BIS568: Homework 3

Rong Sun

2025-04-01

1 Introduction

In this assignment, we focus on MLOps and model interpretability while predicting urinary tract infections (UTIs) using the dataset from Yun et al. (2018). Building upon prior work from Programming Assignment #2, we aim to refine our machine learning workflow by implementing structured pipelines for training, validation, and testing. We compare the performance of Logistic Regression and XGBoost models while leveraging MLflow to track model parameters and evaluation metrics.

Beyond model development, we emphasize interpretability by analyzing feature importance. Using the Logistic Regression model, we report odds ratios for the ten most influential variables. Additionally, we apply SHapley Additive exPlanations (SHAP) to generate a feature importance summary, highlight the top 20 contributing features, and visualize the dependence of white blood cell count on model predictions. These insights, along with the generated plots, are logged in MLflow to ensure transparency and reproducibility in our machine learning pipeline.

This assignment integrates predictive modeling with explainability, aligning with best practices in MLOps for robust, interpretable machine learning solutions.

2 Features selected for analysis

2.1 Data Splitting Process

To ensure robust model evaluation, the dataset was split into three subsets:

- Training Set (70%) – Used to fit the model.
- Validation Set (15%) – Used for hyperparameter tuning and model selection.
- Test Set (15%) – Used for final model evaluation.

The data was initially split into 70% training data and 30% temporary data. The temporary data was then further divided equally into 15% validation data and 15% test data using stratified sampling based on the UTI diagnosis variable. This approach ensures that the class distribution remains consistent across all subsets. Additionally, missing values were handled by replacing `NA` values in categorical variables with `not_reported` to maintain interpretability, and extreme outliers in urine specific gravity and age were removed using the 1st and 99th percentiles.

2.2 Feature Selection Process

Feature selection was performed through a systematic, multi-stage process to retain the most relevant predictors while ensuring model interpretability and performance. Initially, a logistic regression model was fitted using all available features, including demographic, clinical, and categorical variables. Maximum likelihood estimation (MLE) was used to estimate coefficients, providing an initial assessment of variable significance through p-values and effect sizes. To refine the model, variables with p-values below 0.1 were retained, reducing the risk of prematurely excluding potential predictors.

Further refinement involved backward elimination using likelihood ratio tests (LRT), where the least significant predictors were sequentially removed, provided their exclusion did not significantly worsen model fit. Categorical variables, such as ethnicity and insurance status, were transformed into binary indicators to improve interpretability and avoid misleading ordinal assumptions. Finally, model performance was evaluated using AUC, multicollinearity checks, and predictive accuracy metrics. The final set of predictors was selected based on statistical significance and overall model performance, ensuring a balance between parsimony and predictive power.

2.3 Final Selected Features

Based on Feature importance in the glm we picked these key variables: “ua_bacteria”, “ua_clarity”, “ua_epi”, “ua_wbc”, “CVA_tenderness”, “psychiatric_confusion”, “flank_pain”, “age”, “gender”, “ethnicity_Non_Hispanic”, “patid” (retained for potential interpretability). This streamlined feature set ensures that the model remains interpretable while maintaining strong predictive performance (AUC $\approx$ 0.81).

The features given by XGBoost were fairly similar to those selected in the logistic regression model, reinforcing their importance in predicting the target outcome. Based on feature importance ranking from `xgb.importance()`, the top 10 selected features for the XGBoost model include: “abx”, “patid”, “ua_wbc”, “ua_leuk”, “ua_nitrite”, “ua_bacteria”, “dispo”, “antibiotics”, “age”, “chief_complaint”. These variables emerged as the most influential in the XGBoost model, aligning closely with the predictors identified through logistic regression, thereby validating their relevance in modeling the outcome.

3 Model Performance

3.1 GLM Model Performance Evaluation

The logistic regression (GLM) model was assessed using AUC-ROC, Precision-Recall AUC, and Calibration Curve Analysis, with key results summarized below.

1. AUC-ROC Performance (Discrimination Ability): The logistic regression model exhibits strong discriminative performance in classifying UTI-positive and UTI-negative cases, as reflected in the AUC-ROC values of 0.8106 (validation) and 0.8117 (test). Since an AUC above 0.80 is generally considered excellent, these results indicate that the model effectively differentiates between the two classes. The minimal difference between validation and test AUC suggests good generalization and low risk of overfitting. Additionally, the smooth ROC curves further support the model's stability and reliability in distinguishing cases.

