Capstone Project

TED Talk Views Prediction

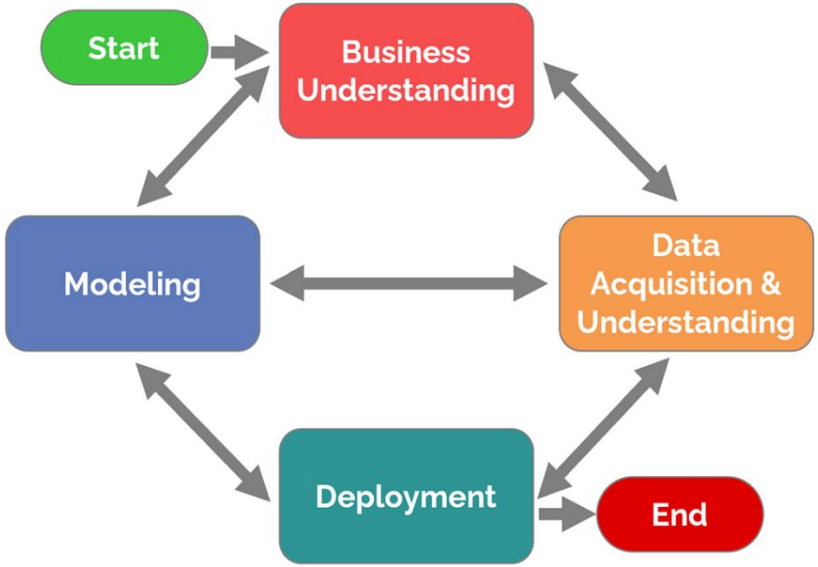
Team Power

Hariom Bhardwaj
Mayank Kumar
Shivam Mishra
Saifuddin Raja
Sarvesh Kumar Yadav

TDSP: TEAM DATA SCIENCE PROCESS

Microsoft approach

1. Business Understanding
2. Data Acquisition and Understanding
3. Modeling
4. Deployment
5. Customer Acceptance



ABOUT TED

- TED is a **nonprofit** devoted to spreading ideas.
- Founded in 1984 by Harry Marks and Richard Saul Wurman, as a conference where Technology, Entertainment and Design converged, in more than 100 languages.
- TED makes money through conference attendance fees, sponsorships, foundation support, licensing fees and book sales.
- TED Talks on the web are also supported by partnerships with carefully selected organizations; their ads on the videos.

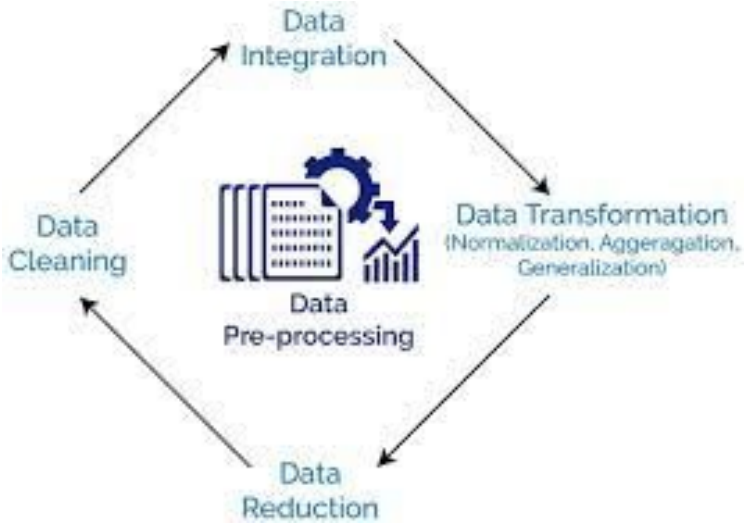
TED Ideas worth spreading

Problem Statement

- This Dataset contains over 4,000 TED talks, including transcripts in many languages.
- The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website.
- As TED is a non-profit organization, we need to think of some unique ways to make the most of the available data so that we can attract more viewers and sponsors in such a way that we can invite some fantastic speakers to share their knowledge and experiences.

DATA PIPELINE:

- **Understanding the Data**
- **EDA/Cleaning the Data:** The data was checked for duplicate values, null and missing values, and primary inspection was performed. Exploratory data analysis was performed to analyse and visualize the data.
- **Feature Engineering:** Creating insightful features and transforming data.



Understanding the Data

DATASET NAME:

- Data_ted_talks

SHAPE:

- Data Points (Rows) : 4,005
- Features (Columns) : 18

TARGET VARIABLE:

- ‘Views’

FEATURES:

Attribute (Type)	Description Data
talk_id (int)	Talk identification number provided by TED
title (string)	Title of the talk
speaker_1 (string)	First speaker in TED's speaker list
speakers (dictionary)	Speakers in the talk
occupations (dictionary)	*Occupations of the speakers
about_speakers (dictionary)	*Blurb about each speaker
views(Dependent Variable) (int)	Count of views
recorded_date (string)	Date the talk was recorded
published_date (string)	Date the talk was published to TED.com
event (string)	Event or medium in which the talk was given
native_lang (string)	Language the talk was given in
available_lang (list)	All available languages (lang_code) for a talk
comments (int)	Count of comments
duration (int)	Duration in seconds
topics (list)	Related tags or topics for the talk
related_talks (Dictionary - key='talk_id', value='title')	Related talks
url (string)	URL of the talk
description (string)	Description of the talk
transcript (string)	Full transcript of the talk

