Predicting TED Talks Performances

Imperial College London MSc Business Analytics 2017-18 Report

AUTHOR: Xiaoyan Zhou

CID: 01384662

Word count: 4623

Abstract

When trying to predict the TED Talks performances, it is vital to use the right variables to do the prediction. This report uses two datasets of TED Talks videos released in Kaggle to find the important variables affecting the performance of TED Talks. Methods: 9 variables are generated and put into the random forest algorithm to get the importance of them. Negative binomial model is used later to predict the views of TED Talks with important variables selected. Results: This research successfully finds out some important variables having influences on the views of TED Talks, including theme, incitement, description positiveness, in-degree of recommendation network, TED event type and gender of the main speaker. Both TED Talks speaker and TED.com can utilize the results and the model built to improve the performance of TED Talks, and use the relevant managerial recommendation as a guidance.

Table of Content

1. Introduction

When talking about using speeches to spread ideas and change the world, TED Talks is one of the most successful formats in this area. TED Talks are short and powerful talks. Established in 1984 as a conference where Technology, Entertainment, and Design converged, TED now involves almost all topics — from science to business to global issues — in over a hundred languages. Except for the annual TED conferences and TED Global conference, it also runs TEDx events to help share ideas in communities all over the world (Early, 2017).

Today, TED Talk is changing the world as more and more people are watching them and be influenced by them. In 1991, the total number of views of TED Talks was merely three hundred thousand, but in 2016, the number changed to more than thirty-six million, 100 times more than that in 1991.

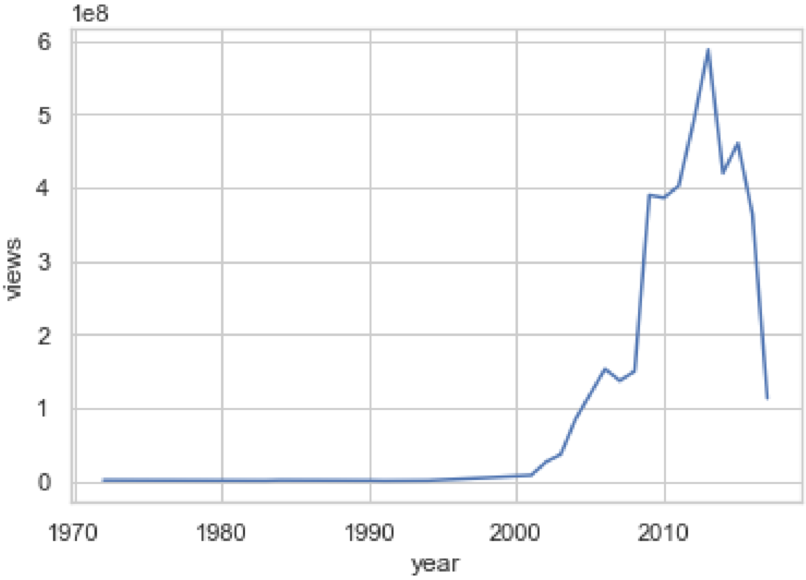


Figure 1.1 TED Talks view from 1991 to 2017

TED Talk producers are always trying to improve the performances of TED Talks. Most of them train hard and even pay consultants or dramaturges to coach them or write the speech draft (Aarons-Mele, 2018). Research shows that in order to improve the engagement of audiences and breach the expert/audience barrier, the speaker of TED Talks improves proximity to establish an 'alignment' between themselves and the speaker (Scotto Di Carlo, 2014). However, the audiences in TED.com is different from other video platforms. A research doing the content analysis of comments left on TED website and the YouTube platform has found that commenters were more likely to engage with the talk content on the TED.com, while commenters tend to talks about the speakers' characteristics on YouTube (Tsou et al., 2014). This phenomenon implies that it is important to use the data in TED.com itself to gain insights for TED talks performances.

It would be very helpful to find out what are the important variables affecting the views of TED Talks so that the TED Talks producers can better organize their talks and TED.com can improve the demonstration of the talks by optimizing relevant attributions.

This report aims to find the important variables when trying to predict TED Talks performances in TED.com. Section 2 will describe the data used in this report and disclose some insights on the original dataset. Section 3 explains how and why generating 9 sorts of variables, including duration, event type (TED or TED Global), description positiveness, incitement, and main speaker gender, the major theme of the talks, and network attributes in the TED.com (in degree, eigenvector centrality). After generating the variables, Variance inflation factor is used in this part to detect sources of multicollinearity and random forest algorithm is used to test the importance of the 9 variables. In section 4, negative binomial regression is used to find the relationship between the dependent variables and independent variables. Also, a final model predicting the views is presented. The interpretation of the model is also provided in this part, from which the marginal effect of each variable can be seen. Section 5 illustrates the conclusions of this research and give managerial recommendations to both TED Talks producers and TED.com, giving them guidance on how to effectively improve the views. The final section of this report will discuss the limitation of this research and come up with some future research directions.

2. Data description and insights

2.1 Data description

The datasets used are download form Kaggle, a platform that provides data source and held predictive modeling and analytics competitions. The TED main dataset contains information about TED Talks audio-video uploaded to the official TED.com website until September 21st, 2017. This paper research the TED Talks released in the year between 2007 and 2017. The TED transcripts dataset contains the transcripts for all talks. The information contains in the TED main dataset can be seen in the Table 2.1.1 (Rounak, 2017). The TED transcript dataset contains the transcript and URL information for TED Talks.

Columns	Descriptions
comments	The number of comments of the video.
description	A brief introduction of what the talk is about.
duration	The total seconds of the talk.
event	The TED/TEDx event where the talk took place.
film_date	The Unix timestamp[1] of the filming.
languages	The number of languages that the talks is available.
main_speaker	The first name speaker of the talk.
num_speaker	The number of speakers in the talk.
publish_date	The Unix timestamp when the talk was published in TED.com
ratings	Ratings given to the talk, including the name of the ratings (Funny, Beautiful, Obnoxious, etc.) and the count of each sort of rating.
related_talks	Recommended talks to watch next.
speaker_occupation	The occupation of the first name speaker.
tags	The themes associated to the talk.
title	The title of the talk.
urls	The URL of the talk.
views	The number of views on the talk.

Table 2.1.1 Description of TED main dataset

The total number of unique TED Talks from 2007 to 2017 in the main dataset is 2324. After generating all the necessary variables and filtering NA values, the total number of observations is 2247.

2.2. Insights

(1) The popular theme of ted talks

There are 416 themes in total involved in all the TED Talks, and the most popular 8 themes (based on the count of each theme) chosen by speakers are technology, science, global issues, TEDx, cultures, designs, business, entertainment. From Figure 2.2.1, it can be seen that the total number of theme technology appeared more than 600 times.

