## ADS508 Data Science Design Document

**Authors: Andrew Zazueta and Jonathan Yoon**

**Company Name: Viewer Boost**

**Company Industry: Social Media Advertising/Marketing**

**Company Size: 30**

### **Abstract:**

Our company, Viewer Boost, is a startup which helps Youtubers gain insights on how they can improve their video performance. This increase in performance is measured by how many new subscribers are gained/lost, how many views and comments the video achieves, and how many likes/dislikes the video receives.

### **Problem Statement:**

TV and radio broadcasting, which were at one point the most influential media platforms, limited the work to large broadcasting corporations and radio stations with full-time professionals having expert-level experience. However, as the Internet and smartphone technologies evolved, YouTube has become one of the most influential media platforms where non-professionals can still broadcast their ideas in a video format without expert-level skills. It is easier to view this media and viewers have variety of what to watch based on their interests. Viewers may like or dislike the video and can also subscribe to channels catered to these preferences. YouTube, owned by Google (2022), generates their profits from paid advertisements (called AdSense) that viewers watch. A portion of this AdSense then gets distributed to YouTube’s content creators based off the number of views received. As a creator gains views, more ads are watched, thus increasing revenue. This viewer-profit system has become integral to being successful on the platform.

Knowing if a video idea will achieve a larger number of views prior to production is highly profitable. Our purpose for this project is to generate a model that predicts the success of a video based on current performance statistics. With these predictions, creators would then be able to decide what content to cover and how to reach their audiences.

### **Goals:**

If you work with us, we can provide insights you can leverage to:

1. Increase video revenue
2. Help you gain more subscribers
3. Improve your like to dislike ratio

Success is measured by building a model (from current videos and performance statistics on your channel) which can accurately predict the revenue generated. This way, new video ideas can be ran in the model before time, effort, and money are used in video production to see if the idea would possibly be successful. Overall, our business will show which types of videos lead to the most growth.

### **Non-Goals:**

Our company does not guarantee that our predictions will grow your platform, but rather will show you with historical data how well your videos did with current knowledge. That way, you can see which metrics do best in predicting if a video will do well or not. Our methods cannot promise to predict the future, but rather give you the best information to help you in your creative needs.

#### **Data Sources:**

Our data comes from Kaggle and was posted by Ken Jee, a popular data science YouTuber. With his data, Ken (2022) was interested in seeing “what types of videos titles and thumbnails drive the most traffic”, “what types of videos have led to the most growth”, and more. The data is given in 4 CSV files, and each have a common field which can be linked (Unique Video ID).

CSV file sizes and descriptions:

1. Aggregated Metrics by Video (36.86 kB)
   1. 19 columns and 225 rows
   2. Description: “This has all the topline metrics from my channel from its start (around 2015 to Jan 22, 2022)” (2022).
2. Aggregated Metrics by Country and Subscriber Status (9.77 MB)
   1. 15 columns and 55,293 rows
   2. Description: “This has the similar data as aggregated metrics by video, but it includes dimensions for which country people are viewing from and if the viewers are subscribed to the channel or not” (2022).
3. All Comment Final (2.63 MB)
   1. 7 columns and 10,240 rows
   2. Description: “This is all of my comment data gathered from the YouTube API. I have anonymized the users so don't worry about your name showing up” (2022).
4. Video Performance Over Time (20.17 MB)
   1. 14 columns and 111,858 rows
   2. Description: “This has the daily data from each of my videos” (2022).

Our data will be stored in an S3 bucket for easy retrieval in AWS SageMaker. The link to data source is <https://www.kaggle.com/kenjee/ken-jee-youtube-data>.

### **Implementation:**

#### **Data Exploration:**

The data downloaded from Kaggle is stored in Amazon S3 bucket (‘‘ads508projectbucket’’), which was manually uploaded to S3. For exploration, we are planning to use boto3, pandas, matplotlib, counter from collections, and other libraries as the project goes on.

All four datasets were explored to determine which variables were considered crucial factors for viewer count, like count, etc. so we can produce the most accurate model possible. Data types for each dataset were analyzed, and based on these data type, appropriate visualizations were made to help understand data trends.

The analysis and exploration were performed using pandas\_profiling, and the results are available in the GitHub repository:

* Data Exploration: <https://github.com/hjyoon16/ADS508_Group7/blob/main/Data_Exploration.ipynb>
* Exploration results of each data set
  + User Countries and Subscribers:

With 15 features and 55292 observations, it is noted that only 223 distinct videos were analyzed in the dataset. Some interesting points can be found in correlations between the variables. It includes the highly positive correlation between the number of views and likes and the average watch time and the average view percentage. It is also worthwhile to note that length of the videos have a highly negative correlation with average view percentage.

<https://github.com/hjyoon16/ADS508_Group7/blob/main/Country_and_Subscriber_output.html>

* + Comments on Videos:

7 features with 10240 observations were analyzed. No surprisingly, high cardinalities in comments, comment\_id, dates, and user\_id were noted due to the uniqueness of each comment. Also, it is noted that the number of replies are highly correlated with the number of likes.

