Capstone Project 2

Ted Talks Views Prediction

Team Members

Mandar Khatavkar

Adi Ingrole

Discussion points

- 1. Problem statement
- 2. Data Summary
- 3. Exploratory Data Analysis
- 4. Feature Engineering
- 5. Correlation
- 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. TED is 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

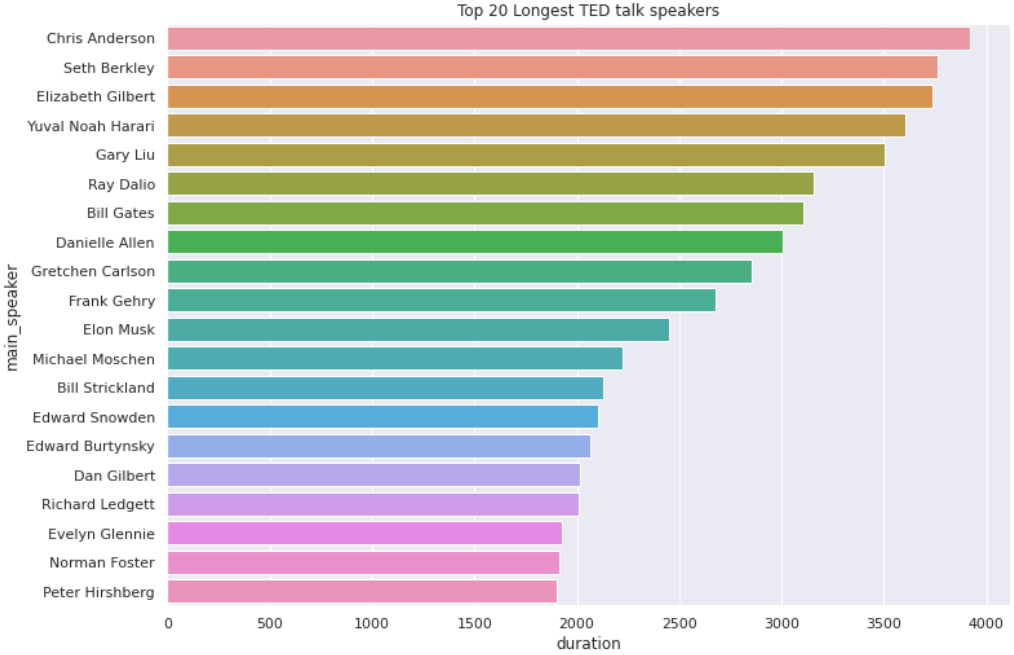
- **For Numerical values:**
- we used KNNImputer to
- impute missing values from numerical values.

- **For categorical values:**
- we used simpleImputer to
- impute missing values from categorical values.

Out[200...

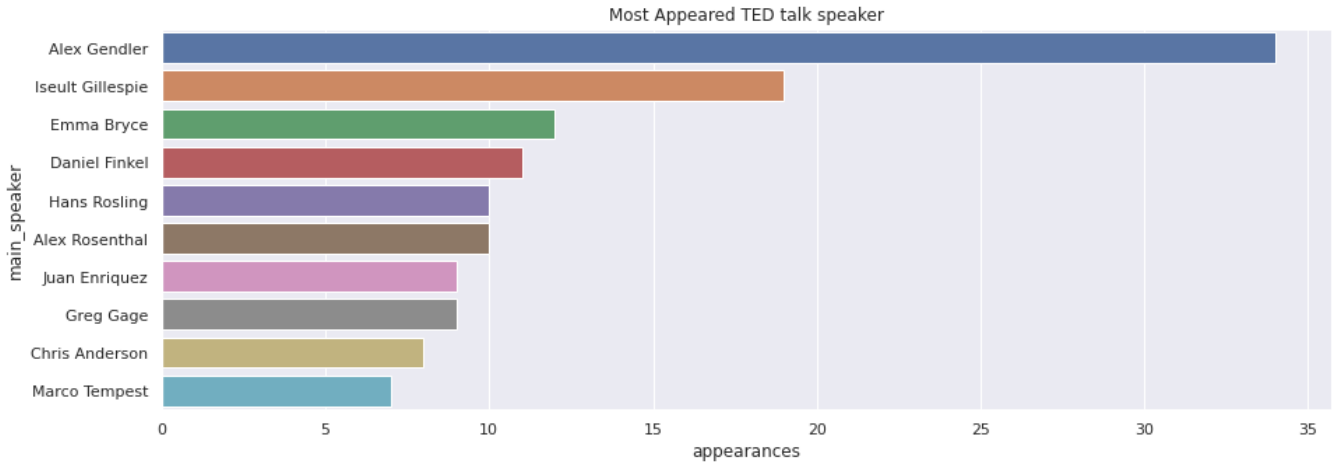
	Total	Percentage
comments	655	0.163546
occupations	522	0.130337
about_speakers	503	0.125593
all_speakers	4	0.000999
recorded_date	1	0.000250
talk_id	0	0.000000
description	0	0.000000
url	0	0.000000
related_talks	0	0.000000
topics	0	0.000000
duration	0	0.000000
event	0	0.000000
available_lang	0	0.000000
native_lang	0	0.000000
title	0	0.000000
published_date	0	0.000000
views	0	0.000000
speaker_1	0	0.000000
transcript	0	0.000000

