How a 35-Variable Model Can Estimate the Views of TED Talks

John Geer jgg150@psu.edu

Abstract

Effective online communication has become essential for many endeavors. This communication could be improved with a better understanding of the features that attract attention online. This analysis built a predictive model of the number of times a TED talk is viewed. This predictive model was based on the title, topic, and publish date of the talks. This model was able to produce predictions with about 20% lower prediction mean squared error on the log of views than a null model. Cross validation techniques were used to prevent overfitting. Variables based on the topic of the TED talks were able to improve predictions the most. Variables based on the title and date when the talks were published online improved predictions to a lesser extent.

Introduction

Why This Research Matters

As of March 2014, approximately 100 hours of video are uploaded to YouTube every minute[i] and 400 million tweets are created every day[ii]. Amazingly, this represents a small part of the content published online.

Given the prevalence of online communication, a better understanding of how the features of a piece affect the attention it receives would be quite valuable. Such an understanding could enable more effective sharing of ideas.

Happily, online communication provides excellent opportunities for analysis. Data on the relative attention that different pieces of content receive is regularly recorded.

This analysis focuses on how some content features predict the amount of attention a piece receives. Content features are particularly interesting because they are easy to change. For example, because headlines are easy to adjust, an experimenter can relatively efficiently test different headline styles to see which tend to attract more attention. Likewise, an author can easily tweak a headline to attract more viewers.

Past Research

Other research has looked into how features of online content are associated with the attention they receive.

Lai and Farbrot investigated the number of clicks that different headlines received.[iii] They found that headlines with self-referential cues (such as "you" or "your"), attracted more attention. They also found that question headlines attracted more attention than non-question headlines.

Other research looked into predictors of changes in Twitter followers based on the content of tweets. Hutto, Sarita, and Gilbert found that negative sentiment predicted decreases in followers while positive sentiment predicted increases in followers.[iv]

Why TED Data

This analysis focuses on TED talks. These are short presentations given at conferences associated with the TED organization. These talks are generally between 7 and 25 minutes long; they cover a broad range of topics including global issues, technology demonstrations, and artistic performances. The TED organization records these talks and publishes the videos online through multiple services. Once published, the number of times these talks are viewed is recorded.

TED talks provide a semi-controlled environment to investigate content features. Compared to a sample of all web content, TED talks have relatively similar non-content features. For example, all of the TED talks are posted to the same site and are associated with the same brand. The talks are all videos. They are also released at relatively regular intervals, which provides them with similar time as a "new" talk.

While many non-content differences remain, TED talks still provide a unique opportunity to investigate the predictive ability of content features on the amount of attention a piece of content receives. Some of the content features that this analysis investigates includes the title, topic, summary, duration, and number of days that a talk has been online.

Aims of This Analysis

This analysis has two primary aims. First, it looks to identify features of TED talks that predict the number of times the talks are viewed. Second, it seeks to provide some perspective on the predictive value of these features, both in comparison to other predictors and with the overall variation in views.

Regarding identifying predictors, particular attention was paid to the content of talks. This analysis considers features noted in previous research, such as whether a title contains a self-referential cue or is a question. However, many other variables are also considered. This allows the possibility of

finding predictors not previously identified.

This analysis was able to accomplish both aims. It found several good predictors of the number of times TED talks were viewed. This analysis then compared the predictive ability of these variables, both with each other and with a null model that used no variables.

Methods

Data Source

Raw Data

Some of the raw data for this analysis was provided in a spreadsheet from the TED organization. This spreadsheet included the title, brief summary, date published online, and duration of all published talks. In addition, a program was written for this analysis that visited each talk's web page on TED.com and recorded the view counts and topics associated with each talk. The view counts recorded from the TED.com website incorporate views from other sources, such as mobile apps, YouTube, and iTunes. Each talk can be associated with multiple topics which were subjects like "Religion", "Dance", or "Economics".

This process collected complete information on 1,608 TED talks.

Extracting Variables

From the raw data, more machine-readable variables were extracted. These variables included time, title, summary, and topic variables.

Two time variables were extracted: the duration of the talk, expressed as an integer of seconds, and the number of days that a talk has been online. The number of days a talk has been online was calculated as the difference between when the talk was published online and when this analysis recorded the number of views the talk had received.

