YouTube Video Analysis for Buzzfeed Tasty

By Nandini Basu, Linyan Dai, Maxine Li

In partial fulfillment of the course BAX 452

Date of Submission: 3/15/2020

Table of Contents

Executive summary

Buzzfeed Tasty is an international media production channel that routinely produces high-quality videos that are primarily related to food and cooking. These videos are hosted on their website, Facebook and YouTube channels. They currently have over 1.7 million subscribers on YouTube. We wanted to help BuzzFeed Tasty evaluate the various factors that impact the number of views that a video gets. Having more views enables Buzzfeed Tasty to not only get more ads on their platform, but also gives them the leverage to negotiate a better price for the ads and sponsorships. Thus BuzzFeed Tasty could increase video views without spending any more money on the video production.

Through web scraping and Youtube API, we got various variables related to a video such as the video's title, description, length, likes, dislikes, captions, comments and views. With this data we were able to perform a lasso linear regression to understand the variables factors that drive the number of views that a video gets. Of our initial 171 variables, the lasso regression selected 57 significant variables. We found that the length of the video is a major factor in determining the video views. We were also able to identify certain words for the title and description that result in more video views. Finally were also able to identify key topics which drive video views. Based on these results we were able to make tangible recommendations to Buzzfeed Tasty. However we do need to keep the limitations of these modeling techniques in mind. Most importantly, correlation doesn't imply causation hence we recommend that Buzzfeed Tasty experiment with our suggestions to better understand their impact.

Introduction

Buzzfeed is a digital media company that has a worldwide following[1], with a web-series on comfort food. In the past five years, the series has produced thousands of videos focused on food and food-related content.[2] These videos have different formats. Some are quick recipe tutorials without any voice overs, some are recipes with detailed voiceover instructions and the others include chefs sharing their favorite foods to eat or cook. While BuzzFeed has many competitors in the media space, such as Vice Media, Mashable, Mic Network and Boston Globe Media, they are still market leaders in the food content creation vertical.

All of Buzzfeed Tasty's videos are produced by an experienced production team and are of high quality. Hence, Buzzfeed Tasty invests significant time, money and energy in producing these videos. These videos have indeed been well received by viewers and Buzzfeed Tasty currently has 1.7 Million subscribers on YouTube and their videos get hundreds of thousands of views. This viewership creates an opportunity for Buzzfeed Tasty to generate revenue by showing video ads on YouTube. Hence the organization is constantly trying to understand what it can do to increase viewership. We want to help Buzzfeed Tasty identify the factors, aside from the actual video production quality, that influence the number of views a video gets. In order to do so, we would have to collect various data points on its existing videos. The end goal would be to understand what the organization can do to organically[3] increase the number of views a video gets.

1*About BuzzFeed*, www.buzzfeed.com/about.
2Ting, Deanna, et al. "With Tasty, BuzzFeed Has a Multi-Revenue Stream Model." *Digiday*, 10 Dec. 2019, digiday.com/media/tasty-buzzfeed-multi-revenue-stream-model/.
3 An organic view is a view that is generated without the use of any sort of promotion.

Data characteristic

We were able to extract the video links for 1300 videos. We chose only 1300 videos because we only wanted to consider the videos produced in 2019. This would allow us to control the quality of video production. Once we had collected all the data, we decided to store it in MongoDB. This was an obvious choice since we were storing comments and captions and didn't need a relational database.

Of the variables we had collected, we decided to use the **video title, description, length, likes, dislikes** and **comments** and quantify their effect on the video views. These variables are divided into text and quantitative data. Most of the quantitative variables have right skewed distributions that need transformation for further analysis, except that the number of comments seems normally distributed. After variable transformation, all these variables have positive correlation with the number of views [Table 1]. Among them, likes, dislikes and length of videos all show apparent positive correlation with the views[Table 2-4], while the relation between like over dislike ratio and the views is relatively random[Table 5].

