# Capstone Project 1: Final Report

## Youtube View Count Prediction

**Introduction:**

All businesses want to offer the right product, at the right place, time, and price. In order to increase their chances, they have to optimize their marketing strategy and increase their brand awareness and online presence. The top three things any S&P 500 company will focus on to reach their customers online:

1. Their website – Fundamental information customers need to see
2. Search engine optimization –find that website
3. Social media- Facebook, Twitter, Youtube, etc. – builds engagement and awareness

This project will focus on one particular channel of number 3 – advertising on Youtube videos to answer the question of how many potential customers can be reached by purchasing advertisements on a particular video. A ranking system was created by me in order to classify videos into categories based on their current level or ability to become viral. The definition of a “black belt” or viral video can be modified and is subjective to the user.

|  |  |  |
| --- | --- | --- |
| **Rank** | **# of Views** | **# of Days Since Creation** |
| White Belt | 20,000 | 300 |
| Blue Belt | 100,000 | 400 |
| Purple Belt | 500,000 | 450 |
| Brown Belt | 2,000,000 | 475 |
| Black Belt | 10,000,000 | 500 |

Since predicting view counts could potentially have multiple applications or use cases, we can look at the problem statement from both the perspective of a potential Youtube customer and also from Youtube themselves. When defining a problem statement, it’s useful to see several dimensions towards how the solution can solve more than one type of problem.

**Problem Statement from Marketer Perspective:**

1. Which youtube video related to my business will achieve greater than 20,000 view counts within 10 months based on title, category ID, description, tags, and other video features?
2. Which youtube videos are going to exceed 10 million views within 17 months (viral video) based on title, category ID, description, tags, and other video features?

The answer to these questions could assist a marketer in making a decision regarding which type of Youtube advertisement to purchase such as pay-per-click, or non-skippable, etc. As an example, on higher view count videos, perhaps a pay-per-click model is more optimal.

**Problem Statement from Youtube Perspective:**

1. Which youtube video should we consider charging a premium for advertisement?
2. How can we optimize the pricing system to charge more for viral videos vs. low rank videos?
3. Which videos have suspicious view counts / could be hacking their view counts?

Note: All of the requirements in the problem statements above can be modified to suit the user’s own interests.

**Constraints / Important Notes:**

1. Due to limitations within API and not using a random word generation, This was not a random sample of searched words. Due to API limits, I could only return 50 entries of data per call. For this project, the API was called 20 times manually to get 1000 rows of data. The search words that were utilized were: dude perfect, music, cats, movies, python, mac, Christmas, Donald trump, vacation, sports, iphone, horror, laugh, gaming, golf, Minecraft, mars, video blogging, speedrun, Siemens.
2. If we step back and think about this problem, probably the most important feature will be the User ID of the youtube video because a particular user will have historic evidence of view counts based on their video production capabilities. Unfortunately, this feature is not available through the Youtube API so we’re going to deal with this problem without this feature.
3. Since individual view counts per video are going to be very scattered, to reduce the scatter of the view counts during EDA, I chose to “bucketize” the viewcounts into buckets of 200,000 views.

**Decision Maker:** Marketing Manager for a company looking to advertise on Youtube or Youtube Sales / Pricing Manager

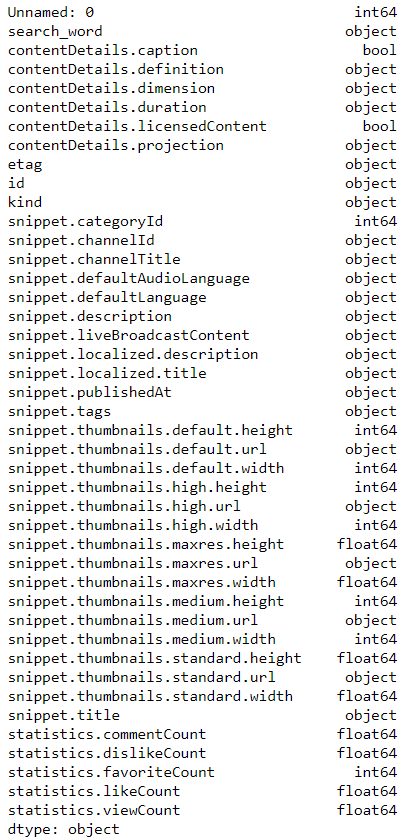
**Success Criteria:** Come up with prediction model with 60-70% accuracy. Since we’re not concerned about false positives, negatives, we will not consider model evaluation metrics such as precision, accuracy, F1 score, ROC, AUROC.