2. Precision-Recall AUC Performance (Handling of Class Imbalance): In imbalanced healthcare datasets, the Precision-Recall (PR) AUC provides critical insights beyond AUC-ROC by focusing on performance in the minority class. The model's PR AUC scores of 0.6349 (validation) and 0.6352 (test) indicate moderate precision and recall, which is reasonable given the lower prevalence of positive cases. The gap between AUC-ROC and PR AUC suggests that while the model ranks observations well, precision could be improved without sacrificing recall. Enhancing positive case detection may require adjusting classification thresholds, incorporating weighted loss functions, or applying resampling techniques to better balance the dataset.

3. Calibration Curve Analysis (Reliability of Predicted Probabilities): The model's probability predictions align well with observed event frequencies, as confirmed by binned calibration plots and val.prob calibration curves. No systematic overestimation or underestimation trends were observed, with the calibration curve closely following the diagonal reference line. This suggests that the model generates well-calibrated probability estimates, making its outputs interpretable and trustworthy for clinical decision-making without requiring extensive post-processing adjustments.

4. Potential Areas for Enhancement: While the model demonstrates strong performance, further optimization could improve precision and recall. Strategies such as threshold tuning, incorporating additional predictive features, or refining feature engineering may enhance classification accuracy. Despite these potential refinements, the current logistic regression model provides a robust and interpretable foundation for UTI diagnosis, balancing discrimination, calibration, and generalizability effectively.

Therefore, the logistic regression model demonstrates strong discrimination ability (AUC-ROC ~ 0.81), reasonable handling of class imbalance (PR AUC ~ 0.63), and well-calibrated probability predictions. The model is generalizable, as shown by consistent performance across validation and test datasets. While the AUC-ROC suggests excellent predictive capability, enhancements in handling class imbalance could improve performance in identifying UTI cases more effectively.

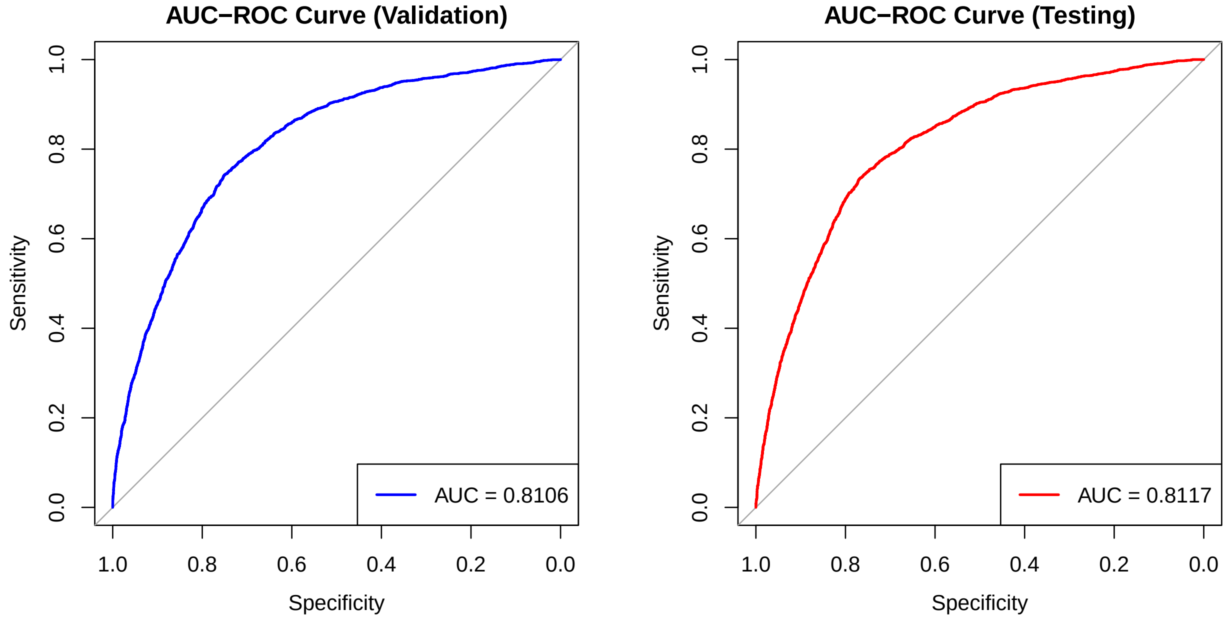

Figure 1: AUC-ROC Curve (Logistic Regression)

