Capstone Project-2

Ted Talk Views Prediction

Shantanu Houzwala

Index

Title
• Problem Statement • Understanding the Data • EDA on Features • Feature Engineering • Feature Selection • Models Used • Which model did I choose and why? • Challenges • 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.
- As of 2015, TED and its sister TEDx chapters have published more than 2000 talks for free consumption by the masses and its speaker list boasts of the likes of Al Gore, Jimmy Wales, Shahrukh Khan, and Bill Gates.
- The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website.

Understanding the Data

❖ Dataset Information:

- Number of instances: 4,005
- Number of attributes: 19

❖ The dataset contains features like:

- **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
- **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

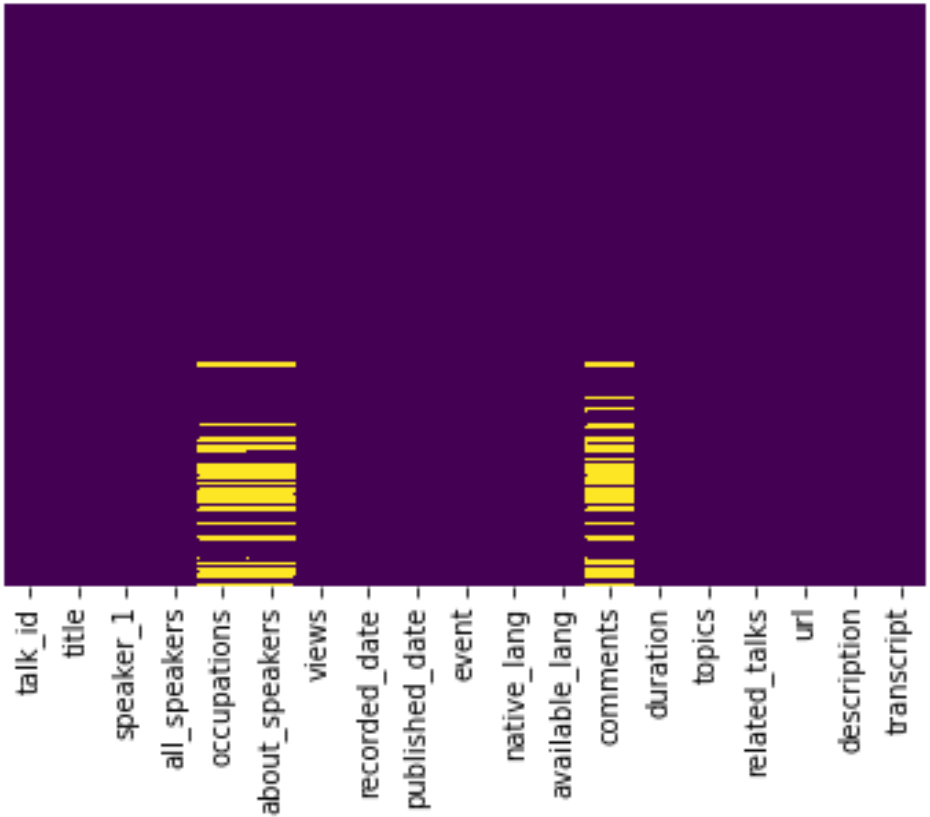
Exploratory Data Analysis on Features

Check Missing Data

- KNN imputation for Numerical Features

- The dataset contains NaN values in few columns like:
  - all_speakers,
  - occupations,
  - about_speakers,
  - comments,
  - recorded_date

