Capstone Project

TED Talk Views Prediction

Ritik Gupta

Contents

Overview

TED is devoted to spreading powerful ideas on just about any topic. 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.

Predicting the potential amount of views has proven to be extremely important for helping organization to understand what type of videos the audience prefers to watch, and to better optimize their tech resources.

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 Description

> Dataset shape –

4005 Rows and 19 columns

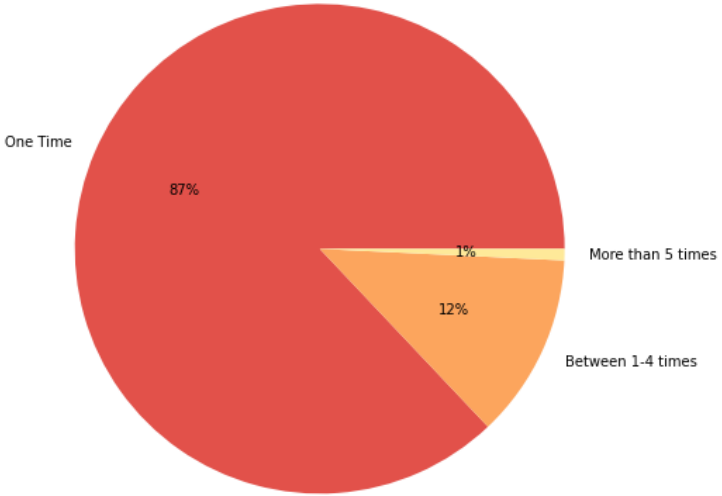
> Given Features-

- ❑ Talk_id
- ❑ Title
- ❑ Speaker_1
- ❑ Speakers
- ❑ Occupations
- ❑ About_speakers
- ❑ Recorded_date
- ❑ Published_date
- ❑ Event
- ❑ Native_lang
- ❑ Available_lang
- ❑ Comments
- ❑ Duration
- ❑ Topics
- ❑ Related_talks
- ❑ URL
- ❑ Description
- ❑ Transcript

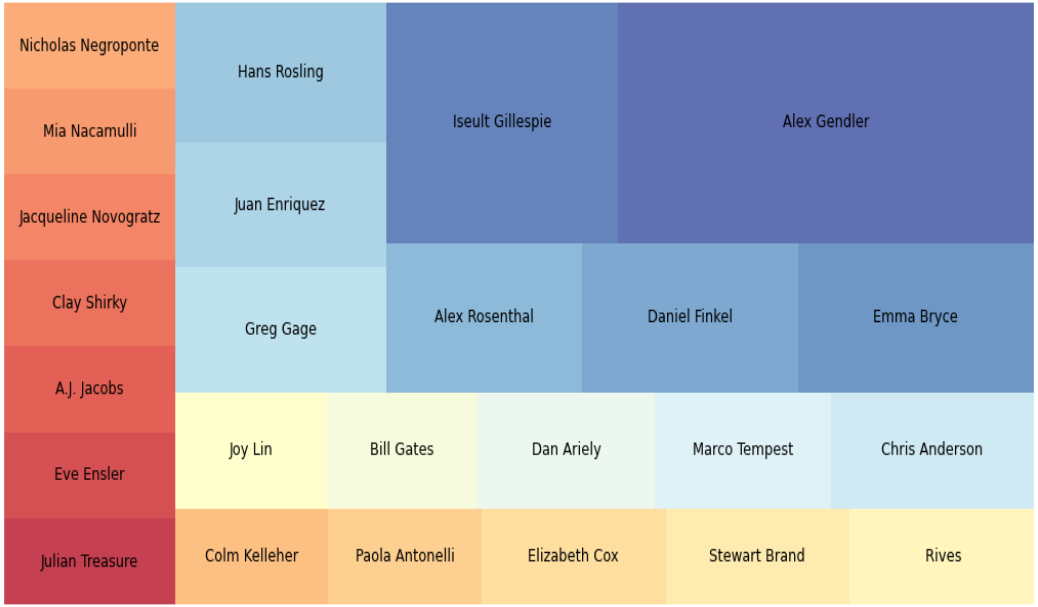
Target Feature:

