Data Mining Final Project Paper

Giancarlo Ayllon, Chris Bernal, Alex Chavez, Ashlyn Sam, Jonathan Wallace

Seton Hall University

CSAS 3010-AA Data Mining

Dr. Hsin-Yi Lin

December 9, 2024

Data Mining Final Project Paper

As of now, almost everyone, if not all, has heard of and used the online video platform YouTube. YouTube is a social media platform that allows its users to upload, view, comment, and watch video content. With almost 2.7 billion users per month, YouTube is the second largest search engine and also the second most popular social media platform. The influence that this platform possesses provides a stronghold for unlimited creativity and connection worldwide.

## Objectives

Due to its billions of users, YouTube generates most of its revenue through advertisements. Additionally, creators who can successfully build and grow their brand on the platform can make a living solely by posting video content. By using data mining concepts and techniques, is it possible for us to forecast YouTube success rate strategies? And more specifically, is there a correlation between upload frequency to video views and subscribers?

## Methodology

To begin, we imported multiple libraries that we learned from throughout our labs such as NumPy, Pandas, Matplotlib, and Scikit Learn before importing our chosen dataset. The very first step was to clean out the data of any outliers; we had to normalize data due to skew using log transformations, we did so for video views, subscribers, and uploads (Figure 1.1 and 1.2). For example, we got rid of any 0 uploading values, viral videos, music channels, as well as music celebrity channels. We removed these outliers because our objectives are targeted more towards YouTube channel creators who were not already famous to begin with. To find these outliers, we used bar charts and quartile checks (Figure 1.3). Once we got rid of these outliers, we input a box plot to make sure we didn’t miss any other anomalies (Figure 1.4).

After our data has been prepared, we had to perform data exploration. To help with this, we performed a correlation heat map (Figure 2.1). We chose a correlation heat map to get a visual representation of how close the variables are in relation to one another. Another data exploration model that was used was scatter plots to add an additional visualization pattern to the relationships (Figure 2.2). Our last data exploration model that we used was bee swarm plots (Figure 2.3). These models helped us show density and potential trends, and it also made it easier to identify clustering.

For our data modeling we used a couple of models like linear regression, KNN models, logistic regression and random forest model. We used linear regression first because not only is it easy to implement, but it also generates predictions and possible relationships between our desired variables (Figure 3.1). Our KNN model represents misclassifications, and if there are any false positives and negatives (Figure 3.2). It was effective for capturing non-linear patterns, but also sensitive to noisy data. The third model that we presented was our logistic regression model (Figure 3.3). We used this model because it predicts the probability of multiple variables. It was a good model for binary classifications, but it struggled with complex patterns. And our last model was the random forest model (Figure 3.4). The purpose of this was to combine multiple predictions and turn it into one big prediction. All in all, the models and methods we used were able to satisfy our objective.

## Results

We performed two linear regression models on upload rates and subscribers (Figure 3.1). As we know, the closer to the dashed line the points are, the greater the relationship. From our upload rate model, we can interpret that there is a positive relationship with viewership. The subscriber count model, on the other hand, didn’t have much of an impact, as the data had a “flat” relationship, instead of following along the regression line. The KNN model we carried out uses the upload count as a predictor (Figure 3.2). From what we gathered from the model, it shows that higher uploads correlate with higher viewership. The model also suggests that there are more channels with high view rates in relation to their upload counts.

In our logistic regression model, the area under the curve (AUC) score measures the overall performance (Figure 3.3). Our AUC score was 0.63. This means that the model has a 63% chance of how effective the probability of high video views to upload count; informing us that there is a moderate correlation between the pair. The random forest model shows that by combining multiple decision trees, it can handle complex data patterns and provide valuable insights into which features were most crucial for predicting YouTube success (Figure 3.4). This model had the best overall performance, informing us of important features like subscribers, rank, and upload frequency.

## Conclusion

Through our research and the models that were presented, we found that our last model that we performed, the random forest model, was the key contributor to our findings. What we can answer from our research questions through our interpretation of the random forest model is through this analysis, it shows that upload count doesn’t have a *huge* role in a YouTube channel’s video views and subscriber count, but it does have a small and indirect impact. Therefore, we can conclude that though consistent uploads don’t significantly influence a channel’s video views and subscriber count, there is still a positive relationship with it.

## Ethical Discussion

Data mining has proven to be a useful tool in making data-driven decisions through analyzing the patterns and relationships of data. However, with great data comes great responsibility. We extracted our data from Kaggle, a platform that holds datasets targeted towards the data science community. While Kaggle does have implemented policies to prevent and protect user and data privacy, that doesn’t mean that there haven’t been instances where privacy was unintentionally breached.