Several variables were extracted from the title of each talk. These variables included the length of the title (in number of characters), the presence of a question mark, the presence of a self-referencing cue (such as “you” or “your”), the presence of a number, the average word length, the maximum word length, the number of words, and estimates of the number of nouns, verbs, and descriptors (adverbs and adjectives) in the title.

A set of variables were also created that indicated the presence of particular word stems in a title. The stem is the root of a word. For example, the words "stories" and "story" have the same word stem, much like "educate" and "education". A set of 0/1 indicator variables were created for all of the word

stems that appeared in more than ten titles. If a talk had a word with a particular stem in the title, the talk received a 1 for the variable associated with that stem. If the talk lacked a word with that stem the stem variable was 0. For example the talk "My immigration story", would have 1's for the stem variables associated with "my", "immigration"[1], and "story" but a 0 for the stem variable associated with "ocean". These word stem variables accounted for 138 of the variables considered.

With the exception of the word stems, the variables extracted from the title were also extracted from the summary. These variables included estimates of the number of parts of speech (nouns, verbs, and descriptors) and the length of the summary.

Topics associated with the talks were used to create a set of 0/1 indicator variables. Much like with the stem variables, if a given topic was associated with a particular talk, that talk received a 1 for that topic variable. Otherwise the topic variable had a value of 0. Variables were created for all the topics associated with more than five talks, resulting in 229 such indicator variables.

In total, 403 predictor variables were considered in this analysis.

Model Building

Several methods for building predictive models were considered, including linear and poisson regressions as well as regressions with smoothing splines (general additive models). Because of the large number of variables, methods that performed some variable selection were also considered. These methods included the lasso, random forests, and boosted tree algorithms.

Because several of these methods assume a normally distributed error, continuous variables were transformed to better approximate normal distributions. The response variable was also transformed for this reason using a log transformation (figure 1 shows density estimates of this variable). Using the log of views as the response improved the predictions of linear methods as well as tree-based methods, such as random forests and boosted trees.

Model Selection

The largest difficulty this analysis faced was selecting from the many variables. Most of these variables were likely to be poor predictors. The large number of variables makes it easy to fit a given data set too closely, incorporating randomness into one's model. This overfitting results in poor predictions on new data.

In order to limit overfitting, training, validation, and testing data sets were used in combination with

1The stem of "immigration" did not actually appear in more than 10 titles. As a result, a stem variable for “immigration” was not created.

10-fold cross validation. Half of the overall data set was randomly selected to be the training data set. Models were initially built on this data and then the quality of their predictions were compared on the validation data set (0.25 of the overall data). Once some reasonable models were found, 10-fold cross validation was used on the combined training and validation data sets to compare the prediction quality of the models. The test data set (0.25 of the overall data) was set aside to test the final model. This test data provided an estimate of the final model's predictive performance on new data.

Models which used the variable sets selected by the lasso, random forests, and boosted tree methods were compared. While the variables selected by random forest performed best of the three, slightly better predictions were produced by using a constrained version of forward stepwise selection.

