YouTube Analytics: Insights and Predictions

Gesi Morris-Odubo

COMP 4980: Machine Learning

Thompson Rivers University

December 2, 2024

Table of Contents

YouTube Analytics: Insights and Predictions

This dataset focuses on YouTube video analytics, providing insights on video performance, audience engagement and revenue metrics across YouTube videos. It tracks views, likes, comments and other performance indicators to optimize video content performance. The design of this dataset enables the metrics to be used to improve monetization strategies with numerical data as the KPI's: views, likes, comments, revenue, and categorical data as: video category, ad sources, and day of publishment. The file was around 137 KB with 362 rows of data. When looking at the variables included in the dataset, we have several categories. For the video details, there is video duration, video publish time, days since published, and day of week. For the revenue metrics, there is revenue per 1000 views, estimated revenue, ad impressions, and ad revenue sources. Engagement metrics has views, likes, dislikes, shares, comments, average view duration, average view percentage, and view thumbnail ctr. Audience data has new subscribers, unsubscribes, unique viewers, returning viewers, and new viewers. Lastly, monetization metrics has monetized playbacks, playback-based CPM, YouTube premium revenue, orders, and total sales volume. The file is formatted in csv. This allows for easily manipulation and analysis of the data on python. The data is well structured with compatible data types in python and python libraries. Each row and column are well defined, and data is easy to read. In terms of performing machine learning, there are many analytics that are suitable with this dataset. The variety of metrics it has makes it great for supervised learning models. The dataset is taken from Kaggle which is a trusted source for data. It also uses real YouTube data that comes from a credible YouTube analytics Api. Data cleaning could be done still to ensure the dataset is more reliable. There are many columns that only contain 0s, but these metrics can be overlooked when doing the analysis. On Kaggle, the dataset has a 10-usability score and is licensed. Looking at the last update, it was updates 25 days ago. Overall, this dataset is valuable for analyzing various

factors that affect monetization and viewer retention on YouTube. It can help gain insights on the impact of content on audience retention and channel growth, trends in audience behavior, effectiveness of different monetization strategies, and how content strategies can maximize views, engagement and revenue.

Data Analysis

The dataset provides insights into YouTube video performance through a variety of metrics. I will be analyzing the views, watch time, revenue, impressions, likes, dislikes, shares, new comments, new subscribers and unsubscribes metrics to derive a conclusion about the relationships between them. Analyzing these metrics together provide a holistic view on the engagement, monetization and visibility of YouTube content and identifies potential improvement in content strategy. A descriptive analysis of these metrics reveals each video has a substantial average number of views however, the average watch time varies with a minimum watch time of 12.69 hours and a maximum of 53,794.66 hours. This indicates diversity amongst the videos in audience engagement. We can illustrate this with a scatterplot to visualize the relationship between the average number of views and watch time.

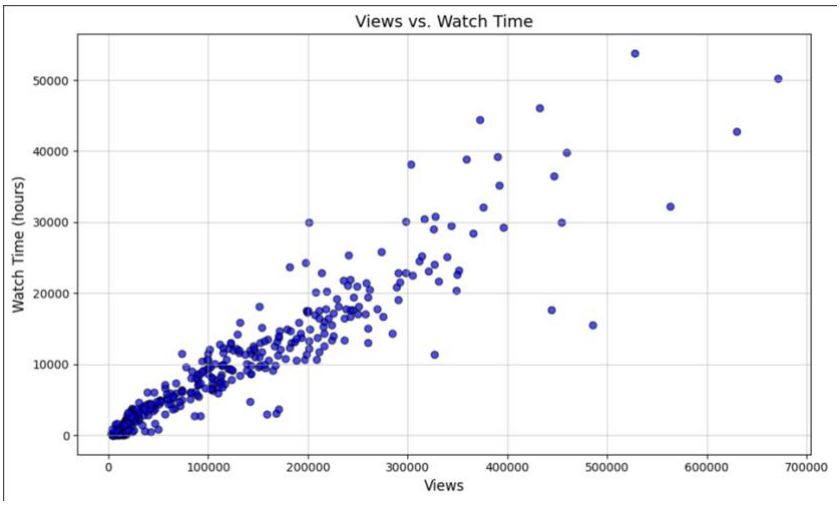

Figure 1: Scatterplot of views vs watch time (hours)

Figure 1 shows a positive relationship between the views and watch time. This relationship is to be expected since videos with larger view counts would have more watch time since a recorded view means the video was watched for a longer period.