[1] Unix timestamp: The number of seconds that have elapsed since January 1, 1970 00:00 UTC.

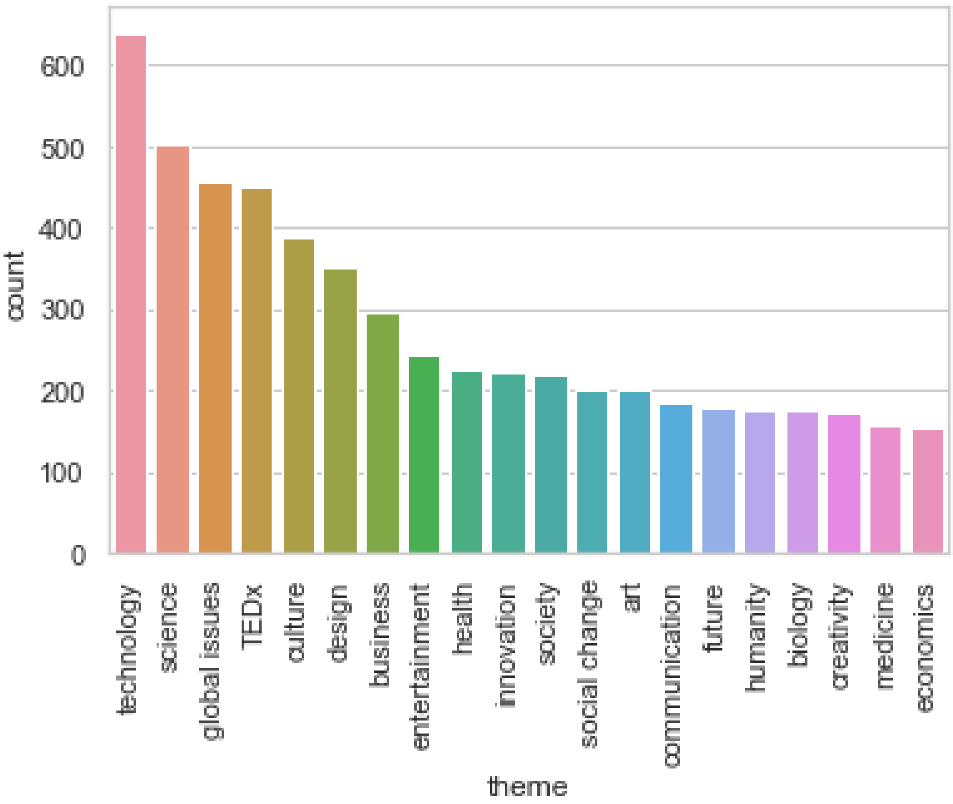


Figure 2.2.1 Theme Popularity- Count

When looking at the views of different themes, the most popular 8 themes are the same with the 8 themes chosen by the speakers, while the ranks change slightly. From Figure 2.2.2, the top 3 themes are technology, culture, and science. However, the most popular 20 themes regarding views and counts are not the same, which means that some themes are popular among speakers, but they are not so favored by the audiences.

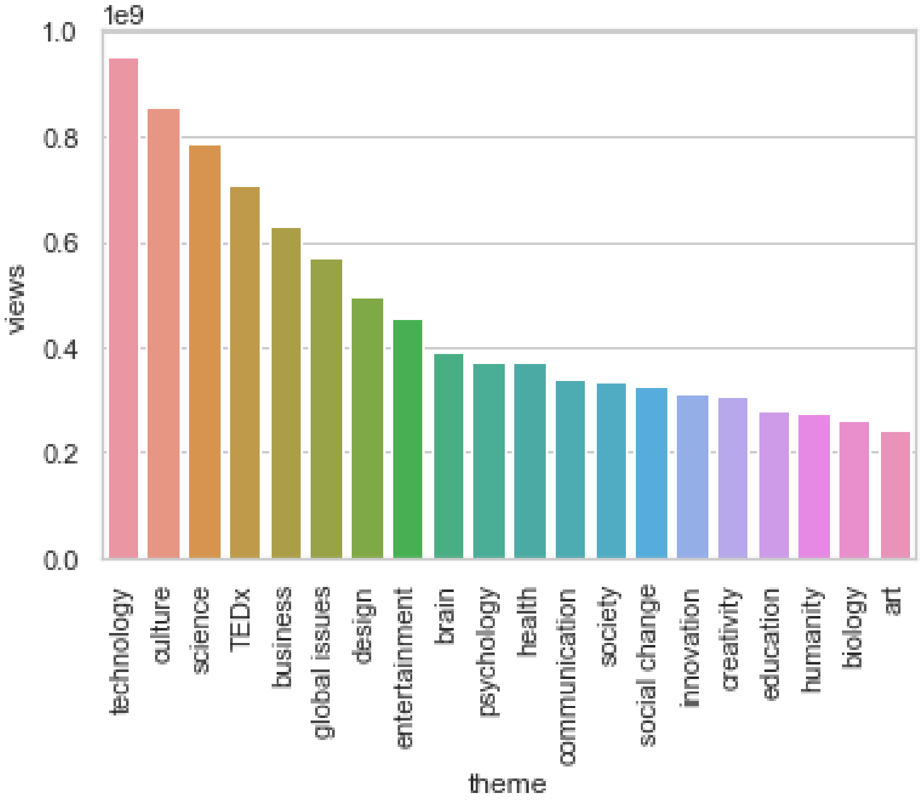
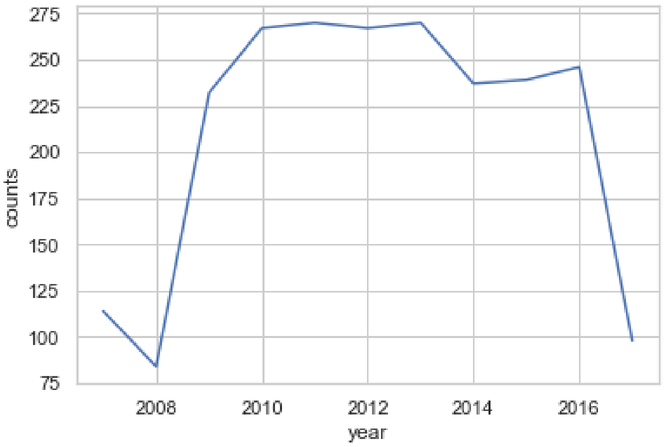


Figure 2.2.2 Theme Popularity- Views

(2) The number of TED Talks and views in each year

Figure 2.2.3 and figure 2.2.4 shows the number of TED Talks released in TED.com and total views from 2007 to 2017. Since the data only records until September 2017, it is understandable to see that both the counts and views are low in 2017. Before 2009, the number of TED Talks released each year are relatively small, while after 2009, the number of TED Talks released each year are above 225. TED Talks views increase quickly after the year 2008, peaking in the year 2013, and start to decrease afterward.



