Capstone Project - 2

TED Talk Views Prediction

Individual Contributor

Ammar Anjum

Content

pTHeyOlem

Title

- Problem Statement
- Data Summary
- EDA on features
- Feature Engineering
- Feature Selection
- Models used
- 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 is 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, they 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.

Data Summary:

Data set name: data_ted_talks

Shape:

- Rows -- 4005
- Columns--19

Features:

'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'

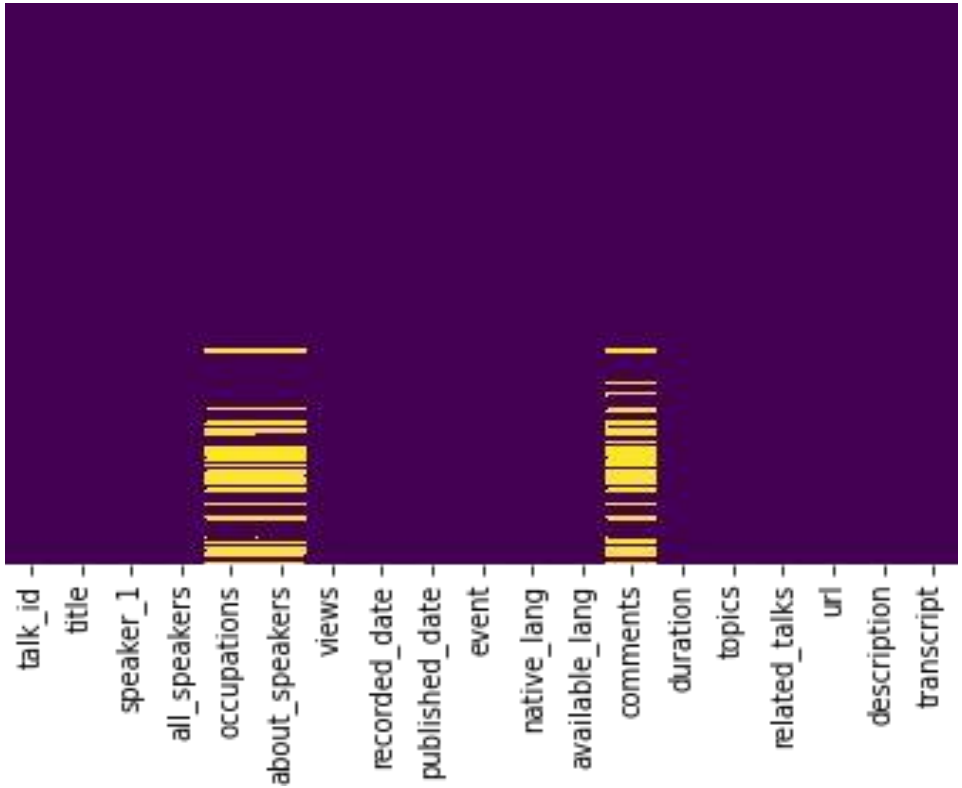
Target Variable: 'views'