Overview of Speaker column



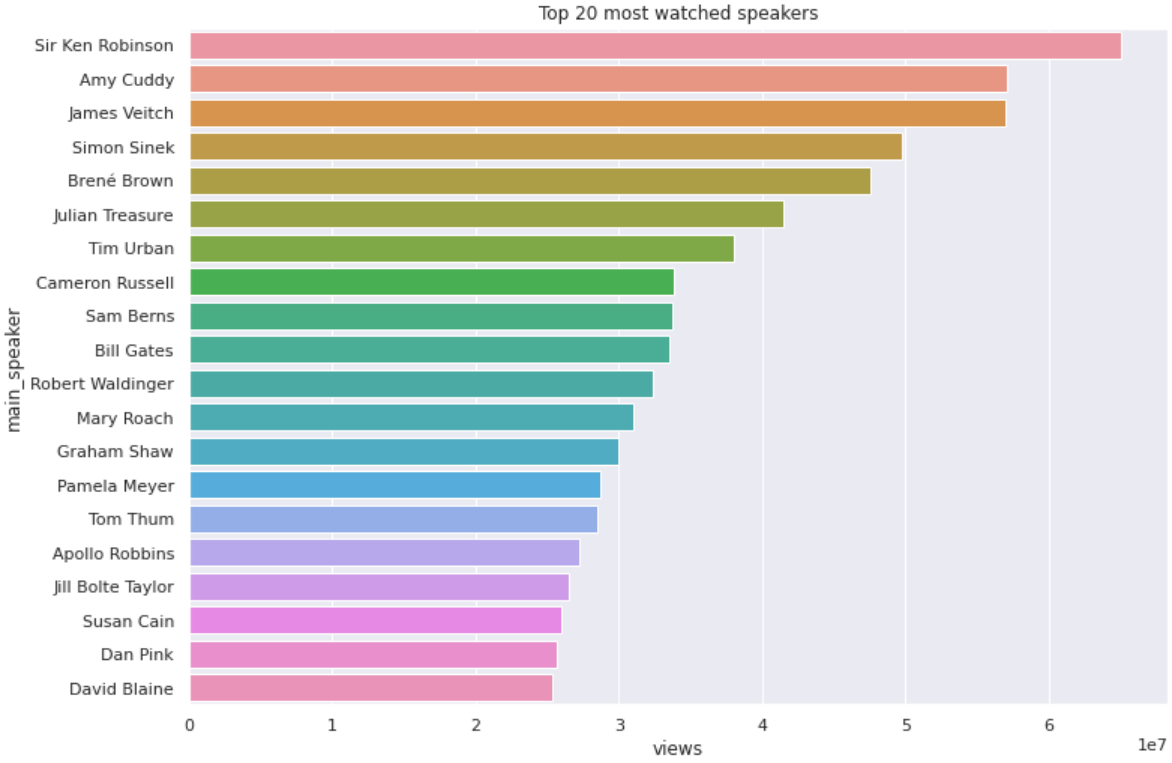
- The above plot shows, Chris Anderson talked for really long time. A lot people do not like to watch longer videos unless it is very interesting. Let’s find out how many views these speaker has got.

Most appeared speaker



- The above plot shows, Alex Gendler is most appeared speaker in TED so lets see if he is in most famous speakers or not.

