TED Talk Views Prediction

Sanchita Paul,
Data science trainee,
AlmaBetter, Bangalore

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 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.

Keywords: machine learning, views prediction, TED Talk, Regression model

1.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.

Dataset Information

- Number of instances: 4,005
- Number of attributes: 19

Features information:

The dataset contains features like:

- **talk_id**: Talk identification number provided by TED
- **title**: Title of the talk
- **speaker_1**: First speaker in TED's speaker list
- **all_speakers**: Speakers in the talk
- **occupations**: Occupations of the speakers
- **about_speakers**: Blurb about each speaker
- **recorded_date**: Date the talk was recorded
- **published_date**: Date the talk was published to TED.com
- **event**: Event or medium in which the talk was given
- **native_lang**: Language the talk was given in
- **available_lang**: All available languages (lang_code) for a talk
- **comments**: Count of comments
- **duration**: Duration in seconds
- **topics**: Related tags or topics for the talk
- **related_talks**: Related talks (key='talk_id',value='title')
- **url**: URL of the talk
- **description**: Description of the talk
- **transcript**: Full transcript of the talk

Target Variable :

- **Views**: The number of views for each talk

2. Introduction

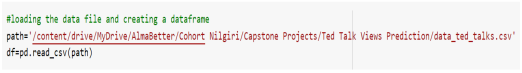
Ted Talks are one of the institutions that are constantly using Machine Learning algorithms to optimize the number of views the videos receive. They do this by understanding the dependency of views with other relevant features to increase the viewers satisfaction by recommending videos according to the average views that have been affected by features like topics,comments, etc.

In order to help Ted Talks to increase their views we are helping them find relevant features and their dependencies on the views.

3. Steps involved:

- **Data Collection**

  To proceed with the problem dealing first we will load our dataset that is given to us in .csv file into a dataframe.Mount the drive and load the csv file into a dataframe.

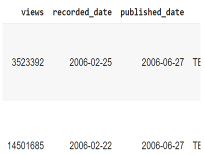


```
#loading the data file and creating a dataframe
path='/content/drive/MyDrive/AlmaBetter/Cohort Nilgiri/Capstone Projects/Ted Talk Views Prediction/data_ted_talks.csv'
df=pd.read_csv(path)
```

- **Exploratory Data Analysis**

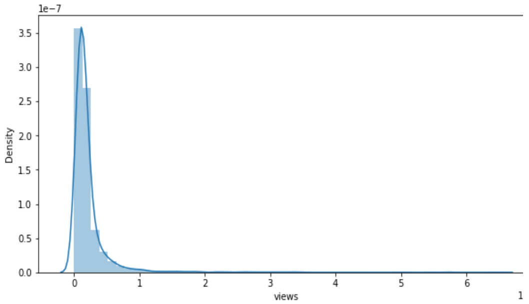
  After loading the dataset we looked for duplicate values in the 'talk_id' column. There were none. So We performed EDA by comparing our target variable that is Views with other independent variables. This process helped us figuring out various aspects and relationships among the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

1. **Numerical Variables:**
   - Talk_id
   - Views
   - Comments
   - duration
2. **Textual Variables:**
   - Title
   - Speaker_1
   - Recorded_date
   - Published_date
   - Event
   - Native_lang
   - Url
   - Description
3. **Dictionaries:**
   - Speakers
   - Occupations
   - About_speakers
   - Related_talks
4. **List:**
   - topics



views	recorded_date	published_date	
3523392	2006-02-25	2006-06-27	TE
14501685	2006-02-22	2006-06-27	TE

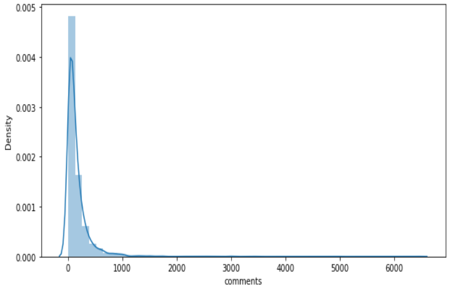
Out of all the continuous variables, 'views' is the target variable.



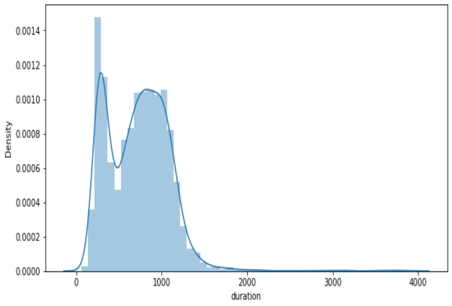
The target variable 'views' was a skewed variable.

The other continuous variables have distributions as:

1. Comments



2. Duration



All of the data had very skewed continuous variable distributions.

- **Null values Treatment**

  Our dataset contains around 400 null values which might tend to disturb our mean absolute score hence we have performed KNN nan value imputer for numerical features and replaced categorical features nan values with the value 'Other'. We chose to impute nan values and not drop them due to the size of the data set

- **Encoding of categorical columns**

  We used Target Encoding for replacing the values of categorical variables with the mean of the views. This was done to not increase the dimensions to the data set while also keeping the relationship of variables with views into consideration.

