## Capstone Project

Supervised ML
(Regression)
On
Ted Talks View Prediction

#### Roadmap of the Presentation



#### Introduction

A brief note about the project.

#### EDA and Feature Engineering

Analysing and manipulating data into features for supervised learning.

Conclusion and References

# Introduction What is machine learning?

Machine learning (ML) is a subset of artificial intelligence (Al) 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.



## 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

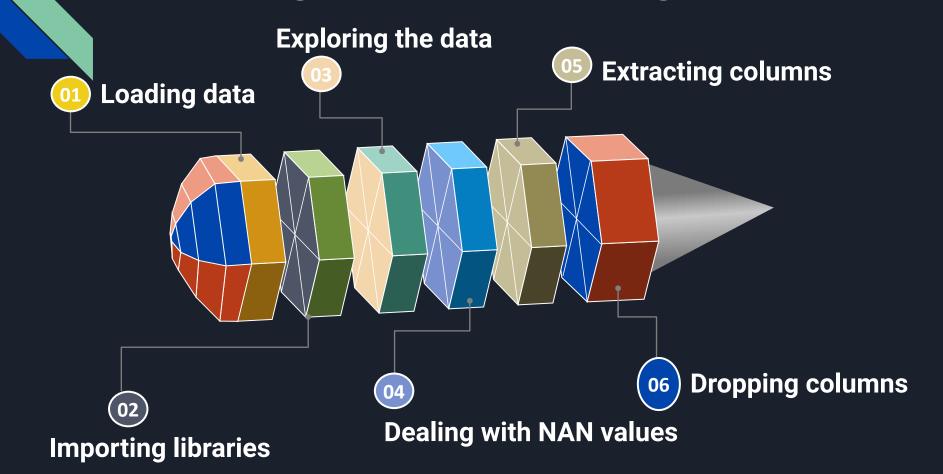
## Data cleaning

Q

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**



## 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 ( )



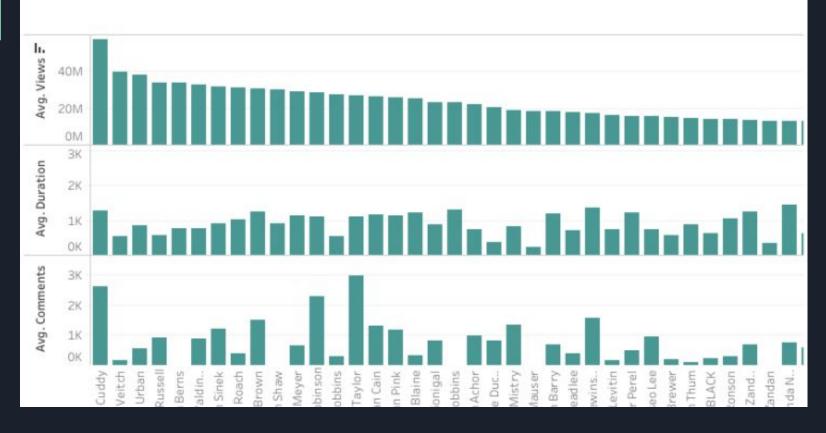
Analysis is done using systematic methods to look for trends, groupings, or other relationships between different types of data.



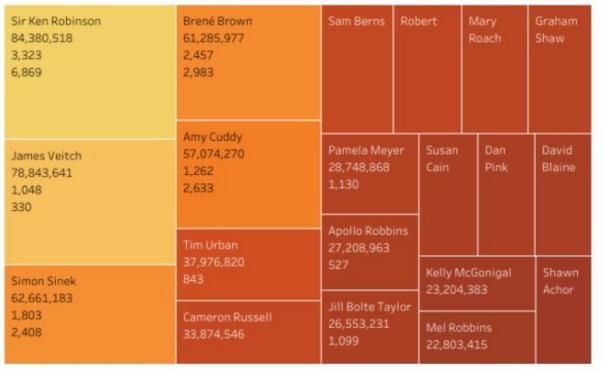
We then use 'Data visualization', to create charts, graphs, or other forms of visualization, which makes information easier to analyse and interpret.



