Capstone Project - 2

TED Talk Views Prediction

Team Member

Aamir Sohail

Content

1. Problem Statement
2. Data Summary
3. EDA on features
4. Feature Engineering
5. Feature Selection
6. Models used
7. Choice of model and reason of choosing
8. Challenges
9. Conclusion

Problem Statement

- TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages Founded in 1984 by Richard Salman as a nonprofit organization that aimed at bringing experts from the fields of Technology, Entertainment and Design together.

- TED Conferences have gone on to become the Mecca of ideas from virtually all walks of life.

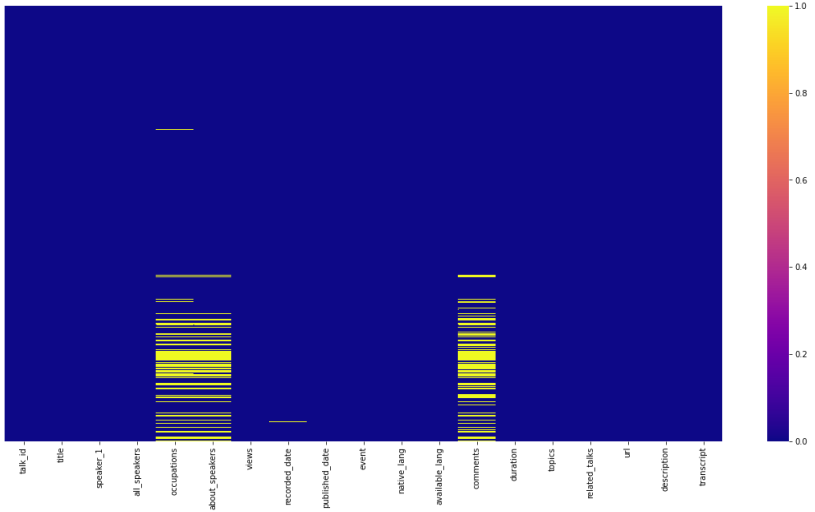
- The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website.

Data Summary

- The dataset name is data_ted_talks and it contains above 4000 talks.
- The dataset contains 4005 rows and 19 columns.
- It contains only 4 numerical variables, others are object dtype.
- Features are: ‘talk_id’, ‘title’, ‘speaker_1’, ’all_speakers’, ‘occupations’, ‘about_speakers’, ‘recorded_date’, ‘published_date’, ‘event’, ‘native_lang’, available_lang’, ‘comments’, ‘duration’, ‘topics’, ‘related_talks’, ‘url’, ‘description’, ‘transcript’.
- Target variable is ‘views’.

