ML-Regression-TED-talks-Project

Machine learning project to predict views for a TED talks using Regression Model.

Problem Statement -

TED is all about spreading powerful ideas on any topic. This dataset contains over 4,005 TED talks, including transcripts in multiple languages.

Founded in 1984 by Richard Salman, the nonprofit organization dedicated to bringing together experts in the fields of technology, entertainment, and design, TED talks have become a mecca for ideas from nearly every industry. In 2015, TED and its sister TEDx chapters published over 2,000 talks that are freely available to the general public, with a list of speakers including Al Gore, Jimmy Wales, Shah Rukh Khan and Bill Gates.

TED Talks has been using “ideas worth spreading” as a platform for years. In the digital world we live in today, TED is a great platform to spread your ideas. But how do you know if your ideas will be heard or appreciated.

The main goal is to build a predictive model that can help predict the number of views of videos uploaded to the TEDx website.

Project Summary -

In this project, the objective is to predict the number of views a TED talk session can have using the dataset provided, by applying specific methods to draw out heuristics after doing exploratory data analysis on the dataset. Before beginning with the data analysis, data wrangling will be done to get the data in a format ready for analysis. Hence after completing the wrangling and analysis statistically and visually, we will continue doing transformations on the dataset, if required. Transformations, such as encoding, normalization, and regularization of the dataset. After the data is ready is ready to be entered in to a model, it will be split into a proportion and a portion of it will be kept for testing the model chosen. Hence, multiple models will be applied to get acknowledged about the model that fits the dataset best for predictions, which will also be followed by parameter tuning, if required to get the highest accuracy possible.

Variables Description

- **talk_id** - identification number provided by TED (int)
- **title** - Title of TED talk (string)
- **speaker_1** - First speaker in TED speaker list (string)
- **all_speakers** - All the speakers in the TED session (dictionary)
  ```
  [FORMAT - {'speaker' : 'speaker_name'}
  (speaker - speaker number of session, speaker_name - name of speaker)]
  ```
- **occupations** - Occupation of the speakers of the session (dictionary)
  ```
  [FORMAT - {'speaker' : 'speaker_occupation'}
  (speaker - speaker number of session, speaker_occupation - occupation of speaker)]
  ```
- **about_speakers** - Descriptive text about speakers (dictionary)
  ```
  [FORMAT - {'speaker' : 'speaker_description'}
  (speaker - speaker number of session, speaker_description - About the speaker)]
  ```
- **views**(**Dependent Variable**} - Number of views (int)
- **recorded_date** - Date of the TED session (string)
- **published_date** - Date of TED session publishment (string)
- **event** - Event of the TED session (string)
- **native_lang** - Native language (string)
- **available_lang** - All the available languages (list)
- **comments** - Number of comments received (int)
- **duration** - Duration of TED talk session (in seconds) (int)
- **topics** - Topic of the TED session and tags (list)
- **related_talks** - Related TED talk sessions (dictionary)
  ```
  [FORMAT - {'talk_id' : 'title'}
  (talk_id - column 1 of dataset, title - column 2 of dataset)]
  ```
- **url** - URL link of the TED session (string)
- **description** - Information about the TED talk session (string)
- **transcript** - Complete transcript of the TED session (string)

MODELS USED:

- Linear Regression
- Ridge Regression
- Lasso Regression
- Random Forest Regressor
- eXtreme Gradient Boost Regressor

MODEL METRICS:

index	model_name	train_score	test_score	train_score_cv	test_score_cv	MSE	RMSE	MAE	R2_SCORE
4	XGBRegressor	0.9934261278409272	0.932030624400551	0.9999999999294615	0.938762635871558	1038051732422.2239	1018848.2381700544	230907.41243408204	0.93203062
3	RandomForestRegression	0.9781102368086143	0.9239665329101074	0.9890333234284177	0.9098070854510008	1161209317264.5068	1077594.2266291643	254280.2227105074	0.92396653
0	LinearRegression	0.9077704289921005	0.8699250323297967	0.9077704289921005	0.8699250323297967	1986549741615.0134	1409450.1557752986	529169.1077861005	0.86992503
2	LassoRegression	0.9077704289921005	0.8699250323078986	0.9077704288720964	0.869922842797592	1986549741949.4487	1409450.1558939388	529169.1077303984	0.86992503
1	RidgeRegression	0.907716344510733	0.8696728644141192	0.9036489763793327	0.8640585013731682	1990400936942.6416	1410815.6991409762	523354.9398250869	0.86967286

MODEL ANALYSIS (findings)

NOTE: We will provide more preference to MAE for the process of model selection not RMSE because :
- RMSE can be affected severely by outliers.
- MAE is linear in nature.

- The XGBRegressor has the least MAE and 0.93(default) and 0.94(CV) R2 score and RandomForestRegressor also has similar metrics in regards of MAE but with a R2 score of 0.93 (default maximum) and 0.92 (CV) approximately.
- After tuning the hyperparameters, both of the models provide > 0.90 test score.
- The training score is considered as well, XGB score is 0.99 and RandomForestRegressor is almost 0.99.
- XGBREgressor gives us a least MAE of 9 percent of the actual mean, and RandomForestRegressor gives us a MAE of 10 percent (11 percent max) making both of them provide predictions with > 90% accuracy.

Considering the efficiency of the models, RandomForest provides more simplicity. But among all the regressor models available, **XGBRegressor provides lesser MAE**, and even lesser than RandomForestRegressor.

CONCLUSION

- The views for sessions with longer duration is less on average basis.
- Primary Speaker and comments are the most important features.
- More number of languages available can increase the number of views.
- The recency of the session published, also plays a major role in deciding the number of views.
- Features like occupations, about_speakers, url have the least importance.
- XGBRegressor and RandomForestRegressor have the best MAE and test score among all the models.