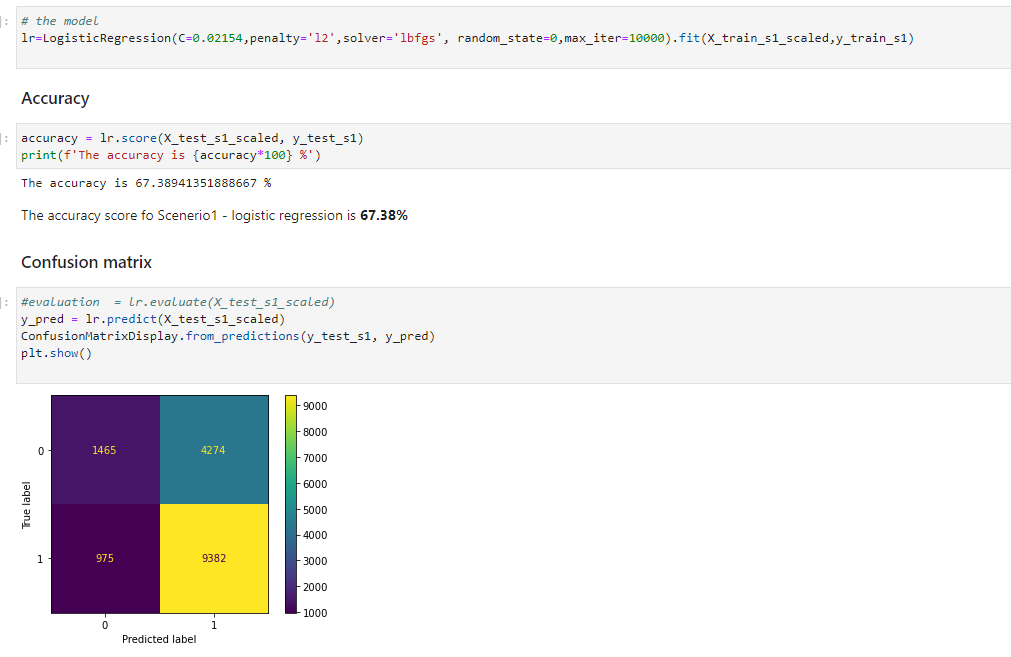
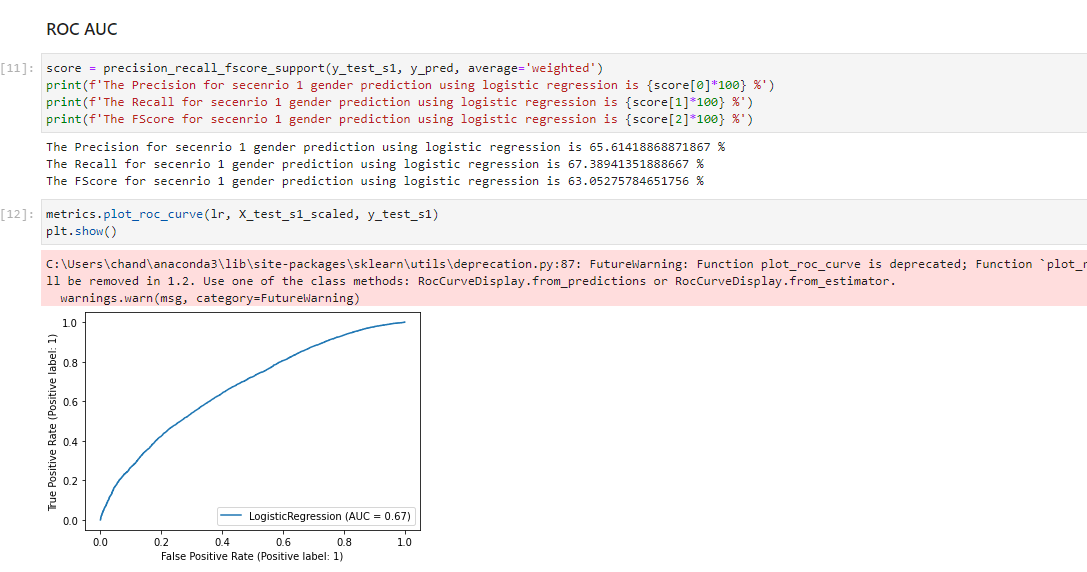
# **6) All the data preparation steps that were used before applying the ML algorithm**

