Capstone Project

Supervised ML
(Regression)
On
Ted Talks View Prediction

Team Members

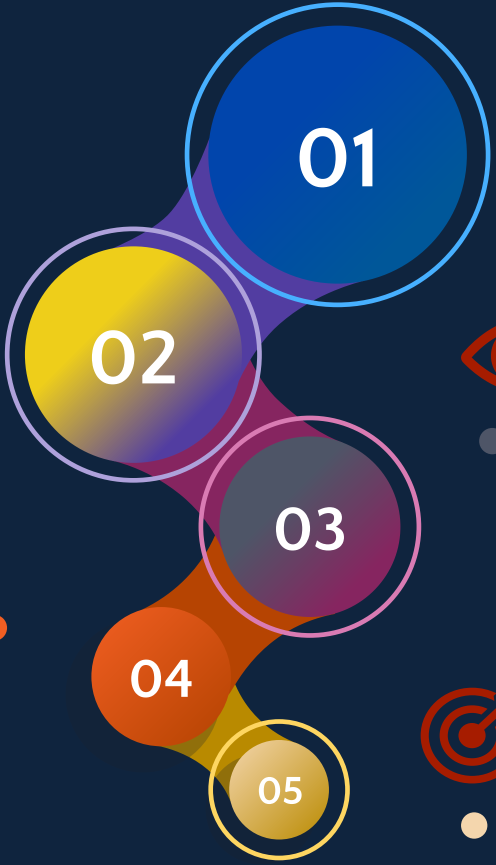
Arvind Krishna

Keshav Sharma

Sahil Ahuja

Jayesh Panchal

Roadmap of the Presentation



Introduction

A brief note about the project.

Data Cleaning

Fixing the given data set and making it ready for the next process.

EDA and Feature Engineering

Analysing and manipulating data into features for supervised learning.

Model Development.

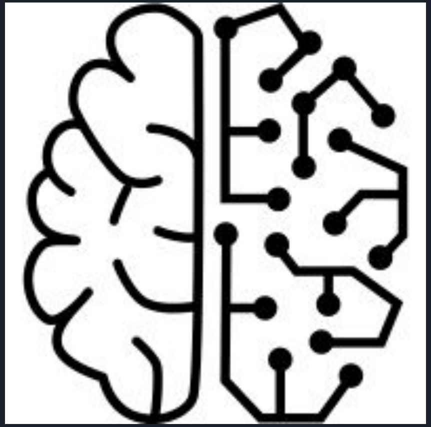
Checking the accuracy with different algorithms.

Conclusion and References

Introduction

What is machine learning ?

Machine learning (ML) is a subset of artificial intelligence (AI) which allows applications to become more accurate in predicting outcomes without being explicitly programmed to do so.



Machine learning algorithms use historical data as input to predict new output values.

Supervised ML (Regression)

Regression is the supervised machine learning technique which is used to predict the continuous values.

Using this technique we are going to predict the Ted Talks views.

About the TED-Talks

Ted Talks are influential videos from expert speakers on education, business, science, tech and creativity, with subtitles in 100+ languages.

TED talk is a recorded public-speaking presentation that was originally given at the main TED (technology, entertainment and design) annual event or one of its many satellite events around the world.

TED
Talks

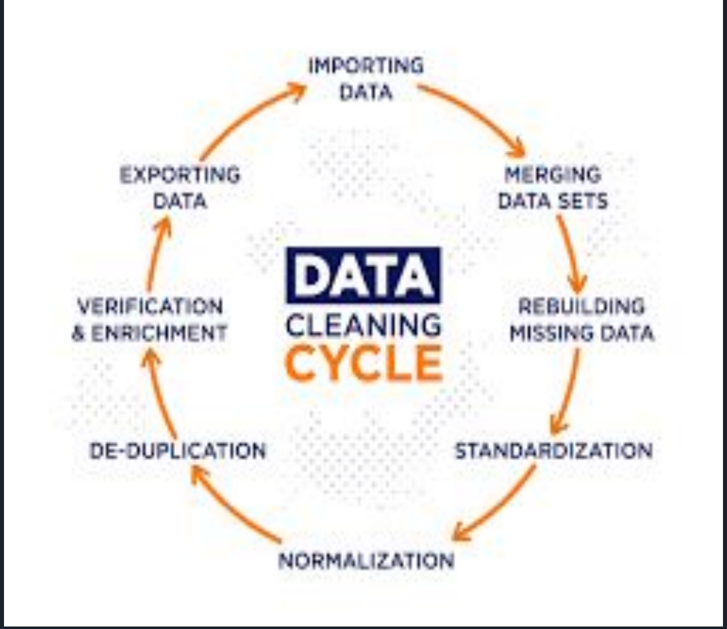
Objective

The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx platform.

