Assessment 2b

Predicting Video posting success on YouTube

Problem Summary:

There is both economic and marketing incentive to improve content engagement. This goal is possible because of the nature of recommendation algorithms. They are governed by predictable patterns. Many creators have sought to exploit this concept, and it is an idea that has been studied extensively (Airoldi, Beraldo & Gandini 2016; Alvarado et al. 2020; Bishop 2019). The goal of this project was to use openly available data from the YouTube API to create a classification tree that could predict whether a video would achieve a like:view ratio better than those in the 50th percentile with an accuracy of between 70-80% using features that are available prior to the videos upload and are unrelated to the specific content of the video (eg, specific tags, music, tone, body language, etc.). This concept hinges on the idea that audience behaviours can be manipulated through the upload style alone. By using a classification tree, we can measure feature importance and node boundaries which will offer specific guidance for optimization.

This project holds relevance to many creators and marketing teams. But the primary stakeholder is Chloe Hayden. Hayden has a largely feminine fanbase of neurodivergent and disabled viewers. An extensive search of the platform uncovered 16 other creators with similar audiences, and their content was used to extend the dataset to 5291 instances. In doing so, we improve the accuracy, validity, and generalizability of the model (Zhou et al. 2022). Admittedly, we had hoped to find more creators but sadly this specific audience has fewer options than initial observations indicated.

Table 1: Updated list of Features available prior to video upload that were used to predict like:view ratio (see part 1a, part 1b, part 2a for more details).

Numerical Features	**Explanation**
Title length	The number of words in title. Brevity vs Specificity – perhaps one video style fosters greater interaction.
Description length	The number of words in description. Brevity vs Specificity (Repetition between videos has not been considered)
Number of tags	The number of tags provided Better SEO will bring in a specific audience who may interact better
Tags in title	The number of tags also present in the video title. SEO in title improves clarity and attracts a more engageable audience.
Emoji count – title	The number of emojis present in the title. Foster an emotional connection and may invoke more engaging responses.
Emoji count - description	The number of emojis in the description. Foster an emotional connection and may invoke more engaging responses.
Description sentiment	The average VADER sentiment of the description. Content heavy videos may encourage more people to read and be influenced by the description.
Title sentiment	The average VADER sentiment of the title. One of the first opportunities to influence engagement
Duration in seconds	The duration of the video in seconds. YouTube 'Shorts' have been removed. A measure of available engagement time and may indicate how content-heavy a video is.
Thumbnail text	If text is present in a thumbnail. Potential to capture audience and an opportunity to influence behaviour.
Thumbnail face	If a frontal face profile is in the image. Faces are known to promote engagement (Weismueller & Harrigan 2021).
Vibrant thumbnail	If the thumbnail had saturation over 50% and brightness above 100 (from 0-255). Vibrant thumbnails may promote engagement (Song et al. 2016).
Categorical Features	**Explanation**
Publish day	The specific weekday a video was uploaded on (one-hot encoded). This feature may indicate how well the video is promoted by the YouTube algorithm and if certain days tend to have more favourable behavioural responses.
Publish month	The specific month a video was uploaded (one-hot encoded). See publish day for explanation.
Target Variable	**Explanation**
Likes per view	The like:view ratio a given video has. This feature is independent of time and provides a good benchmark to predict upload success against.
Potential variables	**Explanation**
Upload count	Engagement often increases with time, this feature captures some informative patterns without specifically introducing time (Page & Lopatka 1999). Not deemed worthwhile with so few creators.
Pseudo-time	Brings all geographical regions into an equivalent time to enhance meaning and be more specific in relation to day/month boundaries. Not deemed worthwhile with so few creators.

Analysis

Updated modifications

Previously, the project's success was dependent on the detection of interactions and non-linear patterns as associations were statistically weak in the more limited initial dataset (Batta, Murthy & Savitri 2022; Halim, Hussain & Ali 2022). To further improve the chance of success, features relating to well understood behaviours were constructed (see the green entries in table 1) (Hoiles, Aprem & Krishnamurthy 2017; Song et al. 2016; Weismueller & Harrigan 2021). Thumbnail features were included using open source CNNs (these should be custom built on large datasets to ensure validity in future). Other previously suggested features were excluded as the dataset is restricted to a very niche subset of available videos to best identify patterns in a specific feminine and neurodivergent audience, but when more creators become available these features may become more enticing. Something that was corrected after initial attempts was the stratification of the target variable to maintain class distributions when splitting the data into test and train sets. Another amendment was the addition of the 'max_features' hyperparameter during the tuning phase to limit overfitting by reducing the features available at each split. To ensure the model was achieving peak performance, a gradient boosting classifier was used as a comparison.

