### *Ted Talk Views Prediction*

*Prashant Gour , Jyotsana Gour,*

*Anuj Gupta*

*Data Science Students*

*AlmaBetter*

**Abstract**

TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many language. Founded in 1984 by Richard Salman as a non-profit organisation that aimed at bringing experts from the field 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 2,000 talks of free consumption by the masses and its speaker list boasts of the likes of Al Gore, Jimmy Wales, Sharukh Khan, and Bill Gates.

***Keywords: Machine learning, Views Prediction, TED Talk, Regression model***

**1.Problem Statement:-**

The main goal of the project is to build a predictive model, that could help us to predict the vies of the vedios uploaded TEDx website.

**Column (Feature) information:**

**1.talk\_id –** Talk id number provided by TED.

**2.title –** Title of the talk.

**3.speaker\_1 –** First speaker in TEDs speaker list.

**4.all\_speakers –** Speakers in the Talk.

**5.occupations –** Occupations of the speakers.

**6.about\_speakers –** Information about all speakers.

**7.recorded\_date –** Date the talk was recorded.

**8.published\_date –** Date the talk was published to TED.com.

**9.event –** Event or medium in which the talk was given.

**10.native\_lang –** Language of the talk.

**11.available\_lang –** All available languages for a talk.

**12.comments –** Comments of the talk.

**13.duration** – Duration in seconds.

**14.topics –** Topics for the talk.

**15.related\_talks –** Related talks.

**16.Url –** URL of the talk.

**17.description –** Description of the talk.

**18.transcript –** Transcript of the talk.

**19.views –** The number of views of each vedio, this will be our target variable.

**2.Introduction:-**

TED talks are one of the most influential vedios on the internet, where experts speak on education, business, science, tech, and creativity. We have data of this TED talk vedios. We have to analyze the data, explore the data, find out the features, apply regression models.

We have to find some features from the data by which TED Talks can increase the views on their vedios.

**3.Project Workflow:-**

* Importing important libraries
* Mount the drive and import the dataset
* Handling null values
* Exploratory Data Analysis (EDA)
* Feature Engineering
* Data cleaning
* Fitting the Regression Models
* Algorithm
* Model Performance
* Conclusion

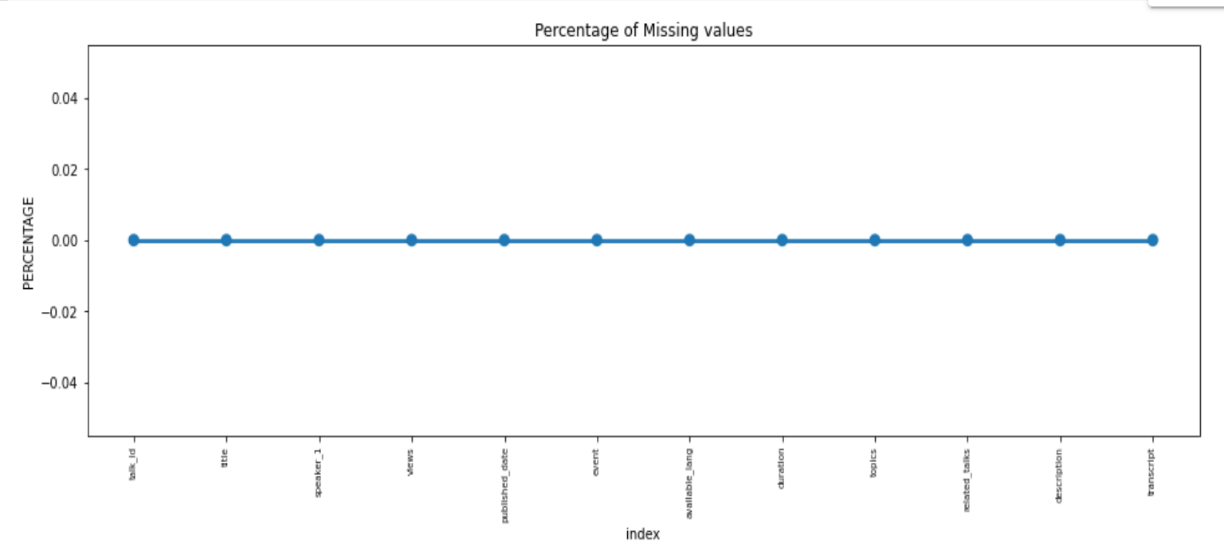
**4.Steps Involved:-**

* **Importing libraries and import the dataset –** We have to import the required libraries and load the data into the python dataframe object.