Most popular speakers:



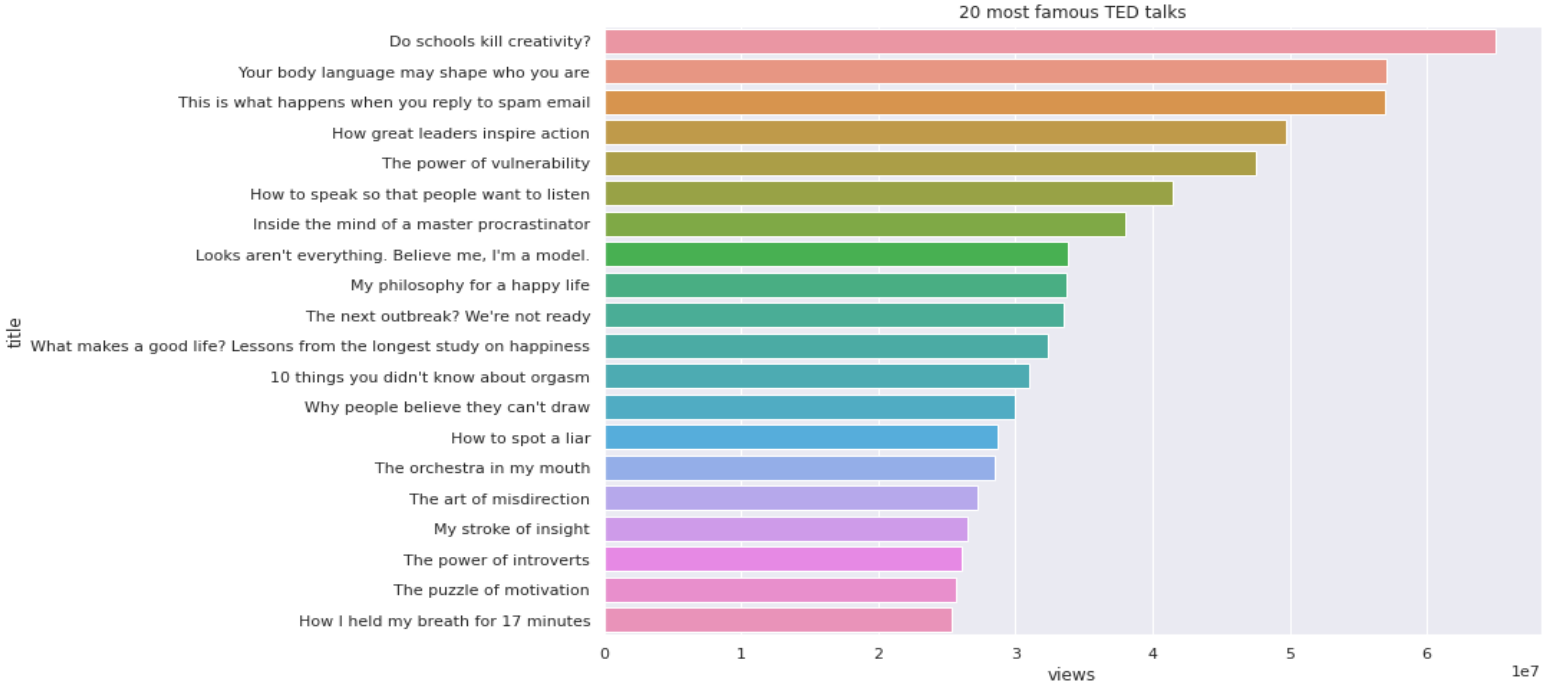
In this plot, as we can see Chris Anderson is not here.

So there is no high correlation between views and duration.

As well as Alex Gendler is not in this list so even if speaker appeared most times doesn't mean he is most popular