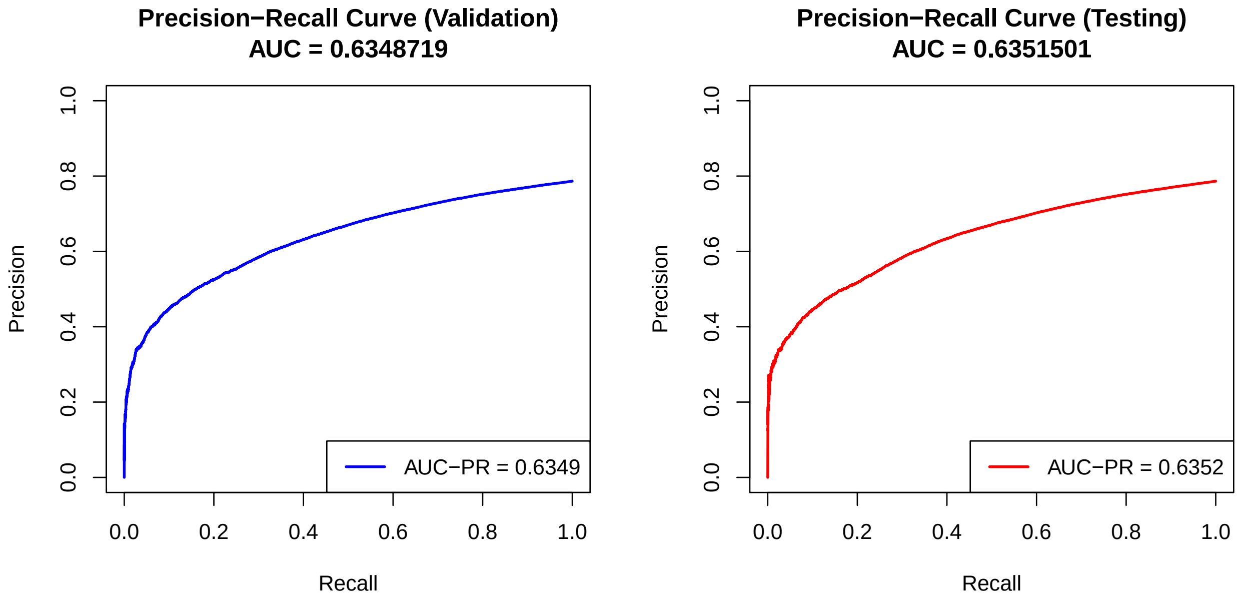

Figure 2: Precision-Recall AUC Curve (Logistic Regression)