Exploratory Data Analysis on Features

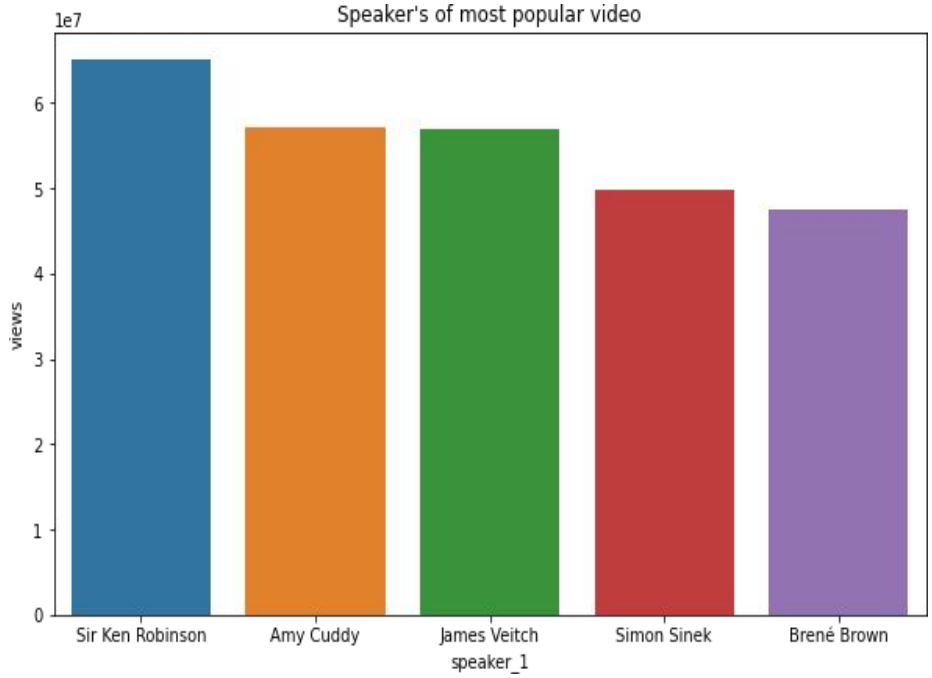
Missing Data Check

- KNN imputation for Numerical Features

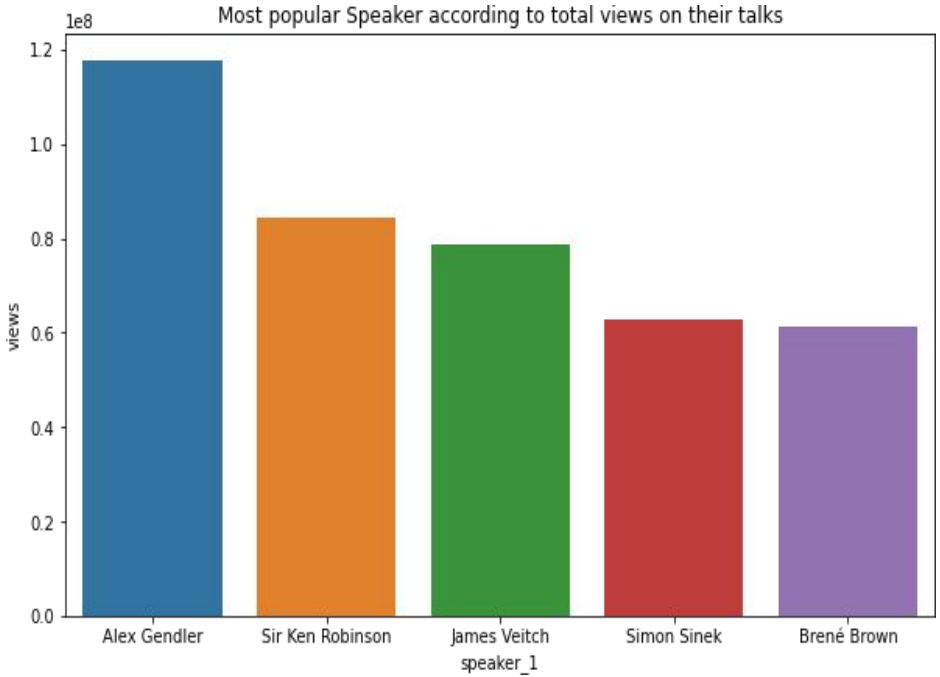
- Replaced Categorical Features Nan values with ‘Unknown’ category



Speakers with Views:

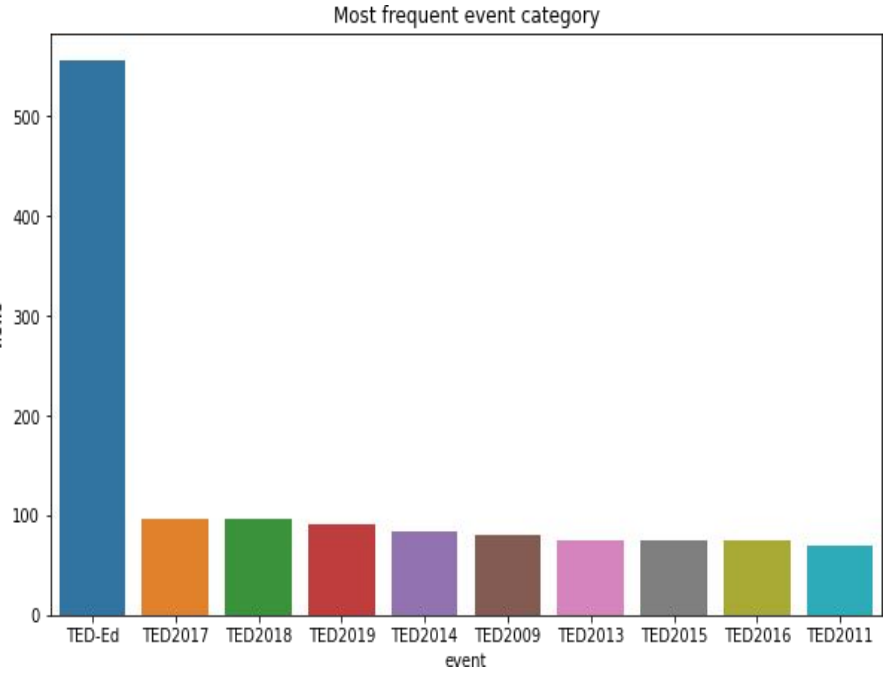


Speakers of most popular video

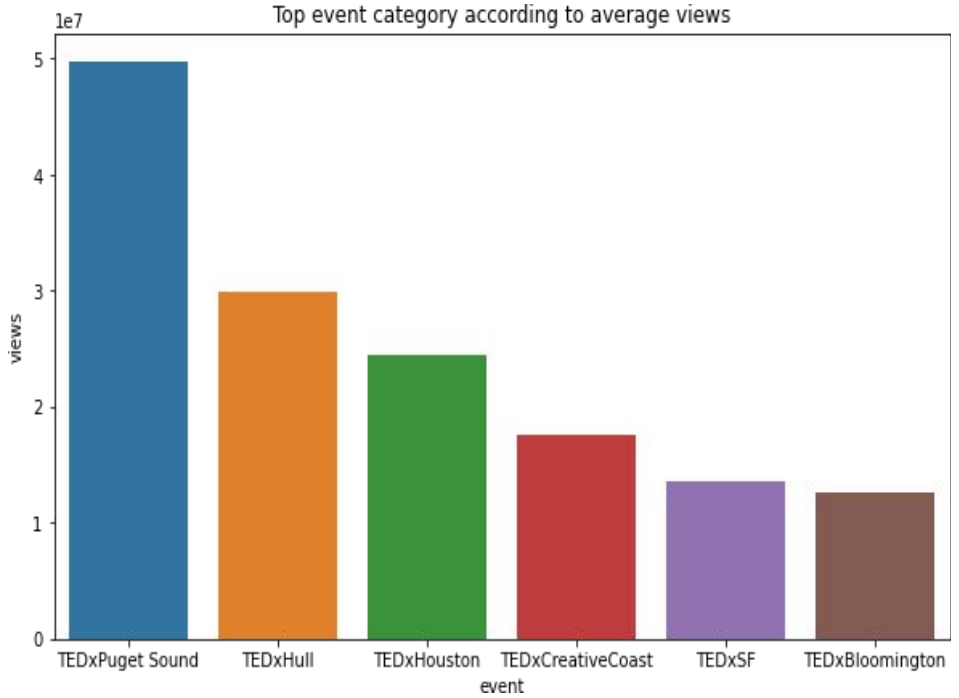


Top Speakers by total Views

Events with Views:

