

**Report On AD Campaign recommender**



|  |  |
| --- | --- |
| **S. NO.** | **Topic** |
| 1 | Introduction |
| 2 | EDA & Date Cleaning and Preprocessing |
| 3 | Model Building |
| 4 | Final Words |

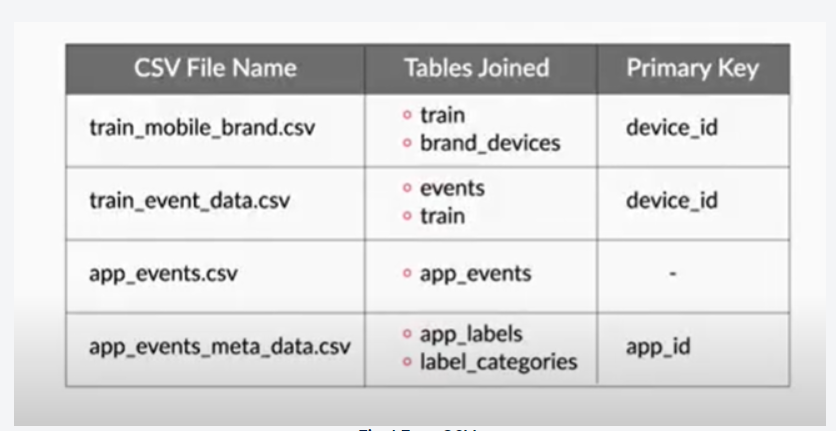
1. **Introduction**

On the basis of the gender and age recommending the campaign to the person. For example, targeting the fashion related campaigns to female and data packs to male customers, smartphones offer to the young age people cashback offers for adults.

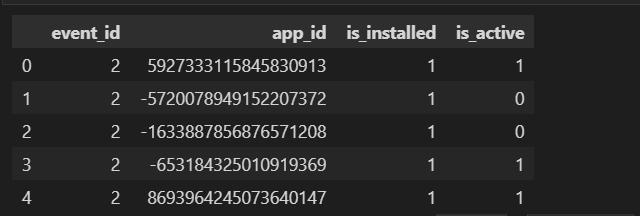
**Main Goal**

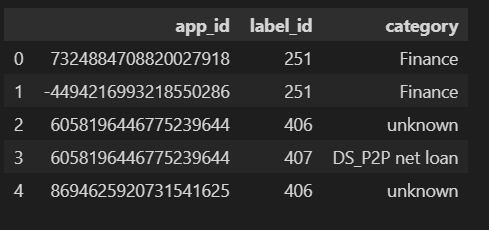
* Dataset contains large amount of data and most of the data is skewed so to fix it first then getting the import features from the data.
* Selecting the important features
* Training the model and to get the good accuracy

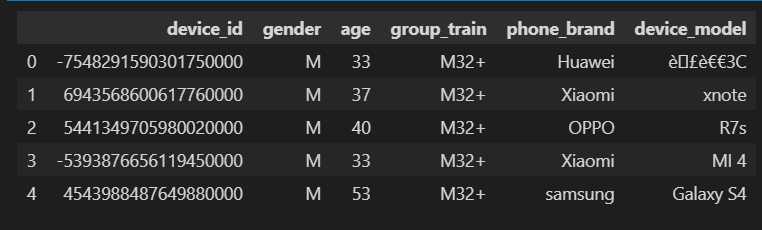
**About Dataset**

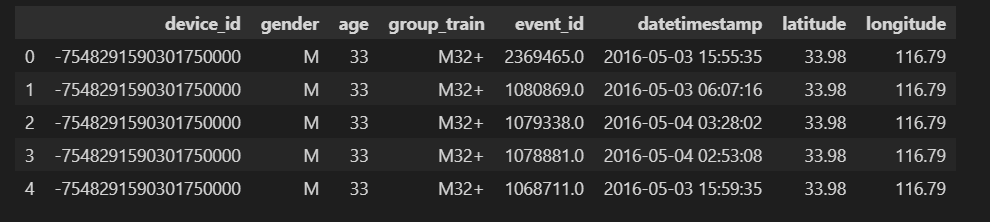
****

Data has four csv having the following fields:

