Youtube Video Performance Project Report

Jelica Bornath

Problem Definition

In 2007, Youtube introduced the Youtube Partner Program as a way to incentivize content creators to continue uploading to the platform through revenue sharing earned on videos via advertisements. There is a business interest in supporting content creators to grow and sustain their audiences, and thus increasing view counts and the associated AdSense collected from their videos. Using data from the daily trending Youtube videos uploaded in 2014, including video performance and channel subscription data, can a creator maximize their video's view count through strategic upload timing?

Success in this exploration will be to establish if a video's performance can be predicted through historical video performance data on trending videos from 2014. This exploration will act as a starting point to identify the feasibility of establishing the optimal upload timing. Key stakeholders for this initiative will be the Youtube Creator Support team as the representatives of the Youtube Partner Program. The team can expect a report outlining the results after modeling, and a presentation outlining the key outcomes for this audience.

Data Wrangling & Exploratory Data Analysis

Through the data cleaning process, the upload time was parsed into days, hours and weekdays. There were 83,769 videos captured in the data for 2014. After removing significant outliers in the data, 78,588 trending videos remained for further analysis.

There were some early indicators in the analysis that the weekday a video was uploaded could impact the view count achieved. This impact is best captured when looking at the Video Categories individually. Each video is assigned a video category, outlined below. The representation of each video category varies in this data so the top 5 with the highest representation were prioritized.

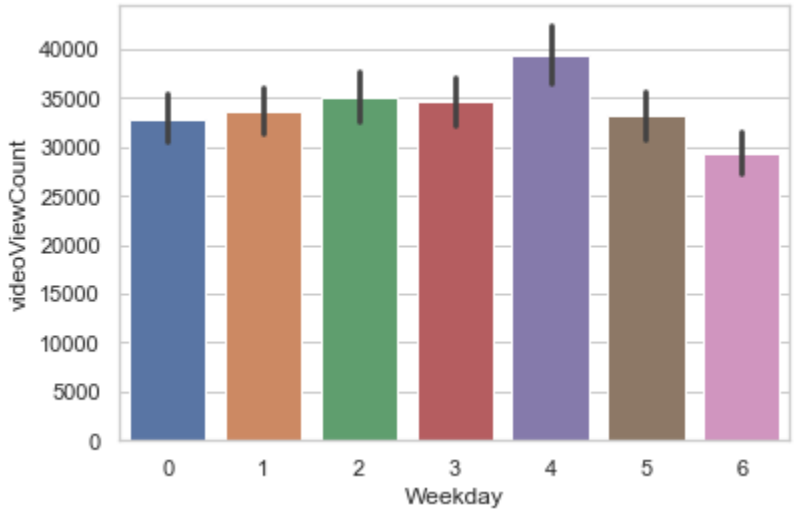
Table 1 - Top 5 Most Frequent Categories in Trending Videos from 2014

Video Category ID	Video Category Name	# of Videos
22	People & Blogs	16795
20	Gaming	12307
10	Music	11052
24	Entertainment	8667
17	Sports	6507

Other	N/A	23260

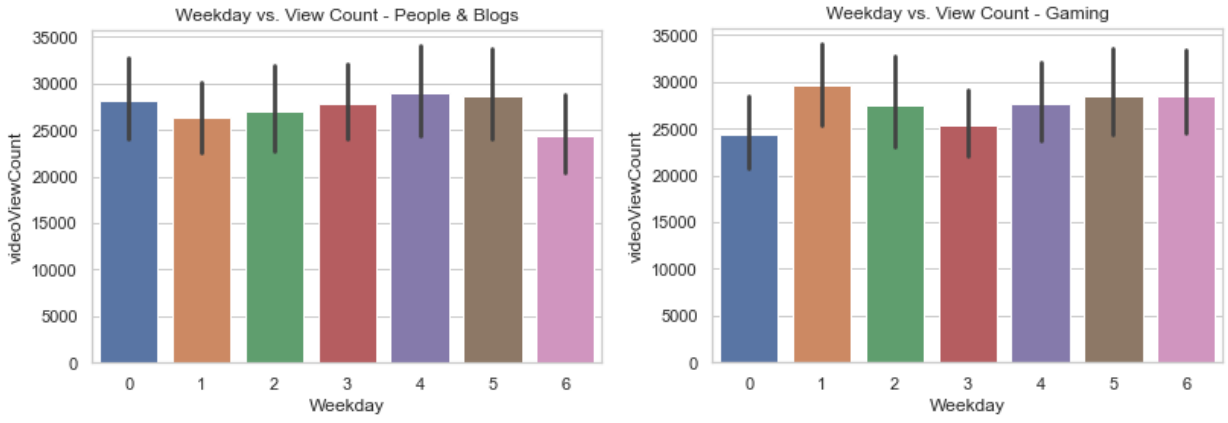
With the weekday definition of 0 = Monday and 6 = Sunday, there was some indication that videos uploaded Friday (4) had higher view counts. This is seen in the figure below.

