Capstone Project-2

Ted Talk Views Prediction

Individual Contributor

Nargis Nasreen

Content

- Problem Statement
- Data Summary
- EDA on features
- Feature Engineering
- Data Cleaning
- Feature Selection
- Models used for training
- Observation
- 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 uploaded videos on TEDx website.

Data Summary

Dataset name: data_ted_talks.csv

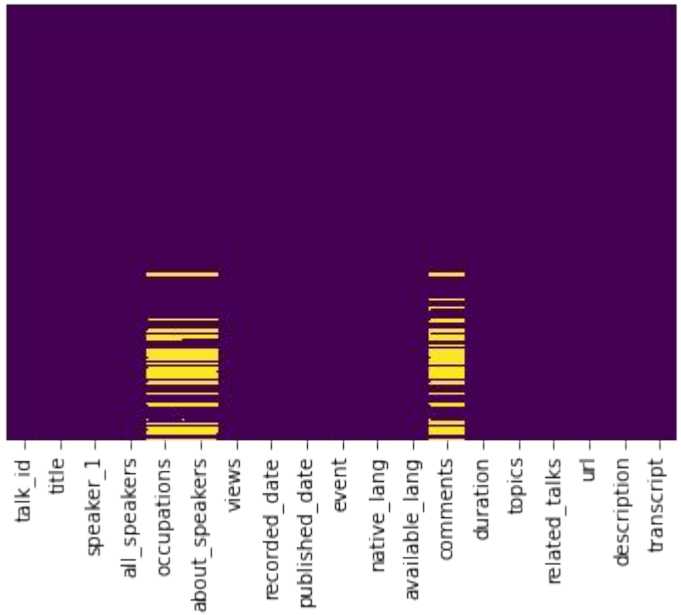
Shape:

- 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

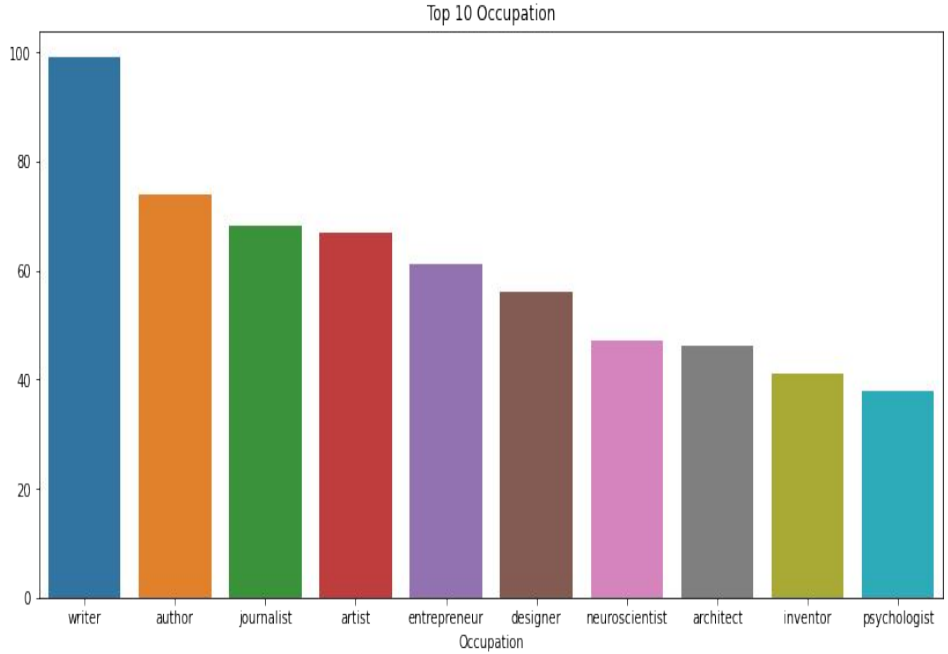
Target Variable: Views

EDA on features



The dataset contains NaN values in 4 columns:

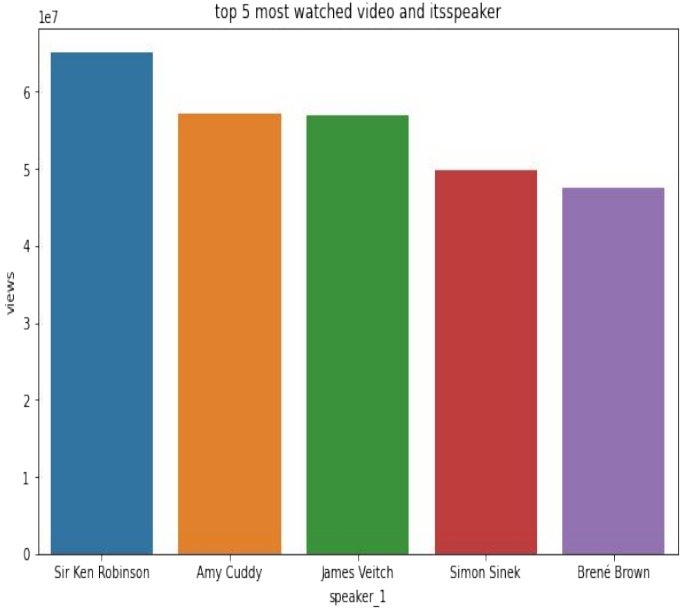
All_speakers, occupations, about_speakers, and comments

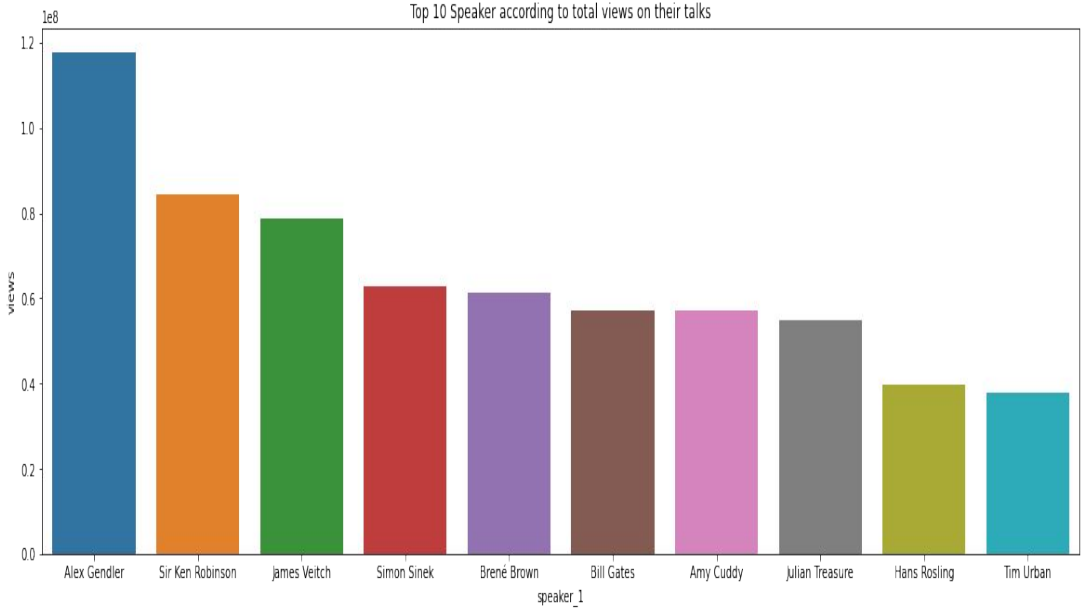


Top 10 occupation of speakers.