- ❑ Views

Exploratory Data Analysis

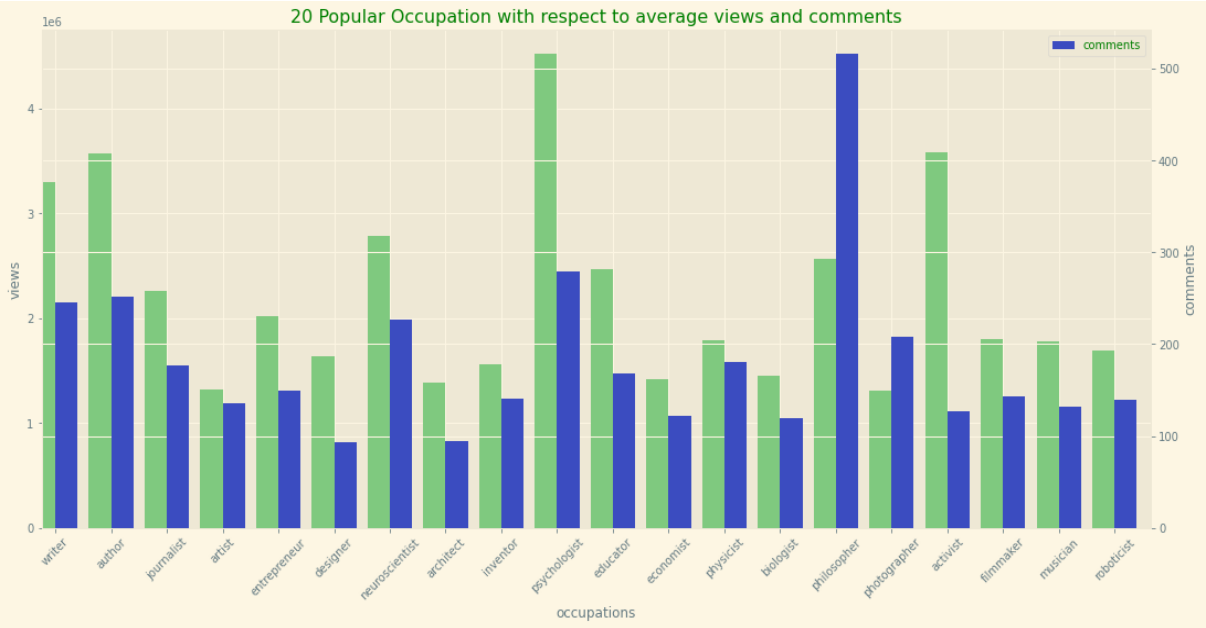


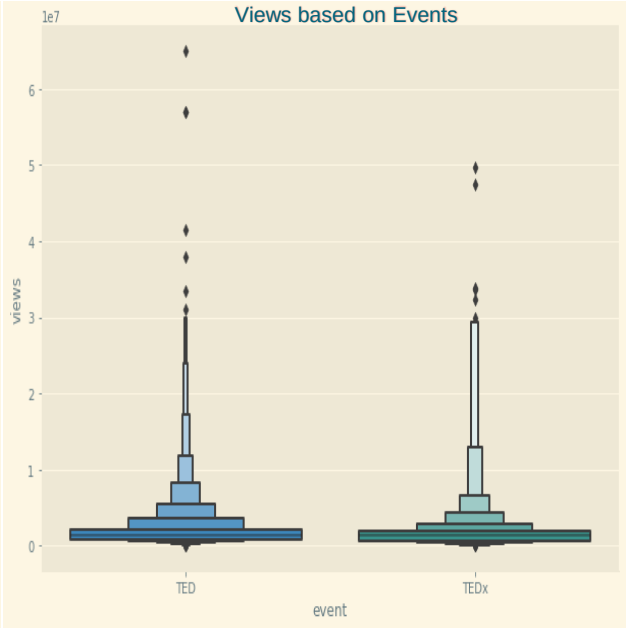
Repetition Rate of Speakers



Top 25 Frequent Speakers

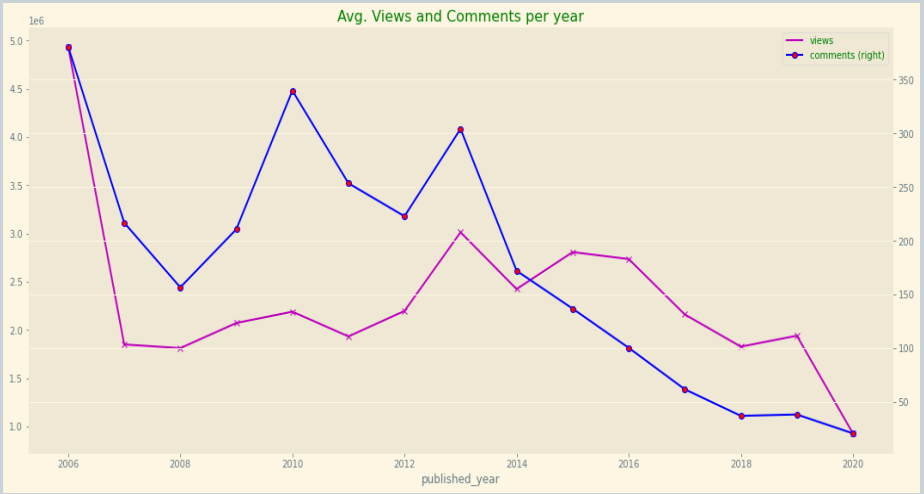
Exploratory Data Analysis





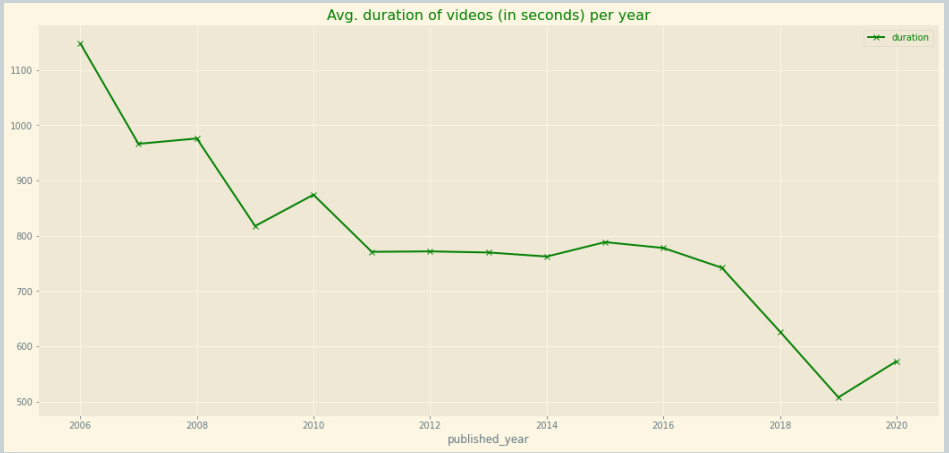
- *It is certain that more views does not means more comments*
- *Psychologist has highest average views*
- *Philosopher has highest comments*
- *TEDx events has slightly more views than TED events*