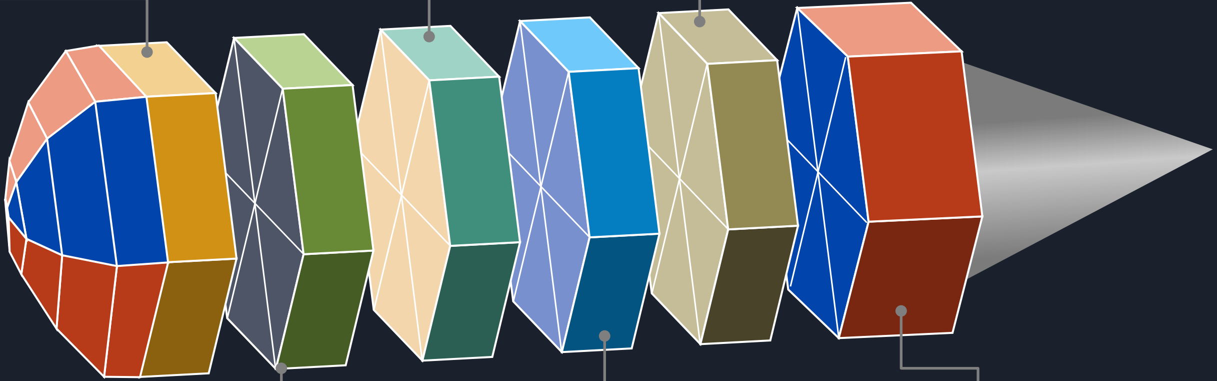
Data Cleaning

Data cleaning

Data cleaning is an important early step in the data analytics process in which you either remove or update information that is incomplete, incorrert, improperly formatted, duplicated, or irrelevant .



Loading and Discovering data



01 Loading data

02 Importing libraries

03 Exploring the data

04 Dealing with NAN values

05 Extracting columns

06 Dropping columns

data_ted_talks.csv

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

The datasets contain over 4,000 TED talks including transcripts in many languages.

Features

Some of the Categorical features include:-

- Topics, related_talks, speaker_1, description, published_date, event, available_lang, etc.

Numeric Features:-

- Comments, Duration

Dependant Variable:

- Views

EDA

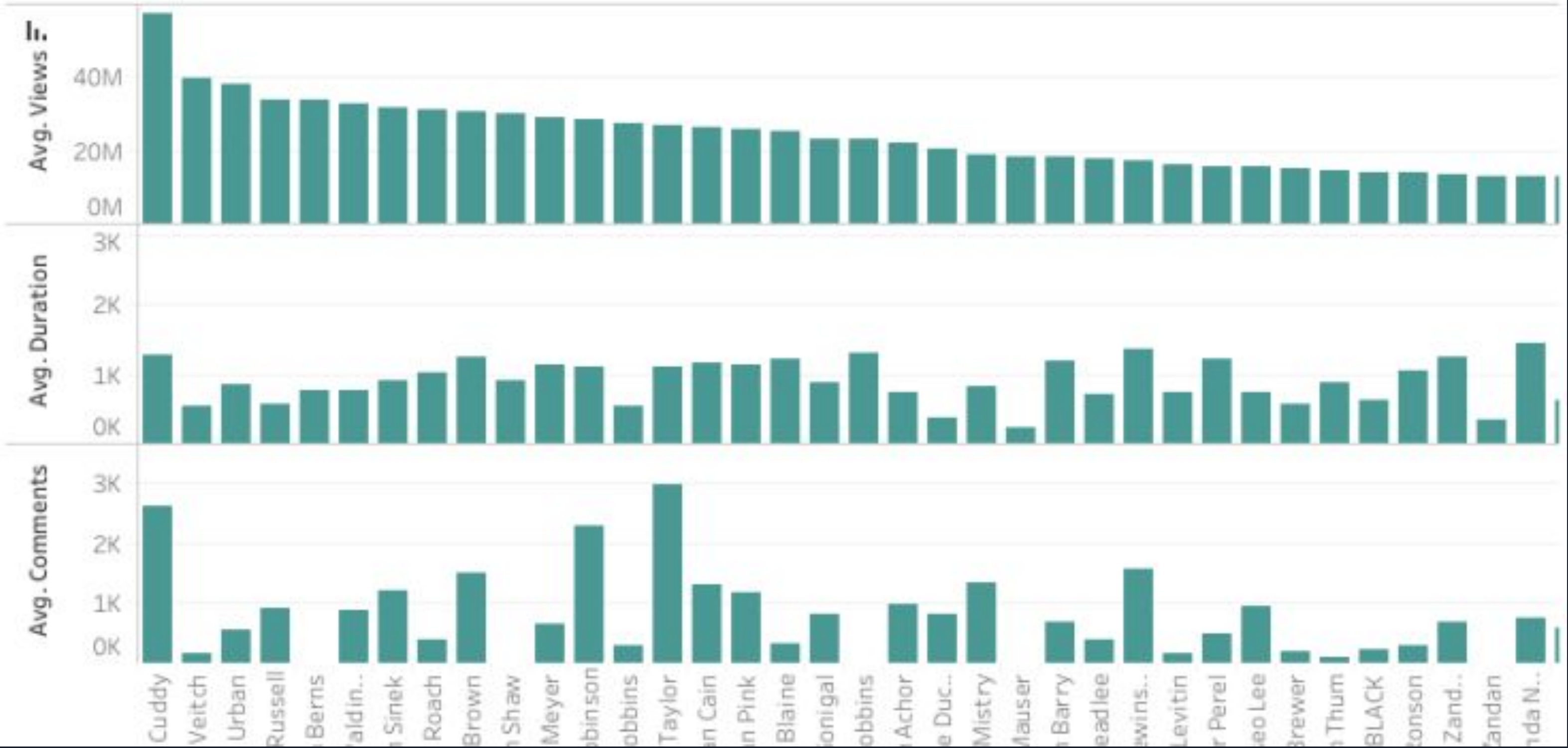
&

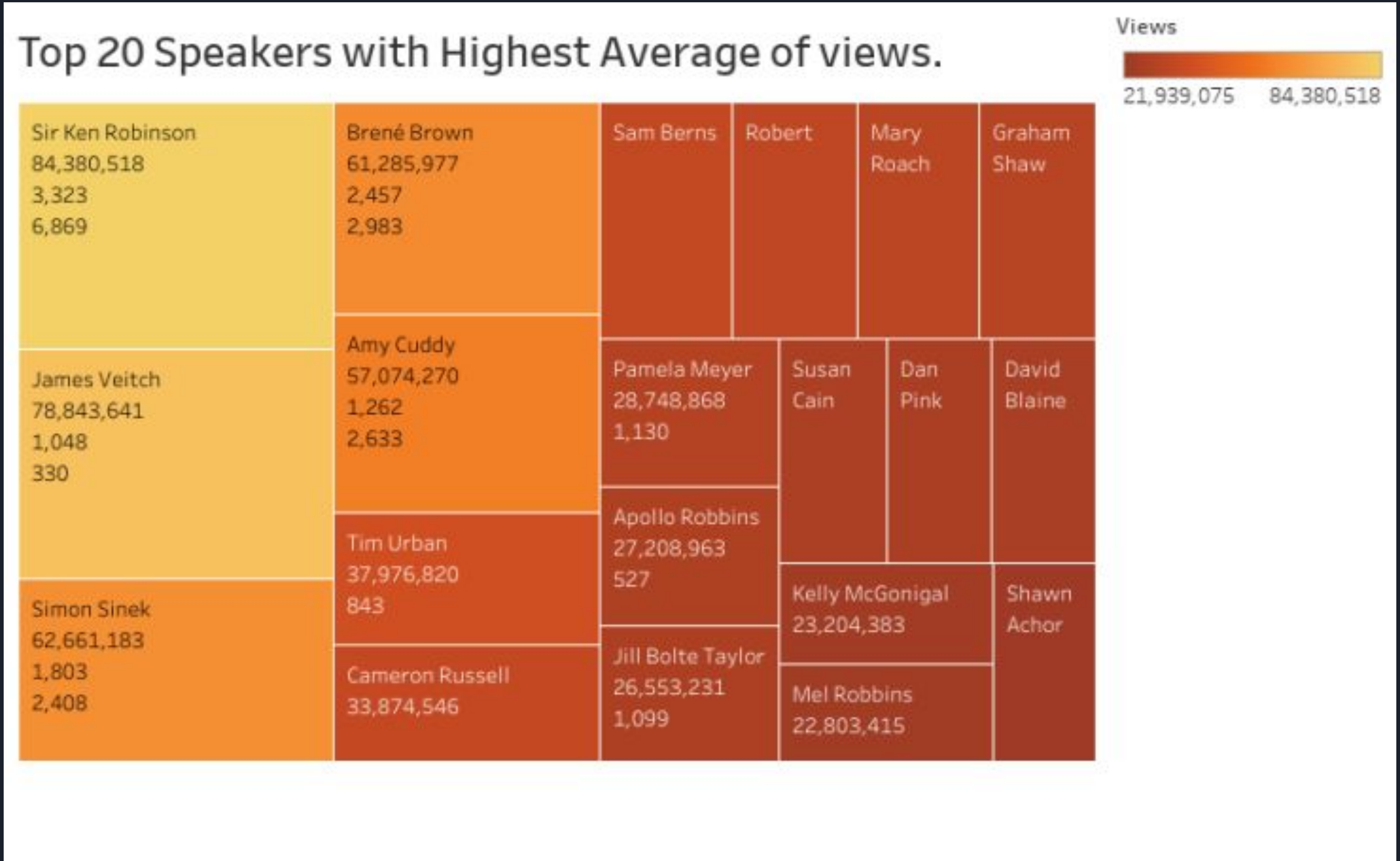
Feature Engineering

Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data.

We then use ‘Data visualization’, to create charts, graphs, or other forms of visualization, which makes information easier to analyse and interpret.