The data provided is in the form of csv file.

* **Handling Null Values –** Occupations, about\_speakers, comments and recorded date coloumn contain the null values.

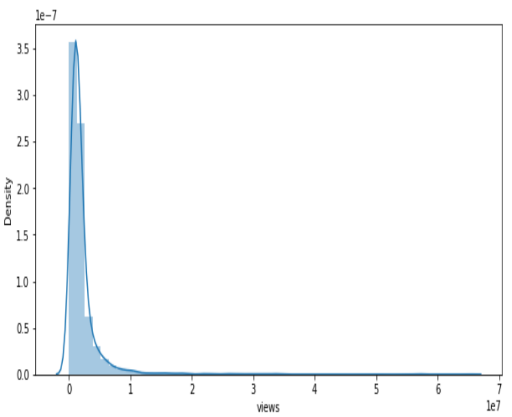
We have to drop some columns as they are of no use. As well as we have to drop occupation, about\_speakers, comments and recorded date column as they contain null values.



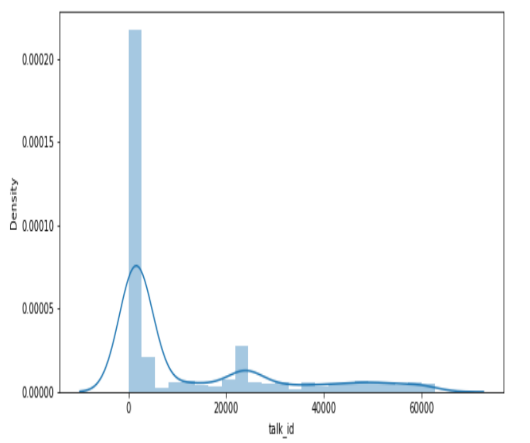
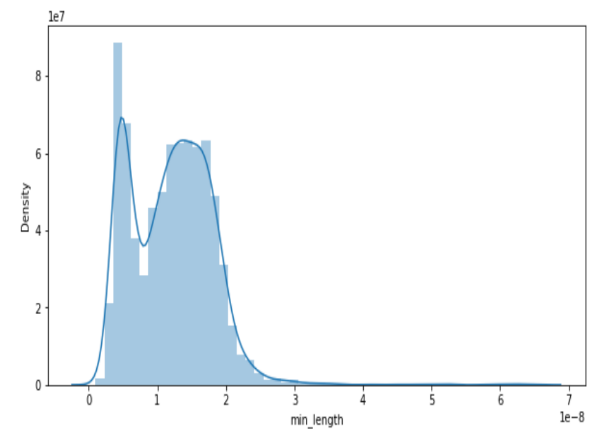
**5. Exploratory Data Analysis:-**

After loading the dataset and dropping the null values. We have to check the size of the dataset. The dataset has 4005 entries with 19 columns.

By performing the dataset.info, dataset.head and dataset.tail operation. We get some inputs about the columns name. We can see views will be our target variable. As we have to perform the analysis, so data type of every column is important to check.



* As we can see, the target variable ‘views’ was a skewed variable.

* All of the data had very skewed and continuous variable distributions.
* URL, all\_speakers, about\_speakers, native\_lang, occupation, recorded\_date, and comments will not be important for analysis, it will be better if we drop it.

**6. Data Cleaning:-**

We had performed outlier treatment to remove the high errors which will be generated by outliers. We have done outlier treatment by replacing the outliers with extreme values.

**7. Fitting the Regression Models:-**

For modelling we tried various regression algorithms like:

* Random Forest Regressor
* Ridge Regression
* Lasso Regression

**8. Algorithm:-**

We have used only non-parametric models for prediction because two of the hypothses such as linearity between output and input variables and errors normally distributed were not met.

