**Assessment 2a**

Predicting Video posting success on YouTube

## **Problem description:**

Youtubers receive revenue based on the number of views a video receives, and this is entirely dependent on the YouTube algorithm. YouTubers use many tactics to increase their odds of being promoted by the YouTube algorithm, likes and subscriber counts are known to play a role and gossip within the community continues to tease out more insights (Bishop 2019). The number of likes a video gets is highly dependent on the quality and content of the video itself, but what if the way the video is presented can influence viewer behaviours?

Youtuber Chloe Hayden has a predominantly feminine fanbase of neurodivergent viewers. To eliminate the time dependent nature of these videos, the like:view ratio is used as a measure of viewer engagement over the lifetime of each video. Features that may influence the likelihood of a viewer liking the video have been considered (*table 1).* Other users creating similar content to a similar audience have been sourced to extend the dataset and improve the accuracy, validity and generalizability of the model (Zhou et al. 2022). Using features available at the time of upload, we seek to predict whether a video will have a like per view ratio above or below the 50th percentile with moderate accuracy (70-80%). The best accuracy of the model is unlikely to exceed 80% because features available before viewing are unlikely to warrant a like, but there may be a set of features that promote the behaviour and this information may give Chloe Hayden a competitive edge.

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

|  |  |
| --- | --- |
| **Numerical Features** | **Explanation** |
| Title length | The number of words in title.  Brevity might attract a general crowd but lengthier titles might attract a more specific crowd |
| Description length | The number of words in description.  Like the title length for explanation. |
| Number of tags | The number of tags provided  More tags attract a more precise audience that may respond positively with a like. |
| Tags in title | The number of tags also present in the video title.  This might drive more precise traffic as well, perhaps even from users that were not initially looking for that content at the time. |
| Emoji count – title | The number of emojis present in the title.  This may foster a more emotional connection and invoke a positive response. |
| Emoji count - description | The number of emojis in the description.  This may foster a more emotional connection and invoke a positive response. |
| Description sentiment | The average VADER sentiment of the description.  On content heavy videos users are more likely to seek more information from the description, this feature might correlate with likes. |
| Title sentiment | The average VADER sentiment of the title.  This feature may influence the behaviour of the user prior to watching the video. |
| Duration in seconds | The duration of the video in seconds.  Video shorts have been removed (videos less than 61 seconds). This feature indicates how long a user typically has to place a like and may indicate how content heavy the video is. |
| **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). |
| Pseudo-time | Brings all geographical regions into an equivalent time to enhance meaning and be more specific in relation to day/month boundaries. |
| Thumbnail | Perhaps a CNN to classify the emotional/ behavioural response of the image. |

## **Methodology**

### *Data Selection, Exploratory Data Analysis, &* Preprocessing

Ten Female Neurodivergent content creators were selected, others were sourced but ultimately had too much audience variation to accurately determine audience responses. No linear associations were apparent in numerical features, and unsupervised clustering techniques (t-SNE and UMAP) identified any patterns. Spearman’s rank coefficient demonstrated that most features had a weakly negative association with p values of less than 0.05. Categorical features were observed to be associated with the target variable with chi-squared tests with p-values of less than 0.05. With weak associations present, the project relies on the detection of interactions and non-linear patterns which is consistent with similar research (Batta, Murthy & Savitri 2022; Halim, Hussain & Ali 2022).

In terms of preprocessing, z-scores were taken for numerical features to remove differences in scale. This will not significantly impact classification trees, but makes the data accessible for other modelling. Categorical features were one-hot encoded.

Modelling & hyperparameter Tuning

Classification trees find patterns in non-linear data, are typically robust against skewness and outliers, and can handle multiple data types (Charbuty & Abdulazeez 2021; Greenwell 2022). They also provide an insight into what features are most important as well as the value that splits each node. These can be exploited by the creator to upload videos that grant the video an advantage. Despite the known accuracy drawbacks mentioned above, it still provides an advantage. To ensure the best model was achieved, a grid search was conducted with five-fold cross validation.

Table 2: Summary of hyperparameters and ranges tested during tuning.

|  |  |  |
| --- | --- | --- |
| **Hyper-parameter** | **Value range** | **Support Of Selection** |
| Criterion | ‘gini’, ‘entropy’ | (Raileanu & Stoffel 2004) |
| Max\_depth | Range(1, 27) | The theoretical maximum range |
| Min\_samples\_split | Range(1, 40) | (R. G. Mantovani 2019) |
| Min\_samples\_leaf | Range(1, 20) | (R. G. Mantovani 2019) |

These parameters were selected for tuning as they will minimize the amount of overfitting by reducing the depth of the tree and by controlling the complexity and generalizability of the nodes.