YouTube offers a premium subscription for users who want to watch videos without having to view ads. Most of youtubers revenue come from the ad revenue so premium users pay youtubers less than regular users. Looking at the AdSense revenue compared to revenue from premium views can show how the number of premium viewers affects the ad revenue.

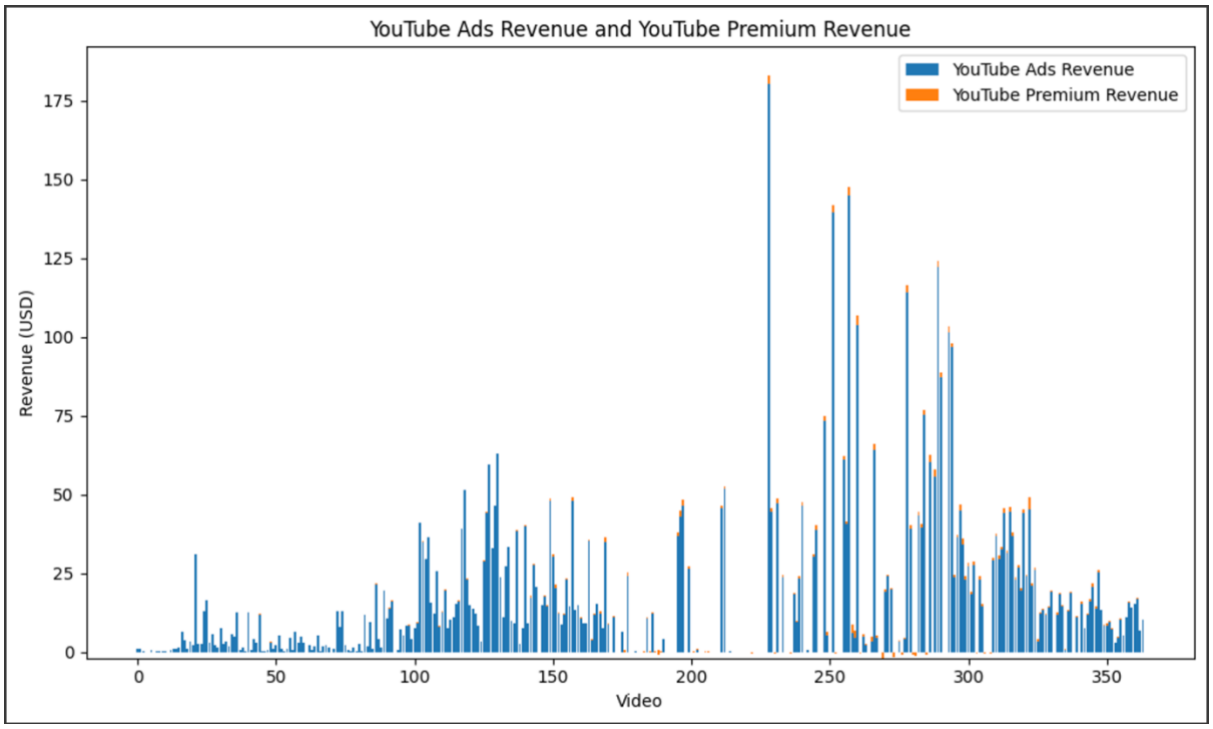

Figure 2: Stacked Bar Chart of YouTube Ads Revenue and premium revenue

Generally, most of the revenue comes from regular users and premium revenue is typically smaller. Some videos show larger portions of premium revenue. Analyzing impressions provides insight on how the visibility of videos can attract more viewers. A mean of 959,528.6 impressions suggests strong visibility for the videos but with such a large range of views and a high standard deviation for the impressions, certain videos are exposed more than other videos. The correlation impressions have with views and revenue are directly related to the CTR. The click through rate is the ratio of viewers clicking on the video vs how

many people see the video thumbnail. So essentially, higher impressions lead to more views if the click through rate doesn't decrease. Knowing this improving the CTR is key to maximize the conversion of impressions to views.