****

****

****

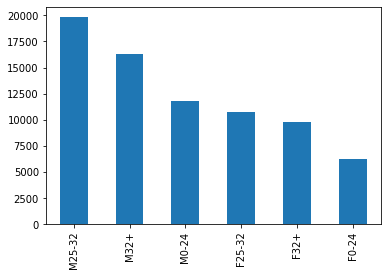
****

1. **EDA & Date Cleaning and Preprocessing**

Observations:

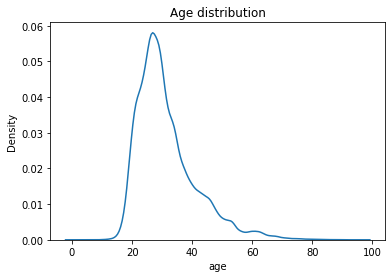
* is\_installed feature is not giving any important information because its value is only one.
* number of inactive apps are more
* some of the major category are Industry tag, Custom label, Tencent, game, Property Industry 2.0
* number of males are approximately double the male

**Age and gender distribution**

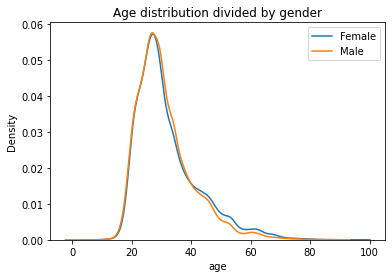




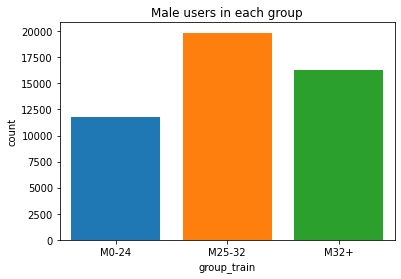
* There are more males than females in this dataset so it is unbalanced.



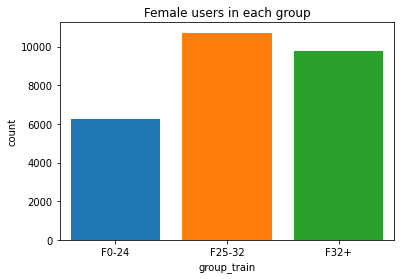
* Age distribution is also unbalanced and most people are in their 20's



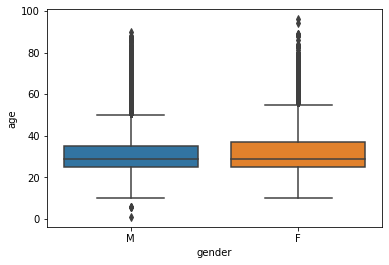
* Age distribution for male and females is almost same.



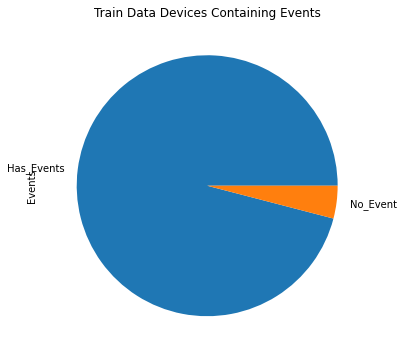
* Distribution of male in each group is also unbalanced

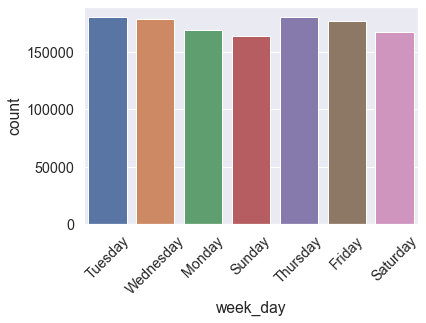


* Distribution of female in each group is also unbalanced
* These plots shows that the classes are unbalanced. Young people are largely more numerous than old, also there is a higher representation of males over females. Among the older, there are more data coming from males rather than females. This result is consistent with Chinese demographic and the fact that the use of technology is more diffused among younger generations. Therefore these data don't have to be resempled in order to create a uniform distribution of the classes.