**Results**

Table 3: The best hyperparameters obtained from the grid search.

|  |  |
| --- | --- |
| **Hyper-parameter** | **Best value** |
| Criterion | ‘entropy’ |
| Max\_depth | 7 |
| Min\_samples\_split | 39 |
| Min\_samples\_leaf | 9 |

s

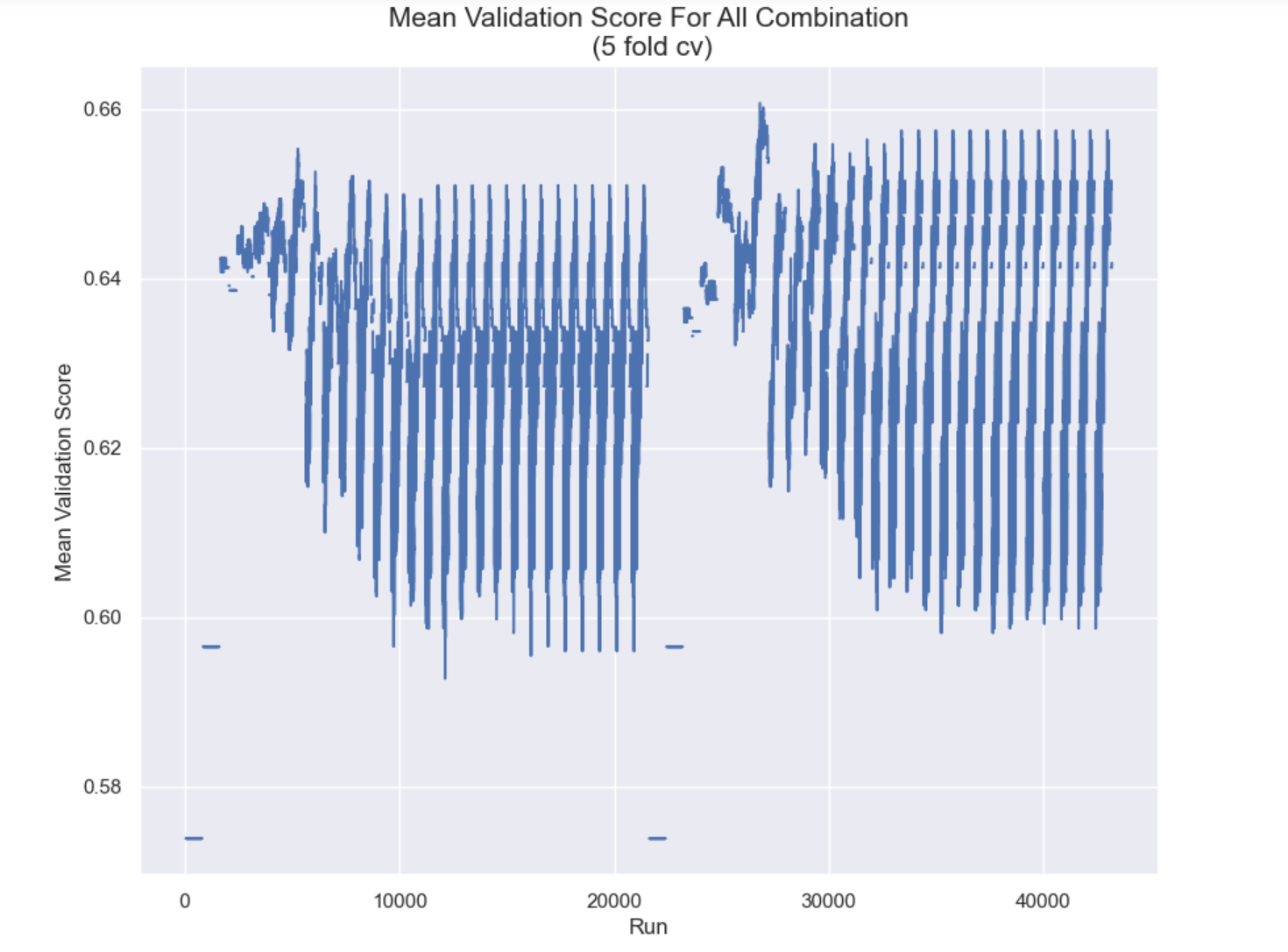


Figure 1: The mean validation scores obtained from each run of combinations, with the maximum existing at 0.6607.

