Predicting the Popularity of TED Talks Using Composite Measures of Popularity and Mixed Models

MATH 6627: Case Study

Consultant: Marcos Cardenas-Zelaya

March 11, 2020

Introduction

- TED is a nonprofit organization that spreads ideas, primarily via short talks that can be accessed on the internet such as YouTube.
- The talks are conference like in style, ranging from themes on science to business to global issues.
- TED talks has an incentive to produce high quality videos that garnish attention.
- Finding what characteristics predict a talks popularity/success is essential to TED Talks growth.

Objective

1. Determine what characteristics predict the success/popularity of a TED talk.
2. Determine whether these characteristics varies by time and themes of the talk.

Data Source

- We used descriptive data on TED Talk videos that has been web scraped from the TED talks website.
- The data contains descriptions of videos created during 2006 to September 21st, 2017.
- The original data set contained 2550 observations.
- Each observation includes descriptions of when and where the video was filmed, when it was published, who is/are in the talk, how many comments and views the video has obtained, the title of the video, duration of the video and other variables to describe the video.

Popularity: Average Views/Day

In this analysis, we used two measures of popularity; the most intuitive being a form of views count, and a composite response.

- The most intuitive measurement of popularity is the number of views.
- However, the flaw of this is that it does not account for age of a video: a day old video with 1 million views is just as popular as a 5 year old video with 1 million views?
- To account for this, we divide the number of views of a video by the number of days since it has been published.
- This allows videos to be comparable across length since publication.
- Number of views are in thousands.

Popularity: A Composite Score

We composed a composite popularity score using the following variables:

- **Number of Languages**: A "popular" talk will be translated into several languages as their is a great demand for the talk.
- **Number of Comments**: A "popular" talk will garnish an active comment section as people discuss/praise the video. To account for the number of comments being a function of how controversial a talk is, we use a comments per views metric. As such, we can capture popular videos with a lot of comments and not unpopular/controversial videos with few views but many comments.
- **Ratings**: To account for the ratings that viewers append to each video, we convert each rating and its frequency into a score of $+1$ if the rating is positive (Funny, Beautiful, Ingenious,...) and a score of -1 if negative (Confusing, Unconvincing,...). We add up the score for each rating times its frequency to get the aggregate rating score.

Popularity: A Composite Score

- **Number of Related Talks**: We can treat the relationship between videos as a graph. A more popular video will be closer to the center of that graph, and popularity drops off from the center. The more related talks, the more popular the video.

We use equal weightings in the construction of the composite variable, however unequal weight could be given if a prior knowledge of a particular variable should be weighted more. Since all of these variables are on vastly different scales of magnitude, we normalize the variables and add them to create our rough composite popularity score.

Predictors

- **Duration**: We include duration of videos as we assume that individuals are more likely to watch a shorter video than a longer video.
- **Number of Speaker**: We assume that more speakers will increase the chance that a viewer can associate with the video and thus watch it.
- **Film Age**: We assume that more recently produced videos are more likely to be seen. Film Age is the videos age in days since being produced.
- **Video Age Group**: We add a categorical variable for when the video was published, labeled as 'old' for videos published prior to 2010 and young after 2010. We assume that viewers also consider when a video was published and have a preference for videos published more recently for the same reason as Film Age. Note: A video made in 1990 but published in 2016 would be considered new.

Predictors

- **Title Sentiment**: As the title is the first thing a viewer will see, we assume that the title plays a crucial role in attracting views. To account for this we applied data clustering with K-Means with TF-IDF on the titles to try to separate titles into three groups that might suggest the titles topic.
- **Title Length**: We assume that shorter titles, like shorter video lengths, will encourage views as a viewer can quickly understand the topic of the video rather than being forced to read a lengthy title which could be potentially cutoff, which could further disincentive a viewer.
- **Themes Label**: The data set provides a list of themes that the video is associated with. However, the videos will have several themes which might not be informative as TED talks has an incentive to apply as many themes to garnish views. Thus, we apply data clustering with K-Means with TF-IDF to determine the most relevant theme of the video. We assume some themes might garnish more views than others.

Variable	Levels	n	%	∑%
Title Label	future	2197	86.2	86.2
	life	85	3.3	89.5
	new	73	2.9	92.3
	world	195	7.7	100.0
	all	2550	100.0	
Video Theme	brain	147	5.8	5.8
	business	184	7.2	13.0
	culture	605	23.7	36.7
	design	329	12.9	49.6
	energy	64	2.5	52.1
	global	354	13.9	66.0
	health	192	7.5	73.5
	music	118	4.6	78.2
	science	346	13.6	91.7
	social	211	8.3	100.0
	all	2550	100.0	
Video Age Label	new	1711	67.1	67.1
	old	839	32.9	100.0
	all	2550	100.0	

: categorical

Statistical Analysis

- **Objective 1:** Determine what characteristics predict the success/popularity of a TED talk.
  - **Model:** Linear and Poisson Regression.
- **Objective 2:** Determine whether these characteristics varies by time and themes of the talk.
  - **Model:** Linear and Poisson Mixed-Models with random slopes and intercepts to account for the differences in time and themes.