Data Wrangling

Due to the requirement of utilizing the Youtube API in order to download the data for this project, several data wrangling procedures had to be utilized in order to access and clean up the data. The steps utilized in cleaning the data are outlined below:

1. Utilized Google / Youtube API online instructions to authenticate an email with Oauth2. These credentials are necessary in order to access the API with a daily quota limit.
2. Through the API, supplied 20 different random words based on categories provided by the Youtube general categories. Each search produces 50 results, therefore the dataset had 1000 rows of data that I saved as a CSV.
3. Once the data was downloaded from the API, utilized json\_normalization to flatten the data and pass this into the pandas dataframe.

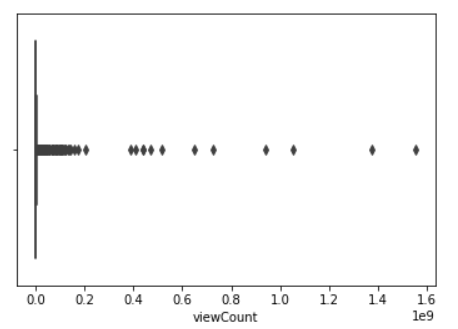
Initial shape of the data from CSV file: 1000 rows (1000 videos collected) and 43 columns (video factors)



Interpretation: Shows all of the original features and data types we are working with

1. Analyzed all of the columns, shape, head of the data to assess the current data and then removed all of the columns that were irrelevant that had no relationship to view Count and could not be used for the prediction.
2. Analysis of the view counts was then performed to assess the outliers. The rationale for this is the concern that there will be a few videos from the search that represent overly huge view counts that is not representative of an average video on youtube. From the analysis, 10 of 11 of the identified outliers were music videos. These videos would skew my model – I made the decision to remove the outliers since predicting view count is not life critical, I chose to discard these values. This action should theoretically improve the overall model.

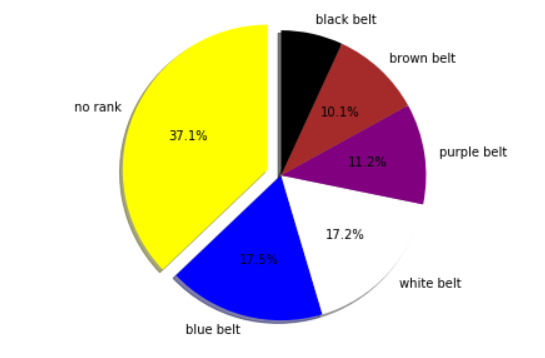
A box plot of view counts shows us 11 outliers. 10 out of these 11 are music videos. This is going to throw off our model so I chose to exclude this data.



Interpretation: The chart above shows a couple of outliers (counted 11) for view counts to be removed. These were mostly all music videos.

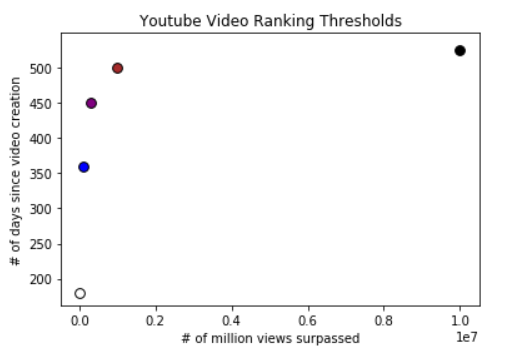
1. Missing values of the continuous data (e.g. likeCount, commentCount, dislikeCount) were filled in using the mean() function. I felt this would be the most appropriate method of addressing the continuous data since I did not want missing values to skew my model. Also, certain columns had missing lists or strings,, therefore I filled in an empty list or string for those columns using the fillna() function.
2. Initially use count based features of snippet tags.
3. Conversion of the video publish date column to age in days was completed.
4. Conversion of a string to calculate the total duration of a video was required with the use of regular expression and a list comprehension plus zip function to multiply / calculate the total time in seconds.
5. Utilize a pie chart to show the distribution of dependent variables.

The target data is well balanced as pictured in the pie chart.



Also, a scatter plot of the ranking thresholds was visualized:

Note: We could draw a line through the points and perhaps a mathematical function could be considered by a future user.

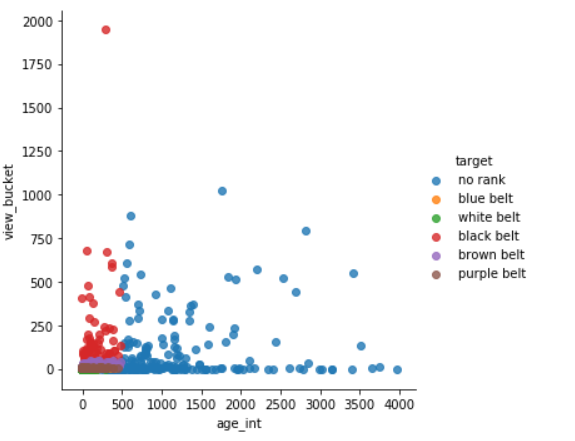


Other Interesting Visualizations:

Plot of Age vs. View Count with Rank as Hue: There are many videos that do not make rank

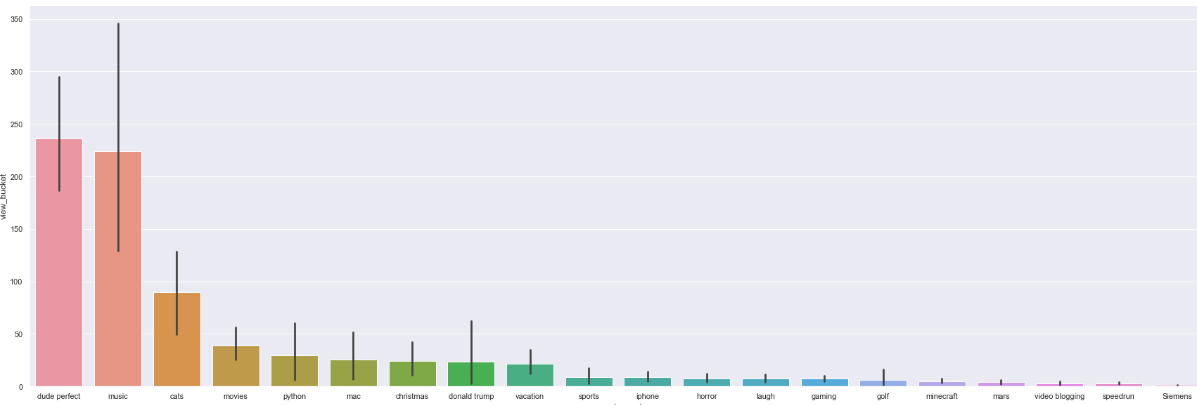
because they did not achieve the target views within the cutoff date but they do have a lot of

views over a long period of time. These videos could have a lower pricing model for ads.



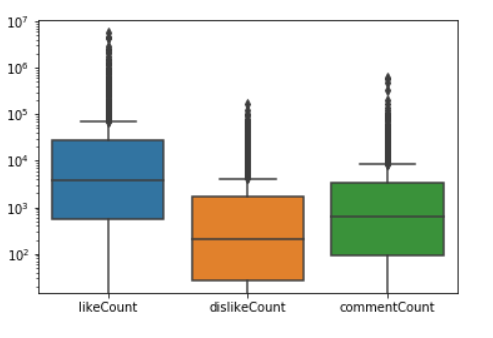
Interpretation: This chart above shows us that perhaps videos that are too old and do not have a lot of views are not ranked / not viral.

Bar plot of the view counts by search word:



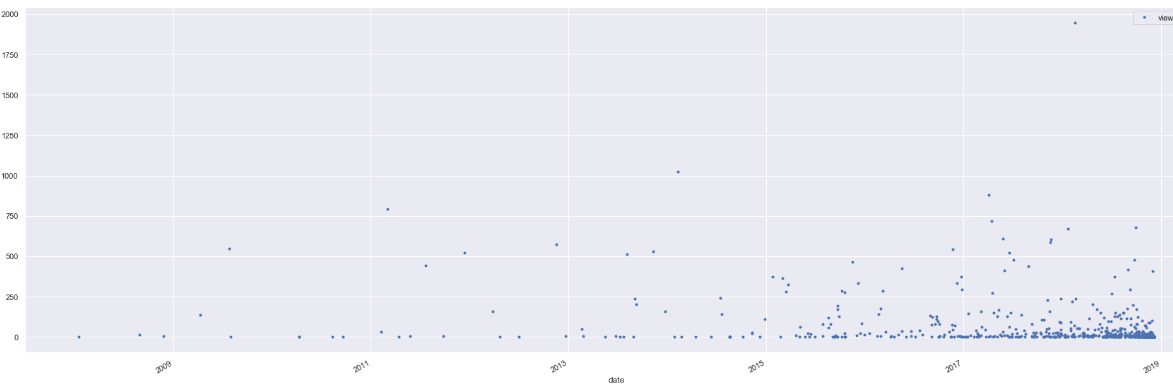
Interpretation: “Dude perfect”, “music”, “cats” were the top three in terms of total view counts

Box plot of like, dislike and comment counts for all videos.



