Capstone Project
TEDtalk Views Prediction

Team Members

Charishma Suddala

Swathi V Hebbar

CONTENTS

INTRODUCTION

TED(Technology, Entertainment, Design) is a non profit devoted to spreading ideas, usually in the form of short, powerful talks. It is an American media organization that posts talks online for free distribution under the slogan “ideas worth spreading:”. These talks address a wide range of topics within the research and practice of science and culture, often through story telling.

The notable programs and initiatives of TED include TED include TED talks, TED Conferences, TED Translators, TED-Ed.

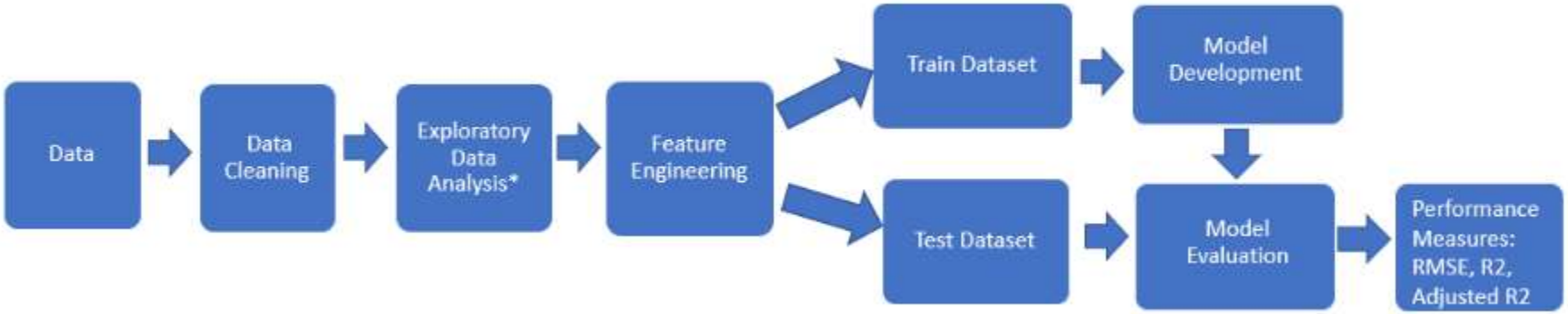
PROBLEM STATEMENT

Our objective is to predict the views of a TED talk that's been uploaded in the TEDx website. 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

METHODOLOGY

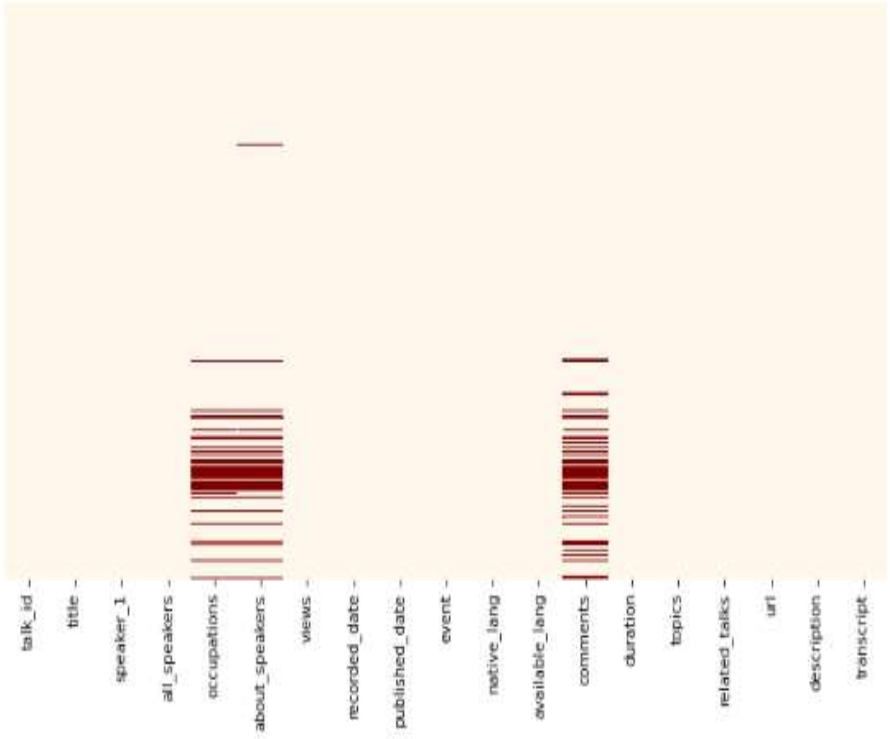


