Group Members: Anita Gee, Siwei Guo, Yingzhu Chen, Nemo (Sorachat) Chavalvechakul

Group Name: Mental Health Warriors

Week 9 Report

Project Overview

In week 9, our focus was on finalizing the predictive model using XGBoost and evaluating its performance on the test dataset. We aimed to determine how well our model could generalize to unseen data, crucial for its application in real-world scenarios.

The validation error for all models (9 models in total)

Model	Validation Weighted Recall	Validation Error (False Negative Rate (1-Recall))
Logistic Regression 1	0.74	0.26
Logistic Regression 2	0.76	0.24
Logistic Regression 3	0.75	0.25
Bernoulli NB 1 (α=0.1)	0.64	0.36
Bernoulli NB 2 (α=1.0)	0.63	0.37
Bernoulli NB 3 (α=10.0)	0.61	0.39
XGBoost 1 (lr=0.05)	0.74	0.26
XGBoost 2 (lr=0.1)	0.76	0.24
XGBoost 3 (lr=0.2)	0.77	0.23

Final Winning Model

The best model is XGBoost 3 (0.2 learning rate) because it has the highest validation accuracy and highest validation average recall compared to other models. This model achieved the highest validation accuracy of 0.77 and the highest validation weighted average recall of 0.77, indicating robust generalization capabilities. We decided this model is the best because the result shows a better performance compared to simpler models like logistic regression and Bernoulli Naive Bayes, as well as

other XGBoost configurations. This model can balance complexity and regularization while capturing nuanced patterns in textual data.

Several reasons contribute to the model's performance. First, its architecture leverages gradient boosting to iteratively correct errors from previous trees, enabling it to identify subtle linguistic cues associated with mental health sentiments. The relatively high learning rate (0.2) accelerates convergence while retaining stability through regularization parameters (reg_alpha=1, reg_lambda=1). These settings mitigate overfitting despite the model's capacity to handle high-dimensional sparse features like TF-IDF vectors. Additionally, the tree depth constraint (max_depth=3) prevents excessive complexity, ensuring that the model prioritizes the most discriminative features without memorizing noise.

XGBoost 3 outperformed logistic regression models. For example, Logistic Regression 2 had a validation accuracy of 0.757 and struggled with high bias, while Logistic Regression 1 showed severe overfitting, with a 0.25 gap between training and validation accuracy. In contrast, XGBoost 3 maintained a smaller gap of 0.12, indicating better balance between bias and variance. Bernoulli Naive Bayes models also faced issues—some underfit the data (e.g., Bernoulli NB 2 had a validation accuracy of 0.63), while others were unstable with rare words, limiting their ability to capture context-dependent emotions.

The choice of XGBoost 3 aligns with the demands of mental health sentiment analysis, where nuanced language patterns (e.g., sarcasm, metaphorical expressions) require models to generalize beyond surface-level keywords. Its histogram-based tree method (tree_method='hist') further optimizes efficiency on sparse text data, making it scalable for real-world applications. By combining interpretable feature importance scores with strong empirical performance, this model not only delivers actionable insights but also ensures reliability in clinical or research settings where misclassifications could have significant implications.