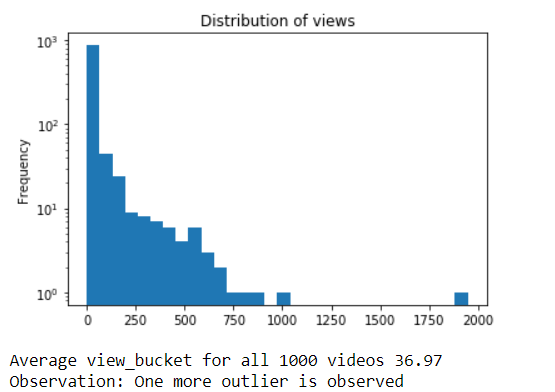
Interpretation: This makes sense since most people will click ‘like’ rather than ‘dislike’ a video.

Further data visualization shows interesting patterns in the data / view count: There appears to be more recent videos based on the search.



Interpretation: This makes sense since youtube is going to try and push the most recent / relevant videos into each search of 50 results.

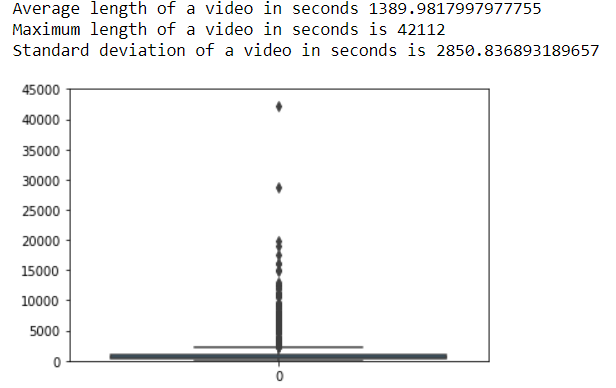
Utilizing the technique of binning, we are going to bucketize the view counts to help look at the distribution of view counts a little better.



Interpretation: We see a right skew and this tells us it’s challenging get a very high view count – most videos are low view counts / low rank.

No further action taken to remove any more outliers.

Average length of video was visualized just for informational purposes.



1. Once the data was cleaned, we can apply inferential statistics on our gathered data

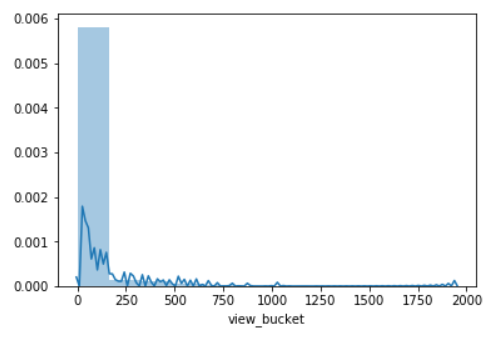
Apply Inferential Statistics

The purpose of my model is to predict the view count of a youtube video depending on attributes about the video:

1. Search word
2. Title
3. Age
4. Description
5. Tags
6. Text based features

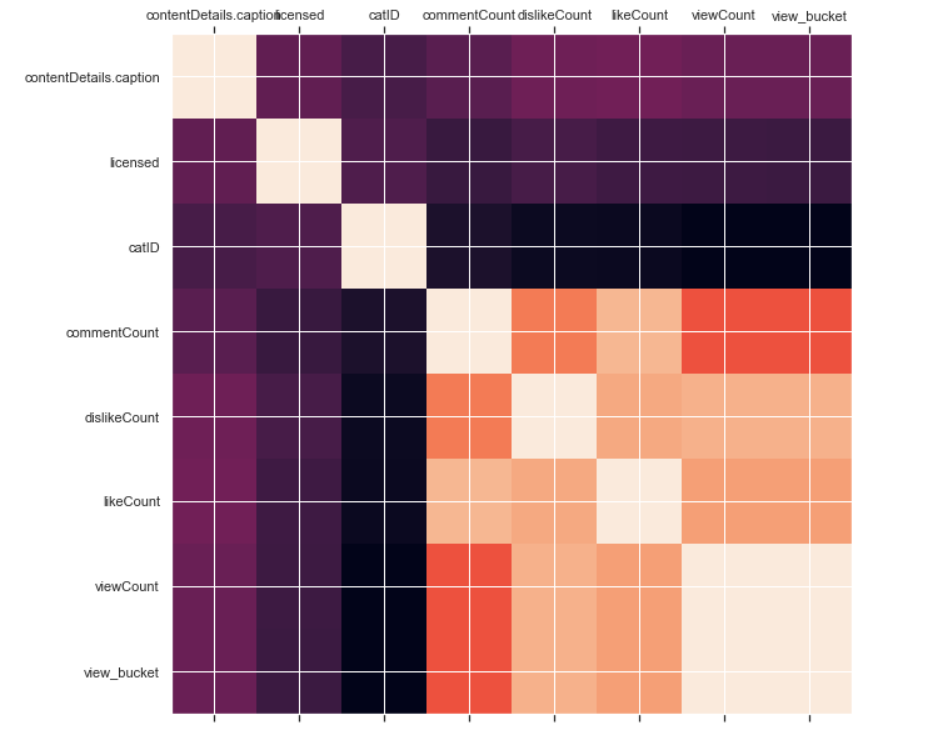
Statistical Techniques:

Distribution Analysis: One of the first things I tried to look for in my data is any normal distribution for the view counts of 1000 videos. It was immediately apparent that a right skew due to outliers occurred in my data due to a number of high view count videos.



