Week 10: Data Centric AI

Yujia Cao, Wendi Yuan, Yuhan Zhao

This week, we focused on enhancing our dataset by implementing new data improvement strategies to further boost model performance and reliability. Building on last week's findings, we confirmed that our XGBoost model, configured with learning_rate: 0.05, n_estimators: 200, and max_depth: 6, and incorporating the "Days Since Published" feature, remains the optimal choice. Our efforts centered on refining the training dataset after splitting, to maximize data quality and relevance.

In earlier stages of our project, we implemented several key improvements to streamline and optimize the dataset. We filtered the raw data to include only English-language videos to ensure linguistic consistency and dropped unnecessary columns to reduce feature complexity, focusing on relevant data. We applied normalization and standardization across numerical features for consistency, and performed text preprocessing by removing stopwords from titles and descriptions, refining our text features for enhanced model interpretation. After splitting the dataset into an 80/20 training and test set, we engineered additional features to add valuable context. These included the *Days Since Published* metric, capturing time-based insights, and sentiment analysis on descriptions and titles to quantify audience tone. Additionally, we implemented TF-IDF encoding for key words in titles and descriptions, enriching the dataset with informative textual features.

We explored three new strategies to further enhance our dataset, focusing on refining text-based variables to deepen model insights and support future analysis (Appx. 3). Given the nature of our research and the text-rich dataset, we opted to replace the *Days Since Published* feature with two targeted metrics. First, we introduced a *Basic Engagement Ratio Analysis*, creating features that capture engagement dynamics such as the comment-to-view ratio, offering a clearer view of audience interaction. Second, we added *Time-Based Metrics Analysis*, designed to capture temporal trends, including the day of the week a video is likely to gain popularity and the time it typically takes for a video to trend. Emphasizing metrics like engagement ratios and time-based insights allows us to better understand the factors that influence view counts, providing strategic insights into how content timing and audience engagement patterns impact video performance.

These new engagement ratio features offer a granular view of audience interaction, measuring relationships between *views*, *likes*, comments, and dislikes:

- *like_view_ratio*: Proportion of likes relative to views, indicating viewer approval.

- *comment_like_ratio*: Comments relative to likes, potentially reflecting higher viewer engagement.
- *dislike_view_ratio and comment_view_ratio*: Capture both negative (dislikes) and neutral (comments) engagement perspectives.
- *total_engagement_ratio*: Sum of likes, dislikes, and comments divided by views, representing overall engagement.
- *like_dislike_ratio*: Likes relative to dislikes, useful for sentiment analysis.

Features like *category_like_view_ratio* and *relative_category_engagement* provide context by comparing engagement to category averages. Additionally, *like_view_percentile* and *comment_like_percentile* rank engagement ratios, groping videos into quintiles from "Very Low" to "Very High," which reveals patterns in engagement levels and highlights top-performing videos. These features offer insights into the types of interactions that correlate most with video popularity, aiding in the segmentation of videos based on user interest and engagement.

To complement these, we calculated *hours_to_trend*, measuring the time between *publish_time* and *trending_date*, providing insights into how quickly a video gains traction after publishing and assessing its viral potential. Additionally, *publish_weekday* and *is_weekend* identify the day of the week and whether the video was published on a weekend, as engagement may vary by publishing time, offering guidance on optimal release strategies.

These changes made to the model have focused on increasing feature complexity and refining feature connectivity, all based on the performance of the previous model, which was already very good. By improving feature selection - removing irrelevant features, introducing new engagement metrics, and enhancing categorical feature coding - the model can now utilize a richer, more relevant set of predictors.

After retraining the model with these updated features, the model continues to perform well, maintaining strong predictive power. New features, such as *engagement ratio analysis* and *time-based metrics*, are designed to more effectively capture underlying patterns of user engagement. This additional information allows the model to identify subtleties in the data that may contribute to video success or engagement, making it possible to gain a more complete understanding of the factors influencing trends. For example, in order to improve the relationship between variables and to see more clearly the different variables measured in viewer engagement (*likes* and *views*), we chose to look further at the ratio between the two and divided them into five levels from very low to very high.

This process demonstrates how we chose to improve the original model. By introducing features on top of the better performing model, we obtained an improved model that maintains high performance while providing greater interpretability. This not only reinforces the importance of model selection, but also the importance of thoughtful feature engineering and validation through retraining-ensuring that the final model is also effective in the new data environment.

