Week 11: Model Explanation, Risk Analysis, and Ethical Evaluation

Yujia Cao, Wendi Yuan, Yuhan Zhao

We have completed model building using the best-performing variation and the features derived from our dataset. Returning to our research focus, we aim to evaluate the performance of YouTube videos and identify the factors influencing their success. Additionally, we reflect on the insights our analysis has provided. This week, we explored the model by identifying its most important features and analyzing their impacts on predictions. Furthermore, we investigated potential biases in the model and discussed them. Lastly, we addressed the ethical challenges and risks that may affect the stakeholders involved.

To identify the key drivers in our model, we examined the features included in the final iteration (Appendix 4). Notably, many engineered features, such as *Engagement_metrics* and *trending_day_of_the_week*, were prioritized by the model, whereas the TF-IDF features derived from the *tags* and *descriptions* columns were not. The dataset contains 17 features in total, with the top 10 contributors primarily originating from the original dataset (Appendix 5), including *title, tags,* and *likes*. Among these, the top five predictors are *title*, *publish_time*, *channel_title*, *tags*, and *category_id*, which have the most substantial influence on the XGBoost model's predictions.

The most influential feature is *title*, contributing 43.7% to the overall feature importance score (0.437). This underscores its critical role as a performance indicator. Video titles often serve as the first point of interaction, significantly affecting click-through rates (CTR). Catchy and descriptive titles with relevant keywords enhance search visibility, drawing higher engagement through improved *views* and *likes*.

The second most important feature is *publish_time* (0.193), highlighting the importance of timing in video performance. Videos released during peak activity hours or aligned with audience behavior tend to perform better initially, which can create a compounding effect on overall performance metrics.

channel_title ranks third with a feature importance score of 0.085. This reflects the influence of channel authority and reputation, as established channels with loyal subscribers often see higher engagement due to brand recognition and content expectations. A trusted or popular channel name attracts more clicks, regardless of other factors.

tags hold moderate importance (0.051), assisting YouTube's algorithm in categorizing videos and matching them with relevant searches. Effective tagging improves discoverability, increasing views and engagement by connecting videos with their target audiences.

The fifth most significant feature is *category_id* (0.049), which complements *tags* by influencing video categorization and searchability. Categories such as *entertainment* or *gaming* typically draw larger audiences, whereas niche categories may yield smaller but highly engaged viewer bases. Together, *tags* and *category_id* shape content discoverability, affect recommendations, influence search rankings, and define audience reach.

To further interpret the model's predictions, we used SHAP (SHapley Additive exPlanations), a method rooted in cooperative game theory. SHAP values quantify each feature's contribution to a specific prediction, identifying whether a feature increased or decreased the likelihood of a predicted class. As SHAP is well-suited to tree-based models like XGBoost, we calculated SHAP values for a subset of predictions (Appx 6).

For instance, for the randomly selected sample 49053, the top contributing features are *dislikes*, *likes, engagement_metrics, comment_count,* and *category_id*. Each SHAP value explains how a feature influences the prediction, with positive values indicating increased likelihood of a specific outcome and negative values indicating the opposite (Appx 7). The SHAP graph visualizes these contributions, revealing how features drive the model's predictions. To shift predictions from one class to another, increasing the values of features with the most positive SHAP contributions would be beneficial.

These findings offer actionable insights for YouTube content creators by highlighting the features that most significantly impact video performance and suggesting areas to optimize for improved outcomes.

- ***dislikes***: Consider an increase of 1770361.00 SHAP impact units to potentially achieve a prediction flip.
- ***likes***: Consider an increase of 1485569.25 SHAP impact units to potentially achieve a prediction flip.
- ***Engagement_metrics***: Consider an increase of 1251085.25 SHAP impact units to potentially achieve a prediction flip.
- ***comment_count***: Consider a decrease of 1207504.88 SHAP impact units to potentially achieve a prediction flip.
- ***category_id***: Consider an increase of 304174.72 SHAP impact units to potentially achieve a prediction flip.

We have listed other 4 examples in our appendix with similar analysis. (Appx. 8, 9, 10, 11)

Before we go deeper into our chosen dataset to find if predictors include any types of the protected categories, we want to clarify the definition of protected categories/classes to amplify the analysis. Protected class is defined by federal law/executive order, federal agencies The protected classes

include but are not limited to age, ancestry, color, disability, ethnicity, gender, gender identity or expression, genetic information.

