TED Talks Analysis

S. Bertoldo, P. Faraji, S. Mulatu, and T. Ould Amer

University of Pisa, Pisa, Italy
Department of Computer Science

Abstract. This paper discusses datasets that gather information about several TED Talks held around the world and in different types of the TED events from 2006 to 2017.
The main dataset contains general details about the talks and the speakers, while the second one consists of the transcripts of the talks. The first section addresses the motivations behind this project. Then, the second one summarizes the state of the art.
After the summary, the description of the content, the cleaning of the data and the list of the features will follow. The next step will deal with the feature engineering and exploiting extra resources and datasets such as Google Trends, which will eventually lead to the following section discussing the distribution and correlations of the meaningful variables. The detection of outliers in the various features will be detailed afterwards.

In the next parts, the classification models will be detailed and their performance analyzed then compared. The model with the best performance was chosen but then a less precise one was preferred to gain in interpretability.
Thus, the following section of this paper focuses on new ways and approaches for this dataset to have a meaningful use.
The last parts complete the previous one by presenting study cases explained with the previously extracted rules, then concludes the whole analysis.

Keywords: TED Talk · Data Understanding · Feature Engineering · Random Forest · Decision Tree · Classification · Supervised Learning

1 Motivation

The general goal of this project is to create a recommendation system to define how to perform the best Ted Talk to reach more success in public, calculating the right mixture of funniness, words choice, time of the year topics and choice of the best title, etc.

We could also see which is the influence of the reputation of the speakers before their talk and the popularity of the tags on the fame of each talk.

2 State of the art

Most of the previous works have covered various questions regarding this set of data. The spotlight was especially on the main dataset on which correlation

studies were performed with comments on the distribution and statistics provided in the data [1][2][3][4]. Some of these works used the transcripts dataset to extract patterns and keywords from the transcripts and tried to link them with metrics in the main dataset such as number of views or comments[5].

3 Solution's Definition

We need to understand how popular a new talk will be, so the first step is to define what popularity is and how it is correlated with other features.

As regards funniness, we will need to find a way to identify a video as funny, and to check the reputation of the speakers a useful tool could be Google Trends. Each step will be individually investigated in the following sections.

4 Data Understanding

4.1 Content and Cleaning

We have two datasets available: one with metadata about every TED Talk hosted on the TED.com website until August 2017, and the other with the transcripts of these videos. *ted_main.csv* is the former one, while *transcripts.csv* is the latter.

The features of the *ted_main.csv* dataset embrace every aspect related to the video. **name** is the official name of the talk, **title** is the title without the speaker's name. **description** summarizes what the speech is about, while **main_speaker, speaker_occupation** and **num_speaker** describe respectively the name, occupation and number of each speaker of a talk (whose maximum number is 5). **duration** is the number of seconds it lasts, **tags** are the associated tags, **event** is the name of the event the talk took place in, **comments** is the number of comments made on the talk, and **languages** define the number of available subtitle languages for the speech. **film_date** and **published_date** are Unix timestamp for the filming and publication dates and **ratings** is a dictionary of the various words used to describe the talk. **related_talk** is a list of dictionaries of recommended talks to watch next, **url** is the video's URL and **views** is the number of views.

The *transcripts.csv* dataset is more elementary as it contains only two fields: **transcript** and **url**.

In *ted_main.csv* we have 2550 entries, and the only feature in which we have missing values is **speaker_occupation** for a total of 6 talks that were handled by putting the mode (*Writer*) instead. In *transcripts.csv* we don't have null entries but we have 2467 transcriptions, with 3 repeated ones, resulting in 86 untranscripted videos.

Both datasets were downloaded with a CC BY-NC-SA 4.0 Licence from Kaggle website [6].