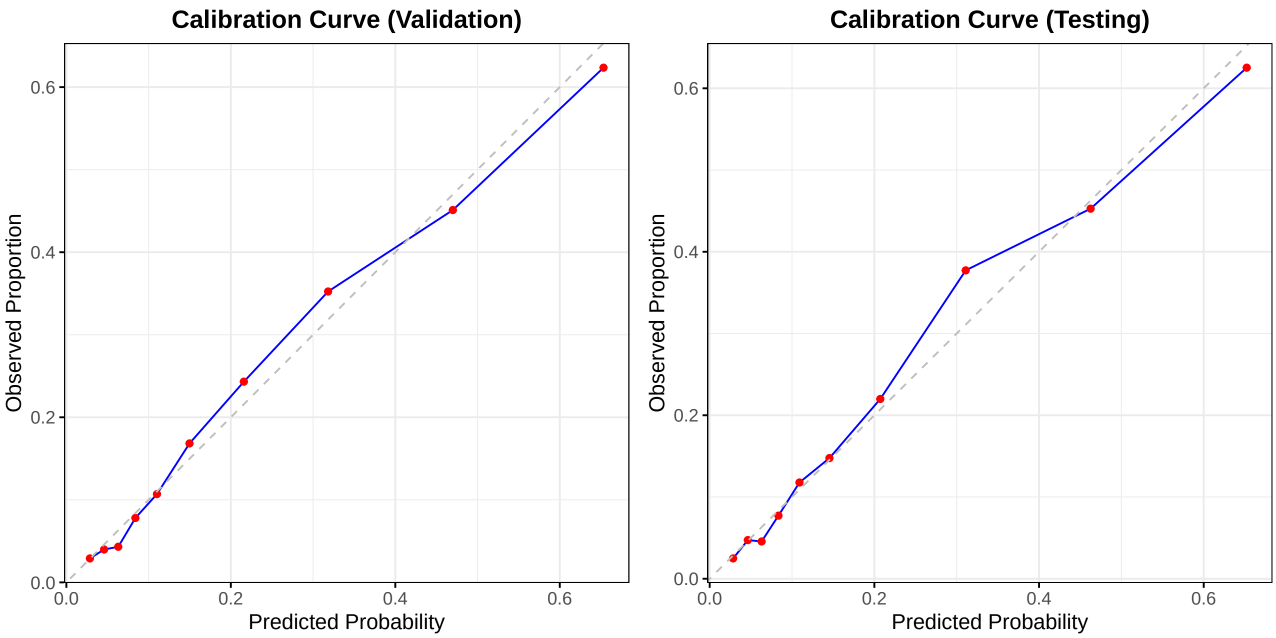

Figure 3: Calibration Curve (Logistic Regression - prob)

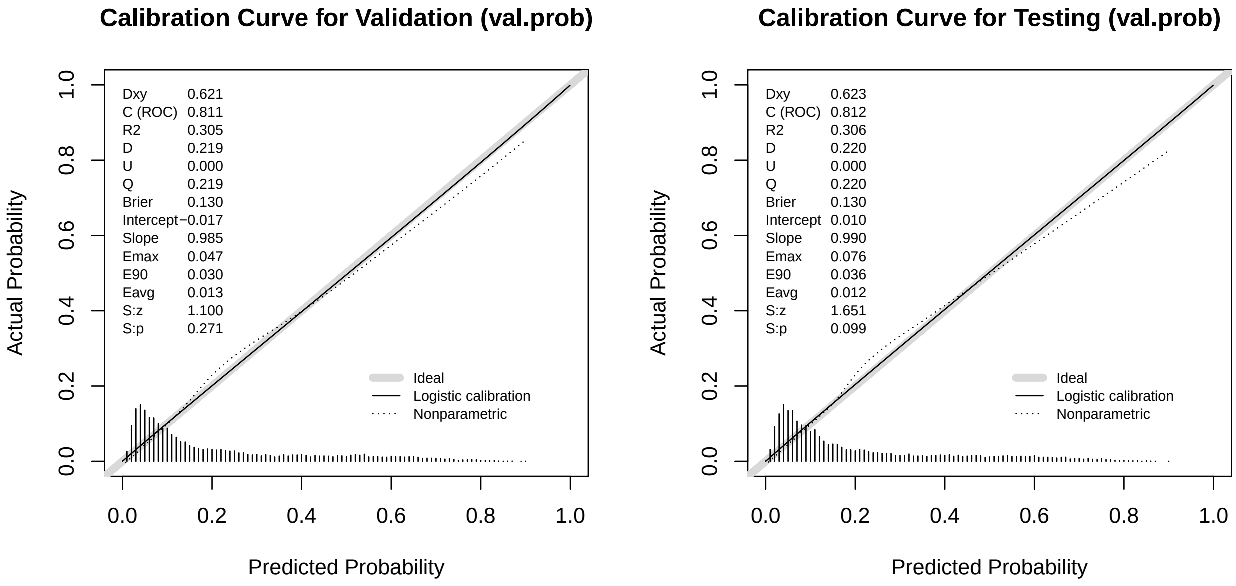

Figure 4: Calibration Curve (Logistic Regression - val.prob)

3.2 XGBoost Model Performance Evaluation

The XGBoost model demonstrates strong classification performance, achieving 88.9% validation accuracy and 89.6% test accuracy, indicating effective generalization to unseen data. The consistency between validation and test accuracy suggests that the model is not overfitting and maintains stable performance across datasets.

However, classification accuracy alone does not fully capture model effectiveness, especially in the presence of class imbalance. A more comprehensive evaluation using AUC-ROC, Precision-Recall AUC, feature importance analysis, and calibration plots provides deeper insights into prediction quality. These additional metrics help assess the model's ability to balance sensitivity and specificity, properly rank predictions, and produce well-calibrated probability estimates.

1. Feature Importance Analysis (Key Drivers of Model Decisions): The top 10 contributing features in the XGBoost model include "abx", "patid", "ua_wbc", "ua_leuk", "ua_nitrite", "ua_bacteria", "dispo", "antibiotics", "age", and "chief_complaint". These variables hold strong clinical relevance, aligning with expectations and reinforcing that the model captures meaningful patterns. The feature importance distribution indicates that a small subset of key predictors dominates decision-making, with diminishing contributions from other variables. This suggests that removing low-importance features may streamline the model while maintaining predictive performance. However, domain expertise remains essential in guiding any feature selection to ensure that clinically relevant factors are retained.