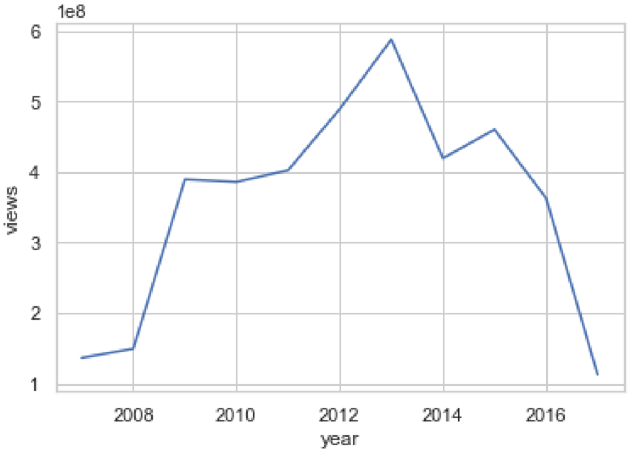


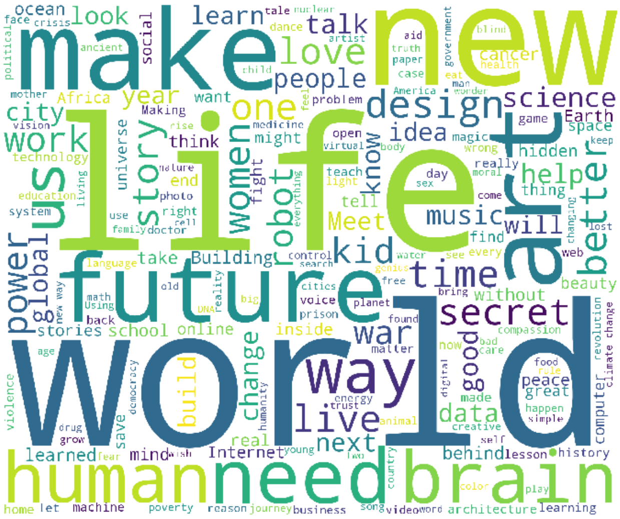
Figure 2.2.3 TED Talks counts by year

Figure 2.2.4 TED Talks views by year

(3) Word cloud of title

The word "life" is the most popular word when TED Talk speakers organize the title of their videos. The other most popular world in titles are "world", "new", "make", "future", "art". It seems that the TED Talks speakers tend to concern about the influencing the life or the world.

The word cloud for the 100 TED Talks videos with the largest views demonstrate that the word appears most is "Power". The other most popular world I titles are "make", "work", "new", "way", "leaders". Those words are more interpersonal.



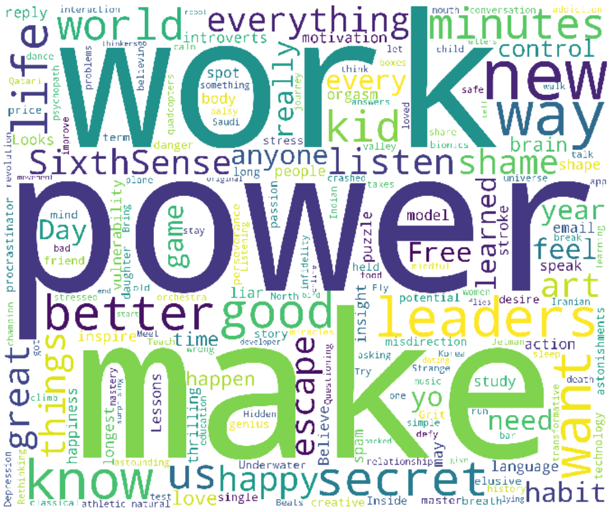


Figure 2.2.5 World Cloud for all TED Talks Title

Figure 2.2.6 World Cloud for 100 Most Viewed TED Talks Title

In this part, it can be seen that one major difference between the speaker and the audience is the theme that they favor. Some themes are very popular among speakers, but they are not enjoying the same level of popularity among the audiences. This phenomenon can also be seen in the word clouds. In the next section, this situation will be considered and relevant variables will be generated.

3. Variable generation and evaluation

3.1 Variable generation

The independent variables generated are the variables that potentially have impacts on the performance of TED Talks. The “views” column is a proxy for the performance of TED Talks. This section will give detail explanations of each variable, including the methodology of variable extraction, justification, and hypothesis.

(1) Duration

Methodology of Variable Extraction

The variable is provided in the original dataset.

Justification

The duration of the speech is important in regarding the views. A very long speech cost audiences too long time, which may hamper the interest of clicking the video.

Hypothesis of Variable

The duration of a TED Talk has a negative influence on the number of views.

(2) Year

Methodology of Variable Extraction

The publish date of TED Talks is provided in the ted_main dataset, in the format of Unix timestamp. The year is extracted and transformed into dummy variables. The year 2007 is excluded as this indicator can be inferred.

Justification

The year dummy variables are important as views should be closely related to the year the video formally filmed, which is also the year of the TED event. Views of a video will increase by year as audiences naturally grow by year, as the frequency of people using the Internet is also increasing from 2007 to 2017. In 2007, there were 1367 million people using the internet, while the number change to 3578 million in 2017 (“Number of Internet Users”, 2018). Since TED.com is an online website, it is highly likely that the more recent the year is, the more likely the talk is viewed by more people.

Hypothesis of Variable

The year dummy variable will have a positive influence on the views. The nearer the year until 2017, the coefficient will be larger.

(3) Description Positivity

Methodology of Variable Extraction

The description of each talk is provided in the ted_main dataset. After removing the English stop words from the description, the description positivity is generated by the following equation:

$$\text{positivity} = \frac{\text{positive words}}{\text{positive words+negative words+neural words}}$$

Justification for Significance of Variable

TED Talks is about spreading good ideas. The more positive the description is, the more likely that people are inspired and get interested in the video, thus they will start to view it.

Hypothesis of Variable

The description positiveness has a positive influence on the views.

(4) Ted Global and TED

Methodology of Variable Extraction

If the TED Talks happens in Ted Global event, then the value will be 1, otherwise 0. If the TED Talks happen in the TED Conference, then the value will be 1, otherwise 0.

Justification

TED Global is a conference that celebrates human ingenuity by exploring ideas, innovation, and creativity from all around the world, while the TED has wider topics and be held on the west coast of North America (“TED Conference”, n.d.). Comparing to TEDx event, TED and TED Global might have a different impact as their selection standard is even higher.

Hypothesis of Variable

The TED Global and TED variable will have a positive influence on the views.

(5) Main Speaker Gender

Methodology of Variable Extraction

Applying the gender detector in python to the first name of the main speaker to get the gender. It has female, male, and neutral. Then transform them into dummy variables. The neutral dummy variable is excluded as it can be inferred by the other two variables.