<https://github.com/hjyoon16/ADS508_Group7/blob/main/Comments_output.html>

* + Video contents:

After analyzing 19 features with 225 observations, high cardinalities in video, video title, video publish time, and average view duration were noted. Also, it was noted that there are high correlations between the number of views and several other features, including the number of shares, the number of likes, subscribers gained, and the estimated revenues (USD). Other highly correlated features include the number of subscriber and the number of subscribers gained.

<https://github.com/hjyoon16/ADS508_Group7/blob/main/Video_output.html>

* + Performance of Videos:

111,857 rows with 14 features were analyzed. High cardinalities were noted in several features, but the interesting point is in the correlations of the features. Highly positive correlations between the number of views and subscriptions added, between the number of likes added and the subscriptions added, and between the number of views and the number of likes. Negative correlation between the average view percentage and the video length is also noticeable.

<https://github.com/hjyoon16/ADS508_Group7/blob/main/Performance_output.html>

Many of the features have distinct values so they have high cardinalities. The focus is to figure out which features need to be added to the models.

The key features being analyzed included:

* Wording of the video titles
* Top countries that viewed the videos
* Watch time
* Number of views
* Whether the viewers are subscribed members
* Revenue of Videos
* Ratio of Like to Dislike
* Comments
* Cumulative views over time
* Cumulative subscribers over time
* Video created date

It is noted that the number of views and subscribers grew exponentially beginning early 2020 so it is expected to normalize data to minimize bias. Also, a tendency of having greater number of likes than dislikes is expected with more views, so a class imbalance bias is expected.

GitHub Repository: <https://github.com/hjyoon16/ADS508_Group7/blob/main/Assignment3.1.ipynb>

#### **Data Preparation:**

For our data set, we utilized multiple data cleaning techniques. We converted categorical data into binary features, created new features from existing features, altered existing features into new formats better for modeling, and normalized features which were continuous to avoid the issue of having different units altering our model performance. For example, average time watched was measured in seconds, but total video watch time was in hours. Normalizing features like these avoided the issue of having one feature having more influence over another due to value magnitude.

In the process of cleaning our data, we also removed an abundance of unnecessary features. From *Comments\_df*, we only kept the columns ‘VidId’ (for merging data down the road), ‘Reply\_Count’, and ‘Like\_Count.’ From this data, we grouped based on Video ID and found the max reply count and like count per video.

From *Country\_and\_Subscriber\_df*, we kept 'External Video ID', 'Country Code', and 'Views.’ Then, we calculated the percentage of views from the United States, India, and the rest of the world per video. This was done due to the United States and India having the largest viewer base by far compared to other countries. We hope to see if the percentages coming from these countries have an influence on the revenue gained per video.

From *Video\_df*, we needed to remove the first row due to it giving column values and not row values. Then, we changed the video publish time to day of the week. We also created a boolean feature which states if 'Data Science' appears in the title of the video or not. We converted the average view duration to seconds rather than minute second format. Lastly, we removed the total likes and dislikes column and replaced it with a 'like\_dislike\_ratio' column. From this data frame, we want to attempt to predict the 'Your estimated revenue (USD)' feature. This feature tells us the estimated revenue of the video and will inform the video creator if they are being paid less or more than predicted by our model from historical data. To remove any features that biased our results, we removed 'RPM' and 'CPM' for giving away monetary information.

From *Performance\_df*, the only useful information was the video length, so the two columns that were kept were 'External Video ID', and 'Video Length.’

The last step in our data preparation stage was to inner join all the data frames based off the video ID and normalize our data as previously mentioned. This left us with a dataset that had 218 rows (1 row per video made), 23 predictor columns, and 1 response column. This produced a unique issue we did not realize we would encounter. SageMaker ML models require data to have at least 500 rows. Since we only had 218 rows, we had to oversample our training set to get to the 500 threshold. This is not good data science due to overfitting becoming a major issue but is unavoidable at this stage of the project. In the future, we would need to ensure the YouTuber has created at least 500+ videos to get the necessary data we need to build our models as accurately as possible. Before we oversampled the training set, we split our data to 80% training, 10% validation, and 10% testing. Since our data requires regression for predictions, no balancing was needed.

GitHub Repository Link to Data Cleaning:

<https://github.com/hjyoon16/ADS508_Group7/blob/main/Assignment4.1.ipynb>

**Model Training:**

The transformed dataset contains 100% numeric values, and the best fit model was thus produced using regression algorithm via SageMaker Autopilot. It passed all features in the transformed dataset, and the “Estimated revenue (USD)” was set as the target feature for this pipeline generation.

Autopliot generated eight pipelines using these features (from dpp0 to dpp7), and the best model was determined by the lowest mean squared error (MSE).

Dpp0-xgboost, Dpp1-xgboost, Dpp2-xgboost, Dpp5-xgboost, and Dpp7-xgboost employed xgboost algorithm to train and tune the data with slight differences in data transformation methods between each other. Dpp3-linear-learner and Dpp4-linear-learner employed linear-learner algorithm with differences in data transformation between them, and Dpp6-mlp employed MLP (multilayer perceptron) algorithm. These transformers and its training pipeline are built using open-sourced sagemaker-scikit-learn-container and sagemaker-scikit-learn-extension.