Exploratory Data Analysis

CHECKING FOR NULL VALUE

Features having NULL values:

comments : 16.35%

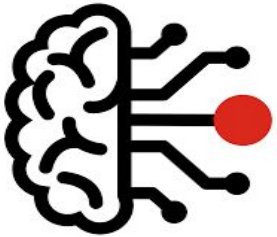
occupations : 13%

about_speakers: 12.5%

all_speakers : 0.0009%

recorded_date : ~0%

Name	dtypes	Missing	Uniques
comments	float64	655	601
occupations	object	522	2049
about_speakers	object	503	2977
all_speakers	object	4	3306
recorded_date	object	1	1334
talk_id	int64	0	4005
description	object	0	4005
url	object	0	4005
related_talks	object	0	4005
topics	object	0	3977
duration	int64	0	1188
event	object	0	459
available_lang	object	0	3902
native_lang	object	0	12
title	object	0	4005
published_date	object	0	2962
views	int64	0	3996
speaker_1	object	0	3274
transcript	object	0	4005

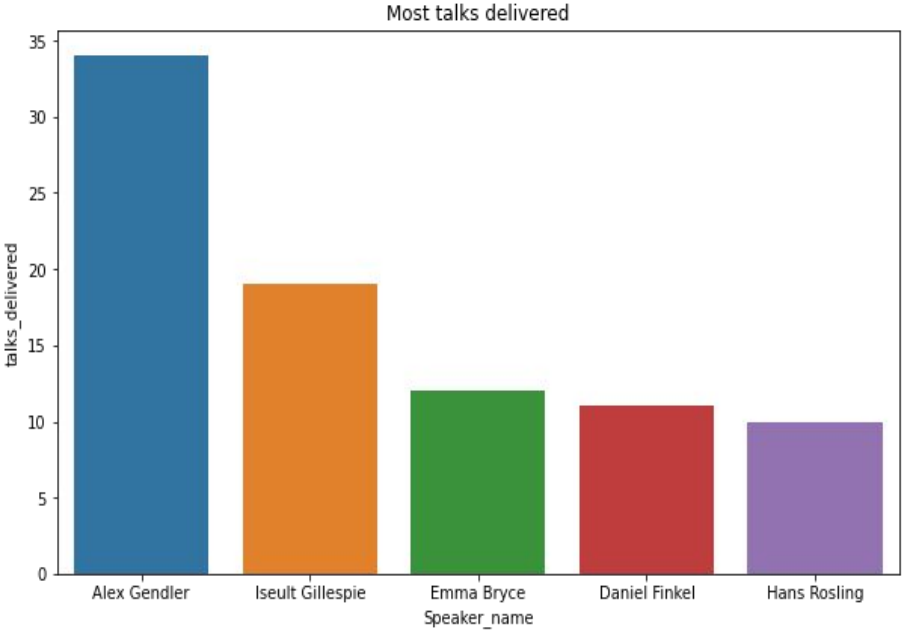


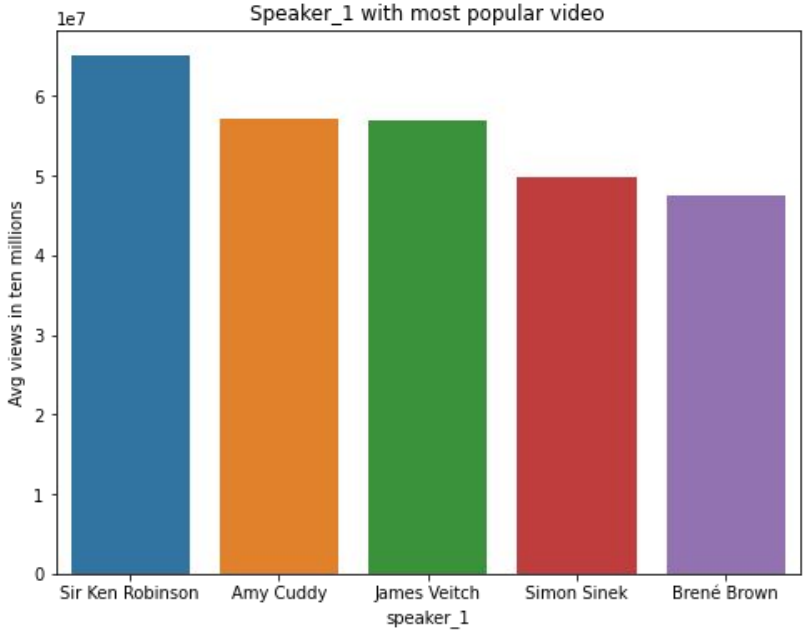
Feature Extraction

Extracted Features	About Feature	Feature Used
time_since_published	Number of days after talk published	published_date
daily_views (new **Target** feat.)	TED talk Views/days	views, time_since_published
speaker_1_avg_views	Average views of each speaker's TED talk	speaker_1
event_wise_avg_views	Average view of each event	events
number_of_lang	Available language count	available_lang
num_of_topics	Number of topics in a particular TED talk	topics
topics_wise_avg_views	Average views of each topic	topics
week_day	Day of week	published_date
month	Month	published_date
year	Year	published_date
day	Day of month	published_date

Speakers with Daily Views:

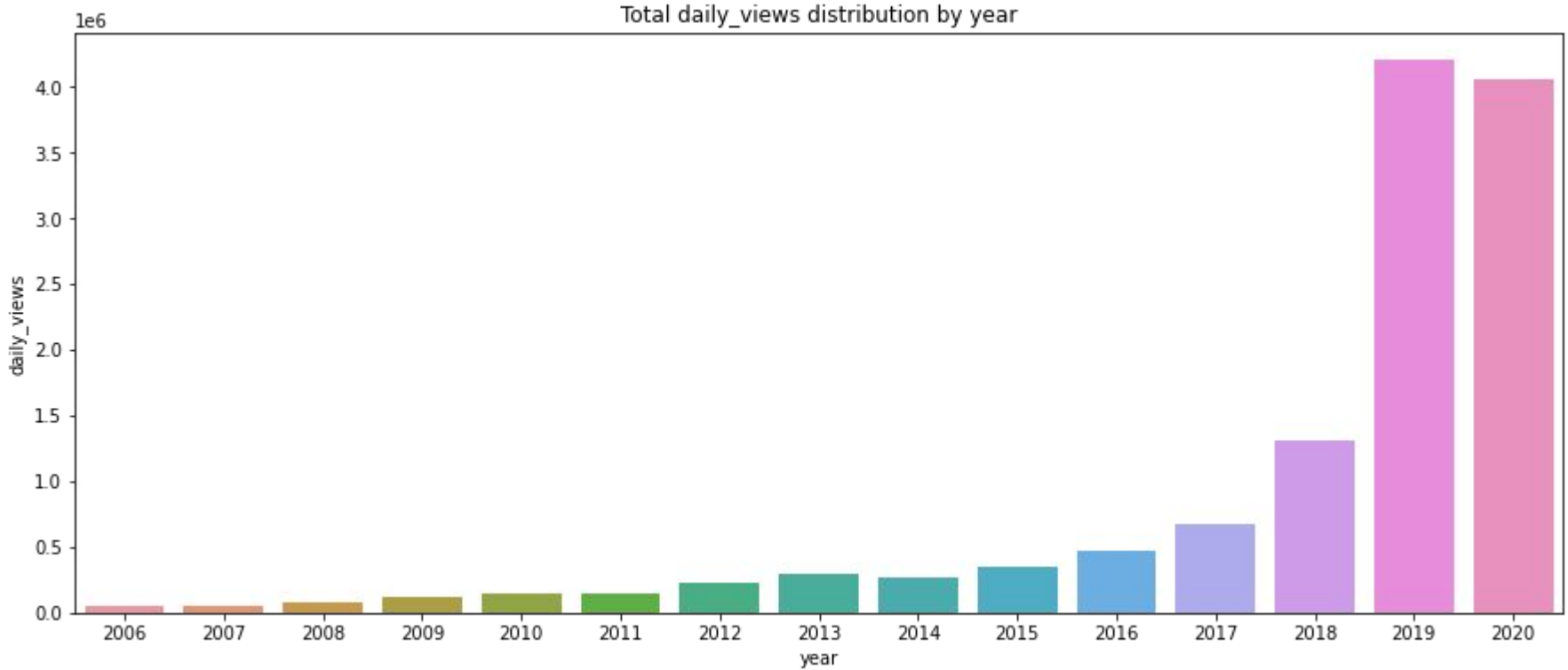
- Alex Gendler is the most invited speaker at TED.
- Sir Ken Robinson is the most popular speaker as per avg views for all talks.





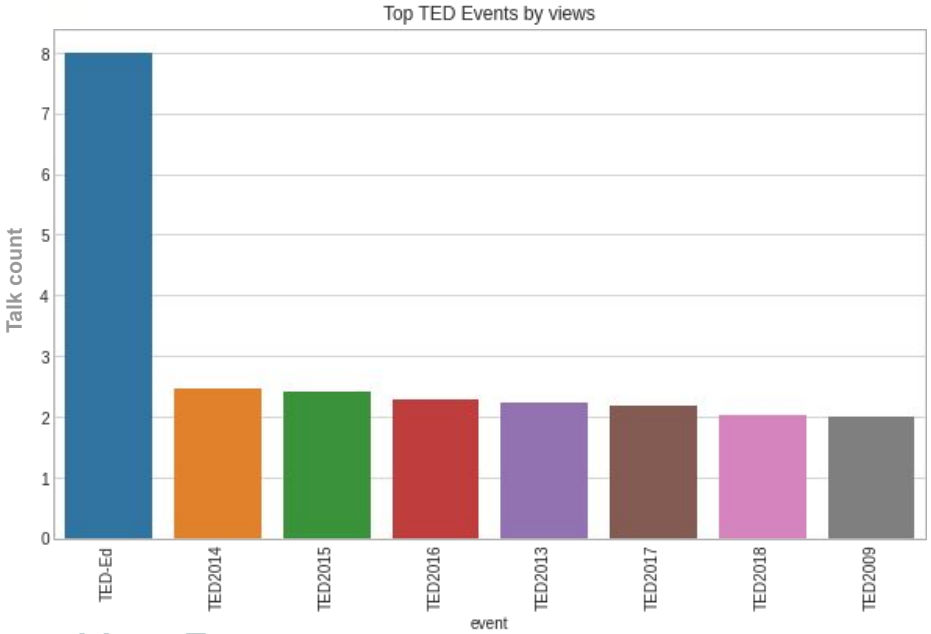
Views varying with Published Year:

- Popularity of TED is continuously growing since 2015.

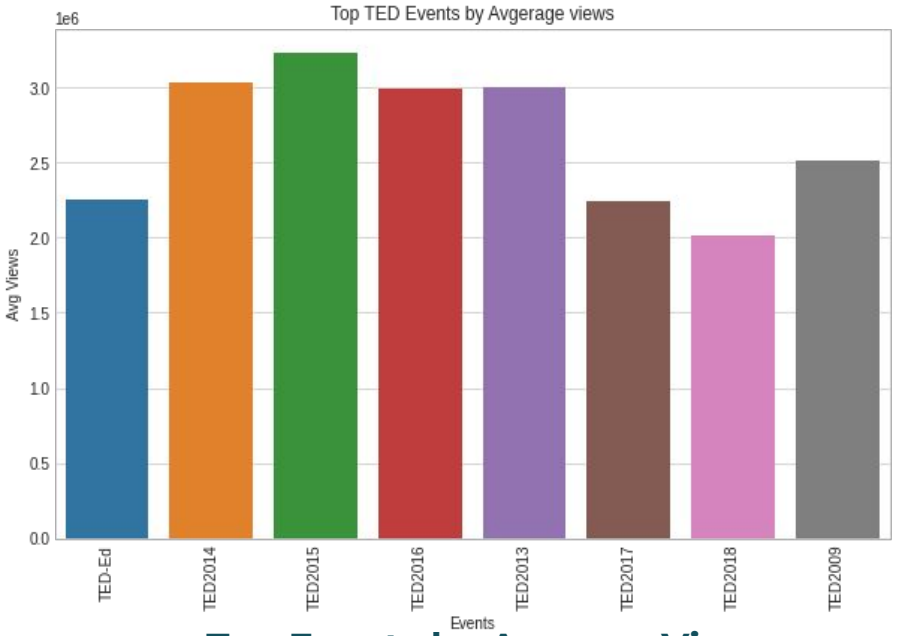


Events with Views:

- TED-Ed is the most frequent event of TED.
- TED-2015 Events has most number of Average views.