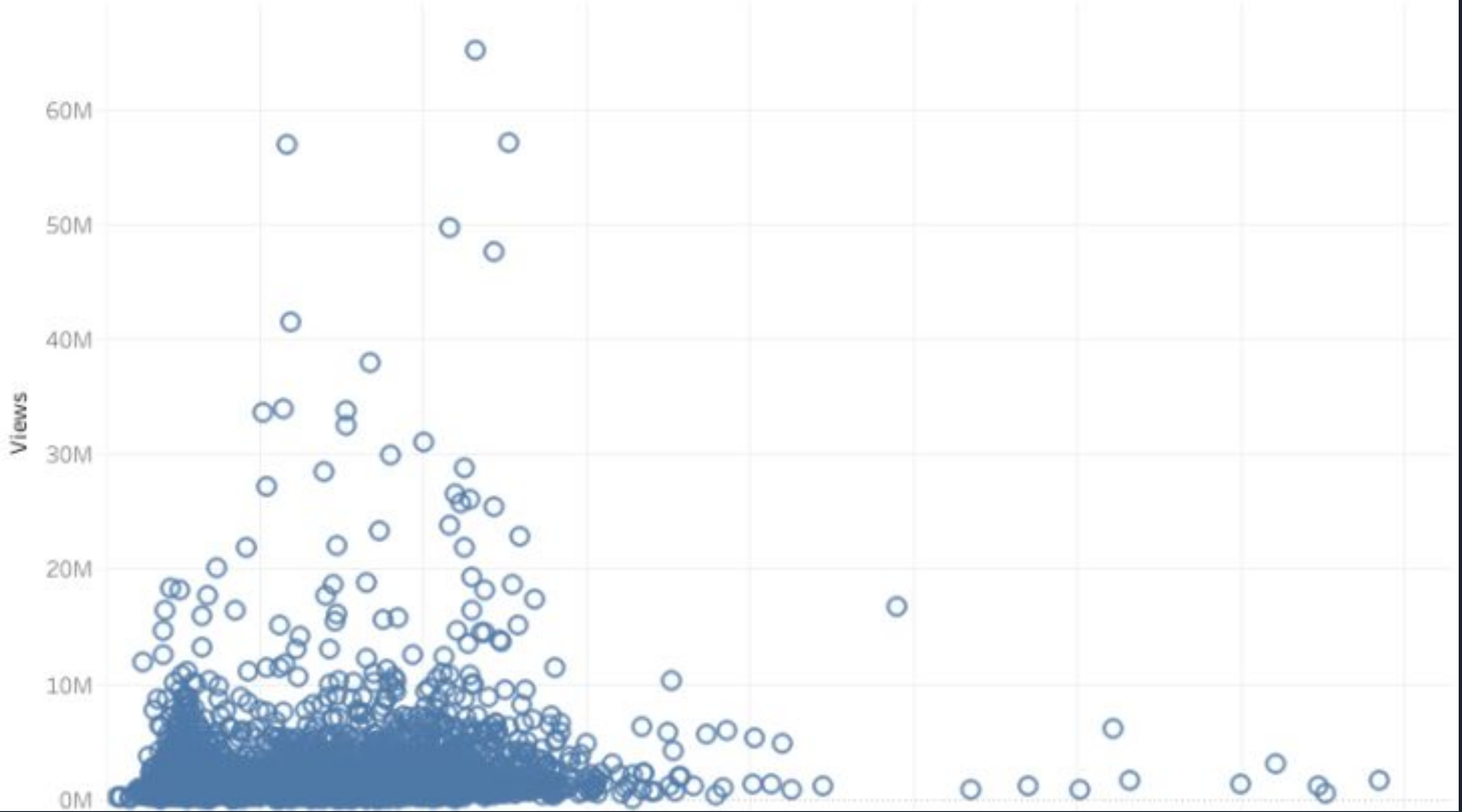
Speakers and their respective Average of views, duration and comments.