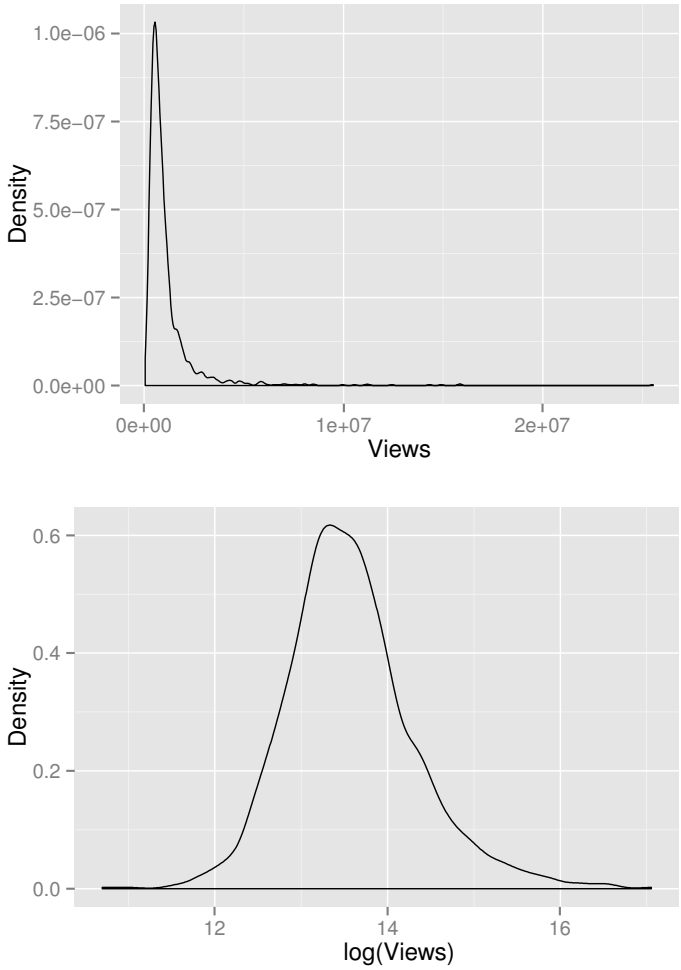

Figure 1: Distribution of the untransformed and transformed response variable, views.

This version of forward stepwise selection added variables one at a time if they improved prediction quality. However, considering all the variables at each step was difficult given the number of them. Instead, an ordering of the variables was used and only the next variable in the ordering was considered at each step. The variable ordering produced by both random forests and boosted trees were tried. However, the best results were found by using a ranking of the variables according to the predictive performance of models which only included the variable in question. The predictive performance of these single variable models was estimated with 10-fold cross validation.

Results

The final model, which had the lowest prediction mean square error (MSE), used just 35 variables in a general additive model. This model included smoothing splines with five degrees of freedom to fit the variables expressing the number of days a talk has been online.

On the test data, the predictions of this model had a mean square error of 0.4273 (for log(views)). This is about a 19% smaller MSE than that of a null model with no predictors (MSE of 0.5292). When evaluated with 10-fold cross validation on the entire data set, this model had an MSE of 0.4268 (for log(views)). This is about 23% better than the null model (MSE of 0.5562).

Of the 35 variables in the final model, 23 were topic variables and 10 were related to the title of the talk. Two of the variables expressed the number of days the talk had been online. One was the number of days the talk was online and the other was the log of that number (removing either or both of these variables resulted in worse predictions). The title and topic variables are listed with their coefficients in figures 2 and 3. These coefficients are also expressed in the table in Appendix A.

Relative Predictive Ability

The predictive ability of these variables was compared by building models with subsets of the final 35 predictors. The predictive ability of these constrained models was compared to each other, to the null model, and to the overall model. This comparison provided an estimate of the relative contribution of these variables to the predictions of the final model. The predictive performance of models using subsets of these variables can be seen in table 1. This table shows that the topic variables seem to provide the greatest improvement in prediction quality, followed by the time and title variables. The fact that the predictive ability of these almost adds up to the total 23% reduction in MSE of the full model suggests that there is minimal redundancy across these groups of variables.

Variables Considered	Improvement over the null model
Entire Final Model	23%
Time Variables Only	5%
Topic Variables Only	15%
Title Variables Only	4%
Null Model	0%

Table 1: Predictive performance of the final model and models using subsets of those variables. Performance calculated with 10-fold cross validation using MSE on log(views).

Discussion

Topic Predictors

The topic variables appear to improve the predictions of log(views) the most. All of these variables are 0/1 indicator variables. As a result, the coefficients indicate the change in log(views) for talks tagged with that topic. Figure 2 lists the topic variables in the final model and shows their relative coefficient sizes.

Interestingly, topics with positive coefficients, suggesting more views, seem related to happy emotions

and excitement. For example, the topics of "magic", "success", "love", and even "demo" are generally upbeat with titles like "The brain in love", "The magic of truth and lies (and iPods)", and "Creating tech marvels out of a $40 wii remote". However, topics with negative coefficients, which suggest fewer views, seem more related to negative emotions. These include topics like "disaster relief", "biosphere", and "health care" with titles like "Haiti's disaster of engineering", "Life science in prison", and "The mystery of chronic pain".

