**Assessment 1a**

Predicting Video posting success on YouTube

## **Problem description:**

I have wrangled data from the YouTube API. This option provides a dataset with volume, variety, velocity, and veracity (it also better replicates real world conditions) (Foster et al. 2020). I will be focussing on my partner’s channel, but will extend the dataset with creators like her with a similar audience base. However, things may be made more difficult by the noise her success might have made in terms of inflated bouts of traffic. It may be necessary to consider the impact of these external events on the data over time. Additionally, changes in the YouTube algorithm or media trends have not been accounted for as that is a much lengthier exercise, but these may improve model accuracy in future iterations (Paleyes, Urma & Lawrence 2023). In the entertainment industry, the goal is usually to produce content that is relevant and engages more of an audience (Munaro et al. 2021). In this project, I will be answering the question ‘how can I post content that will receive high engagement as measured by like to view ratios?’

See the below table for some of the relevant questions we could use this data to answer:

Table : Potential questions this dataset could be used to answer.

|  |  |
| --- | --- |
| **Potential question** | **Potential solution** |
| Will a video surpass a certain threshold of likes, views, or engagement? | Decision tree (Song & Ying 2015) |
| Are there any clusters in the video types uploaded, and is there a niche outside of these that the creator could exploit? | Clustering question, likely density-based method (Ampili & Kanakala 2022) |
| Do certain types of thumbnail result in more likes/views or increase their rate? | CNN question (Amrutha 2022) |
| Does the sentiment of the description/title and/or use of emojis result in more views/likes? | Regression or clustering question depending (Borg & Boldt 2020; Kralj Novak et al. 2015) |
| Does promoting the content on other social media channels result in more views/likes? | Regression question (Arora et al. 2019) |

## **Dataset description:**

There was no missing data or duplicates from the YouTube API. It had the initial shape of 309 rows and 14 columns, but as previously mentioned I have the capacity to pull in other similar YouTube channels to bulk up the dataset by an additional 10,000 rows per day. (This is also more of a real-world application than if I had chosen a stagnant Kaggle dataset).

Table : The original features retrieved from the YouTube API.

|  |  |
| --- | --- |
| **Initial feature** | **Explanation** |
| Video ID | The unique identifier for the specific video in the channel’s library. |
| Publish date | The exact time published, formatted as ‘YYYY-MM-DDTHH:MM:SSZ’ (UTC). |
| Tags | Any tags the author labelled the video with. |
| Comment count | The number of comments a video has received to date. |
| Title | The title of the video. |
| Description | The description posted below the video. |
| Times viewed | The number of times the video has been viewed to date. |
| Times liked | The number of likes the video has received to date. |
| Times favourited | The number of times the video has been favourited by users to date. |
| Duration | The length of the video received in the format ‘PT\_\_H\_\_M\_\_S’ |
| Caption | If captions have been provided for the video (true or false) |
| Privacy status | If the video is publicly viewable. |
| Localization | Language specific meta data associated with the video |
| thumbnail | The URL of the thumbnail image. |

## **Initial data processing:**

Python was used to connect to the API and request all the information above. A few modifications were made to the values present:

* Converting the date from a string into a datetime format
* Converting the duration into a measure of seconds from a time string of hours, minutes, and seconds.

To obtain a greater degree of information from the above list, some features engineering was undertaken:

Table : Additional features obtained through feature engineering the data received from the YouTube API.

|  |  |
| --- | --- |
| **Engineered features** | **Explanation** |
| Title length | How long the title is. This may be correlated with views. |
| Description length | How long the description is. This may indicate if a video is content heavy or more light-hearted and playful. |
| Number of tags | The number of tags a video has. This may correlate with views |
| Days live | The number of days a video has been viewable for. This is necessary to determine rates. |
| Comment count per day | The average number of comments per day. |
| Likes per view | The number of likes per view. |
| Times liked per day | The average number of likes per day. |
| Times viewed per day | The average number of views per day. |
| Images | These are the images found at the thumbnail URL. These may be useful with a CNN (Amrutha 2022). |
| Tags in title | The number of tags also present in the video title. This is relevant as it might better direct traffic. |
| Upload day | The day a video was uploaded on. |
| Month uploaded | The month a video was uploaded |
| Emoji count – title | The number of emojis present in the title. |
| Emoji count description | The number of emojis used in the description. These invoke an emotional response/ relationship and may guide traffic. |
| Description sentiment | The VADER sentiment of the description. |
| Title sentiment | The VADER sentiment of the title. |

Further Exploratory analysis will be needed to observe the strength of relationships and any multicollinearity. Distributions may also need to be modified to be included in any models. This alone will be informative to the creator and is well worth the exercise.