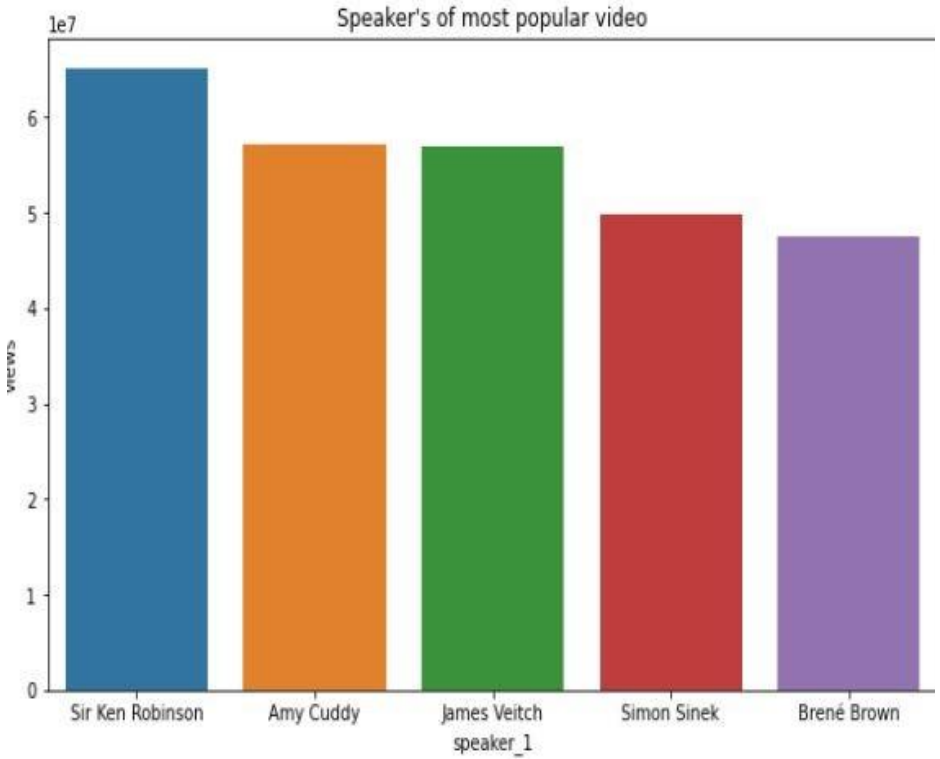
EDA on Features

Missing Data and Outliers

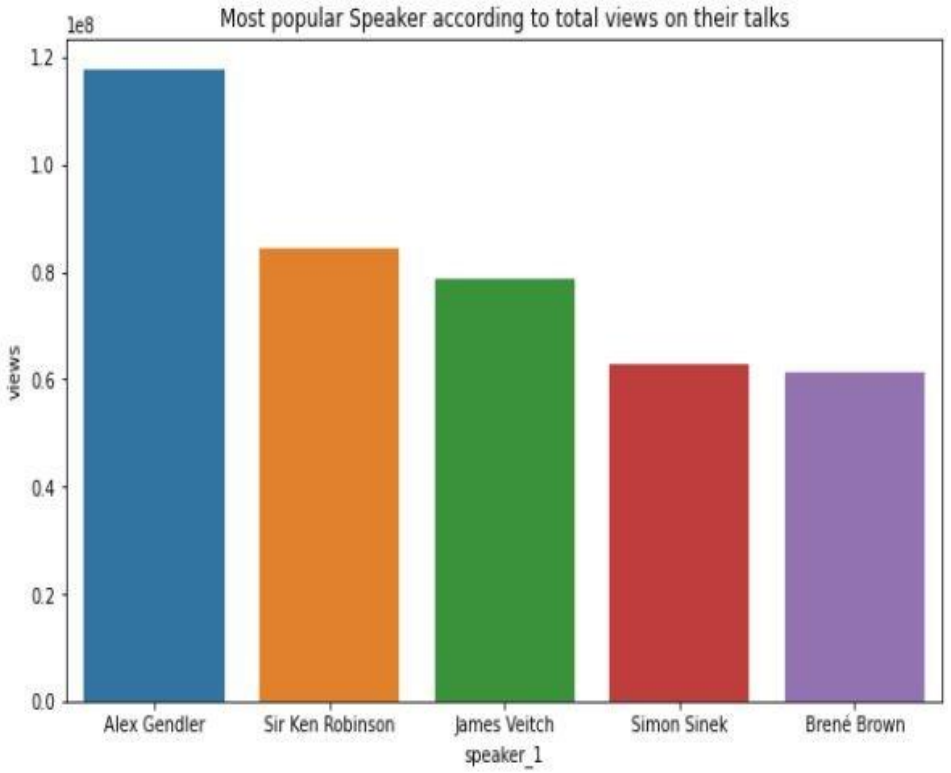


- KNN imputation for numerical features.
- Outliers treatment in numerical features.
- Nan values of Categorical features replaced with ‘Unknown’ category.