- Replaced Categorical Features
  Nan values with ‘Unknown’ category

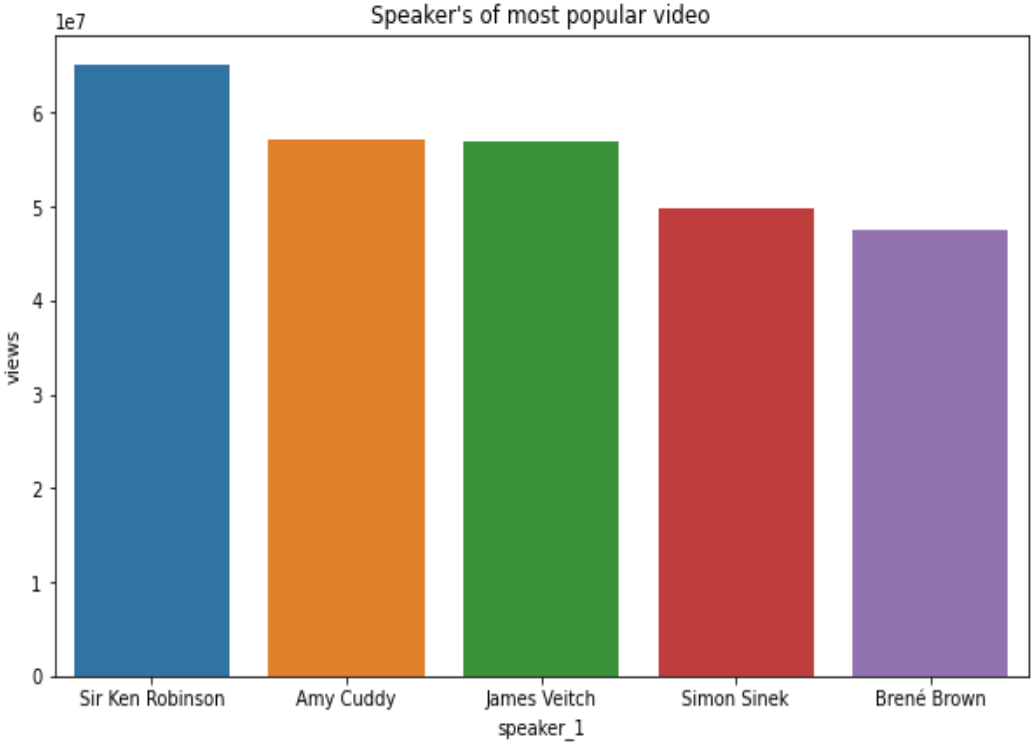


Top Speakers according to Views

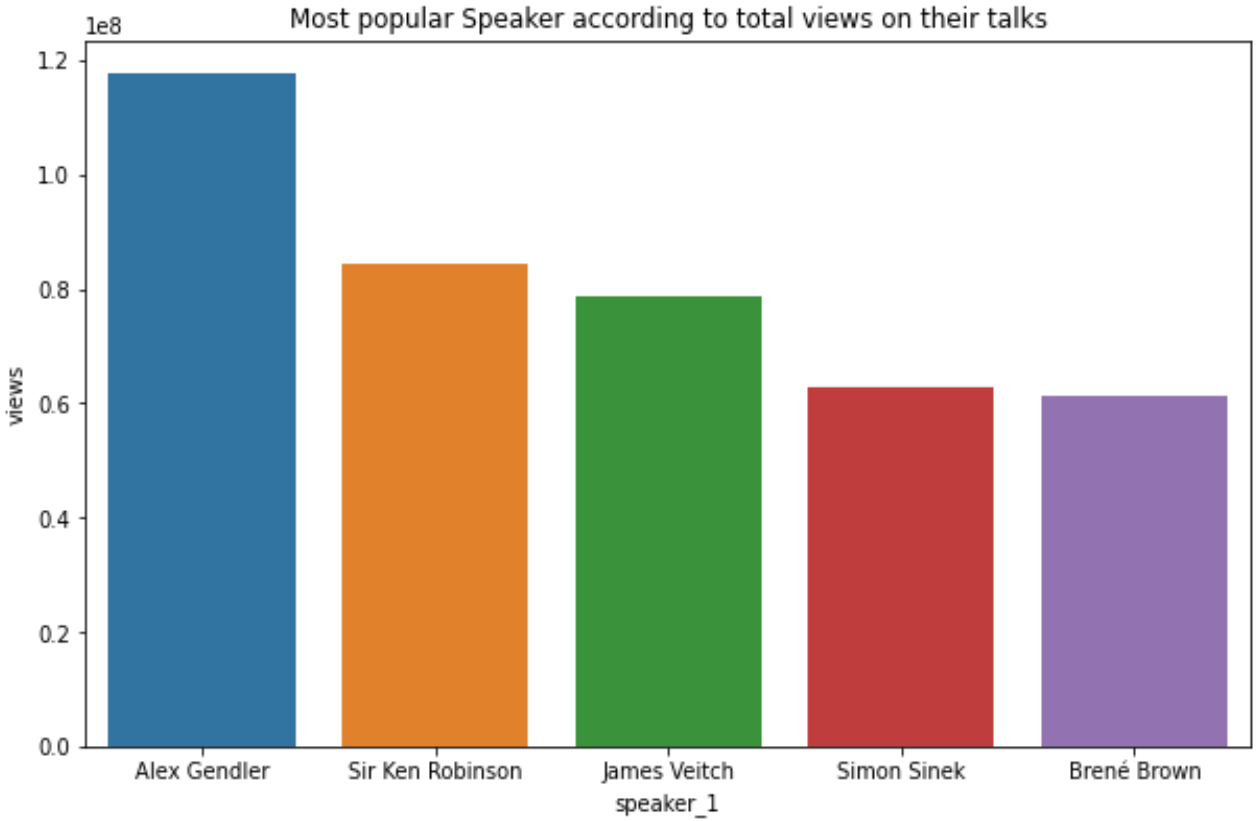
- Sir Ken Robinson's talk on "Do Schools Kill Creativity?" is the most popular TED Talk of all time with more than 65 million views.

- It is closely followed by Amy Cuddy talk on "Your body language may shape who you are" with more than 57 million views.

- There is only one talk that has crossed 60 million mark while 3 talks have crossed 50 million mark.