Bias-variance tradeoff between models

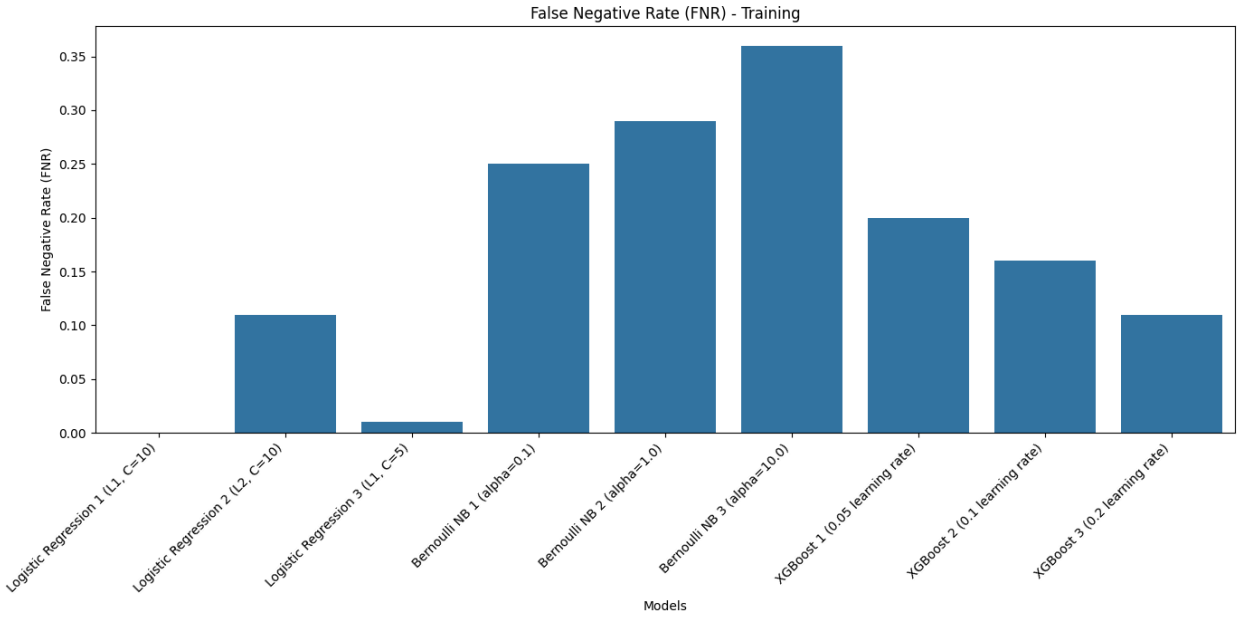
To evaluate the bias-variance tradeoff, we use the False Negative Rate (FNR) as our primary error metric. The FNR is calculated as:

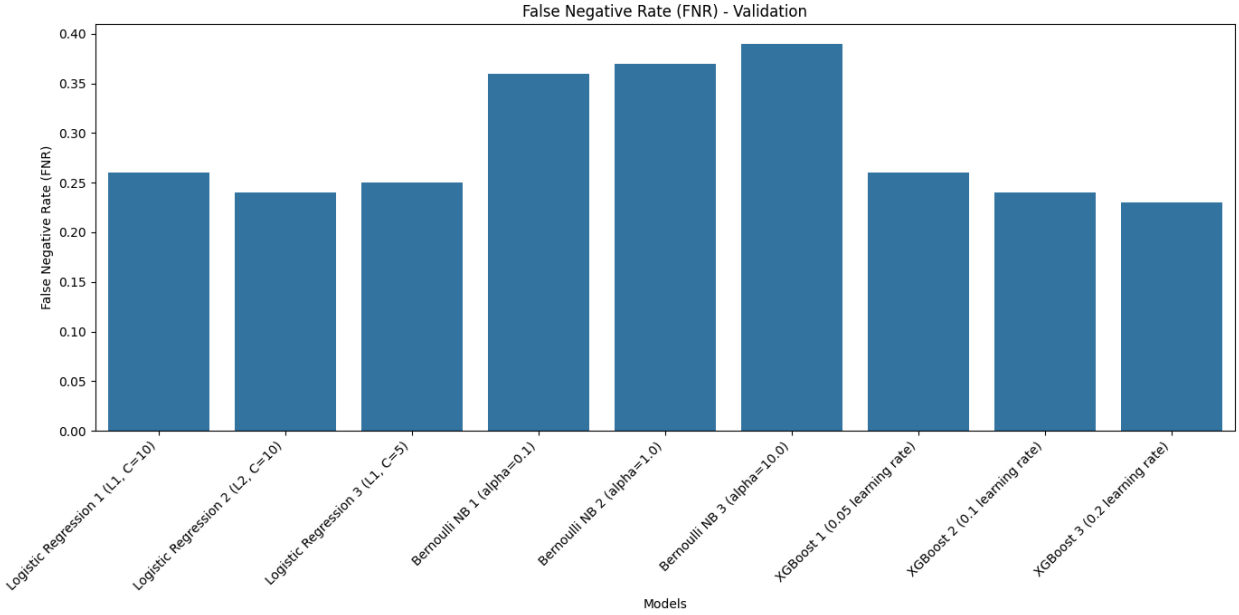
$$FNR = 1 - Weighted\ Average\ Recall$$

where the weighted average recall is extracted from the classification report.