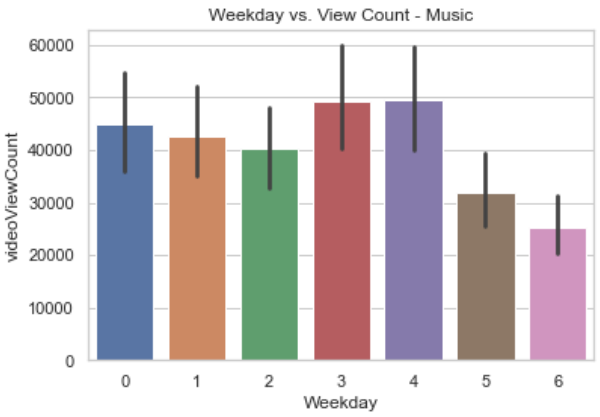
Figure 1 - Relationship between view count and weekday for all 2014 trending videos.

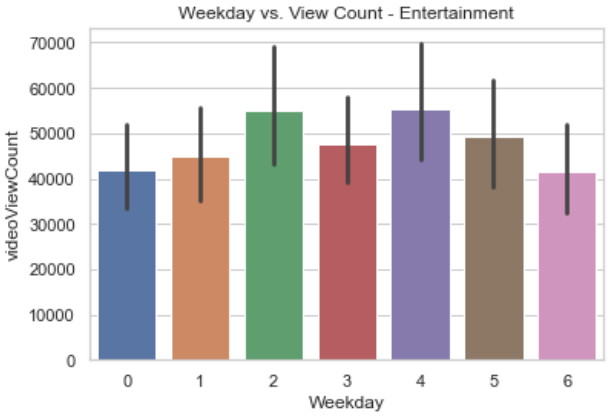


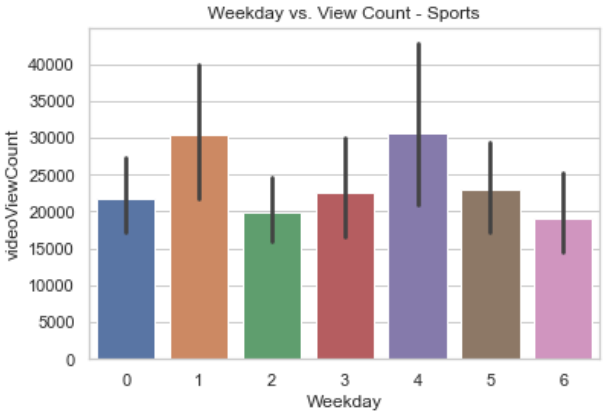
Reviewing the top 5 categories individually gave a better view to these trends within similar video types. While People & Blogs and Gaming videos had relatively consistent performance values across the weekdays, the Music, Entertainment and Sports categories saw larger performance differences across the weekdays. This is shown below.

Figure 2 - Relationship between view count and weekday for top 5 video categories.









