Abstract:

Tedtalks or TED Conferences is the media organization where the influential and successful speakers were called and delivers speeches. Tedtalks posts these video's on their website.

Our experiments can help in understanding how many views can a new video get based on features that we are provided with.

Keywords: Speakers, Eda, Views, Dates, Models, Hyper parameter tuning, Linear Regressor, Random Forest Regressor, XGB Regressor.

1. Problem Statement

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.

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

Views: Count of Views

2. Introduction

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

3. Types of Views:

The views are based on the properties like who is the speaker, when it is published, what events make more views and which language audiences are more watching.

- ### Speaker Views:
  The Views based on the speaker. These views vary from one speaker to another based on the topic and related talks.

- ### Published Date Views:
  These views are based on when the video is published. As per the data analysis certain months having the release date of videos getting more views.

- ### Event Views:
  These views are separated from event to event. Suppose TedTalks event gets more views compared to other events.

- ### Language Views:
  These views are based on the language. There is no rule that the published language has to get more views. The video may be liked and popular on available languages and may be that made the video popular.

4. Steps Involved:

- ### Explorative Data ANALYSIS:
  After loading the dataset, we performed this method 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.

- ### Feature Engineering:
  In Feature Engineering we will convert the Categories inside the column as values into model fit data using eval and other define functions.

- ### Data Cleaning:
  In Data Cleaning we dealt with null values using KNN imputer, Outliners using IQR Method, Transformation of Skewed Data Using Log Transformations.

- ### Feature Selection:

In Feature selection we had selected some important features to fit into model using p values by f regression

- ## Fitting different models:

For modelling we tried various Regression algorithms like:

1. Linear Regression
2. Random Forest Regression
3. XGB Regression

- ## Hyper Parameter Tuning:

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of tree-based models like Random Forest Classifier and XGBoost classifier.

5. Algorithms:

- ## Linear Regression:

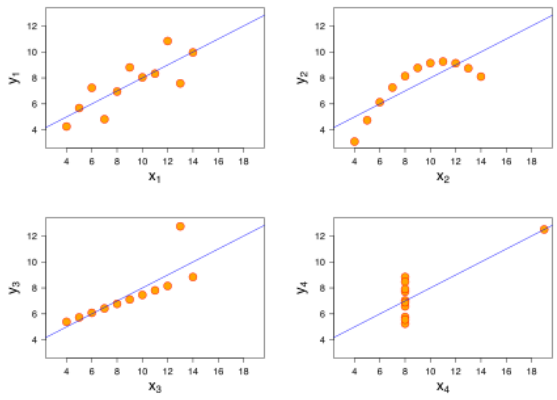
Linear regression is an important tool in analytics. The technique uses statistical calculations to plot a trend line in a set of data points. The trend line could be anything from the number of people diagnosed with skin cancer to the financial performance of a company. Linear regression shows a relationship between an independent variable and a dependent variable being studied.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

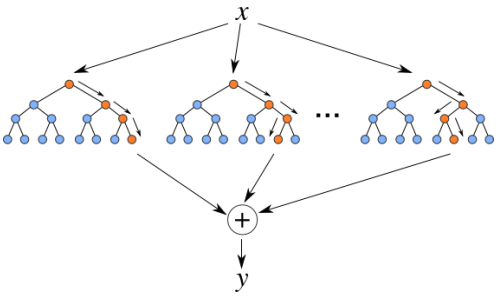
$$y = B0 + B1*x$$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g., B0 and B1 in the above example).

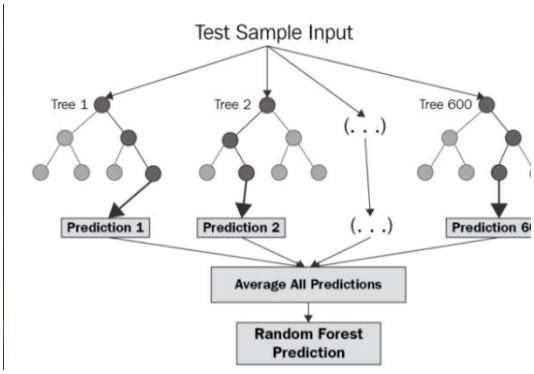
There are a number of ways to calculate linear regression. One of the most common is the ordinary least-squares method, which estimates unknown variables in the data, which visually turns into the sum of the vertical distances between the data points and the trend line.



- ## Random Forest Regressor:



Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



The diagram above shows the structure of a Random Forest. You can notice that the trees run in parallel with no interaction amongst them. A Random Forest operates by constructing several decision trees during training time and outputting the mean of the classes as the prediction of all the trees. To get a better understanding of the Random Forest algorithm, let's walk through the steps:

1. Pick at random k data points from the training set.
2. Build a decision tree associated to these k data points.
3. Choose the number N of trees you want to build and repeat steps 1 and 2.
4. For a new data point, make each one of your N-tree trees predict the value of y for the data point in question and assign the new data point to the average across all of the predicted y values.