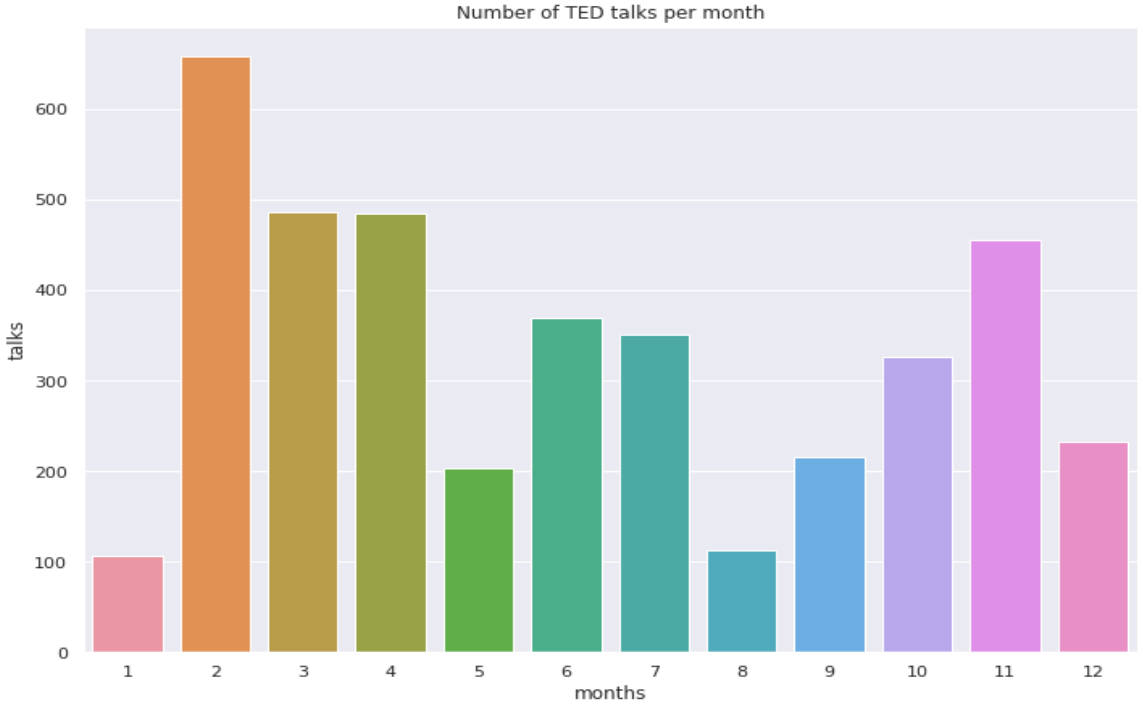
Most popular title?

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



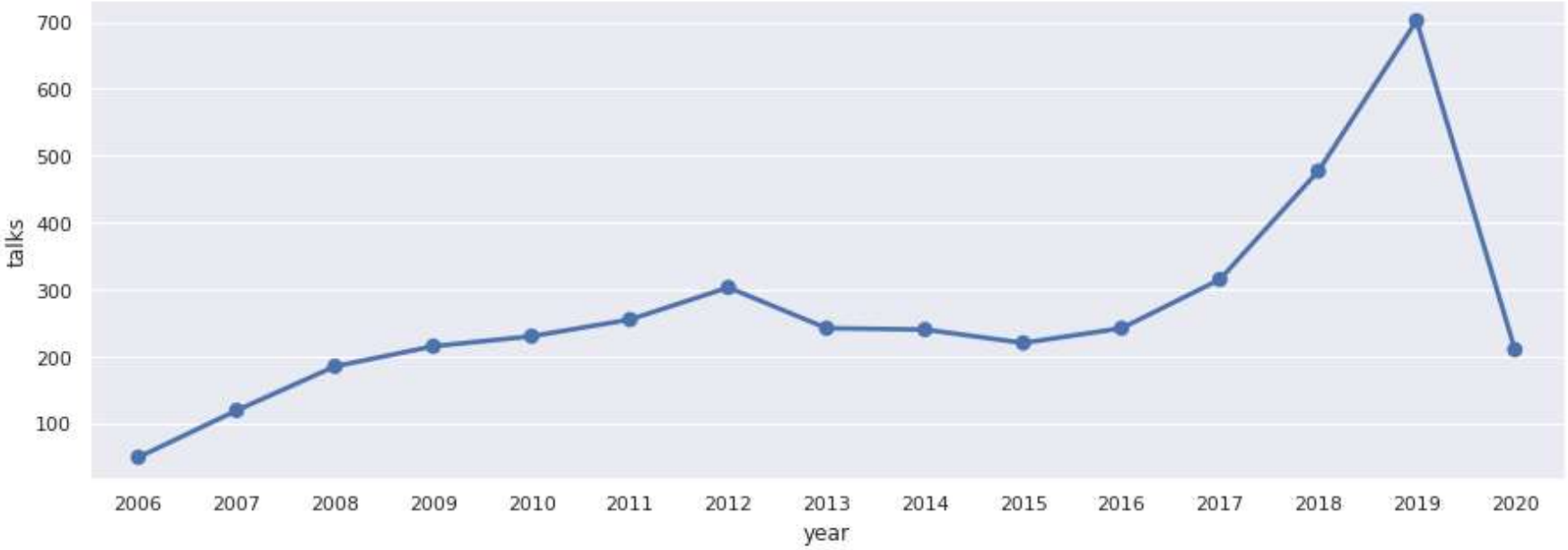
Most popular topic tags?