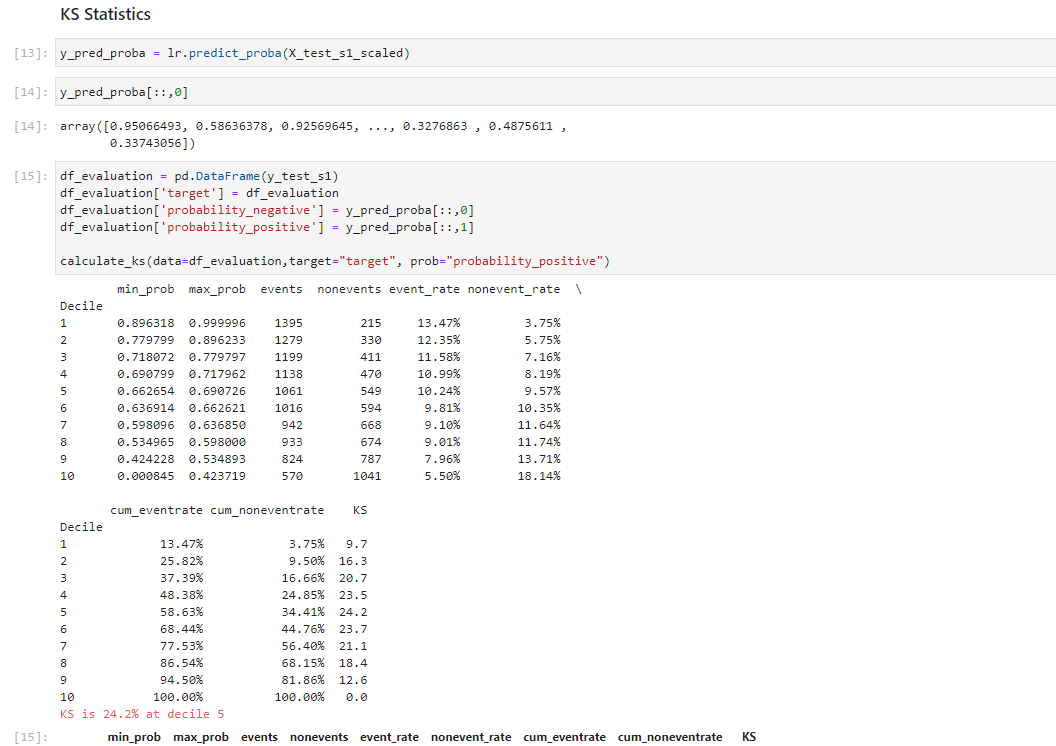
* We extracted label ids for each app. And updated in app events dataset.
* We also collected all the label ids for each event and updated events dataset.
* Finally, the device ids are updated with all the associated label ids.
* Converted the features in One-Hot-Encoding
* Created the train and test split as per the file provided for train-test split
* We created dummy variables for features like ['phone\_brand', 'device\_model', 'group\_cat', 'loc\_cluster']
* After all the above processing we saved the test and train split in separate files for further building models on with them.

# **7) Documentation of all the machine learning models**

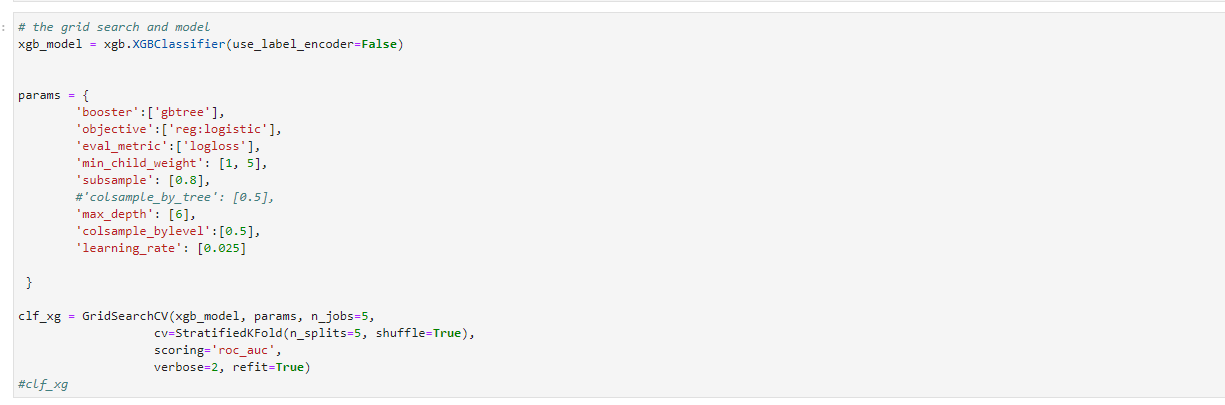
We have built following models as per model building exercise.

