Capstone Project

Ted Talks Views Prediction

SAMEER THETE

GitHub Link

Discussion points

- 1. Problem statement
- 2. Data Summary
- 3. Exploratory Data Analysis
- 4. Feature Engineering
- 5. Feature selection
- 6. Modelling
- 7. Model Selection
- 8. Challenges
- 9. 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 videos uploaded on the TEDx website.

Data Summary

- Dataset name: data_ted_talks
- Shape:
- Rows = 4005
- Columns = 19

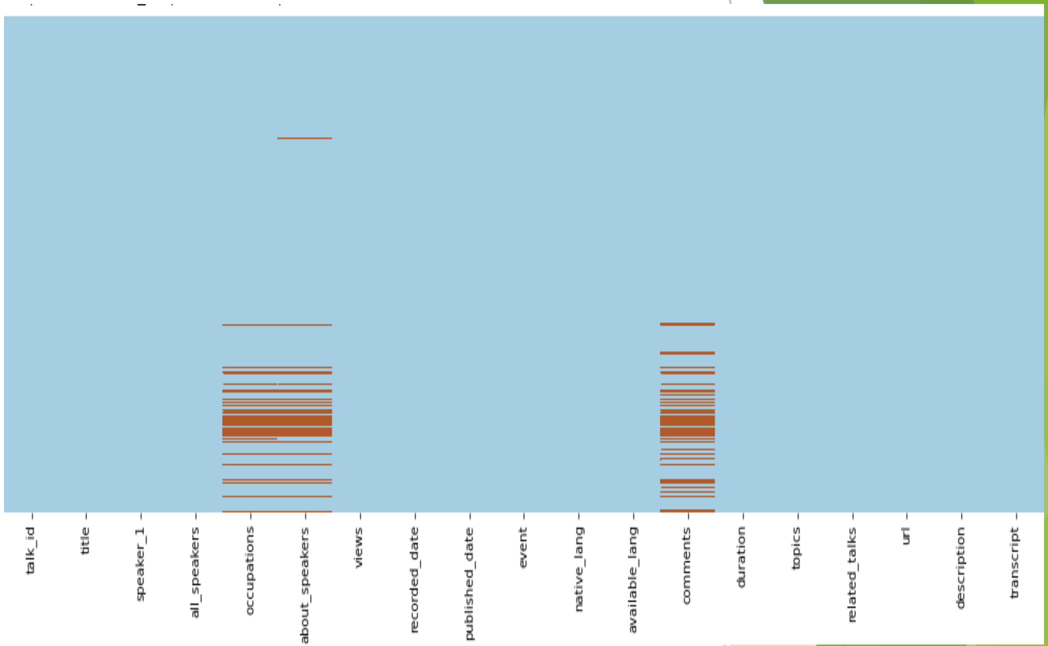
- Features:
- 'talk_id', 'title', 'speaker_1', 'all_speakers', 'occupations’, 'about_speakers', 'views', 'recorded_date', 'published_date', 'event’, 'native_lang', 'available_lang', 'comments', 'duration', 'topics', 'related_talks', 'url', 'description', 'transcript’

- Target variable : ‘views’

Exploratory Data Analysis

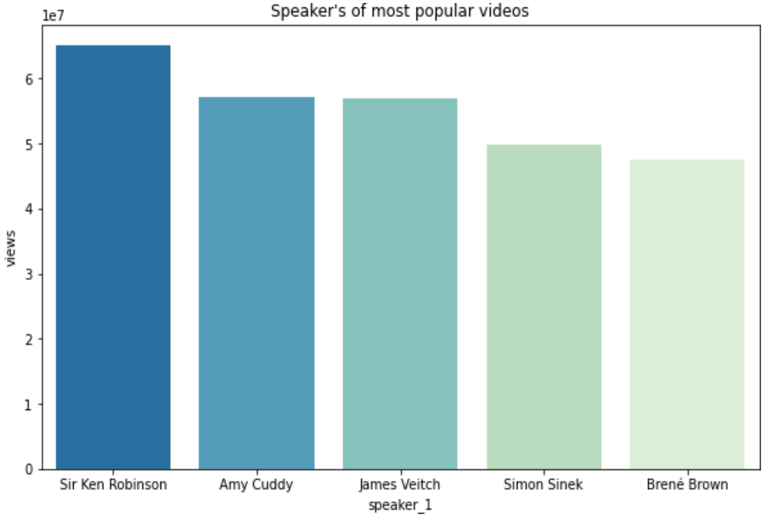
Handling Missing values

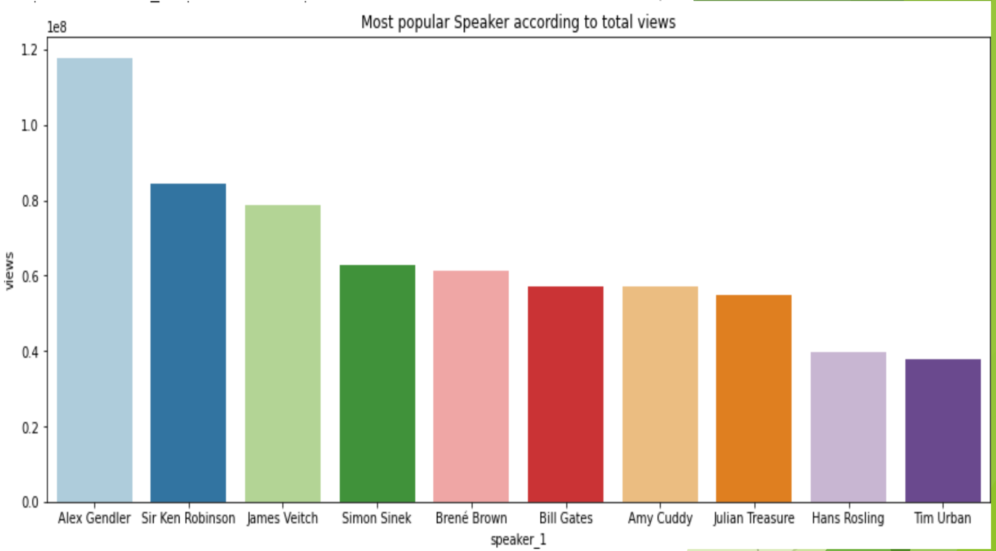
- **For numerical feature:**
- used KNNImputer to
- impute missing values