The definitions of these protected categories are not fixed and need to be judged on a case-by-case basis, but after understanding the definitions, our dataset does not contain any protected categories explicitly. Our dataset is sourced from Kaggle, and it is specifically focused on analyzing the trends and relationships between YouTube video trending metrics and viewership. However, it's important to note that this dataset does not include social, economic, or individual factors, as these elements were not part of the original data collection process. Our research is primarily concerned with identifying and understanding the patterns and variables related to the trending status and view counts of YouTube videos, without considering the broader societal or economic context that might influence these trends.

Our dataset contains variables that are specifically designed to provide an objective representation of the information associated with YouTube videos. These variables act as a comprehensive framework to capture the essential attributes and context of each video, serving as a descriptive shortcut for understanding its details without referencing or inferring any protected categories. Take the variable ‘ *tags*’ as an example to further illustrate the main point.

What the first line of tags (Row 1) shows is :*"The Tonight Show" | "Jimmy Fallon" | "Marshmello"* , and These tags are clearly extracted keywords that describe the video in terms of the content context(*“The Tonight Show”*) which identifies the program.And it includes the Host/Participants of the video (*“Jimmy Fallon” and “Marshmello”*)which point to key figures involved. It also includes genres that might be involved in entertainment/music-related terms. (Appx.1)

The objective nature of the ‘*tags*’ focuses on factual details about the video, capturing its genre, themes, and key features without referencing sensitive information like the identity or background of the creators or audience. Similarly, other variables present information like *tags* with the essence of each YouTube video without referring to the protected categories or demographic classifications.

In recommendation models, biases often arise from the relationships between observed variables and underlying protected categories such as gender, age, ethnicity, or geographical location. Our dataset here does not explicitly include protected categories; the strong correlation between engagement metrics (e.g., *views* and *likes*) indicates that the model may inherently favor content that receives higher interactions. This preference can unintentionally introduce biases if such content is predominantly associated with certain groups or categories (Appx. 2).

Content categories with traditionally higher engagement (e.g., *entertainment* or *gaming*) may overshadow niche categories that align with underrepresented communities, potentially marginalizing

content creators from these groups. A strong positive correlation (0.79) suggests that content with higher views also receives more likes. If certain content types or categories (e.g., mainstream entertainment) dominate in views, the algorithm might amplify this trend, further marginalizing less popular but high-quality content. If protected categories are tied to these content types, this can exacerbate inequality (Appx. 3).

Understanding these correlations is critical for mitigating potential biases in recommendation systems. If protected categories are indirectly tied to these engagement metrics (e.g., creators from certain demographics systematically receiving fewer *views* or *likes*), the system may unintentionally perpetuate inequities. Identifying and addressing such correlations requires incorporating fairness metrics and analyzing how engagement metrics differ across various groups.

In our analysis, we observed three prevalent types of bias: *popularity bias*, *recency bias*, and *content-type bias*. These biases can disproportionately affect certain groups or categories within a dataset, underscoring the importance of understanding and addressing their implications to achieve fairness and accuracy in predictions.

Popularity bias occurs when recommendation systems prioritize globally popular items at the expense of personalized or niche suggestions. This arises from algorithms that optimize for metrics favoring mass appeal, such as total engagement or likes. Popularity bias disproportionately amplifies the reach of already popular items, leading to a lack of diversity in recommendations. This dynamic can disadvantage underrepresented categories, perpetuating inequality in visibility for items that appeal to smaller or more niche audiences. For example, in datasets that include protected categories such as gender or ethnicity, this bias could inadvertently reinforce stereotypes by favoring content that aligns with mainstream preferences. Addressing this bias is critical to fostering equitable recommendations.

Recency bias occurs when trending algorithms prioritize newer content to maintain a fresh and dynamic trending page. While this approach ensures up-to-date recommendations, it often undervalues older or timeless content, even if it is experiencing a resurgence in popularity. This can disproportionately impact protected categories if content tied to specific groups or historical contributions is overlooked in favor of the newest additions. Such biases may inadvertently marginalize valuable contributions from certain communities, highlighting the need for mechanisms to assess long-term content relevance.

