Capstone Project

TEDtalk Views Prediction

Individual contributor

Mohammed Javeed

CONTENTS

- Introduction
- Problem Statement
- Loading the data and data cleaning
- EDS on features
- Feature engineering
- Feature selection
- Models used
- Conculusion
- Challenges

INTRODUCTION

- 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.

PROBLEM STATEMENT

Our objective is to predict the views of a TED talk that’s been uploaded in the TEDxwebsite. For this we are provided with a data set “data_ted_talks”. This data set contains information about:

- talk id and title of the TED talks
- Speakers and their occupations who had given TED talks
- Recorded and published date of TED talks
- Event on which TED talks were held
- Native and available languages for the respective TED talks
- Topics, duration and comments of the TED talks
- URL, description and transcript of the TED talks

LOADING THE DATA AND DATA CLEANING

After loading the data, we can observe that the data frame contains 4005 rows with 19 variables.

Data set name: **data_ted_talks**

Shape:

- **Rows -- 4005**
- **Columns--19**

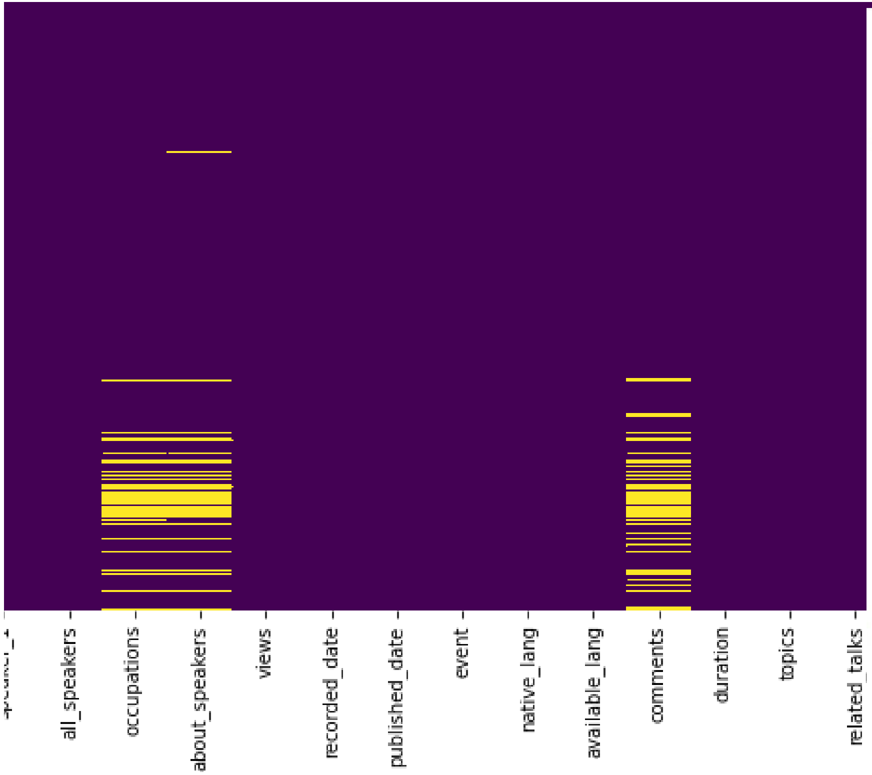
Features:

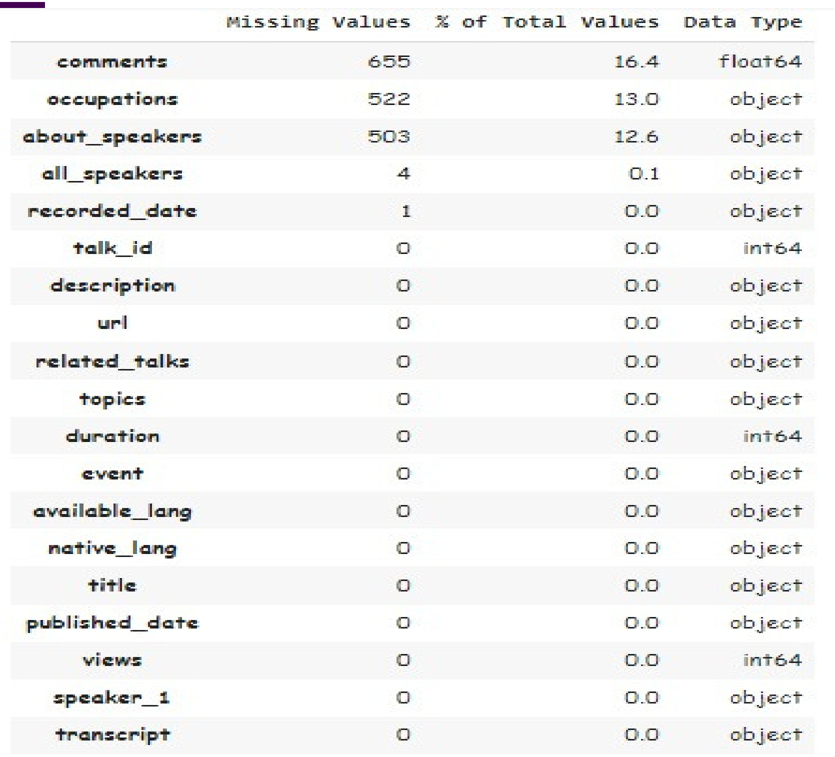
'talk_id', 'title', 'speaker_1', 'all_speakers', 'occupations', 'about_speakers', 'recorded_date', 'published_date', 'event', 'native_lang', 'available_lang', 'comments', 'duration', 'topics', 'related_talks', 'url', 'description', 'transcript'