Time permitting, I may endeavour to access the Instagram/facebook API to obtain datetime indexed information about when she posted content and if it was to promote a video she had released on YouTube as alluded to in one of the suggested questions above.

## **Refined problem and plan:**

The most informative initial question revolves around determining what makes a video engaging. To answer this question, we can no longer use time dependent features because we will not have posted the video yet so making predictions when the counters are at 0 is uninformative. The target variable will be the ‘likes per view’ and we will only include features that are known of the video prior to it being published (including features associated with the title, description, and thumbnail). It would be easiest to categorize the target variable into three groups in a way that avoids data imbalances, their labels will be ‘low’, ‘moderate’, and ‘high’ engagement. The data will be split into training and test sets, and 5-fold cross validation will be used to validate the model and measure any over/underfitting. A decision tree model could be used to determine feature thresholds and importance and allow creators to optimize their posts (Song & Ying 2015). These models do not extrapolate well, however the maximum value in this instance will always be 1 so the target labels should maintain good relevance. A random forest model may also improve the model’s generalizability or accuracy; however, the interpretability of the decision tree is the informative piece as it helps guide future content curation decisions and this is not as informative in random forest models (Kotsiantis 2013). However, an accurate random forest model would provide the creator with a useful tool to predict engagement prior to posting.

Table : Summary of potential issues and methods to overcome those issues.

|  |  |
| --- | --- |
| **Potential issues** | **Solution** |
| More data needed | Expand YouTube dataset with similar creators,  Eg, Page Layle, Aspieworld, Jessie Paige, etc. |
| The model is overfitting/underfitting | Better tune hyperparameters,  Include more instances. |
| Model accuracy is still low | Consider external factors such as YouTube algorithm changes or trending topics (Bärtl 2018; Bishop 2019; Kotsiantis 2013). This could be obtained from google trends with key words.  Consider support vector machine or random forest alternative. |

As the exploratory data analysis continues, a fatal flaw may be identified. In that scenario, I will consider if social media data or general media data could improve the outcome (this is publicly available through Meta’s graph APIs). If it cannot be rectified, then I will shift to answering one of the other questions posed above.

## **References**

Ampili, K & Kanakala, S 2022, 'Tweet Summarization Using Clustering Mechanisms', in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N),* IEEE, pp. 1481-1486.

Amrutha, K 2022, 'Predicting Popularity of Online Videos Using Support Vector Regression with Gaussian Radial Basis Functions', *ECS Transactions*, vol. 107, no. 1, p. 16943.

Arora, A, Bansal, S, Kandpal, C, Aswani, R & Dwivedi, Y 2019, 'Measuring social media influencer index- insights from facebook, Twitter and Instagram', *Journal of Retailing and Consumer Services*, vol. 49, pp. 86-101.

Bärtl, M 2018, 'YouTube channels, uploads and views:A statistical analysis of the past 10 years', *Convergence*, vol. 24, no. 1, pp. 16-32.

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

Borg, A & Boldt, M 2020, 'Using VADER sentiment and SVM for predicting customer response sentiment', *Expert Systems with Applications*, vol. 162, p. 113746.

Foster, I, Ghani, R, Jarmin, RS, Kreuter, F & Lane, J 2020, *Big data and social science: data science methods and tools for research and practice*, CRC Press.

Kotsiantis, SB 2013, 'Decision trees: a recent overview', *Artificial Intelligence Review*, vol. 39, no. 4, pp. 261-283.

Kralj Novak, P, Smailović, J, Sluban, B & Mozetič, I 2015, 'Sentiment of Emojis', *PLOS ONE*, vol. 10, no. 12, p. e0144296.

Munaro, AC, Barcelos, R, Francisco Maffezzolli, EC, Rodrigues, JP & Paraiso, E 2021, 'To engage or not engage? The features of video content on YouTube affecting digital consumer engagement', *Journal of Consumer Behaviour*, vol. 20, no. 5, pp. 1336-1352.

Paleyes, A, Urma, R-G & Lawrence, ND 2023, 'Challenges in Deploying Machine Learning: A Survey of Case Studies', *ACM Computing Surveys*, vol. 55, no. 6, pp. 1-29.

Song, Y-Y & Ying, L 2015, 'Decision tree methods: applications for classification and prediction', *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130.