Capstone Project -2

Ted Talk Views Prediction

(individual)

Rajesh Kumar Patel

Introduction:

TED is a nonprofit devoted to spreading ideas, usually in the form of short, powerful talks. TED began in 1984 as a conference where Technology, Entertainment and Design converged, and today covers almost all topics — from science to business to global issues — in more than 100 languages. Meanwhile, independently run TEDx events help share ideas in communities around the world.

TED offers speakers a platform to provide information directly to millions of people around the world.

Problem Statement:

The problem statement was to build a machine learning model that could predict the views of the videos uploaded on the TEDx website.

About dataset:

This dataset has 4005 observations in it with 19 columns and it is a mix between categorical and numeric values.

Only 4 columns has numerical value and all others are categorical or textual data.

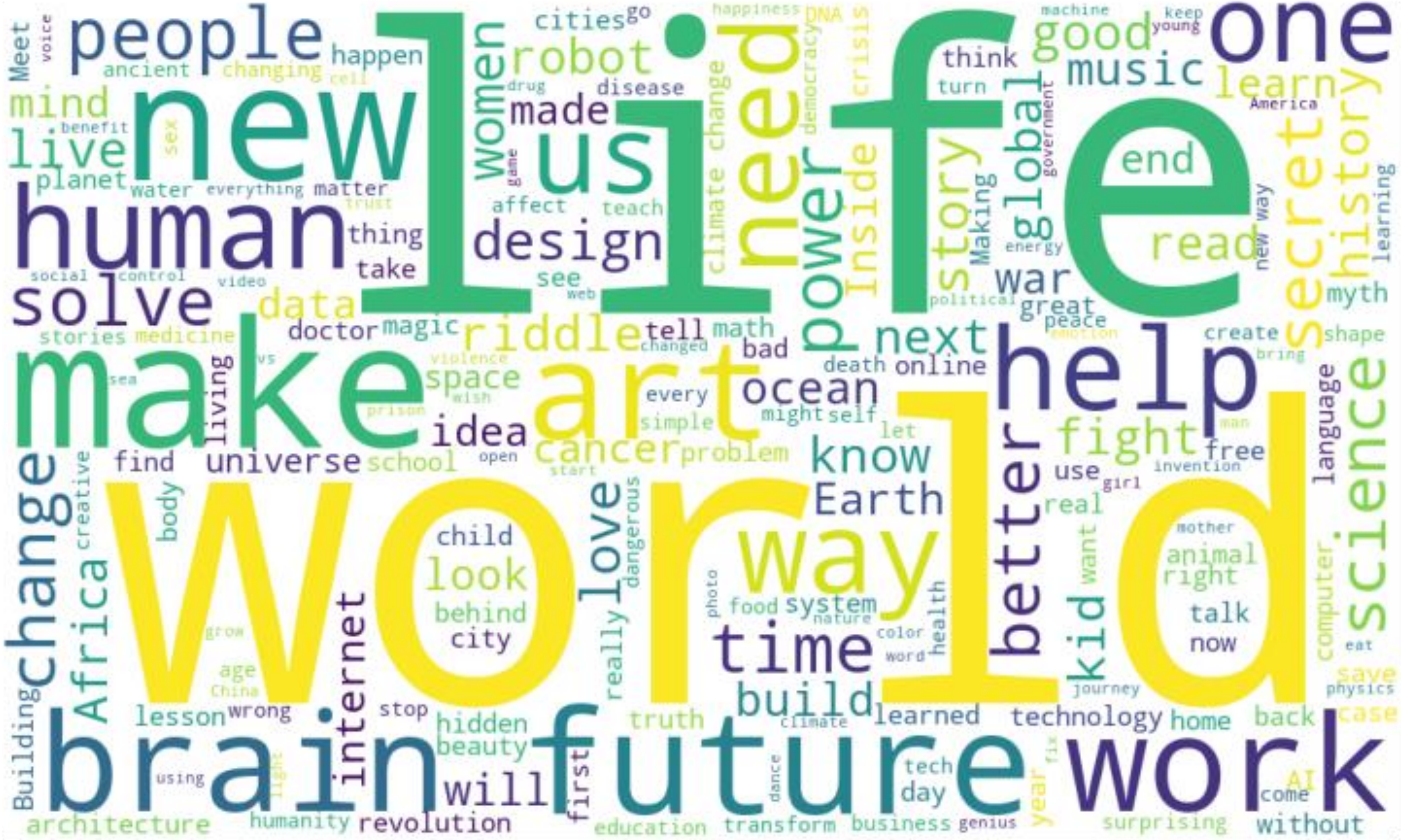
Views column is target column in dataset

Feature information in details

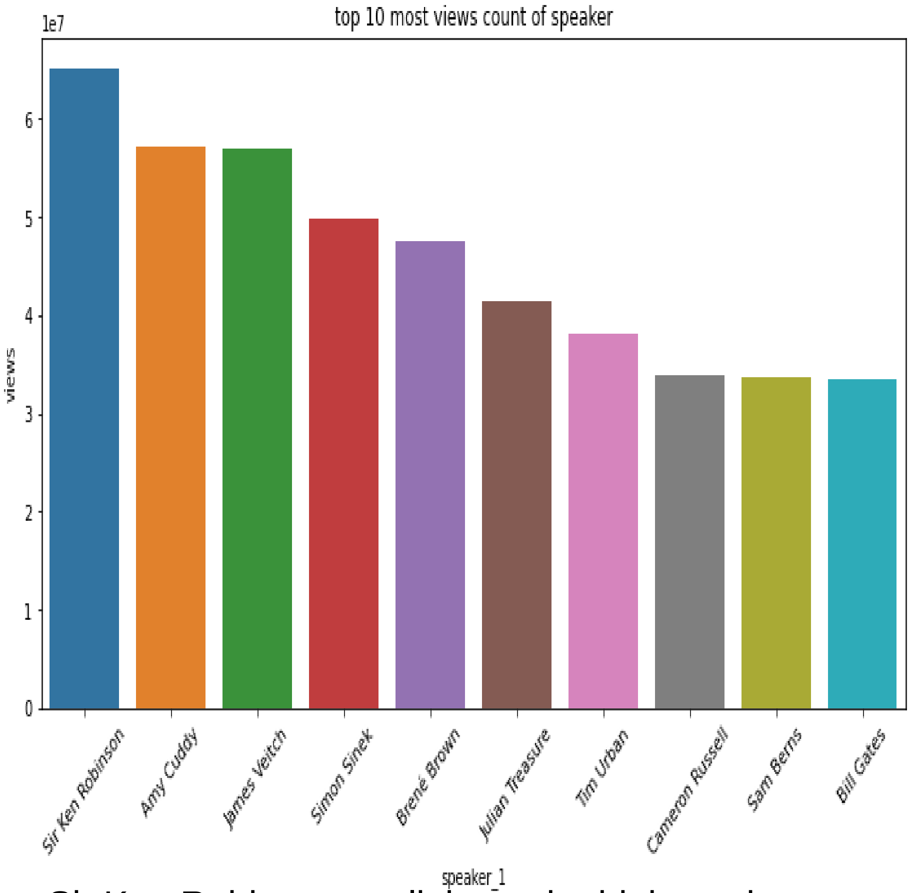
- **talk_id**: Talk identification number provided by TED
- **title**: Title of the talk
- **speaker_1**: First speaker in TED's speaker list
- **all_speakers**: Speakers in the talk
- **occupations**: Occupations of the speakers
- **about_speakers**: Blurb about each speaker
- **views**: Count of view(dependent variable)
- **recorded_date**: Date the talk was recorded
- **published_date**: Date the talk was published to TED.com
- **event**: Event or medium in which the talk was given
- **native_lang**: Language the talk was given in
- **available_lang**: All available languages (lang_code) for a talk
- **comments**: Count of comments
- **duration**: Duration in seconds
- **topics**: Related tags or topics for the talk
- **related_talks**: Related talks (key='talk_id',value='title')
- **url**: URL of the talk
- **description**: Description of the talk
- **transcript**: Full transcript of the talk

