Week 9: Select the Winning Model

Yujia Cao, Wendi Yuan, Yuhan Zhao

This week, we finalized our model selection by comparing the performance of various models and selecting the one with the best metrics. After engineering features from our dataset and evaluating models aligned with our research objectives, we identified XGBoost with TF-IDF features as the winning model. This model, optimized with parameters of learning_rate: 0.05, n_estimators: 200, and max_depth: 6, was initially shortlisted in our first week and ultimately provided the highest R^2 score alongside strong RMSE performance (Appx. 1).

The chosen XGBoost model achieved the lowest test RMSE of approximately 2.91 million and a high test R^2 of 0.9608, meaning it captures about 96% of the variance in the test data. This combination of low error and high R^2 suggests that the model generalizes well without overfitting. RMSE measures the average error between predictions and actual values, offering an intuitive view of the model's accuracy. Meanwhile, the R^2 score indicates the proportion of variability in the target variable explained by the model, reflecting its ability to capture the data's underlying trends and complexity.

Ultimately, we selected the model with the lowest test RMSE, emphasizing accuracy on unseen data, and the highest test R^2, showcasing strong predictive power. Balancing these metrics ensures the model's ability to perform reliably and generalize well, highlighting it as the optimal choice for our analysis.

This week, we are examining the validation errors of different model configurations to determine the best-performing setup for our task. Notably, the XGBoost model with TF-IDF features, Variation 2, stands out due to its hyperparameters of a learning rate of 0.05, 200 estimators, and a max depth of 6. This configuration yields a train RMSE of approximately 2.12 million and a test RMSE of about 2.91 million, with high R^2 values on both training (0.978) and test data (0.961), indicating strong generalization. The performance on validation data further reinforces this configuration as a balanced and effective choice among the XGBoost models(Appx. 1).

The CNN models, configured with 32, 64, and 128 neurons, show stable validation MSE values close to 37,000, regardless of neuron count. Variation 2 (64 neurons) achieves the lowest validation MSE at 36,520.55, but the minor difference suggests that increasing the number of neurons in this particular problem does not significantly improve performance. This finding is echoed in the validation R^2 values,

which are consistently low across all CNN configurations, indicating limited variance in performance with neuron adjustments(Appx. 2).

We are analyzing the XGBoost models with LDA features. Similar to the TF-IDF-based models, Variation 2, with a learning rate of 0.05, 200 estimators, and max depth of 6, performs best among the LDA configurations. Although its validation MSE (around 75) is relatively higher compared to TF-IDF configurations, this result suggests that LDA features may not be as effective for this particular task. Overall, this week's analysis highlights the advantage of TF-IDF over LDA features and shows that fine-tuning hyperparameters in XGBoost yields the best validation performance across all tested models(Appx. 5).

The criteria to find the winning model is a synthesization of the model choice, the hyperparameter variation, the performing metric, so from these factors, the winning model is the variation 2 in XGBoost model with TF-IDF features. (Appx.6) The test outcomes revealed that the final model delivered the strongest performance, achieving a test RMSE of 2.909 million and an R^2 score of 0.9608. These metrics underscore the model's predictive reliability, with the R^2 score demonstrating that approximately 96% of the variance in the test data is captured, confirming its suitability for predictive analysis. The low RMSE value highlights the model's strong predictive accuracy, particularly in forecasting actual engagement metrics for YouTube videos.

This optimal performance was achieved with a learning rate of 0.05, 200 estimators, and a max depth of 6—showing a balance between model complexity and predictive accuracy, which supports effective generalization on unseen data. By identifying variation 2 as the superior model, this XGBoost-based approach provides a foundation for further exploration into the features influencing engagement. Future studies could delve into factors like sentiment or engagement trends, potentially guiding strategies for maximizing content reach and impact.

These insights can help content creators and marketers understand which video features are associated with higher engagement, which in turn can help refine content strategies to enhance audience interaction on platforms like YouTube, our platform of choice. When we begin a research project, we always return to the topic and project questions to examine how we approached our research topic at each step of the modeling process. The final model chosen was simply based on a comparison of the data available upfront, after we get this result, we will expand more in practical application.

The bias-variance chart provides valuable insights into how model complexity impacts the error for each of our nine model variations. The Y-axis represents the error (or loss) metric, specifically RMSE and MSE, while the X-axis arranges the models from simplest to most complex, moving left to right(Appx. 3). In examining the XGBoost models, we observe that increasing complexity initially leads to a significant decrease in both train and test errors. The variation 2 achieves the lowest test error, suggesting an ideal balance where the model captures the underlying patterns without overfitting. This reduction in error with increased complexity indicates a decrease in bias, as the model becomes better equipped to learn from the data.

