

Regression Project on

Ted talk Views Prediction

Project performed by

Sunil Kumar



**Table of Contents**

* Summary...........................................................................1)
* Problem statement.............................................................2)
* Introduction.........................................................................3)
* Technical Work...................................................................4)

1. Library used.
2. Function and Method used.

* Step involved in exploratory data analysis..........................5)
* Observations.......................................................................6)

**Summary:**

TED is devoted to spreading powerful ideas on just about any topic. These datasets contain over 4,000 TED talks including transcripts in many languages. Founded in 1984 by Richard Salman as a nonprofit 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.

**Problem Statement**

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

**Technical Work**

* ***Library used***
* ***Functions and methods used.***

**Library used**

* **Pandas** This library is used in data analysis and manipulation and importing files.
* **Matplotlib** Used infor data visualization, graphs & plotting.
* **Seaborn** Used inmaking data visualisation in a more colourful and meaningful way.
* **Sklearn:** useful and robust library for machine learning in Python.
* **Catboost:** It provides a gradient boosting framework that attempts to solve for Categorical features using a permutation driven alternative compared to the classical algorithm.
* **os** Used in provides functions for creating and removing a directory folder.

**Common functions and methods used**

**pd.read\_csv():**used in to import data in CSV format. This function has a number of arguments, but the only essential argument is file, which specifies the location and filename.

**summary():**function return a summarized representation of the Index.

**pwd():** Finds present working directory.

**os.chdir():** Changes your directory to your data directory.

**head():** Used in checking the first rows of the dataset.

**tail():** Used in fetching the last rows of the dataset.

**shape():** Used in getting the total no. Of columns and rows, a dataset contains.

**Type():** Used in knowing the data structure of the dataset.

**df.dtypes:** Used in understanding the data types of the dataset.

**IsNull():** used in checking if any null values are available.

**Sum():** used in giving sum of the values.

**value\_counts() :** Used in providing counts of particular values present in data.

**rename():** used to rename a file or directory or data columns.

**corr():** aggregate function returns a coefficient of correlation between two numbers.

**unique():** function is used to find the unique elements of an array.

**Steps involved:**

* *Setting directory/path.*
* *Loading dataset.*
* *Data cleaning.*
* *Data wrangling.*
* *Visualizations.*
* *Data pre-processing- feature selection..*
* *Modelling.*
* *Performance improvement.*
* *Model selection and evaluation.*
* *Deployment.*

**Setting directory/path**

Before we proceed with python to understand the data, it is important to let python understand first that where our data available is so that, we can perform the operation on data using python. The setting directory is similar to that and helps in understanding the presence of the data.

**os.chdir(your path)**

**Loading dataset**

Once the directory is set now we need to load the dataset which is commonly done using:

**Pd.read\_csv(‘file\_name’)**

**Data cleaning**

Now once we have our dataset loaded it is important that we look for data in good format and it is not containing anything which is not required for our analysis or that contains values that make no sense.

Data that has null values in this step of cleaning the data and treating the null values or empty columns is called data cleaning.

Essentially data cleaning has below operations to perform:

1. Dropping and selecting columns in a DataFrame.
2. Changing the Index of a DataFrame.
3. Tidying up Fields in the Data.
4. Cleaning the Entire Dataset in terms of null values.
5. Renaming columns and skipping rows.
6. Looking for duplicates.
7. Searching for outliers and treating accordingly.

**Data wrangling**

Process of cleaning and unifying messy and complex data sets for easy access and analysis. Converting the format of data to a suitable format.

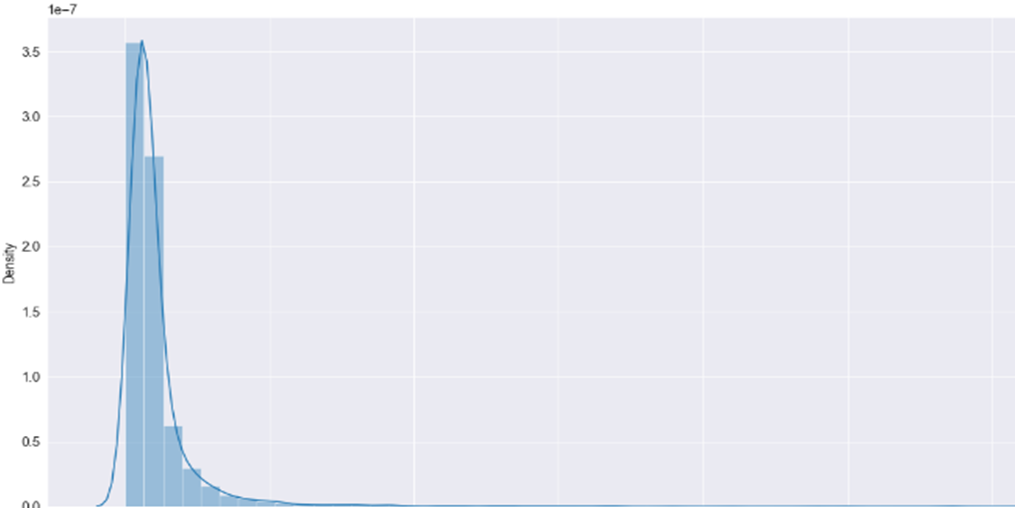
We can rename our variables using the .rename() function put your columns inside a list of a dictionary in case columns are more and which requires renaming to a more sensible form.

