TED TALK VIEW PREDICTION

Pushkar srivastava, Rahul pandey, Hritik sharma
Data science trainees,
AlmaBetter, Bangalore

Abstract

The main objective is to build a predictive model, which could help in predicting the views of the videos uploaded on the TEDx website. We performed a Statistical Analysis of the dataset to discover relationships between each feature and the target variable. So that this relationship information can be used by the management in making better business decisions and then created a Machine Learning Pipeline, that can take in the data of any new video and predict how many views it will generate daily. It was required to keep this pipeline modular, such that it can be retrained often when new data is collected.

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

Introduction

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 ideas out there. But how do you know if your idea will be heard or appreciated? We aim to perform a comprehensive analysis of TED talks to determine what it is that makes an idea powerful.

To achieve our goal of estimating the power of an idea, we will first perform data cleaning and exploration. Then, we will begin modelling to identify what aspects of a TED talk are most useful for prediction. TED could use this information to determine what kinds of talks will become most popular and thus increase its influence. We measure our results based on views, comments, and positive ratings.

Steps Involved:

- **Exploratory Data Analysis**

  After loading the dataset we performed this method by comparing our target variable which is daily_view with other independent variables. This process helped us figure out various aspects and relationships between the target and the independent variables. It gave us a better idea of which feature behaves in which manner compared to the target variable.

- **Target Encoding**

  Target encoding is the process of replacing a categorical value with the mean of the target variable. This can help improve machine learning accuracy since algorithms tend to have a hard time dealing with high cardinality categorical features.

- **Null Values Treatment**

  Our dataset contains a large number of null values which might tend to disturb our accuracy hence we dropped them in order to get a better result.

- **Preprocessing and Feature Engineering**

  We checked for OLS Assumptions in our dataset and transformed our data accordingly to better the accuracy of our model. For removing correlation among predictors we performed Variance Inflation Factor Analysis and removed all highly correlated features.

- **One-Hot Encoding**

  We used One Hot Encoding to produce binary integers of 0 and 1 to encode our categorical features because categorical features that are in string format cannot be understood by the machine and needs to be converted to the numerical format.

- **Standardization of Features**

  Our main motive through this step was to scale our data into a uniform format that would allow us to utilize the data in a better way while performing fitting and applying different algorithms to it. The basic goal was to enforce a level of consistency or uniformity to certain practices or operations within the selected environment.

- **Splitting Dataset into Train and Test**

  We split the dataset into training and test sets. The test data will be later used to validate the ML models on unseen data.

- **Fitting Different Models**

  For modelling our prediction model, we tried :

  1. **Linear Regression**
  2. **Lasso Regression**
  3. **Ridge Regression**
  4. **ElaticNet Regression**

- **Tuning the hyperparameters for Better Accuracy**

  Tuning the hyperparameters algorithms is necessary for getting better accuracy and avoiding overfitting. We used Grid Search to optimize the values of hyperparameters.

- **Error Matrices**

  We used Mean Absolute Error (MAE) to measure the error as it will give us a more intuitive understanding of how accurate the model is. Additionally, using Mean Squared Error (MSE) to predict target variables with large values (such as the TED Talks views) can lead to problems. Together with the MAE of the model we also analyzed the variance of the target variable. Judging by the high variance of the data, it's safe to conclude that the model is performing reasonably well.

Algorithm

- **Linear Regression**

  Linear regression is the simplest and most widely used statistical technique for predictive modelling. It gives us an equation, where we have our features as independent variables, on which our target variable [views in our case] is dependent.

  The linear regression equation looks like this:

$$Y = \theta_1 X_1 + \theta_2 X_2 + \dots \theta_n X_n$$

Here, we have Y as our dependent variable, X's are the independent variables and all thetas are the coefficients. Coefficients are the weights assigned to the features, based on their importance.