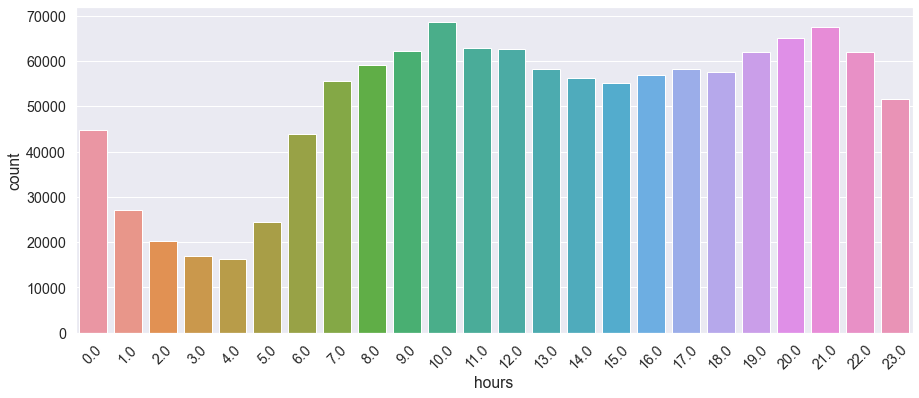


### **Trends in event data (devices, days of the week, hours, gender, and age groups):**



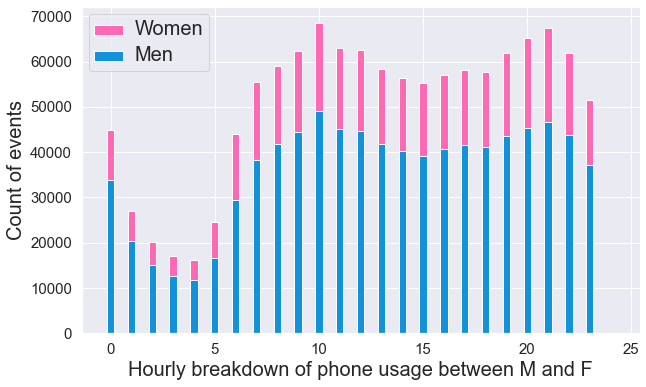


### **Graph representing the distribution of events per hour (for one-week data)**

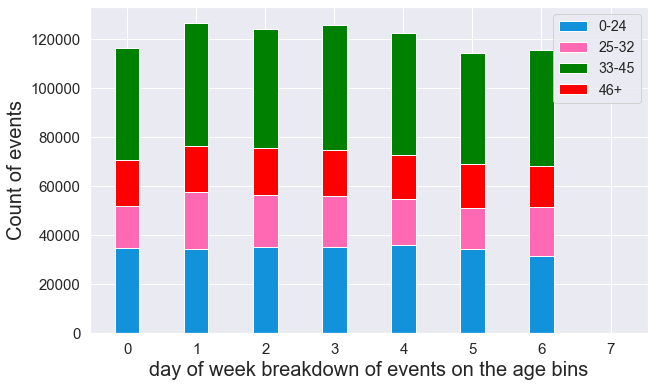


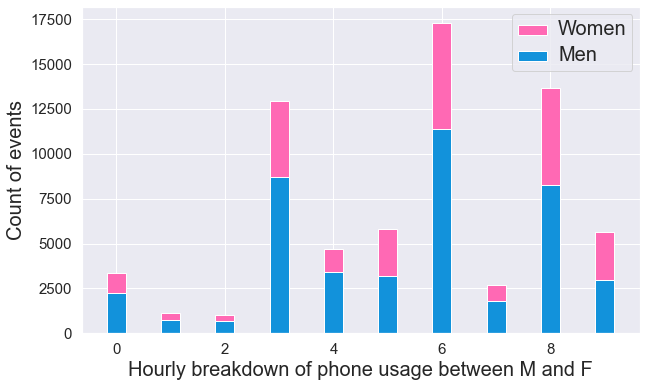
* We can see usage of apps are more in the day time and in the night time usage is just half it's quite acceptable because adults dont use mobiles in the night.

