ENGAGEMENT SCORE PREDICTION

Girish Sahu

JOB-A-THON

11-13 Feb 2022

THE WAY TO GET STARTED IS TO QUIT TALKING AND BEGIN DOING.

Walt Disney

Agenda

Problem Statement

Data Description

Exploratory Data Analysis

Model Building

Summary

Problem Statement

ABC is an online content sharing platform that enables users to create, upload and share the content in the form of videos. It includes videos from different genres like entertainment, education, sports, technology and so on. The maximum duration of video is 10 minutes.

Users can like, comment and share the videos on the platform.

Based on the user’s interaction with the videos, **engagement score** is assigned to the video with respect to each user. **Engagement score** defines how engaging the content of the video is.

Understanding the **engagement score** of the video improves the user’s interaction with the platform. It defines the type of content that is appealing to the user and engages the larger audience.

The main objective of the problem is to develop the machine learning approach to predict the **engagement score** of the video on the user level.

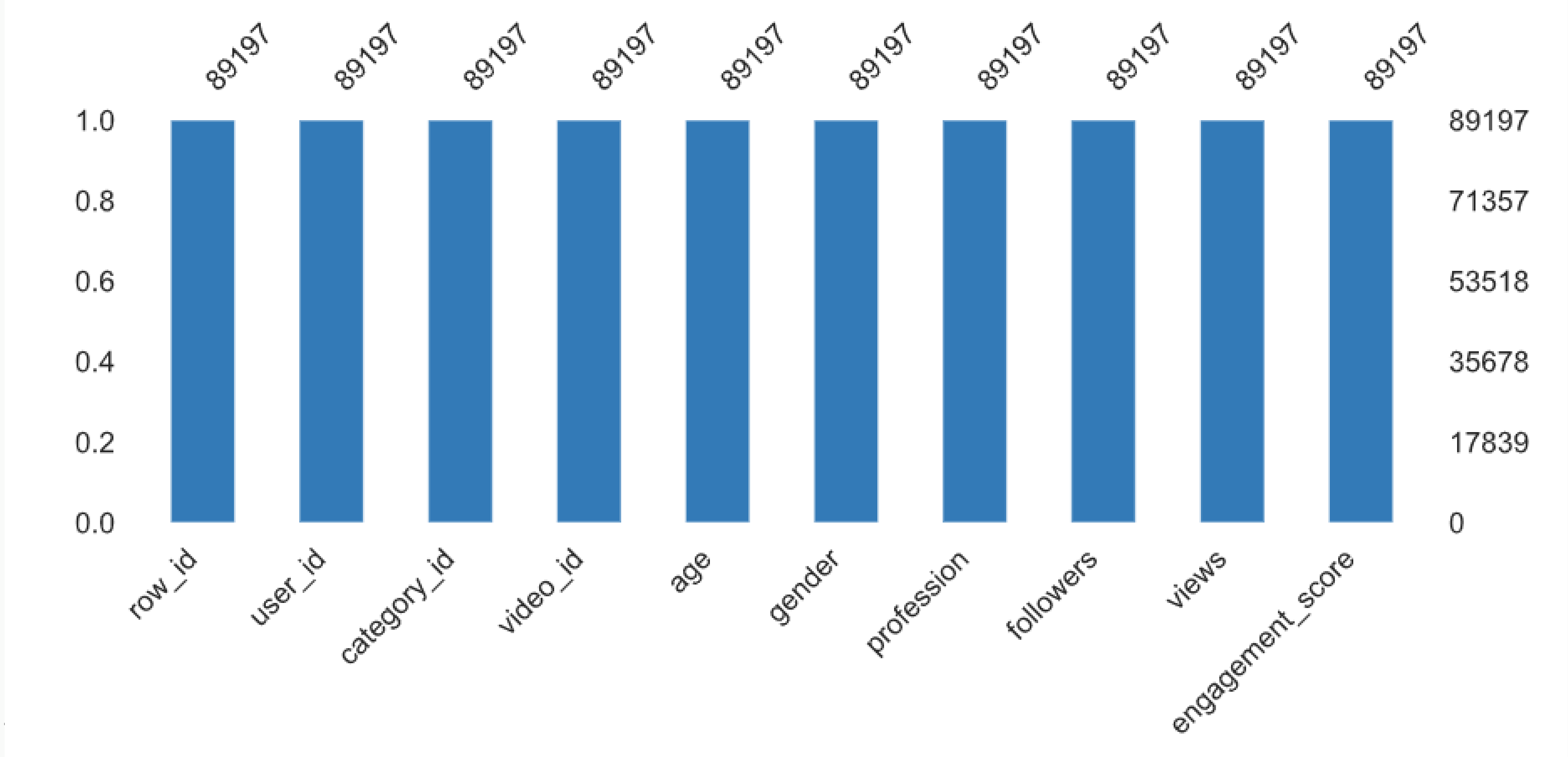
Data Description

Dataset

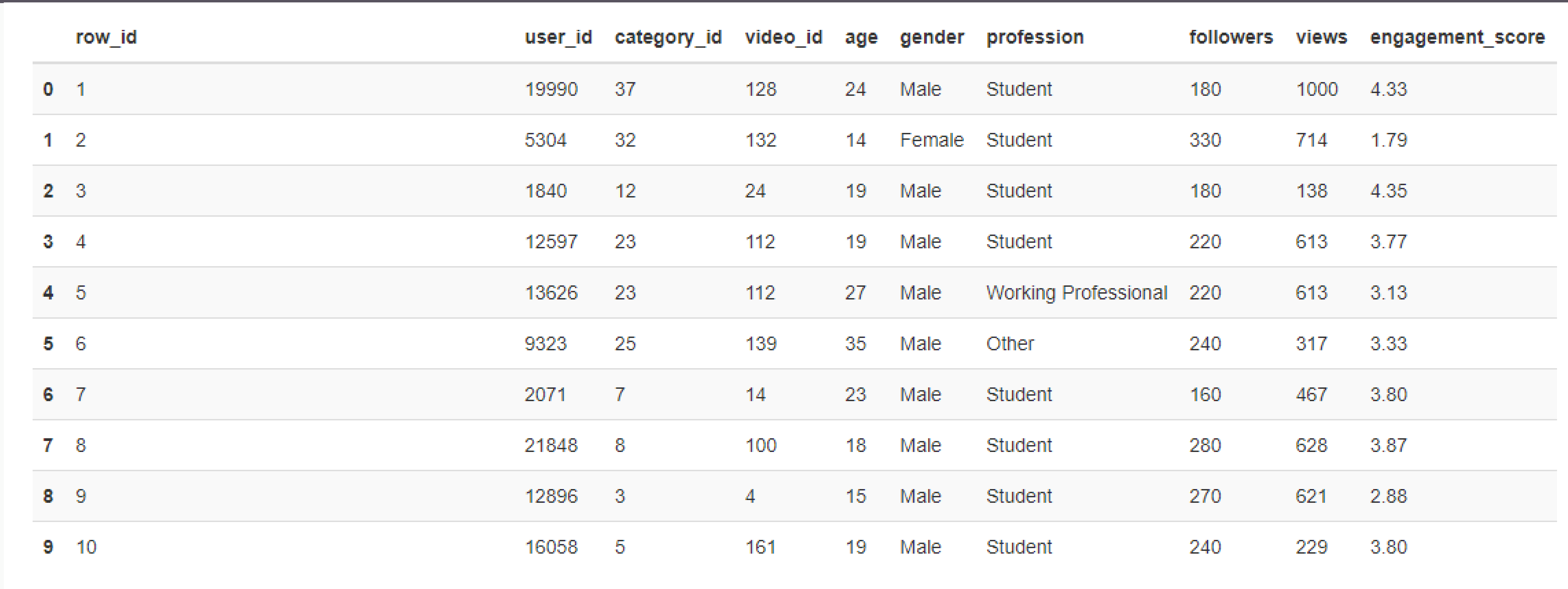
DATASET DESCRIPTION - OVERVIEW

Number of Variables	**10**
Number of Observations	89197
Missing Cell	0
Missing Cell (%)	0.0%
Duplicate Rows	0
Duplicate Rows (%)	0.0%
Numeric Variable	8
Categorical Variable	2

Dataset Description - Count



Dataset Description- Sample Data



	row_id	user_id	category_id	video_id	age	gender	profession	followers	views	engagement_score
0	1	19990	37	128	24	Male	Student	180	1000	4.33
1	2	5304	32	132	14	Female	Student	330	714	1.79
2	3	1840	12	24	19	Male	Student	180	138	4.35
3	4	12597	23	112	19	Male	Student	220	613	3.77
4	5	13626	23	112	27	Male	Working Professional	220	613	3.13
5	6	9323	25	139	35	Male	Other	240	317	3.33
6	7	2071	7	14	23	Male	Student	160	467	3.80
7	8	21848	8	100	18	Male	Student	280	628	3.87
8	9	12896	3	4	15	Male	Student	270	621	2.88
9	10	16058	5	161	19	Male	Student	240	229	3.80