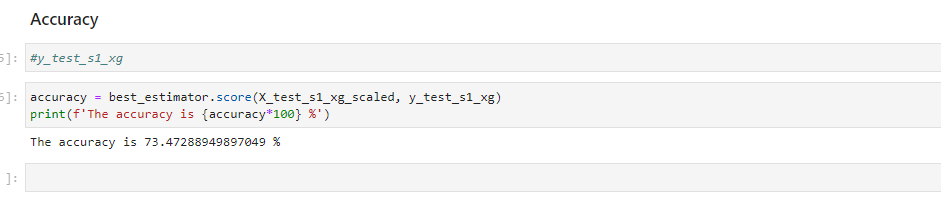


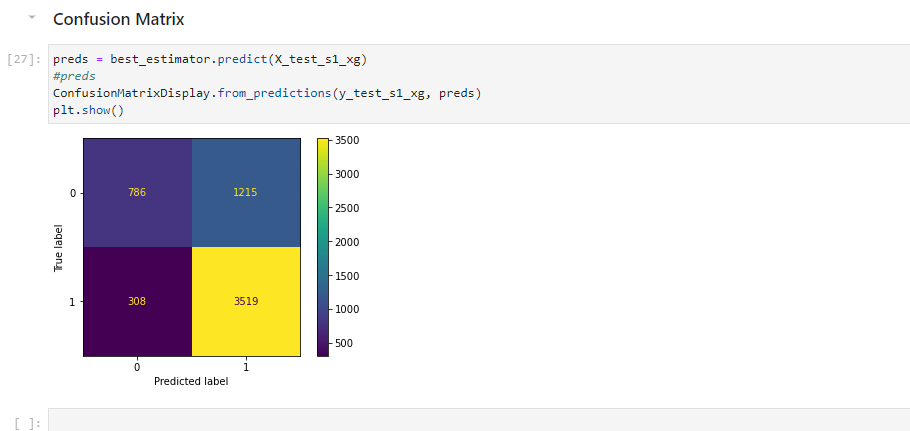


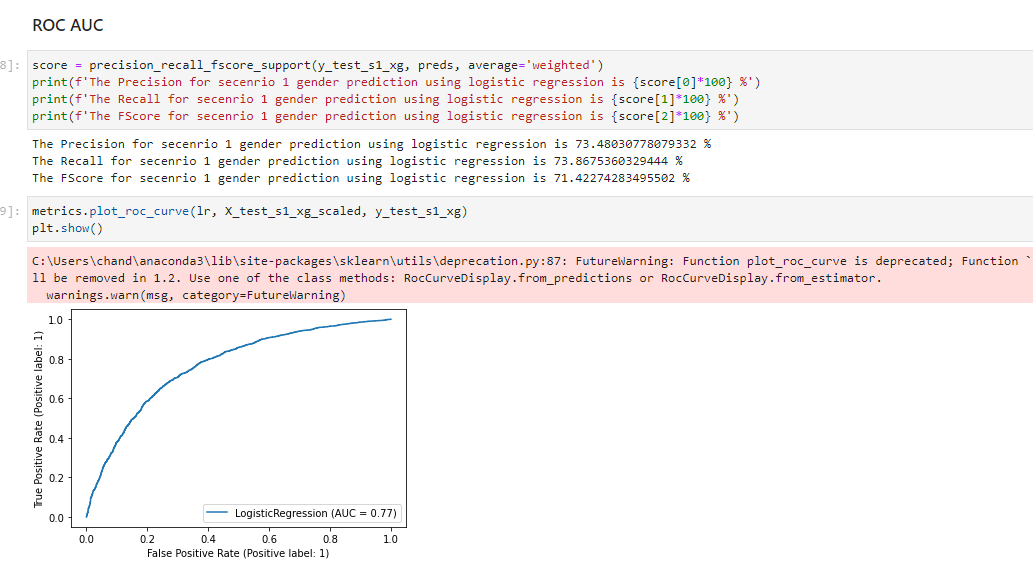


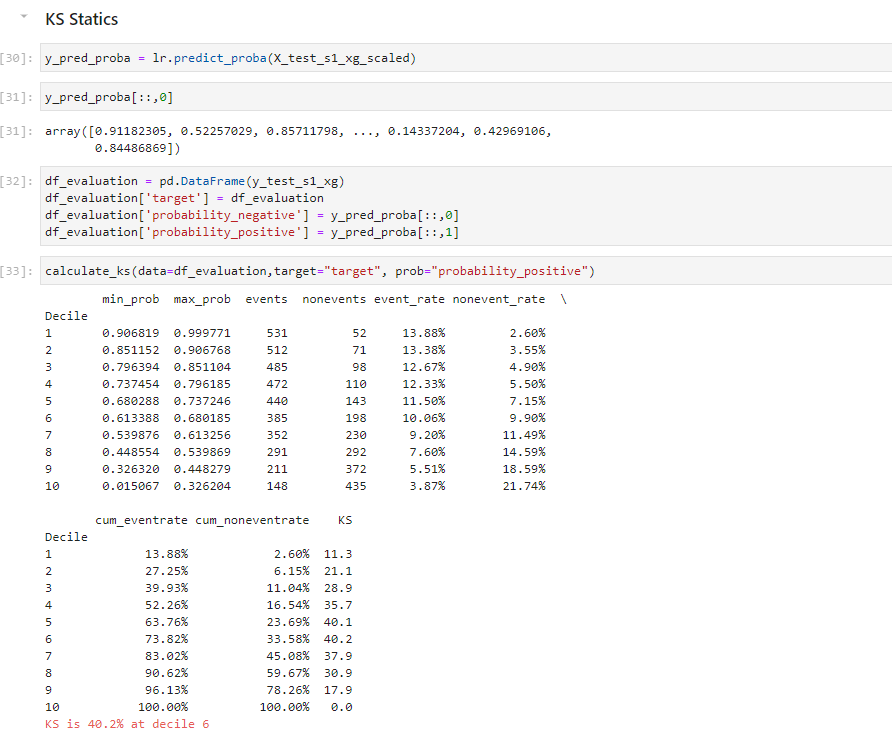
**Observation:** As we found out that the accuracy was not very good for this model we will keep trying with other models.