Descriptive measures: The mean, and max of video length was looked at to see the average and max length of a video. The mode of the view bucket was looked at to see most of the videos fall under the first bucket. In other words, most of the videos have small view counts.

Correlation Matrix of numeric features:



Interpretation: Observed higher correction with comment, dislike, with view count. This is an obvious correlation because the more view counts a video has, the more likes, dislikes and comments there will be.

Statistical problem encountered: A random sample was not utilized because I was searching for specific search terms that I knew about.

Possible Solution: Utilize a random word generator to search for truly random terms for whatever the random generator produced. It would be great to automate this in the call to Youtube API along with increasing each search size beyond 50 entries. Right now, I had to manually search 20 words to get 1000 entries of data.

Dealing with the data I have: I started looking for any relationship between the independent and dependent variables such as video duration, licensed content, category ID, like, dislike, comment count.

1. For ordinal data, I utilized Spearmans correlation coefficient and for numeric data I utilized Pearson’s correlation coefficient.
2. The significance level that was chosen for all features was 5%.
3. The p values were calculated for each of the features in order to assess their statistical significance.
4. If null hypothesis was rejected, then the feature was included for further modeling efforts. If the null hypothesis failed to be rejected, then the feature was excluded from further experiments.

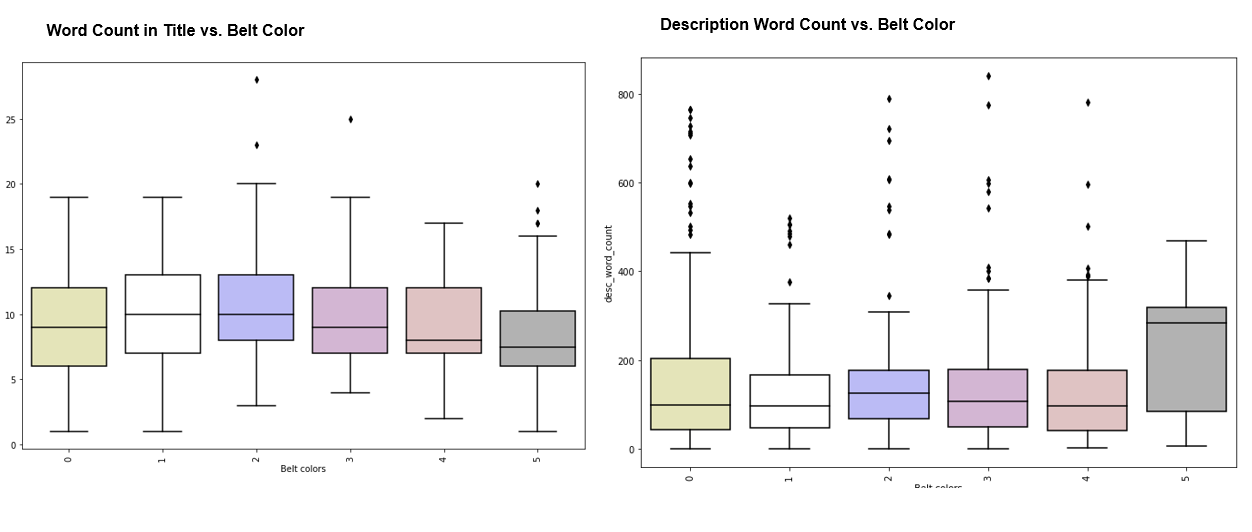
Outcome of the PCC and SCC, I will include category ID, like count, dislike count, comment count in the numerical modeling. For the NLP modeling, I will include only the category ID.

