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 rgaization that speads 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 from 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.

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.