Exploratory Data Analysis

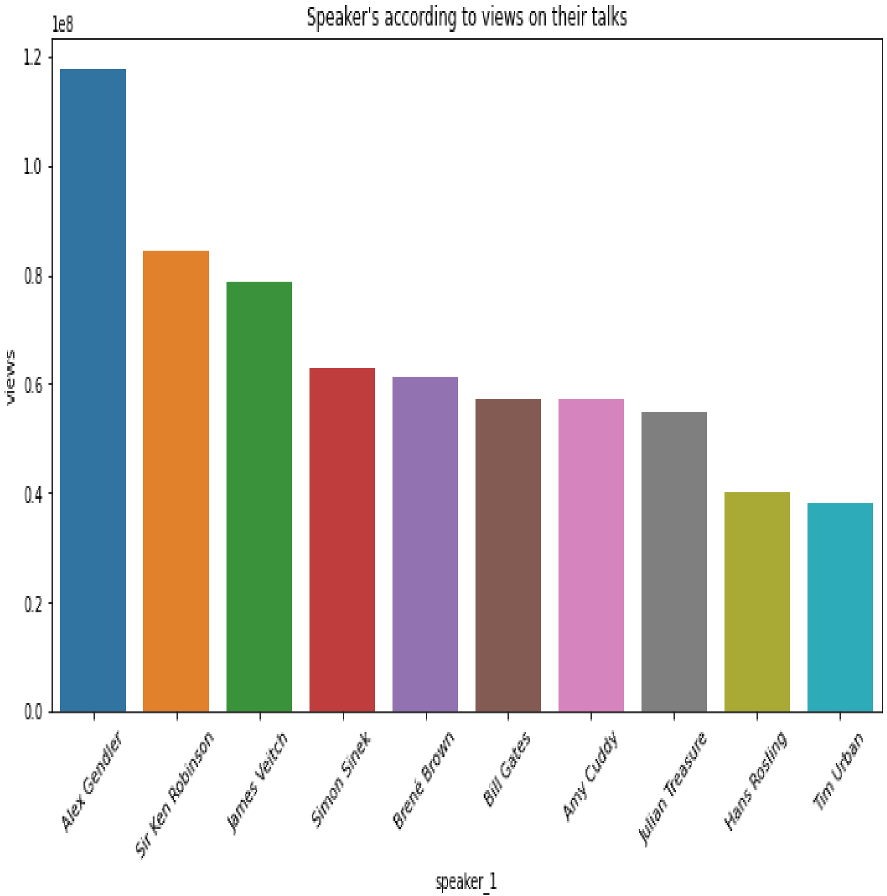
Title feature



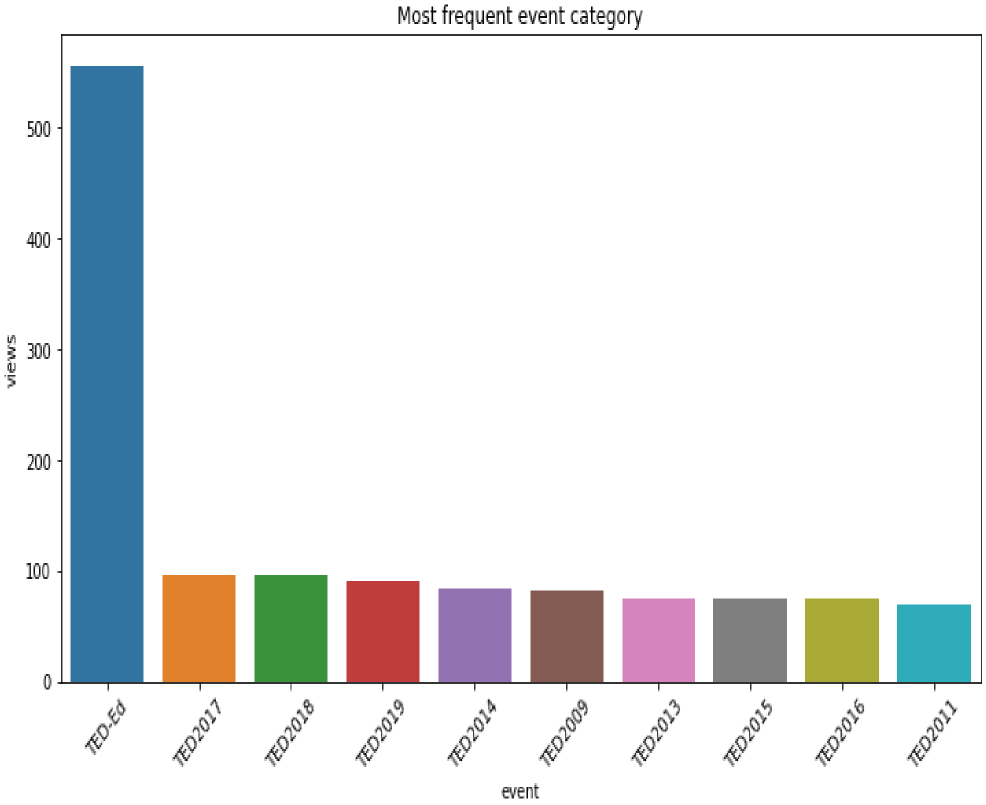
Majority of titles contains words: life, world, make, new, future, art, brain, work, human, science.