These findings do not directly link the upload day to performance as there could be external factors. For example, uploading a Sports video might see better performance if it is posted soon after the footage was taken (i.e. Football Sunday, Saturday Night Hockey etc.), regardless of when during the week that is. Identifying the impact from those external factors would require additional exploration. Recognizing these limitations, the preprocessing and modelling phase was still focused on using the Weekday of upload, channel Subscriber Count, and Video Category in an effort to predict the view performance of a trending video.

Preprocessing & Modelling

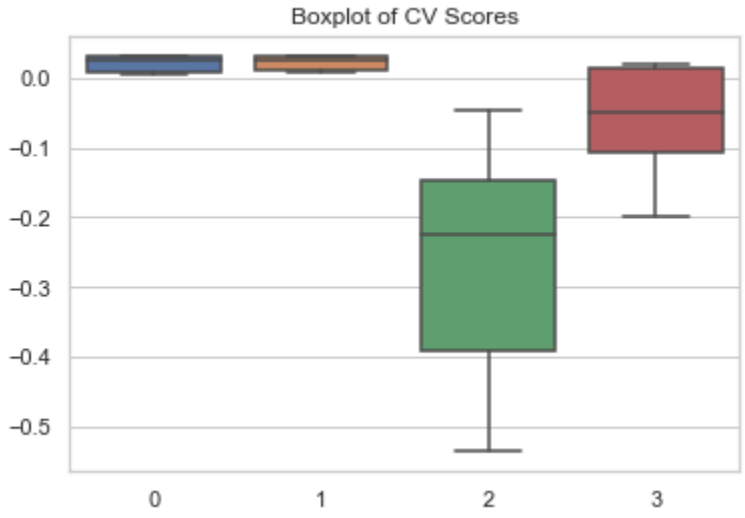
This challenge was approached as a regression problem, with the target variable being predicting the Video View Count. Linear Regression and Random Forest Regression models were used to identify which had the best performance with respect to cross-validation scores. The Weekday attribute was converted into a categorical variable and the data was sectioned into the 5 major categories outlined in the Exploratory Data Analysis Section - People & Blogs, Gaming, Music, Entertainment, and Sports. Each category of interest was tested on the models and then hypertuned to improve performance further.

The first attempt was to remove all features outside of Subscriber Count and Weekday but this garnered poor performance from the models in predicting the potential view count, seen in Figure 3 for the Music category.