Target Variable: **'views'**

Exploratory Data Analysis on Features

SPREAD OF MISSING VALUES

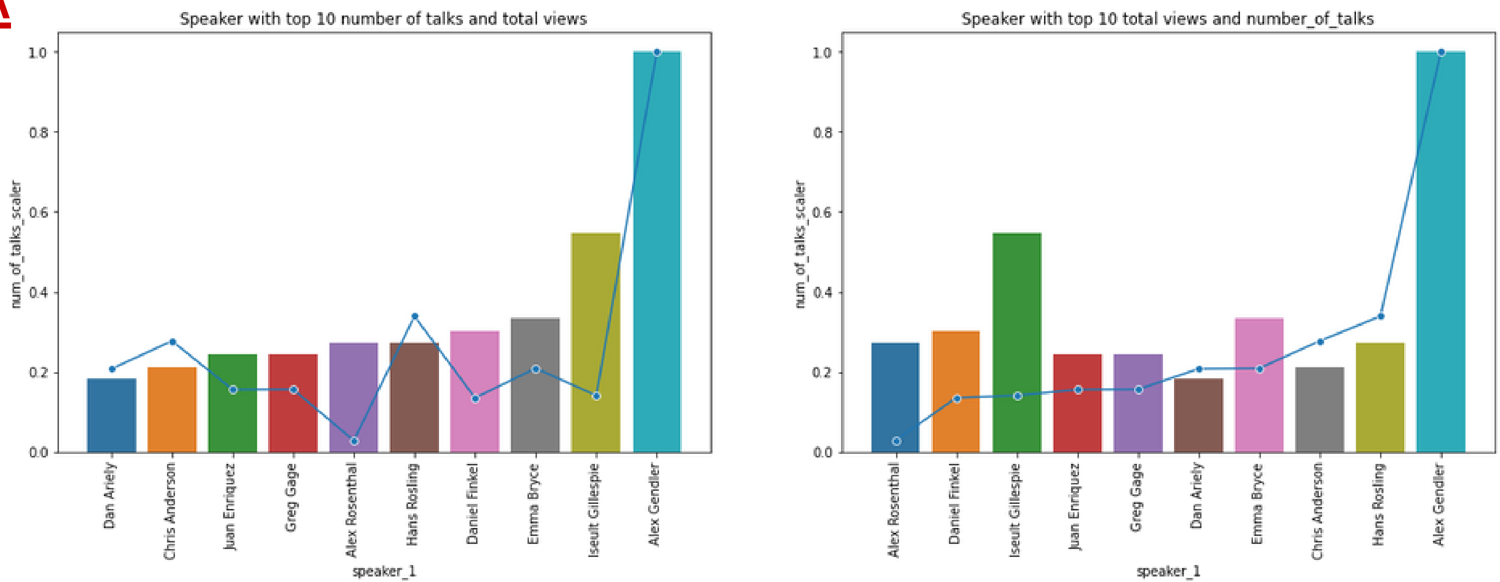




	Missing Values	% of Total Values	Data Type
comments	655	16.4	float64
occupations	522	13.0	object
about_speakers	503	12.6	object
all_speakers	4	0.1	object
recorded_date	1	0.0	object
talk_id	0	0.0	int64
description	0	0.0	object
url	0	0.0	object
related_talks	0	0.0	object
topics	0	0.0	object
duration	0	0.0	int64
event	0	0.0	object
available_lang	0	0.0	object
native_lang	0	0.0	object
title	0	0.0	object
published_date	0	0.0	object
views	0	0.0	int64
speaker_1	0	0.0	object
transcript	0	0.0	object

EDA

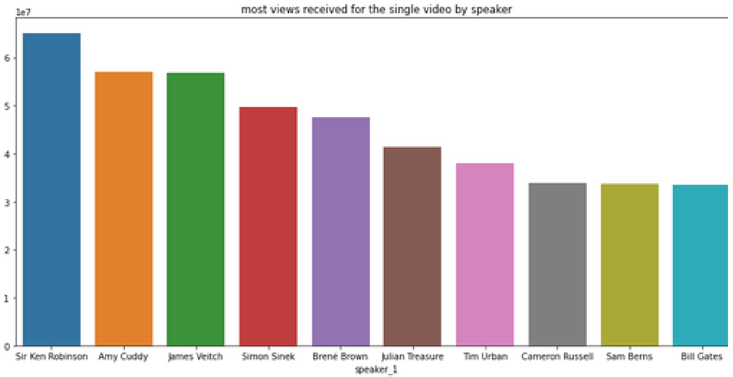
Speakers with top 10 total views with respect to number of talks



Speaker with top 10 number of talks and total views

Speaker with top 10 total views and number_of_talks

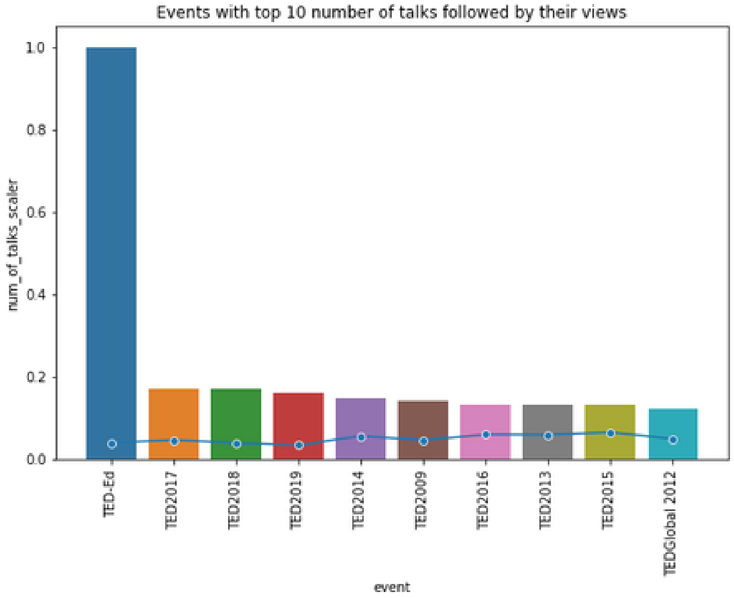
Speakers with max number of views for a single TED talk



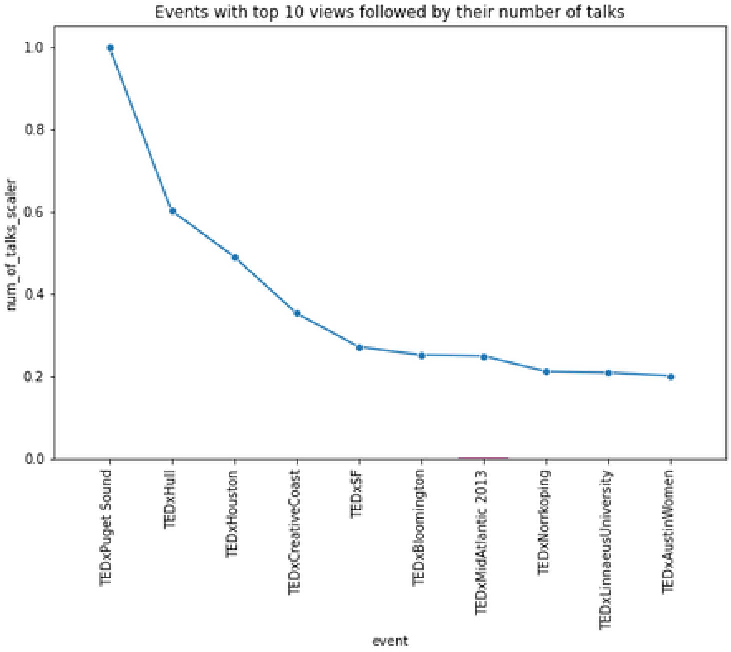
most views received for the single video by speaker

EDA (continued)

