**Evergreen Content Analysis**

**Objective**

To define which pieces of content produced will remain evergreen. In this project, data analysis/science techniques are implemented to define evergreen videos across three YouTube Channels (Carli Bybel, FamousTubeFamily and The Rush Fam).

**Evergreen Definition**

We define a video as evergreen if it remains relevant despite of its age on the YouTube. Metrics like monthly revenue, likes and views are used as tools to aid in calculating relevance. Together with the video’s rate of change (decay/growth), it is possible to determine which videos will be most evergreen.

A graph with numbers and a line

Description automatically generatedA graph with numbers and a red line

Description automatically generatedIdeally, we would like to see videos that have growth and generate significant revenue from its posted date to present day. But realistically the opposite happens. Most, if not all videos will decay and generate less revenue over time. The evergreen ones are those that show a gentle descend instead of a steep one.

Fig 2 – Steep Decay

Fig 1 – Gentle Decay

In the above graphs, we can clearly discern that Fig 1 is better as it is decaying at a slower rate compared to Fig 2.

**Classification Methodology**

To classify how evergreen a video is, we have come up with an equation that returns an evergreen score.

Median Rate of Change (percentage) represents the video’s month on month decay. Taking the median gives us a better representation of the video’s performance as it removes outliers. The same reasoning is applied to median revenue. A positive number represents decay and negative for growth. We then subtract it from 1 to derive inverse percentage for multiplication with median revenue. Doing so allows an accurate “effect of decay” representation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Video | Median Rate of Change | Inv Rate of Change | Median Revenue | Evergreen Score |
| Video 1 | 0.8 (decay) | 0.2 | 15 | 3 |
| Video 2 | 0.4 (decay) | 0.6 | 10 | 6 |
| Video 3 | -0.1 (growth) | 1.1 | 8 | 8.8 |

Table 1 – Evergreen Equation Proof of Concept

Table 1 breaks down and explains the equation. We can see that video 1, with higher decay and revenue returns a lower score than video 2, which has lower decay and revenue. Better still if the video displays growth.

Next, these data are then run through a clustering algorithm known as K-means clustering. The algorithm splits the data into N number of clusters (N = the number of clusters for the algorithm to define. E.g. If N = 4, data will be split into 4 clusters). Each cluster will have a centroid aka cluster average. These centroids label the cluster as either good or bad, thus any data associated with the cluster will also be given the same label. Essentially, the K-means algorithm classifies and assigns labels to unlabeled data, hence it is an unsupervised learning method.

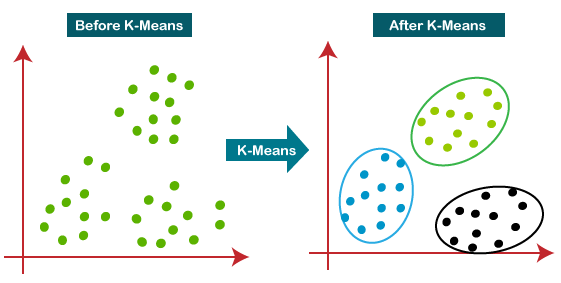


Fig 3 – K Means Clustering

**Implementation**

1. **Data Retrieval**

The data was pulled from the Bankeble database using SqlAlchemy. Columns pulled includes ‘channel\_id, video\_id, period, likes, views, estimated\_revenue and estimated\_revenue\_total’. The data pertaining to each channel were stored in three different pandas data frames.

A screenshot of a computer

Description automatically generated

Fig 4 – First five video records for Carli Bybel

1. **Data Preprocessing**
   1. **Cleaning Null Values**

In this step, data cleaning is required before any calculations can be performed. This is to ensure that data being fed into calculations are of good quality. For example, some records may contain null values. These null values if left in the dataset will lead to errors in calculation.

* 1. **Populating new columns and missing data**

Other steps include preparing additional columns for future use like cumulative views and like, sorting of records according to the date created which is important when working with time series data. For channels like FamousTubeFamily and The Rush Fam, there are missing revenue data due to the channels outsourcing their income. To account for this, we populate these missing data with estimations based on the last known month with revenue data as we still have access to the views data.

|  |  |  |
| --- | --- | --- |
|  | Views | Revenue |
| Month with no missing data | 100 | $10 |
| Month with missing data | 150 | $10/100\*150 = $15 |

Table 2 – Revenue estimation explained. Estimated value highlighted.

* 1. **Data Reduction**

Next, we filtered the dataset to only include videos that are at least 13 months old and make up 80% of the channel’s total revenue. By doing so, we can reduce our data to concentrate on the more important videos that affect the channel’s performance. This based on the pareto principle whereby the 80% of the result is generated by 20% of the inputs.

* 1. **Rate of Change**

After data reduction, we calculate the median rate of change for each video. Rate of change is calculated with this equation.

The median is derived using a python method and returned as output.

