Predictive Model for Views In YouTube Beauty Community

Lavanya Sunder
Duke University
Spring 2016

April 8, 2016

Abstract

The YouTube makeup community is a large and growing niche of YouTube, as well as an influential force in the overall beauty space. Recently, the YouTube makeup community has been saturated with new content-creators, "beauty gurus," who want to successfully create content for subscribers. However, it is not widely certain what specific aspects of makeup videos make them appealing to subscribers. It is also uncertain what the impact of brand sponsorship is on a makeup video, particularly as it pertains to views and viewer sentiment. In this paper, a sample of makeup videos is used to create a predictive model for views, based on a variety of factors including: video content, tags, comments, and the description box. The results provide insight on the dominant variables of a YouTube video, and indicate areas for further discussion and exploration.

1 Introduction

The collection of YouTube users producing makeup and beauty related videos, known collectively as the "YouTube Beauty Community," is a large force in the world of online and offline beauty. As of April 2015, there were more than 45.3 billion total views on beauty related videos, a growth of over 50% from the year prior.[12] The YouTube Beauty Community is primarily comprised of makeup videos, which account for over 50% of all videos in the beauty space (which also includes hair, skincare, etc).[12] Content creators known as beauty or makeup "gurus," create a variety of videos: tutorials, reviews, hauls, and other subcategories. The community not only has a strong influence on YouTube, but also drives much of the national dialogue on makeup and beauty. In fact, after a popular beauty guru created a video on Nivea After Shave Balm as a makeup primer, the makeup space exploded with videos and articles regarding the product, including mainstream publications such as The Sun UK, The Daily Mail UK, and Bustle.[6]

As the influence of YouTube has grown, brands and advertisers have become increasingly involved in the makeup community. In lieu of brand-affiliated channels, which comprise only 3% of the beauty space and rarely compete against non brand-affiliated channels, brands are enabling gurus to monetize their videos through advertising deals, sponsorship, and even through their own makeup products.[12] Michelle Phan, currently the highest-grossing beauty guru (and one of the highest-grossing YouTubers) in the world, is possibly the best example

of this monetization. She often creates brand sponsored videos, has been in a Dr. Pepper commercial, and even has her own beauty line, backed by Lancome; Phan made over $5 million in 2013.[10] While the beauty community is by no means dominated by Phans, fairly successful gurus with more than a million followers can make six-figure salaries.[10] The most common monetization tool used by these gurus are brand sponsored videos; gurus can get anywhere between $10,000 and $15,000 per sponsored product, and brands benefit through increased sales as well as comments from viewers.[10]

Unfortunately, this is not the case for a large number of beauty gurus, YouTube's so called "middle class," who are in a limbo of fame, often "too small to sponsor... [and] too big for donations."[7] Fans encourage authentic and original content, and often spur videos with sponsored content. A pair of self-described middle-class gurus note that, every time they post a branded video, "we make money but lose subscribers."[7] This notion is felt across the beauty community, particularly with content creators who have not yet established a large subscriber base.

Thus, the motivation behind this project is two fold: first, to largely understand what makes makeup videos on YouTube successful, in terms of viewership, and second, to understand how sponsorship specifically affects viewership. Data was primarily collected through web scraping and use of the YouTube API, which is further explained in Section 2.1. Additionally, two predictive models were created to estimate views: Model A, which contains only variables which exist at the time of the video being published, and Model B, which also contains dynamic variables such as likes, dislikes, and comment sentiment. Lasso and ridge regression were used to fit both models, as explained in Section 2.2. Finally, the results of these regressions are explained in Section 3, along with a discussion of limitations and next steps.

2 Methods

2.1 Data Collection and Preprocessing

The data set was created using a snowball sampling methodology to create a set of 752 videos. Some videos were discarded because they did not fit the category for YouTube makeup videos; others were discarded because comments were disabled or the video did not have an auto-generated transcript. After a series of preprocessing steps, the final video set consisted of 707 videos and 106 variables, including all values for categorical variables (see Appendix D). The following sections fully explain the steps taken for data collection and preprocessing.

Snowball Sampling

There currently does not exist a freely available source of all content creators within the YouTube beauty community; therefore, a snowball sampling technique was utilized in order to get a data set of makeup videos on YouTube. Snowball sampling is a sampling techniques that identifies qualifying respondents (or in this case, videos) who are then used to identify other qualifying respondents.[2] The technique is traditionally used for more impenetrable social groups, such as drug addicts, and was implemented primarily to mirror the way users navigate YouTube.[2]

A set of 15 starting makeup videos (see Appendix B) was created by looking at a list of the most popular type of makeup videos, searching those types of makeup videos on

YouTube, and then randomly selecting a video from the first page of search results.

These results were supplied to an algorithm that then recursively sampled the "Related Videos" side-bar (see Appendix A) of these videos up to a certain depth. The side-bar videos X_i were sampled using a Bernoulli distribution:

$$X_i \sim Bern\,(p_i)$$

$$p_i = .8 - (i * .7/19)$$

where i ={0,1,2...19} and represents the order of the video in the side-bar. The side-bar was not expanded to include additional related videos. This assumes that users are more likely to click a video at the top of the side-bar than at the bottom, and that the probability decreases linearly from .80 to .10.

With a starting video list of 15 videos and depth of 3, the snowball sampling procedure produced an initial collection of 841 videos (with repeats) that was then reduced to a set of 752 videos.

Web Scraping/YouTube API

Using the initial set of 752 videos, the following variables were obtained by web scraping YouTube: views, author, likes, dislikes, subscribers, description box, tags, title, date published, and video length (see Appendix A). Some values (views, likes, dislikes, subscribers) are dynamic, and thus the data set is technically valid as of 4:05 am on March 29th, 2016. Additionally, the YouTube API was used to obtain the comments on the makeup video. Videos with disabled comments were discarded, and neither the commenter's username nor the relation between comments was obtained. Because the process was computationally taxing, the number of threads retrieved from the video was capped, at around 300 threads, biasing the comments retrieved from the video. The algorithm implemented downloads threads that are most popular first, so the comments and threads that were not downloaded because of the imposed cap were likely early comments, or comments with no replies.

Thumbnail Images

Every YouTube video has four generated images (one full size image and three thumbnail images) which were scraped. These images were then run through a program, using OpenCV, that determined how many faces the images contained.[5] This algorithm was effective, but certainly not fully reliable: a quick analysis of images and predicted number of faces in the images showed that the algorithm was generally effective at determining faces that were straight on (see Figure 1), but not particularly effective at determining faces in profile, or mid-makeup application.