Method

In order to proceed we had to treat these variables so that they can be inputted into a regression model. For the title and description we decided to use the **bag of words** which is a way of representing text data using a vector. It allows us to extract features from text so we can quantify if these features are responsible for driving video views. It captures the occurences of words in a document, in this case the titles and descriptions. For the title, we decided to not remove the stopwords as one usually does because we felt that these would be meaningful in the video titles. For example, 'I' could be used to make people feel closer to the video maker and 'to' shows intention. These can all be words that attract people to click on the video. Also, we

decided to not remove stop words because the titles were anyway short. The only two words removed from word vector were '•' and 'tasti', which show up in almost all titles. We picked the 100 most frequent word from title and removed '•' and 'tasti' from it leaving us with 98 word variables for analysis. Also, since there were a lot of numbers in titles such as 'xx ways to do something', we decided to replace all numbers with 'NUMBER' using Regex. In this way, we can analyze how numbers in titles affect views.

We decided to log the views and the length of the video because the data was skewed and we wanted to normalise it. We also realised that likes and dislikes were both positively correlated with the views which is intuitive because a video with more views would have more engagement in general. In order to use these variables in the model, we decided to calculate the like to dislike ratio. This would represent how positively viewers have responded to the video and we can then use this feature in our final regression model. Finally, we decided to do topic modeling on each video's comments to better understand the content of each video and if this would have an impact on the video views. We had initially planned on combining the video captions with the comments but only very few videos had captions. Also, we did not want to combine user generated and business generated variables. We wanted to understand how users were responding to the videos and use the content of their responses as an input in our regression model. We were able to come up with 4 topics for the comments. Topic one focuses on the one of the Tasty chefs, Rie and desserts. The second topic focuses on chef Alix and chef Rie and the weekend. The third topic focuses on chicken, cheese and cream so maybe indulgent foods. The last topic is about che Rie again but this time with savory food.

Model and analysis

We wanted to run a linear regression model to better understand the effect of our independent variables on the video views. Once we had finished transforming the variables we had 171 independent variables and only around 1300 observations. Hence we had to do variable selection and decided to do a lasso CV to identify only the most influential variables for the model. Lasso puts a constraint on the number of model parameters. This constraint results in the regression coefficients shrinking towards zero. We put all variables we got from the bag of words for titles and description, and topic modeling for comments along with video length and like to dislike ratio into a lasso regression. After lasso, we were left with only 57 variables. These can be divided into variables that have a positive influence on views and variables that have a negative influence on views. The model had a 58% out of sample R-squared.

Limitations

Our biggest limitation with using a linear regression is that we're only measuring correlation and not causation. However it would be meaningful for Buzzfeed Tasty to experiment with our recommendation to see if these actually do increase the video views. Additionally, the following are some of the limitations associated with some of the modeling techniques we have used:

Limitations for Bag of Words:

1. It ignores the location of the word: The location of words can be really important. Different orders of words can have very different meanings. For example, 'French toast made by chef' and 'chef eat things made by French toast' can have different meanings

but they might have the same word vector. Even though they attract different audiences, the algorithm can not catch it.

2. It can not tell from common words and keywords: Common words like ‘and’,’the’ and keywords like ‘pizza’ may look the same to the algorithm because they both appear many times. But they have different importance, keywords clearly are more important that common words. So we have to manually pick the right stop words.

Limitations for LDA:

1. It is hard to choose the K for topics: Since this is unsupervised learning, we don’t know how many topics are there in all the documents, we have to test out different K and see which K makes the most sense.
2. The results may be different every time we run: Same with K-Means, it starts by randomly assigning topics. So different start points may lead to different results.
3. Same problem as with the bag of words, it ignores the orders of word.
4. It is static, so it does not have any evolution of topics over time: People might have new thoughts when talking about the same topic, so the word for each topic might change overtime.
5. It can not capture the correlation of topics: It only shows that there are different topics, but sometimes several topics are related to each other.