The root means square error (RMSE) The most common metric for evaluating linear regression model performance is called root mean squared error, or RMSE. The basic idea is to measure how bad/erroneous the model's predictions are when compared to actual observed values. So a high RMSE is "bad" and a low RMSE is" good"

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}$$

In the formula, the difference between the observed and predicted values is called the residual. The mean squared error (MSE) is the average of all the squared residuals. Then the RMSE just takes the square root of that, which puts the metric back in the response variable scale.

Mean absolute error (MAE) is a loss function used for regression. Use MAE when you are doing regression and don't want outliers to play a big role. The loss is the mean over the absolute differences between true and predicted values, deviations in either direction from the true value are treated the same way.

MAE is not sensitive towards outliers and given several examples with the same input feature values, the optimal

prediction will be their median target value. This should be compared with Mean squared error, where the optimal prediction is the mean Use MAE when you are doing regression and don't want outliers to play a big role.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}|$$

Coefficient of determination (R^2)
The coefficient of determination — more commonly known as R^2 — allows us to measure the strength of the relationship between the response and predictor variables in the model. It's just the square of the correlation coefficient R, so its values are in the range of 0.0–1.0. Say for example that R^2=0.65. This means that the predictor variables explain about 65% of the variance in the response variable. The R^2 calculation depends on whether you're dealing with a sample or the entire population. I'll just give the sample version here. First, here's R:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i -}$$

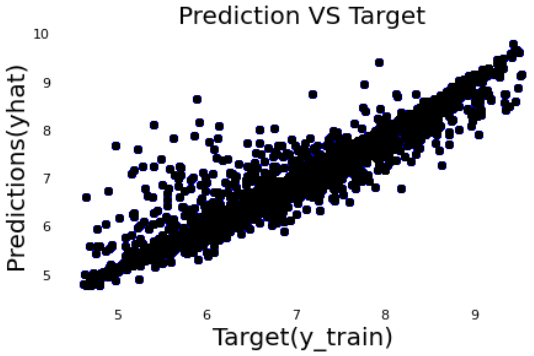
In this context, x and y are the predicted and observed values. To get the model's R^2, just square the above.

Comparing RMSE to R^2, both measure a model's goodness of fit, but they have different ideas about what "good" means. For RMSE, "good" means that the model generates accurate predictions (small residuals). For R^2, "good" means that it's the predictor variable doing the actual predictive work, as opposed to the response variable simply having low variance and being easy to predict even without the predictor variable.

Observation:

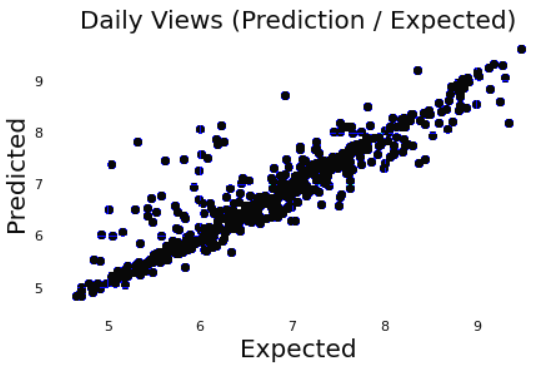
We started with implementing Linear Regression Training Model by plotting prediction vs target scatter plot

- **Training**



- **Testing**

We tested our model by using test values, and we plot to scatter plot between predicted and expected values



Conclusion:

That's it we have reached the end of this project. We started the project with Data Exploration Following EDA we performed feature engineering, data cleaning, target encoding and one hot encoding of categorical columns, feature selection and then model building.

Then we checked our model for overfitting by comparing it with the Lasso Regression model, Ridge Regression model, and ElasticNet Regression model. We found that our original base model was overfitting and Lasso Regression has the best accuracy.

In all of these models mean error is 13 %, That implies we have been able to correctly predict views 87 % of the time. Our models have performed very well on unseen data due to various factors like effective EDA, feature selection, and correct model selection.

Among all the features speaker_1_avg_views is the most important this implies that speakers are directly impacting the views.

Future Work

- Training our data on other models (XGB, Random Forest, etc)
- More efficient Hyperparameter Tuning through techniques like Random Search

References

- GeeksforGeeks
- Researchgate.net
- H2O.ai