**Capstone Project**

**Supervised Learning-Regression**

**TED-Talks Views Prediction**

**Technical Documentation**

**Submitted by**

**Mayur . H . Chougule**

****

**Date: 20th January 2023**

|  |  |
| --- | --- |
| **TABLE OF CONTENTS** |  |
| **Abstract** | 3 |
| **Data Summary** | 3 |
| **Steps Implemented** |  |
| 1. Exploratory Data Analysis | 3 |
| 1. Data Cleaning | 3 |
| 1. Feature Engineering | 3 |
| 1. Checking Correleation | 3 |
| 1. Outlier Treatment | 3 |
| 1. One Hot Encoding | 3 |
| 1. Model Implementation | 4 |
| 1. Model Comparison | 4 |
| 1. Model Explainability | 4 |
| **Algorithms** |  |
| 1. Linear Regression | 4 |
| 1. Regrularizations for Linear Regression | 4 |
| 1. Decsision Tree Regression | 4 |
| 1. Bagging Regression | 4 |
| 1. Random Forest Regression | 5 |
| 1. Gradient Boosting Regression | 5 |
| 1. XG-Boosting Regression | 5 |
| 1. KNN Regression | 5 |
| **Model Performance** |  |
| 1. Mean Absolute Error (MAE) | 6 |
| 1. Mean Squared Error (MSE) | 6 |
| 1. Root Mean Squared Error (RMSE) | 6 |
| 1. R2 Score | 6 |
| 1. Adjusted R2 Score | 6 |
| **Summary** | 7 |
| **Conclusion** | 7 |

**Abstract**

TED is a devoted to spreading powerful ideas on just about any topic. These datasets contain over 4000 TED talks including transcripts in many languages. The regression algorithms provide a way of making prediction for views of TED videos uploaded on website. Pre-processing the columns reveals various informations that helps in improving predictions. Some models prove that even a simple model without much Hyperparameter tuning yield a good results. Different regression evaluation metrics are implemented for comparing various Machine Learning models and obtain efficient results. The split method is utilized to validate the results in which data has been split in training and test sets.

# **Problem Statement**

The main objective of the project is to build a predictive model , which could help in predicting the views of the videos uploaded on the TEDx website and find the best algorithm that gives best prediction based on different evaluation metrics .

# **Data Summary**

The dataset contains information about various TED events over a span of 3 decades and contains following features :

1. **talk\_id:** Talk identification number provided by TED

2. **title:** Title of the talk

3. **speaker\_1:** First speaker in TED's speaker list

4. **all\_speakers:** Speakers in the talk

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

6. **about\_speakers:** Blurb about each speaker

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 the talk was given in

11. **available\_lang:**All available languages (lang\_code) for a talk

12. **comments:**Count of comments

13. **duration:**Duration in seconds

14. **topics:** Related tags or topics for the talk

15. **related\_talks:** Related talks (key='talk\_id',value='title')116. **url:** URL of the talk

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

18. **transcript:**Full transcript of the talk

# **Steps Implemented :**

## Exploratory Data Analysis

After loading the dataset , the target variable ‘Views’ was compared with other independent variables. This process helped in the understanding of various aspects and overview of the dataset and relationship among the dependent and independent variables [ Skewness, mean , median etc] .It gave a better idea of which feature behaves in which manner compared to the dependent variable.

## Data Cleaning

The dataset didn’t had any duplicate rows or columns. But included columns with considerable null\_values .The null values in columns with ‘object64’ data-type where replaced with ‘NA’ , where as those in columns with ‘int64’ data-type where replaced with ‘0’

## Feature Engineering

New features are extracted from dataset namely ‘Number of unique topics’, ‘Number of available languages ‘while others included information about ‘Event Category’, ‘Date-time features’ ,‘Average views ‘.These features improves the interpretability of dataset and also helps in better model prediction.

## Checking Correlation

The dataset didn’t had any considerable correlation between the variables.

## Outlier Treatment

The outliers in some of the important columns which affected the ML algorithms results where treated using Inter-Quartile Range (IQR) method.