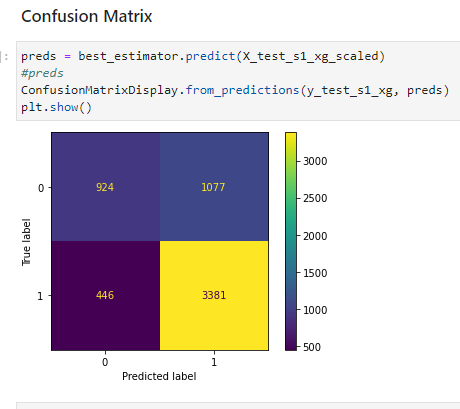


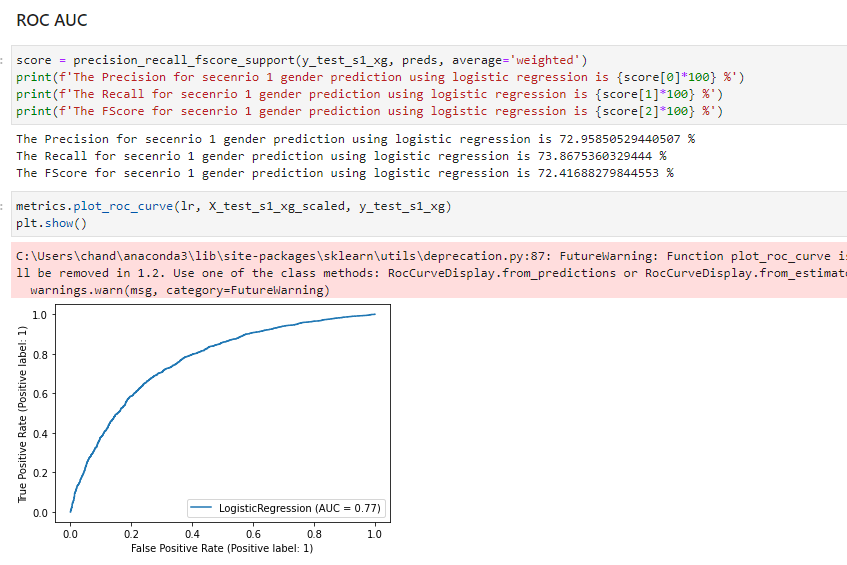


**Observation**: Here we observed that the accuracy of the model has been improved also the other matrices are much better.

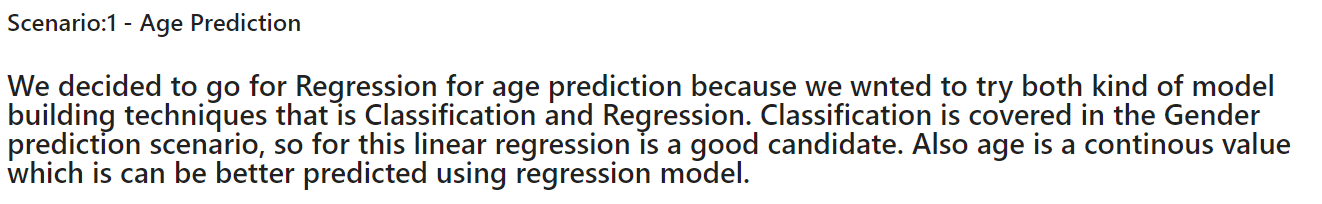




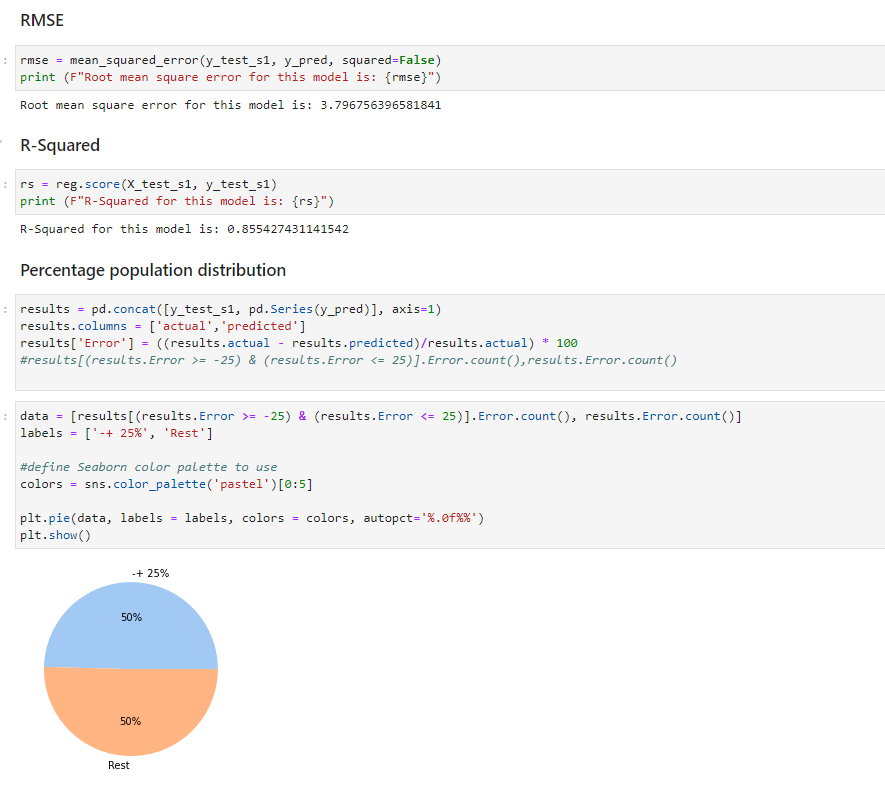




**Observation**: We have found that there is not a much difference in accuracy, however was reduce. Hence, we can confirm that XGBoost was the best model and same has been used to generate pickle file at later stage for flask application.