In this case, we can critique that the data we utilized, while it does say *who* created the dataset, it doesn’t specify *how* he did it. This can contribute to the ethical issue of data privacy. Technically YouTube statistics are publicly visible, however creators may not expect or consent their data to be analyzed and repurposed in the way that they didn’t initially intend for it to be.

Another potential ethical issue may impact present or potential YouTube creators. We’re extracting data to figure out if there is a data-proven strategy that determines the success of a channel. The answer that we found could discourage creators and pressure them into following that strategy, which can ultimately lead to less creativity and diversity of content. And although our findings may be insightful that doesn’t mean that it should be the standard.

## CRISP-DM Standard

Our project aligns with the 6 stages of the data mining life cycle. Stage 1 was business understanding. We were able to brainstorm and mutually agree that YouTube be our focus. We then narrowed down and came up with the objective and asked ourselves if there are any patterns in data that can figure out if there is a relationship between the number of content put out with the number of subscriber growth. Stage 2 is data understanding. Once we figured out what our objectives were, we had to search and assess the right dataset that tied into our research question. Data preparation was step 3 of the data mining life cycle. After downloading our dataset, we had to clean and filter out our data. For example, we had to normalize it by taking out any outliers, in this case we took out viral YouTube videos as it didn’t pertain to our research question. Another data cleanse that we performed was filtering out any music content, like Vevo or famous singer channels, as our focus was targeted on non-musical YouTube channels. The 4th stage is modeling, and we used data mining techniques like regression, heat maps, and cross tabulation to find a predictive and descriptive model that shows the impact of upload frequency.

Evaluation is the 5th stage where we are able to assess the model and make any changes to adjust them to our objective. Once we are satisfied with the evaluation we can move on to the next stage. The 6th and final stage of the data mining life cycle is deployment where we were able to share our findings and insights by writing a summary paper, saving and showing our methods through a workbook and giving a presentation to our peers. By following the data mining life cycle, CRISP-DM Standards, we can follow a road map ensuring that our project stays relevant and structured.

## Teammate Contributions

After we all mutually agreed on our topic and objective, the project was split into three main parts. Chris and Jonathan were in charge of data preparation and exploration. Following steps 2 and 3 of the data mining life cycle, they were able to format and clean out any data that wasn’t going to contribute anything to our objective. Giancarlo and Alex were responsible for the data modeling portion of the project which were steps 4 and 5 of the life cycle. They used data mining techniques and concepts such as KNN models, random forest model, linear, and logistical models. Ashlyn was in charge of writing the summary ethics paper and powerpoint presentation. Through organization and knowing each member’s roles in the project, was the last step of the data mining life cycle: putting all the work into a documentation and visual representation for our findings to be shared and discussed among others.

