**REPORT ON CAPSTONE PROJECT**

**LIVE STREAMING DATASET ANALYSIS**

**INTRODUCTION**

Problem statement: A streaming platform company allows different streamers to use its platform to conduct/deliver a live stream session.

Now, the company wants to know which are its top streamers. But the company doesn’t have a labeled dataset that indicates which are the top streamers and which are not.

This dataset is generated from the live-streaming platform. It consists of all the streams that were conducted on the platform and the information related to each stream - its duration, country it was streamed from, no. of comments received during the stream, no. of viewers who attended the live stream etc.

**Tasks:**  1. Analyse the dataset to come up with the top 20% streamers.

2. Label these top 20% streamers as “good” streamers, and the remaining as “bad” streamers. This will become your target variable. Now create a binary classification ML model that can classify whether any streamer is a good streamer or not.

**Evaluation Metric:** Use F1-Score

**TOOLS\SKILLS USED**

* Python Programming
* Python Libraries and Packages Numpy, Pandas, Matloplib, Seaborn
* Exploratory Data Analysis & Data Visualisation
* Machine Learning

**OBJECTIVE**

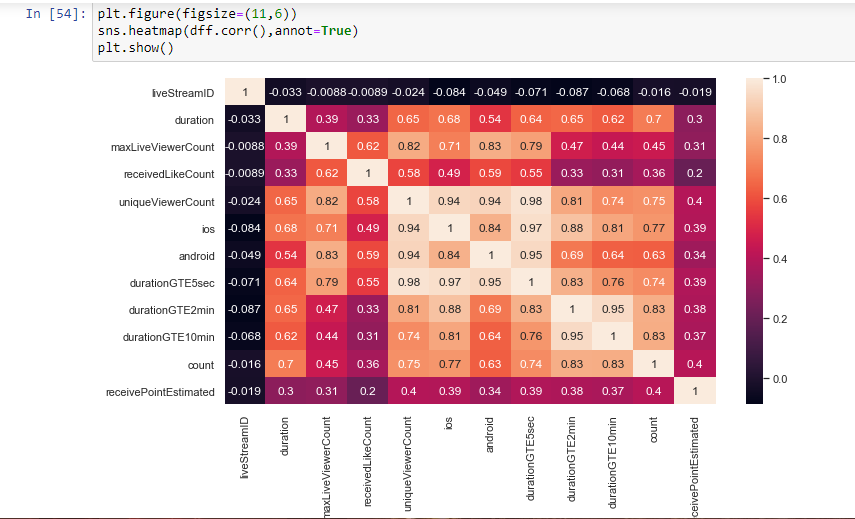
Analyse the dataset to come up with the top 20% streamers & create a binary classification ML model that can classify whether any streamer is a good streamer or not.

**METHODOLOGY**

* Data Description
* EDA
* Visualisation of the Data
* Model Building &Finding Insights

**VISUALISATION OF THE DATA**

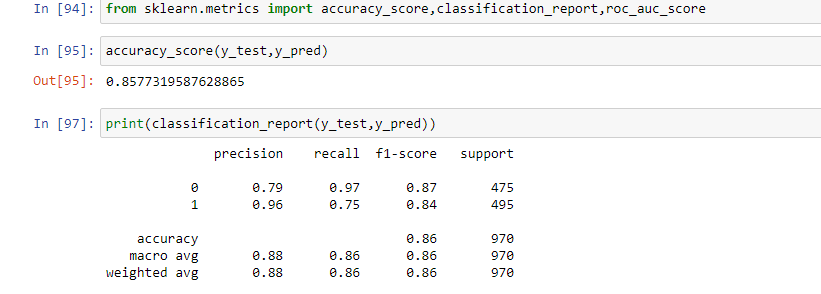
**Visualisation Snippets:**



From above plot, we can see how different features are correlated with each other. “Duration” is mostly co-related to every variable.

**MODELLING & CONCLUSION**

* After the analysis we found that duration is the most important column for creating target variable. So, target variable is created on the basis of duration. Top 20% streamers with highest duration is labeled as “good streamers” & rest “bad streamers”.
* In order the make dataset balance in nature we have used the Synthetic Minority Oversample Technique ( SMOTE ) .
* A Classification Machine Learning Model is created with the Logistic regression algorithm
* The accuracy of the model is 85.7%.

****

* From the above Screenshot we can easily see f1-score is quite well & also accuracy of the model is 85.7%.

**FUTURE SCOPE**

In future, the model can be upgraded with some better techniques in terms of getting higher and better quality Analysis.