Exploratory Data Analysis



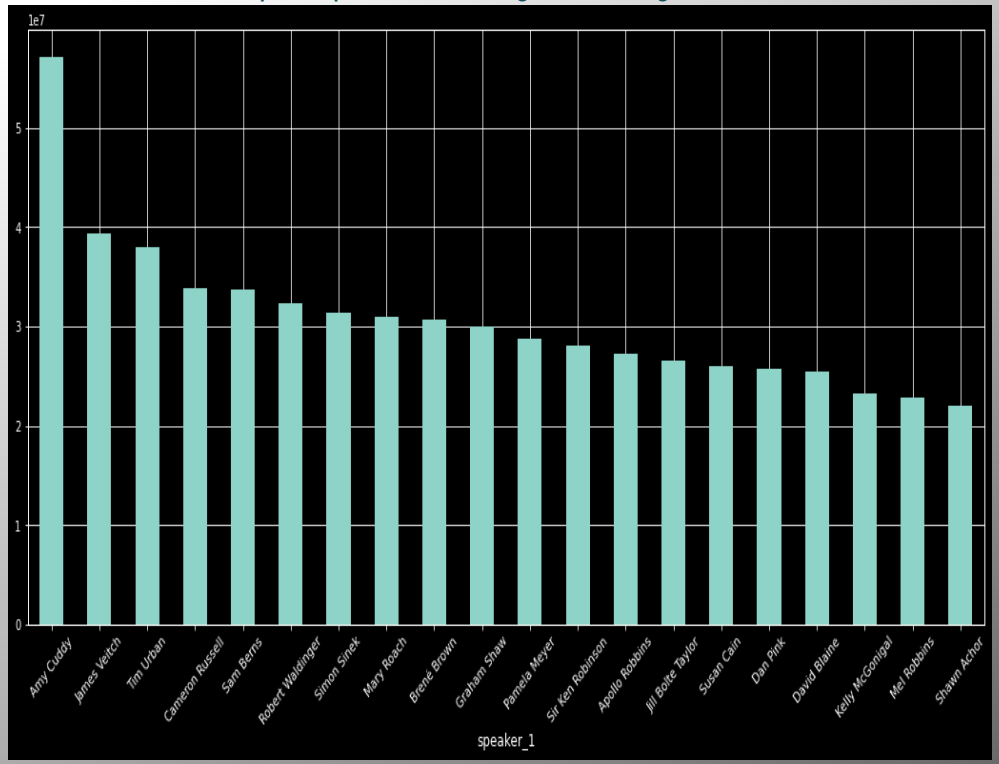
- *After 2013, Comments started declining*
- *After 2016 views are declining*
- *Year 2006 has most average views as well as comments*

- *Average Duration is slowly reducing throughout the years*
- *Duration is reducing as the attention span of users are decreasing with time.*