**Modeling Steps**

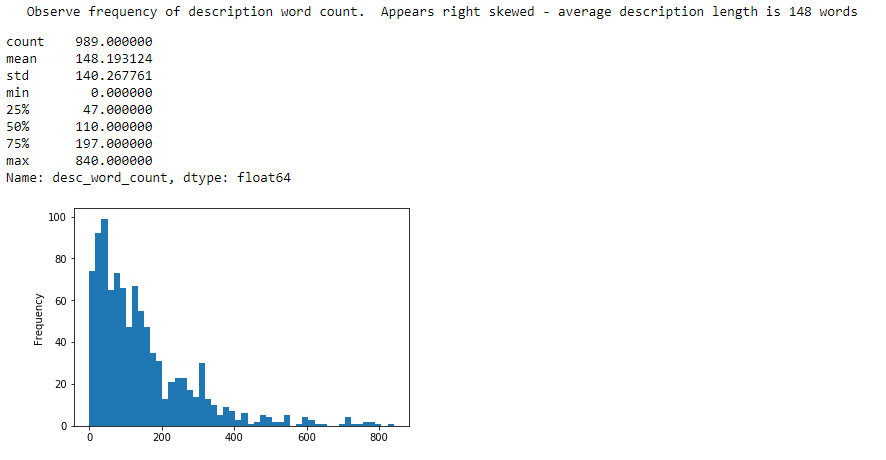
1. My approach to experimenting with the NLP data is to use a shotgun approach to trying out different models and preprocessing techniques.
   1. Utilize both CountVectorizer and TFIDVectorizer for vectorization
   2. Utilize a pipeline to run a slew of ML models (number of experiments would be 2 \* 6 + hyperparameter tuning experiments if applicable) :
      1. Logit
      2. MultinomialNB
      3. SVM
      4. GBM
      5. Random Forest
   3. Assess the accuracy of each model - look for the highest score from each of the experiments.
2. **Integrate the numeric + NLP data** to rerun the top three pipelines to assess the performance and pick the best model.
3. **Conclusion** / next steps beyond the project

## Text Data Visualization and Summary :

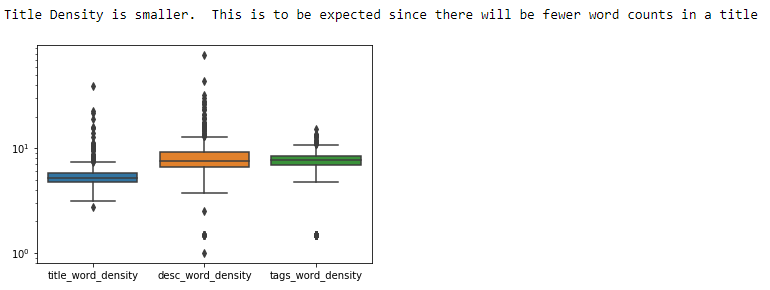
Examples of some of the visualizations with text data:



Interpretation: Very interesting to see specific patterns, upside down parabola on the title’s word counts. Black belt videos tended to have shorter titles which implies the theory that if a title is more direct, it will catch the attention of more people. The chart on the right above, shows the description word count and the black belt videos have significantly higher word count. Theory regarding the description is that black belt videos are more descriptive so viewers can understand better the video content.



Interpretation: The chart above shows the frequency of description word count. We see a right skew and this is normal – The average description length is 148 words



The table below represents an summary / analysis of all of the count, density based features of the text data.

Note: If I was more concerned about the high dimensionality of the data, I would have conducted hypothesis testing on the features to keep the significant features only. In the interest of time, I just kept almost all of the features.

|  |  |  |
| --- | --- | --- |
| **Count Based Features of Text Data** | **Brief Description** | **Notes:** |
| title\_char\_count | # of characters in title | black belts have fewer chars |
| title\_word\_count | # of words in title | black belt videos have fewer words |
| title\_word\_density | character count / word count + 1 | Disregard |
| title\_punctuation\_count | # of punctuations in title | black belt videos have punctuations |
| title\_title\_word\_count | # of words that have first letter capitalized | Disregard |
| title\_upper\_case\_word\_count | # of words that are completely capitilized | Upside down parabola |
| title\_stopwords\_count | # of stop words in title | black belt video titles have lowest stop word avg |
| desc\_char\_count | # of characters in description | black belt videos have highest description char count avg |
| desc\_word\_count | # of words in description | black belt videos have highest description word count avg and no outliers |
| desc\_word\_density | character count / word count + 1 | black belt videos have smallest range |
| desc\_punctuation\_count | # of punctuations in title | black belt videos have highest punctuation count |
| desc\_title\_word\_count | # of words that have first letter capitalized | black belt videos have highest average words for first letter capitalized in desc |
| desc\_upper\_case\_word\_count | # of words that are completely capitilized | black belt videos have fewer completely upper case description words |
| desc\_stopwords\_count | # of stop words in description | black belt videos have MORE stop words on average inside description |
| tags\_char\_count | # of characters in tags | black belt videos have higher average for character tags count |
| tags\_word\_count | # of words in tags | black belt videos have higher average for # of words |
| tags\_word\_density | character count / word count + 1 | black belt videos have slightly higher word density average with a smaller range |
| tags\_punctuation\_count | # of punctuations in tags | black belt videos have slightly higher average for punctuation count in tags |
| tags\_title\_word\_count | # of words that have first letter capitalized | Disregard |
| tags\_upper\_case\_word\_count | # of words that are completely capitilized | black belt videos have fewer completely upper case tags |
| tags\_stopwords\_count | # of stop words in tags | Upside down parabola |