Content-type bias arises when algorithms favor specific genres, such as entertainment, music, or gaming, over less mainstream categories like educational or niche content. This bias skews visibility toward highly engaging or "trendy" content, often neglecting high-quality content in less popular

categories. For datasets including protected categories, such biases could inadvertently overshadow valuable contributions from underrepresented groups. For instance, educational or cultural content tied to marginalized communities might struggle to gain the same traction as widely popular entertainment categories.

Our group is dedicated to removing biases in our model. Based on our model, we attempted to address popularity bias by using normalized metrics, such as views per subscriber or likes per minute, to level the playing field for smaller creators. However, our current metrics are outdated, and our dataset lacks sufficient content for comprehensive analysis. If we had access to instantaneous data, we would eagerly implement these strategies to reduce bias.

In addition to popularity bias, we also considered recency bias. To mitigate this, we proposed introducing an "evergreen" category in the trending section that surfaces high-quality older videos with sustained or renewed engagement. However, our dataset is small and originates from 2018, limiting our ability to perform robust analyses. With real-time data, we could better refine our approach to address these biases.

To tackle content-type bias, our group explored developing scoring systems that assess quality, relevance, and viewer impact across all content types, rather than favoring specific genres based on historical trends. However, we encountered challenges in refining our scoring system and plan to focus on improving it in the upcoming weeks. These biases collectively skew recommendations, reducing diversity and potentially excluding certain voices or categories from visibility. The root of these biases often lies in incomplete or imbalanced data collection, as datasets may fail to represent all content types, demographic groups, or temporal patterns adequately. This leads the model to generalize based on limited perspectives.

These strategies not only reduce systemic biases but also improve overall prediction accuracy by providing a more comprehensive view of the dataset. Ensuring data completeness and diversity is essential for empowering the model to make balanced recommendations that reflect the full breadth of user preferences and content quality. Ultimately, this fosters inclusivity, fairness, and a more equitable platform for all creators and users.

Due to the speciality of our dataset, research orientation, and the model, we think there will be conflicts between viewers, citizens, and the entertainment business. Viewers, as the most directly exposed stakeholders to the YouTube videos, seek personalized, diverse content, but algorithms that prioritize popular or commercial content may limit their exposure to new or varied ideas. Citizens care about the social impact and cultural diversity of content, but if the model promotes biased or manipulated content, it

could skew public discourse. The entertainment business prioritizes engagement and profits, but this focus can suppress smaller creators and favor commercial content, potentially undermining diverse or independent voices.

For viewers, which are also the end users of the YouTube platform, they are exposed to the contextual and personalized risks. If the model relies on trends in a particular variable, such as the data-type variables we overuse in our analysis, it creates an echo chamber that limits the diversity of content presented to users. Users may only see content that matches their existing preferences, thus hindering exposure to new ideas or perspectives. Model advancement requires large amounts of data as a basis for analysis, and such a process may involve collecting sensitive user information (e.g., browsing patterns, interactions with content). This presents potential privacy risks, especially if the data is not anonymized or properly protected. If the model generates popular content that is consistently pushed to the forefront, already popular videos will attract more viewers to watch them, e.g., a particular type of channel popularity attracts a large number of viewers getting higher exposure, and thus customers may become fatigued by overexposure to similar types of content. Over time, this may reduce user satisfaction.

If the model is used in a way that allows certain entities (e.g., corporations, political groups) to influence trending topics or rankings, it could result in manipulation of public opinion or priorities. Citizens may be exposed to biased or skewed content without realizing it or misinterpret some contents that are not correct or related to sensitive agendas. The model's ranking of content based on different variables could unintentionally suppress certain voices or topics if the popularity of the video is the ultimate goal. For example, underrepresented communities may be overshadowed by more dominant voices in the algorithm's rankings, and videos that represent majority voices prevent fair representation in public conversations.

If the entertainment business applies this model, one key risk is over-reliance on trending content that might prioritize sensational or commercially viable content at the expense of diversity and creativity. This could lead to audience fatigue, as viewers may become tired of seeing the same types of content repeatedly. In addition, deviation in the actual application of the model may result in certain content being favored over others, potentially alienating certain audience segments that do not fit the model's preferences. If the entertainment industry promotes specific content or produces spin-offs based on the popularity of videos and what the model projects, this could suppress less commercially viable content or limit the diversity of content, which in turn could reduce long-term viewer engagement. For entertainment business, they bear part of the responsibility for promoting trends, and if they simply use the model without doing user preference surveys or constantly updating the variables in the model, such as adding

the types of channels that videos are made available on and changes in the time period in which users are watching YouTube videos, the corresponding revenues of the entertainment platforms may also decline, and their interests are put at risk.