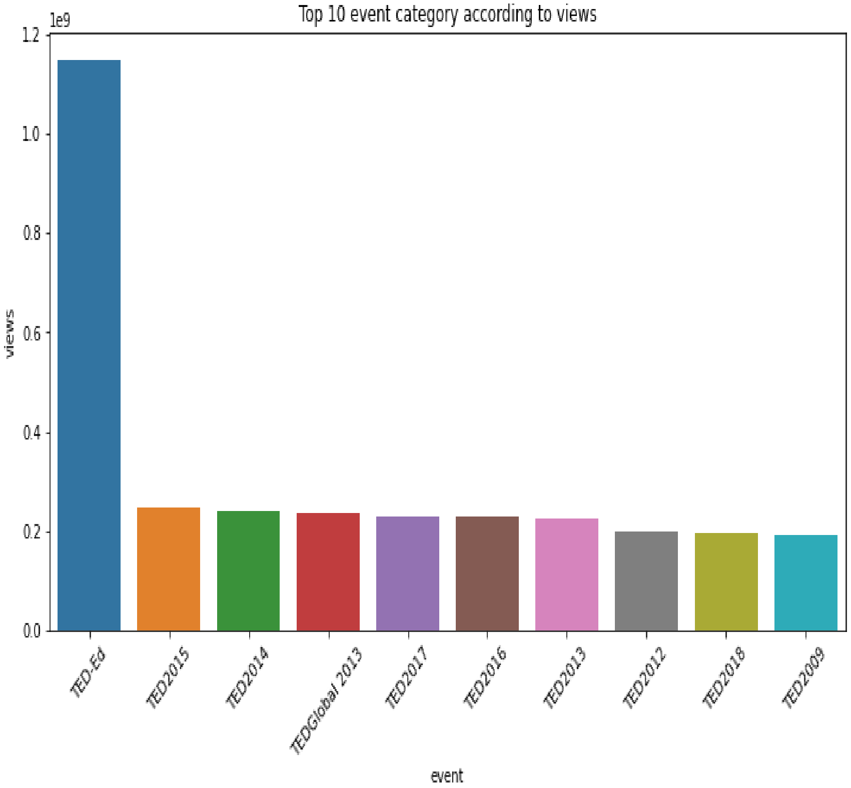
Sir Ken Robinson,s talk have the highest view count.



Alex Gendler is the most popular speaker.