Events with top 10 number of talks with respect to views

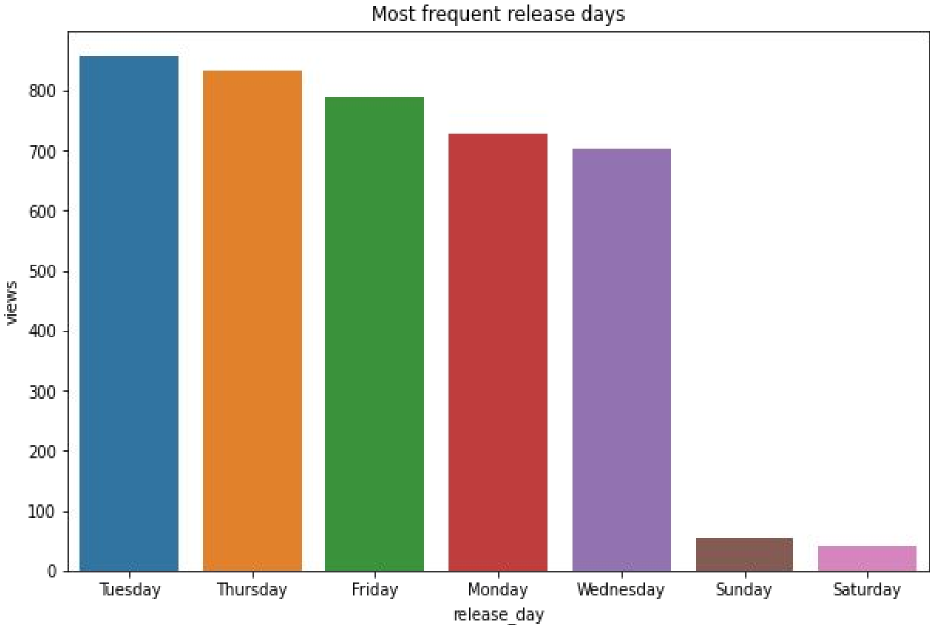


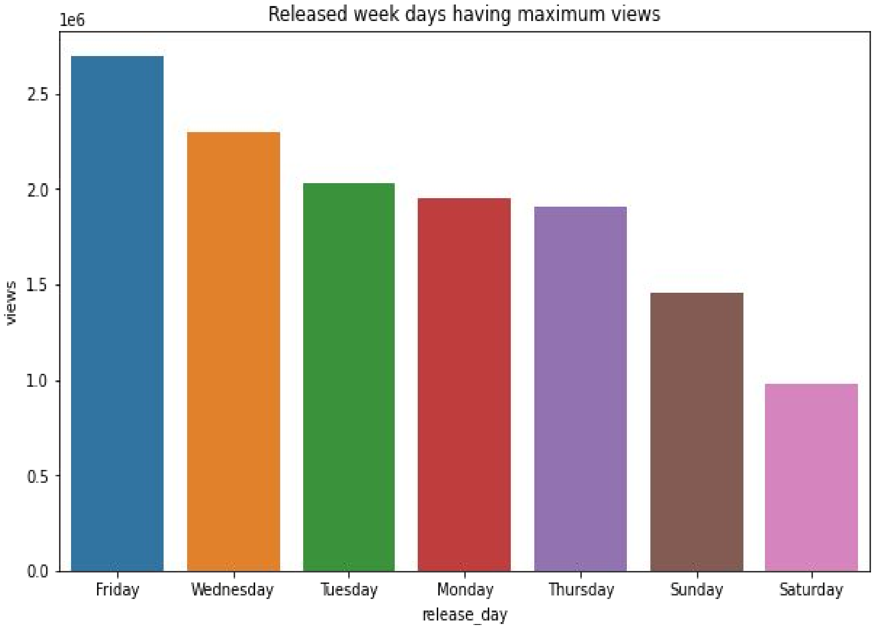
Events with top 10 number of talks followed by their views



Events with top 10 views followed by their number of talks

Published Days with Views:



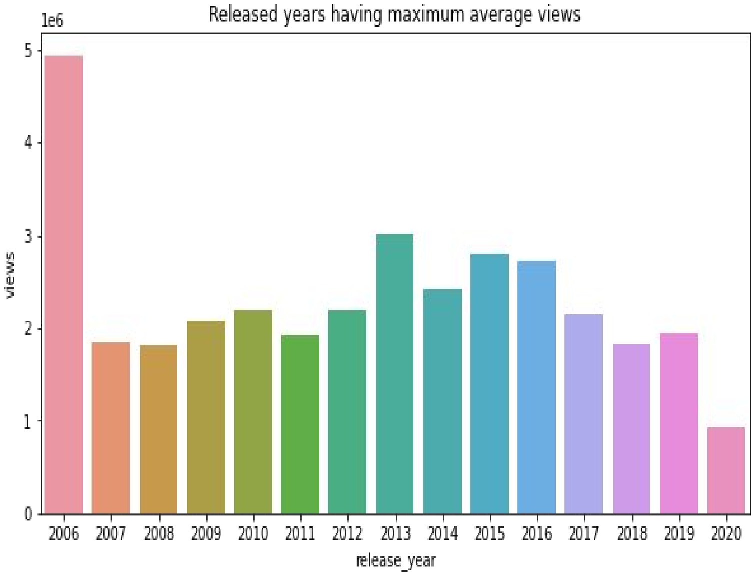


Frequent Released Days

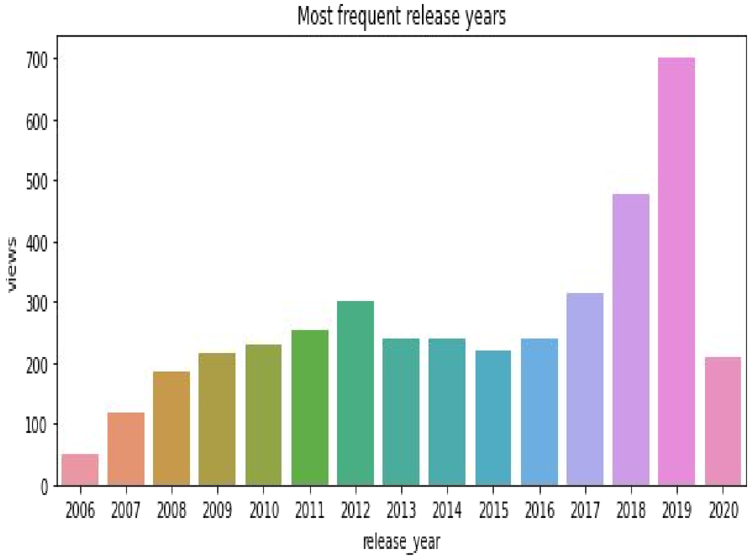
Released Days by avg Views

- **Friday release is impacting the views of the video**

Published Year with Views:



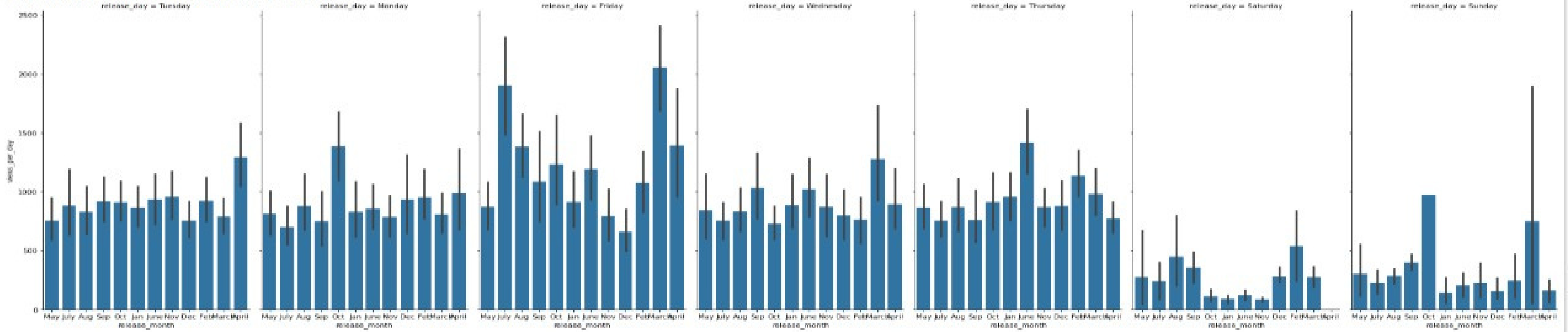
Released years having maximum average views



Most frequent release years

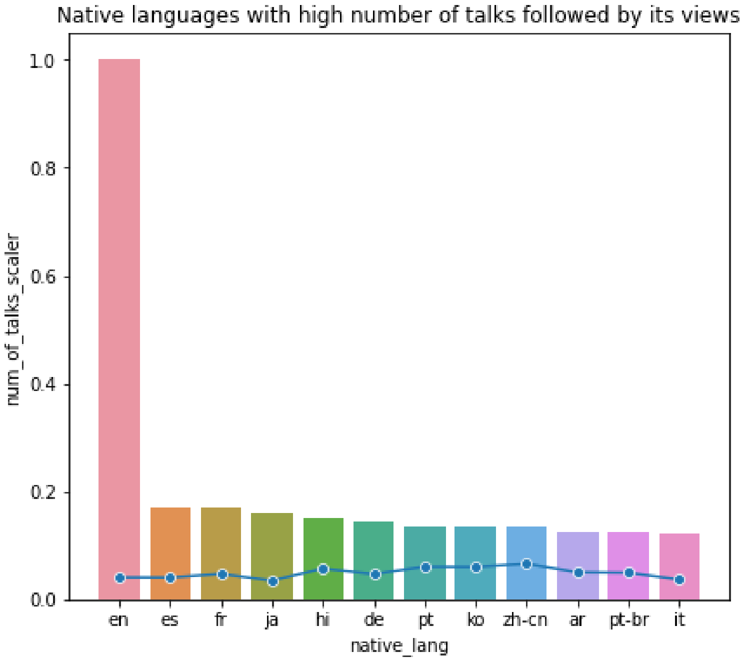
Views per day with respect to every day on monthly basis

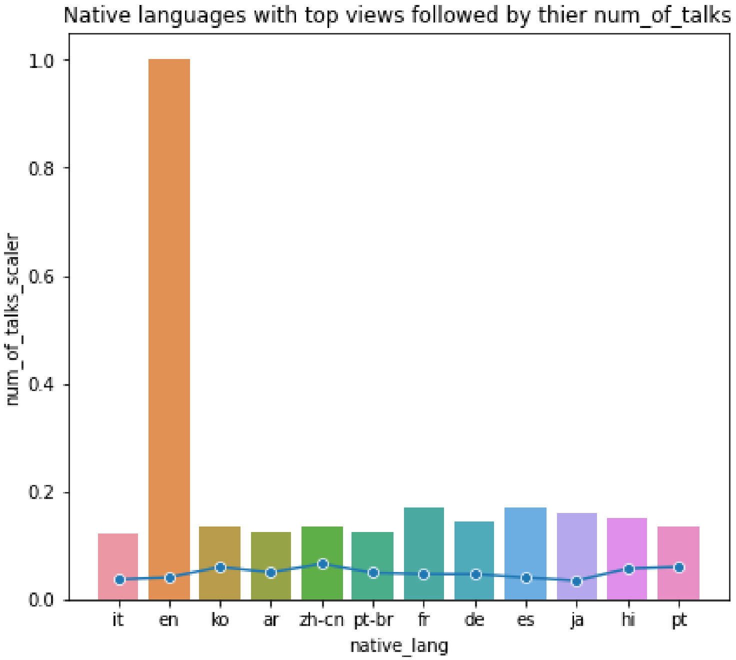
```
<seaborn.axisgrid.FacetGrid at 0x7f59d458b4d0>
```



EDA (continued)

Native languages with views and number of talks





EDA CONCLUSIONS:

- The number of views depends on many points,but the more number of talks does not give the more number of videos.
- Speaker and occupation of the speaker alters the number of views.
- Some speakers who are influencers will contribute a lot for the maximum number of views and thus the occupation.
- People tend to look at the video which was delivered by Psychiatrist, Activist and Authors.
- On weekend and on the month of March there will be surge on number of views. We can see that most of the videos are on the topic 'Science' and 'Technology' English is the language which is available as main language and as well as the subtitles for many of the videos.