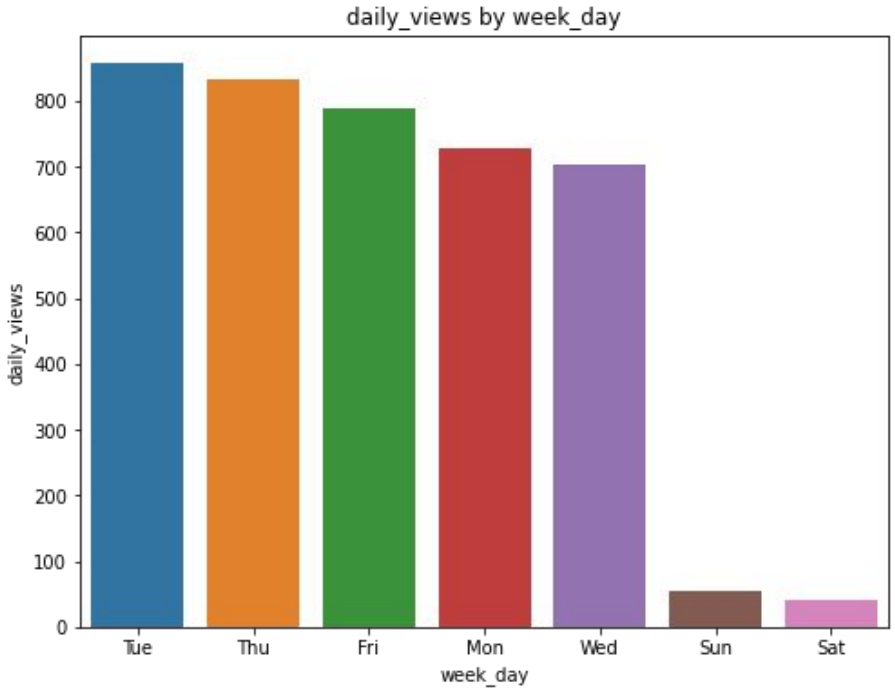


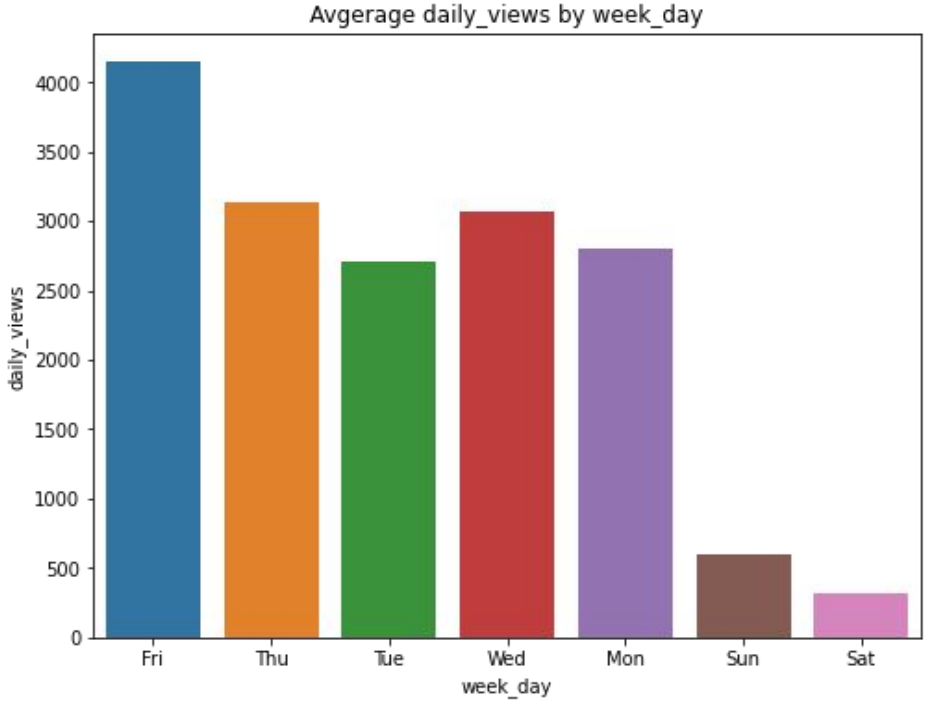
Most Frequent event category



Top Events by Average Views

Published WeekDay vs. Daily Views:





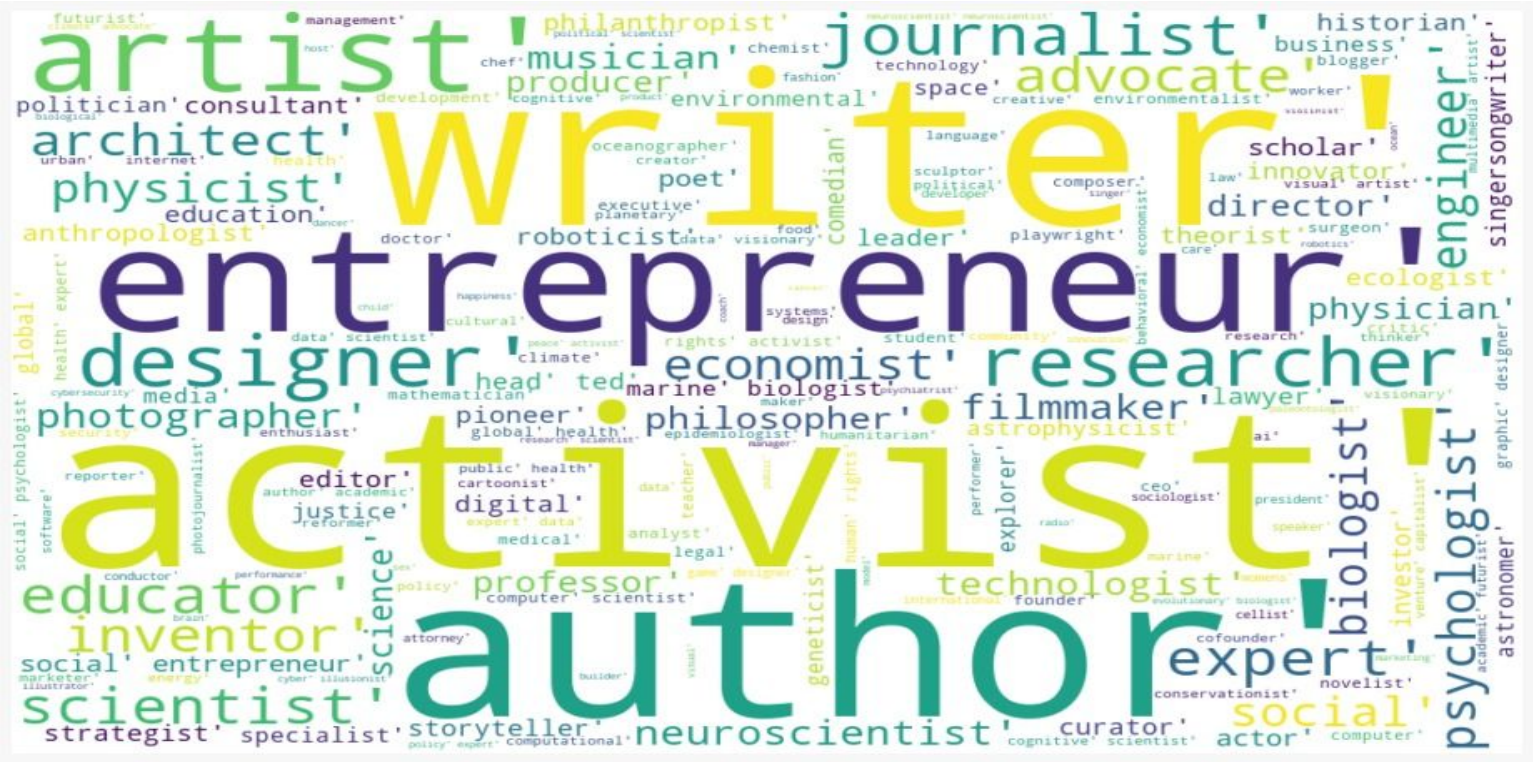
- **Friday release is impacting the views of the video**

Most popular Titles:

- Most used words in the title are World and Life.

Most Popular Occupations:

- Most of the talks are delivered by activist, writer, entrepreneur and author.



Most Popular Topics:

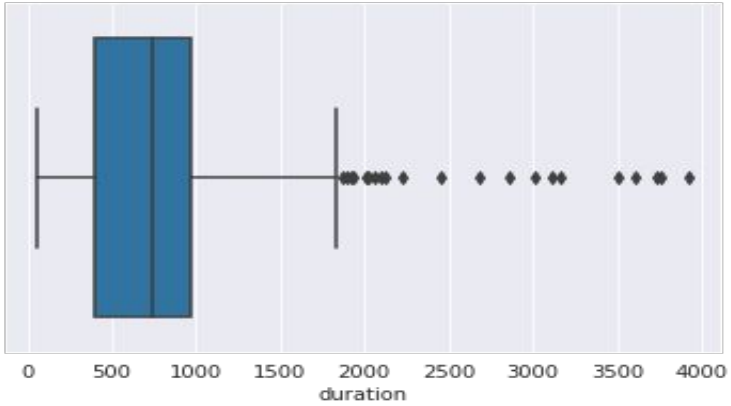
- Most talks are delivered for topics social, technology, humanity, science, etc.

Feature Engineering

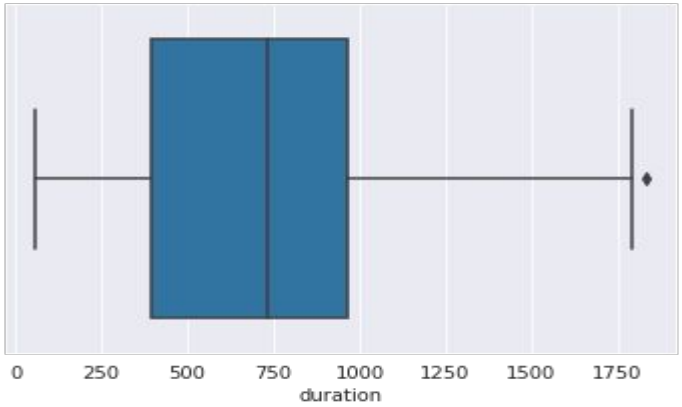
DURATION :

- 0.53 % outlier is detected in duration column.

MITIGATION: Imputed outliers by mean of duration.