Visualizations

Table 2: The metrics observed for all tuned models. Note: five-fold cross validation was used in all cases.

Metric	Initial tree	Updated tree	GB Classifier
CV Mean Accuracy	0.66	0.68	0.71
Accuracy	0.64	0.68	0.635
Precision	0.661	0.69	0.63
Recall	0.65	0.65	0.635
ROC-AUC score	0.65	0.68	0.635

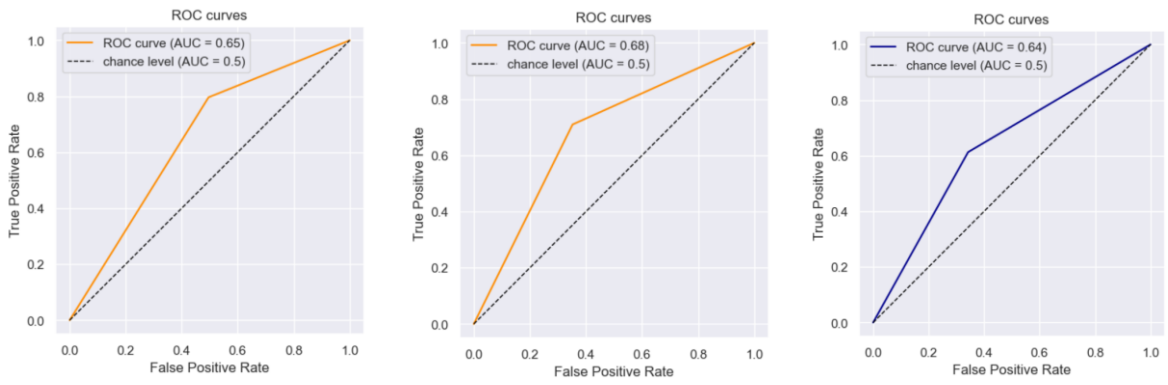

Figure 1: The receiver operating characteristics curve for each of the tuned models. (Left: Tuned model from 2a, centre: final tuned model, right: Bosted classifier.

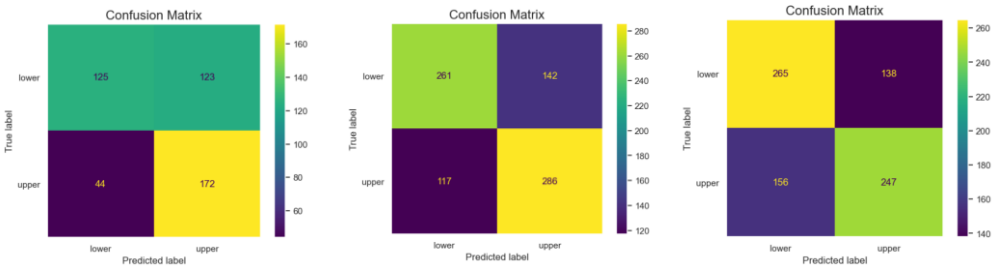

Figure 2: The confusion matrix output from the tuned 2a model (left), tuned final model (centre), and the boosted classifier (right).

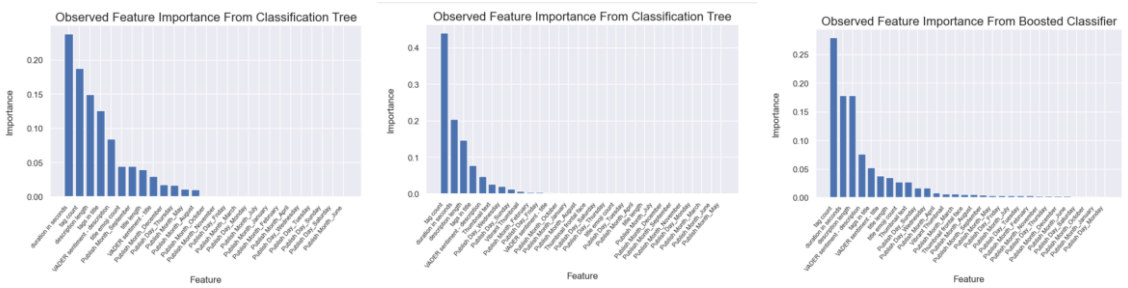

Figure 3: The feature importance's obtained from the tuned 2a model (left), the final model (centre) and the boosted classifier (right).

Improvement

Analysis

The final classifier was ~4% more accurate than the initial classifier presented in 2a, representing a 6.25% improvement. The boosted classifier showed signs of overfitting despite a wide range of tuning parameters and different subsample ratios. In other words, the less expensive model seems to have generalized with higher accuracy and provides more granulated information. The final classification tree may prove useful in assessing posts prior to upload, but they may not generate the most optimal results.