Justification

Researches have shown that people tended to be more emotional in comments when the TED Talk speaker was a woman (Tsou et al., 2014). This phenomenon reveals that women speaker can better increase the engagement of the audience during the speech. Consequently, it might also make the talks more attractive and gain more views.

Hypothesis of Variable

The Female main speaker has a positive influence on the views.

（6）20 theme dummy variables

Methodology of Variable Extraction

The 20 topic dummy variables are the 20 most popular themes chosen by the TED speaker, including 'technology', 'science', 'global issues', 'culture', 'design', 'business', 'entertainment', 'health', 'innovation', 'society', 'art', 'social change', 'future', 'communication', 'biology', 'creativity', 'humanity', 'economics', 'environment', and 'medicine'. The “TEDx” theme is excluded as it does not represent the theme of each talk. Each observation will have 20 columns indicating whether it belongs to them or not. If one TED Talks have none of the 20 tags, all 20 columns will have value 0.

Justification

The 20 topic dummy variables created here is to test whether some themes are favored by audiences, while some are not, even though they are all favored by the speakers. On the one hand, the audiences might love certain themes, leading to a larger number of views than those without the tags. On the other hand, some themes might not as attractive as the others, getting fewer views.

Hypothesis of Variable

Some of them have positive influences on the views, and some of them have negative influences on the views.

(7) Incitement

Methodology of Variable Extraction

The Incitement variable shows how many times the speaker triggers the audiences to laugh or applause during the talk. The transcript of each talk uses "(Laughter)" and "(Applause)" to label every time when the audiences laugh and applause. Counting the number of "(Laughter)" and "(Applause)" in the transcript can be used to represent the incitement of each TED Talks.

Justification

Incitement can be used to measure how fascinating or humorous the audience think of the speech. If many laughter or applause appear during one talk, then the talks must be fully catching attendances' attention and delivering something interesting. Such a video can also grasp the attention of the online audiences and earn larger views.

Hypothesis of Variable

The incitement has a positive influence on the views.

(8) In-degree

Methodology of Variable Extraction

In-degree is the number of inward directed graph edges from a given graph vertex in a directed graph (Venkatakrishnan, 2016). After using the related talks to produce a network of TED Talks, the in-degree of each TED Talks can be calculated by counting the number of TED Talks recommending the targeted TED Talks to show next.

Justification

The in-degree variables is created to measure how many other TED Talks recommends this TED Talk to play next at the end. With more TED Talks' recommendation, a TED Talk are more likely to be shows up in audiences' screen, thus has a higher odd to be played.

Hypothesis of Variable

The in-degree has a positive influence on the views.

(9) Eigenvector centrality

Methodology of Variable Extraction

The eigenvector centrality of each TED Talk can be calculated using the eigenvector_centrality_numpy function in networkx packge in Python. The network of the TED Talks can be represented by an adjacency matrix. The eigenvector of the largest eigenvalue given by this matrix can be used to calculate the eigenvector centrality (Franks at el. 2014).

Justification

In graph theory, eigenvector centrality is a measure of the influence of a node in a network. Scores will be assigned to each node basing on that high scores will be assigned to nodes that connect to nodes that with high scores, while low scores will be assigned to nodes that connect to low scores nodes. A high eigenvector score means that a node is connected to many nodes whom themselves have high eigenvector centrality (Newman, 2010). A TED Talk with high eigenvector centrality means that it is recommended by many other TED Talks who are recommended by many other TED Talks, which means it has a higher odds of getting exposed and viewed.

Hypothesis of Variable

The eigenvector centrality has a positive influence on the views.

3.2 Variable evaluation

After generating the variables interested, it is important to check whether multicollinearity exists among those variables. Since multicollinearity will hamper the regression process, leading to incorrect prediction results, multicollinearity should be excluded before putting the variables into the regression model.

Once the multicollinearity is excluded, random forest algorithm can be used to get the importance of each variable.

3.2.1 Detecting Sources of Multicollinearity

the variance inflation factor (VIF) is the reciprocal of tolerance. It measures the how severe the multicollinearity in the model (O'brien, 2007). From Appendix A, it can be seen that the VIF of all variables created are below 10, according to Hair et al. (1995), this level of VIF is acceptable. Thus, all variables will be kept to find the feature importance of them in the random forest algorithm.

3.2.2 Feature importance

After putting the dependent variables and independent variables into the random forest algorithm, the Inc node purity can be used to measure the feature importance. At each split, the algorithm calculates how much this split reduces node impurity (for regression trees, indeed, the difference between RSS before and after the split). This is summed over all splits for that variable, over all trees. The larger the importance, the more important the variable is.

Figure 3.2.2.1 shows that in the random forest, the incitement variable is considered most important in splitting the node, with the importance value of 0.35. The importance of duration and description positiveness is also larger than 0.1. The importance of "business", "eigenvector centrality", "in degree", "2013", "culture", "male", "2009", "2014", "2015", "female", "global issues", "TED" and "TED Global" are all between 0.01 to 0.1. The importance of the rest of the variables is all lower than 0.01.

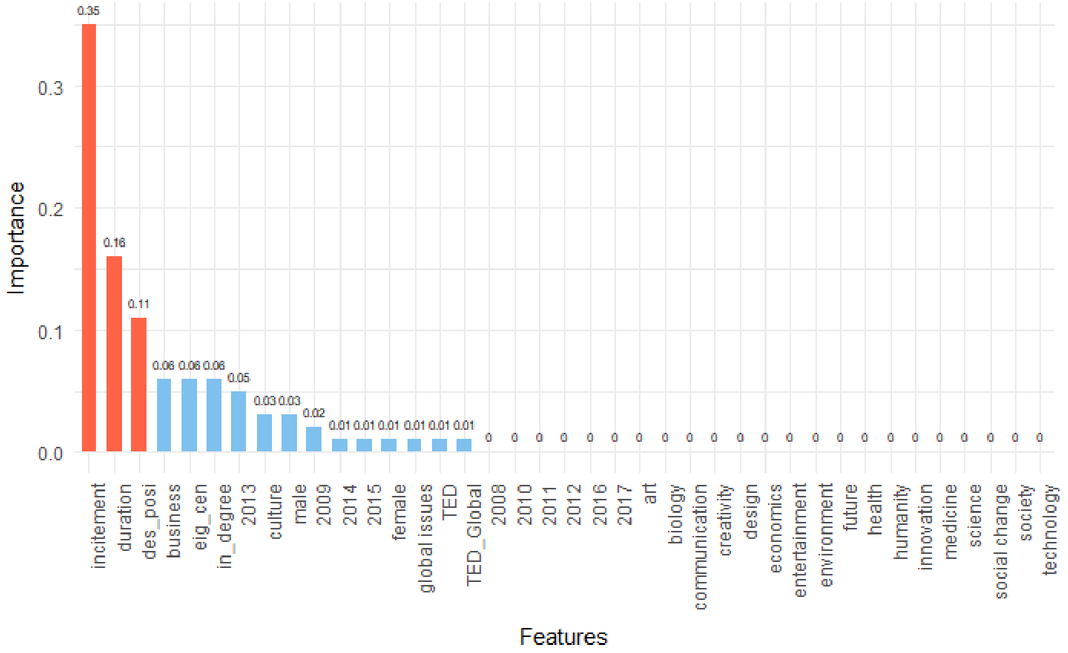


Figure 3.2.2.1 Feature Importance in Random Forest Algorithm

5. Negative Binomial Regression model

In this part, Negative Binomial regression is used to predict the views, because the dependent variable is count data, and it is heavily right-skewed. It is also because the views are over dispersed, having way higher variance (5469585245594) compared to the mean (1722348).