LOADING THE DATA AND DATA CLEANING

After loading the data, we can observe that the data frame contains
4005 rows with 19 variables. And we are trying to have an insight on missing values

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4005 entries, 0 to 4004
Data columns (total 19 columns):
 #   Column           Non-Null Count  Dtype  
---  ------           --------------  -----  
 0   talk_id          4005 non-null   int64  
 1   title            4005 non-null   object 
 2   speaker_1        4005 non-null   object 
 3   all_speakers     4001 non-null   object 
 4   occupations      3483 non-null   object 
 5   about_speakers   3502 non-null   object 
 6   views            4005 non-null   int64  
 7   recorded_date    4004 non-null   object 
 8   published_date   4005 non-null   object 
 9   event            4005 non-null   object 
 10  native_lang      4005 non-null   object 
 11  available_lang   4005 non-null   object 
 12  comments         3350 non-null   float64
 13  duration         4005 non-null   int64  
 14  topics           4005 non-null   object 
 15  related_talks    4005 non-null   object 
 16  url              4005 non-null   object 
 17  description      4005 non-null   object 
 18  transcript       4005 non-null   object 
dtypes: float64(1), int64(3), object(15)
memory usage: 594.6+ KB
```

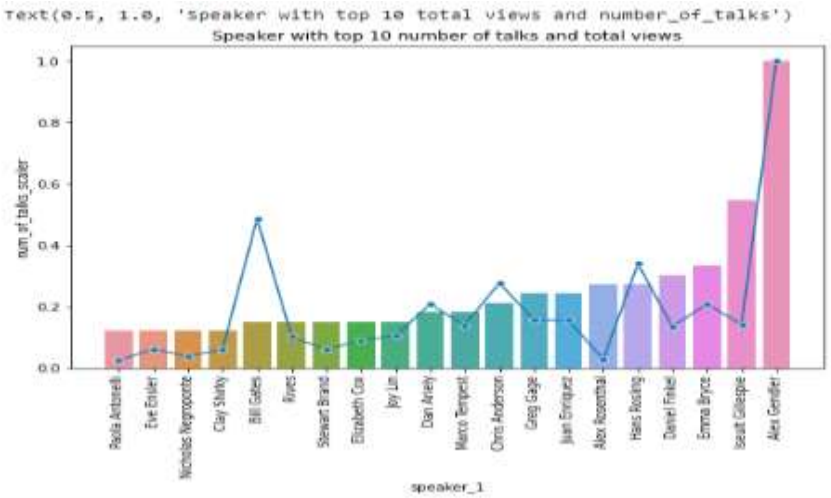
SPREAD OF MISSING VALUES

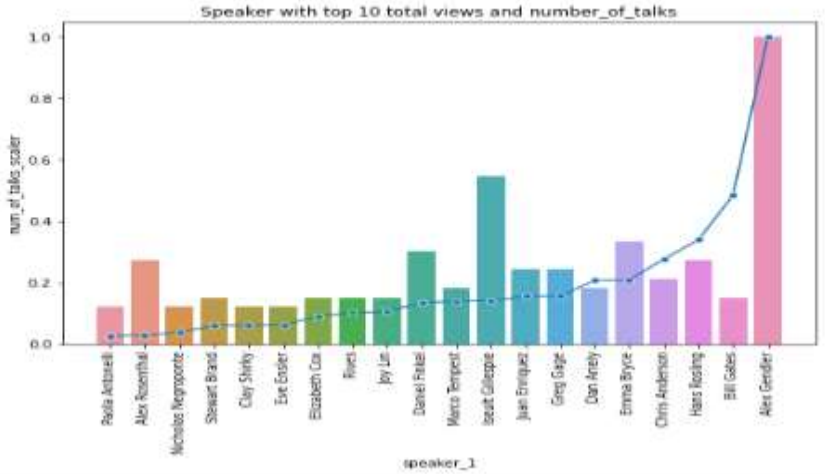


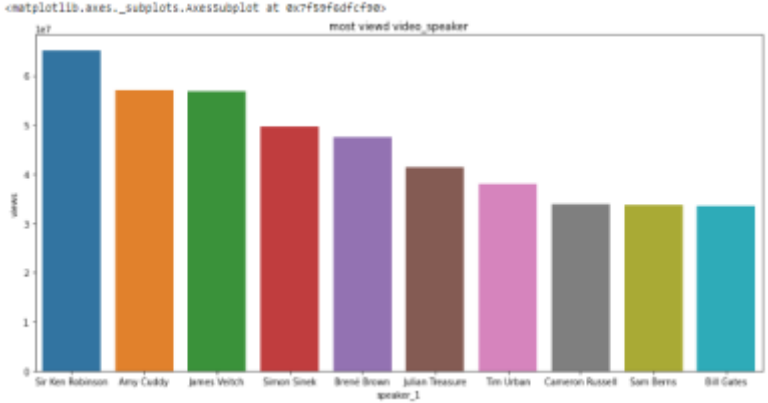
	Missing Values	% of Total Values	Data Type
comments	655	16.4	float64
occupations	522	13.0	object
about_speakers	503	12.6	object
all_speakers	4	0.1	object
recorded_date	1	0.0	object
talk_id	0	0.0	int64
description	0	0.0	object
url	0	0.0	object
related_talks	0	0.0	object
topics	0	0.0	object