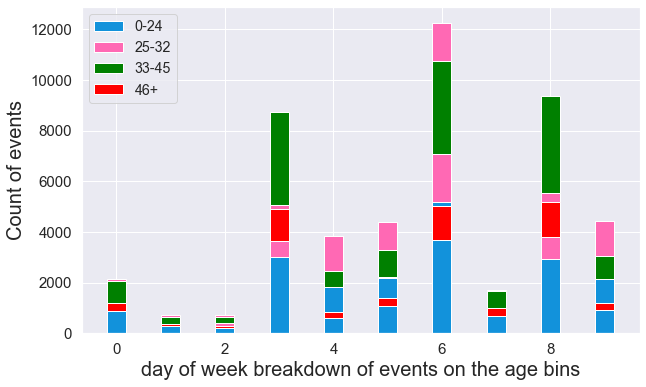
### **The difference in the distribution of events per hour for males and females (Show the difference using an appropriate chart for one week’s data)**

* 

### **Is there any difference in the distribution of events for different age groups over different days of week? (Consider the age groups as 0–24, 25–32, 33–45, 46+]**



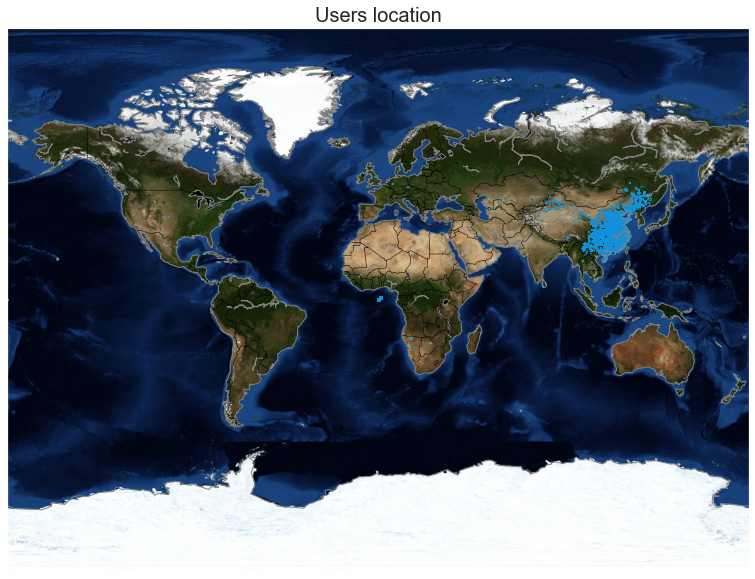


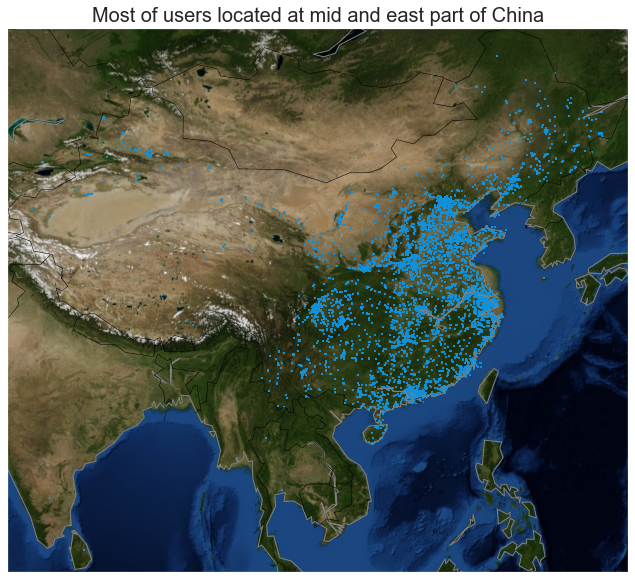


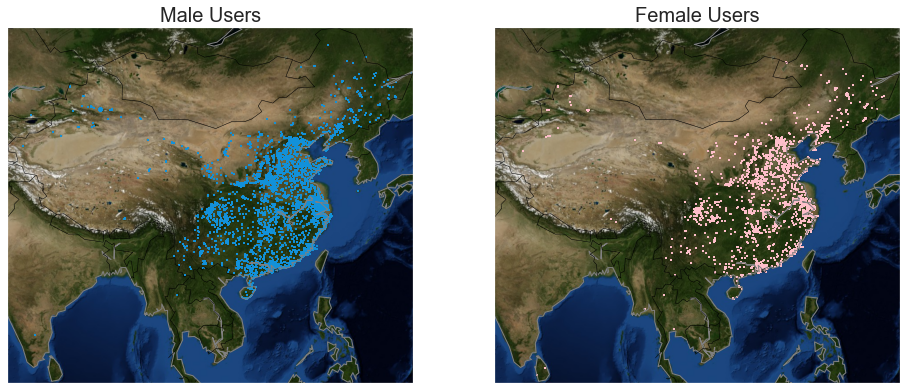
**Feature Engineering**

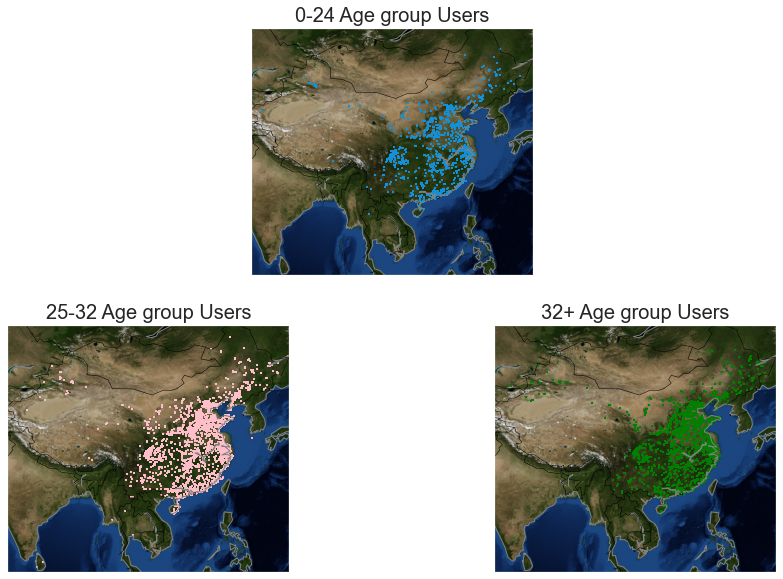
* **Label encoding following columns:**
  + Phone Brand
  + Pone Model
  + App Id’s
* **Clustering**
  + Kmeans Clustering
  + DBScan Clustering