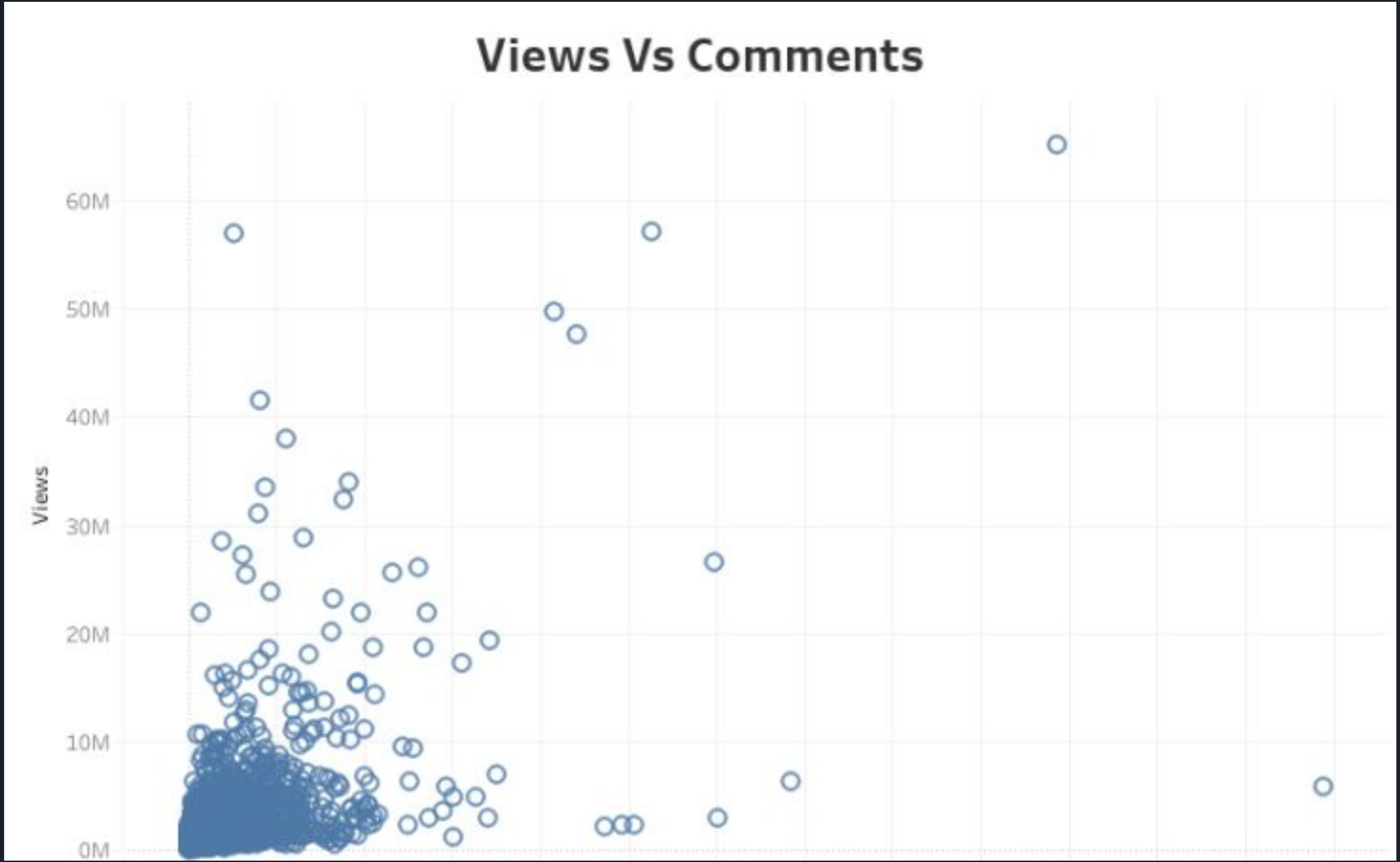




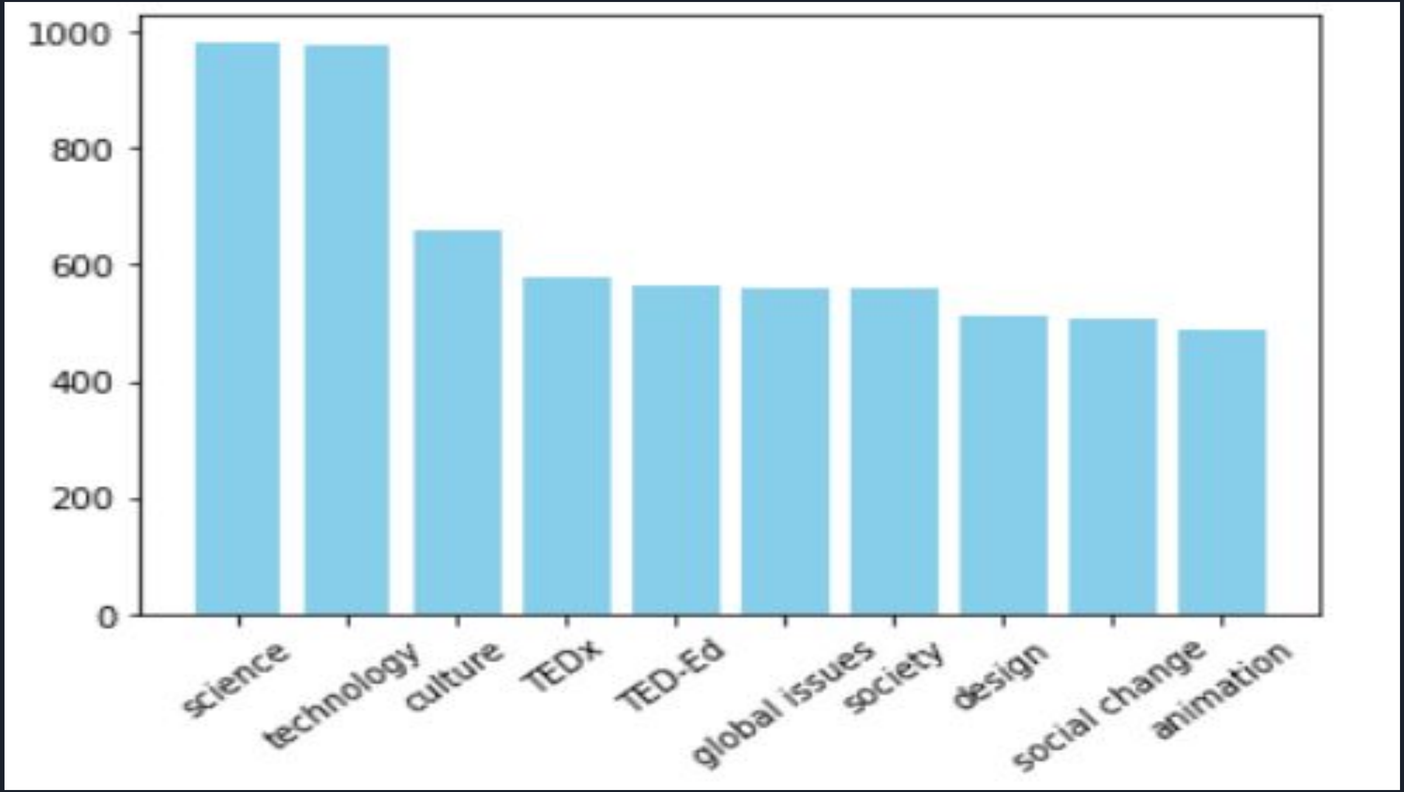
Top 20 Speakers with Highest Average of views.