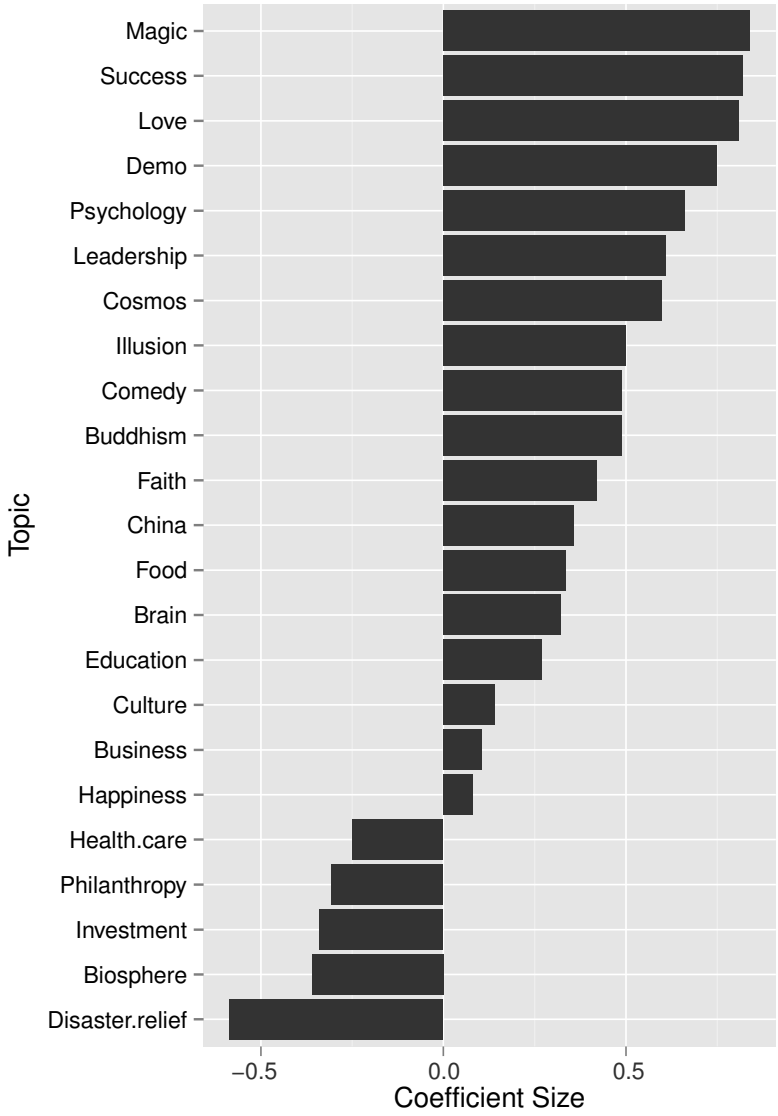

Figure 2: The topic variables in the final model and their associated coefficients.

This possible association between emotions and views is in line with previous research regarding tweets with positive sentiment garnering more followers[1]. However, it is worth noting that the predictive ability of positive or negative sentiment is not entirely clear in this data. There are many other topics with positive ("play" or "innovation") or negative ("war" or "inequality") associations that did not appear to be strong predictors of the number of times TED talks were viewed.

Title Predictors

Figure 3 lists the 10 title related variables and their coefficients. As with the topic variables, these are all 0/1 indicator variables. As a result, the size of their coefficient indicates the change in log(views) between a talk with or without these words or features in its title.

You and We

Of particular note are the "you" and "we" variables. The coefficients for these variables are both

positive, suggesting that talks with these words in the title tend to receive more views. Titles with "you" or "we" in them seem to focus attention on one's own needs and perspectives with titles like "Why we do what we do, and how we can do it better", "Yes, design can make you happy", and "Are we born to run?".

This association between self-referential words and increased views agrees with research on headlines by Lai and Farbrot[i]. However, it is also interesting to see this in relation to the topic variables that we considered. A model only fit on the topic variables reduced MSE more than a model that only considered the title variables. This suggests that while self-referential words in the title do seem to be predictive of more views, what the title expresses, such as its topic, may be more important.