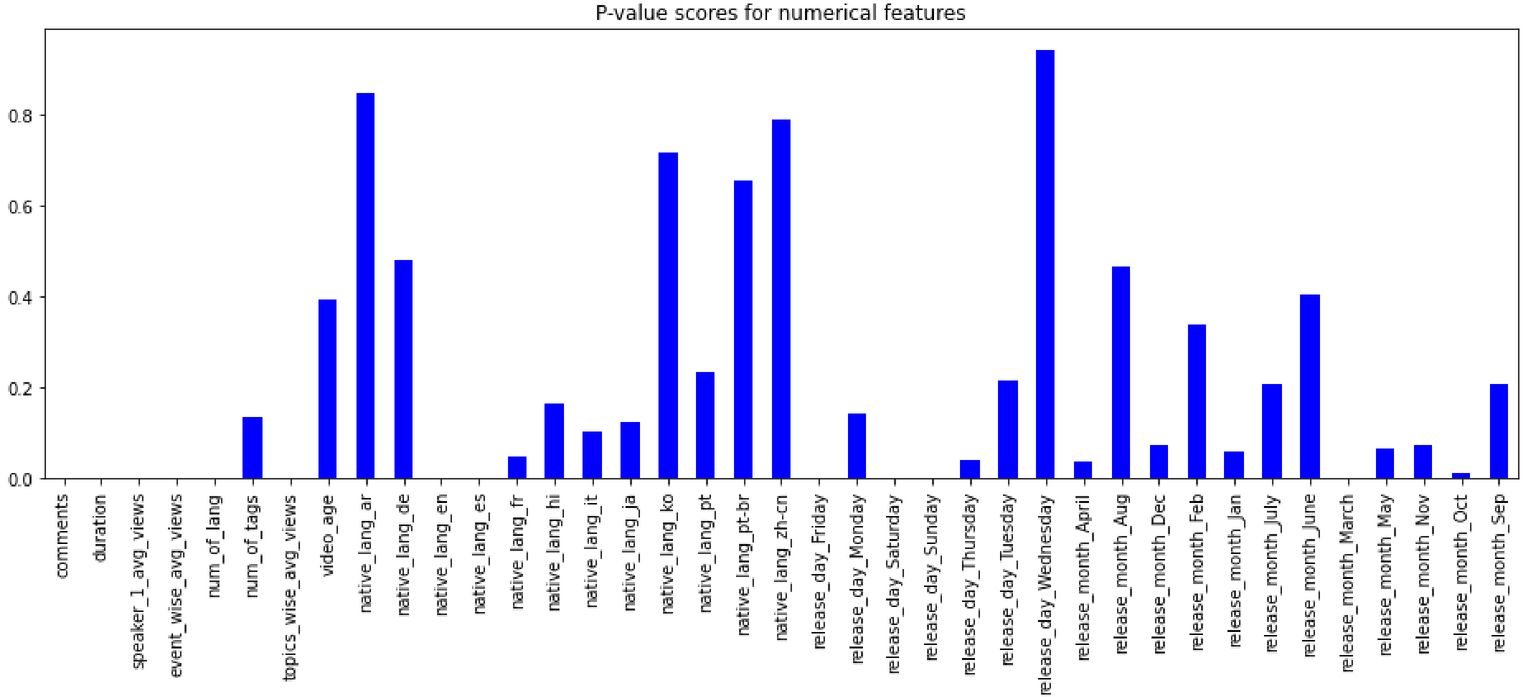
FEATURE ENGINEERING

Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning

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

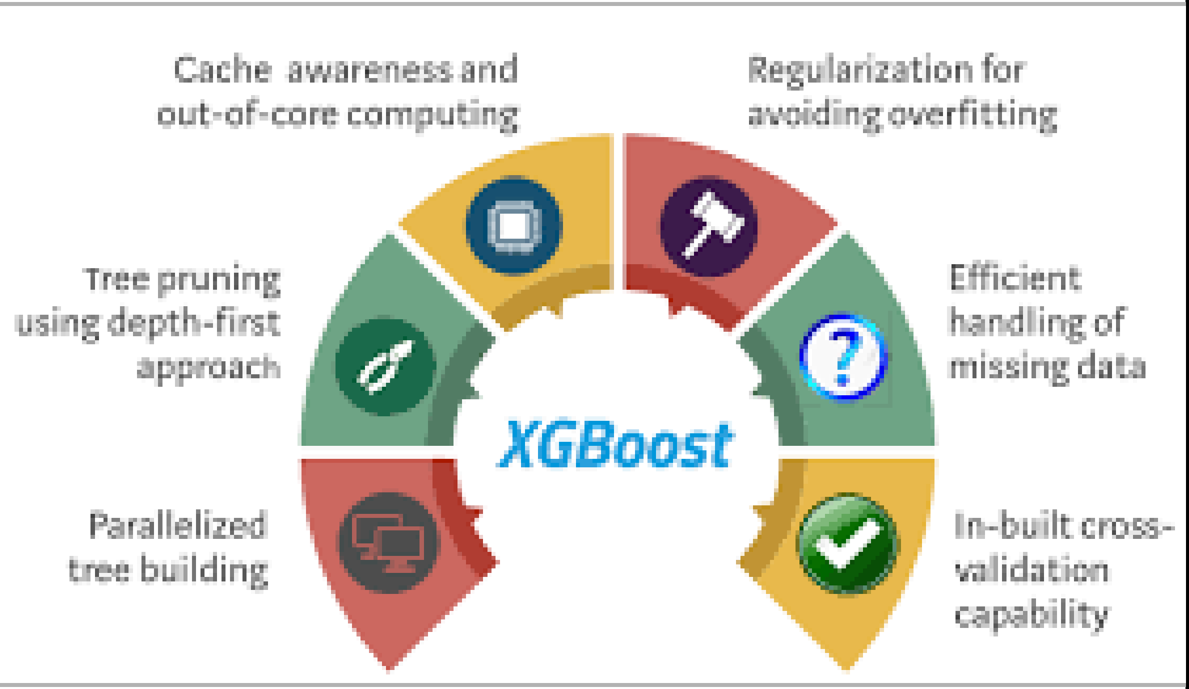
Features selection(f regression):



Feature selection models used:

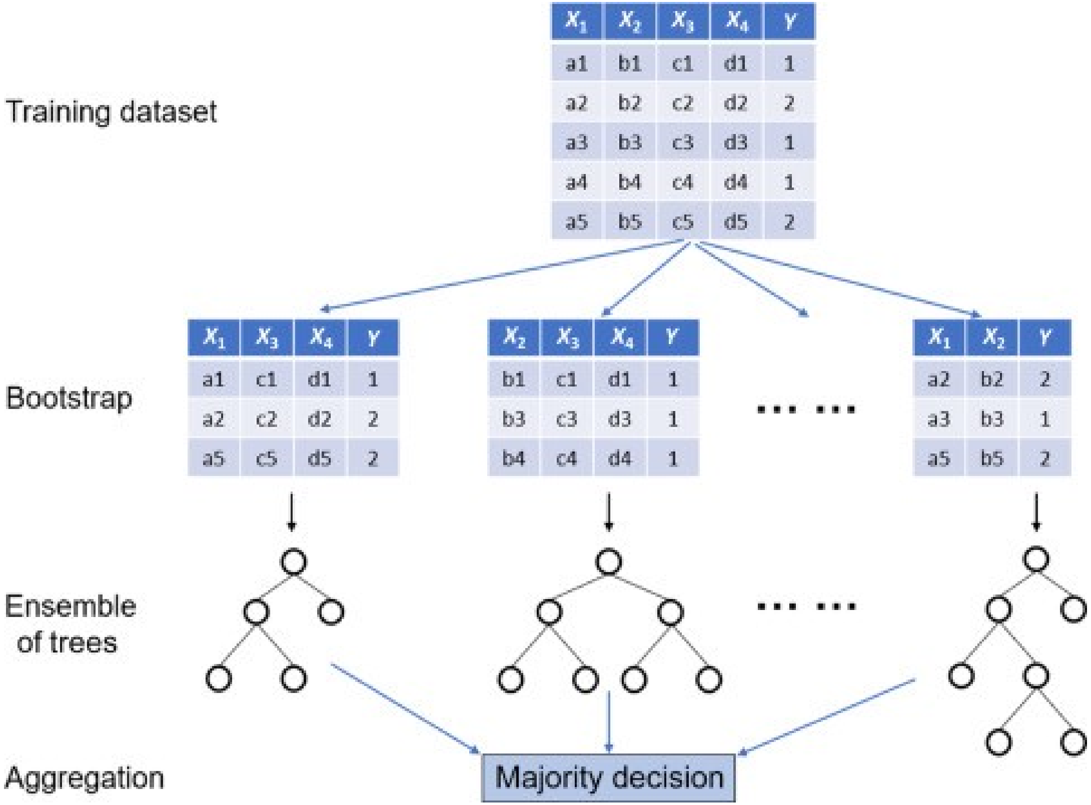
- XGBoost Regressor
- Extra Trees Regressor
- Random Forest Regressor