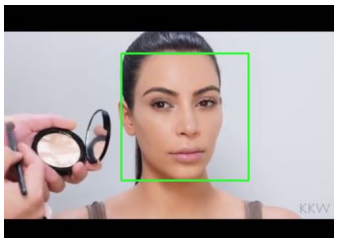


Figure 1: Example Face Detect

Auto-Generated Transcript

Most YouTube videos have auto-generated transcripts generated by Google voice.[9] They are not completely accurate transcriptions of the YouTube videos, and there are common errors, particularly with nouns and oftentimes with makeup brands. However, they are a good proxy for the actual content of the YouTube videos. These transcripts were downloaded using a command-line program, youtube-dl, and then converted to a text format, which discarded any information regarding the timing of words.[4]

Removing Emojis

Figure 2: Smile emoji

In order to prepare the comments, title, and description box for analysis, the texts were cleaned of emojis. The unicode strings indicating an emoji were replaced with the word describing the emoji, using an online data base.[1] For example, "smile" would replace "\U0001f604" which is the emoji in Figure 2. The data base provided descriptive titles for the image of the emojis rather than titles that described what the meaning of the emoji would be. Additionally, emojis that are able to be different colors/races were not indicated as such when replaced. For instance, a Caucasian running man was replaced with just "running man."

Bag Of Words

Added Words	
Word	Classification
fleek	Positive
yas	Positive
yaas	Positive
yasss	Positive
bae	Positive
slay	Positive
fierce	Positive
sick	Positive
obsessed	Positive

Table 1: Words added to the sentiment dictionary to reflect natural language on YouTube.

Additional variables were added to the data set by reducing the title, description box, transcript, and tags to a bag of words. These words were cleaned for stop words, stemmed, and transformed into a bag of words and their word frequencies. The bag of words for title, description box, and transcript were analyzed, and categorical variables were added to represent the presence of certain words within the top 5% of words by word frequencies (discarding certain words that were uninformative). A majority of these variables represented unigrams, such as "Maybelline" or "Subscribe." However, some bigrams were added, such as "provided by" or "sponsored by." Additionally, some word frequences were combined with others that indicated the same concept: the frequency of "facebook"and "f a c e b o o k" (a common formatting quirk used by gurus) was combined into one variable.

Sentiment Analysis

The description box, title, comments, and transcript were analyzed for their sentiment value. First, the words shown in Table 1 were added to the sentiment dictionary to reflect the natural language used by beauty gurus and viewers on YouTube, particularly in the comment section. Then, for each polarized word found in the sentiment dictionary, the value of that word, w, found in the sentiment dictionary is additionally weighted by the surrounding words. Essentially, for each word, a context cluster (x_i^T) is pulled from around the word to be used as valence shifters. The words in this context cluster are tagged as neutral (x_i^0), negator (x_i^N), amplifier (x_i^A), or de-amplifier (x_i^D).[13]

The polarity score of that context cluster is then equal to:

$$\sum((1 + c * (x_i^A - x_i^D)) * w(-1)^{(\sum x_i^N)})$$

where

$$x_i^A = (w_{neg} * x_i^A)$$
$$x_i^D = max(x_i^{D'}, -1)$$
$$x_i^{D'} = \sum(-w_{neg} * x_i^A + x_i^D)$$
$$w_{neg} = (\sum x_i^N)2$$

The overall polarity score then given to the word, C, then becomes:

$$C = x_i^T/\sqrt{n}$$

where n is the number of words in the context cluster.[13]

For each item of text, the text becomes an average of the polarity scores for each context cluster surrounding each polarizing word in the text. For the transcript, title, and description box, this average was used as a variable. For the comments, the average of these averages was used, in addition to the standard deviation.

2.2 Predicting Views

The variables were split into two groups: variables that are intrinsic to the video and were available at the time it was posted, and variables that are dynamic and change over time. The latter group consisted only of likes, dislikes, average comment sentiment and standard deviation of comment sentiment. As an additional note, the subscriber count of the YouTube guru is dynamic and changes over time, but was treated as an intrinsic variable solely because subscriber count is available at the time the video is posted, though it is likely not exactly the value scraped. The value scraped is therefore treated as a proxy for the subscriber count at the time of the video being published. Hereafter, the model containing only variables intrinsic to the video will be referred to as "Model A," and the model containing those variables as well as additional dynamic variables will be referred to as "Model B."

To predict views using the two sets of variables, a generalized linear model with a penalized maximum likelihood was used. The regression formula is shown below:

$$\min_{B_0,B} \frac{1}{N} \sum w_i l(y_i, B_0 + B^T x_i) + \lambda[(1-a)||B||_2^2/2 + a||B||_1]$$

Two parameters in the model are of interest to the analysis, the a value and λ value. The generalized linear model was run for $a = 0$ and $a = 1$, hereafter referred to as ridge regression and lasso regression. Moreover, the optimal λ value was optimized using 10-fold cross validation to find the λ that minimized mean squared error; a tool built in to the "glmnet" package used.[14] Thus, four predictive models were analyzed: Model A run with lasso regression and ridge regression, and Model B run with lasso regression and ridge regression.

Model Evaluation

Each model was assessed for predictive accuracy by using K-fold cross validation. The data set was divided into K parts, and each model was trained on K-1 parts, and tested on the remaining part. The models were analyzed with $K = 10$. The final cross validated mean squared error, $CV_{(10)}$ is a weighted average of the mean squared errors of each iteration:

$$CV_{(K)} = \sum_{k=1}^{K} \frac{n_k}{n} MSE_k$$

where:

$$MSE_k = \sum_{i \in C_k} (y_i - \hat{y}_i)^2 / n_k$$

Additionally, the estimate of the standard deviation of $CV_{(10)}$ was calculated as:

$$\hat{\sigma}_{CV_{(10)}} = \sigma_{MSE} * \sqrt{\sum_{k=1}^{K} (n_k/n)^2}$$

which is the standard deviation of a weighted average, in this case, with weights n_k/n. This is a biased estimate of the standard deviation of $CV_{(10)}$, and it likely underestimates the true standard deviation of the statistic.[3] However, there does not exist an unbiased estimator of the variance or standard deviation of K-Fold cross validation, so $\hat{\sigma}_{CV_{(10)}}$ will be utilized.[3]