5.1 Model building

Firstly, the whole dataset is split into a training set and a test set, with the ratio of 7:3. After putting all variables generated into the negative binomial regression model and gradually removing the insignificant variables, the final model for predicting the TED Talks performances is presented in Figure 5.2.1. (The full model can be seen in Appendix B)

The AIC for the final model is 48020, deviance residual is 663 and the log-likelihood is -23984.

5.2 Model interpretation

The negative binomial regression model for predicting TED Talks views is:

$$\begin{aligned}\log(\text{views}) = {} & 13.15 + 0.15\text{TED\_GLOBAL} + \ 0.21\text{TED} + 0.04\text{Incitement} + 1.39\text{des\_posi} \\ & + 0.04\text{in\_degree} - 3.88\text{eig\_cen} - 0.34\text{global\_issues} + 0.22\text{culture} + 0.34\text{business} \\ & + 0.21\text{health} - 0.24\text{art} - 0.20\text{enironment} - 0.30\text{medicine} \ + 0.19\text{year\_2008} \\ & + 0.27\text{year\_2009} + 0.2\text{year\_2010} + 0.33\text{year\_2011} + 0.47\text{year\_2012} \\ & + 0.71\text{year\_2013} \ + 0.58\text{year\_2014} + 0.61\text{year\_2015} + 0.34\text{year\_2016} \\ & - 0.01\text{year\_2017} + 0.13\text{male} + 0.16\text{female}\end{aligned}$$

It can be seen from Figure 5.2.1 that the p-value for variables “TED_Global”, “TED”, “Incitement”, “des_posi”, “in_degree”, “eig_cen”, ”global_issues”, “culture”, “business”, “health”, “art”, “environment”, “medicine”, “year_2009”, “year_2011”, “year_2012”, “year_2013”, “year_2014”, “year_2015”, “year_2016”, ”female” are all statistically significant, using 95% confidential level, thus they are all significantly influencing the dependent variable, views.

The marginal effect of “TED_Global” is that compared to the views of TED Talks that happened in the TEDx event, the views of TED Talks happened in the TED Global event will increase by 15%.

The marginal effect of “TED” is that compared to the views of TED Talks that happened in the TEDx event, the views of TED Talks happened in the TED event will increase by 21%.

The marginal effect of “Incitement” is that one unit increase in the laughter or applause during the talk will increase the views by 4%.

The coefficient of description positiveness is positive, which means that the more positive the description is, the more views will be.

The marginal effect of indegree is that one unit increase in the indegree of the TED Talks will increase the views by 4%.

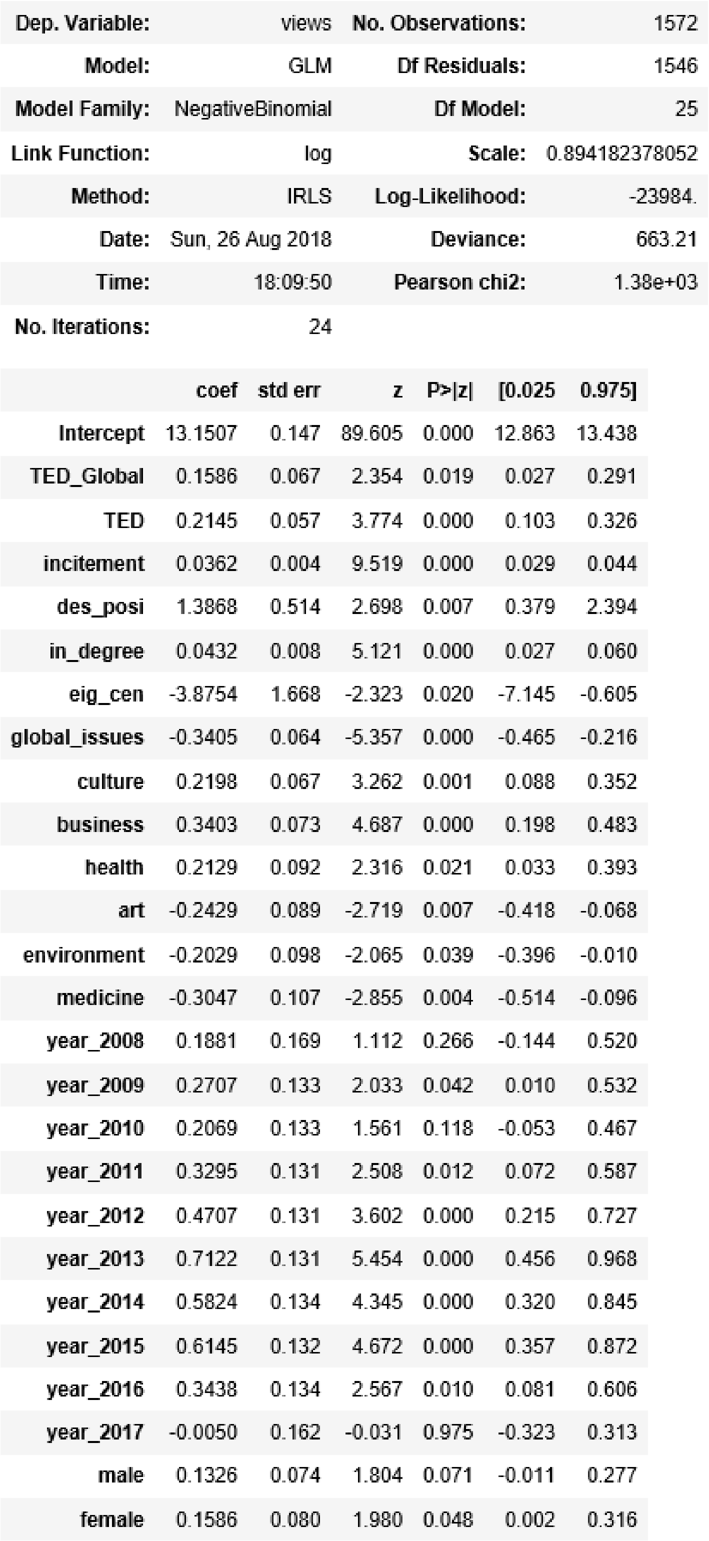
The coefficient of eigenvector centrality is negative, which indicates that when linking to the talks with many linkages can have a negative impact on the views.