2. AUC-ROC Performance (Discrimination Ability): The model demonstrated excellent discrimination ability, with AUC-ROC values of 0.9348 (validation) and 0.9564 (test), indicating strong performance in distinguishing between positive and negative cases. An AUC above 0.90 is considered outstanding, and the higher AUC on the test set suggests good generalization with minimal overfitting. Both the validation and test ROC curves were smooth, reflecting stable decision boundaries and consistent performance. The minimal difference between AUC values across datasets further supports the model's robustness. However, while AUC-ROC highlights overall discrimination, it does not address class imbalance, necessitating a closer look at Precision-Recall metrics to assess performance in identifying positive cases.

3. Precision-Recall AUC Performance (Handling of Class Imbalance): The model's Precision-Recall (PR) AUC values were 0.5934 on the validation set and 0.5878 on the test set, reflecting moderate performance in handling class imbalance. These values indicate that while the model ranks cases effectively, its ability to accurately classify the minority class (likely positive cases) remains limited. The gap between the AUC-ROC and PR AUC suggests that while the model is strong in overall discrimination, it faces challenges with precision when identifying the minority class. This highlights the need for further improvements, such as class weighting, threshold tuning, or data balancing techniques like SMOTE, to better capture and classify positive cases.

4. Calibration Curve Analysis (Reliability of Predicted Probabilities): The calibration curves compare predicted probabilities to observed event frequencies, assessing the alignment

between the model's confidence and reality. The model is generally well-calibrated, with reliable probability estimates, though slight deviations from the diagonal indicate minor miscalibration. When the curve is below the diagonal, the model is overconfident, assigning probabilities that are too high; when above, it is underconfident, assigning probabilities that are too low. Post-processing techniques like Platt scaling or isotonic regression can improve calibration, ensuring better alignment with actual outcomes. This is crucial for clinical decision-making, where accurate probability estimates guide risk assessments and treatment choices.

Therefore, the XGBoost model demonstrates excellent discrimination ability (AUC-ROC ~ 0.94), moderate handling of class imbalance (PR AUC ~ 0.59), and generally reliable probability predictions. The model is generalizable, as shown by consistent performance across validation and test datasets. While the AUC-ROC suggests strong overall predictive capability, improvements in handling class imbalance could enhance the model's ability to more effectively classify minority cases, such as positive outcomes.