However, as we move to even more complex configurations, such as Variation 3 and beyond, we see little to no additional improvement in test error, and the values begin to plateau. This pattern suggests that adding further complexity may not help—and could even lead to overfitting, where the model becomes too tailored to the training data at the expense of generalizing to new data(Appx. 3). For the CNN models, we see a consistent trend where increasing the number of neurons from 32 to 128 has minimal impact on validation error. This stability suggests that the CNN model might already be at its optimal capacity for this prediction task, and additional neurons do not provide a meaningful reduction in bias. The chart underscores the importance of finding an optimal level of complexity, showing that while increased complexity can initially improve model performance, excessive complexity does not always yield better generalization and can even hinder it.

The bias-variance plot also shows the training root-mean-square error and the test root-mean-square error (or validation root-mean-square error) for the different model variants, and help to provide insights into the performance of the model in terms of bias and variance.

From the view of model complexity, The graphs show that the initial models (variations 1 through 3) from two types of models (LDA embedded in XGBoost and XGBoost with TF-IDF features) have relatively high RMSE values for both the training and test data. The data on CNN is void because the performing metrics don't work for CNN model since it performs poorly on dataset, and performing metrics for CNN mostly are precision, recall or the f-1 score which are more useful in image classification tasks. This indicates underfitting of the model, i.e., the model is too simple to capture the underlying patterns in the data, which results in high bias. As the number of variants increases, the RMSE values decrease significantly, especially for variation 2, which may represent an optimal balance between bias and variance.(Appx.4)

Starting with variation 3, the RMSE decreases significantly, and the RMSE values in subsequent variants are almost identical, which may indicate that the models fit the data well, but are not overly complex. The test root mean square error is closely related to the training root mean square error, which means that these models generalize well to new data without overfitting (low variance). From the bias-variance tradeoff chart, the variation 2 in the XGBoost(with TF-IDF features) model appears to be the best tradeoff between bias and variance. The low training and testing RMSE values indicate that the model is well enough to learn from the training data while remaining generalizable, which means adaptable for future data. This suggests that variation 2 effectively captures the underlying relationships in the data with minimal error, balancing model complexity with predictive performance.

The graphs confirm that the RMSE values for the training and test data remain stable after Variable 2, indicating diminishing gains in model complexity. The stabilized performance suggests that these later models have reached optimal complexity, avoiding overfitting while reducing prediction error. This is the basic standard that how we choose the optimal model for our analysis, to see how performing metrics vary under different model types and hyperparameters.

Our final selected model, XGBoost with Variation 2, achieved strong performance metrics with an R^2 of 0.9608 and an RMSE of 2,909,245 on the test dataset. This Variation 2 model had the lowest test RMSE among the options, meaning it produces the most accurate predictions. A lower RMSE reflects better prediction accuracy in the same units as the target variable, making it easier to interpret. The model's high test R^2 score indicates it explains approximately 96% of the variance in the test data, showcasing its predictive power and reliability.

In comparison, Variations 1 and 3 have lower R^2 scores and higher test RMSEs, suggesting they may be underfitting or overfitting slightly relative to Variation 2(Appx.5). Variation 2 strikes a strong balance between complexity and performance, fitting both training and test data effectively without significant overfitting. The optimal hyperparameters for this model include a learning rate of 0.05 (to encourage generalization), 200 estimators (providing sufficient learning rounds without overfitting), and a max depth of 6 (balancing complexity with generalization).

RMSE helps gauge prediction accuracy, important for assessing video engagement metrics, while R^2 assesses how well the model explains variance in engagement, giving insight into the audience behaviors the model can capture. Identifying Variation 2 as the best model enables further research, such as analyzing sentiment or engagement patterns linked to higher view counts, helping us understand which

factors contribute to higher engagement. Notably, the similarity between training and test metrics demonstrates the model's robustness, generalizing well across datasets without overfitting or underfitting.

