YouTube Views Predictor

Maxence Frenette, Casey Hanh, Michael Lu

Background

- YouTube is the world's largest video hosting service, 2nd largest social media platform, and 2nd largest search engine
- Every day 2.5 billion users consume 1 billion hours of content
- Total revenue of $29B+ in 2022
- **Problem**: What features make a video more likely to get views?
  - Advertisers can improve ad placement to maximize exposure
  - Content creators earn more with more views
- **Goal**: Build a machine learning model to predict the number of views a video will receive based on its features & attributes

Data

- Retrieved data using Google's YouTube Data API
  - Pulled data on March 4th and April 3rd (30 days apart)
  - Top 50 videos in the US, plus 25 related videos for each
  - Data about the channels that published each video
  - De-duplicated records by unique video ID
- Data Fields
  - Text: title, description
  - Image: thumbnails
  - Date/Time: video published date, channel published date
  - Numeric:
    - Video: duration, # of views, # of likes, # of comments
    - Channel: # of subscribers, # of views, # of videos
- Final Data Frame - **2,324** records with **24** columns
  - 70/10/20 - Train/Validation/Test Split
- Updated data visualizations can be found in the Appendix

Steps taken since mid-term presentation

- Pulled more data points from the YouTube API
- Deduplicated our data
- Refactor code to allow easier experimenting
- Train models that use the thumbnail images
  - Neural networks
  - Transfer learning using pre-trained vision model
- Used NLP models on the video title and description
  - Bag of Words
  - Embeddings
  - Transformers
- Ensemble models

Models

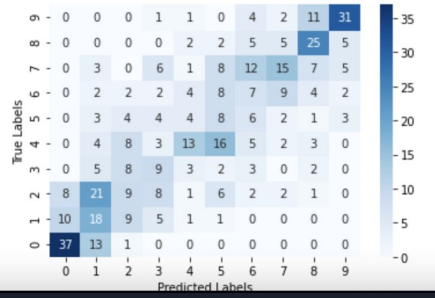
Data: Video Attributes Only

Model	Accuracy	Precision	Recall	F-1 Score
Baseline Model: Random Number	0.10	0.10	0.10	0.10
Logistic Regression	0.11	0.13	0.11	0.07
Gradient Boosted Decision Trees	**0.37**	**0.37**	**0.37**	**0.36**
Random Forest (TPOT Model)	0.37	0.35	0.36	0.35

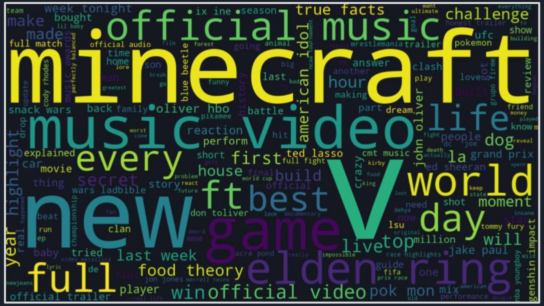
Classification Report:

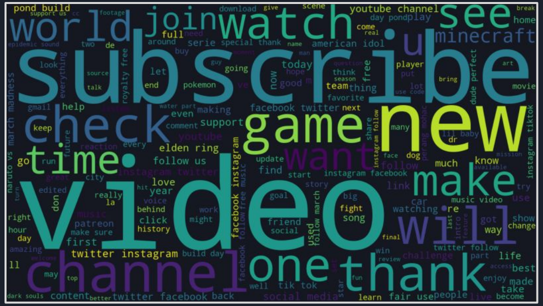
	precision	recall	f1-score	support
0	0.67	0.73	0.70	51
10	0.26	0.41	0.32	44
20	0.22	0.16	0.18	58
30	0.24	0.28	0.26	32
40	0.43	0.24	0.31	54
50	0.16	0.23	0.19	35
60	0.16	0.17	0.17	40
70	0.41	0.26	0.32	57
80	0.46	0.57	0.51	44
90	0.67	0.62	0.65	50
accuracy			0.37	465
macro avg	0.37	0.37	0.36	465
weighted avg	0.38	0.37	0.37	465

Confusion Matrix:



NLP Models


minecraft
official music
music video
new
v
game
elden ring
full
world
best
ft
day
every
life
first
house
challenge
true facts
official video
food theory
last week
top
live
build
secret
dog
week tonight


subscribe
video
channel
new
game
check
time
world
join
watch
see
make
will
thank
one
want
minecraft
twitter instagram
youtube channel
facebook instagram
elden ring
follow us

NLP Models

- Data Normalization & Cleanup
  - Remove hyperlinks, non-alpha characters, single character “words”
  - Make all text lowercase
  - Filtered out “stop words” from NLTK’s English set

	Before	After
Title	Food Theory: Did MrBeast’s Video Just BREAK the Law?	food theory mrbeast video break law
Description	Be one of the first to subscribe to Style Theory! ► https://bit.ly/styletheorysub If @MrBeast offered you a FREE trip to Paris, would you take it, Loyal Theorists? The only contingency is that you must bring him back one baguette…	one first subscribe style theory mrbeast offered free trip paris would take loyal theorists contingency must bring back one baguette

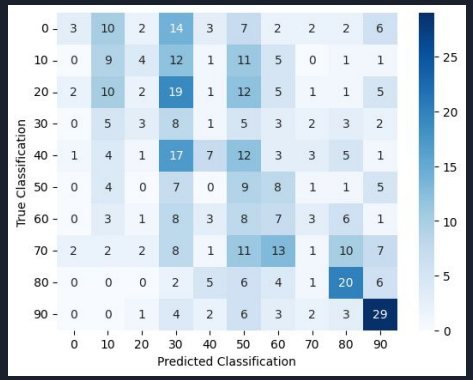
- **Titles**: 12,656 unique words, Avg. of 6 words in each title
- **Descriptions**: 26,640 unique words; Avg. of 80 words in each description

NLP Models: Video Description

Bag of Words

- One-hot encoding
- Applied tf-idf weighting

- **Test Accuracy: 0.20**

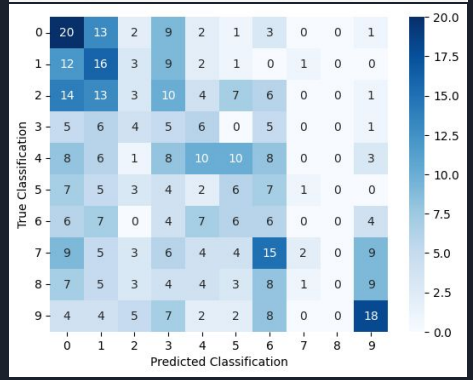


Embeddings

- Tokenized & padded
- Layers: Embeddings, Convolution, & Avg Pooling

- **Test Accuracy: 0.18**

Vocab Size	Sequence Len	Embedding Dims	# of Epochs	Test Loss	Test Accuracy	Val Loss	Val Accuracy
1000	50	8	5	2.2760	0.1494	2.2834	0.1344
1000	100	16	10	2.1693	0.2516	2.2183	0.1989
2000	100	32	10	1.9823	0.3640	2.1688	0.1667
2500	150	32	10	1.9964	0.3652	2.1745	0.2419
5000	300	64	20	1.1989	0.6748	2.3266	0.3118



Transformers

- "bert-base-uncased" model with vocab size of 30,522
- Layers: BERT, dropout, linear, & Softmax
- **Test Accuracy: 0.14**