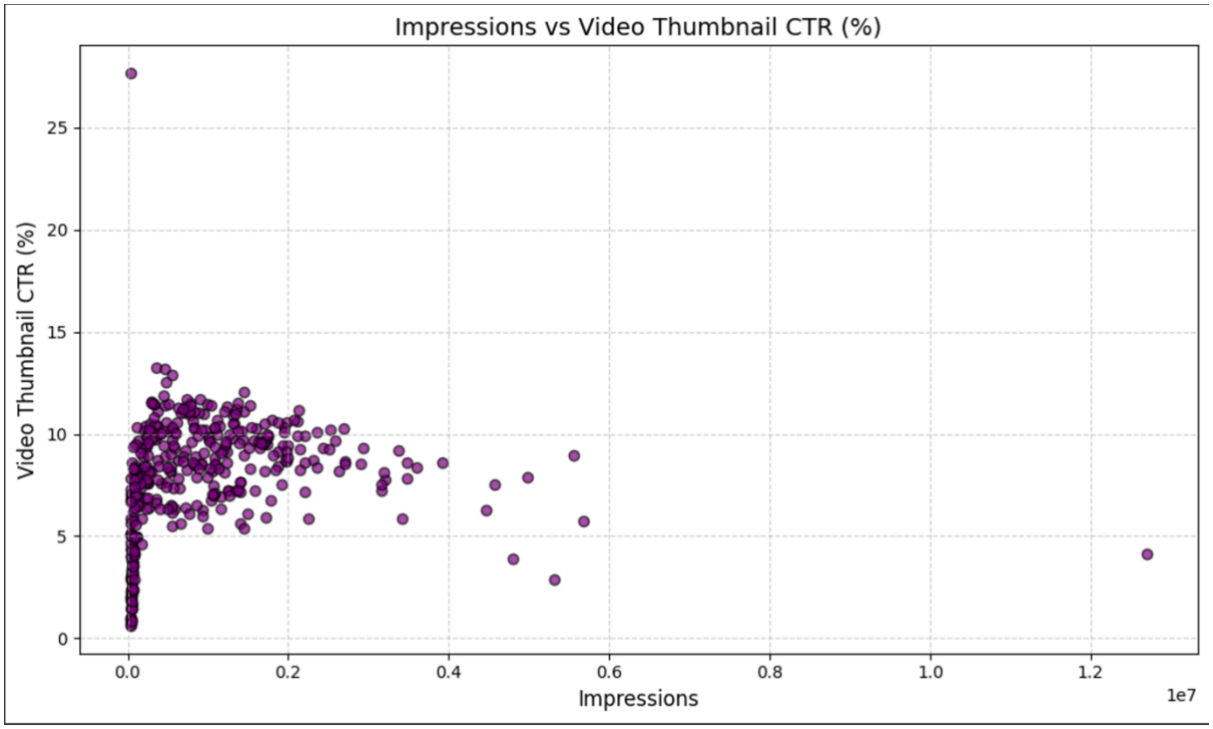

Figure 3: Scatterplot of Impressions vs CTR

We can see most of the points have low impressions and low CTR. The impressions are measured in 1e7 so in this context, these are large impressions. However, we can see an outlier for the impressions and an outlier for the CTR. This lets us know that one video did well with impressions, but its CTR remained relative to the other videos, and another video didn't have many impressions but had a high CTR. These outliers don't represent the entire data set as most CTR stays the same as impressions increase.

Views generate revenue through ads so impressions are indirectly linked to revenue. Videos with high impressions but low revenue per 1000 views could be linked to the number of those views that are from premium users as that could reduce ad revenue.