Most of the speakers are writer followed by author and journalist

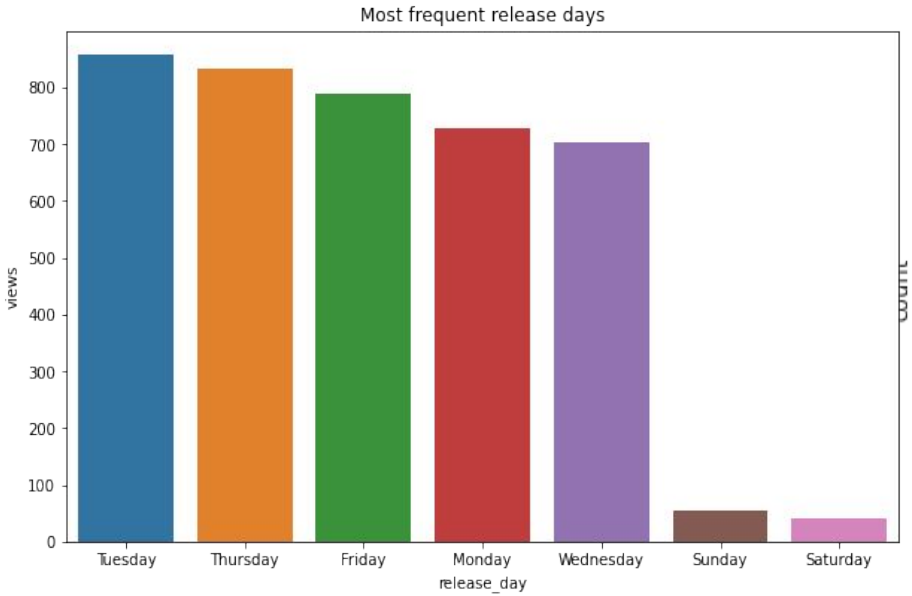
EDA on features

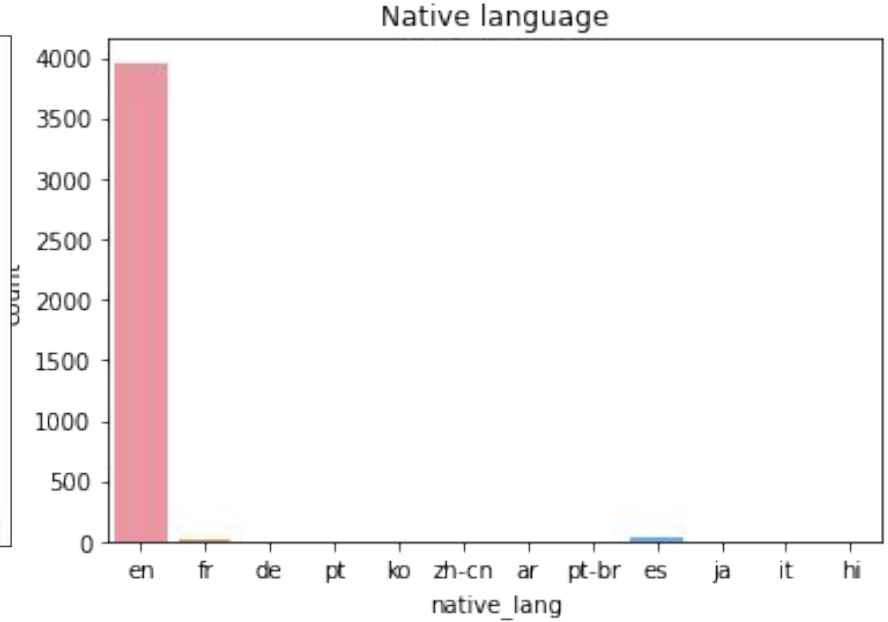




- **Sir Ken Robinson's talk is the most popular TED Talk of with more than 65 million views.**
- **It is only talk that has crossed 60 million views.**
- **Alex Gendler is the most popular speaker followed by Sir Ken Robinson.**

