YouTube Viewcount Prediction

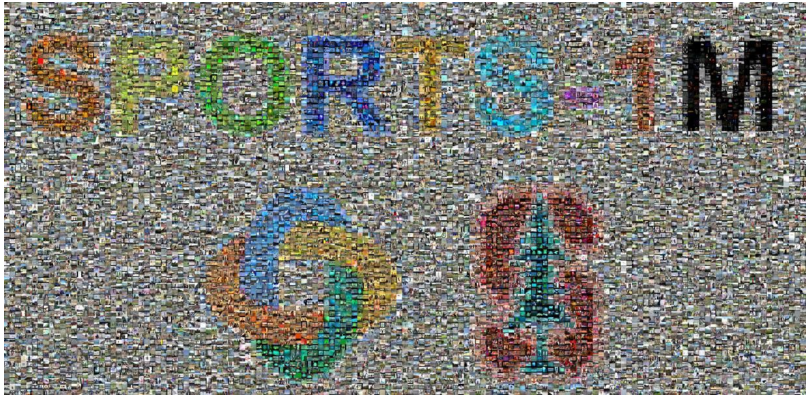
Brett Bass and Evan Ezell

Problem Description

- Since inception in 2005, YouTube has rapidly increased in popularity
- Becoming very lucrative to content creators as it has become a great marketing avenue for companies
- Predicting the number of views a video will receive could have great impacts on where these companies will allocate their resources

Data

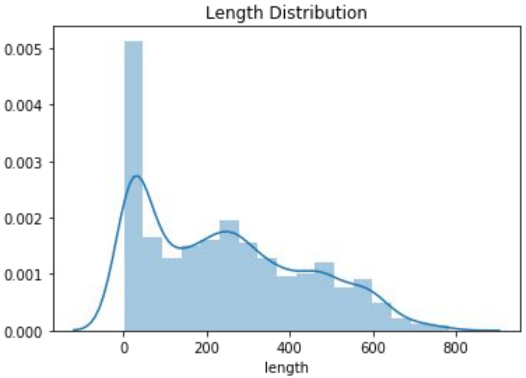
- 3,000 total sports videos selected from Stanford Sports 1M Dataset
  - 1,000 Table Tennis
  - 1,000 Bowling
  - 1,000 American Football
- Used 2,730 videos as many of the videos had been removed from YouTube

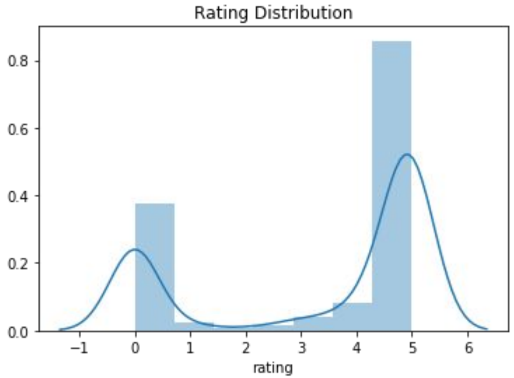


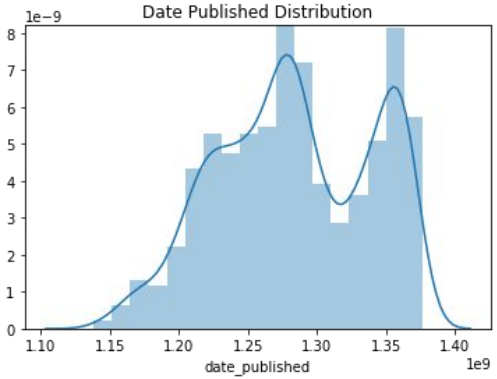
Scraping

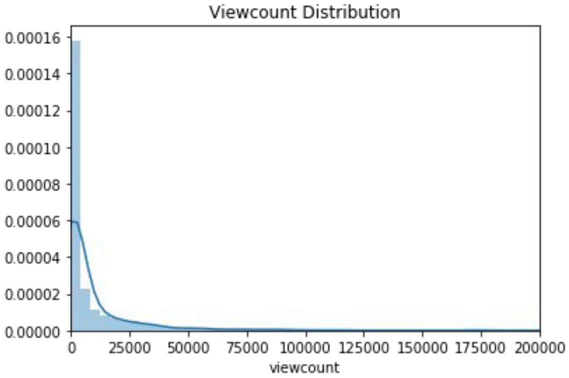
- PAFY Python library was used to scrape the majority of YouTube metadata
  - # Likes/Dislikes
  - Rating
  - Publish Date
  - Title
  - Description
  - Length
- The YouTube API was also used to collect some additional data that could not be retrieved using PAFY
  - Subscriber Count
  - View Count