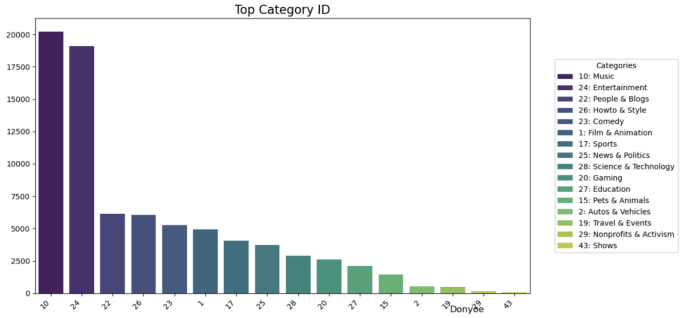
Appendix

```
       trending_date                                              title  \
23604    2018-03-14                      Marshmello & Anne-Marie: Friends
25630    2018-03-24  Kirby Star Allies' Surprising HD Rumble Secret...
68698    2018-04-20  Stephen A.: Kevin Hart 'got his feelings hurt'...
39559    2017-11-17                                How to be an Aquarius
62877    2018-03-16  Charlie Puth - Done For Me (feat. Kehlani) [Of...

                            channel_title  category_id  \
23604  The Tonight Show Starring Jimmy Fallon           23
25630                        GameXplain           20
68698                              ESPN           17
39559                          Sailor J           24
62877                       Charlie Puth           10

             publish_time                                               tags  \
23604 2018-03-07 14:00:03  The Tonight Show|"Jimmy Fallon"|"Marshmello"|"...
25630 2018-03-16 04:00:01  Kirby|"Kirby Star Allies"|"Dedede"|"Meta Knigh...
68698 2018-04-17 14:55:31  espn|"dwyane wade"|"dwayne wade"|"d wade"|"76e...
39559 2017-11-15 13:29:28                  Zodiac|"makeup"|"comedy"|"aquarius"
62877 2018-03-15 16:02:17  Charlie|"Puth"|"charlie puth"|"Charlie Puth - ...

       likes  dislikes  comment_count  \
23604  45011      1156           2365
25630   2716        52            450
68698   6829       537           1445
39559   5172       453            976
62877  84227       739           8663

                                             description  ...     rank  \
23604  Music guest Marshmello & Anne-Marie performs F...  ...  27872.0
25630  Kirby Star Allies does something pretty fun wi...  ...  67076.0
68698  First Take's Stephen A. Smith says Kevin Hart ...  ...  57168.0
39559  Ya'll asked lol. What sign should I do next? D...  ...  60832.0
62877  Download & Stream Done For Me (feat. Kehlani):...  ...  18947.0

                                          new_text    views  \
23604                      marshmello annemarie friends  1443792
25630  kirby star allies surprising hd rumble secret ...   106398
68698  stephen kevin hart got feelings hurt dwyane wa...   976783
39559                                        aquarius    88644
62877     charlie puth done feat kehlani official audio   722009
```

Appx.1

	views	likes	dislikes	comment_count
views	1.000000	0.791670	0.405290	0.485986
likes	0.791670	1.000000	0.448010	0.763192
dislikes	0.405290	0.448010	1.000000	0.745064
comment_count	0.485986	0.763192	0.745064	1.000000

Appx.2



Appx.3

```
X_train columns: Index(['category_id', 'likes', 'dislikes', 'comment_count',
       'Engagement Metrics', 'score', 'rank', 'trending_day_of_week_Monday',
       'trending_day_of_week_Saturday', 'trending_day_of_week_Sunday',
       'trending_day_of_week_Thursday', 'trending_day_of_week_Tuesday',
       'trending_day_of_week_Wednesday', 'day_of_week_Monday',
       'day_of_week_Saturday', 'day_of_week_Sunday', 'day_of_week_Thursday',
       'day_of_week_Tuesday', 'day_of_week_Wednesday', 'trending_year',
       'trending_month', 'trending_day'],
      dtype='object')
```