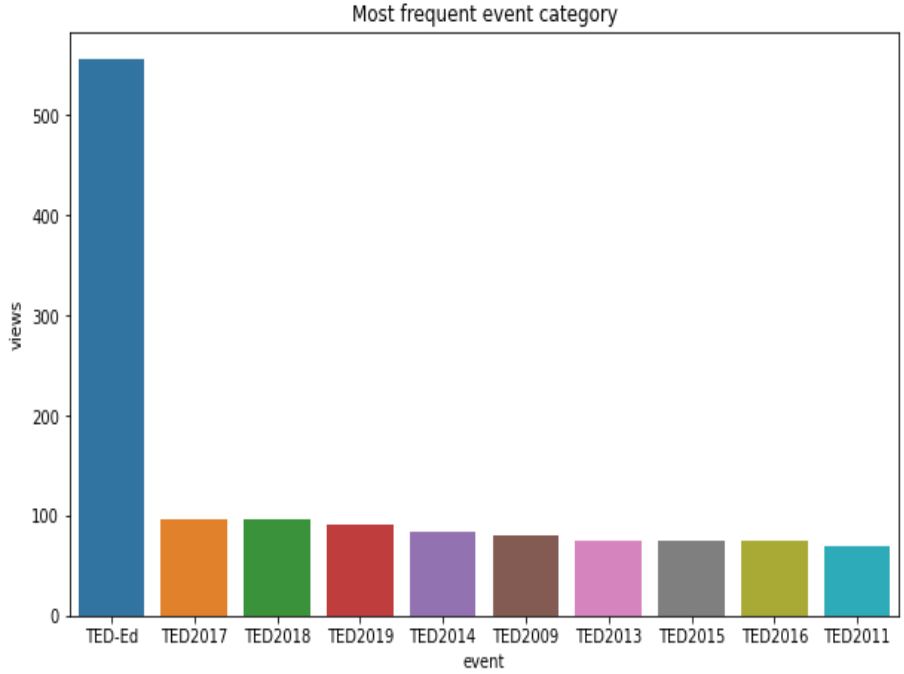


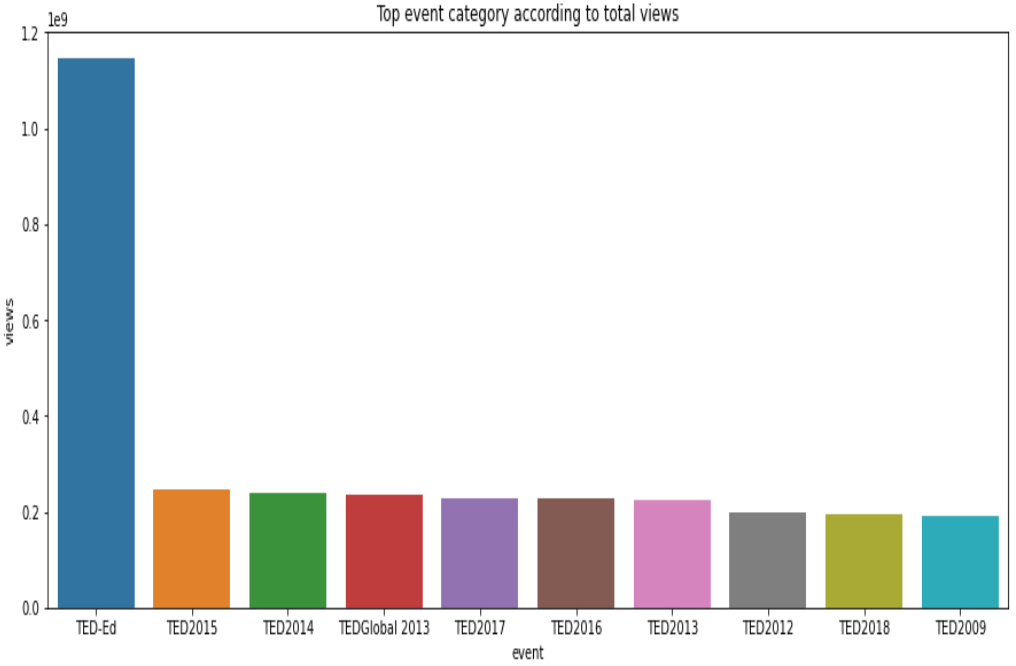
Top Speakers according to Views



- Alex Gendler is the most popular speaker followed by Sir Ken Robinson

Top Event Category

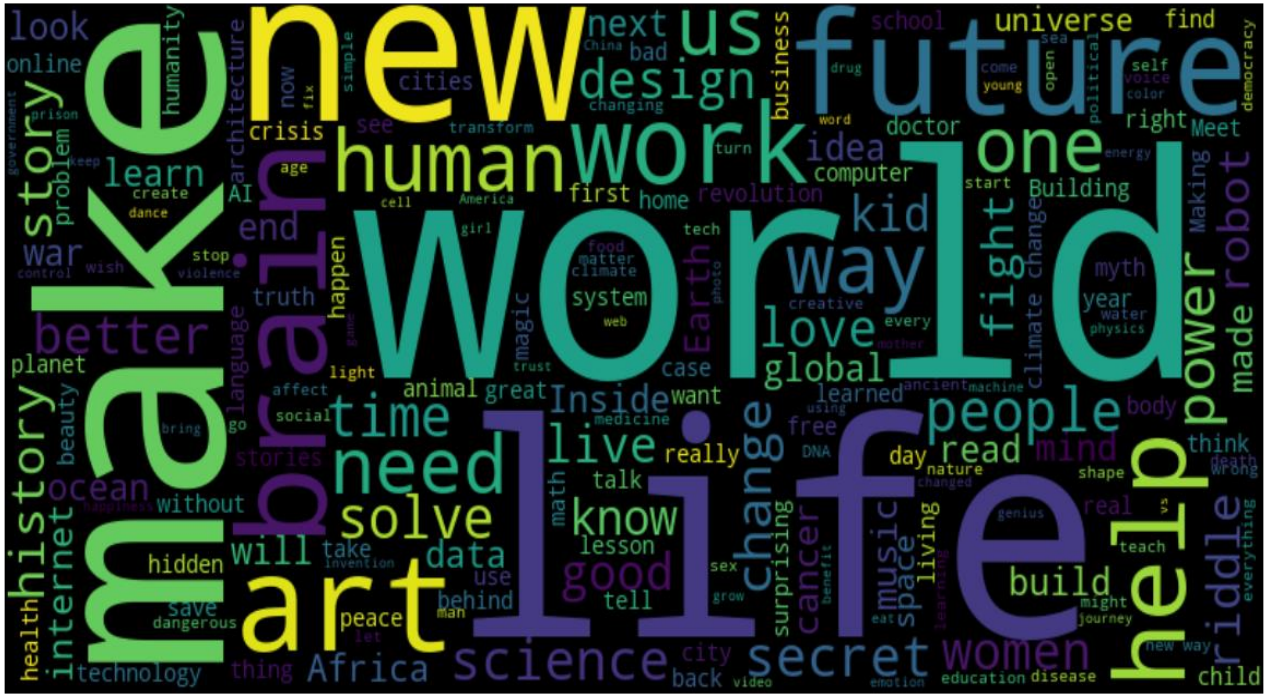




- **TED-Ed is the most frequent event category with 556 entries followed by TED2017 and TED2018**
- **TED-Ed is the most popular event category having maximum number of total views followed by TED2015**

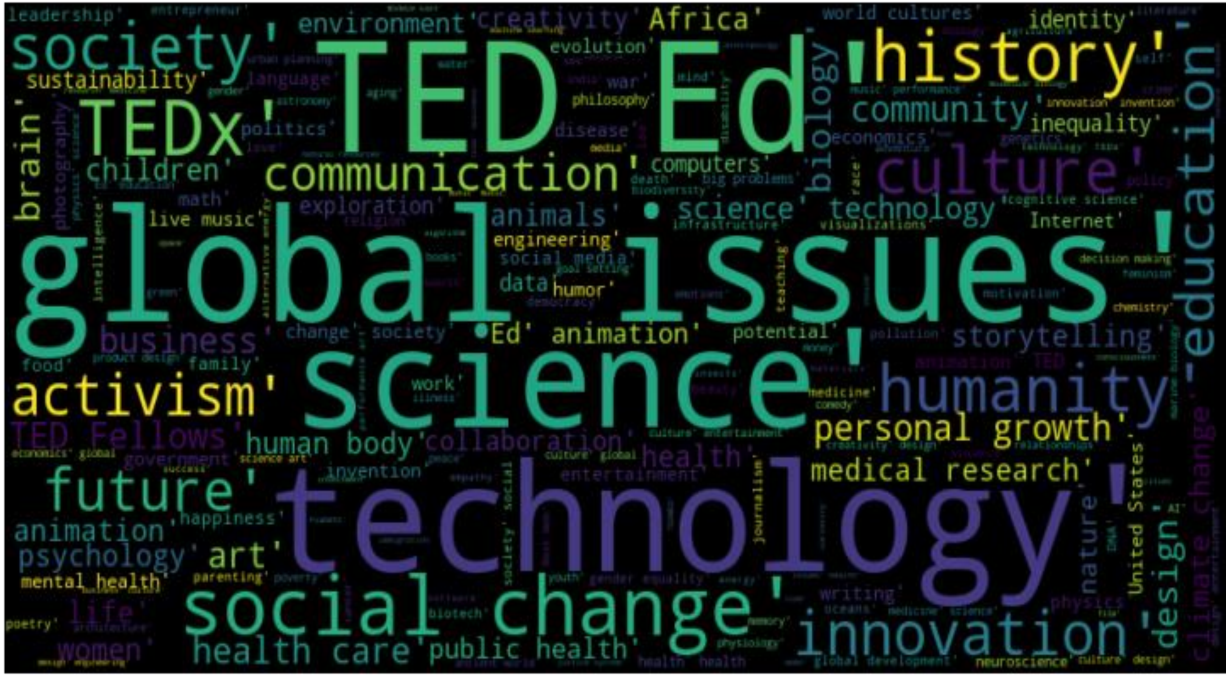
Most Popular Titles

- There are 1,57,160 words in the combination of all titles
- Majority of Titles contains words like life, world, make, new, future, art.