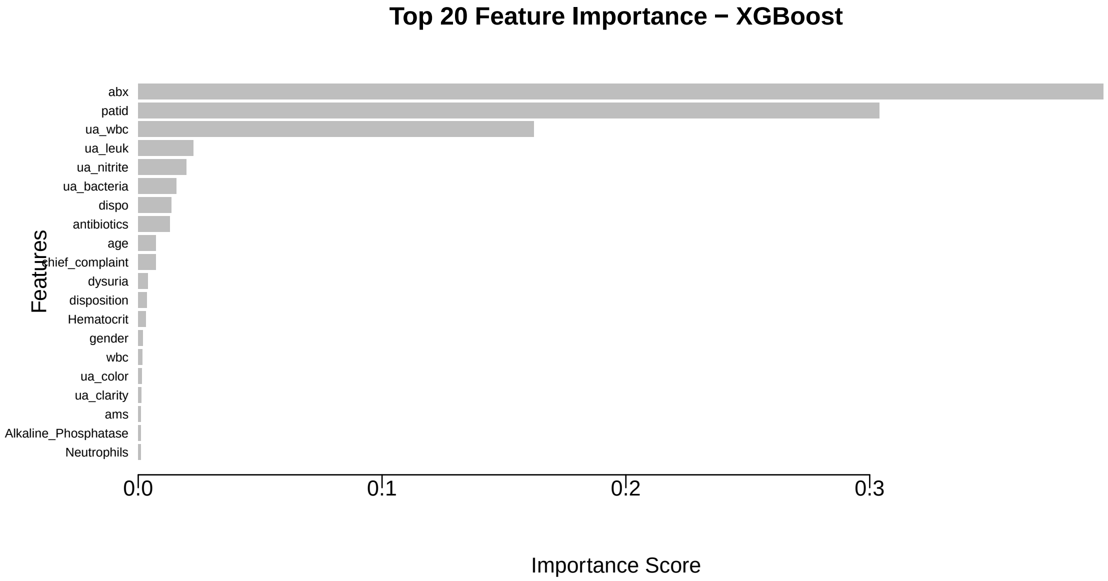

Figure 5: Top 20 Feature Importance - XGBoost

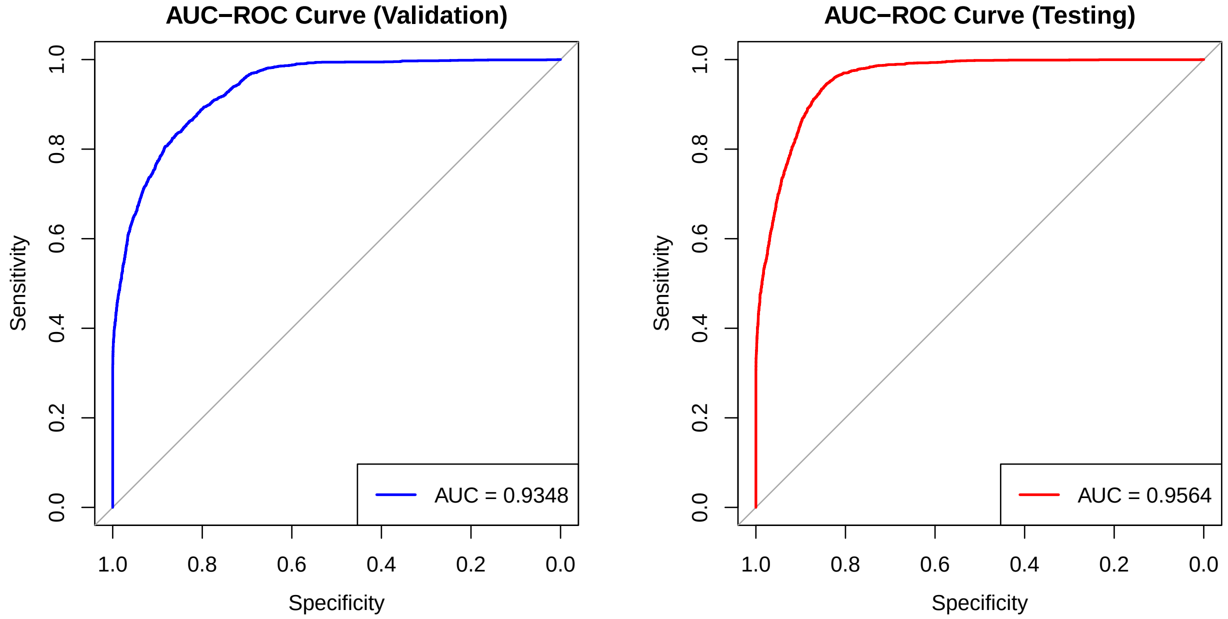

Figure 6: AUC-ROC Curve (XGBoost)

