# An Analysis of YouTube Trending Videos

A Fall 2021 ORIE 5741 Project Midterm Report

Jeremy Wang (jw2363) and Lorenzo Scotto di Vettimo (ls769)

YouTube Data Analytics
Cornell University
Ithaca, NY 14850

Dear YouTube Data Analytics Team,

We hope all is well with you. We wish to update you on the progress of our project by offering a summary of the dataset with a few preliminary analyses and a plan for how the project will be developed in the upcoming month.

### 1. Insights about Dataset

The dataset is a daily record of up to 200 top trending YouTube videos per day collected using the YouTube API. As the data included offers various regions including Great Britain, India, and Japan, we choose to focus our study to daily trending videos in the USA to narrow our focus analysis. For the purposes of our exploration, we will refer to the dataset as YouTube trending videos within the US. There are a total of 89791 example videos, all with 16 features. These features are described as follows divided into the appropriate category:

**Text Features**

Feature 1. Video_id - Video ID associated with video that can be used to access link to video
Feature 2. Title - name of the YouTube video posted at the point video is trending
Feature 4. ChannelID - ChannelID associated with the video uploader
Feature 5. ChannelTitle - Name of video uploader channel
Feature 8. Tags - Tags the users decides to use for the video
Feature 13. Thumbnail_link - A link to a picture of the thumbnail for the video
Feature 16. Description - The description of the video written by the uploader

**Numerical Features**

Feature 9. View_count - The number of views the video has to reach trending
Feature 10. Likes - The number of likes the video has once reached trending
Feature 11. Dislikes - The number of dislikes the video has once reached trending
Feature 12. Comment_count - The number of comments the video has once reached trending

**Boolean Features**

Feature 14. Comments_disabled - A boolean value to indicate if comments were disabled
Feature 15. Ratings_disabled - A boolean value to indicate if the ratings were disabled

**Continuous Features**

Feature 3. PublishedAt - Date and time when video was published
Feature 7. Trending_Date - Date video reached “Trending” status on YouTube

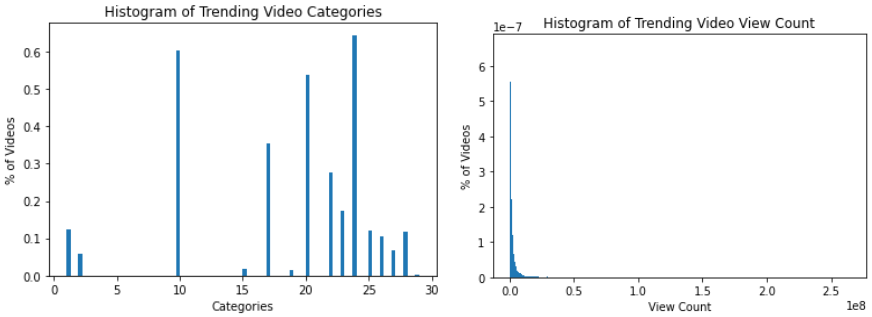
**Categorical Features**

Feature 6. CategoryID - Integer digit used to refer to one of 31 possible categories of videos

We know we have some missing data by counting the amount of NaN and Null values in the dataset. These only occur within the “Description” feature, and this is most likely due to content creators not adding any words to the description. This is the case for 1.5% of the data (a total of 1316 videos). As the

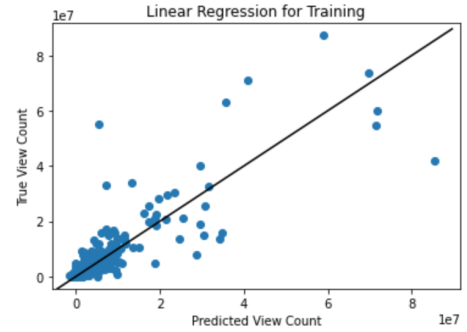
dataset was created and is automatically updated daily using the YouTube API, it is highly unlikely a mistake would come from the API request not capturing that feature.

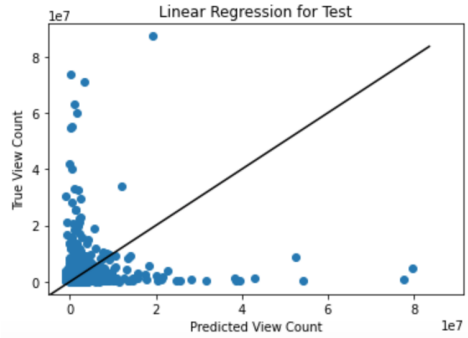
For histograms that are shown for this report, we highlight the ones created for the “View_count” and the “CategoryID” features. As can be seen for the trending video categories, most of them seem to relate to category number 24 and 10, which are Family and Comedy respectively. This somewhat makes sense, as the “Family” category reflects how YouTube features many content creators that encourage user engagement. As they want watchers to become friendly and think of the creator as a welcoming friend, the popularity of the “Family” category seems to reflect this. The “Comedy” category also seems logical with the popularity of many videos to provide entertainment and laughter to the audience. We can see from the right histogram that a majority of the views fall below the 50 million view mark with a heavy tail to the right as some views are able to reach 200 million views.