4.2 Statistics

Regarding the statistics of some variables, the most viewed talks count more than 47M **views**, the least one has 50k views, with a mean of 1.7M views. The maximum number of **comments** is 6404, and the minimum is 2, the distribution of these values nearly respect Pareto's distribution, since 20% of the videos contains 56% of the comments: the biggest part of the videos contain very few comments in contrast with a little part of the videos which contain most of the comments. Pareto's distribution can be found also in the **event** distribution: on 355 total events, 78% of the speeches were filmed in 20% of the events. The most present event is TED2014 with 84 talks.

The mean **duration** of the talks is about 14 minutes. The **film_date** range from 1972 to 2017. The reason why we have Ted videos even preceding Ted creation is because they are historical filming records. Most of the videos have around 30 **languages** and the **main_speaker** who did the maximum number of talks is Hans Rosling with 9 appearances.

As regards **speaker_occupation**, after having divided the various occupations (since some speakers define themselves with more than one job), we obtained that the most frequent job is "writer", followed by "author" and "activist".

4.3 Feature Engineering

In order to have a better evaluation of talks, we can have some scores indicating how popular or how funny each talk is. Each of these scores are between 0 and 1. Moreover, high values mean higher popularity or fun and low values mean the opposite. To say if a talk is popular or not we should look at three main features: number of comments, number of views, and sum of ratings. Thus, for popularity score we used the following formula:

$$\begin{aligned} x &= Number\ of\ Comments \\ y &= Number\ of\ Views \\ z &= Sum\ of\ Ratings \\ popularity\ factor &= \frac{normalized(x) + normalized(y) + normalized(z)}{3} \end{aligned} \tag{1}$$

We can also say if a talk is funny or not based on two features: number of the word *"laughter"* in the transcript and number of funny in the ratings.

$$\begin{aligned} x &= Number\ of\ "laughter"\ in\ transcript \\ y &= funny\ rating \\ funny\ factor &= \frac{normalized(x) + normalized(y)}{2} \end{aligned} \tag{2}$$

All normalizations used in the above formulas are Min-Max Normalization.

4.4 Extra resources

For further investigation about the talks, we can use extra resources like Google Trends. By using Trends, we hope to find probable meaningful relations between speakers' "interest over time" and popularity of their talks. Not only speakers' trends but also subjects' trends can be helpful.

"Interest over time" of all speakers and all subjects from 2006 to 2017 (beginning and end of dataset) were retrieved and it shows a number in range [0, 100] for each year, which represents search interest for that specific year. A value of 100 is the peak popularity and a value of 50 means that the term is half as popular. A score of 0 means there was not enough data for this term.

4.5 Variables Distribution and Correlations

The Pearson correlation coefficient matrix **Fig 1** between pair of all numeric attributes shows that there is a positive correlation among: (popularity and no. comments, no. views and ratings count) which is normal since it is derived from them, (Rating count, comment), and (rating count and views) are also correlated as we could expect.