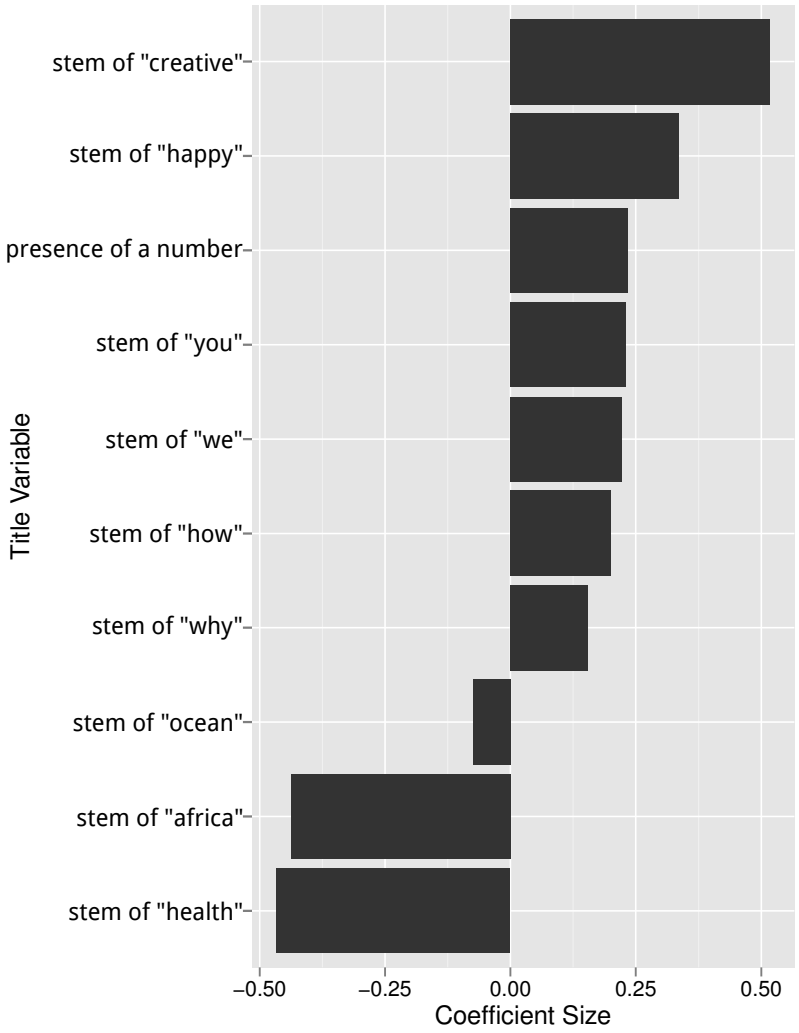

Figure 3: The title variables in the final model and their coefficients

How and Why

The stem variables for "how" and "why" are also noteworthy predictors. Their coefficients are both positive, suggesting titles with these words are associated with more views. This largely agrees with research by Lai and Farbrot which found question titles were associated with more clicks[1].

It may be worth noting that this analysis included another variable which indicated the presence of a question mark. When the "how" and "why" stem variables are replaced by this question mark variable, the model produces slightly worse predictions. In fact, it produces very similar predictions if the "how" and "why" stem variables are removed without replacing them at all. This may indicate that a feature beyond simply being a question is predicting views. One possibility is that titles with "how"

or "why" in them seem to indicate solutions or explanations. For example, "How great leaders inspire action", "Why we will rely on robots", or "How a fly flies". However, the TED titles with a question mark can simply indicate an open question such as "Why is 'x' the unknown?", "What will future jobs look like?", or "Is anatomy destiny?". Perhaps it is not the question that attracts attention, but the suggestion of a solution.

Numbers

Whether or not a title has a number in it is also a notable predictor. The large positive value of this indicator variable's coefficient suggests that titles like "How I held my breath for 17 minutes", "A virtual choir 2,000 voices strong" and "7 rules for making more happiness" are likely to garner more views than talks without numbers in their title. This may indicate that lists tend to receive more attention, because many of the titles with numbers are lists. However, this variable includes many talks whose titles aren't lists such as "Life at 30,000 feet". It is possible that numbers have predictive ability beyond indicating lists.