### **Advanced Visualization and Clustering**

****

****

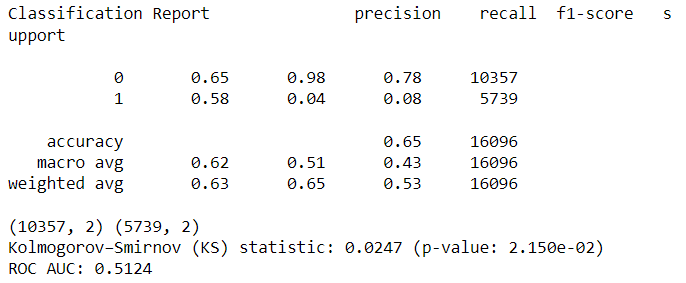
****

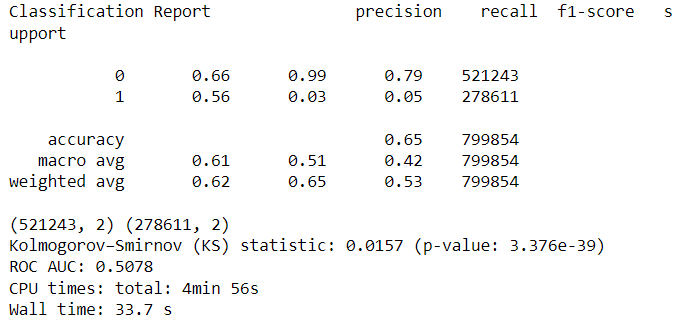
****

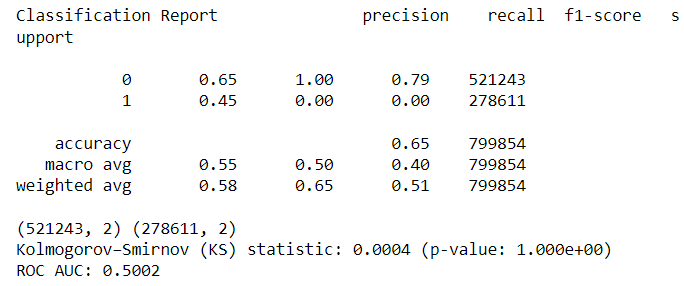
**Model Building**

**Scenario 1 – Data having Events data**

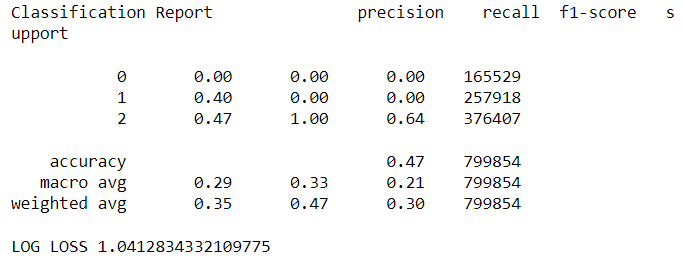
* **Gender Prediction**
  + Applying the one hot encoding on the categorical variable using the column transformer then training the machine learning model
    - **Logistic Regression**

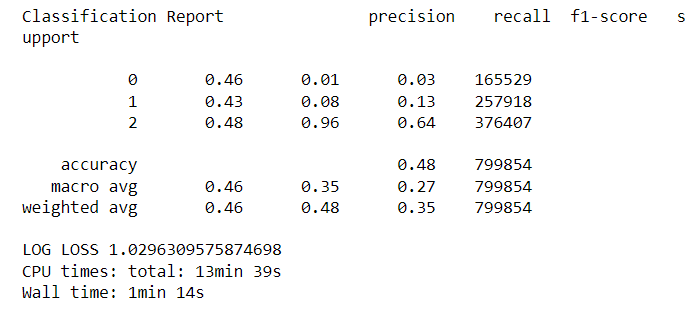


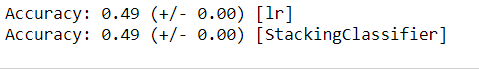
* + - **XGBoost**
* ****
  + - **Stacked model**