- **For categorical features:**
- Replaced Nan values with
- ‘Unknown’ category

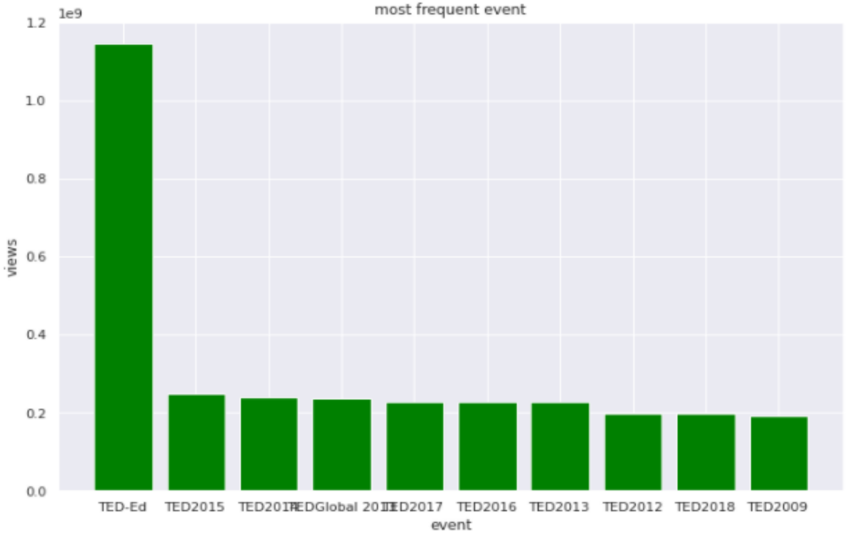
Overview of Speaker column

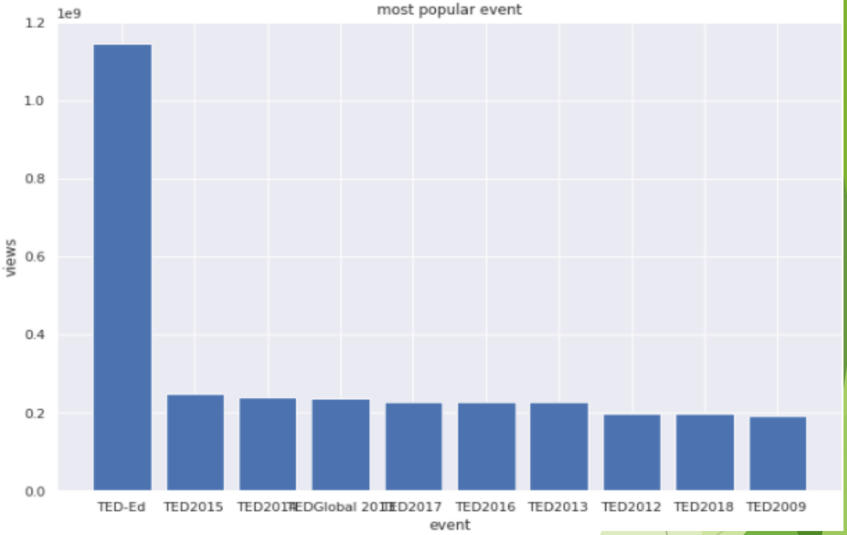




- Sir Ken Robinson's talk on "Do Schools Kill Creativity?" is the most popular TED Talk with more than 65 million views.
- Alex Gendler is the most popular speaker wrt to total views followed by Sir Ken Robinson

Which is the most popular and frequent event?