**2. Insights about Dataset**

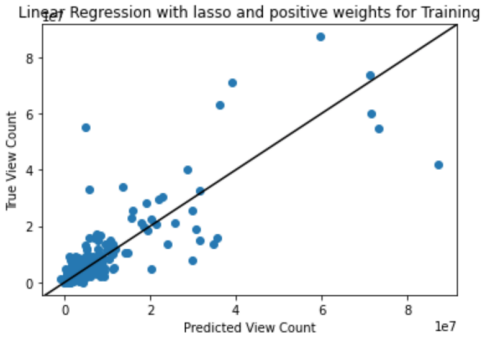
We run numerous preliminary regression models on the data splitting 80% training and 20% testing and achieve the following plots for linear regressions with lasso and having positive weights only. We also compute the mean squared errors for each regression below. The errors may seem very high, but since our labels are views in the millions, the errors we are seeing below are reasonable. For lasso, we tried many values for alpha (0.0001, 0.001, 0.01, 0.1, 0.5, 0.7, 1.0) which all had similar errors, so we chose alpha = 0.001 for now. We will perform cross-validation on the hyperparameter alpha in the future. With respect to linear regression, we also tried restricting the weights to be positive since all our features only contain positive values.

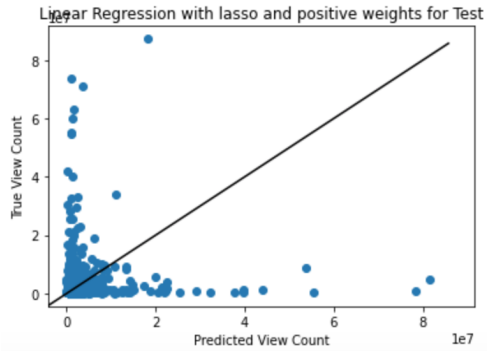



**Linear Models Mean Squared Errors (MSE)**

Linear Regression Training: 11806351946974.607
Linear Regression Test: 12012601851291.504
Linear Regression with positive weights Training: 13093955663167.113
Linear Regression with positive weights Test: 15561183696948.918
Linear Regression with lasso Training: 12878819766859.775
Linear Regression with lasso Test: 12868158484967.846
Linear Regression with lasso and positive weights Training: 12428152919998.734
Linear Regression with lasso and positive weights Test: 12868157743111.736




With both our training and testing errors, our model currently overfits. To avoid this risk, we plan to use cross validation (k-fold) and incorporate L2 regularization. We also want to avoid overfitting by changing our features to incorporate the text futures. We speculate a reason for the poor performance is because our text features for “Tags” and “Description” may provide valuable insight into a trending status. To avoid underfitting, we also plan to increase the complexity of our model by incorporating our text features. One feature engineering approach is to perform sentiment analysis over the “Description” feature to create an additional ordinal data feature. We could also perform term frequency-inverse document frequency (TF-IDF) as a way of highlighting important words for the content of each “Description” feature. Here, we decrease the weight for commonly used words that may be unimportant and increase weights for uncommon words. We may also lower the number of text features as well if our method for decreasing underfitting results in overfitting. Another way of using text is to use the “Title” feature, which we also have currently dropped, to count how many words are fully capitalized as this is a common trend seen on YouTube’s trending page.

One way we have already started to make our model more complex was to parse “trending_date” and “publishedAt” features from ISO 8601 time (ex: "2011-08-12T20:17:46.384Z") to a more readable and integer based set of features: “trending_year”, “trending_month”, “trending_day”, “trending_hour”, “trending_minute”, “trending_second”, “uploaded_year”, “uploaded_month”, “uploaded_day”, “uploaded_hour”, “uploaded_minute”, and “uploaded_second”. Here, we drop “trending_date” and “publishedAt” for our newly constructed features.

**3. Next Steps**

In the future, we will test out different metrics to make sense of which ones are the most reasonable in our setting. We will also try more complex models like ensemble tree methods. Another direction we could try is to incorporate feature engineering through sentiment analysis and TF-IDF techniques on the “Description” and “Tags” features. As tags are important descriptive keywords that can help viewers find the video, this feature should be studied in greater depth to understand if more videos use these features or not. In addition, we want to eventually present our final regression model with new top trending videos in November to understand with new data and time how the model performs at predicting views. This test dataset would include non-trending videos to see how the model does at predicting whether a video is trending or not. We hope our project can provide valuable insights in the upcoming month, and please reach out to us if you have any comments or concerns about our midterm progress.

Sincerely,
Jeremy Wang and Lorenzo Scotto di Vettimo