Before outlier treatment

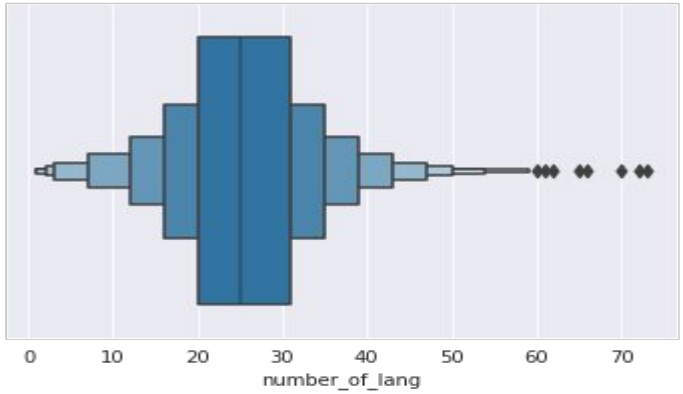


After outlier treatment

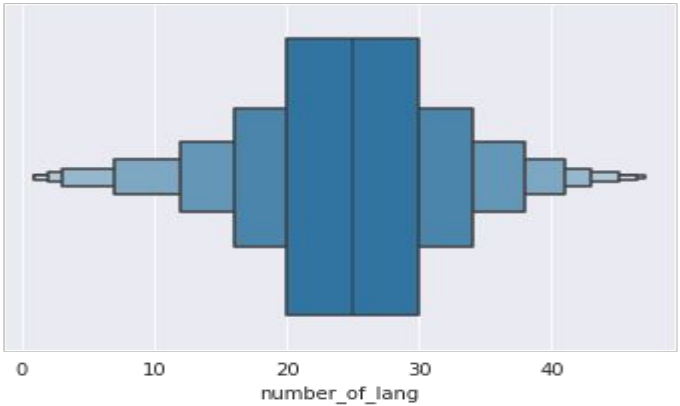
Number of language:

- 3.02 % outlier is detected in number of language column.

MITIGATION: Imputed outliers by mean of number of language .



Before outlier treatment



After outlier treatment

Treating Missing/Null Values

- Comments column has 655 missing values i.e 16.35 % of total records.

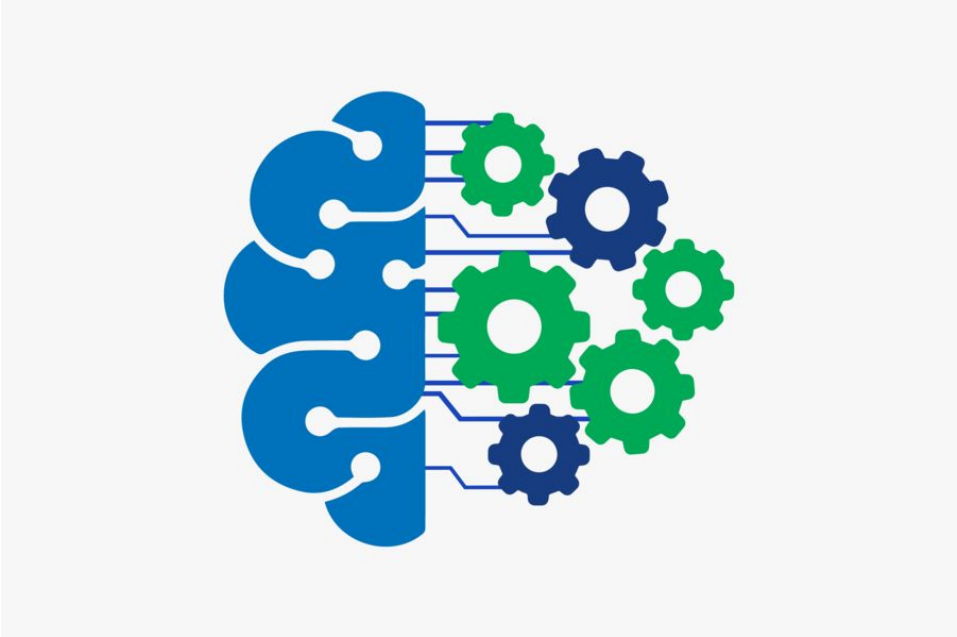
MITIGATION: Imputed the null values using KNN Imputer.

Encoding:

We have used dummy encoder on the below features:

- **time_since_published**
- **speaker_1_avg_views**
- **event_wise_avg_views**
- **number_of_lang**
- **num_of_topics**
- **topics_wise_avg_views**