Data Distributions





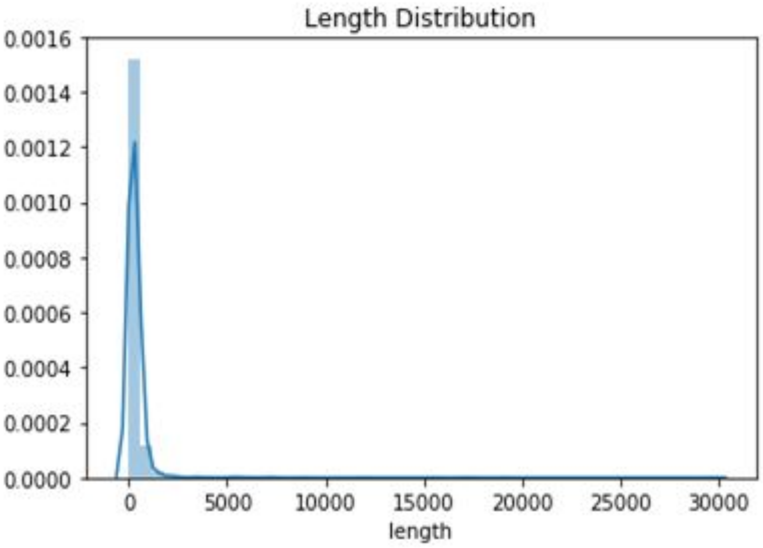


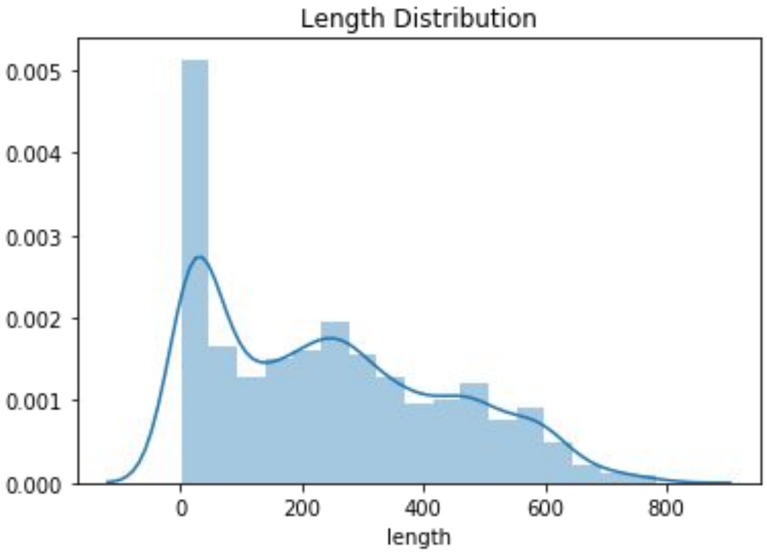


Missing Data

- Only 68 total missing values
  - 34 missing “likes” values
  - 34 missing “dislikes” values
- Values were imputed with the average ratio of views to likes and dislikes