This step involves finding manually the data and making a suitable alteration for that it make more sense.

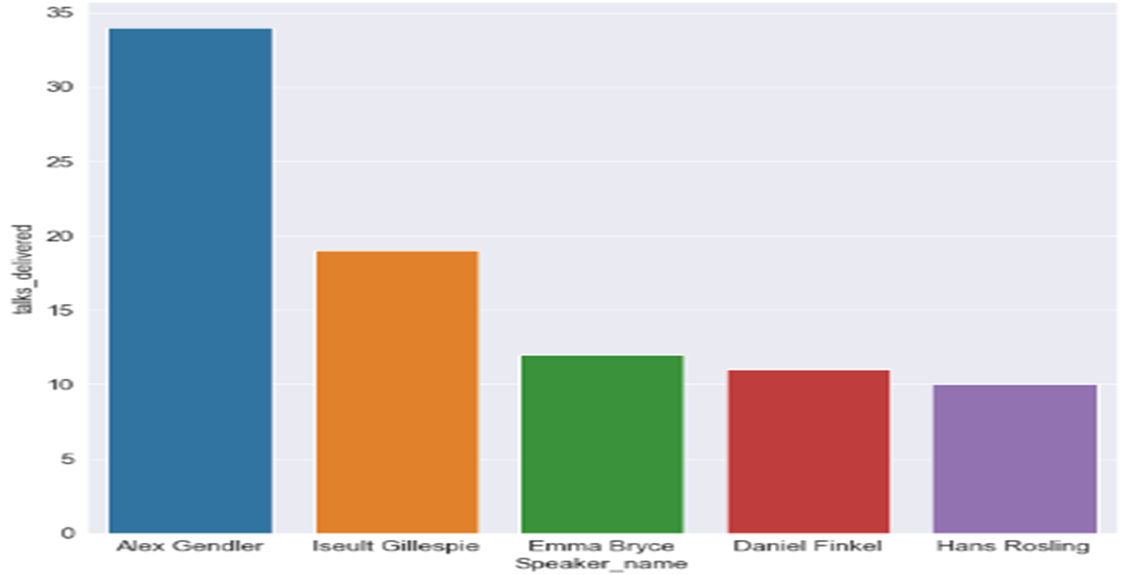
**Data Visualization**

This step of visualization involves checking the data in pictorial form. We can see outliers explore null values and also we can understand the variables which are correlated to each other’s and how data is distributed.

Apart from this, visualization helps in easy understanding of the data and representation of most of the data information in right and quick fashion.



Views are right skewed. Views have positive skewed distribution of data. Mean is greater than mode. Mean is greater than median as well.



More talks session delivered shows more views and popularity

**Modelling**

Before we proceed with model building, it is important that we have the right amount of data and we understand the features/variables very well. Almost every model takes the numerical data for further processing in the model part so it is important to convert the entire data into numerical columns so that we can process it further.

It is equally important to classify your columns into dependent and independent variables. Once this task is performed we are ready for the next step in modelling which can be splitting the training and testing data.

Next in this process, we would need some of the machine learning libraries and suitable models for the prediction since in this case we are predicting the output of views.

We can start with a simple model like Linear regression and see the scores after fitting the data. Look for the scope of improvement if required we can use regularization for overfitting problems as it decreases the magnitude of coefficients and gives better results. Hyperparameter tuning is the step that chooses the optimal hyperparameter for learning of model So that we can have good prediction scores. Finally Model has to undergo a thorough evaluation which checks whether your model is doing good in terms of accuracy and errors in prediction.

**Model Used**

Linear Regression.

Catboost.

XGBoost.

Random forest.

**Observations**

Starting with loading the data so far I have done EDA, feature engineering, data cleaning, target encoding and one-hot encoding of categorical columns, feature selection and then model building and hyperparameter tuning.

* The results for models seems fair enough and I have been able to correctly predict views 80% of the time. After hyperparameter tuning, I have prevented overfitting and decreased errors by reducing the learning rate. ###Given that only have 20% errors, my models have performed very well on unseen data due to various factors like feature selection, correct model selection and hyperparameter tunning etc.
* I have used four models for the views prediction started including Linear regression, Random forest, XtremeGradientBoost, Random forest and Cat boost.
* Out of my 4 models including the hyperparameter also I have seen that the Catboost model which is basically meant for categorical data( categorical boost) have shown a good Training accuracy of 99.99% and testing accuracy of 81% keeping in mind the score of mean absolute errors and RMSE since I have to use the MAE for the measurement of the good model and not the RMSE because RMSE can heavily be impacted by outlier but MAE is not impacted by outliers also inn term of MAE Random forest and XGBoost has good scores.
* I have explored the data and looked for the trend in bi-variate analysis.
* I have checked that views are correlated with the speakers as shown top 5 speakers with daily views exceeding 100000.
* More talks session delivered is showing that more views and popularity also the duration is not much influenced when there are poplar speakers.

**Challenges.**

* I have faced many challenges during this project starting with looking for converting the data into numerical and standardization and selecting the right amount of features and encoding including target encoding and one\_hot\_encoding.
* I have made a few observations that keeping the right amount of featuresand right feature selection is very important and finally are responsible for your models' betterment and good accuracy.