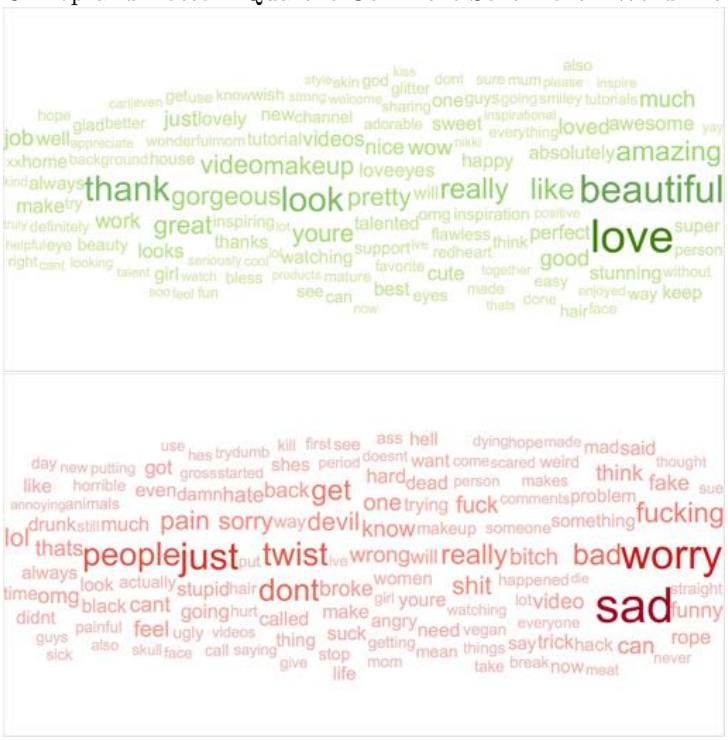
3 Results

Exploratory data analysis indicates that there are some biases in the data set due to the sampling procedure, and difficulty in identifying sponsorship in videos. These findings, and others as a result of exploratory analysis, are discussed in Section 3.1. The predictive capabilities and features selected by the lasso regressions for Model A and Model B are discussed in Section 3.2. The cross-validated MSE of the regressions indicate that no model is significantly better at predicting views.

3.1 Exploratory Data Analysis

Initial data analysis indicated that there was a group of highly influential beauty gurus within the sample; 9 beauty gurus, (out of 208 in the sample) accounted for more than 45% of videos, 13.7% of total subscribers, and 46.8% of total views in the sample. As an attempt to align this with the entire beauty community, it was determined that 6 of the 9 are ranked in the 2015 Top 25 YouTube Beauty Creator Channels by audience engagement

Figure 3: Top and Bottom Quartile Comment Sentiment - Word Frequency



(which measures likes, dislikes, comments, and shares).[12] It is unclear how much of this concentration is influenced by the sampling method or influenced by the actual dominance of certain beauty gurus on YouTube, but there is certainly a concentration of influential gurus in the sample.

It was also very difficult to determine whether videos were sponsored or not sponsored. The variables in question, "Not Sponsored" and "Potentially Sponsored," examine the video description boxes for certain phrases that indicate sponsorship, like "sponsored by" or "provided by", or lack of sponsorship, like "not sponsored" or "not a sponsored video". Ideally, most videos would be classified as either "Not Sponsored" or "Potentially Sponsored." However, in the sample, only 29 videos were classified as "Potentially Sponsored", 169 videos classified as "Not Sponsored", 3 videos classified as both, and 512 videos classified as neither.

This made it difficult to analyze whether sponsorship (or in this case "potential" sponsorship) had an impact on views. Anecdotally, beauty gurus often mention sponsorship only in the content of the video, and often so subtly that it is not even flagged with the word "sponsorship." This contributed to the difficulty in determining sponsored and not sponsored videos.

Additionally, for the few videos that were classified as both "Not Sponsored" and "Potentially Sponsored," there were conflicting ideas in the description box. For instance, for the video, "Rose Toned Makeup + Spring 2016 Newness — Melissa Alatorre," (one of the 3 videos dually classified) the description box reads, "this video is not sponsored. some of the items included in this video were sent to me," which does suggest a gray area of

sponsorship.

The analysis of comment sentiment indicated that the sentiment tool was fairly accurate at identifying extremely positive or negative comments, and word frequencies in the top and bottom quartile where significantly different. Figure 3 shows the word frequencies for the top and bottom quartile of comments, and the color and size of the word indicates its frequency. As shown, there is a clear difference in the top word frequencies; heuristically the top quartile definitely seems more "positive" than the bottom quartile. Additionally, some of the top words in the top quantile are emojis: "loveeyes" and "redheart" (see Appendix C for the emoji images). These emojis were not added to the sentiment dictionary, so are therefore associated with more positive comments.

Additionally, primarily because of computational issues, the only variables that were added regarding comment sentiment were average sentiment and standard deviation of sentiment, so some of the granularity of specific comments was lost. For instance, average sentiment across all videos is .11, minimum average sentiment is -.6030, and maximum average sentiment is .5158. However, the minimum sentiment value of all comments in the sample is -16.941 and the maximum sentiment value of all comments in the sample is 3.51. Naturally, using the average will eliminate outliers and smooth the differences, but there was a notably large variance in individual comment sentiment throughout the sample.

3.2 Predictive Accuracy and Relevant Variables

The results show that the predictive capabilities of lasso and ridge regressions for Model A and Model B overlap and are within a standard deviation of one and other. Table 2 shows the results of the cross-validated MSE (CV_K), an estimate of its standard deviation ($\hat{\sigma}_{CV_K}$), as well as the optimized λ value. The dependent variable, views, was scaled to be in millions of views, so the results should be interpreted as such.

Model A

Cross-Validated Mean Squared Error				
Model	Regression	λ	CV_K	$\hat{\sigma}_{CV_K}$
Model A	Lasso	0.114	5.25	1.82
Model A	Ridge	2.52	5.29	1.78
Model B	Lasso	0.423	4.38	1.44
Model B	Ridge	2.31	3.85	1.37

Table 2: Shows the different combinations of models and regression types, with cross-validated mean squared error and the standard deviation of that statistic.

The results indicate that for Model A there are not any variables that are particularly dominant in the data set, as far as predicting views is concerned, and the CV_K values for both regressions are not significantly different from one another. The lasso regression selected 28 out of 106 variables; this could suggest that groups of variables were highly correlated, and lasso chose to eliminate all but one variable in the group (see Appendix G). It is, however, still meaningful to analyze the features selected by Model A lasso regression.