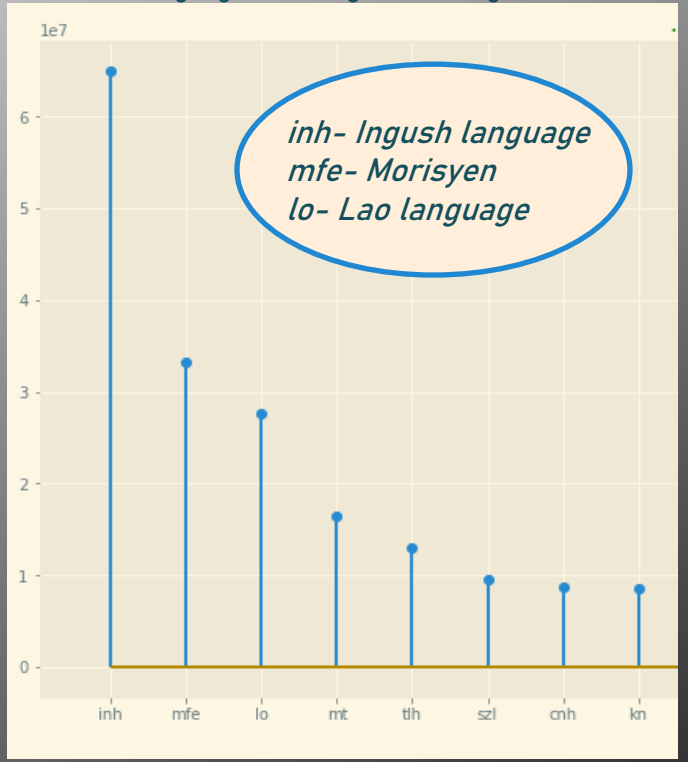


Exploratory Data Analysis

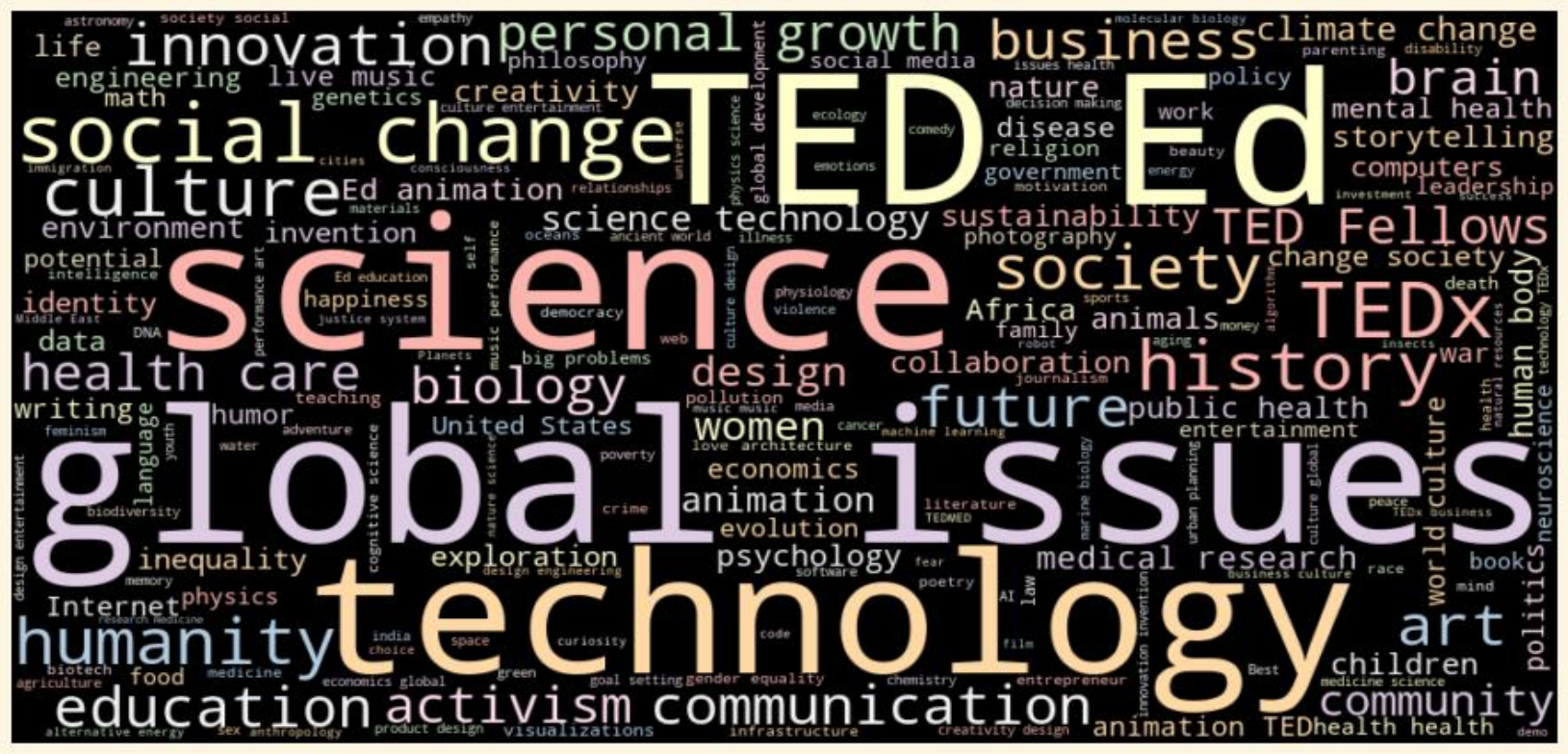
Top 20 Speakers with highest average views