Table 2 - Model Identification for Figure 3 & 4

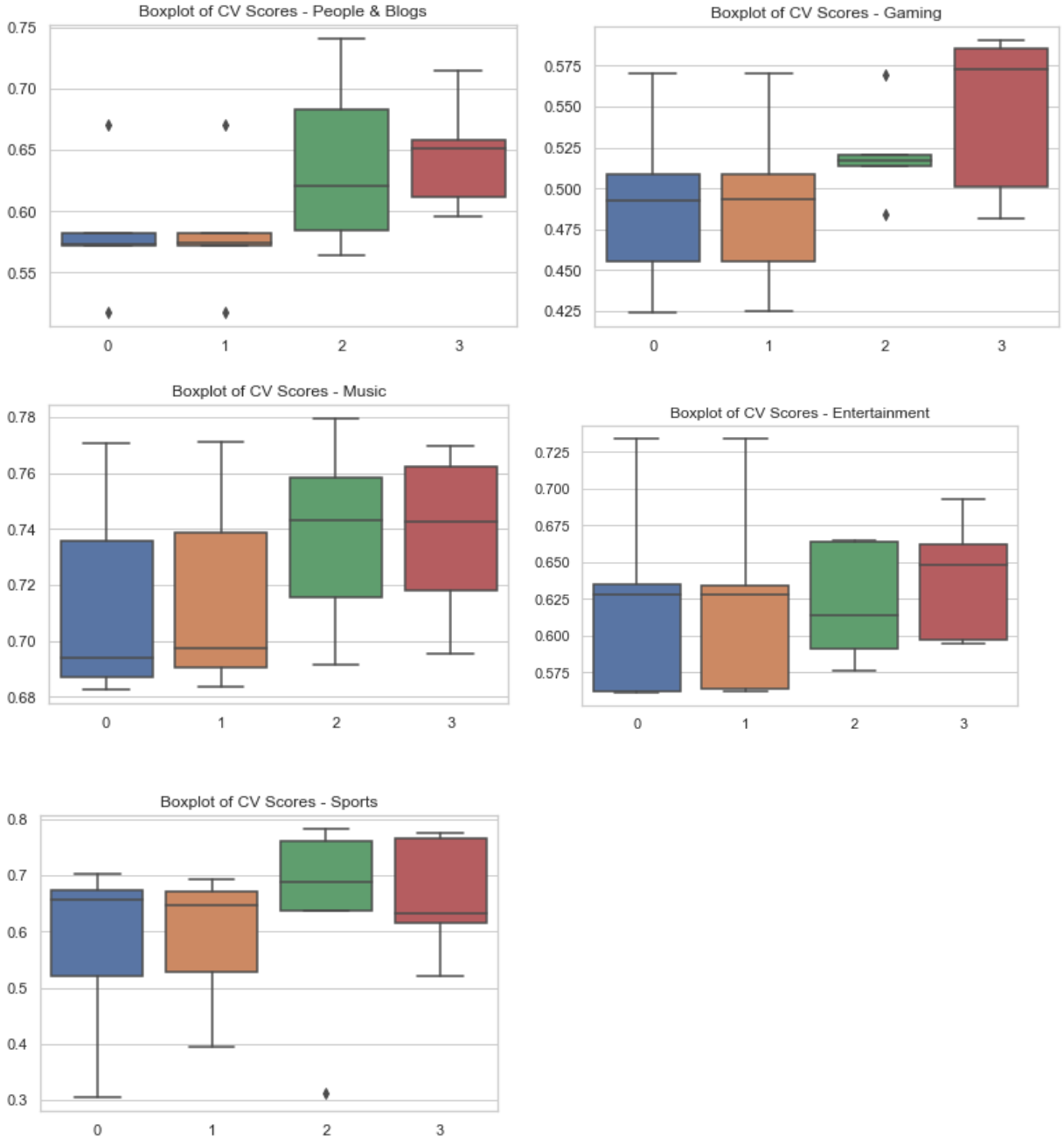
X-Axis Label	Model
0	Linear Regression
1	Regression using K Best
2	Random Forest
3	Hypertuned Random Forest

Figure 3 - CV scores for predicting View Count using Subscriber Count and Weekday features; Music category



Introducing the basic video metrics like number of likes, number of dislikes and video comment count greatly improved the performance of the models in predicting view count but was a recognized departure from the original scope.

Figure 4 - CV Scores for predicting View Count across the models for each of the top 5 categories.



Conclusion & Future Exploration

The model performance when focusing on just Weekday, Subscriber Count and Video Category features indicated that these features were not sufficient to determine a video potential view count. Despite there being some variation in the performance by Weekday in specific

categories, it is unclear if these differences can be attributed to the Weekday that it was uploaded instead of external factors.

When basic video performance features such as likes, dislikes and comment count are included in the data, the models perform better as these features are also indicators of performance. That said, these features are not available for a video prior to its upload and are therefore a scope creep for resolving the original business interest.

An alternative approach to this problem could be treating it as a classification problem, with Weekday being the target variable. By exploring the breadth of the video attributes, can a model predict which day that video was uploaded? If it can, does this lend itself to identifying an upload day best suited to the video? The conclusions reached during the exploration in this report indicated that this might be a challenge, but this alternative approach would open up more features for consideration.

A further exploration could also be to use data specific to a creator's channel instead of focusing on all trending videos. This could personalize the results to the creator's audience and give a sense of the optimal upload timing based on their existing metrics.

The original criteria for success in this project was to establish if a video's performance can be predicted through historical data on trending videos from 2014. The outcome of this effort is that it may be possible to use historical data to determine a video's potential success, but the features required for it can only be collected after video is published (i.e. likes, dislikes, comment count). The recommendation is to perform further exploration using channel specific data, or transitioning to a classification model for Weekday.