Speakers v/s Views

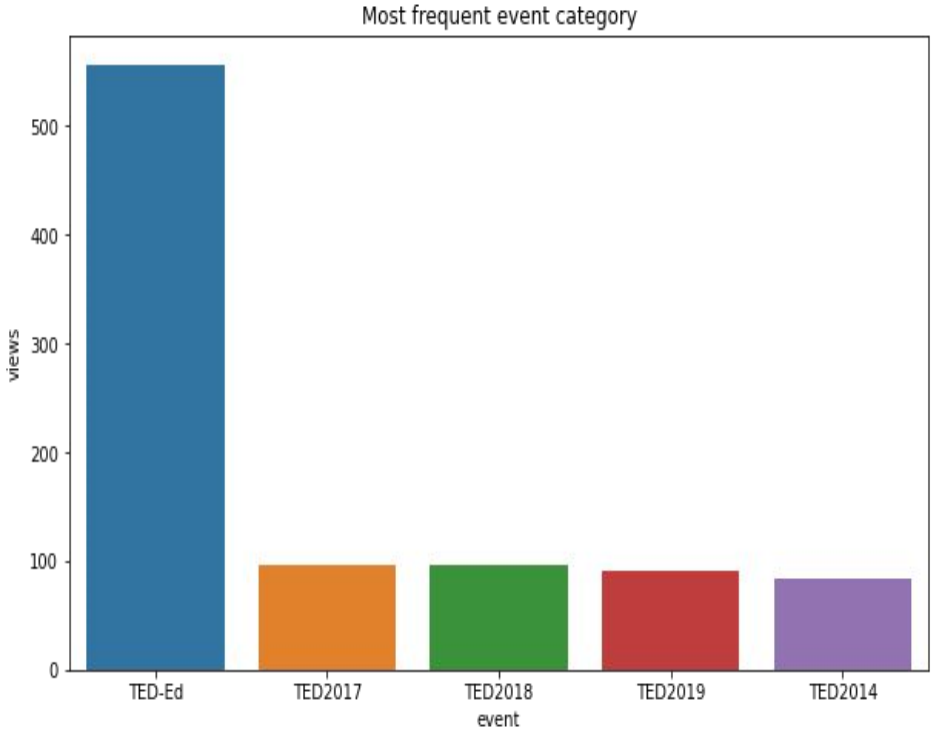


Speakers of most popular video

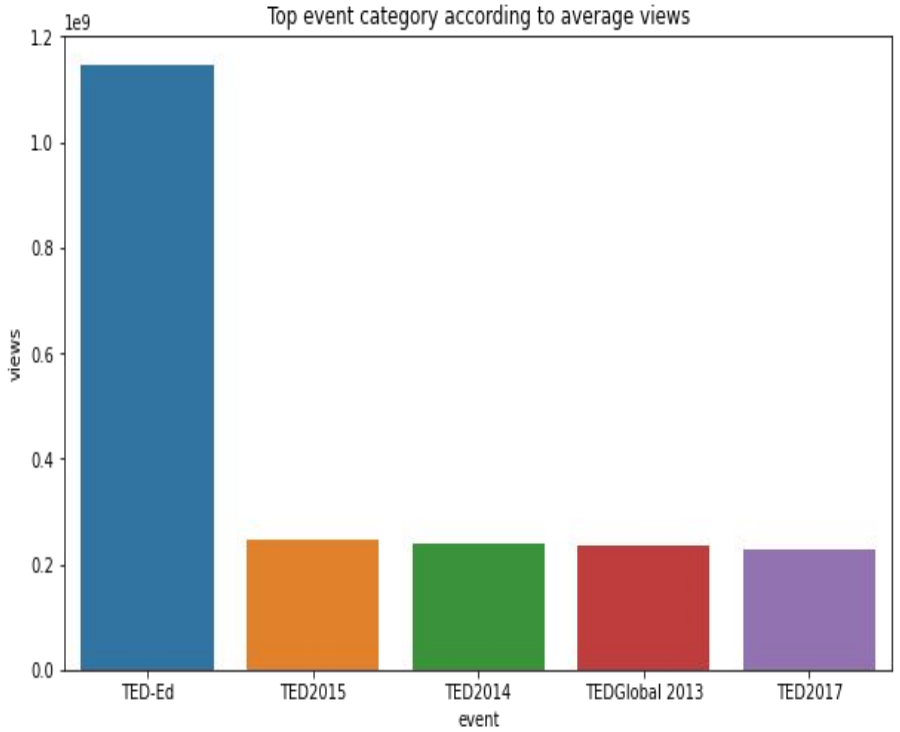


Top Speakers by total Views

Event v/s Views

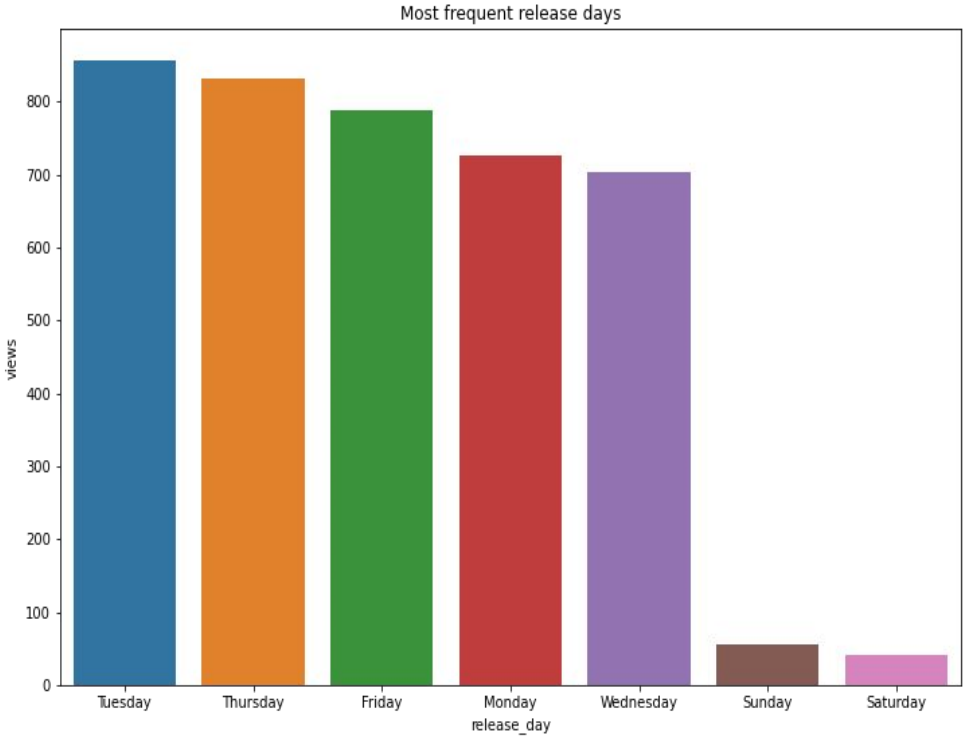


Most frequent event category

