TED Talk: Insight and Key Success Factor Study

Abstract — This research explores the factors that contribute to the success of TED Talks by analysing video performance through a unique metric of popularity. Popularity is defined as the total views divided by the duration the talk has been available online, calculated as the difference between the scraping date and the release date. While the number of reactions to a video is acknowledged, the primary focus is on this popularity metric to provide a consistent and objective measure for comparing talks. The study aims to uncover actionable insights into the elements that drive a talk's success, offering valuable guidance for speakers and content creators seeking to maximize impact and engagement.

Introduction

Today, an abundance of online content is available on virtually any topic. Therefore, content creator or a speaker who wants to connect with their audience must understand strategies to make their talks more engaging and attract a larger audience.

TED Talks are freely available on popular platforms such as YouTube in video format and on Spotify and other podcast platforms in audio format. Analysing TED Talk data provides valuable insights into the general trends in the popularity of these talks.

This analysis provides insight into the factors which contribute to the success of a TED talk, and which can be used by speakers and content creators to make better decisions about the way they present their speech and on which topic they should focus on to make their talks be heard and well received by more audience.

1 Analysis Domain, Question and Plan

1.1 Data and Domain

We source our data from Kaggle. The datasets contain information about all audio-video recordings of TED Talks uploaded to the official TED.com website until September 23rd, 2017[1].

1.2 Research Questions

Several factors contribute to a talk's success, including capturing the viewer's curiosity to draw them in, maintaining their attention throughout the speech, and ultimately receiving a positive response from the audience. Our research question aims at finding those features.

The research questions we are going to address here are:

- Do certain words, description style make talk more popular?
- Are certain topics inherently more popular?
- Does narrative style, speech pace, talk's length and language complexity of the TED talk influence popularity?
- Is it possible to predict talk popularity using data available prior to the talk?

1.3 Analysis Plan

The objective is to find factors that effect TED talk's success and find best ways to leverage them to improve future talks. Thus, the following plan was implemented to attain our objective.

1. Importing data files (.csv) form Kaggle and merging them together.
2. Data cleaning: removing outliers and handling missing values.
3. Define matrix, grades and categories used for testing and comparison.
4. Feature engineering: creating features which will help in our

analysis and binning features if necessary.

5. Analysing the data in two parts: first analysing the descriptive variables (features which are used to draw users' attention prior to the talk like title and description) and analysing the feature of the talk.
6. Developing model and gaining insight about the feature using it.
7. Comparing our results and concluding.

2. Analysis

2.1 Initial data preparation

Initial exploratory data analysis (EDA) was conducted to identify outliers and address missing values. Numeric outliers were handled using the interquartile range (IQR) method[2], ensuring robust central tendency estimates and preventing undue influence on model performance. For missing data, rows with missing target values (popularity) were removed, while numerical feature gaps were imputed using the mean.

2.2 Feature Engineering and Encoding

Key features were engineered to extract meaningful insights from the dataset:

1. **Popularity Metric**: A normalized metric, views per day (views/days_out), was created to account for the time a talk had been available.
2. **Feedback Score**: A composite feature combining comments and total ratings to capture audience engagement.
3. **Sentiment and Readability**: Sentiment polarity[3] and Flesch-Kincaid Grade Level scores[4] for titles, descriptions, and transcripts were computed using TextBlob and textstat, evaluating tone and complexity.
4. **Tag Analysis**: Tag combinations were assessed for thematic influence, with tag count and a weighted rating score (positive, neutral, negative) introduced to reflect audience preferences.
5. **Rating Categories**: Ratings were grouped into positive (e.g., "Inspiring"), neutral (e.g., "OK"), and negative (e.g., "Confusing").

2.3 Title analysis

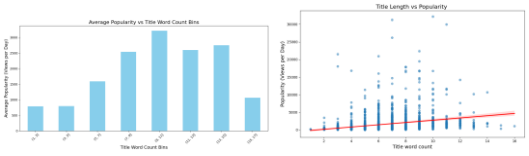

Figure 1.

The analysis of TED Talk titles explored their influence on popularity:

1. **Title Length**: Titles were binned by word count, and ANOVA tests examined significance in popularity differences (Figure 1). Titles of medium length were slightly more popular on average.
2. **Starting Words**: Frequently used starting words (e.g., questions or numerals) were analysed for their effect. Titles beginning with "3" or questions were associated with higher average popularity (Figure 2).
3. **Complexity and Sentiment**: Flesch-Kincaid scores and sentiment polarity were compared, finding that simpler titles generally outperformed complex ones, while sentiment had no statistical significance (Figure 3).

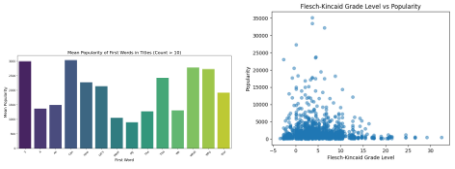

Figure 2. figure 3.