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

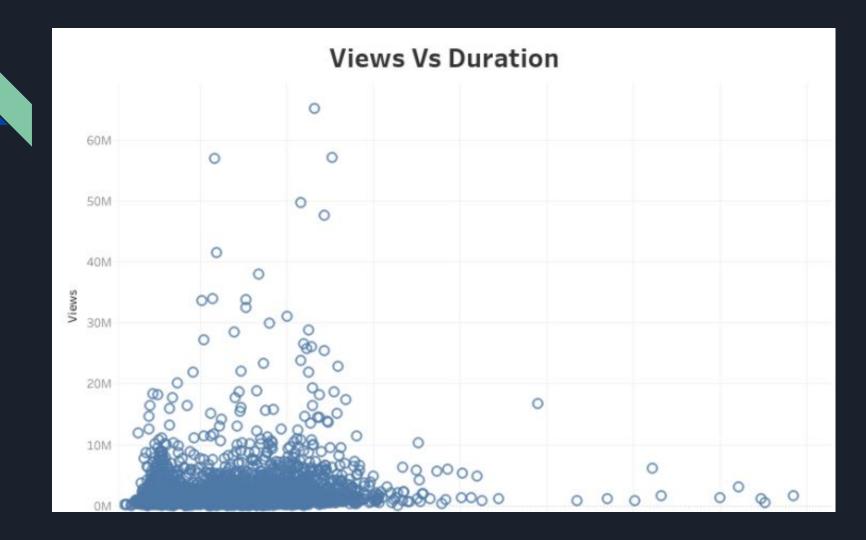


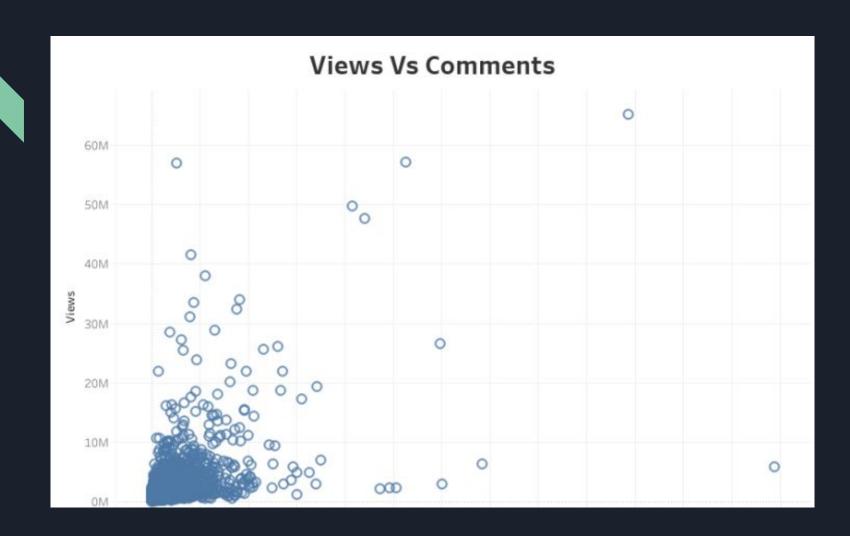
#### Top 20 Speakers with Highest Average of views.