Languages with highest average views



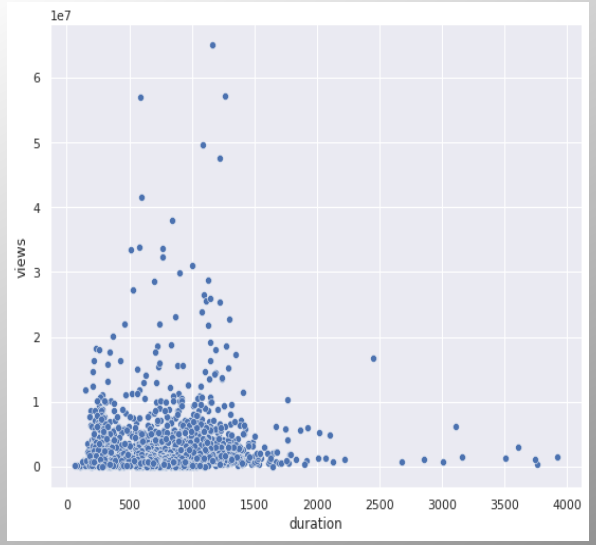
Exploratory Data Analysis

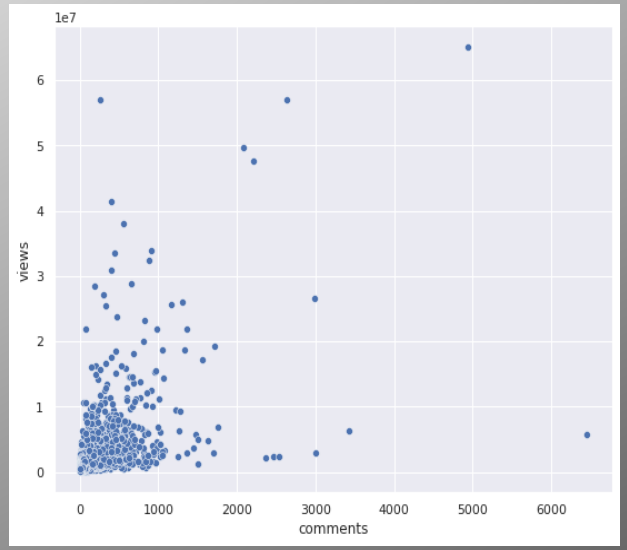


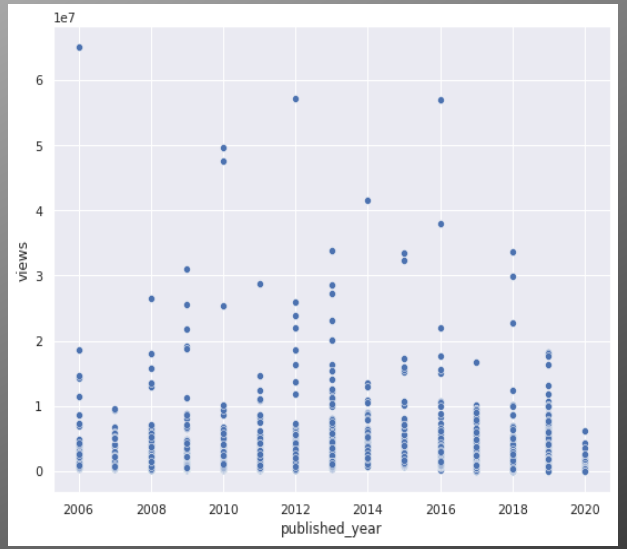
Popular Topics in Talk Events

Exploratory Data Analysis

Numerical Features with respect to Views







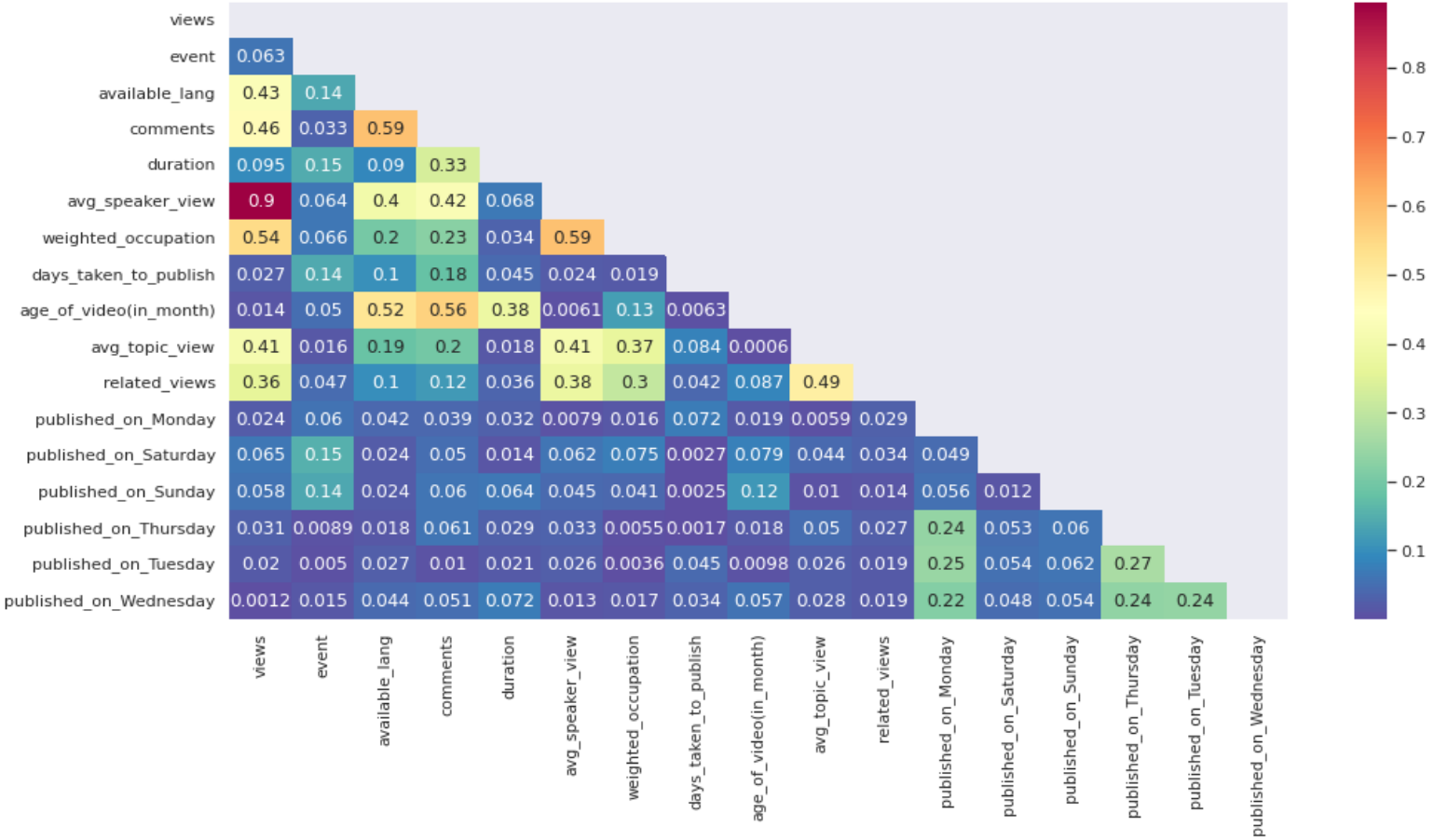
- No given feature has any clear relation with our target variable hence, in this case feature engineering is certainly needed to build a good model, more than improving it.

Feature Engineering

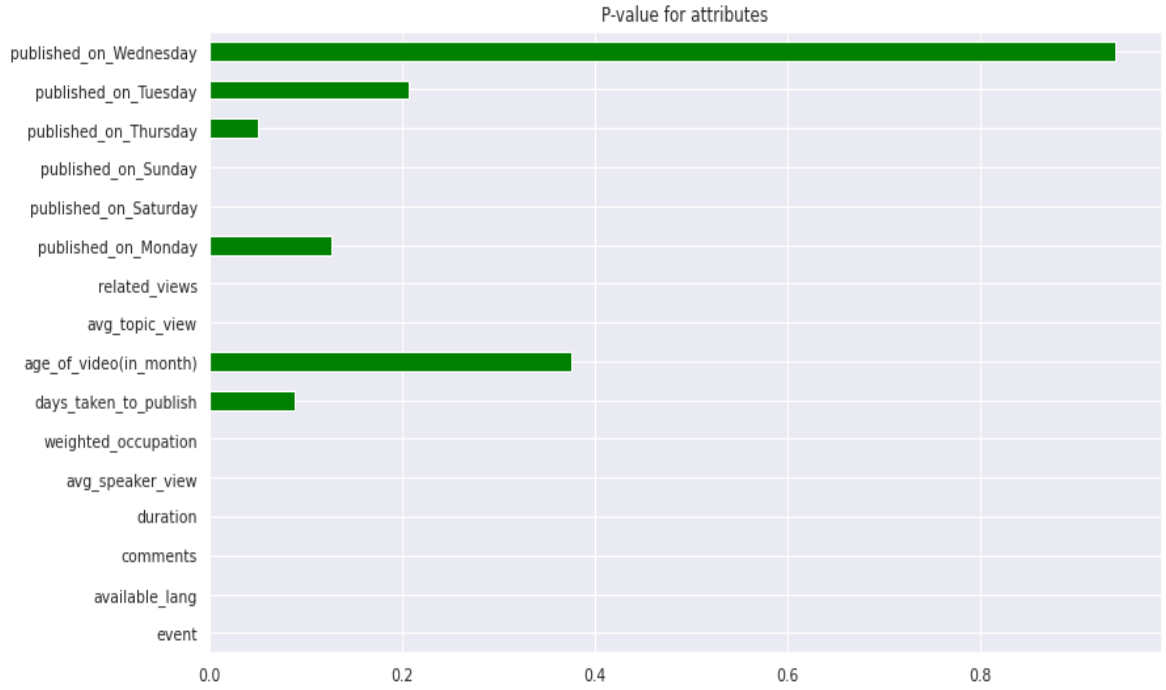
Feature	Details
avg_speaker_view	Derived from speakers as average views of each unique speaker.
weighted_occupation	Given rank from low to high, from average views of each unique occupation
days_taken_to_publish	Difference of days between recorded date and published date
age_of_video(in_month)	Age of video, till the last published video in dataset
published_weekday	Week day of the published date (categorical values)
event	Considering 2 main categories from different events and label encoded them.
available_lang	Count of languages in which the video is available
avg_topic_views	Average of average each unique topic views given for video
related_views	Average views of related talks given for each unique video