The coefficients of “global issue”, “art”, “environment”, “medicine” are all negative, which means that even though they are very popular topics among the producer of TED Talks, they are not necessarily considered attractive by the audiences.

The coefficients of “culture”, “business”, “health” are all positive, which indicates that these three topics are popular among both audiences and speakers, having higher views.

The coefficients of all year dummy variables are positive, which shows that compared to the views of TED Talks happened in 2007, those happened in later year enjoys a higher amount of hit.

The coefficient of the female is positive and the variable “female” is statistically significant, while the p-value of “male” is not statistically significant, thus it means that TED Talks given by female speaker is more popular than those given by male speaker or neutral speaker.



Dep. Variable:	views	**No. Observations:**	1572
Model:	GLM	**Df Residuals:**	1546
Model Family:	NegativeBinomial	**Df Model:**	25
Link Function:	log	**Scale:**	0.894182378052
Method:	IRLS	**Log-Likelihood:**	-23984.
Date:	Sun, 26 Aug 2018	**Deviance:**	663.21
Time:	18:09:50	**Pearson chi2:**	1.38e+03
No. Iterations:	24		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.1507	0.147	89.605	0.000	12.863	13.438
TED_Global	0.1586	0.067	2.354	0.019	0.027	0.291
TED	0.2145	0.057	3.774	0.000	0.103	0.326
incitement	0.0362	0.004	9.519	0.000	0.029	0.044
des_posi	1.3868	0.514	2.698	0.007	0.379	2.394
in_degree	0.0432	0.008	5.121	0.000	0.027	0.060
eig_cen	-3.8754	1.668	-2.323	0.020	-7.145	-0.605
global_issues	-0.3405	0.064	-5.357	0.000	-0.465	-0.216
culture	0.2198	0.067	3.262	0.001	0.088	0.352
business	0.3403	0.073	4.687	0.000	0.198	0.483
health	0.2129	0.092	2.316	0.021	0.033	0.393
art	-0.2429	0.089	-2.719	0.007	-0.418	-0.068
environment	-0.2029	0.098	-2.065	0.039	-0.396	-0.010
medicine	-0.3047	0.107	-2.855	0.004	-0.514	-0.096
year_2008	0.1881	0.169	1.112	0.266	-0.144	0.520
year_2009	0.2707	0.133	2.033	0.042	0.010	0.532
year_2010	0.2069	0.133	1.561	0.118	-0.053	0.467
year_2011	0.3295	0.131	2.508	0.012	0.072	0.587
year_2012	0.4707	0.131	3.602	0.000	0.215	0.727
year_2013	0.7122	0.131	5.454	0.000	0.456	0.968
year_2014	0.5824	0.134	4.345	0.000	0.320	0.845
year_2015	0.6145	0.132	4.672	0.000	0.357	0.872
year_2016	0.3438	0.134	2.567	0.010	0.081	0.606
year_2017	-0.0050	0.162	-0.031	0.975	-0.323	0.313
male	0.1326	0.074	1.804	0.071	-0.011	0.277
female	0.1586	0.080	1.980	0.048	0.002	0.316

Figure 5.2.1 Summary for the final Negative Binomial Model

6. Conclusion and managerial recommendations

6.1 Conclusion

When predicting the TED Talks views, it is vital to find out which variables can significantly influence it. This research has studied the attributes (duration, event, description positiveness, incitement, and main speaker gender) of different TED Talks, the theme of the talks, and network attributes in the TED.com (in degree, eigenvector centrality). After putting those variables into the negative binomial regression model, according to the significant level and coefficient of each variable, it can be concluded that whether the variables have impacts on the views or not.

Firstly, the duration of a TED talks does not affect the views when controlling other variables. However, the event of the TED Talks does have a different influence on the views. People love the annual TED conference most and they also put much attention to the TED Global event, when comparing to the TEDx event. The description positiveness and incitement of TED Talks can influence the views too. The audience tends to play the videos with a positive description, and they love to see the TED Talks that can cause more laugh and applause. Besides, the views of TED Talks given by female speakers is higher than the views of TED Talks given by the male speaker.

Secondly, when looking at the theme of the talks, it can be found out that some themes are particularly favored by the audience, including “culture”, “business”, and “health”, while the talks with theme “global issues”, “art”, “environment”, and “medicine” gain fewer views than the others,

Lastly, the network attributes of the TED Talks in the website can also have impacts on the views. Those TED Talks with more other talks’ recommendation to play next tend to gain more views. However, the TED Talks have a high eigenvector centrality cannot attract more views, this phenomenon is counterintuitive, needing further study.

6.2 Managerial recommendations

For the TED Talk speaker, before they organize a talk, it is recommended to consider the topics that are favored by the audiences or make the talk more relevant to those popular topics. For example, the “medicine” theme is not as popular as the “health” theme, but the speaker can add example addressing how certain medicines can help to improve health, making it more attractive.

When the topic is settled and the speaker needs to organize the speech, it is recommended to articulate the expression, causing more laugh or applause during the speech. This can definitely inspire the atmosphere and make the talk more enjoyable. Once the talk is filmed and ready to be published in the TED.com, the speaker can also give more thought to the description of the video. It would be better to use more positive words to describe the talk.

For TED.com, if they want to increase the views of the TED videos, it is recommended to expand the watch next list. The current watch next list has 6 videos, so it still has rooms to expand. Besides, TED can also build a recommendation system, taking the attributes of the TED talks and the preferences of customers into consideration at the same time to improve the views. Using a dataset of user profiles and the TED Talks details, Nikolaos and Andrei compare several methods for the recommendation of lectures form the TED.com and proposed a combined method to improve the performance (Pappas, N., & Popescu-Belis, A., 2013).

The TED.com can also make some effort to promote the TEDx videos. According to the result of the model, TED and TED Global talks enjoy a larger amount of views than TEDx videos. The quality of TEDx talks are not necessarily poorer than TED or TED Global talks, but TEDx event does not gain as much attention as TED or TED Global event. TED.com can do a survey to find out the reasons behind this phenomenon and react correspondingly.

Last but not the least, from 2007 to 2017, the total number of the female main speaker is only 649, while the male main speaker is 1307, with 291 speaker's gender is neutral according to the gender guesser algorithm. As the female speaker seems to be more popular, it is recommended to encourage and introduce more female speaker to present the talks.