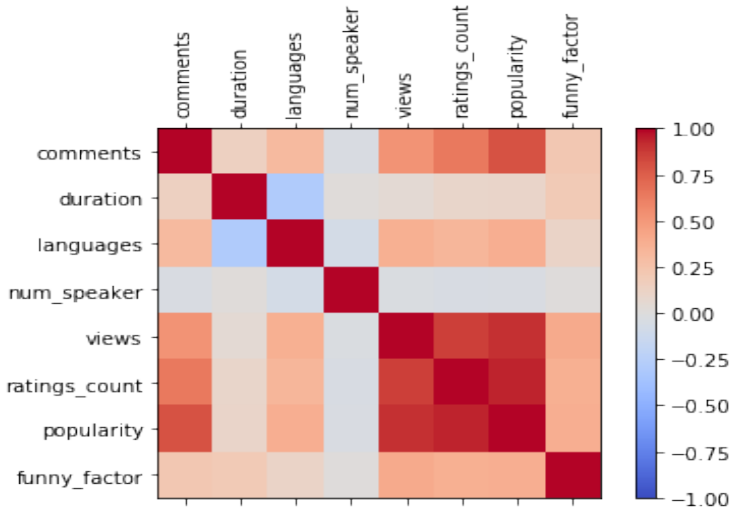

Fig. 1: **Pairwise correlation matrix**

Popularity, funny factor and Publication day:- Talks published on 'Tuesday' have a higher number of views and high frequency of 'laughter'. The ones published on weekend days have less views as shown below in Fig.2. Hence, number of views and frequency of 'laughter' in the the talk are highly correlated to popularity and funny factor respectively. The publication date of the talk is positively correlated to the popularity and the funny factor of the talk.

Using boxplots we performed an outliers' detection and a data distribution analysis. The distribution of the number of comments, views and rating count

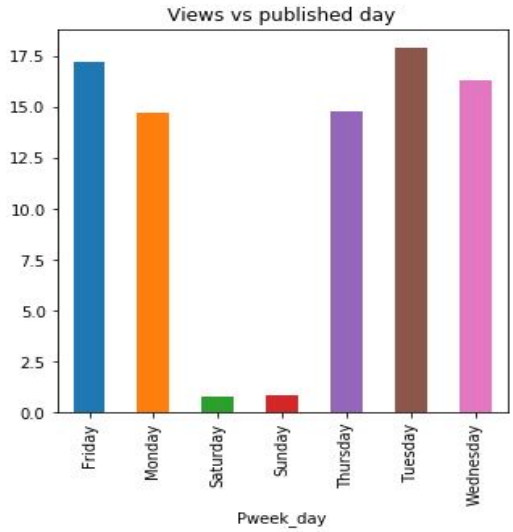

Fig. 2: **Published Date vs Views**

is almost uniform. However we are able to see that popularity and number of views have a similar distribution compared to the one of number of comments, its median value is lower. The boxplot below shows the distribution and outliers regarding the popularity of talks.

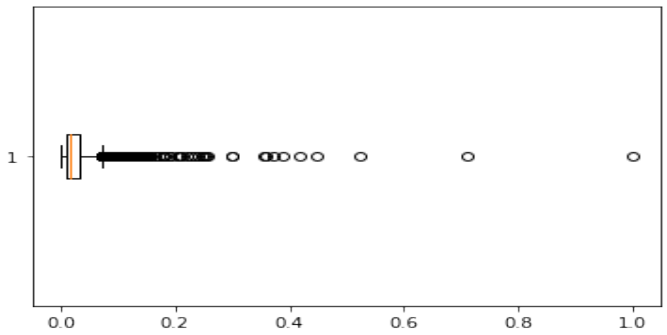

Fig. 3: **Comments boxplot**

Regarding being funny of a talk, from the box plot we can see that viewers have a slightly different opinion about the talk though frequency distribution of 'laughter' in each talk is slightly different compared to funny factor. This could suggest that viewers has a slightly different opinion on being funny of a particular talk in which it has high frequency of 'laughter'. W.r.t outlier detection considering values less than $Q1-1.5*IQR$ and values greater that $Q3+1.5*IQR$ funny count has high number of outliers , Table 1 summarizes the no. outliers.

Table 1: **Outliers**

Feature Name	Views	Popularity	Comments	Funny Count	Rating Count	Funny Factor	Laughter
No. outilers	239	228	195	336	230	164	183

5 Classification

5.1 Dummification of variables

Before creating a prediction model, some categorical variables needed to be converted into numerical features because models such as decision tree and random forest accept only this kind of values. Therefore, **occupations, events, publication day**, and **ratings** have been converted into numerical values.

Since **speaker_occupation** contained a lot of different entries, we simplified grouping them in different fields: we created **Literature, Art, Economy_Politics, Medicine, Engineering_Science** and **Other**, in a dummy way (containing only binary values).