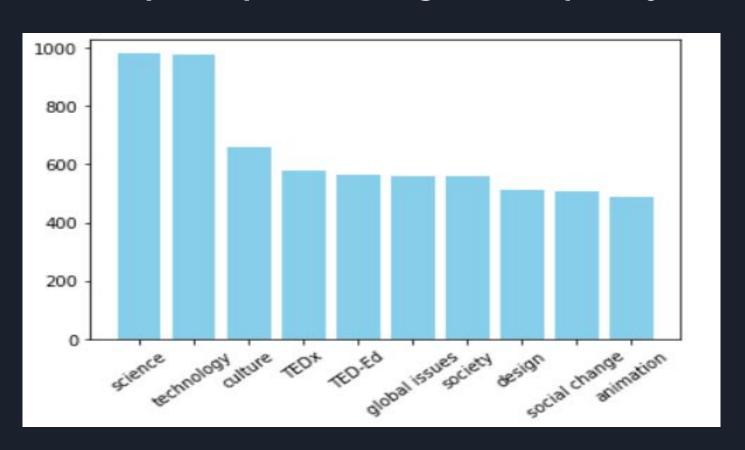
Views

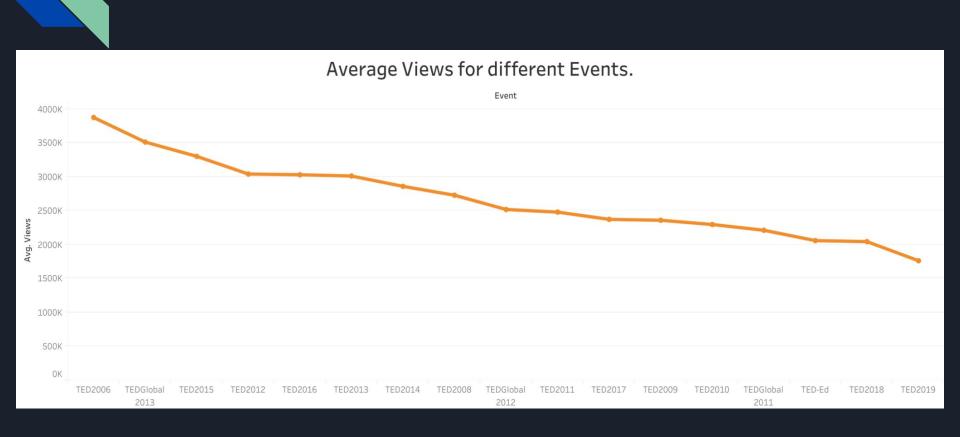
21,939,075 84,380,518





#### Top 10 topics with highest frequency





## 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.



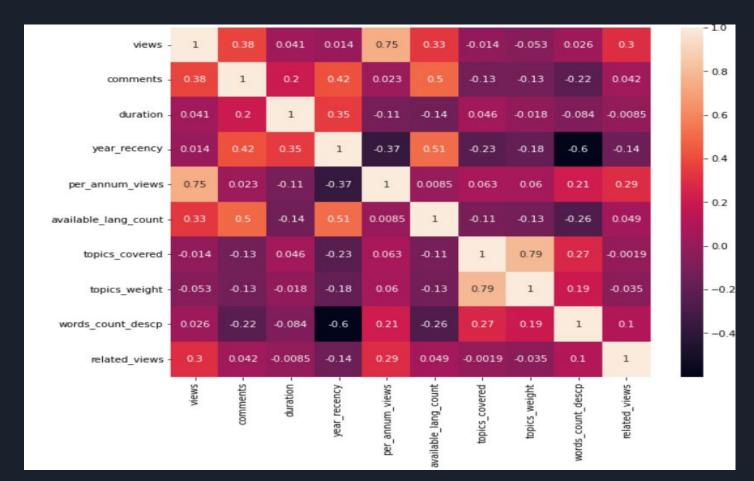
### Feature Engineering consists of:

- 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**



## 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 6

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
   <a href="https://pandas.pydata.org/pandas-docs/stable">https://pandas.pydata.org/pandas-docs/stable</a>
- Python MatPlotLib Documentation <u>https://matplotlib.org/stable/index.html</u>
- Python Seaborn Documentation <a href="https://seaborn.pydata.org">https://seaborn.pydata.org</a>
- Python SkLearn Documentation <u>https://scikit-learn.org/stable</u>
- TED website <a href="https://ted.com">https://ted.com</a>

## Thank You