Machine Learning



Machine Learning Model Used

1. XGB Regressor
2. CatBoost Regressor

Metrics Used

- R2
- RMSE
- MAE

Hyper-Parameter tuning

- Random Search CV

- Criterion = R_Square

- R_Square for train = 0.94
- R_Square for test = 0.81

- MAE Train = 0.08
- MAE Test = 0.11

- RMSE for Train = 0.25
- RMSE for Test = 0.38

- Criterion = R_Square

- R_Square for train = 0.99
- R_Square for test = 0.73

- MAE Train = 0.04
- MAE Test = 0.14

- RMSE for Train = 0.07
- RMSE for Test = 0.44

Model Selection

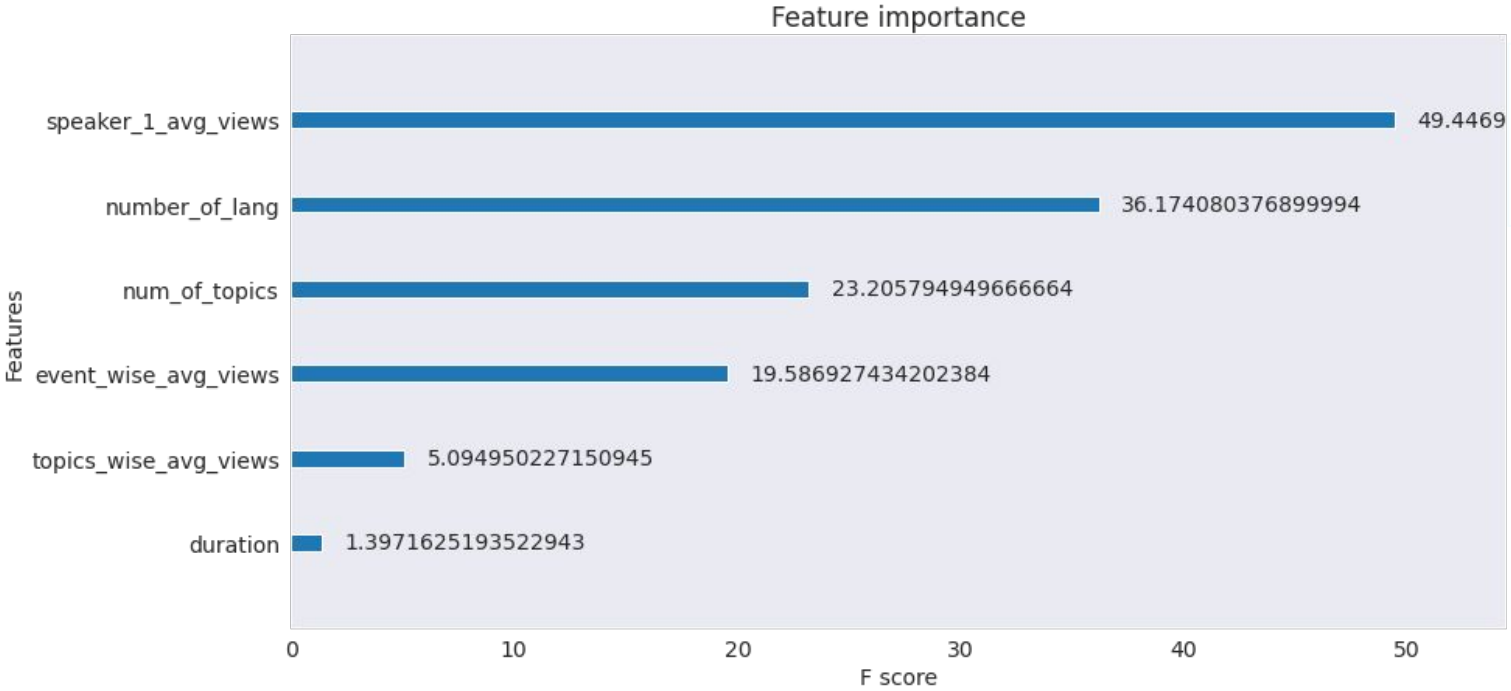
Base on Final tuning results

On the criterion of R_Square

XGBoost

is the best performing model

Feature importance



Natural Language Processing

FEATURES USED FOR NLP

Features in dataset that are continuous and contain information in text

FEATURES	Description
title	Title of the talk
occupations	occupation of the speaker
topics	Related tags or topic of the talk
description	Description of the talk

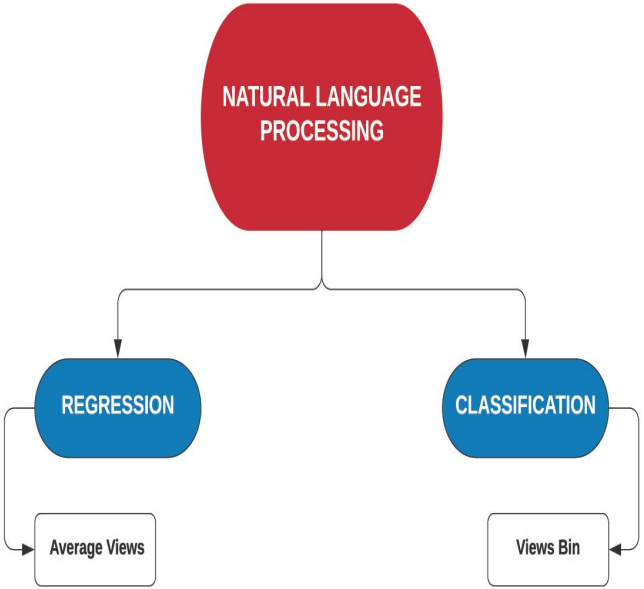
Target variable:

Views

Preparing and Cleaning Data for NLP

STEPS USED:

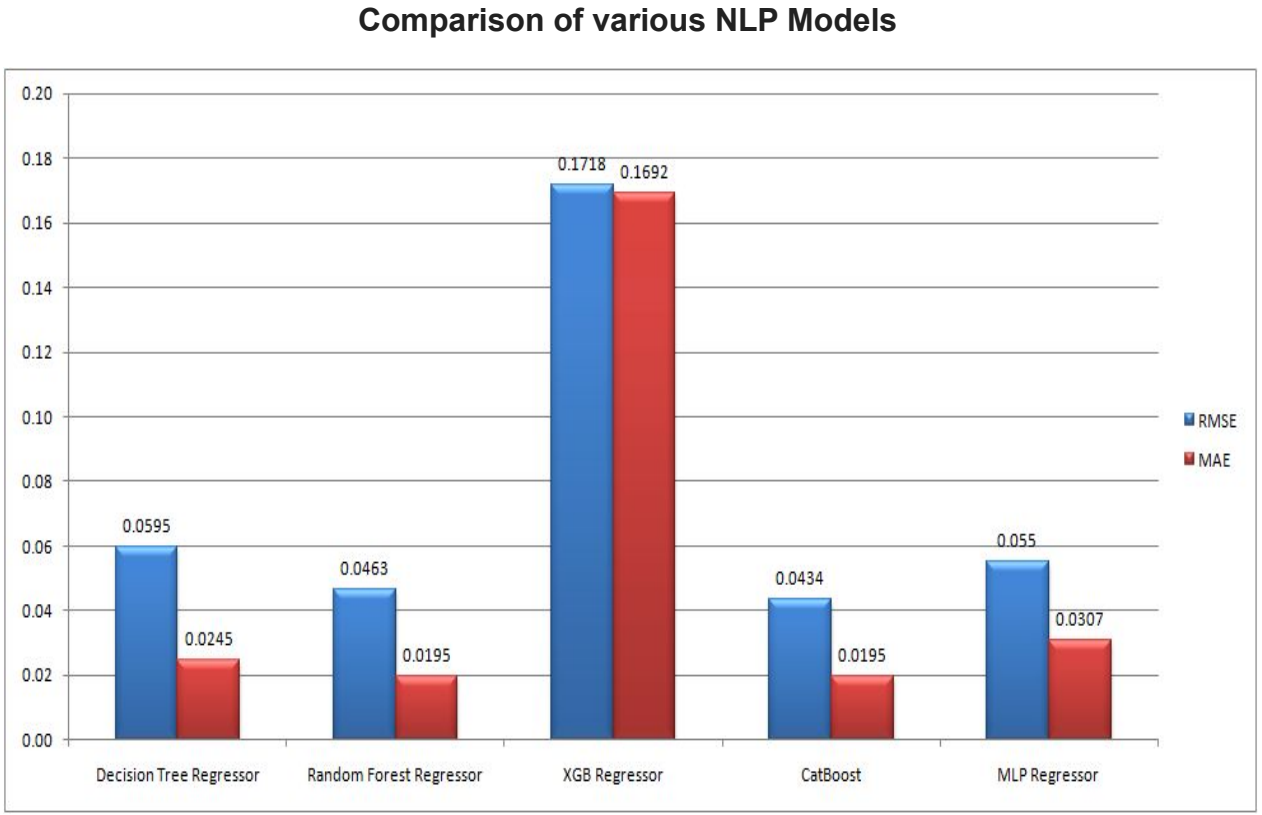
1. Convert all the words in to its lower case
2. Removed Punctuations
3. Removed Stopwords
4. Word Clouds
5. Split Data into Train and Test set.
6. Count Vectorization / TF-IDF vectorization
7. ML Model