The variable **Events** was converted into dummy columns using the most frequent events and creating **TED**, **TEDx**, **TED_Global**, **TED_Other**, **Non_TED_University** (non-Ted events in Universities), and **Non_TED_Other** (other non-Ted events). Regarding the **publication_day**, we extracted the day of the week were the video was released and we made a new column for each day, so we added seven new columns. For **ratings**, we divided them into three groups of **positive_rating, negative_rating,** and **neutral_rating**.

Regarding **Google Trends** data and **tags**, we computed the average of the value of each speaker and of each tag from 2006 to 2017 and created the corresponding new columns in the dataset. Furthermore, for the tags we did another average keeping only one value per talk.

5.2 Target Prediction and Baselines

Our target variable is **popularity**, which is a number from 0 to 1 that we decided to divide into three main groups: **low, normal**, and **high**. Thus, we created a new column consisting in only three values (0, 1, and 2). An important fact to consider is that the splitting was made in a way to get the same number of items in each group, hence the threshold values between [0, 1] were not the simple division of the range by 3. This approach solved our imbalance problem for popularity.

Our goal in predicting the popularity class is that we should predict if a talk is going to be popular or not, for this reason the metric we must use to evaluate our model is precision.

We then created different kinds of baselines to be compared with our model. The first one to be used was *dummy classifier*, which assigned to each talk a completely random value 0, 1 or 2.

Moreover, we decided to use *trends* of the speakers to create another baseline because we thought that it might be good indicator of popularity. The results of all these baselines are displayed in table 2.

Table 2: **Baselines results**

Baseline	**Precision**
Dummy Classifier	0.340
Speaker Trends	0.370

5.3 Tested models

The models we decided to use were Decision Tree and Random Forest. In the following section we will analyze the results.

For **Decision Tree**, we first used all numerical variables we had as input. But since our target class is derived from ratings (see 4.3) we decided to remove them from our model. The parameters for testing are chosen from top three models of GridSearch method for each test size.
The results of different parameters used in our analysis are detailed in Table 3.

Table 3: **Decision Trees results.**
***MinSS**: Min Samples Split, **MinSL**: Min Samples Leaf,*
***TrnP**: Train Precision, **TestP**: Test Precision*

Test Size	**Criterion**	**Max Depth**	**Min SS**	**Min SL**	**TrnP**	**TestP**
80 %	gini	4	2	5	0.494	0.407
80 %	gini	4	50	5	0.476	0.407
80 %	gini	None	2	30	0.527	0.419
50 %	entropy	5	10	5	0.47	0.42
50 %	**entropy**	**5**	**60**	**5**	0.47	0.427
50 %	entropy	5	60	1	0.46	0.424

Concerning **Random Forest**, we did several experiments by considering different parameters, features, train-test split and cross validation with number of folds 10, then we validated our model on both training and test set for each model. We implemented Hyperparameter Tuning with cross validation randomized search to narrow down the range for each parameter we should look while building our model. Moreover, to choose the most relevant features for our classification model we studied the importance of each feature. Table 4 summarizes the selected results of the analysis.

Table 4: **Random Forest results.**
***TS**: Test Size, **MaxF**: Max Features, **MaxD**: Max Depth,*
***MinSS**: Min Samples Split, **MinSL**: Min Samples Leaf,*
***TrnP**: Train Precision, **TestP**: Test Precision*

TS	Criterion	N_estim.	MaxF	MaxD	MinSS	MinSL	Bootstrap	TrnP	TestP
80 %	entropy	110	Auto	110	10	1	False	0.99	0.56
80 %	gini	100	Auto	110	10	2	False	0.98	0.57
80 %	entropy	100	None	50	10	2	True	0.94	0.58
50 %	entropy	110	Auto	110	10	1	False	0.99	0.55
50 %	gini	100	Auto	110	10	2	False	0.98	0.57
50 %	entropy	100	None	50	10	2	True	0.94	0.56

5.4 Final Model

According to the results in Table 3 and 4 we should choose the model with the highest precision which is **Random Forest**, but we will use **Decision Tree** instead to improve the interpretability. The corresponding parameters for DT are bold values in Table 3.