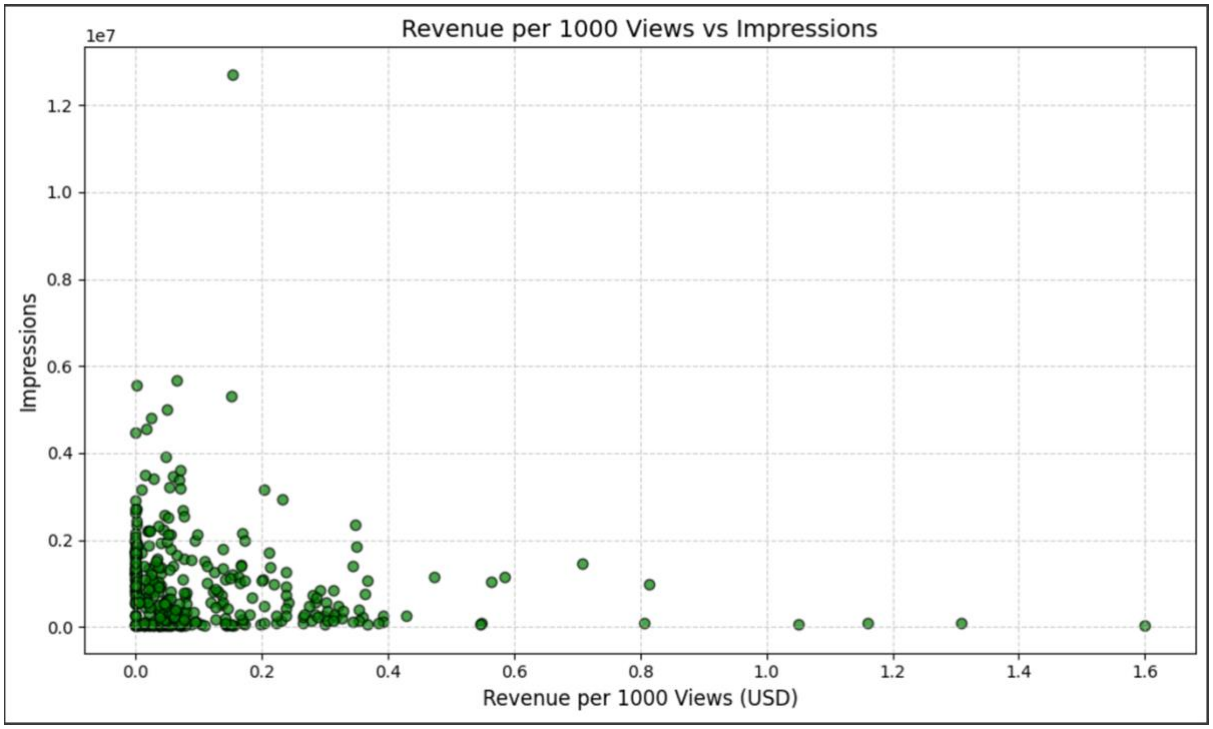

Figure 4: Scatterplot of Revenue per 1000 views vs Impressions

Most of the points are close to the origin meaning most points have low revenue per 1000 views and low impressions. It's important to understand the y-axis of the graph is 1e7 so these impression values are small when looking at the graph but aren't small numbers given the context of the data. The skewness can be a result of this we can see an outlier with low revenue per 1000 views but high impression. This skewness can be fixed by getting the log of the impressions; however, this doesn't change the interpretation that most of the data in the dataset have low revenue per 1000 views and other factors affect impressions more.

The discrepancies between the estimated ad revenue and the actual ad revenue show how users who use YouTube premium play a role in revenue from ads. Additionally, videos with high impressions but low CTR could be labelled as outliers that suggest opportunities to improve engagement. Overall, these metrics give insights in enhancing the channels video performance through engagement and monetization.

Data Exploration

The PCA analysis done using the sklearn decomposition kit helps determine the importance of the principal components and how many are necessary in explaining 96% of the variance. The metrics that are focused on for this exploration are first standardized to ensure that all metrics are on the same scale making it easier to derive meaningful PCA results. To figure out how many components are needed, getting the cumulative sum of the explained variance ratio allows us to see how many components are needed to reach the 96% variance. For these metrics, we need five components to explain 96% of the variance. This suggests that the data can be reduced to 5 dimensions without losing a significant amount of information.