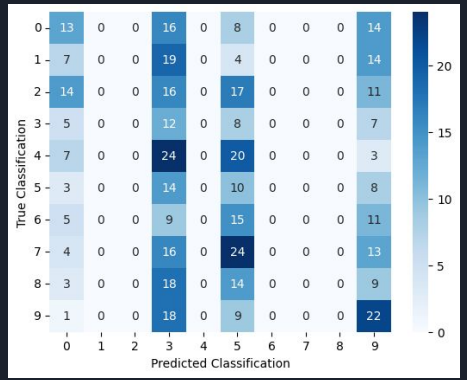


Image Models

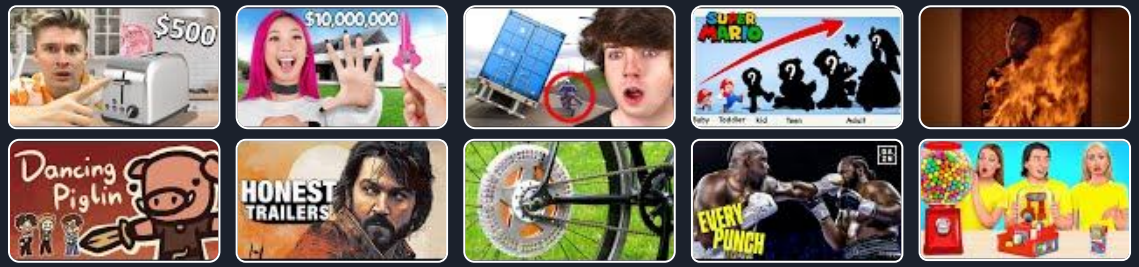


Image Models: Object Detection

- Pre-trained object detection models: vgg16, Xception, mrcnn, YOLO
- Classified objects found in the images, and then used identified object labels to predict video views
- Object detection label with highest probability used as feature to predict views.
- 423 unique object detection labels

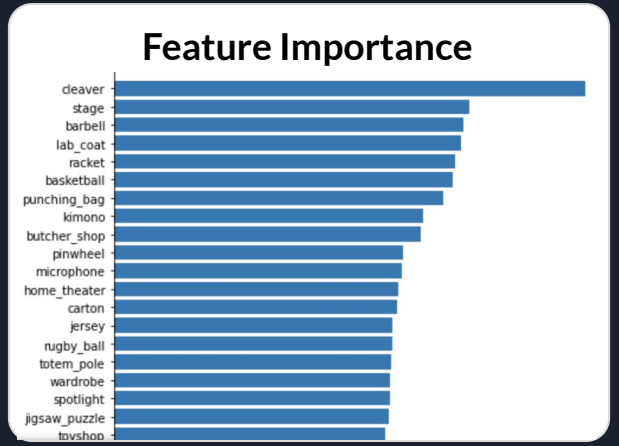


Image Models: NSFW Detection

- The NSFW Detection model is designed to identify potentially inappropriate content, such as nudity, violence, or drug use, in images and videos.
- By integrating NSFW detection into our YouTube video analysis, we can get a more accurate picture of how a video is likely to perform and make more responsible decisions about what content to promote and market.

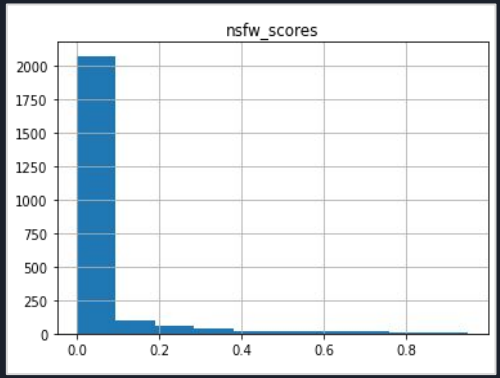
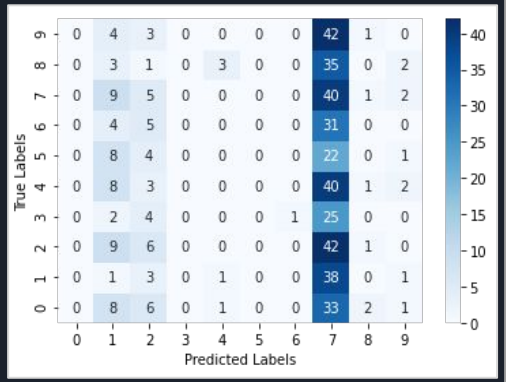


Image Models: Summary

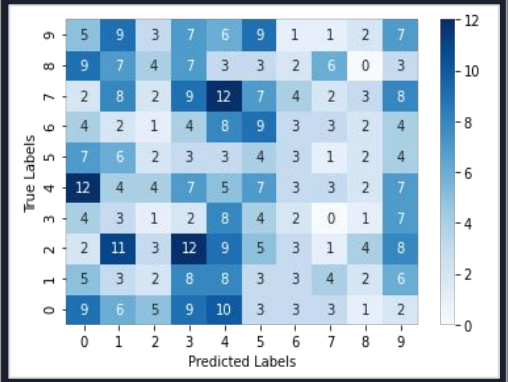
Deep Learning Models

- Convolutional Neural Network
  - Layers: 172k Pixels, Convolution + Avg Pooling
  - Training augmentation - adjust contrast/flip images
- **Test Accuracy: 0.10**