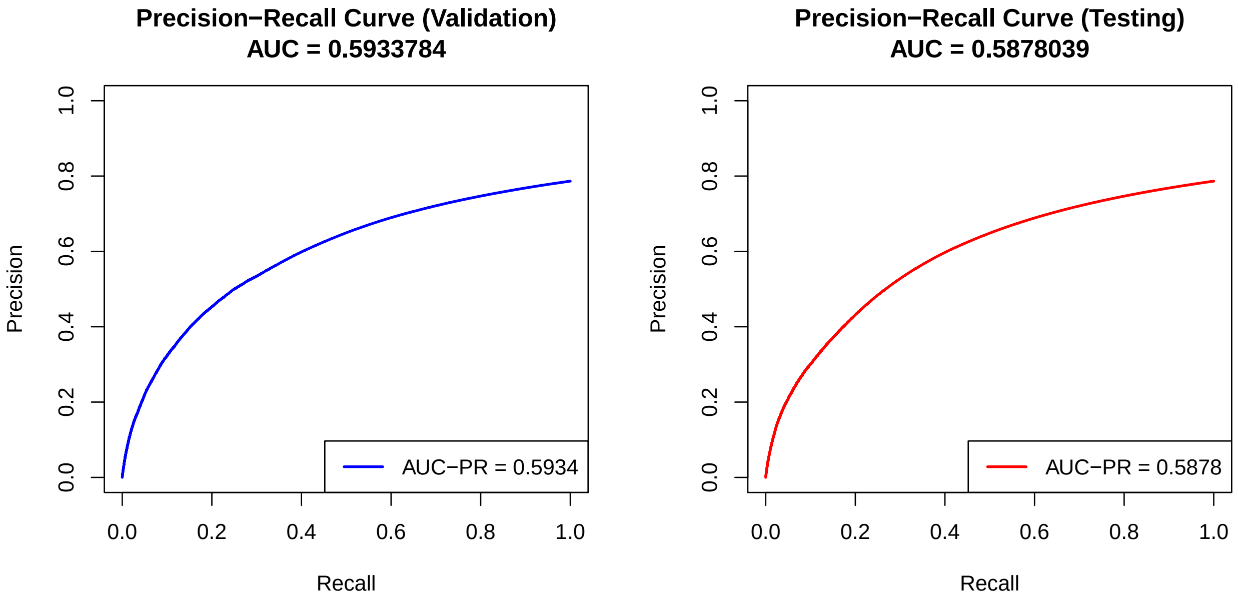

Figure 7: Precision-Recall AUC Curve (XGBoost)

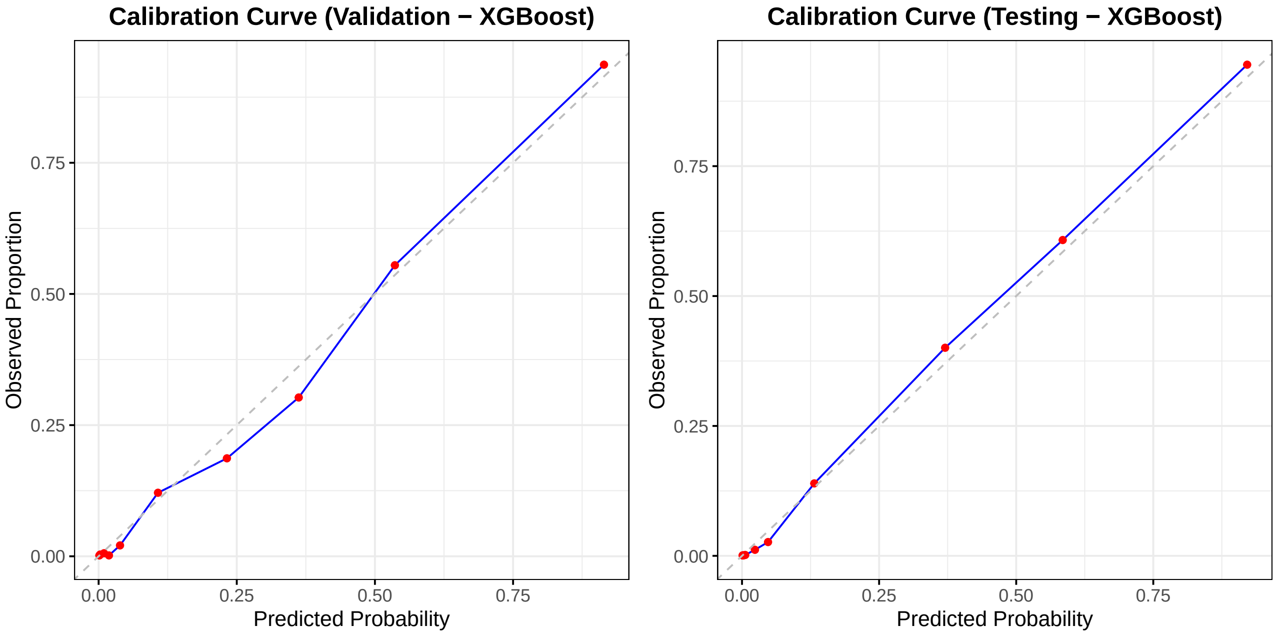

Figure 8: Calibration Curve (XGBoost - prob)