## One Hot Encoding

One Hot Encoding ( converting to dummy variable ) was performed on Categorical features in string format to produce binary integers of 0 & 1 since, machine needs numerical format to understand the data.

## Model Implementation

The Regression models implemented for evaluation are

1. Linear Regression

* Ridge Regularization
* Lasso Regularization
* Elastic-Net Regression

1. Decision Tree Regression
2. Bagging Regression
3. Random-Forest Regression
4. Gradient Boosting Regression
5. XG-Boost Regression
6. KNN Regression

## Model Comparison

For evaluating the model performances, models were compared on the basis of following evaluation metrics :

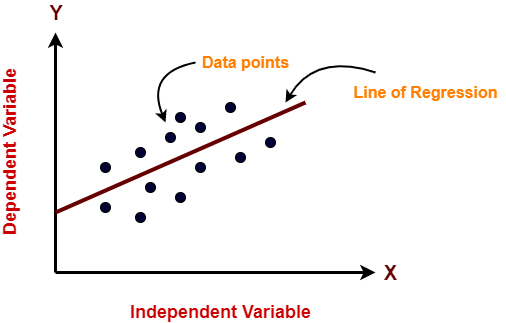
1. Mean Absolute Error (MAE)
2. Mean Squared Error (MSE)
3. Root Mean Squared Error (RMSE)
4. R2 Score
5. Adjusted R2 Score

## Model Explainability

To explain the feature importance & their contribution towards model performance is explained using SHAPLEY method, were in different plots explain whether a particular feature increases or decreases the model prediction from the base value of model.

# **Algorithms**

## Linear Regression

Linear Regression predicts a real-valued output (dependent variable) based on input values (Independent variables). It finds out linear relationship between dependent & independent variables. The representation of linear regression is y = b\*x + c. 

## Regularizations for Linear Regression

Regularizations keep the same number of features and reeduce the magnitude of coefficients to improve the model prediction and performance .

* **Ridge :** (L2 regularization) shrinks magnitude of coefficient close to zero
* **Lasso :** (L1 regularization)shrinks coefficient all the way to zero, thus removing them from the model.

Model comlexity & performance of Ridge & Lasso is based on the value of hyperparameter alpha (α )

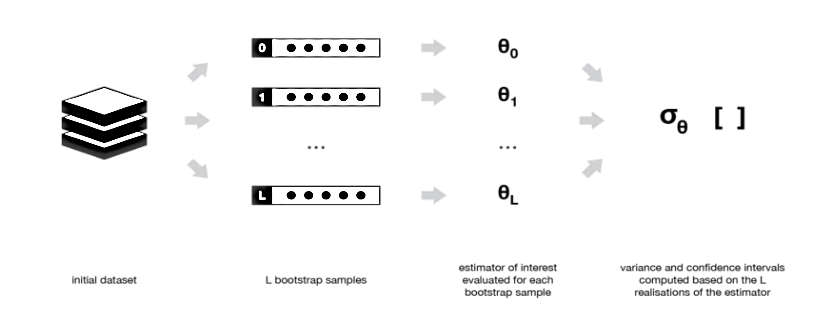
* **Elastic-Net :** Combination of Ridge & Lasso **,** uses both L1 & L2 penalty .It uses alpha (α) and l1\_ratio as parameters to control L1 and L2 penalty separately.

## Decision Trees Regression

Decision Trees refers to statistical modelling that uses a form of a tree where each node represents a feature, each branch represents a decision and each leaf represents an outcome (numerical value for regression ) .They are good at capturing non-linear relationship between the features and target variables (i.e. require very less data pre-processing) 