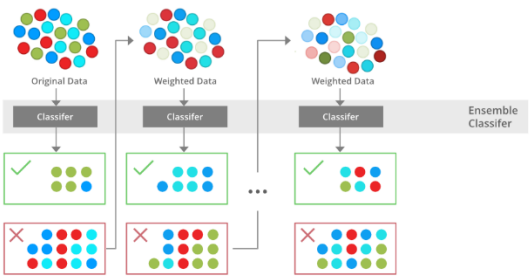
A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

- ## XGBOOST Regressor:

XGBoost is an implementation of Gradient Boosted decision trees.

XGBoost models majorly dominate in many Kaggle Competitions.

In this algorithm, decision trees are created in sequential form. Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results. The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree. These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems.



6.1 Model Performance:

Model can be evaluated by various metrics such as:

- ## Mean Absolute Error:

  In the context of machine learning, absolute error refers to the magnitude of difference between the prediction of an observation and the true value of that observation. MAE takes the average of absolute errors for a group of predictions and observations as a measurement of the magnitude of errors for the entire group. MAE can also be referred as L1 loss function.

  What exactly does 'ERROR' in this metric mean? We do a subtraction of Predicted value from Actual Value as below.

  Prediction Error → Actual Value - Predicted Value

  This prediction error is taking for each record after which we convert all error to positive. This is achieved by taking Absolute value for each error as below;

  Absolute Error → |Prediction Error|

  Finally, we calculate the mean for all recorded absolute errors (Average sum of all absolute errors).

  MAE = Average of All absolute errors

$$mae = \frac{\sum_{i=1}^{n} abs\left(y_i - \lambda(x_i)\right)}{n}$$

- ## Mean Squared Error:

  The Mean Squared Error (MSE) is perhaps the simplest and most common loss function, often taught in introductory Machine Learning courses. To calculate the MSE, you

take the difference between your model's predictions and the ground truth, square it, and average it out across the whole dataset.

The MSE will never be negative, since we are always squaring the errors. The MSE is formally defined by the following equation:

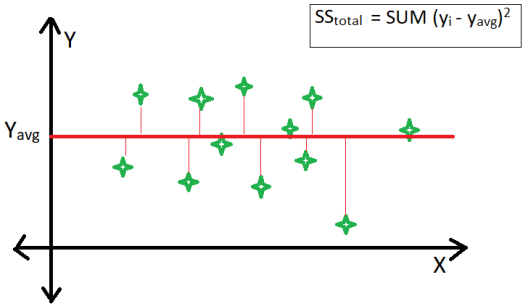
$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$$

Where N is the number of samples we are testing against.

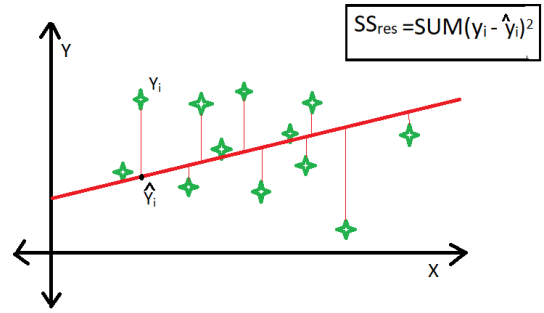
- ## R Squared(R2) Error:

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. The closer the value of r-square to 1, the better is the model fitted.

R-square is a comparison of residual sum of squares (SSres) with total sum of squares (SStot). Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.



Residual sum of squares in calculated by the summation of squares of perpendicular distance between data points and the best fitted line.



R square is calculated by using the following formula:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SSres is the residual sum of squares and SStot is the total sum of squares.

The goodness of fit of regression models can be analyzed on the basis of R-square method. The more the value of r-square near to 1, the better is the model.

Note : The value of R-square can also be negative when the models fitted is worse than the average fitted model.

6.2. Hyper Parameters:

- ### Randomized Search Cv:

  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.

7. Conclusion:

Linear Regressor, Random Forest Regressor, XGB Regressor are the models used on this project and evaluated on MSE, RMSE, MAE ,R2 score and Adjusted R2 scores and finally selected the RandomForest Model as it gaining the best score in Mean Square Error i.e. MSE is robust to outliners.

In all these 3 models our errors are ranging2,00,000 which is around 10% of the average views. The model has 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.

References-

1. MachineLearningMastery
2. GeeksforGeeks
3. Medium