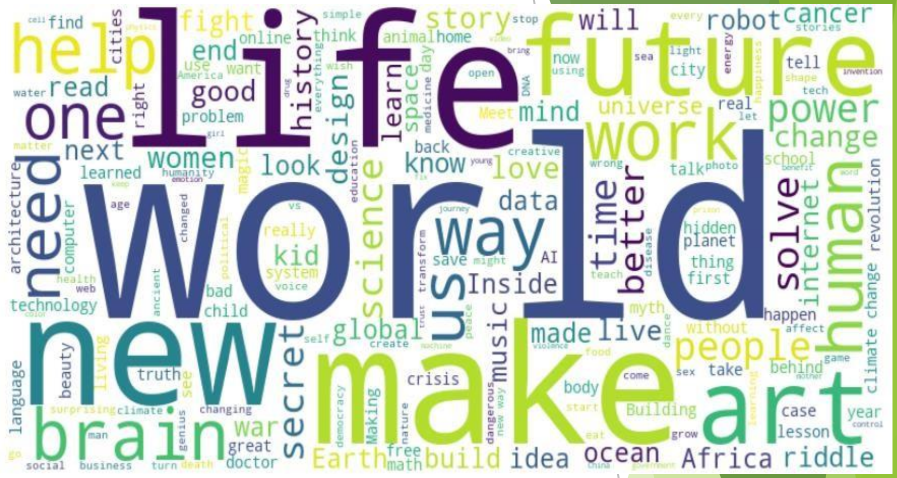


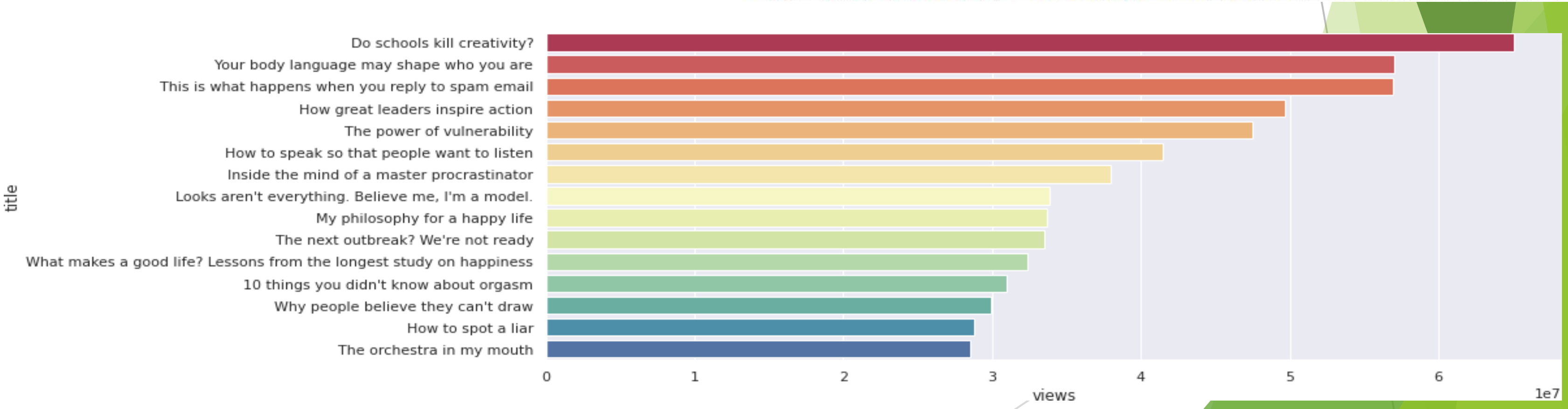


- TED-Ed is the most popular and frequent event

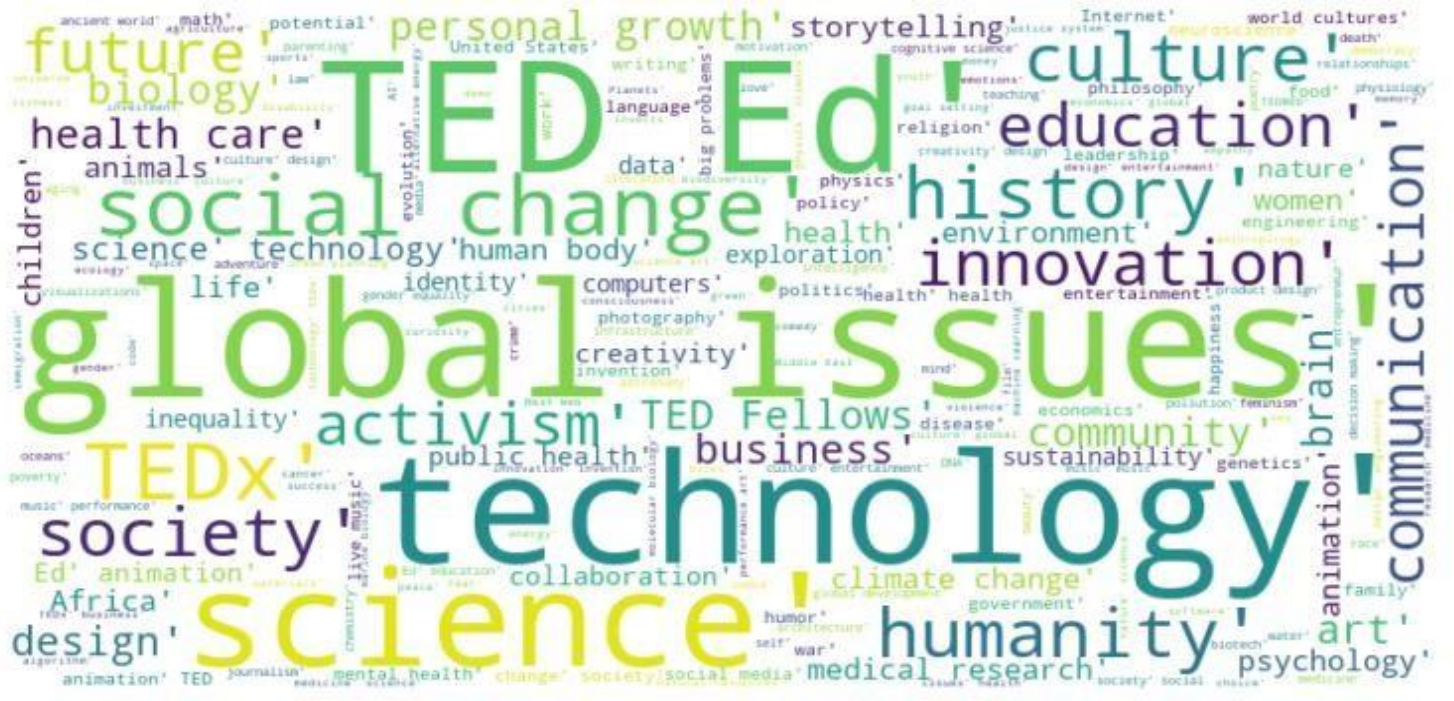
Most popular title?