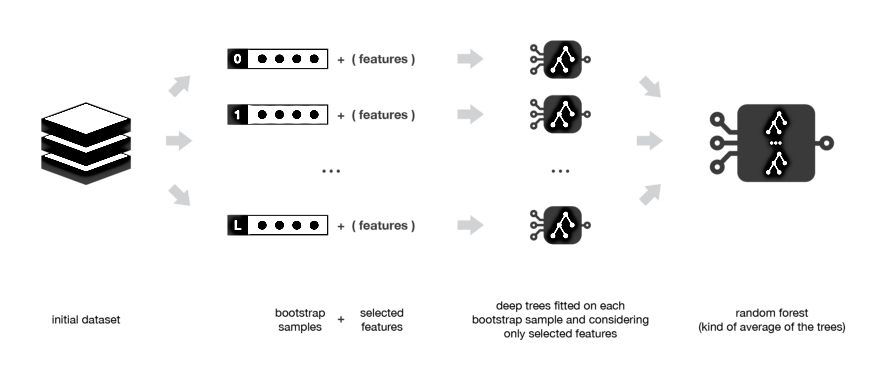
## Bagging Regression

Bagging is an ensemble technique of decision trees that combines the result of multiple subsets of the data to get a generalized result. This process is called Bootstrapping hence, the algorithm is also known as ‘Bootstrap Aggregation’

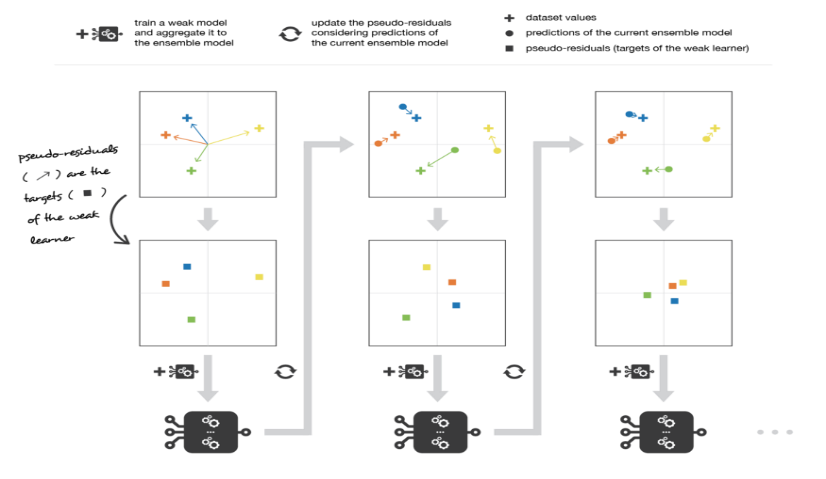


## Random Forest Regression

Random Forest is an implementation of Bagging technique, where we grow multiple decision trees and calculate the predictions based on aggregate prediction for those sub-trees. Instead of only sampling over observations in the dataset to generate bootstrap samples it also samples over features, this reduces correlation between different returned output .It is robust to outliers and automatically handles missing values.



## Gradient Boosting Regression

Boosting is a technique that consists in fitting sequentially multiple weak learners (i.e each subsequent model attempts to correct the error of previous models). Gardient Boosting is an implementation of Boosting where the objective is to minimize the loss function by adding weak learners using gradient descent. 

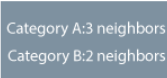
## XG-Boost Regression

Extreme Gradient Boosting (XG-Boost) is an extension of Gradient Boosting algorithm, with some optimizations and algorithmic enhancements.

* **Parallelization :** It implements parallel processing for sequentiall tree building.
* **Regularization :** It penalizes more complex models through both Lasso(L1) & Ridge(L2) regularizations to prevent overfitting.
* **Handling missing values :** It has an inbuilt routine to handle missing values
* **Cross-validation :** The algorithm comes with a built-in cross-validation method at each iteration , taking away the need to explicitly program this search.

## KNN Regression

K-Nearest Neighbors (KNN) is a supervised machine learning algorithm that can be used for classification and regression problems. It is one of the simplest algorithms to learn. It is non-parametric (i,e. It does not make any assumptions for underlying data assumptions). It is also termed as a lazy algorithm as it does not learn during the training phase rather it stores the data points but learns during the testing phase. It is a distance-based algorithm.