duration	0	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 20 total views with respect to number of talks



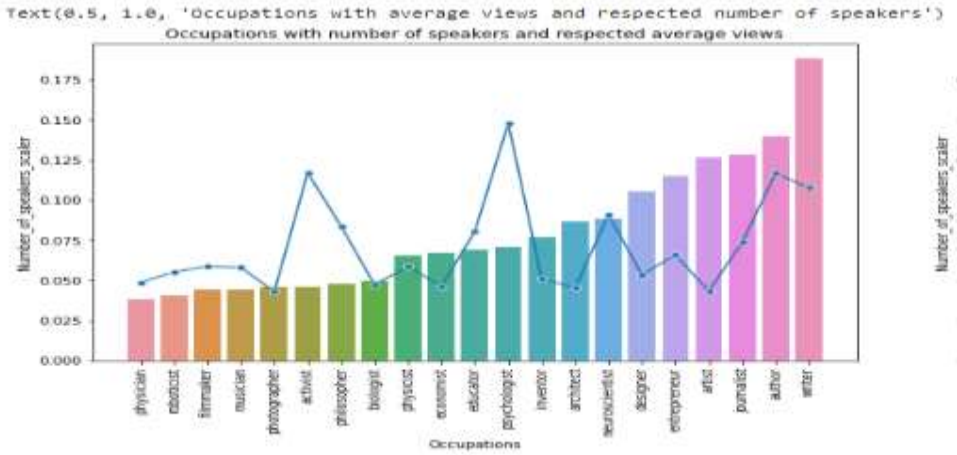


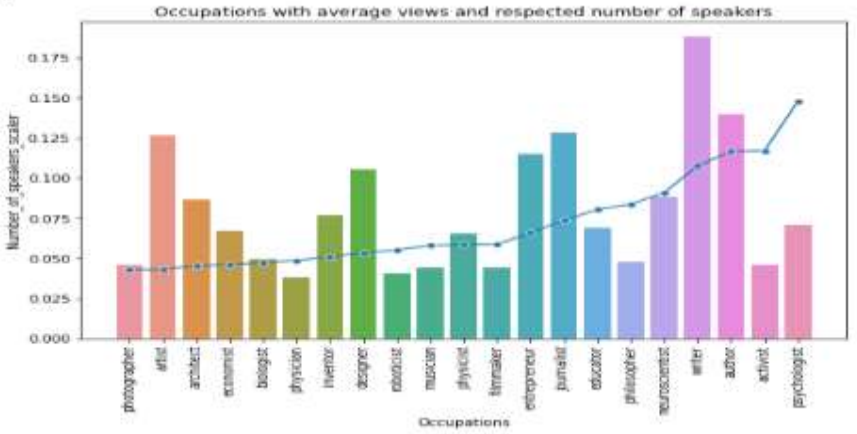


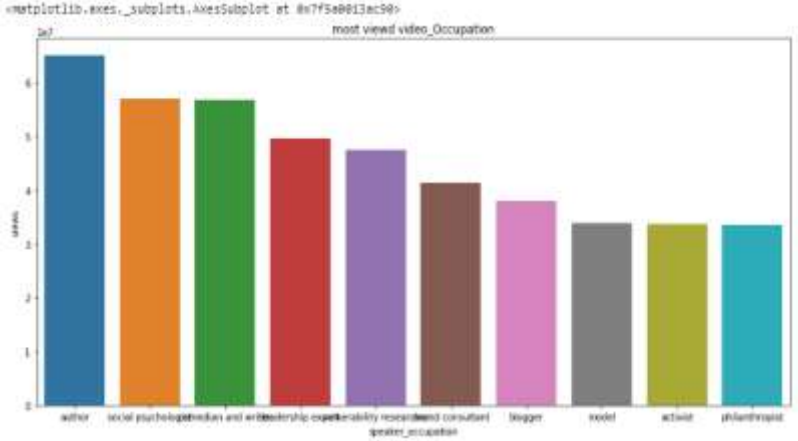
Speakers with max number of views for a single TED talk

EDA (continued)

Top occupations of the speakers with respect to number of talks and views



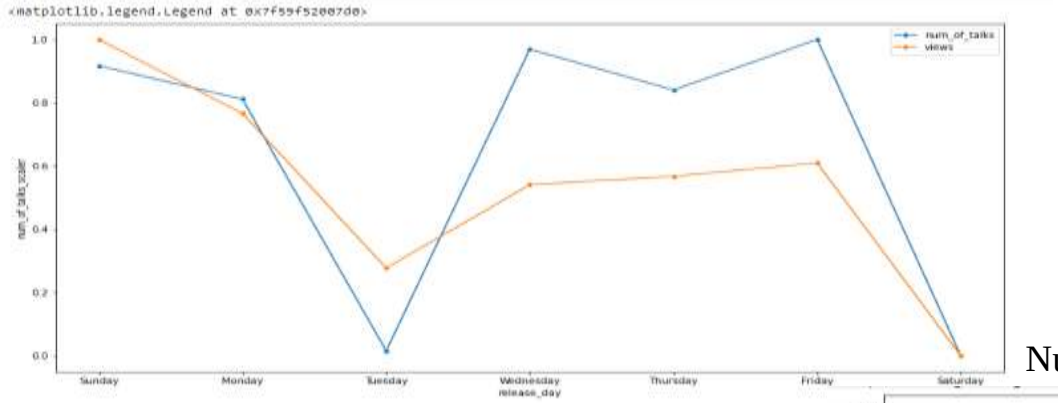




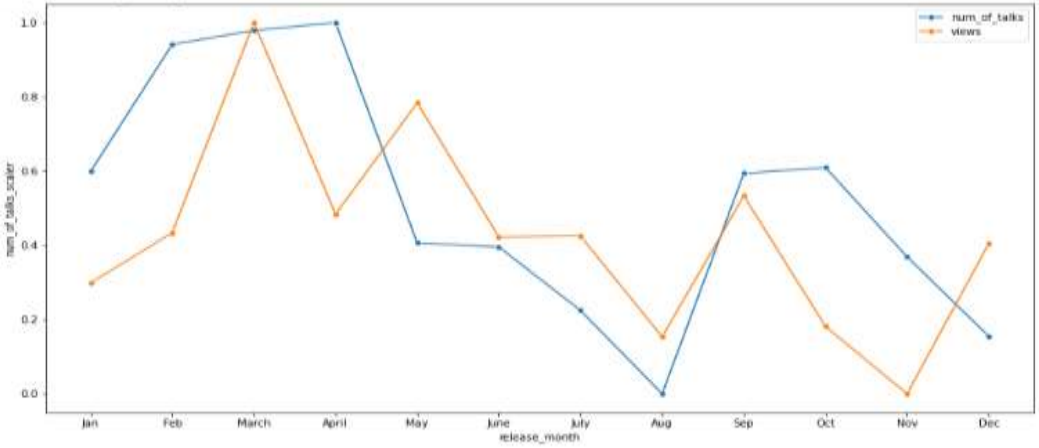
Occupations of speakers which got maximum number of views for a single TED talk

EDA (continued)

Number of talks and average views received with respect to weekdays

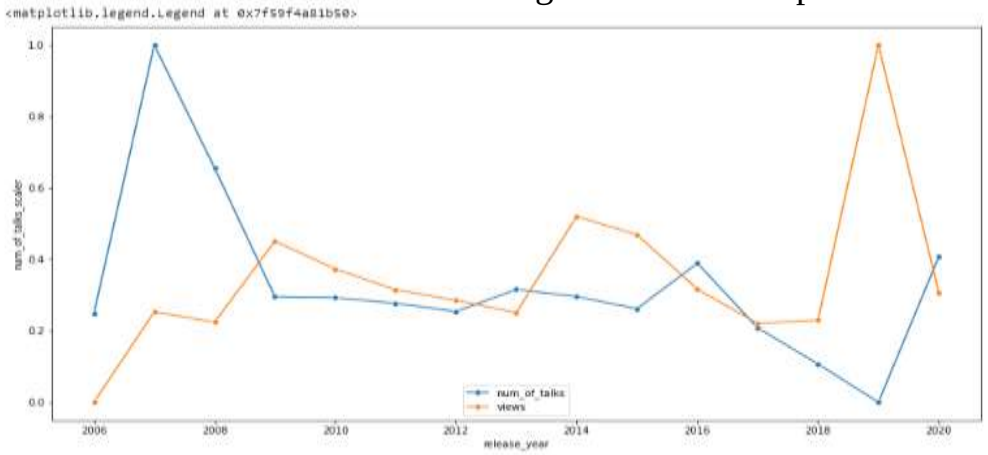


Number of talks and average views with respect to Month

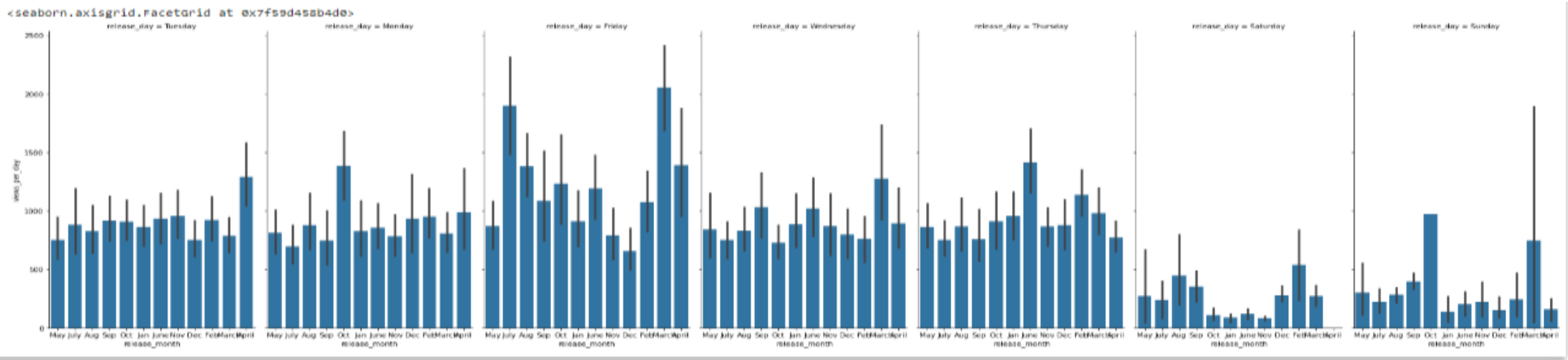


EDA (continued)

Number of talks and average views with respect to Year

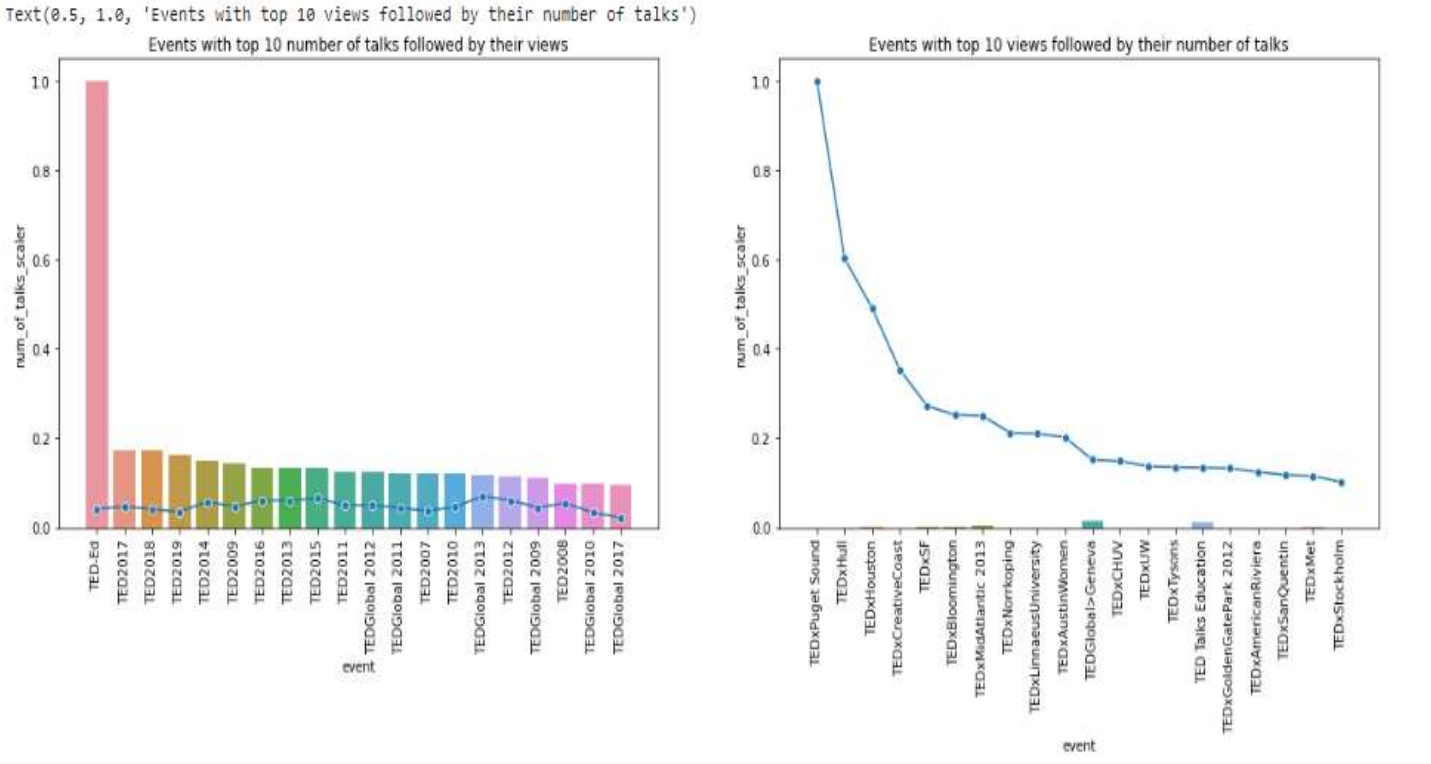