- **Feature Selection**

  For Feature Selection we have done the following: we have introduced new numerical features from the categorical features,combined features and also we have used f_regression in which we have taken the features with the maximum f-scores.

- **Outlier Treatment**

  We have done outlier treatment on variables like duration and occupation. This was done by replacing outliers with the extreme values at the first and third quartiles. We have done outlier treatment to prevent high errors that were influenced by outliers.

- **Fitting different models**

  For modelling we tried various regression algorithms like:

1. **Decision Tree Regressor**
2. **Random Forest Regressor**

- **Tuning the hyperparameters for better accuracy**

  Tuning the hyperparameters of respective algorithms is necessary for less error values,regularization and to avoid overfitting in case of tree based models.

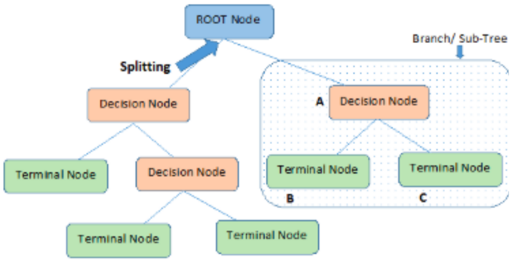
4.1. Algorithms:

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

1. Decision Tree Regression:

Decision tree is a type of **supervised learning algorithm** that is mostly used in classification problems. It works for both categorical and continuous input and output variables.

Let's look at the basic terminology used to study decision trees:



Note:- A is parent node of B and C.

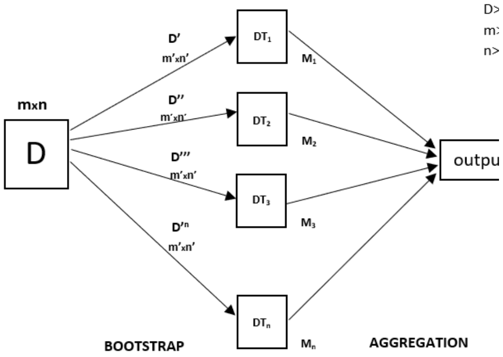
The decision of making strategic splits heavily affects a tree's accuracy. The decision criteria is different for classification and regression trees.

Decision trees use multiple algorithms to decide to split a node in two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. In other words, we can say that purity of the node increases with respect to the target variable. Decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

SMAPE train: 21%
SMAPE test: 20%

2. Random Forest Regressor:

Every decision tree has high variance, but when we combine all of them together in parallel then the resultant variance is low as each decision tree gets perfectly trained on that particular sample data and hence the output doesn't depend on one decision tree but multiple decision trees. In the case of a classification problem, the final output is taken by using the majority voting classifier. In the case of a regression problem, the final output is the mean of all the outputs.



A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as **bagging**. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Random Forest has multiple decision trees as base learning models. We randomly perform row sampling and feature sampling from the dataset forming sample datasets for every model. This part is called Bootstrap.

SMAPE train: 13.5%
SMAPE test: 13%

4.2. Model performance:

Model can be evaluated by various metrics such as:

1. **SMAPE:**
   Symmetric mean absolute percentage error (SMAPE or sMAPE) is an accuracy measure based on percentage (or relative) errors. It is usually defined as follows:

$$\text{SMAPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|F_t - A_t|}{(|A_t| + |F_t|)/2}$$

where A_t is the actual value and F_t is the forecast value.

where A_t is the actual value and F_t is the forecast value.

4.3. Hyper parameter tuning:

Hyperparameters are sets of information that are used to control the way of learning an algorithm. Their definitions impact parameters of the models, seen as a way of learning, change from the new hyperparameters. This set of values affects performance, stability and interpretation of a model. Each algorithm requires a specific hyperparameters grid that can be adjusted according to the business problem. Hyperparameters alter the way a model learns to trigger this training algorithm after parameters to generate outputs.

We used Randomized Search CV for hyperparameter tuning. This also results in cross validation and in our case we divided the dataset into different folds.

1. **Randomized Search CV-** In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions

of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist's control

5. Conclusion:

That's it! We reached the end of our exercise.
Starting with loading the data so far we have done EDA , null values treatment, encoding of categorical columns, feature selection and then model building.
In all of these models our errors have been in the range of 2,00,000 which is around 10% of the average views. We have been After hyper parameter tuning, we have prevented overfitting and decreased errors by regularizing and reducing learning rate. Given that only 10% is errors, our models have performed very well on unseen data due to various factors like feature selection,correct model selection,etc.

Future work:

1. We can do a dynamic regression time series modelling due to the availability of the time features.
2. We can improve the views on the less popular topics by inviting more popular speakers.
3. We can use topic modelling to tackle views in each topic separately.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Analytics Vidhya