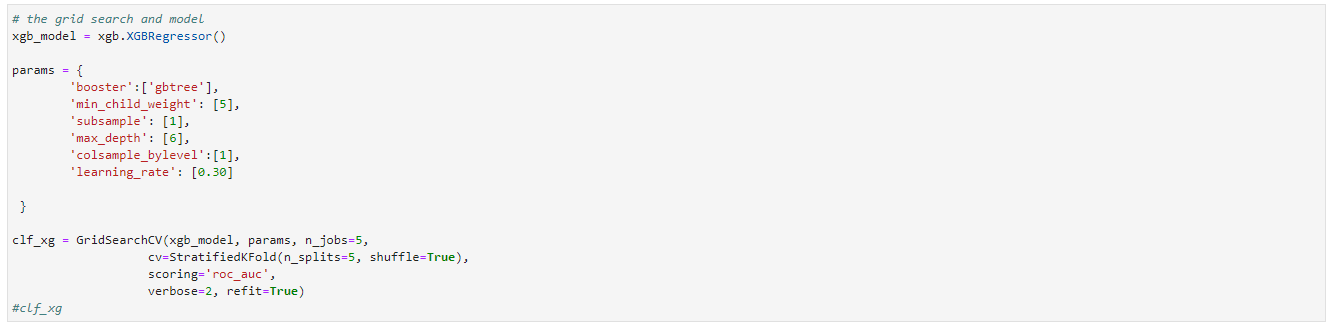


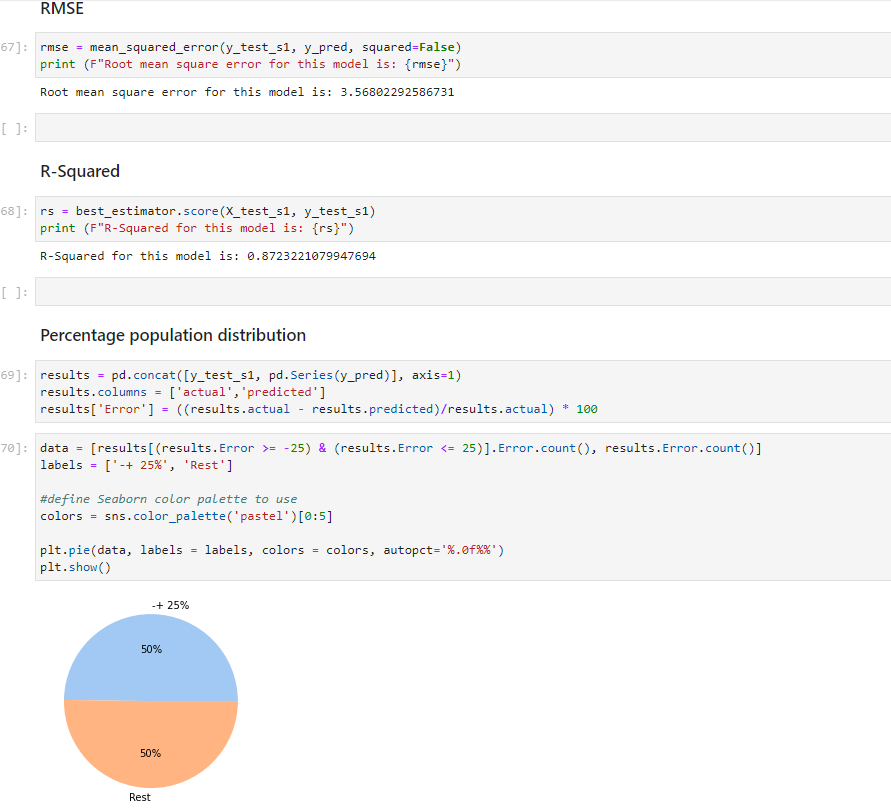




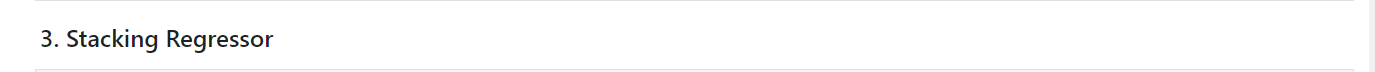
**Observation:** We have seen that the model has a RSquared of 0.855. We can try with other techniques to see improvement.