The results of the lasso regression are influenced by both the sampling methodology and the actual underlying signal in the data. It appears that 2013 is associated with higher views than 2016, which is contradictory to most external analyses of the YouTube beauty

community.[12] Upon analyzing the videos that were from 2016 (n= 105) and 2013 (n=119), the videos sampled from 2013 have more views, but this appears to be a function of the sampling methodology. YouTube's "Related Videos" section is comprised using an algorithm that essentially looks at what videos users would naturally view next.[8] They recently updated their algorithm to include views as well as session time (the amount of time a user watches a video) to create the "Related Videos" section.[8] Thus, if there were videos from 2013 in the "Related Videos" section, they most likely are extremely popular videos from 2013, because it is less likely that a user would naturally view an older video. This appears to a be a bias due to the sampling methodology.

Another interesting finding is regarding the presence of "Like", "Follow", and "Subscribe" in the description box, which is a technique used by gurus to remind viewers to engage in the video/channel. This was a common variable; 566 videos had either "Like", "Follow" or "Subscribe" in their description box, and 105 videos had all instances in their description box. All three features were selected by Model A lasso regression, and were all associated with lower views. The presence of an "FTC" notice (Federal Trade Commission notice) was also associated with lower views. Some videos used the FTC notice to declare non-sponsorship (n=44) and few used the FTC notice to declare sponsorship (n=5). Finally, the only tags selected were "tutorial", "cosmetics", "eye", and "smokey"; all but "cosmetics" were associated with higher views.

Model B

The inclusion of likes, dislikes, average comment sentiment, and the standard deviation of average comment sentiment did not result in either a lasso or ridge model that was significantly better at predicting views. Additionally, the lasso regression for Model B confirmed the dominance of likes and dislikes (coefficients of $1.14*10^{-5}$ and $3.42*10^{-4}$ respectively) in predicting views, as those were the only features selected among the 106 features provided. The relationship between likes/dislikes and views is obviously highly correlated, because a like or dislike automatically registers as a single view, and this predictability dominated all variables in the data set.

Additionally, when likes and dislikes were removed but all other variables kept in, both average comment sentiment and standard deviation of average comment sentiment were features selected by the lasso regression, with coefficients of $-3.64*10^{-4}$ and $1.76*10^{-4}$ respectively. This appears to indicate that videos that were more "controversial" (more negative comments and more variance in the comments) are associated with higher views. However, as mentioned in Section 2.1, the imposed limit of comment threads downloaded likely biased the sentiment.

Predicting Viral Videos

The results of all four predictive models also exhibit an interest trend; predictability becomes less accurate as views increase and as the ratio of views to subscribers increases. This seems to indicate that the more "viral" a video is, the harder it is to predict views with the variables available.

Though there does not exist a universal definition of a "viral" video, a common definition is a video that receives 5 million views in 3-7 days.[11] Because the data set did not include any information about when the views were received, a proxy for a video being viral

was if the video had more than 5 million views.

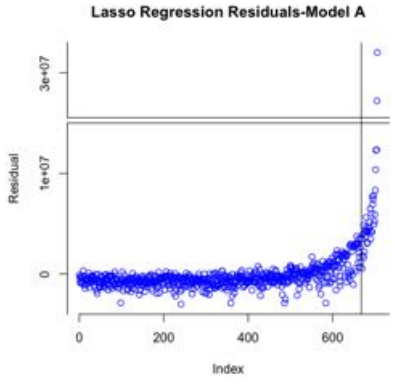


Figure 4 shows the residuals for Model A lasso, where the index of videos is ordered by total views, and the vertical line marks 5 million views (views scaled back to normal units). As shown, there is a clear spike in residuals as videos have more than 5 million views and become more "viral". This was the case for all regressions run (see Appendix E). Additionally, the ratio of views to subscribers offers another nuanced statistic of a video being viral for a specific channel. An analysis of residuals against the ratio of views to subscribers also shows that predictability is hampered by a video being viral (see Appendix F).

Figure 4: Residuals, videos indexed by views

4 Discussion

The final section will elaborate on some limitations of the analysis mentioned throughout the paper. These limitations are primarily in the sampling method, as well as the analysis of text variables, such as comments and description box. Additionally, future work will be discussed, as it pertains to further relevant analysis in the YouTube makeup community.

Limitation of Sampling Method

There are a number of limitations in the analysis conducted that encourage future work. The initial limitation with this analysis is obtaining an appropriate sample of YouTube makeup videos. While snowball sampling was an effective way to sample videos as far as the user perspective is concerned, it has a number of biasing qualities. There is an inherent sampling bias with snowball sampling, based on the original respondents, or in this case, link set.[2] It is therefore very difficult to make generalized statements about the YouTube makeup community as a whole while using snowball sampling.

Additionally, the snowball sampling method used relies on the connection between nodes (in this case starting links), and therefore undervalues videos that are less "interconnected."[2] The impact of this in the makeup space is that beauty gurus who might represent more isolated, marginalized portions of the population (racial/ethnic minorities, members of the LGBTQ+ community) may not be represented in the sample. Moreover, as mentioned earlier, the YouTube "Recommended Videos" algorithm biases the videos on the sidebar to represent videos that are interconnected, and that are popular in relation to the current video. This biases the sample to generally popular videos, or videos that are "sticky" and encourage long session times.

Finally, it is difficult with this particular implementation of snowball sampling to determine which videos actually qualify within the guidelines of the thesis. Because there are no specific demographic qualities of a video that make it a "makeup video" (at least with the data available to gather), any culling of the final sample had to be done manually, which was computationally inefficient.

Limitation of Text Analysis

There were also a number of limitations regarding language processing and sentiment analysis. First, the language used on YouTube both by gurus and subscribers was very difficult to process, as it was comprised largely of slang, acronyms, and invented spelling. This made it quite difficult to judge the sentiment of items of text, even with the words added to the sentiment dictionary. Additionally, while emojis were replaced with their appropriate "description," this was not a complete solution, and does not often accurately indicate what an emoji means on YouTube. For instance, "redheart" (see Appendix B) does not necessarily relay the meaning of a red heart emoji; perhaps "like" or "love" would be more appropriate.

Secondly, there were a number of computational and data issues regarding language processing. Processing the transcripts from the videos did not produce nearly exact transcripts of the videos. There were a number of visible errors in the auto-generated captions, and although there were distinct patterns in word frequency, it was difficult to make more nuanced analysis. This was particularly difficult when trying to asses the number of times a brand was mentioned in a video; this data was available, but sparse likely due to the imprecise transcripts. Additionally, for computational reasons as stated earlier, the number of comment threads had to be capped as around 300, which biased the text obtained to likely be the most "contentious" comments of the comment threads.