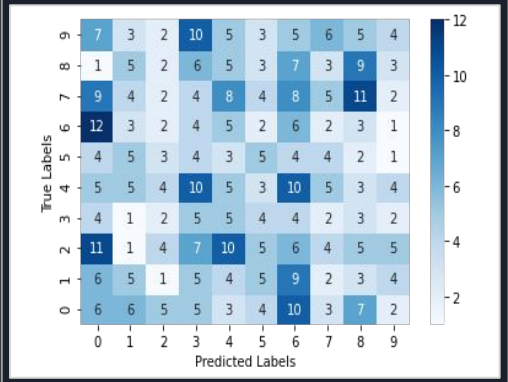
Feature Extraction: Object Detection

- Feature Extraction: VGG16, Xception, YOLO
- Logistic regression, Decision Trees, Neural Network
- **Test Accuracy: 0.08**



Feature Extraction: NSFW Score

- Feature Extraction: Open-NSFW2 model (not suitable for work)
- Logistic regression, Decision Trees, Neural Network
- **Test Accuracy: 0.12**



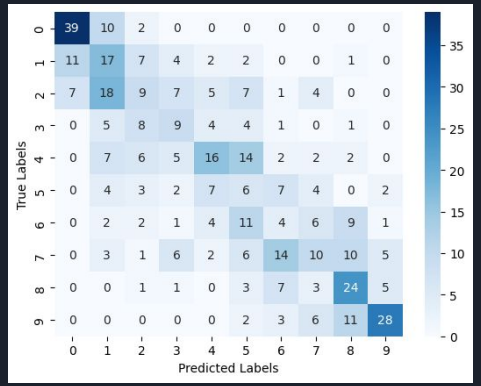
Ensemble Model

Combined Model with all Image Features

Gradient Boosted Decision Tree

- Base Data
- Image Features

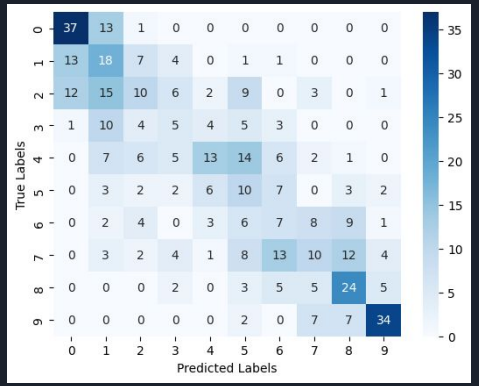
- **Test Accuracy: 0.37**



Random Forest

- Base Data
- Image Features

- **Test Accuracy: 0.35**

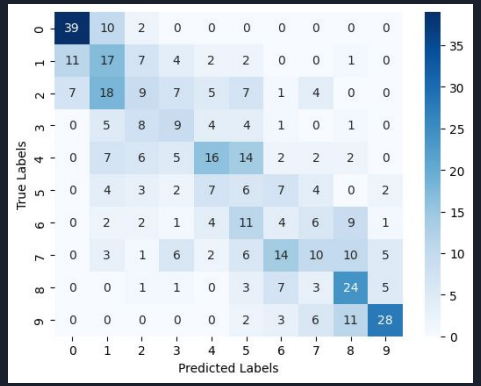


Full Ensemble Models

Base Data Ensemble Model

- Gradient Boosted Tree
- Random Forest

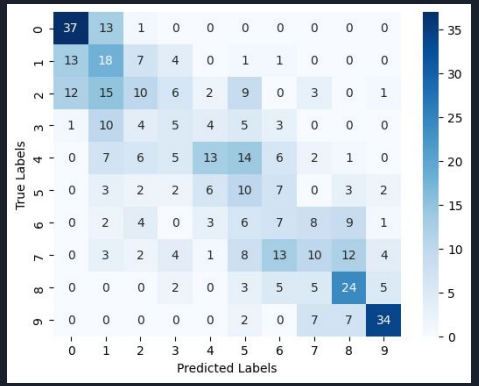
- **Test Accuracy: 0.38**



Full Ensemble Model

- Gradient Boosted Tree
- Random Forest
- Title Bag of Words
- Description Bag of Words