Poisson Regression on Average Views Per Day

To predict the average views per day, which is a count variable, we use a Poisson regression with a log link.

$$\begin{aligned} log(\text{Video } i \text{ Average Num.Views/Day}) = &\text{ Video Duration}_i + \text{Num. Speakers}_i \\ &+ \text{Film Age}_i + \text{Title Length}_i \\ &+ \text{Video Themes}_i + \text{Titles Content}_i \\ &+ \text{Video Age Group}_i + \epsilon_i \end{aligned} \tag{1}$$

$$i = \{1, .., 2550\}$$

This model indicates the log of the average number of views a video will obtain on any given day given the predictors.

Linear Regression on Composite Popularity Score

To predict the popularity, which is a normally distributed number, we use a Linear regression model.

$$\begin{aligned} \text{Video } i \text{ Popularity} = &\ \text{Video Duration}_i + \text{Num. Speakers}_i \\ &+ \text{Film Age}_i + \text{Title Length}_i \\ &+ \text{Video Themes}_i + \text{Titles Content}_i \\ &+ \text{Video Age Group}_i + \epsilon_i \end{aligned} \tag{2}$$

$$i = \{1, .., 2550\}$$

This model indicates the popularity score a video will obtain given the predictors.

Statistical Analysis With Mixed Models: Time Variation

- To determine whether the characteristics that predict popularity change over time, we define time into two groups, old (videos published prior to 2010) and new (after 2010).
- **Random intercept:** Viewers of TED talk videos in 1990 when personal computers where not readily available were most likely from a higher socio-demographic population than the average viewer in 2017 when the vast majority of individuals have personal computers. As such, 'old' and 'new' videos will have different intercepts (different average popularity or average views/day).
- **Random slope:** An individual in the early 2000s with a computer was most likely more educated (due to the general lack of widespread computers they would need it for very specific reasons like work/school) than the average computer user in 2017 (where everyone owns a computer for both pleasure and work), and as such the former group might rate a longer video better than the latter group due to differences in attention span, for example.

Statistical Analysis With Mixed Models: Time Variation

We apply Mixed Models to the Poisson regression to model the variation in time for the average views per day. $log(\text{Average Num.Views/Day}_{ij})$ denotes the log number of average views per day.

$$\begin{aligned} log(\text{Average Num.Views/Day}_{ij}) = \beta_0 + b_{0j} + (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ + (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} \\ + (\beta_3 + b_{3j})\text{Film Age}_{ij} \\ + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ + (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij} \end{aligned} \tag{3}$$

$$i = \{1, .., n_j\}, j = \{old, new\}$$

In this model we have b's as the random slope/intercept for i observations from each time group j.

Statistical Analysis With Mixed Models: Time Variation

We apply Mixed Models to the Linear regression to model the variation in time for the popularity score. Popularity_{ij} denotes the popularity score.

$$\begin{aligned}\text{Popularity}_{ij} = \beta_0 + b_{0j} + (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ + (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} + (\beta_3 + b_{3j})\text{Film Age}_{ij} \\ + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ + (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij}\end{aligned} \tag{4}$$

$$i = \{1, .., n_j\}, j = \{old, new\}$$

In this model we have b's as the random slope/intercept for i observations from each time group j.

Statistical Analysis With Mixed Models: Theme Variation

- To determine whether the characteristics that predict popularity vary by the talks themes, we define themes into ten groups, as determined by the K-means clustering with TF-IDF, as: brain, business, culture, design, energy, global, health, music, science, social.
- **Random intercept:** We assume each theme attracts viewers from different populations. As such, we assume each theme will attract different populations and thus videos from each theme will have different intercepts (different average popularity or average views/day).
- **Random slope:** To account for potential differences in populations that might have different video viewing behaviors. Indeed, a computer science student that watches science related videos will most likely have a different attention span than the average viewer who watches music related videos. As such, the former group might rate a longer video better than the latter group due to differences in attention spans.

Statistical Analysis With Mixed Models: Theme Variation

We apply Mixed Models to the Poisson regression to model the variation in themes for the average views per day. $log(\text{Average Num.Views/Day}_{ij})$ denotes the log number of average views per day.

$$\begin{aligned} log(\text{Average Num.Views/Day}_{ij}) = \beta_0 + b_{0j} + (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ + (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} \\ + (\beta_3 + b_{3j})\text{Film Age}_{ij} \\ + (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ + (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij} \end{aligned} \tag{5}$$

$$i = \{1, .., n_j\}, j = \{brain, business, culture, ..., social\}$$

In this model we have b's as the random slope/intercept for i observations from each theme j.