XGBoost Regressor:


Cache awareness and out-of-core computing
Regularization for avoiding overfitting
Tree pruning using depth-first approach
Efficient handling of missing data
XGBoost
Parallelized tree building
In-built cross-validation capability

Criterion	MAE
R_Square for train	0.9
R_Square for test	0.83
MAE train	166772.39
MAE test	222746.49
RMSE train	321406.05
RMSE test	449587.95

Extra Trees Regressor:



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

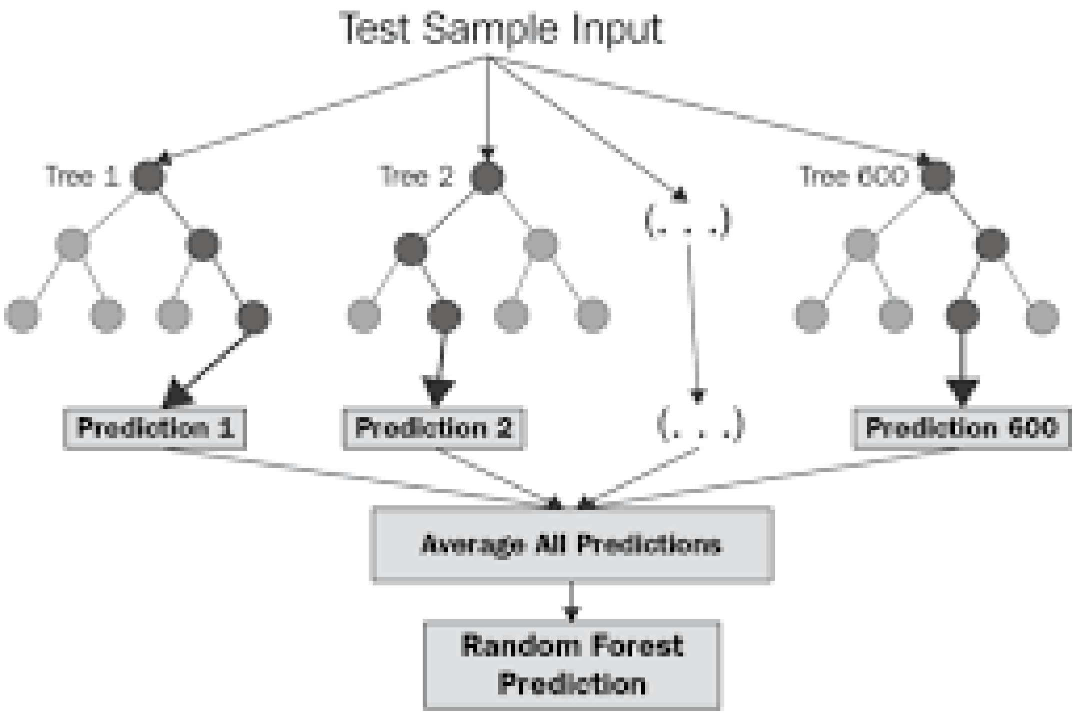
X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

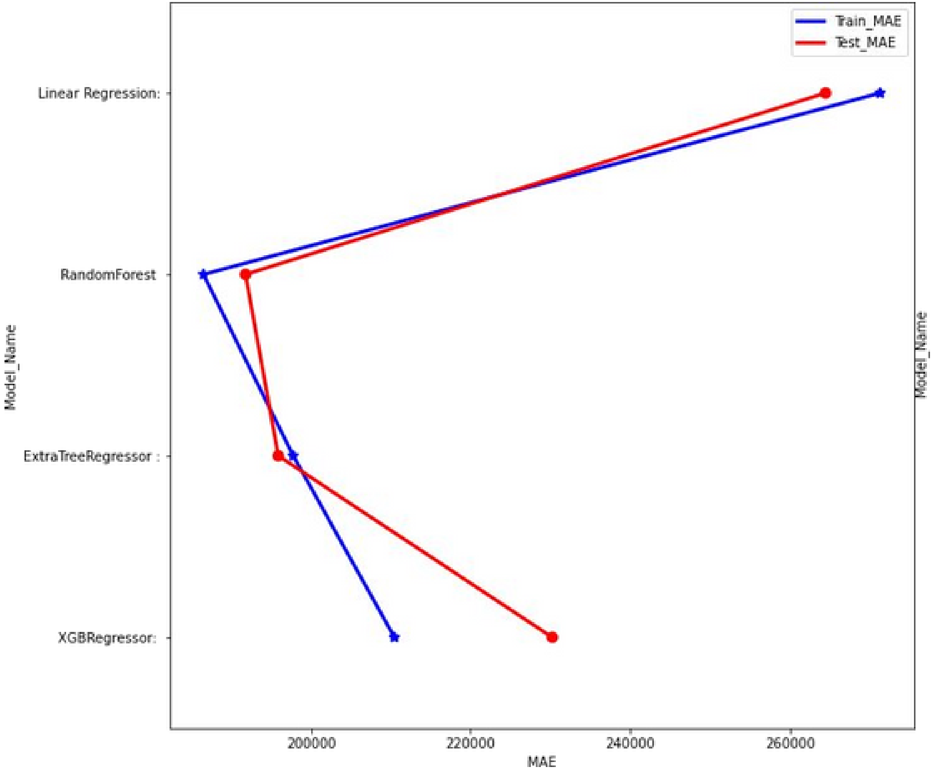
Criterion	MAE
R_Square for train	0.79
R_Square for test	0.8
MAE train	206746.77
MAE test	203522.48
RMSE train	496441.73
RMSE test	484593.54