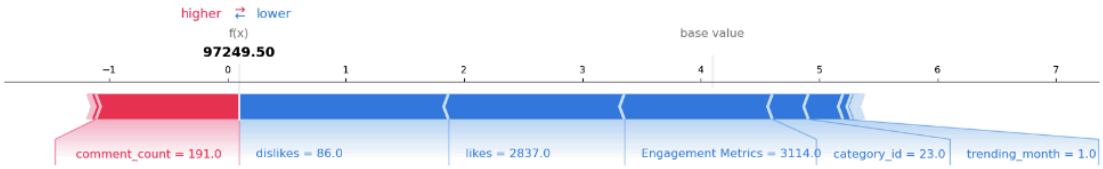
Appx. 4

```
Top 10 Features:
           Feature  Importance
1            title    0.437080
4     publish_time    0.193274
2    channel_title    0.085307
5             tags    0.051353
3      category_id    0.048652
14            rank    0.040345
15        new_text    0.027795
6            likes    0.009272
13           score    0.009106
0    trending_date    0.008889
```

Appx. 5

```
Top contributing features:
                      shap_value
feature
dislikes           -1.770361e+06
likes              -1.485569e+06
Engagement Metrics -1.251085e+06
comment_count       1.207505e+06
category_id        -3.041747e+05
```

Appx. 6 Index #49053 SHAP value



Appx. 7 Index #49053 SHAP force plot

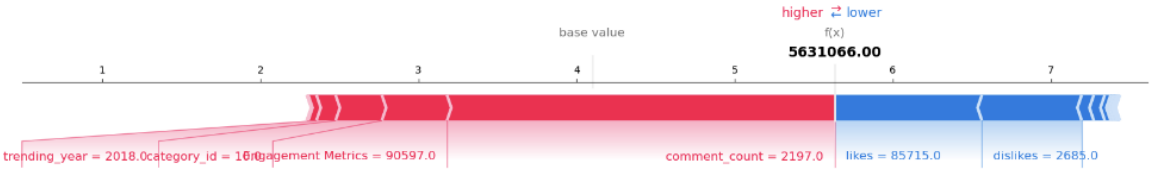
```
Top contributing features:
                      shap_value
feature
Engagement Metrics  5.385756e+06
dislikes           -4.143052e+06
comment_count       3.271754e+06
likes               1.796564e+06
trending_month     -2.142298e+05
```

\

Appx. 8: Index #71003 SHAP value and SHAP force plot

Top contributing features:

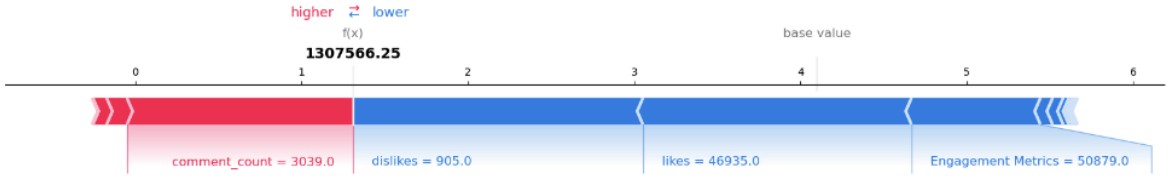
feature	shap_value
comment_count	2.451082e+06
likes	-9.323581e+05
dislikes	-6.310359e+05
Engagement Metrics	4.119252e+05
category_id	2.924393e+05



Appx. 9: Index #31981 SHAP value and SHAP force plot

Top contributing features:

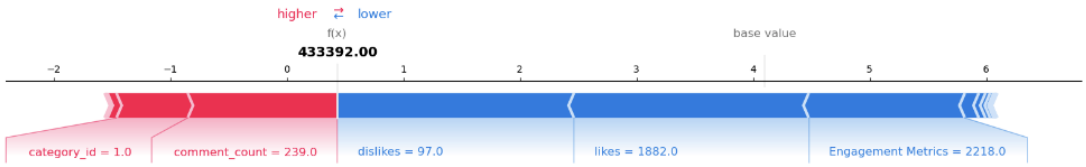
feature	shap_value
dislikes	-1.744691e+06
likes	-1.615474e+06
comment_count	1.359217e+06
Engagement Metrics	-7.754625e+05
trending_month	1.220741e+05



Appx. 10: Index #17588 SHAP value and SHAP force plot

Top contributing features:

feature	shap_value
dislikes	-2030229.750
likes	-2014201.250
Engagement Metrics	-1338711.000
comment_count	1282602.625
category_id	611415.875



Appx. 11: Index #76541 SHAP value and SHAP force plot