Overview of recorded_date



- Most videos are recorded in February but January and August has Minimum number of talks

Published year overview



As expected, the number of TED Talks have gradually increased over the years since its inception in 2006.

2018 has most number of TED talks per year

After 2019 we can see big drop in chart due to pandemic

Feature Engineering

- 1. total_available_lang : number of available languages for video
- 2. total_related_talks : number of related talks
- 3. all_speakers_count : number of all speakers in the talk
- 4. total_topics : number of topic covered in a talk
- 5. occupation_count : number of occupations of speaker

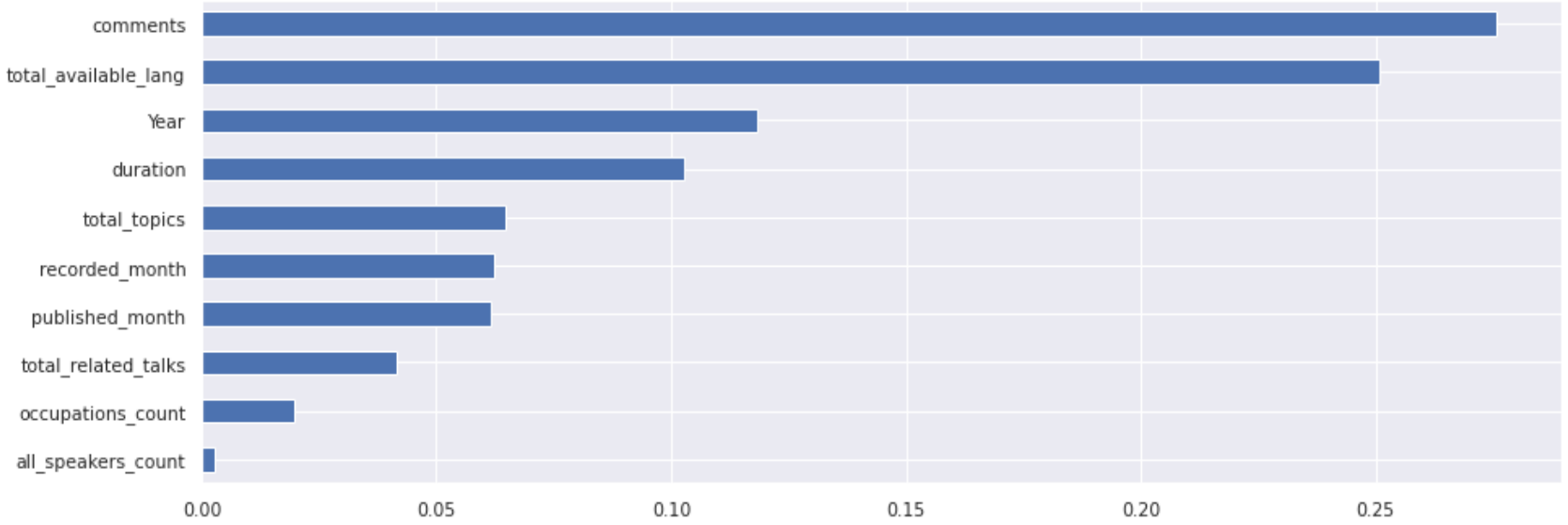
Correlation

Modelling

We used Two ML models for this project those are

- Linear Regression
- Random Forest Regressor

Feature Importance



Model Selection

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

Out[270...

	Model	MSE	RMSE	MAE	r2e
0	Linear Regression	8.255358e+12	2.873214e+06	1.425220e+06	0.222831
1	Random Forest Regressor	6.918926e+12	2.630385e+06	1.123729e+06	0.348645

Challenges

- 1. Dataset has lot of categorical features with high cardinality. So, its conversion to meaningful numerical data was a tedious task.
- 2. 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.
- Performed Exploratory data analysis on various features, then carried out feature engineering and encoding of categorical columns, handled missing values in the dataset then carried out feature selection and build various models
- Following models have been used:

1. 1) Linear Regression
- 2) Random Forest Regressor
- Evaluated these models on various metrics like MSE, RMSE, MAE ,R2 score. Finally selected the best model out of these two.
- After evaluating the performance of all the models, the best model is Random Forest Regressor.

Thank You !

Mandar Khatavkar
Adi Ingrole