Correlation & Selection

Correlation - Heatmap



Feature Selection



- *Higher p-value is a concern*
- *We have to drop:-*
  - *Published_on_wedneday*
  - *Age_of_video(in_months)*

ML Models Used

- Lasso Regression
- Ridge Regression
- Random Forest
- Gradient Boosting
- XG Boost
- K Neighbor Regressor
- Support Vector Machine

Models used

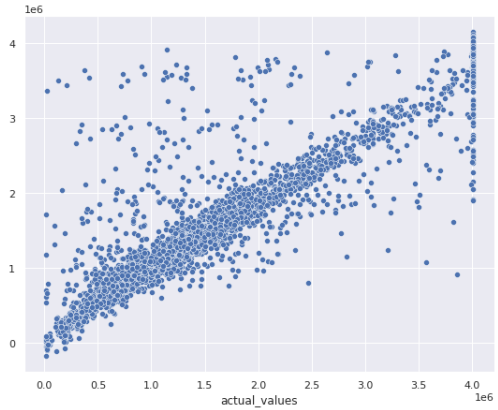
Lasso Regression-

- Train

R2 Score = 0.815646
RMSE Score = 475099.602720
MAE Score = 264386.647886

- Test

R2 Score = 0.816307
RMSE Score = 463941.933932
MAE Score = 261256.744277



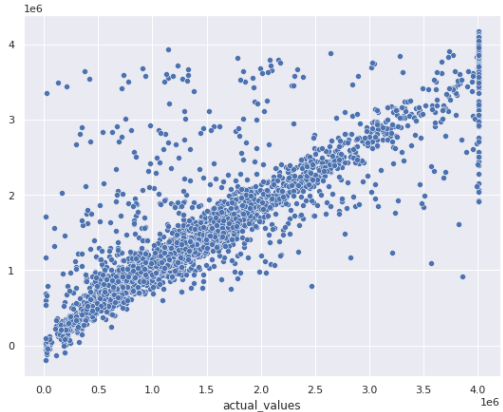
Ridge Regression-

- Train

R2 Score = 0.815646
RMSE Score = 475099.756286
MAE Score = 264525.067132

- Test

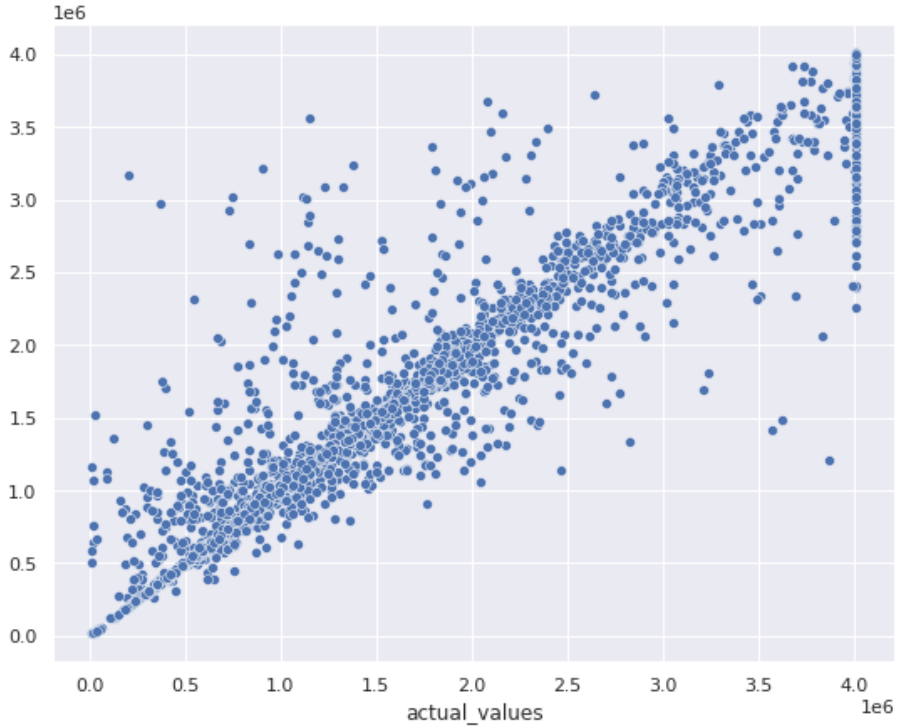
R2 Score = 0.816322
RMSE Score = 463922.997033
MAE Score = 261385.290661