- Most popular title:
- Do schools kill creativity with 65M views



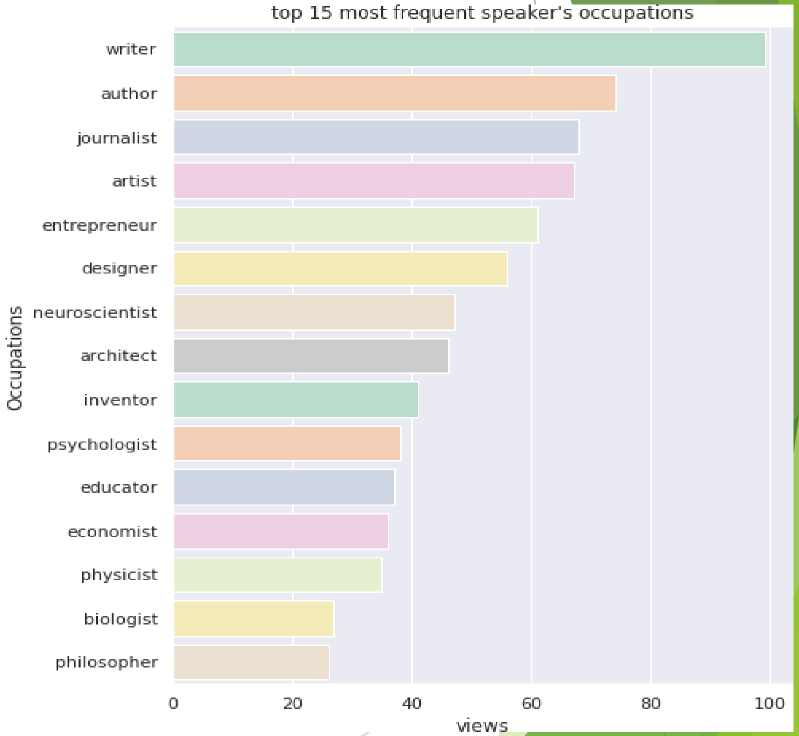


Most popular topic tags?

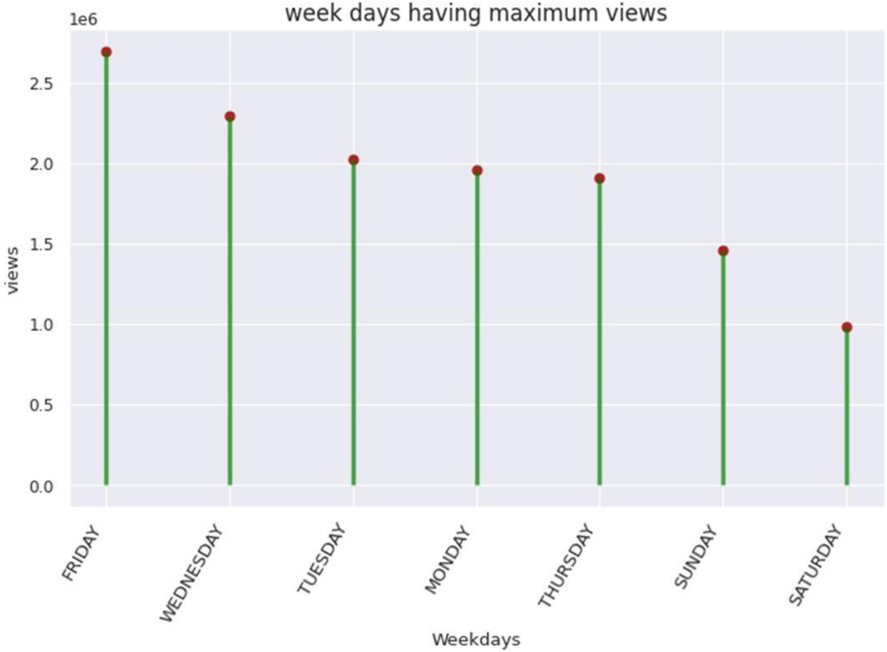


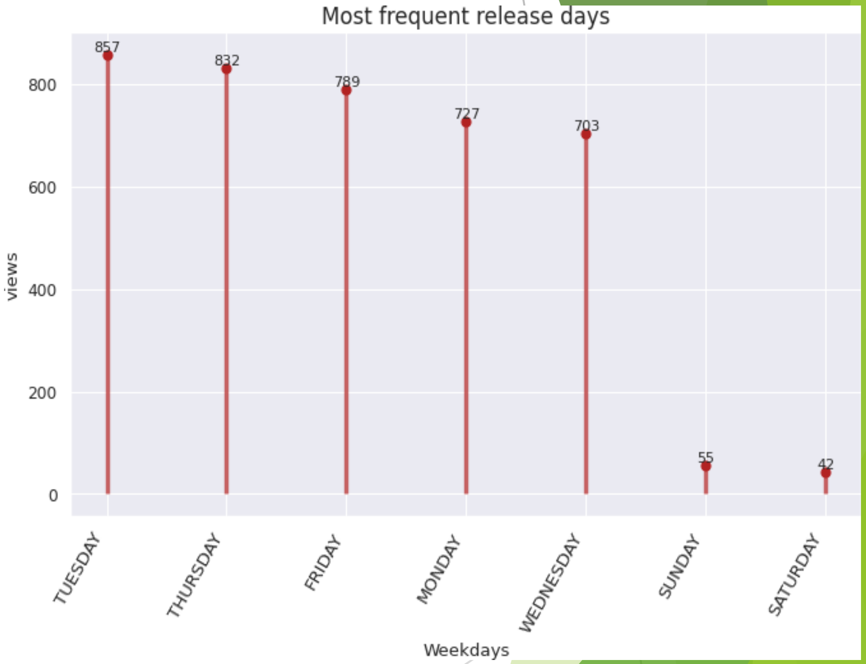
Most frequent Speaker’s occupations

- Writer is the most frequent speaker’s occupation followed by author and journalist