2.4 Tag Analysis

The significance of tags was evaluated using the Mann-Whitney U Test[5], comparing popularity scores of talks with and without specific tags. False Discovery Rate (FDR) [6] correction using the Benjamini-Hochberg method [7] reduced false positives, revealing tags like "society" and "identity" with weak positive correlations to popularity (Figure 4). Additionally, talks with more than 10 tags showed significantly higher average popularity (Figure 5).

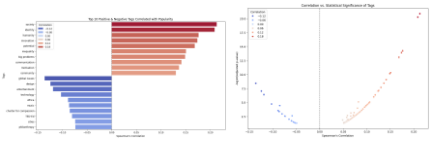

Figure 4.

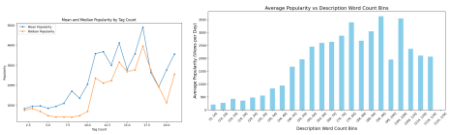

Figure 5. figure 6.

2.5 Description analysis

1. **Word Count**: Description lengths were visually analysed, showing a positive correlation between longer descriptions and popularity (Figure 6).
2. **Complexity**: Description complexity was assessed using Flesch-Kincaid scores, but no significant impact on popularity was observed.

2.6 Transcript and talk analysis

1. **Transcript Length**: Moderately long transcripts (1,700-1,800 words) were most popular, confirmed through visual analysis.
2. **Talk Duration**: Talks lasting 10-14 minutes were associated with higher popularity, aligning with TED's recommended average duration.[8]
3. **Speech Pace**: Talks delivered at a slower pace (fewer words per minute) were more engaging, possibly due to improved clarity.
4. **Language Complexity**: Transcript complexity, measured using readability metrics, did not significantly affect popularity.

2.7 Model Training

In this section, we trained multiple machine learning models to predict the popularity of TED Talks using various features from the dataset. The models selected for evaluation include **Ridge Regression**, **Lasso Regression**, **Random Forest Regressor**, and **Gradient Boosting Regressor**. Each model was tuned using **GridSearchCV** to optimize hyperparameters and achieve the best performance. Below is a brief explanation of each step, the models used, and their evaluation results.

Step 1: Data Preprocessing

1. **Handling Missing Values**:

   Rows with missing target values (Popularity) were removed, while missing numerical values were imputed using the mean of each feature.

   Remaining missing values were dropped to ensure a clean dataset for model training.

2. **Feature Engineering**:

   Features were divided into **categorical** and **numerical** types. Numerical features were standardized using StandardScaler [9], and categorical features were one-hot encoded using OneHotEncoder [10].

The target variable (Popularity) was log-transformed (log1p) to reduce the impact of extreme values and improve model stability.

3. **Data Splitting**:

   The data was split into training, validation, and test sets with a 70/15/15% split ratio. This allowed us to train the models, validate them using unseen data, and finally test their generalization performance.

Step 2: Model Training

We tested and tuned four models: **Ridge Regression**, **Lasso Regression**, **Random Forest Regressor**, and **Gradient Boosting Regressor**.

1. **Ridge Regression**:

   Ridge regression is a linear regression model with **L2 regularization** to prevent overfitting. It penalizes large coefficients in the model.

   Hyperparameter tuning: We tuned the alpha parameter to control the regularization strength.

   Best Model:

   Training - MSE of 0.4611

   R2 score of 0.653

   validation- MSE of 0.593

   R2 score of 0.518

2. **Lasso Regression**:

   Lasso regression is like Ridge but uses **L1 regularization**, which can also drive some feature coefficients to zero, aiding in feature selection.

   Hyperparameter tuning: We optimized the alpha parameter, which controls the strength of regularization.

   Best Model:

   Training - MSE of 0.785

   R2 score of 0.409

   validation- MSE of 0.693

   R2 score of 0.436

3. **Random Forest Regressor**:

   Random Forest is an ensemble method that creates multiple decision trees and averages their predictions. It handles non-linearity well and is robust against overfitting.

   Hyperparameter tuning: We optimized the number of trees (n estimators) and the depth of the trees (max depth).

   Best Model:

   Training - MSE of 0.098

   R2 score of 0.926

   validation- MSE of 0.546

   R2 score of 0.555

4. **Gradient Boosting Regressor**:

   Gradient Boosting builds decision trees sequentially, where each tree corrects the errors of the previous one. It is a powerful model for complex data.

Hyperparameter tuning: We optimized the number of estimators (n estimators) and the learning rate.

Best Model:

Training - MSE of 0.331

R2 score of 0.751

validation- MSE of 0.586

R2 score of 0.522

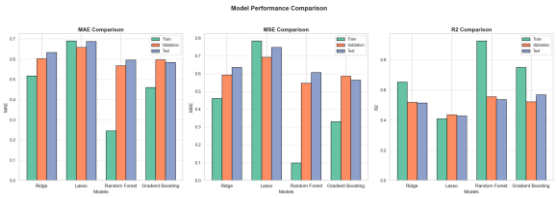

Figure 7.