Random Forest

➢Train
R2 Score = 0.978156
RMSE Score = 163540.862529
MAE Score = 79686.155437

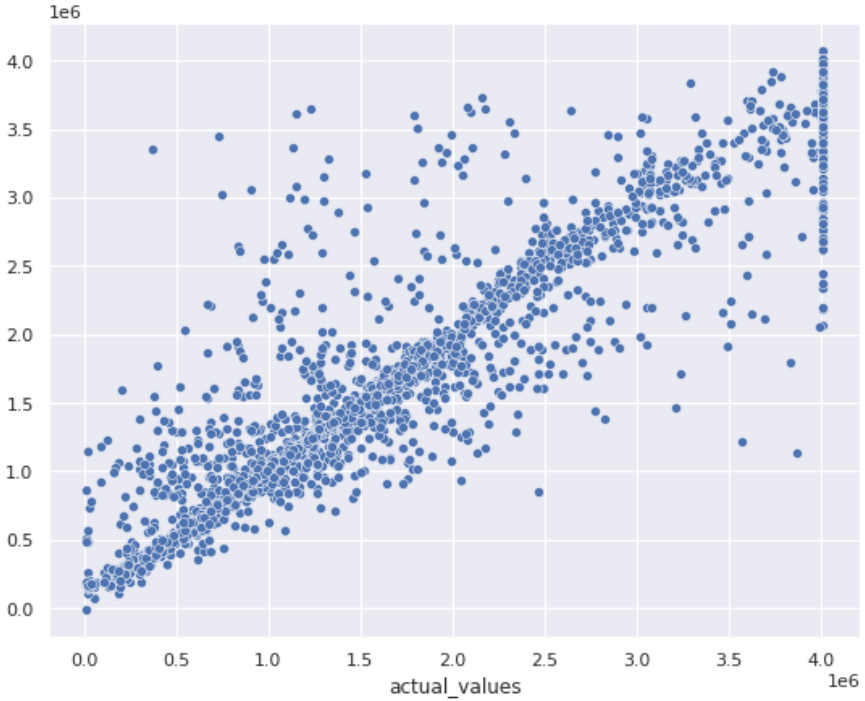
➢Test
R2 Score = 0.855799
RMSE Score = 411056.076421
MAE Score = 204731.737881



Gradient Boosting

➢Train
R2 Score = 0.901707
RMSE Score = 346912.935203
MAE Score = 179431.161659

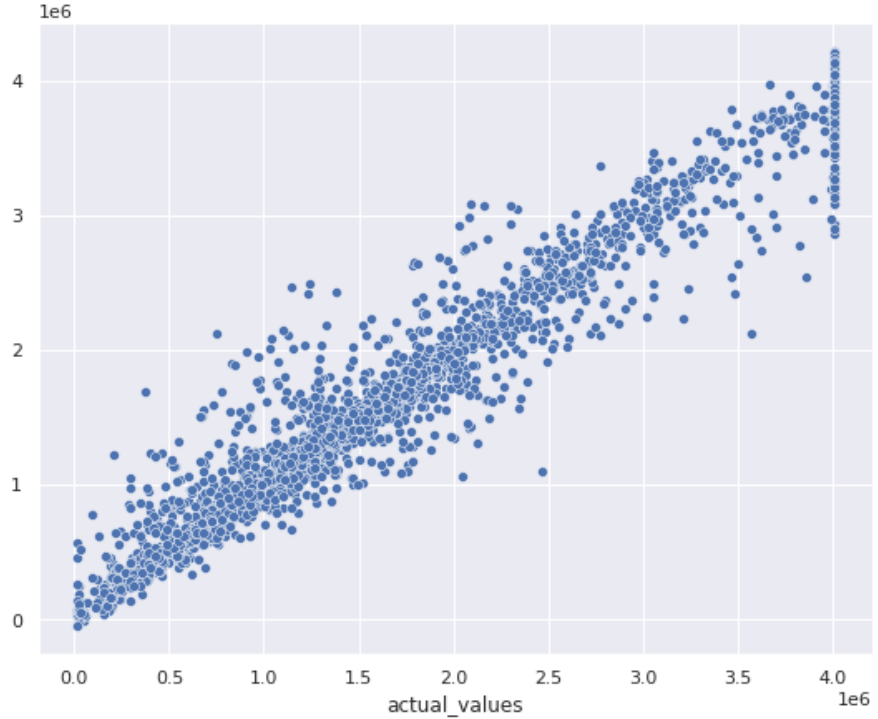
➢Test
R2 Score = 0.856578
RMSE Score = 409944.554402
MAE Score = 209046.949291