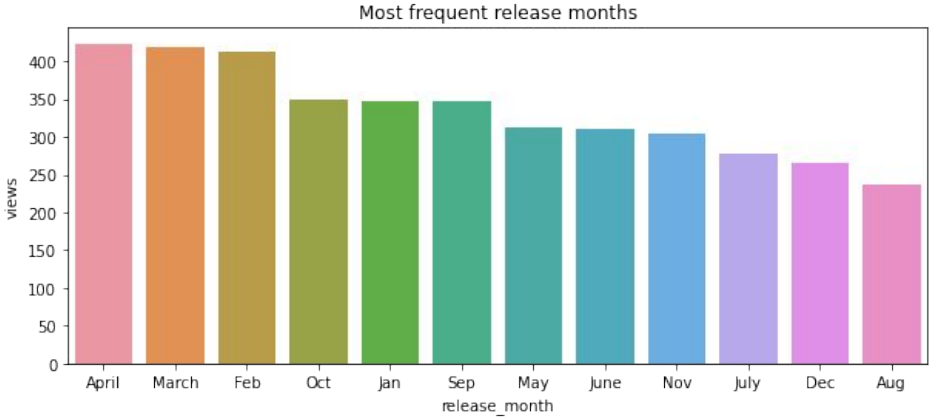
EDA on features





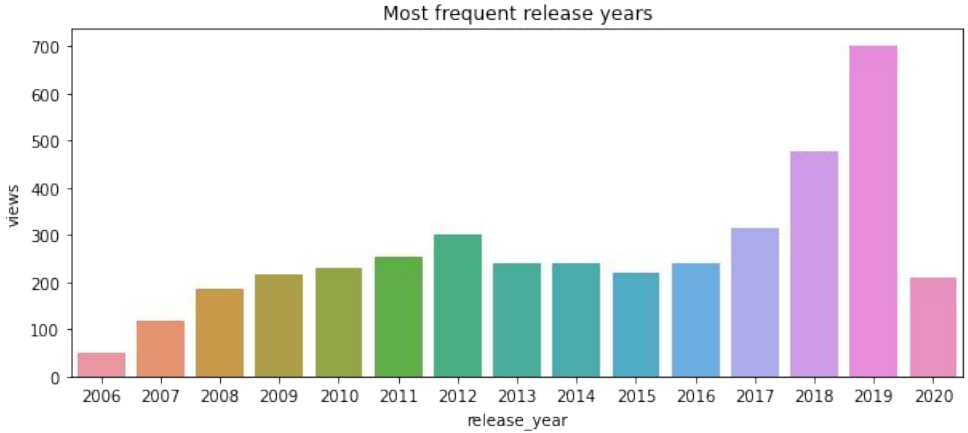
- Here, approx 99% videos having English as native language.
- Most of the videos are released on Tuesday followed by Thursday and Friday.

EDA on features

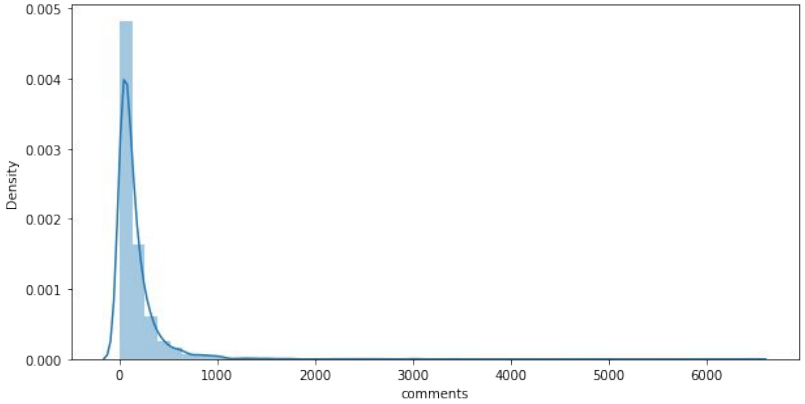


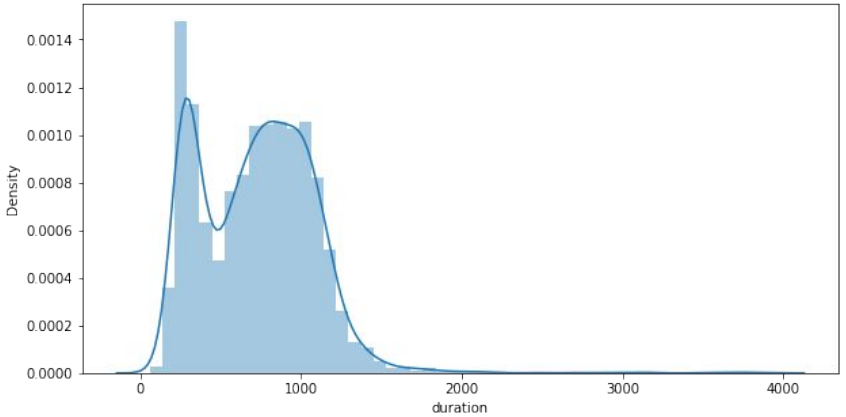
Most videos are released in April followed by March and Feb.

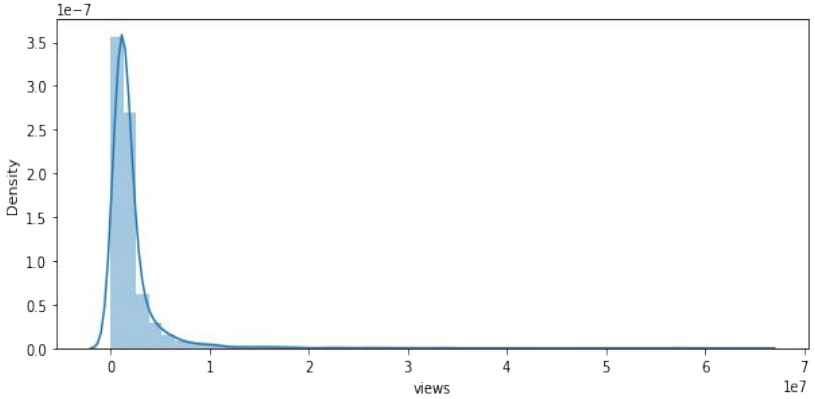
2019 was the most frequent release year.