Views Vs Duration

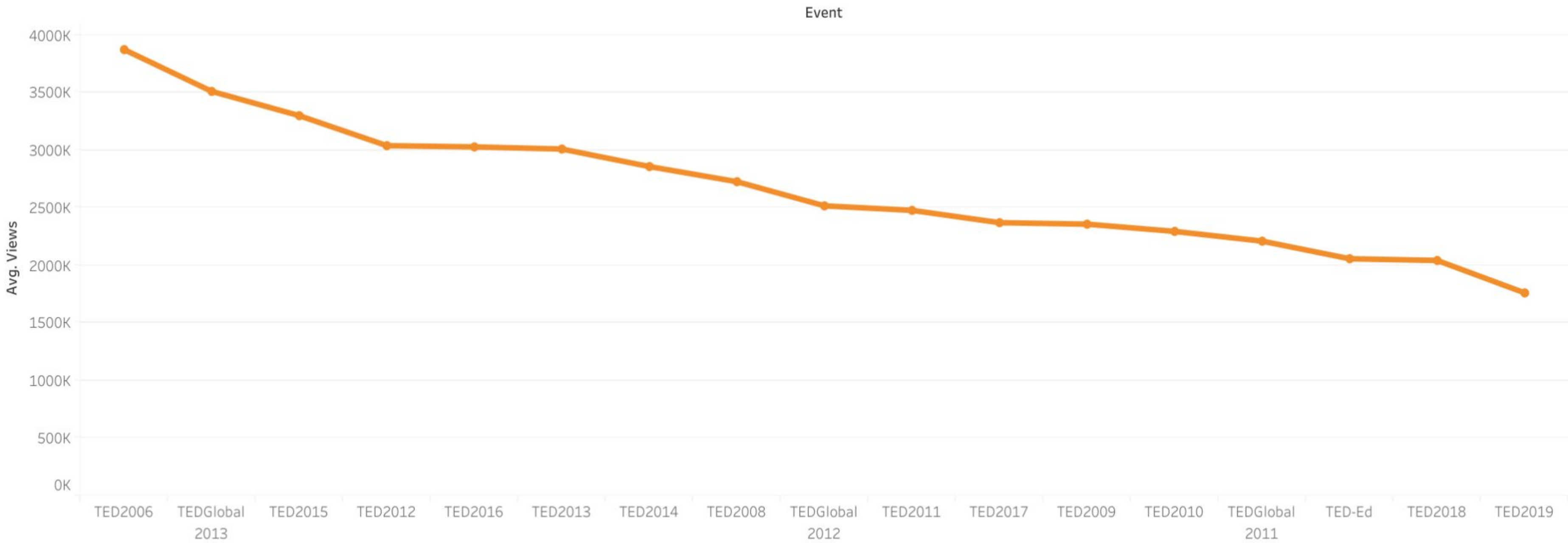



Views Vs Comments
Views
60M
50M
40M
30M
20M
10M
0M

Top 10 topics with highest frequency



Average Views for different Events.



Feature Engineering

Feature engineering refers to manipulation — addition, deletion, combination, mutation — of the data set to improve machine learning model training, leading to better performance and greater accuracy.

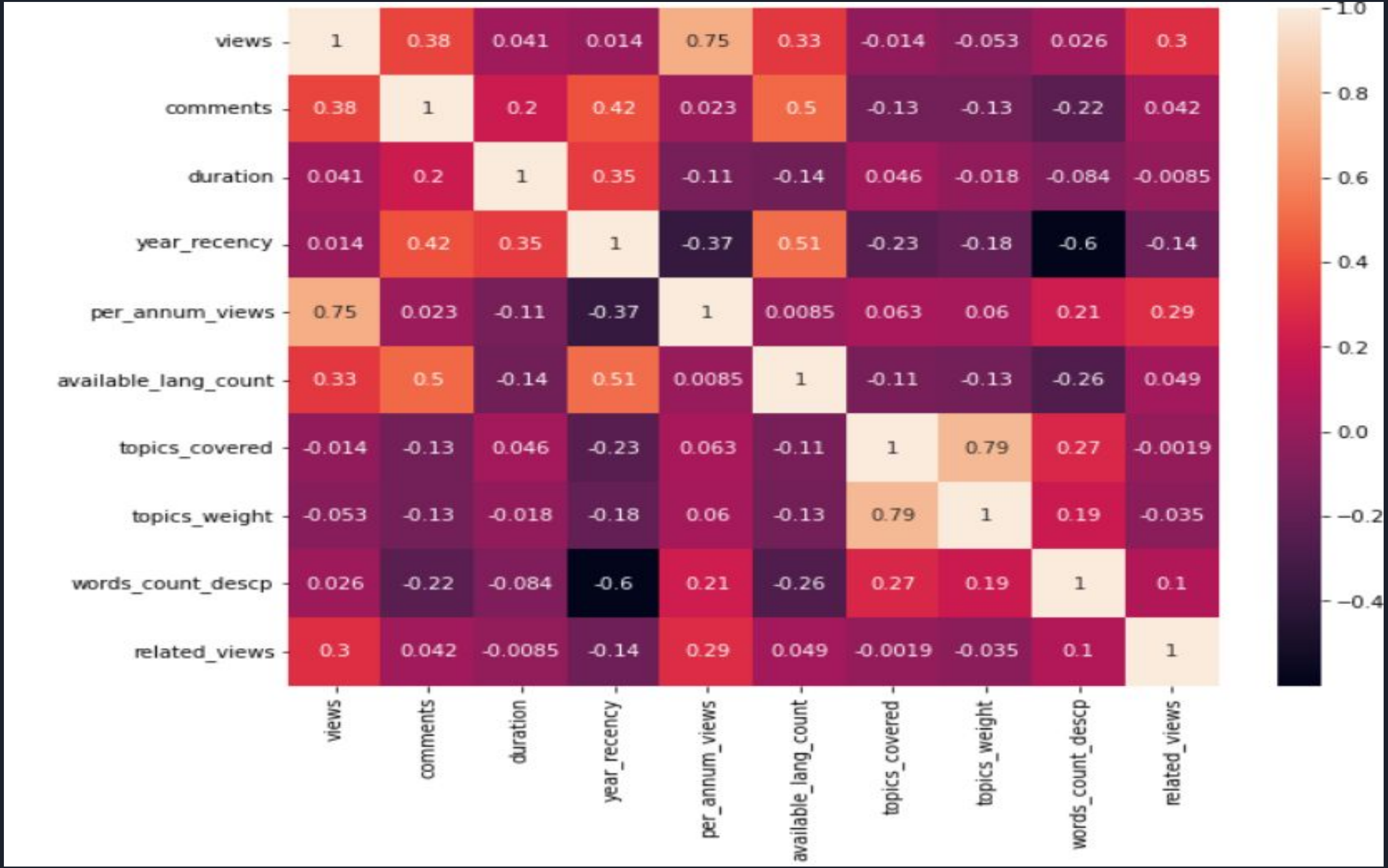
Feature Engineering consists of :

- Feature creation - creating features involves creating new variables.
- Transformation - transformation of features from one representation to another.
- Feature extraction - extracting feature from dataset to identify useful information.
- Benchmark Model - for comparing the performances between different machine learning models.