6 Features Study

In this part, we decided to focus on the impact and possible relations between some specific features and the popularity class their members end up in.

6.1 Speaker Occupation

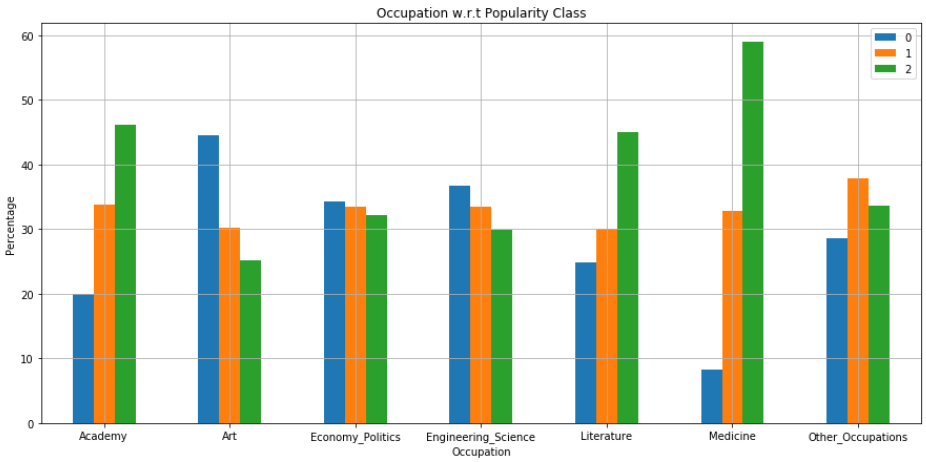

Fig. 4: **Percentages of Occupations in the three classes of popularity.**

Medicine: the bar chart shows that talks by a physician or a speaker in health and wellness professional would get a high number of views, comments and ratings than any other professional speakers. This could be because of an alarming

increase rate of chronic disease nowadays people care a lot about their health and used to seek for tips and advice online.

Academy: talks by academicians also had a high probability of getting high number of views, comments and ratings than other speakers. This capture the intuitive insight that most of the viewers could be students which were looking for inspirational and motivational talks.

Economy_Politics: chance of getting popular of a talk by an economist or politician were almost equi-probable. This may be due to existing political and economic ideology difference among humans. Also, it could be due to the seasonality of political and economic issues.

Literature and Art: talks from literature occupation were more popular compared to speakers from Art occupation such as graphic designer, film maker, chef, and cartoonists. This may be because professionals from literature such as journalists or poets have good soft skills and were able to engage the viewer easily. Whereas artists, such as graphic designers or chefs are more hard skill oriented most of the time.