## Modeling Results: Numeric Data

Modeling results of the numeric data only. Best scores highlighted in yellow.

|  |  |  |
| --- | --- | --- |
| **Model** | **Notes** | **Accuracy** |
| Linear Regression |  | 0.07 |
| Logit - C hyperparameter tuning | Convergence problem when tuning C: Analyzed data, C, and increased max\_iter. Decided to not pursue this model further | 0.3737 |
| SVM |  | 0.4242 |
| Logit |  | 0.44 |
| KNN | Tested nn range 1-20, best =3 | 0.5252 |
| Decision Tree Classifier |  | 0.5959 |
| Gradient Boosting - default |  | 0.6464 |
| Gradient Boosting - random search | {'n\_estimators': [100, 311, 522, 733, 944, 1155, 1366, 1577, 1788, 2000], 'max\_features': ['auto', 'sqrt'], 'max\_depth': [2, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, None], 'min\_samples\_split': [2, 5, 10], 'min\_samples\_leaf': [1, 2, 4]} | 0.6565 |
| Gradient Boosting - grid search | param\_grid = {  'max\_depth': [30, 40, 50, 60, 70],  'max\_features': [2, 3],  'min\_samples\_leaf': [1, 2, 3],  'min\_samples\_split': [2, 3, 4, 5],  'n\_estimators': [100, 200, 300, 1000] } | 0.6666 |
| Random Forest - default |  | 0.6869 |
| Random Forest - random search | {'bootstrap': True, 'class\_weight': None, 'criterion': 'gini', 'max\_depth': None, 'max\_features': 'auto', 'max\_leaf\_nodes': None, 'min\_impurity\_decrease': 0.0, 'min\_impurity\_split': None, 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'min\_weight\_fraction\_leaf': 0.0, 'n\_estimators': 10, 'n\_jobs': None, 'oob\_score': False, 'random\_state': 42, 'verbose': 0, 'warm\_start': False} {'n\_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], 'max\_features': ['auto', 'sqrt'], 'max\_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None], 'min\_samples\_split': [2, 5, 10], 'min\_samples\_leaf': [1, 2, 4], 'bootstrap': [True, False]} | 0.6969 |
| Random Forest - grid search | param\_grid = {  'bootstrap': [True],  'max\_depth': [30, 40, 50, 60, 70],  'max\_features': [2, 3],  'min\_samples\_leaf': [1, 2, 3],  'min\_samples\_split': [2, 3, 4, 5],  'n\_estimators': [100, 200, 300, 1000] } | 0.6969 |

## Modeling Results: Text Data

Modeling results of the text data only. Best scores highlighted in yellow.

|  |  |  |  |
| --- | --- | --- | --- |
| **NLP Features** | **Model** | **Notes** | **Accuracy** |
| Train W2V Scratch, MeanEmbeddingVectorizer, unstemmed | Log Reg |  | 0.303 |
| BoW TFIDF | Random Forest( n\_estimators=100, max\_depth=5 |  | 0.3131 |
| Train W2V Scratch, TFIDFEmbeddingVectorizer, unstemmed | Log Reg |  | 0.3131 |
| 1-2 gram, TFIDF, stemmed, stop words | XGboost |  | 0.3232 |
| Count and Density Based | Log Reg (C=1) |  | 0.3333 |
| BoW, TFIDF | MNB |  | 0.3333 |
| BoW, TFIDF | Stochastic Gradient Descent (loss='hinge', penalty='l2', alpha=1e-3, n\_iter=5) | tested w/ n\_iter, alpha, no changes | 0.3333 |
| BoW, TFIDF | KNN (nn range 1-20) | no change based on nn | 0.3434 |
| 1-2, gram TFIDF, stop words | MNB |  | 0.3535 |
| 1-2 gram, TFIDF, stemmed, stop words | Log Reg |  | 0.3737 |
| 1-2 gram, TFIDF, stop words | XGBoost |  | 0.3737 |
| BoW, TFIDF | XGBoost (objective ='multi:softmax', colsample\_bytree = 0.3, learning\_rate = 0.1, max\_depth = 5, alpha = 10, n\_estimators = 10) |  | 0.3838 |
| 1-2 gram, TFIDF, stemmed, stop words | XGBoost |  | 0.3838 |
| BoW TFIDF | Random Forest - random search | {'n\_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000], 'max\_features': ['auto', 'sqrt'], 'max\_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None], 'min\_samples\_split': [2, 5, 10], 'min\_samples\_leaf': [1, 2, 4], 'bootstrap': [True, False]} | 0.3992 |
| BoW, TFIDF | Log Reg |  | 0.404 |
| BoW, CountVectorizer | Log Reg | Same results as TFIDF because TFIDF is weighted with Log(N/n) | 0.404 |
| 1-2 gram TFIDF, stop words | Log Reg |  | 0.404 |
| 1-2 gram, TFIDF, stemmed, stop words | Log Reg |  | 0.404 |
| 1-2 gram, TFIDF, stemmed, stop words | Random Forest |  | 0.4242 |