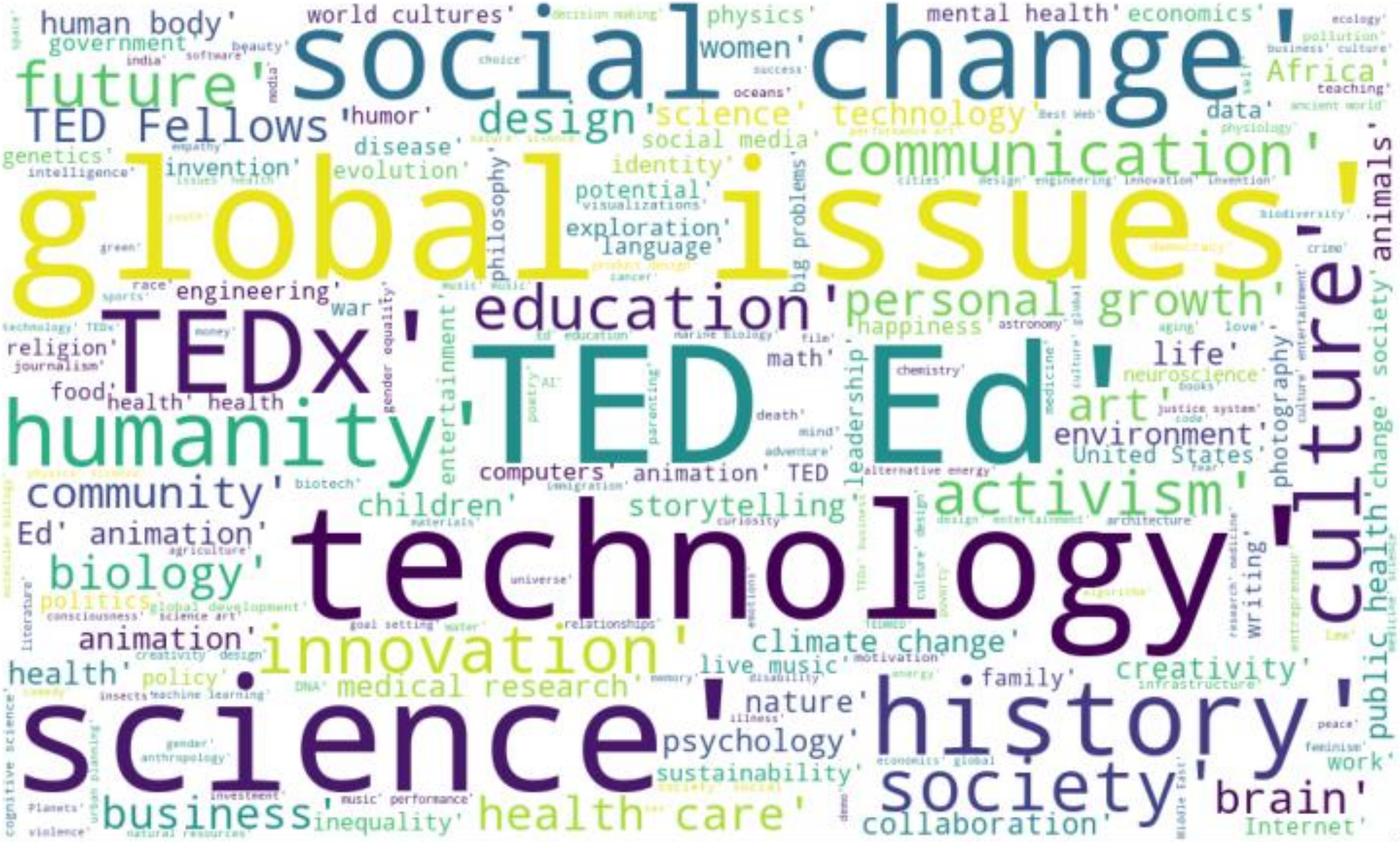


TED-Ed is the most frequent event

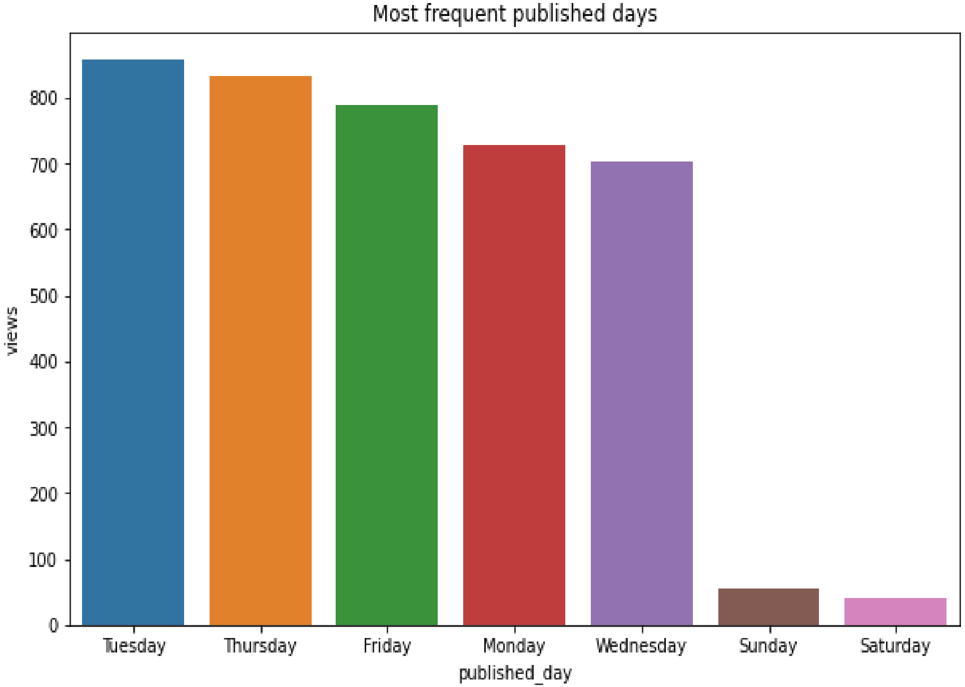


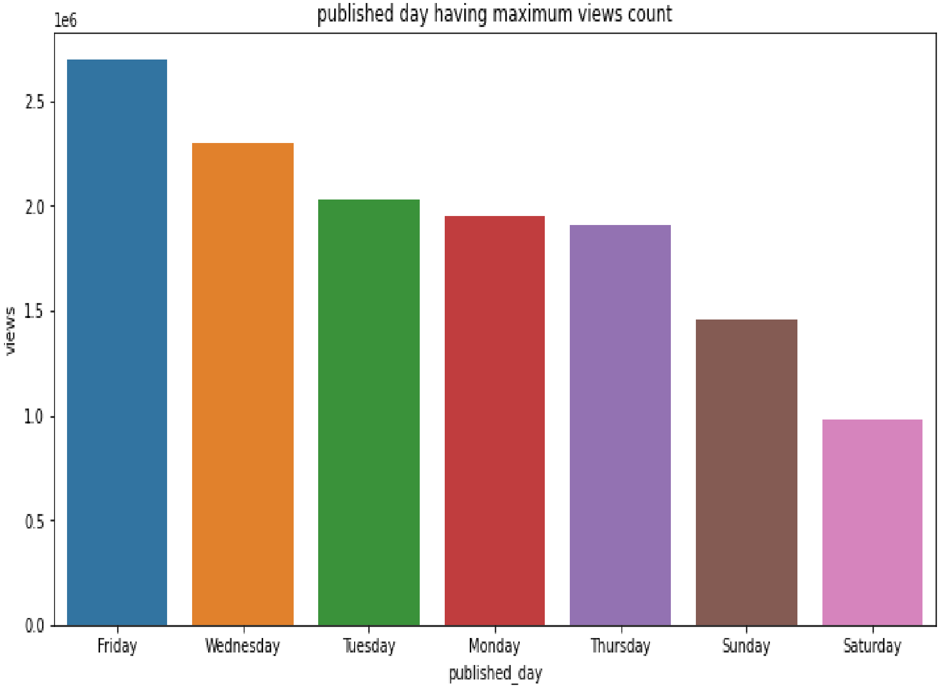
Most popular event is TED-Ed having highest number of total views.

Topics feature



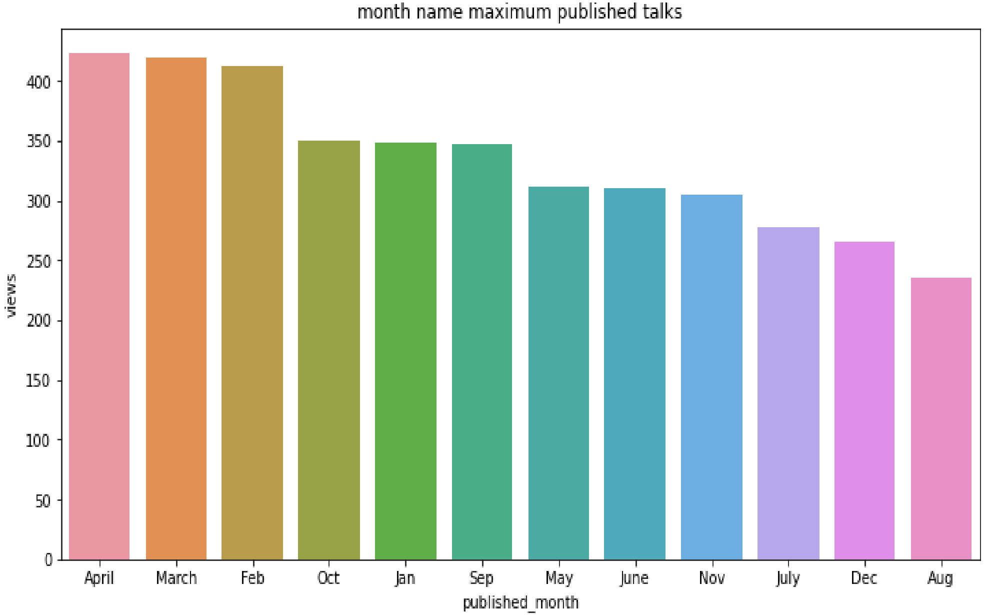
Most popular topic tags are TED Ed, technology, global issues, science, TEDx, Social change, humanity, society, activism, education, communication.