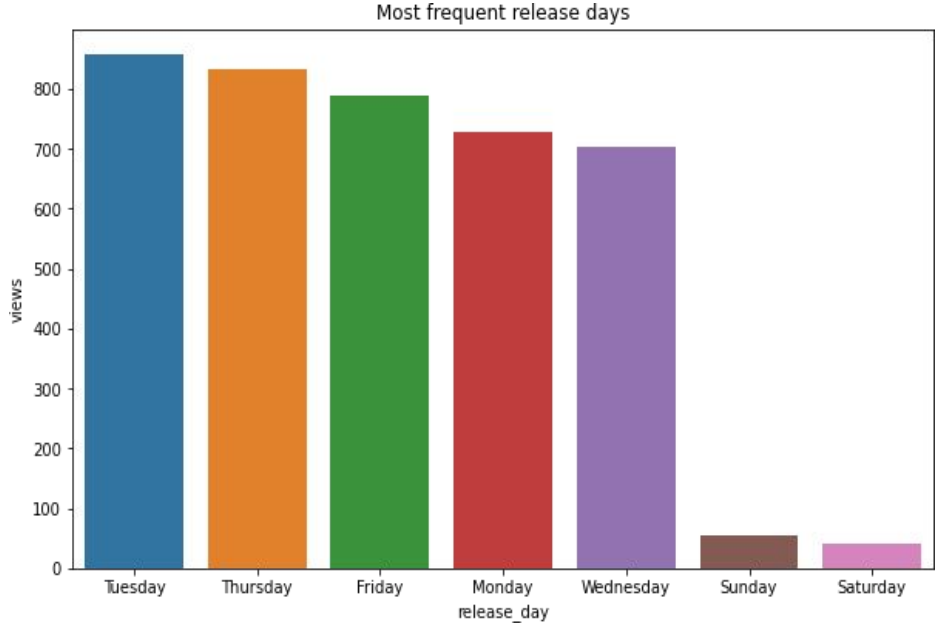


Most Frequent event category

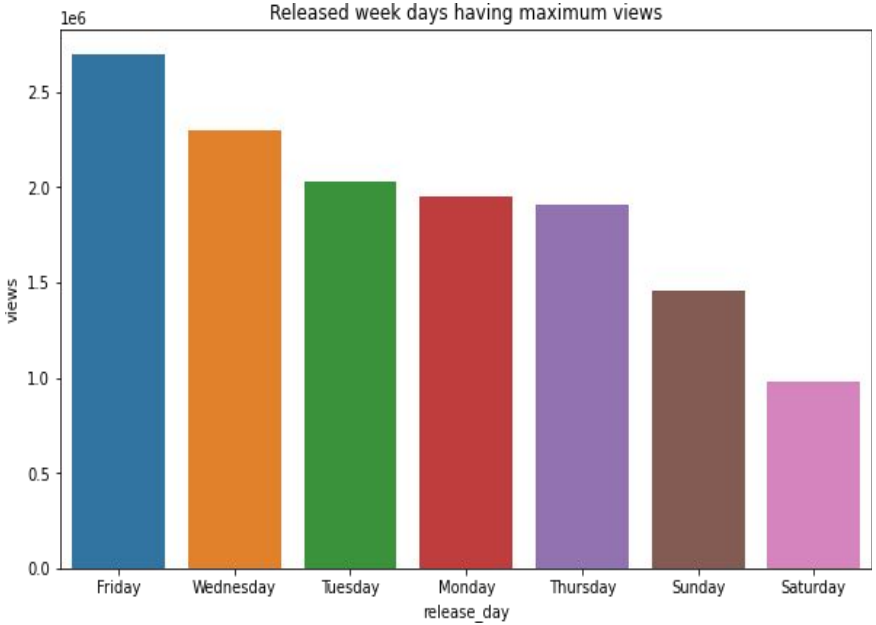


Top Events by Average Views

Published Days with Views:



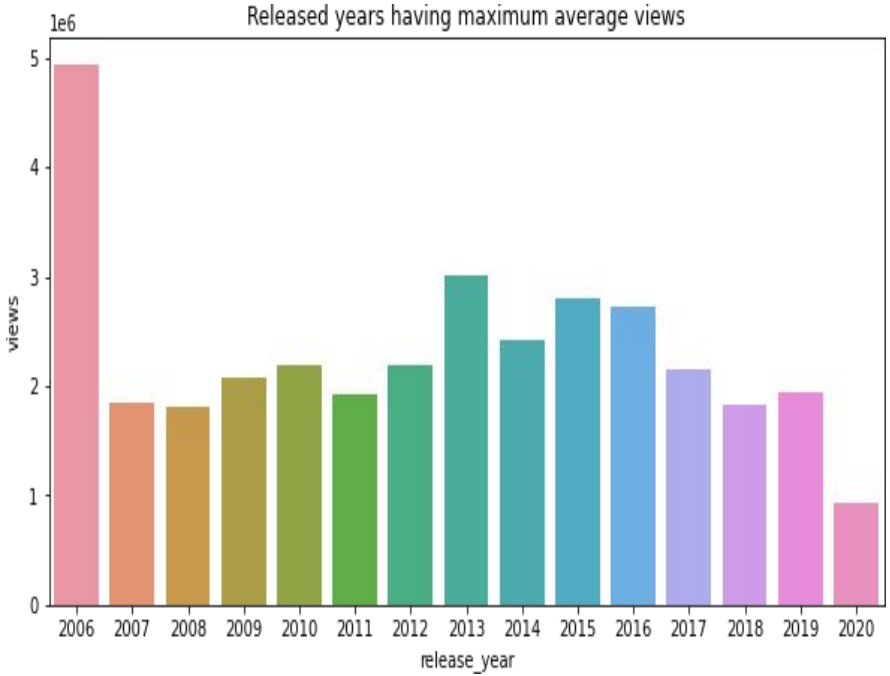
Frequent Released Days



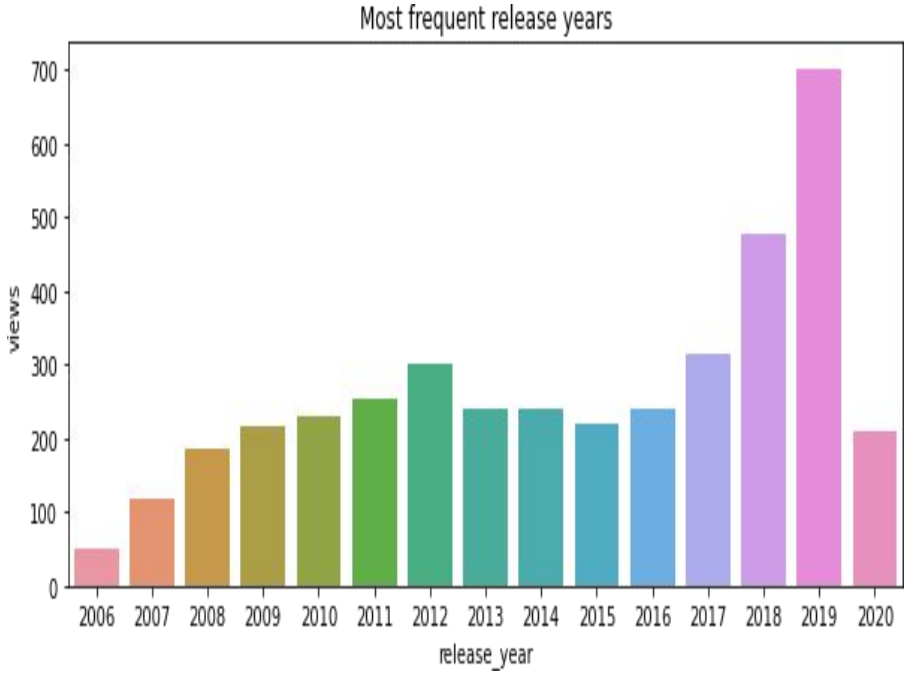
Released Days by avg Views

- Friday release is impacting the views of the video

Published Year with Views:



Released Year with Max average views

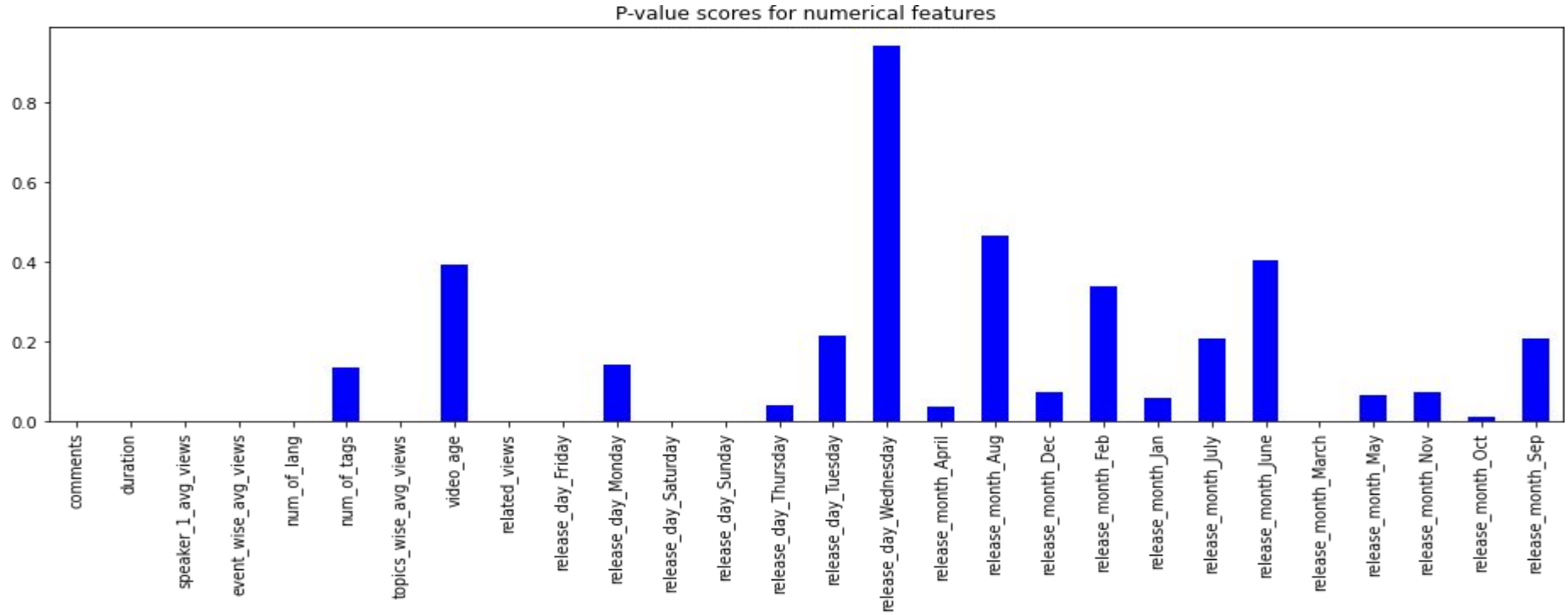


Most Frequent Released Year

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

Features selection(f regression):