With a global audience, predicting behaviour accurately can be difficult. Thus, a moderate accuracy range was selected. The model failed to achieve the target of 70-80% accuracy with the selected features, but the improvements brought the model much closer. There are additional, but more costly, features that may provide more informative patterns that improve predictive power (*table 2*). However, the model does not require 100% accuracy in order to be informative.

The final classification model implies a few things, whether these are promoted explicitly by human behaviour or by the YouTube algorithm will require more research:

1. Excessive use of tags, or too few tags, may be detrimental.
2. More positive title sentiment scores promote engagement.
3. More positive descriptions that are longer tend to be more favourable.
4. Longer videos perform more poorly, particularly when published on Friday.
5. There are complex relationships present as trees require multiple branches and combinations of features.

Adding additional feature may introduce issues with extrapolation, particularly for Hayden as she has already become something of an outlier within her niche (Malistov & Trushin 2019). However, it is our belief that the concept has been proven effective and can provide advantageous economic and strategic opportunities in this domain, and may be more effective within more generalised categories (eg, gaming videos, makeup tutorial videos, DIY videos, etc.).

Future work

While we elected to use an arbitrary split at the 50th percentile (because data was limited), this may not be the best boundary. Trailing alternatives splits with the available data may be an effective strategy. Additionally, we have previously commented on features we could include to capture information surrounding time. Some have commented that the impact of COVID-19 may alter engagement. However, we believe that efforts would be better spent including features that capture trending media topics/sentiment. Events themselves do not intrinsically change outcomes, but the associated emotional response is reproducibly informative. Please see the table below for more potential features:

Table 2: Suggested features to consider to maximise acuracy in more complex models.

Information content	Features Ordered by Construction Difficulty
High	Upload count Creator popularity “Like and subscribe” call to action (present not present binary) Trending tag (Trending not trending binary) Geographically-weighted Media Sentiment Score Tone of Voice Music (present not present binary)
Moderate	Creator locality Trending music present Vocal features Body language (Open closed binary) Ratio of ‘interaction’ (% creator is seen or heard in a video)
Low	Duplicate Description (Old or fresh binary) Pseudo-time/locality correction Inclusivity, Exclusivity (Use of Jargon, etc.)

Conclusion

With the available data, the classification tree introduced in 2B did not achieve the targeted accuracy. Nonetheless, it outperformed the more intricate alternative and provided insight into the YouTube algorithm and audience behaviours. Some structural adjustments have been proposed, along with various additional features of different complexities and informational value. While this has performed well as a proof of concept, we recommend stakeholders seek to improve accuracy before making decisions based solely on this model's output.

References

Airoldi, M, Beraldo, D & Gandini, A 2016, 'Follow the algorithm: An exploratory investigation of music on YouTube', *Poetics*, vol. 57, pp. 1-13.

Alvarado, O, Heuer, H, Vanden Abeele, V, Breiter, A & Verbert, K 2020, 'Middle-aged video consumers' beliefs about algorithmic recommendations on YouTube', *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW2, pp. 1-24.

Batta, H, Murthy, AV & Savitri, S 2022, 'Predicting Popularity of YouTube videos using Viewer Engagement Features', in *12th International Conference on Cloud Computing, Data Science and Engineering* IEEE, India, pp. 77-81.

Bishop, S 2019, 'Managing visibility on YouTube through algorithmic gossip', *New Media & Society*, vol. 21, no. 11-12, pp. 2589-2606.

Halim, Z, Hussain, S & Ali, RH 2022, 'Identifying content unaware features influencing popularity of videos on youtube: A study based on seven regions', *Expert Systems with Applications*, vol. 206, p. 117836.

Hoiles, W, Aprem, A & Krishnamurthy, V 2017, 'Engagement and popularity dynamics of YouTube videos and sensitivity to meta-data', *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1426-1437.

Malistov, A & Trushin, A 2019, 'Gradient boosted trees with extrapolation', in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA),* IEEE, pp. 783-789.

Page, WH & Lopatka, JE 1999, 'Network externalities', *Encyclopedia of law and economics*, vol. 760, pp. 952-980.

Song, Y, Redi, M, Vallmitjana, J & Jaimes, A 2016, 'To click or not to click: Automatic selection of beautiful thumbnails from videos', in *Proceedings of the 25th ACM international on conference on information and knowledge management,* pp. 659-668.

Weismueller, J & Harrigan, P 2021, 'Organic Reach on YouTube: What Makes People Click on Videos from Social Media Influencers?', in Springer International Publishing, pp. 153-160.

Zhou, K, Liu, Z, Qiao, Y, Xiang, T & Loy, CC 2022, 'Domain Generalization: A Survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 1-20.