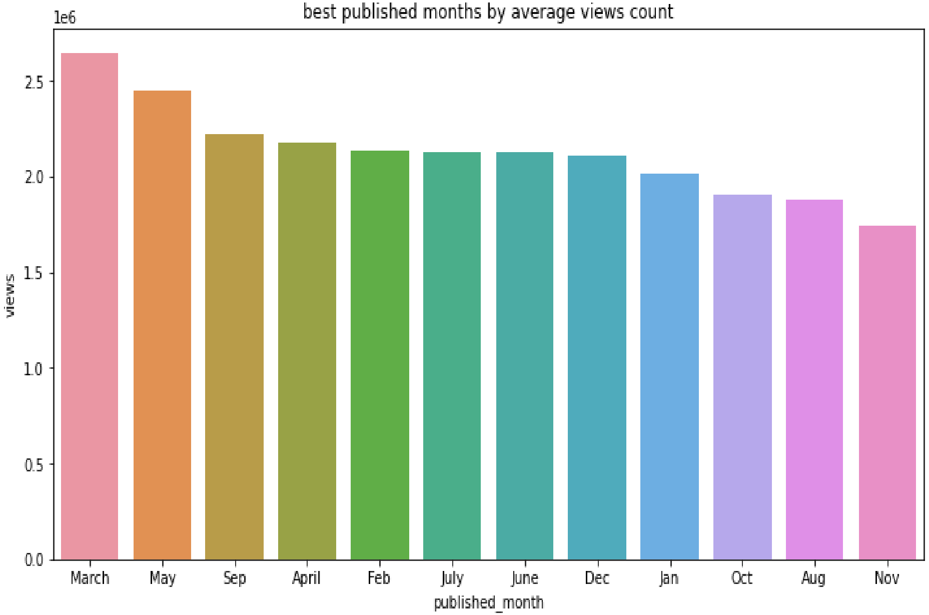




Most Talks are published on 5 days in week Tuesday, Thursday, Friday, Monday, Wednesday.

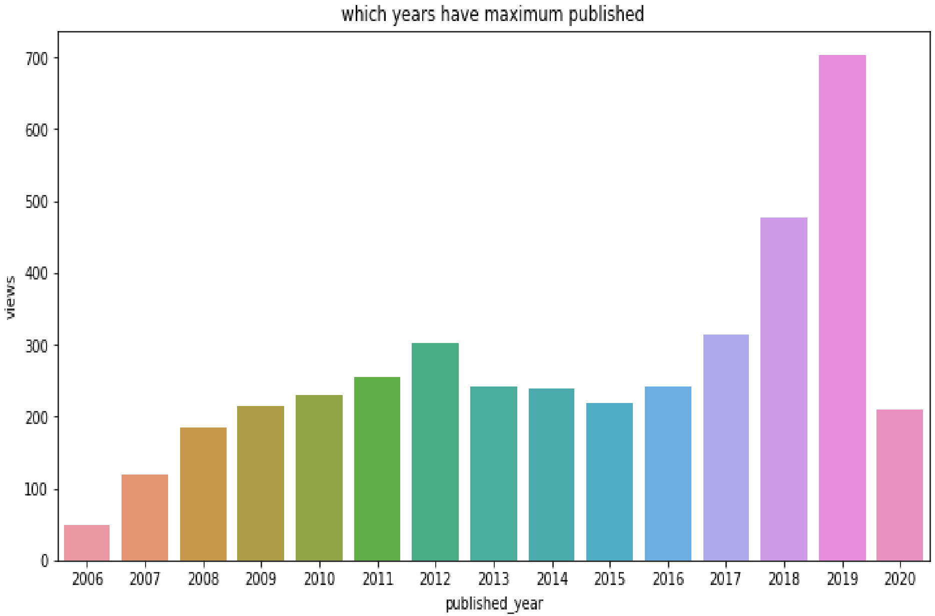
Friday published talks have more average views count