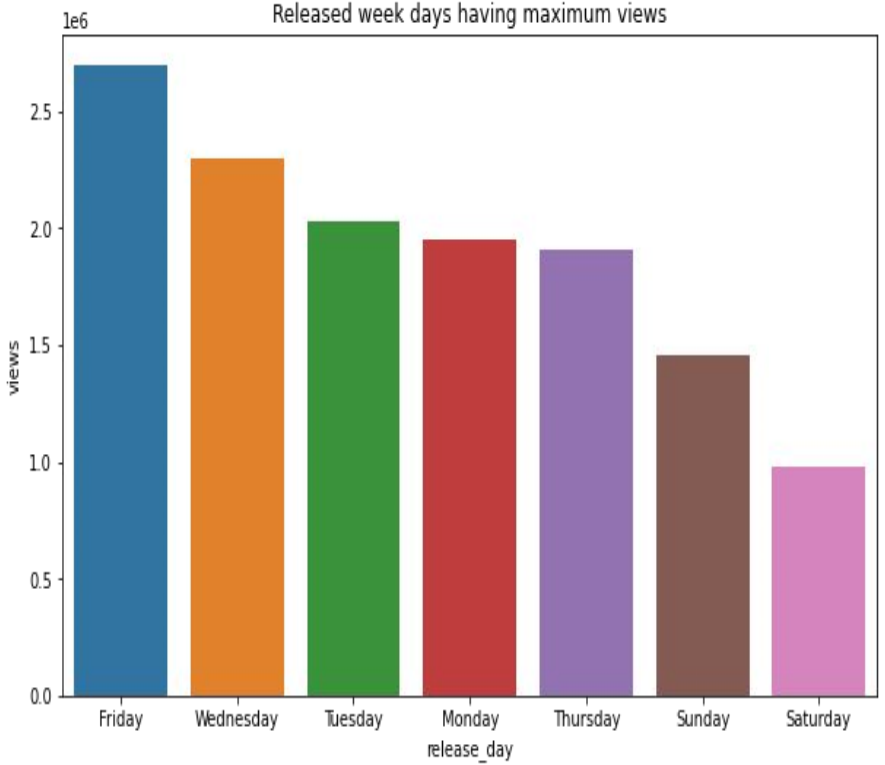


Top events by average views

Published Day v/s Views

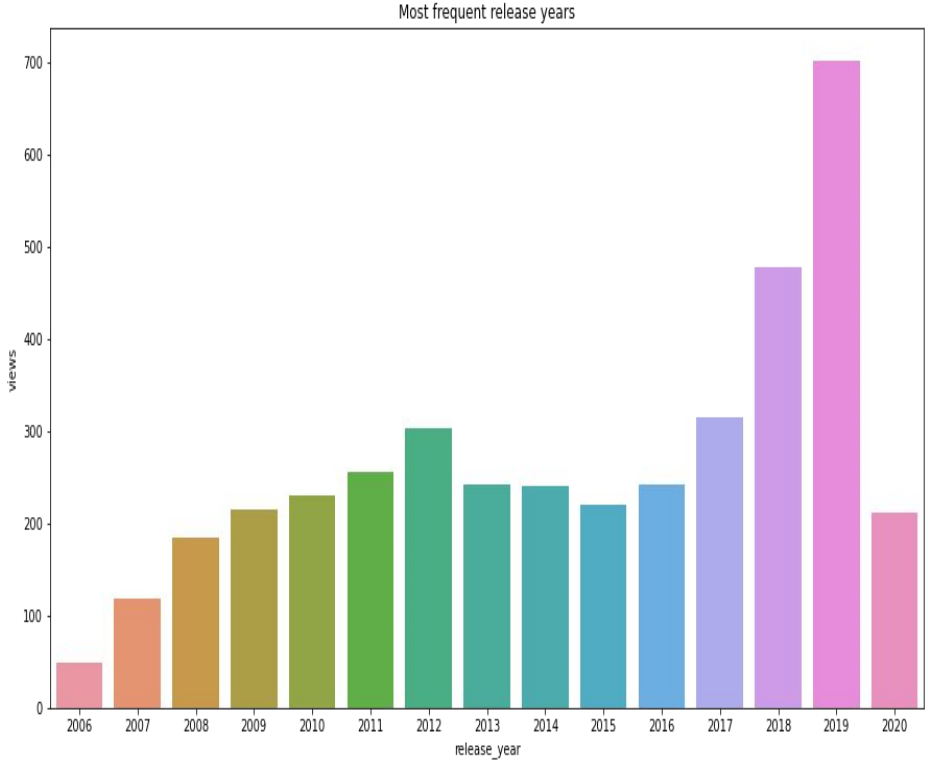


Frequent released days

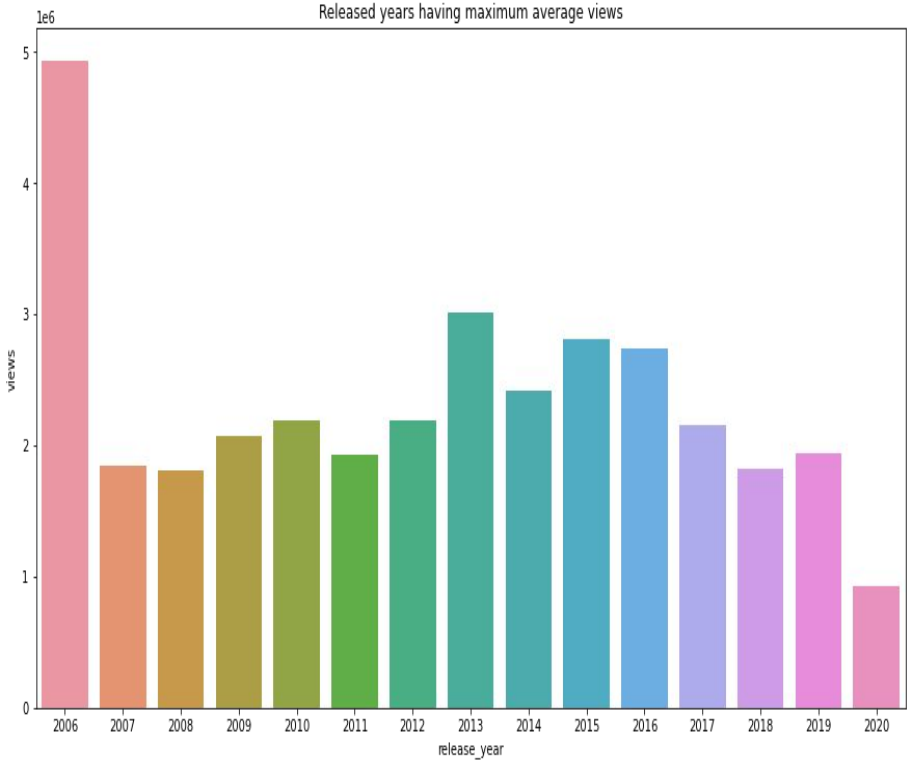


Released days by average views