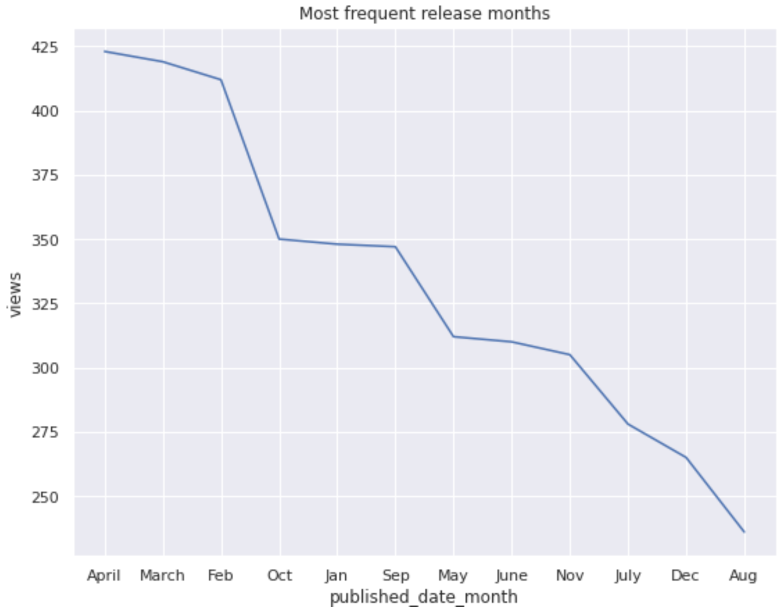
Overview of published_date

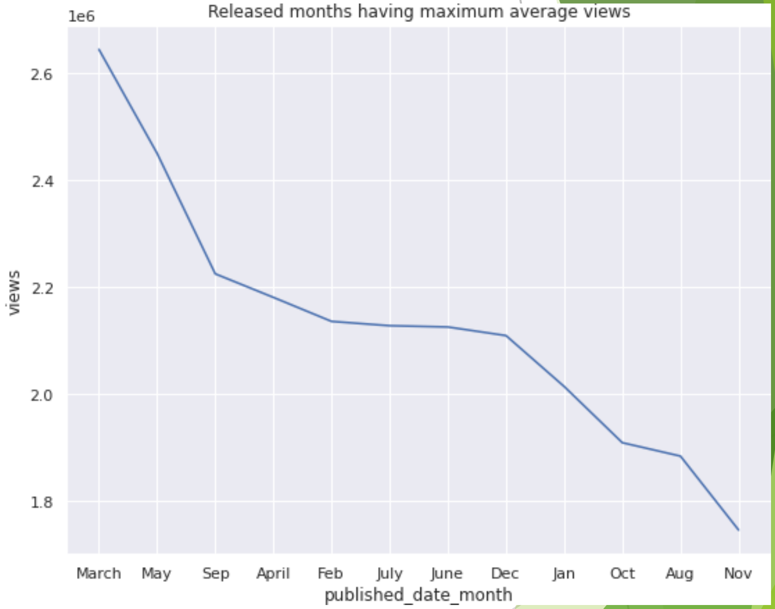




- Most videos are published on Tuesday followed by Thursday
- But the videos published on Friday are more popular

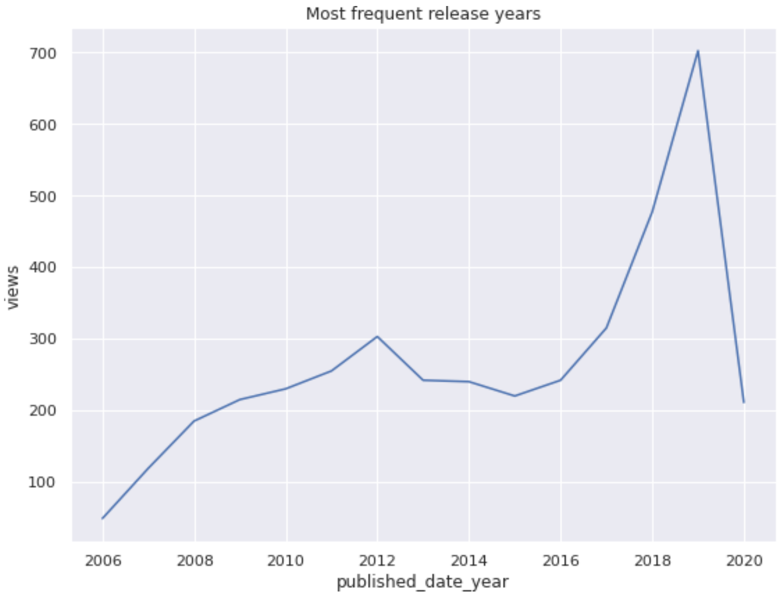
Published month overview

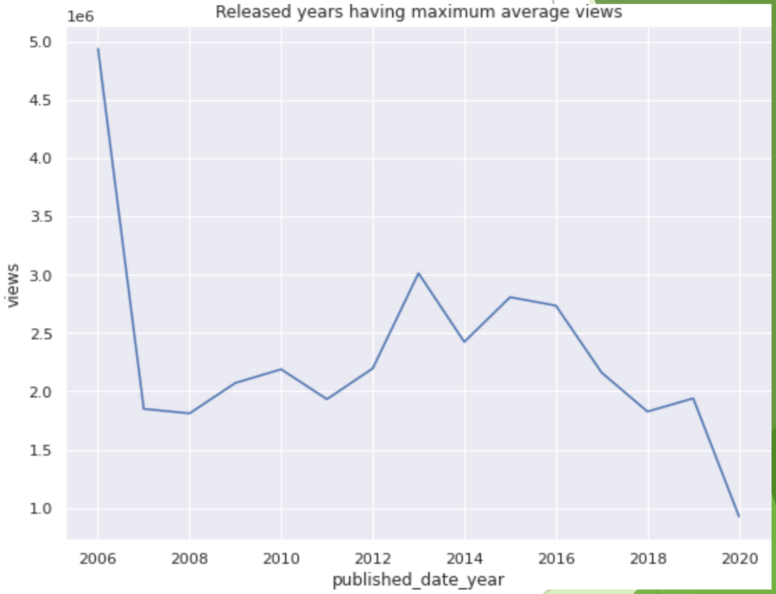




- April have maximum released videos. But the videos released in March are more popular