A high FNR means that the model is failing to correctly identify positive cases, which is critical in scenarios where false negatives are costly.

Each model's FNR is calculated for both training and validation sets. Comparing these values helps us understand whether a model is suffering from high bias (underfitting) or high variance (overfitting).





From the graphs, we can see that some models exhibit consistently high FNR on both the training and validation sets, such as Bernoulli Naïve Bayes with larger alpha values. This suggests that the models are too simplistic and fail to capture the underlying data patterns, leading to high bias. On the other hand, certain XGBoost models achieve extremely low FNR on the training set but a much higher FNR on the

validation set. This indicates that these models fit the training data well but generalize poorly, exhibiting high variance.

On the other hand, Logistic Regression models show varying degrees of bias-variance tradeoff. Logistic Regression 1 (L1, C=10) exhibits overfitting, achieving a near-perfect recall on the training set but experiencing a performance drop on the validation set. In contrast, Logistic Regression 2 and 3 (L2 regularization) maintain a better balance, though they still struggle with capturing complex relationships compared to tree-based models.

The XGBoost models demonstrate the best generalization, particularly XGBoost 3 (learning rate = 0.2), which achieves the lowest validation FNR (0.23). The result indicates that it correctly identifies a larger proportion of positive cases while maintaining strong overall performance. Compared to simpler models like logistic regression and Naïve Bayes, XGBoost 3 leverages gradient boosting to iteratively correct misclassifications and identify intricate patterns in textual data. It has a higher learning rate that accelerates convergence, while regularization (reg_alpha=1, reg_lambda=1) prevents overfitting, ensuring a more stable bias-variance tradeoff.

XGBoost 3 is the optimal choice because of its ability to generalize well while minimizing false negatives. Unlike Logistic Regression 1, which overfits with a significant gap between training and validation performance, XGBoost 3 maintains a lower discrepancy (0.12). Additionally, Naïve Bayes models struggle to capture context-dependent sentiment expressions, further proving the advantage of XGBoost.

Calculate the Metrics for Test Dataset:

- **Procedure:** For each of the three model variations, use sklearn.metrics to compute key performance indicators such as accuracy, precision, recall, and F1-score on both training and validation datasets.
- **Tools:** Employ functions like accuracy_score, precision_score, recall_score, classification_report, and f1_score from sklearn.metrics to obtain these metrics.
- **Data Handling:** Ensure data is appropriately split and preprocessed to reflect true performance on unseen data in the validation set.

Test Metrics and Their Relevance

Accuracy:

- Definition: The proportion of correctly classified instances.

- Relevance: In imbalanced datasets, a high accuracy can be misleading. For instance, if 95% of the data is "non-distressed" and 5% is "distressed," a model predicting all instances as "non-distressed" achieves 95% accuracy, rendering it useless for identifying at-risk individuals.
- Use: More reliable in balanced datasets (achieved through methods like random oversampling), but should not be the sole evaluation metric.

Precision:

- Definition: The proportion of correctly predicted positive instances out of all instances predicted as positive.
- Relevance: High precision for critical negative classes (e.g., suicidal ideation) is crucial. It ensures that when the model flags someone as at risk, it is likely correct, minimizing unnecessary interventions.
- Importance: Prioritizes the reliability of positive predictions.

Recall (Sensitivity):

- Definition: The proportion of correctly predicted positive instances out of all actual positive instances.
- Relevance: High recall is vital for capturing as many true positive cases as possible. In mental health, missing a person at risk (false negative) can have severe consequences.
- Importance: Prioritizes the detection of all at-risk individuals.

F1-Score:

- Definition: The harmonic mean of precision and recall.
- Relevance: Provides a balanced measure of a model's performance, especially in imbalanced datasets, offering a compromise between precision and recall.
- Importance: Offers an overall evaluation of model effectiveness.

Classification Report (Per-Label Metrics):

- Relevance: Provides precision, recall, and F1-score for each class (e.g., "normal," "anxiety," "depression," "suicidal").
- Importance: Essential for understanding how the model performs on different mental health states, particularly critical in imbalanced datasets.