Published Year v/s Views

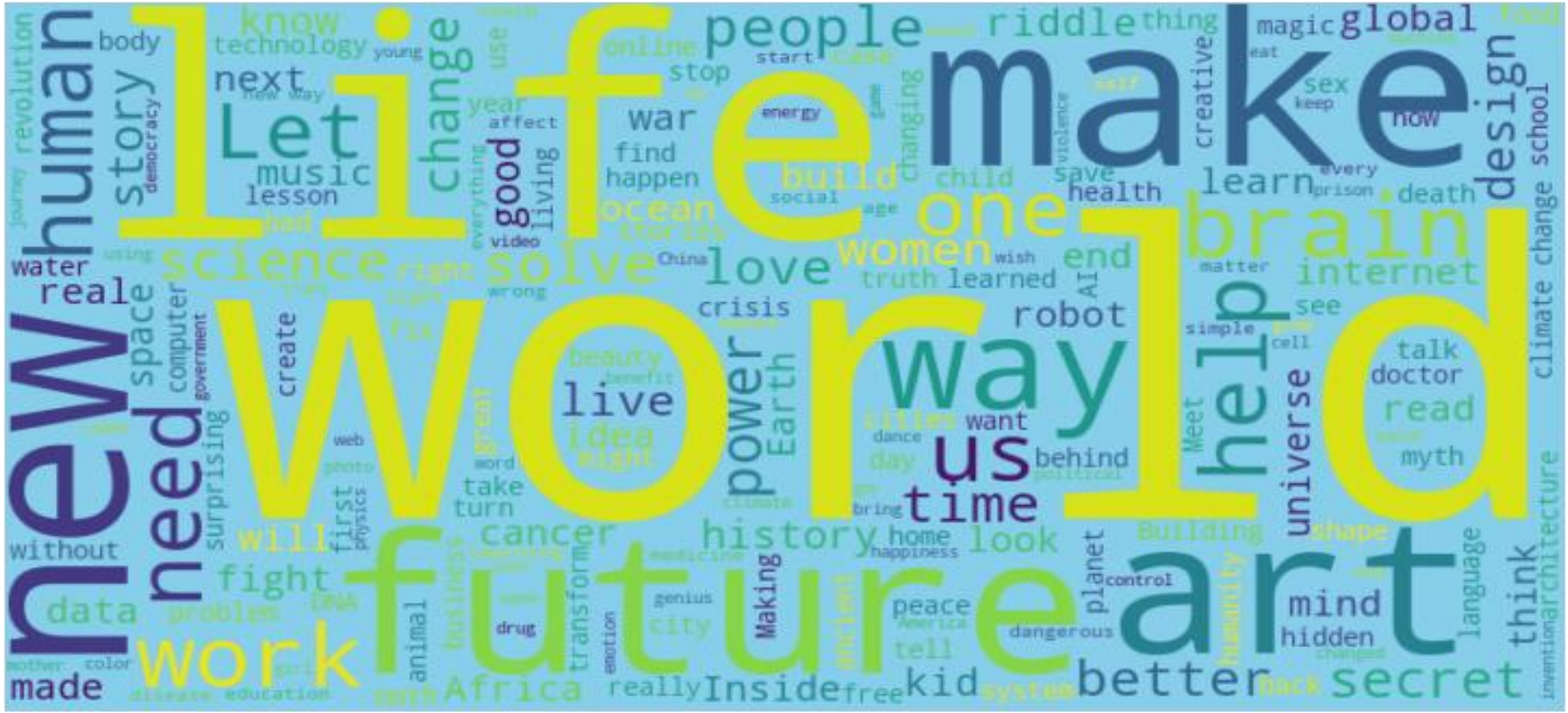


Most frequent released year

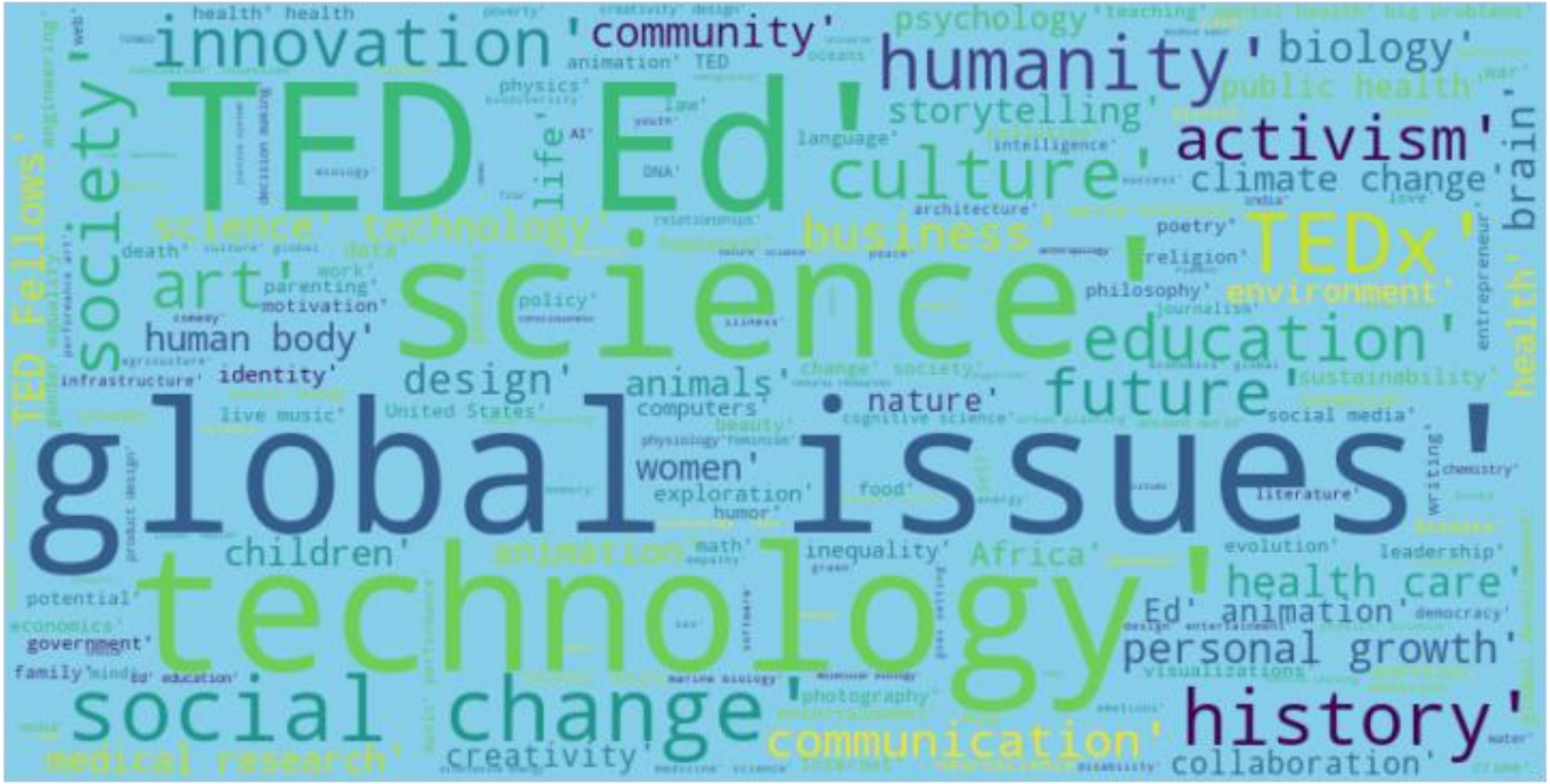


Released year with max average views

Most popular Titles



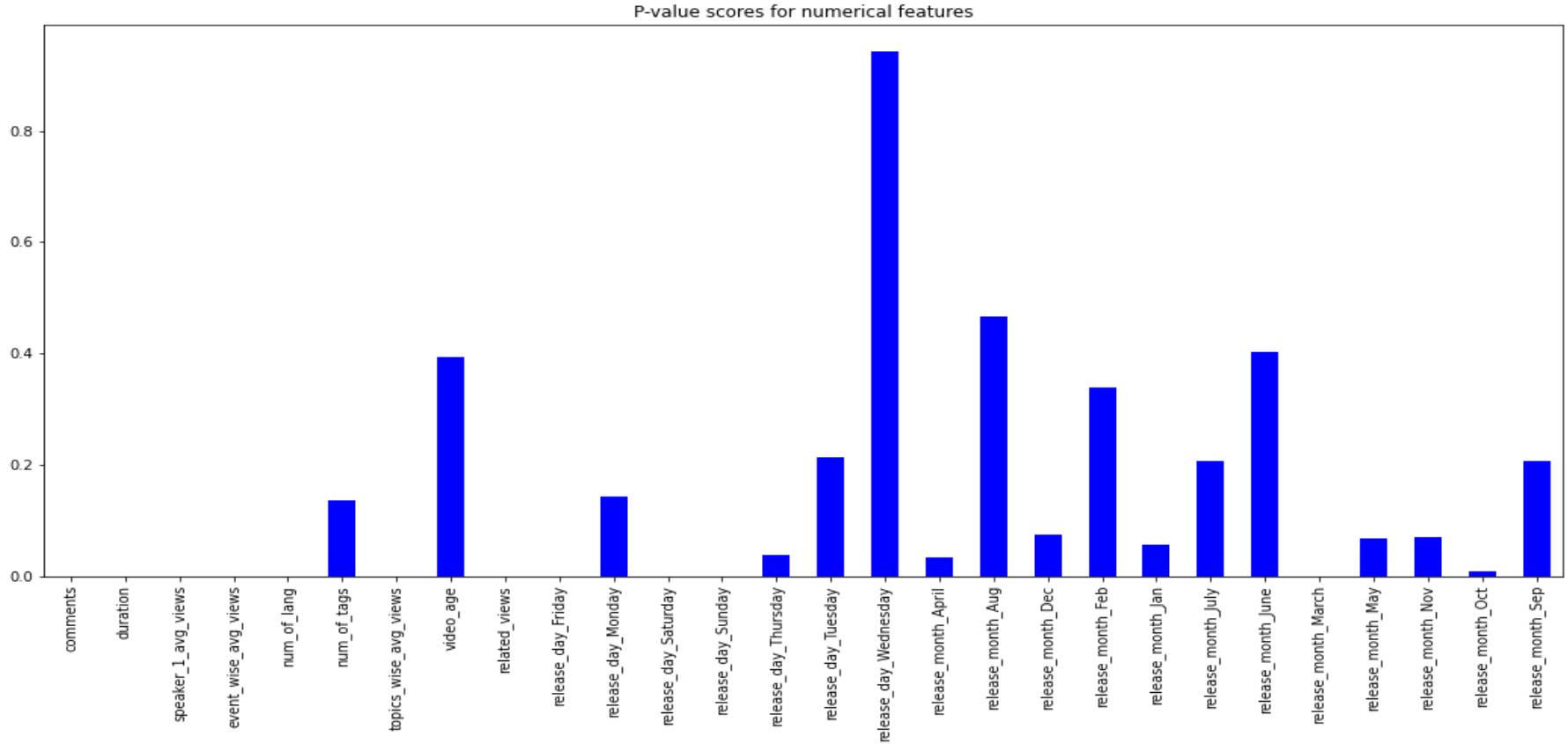
Most popular Topics



Feature Engineering

- Speaker_1_avg_views
- Event_wise_avg_views
- Related_views
- Topic_wise_avg_views
- Num_of_languages
- Num_of_tags
- Release_day
- Release_month
- Video_age

Features selection(f regression)