Statistical Analysis With Mixed Models: Theme Variation

We apply Mixed Models to the Linear regression to model the variation in themes for the popularity score. Popularity_{ij} denotes the popularity score.

$$\begin{aligned}\text{Popularity}_{ij} = \beta_0 + b_{0j} &+ (\beta_1 + b_{1j})\text{Video Duration}_{ij} \\ &+ (\beta_2 + b_{2j})\text{Num. Speakers}_{ij} \\ &+ (\beta_3 + b_{3j})\text{Film Age}_{ij} \\ &+ (\beta_4 + b_{4j})\text{Title Length}_{ij} \\ &+ (\beta_5 + b_{5j})\text{Titles Content}_{ij} + \epsilon_{ij}\end{aligned} \tag{6}$$

$i = \{1, .., n_j\}, j = \{brain, business, culture, design, energy, global, health, musi$

In this model we have b's as the random slope/intercept for i observations from each time group j.

Results: Linear and Poisson Regressions

: Poisson and Linear Regression

	Dependent variable:	
	avg_views_per_day	popularity
	Poisson	Linear
Duration	1.387*** (0.008)	0.132** (0.056)
Num. Speaker	-0.834*** (0.009)	-0.083(0.076)
Film Age in Days	-16.303*** (0.016)	-2.849*** (0.118)
Title Label: Life	0.149*** (0.002)	-0.001(0.022)
Title Label: New	-0.142*** (0.003)	-0.007(0.023)
Title Label: World	-0.099*** (0.002)	-0.044*** (0.015)
Title Length	0.053*** (0.004)	0.087*** (0.030)
Note:		*p<0.1; **p<0.05; ***p<0.01

Results: Linear and Poisson Regressions

	avg_views_per_day Poisson	popularity Linear
Theme Label: Business	-0.135*** (0.002)	-0.002(0.022)
Theme Label: Culture	-0.314*** (0.002)	-0.043** (0.018)
Theme Label: Design	-0.537*** (0.002)	-0.057*** (0.019)
Theme Label: Energy	-0.589*** (0.004)	-0.082*** (0.029)
Theme Label: Global	-0.845*** (0.003)	-0.059*** (0.019)
Theme Label: Health	-0.692*** (0.003)	-0.004(0.021)
Theme Label: Music	-0.523*** (0.003)	-0.130*** (0.025)
Theme Label: Science	-0.669*** (0.002)	-0.030(0.019)
Theme Label: Social	-0.552*** (0.002)	-0.047** (0.021)
Video Age Group: Old	-1.841*** (0.006)	-0.317*** (0.026)
Film Age:Video Age Group: Old	13.460*** (0.031)	2.434*** (0.153)
Intercept	8.952*** (0.003)	2.038*** (0.025)
Observations	2,550	2,550

Results: Linear and Poisson Regressions

- **Duration**: Both models report an increase in the response, which is counter intuitive as we initially believed that viewers would be turned off from longer videos.
- **Number of Speaker**: Both models reported a decrease in the response, however the linear model did not consider it significant. This goes against our initial beliefs that more speakers would increase the likelihood of a viewer seeing a speaker they enjoy.
- **Film Age**: Both models report significant decreases in the response, which agrees with our initial hypothesis that viewers prefer videos that have been published more recently.
- **Video Age Group**: Both models report decreases in the response. This supports our initial hypothesis that viewers prefer videos that have been produced more recently.

Results: Linear and Poisson Regressions

- **Film Age:Video Age Group: Old**: We also model the interaction between these two terms to see how talks produced prior to 2010 but published after 2010 performs. Both models report an increase in the responses, which suggests that an 'old' video but published recently will perform just as well as a 'new' video published recently. This suggest that the age of the video in days since posted is a stronger predictor for success than the actual content of the talk.
- **Title Sentiment**: Both models agree that titles regarding the topic 'world' decrease the response and are significant. The poisson model reports a decrease in average views/day for the title topic 'new', but reports an increase for the topic 'life'.
- **Title Length**: Both models report increases in the response and are significant. This is surprising as we thought verbose titles would discourage viewers from watching the video.
- **Themes Label**: Both models report all the labels as significant and negative which suggests that the most popular talks do not fall under any of these themes.