Views per day with respect to every day on monthly basis



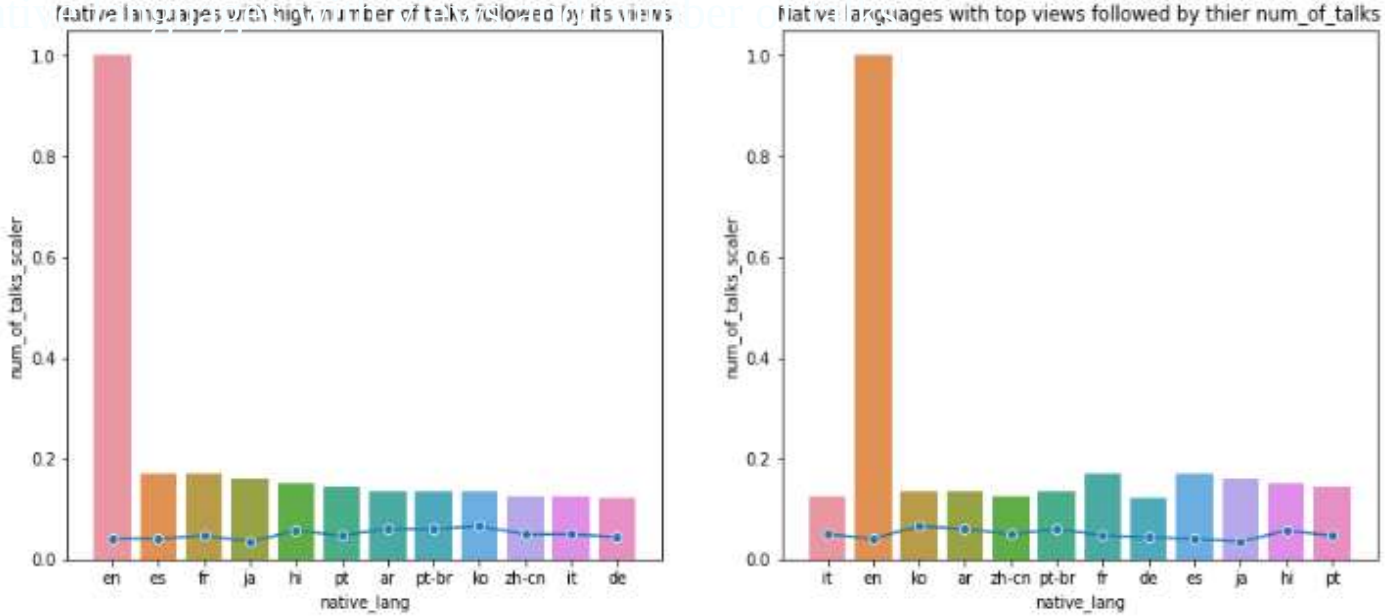
EDA (continued)

Events with top 10 number of talks with respect to views

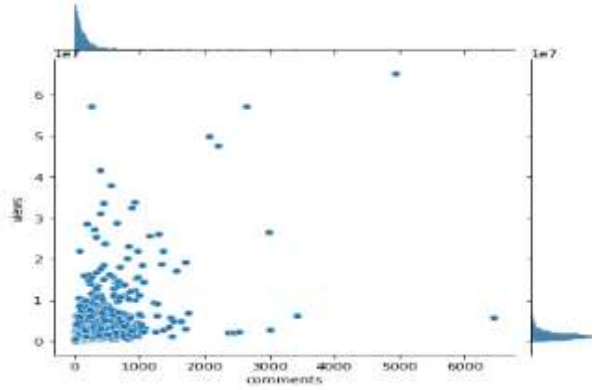


EDA (continued)

Native languages with views and number of talks

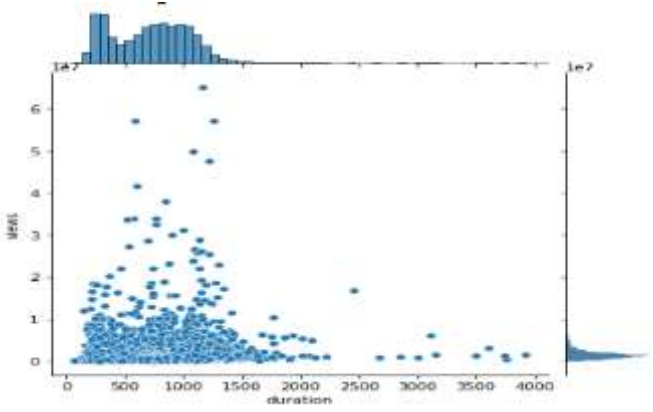


EDA (continued)

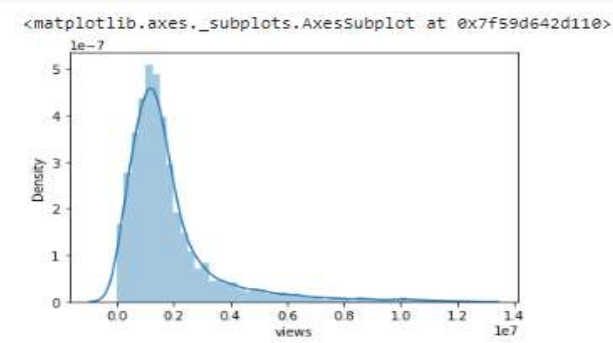


Comments and views

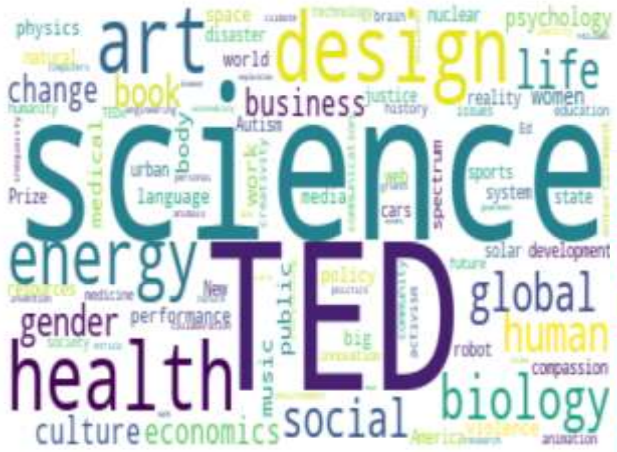
Duration and views



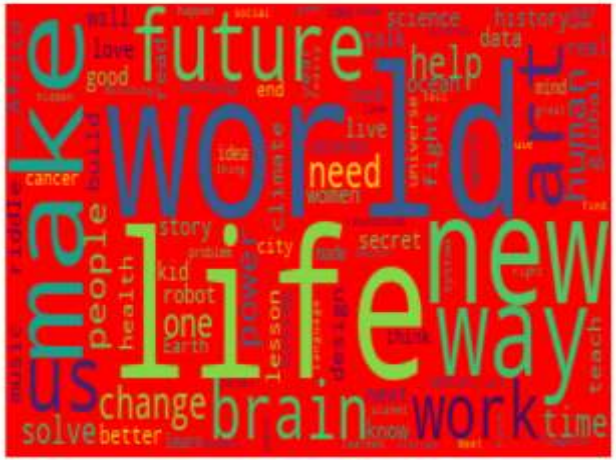
Distribution of views



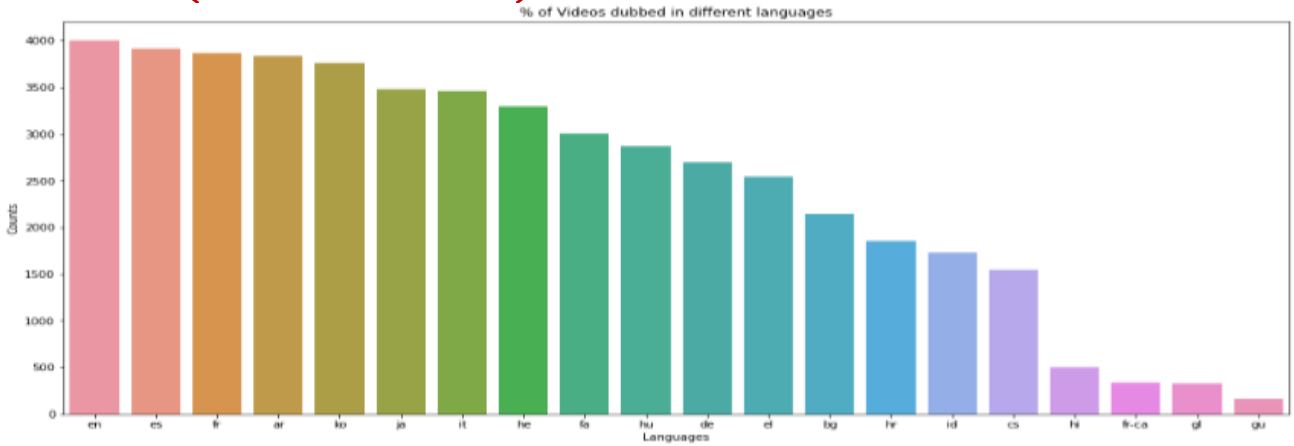
EDA (continued)



Word cloud for Topics , Description and Title columns

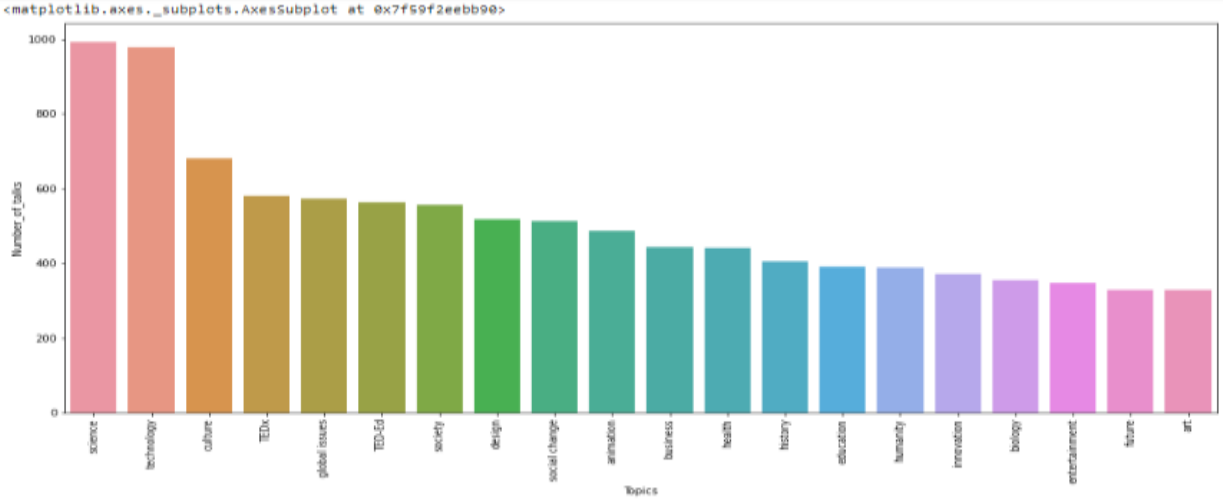


EDA (Continued)



Count of available languages

Topics with respect to number of talks



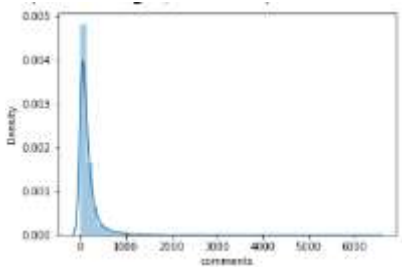
EDA CONCLUSIONS:

- 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.