If we focus on the performing metrics measurement, RMSE compared to this week's configuration, the winning model has a lower RMSE on both the training and test datasets. The lower RMSE values indicate that the winning model is better able to minimize errors and therefore more accurately predict the target value. On the other hand, the R^2 in the winning model also had higher R^2 values on both the training and test sets, indicating that it explains more of the variance in the data compared to this week's model. This greater explanatory power suggests that the winning model is more consistent with the underlying data patterns(Appx.1 and Appx.2). Numerically, the XGBoost model we chose last week performed better than this week's improved XGBoost model.

The winning model with previous configuration achieved a good balance between training and testing performance, minimizing overfitting. Its R^2 and RMSE scores were similar on both datasets, indicating strong generalization and consistent performance on both seen and unseen data. This may be useful for us to develop practical applications of the model later. In contrast, this suggests that the increased complexity may have introduced noise or led to overfitting, which reduces the generalization ability.

Also, it benefited from selective feature engineering, which emphasized relevant predictors without overcomplicating the model. By removing irrelevant temporal features and focusing on key engagement metrics, the model captured essential information without redundancy.While this week's model is embedded with more complex features, it may also diminish the impact of truly predictive features due to increased complexity. This additional complexity may make it more difficult for the model to discern the most meaningful patterns.

The week 9's XGBoost model has already been completed after we have processed the missing data, performed feature engineering and feature culling, which for our study has included more influential features such as engagement metrics, with the ability to analyze video popularity. This allows it to focus on the most relevant predictors without being influenced by irrelevant temporal features or redundant information.

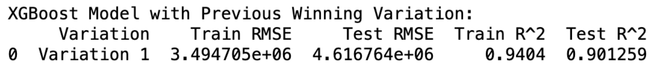
Our group made an analysis based on the XGBoost model with the best-performing variation after evaluating its predictive accuracy and consistency across different datasets. Since the winning model exhibits a training RMSE of approximately 2,120,971 and a test RMSE of around 2,909,425 from last week. This close alignment between training and test errors suggests that the model performs consistently on both known and unseen data, indicating minimal overfitting. Furthermore, the high R^2 values 0.978 for training and 0.961 for testing, indicate that the model captures a substantial portion of the variance in the data, making it highly reliable for our predictive tasks.

The minimal increase in RMSE between training and testing highlights that our model generalizes well and retains its predictive power even when applied to new data. This small variance between datasets implies that the model has effectively captured underlying patterns without becoming overly specialized to the training data. The high R^2 values across both datasets reinforce that our model is neither too simple nor excessively complex, allowing it to achieve a good balance between fitting the data accurately and avoiding overfitting.

We noted the optimized hyperparameters is learning rate of 0.05, 200 estimators, and a maximum depth of 6. Our group recognized this model's robustness and believes it is well suited for practical deployment in real-world applications. Moving forward, we recommend incorporating cross validation to confirm the model's consistency across various validation to confirm the model's consistency across various data subsets, ensuring that it continues to deliver accurate results. This XGBoost model stands as a reliable solution for predictive tasks, with strong generalization capabilities that align well with our group's objectives.(Appx.1 and Appx.2)

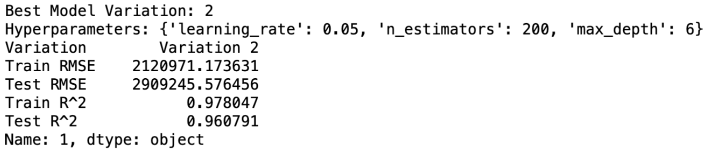
Finally, the model can handle a mix of different types of features (both numbers and categories) because of the preprocessing steps we took. This flexibility is a big strength because it allows the model to adapt to all kinds of patterns in the data. With the balanced performance on training and test sets, we can say that the model has successfully picked up general trends in video engagement without getting distracted by noise or random quirks in the training data. Considering these factors, we still picked the week 9's XGBoost configuration as the final model to be deployed. It not only maintains a high level of accuracy, but also strikes a balance between simplicity and predictive power. The aim of our research is not to be extremely complex and inaccessible, and by focusing on the essential features and avoiding unnecessary complexity, the model can prove to be more effective and robust in real-world applications. This balance is essential for deploying an interpretable model that performs well under different data conditions.