Exploratory Data Analysis

Data Visualization

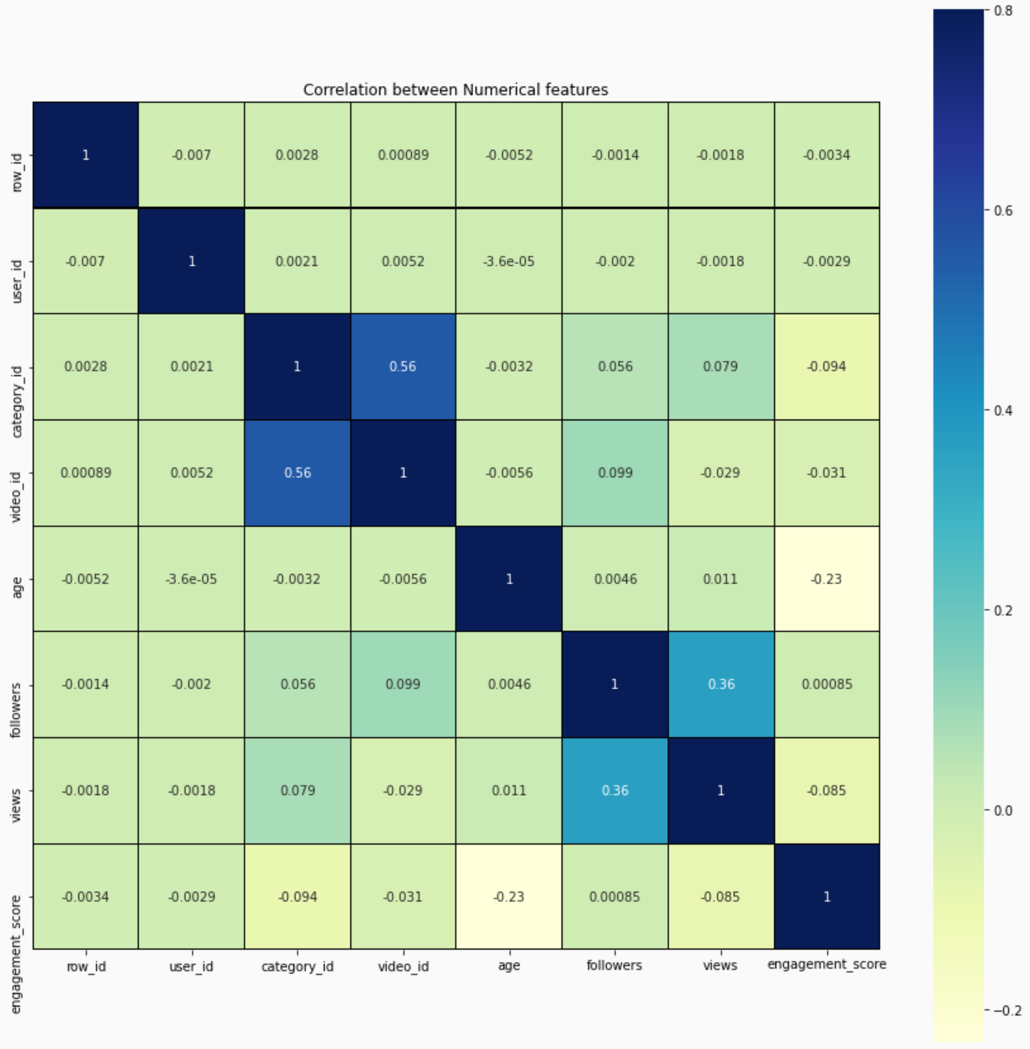
Exploratory Data Analysis

There were no missing values, and data distribution appeared to be normal which allowed to move to EDA without hesitation.

Used Seaborn plot for Univariate, Bivariate and Multi-variate Analysis. Please refer my Jupyter Notebook as reference.

There were no major correlation observed between variables as shown in Heatmap and Pair Plot in next few slides.