Finally, as mentioned earlier, there was a fundamental difficulty in determining whether or not a video was sponsored. While some videos used a Federal Trade Commission notice (often indicated by "FTC:"), others simply mentioned in the description box that the video was sponsored by thanking the sponsoring company. Most problematic, however, were the videos whose information regarding sponsorship solely resided in the video itself, which made it very hard to identify. Not only were the transcripts not ideal quality, but information regarding sponsorship in a transcript would be framed as a sentence. As functionally only unigrams were analyzed, this made it particularly difficult to parse out information regarding sponsorship.

Future Work

The current analysis identified areas in which future work regarding the YouTube makeup community would be useful. Beyond rectifying the issues with the sampling methodology and text analysis, it would be very useful to create a classification tool that would be able to predict whether or not a video was sponsored. Such a logistic regression would work off of variables such as "Not sponsored", "Potentially sponsored", "FTC notice" and "Sponsored- transcript." The primary difficulty in this would be to easily determine whether or not a video is sponsored, locating highly accurate transcripts of videos, as well as using a text analysis methodology that incorporated sentences and phrases rather than just unigrams and bigrams.

Additionally, it would be interesting from a brand and advertisers point of view to

look at dependent variables other than views. While views are certainly important, they do not necessarily measure the audience's engagement with a particularly video, like the amount of time spent on the video, any comments, likes, or even shares on Facebook and Twitter can.

Finally, it would be very useful to create a predictive model for whether a not a video will go viral, based on inherent characteristics of the video. This model would be highly useful for gurus, brands, and truly anyone involved in the makeup space. As mentioned earlier, a common definition for a viral video is a video that gets more than 5 million views in a 3-7 day period. However, there are other factors that influence what specifically is a viral video: buzz, how much people are talking about a video on other social media sites, parody, how much the video is being imitated or "memed" by other sites, and longevity, how long after the initial spike in views is the video mentioned in popular culture.[11] Thus, one would need to access information from Reddit, Facebook, Twitter, talk shows, cable TV, etc. to properly inform the predictive model. However, such a model could be well utilized by gurus to create consistently relevant content.

Acknowledgement

I'd like to thank Professor Colin Rundel for really helping me with the web scraping, YouTube API work, and all of my computational issues. I'd also like to thank Professor Çetinkaya-Rundel, Professor David Banks, and Professor Galen Reeves for serving on my thesis committee. Finally, I'd like to thank Professor Sayan Mukherjee for advising me and directing me throughout this entire process.

References

[1] The unicode standard. http://unicode.org/emoji/charts/full-emoji-list.html, 2016.

[2] Rowland Atkinson and John Flint. Accessing hidden and hard-to-reach populations: Snowball research strategies. *Sociology at Surrey*, 33, 2011.

[3] Yoshua Bengio and Yves Grandvalet. No unbiased estimator of the variance of k-fold cross-validation. *Journal of Machine Learning Research*, 5, 2004.

[4] Daniel Bolton. Youtube-dl. https://github.com/rg3/youtube-dl, 2015.

[5] G. Bradski. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] Bustle Magazine. *The One Men's Beauty Product You Need In Your Makeup Routine Right Away*, 2015.

[7] Gaby Dunn. Get rich or die vlogging: The sad economics of internet fame. *Fusible*, 2015.

[8] Dane Golden. How to optimize youtube related videos. 2015.

[9] Ken Harrenstien. Automatic captions in youtube. 2009.

[10] Dhani Mau. How the fastest-rising beauty vloggers found youtube success. *Fashionista*, 2014.

[11] Megan O'Neill. What makes a video "viral"? 2011.

[12] Pixability, Boston, Massachusetts. *Beauty on YouTube 2015*, 2015.

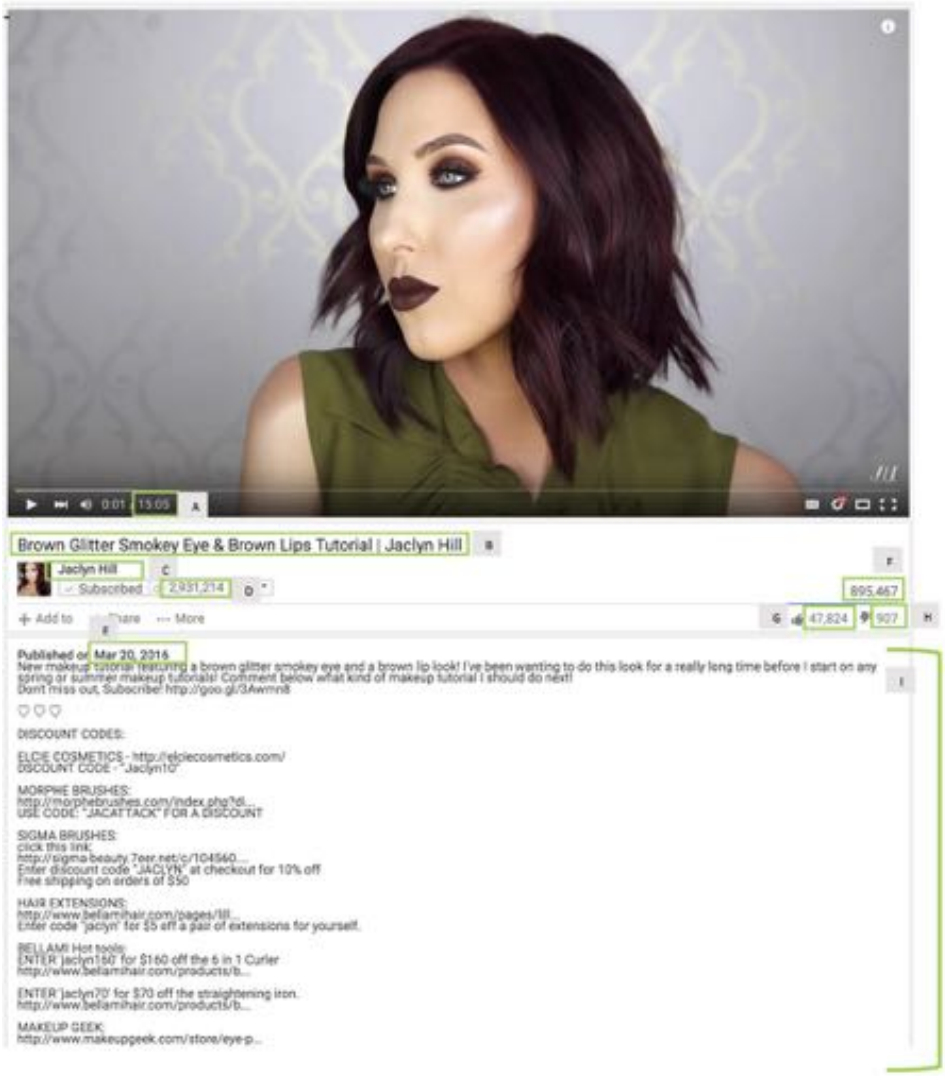
[13] Tyler W. Rinker. *qdap: Quantitative Discourse Analysis Package*. University at Buffalo/SUNY, Buffalo, New York, 2013. 2.2.4.