- **Test Accuracy: 0.39**



Ethics

Issue	Mitigation Strategy
Data collection subject to limitations of free API	Use other methods to collect broader set of data
NLP models filtered non-English characters; for example: **【サーキットレビュー】アストンマーティン ヴァルキリー**	Enhance NLP models to detect non-English languages and use appropriate vocabulary to encode; Fully disclose which languages are not supported
Pre-trained object recognition models which may have inherent bias	Identify alternative pre-trained models; Fully disclose the pre-trained models used for complete transparency
Model may unintentionally encourage harmful or inappropriate content	Remove data about harmful or inappropriate videos from the training dataset; Fully disclose to model users what types of content are excluded
YouTube viewership data is the result/product of any biases in YouTube's recommendation algorithms	Our model aims to demystify what makes a video attract high viewership and better equip users to "use" YouTube's algorithm to their advantage

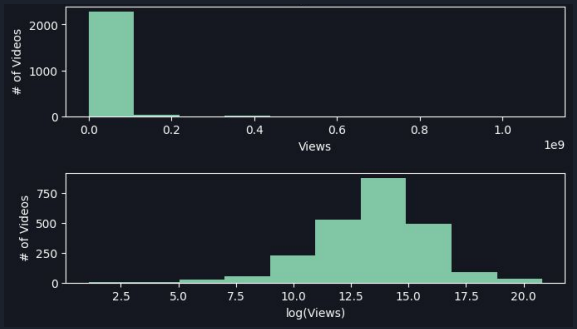
Next Steps

- Pull more videos
- Perform hyperparameter optimization on the sklearn models
  - Manually
  - Using TPOT
- Group the object detection labels from the thumbnails into smaller number of categories to reduce noise
- Train a model to process the actual video
- Train a bigger model (likely a neural net) to combine all the inputs in one model
- Train a smaller CNN (less parameters) on the thumbnails