EDA on features







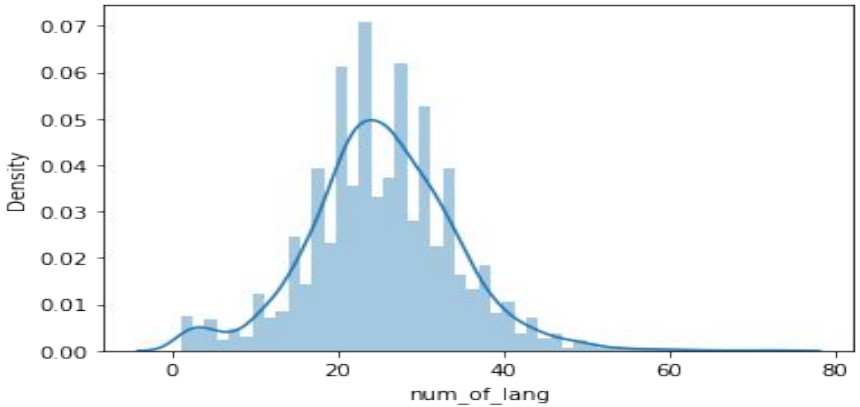
Comments, views and duration are right skewed.

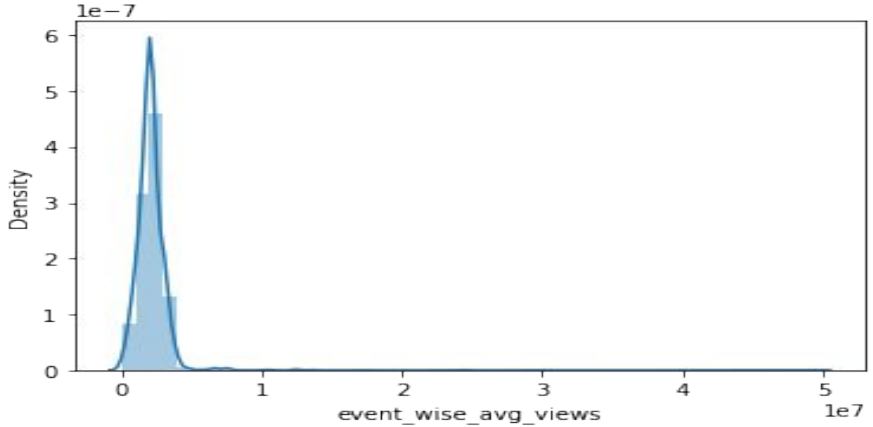
Range of count of most of the comments are 0-1000.

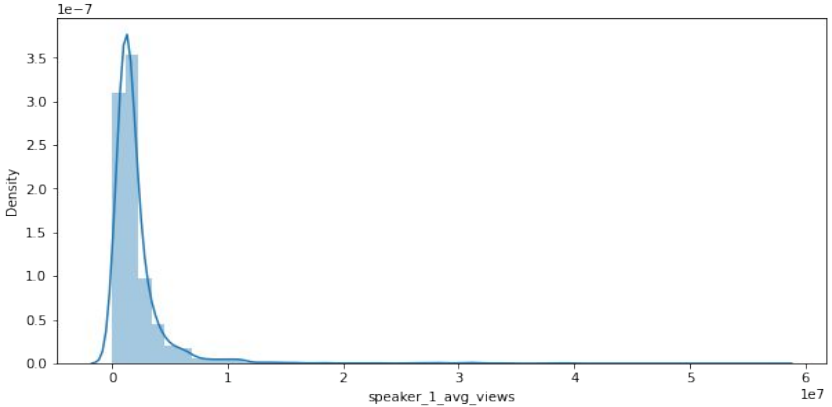
Range of views of most of the videos are 0-1M.

Range of duration of most of the videos are 0-2000 minutes.

EDA on features







Most of the videos are available in 20-25 different languages as num_of_lang follows gaussian distribution.

There are few events that is affecting the number of views and videos get more than 1M views.