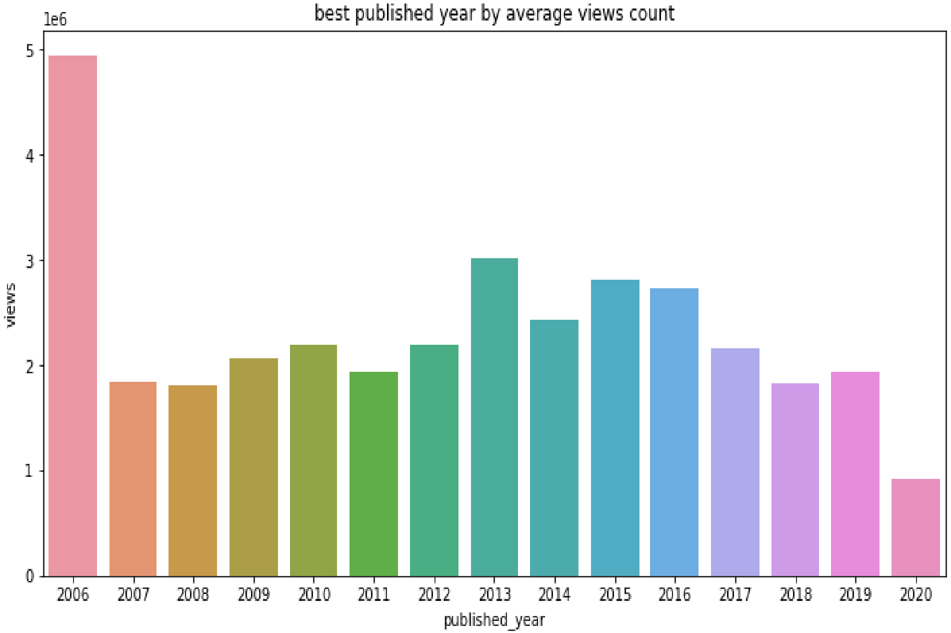




April, March, and Feb months have highest frequency of published talks

March month have more average views count





In 2019 have Published maximum talks

in 2016 have highest number of average view count

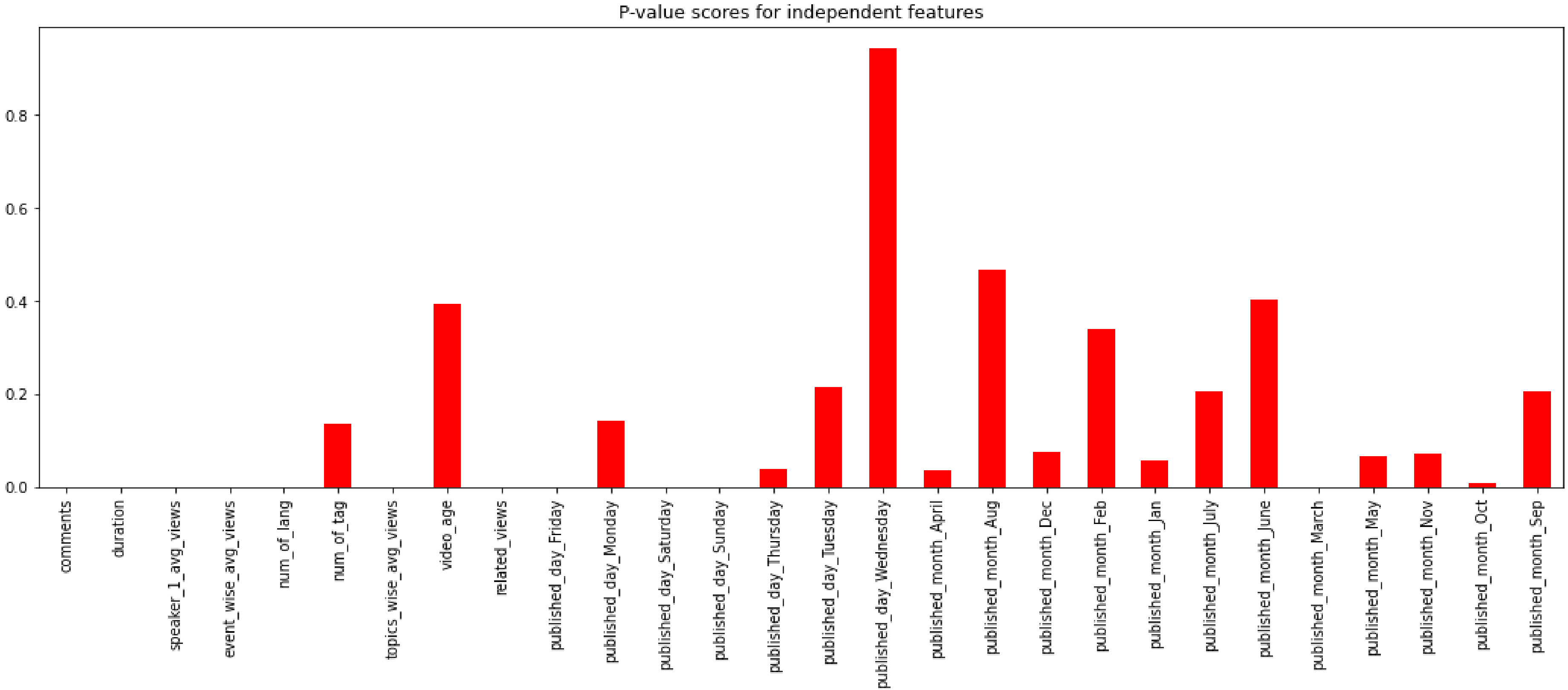
Feature engineering

- speaker_1_avg_views
- event_wise_avg_views
- num_of_lang
- topics_wise_avg_views
- video_age
- related_views
- published_day
- published_month
- published_year

Data Cleaning

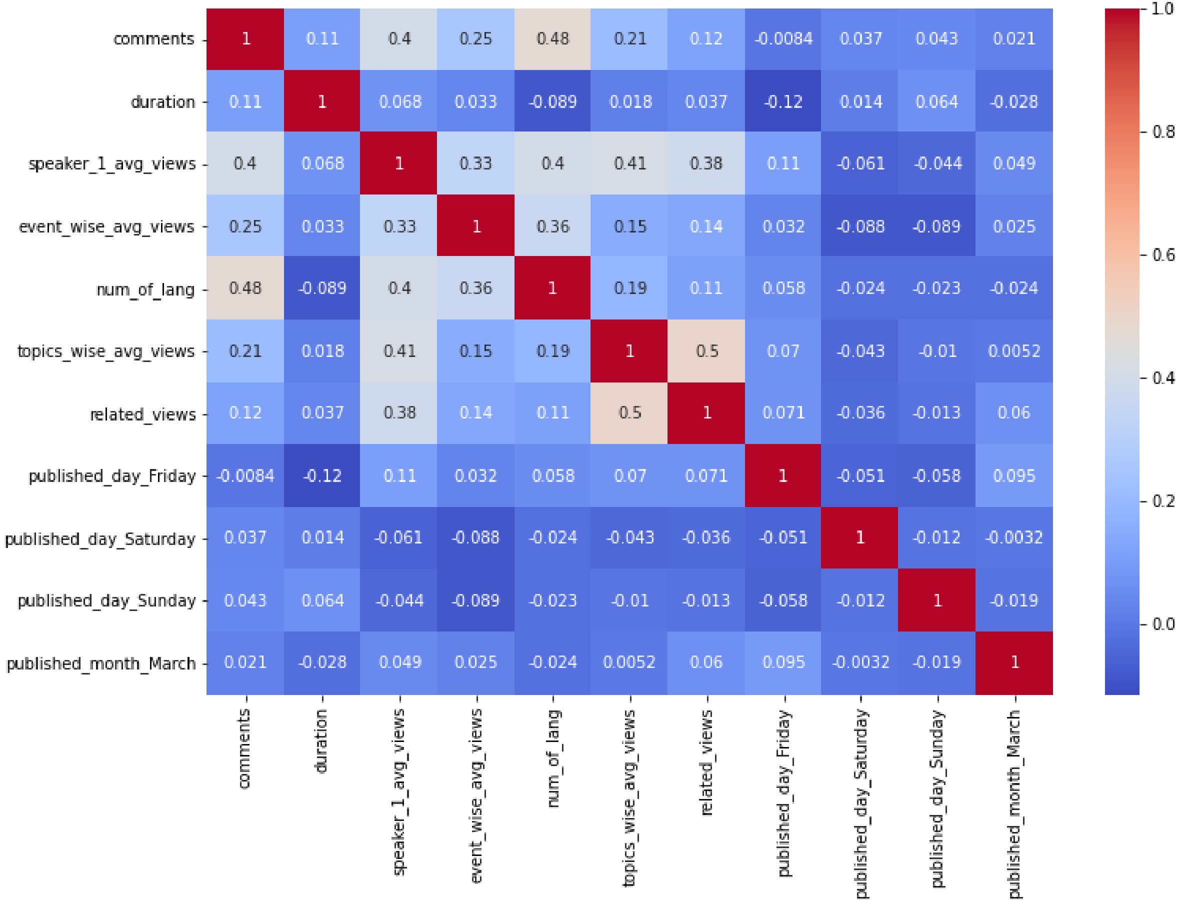
- Treated NaN value by KNN imputer
- Treated outliers by IQR with replacing extreme value