**Charts and Graphs**

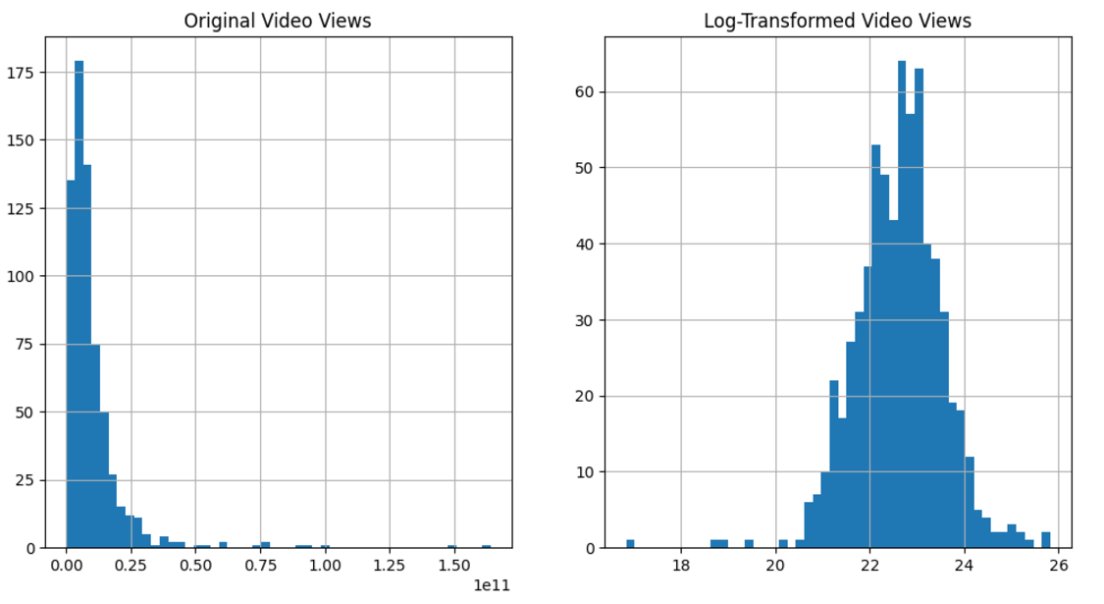


Figure 1.1

A group of blue and white graphs

Description automatically generated

Figure 1.2

A diagram of a graph

Description automatically generatedA diagram of a graph

Description automatically generated

Figure 1.3

A screen shot of a graph

Description automatically generated

Figure 1.4

A screenshot of a graph

Description automatically generated

Figure 2.1

A diagram of a diagram with a red line and blue dots

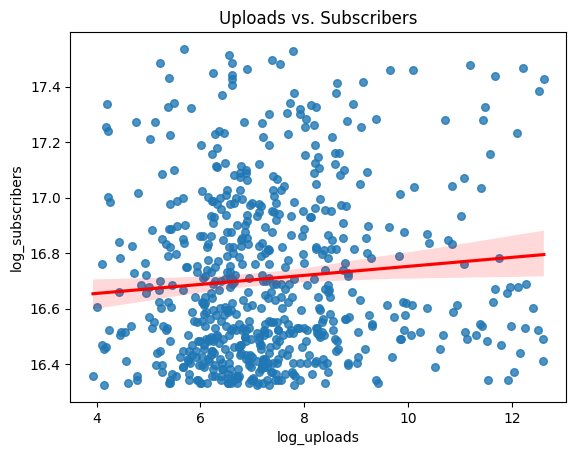
Description automatically generated

Figure 2.2

A screen shot of a screen

Description automatically generatedA screen shot of a screen

Description automatically generated

Figure 2.3

A graph with blue dots

Description automatically generatedA green and black line graph

Description automatically generated

Figure 3.1

A blue squares with numbers and lines

Description automatically generated

Figure 3.2

A graph of a logistic curve

Description automatically generated

Figure 3.3

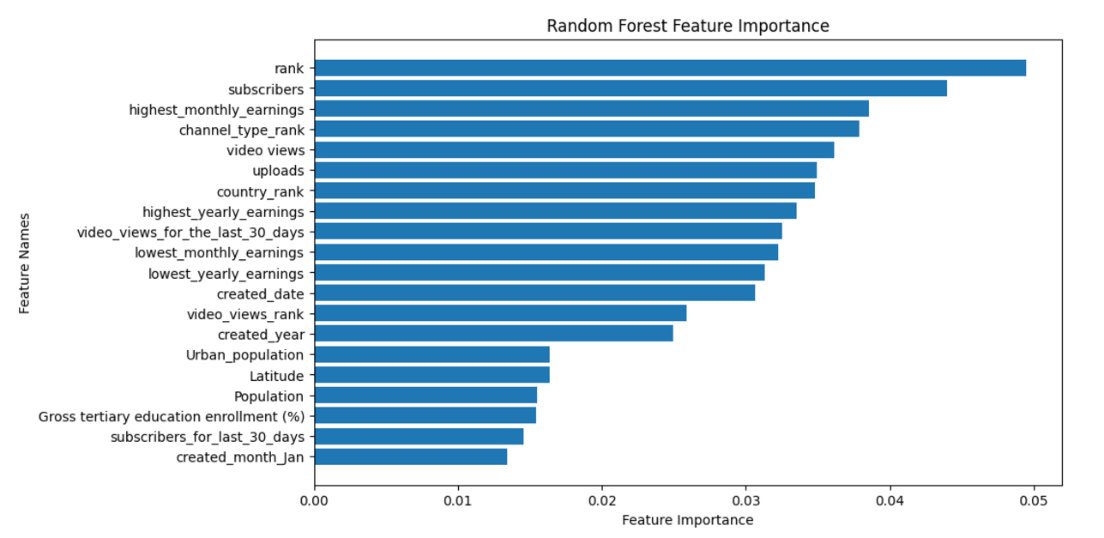
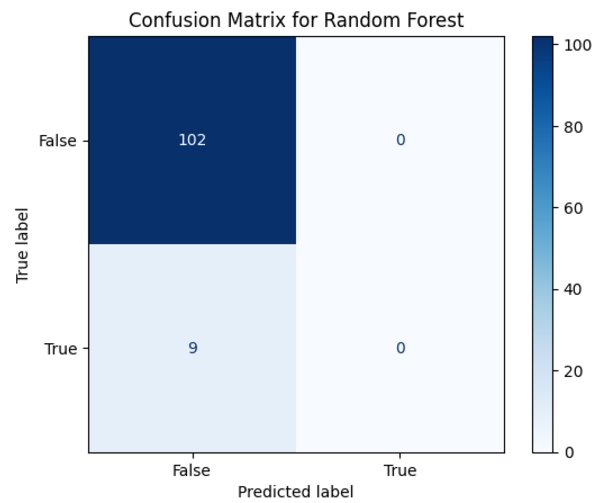


Figure 3.4