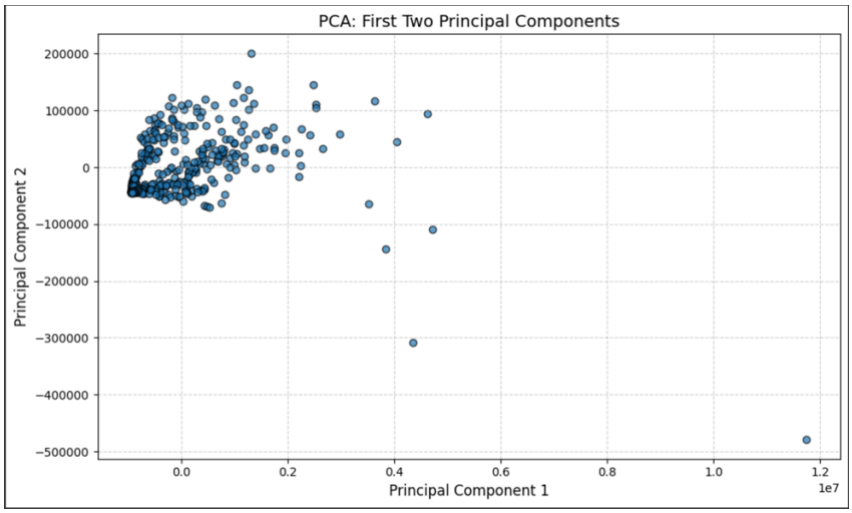

Figure 5: Plot of first two principal components

For this report, our goal is to suggest methods to increase revenue for youtubers using certain video metrics. This means the YouTube ads revenue column in our dataset would be a good numerical value to predict. We can build a decision tree to predict the estimated revenue based on metrics such as watch time, views, impressions and premium revenue. Firstly, our data needs to be split into training data and test data. Using the sklearn tree kit, we can build our decision tree.

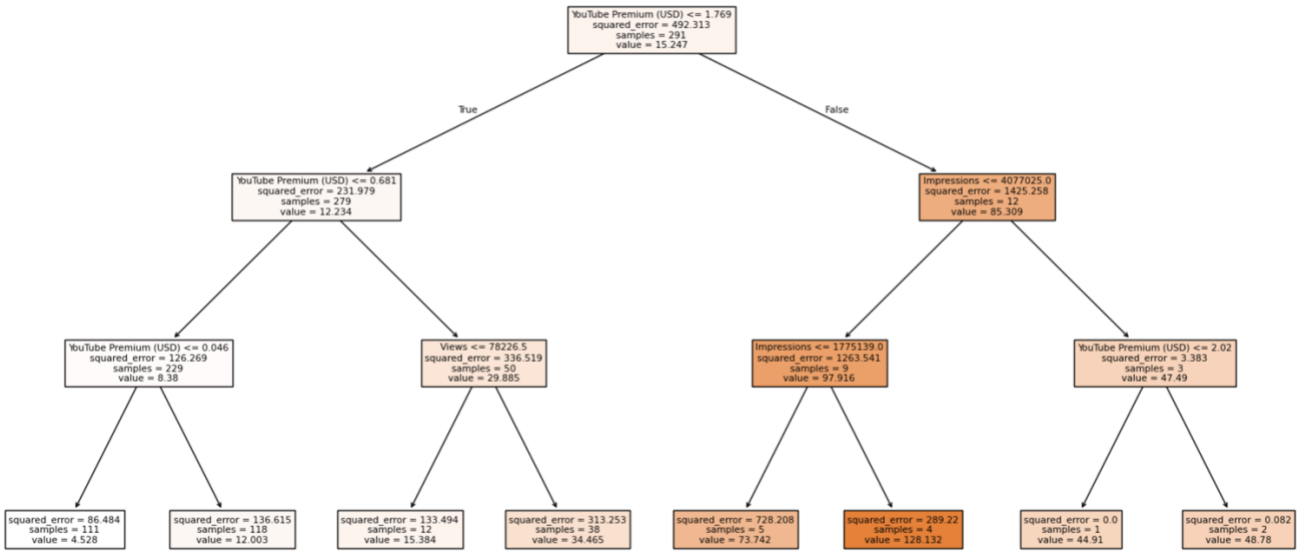

Figure 6: Decision tree for Regression

The root node displays that YouTube premium (USD) is the most valuable predictor in our dataset. The root node splits the data showing most of the data had YouTube premium revenue less than 1.769. The next split first splits the tree again is the YouTube premium revenue is less than 0.681 which again, most of the data falls under. The leaf nodes show the final predictions. Looking at the other splits, on the left side's right side, its split by views being less than 78226.5 which most of the data is more than this. The right side shows high premium revenue and splits first based on impressions less than 4077025. Then splits again by impressions being less than 1775139 and YouTube premium less than 2.02. This tree provides insights that high impressions and high premium revenue predict higher values for ad revenue. Therefore, high impression videos are meant to perform well.