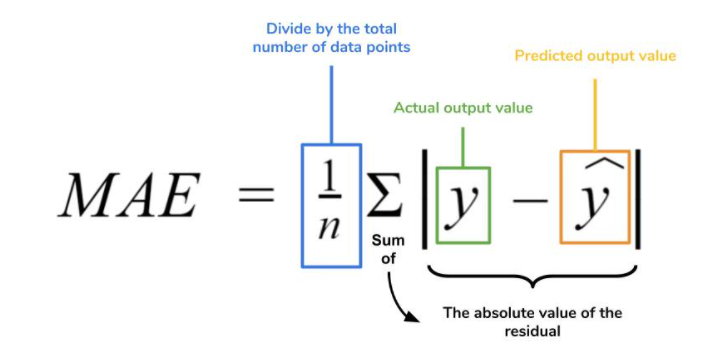


# **Model Performance**

The models can be evaluated using various Regression evaluation metrics :

## Mean Absolute Error (MAE)

MAE calculates absolute difference between actual and predicted values.



**Advantage :**

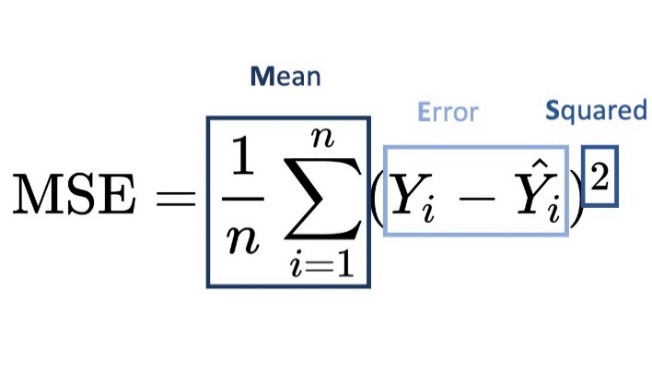
* The MAE you get is in the same unit as the output variable.
* It is most robust to outliers.

**Disadvantage :**

* The graph of MAE is not differentiable so we have to use various optimizers like Gardient-descent which are differentiable.
* It won’t punish large outliers.

## Mean Squared Error (MSE)

MSE calculates squared difference between actual and predicted values.



**Advantage :**

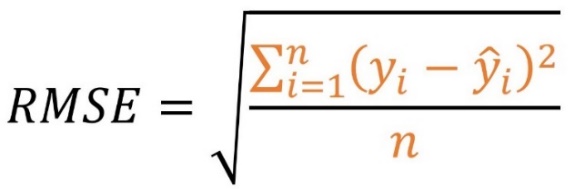
* The graph of MSE is differentiable , so can be used as a loss-function.

**Disadvantage :**

* The value after calculating MSE is a squared unit of output.
* Not robust to outliers.

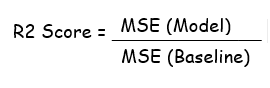
## Root Mean Squared Error (RMSE)

RMSE is a square root of Mean Squared Error (MSE).



## R2 Score

R2 Score tells the performance of the models , not the loss. It is a statistical measure that represents the proportion of variance for a dependent variable (Output) which is predictable from the independent variables(Input).

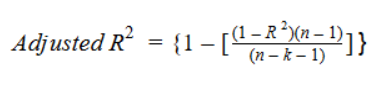


**Disadvantage :**

While adding new features in data the R2 satrts increasing or remains constant but it never decreases because it assumes that adding more data increases variance which can be false when adding irrelevant data

## Adjusted - R2 Score

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.



Here,

n = number of datapoints in our dataset.

k = number of independent variables.

R = R2 score of model.

# **Summary**

* We trained 8 Machine Learning models on training dataset by considering the best parameters for each model.
* The performance of each model was evaluated using comparison graph between Predicted and Actual values and some Regression evaluation metrics.
* We started with Linear regression, and further implemented regularizations to the same.
* To further evaluate our dataset on more complex and restricted parameters , we used Decision Trees and its ensemble techniques.

# **Conclusion**

* Bagging, Gradient-Boosting and XG-Boost have almost similar scores.
* Considering the Test\_score XG-Boost gives better predictions with slightly low error than Gradient-Boosting.
* But, considering the overall optimal values from errors and R2 score, Random-Forest has least errors and optimal R2 score for training and test dataset.