Feature selection using f_regression



we have selected only 11 important independent feature according to the p_value

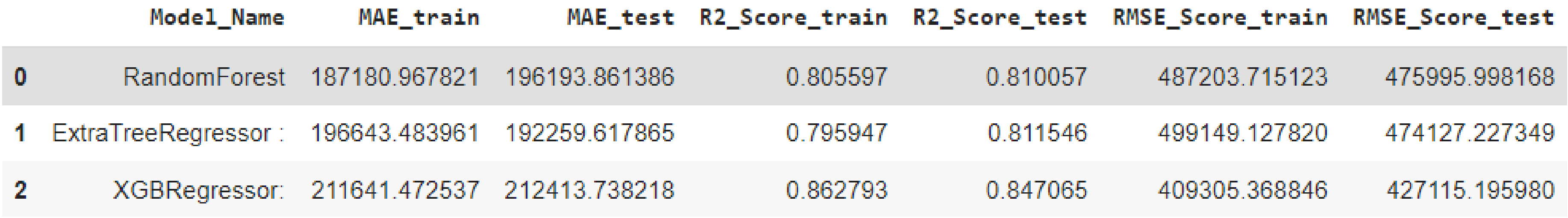
correlation matrix of selected features



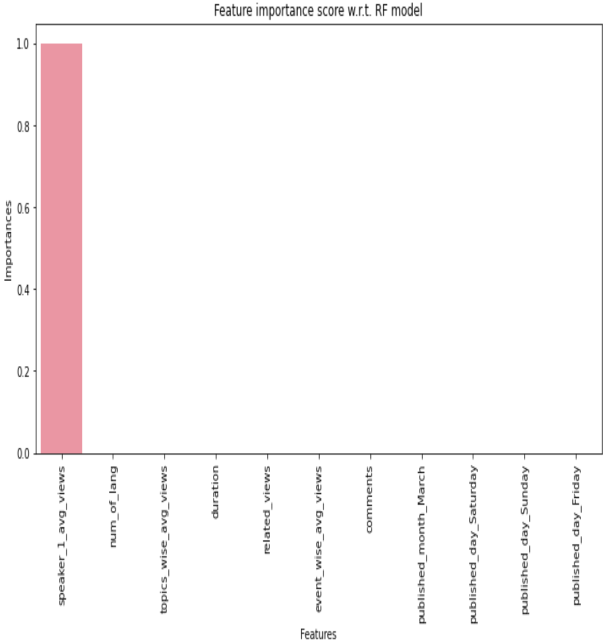
	comments	duration	speaker_1_avg_views	event_wise_avg_views	num_of_lang	topics_wise_avg_views	related_views	published_day_Friday	published_day_Saturday	published_day_Sunday	published_month_March
comments	1	0.11	0.4	0.25	0.48	0.21	0.12	-0.0084	0.037	0.043	0.021
duration	0.11	1	0.068	0.033	-0.089	0.018	0.037	-0.12	0.014	0.064	-0.028
speaker_1_avg_views	0.4	0.068	1	0.33	0.4	0.41	0.38	0.11	-0.061	-0.044	0.049
event_wise_avg_views	0.25	0.033	0.33	1	0.36	0.15	0.14	0.032	-0.088	-0.089	0.025
num_of_lang	0.48	-0.089	0.4	0.36	1	0.19	0.11	0.058	-0.024	-0.023	-0.024
topics_wise_avg_views	0.21	0.018	0.41	0.15	0.19	1	0.5	0.07	-0.043	-0.01	0.0052
related_views	0.12	0.037	0.38	0.14	0.11	0.5	1	0.071	-0.036	-0.013	0.06
published_day_Friday	-0.0084	-0.12	0.11	0.032	0.058	0.07	0.071	1	-0.051	-0.058	0.095
published_day_Saturday	0.037	0.014	-0.061	-0.088	-0.024	-0.043	-0.036	-0.051	1	-0.012	-0.0032
published_day_Sunday	0.043	0.064	-0.044	-0.089	-0.023	-0.01	-0.013	-0.058	-0.012	1	-0.019
published_month_March	0.021	-0.028	0.049	0.025	-0.024	0.0052	0.06	0.095	-0.0032	-0.019	1