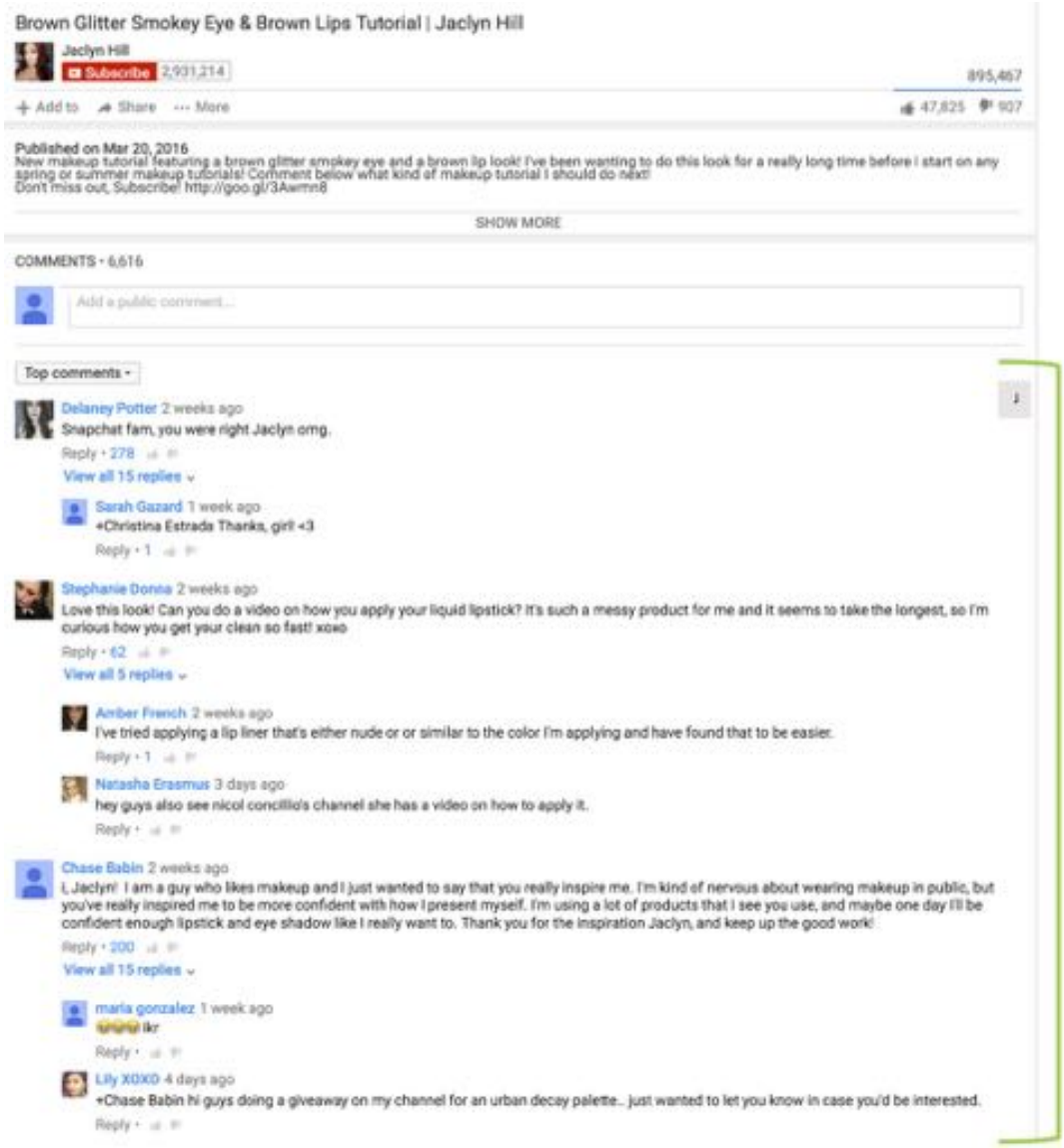
[14] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.

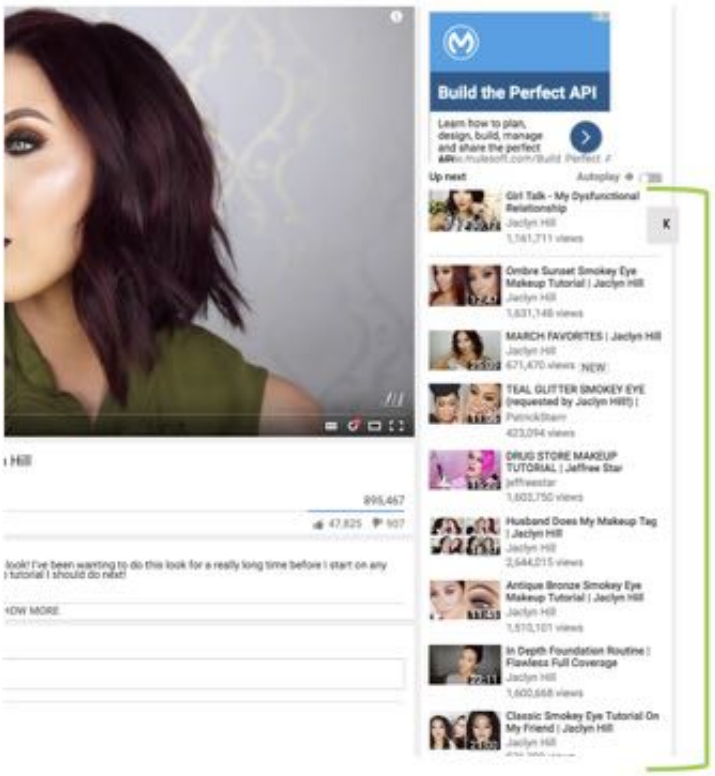
Appendices

A Example Video Screens

A Length	D Subscribers	G Likes	J Comments
B Title	E Date	H Dislikes	
C Author	F Views	I Description box	K Related Videos