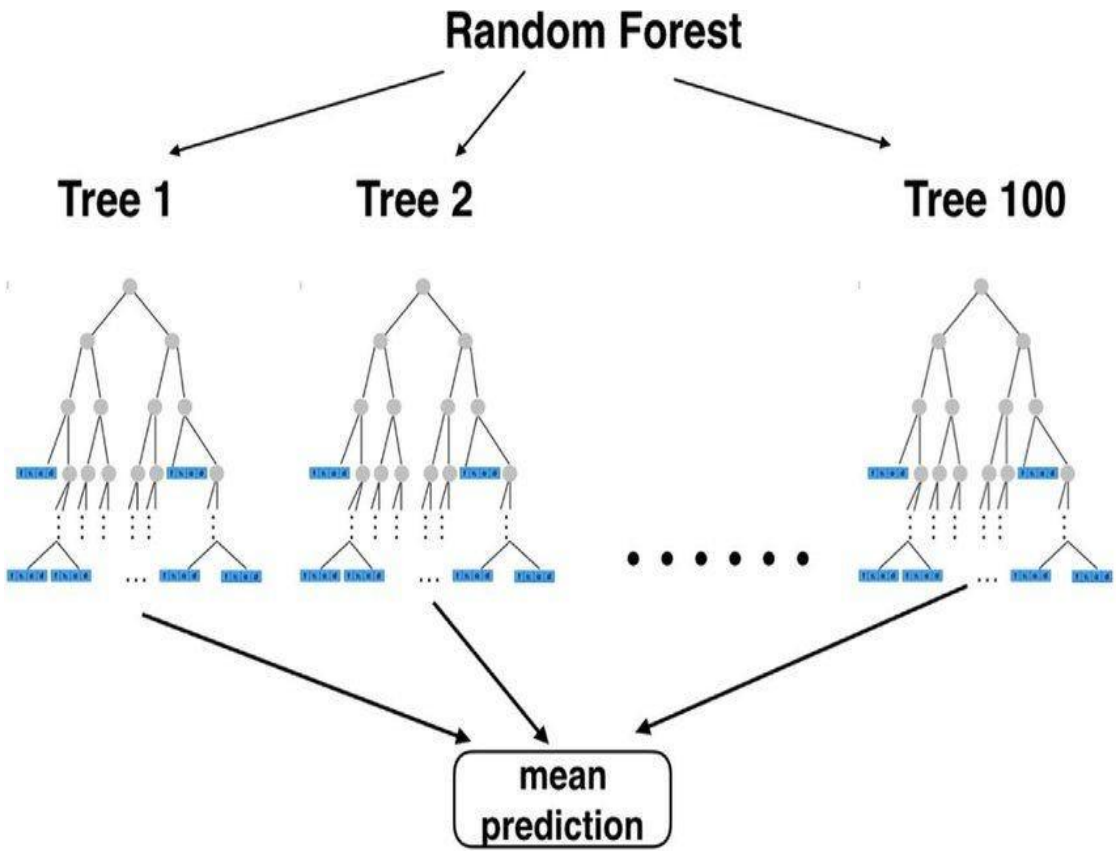


Models used

- Random Forest Regressor
- Extra Trees Regressor
- XGBoost Regressor

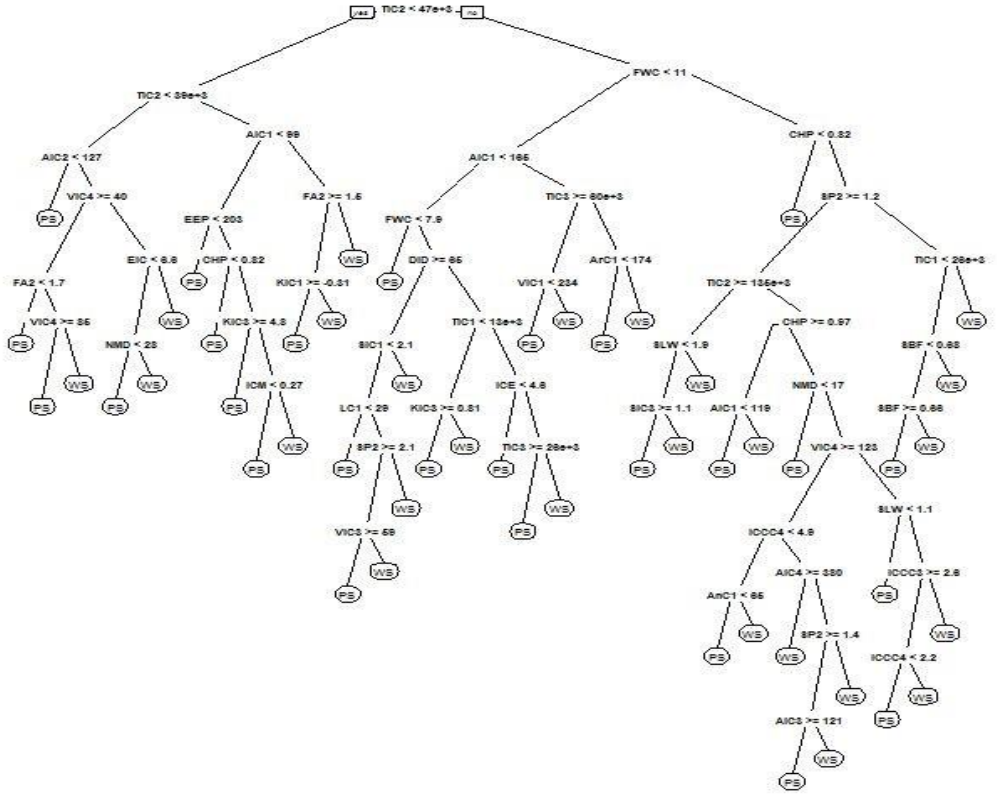
Random Forest Regressor

- Criterion = MAE
- MAE train = 186638.460
- MAE test = 192011.316
- RMSE train = 485112.674
- RMSE test = 489031.640
- R_Square for train = 0.80
- R_Square for test = 0.80



Extra Trees Regressor

- Criterion = MAE
- MAE train = 197961.601
- MAE test = 195381.403
- RMSE train = 497703.788
- RMSE test = 484983.269
- R_Square for train = 0.79
- R_Square for test = 0.80