There are few speakers that is affecting the number of views and videos get more than 1M views.

Feature Engineering

I did data manipulation and added some new feature for better analysis:

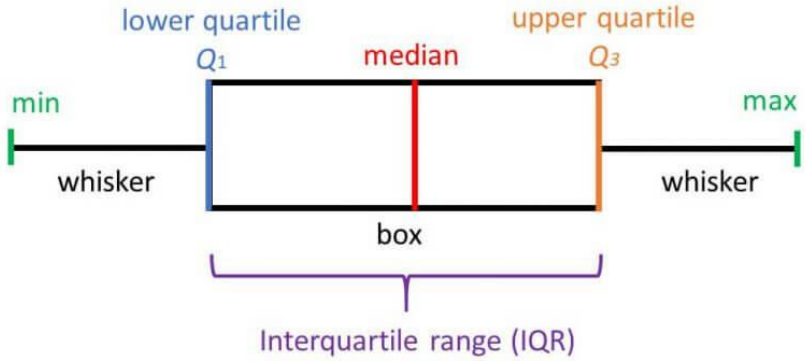
- speaker_1_avg_views
- event_wise_avg_views
- num_of_lang
- video_age
- release_month
- release_day

Data Cleaning

Steps followed for data cleaning:

1. Replaced the NaN values of ‘Occupation’ columns with ‘Others’.
2. Replaced the NaN values of ‘comments’ column using ‘KNNImputer’.
3. Then, I handled the outliers present in columns: 'comments','duration','num_of_lang','views','speaker_1_avg_views','event_wise_avg_views' using IQR and replaced the outliers with extreme values.

IQR=Q3-Q1
Lower Bound=Q1-(1.5*IQR)
Upper Bound=Q3+(1.5*IQR)



Data Cleaning

4. After handling the outliers I removed the unnecessary columns like:

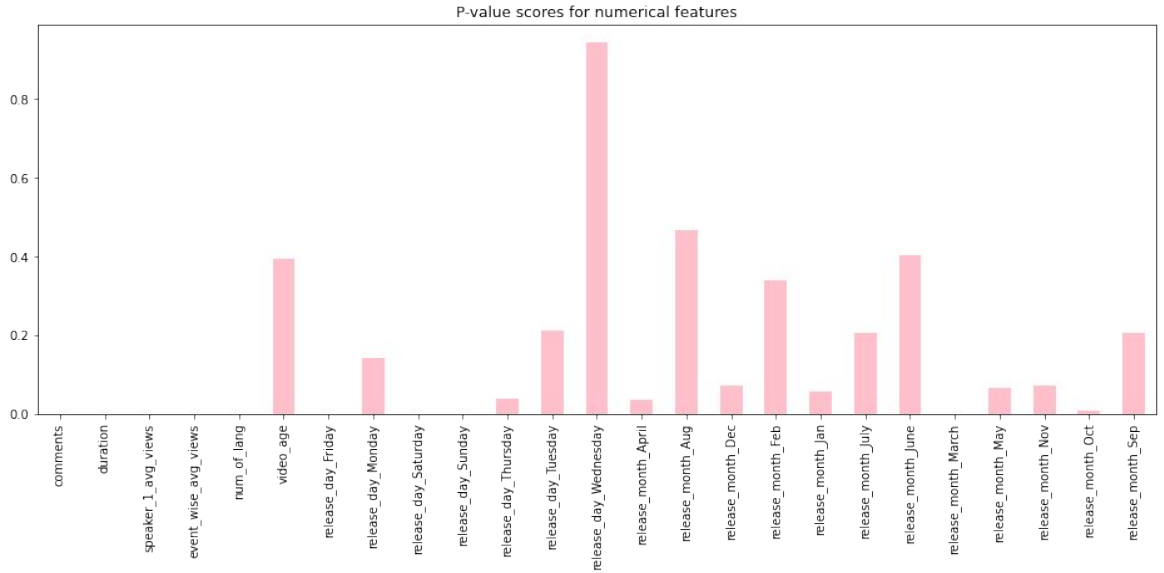
- 'talk_id'
- 'title'
- 'speaker_1'
- 'all_speakers'
- 'occupations'
- 'event'
- 'recorded_date'
- 'topics'
- 'related_talks'
- 'native_lang'

- 'transcript',
- 'description',
- 'occupation',
- 'release_year',
- 'about_speakers'
- 'url'
- 'available_lang'
- 'published_date'

Feature selection

To do feature selection I used f-regression technique

After plotting p-values with features I selected the features with small p-values and leave the remaining.