Predicting Views with Title

Metric used for evaluation:

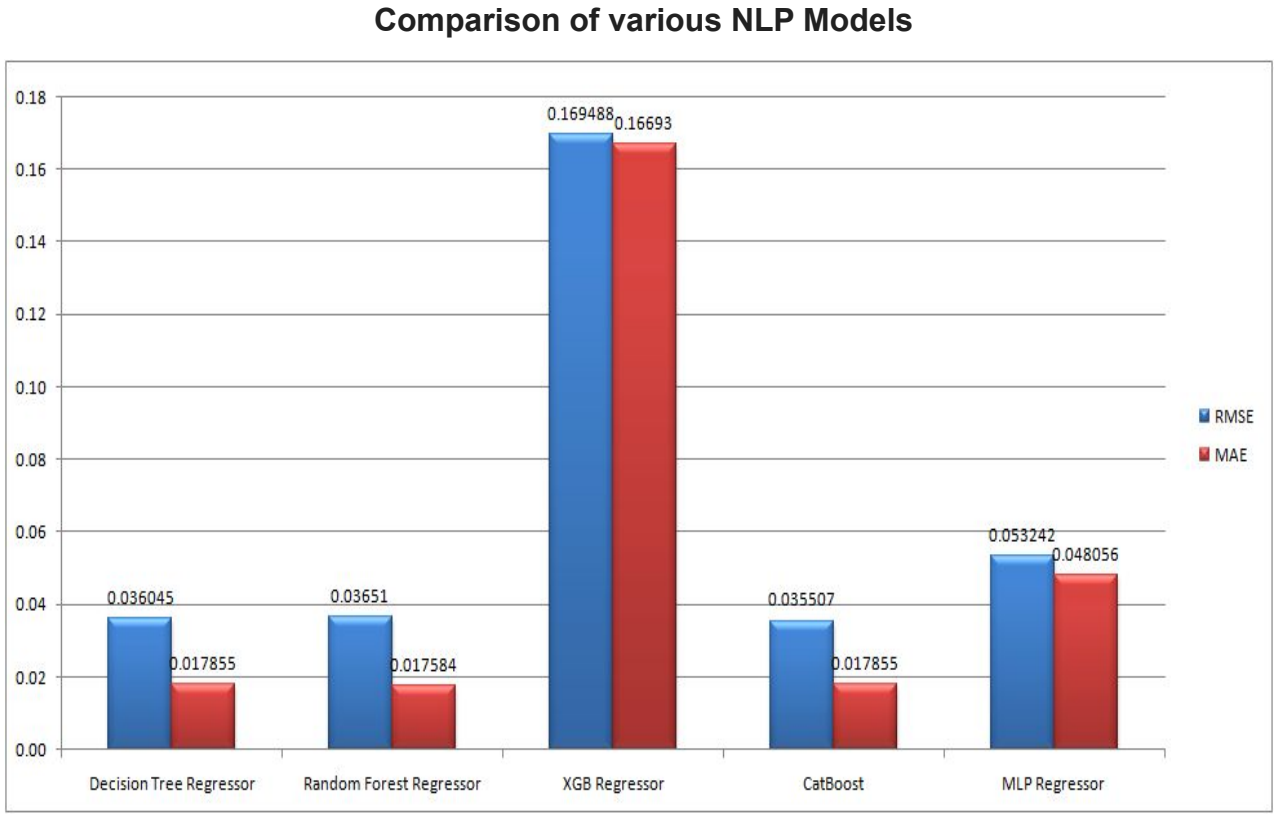
1. **RMSE**
2. **MAE**



Predicting Views with All Text Data

Metric used for evaluation:

1. **RMSE**
2. **MAE**



Conclusion

- Through our Analysis, we have discovered key insights about what factors influence the Views gained by a video .
- Topics like Technology , Science , Education , Biology attract the attention of viewers more than other topics .
- Entrepreneurs and Activists are the most engaging speakers

- For the ML Pipeline , the XGBoost Model performed the best
- For the NLP Pipeline , the Random Forest Model performed the Best
- Feature Engineering and Feature Extraction helped in increasing the model performance

Challenges and Future Scope

- Dataset have lots of textual and categorical data having high cardinality .
  - So the conversion to meaningful numerical data was a challenge.
- NLP did not perform well on the given dataset ,
  - because the number of rows is very less , which led to high variance
  - Deep learning (RNN and LSTM) and Auto-Transformer models can help
- Feature Engineering and Feature Extraction
  - can always be improved in creative ways
  - We can explore more advanced feature encoding