Recommendations

Hence we can accurately pinpoint the variables that are likely to result in an increase in video views. The top ten variables that had the largest coefficients and hence resulted in the largest increase in video views were the words “giant”, “snack”, “i”, “vs.”, “made” and “?” in

the title and the words “networkfx”, “season” and “audiobook” in the description. Hence we would encourage Buzzfeed Tasty to use more of these words in their titles and descriptions. The variables with the lowest negative coefficients were the words “LG”, “as”, “by”, “pie”, “and” and “!” in the title and the words “shop”, “holiday”, “merch” and “get” in the description. Topic one that focuses on chef Rie and desserts was also in this list. Hence we would recommend that Buzzfeed Tasty reduce the usage of those words in their title and description and also maybe make fewer videos of Rie making desserts. Finally we also found that the video length has the highest coefficient which would mean that longer videos on average generate more views. Hence we would encourage Buzzfeed Tasty to make longer videos as this is what views seem to prefer. Additionally, we also found that Topic 3, which focuses on chicken, cheese and cream has a positive coefficient and we would encourage Buzzfeed Tasty to produce more videos on these items. Tasty also sells its own merchandise such as t-shirts, and whenever they advertise for their merchandise, they will put a link in description so people can buy their product for the link. ‘descrip_shop’ , ‘descrip_holiday’, ‘descrip_merch’ and ‘descrip_get’ are all related to the merchandise and it shows from the result that it has a significant negative influence on views.

Conclusion

In conclusion, there are many ways that Buzzfeed Tasty can increase the number of views a video gets. They can focus on creating longer videos as these have more views and they can also add certain keywords in their titles and description to act as a ‘click bait’. By using our recommendation, Buzzfeed Tasty can increase their video views without investing more money or effort in the actual video production. As mentioned before, an increase in video views will enable Buzzfeed Tasty to not only get more ads on their platform but also negotiate a better price for those ads. Hence our recommendations have tangible monetary gains associated with them.