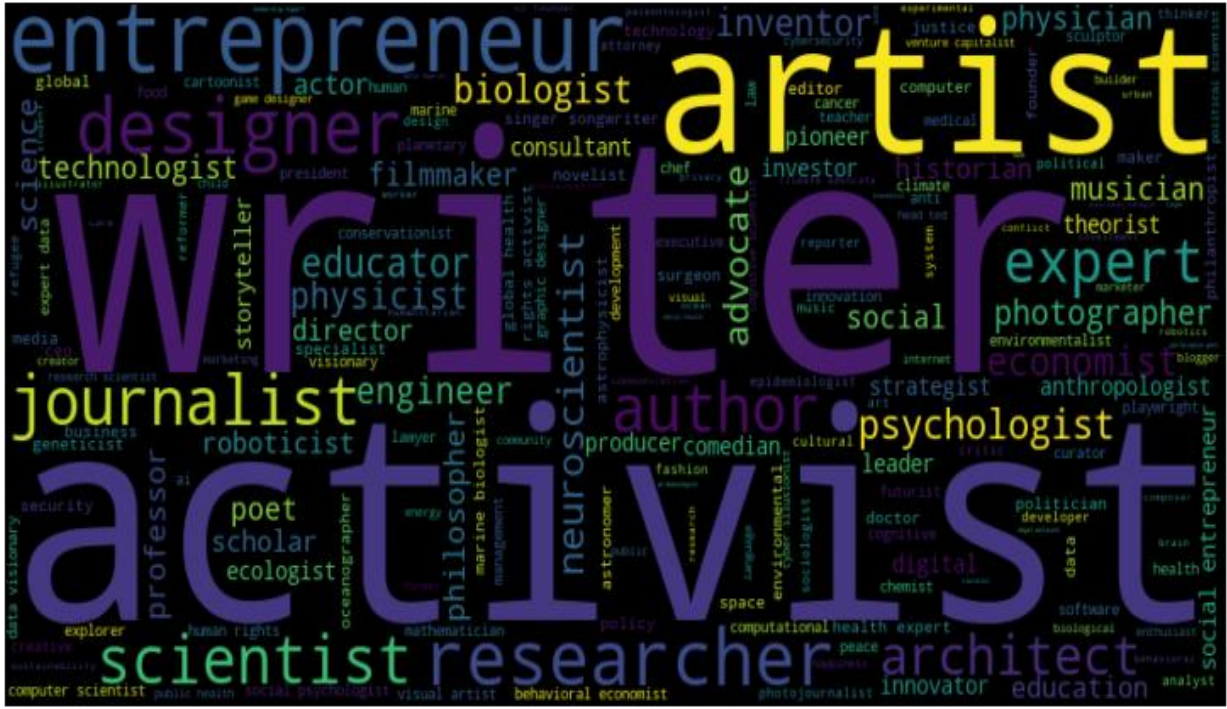
Most Popular Topic Tags

- There are 4,04,126 words in the combination of all topics.
- Majority of topic tags contain words like technology, global, science, issues, social change.