Appendix



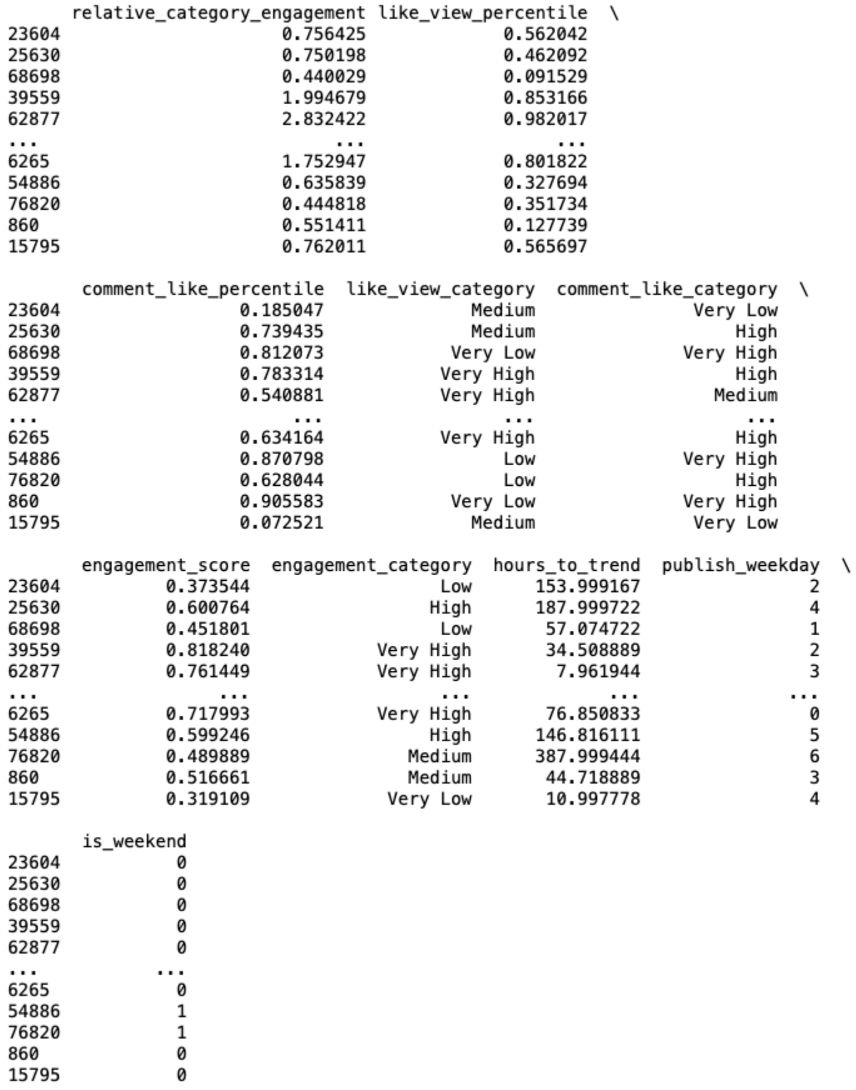
```
XGBoost Model with Previous Winning Variation:
     Variation    Train RMSE     Test RMSE  Train R^2  Test R^2
0  Variation 1  3.494705e+06  4.616764e+06     0.9404  0.901259
```

Appx.1



```
Best Model Variation: 2
Hyperparameters: {'learning_rate': 0.05, 'n_estimators': 200, 'max_depth': 6}
Variation        Variation 2
Train RMSE    2120971.173631
Test RMSE     2909245.576456
Train R^2           0.978047
Test R^2            0.960791
Name: 1, dtype: object
```

Appx.2



```
       relative_category_engagement  like_view_percentile  \
23604                      0.756425              0.562042
25630                      0.750198              0.462092
68698                      0.440029              0.091529
39559                      1.994679              0.853166
62877                      2.832422              0.982017
...                             ...                   ...
6265                       1.752947              0.801822
54886                      0.635839              0.327694
76820                      0.444818              0.351734
860                        0.551411              0.127739
15795                      0.762011              0.565697

       comment_like_percentile  like_view_category  comment_like_category  \
23604                 0.185047              Medium               Very Low
25630                 0.739435              Medium                   High
68698                 0.812073            Very Low              Very High
39559                 0.783314           Very High                   High
62877                 0.540881           Very High                 Medium
...                        ...                 ...                    ...
6265                  0.634164           Very High                   High
54886                 0.870798                 Low              Very High
76820                 0.628044                 Low                   High
860                   0.905583            Very Low              Very High
15795                 0.072521              Medium               Very Low

       engagement_score  engagement_category  hours_to_trend  publish_weekday  \
23604          0.373544                  Low      153.999167                2
25630          0.600764                 High      187.999722                4
68698          0.451801                  Low       57.074722                1
39559          0.818240            Very High       34.508889                2
62877          0.761449            Very High        7.961944                3
...                 ...                  ...             ...              ...
6265           0.717993            Very High       76.850833                0
54886          0.599246                 High      146.816111                5
76820          0.489889               Medium      387.999444                6
860            0.516661               Medium       44.718889                3
15795          0.319109             Very Low       10.997778                4

       is_weekend
23604           0
25630           0
68698           0
39559           0
62877           0
...           ...
6265            0
54886           1
76820           1
860             0
15795           0
```

Appx. 3