* **Random Forest Regressor** – Bootstrapping is the process of randomly sampling subsets of a dataset over a given number of iterations and a given number of variables. These results are then average together to obtain a more powerful result, Bootstrapping is an example of an applied esemble model.

The Bootstrapping random forest algorithm combines esemble learning method with the decision tree framework to create multiple randomly drawn decision tree from the data, averaging the results to output a new result that often leads to strong prediction.

Random Forest has multiple decision trees as base learning models. We randomly perform now sampling and feature sampling datasets for every model. This part is called Bootstrap.

The result of this regressor given below:

MSE: 53025015132.822495

MAE: 94912.11522784022

RMSE: 23027161165202822

R2 : 0.9565726968623682

Adjusted R2: 0.9551138985657552

* **Ridge Regression –** Ridge regression is a way to create a parsimonious model when the number of predictor variables in a set exceeds the number of observation, or when a dataset has multicollinearity (corellations between predictor variables).

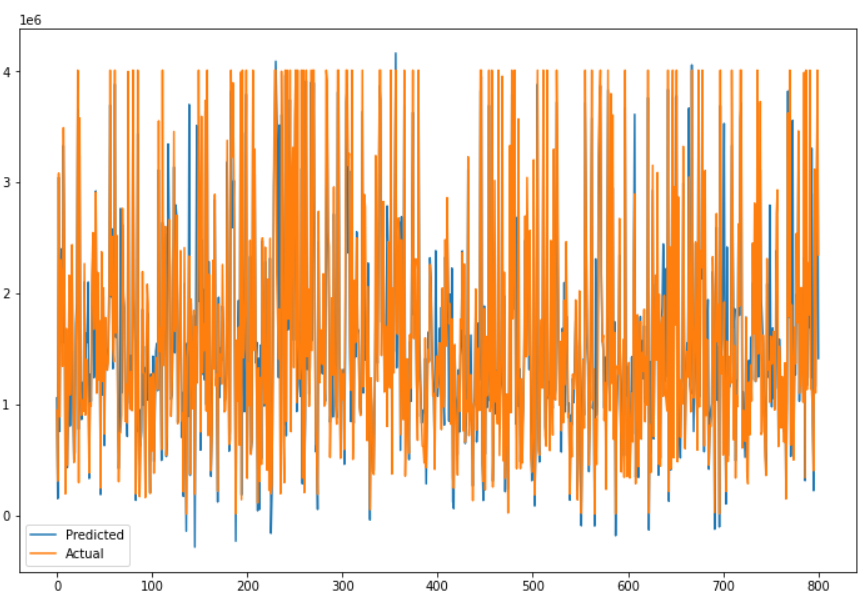
MSE : 197012475172.63742

MAE : 256065.34238410502

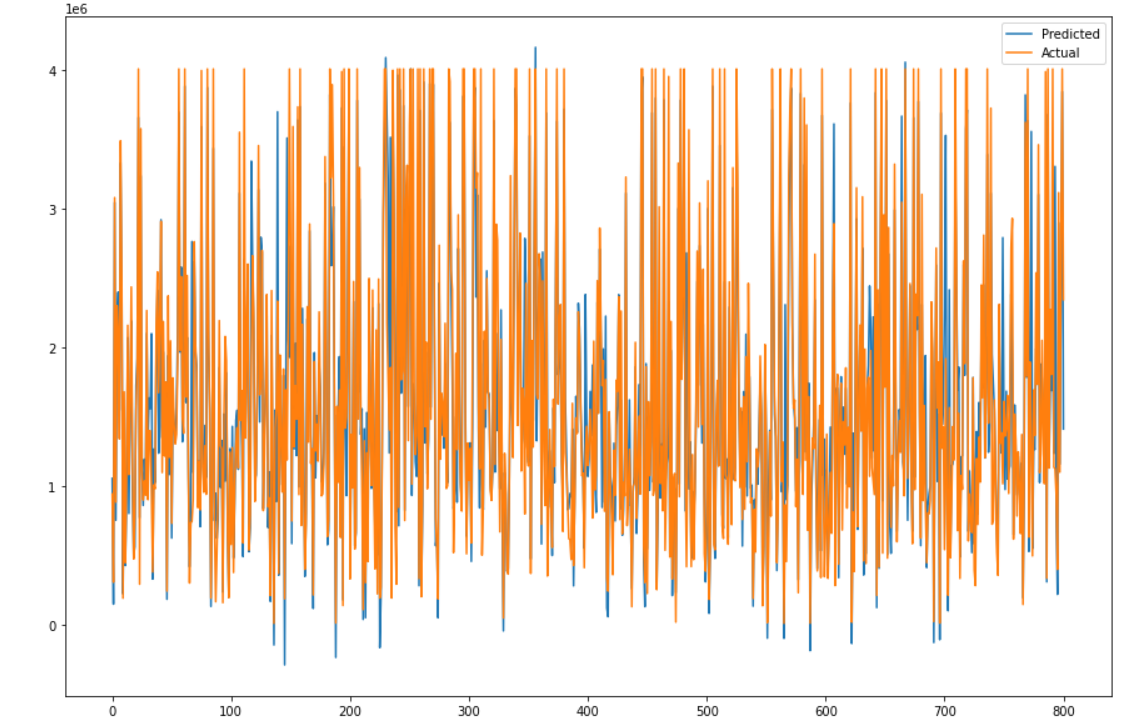
RMSE : 443860.84366723083

R2 : 0.8348381490304355

Adjusted R2 : 0.8292900765172461



* **Lasso Regression -** Lasso Regression is a type of linear regression that usage shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters).

****

**9. Model Performance:-**

Model can be evaluated by various matrices such as:

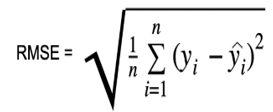
* **Root Mean Square Error –** RMSE is computed by taking the square root of MSE. RMSE is also called the Root Mean Square.

Deviation. It measures the average magnitude of the errors and is concerned with the deviations from the actual value. RMSE value with zero indicates that the model has a perfect fit. The lower the RMSE, the better the model and its predictions. A higher RMSE indicates that there is a large deviation from the residual to the ground truth.

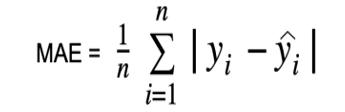
One major drawback of RMSE is it sensitivity to outliers and the outliers have to be removed for it to function properly.