6.2 Event type

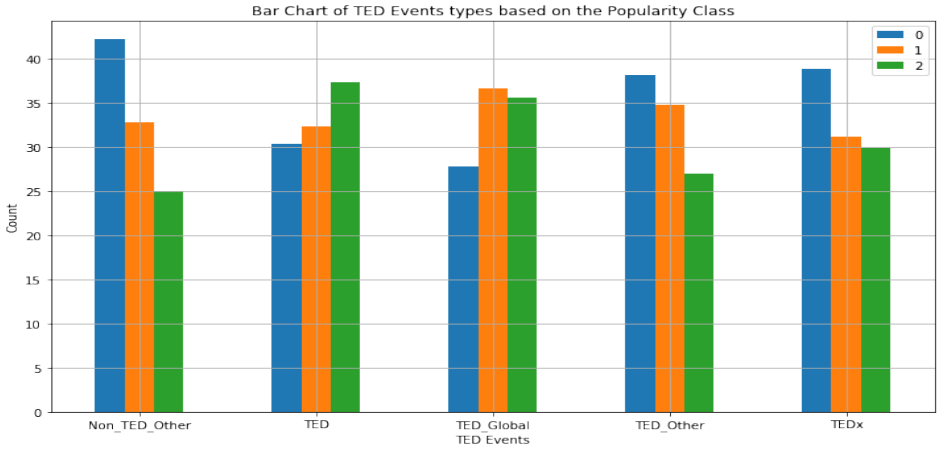

Fig. 5: **Percentages of each Event type in the three classes of popularity.**

This bar chart represents the percentage of each TED event type in the different classes of popularity. It means that for the official TED events (named TED) we overall have a bit more than 30% in popularity 0, around 32% in popularity 1, and around 37% in the highest popularity class. In the step prior to this one, we had 6 different classes. Five are displayed here and the sixth one, Non_TED_University, has been merged with Non_TED_Other because we found only five observations in it.

It is interesting to observe the distributions while taking the number of observations into account.

Table 5: **Number of talks divided by events.**

TED Event type	Non_TED_Other	TED	TED_Global	TED_Other	TEDx
Total	128	1065	464	422	471

The best cases are observed when the event type is either TED or TED Global because, even though we have more events of these two types (especially TED) we can still notice a big representation in the 2 highest classes of popularity, whereas TEDx, TED_Other and Non_TED_Other have more chances to end up in the class 0 than 1 and 2.

6.3 Title of the talk

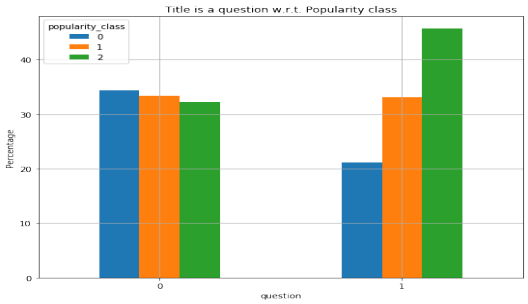

Fig. 6: **Distribution of the type of title (if it is a question or not) in the three classes of popularity.**

We created a new feature to check if the title is a question or not. Looking at the results comparing the popularity class, we can see that there exist correlation. Titles as question are more catchy, resulting in more popular videos.

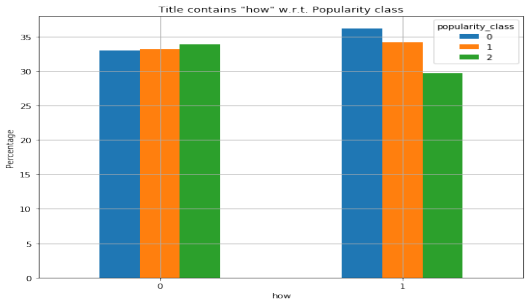

Fig. 7: **Distribution of the type of title (if it contains the word "how") in the three classes of popularity.**

We also created a new feature to check if the title contains the word "how" or not, because this word can help us detect educational speeches. Looking at the results comparing the popularity class, we can see that if the title contains the word, there exist negative correlation. Titles having "how" are contrary to our intuition, less catchy, resulting in less popular videos.

6.4 Tags

From the distribution of tags based on the different popularity classes, we observed that tags don't indicate how popular a talk is going to be, instead we just notice that there are more used tags and less used ones; meaning that the most used tags in each popularity class were the same.

Table 6: **Distribution of tags .r.t. Popularity Class.**

Tags/Popularity class	**0**	**1**	**2**
Culture	108	134	244
Technology	250	263	214
Science	201	179	187
Global issues	174	173	154
Business	85	123	140

6.5 Publication Day

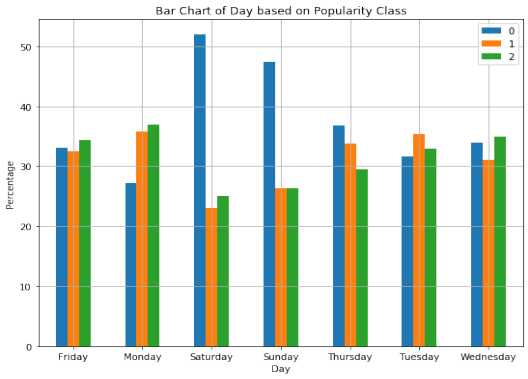

Fig. 8: **Bar chart according to the publication day w.r.t. Popularity class.**

According to the chart it is clear that publishing the video during weekend is a bad idea because it brings less popularity. On the other hand, days like Monday or Wednesday are better choices.