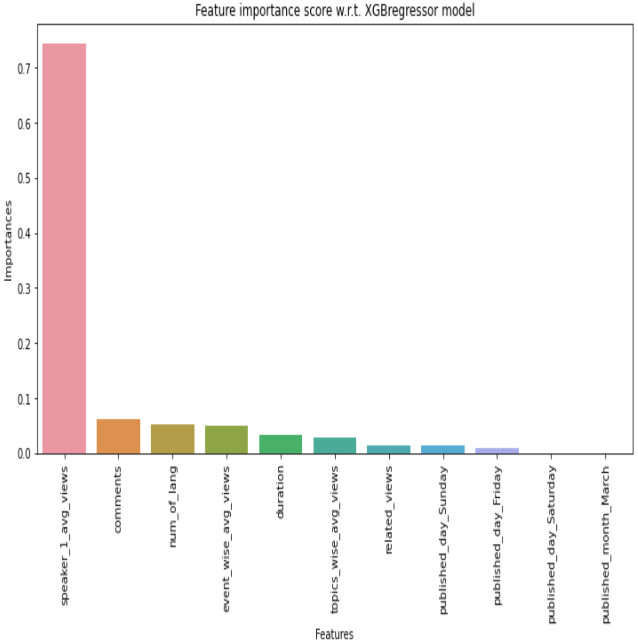
Regression models using

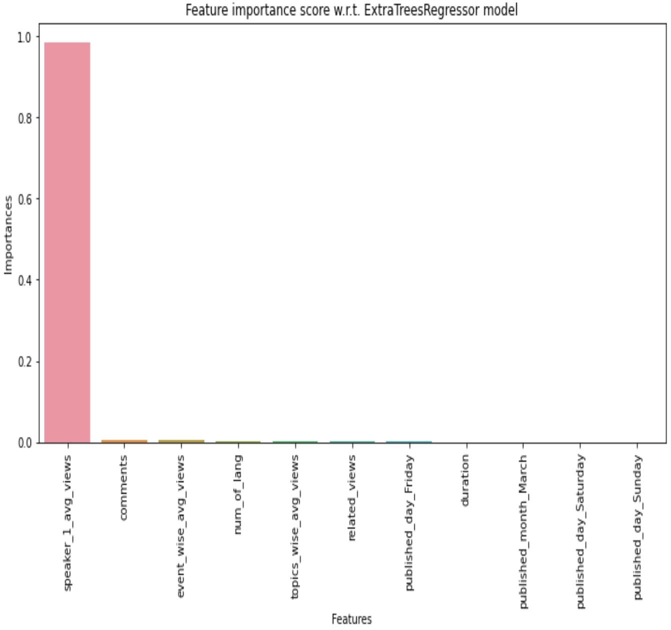
1. Random Forest Regressor
2. XGboost Regressor
3. ExtraTrees Regressor



	Model_Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	RandomForest	187180.967821	196193.861386	0.805597	0.810057	487203.715123	475995.998168
1	ExtraTreeRegressor :	196643.483961	192259.617865	0.795947	0.811546	499149.127820	474127.227349
2	XGBRegressor:	211641.472537	212413.738218	0.862793	0.847065	409305.368846	427115.195980

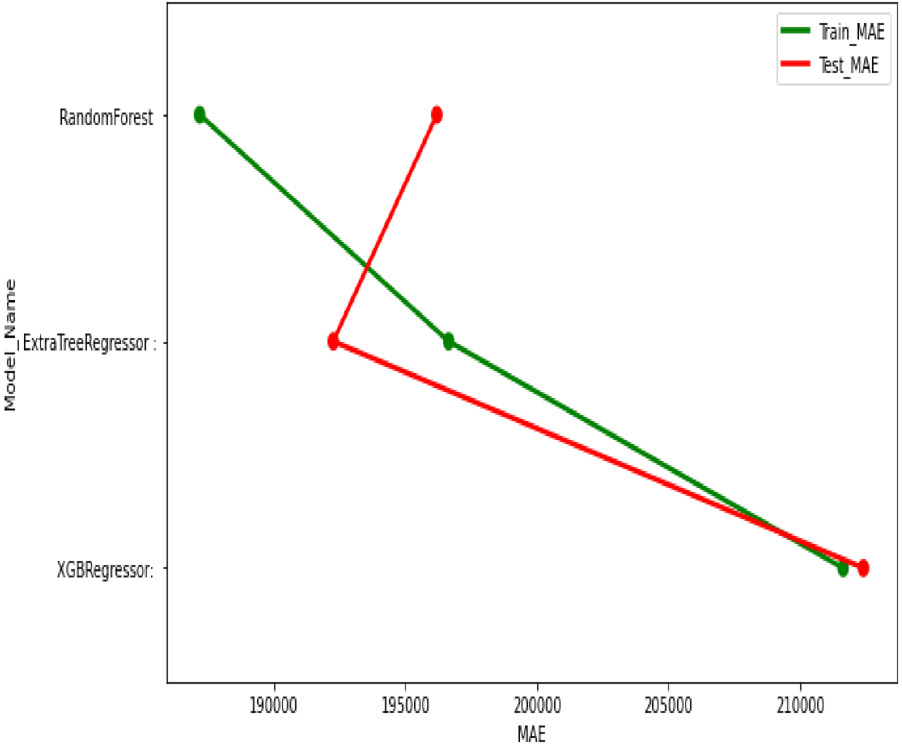


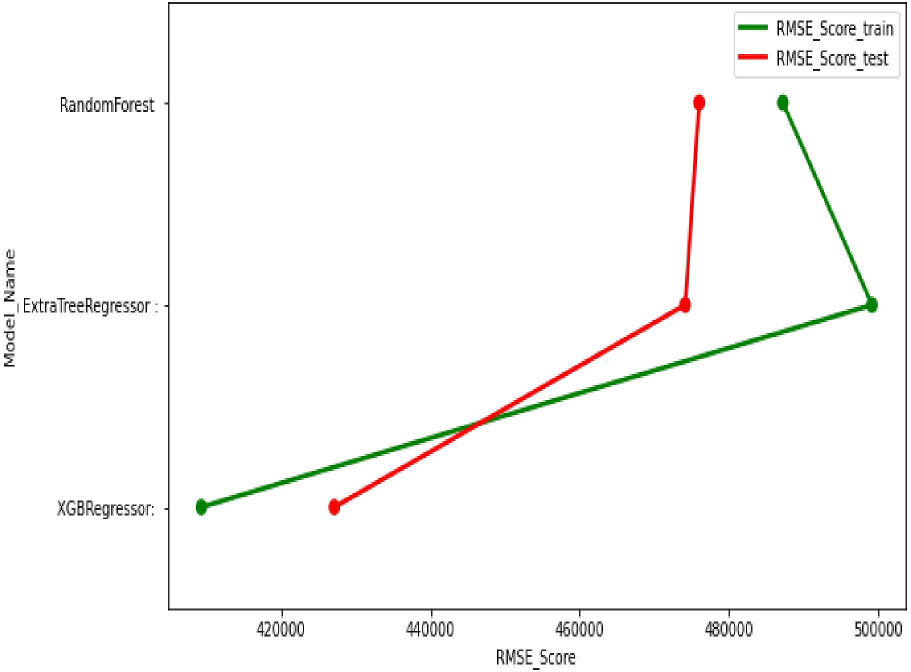




speaker1_avg_views is the most important feature in all three Models

Comparison of all Model





Selection of the model

We choose MAE and not RMSE as the deciding factor of our model selection because of the following reasons:

RMSE is heavily influenced by outliers as in the higher the values get the more the RMSE increases

MAE doesn’t increase with outliers, MAE is linear and RMSE is quadratically increasing

The best performing regressor model for this dataset is Random Forest Regressor on the basis of MAE

Conclusion

We build a predictive model, which could help TED in predicting the views of the talks uploaded on the TEDx website.

In all the features speaker_wise_avg_views is most important this implies that speakers are directly impacting the views.

TED can increase their views and popularity by increasing videos on sections like Technology and Science.

Increasing the number of languages the talk is available in, increases the views of the TED talks.

Thank you!

Presented by: Rajesh Kumar Patel