Published year overview


Most frequent release years


Released years having maximum average views

- Most videos are published in 2019. But videos in 2006 are most viewed

Feature Engineering

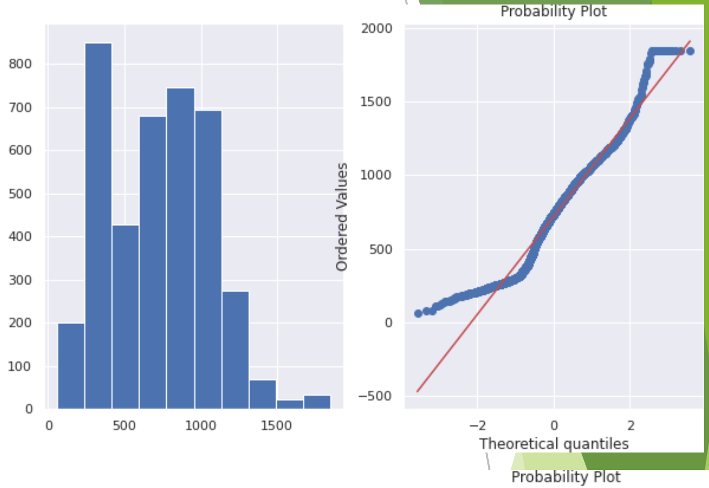
- 1. speaker_1_avg_views
- 2. event_wise_avg_views
- 3. num_of_tags
- 4. topics_wise_avg_views
- 5. num_of_lang
- 6. video_age
- 7. related_talks_views

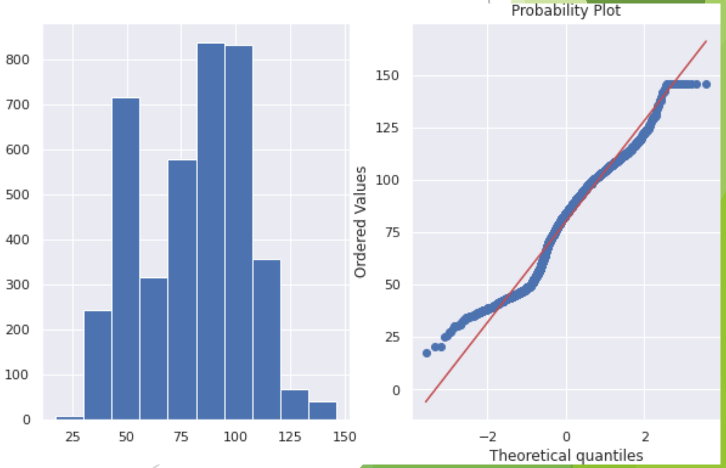
Due to high number of cardinality in Speaker_1 and event Column, therefore applied mean encoding

In Mean Encoding each distinct value of categorical value is replaced with average value of target variable

