**SUMMARY OF TED TALK VIEW PREDICTION**

**Logo

Description automatically generatedA picture containing night sky

Description automatically generatedLINEAR REGRESSION**

* **INTRODUCTION**

TED is devoted to spreading powerful ideas on just about any topic. Founded in 1984 by Richard Salman as a non-profit 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.

TED talks have been given for many years with the platform of "Ideas Worth Spreading". In the digital world we live in today, TED is a great platform to get your idea out there.

* **PROBLEM**

But how do you know if your idea will be heard or appreciated.

* **OBJECTIVE**

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

* **TOOLS REQUIRED**

To do analysis on Ted Talk data we used language python. In python we have various libraries which we must import like NumPy, Pandas, Matplotlib and Seaborn**,** Scikit-Learn.

* **ACTIVITY**

As the first step, perform data wrangling over the raw data. With this came to know about some basic features of the data, identifying the dependent and independent variables and the columns with Nan values. However, most of the features are categorical but a lot of insights can be drawn with these data. Some of the important features which can help in predictions are title, views, duration, topics. Looking at the description of numeric features we can conclude that, Comments have some null values. Duration values are more spread as it has high value of standard deviation from the mean. We can also see the minimum and maximum values of each column and the percentile.

As the second step, performing data visualization. The distribution plot of views and comments are rightly skewed. However, the correlation between the numeric features is very less. Sir Ken Robinson and Amu Cuddy have got the most views among others and they are author and psychologist respectively. For better inspection adding some more column to the data frame like daily views, event wise average views, topic wise average views, number of languages, number of topics, then removing irrelevant columns from the dataset.

As the last step, identifying dependent and independent variables, one hot encoding using get dummies function. The values of features are standardised using minmaxscaler, then cleaned data is now put into Model. Also, Ridge and Lasso to avoid overfitting and underfitting.

* **CONCLUSION**

In all these models our errors have been in the range of 2,00,000 which is around 10% of the average views. We have been able to correctly predict views 90% of the time. After hyper parameter tuning, we have prevented overfitting and decreased errors by regularizing and reducing learning rate. Given that only have 10% errors, our models have performed very well on unseen data due to various factors like feature selection, correct model selection, etc.

In all the features speaker wise average views is most important this implies that speakers are directly impacting the views.