* **Age prediction**
  + Training the model
    - **Logistic regression**

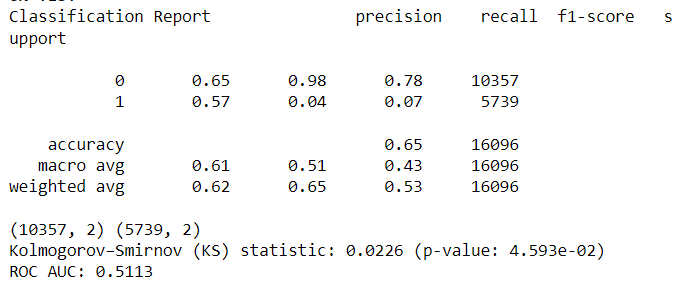
**D**

* + - **XGBoost**
* ****
  + - **Stacked model**

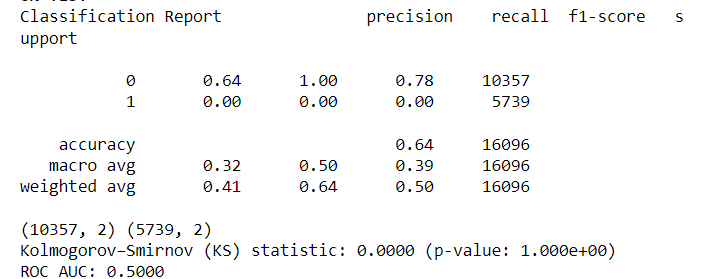
****

**Scenario 2 – Only having the Phone brand and model data**

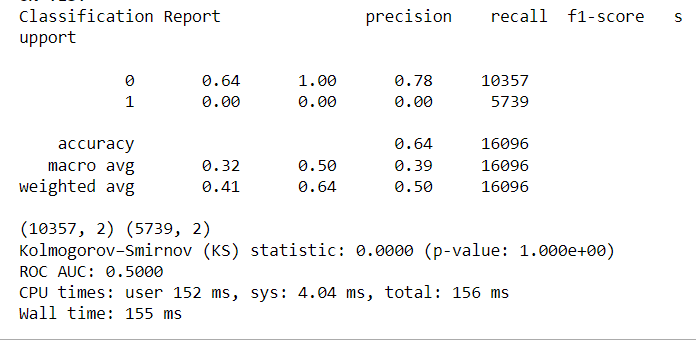
* **Gender Prediction**
  + Training the model and one hot encoding on the phone model and brand
    - **Logistic regression**



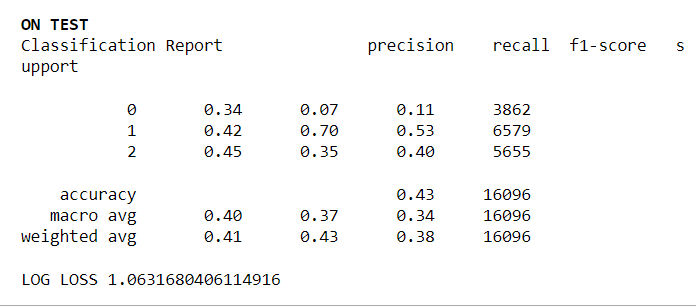
* + - **XGBoost**



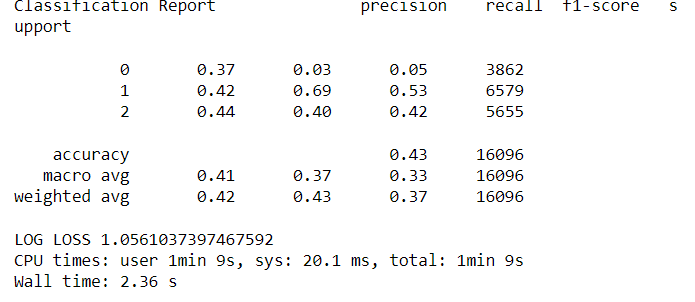
* + - **Stacked model**



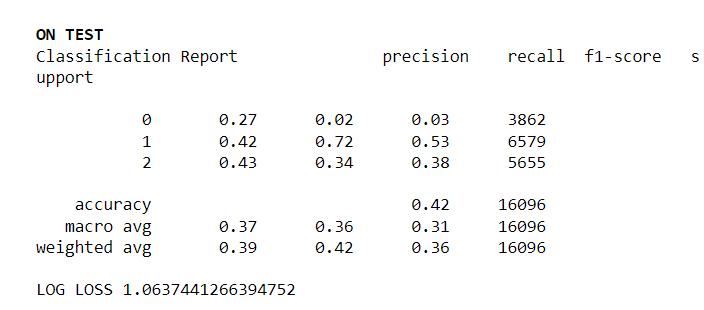
* **Age prediction**
  + Training the model
    - **Logistic regression**



* + - **XGBoost**

****

* + - **Stacked model**

****

**Final Recommendations:**

We can try Neural network for getting better accuracy and we can also try doing more feature engineering on the data and normalizing the data.