Brown Glitter Smokey Eye & Brown Lips Tutorial | Jaclyn Hill
Jaclyn Hill
Subscribe 2,931,214
895,467
Add to Share More
47,825 907
Published on Mar 20, 2016
New makeup tutorial featuring a brown glitter smokey eye and a brown lip look! I've been wanting to do this look for a really long time before I start on any spring or summer makeup tutorials! Comment below what kind of makeup tutorial I should do next!
Don't miss out, Subscribe! http://goo.gl/3Awmn8
SHOW MORE
COMMENTS • 6,616
Add a public comment...
Top comments
Delaney Potter 2 weeks ago
Snapchat fam, you were right Jaclyn omg.
Reply • 278
View all 15 replies
Sarah Gazard 1 week ago
+Christina Estrada Thanks, girl! <3
Reply • 1
Stephanie Donna 2 weeks ago
Love this look! Can you do a video on how you apply your liquid lipstick? It's such a messy product for me and it seems to take the longest, so I'm curious how you get your clean so fast! xoxo
Reply • 62
View all 5 replies
Amber French 2 weeks ago
I've tried applying a lip liner that's either nude or or similar to the color I'm applying and have found that to be easier.
Reply • 1
Natasha Erasmus 3 days ago
hey guys also see nicol concillio's channel she has a video on how to apply it.
Reply •
Chase Babin 2 weeks ago
I, Jaclyn! I am a guy who likes makeup and I just wanted to say that you really inspire me. I'm kind of nervous about wearing makeup in public, but you've really inspired me to be more confident with how I present myself. I'm using a lot of products that I see you use, and maybe one day I'll be confident enough lipstick and eye shadow like I really want to. Thank you for the inspiration Jaclyn, and keep up the good work!
Reply • 200
View all 15 replies
maria gonzalez 1 week ago
ikr
Reply •
Lily XOXO 4 days ago
+Chase Babin hi guys doing a giveaway on my channel for an urban decay palette.. just wanted to let you know in case you'd be interested.
Reply •


Build the Perfect API
Learn how to plan, design, build, manage and share the perfect
Up next
Autoplay
Girl Talk - My Dysfunctional Relationship
Jaclyn Hill
1,161,711 views
Ombre Sunset Smokey Eye Makeup Tutorial | Jaclyn Hill
Jaclyn Hill
1,631,148 views
MARCH FAVORITES | Jaclyn Hill
Jaclyn Hill
671,470 views
NEW
TEAL GLITTER SMOKEY EYE (requested by Jaclyn Hill!) |
PatrickStarrr
423,094 views
DRUG STORE MAKEUP TUTORIAL | Jeffree Star
jeffreestar
1,603,750 views
Husband Does My Makeup Tag | Jaclyn Hill
Jaclyn Hill
2,644,015 views
Antique Bronze Smokey Eye Makeup Tutorial | Jaclyn Hill
Jaclyn Hill
1,510,101 views
In Depth Foundation Routine | Flawless Full Coverage
Jaclyn Hill
1,600,668 views
Classic Smokey Eye Tutorial On My Friend | Jaclyn Hill
Jaclyn Hill
Hill
895,467
47,825
907
look! I've been wanting to do this look for a really long time before I start on any tutorial I should do next!
HOW MORE

B Starting Link List

VideoID	Title	Author
foSHlt3rFaU	Get UNReady W/ Me! My Night Time Routine — Blair Fowler	Blair Fowler
AdpPDpM1tIA	Daytime Glam For Every Woman - Makeup Tutorial — Jaclyn Hill	Jaclyn Hill
uCingWMa_ek	Grunge Glam Makeup Tutorial	Carli Bybel
LS6gI67U_fw	Everyday Makeup Tutorial for HOODED EYES — Stephanie Lange	Stephanie Lange
vIOLEujuEJk	INDIAN BRIDAL MAKEUP TUTORIAL - GREEN and GOLD GLITTER EYES	Binny Khan
s1ApPspRLs4	Kylie Jenner Makeup Tutorial — Tori Sterling	Tori Sterling
lmDVWGX0pv0	How To Color Correct: Color Correction Makeup Tutorial	AlexandrasGirlyTalk
AZYpg1_Sd54	"Parisian Night Look from "Rouge in Love"""	Michelle Phan
Hh_HeFjrH8k	Get Ready With Me — COPPER EYE MAKEUP	Cydnee Black
7_rR06bo9Jc	COBALT BLUE SMOKEY EYE — DESI PERKINS	Desi Perkins
3qU_7o6dgi8	QUICK HOLIDAY MAKEUP — Talk Thru Tutorial	Tati
UNI-HXwR9rc	Get Ready With ME / Go-To Fall Makeup Tutorial! — Casey Holmes	Casey Holmes
APNVNQr2Ris	Huge Collective Makeup Haul_ Mac -NARS-Nordstrom-Colourpop & More	Shaaanxo