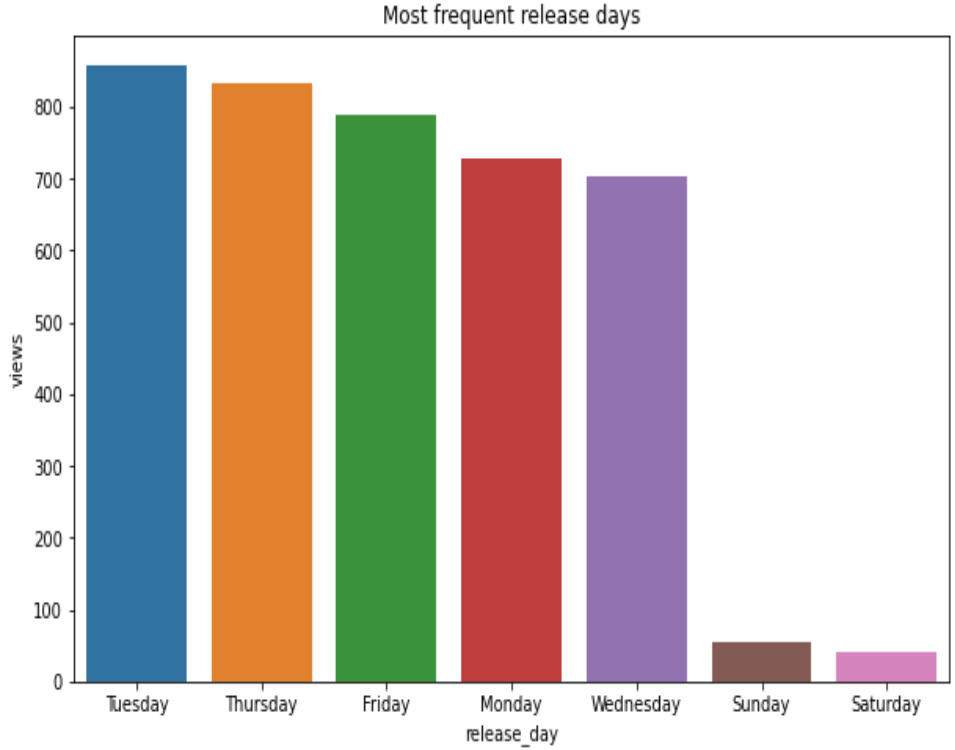


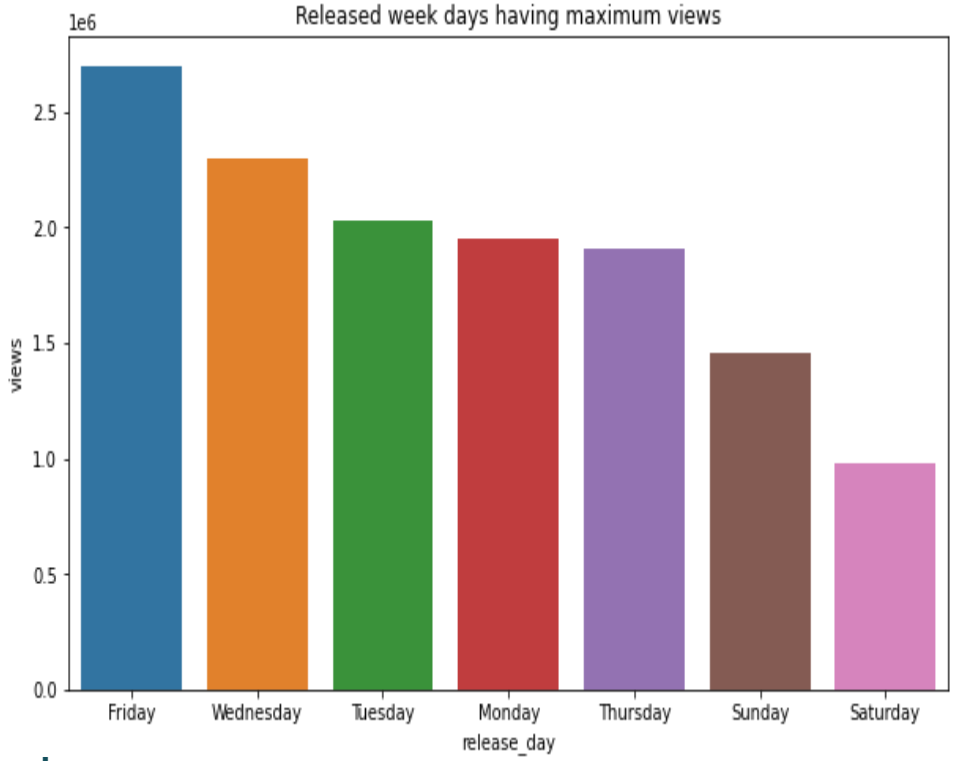
Most Popular Occupation of Speakers

- There are 55,651 words in the combination of all occupations.
- Most of the speakers are writers followed by activists, artists and journalists.



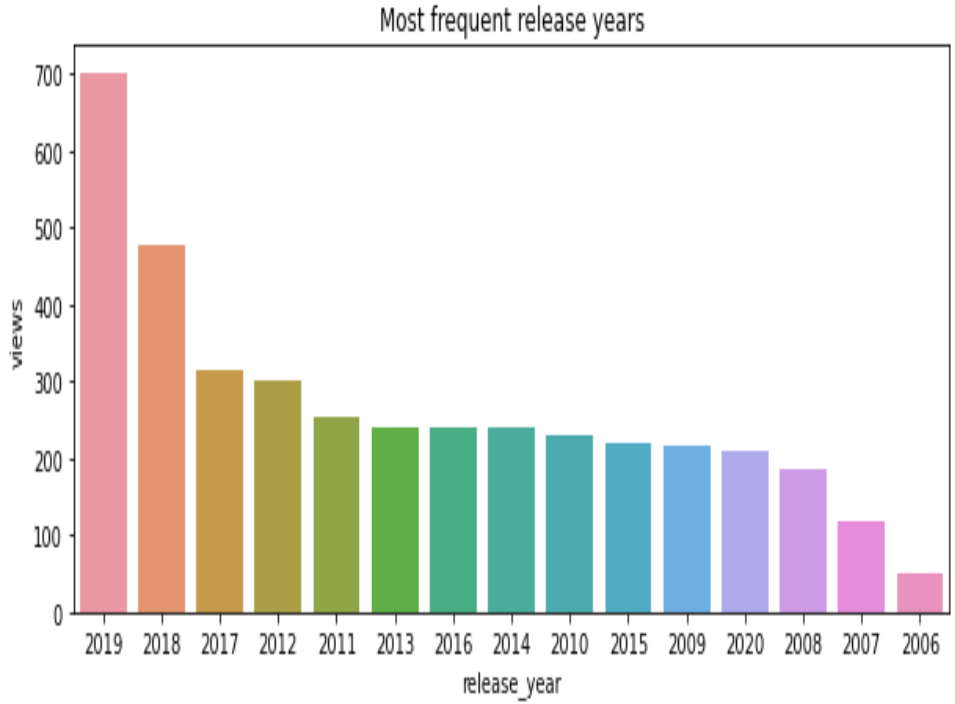
Published Days with Views

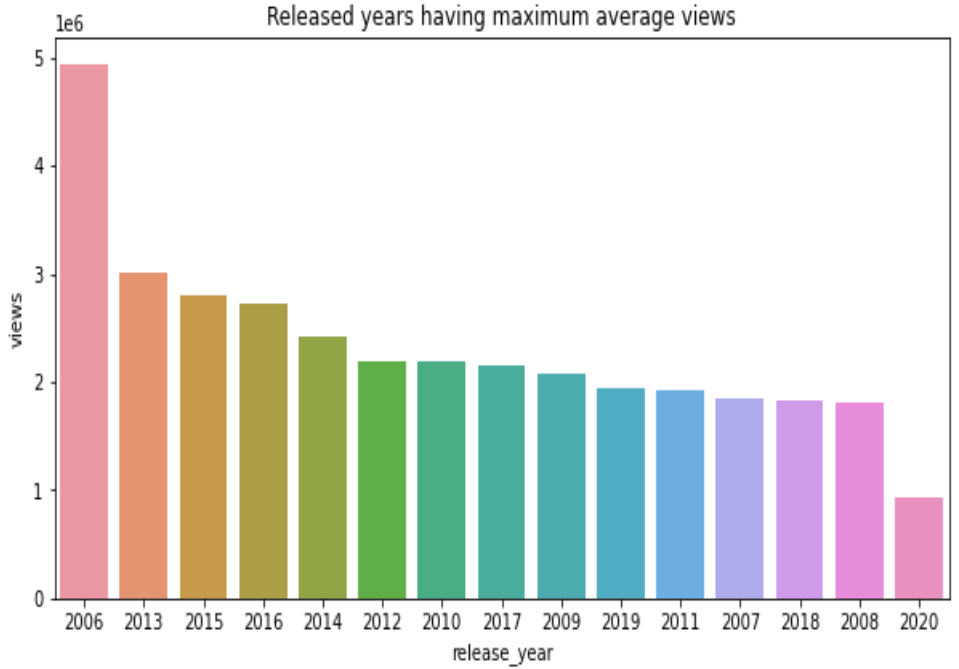




- **Most videos are published on Tuesday followed by Thursday.**
- **But the videos published on Friday are more popular (i.e. have more average views) followed by Wednesday.**
- **Friday release is impacting the views of the video**