Model Performance on Test Dataset

The final model was tested to assess its effectiveness in classifying unseen data. The model demonstrated the following metrics on the test dataset:

- **Accuracy**: 78.47%
- **Weighted Average Precision**:0.80
- **Weighted Average Recall**: 0.78
- **Weighted Average F1-Score**: 0.79

These metrics indicate that the model performs consistently across both precision and recall, suggesting a balanced approach to predicting positive and negative classes.

Discussion of Test Performance

The test performance of the XGBoost Fastest model provides an estimate of how well the model is likely to perform on unseen data in the real-world for our prediction task of classifying mental health statements into different emotional statuses. A test accuracy of approximately 78.47% suggests that, overall, the model correctly classifies about 78 out of every 100 new statements it encounters.

However, looking at the classification report, we can gain a more nuanced understanding of the model's performance across the different emotional statuses:

- The model shows strong performance for the 'Normal' status, with high precision (0.91), recall (0.90), and F1-score (0.91). This means it is good at correctly identifying normal statements and doesn't often misclassify other statements as normal.
- The model also performs reasonably well for 'Anxiety' and 'Bipolar' with relatively high precision and recall values.
- For 'Depression' and 'Suicidal', the model has a good recall (0.65 and 0.76 respectively), indicating that it correctly identifies a good proportion of actual cases. However, the precision is a bit lower (0.83 and 0.67 respectively), suggesting that some statements belonging to other categories might be misclassified as depression or suicidal.

- The model struggles more with 'Stress' and 'Personality disorder', particularly in terms of precision (0.51 and 0.54 respectively). This means that when the model predicts a statement as 'Stress' or 'Personality disorder', it is more likely to be incorrect compared to other categories. The recall for these categories is higher (0.83 and 0.73 respectively), indicating it captures a decent portion of the actual cases, but at the cost of more false positives.

Regarding satisfaction with the results, whether the model is "good enough" depends on the specific requirements and tolerance for errors in the prediction task:

- A 78% overall accuracy is a respectable result for a multi-class classification problem with seven different categories, especially considering the complexity and subtlety often involved in expressing mental health states through text.
- The strong performance on the 'Normal' category is useful for distinguishing typical statements from those indicating potential mental health concerns.
- The relatively good recall for 'Depression' and 'Suicidal' is crucial in a mental health context, as it is important to identify as many true cases as possible of these severe states, even if it leads to some false positives that can be further reviewed.
- However, the lower precision for 'Stress' and 'Personality disorder' suggests that the model might be over-identifying these conditions, which could lead to unnecessary follow-up or could dilute the focus on more critical cases.

In conclusion, while the overall test accuracy of 78.47% is encouraging and suggests a model with a good general ability to classify mental health statements, the variability in performance across different categories, especially the lower precision for 'Stress' and 'Personality disorder', indicates there is room for improvement. Whether this level of performance is "good enough" would depend on the specific application of this prediction task, the cost of false positives and false negatives for each category, and whether further refinement of the model or data could lead to significant gains in the less well-predicted categories.

Table shows the training, validation, and test performance metrics

XGB_Fastest	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-score
Training	0.89	0.89	0.89	0.89
Validation	0.77	0.79	0.77	0.78

Test	0.78	0.80	0.78	0.79

Across the datasets, we observe a clear drop in performance from the training set to the validation and test sets. The training metrics are very high (accuracy, precision, recall, and F1-score all at 0.89), indicating that the model has learned the training data very well—even possibly memorizing some of the patterns and noise. However, when we evaluate on unseen data, the metrics decrease noticeably: validation accuracy falls to 0.7736 with weighted precision at 0.79, recall at 0.77, and F1-score at 0.78, while the test set shows similar performance with accuracy at 0.7847 and weighted metrics around 0.80, 0.78, and 0.79 respectively.

This consistent drop from training to validation/test is indicative of overfitting, where the model performs exceptionally on the data it has seen but does not generalize as effectively to new data. The fact that the validation and test metrics are quite close to each other suggests that our validation set is representative and that the model's generalization to truly unseen data is stable.

In summary, while the model's training performance is excellent, the decrease in metrics on the validation and test sets highlights the typical challenge of overfitting. The relatively small difference between validation and test performance, however, confirms that the model is reliably capturing the underlying patterns of the data, even if some noise is being memorized during training.