Both the producer and the TED.com can use the negative binomial model trained in this report to predict views for a talk before publishing it. Once they get the estimation for the views, they can adjust the talk by improving some metrics to attract the potential audience.

7. limitation and future research

When trying to find the key variables to predict the performances of TED Talks, there are five main limitations exists in this research, and researches can pay attention to these problems and help to tackle them in further studies.

In this research, the views of TED Talks is considered as a proxy of the TED Talks performances. However, this is just one aspect of the performances of the TED Talks. There are many other metrics measuring the performances of TED Talks. For example, the positiveness of comment under each TED Talk can also be such a metric, representing how satisfied the audiences feel after watching the video. By accessing how positive the comments are, a numeric dependent variable called “rating positiveness” can be generated. But researchers should be careful when dealing with this variable, as it could be highly correlated with the views. Another proxy variable for TED Talks performances is how many audiences actually change their thoughts and behavior after watching one TED Talk. This variable can be named “effectiveness”, and it can measure the power of one TED Talk. Unfortunately, this data is missing. In the future research, it is recommended to do a survey among audiences to get this data.

When trying to predict views of TED Talks, information of the audiences is also interesting, but they cannot be generated in the given datasets. Different kind of audiences can have different preference on the themes of TED Talks. It would be helpful to generate variables representing such information, including the average income of audiences, the average age of audiences, etc.

Another thing needs to mention is that the gender variables generated in this report are done by the gender guesser algorithm in Python. Since the gender is generated according to the probability of being male or female or neutral, it is not 100% correct. In order to get more accurate results, further data including the main speakers' gender is required.

The eigenvector centrality has a negative influence on the TED Talks views, this result is counterintuitive. It shows that a video that recommended by videos that get more recommendations tends to have a lower amount of views. The reason behind this might be that TED.com tend to recommend those unpopular TED Talks after one TED Talk is finished playing. However, further research is needed to figure out the causal relationship.

Though this report focuses on finding the important variables in predicting the TED Talks performances, it is also important to actually implementing the result of the variables into suitable models and predict for the views. Due to the limitation of the words, this report only uses the random forest to test the importance of features and implements the negative binomial regression model, but there are many other machine learning models can be used. One example is that when trying to predict the views, K nearest neighboring methods can be used, as this model can match similar TED Talks and predict for the views according to their attributes.

References

Aatons-Mele, M. (2018). The Myth of The TED Talk. [online] forbes.com. Available from: https://www.forbes.com/sites/morraaaronsmele/2018/01/12/the-myth-of-the-ted-talk/#a06654265f42 [Accessed 29th Aug, 2018].

Early, C. (2017). What is a TED Talk? [online] bt.com. Available from: http://home.bt.com/lifestyle/what-is-a-ted-talk-11364177580250 [Accessed 29th Aug, 2018].

Franks, H., Griffiths, N., & Anand, S. (2014). Learning agent influence in MAS with complex social networks. Autonomous Agents and Multi-Agent Systems, 28(5), 836-866.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L. & Black, W. C. (1995). Multivariate Data Analysis (3rd ed). New York: Macmillan.

Number of Internet Users Worldwide From 2005 to 2017 (In Millions). (2018). [online] Statista.com. Available from: https://www.statista.com/statistics/273018/number-of-internet-users-worldwide/ [Accessed 29th Aug, 2018].

Newman, M. (2010). Mathematics of networks: An introduction to the mathematical tools used in the study of networks, tools that will be important to many subsequent developments. In Networks(p. Networks, Chapter 6). Oxford University Press.

O'brien, R. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673-690.

Pappas, N., & Popescu-Belis, A. (2013). Combining content with user preferences for TED lecture recommendation. Content-Based Multimedia Indexing (CBMI), 2013 11th International Workshop on, 47-52.

Rounak B. (2017). TED Data Analysis. [online] Kaggle.com. Available from: https://www.kaggle.com/rounakbanik/ted-data-analysis [Accessed 15th Aug, 2018].

Tsou A, Thelwall M, Mongeon P, Sugimoto CR (2014) A Community of Curious Souls: An Analysis of Commenting Behavior on TED Talks Videos. PLoSONE 9(4): e93609.

TED Conference. (n.d.). [online] TED.com. Available from: https://www.ted.com/attend/conferences/ted-conference [Accessed 29th Aug, 2018].

Scotto Di Carlo, G. (2014). The role of proximity in online popularizations: The case of TED talks. Discourse Studies, 16(5), 591-606.

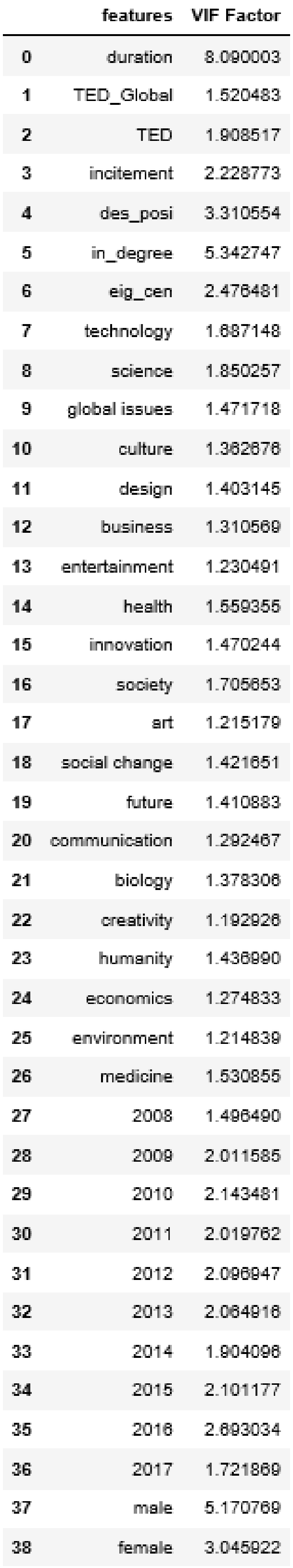
Venkatakrishnan, R. (2016). What is the indegree and outdegree of a graph? [online] quora.com. Available from: https://www.quora.com/What-is-the-indegree-and-outdegree-of-a-graph [Accessed 29th Aug, 2018]

Data sources

Rounak B. (2017). Data about TED Talks on the TED.com website until September 21st, 2017. [online] Kaggle.com. Available from: https://www.kaggle.com/rounakbanik/ted-talks [Accessed 29th Aug, 2018].