- Portuguese is the language which received maximum average views

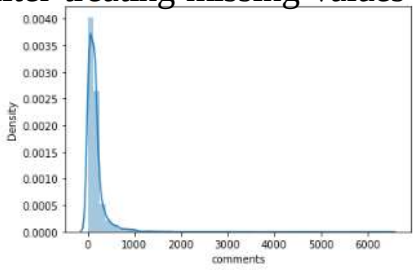
TREATING MISSING VALUES AND OUTLIERS

- While dealing with the missing values, we observed that there exists a column ‘comments’. It has more than 16% of missing values and has skewed data, so we tried predicting the missing values by KNN Imputer.
- Later Outliers in the independent numerical columns were treated by IQR and dependent numerical columns outliers were treated by Z score.

Before treating missing values

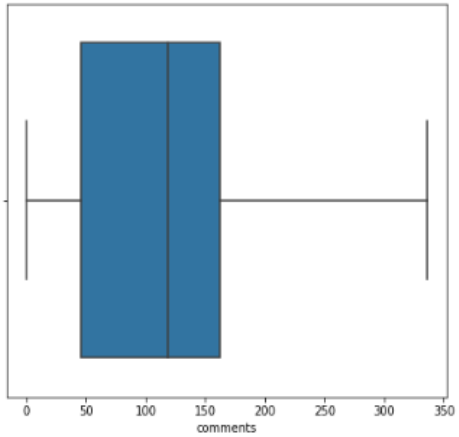


After treating missing values

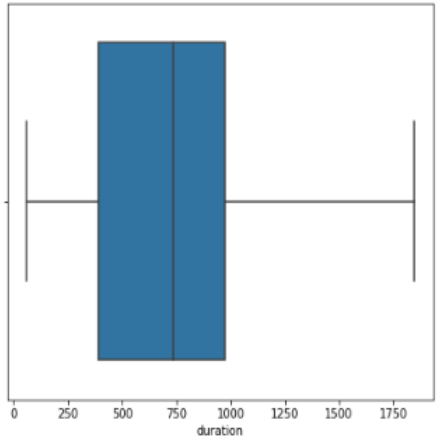


Box plots after applying Outlier treatment for the numerical variables

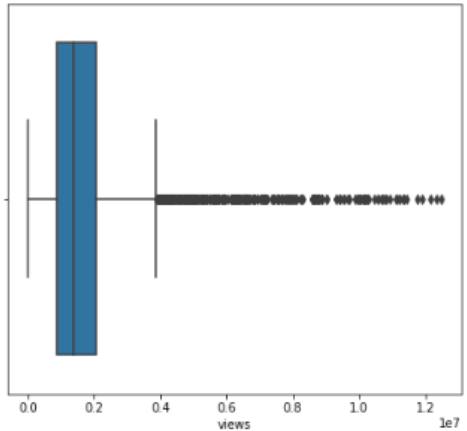
Comments



Duration



Views



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 consists of various process :
  (1) Feature Creation (2) Transformation (3) Feature Extraction

(1) Feature Extraction: Feature extraction is the process of extracting features from a data set to identify useful information.

(2) Feature Creation: Creating features involves creating new variables which will be most helpful for our model.

(3) Transformations: Feature transformation is simply a function that transforms features from one representation to another.

FEATURE ENGINEERING (Continued)

- When feature engineering activities are done correctly, the resulting dataset is optimal and contains all of the important factors that affect the business problem. As a result of these datasets, the most accurate predictive models and the most useful insights are produced.
- Once we were done with creating/altering existing/new variables, we try the conditions which are necessary for Linear regression models.
- Those are –

(1) Linearity: The relationship between the independent and dependent variables must be linear..

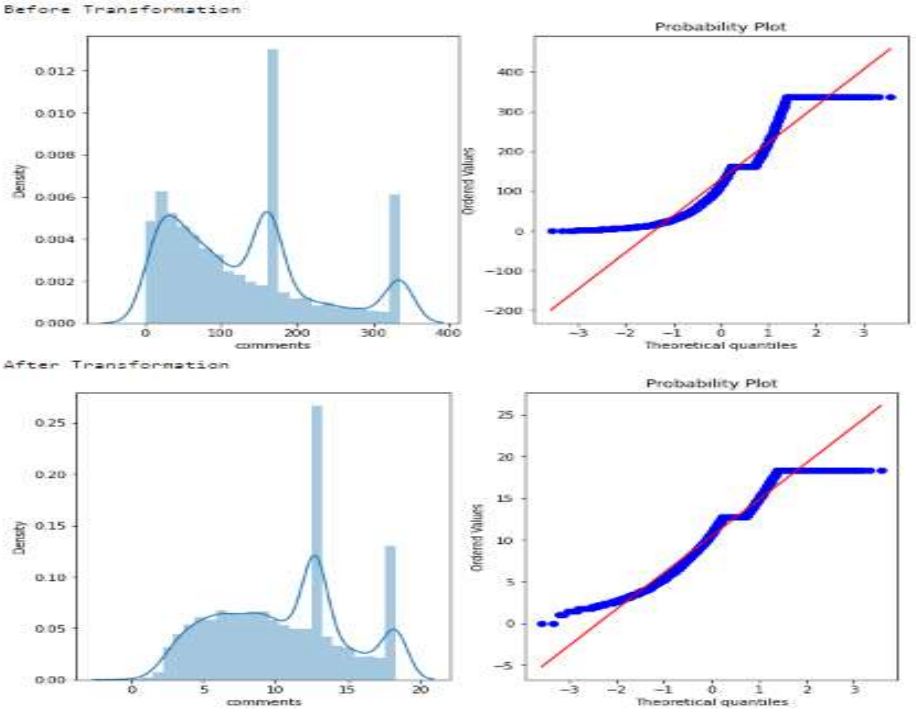
(2)There should be no or little multi-collinearity: Multi-collinearity is the phenomenon when a number of the explanatory variables are strongly correlated.