****

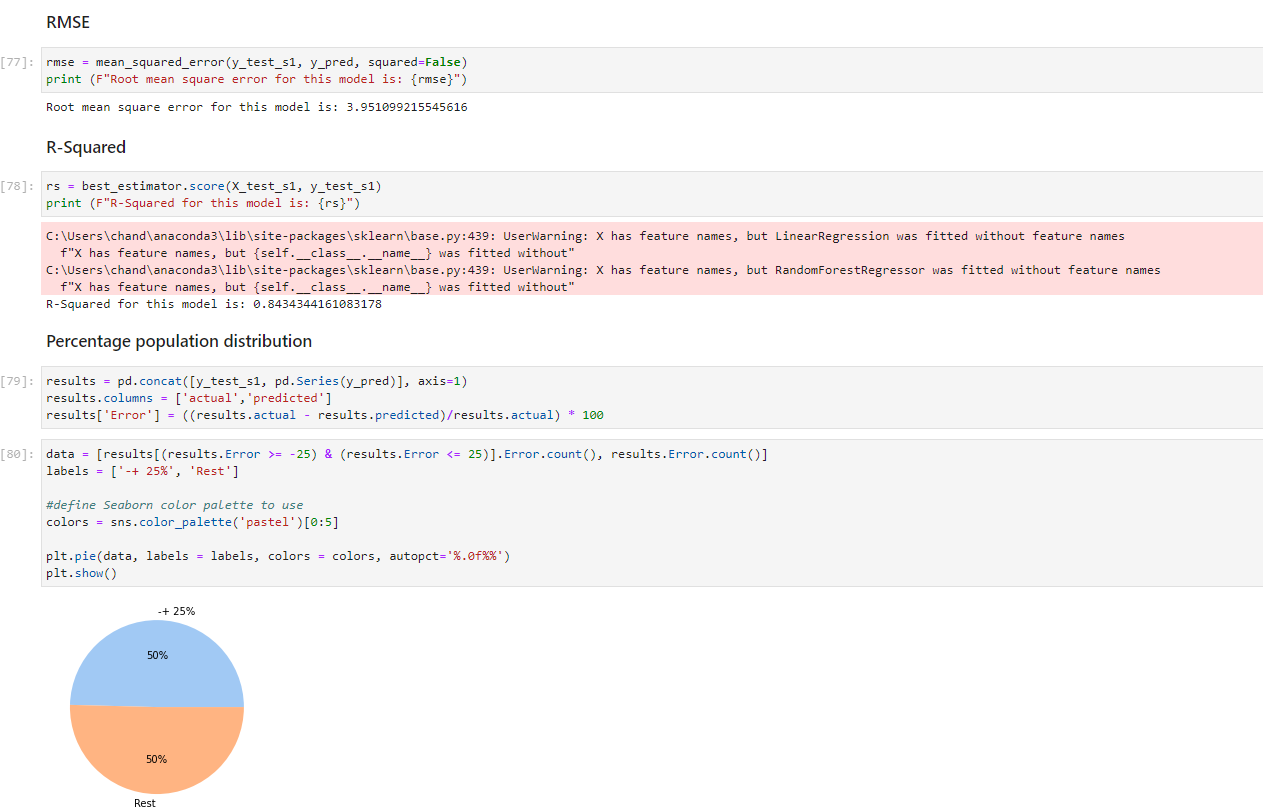
****

****

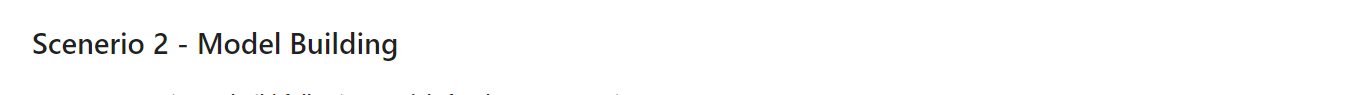
**Observation:** We found that this model performed better than the linear regression. We will try one more model to evaluate if we can get better results.

****

****

****

**Observation**: We have found that there is not a much improvement. Hence, we can confirm that XGBoost was the best model and same has been used to generate pickle file at later stage for flask application.

****

**For scenario 2 model building was similar to the scenario 1. The only difference was data so we are going to display the matrices in a tabular form. The detailed output is shown in the notebook “scenerio2\_model\_building.ipynb”.**





**Observation:** For scenario 2 we see that XG boost classifier and regressor are the best models.