Selected features:

'comments', 'duration', 'num_of_lang',

'release_day_Friday','speaker_1_avg_views', 'event_wise_avg_views'

Models used for training

- Linear Regression
- Lasso Regression
- Ridge Regression
- Random Forest Regression
- Xgboost Regression

I applied hyperparameter tuning on random forest regression and xgboost regression to get better performance.

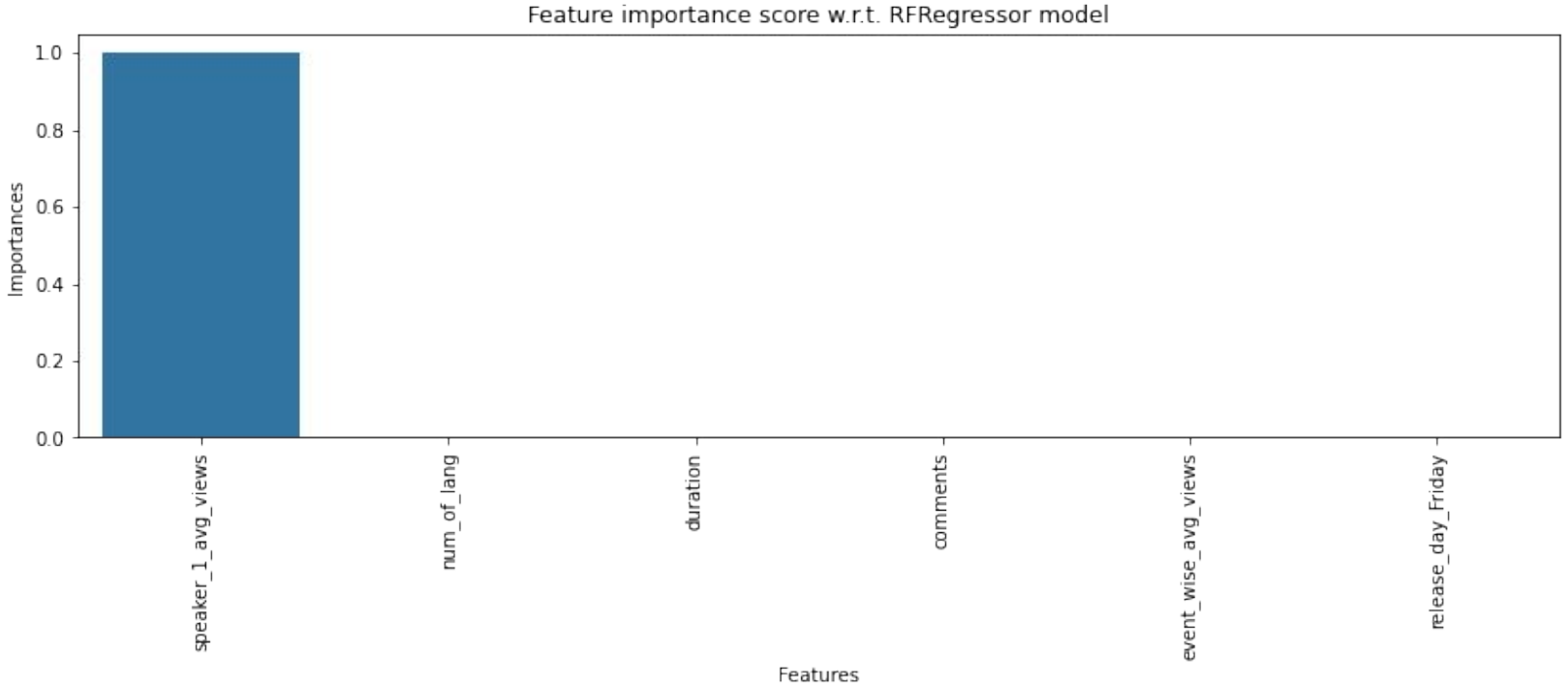
Models used for training

After hyperparameter tuning:

Estimator	Random Forest Regressor	Xgboost Regressor
R_sq. for train	0.805132	0.908343
R_sq for test	0.800267	0.835552
MAE train	203703.785753	169262.805385
MAE test	210610.538512	220553.085433
RMSE train	486698.622559	333788.787008
RMSE test	492614.788850	446989.161918