(3)Normality: All residuals should follow a normal distribution in Linear Regression

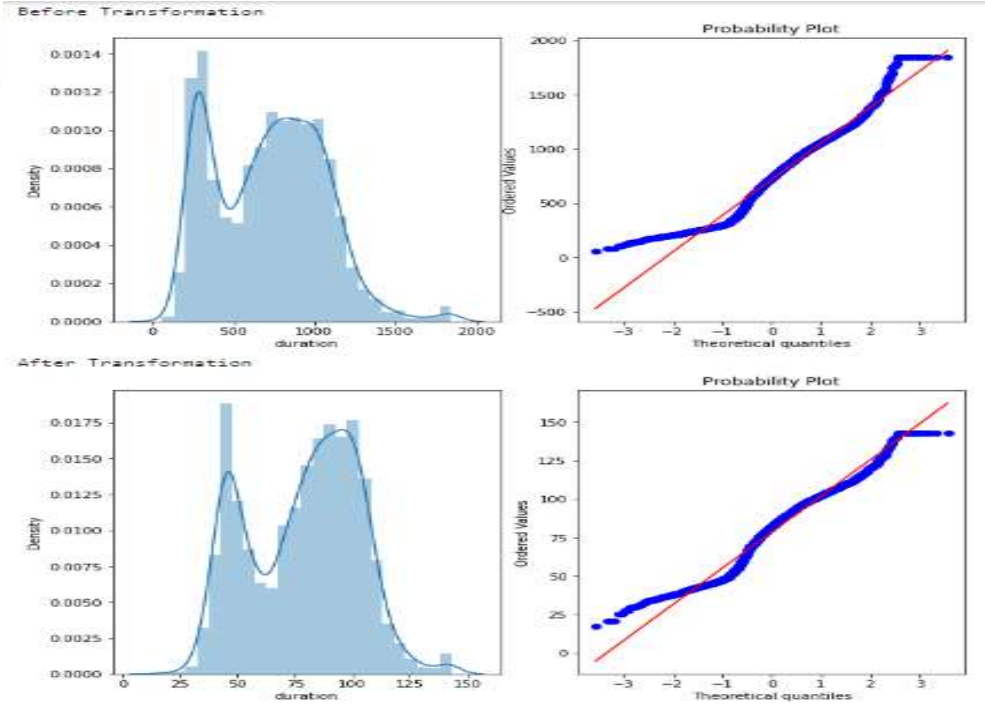
FEATURE ENGINEERING (continued)

Applying Transformations for the featured columns to convert it into normal distribution.

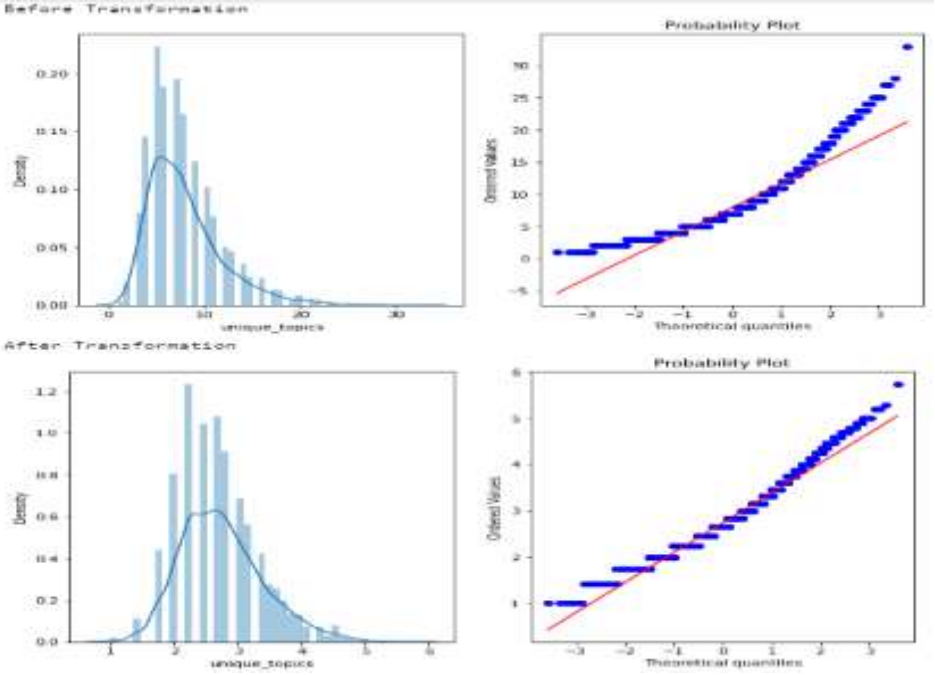
Comments



Duration

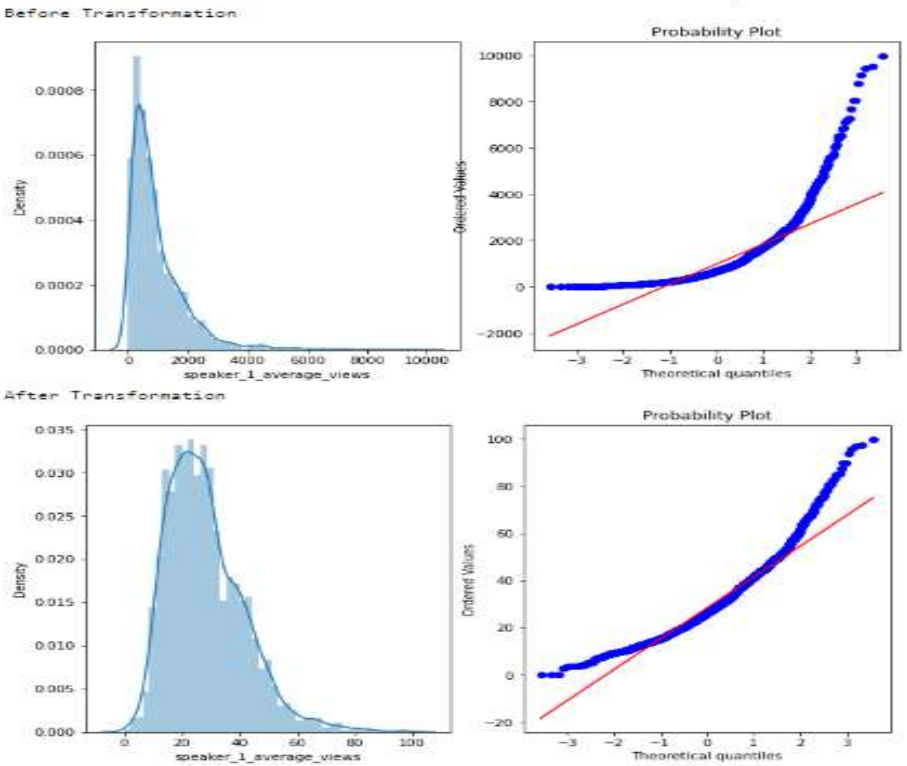


Feature Engineering (Continued)



Speaker average Views

Unique topics



Before Building a model…

After finished with Feature engineering, this is the data frame that we have.

	comments	duration	video_age	speaker_1_average_views	unique_topics	release_day_Friday	release_day_Monday	release_day_Saturday	release_day_Sunday	release_day_Thursday
0	0.899735	0.638816	1.000000	0.236269	0.421535	0.0	0.0	0.0	0.0	0.0
1	0.607493	0.778386	1.000000	0.238352	0.421535	0.0	0.0	0.0	0.0	0.0
2	0.807332	0.703545	1.000000	0.198906	0.421535	0.0	0.0	0.0	0.0	0.0
3	0.377964	0.740186	0.997429	0.145179	0.260523	0.0	1.0	0.0	0.0	0.0
4	1.000000	0.645974	0.997429	0.291424	0.421535	0.0	1.0	0.0	0.0	0.0

Correlation matrix for the featured numerical variables

	comments	duration	video_age	speaker_1_average_views	unique_topics
comments	1.000000	0.044590	0.380605	0.016621	-0.173275
duration	0.044590	1.000000	0.371631	-0.171023	0.034621
video_age	0.380605	0.371631	1.000000	-0.519892	-0.239105
speaker_1_average_views	0.016621	-0.171023	-0.519892	1.000000	0.159571
unique_topics	-0.173275	0.034621	-0.239105	0.159571	1.000000

Now that we have finished EDA, Feature Engineering, we have only the variables which are important.