Table 4: The scores obtained for a variety of metrics on the tuned model.

|  |  |
| --- | --- |
| **Metric** | **Score** |
| Accuracy | 0.64 |
| Precision | 0.661 |
| Recall | 0.65 |
| ROC-AUC score | 0.65 |

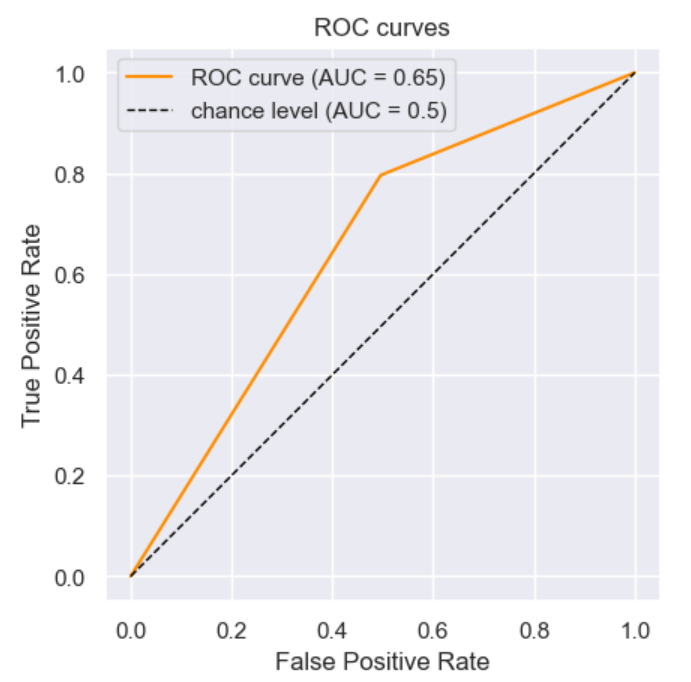


Figure 2: The receiver operating characteristics curve of the tuned model.

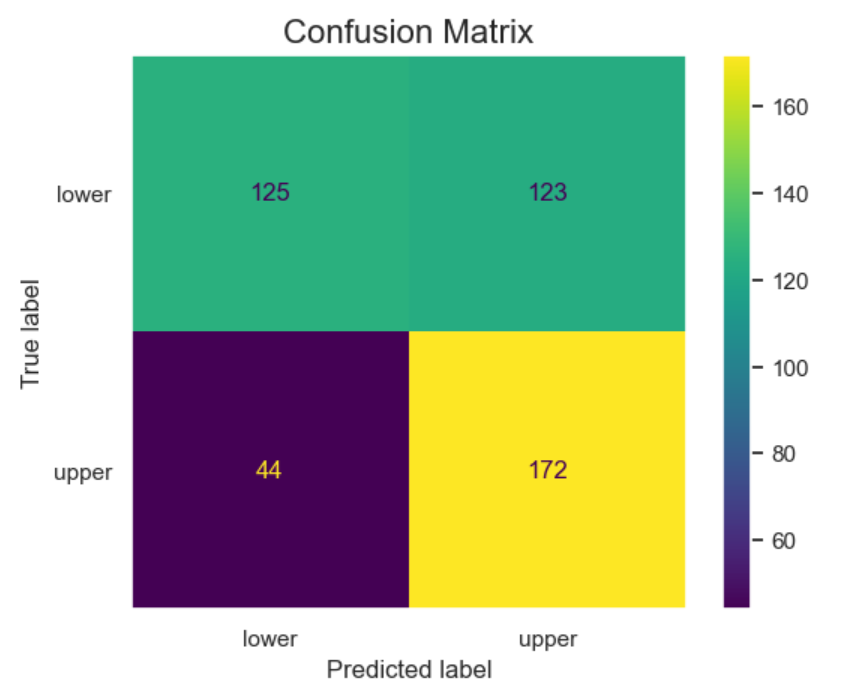


Figure 3: The confusion matrix for the tuned model.