Appendix

```
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
```

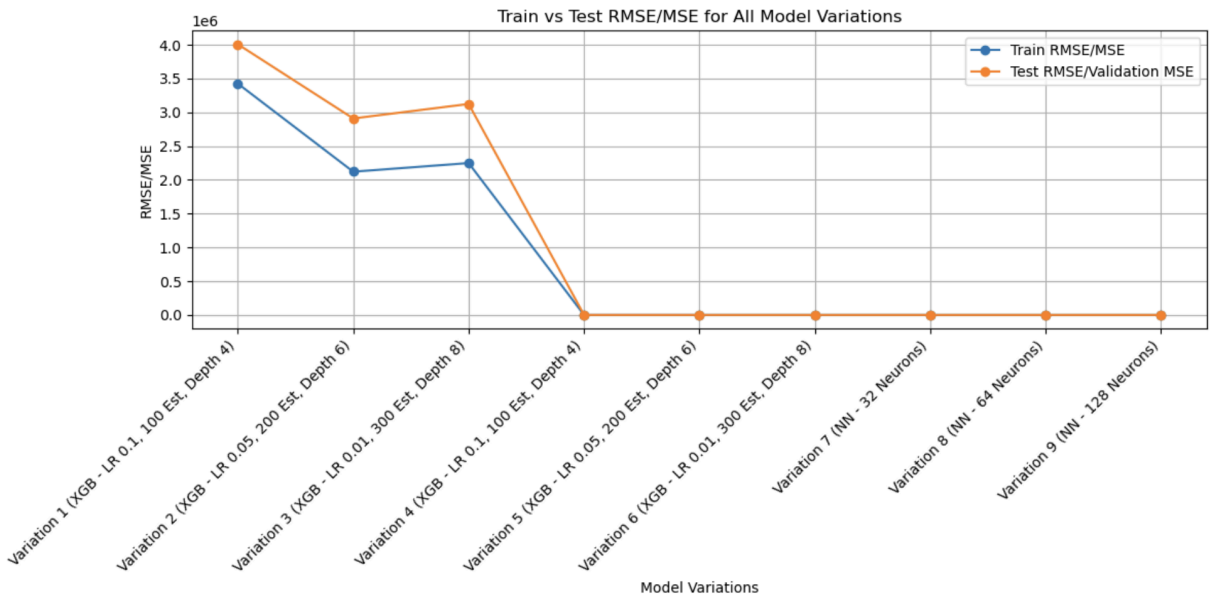
Appx.1 XGBoost #1 Model

```
Best Model:
Variation          Variation 2 (64 Neurons)
Train MSE                          37651.89
Validation MSE                     36520.55
Train R²                             0.0473
Validation R²                        0.0358
Name: 1, dtype: object
```

Appx.2 CNN Model

```
Best Model Variation: 2
Hyperparameters: {'learning_rate': 0.05, 'n_estimators': 200, 'max_depth': 6}
Variation     Variation 2
Train RMSE      75.719178
Test RMSE       95.550469
Train R^2        0.855187
Test R^2         0.764912
Name: 1, dtype: object
```

Appx.3 XGBoost #2 Model



Appx. 4 Bia-Variance Chart

```
Comparison of XGBoost Model Variations:
     Variation    Train RMSE     Test RMSE  Train R^2  Test R^2
0  Variation 1  3.425430e+06  4.007967e+06   0.942740  0.925583
1  Variation 2  2.120971e+06  2.909246e+06   0.978047  0.960791
2  Variation 3  2.248698e+06  3.123748e+06   0.975323  0.954796
```

Appx. 5 XGBoost Variations

Model Name	Numbers	Variations	Train RMSE	Test RMSE	Train R²	Test R²	Train MSE	Validation MSE	Validation R²
XGBoost TF-IDF	#1	LR 0.1, 100 Est, Depth 4	3.43E+06	4.01E+06	0.94274	0.925583			
	#2	LR 0.05, 200 Est, Depth 6	2.12E+06	2.91E+06	0.978047	0.960791			
	#3	LR 0.01, 300 Est, Depth 8	2.25E+06	3.12E+06	0.975323	0.954796			
CNN	#1	32 Neurons			0.0397		38195.26	37291.3	0.0302
	#2	64 Neurons			0.0473		37651.89	36520.55	0.0358
	#3	128 Neurons			0.0551		37213.45	37321.47	0.0319
XGBoost LDA	#1	LR 0.1, 100 Est, Depth 4	1.32E+02	1.38E+02	0.558682	0.511415			
	#2	LR 0.05, 200 Est, Depth 6	7.57E+01	9.56E+01	0.855187	0.764912			
	#3	LR 0.01, 300 Est, Depth 8	7.34E+01	9.93E+01	0.86376	0.746352			

Appx. 6 Model's Metrics Performance