Appendix

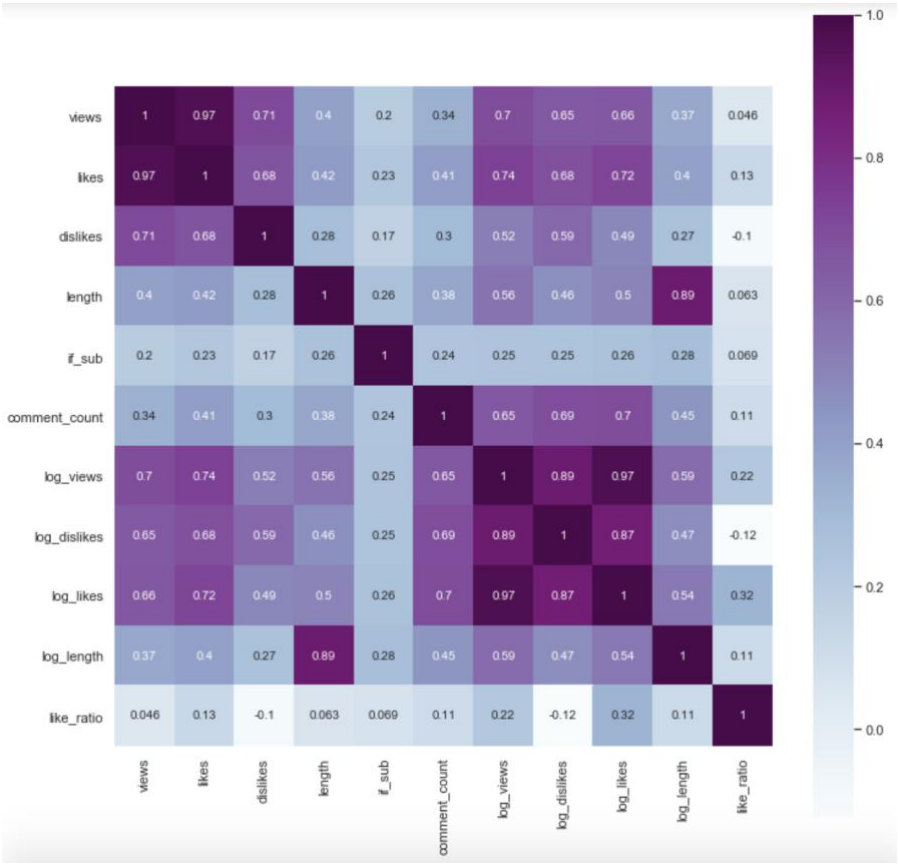

Table 1

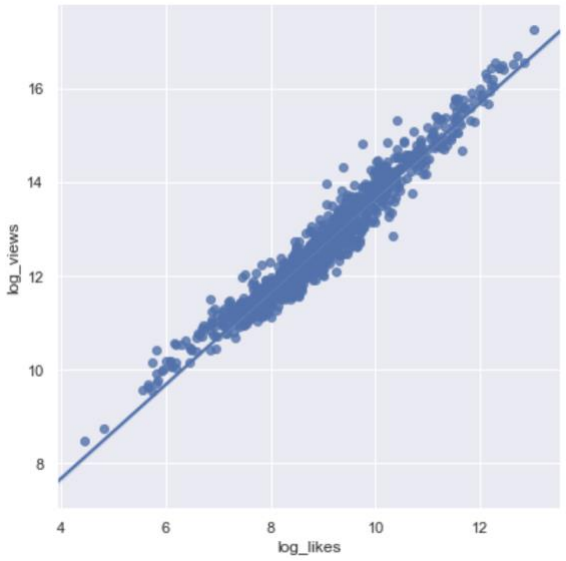

Table 2

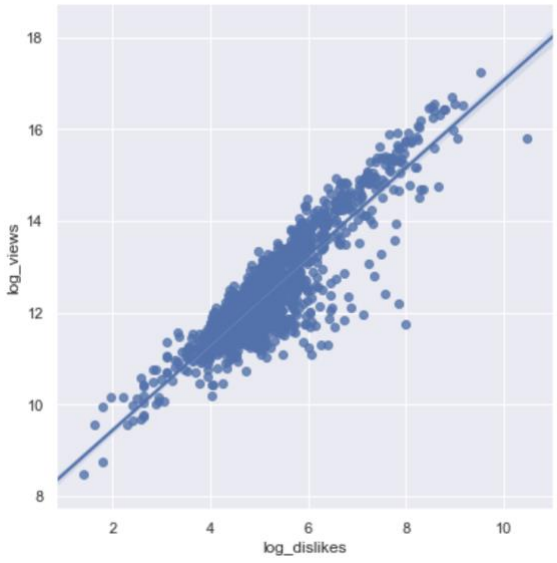

Table 3

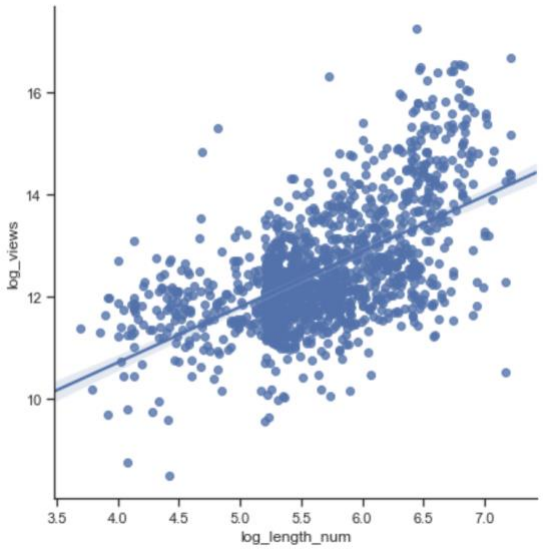

Table 4

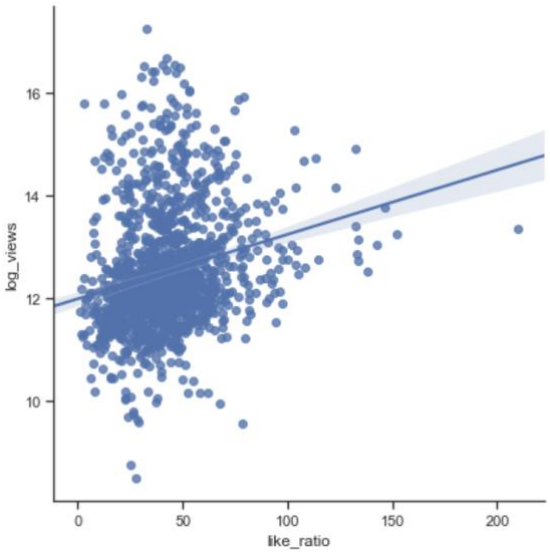

Table 5