## Modeling Results: Numeric + Text Data

Modeling results of combined numeric and text data

|  |  |  |  |
| --- | --- | --- | --- |
| **NLP Features** | **Model** | **Notes / Parameters** | **Accuracy** |
| Count features, BoW, CountVectorizer | OneVsrest(Log Reg) |  | 0.4678 |
| Count features, BoW, CountVectorizer | Random Forest |  | 0.4355 |
| Count features, BoW TFIDF | Stochastic Gradient Descent (loss='hinge', penalty='l2', alpha=1e-3, n\_iter=5, ) |  | 0.2621 |
| Count features, BoW TFIDFVectorizer | Random Forest - Default |  | 0.4516 |
| Count features, BoW TfidfVectorizer(1-2 gram, stop words) | KNN (nn =11) |  | 0.5323 |
| Count features, BoW TfidfVectorizer(1-2 gram, stop words) | XGBoost(objective ='multi:softmax', colsample\_bytree = 0.3, learning\_rate = 0.1, max\_depth = 5, alpha = 10, n\_estimators = 10) |  | 0.7581 |
| Count features, BoW TfidfVectorizer | XGBoost(learning\_rate =0.1, n\_estimators=1000, max\_depth=5, min\_child\_weight=1, gamma=0, subsample=0.8,  colsample\_bytree=0.8, objective= 'binary:logistic', nthread=4, scale\_pos\_weight=1,seed=27) | increasing number of estimators to 1000 helped | 0.8145 |
| Count features, BoW TfidfVectorizer(1-2 gram, stop words) | XGBoost Grid Search Tuned params | Grid Search ((xgb\_\_max\_depth': 3, 'xgb\_\_min\_child\_weight': 1)) | 0.8064 |
| Count features, BoW TfidfVectorizer(1-2 gram, stop words) | Gradient Boosting Classifier | Random Search (n\_estimators = [100, 500, 900, 1100, 1500] max\_depth = [2, 3, 5, 10, 15] min\_samples\_leaf = [1, 2, 4, 6, 8]  min\_samples\_split = [2, 4, 6, 10] max\_features = ['auto', 'sqrt', 'log2', None]) | 0.8387 |
| Count features, BoW TfidfVectorizer(1-2 gram, stop words) | Gradient Boosting (default) |  | 0.8548 |

## Conclusion

* As our initial EDA indicated, numeric data proved to be more accurate in predicting virality. Text data accuracy was low. If we can collect even more text data, we can enhance our word2vec modeling.

## Random Forest seemed to work better on numeric data initially

## Log Reg and Random Forest worked best on test data

## Best accuracy had 85% accuracy

## Gradient Boosting came out as the final winner. It was interesting to note that Gradient Boosting beat XGBoost on the final model.

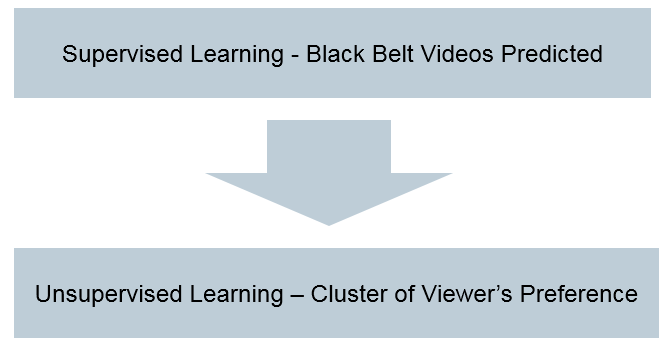
## Next Steps

## Obtain additional random data from Youtube API to improve performance and W2V models

* Explore additional NLP preprocessing / features – removing special characters, removing additional stop words (i.e. “html”), POS tagging, chunking, constituency, and dependency parsing, sentiment analysis

## Consider external sources that could cause virality scraped on the internet

## Model stacking - Feed the outcome of this model to an unsupervised learning problem which clusters of black belt videos are sent to user feeds.



Use Cases for this Project

* A Youtube marketer could utilize the text + numeric models to evaluate the virality ranking of a video. If they want to reach the most customers within the shortest period time, they can choose only black belt videos to advertise on.
* Youtube could utilize the text only models to assess if a video shows any potential signs of view count hacking instead of manually checking videos. Youtube could also utilize text + numeric models to do strategic pricing of their advertising program.