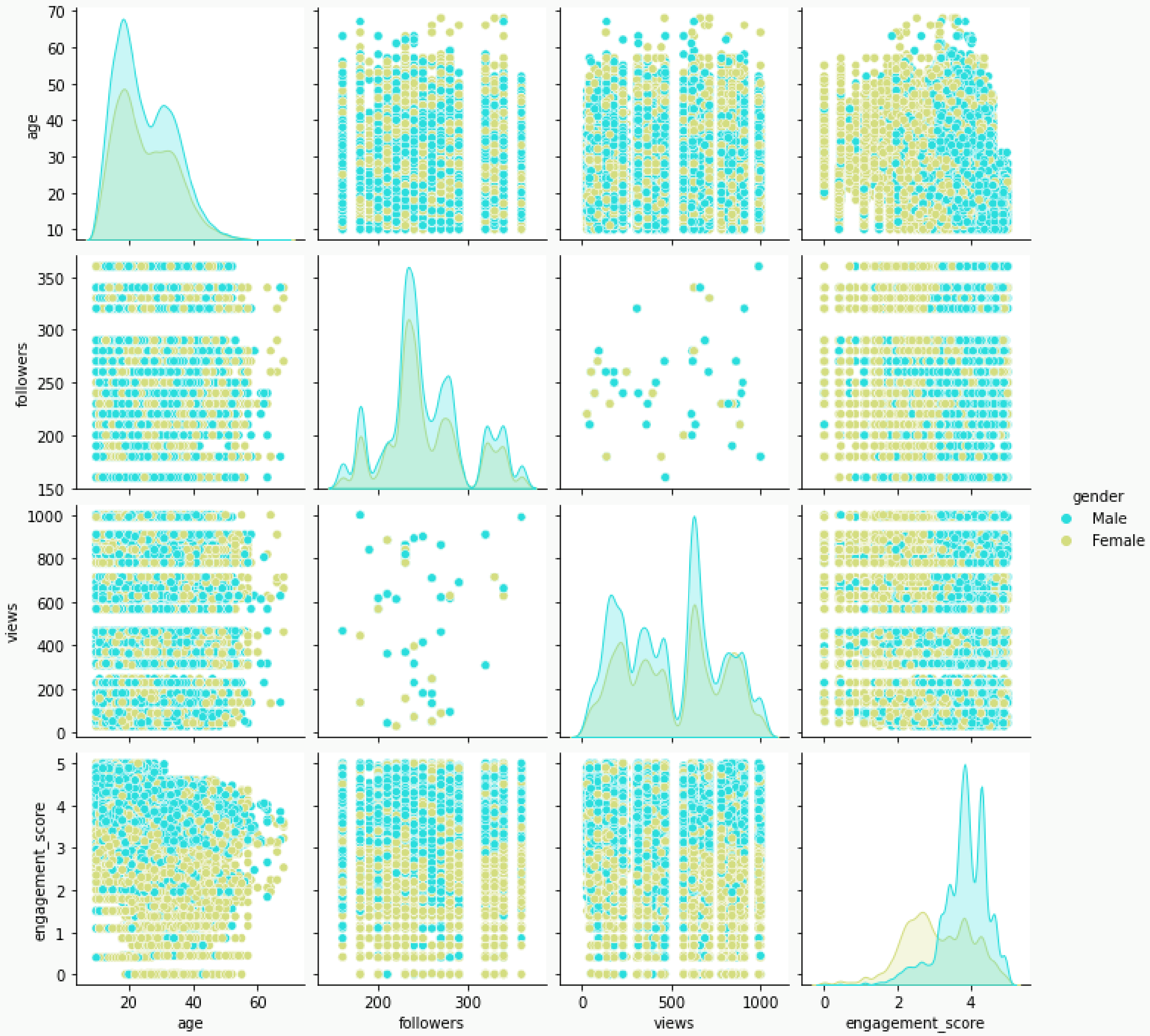
EDA- HEATMAP



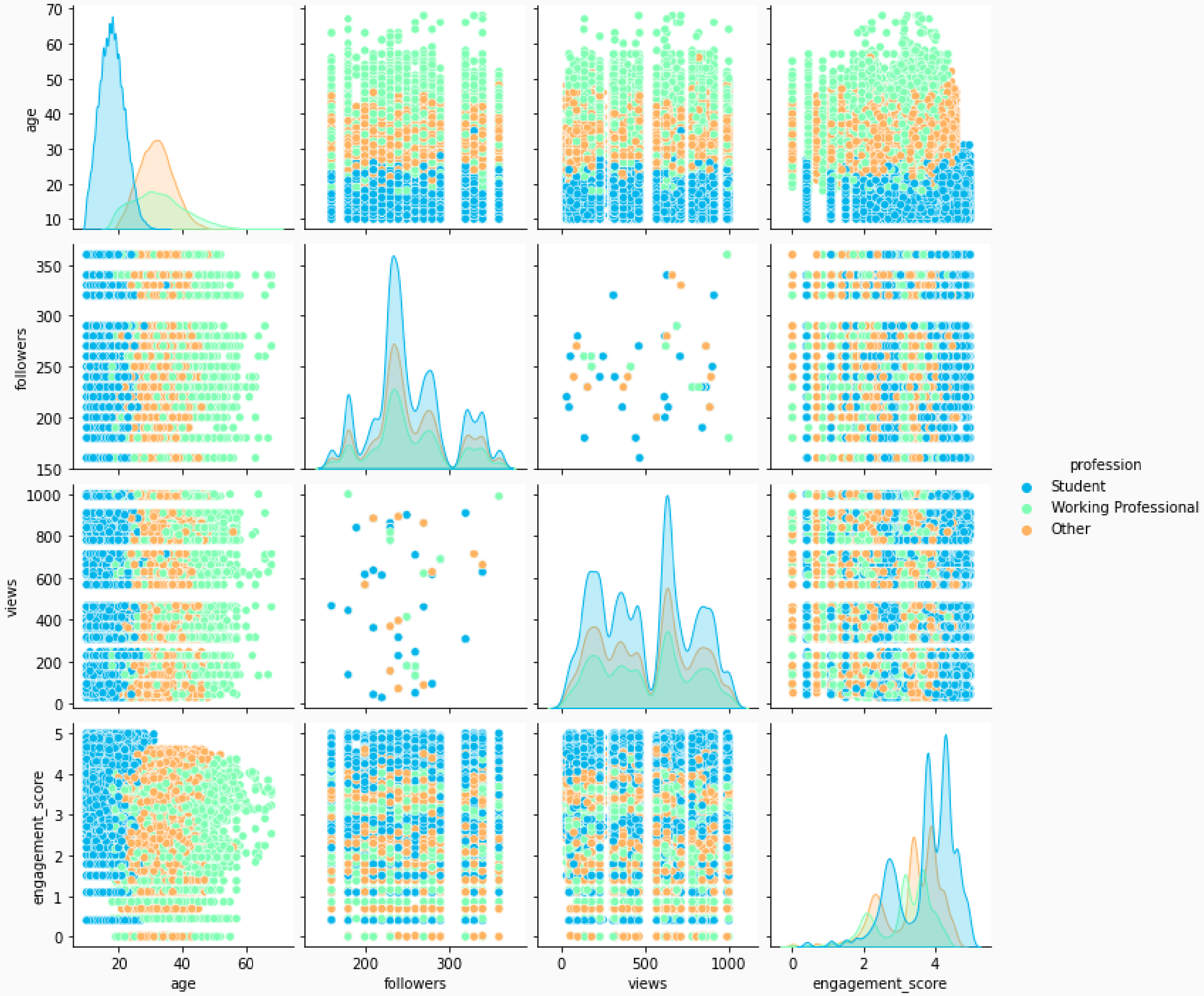
Correlation between Numerical features

	row_id	user_id	category_id	video_id	age	followers	views	engagement_score
row_id	1	-0.007	0.0028	0.00089	-0.0052	-0.0014	-0.0018	-0.0034
user_id	-0.007	1	0.0021	0.0052	-3.6e-05	-0.002	-0.0018	-0.0029
category_id	0.0028	0.0021	1	0.56	-0.0032	0.056	0.079	-0.094
video_id	0.00089	0.0052	0.56	1	-0.0056	0.099	-0.029	-0.031
age	-0.0052	-3.6e-05	-0.0032	-0.0056	1	0.0046	0.011	-0.23
followers	-0.0014	-0.002	0.056	0.099	0.0046	1	0.36	0.00085
views	-0.0018	-0.0018	0.079	-0.029	0.011	0.36	1	-0.085
engagement_score	-0.0034	-0.0029	-0.094	-0.031	-0.23	0.00085	-0.085	1

EDA-PAIRPLOT1



EDA-PAIRPLOT2



Model Building

Regression

Model Building

Categorical variable was converted to Numeric variable using encoding technique.

Feature selection was done using SelectKBest technique.

Data scaling was done using MinMaxScaler but at the end also tried model prediction without data scaling.

LinearRegression, RandomForestRegressor and XGBRegressor model was used and at the end RandomForestRegressor appears to be giving better prediction based on score generated in Leaderboard.

Summary

Best prediction appears to be coming out of RandomForestRegressor based on score generated at Leaderboard.

Features such as Gender, Profession and Age appears to have greater influence towards Engagement Score but at same time number of Followers and number of Views also play a critical role in determining engagement score.

THANK YOU

Girish Sahu
girishksahu@gmail.com