Days Online

The number of days that a talk had been online was the only continuous variable in the final set of predictors. In the final model this variable was fit with a smoothing spline, which allowed a curved relationship between the log of views and the number of days a talk has been online. Figure 4 shows the relationship between days online and log(views) in the final model.

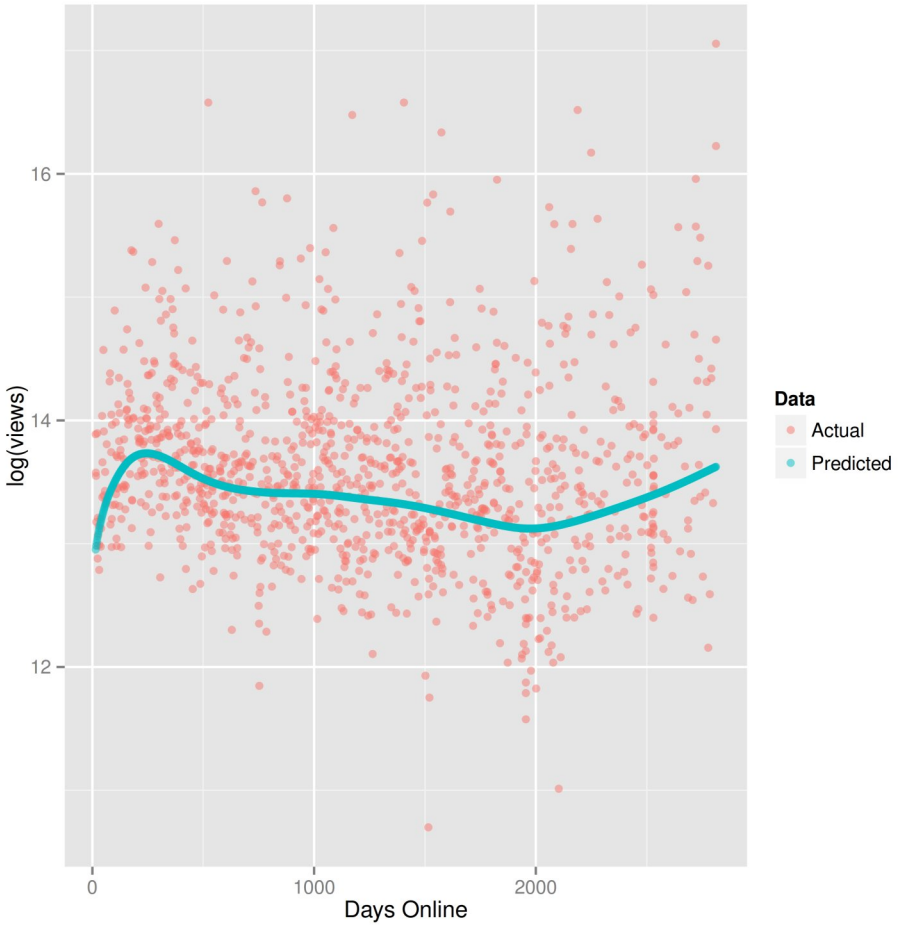

Figure 4: The final models fit to the number of days online and the values of the actual data.

This variable was calculated as the number of days between when the

view count was recorded for this analysis and when a talk was published online by the TED organization. Because this analysis recorded the view counts on the same day for all the talks, this unfortunately conflates two features of the talks. The first feature is how long the talks have been available to circulate online. The second is when the talks were posted online. When the talk was published incorporates features such as time of year, day of week, the popularity of TED when this talk was featured as a “new release”, and other possible influences.

Figure 4 shows evidence of both of these features in this variable. There is a steep increase in views for low values of days online. This may be due to increased attention placed on new talks. However, for large numbers of days online we see a different pattern. This long term pattern may indicate different designs of the TED website, changes in distribution, changes in the quality of talks posted, or simply chance.

Summary Predictors

None of the predictors derived from the summary made it into the final model. This suggests that the features of the summary that this analysis was able to extract did not have a particularly strong effect on the number of views a talk received.

Results of Using the Log Transform of Views

The models which performed best predicted the log of views. For a linear model, like the final model used, this means that the predictors are multiplicatively related to the raw view count. So having more of these positive predictive features indicates an exponentially larger number of views.