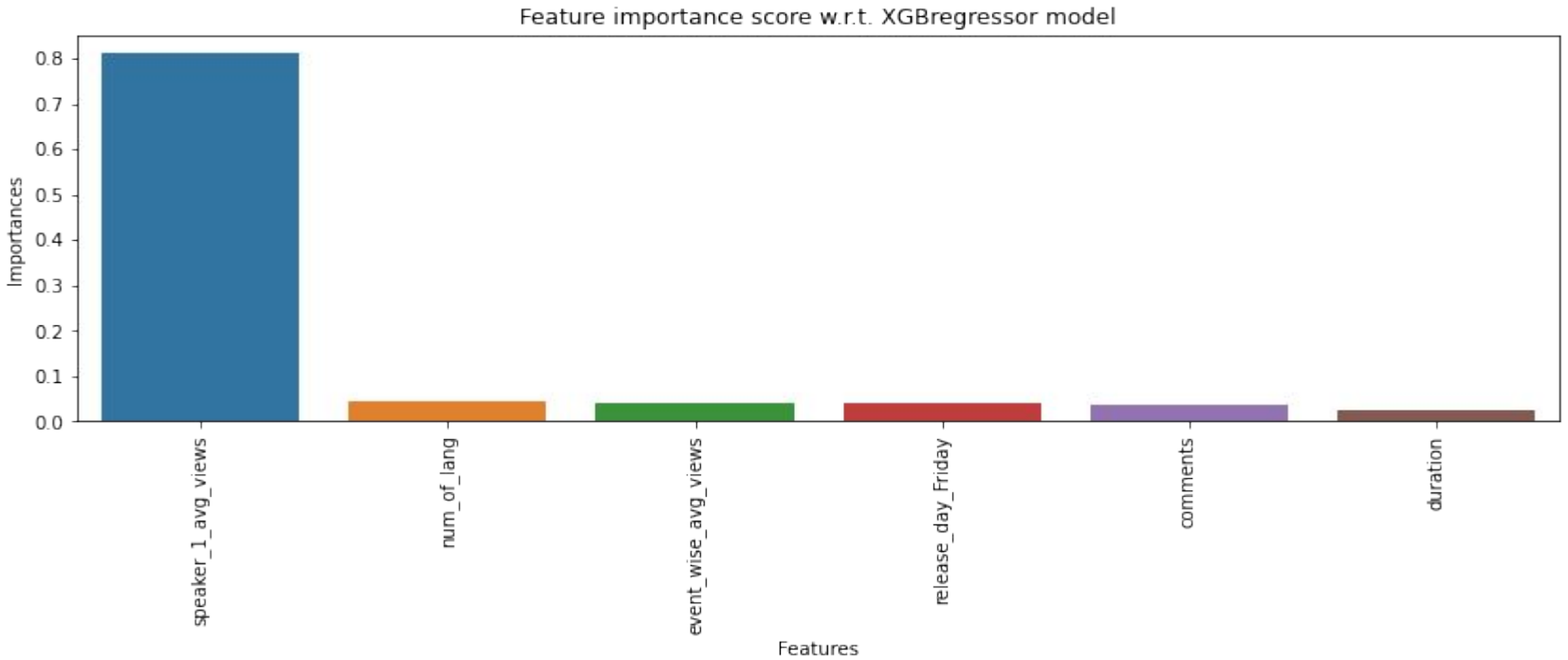
Models used for training

Feature Importance w.r.t Random Forest Regressor:



Models used for training

Feature Importance w.r.t Xgboost Regressor:



Observation

- Out of all these models RandomForestRegressor is the best performer in terms of MAE.
- MAE is not affected by outliers.
- MAE is linear.
- In all of these models our errors have been in the range of 2,00,000 which is around 10% of the average views. We have been able to correctly predict views 90% of the time.
- After hyper parameter tuning, we have prevented overfitting and decreased errors by regularizing and reducing learning rate.
- Given that only have 10% errors, our models have performed very well on unseen data due to various factors like feature selection,correct model selection,etc.
- In all the features speaker_wise_avg_views is most important this implies that speakers are directly impacting the views.

Conclusion

- I built a predictive model which will help to predict the number of views for the video uploaded on TEDx website.

- TEDx can increase views and popularity by increasing videos on section like science, technology, entertainment, etc.

Thank you