* 1. **Seasonal Detection**

Some videos created by the channel may be seasonal. Seasonal videos often display a pattern to when they are most effective.

A graph with numbers and a line

Description automatically generated

Fig 5 – Halloween video displaying season pattern.

In the above example chart, we can clearly see a consistent spike every October as it is a video related to Halloween. To computationally detect similar patterns, we break the data into groups of twelve. These groups represent each month in every year since the video was posted. The month with the highest revenue in each group is identified. If the month with the highest revenue is always the same month in each group, the video is deemed as seasonal. **Note that the video must be at least three years old to avoid false positives.**

* 1. **Collating final preprocessed dataset**

To create the finalized dataset, we create groups by unique video\_id to represent each video. Earlier mentioned rate of change and seasonal detection is calculated for each group. The median revenue is derived using a python method. Once all these prerequisite metrics are derived, we can calculate the evergreen score for each video with the previously mentioned evergreen score equation. All metrics are then populated into a new dataframe.

A screenshot of a graph

Description automatically generated

Fig 6 – Sample dataframe of The Rush Fam

1. **Data Visualization**

Using median rate of change and revenue as our axes, we plot the data onto a scatter plot to visualize each channels video performance distribution.

* 1. Carli Bybel

A graph with red dots

Description automatically generated

Fig 7 – Carli Bybel video distribution

A screenshot of a graph

Description automatically generated

Fig 8 – Carli Bybel top 5 evergreen videos

* 1. The Rush Fam

A graph with red dots

Description automatically generated

Fig 9 – The Rush Fam video distribution

A screenshot of a graph

Description automatically generated

Fig 10 – The Rush Fam top 5 evergreen videos

* 1. A graph with red dots

     Description automatically generatedFamousTubeFamily

Fig 11 – FamousTubeFamily video distribution

A screenshot of a graph

Description automatically generated

Fig 12 – FamousTubeFamily top 5 evergreen videos

The scatter plots give us rough insights into the channel’s performance. For example, it shows us that only a handful of videos in each channel generate significant revenue. However, the scatterplot does not classify what is considered a good or bad video.

1. **K-Means clustering**

To achieve classification, median rate of change and revenue are used as two features to train the K-means model. The model is instructed to identify four different clusters. These clusters represent “Very Good, Good, Bad and Very Bad”. As explained earlier, videos in the “Very Good” cluster will be labeled the same, likewise for the different clusters and their associated videos. This will then be updated into a new and final dataframe.

Do note that when generating the clustering results from the algorithm, the cluster labels are inconsistent. For example, the graph for Carli Bybel will show cluster 2 as the best cluster, but the graph for FamousTubeFamily shows cluster 0 as the best cluster. To standardize this inconsistency, code has been written to rename the cluster into the more readable representation as mentioned earlier.

These graphs show the results generated by the K-means algorithm for all three channels.

1. Carli Bybel

A graph with red and blue dots

Description automatically generated

Fig 13 – Carli Bybel Clusters

|  |  |
| --- | --- |
| Cluster | Centroid Evergreen Score |
| 0 | 4.02192 |
| 1 | 68.82091 |
| 2 | 143.36516 |
| 3 | 24.87232 |

Table 3 – Carli Bybel best perfoming cluster

A screenshot of a graph

Description automatically generated

Fig 14 – Carli Bybel Top 5 videos. Note the new cluster column.

1. The Rush Fam

A graph with red blue and gray dots

Description automatically generated

Fig 15 – The Rush Fam Clusters

|  |  |
| --- | --- |
| Cluster | Centroid Evergreen Score |
| 0 | 400.42599 |
| 1 | 29.35987 |
| 2 | 830.59412 |
| 3 | 165.53630 |

Table 4 – The Rush Fam best perfoming cluster

A screenshot of a graph

Description automatically generated

Fig 16 – The Rush Fam Top 5 videos. Note the new cluster column.

1. FamousTubeFamily

A graph with different colored dots

Description automatically generated

Fig 17 – FamousTubeFamily Clusters

|  |  |
| --- | --- |
| Cluster | Centroid Evergreen Score |
| 0 | 1889.93357 |
| 1 | 429.85644 |
| 2 | 1009.69790 |
| 3 | 72.42047 |

Table 5 – FamousTubeFamily best perfoming cluster

A screenshot of a graph

Description automatically generated

Fig 18 – FamousTubeFamily Top 5 videos. Note the new cluster column.

**Future Uses**

The workflow of evergreen score and evergreen classification can be added as new metrics in the bankeble system. Each video evergreen score shows its potential in being sustainable and the classification label makes identifying evergreen videos easier. Essentially the videos will have a new evergreen rating.

As for channel-to-channel comparison, the percentage of videos in the very good cluster will be a good telltale sign of creator performance. For example, if 15% of channel A’s videos are in the very good cluster as compared to channel B, which has 5% of its videos in the same cluster, channel A will be deemed as better performing.