Dpp1-xgboost model gave the lowest MSE of 2.535, and this presumes that dpp1-xgboost model would be the ideal model with the best performance. With consideration of potential overfitting issue, all 8 models will be considered and validated with validation and testing to determine the final best model.

Link to Data Training Notebook: <https://github.com/hjyoon16/ADS508_Group7/blob/main/SageMakerAutopilotCandidateDefinitionNotebook.ipynb>

### **Measuring Impact:**

Before, one of our goals was to be able to spot what types of videos perform the best with the deciding metrics being views, likes, and new subscribers gained. This task, however, does not necessarily need data science to complete; a lot of these findings can be done through data analysis alone. This was discovered during our exploratory data analysis phase, so we decided to make a small pivot with what we want to accomplish.

Instead, our new goal is to try and predict revenue gained in a video using features provided and new features created. For example, we discovered that a good metric to create was the “Like to Dislike ratio” rather than using likes and dislikes counts. This way, newer, more popular videos will not skew the data due to having a larger number of likes and dislikes compared to older videos created. Another feature to be added is a Boolean that specifies if “Data Science” appeared in the title of the video or not. The reasoning for this addition is because the top 10 videos on the YouTubers channel all included “Data Science” in the title, so it might be a good predictor if a video will perform better or not. Current subscriber count is another good predator that will be added by finding the cumulative sum of net subscriber growth (subscribers gained – subscribers lost) over time. Lastly, the date the video was created can cause issues due to the channel’s exponential growth since 2020. We will remove this feature and replace it with a day of the week tracker. For example, if a video was created March 21, 2022, the new column would include “Monday” instead of the date (incase the day of the week the video was uploaded can influence views). This is because performing time series analysis will be a complicated task on top of the other methods we would like to try.

### **Security Checklist, Privacy and Other Risks:**

There are no personal health information (PHI) and personal identification information (PII) included in the dataset. Only user comments are tracked, and the usernames are replaced by a random IDs. The S3 bucket ‘ads508projectbucket’ will be used to read the dataset and write the results to.

Some bias that was discovered was the distribution of viewers by country. The United States and India make up the largest portion of viewers, despite viewers watching from over 230 countries. This can have an impact on our model making predictions due to data mostly coming from the USA or India, which is not representative of all possible viewers. Another area of bias is video creation date. After 2020, the number of views and subscribers the YouTuber gained was exponential, while videos they created from 2017 to the beginning of 2020 were less successful. Therefore, any new videos created might have a bias that they will instantly perform well due to the date created rather than the contents of the title.

No other ethical concerns are related to this data. It contains no sensitive data about any person’s race, gender, socioeconomic class, etc. The only information that could be seen as personal is the revenue generated by each video, but since the results are only shared with the Video creator, this is not an issue.

### **Future Enhancements:**

Several improvements can be made to enhance the model pipeline. First, we would want to gather a greater quantity of data to build our models. Only 218 rows were obtained after inner joining the data sets by video IDs. Having so few rows resulted in our data not being usable for SageMaker Autopilot. This forced us to duplicate our data to meet the 500-record threshold, which hurts our model from overfitting. From a business standpoint, we would need to advertise that we only help YouTubers who have made 500 videos or more so that our models can perform the way they should.

Another potential improvement would be deploying more models at different endpoints so there are more options to study, test, and select. This would allow us to choose a more accurate model since the project would have gone through a greater number of tests. Realistically, models are developed and trained multiple times throughout multiple weeks in the project lifespan, which was not doable in our scenario.

With a $100 budget, SageMaker lasted only 5 days due to our funds depleting. Also, we could only use SageMaker Autopilot to generate models because other applications such as SageMaker Jumpstart were not able to use our dataset. These issues would have been avoided by 1) having a larger budget and 2) having a higher level of SageMaker knowledge before selecting a data set.

Lastly, if given more time, we would have liked to have spent more effort transforming our data set. For example, instead of only having three features dedicated to viewer country location data, we would use 6 binary features (one for each continent the countries resided in) to give more detail on where viewers are located. This enhancement would have been timely due to having to individually go through over 220 countries and assign them to the continent they belong to.

Another change to the features we would have liked to make would be to specifically format them for certain models. Linear models, like the one we used, should not have features which are highly correlated to one another. We did not have the opportunity to examine the correlation between features via a correlation matrix, which could have resulted in our model training on redundant features. This caused our model to be more costly and hindered its performance.

It is important to be able to adjust features before each model run after examining how well they performed from the previous model. Feature transformation is dynamic and is constantly being done throughout a traditional work project, and only being able to have one week to dedicate to this phase results in less-than-optimal results. In conclusion, a model can only perform as well as the data it receives, and not being able to perfect the feature selection and transformation is something we would have liked to have done.

**References:**

Google. (2022). *Google AdSense*. https://www.google.com/adsense/start/

Jee, K. (2022, January). *Ken Jee YouTube Data.* Kaggle.

https://www.kaggle.com/datasets/kenjee/ken-jee-youtube-data