Appendix

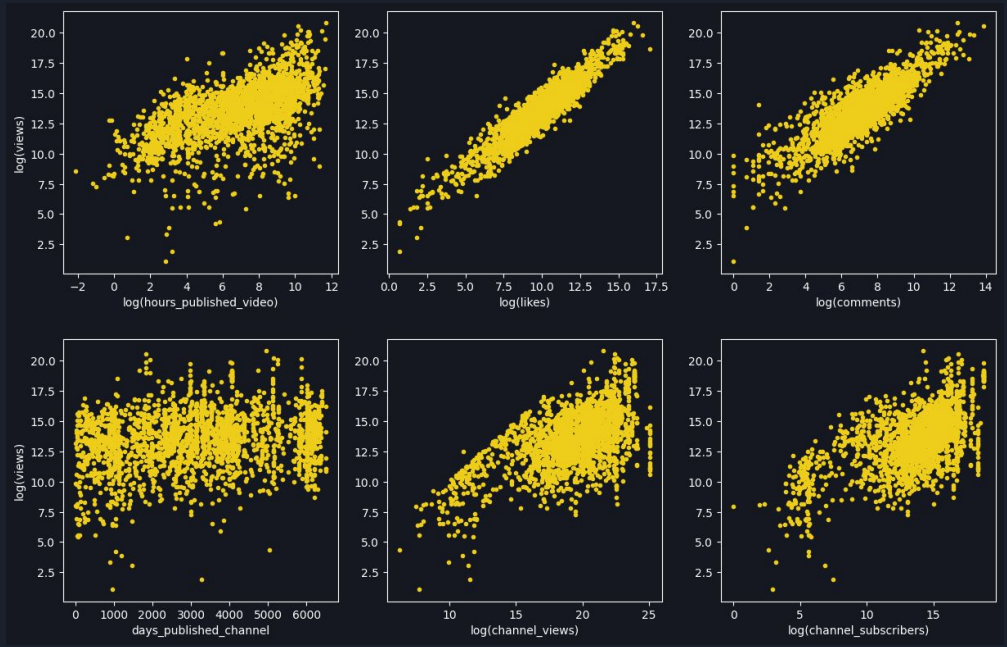
Data Visualizations

Figure 1: # of Videos by Total Views

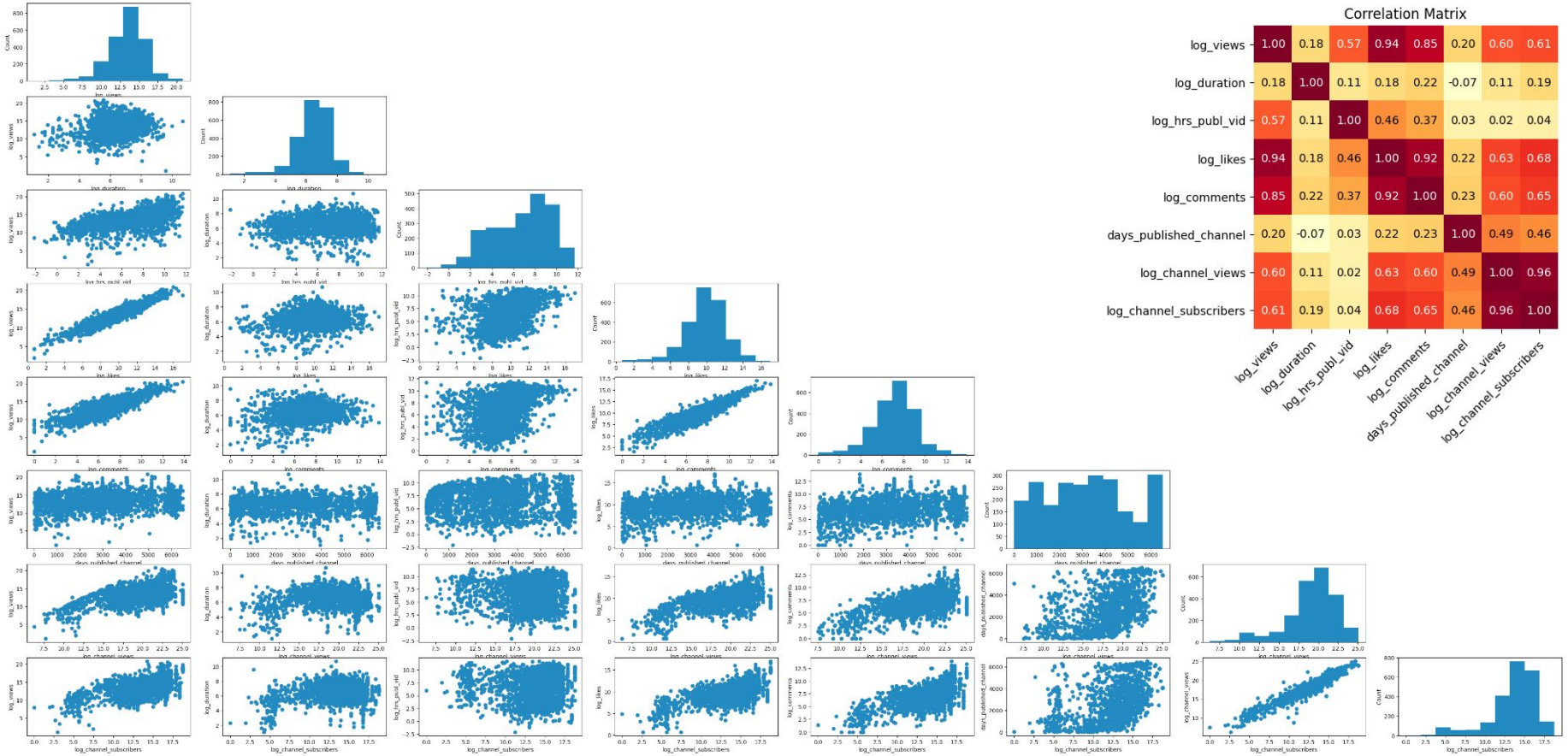


- Log transformations applied to several of the numeric features to unskew the data distribution
- Additional scatterplots, histograms, & correlation matrix in the Appendix