Published Year with Views



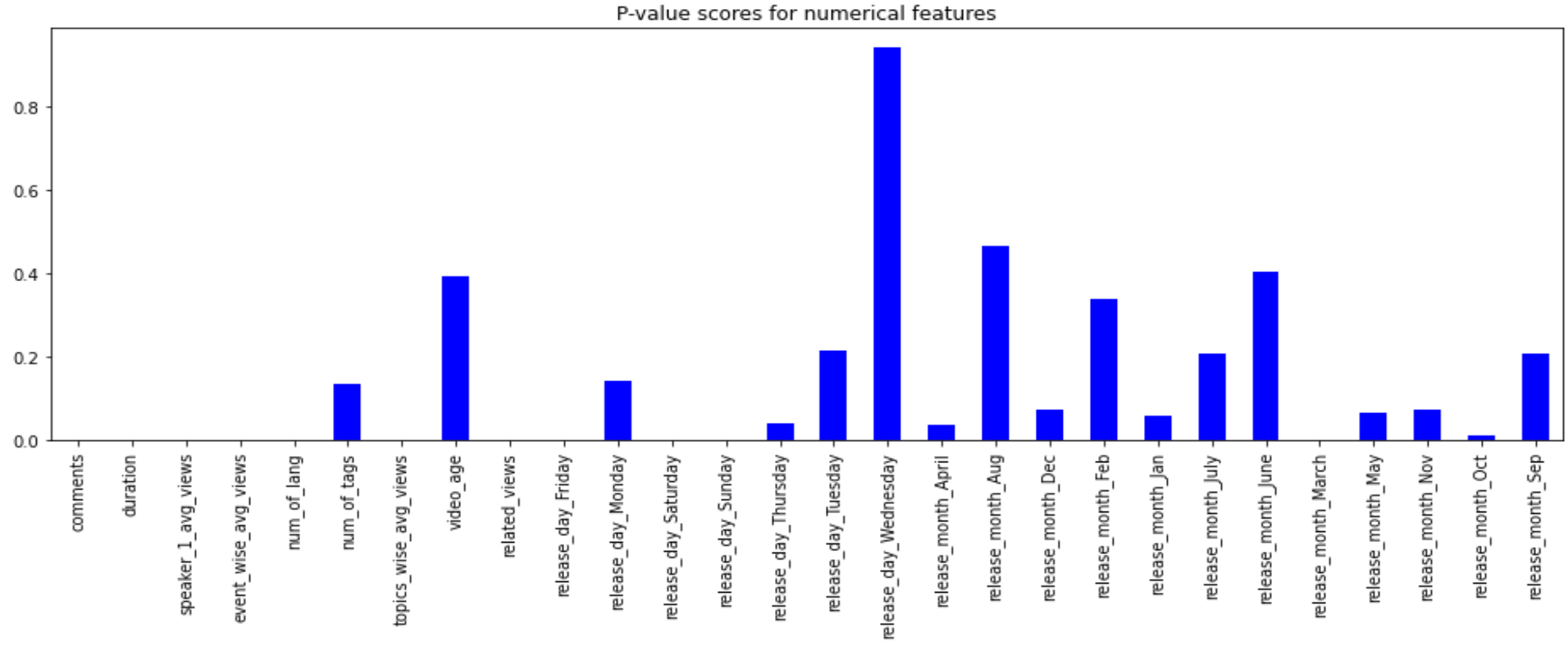


- **Most videos are published in 2019 followed by 2018 and 2017.**
- **But the videos published in 2006 are most viewed followed by 2013 and 2015.**

Feature Engineering

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

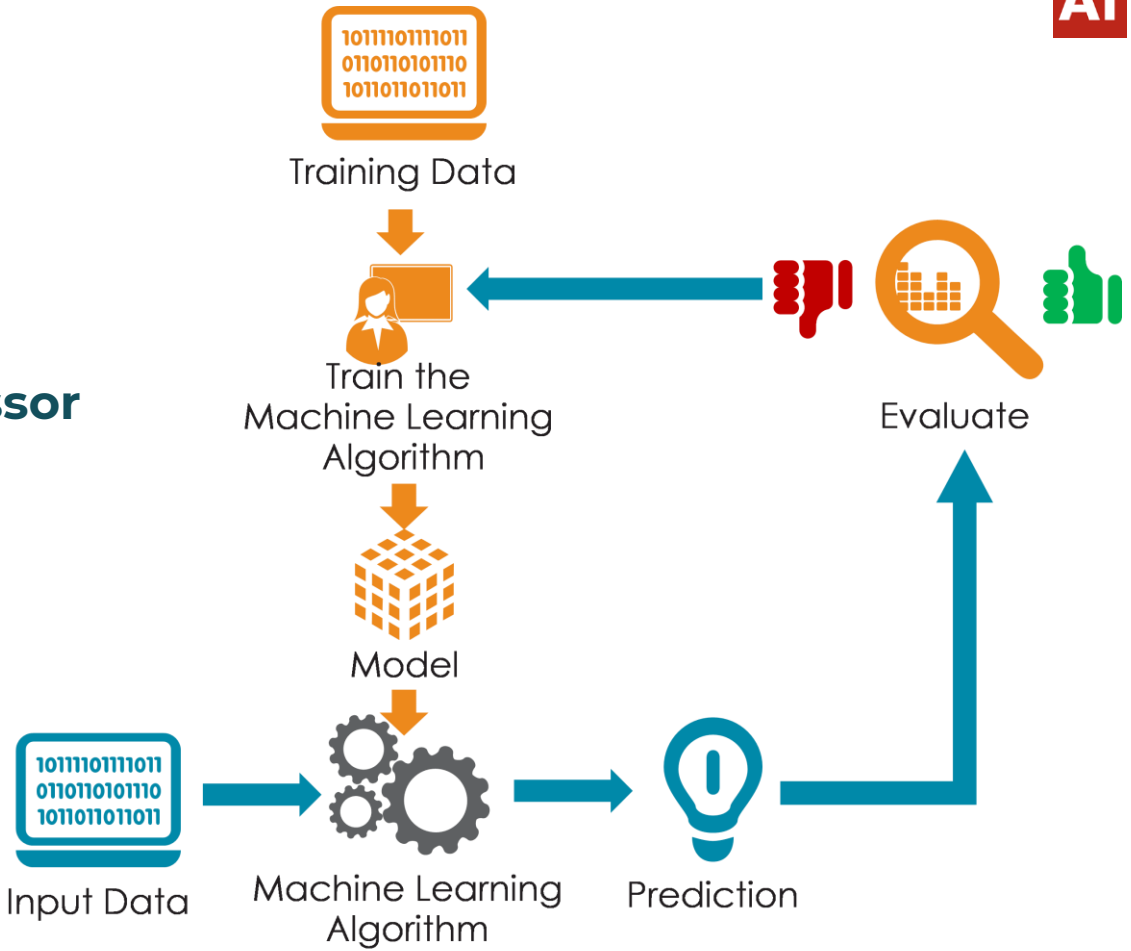
Feature Selection (f regression)



- From here using p value analysis we can drop those features having high p values