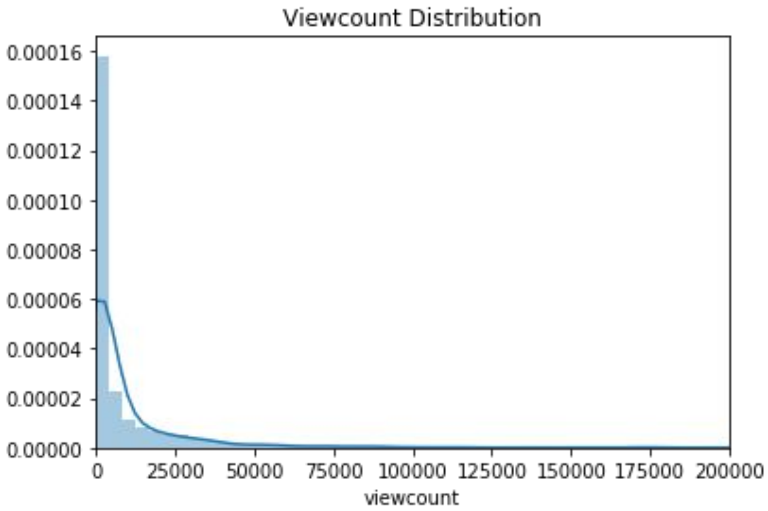
Removing Outliers





Transformations

- Due to the skew of many of the attributes, a log transform was used



Sentiment Analysis

- Sentiment analysis was used for the Title and Description of each video
- VADER (Valence Aware Dictionary and sEntiment Reasoner) package was used
- Social media context for text analysis
- Sensitive to both polarity and intensity
- Used scores would be used as inputs to the final model

https://github.com/cjhutto/vaderSentiment

Random Forest

- Very robust
- Default hyper-parameters often lead to very good results
- Less prone to overfitting
- Interpretation is made easier with relative feature importance

Training/Tuning

- 75% train and 25% test set
- 5-fold cross validation on the training set
- Random grid search to gain insight of hyper-parameter space
- Fine grid search based on values obtained from random grid output
- Optimal values and more details are covered in our report

Results

- We assessed the model quality on the 25% test set
- We iteratively made changes throughout the process
- Log transformation, adding subscriber count, and removing outliers

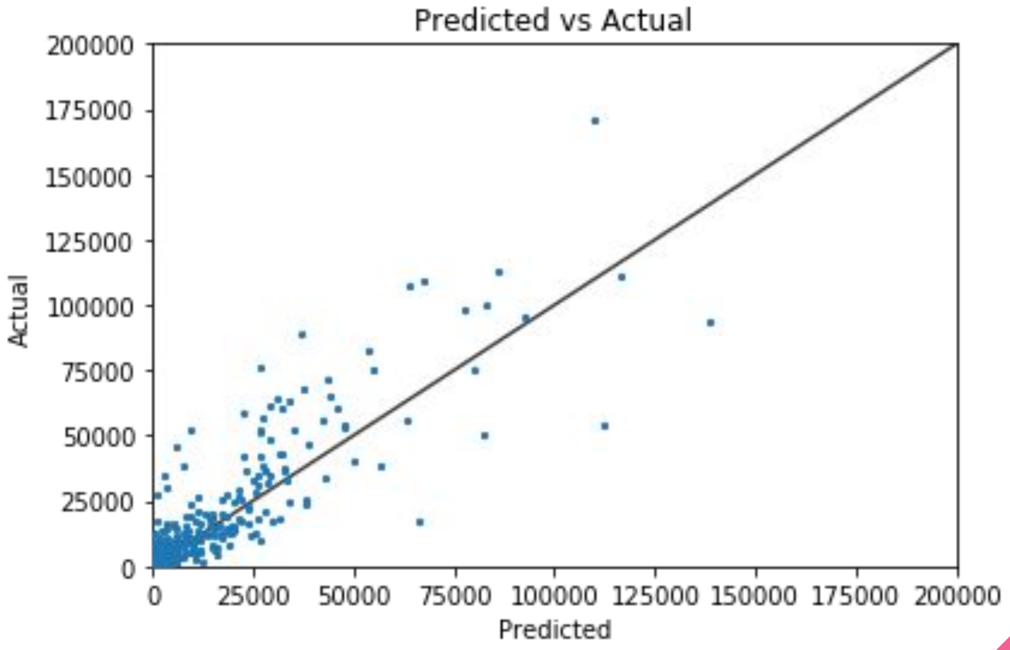
Mean Absolute Error		
Initial	Removed Outliers and Added Subscriber Count	Log Transform (Final)
9827.55	5203.4	3795.33

Results

- We binned the results to see how performance looked for different number of views

Regression Metrics			
R^2	MAE (Full Dataset)	MAE (0-1000 Views)	MAE (0-5000 Views)
0.8	3795.33	735.77	313.45

Results


Predicted vs Actual
Actual
Predicted
0
25000
50000
75000
100000
125000
150000
175000
200000

Limitations

- Time
- Google Translate
- Available storage for videos

Future Recommendations

- Pull more videos and use several different categories
- Use CNN to see impact of actual video
- Use more explanatory variables from Youtube API
- Use google translate for sentiment analysis
- Explore more models
- Look at how number of views grows over time