7 Model Modification

We will use only a decision tree to make our model easily "interpretable". We also dropped the "languages" feature from the considered variables in the decision tree because we know that besides the script in English that can be made before posting the talk, the transcripts are done after publication. Number of languages as said before, is positively correlated to popularity, and we understood from the metadata that the translations were made by the people who watch the talks. So

it cannot be considered when our target is the establishment of general guidelines to have a successful talk. For the same reason the definition of *funny_factor* has also changed not taking into account the funny ratings but only the number of "laughter" in the transcripts.

With these elements taken into account, we have this feature importance distribution:

Table 7: **Feature importance.**

Feature	**Importance**
Funny_factor	0.39
Art (Occupation)	0.13
Tags_trend	0.12
Duration	0.11
Monday (Day of publication)	0.09
Speaker trend	0.07
TEDx (Event)	0.06
TED (Event)	0.03

All the other columns' importance is equal to 0.
The Decision Tree we got with these elements is the following:

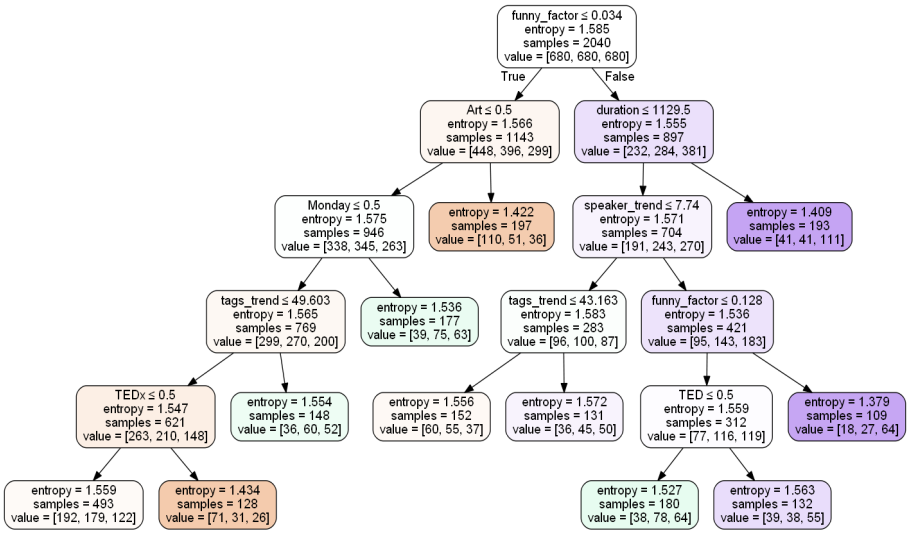

Fig. 9: **Decision Tree with the features considered.**

From these points, we will extract a set of recommendations to follow in the next section.

8 Recommendation System

Considering all the previous analysis and results, we can summarize the following points.

The best practices, intended for **TED Talks organisers**, to produce popular videos, are the following:

Theme and tags choice: Culture seems to be the kind of subject that will most likely lead to a more popular talk. It is also very important to check the trends of some tags before including them since they help making a video more popular when they are more trendy.

Type of holding event: It is obviously better when a TED Talk is actually held by TED. This element being independent from the will of the organisers, we suggest that, at least, if there is a sponsorship or partnership with TED, it should be labelled as TEDx rather than without it. As shown in Figure 5, The TEDx events score better in popularity than the Non_TED events, even if they are also featured in TED website.