Experimental Method

This report will explore the hypothesis that YouTube videos with higher views, impressions and watch time generate higher YouTube ad revenue with features like premium views influencing this prediction. To test this hypothesis, the random forest model will be

used with the features: Views, Watch time, Impressions, and Premium revenue. The random forest regressor produces an R2 score of 0.9288 meaning the model performs well in capturing the variability in the data. This indicates the features views, watch time, impressions and premium revenue have strong correlations with ad revenue. Although the R2 score is high, correlation doesn't always equal causation, so we need to look at other metrics to conclude. The mean absolute error shows absolute difference between the predicted and actual values while the mean squared error shows the squared difference. This model produces a MAE of 2.02 and a MSE of 51.23. The low MAE suggests the model's predicted values are more accurate on average. The MSE suggests the same, the model is performing well overall even if there are some errors. For further testing, let's look at tuning the hyperparameters using a grid search. The best parameters for the random forest included 100 estimators, max depth of 10, minimum of 2 sample per leaf, minimum of 2 samples per split and an r2 score of 0.922.

A gradient boost regressor shows even smaller mean squared error and mean absolute error with a high r2 score. Performing the grid search on the gradient boost regressor suggests the gradient boost regressor is a better model to use. The best parameters included a learning rate of 0.1 and 200 estimators. This indicates that the model only makes small changes during training and uses 200 decision trees. This gradually improves the predictions as this model focuses on creating deeper trees to assess the relationships more effectively.

The MLP regressor performed on this dataset struggles to assess the relationships effectively. The initial r2 score was a negative value, indicating the hyperparameters failed to model the data effectively. The grid search makes it a better fit as the r2 score is now positive and the best parameters chosen include the tangent activation function and adaptive learning rate. However, it only explains less than 1% of variance in ad revenue so it is not satisfactory for regression. To improve these results, we can standardize the training and testing data to

ensure all the features contribute equally. Earlier we standardized our data before splitting it into training and test data. After the standardization, the r2 score for the MLP regressor using the grid search increased to around 13% of variance in ad revenue indicating more consistency between training and testing data. So, the relationships between the features and ad revenue are more accurate.

Based on these models, we can conclude that higher views, impressions and watch time generate higher ad revenue. There are other factors that affect the ad revenue, but these metrics are the most important out of the metrics in the dataset. Let's dive further into this hypothesis by asking a new hypothesis. The new hypothesis is views, watch time, and impressions can predict the day of the week the YouTube video was published. More specifically, certain days are associated with having higher ad revenues. Exploring this hypothesis allows us to determine if the day the video is uploaded affects performance and revenue. The classification results from the random forest, gradient booster and MLP classifiers indicate that all the models have low accuracy scores most likely because the features selected don't vary based on the day. To improve these results, revenue per 1000 views and playback-based CPM were added to the feature list. These features improved the accuracy score to around 0.26 which is an improvement but still relatively low.

Results & Analysis

After standardization, feature engineering and cross validation, the gradient booster regressor was the best regression model to use while the MLP Classifier was the best classification model to use. The gradient booster regressor producing a r2 score of 0.928 after the grid search means the model explains 92.8% of the variance in ad revenue. The mean squared error of 74.63 and mean absolute error of 3.47 confirm the reliability of the model. The MLP regressor struggles to explain the variance in ad revenue after cross validation

however this could be because tree-based models handle non-linear data better. For classification, the MLP classifier produced the best accuracy score of 0.26 before cross validation and 0.19 after cross validation. All the models struggled in making predictions even when adding new features and performing cross validations. The confusion matrix tells us Type I and II errors were common, so the metrics do not vary based on the day of the week.