3. Findings, reflection and further work

A) Do certain words, description style make talk more popular?

In our study, we found that talks which has moderate word count (around 9 to 11) or start with "3" or is phrased like a question is more popular on average. Besides that, title should not to too complex or vague. Also talks with longer descriptions also performed better. (Figure 8. and 9.)

B) Are certain topics inherently more popular?

Figure 4. show the top 10 features with highest positive and negative correlation with respect to popularity, from that graph we can see that some topics tends to be inherently more popular than the other, namely "society" and "identity". But the correlation is not strong.

C) Does narrative style, speech pace, talk's length and language complexity of the TED talk influence popularity?

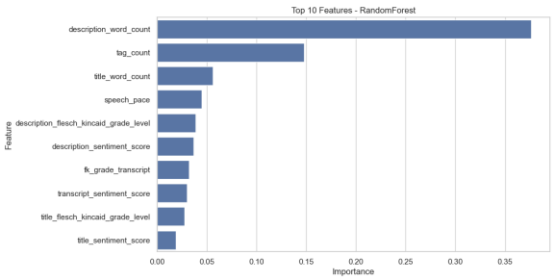

Figure 8.

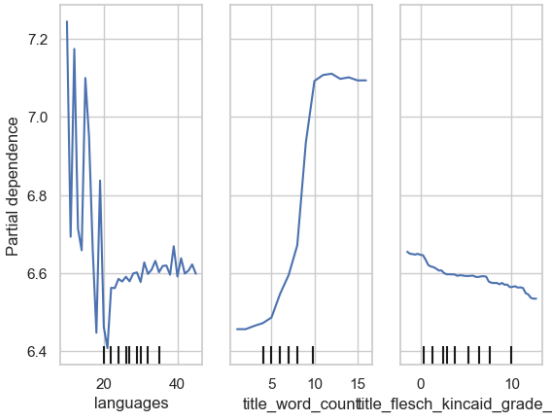

Figure 9.

Figure 8. show the top 10 important features for the random forest model we trained. We can see from it that speech pace in also an important feature when it comes to predicting talk's popularity. From of study we observed that speech with speaking speed less than the average 150-160 words per minute [11] where more popular. The complexity of the speech did not have any impact on the popularity.

D) Is it possible to predict talk popularity using data available prior to the talk?

Figure 7. show our model performance, our best model prediction rate was just above 0.5, so it prediction was no better then a random guess.

So, for further work, the matrix for analysis can be improved, new features can be engineered to improve model performance, perform transcript subject based analysis, use K mean clustering to create new features and use word cloud to distinguish them.

4. References

[1] Banik, R. "TED Talks Dataset." *Kaggle*, https://www.kaggle.com/datasets/rounakbanik/ted-talks?select=transcripts.csv, accessed 20 December 2024.

[2] "Identifying outliers with the 1.5xIQR rule." *Khan Academy*. https://www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule.

[3] "Sentiment Analysis." *Wikipedia*, last modified 21 October 2023, https://en.wikipedia.org/wiki/Sentiment_analysis.

[4] "Flesch–Kincaid Readability Tests." *Wikipedia*, last modified 21 December 2023, https://en.wikipedia.org/wiki/Flesch%E2%80%93Kincaid_readability_tests.

[5] "Mann–Whitney U test." *Wikipedia*, last modified 20 September 2003, https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test.

[6] "False discovery rate." *Wikipedia*, last modified 20 September 2003, https://en.wikipedia.org/wiki/False_discovery_rate.

[7] Benjamini, Y., and Hochberg, Y., 'Benjamini–Hochberg Method', *Encyclopedia of Statistics in Quality and Reliability*, 2nd edn (Springer, 2017), https://link.springer.com/referenceworkentry/10.1007/978-1-4419-9863-7_1215 (last accessed 22 December 2024).

[8] Eldor, T. "Data Reveals: What Makes a TED Talk Popular?" *Towards Data Science*, 27 January 2018. https://towardsdatascience.com/data-reveals-what-makes-a-ted-talk-popular-6bc15540b995.

[9] Pedregosa, F. et al., 'sklearn.preprocessing.StandardScaler', *Scikit-learn: Machine Learning in Python*, (2023), https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.StandardScaler.html (last accessed 22 December 2024).

[10] Pedregosa, F. et al., 'sklearn.preprocessing.OneHotEncoder', *Scikit-learn: Machine Learning in Python*, (2023), https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.OneHotEncoder.html (last accessed 22 December 2024).

[11] "Speaking Speed Test - Test your speech rate in a minute (WPM)." *TypingMaster*, https://www.typingmaster.com/speech-speed-test/.

Word count:

Abstract: 110

Introduction: 138

Analysis Domain, Question and Plan: 268

Analysis: 925

Findings, reflection and further work:286

Reference: 164