Figure 2: Scatter Plots of Views vs. Numeric Features



Scatterplot & Correlation Matrix

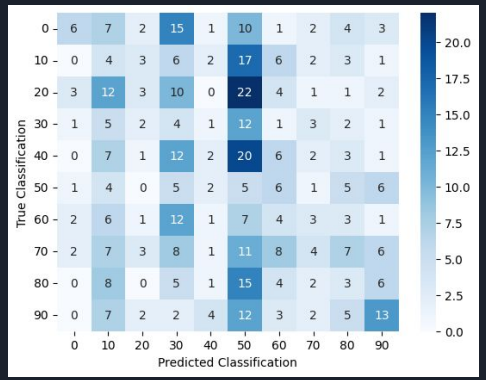


NLP Models: Video Title

Bag of Words

- One-hot encoding
- Applied tf-idf weighting

- **Test Accuracy: 0.10**

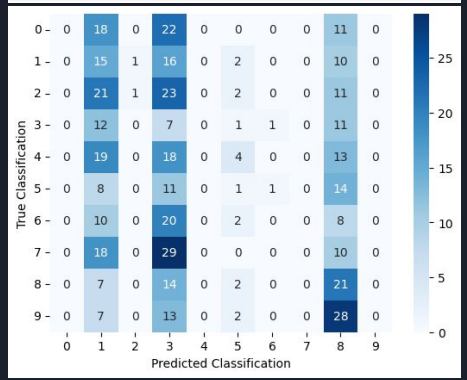


Embeddings

- Tokenized & padded
- Layers: Embeddings, Convolution, & Avg Pooling

- **Test Accuracy: 0.10**

Vocab Size	Sequence Len	Embedding Dims	# of Epochs	Test Loss	Test Accuracy	Val Loss	Val Accuracy
500	12	4	5	2.2913	0.1524	2.3002	0.1075
1000	12	4	10	2.2059	0.2343	2.2781	0.1344
1000	24	4	10	2.2675	0.1620	2.2961	0.1290
2000	24	8	10	2.1992	0.2140	2.2769	0.1398
2000	32	8	10	2.2401	0.1841	2.2881	0.1290



Transformers

- "bert-base-uncased" model with vocab size of 30,522
- Layers: BERT, dropout, linear, & Softmax
- **Test Accuracy: 0.15**

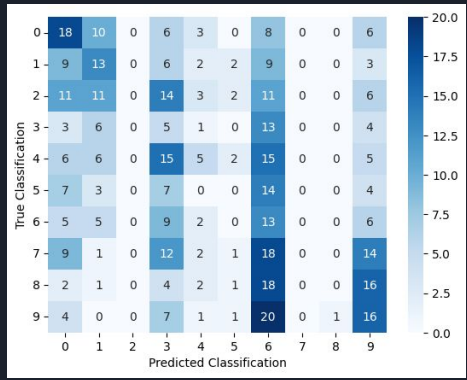


Image Models: Thumbnails - Raw Pixel Data

Model Performance Additional Detail

Model	Accuracy	Precision	Recall	F-1 Score
Baseline Model: Random Classification	0.10	0.10	0.10	0.10
Neural Net	0.09	0.01	0.10	0.02
CNN	0.10	0.03	0.08	0.03

Image Models: Thumbnails - Object Detection

Model Performance Additional Detail

Model	Accuracy	Precision	Recall	F-1 Score
Baseline Model: Random Classification	0.10	0.10	0.10	0.10
Logistic Regression	0.07	0.07	0.08	0.06
Gradient Boosted Decision Trees	0.08	0.08	0.09	0.07
Random Forest (TPOT Model)	0.08	0.08	0.08	0.08
Neural Net	0.07	0.06	0.07	0.06

Image Models: Thumbnails - NSFW Score

Model Performance Additional Detail

Model	Accuracy	Precision	Recall	F-1 Score
Logistic Regression	0.07	0.03	0.10	0.02
Gradient Boosted Decision Trees	0.12	0.12	0.12	0.12
Random Forest (TPOT Model)	0.09	0.10	0.09	0.09

Combined Models and Ensemble Models

Model Performance Additional Detail

Model	Accuracy	Precision	Recall	F-1 Score
Gradient Boosted Decision Trees with Image Data	0.37	0.37	0.37	0.36
Random Forest with Image Data	0.35	0.36	0.35	0.34
Basic Data Ensemble	0.38	0.39	0.38	0.37
Full Ensemble (with NLP)	0.39	0.39	0.39	0.37