XG Boost

➢Train
R2 Score = 0.898855
RMSE Score = 351910.312477
MAE Score = 180496.695525

➢Test
R2 Score = 0.863784
RMSE Score = 399514.120264
MAE Score = 204730.060437



Models used

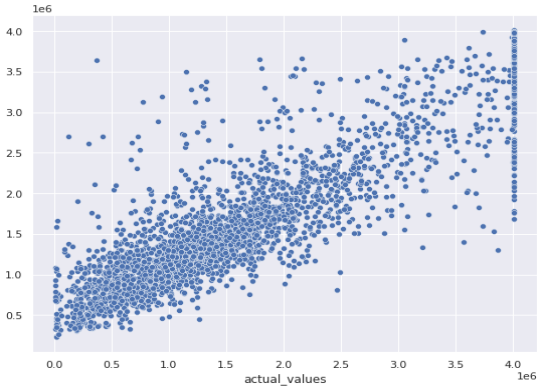
K Neighbor Regressor-

- Train

R2 Score = 0.832810
RMSE Score = 452443.261639
MAE Score = 304780.633276

- Test

R2 Score = 0.725436
RMSE Score = 567204.573574
MAE Score = 392778.158500



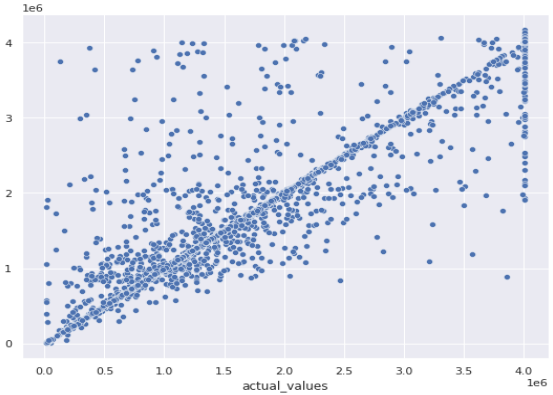
Support Vector Regressor-

- Train

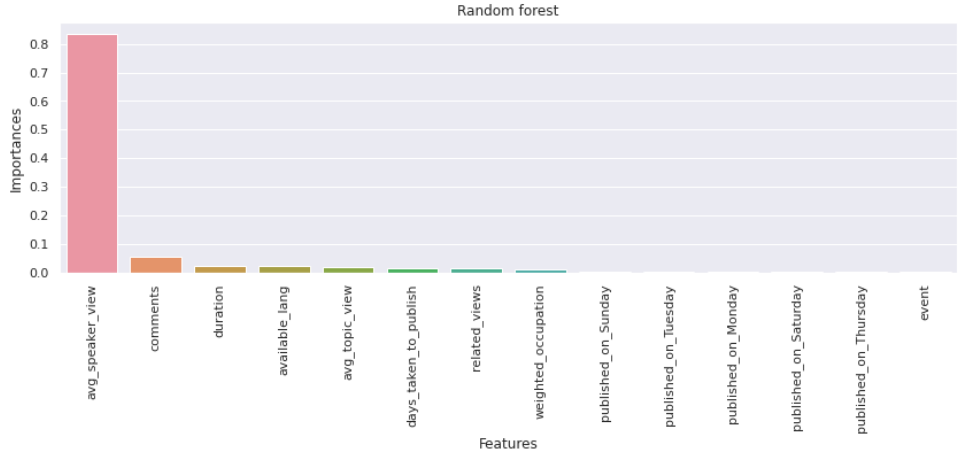
R2 Score = -0.071376
RMSE Score = 1145328.028768
MAE Score = 835985.802941

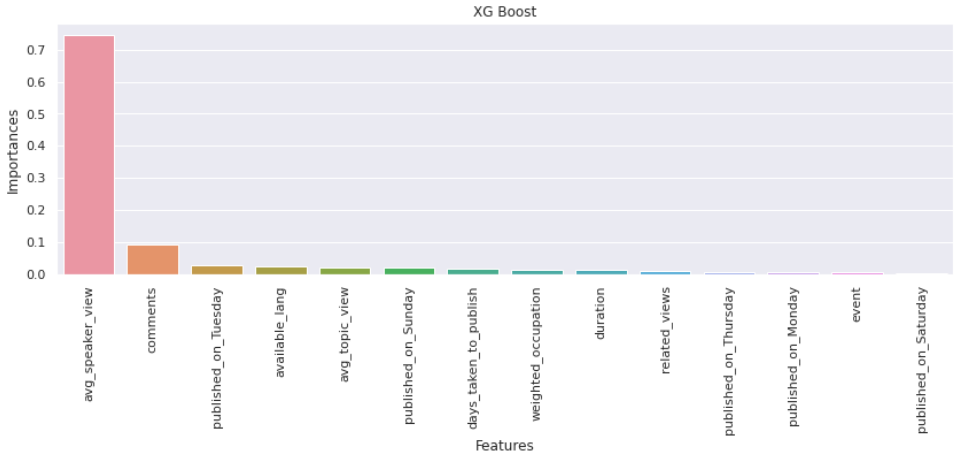
- Test

R2 Score = -0.066677
RMSE Score = 1117979.756202
MAE Score = 825600.070133



Feature Importance





- *Both Models given the importance of more than 0.8 for the feature named avg_speaker_view.*
- *Also both models has given some importance to the feature which we created from feature engineering process.*

Challenges and Solutions

- More than 600 values as Nulls in Comments attribute was a concern, KNN imputer helped us to treat it carefully.
- Less than 6 features are numerical from total 19 features and it's a challenge to find a way to convert them into valuable numerical features.
- Outliers are big problem for the models, so we have to explicitly set the upper limit as upper quartile.
- Features doesn't follow any clear linear relation with the target variable, that's why black box models helped us to perform better.
- GridsearchCV is computationally expensive process, so we have limit our parameters in gridsearch, that may have resulted in unoptimized performance.

Conclusion

✓ *EDA:-*

- *Repetition percentage of speakers is low, only 12% of speakers repeated between 1-4 times and 87% are unique speakers.*
- *Views and Comments has no direct influence on each other, they can vary according to other factors.*
- *Analysis shows a decline in views on website, year by year, possible reason would be Youtube, as they also published their video on this platform.*

✓ *ML Models:-*

- *Linear Models able to set-up balance in train and test scores, also without a clear linear relation they performed better than KNN and SVM, and SVM is not even in competition since its worst performer with most time consuming than any other model.*
- *if model explainability is not our priority then Random forest, Gradient boost and XG boost are serving their purpose fairly, and are reliable, but if we need to get more detail on how its happening, then we can go for linear models or simple tree based models.*
- *Avg_speaker_views variable is given high importance by different models, meaning speakers are influencing the views.*


Feedback