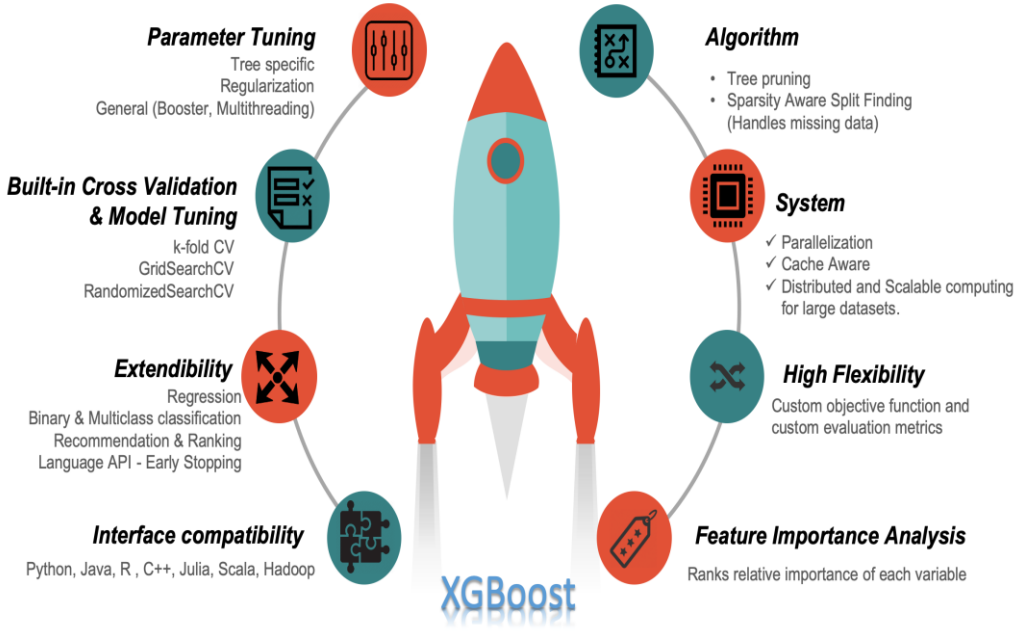
Models Used:

- **XGBoost Regressor**
- **Random Forest Regressor**



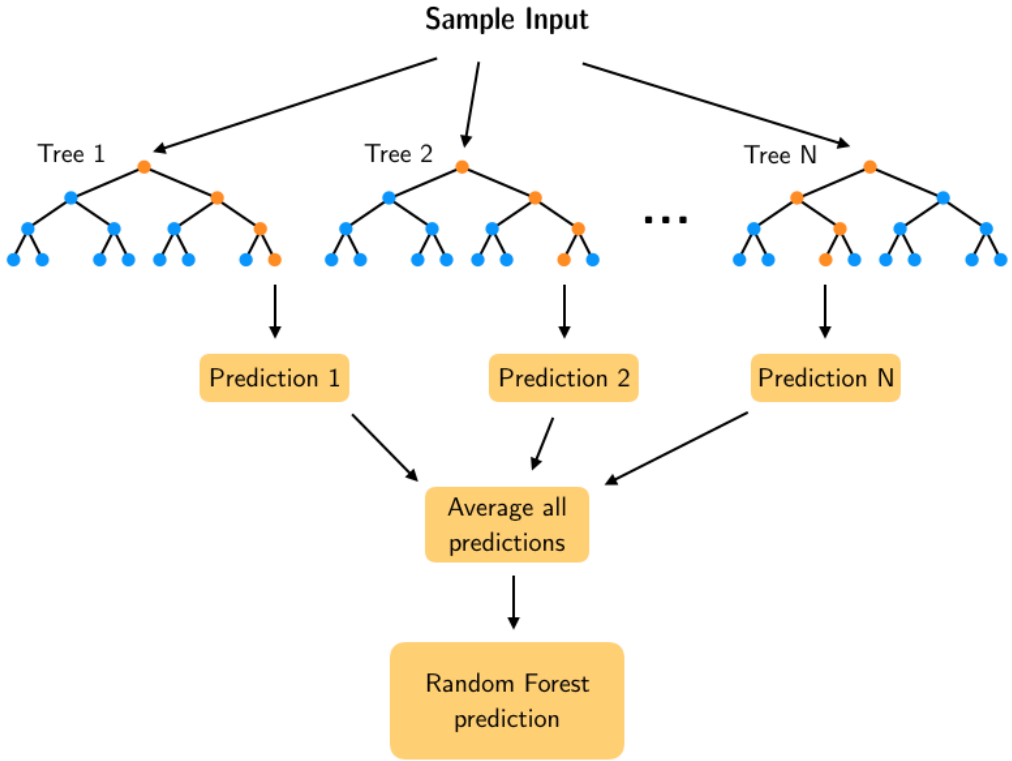
XGBoost Regressor

- MAE train: 164091.332037
- MAE test: 226944.860549
- R2_Score train: 0.918158
- R2_Score test: 0.830151
- RMSE_Score train: 315411.385197
- RMSE_Score test: 454270.753145

