**TED TALK VIEW PREDICTION**

**Md Amanatullah**

**Data science trainees,**

**Alma Better, Bangalore**

In this document, we explore how to predict a TED talk’s popularity by its inherent features via machine learning techniques. We quantify a TED talk’s popularity as transformation of its daily views and daily comments and include 19 features as predictors. We find that the ordinary least squares regression, ridge regression, and LASSO regression models perform well, and predictors such as a talk’s number of language translations, duration, main speaker’s occupation, as well as the timing it being uploaded have essential effects on its popularity. In the end, we also provide our suggestion on how to improve TED talks’ popularity within and beyond the scope of machine learning.

**Introduction:**

Founded by Richard Saul Wurman in 1984 as a conference and under the slogan “ideas worth spreading” (“TED (conference)”2, 2020), TED is one of the most well-known non-profit organizations in the world for its powerful impact on education. Based on the data from TED talks, TED has offered more than 4,000 talks covering various topics from science to humanities to daily lives in over 110 languages as of June 2020. According to TED blog, TED surpassed a billion video views in total back to November 2012. Also based on the data from talks, until March 2020, the most popular talk on TED has gained over 64 million views, and the median views of all the talks have also reached 1.2 million. It is not surprising to acknowledge that TED Talk, an online educational platform devoted to spreading ideas, has valued and will be valuing the popularity of its content (web-based talks) all the time. As Pinto, Almeida, & Gonçalves (2013) points out, “web content popularity is of great importance to support and drive the design and management of various services.” In TED’s case, “various services” are reflected on its slogan “ideas worth spreading” including producing high-quality videos, finding sponsorship from partnerships, establishing TED Fellows programs to support new voices, etc.

Then, how can we approach the myth behind TED talks’ popularity from the aspect of information science? Machine learning is an ideal solution. Murphy (2012) defines machine learning as “a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty.”

**Headings:**

* TED talks popularity
* Machine learning
* Linear regression
* Ordinary least squares (OLS)
* Best feature subset
* Ridge regression
* LASSO regression

**Approach:**

We find that OLS, Ridge, and LASSO models perform well in our prediction, and features such as the number of language translations, average Internet connection speed, duration, posting gap between the date of being filmed and the date of being published, main speaker’s occupation as writer or psychologist, being themed on “culture” or “design”, being published on Friday, Saturday or March are all powerful predictors.

**Steps involved:**

* **Exploratory Data Analysis**

After loading the dataset, we performed this method by comparing our target variable that is Daily 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.

* **Null values Treatment**

Our dataset contains many null values which might tend to disturb our accuracy hence we dropped them at the beginning of our project in order to get a better result.

* **Encoding of categorical columns**

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 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.

* **Fitting different models**

For modelling we tried various regression algorithms like:

1. **Linear Regression**

Linear Regression Linear regression assumes linear functional dependency between the independent variables and the dependent variables.

1. **Ordinary Least Squares (OLS)**

OLS is the simplest type of linear regression without regularization or feature selection. In other words, we will put all 19 independent variables to fit the linear model by the principle of least squares, which refers to “choosing the regression coefficients so that the estimated regression line is as close as possible to the observed data, where closeness is measured by the sum of the squared mistakes made in predicting Y given X. OLS has its advantages for it is efficient to operate the model building process with reasonable computation and the results are easy to interpret. However, its disadvantages are also evident, for example, without regularization or feature selection, we can include some useless variables in the model since we have no ideas on how to distinguish which of these 19 predictors are useful for the model building and this will lead to an overly complex model or overfitting issues. After all, OLS could function as the baseline method for others to compare with

1. **Ridge Regression**

Ridge regression is also invented for controlling model complexity based on OLS. Instead of directly minimizing OLS’s least squares, ridge regression adds the regularization/ penalty term λ ∑ j2 β (where λ ≥ 0 is a tuning hyperparameter and βj refers to any coefficient given p-dimensional model, which in our case p = 19). We will use defaulted values of λ and cross-validate them to find the most reasonable hyperparameter.

Although ridge regression helps us control model complexity via λ, it suffers from the problem of interpretability from the shrunken coefficients. Plus, it will include all 19 variables without doing any feature selection so that it won’t apply well to the cases when many of the independent variables are useless, which we will never know before running any models.

1. **Lasso Regression**

With a similar idea of shrinking coefficients, LASSO regression can be regarded as a transformation from ridge regression. According to James et al. (2013), the only difference between these two is LASSO regression adds the term λ ∑ | βj| (where λ ≥ 0 is a tuning hyperparameter, and βj refers to any coefficient given p-dimensional model, which in our case p = 19).

LASSO regression also controls model complexity via λ like ridge regression does, while not like ridge, it does feature selection by yielding zero coefficients for some variables. Therefore, LASSO regression usually outperforms ridge regression if the case is many of the independent variables are useless. However, as mentioned, we will never know how many of our independent variables are useful before running any models. Therefore, it is better to experiment with both ridge and LASSO.

* **Tuning the hyperparameters for better accuracy**

Tuning the hyperparameters of respective algorithms is necessary for getting better accuracy and to avoid overfitting in case of linear regression models. Hence, we use regularized linear regression like Ridge and Lasso.

1. **Grid Search CV-**Grid Search combines a selection of hyperparameters established by the scientist and runs through all of them to evaluate the model’s performance. Its advantage is that it is a simple technique that will go through all the programmed.

combinations. The biggest disadvantage is that it traverses a specific region of the parameter space and cannot understand which movement or which region of the space is important to optimize the model.

**2.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.

* **Evaluation Metrics**

An integral part of machine learning is doing the evaluation. In our research’s case, the evaluation question would be, how can we decide whether a model with certain features outperform the other one? Specifically speaking, a fit evaluation metric is mainly determined by the dependent variable. If the dependent variable is “number of views” or “number of comments,” then it’s a numerical prediction task. In that case, “mean squared error” or “mean absolute error” or R square or adjusted R square. Given our dataset supports us in conducting predictions on a numerical dependent variable, mean squared error (MSE) or mean absolute error (MAE) will be the primary metric for us to evaluate a model’s performance.

* **Conclusion:**

Conclusion In a nutshell, a TED talk’s popularity can be predicted by its inherent features via machine learning techniques. We found that the OLS, Ridge, and LASSO models performed well in the prediction, and we also learned several powerful predictors such as a talk’s number of language translations, duration, main speaker’s occupation, as well as its being published timing.

With the support of our experimented models and their corresponding predictors, we detected that a TED talk’s views or comments are highly related and can either function well as the indicator of “popularity.”

Furthermore, we also investigated how to predict future data and make sound decisions based on our trained models.