This transformation also results in predictions that are less precise for higher view counts. Errors for low values of log(views) and large values of log(views) were treated the same when fitting this model. However, the same one unit difference of log(views) represents a much larger difference in raw views for higher numbers of views.

Limitations of Observational Approach

The observational approach used in this analysis did not show causation. It was only able to find features in TED talks that seem to predict the number of views those talks receive. The difficulty of variable selection in this process means that it would be quite easy to miss relationships or to mistakenly include unrelated variables. Though the cross validation techniques used help limit these problems, true predictors were almost certainly missed or misleading ones included in the final model. That said, the separation of the data used to fit the model and the data used to asses the model helped the analysis look beyond simple associations and consider predictive relationships.

Unmeasured Features & Future Research

The final model in this analysis explained only a modest amount of the variation in the views. This is probably because this analysis only considered a few features of the talks. Features of TED talks such as the fame of the presenter, the topic's relation to current events, or the publicity particular talks have received are likely to influence the number of times talks are viewed. However, none of these features were considered in this analysis.

The variables this analysis focused on are all features of the content. Because such features are easy to change in an experimental setting, they present excellent targets for future research. For example, an interesting future experiment would be to further investigate the effect of a number in the title. Such an experiment could compare the impact of using a number as a list ("4 lessons in creativity") with using a number as a detail ("Try something new for 30 days"). Another interesting experiment would be to further investigate question words in titles. This experiment could compare responses to titles that are questions ("Why do we sleep?") with titles that suggest an explanation ("Why we need the explorers").

Conclusion

This analysis identified several good predictors of the number of times TED talks are viewed and provided perspective regarding the predictive ability of these variables. The predictive variables included features of the talks' title, topics, and publish date. Among these categories of predictors, the topics variables appeared to provide the greatest increase in prediction accuracy. The title variables and number of days a talk has been online provided lesser improvements in predictions. Overall, these variables were able to predict around 20% of the variation in the log of TED talk views.

Beyond TED talks, the features used in the final model offer a good starting point for future investigations into how the features of a piece of content affect the attention it receives online.

References

i "Viewership Statistics." Youtube.com, n.d. Web. 15 Feb. 2014. <https://www.youtube.com/yt/press/statistics.html>.

ii "Twitter Turns 7: Users Send over 400 Million Tweets per Day." Washington Post, n.d. Web. 15 Feb. 2014. <http://www.washingtonpost.com/business/technology/twitter-turns-7-users-send-over-400-million-tweets-per-day/2013/03/21/2925ef60-9222-11e2-bdea-e32ad90da239_story.html>.

iii Lai, Linda, and Audun Farbrot. "What Makes You Click? the Effect of Question Headlines on Readership in Computer-mediated Communication." Social Influence (2013): N. pag. Web. 25 Oct. 2013.

iv Hutto, C J., Sarita Yardi, and Eric Gilbert. "A Longitudinal Study of Follow Predictors on Twitter." Georgia Institute Of Technology, n.d. Web. 15 Feb. 2014. <http://comp.social.gatech.edu/papers/follow_chi13_final.pdf>.

Appendix A: Coefficients of final model

(Intercept)	13.41	headline_has_number	0.23
topic_Magic	0.84	stem_you	0.23
topic_Success	0.82	stem_we	0.22
topic_Love	0.81	stem_how	0.20
topic_Demo	0.75	stem_whi	0.15
topic_Psychology	0.66	topic_Culture	0.14
topic_Leadership	0.61	topic_Business	0.11
topic_Cosmos	0.60	topic_Happiness	0.08
stem_creativ	0.52	stem_ocean	-0.08
topic_Illusion	0.50	topic_Health.care	-0.25
topic_Comedy	0.49	topic_Philanthropy	-0.31
topic_Buddhism	0.49	topic_Investment	-0.34
topic_Faith	0.42	topic_Biosphere	-0.36
topic_China	0.36	stem_africa	-0.44
stem_happi	0.34	stem_health	-0.47
topic_Food	0.33	topic_Disaster.relief	-0.59
topic_Brain	0.32	time_online	NA
topic_Education	0.27	time_online_log	NA

Appendix B: Reference to code and data

This analysis used python 2.7.3 and R 3.1.0. The code and data used in this analysis are attached in a separate file, due to size.