Some new derived features :

- Available_lang_count
- Related_views
- Topics_weight
- Year_recency
- Per_annum_views

Correlation Matrix



	views	comments	duration	year_recency	per_annum_views	available_lang_count	topics_covered	topics_weight	words_count_descp	related_views
views	1	0.38	0.041	0.014	0.75	0.33	-0.014	-0.053	0.026	0.3
comments	0.38	1	0.2	0.42	0.023	0.5	-0.13	-0.13	-0.22	0.042
duration	0.041	0.2	1	0.35	-0.11	-0.14	0.046	-0.018	-0.084	-0.0085
year_recency	0.014	0.42	0.35	1	-0.37	0.51	-0.23	-0.18	-0.6	-0.14
per_annum_views	0.75	0.023	-0.11	-0.37	1	0.0085	0.063	0.06	0.21	0.29
available_lang_count	0.33	0.5	-0.14	0.51	0.0085	1	-0.11	-0.13	-0.26	0.049
topics_covered	-0.014	-0.13	0.046	-0.23	0.063	-0.11	1	0.79	0.27	-0.0019
topics_weight	-0.053	-0.13	-0.018	-0.18	0.06	-0.13	0.79	1	0.19	-0.035
words_count_descp	0.026	-0.22	-0.084	-0.6	0.21	-0.26	0.27	0.19	1	0.1
related_views	0.3	0.042	-0.0085	-0.14	0.29	0.049	-0.0019	-0.035	0.1	1

Model Development

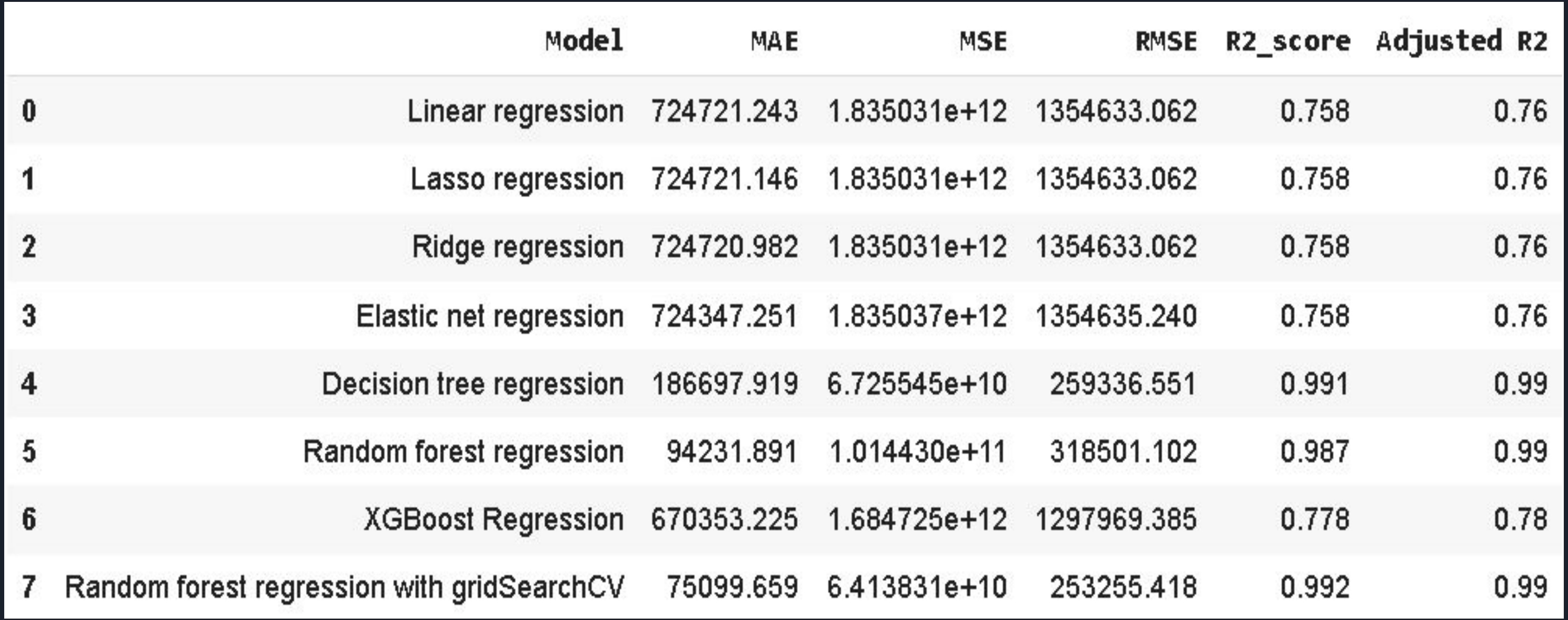
Algorithm ?

A machine learning algorithm is the method by which the AI system conducts its task, generally predicting output values from given input data.
The two main processes of machine learning algorithms are classification and regression.

