## ADS508 Data Science Design Document Template

**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, 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 number views a video will gain. 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 lead 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).”
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.”
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!”
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.”

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:**

Details on how the solution will be implemented. This should be the longest section of the document.

#### **Data Exploration:**

Detail what you will look for in the data during the exploration phase. Consider data quality, potential data bias, key fields, data types, etc.

* Where will you be storing your data, how will you get ingest it there?
* What tools will you use to ingest and explore the data?
* Detail what you will look for in the data during the exploration phase. Consider data quality, potential data bias, key fields, data types, etc.
* Implement your data ingestion (if done via code) and data exploration using SageMaker studio notebook. Store your code in a github repository and provide a link to the code used for ingestion and exploration.

#### **Data Preparation:**

How will you transform the raw data so that it is ready for training?

* Data Scrubbing: What data cleansing techniques will you apply?
* Feature Selection: Which fields from your data will you use/exclude?
* Feature Creation: Which fields will be combined, or bucketed?
* Feature Transformation: What other transformations (such as one hot encoding, etc) will you apply to your data?
* How will you balance your data set?
* How will you split your dataset?

**Model Training:**

How will you train your model, what tools will you use?

* Will you be using SageMaker Jumpstart, built-in-algorithms, bring-your-own-script or bring-your-own container?
* Which algorithm will you be using?
* Which parameters will you be passing?
* Which instance size/count will you be using?
* How will you evaluate your model?

**Model Tuning**

* Which hyper-parameter optimization strategy will you use, why?
* Which hyper-parameters and ranges will you use?
  + How did you determine these were the hyper-parameters to test?
  + What are the types for each hyper-parameter you have chosen?
  + What scaling-type will you use for each hyper-parameter?
* How many jobs and max parallel jobs will you use? How did you determine this number?
* What early\_stopping\_type setting will you use? Why?
* Will you be using warm-start? Why or why not?

**Model Evaluation:**

How will you evaluate your model and determine if the results are accurate?

* What is the objective metric?
  + How will you determine when you have achieved the optimal metric for your model?
  + Include a screenshot of the results of your final HyperParameterTuningJobAnalytics.

**Model Inference:**

If the board approves your suggestions, what steps will you take to put your model into production?

* Will you use batch or real-time inference? Why?
* Which deployment pattern will you use (canary, etc)?
* Describe how you will auto scale your endpoint.
* Describe how you will monitor your endpoint.

### **Measuring Impact:**

List at least two specific metrics you expect to change with this project. These should tie back to the goals above.

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

* Will this store or process [PHI](https://www.hhs.gov/answers/hipaa/what-is-phi/index.html) data?
* Will this store or process [PII](https://www.dhs.gov/privacy-training/what-personally-identifiable-information#:~:text=%E2%80%9D%2C%20or%20PII%3A-,DHS%20defines%20PII%20as%20any%20information%20that%20permits%20the%20identity,U.S.%2C%20or%20employee%20or%20contractor) data?
* Will user behavior be tracked and stored?
* Will this store or process credit card data?
* What S3 buckets will this application read or write to?
* What data bias should be considered?
* Are there any ethical concerns with the data or business problem that should be addressed?

### **Future Enhancements:**

Provide at least 3 ways you would improve your model pipeline if you had more time/ additional resources?