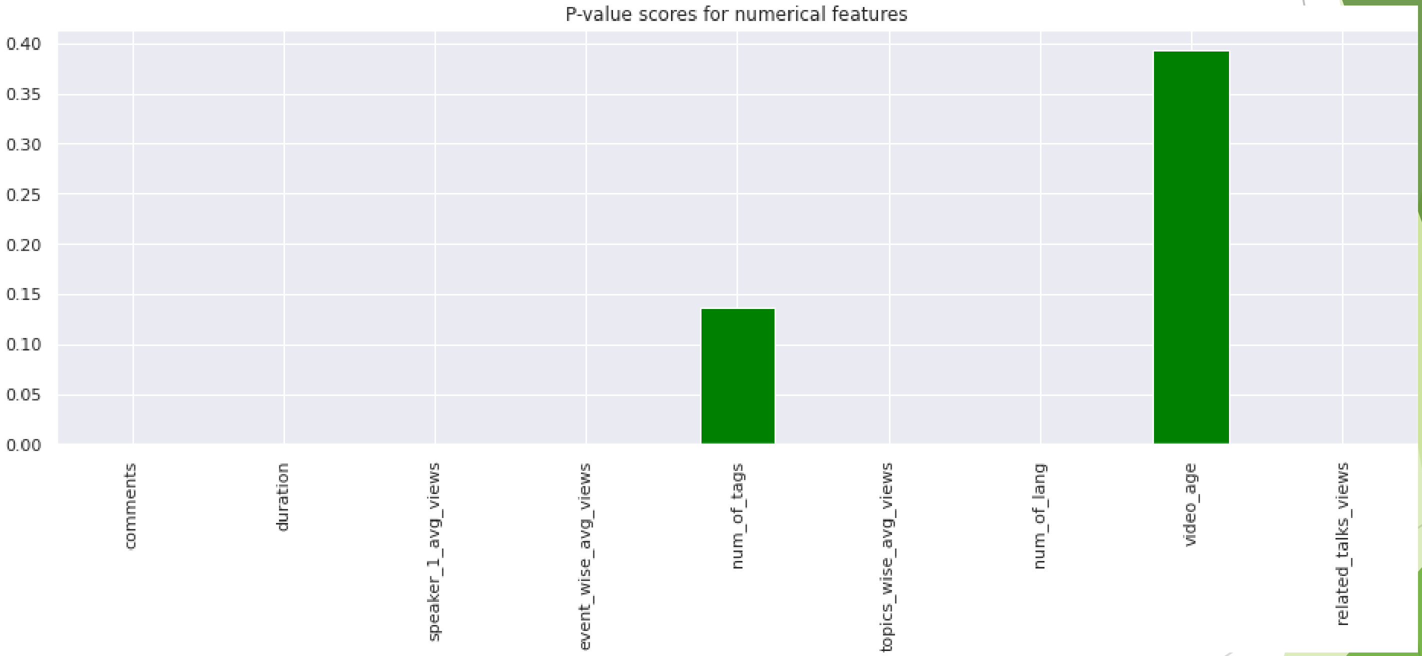
Transformations

- Applied on the following features:
- Comments
- Duration
- event_wise_avg_views
- num_of_tags
- topics_wise_avg_views
- num_of_lang
- video_age
- related_talks_views
- speaker_1_avg_views

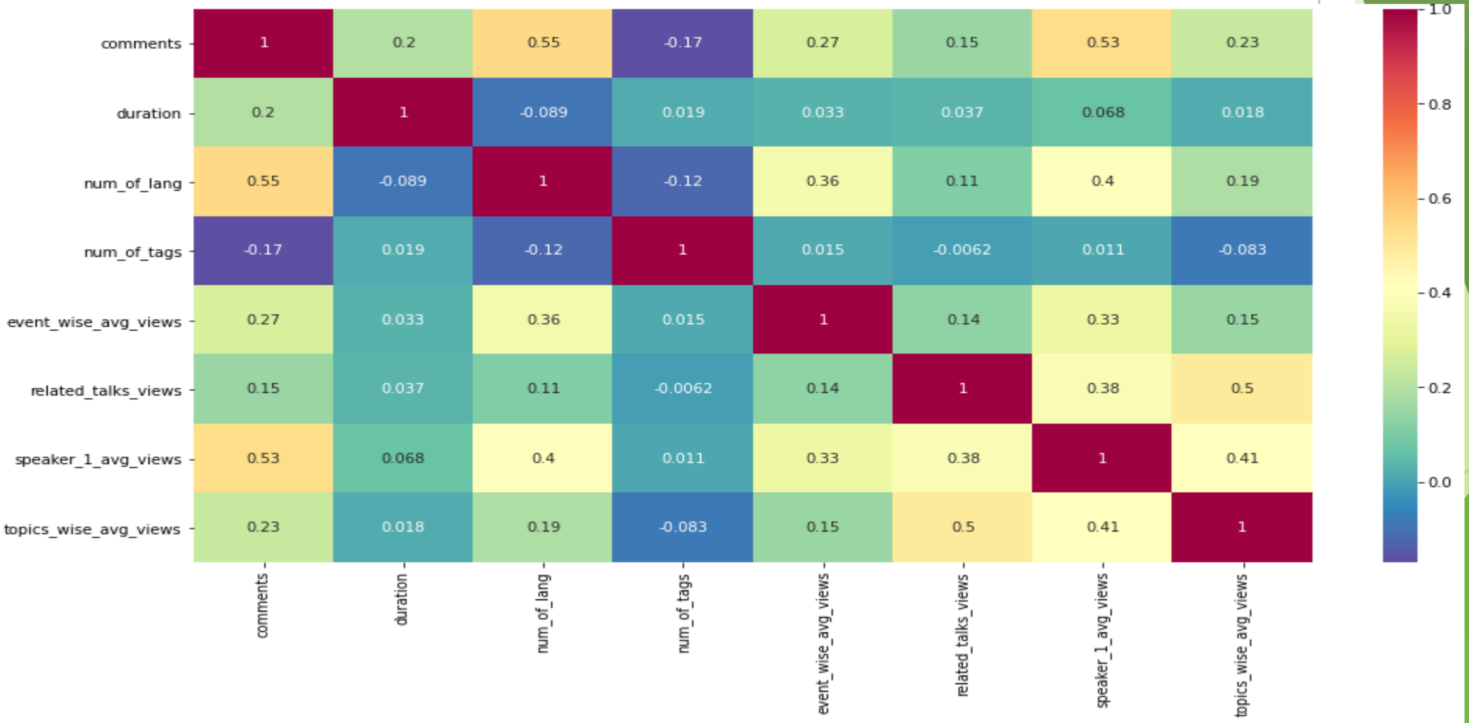




Feature selection



Collinearity



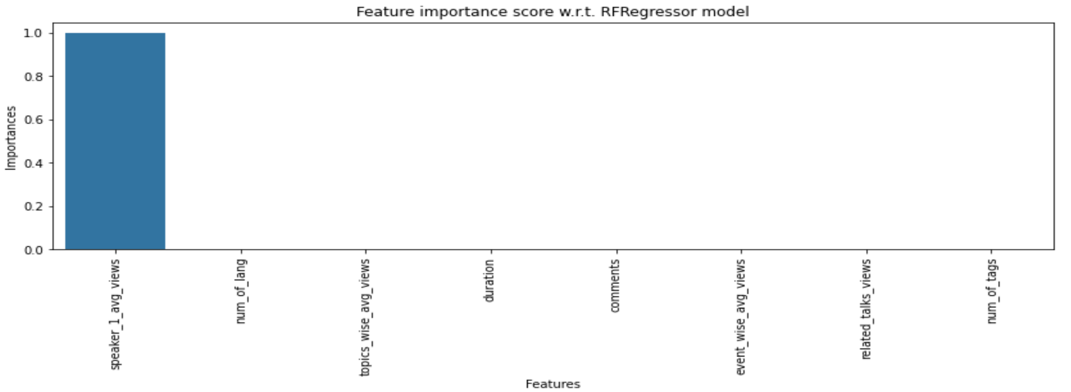
	comments	duration	num_of_lang	num_of_tags	event_wise_avg_views	related_talks_views	speaker_1_avg_views	topics_wise_avg_views
comments	1	0.2	0.55	-0.17	0.27	0.15	0.53	0.23
duration	0.2	1	-0.089	0.019	0.033	0.037	0.068	0.018
num_of_lang	0.55	-0.089	1	-0.12	0.36	0.11	0.4	0.19
num_of_tags	-0.17	0.019	-0.12	1	0.015	-0.0062	0.011	-0.083
event_wise_avg_views	0.27	0.033	0.36	0.015	1	0.14	0.33	0.15
related_talks_views	0.15	0.037	0.11	-0.0062	0.14	1	0.38	0.5
speaker_1_avg_views	0.53	0.068	0.4	0.011	0.33	0.38	1	0.41
topics_wise_avg_views	0.23	0.018	0.19	-0.083	0.15	0.5	0.41	1