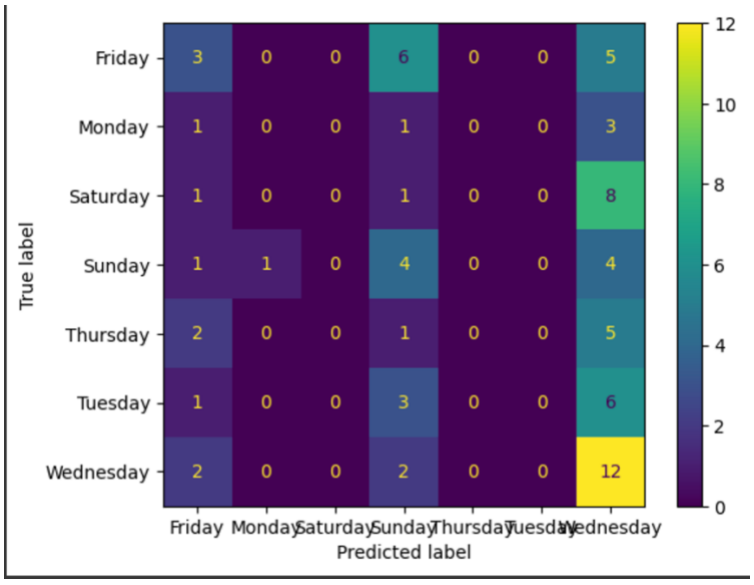

Figure 7: Confusion Matrix

Looking at the classification report, Friday had 27% of predictions correct, 21% of actual Fridays correctly predicted, and an f1 score of 0.24. Sunday had 22% of predictions correct, 40% of actual Sundays correctly predicted, and an f1 score of 0.29. Wednesday had 28% of predictions correct, 75% of actual Wednesdays correctly predicted, and an f1 score of 0.41. All other days had 0 for their scores meaning the classifier failed to predict these days. Wednesday had the best performance, likely due to the number of samples being the largest.

The regression models suggest that YouTube ad revenue generation can be predicted by views, watch time and impressions. These features have a positive relationship with ad revenue. YouTube creators can use these results to understand which strategy they should implement to optimize revenue. Focusing on metrics like CTR can maximize their ad

revenue. On the other hand, the classification models suggest these features aren't enough to classify the day of the week the video was uploaded on. Therefore, there may be patterns by day of week but the model, even with cross validation, still performs poorly. YouTube creators based on these findings can prioritize the quality of their content to increase viewership, impressions and watch time rather than upload timing. These results make no suggestion that the day of upload can increase these metrics. The regression model tells us that these metrics are related to higher ad revenues, therefore we derive weak relationships between the metrics and day of upload.

References

Alex. “YouTube CPM: All You Need to Know in 2024 (Rates by Country Based on 1.5M Views + Increasing RPM).” *Isthischannelmonetized.com*, 6 June 2024, isthischannelmonetized.com/data/youtube-cpm/. Accessed 2 Dec. 2024.

DrStephPowers. “BIA/Ch10_ANN.ipynb at Main · DrStephPowers/BIA.” *GitHub*, 2024, github.com/DrStephPowers/BIA/blob/main/Ch10_ANN.ipynb. Accessed 2 Dec. 2024.

“Implementing PCA in Python with Scikit-Learn.” *GeeksforGeeks*, GeeksforGeeks, 16 Feb. 2021, www.geeksforgeeks.org/implementing-pca-in-python-with-scikit-learn/. Accessed 2 Dec. 2024.

Vaca, Julian R. “8 Steps to Improve Your Video Click-through Rate and Visibility.” *Switcherstudio.com*, Switcher Inc., 16 Apr. 2024, www.switcherstudio.com/blog/video-click-through-rate. Accessed 1 Dec. 2024.

Notebook:

https://colab.research.google.com/drive/1QYalgjnimiRrYweLoTgFtZjeY_uDcQ90#scrollTo=x4UDgZ_tQVcU

Dataset:

https://www.kaggle.com/datasets/positivealexey/youtube-channel-performance-analytics