These variables are transformed into normal distribution and been scaled. We can observe that even **One hot encoding** is done for the categorical variable.
We are all set to build models.
So lets just start with the train test split.

TRAIN – TEST SPLIT

After cleaning the data, the dataset is split into Train – Test datasets. This is done to ensure that our test dataset is completely isolated and there is no information leakage during the training process of machine learning models

DATA MODELING

- In Machine Learning, we use various kinds of algorithms to allow machines to learn the relationships within the data provided and make predictions based on patterns or rules identified from the dataset.
- So, regression analysis is a machine learning technique where the model predicts the output as a continuous numerical value.
- Many models were trained, from simple parametric models like Linear Regression to tree based models.

DATA MODELING (continued)

Few of the important Regression models which we have used in here are,

(1) **Linear regression** – This technique finds out a linear relationship between a dependent variable and the other given independent variables.

(2) **Random Forest Regressor** – Random Forests are an ensemble(combination) of decision trees. It executes by constructing a different number of decision trees at training time and mean prediction (for regression) of the individual trees.

(3) **Catboost Regressor –** It provides Machine Learning algorithms under gradient boost framework. It supports both numerical and categorical features.

DATA MODELING (continued)

4) LGBM Regressor – It is a boosting technique that uses tree based learning algorithm. It grows tree leaf wise rather than level wise.

5) XGBoost Regressor – It is also a boosting technique that uses gradient descent algorithm to minimize the loss when adding new tree models.

6) Extra Trees (Extremely Randomized Trees) - The ensemble learning algorithms. It constructs the set of decision trees. During tree construction the decision rule is randomly selected. This algorithm is very similar to Random Forest except random selection of split values.

EVALUATION OF MODELS

	Name	MAE_train	MAE_test	MSE_train	MSE_test	R2_Score_train	R2_Score_test	Adjusted_R2_score_train	Adjusted_R2_score_test	RMSE_Score_train	RMSE_Score_test
0	Regularized Linear Regression	0.6144	0.6071	1.2976	1.1652	0.8565	0.8773	0.855028	0.876106	1.1391	1.0794
1	Optimal Random Forest	0.4869	0.5228	1.0543	1.1324	0.8834	0.8808	0.882216	0.879594	1.0268	1.0641
2	LGBM	0.4874	0.5587	0.8714	1.0513	0.9036	0.8893	0.902643	0.888213	0.9335	1.0253
3	Catboost	0.5009	0.5828	0.8395	1.0686	0.9071	0.8875	0.906217	0.886368	0.9162	1.0338
4	XGBoost	0.5004	0.5783	0.9259	1.1052	0.8976	0.8836	0.896562	0.882478	0.9622	1.0513
5	Extra tree regressor	0.7418	0.7460	1.5823	1.5128	0.8250	0.8407	0.823228	0.839141	1.2579	1.2300

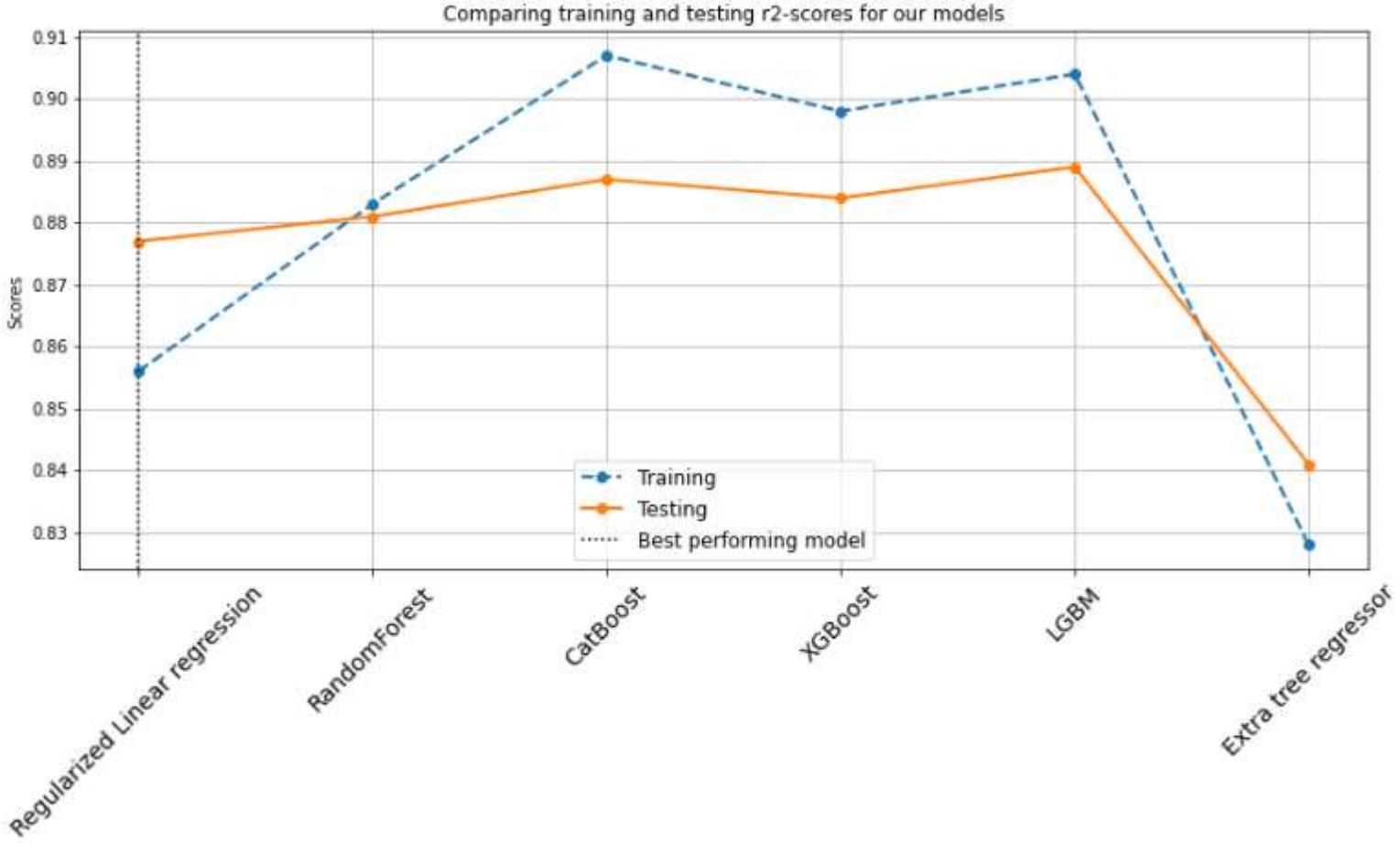
R-squared (R2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.

Mean absolute error (MAE) is the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.

Mean Squared Error (MSE) or **Mean Squared Deviation (MSD)** represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals. It is always non-negative, and values closer to zero are better.

Root Mean Squared Error (RMSE) is a common way of measuring the quality of the fit of the model. A value of zero would indicate a perfect fit to the data.

TRAIN AND TEST EVALUATION



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.

Thank You