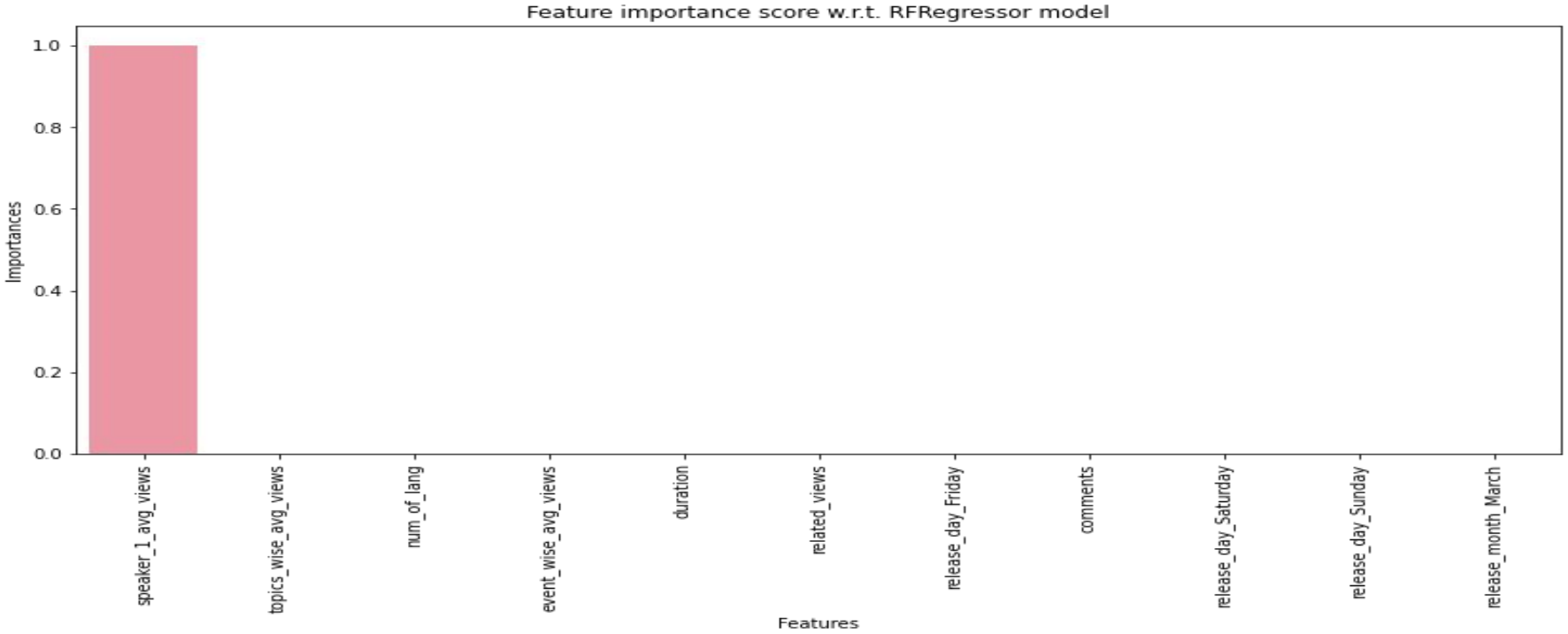
XGBoost Regressor

- Criterion = MAE
- MAE train = 211051.556 ●

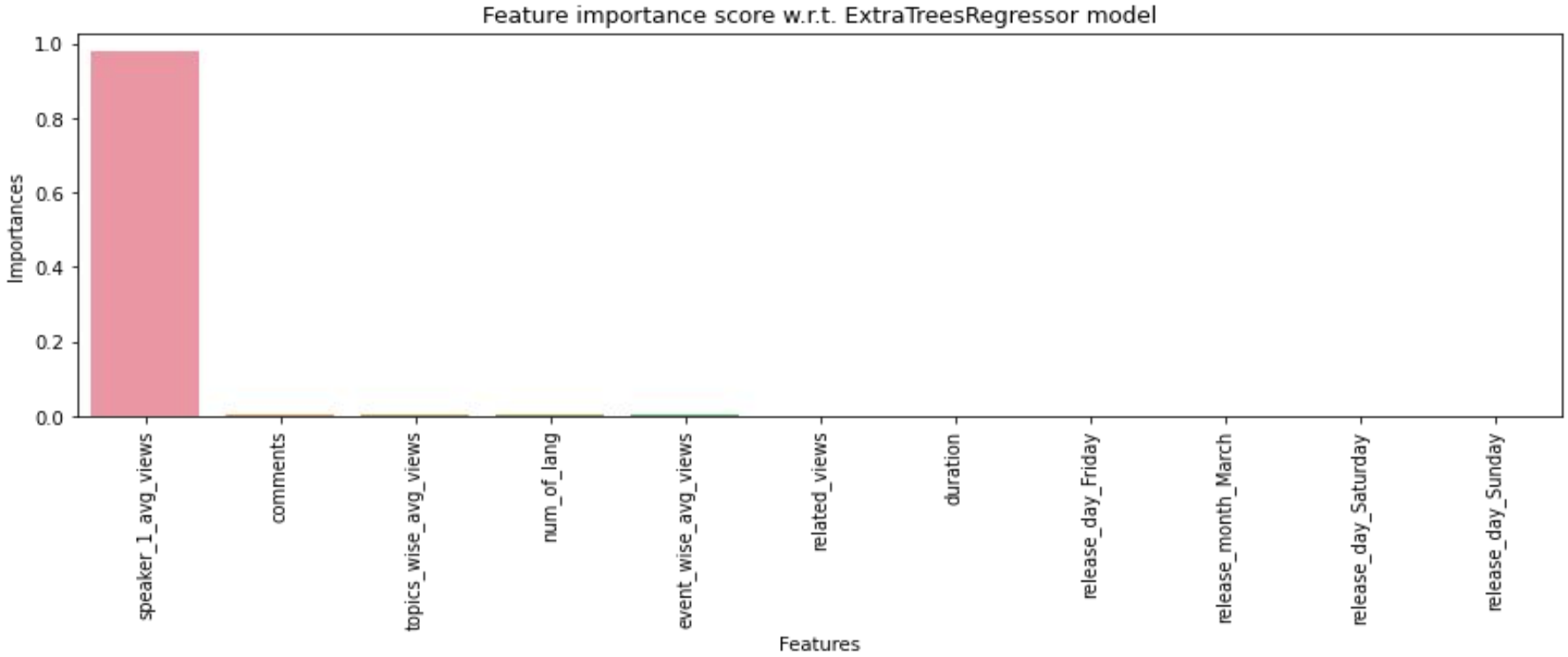
  MAE test = 228812.768
- RMSE train = 403273.960 ●

  RMSE test = 451028.530
- R_Square for train = 0.86
- R_Square for test = 0.83

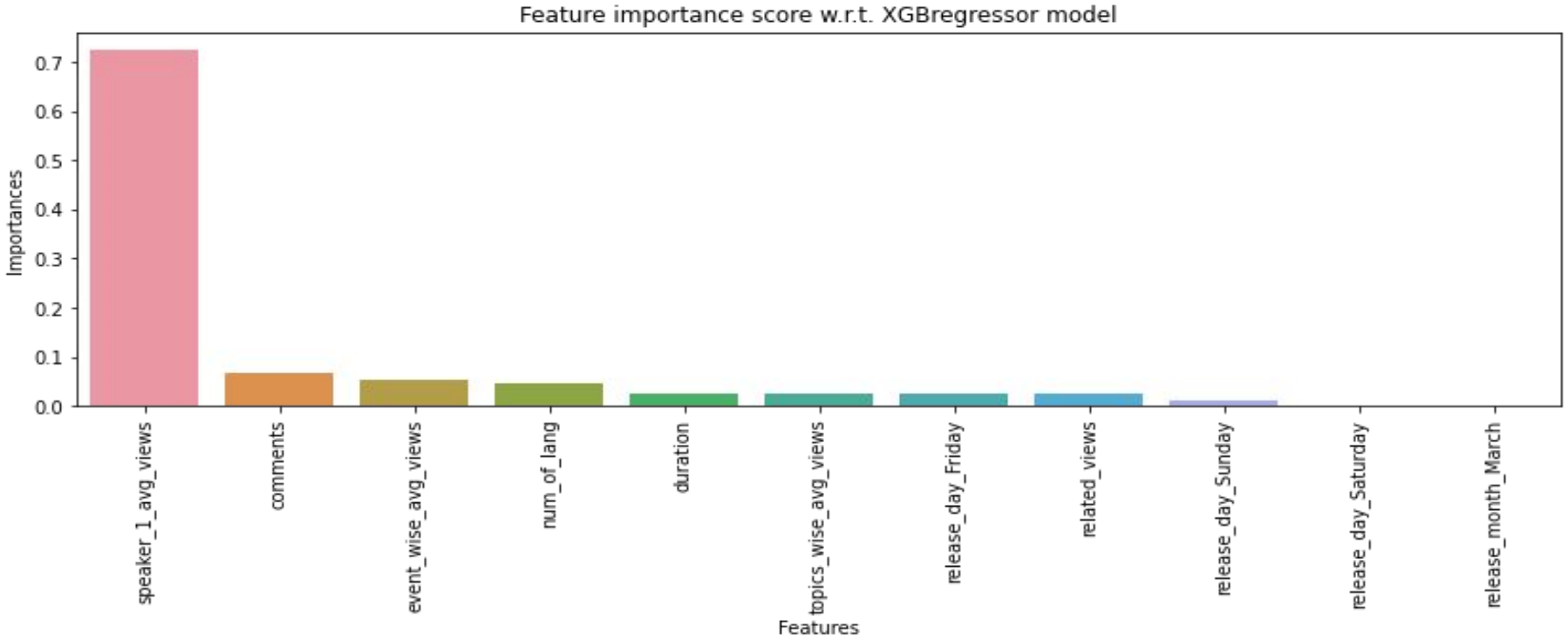
Feature importance w.r.t. Random Forest Regressor



Feature importance w.r.t. Extra Trees Regressor



Feature importance w.r.t. XGBoost Regressor



Choice of model and reason of choosing

- **Among all the models, Random Forest Regressor is the best performer in terms of MAE.**
- **We consider MAE over RMSE to choose our model because :**
  - **MAE is linear and RMSE is quadratically increasing.**
  - **MAE is best deciding factor because it isn’t affected by outliers.**

Challenges

- Dataset have lots of textual and categorical data having high ordinal number. So the conversion to meaningful numerical data was a challenge.
- Treating the outliers in numerical features.
- Generation of new features which need to be added in the model.
- Choosing the right features for modelling.
- Choosing the right models to get the best scores.

Conclusions

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

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

- The popularity of TED talk is dependent on number of languages it is available in. So, TED can increase their views by making every videos available in large number of different languages.