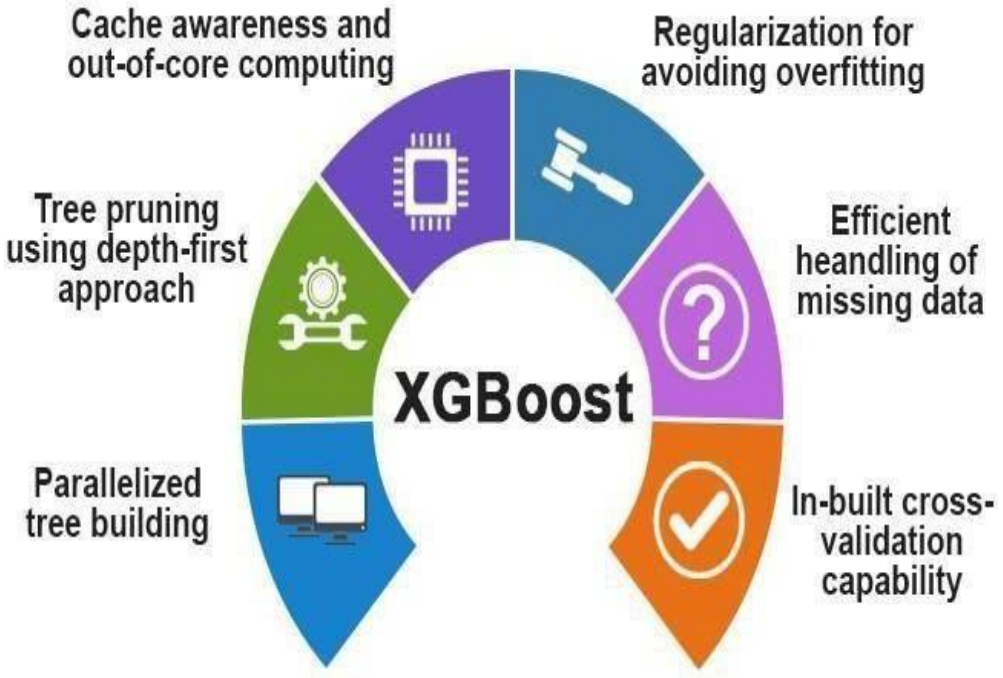


Models used:

- XGBoost Regressor
- Extra Trees Regressor
- Random Forest Regressor

XGBoost Regressor:

- Criterion = MAE
- R_Square for train= 0.9
- R_Square for test= 0.83
- MAE train = 164091.33
- MAE test= 226944.86
- RMSE train= 315411.38
- RMSE test= 454270.75

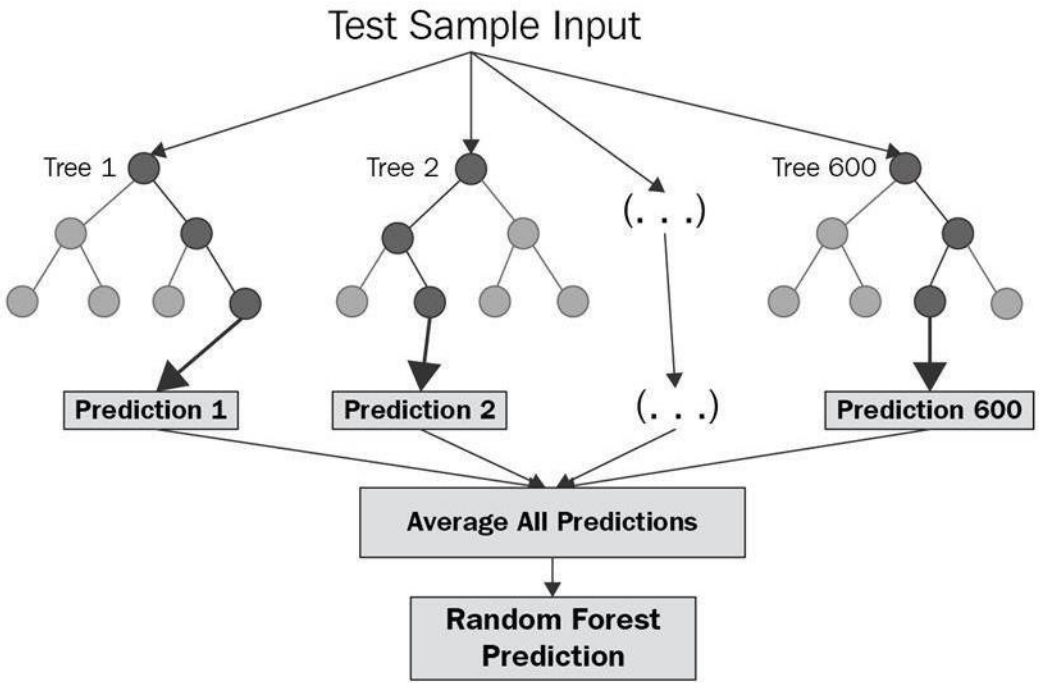


Extra Trees Regressor:

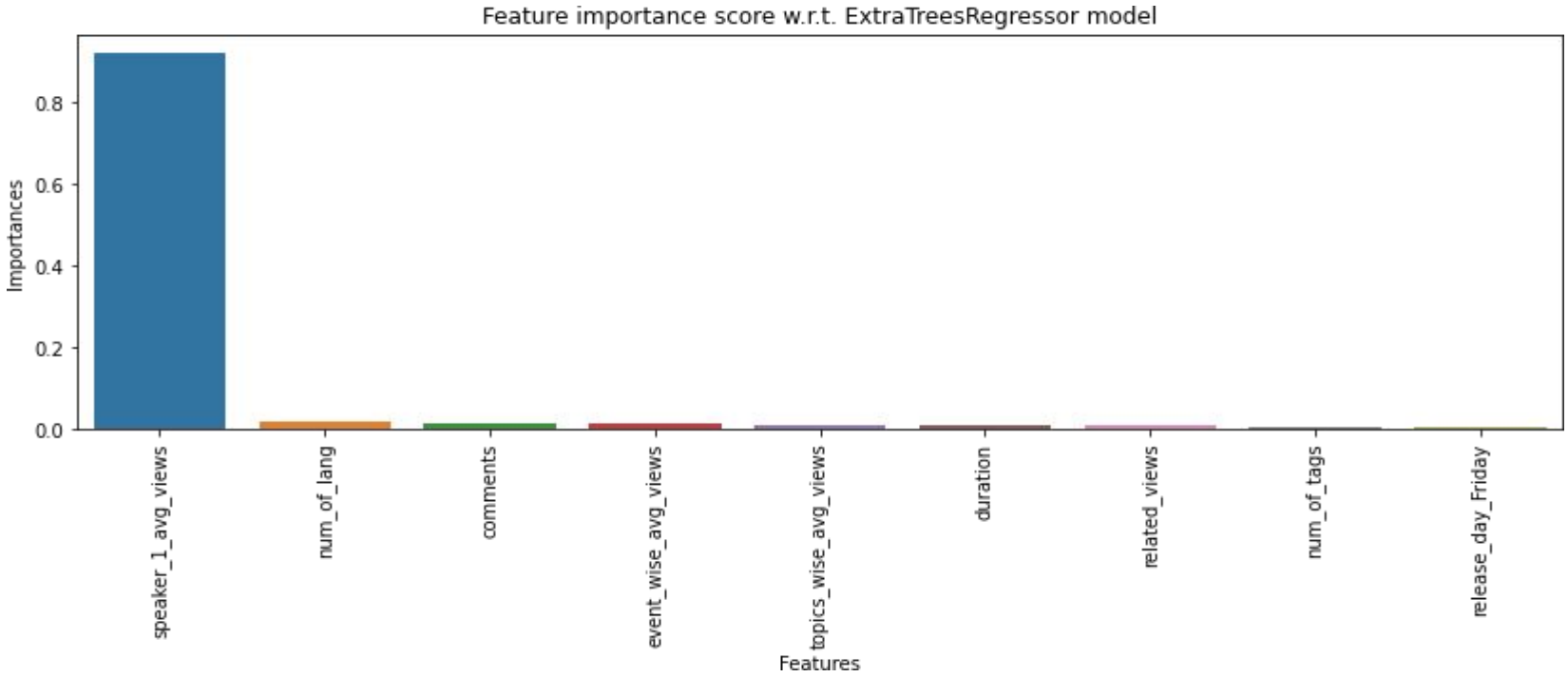
- Criterion = MAE
- R_Square for train= 0.79
- R_Square for test= 0.83
- MAE train = 207304.04
- MAE test= 204793.75
- RMSE train= 497317.34
- RMSE test= 484832.84

Random Forest Regressor:

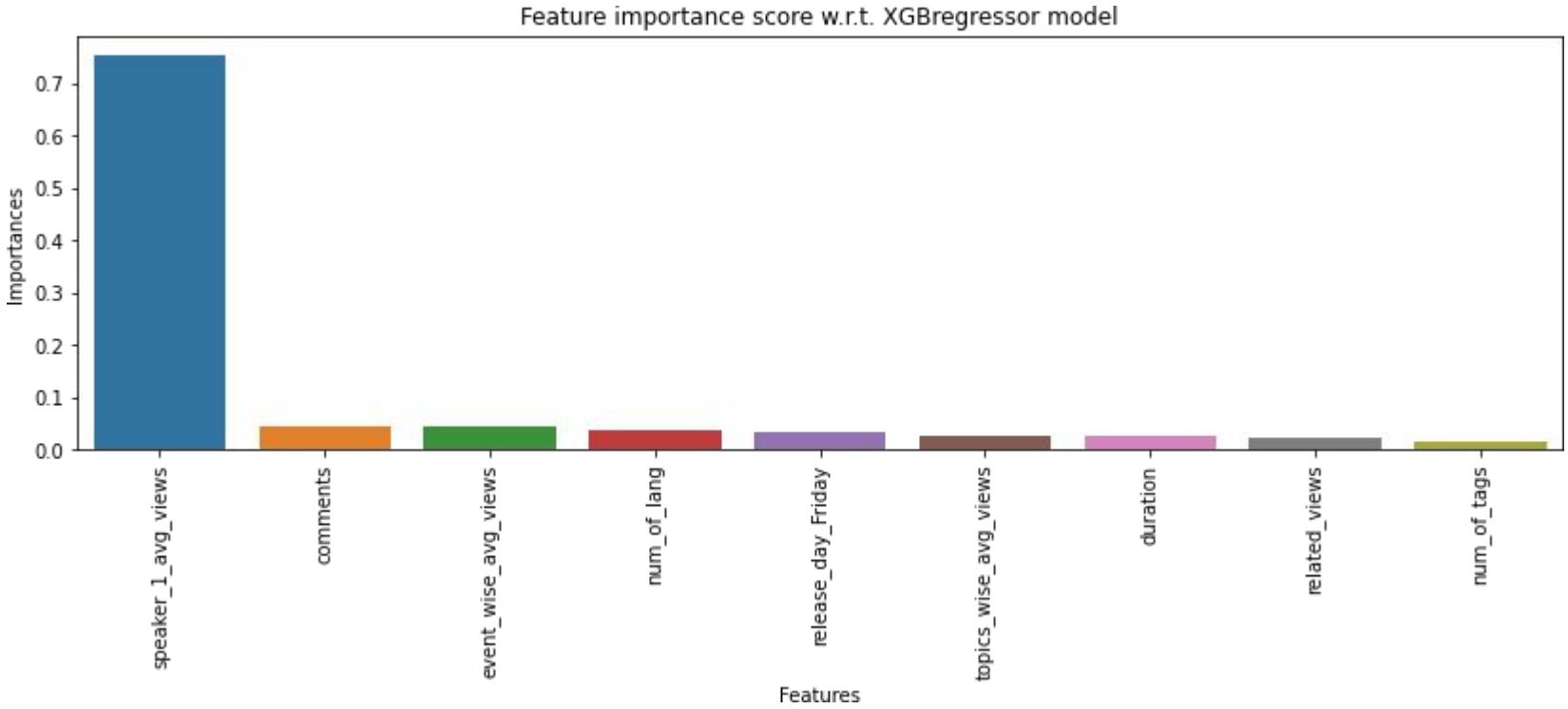
- Criterion = MAE
- R_Square for train= 0.80
- R_Square for test= 0.80
- MAE train = 186583.31
- MAE test= 191844.53
- RMSE train= 485371.33
- RMSE test= 488927.13



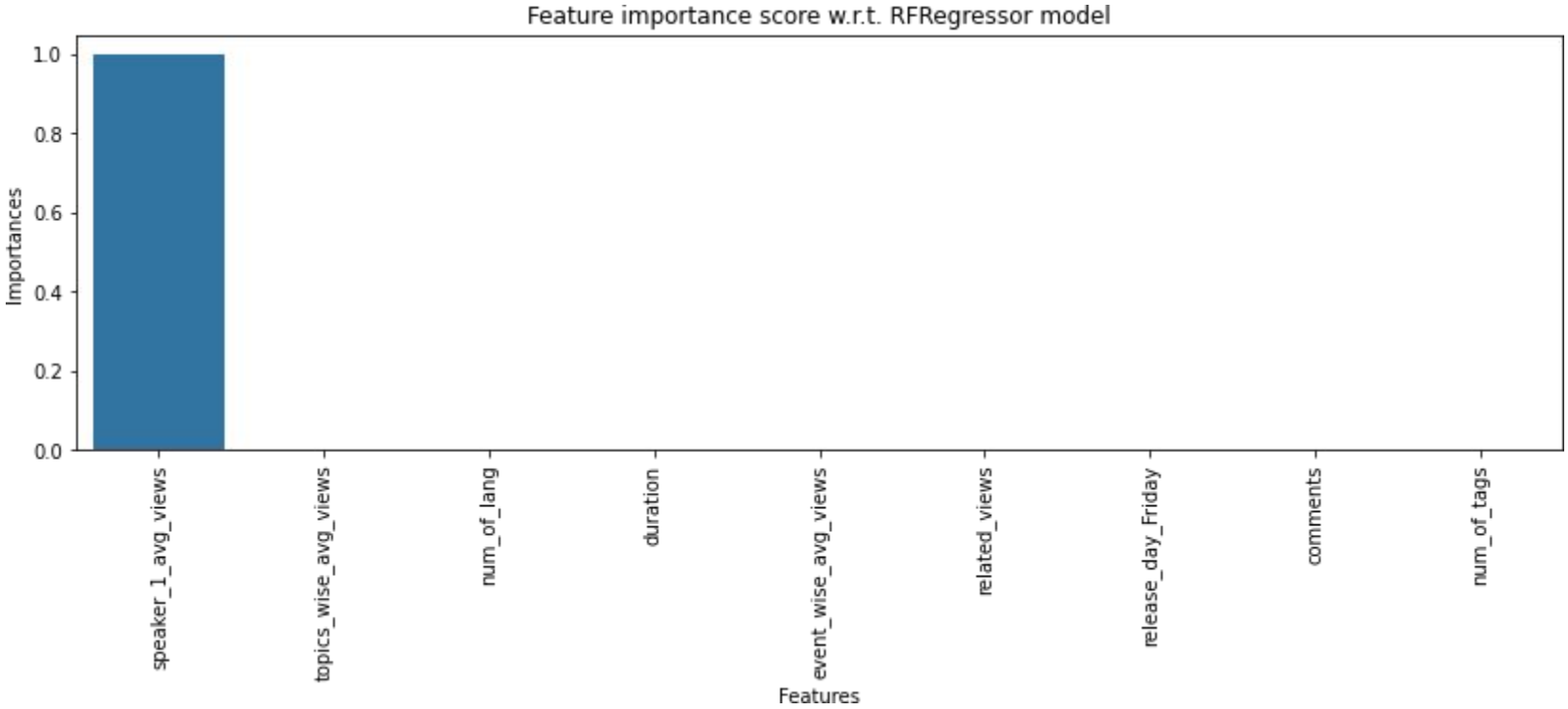
Feature importance wrt Extra Trees Regressor:



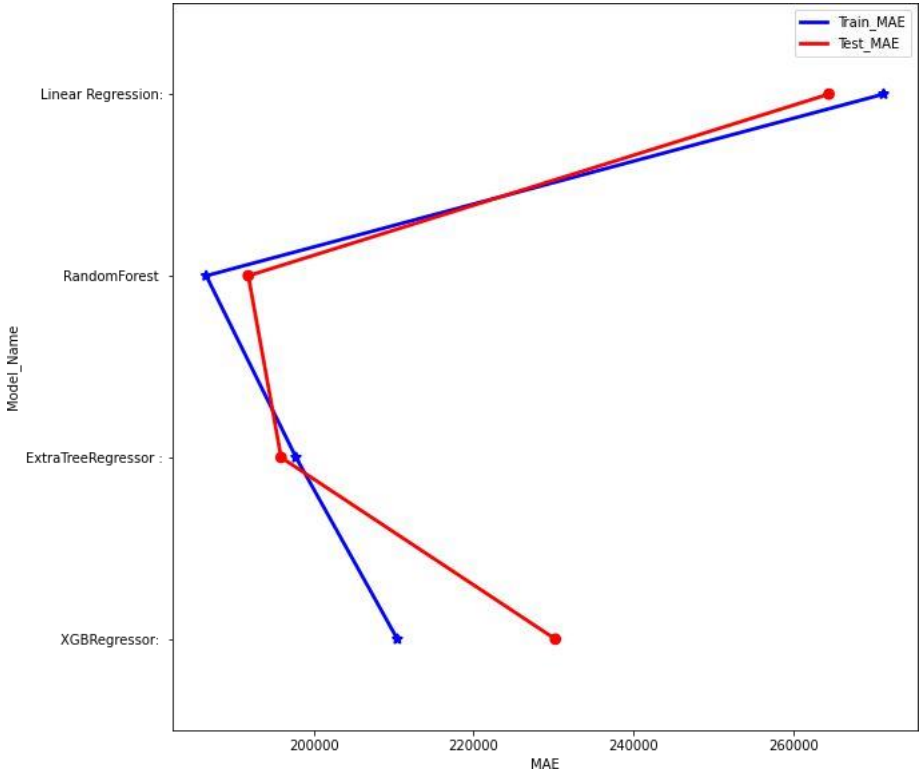
Feature importance wrt XGBoost Regressor:

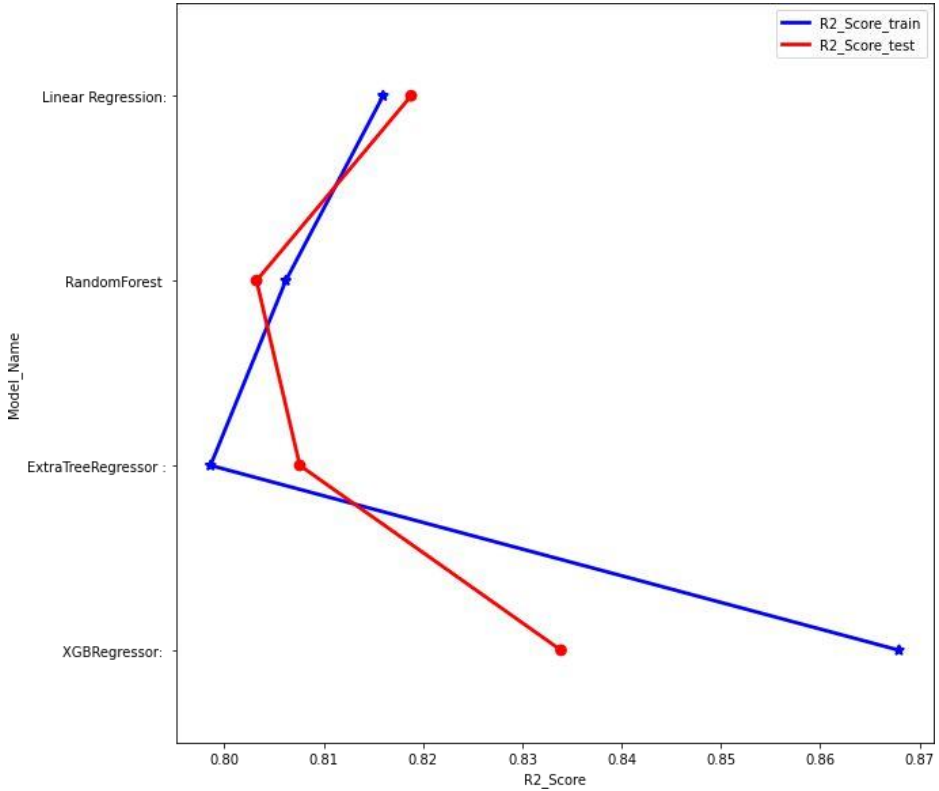


Feature importance wrt Random Forest Regressor:



Model Comparison:





Which model did we choose and why?

- Out of all these models RandomForestRegressor is the best performer in terms of MAE.
- MAE is the best deciding factor because it isn’t affected by outliers.
- MAE is linear and RMSE is quadratically increasing.

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 needs to be added in the model.
- Choosing the right features for modelling.
- Choosing the right models to get the best scores.

Conclusion

- 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.
- TED can tackle the sectors like Music by inviting more popular speakers in this sectors like ‘OK GO’ in this category.