RMSE increases with an increase in the size of the test sample. This is an issue when we calculate the results on different test samples.

Like RMSE,MSE is dependent on the scale of the data. It increases in magnitude if the scale of the error increases.



* **Mean Absolute Error –** Mean absolute Error , also known as L1 loss is one of the simplest loss functions and an easy-to-understand evaluation metric it is calculated by taking the absolute difference between the predicted values and the actual values and averaging at across the dataset. Mathmatically speaking, it is the arithmetic average of absolute errors. MAE measures only the magnitude of the errors and dosen’t concern itself with their direction. The lower the MAE, the higher the accuracy of a model. Mathematically. MAE can be expressed as follows:

****

**10.Conclusion:-**

* TED Talks mainly provide inspirational and educational content that helps individuals look at things from a diffrent prespective. There is a wide variety of TED Talks, ranging from technology to comedy. From the given dataset, We can infer that the popularity of a TED Talk depends on the number of languages available. It also marked with a good sense of humor, presentation skills and mostimportantly, choice of topic etc.
* According to analysis We have highest number of views on ted talk in the year **2019** followed by 2018.Also we have highest number of views on ted talk in the **April** followed by **March**.
* **TED-ED** is the most frequent event category with 556 entries followed by **TED2017** and **TED2018**.
* **Alex Gendler** is the most popular speaker followed by sir **Ken Robinson** according to total views on their talks.
* **Sir Ken Robinson's** talk on "Do School Kill Creativity?" is the most popular TED Talk of all time with more than 6.5 million view.
* As we can see that mejority of titles contain words: **Life , World , Make , New , Human , Brain and Future**
* Applying train test split and we found linear regression 83%.As well as 83% with lasso regression and ridge regression.
* Random Forest Regressor with 95%.
* The objective of this work was to understand a little about the TED talks ratings. They are intersting metrics to evaluate a TED Talk and this work explored the many results,how people evaluate them and what may influence these results.