Algorithm for regression

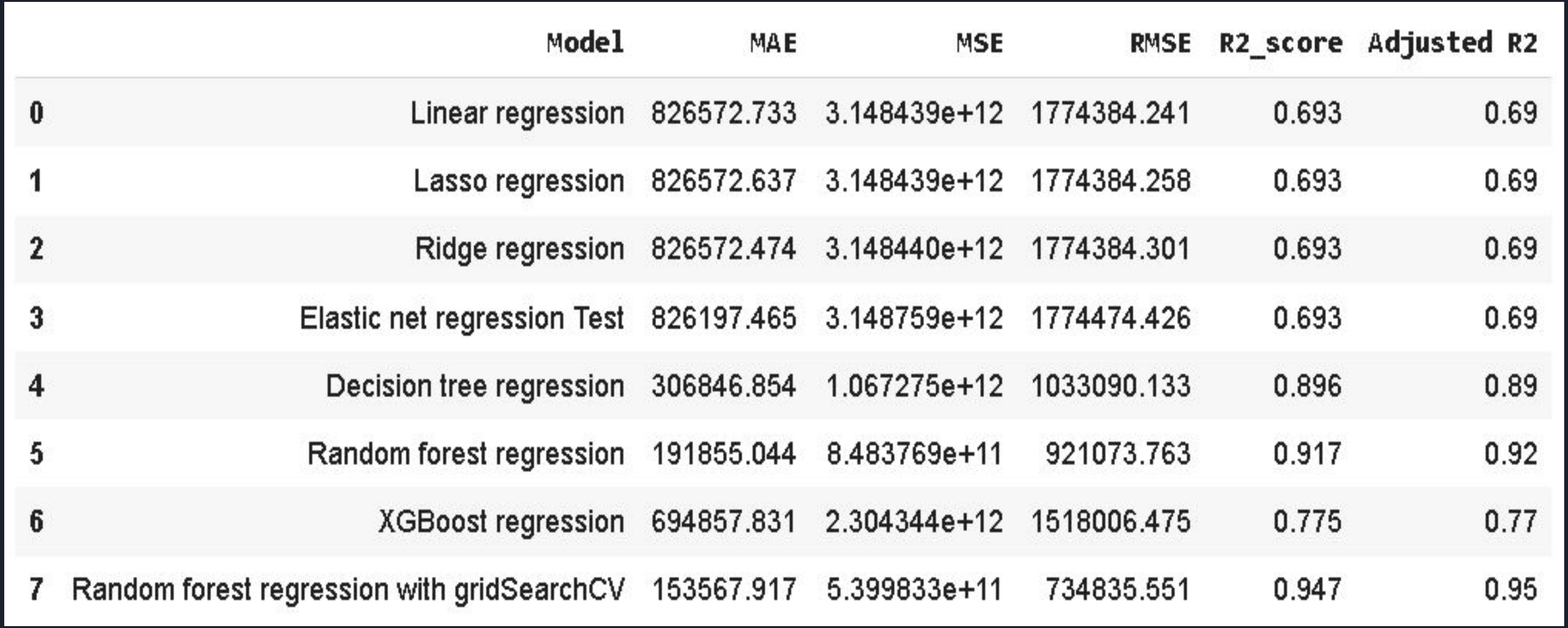
- Linear Regression
- Lasso Regression
- Ridge Regression
- Elastic - Net Regression
- Decision tree
- Random forest
- XGB Regression
- Random forest with Grid search CV

Training score



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	724721.243	1.835031e+12	1354633.062	0.758	0.76
1	Lasso regression	724721.146	1.835031e+12	1354633.062	0.758	0.76
2	Ridge regression	724720.982	1.835031e+12	1354633.062	0.758	0.76
3	Elastic net regression	724347.251	1.835037e+12	1354635.240	0.758	0.76
4	Decision tree regression	186697.919	6.725545e+10	259336.551	0.991	0.99
5	Random forest regression	94231.891	1.014430e+11	318501.102	0.987	0.99
6	XGBoost Regression	670353.225	1.684725e+12	1297969.385	0.778	0.78
7	Random forest regression with gridSearchCV	75099.659	6.413831e+10	253255.418	0.992	0.99

Test score



	Model	MAE	MSE	RMSE	R2_score	Adjusted R2
0	Linear regression	826572.733	3.148439e+12	1774384.241	0.693	0.69
1	Lasso regression	826572.637	3.148439e+12	1774384.258	0.693	0.69
2	Ridge regression	826572.474	3.148440e+12	1774384.301	0.693	0.69
3	Elastic net regression Test	826197.465	3.148759e+12	1774474.426	0.693	0.69
4	Decision tree regression	306846.854	1.067275e+12	1033090.133	0.896	0.89
5	Random forest regression	191855.044	8.483769e+11	921073.763	0.917	0.92
6	XGBoost regression	694857.831	2.304344e+12	1518006.475	0.775	0.77
7	Random forest regression with gridSearchCV	153567.917	5.399833e+11	734835.551	0.947	0.95

Conclusion

We were able to see that the linear algorithms were not performing optimally even with Gradient Boosting optimization, and the tree-based algorithms performed significantly better.

Out of the tree-based algorithms, the Random Forest Regressor was providing an optimal solution towards achieving our Objective. We were able to achieve an R2 score of 0.99 in the train split, and 0.92 in the test split. We also noticed that even in the case of Decision tree, we were able to achieve an R2 score of 0.89 in the test split.

We then implemented Grid Search Cross Validation on the Random Forest Regressor, to further optimize the model, and were able to achieve an R2 score of 0.99 in the train split, and 0.95 in the test split.

Finally, we conclude Random Forest with GridSearchCV to be the best model to achieve our objective.

References

- Python Pandas Documentation
https://pandas.pydata.org/pandas-docs/stable

- Python MatPlotLib Documentation
https://matplotlib.org/stable/index.html

- Python Seaborn Documentation
https://seaborn.pydata.org

- Python SkLearn Documentation
https://scikit-learn.org/stable

- TED website
https://ted.com

Thank You