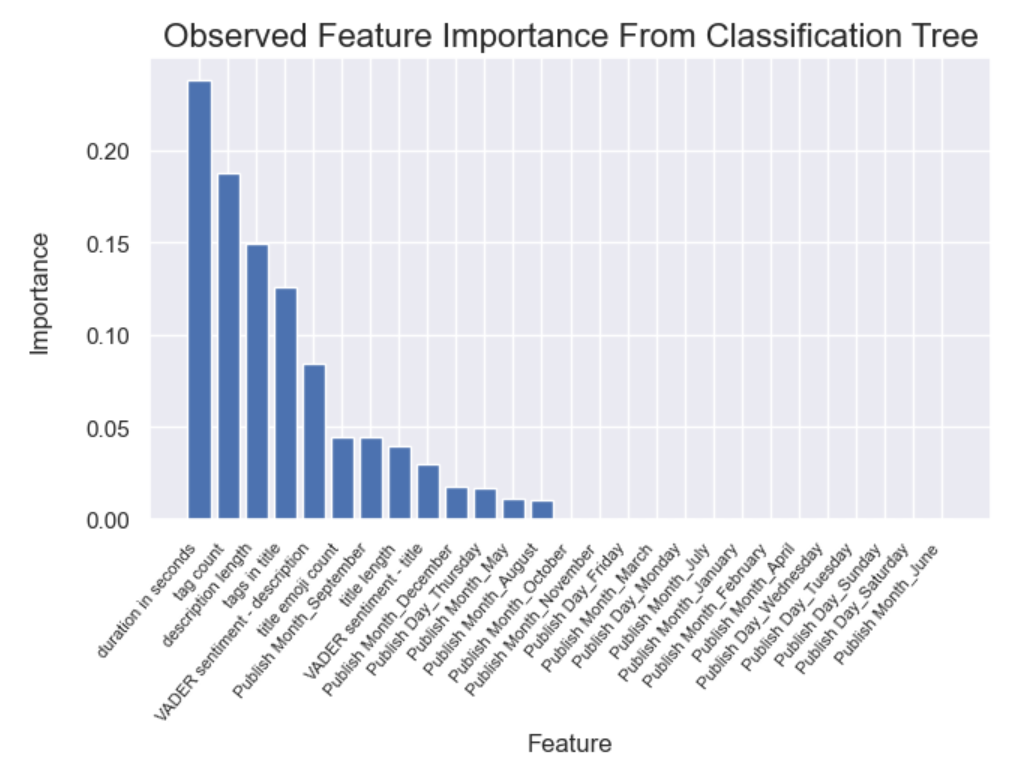


Figure 4: The Feature importance’s obtained from the tuned classification tree.

**Discussion**

Obtaining data with a fixed amount of API tokens has been a continual difficulty in this project. The model may perform better with additional data. With less data than hoped, a lot of attention was given to reducing underfitting and overfitting. An extensive grid search was used to determine hyperparameter values that are known to control overfitting (*table 2*). Five-fold cross validation was used to determine the values based on the best mean validation score (*table 3*). The mean validation score was comparable with the accuracy of the model on the test set (~0.02 difference). However, an accuracy of 0.64 is lower than we expected and hope to achieve. Furthermore, the precision of the model is only 0.661, this indicates the model can accurately predict the positive case only 66.1% of the time. There are more informative features that could improve these metrics but they are either not available to us at the time of prediction/upload, or are more difficult to extract (e.g., number of likes/comments, use of popular music, sentiment of video, subject matter, etc.). Although the model did not perform as well as hypothesized, the output is still more useful than a random selection indicating some of the complex patterns have been identified by the classification tree (*figure 2*). The decision node boundaries can be used to optimize video uploads so they match that of other successful videos, however more time should be put into improving the model before Chloe Hayden can reliably achieve a competitive edge from this model’s output. While the accuracy of the prediction is not the informative piece, it is essential that it is sufficiently larger than the chance level otherwise any benefit may not be worth the time and effort it takes to interpret the decision nodes. The tree itself has been provided separately as a PDF.

Plans for Improvement:

1. Extend the dataset (best expected return on time now that the pipeline is built).
2. Consider the additional features suggested in table 1.
3. Consider ensemble methods such as Extreme gradient boosting, or XGBoost, to improve the model’s ability to identify patterns.

Notes:

1. Altering the datasets included could obtain optimal precision scores of 74% with the same accuracy.

**Summary**

The model was found to perform better than chance, but less optimally than desired. The information obtained and the framework used were able to achieve all that was hoped for. With more time, this model may be able to achieve a level of accuracy which fosters confidence in stakeholders. Some additional variables and other improvements have been suggested.

## **References**

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 &amp; Society*, vol. 21, no. 11-12, pp. 2589-2606.

Charbuty, B & Abdulazeez, A 2021, 'Classification Based on Decision Tree Algorithm for Machine Learning', *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20-28.

Greenwell, BM 2022, *Tree-Based Methods for Statistical Learning in R*, CRC Press.

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.

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

R. G. Mantovani, TH, R. Cerri, S. B. Junior, J. Vanschoren and A. C. P. d. L. F. de Carvalho 2019, 'An empirical study on hyperparameter tuning of decision trees', *arXiv pre-print server*, 2019-02-12.

Raileanu, LE & Stoffel, K 2004, 'Theoretical comparison between the gini index and information gain criteria', *Annals of Mathematics and Artificial Intelligence*, vol. 41, pp. 77-93.

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.