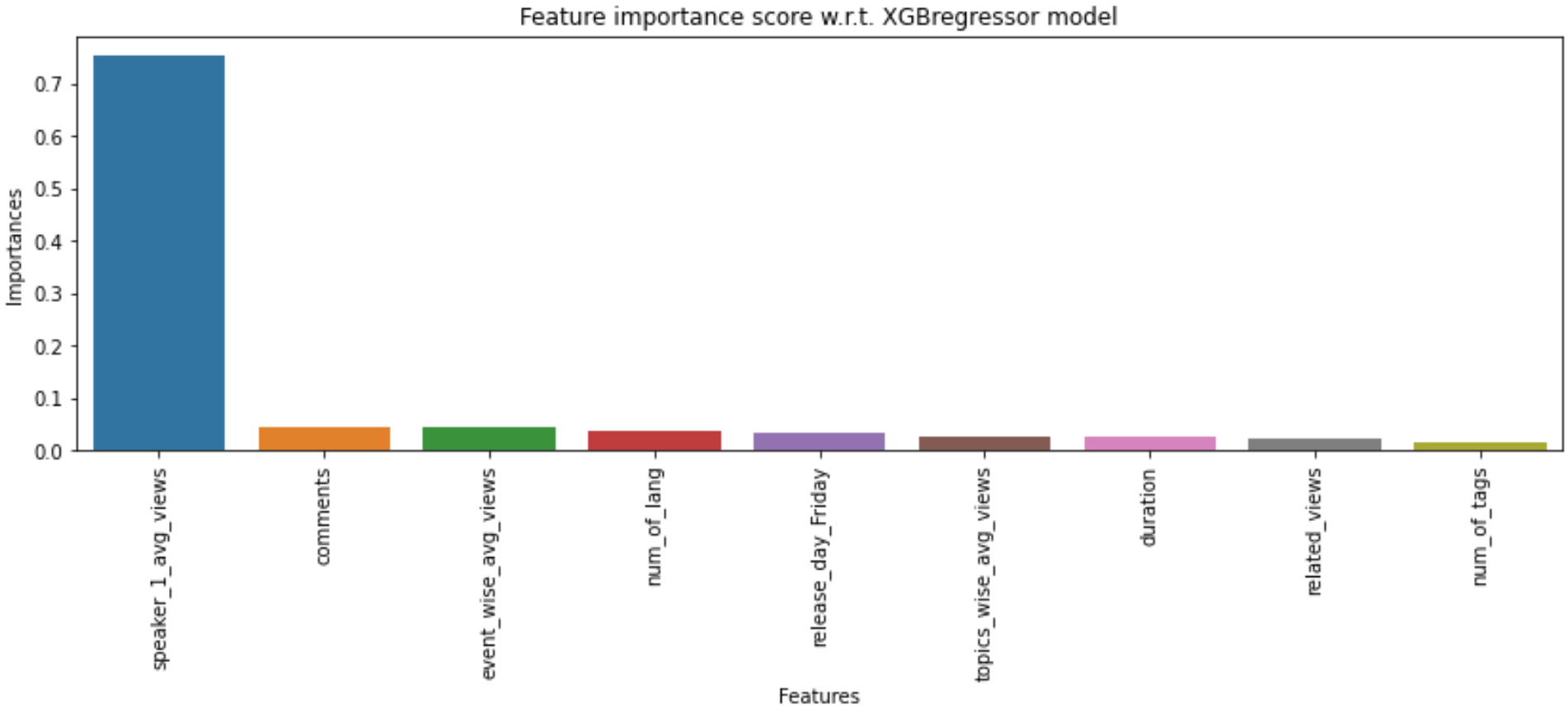


Random Forest Regressor

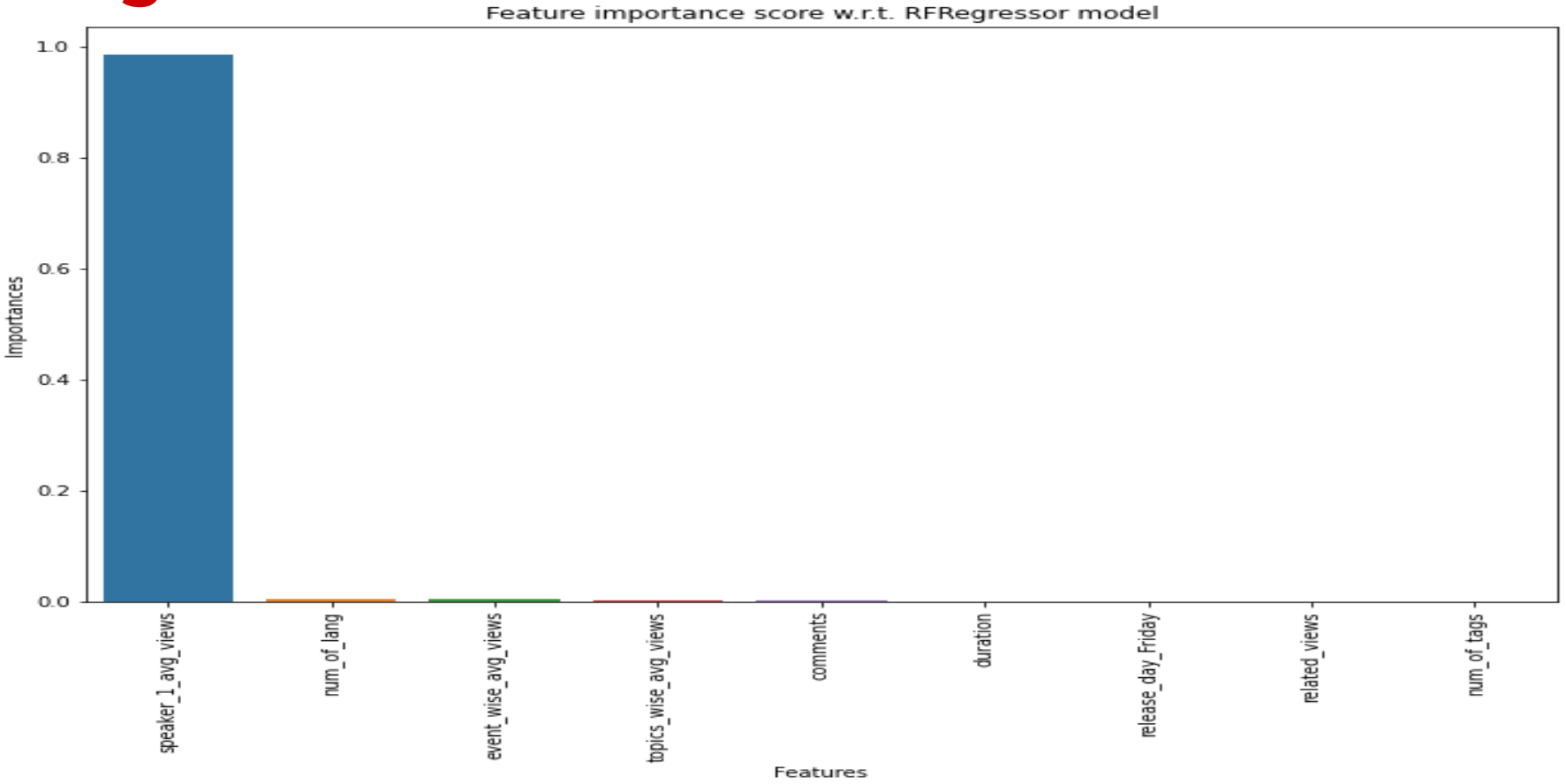
- MAE train: 186583.315347
- MAE test: 191844.536467
- R2_Score train: 0.806193
- R2_Score test: 0.803246
- RMSE_Score_train: 485371.330401
- RMSE_Score_test: 488927.132141



Feature Importance wrt XGBoost Regressor



Feature Importance wrt Random Forest Regressor



Which Model Did we Choose and Why?

- Out of all the models Random Forest Regressor is the best performer.

- We choose MAE and not RMSE as the deciding factor. RMSE is heavily influenced by outliers as the higher the values get the more the RMSE increases.

- MAE is the best deciding factor because it doesn’t increase with outliers, thus MAE remains linear.

Challenges Faced

- Dataset had lots of textual and categorical data. So the conversion to meaningful data was a challenge.
- Treating the outliers in numerical features.
- Generation of new features needed for models.
- Choosing right features
- Choosing right model to get best results.

Conclusion

- Successfully build a predictive model, which could help TED in predicting the views of the talks uploaded on the TEDx website.
- TED can improve the views on the less popular topics by inviting more popular speakers.
- TED can increase their views and popularity by increasing videos on sections like Technology and Science.
- TED can use topic modelling to tackle views in each topic separately.

Thank You