Results: Poisson Mixed Models

: Poisson Mixed Effects Results

	Dependent variable:	
	avg_views_per_day	
	Poisson Themes	Poisson Times
Duration	1.976*** (0.490)	1.386*** (0.149)
Title Length	0.155(0.169)	-0.097(0.604)
Title Label: Life	0.064(0.093)	-0.147(0.729)
Title Label: New	-0.166(0.147)	-0.218(0.629)
Title Label: World	-0.129** (0.054)	-0.226(0.400)
Num. Speaker	-0.689(0.755)	-0.887*** (0.231)
Film Age in Days	-12.327*** (0.698)	-9.810*** (2.759)
Intercept	8.105*** (0.104)	7.629*** (0.951)
Observations	2,550	2,550
Log Likelihood	-1,075,431.000	-1,131,767.000
Akaike Inf. Crit.	2,150,950.000	2,263,623.000
Bayesian Inf. Crit.	2,151,207.000	2,263,880.000

Results: Linear Mixed Models

: Normal Mixed Effects Results

	Dependent variable:	
	popularity	
	Normal Themes	Normal Times
Duration	0.399*** (0.136)	0.171(0.107)
Title Length	-0.047(0.124)	-0.238* (0.139)
Title Label: Life	0.178*** (0.050)	0.107*** (0.030)
Title Label: New	0.023(0.032)	0.001(0.022)
Title Label: World	-0.0002(0.027)	-0.006(0.023)
Num. Speaker	-0.048** (0.023)	-0.046*** (0.015)
Film Age in Days	-1.330*** (0.145)	-1.615(1.191)
Intercept	1.797*** (0.035)	1.823*** (0.146)
Observations	2,550	2,550
Log Likelihood	436.955	515.918
Akaike Inf. Crit.	-783.910	-993.836
Bayesian Inf. Crit.	-520.937	-882.802

Model Selection: Testing Random Intercept

To compare whether these groups have different average responses, we test the random intercept of these four models using AIC for model selection.

Table: Testing Random Intercept

	df	AIC
Linear Popularity Themes null	9.00	-810.37
Linear Popularity Themes full	10.00	-789.73
Linear Popularity Times null	9.00	-810.37
Linear Popularity Times full	10.00	-767.85
Poisson Avg. Views/day Themes null	8.00	2456258.87
Poisson Avg. Views/day Themes full	9.00	2262017.56
Poisson Avg. Views/day Times null	8.00	2456258.87
Poisson Avg. Views/day Times full	9.00	2418695.71

Model Selection: Testing Random Intercept

- **Linear Model with Random Intercept for Themes**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average popularity score across themes.
- **Linear Model with Random Intercept for Time**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average popularity score across time.
- **Poisson Model with Random Intercept for Themes**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average avg. views per day across themes.
- **Poisson Model with Random Intercept for Time**: The AIC for the mixed model with random intercept is lower than the null, so it is preferred. This suggest that there exists differences in the average avg. views per day across time.

Model Selection: Testing Random Slope

Similarly, we test whether individuals between the four models differ in their odds of having OA. We test these random slopes using χ^2 model selection.

Table: Testing Random Slopes

	Df	Chisq	Chi Df	Pr(>Chisq)
Lin. Popularity Themes simple	10			
Lin. Popularity Themes full	45	59.91	35	0.0055
Lin. Popularity Times simple	10			
Lin. Popularity Times full	19	240.34	9	0.0000
Pois. Avg. Views Themes simple	9			
Pois. Avg. Views Themes full	44	111138.01	35	0.0000
Pois. Avg. Views Times simple	9			
Pois. Avg. Views Times full	44	155142.81	35	0.0000

Model Selection: Testing Random Slope

- **Linear Model with Random Slope for Themes**: The p-value is very small; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect popularity score across themes.
- **Linear Model with Random Slope for Time**: The p-value is very small; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect popularity score across time.
- **Poisson Model with Random Slope for Themes**: The p-value is very small; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect avg. views per day across themes.
- **Poisson Model with Random Slope for Time**: The p-value is very small; the preferred model includes the random intercept. This suggest that there exists differences in how the predictors effect avg. views per day across time.

Model Selection

- The best models include both random intercepts and slopes for both Poisson and Linear models.
- Which suggests that a lot of the variation in the data can be explained by timing of when a video is published and its theme rather than the actual form/content.
- The main predictors in these models are the talks duration and age (in terms of how recent since it has been published).
- Which suggests that the best way to create a popular video is by having a long video and re-uploading it to escape the drop in popularity from aging.

Conclusion

- In this analysis we used web scraped data from the TED talks website to determine what predicts talks popularity.
- To measure popularity we used average views per day and a composite score that encompasses several intuitive measures of popularity.
- We used Linear and Poisson regression with Lasso penalization and Linear and Poisson Mixed Models with random intercepts and slopes to exploit variation in talks across time and themes.
- We found that the majority of variation in the data can be explained by variation across time and themes.
- We found that the strongest predictors of popularity, in all models, was a talks duration and how long it has been uploaded for.

Conclusion

- Future research will benefit from using the transcripts of the talks as a measure of the talks content.
- Due to computational restrictions we were unable to run a model with both time and themes as random components. Future research will benefit from running these models to see which component explains the data the most.