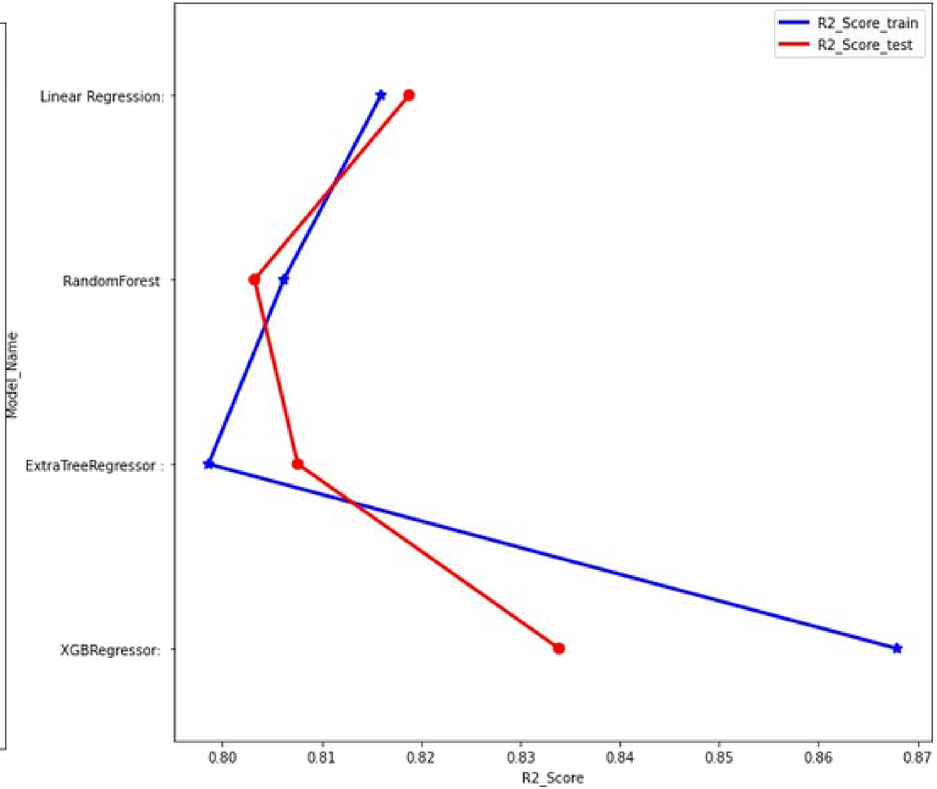
Random Forest Regressor:



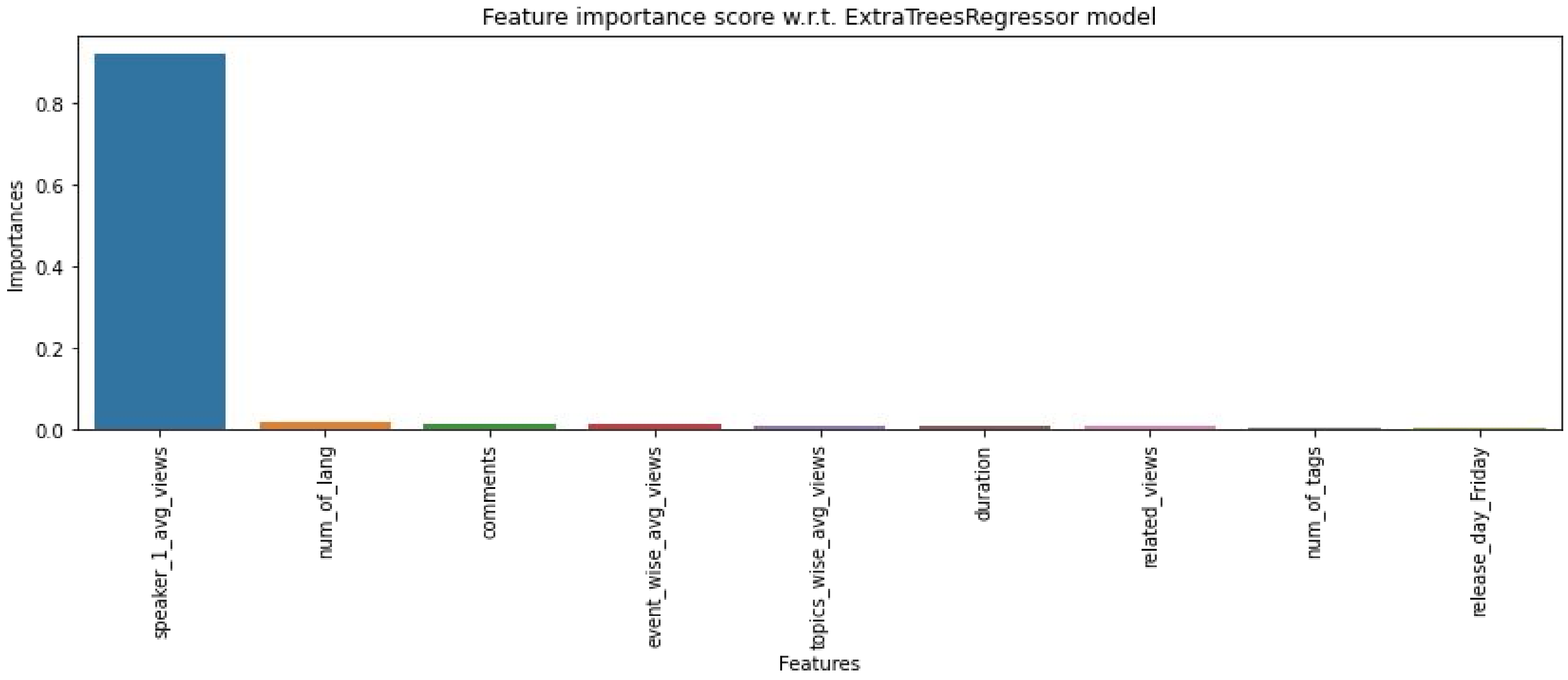
Criterion	MAE
R_Square for train	0.8
R_Square for test	0.8
MAE train	186708.08
MAE test	191720.5
RMSE train	485534.03
RMSE test	488631.91

Model Comparison:

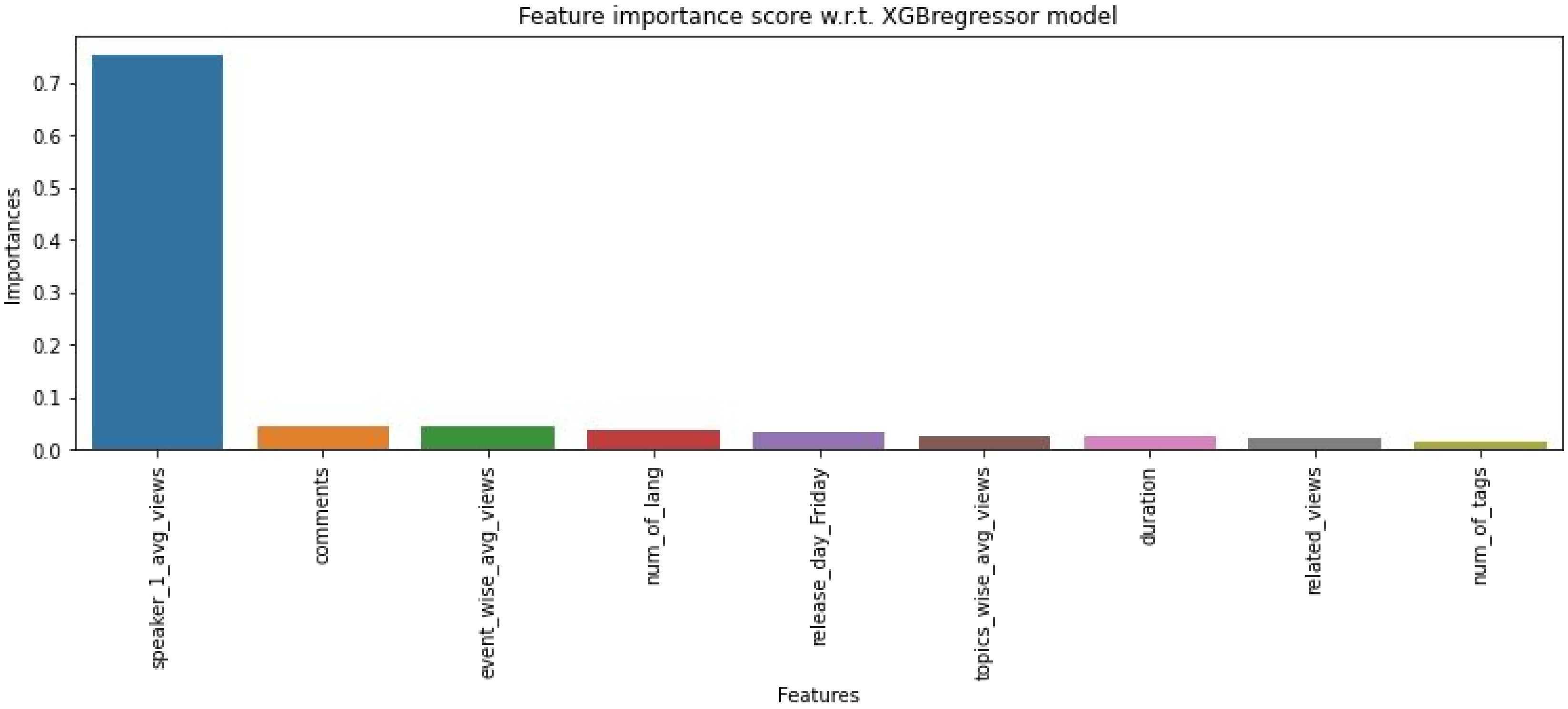




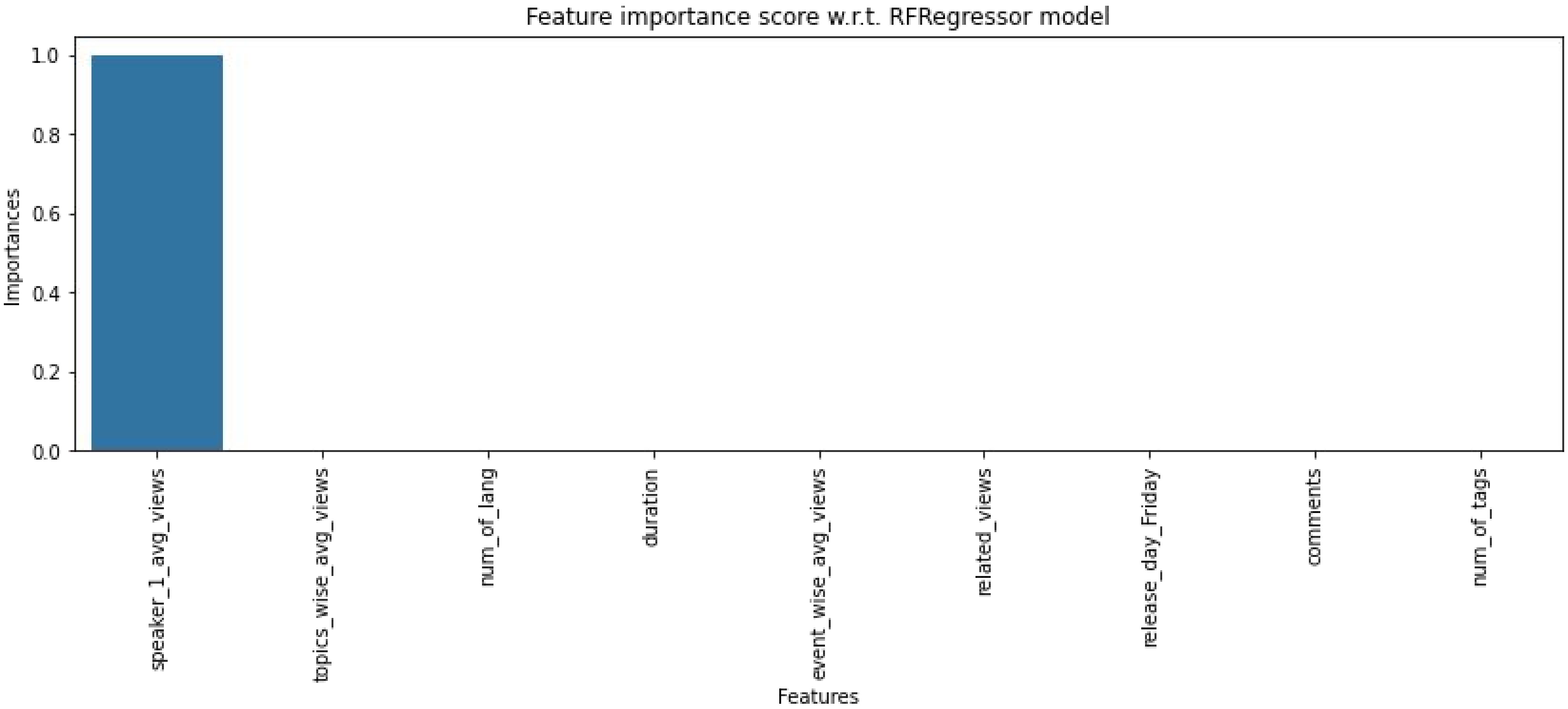
Feature importance wrt Extra Trees Regressor:



Feature importance wrt XGBoost Regressor:



Feature importance wrt Random Forest Regressor:



CONCLUSION

- If we try comparing the prediction accuracy among different linear regression (LR) models then RMSE is a better option as it is simple to calculate and differentiable.And the number of predictor variables in a linear regression model is determined by adjusted R squared.
- As we are more concerned about evaluating prediction accuracy among different LR models we can choose RMSE over adjusted R squared.
- If we compare RMSE, **Optimal Random Forest and as well as Extra Tree** is performing well. But if we consider RMSE along with the adjusted R squared, **Optimal Random Forest** is best performer.

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.

Thank You.