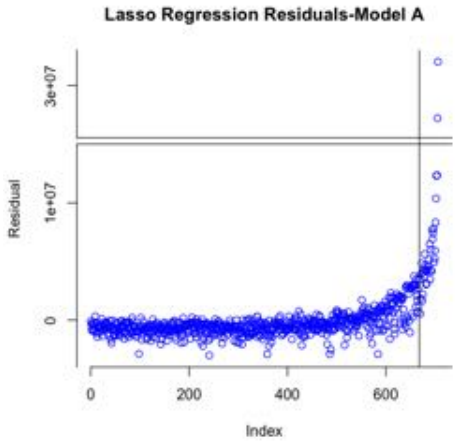
C "Loveeyes" and "Redheart" emojis

D Final Variable List

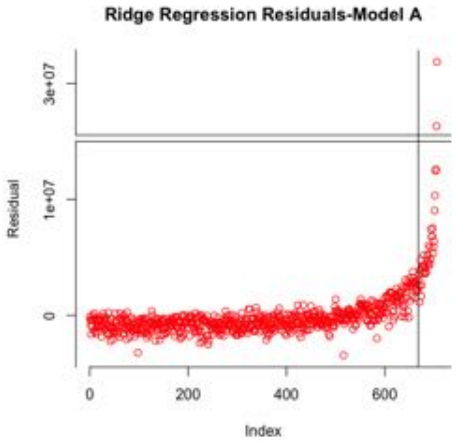
- Views
- Title sentiment
- Des. sentiment
- Ave comment sentiment
- SD comment sentiment
- Transcript sentiment
- Likes
- Dislikes
- Day published
- Month published
- Year published
- Description box
- Subscribers
- Video length
- Author
- Day of week
- "Get Ready With Me" in title
- "Makeup" in title
- "Fall" in title
- "Routine" in title
- "Drugstore" in title
- "Tutorial" in title
- "Favorite" in title
- "FTC " in des
- "Not sponsored" in des
- "Sponsored by" in des
- "Affiliate links" in des
- "Discount"in des
- Number of discount in des
- "Instagram" in des
- "Facebook" in des
- "Snapchat" in des
- "Twitter" in des
- "Thanks" in des
- "Subscribe" in des
- "Like" in des
- "Follow" in des
- "Tutorial" in tags
- "Beauty" in tags
- "Cosmetics" in tags
- "Get Ready" in tags
- "Beauty" in tags
- "Lipstick" in tags
- "Routine" in tags
- "Contour" in tags
- "Eye" in tags
- "Beauty" in tags
- "Smokey" in tags
- "Haul" in tags
- "Natural" in tags
- "Kylie" in tags
- "Glam" in tags
- "Beauty" in tags

E Residuals for all Models, Index sorted by Views

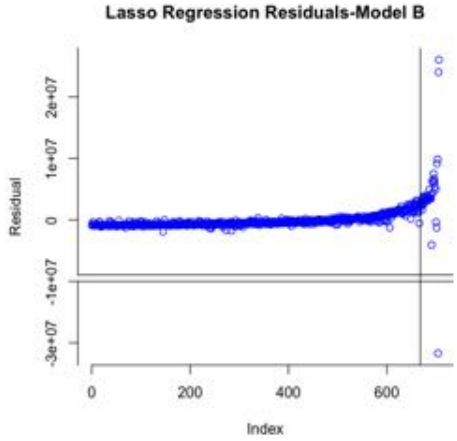
Vertical line at 5 million views.



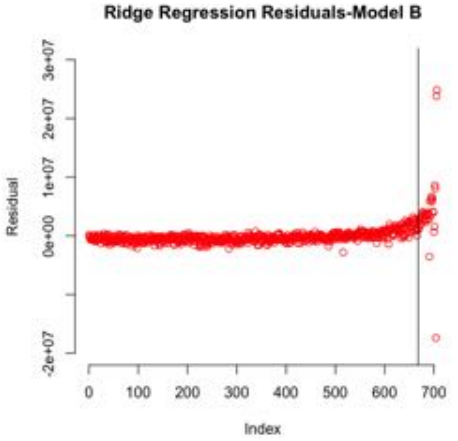
a



b

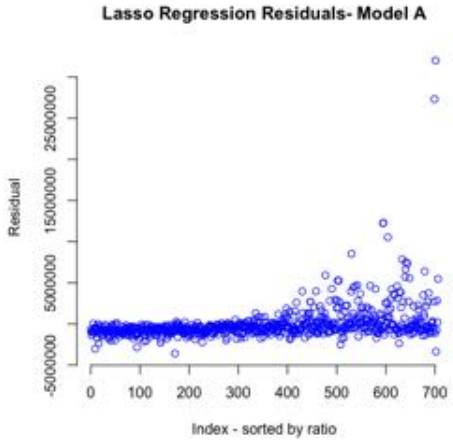


c

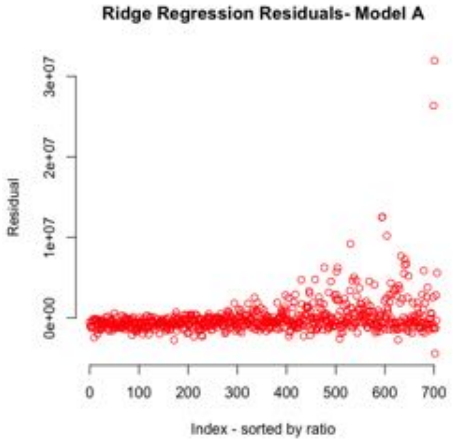


d

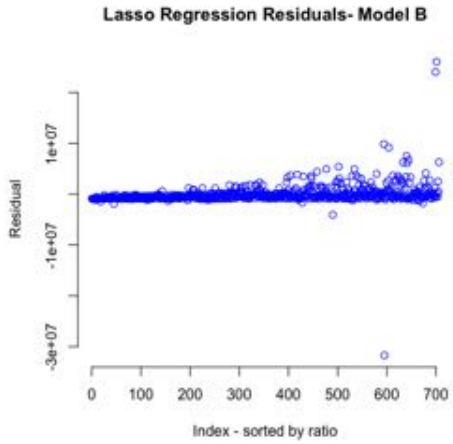
F Residuals for all Models, Index sorted by Views/Subscribers



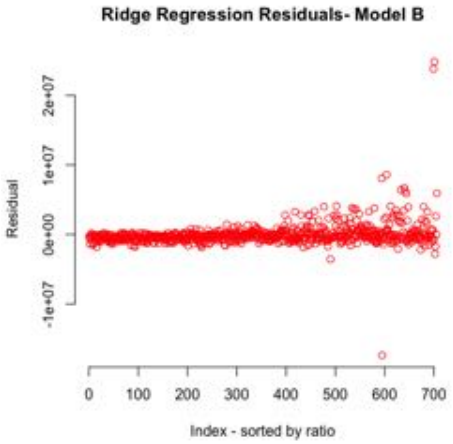
a



b



c



d

G Model A Lasso: Selected Features and Coefficients

Variable	Coefficient
intercept	1.598
subs	0.000002
length	-0.0001
titlesen	-0.305
all_faces	-0.071
contour_tran	0.066
year2013	0.134
year2016	-0.391
month02	0.175
month03	-0.060
month06	0.155
month07	0.275
month08	0.750
month12	-0.306
day12	0.091
day13	0.665
day19	0.873
day30	0.180
FTCTRUE	-0.199
discountpercentTRUE	-0.396
fbookTRUE	-0.059
likeTRUE	-0.037
subscribeTRUE	-0.145
followTRUE	-0.194
socialTRUE	0.136
tagtutorialTRUE	0.006
tagcosmeticsTRUE	-0.233
tageyeTRUE	0.238
tagsmokeyTRUE	0.899