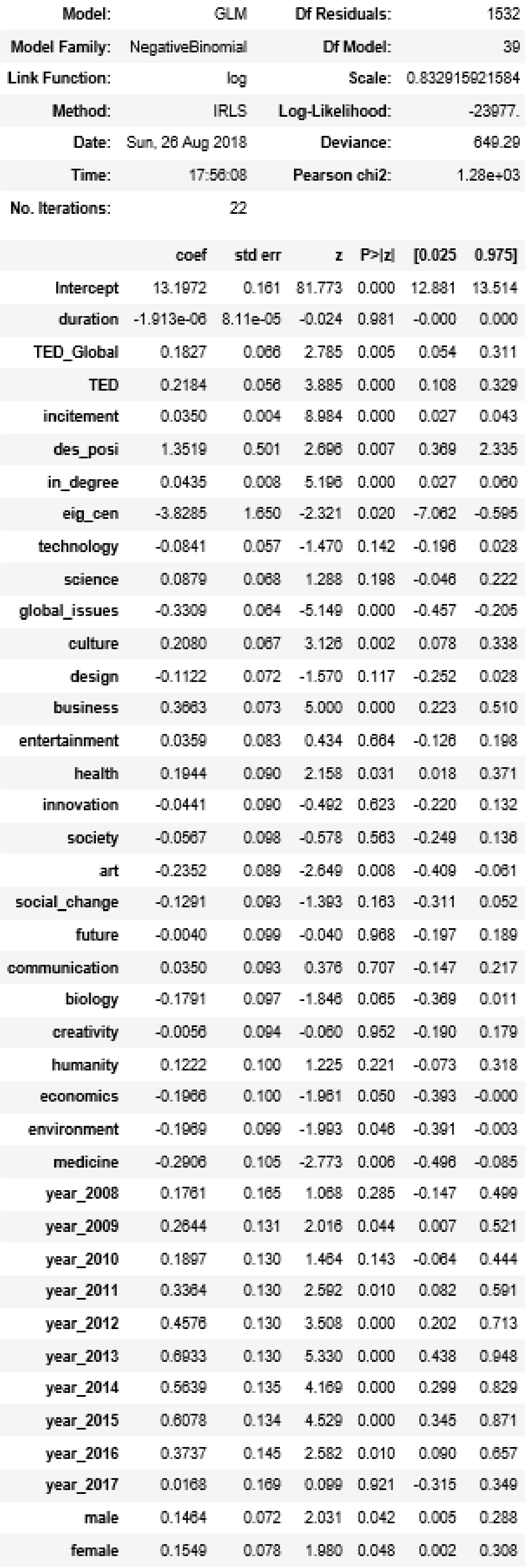
Appendices

A. VIF for Variables created



	features	VIF Factor
0	duration	8.090003
1	TED_Global	1.520483
2	TED	1.908517
3	incitement	2.228773
4	des_posi	3.310554
5	in_degree	5.342747
6	eig_cen	2.476481
7	technology	1.687148
8	science	1.850257
9	global issues	1.471718
10	culture	1.362676
11	design	1.403145
12	business	1.310569
13	entertainment	1.230491
14	health	1.559355
15	innovation	1.470244
16	society	1.705653
17	art	1.215179
18	social change	1.421651
19	future	1.410883
20	communication	1.292467
21	biology	1.378306
22	creativity	1.192926
23	humanity	1.436990
24	economics	1.274833
25	environment	1.214839
26	medicine	1.530855
27	2008	1.496490
28	2009	2.011585
29	2010	2.143481
30	2011	2.019762
31	2012	2.096947
32	2013	2.064916
33	2014	1.904096
34	2015	2.101177
35	2016	2.693034
36	2017	1.721869
37	male	5.170769
38	female	3.045922

B. Summary of the Full Model



Model:	GLM	Df Residuals:	1532
Model Family:	NegativeBinomial	Df Model:	39
Link Function:	log	Scale:	0.832915921584
Method:	IRLS	Log-Likelihood:	-23977.
Date:	Sun, 26 Aug 2018	Deviance:	649.29
Time:	17:56:08	Pearson chi2:	1.28e+03
No. Iterations:	22		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.1972	0.161	81.773	0.000	12.881	13.514
duration	-1.913e-06	8.11e-05	-0.024	0.981	-0.000	0.000
TED_Global	0.1827	0.066	2.785	0.005	0.054	0.311
TED	0.2184	0.056	3.885	0.000	0.108	0.329
incitement	0.0350	0.004	8.984	0.000	0.027	0.043
des_posi	1.3519	0.501	2.696	0.007	0.369	2.335
in_degree	0.0435	0.008	5.196	0.000	0.027	0.060
eig_cen	-3.8285	1.650	-2.321	0.020	-7.062	-0.595
technology	-0.0841	0.057	-1.470	0.142	-0.196	0.028
science	0.0879	0.068	1.288	0.198	-0.046	0.222
global_issues	-0.3309	0.064	-5.149	0.000	-0.457	-0.205
culture	0.2080	0.067	3.126	0.002	0.078	0.338
design	-0.1122	0.072	-1.570	0.117	-0.252	0.028
business	0.3663	0.073	5.000	0.000	0.223	0.510
entertainment	0.0359	0.083	0.434	0.664	-0.126	0.198
health	0.1944	0.090	2.158	0.031	0.018	0.371
innovation	-0.0441	0.090	-0.492	0.623	-0.220	0.132
society	-0.0567	0.098	-0.578	0.563	-0.249	0.136
art	-0.2352	0.089	-2.649	0.008	-0.409	-0.061
social_change	-0.1291	0.093	-1.393	0.163	-0.311	0.052
future	-0.0040	0.099	-0.040	0.968	-0.197	0.189
communication	0.0350	0.093	0.376	0.707	-0.147	0.217
biology	-0.1791	0.097	-1.846	0.065	-0.369	0.011
creativity	-0.0056	0.094	-0.060	0.952	-0.190	0.179
humanity	0.1222	0.100	1.225	0.221	-0.073	0.318
economics	-0.1966	0.100	-1.961	0.050	-0.393	-0.000
environment	-0.1969	0.099	-1.993	0.046	-0.391	-0.003
medicine	-0.2906	0.105	-2.773	0.006	-0.496	-0.085
year_2008	0.1761	0.165	1.068	0.285	-0.147	0.499
year_2009	0.2644	0.131	2.016	0.044	0.007	0.521
year_2010	0.1897	0.130	1.464	0.143	-0.064	0.444
year_2011	0.3364	0.130	2.592	0.010	0.082	0.591
year_2012	0.4576	0.130	3.508	0.000	0.202	0.713
year_2013	0.6933	0.130	5.330	0.000	0.438	0.948
year_2014	0.5639	0.135	4.169	0.000	0.299	0.829
year_2015	0.6078	0.134	4.529	0.000	0.345	0.871
year_2016	0.3737	0.145	2.582	0.010	0.090	0.657
year_2017	0.0168	0.169	0.099	0.921	-0.315	0.349
male	0.1464	0.072	2.031	0.042	0.005	0.288
female	0.1549	0.078	1.980	0.048	0.002	0.308