Modelling

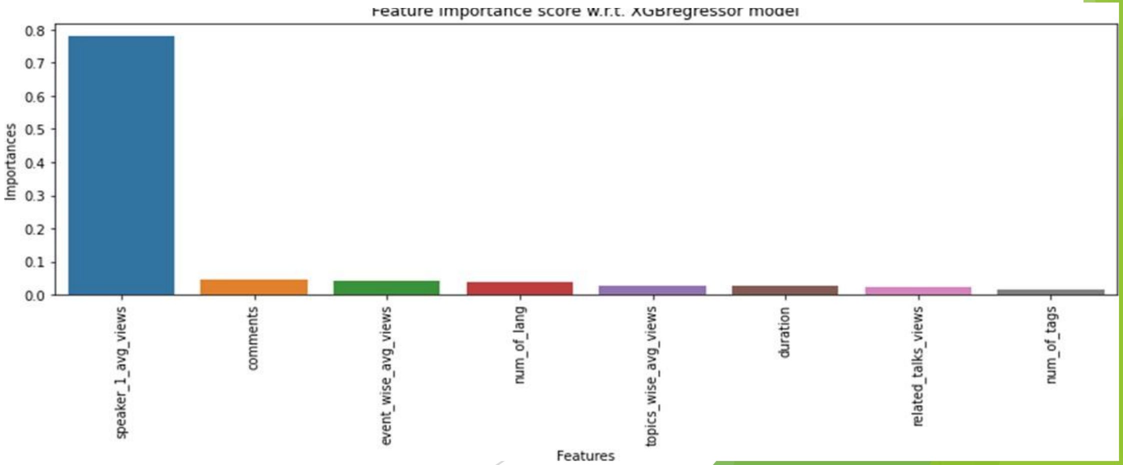
- Linear Regression
- Random Forest Regressor
- XGB Regressor

Feature Importance

- Random Forest
- Regressor

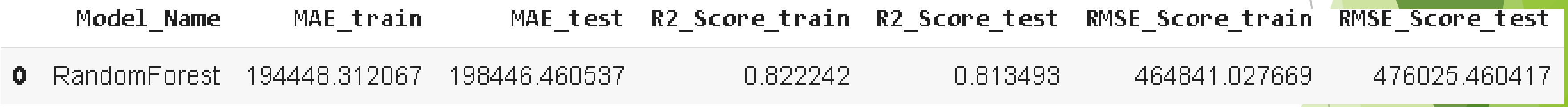


- XGBoost Regressor

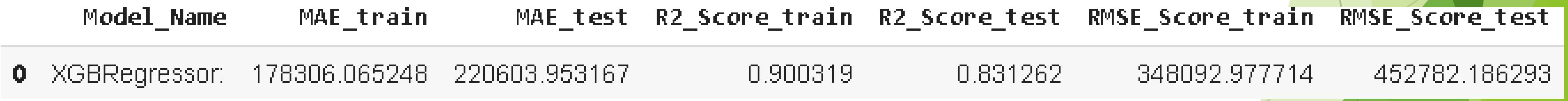


Model Selection

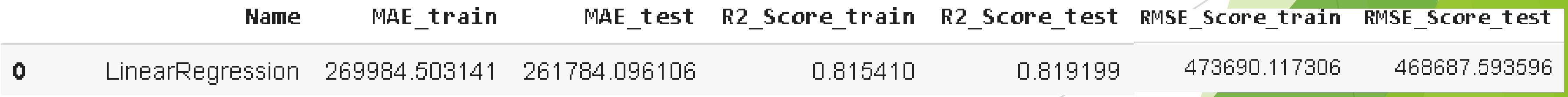
- Out of all the models Random Forest Regressor is the best model according to MAE
- MAE is the best deciding factor because it is linear, and it is not affected by outliers.



	Model_Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	RandomForest	194448.312067	198446.460537	0.822242	0.813493	464841.027669	476025.460417



	Model_Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	XGBRegressor:	178306.065248	220603.953167	0.900319	0.831262	348092.977714	452782.186293



	Name	MAE_train	MAE_test	R2_Score_train	R2_Score_test	RMSE_Score_train	RMSE_Score_test
0	LinearRegression	269984.503141	261784.096106	0.815410	0.819199	473690.117306	468687.593596

Challenges

- 1. Dataset has lot of categorical features with high cardinality. So,its conversion to meaningful numerical data was a tedious task.
- 2. Treatment of outliers in numerical features
- 3. Creation of new features to be added in the model
- 4. Selection of right features for modelling
- 5. Selection of right model with best scores

Conclusion

- We have built a predictive model, which could help TED in predicting the views on the talks uploaded on TEDx website.
- In all 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.

THANK YOU !!