Title of the talk: It is better when the title doesn't includes "How" and is formulated as a question.

Mood/features of the talk: The funnier is the talk, the more popular it gets.

Speaker choice: The best occupation seems to be Medicine. Although, Literature and an Academic position are also privileged, we can see in the DT that the only occupation that has a big influence on the classification is Art (that leads mostly to popularity 0, so before thinking about the best occupation to bring, it can be wise to think about which one not to bring (Art). Moreover, it is better to have a popular speaker (Trends).

Duration: The best duration was around 1100 seconds (18.3 minutes approximately).

Publication day on internet: Publish it preferably on Monday. Avoid the weekend.

9 Case Study

In this section we illustrate the correctness of our model on specific talks. This exercise helps us understand the algorithm, together with its limitations, offering an intuition of how different features contribute to the success of a particular talk.

Considering for example the speech *M' Bifo*, it is correctly classified as 0. In fact, it doesn't follow our recommendation system. Additionally, the speech *How books can open your mind* is correctly detected as 2 and we noticed that it fitted perfectly in our recommendation system and succeeded. In both cases, the value of the features allow us to detect the reality and therefore to predict correctly the result.

Unfortunately, this does not always work, as in the case of *How I swam the North Pole* that was identified as 2 but it's instead 0; or with *Meet the*

dazzling flying machine of the future that is 2 but detected as 0. Both of them don't follow the recommendation system and our model, working only with the feature values, wasn't able to understand the true quality of these videos.

We conclude that our model succeed in prediction of the true class of the majority of the speeches, and the recommendation system has an important role. Beside this, if the speaker is a great lecturer, he will reach the public and have a good feedback even without following our rules; and vice versa, even in the case of a very strict compliance, if the speaker doesn't have great presentation skills he won't create high popular content. There is also always an unpredictable factor on internet where videos can blow up and become famous for no reason, or because a famous person talked about... etc.

10 Conclusion

In this paper, our goal was to create a recommendation system to produce popular TED talks. To do this, we created a model capable of predicting video popularity prior to its publication, helping us understand what factors contribute to its success. In order to achieve that, we performed various feature modifications on the dataset available, combined with the creation of new features and consideration of another dataset (Google Trends).

An important challenge of our prediction task is that we have far more low popular videos than high popular ones, resulting in an imbalanced situation. The idea was to transform the problem from regression to classification, allowing us to use a Decision Tree, obtaining a well working model.

We found out that the funniness of a speech, also with speaker occupation and duration are very important features to consider when organizing a talk.

We hope our rules serve as starting point to design successful speeches and that our work could inspire further investigation in similar fields.

References

1. Rounak Banik: TED Data Analysis, https://www.kaggle.com/rounakbanik/ted-data-analysis. Last access: October 22, 2019.
2. GSD: Let's talk about TED Talks, https://www.kaggle.com/gsdeepakkumar/lets-talk-about-ted-talks. Last accessed 22 Oct 2019.
3. Adelson Araujo Junior: TED-Talks topic models, https://www.kaggle.com/adelsondias/ted-talks-topic-models. Last accessed 22 Oct 2019.
4. TED TALKS - LESSON WORTH SHARING, https://www.kaggle.com/ashishpatel26/ted-talks-lesson-worth-sharing. Last accessed 22 Oct 2019.
5. Tomer Eldor: Data Reveals: What Makes a Ted Talk Popular?, https://towardsdatascience.com/data-reveals-what-makes-a-ted-talk-popular-6bc15540b995. Last accessed 22 Oct 2019.
6. Kaggle TED Talk page, https://www.kaggle.com/rounakbanik/ted-talks. Last accessed 22 Oct 2019.
7. The project's GitHub repository, https://github.com/thizirie/BigDataAnalyticsProject_2019_2020. Last accessed October 23rd, 2019.