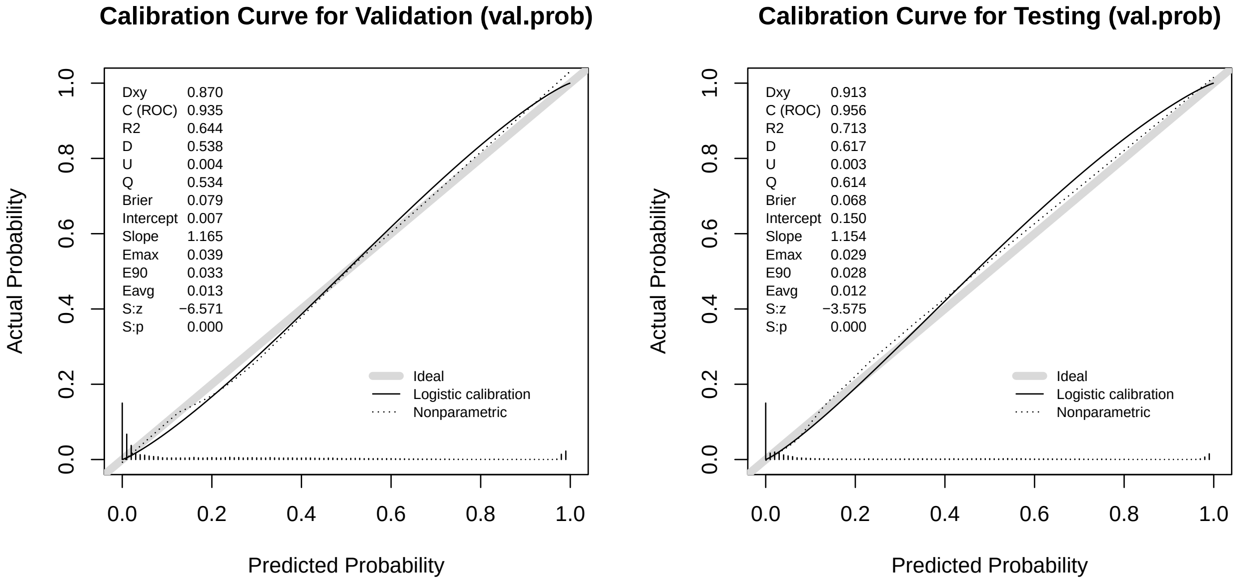

Figure 9: Calibration Curve (XGBoost - val.prob)

4 Interpretability and Explainability

4.1 Odds Ratios (OR)

Here, we present the odds ratios (OR) along with their 95% confidence intervals for the selected variables in the logistic regression model, highlighting the relative impact of each predictor on the outcome:

Variable	Odds Ratio (OR)	Lower 95% CI	Upper 95% CI
(Intercept)	1.0088670	0.8343131	1.2194250
patid	0.9999724	0.9999708	0.9999739
ua_bacteriamany	1.3795866	1.2676226	1.5013886
ua_bacteriamarked	1.7069987	1.5106642	1.9286915
ua_bacteriamoderate	1.3750973	1.2771083	1.4804458
ua_bacterianone	0.2852436	0.2516283	0.3225098
ua_bacterianot_reported	0.8517616	0.7884875	0.9198778
ua_claritynot_clear	1.1709974	1.1004790	1.2459759
ua_claritynot_reported	0.5850186	0.5409482	0.6324452
ua_epimoderate	1.3882685	1.2283681	1.5703230
ua_epinegative	2.0412792	1.7353770	2.4016610
ua_epinot_reported	2.3959436	2.1052525	2.7292611
ua_epiother	1.4478558	0.1989411	6.5898836
ua_epismall	2.1759289	1.9465410	2.4353977
ua_wbcmoderate	0.6766772	0.6247459	0.7328598
ua_wbcnegative	0.1213986	0.0993691	0.1474784
ua_wbcnot_reported	0.0634834	0.0560026	0.0719299
ua_wbcother	0.3286416	0.0874962	1.0139879
ua_wbcsmall	0.1263746	0.1156772	0.1380237
CVA_tenderness1	0.8286293	0.7211039	0.9511976
CVA_tendernessnot_reported	0.9396103	0.8756972	1.0085858
psychiatric_confusion1	1.6832375	1.4529255	1.9478859
psychiatric_confusionnot_reported	1.4263682	1.3267113	1.5344112
flank_pain1	0.6141675	0.5563708	0.6776513
flank_painnot_reported	0.7722243	0.7250145	0.8226782
age	1.0059572	1.0047455	1.0071711
genderMale	0.8221455	0.7766276	0.8701989
gendernot_reported	0.8301395	0.6672735	1.0268661
ethnicity_Non_HispanicTRUE	0.8730831	0.8217626	0.9277738

Top 10 Variables by Odds Ratios

Variable	Odds Ratio (OR)	Lower 95% CI	Upper 95% CI
ua_epinot_reported	2.395944	2.1052525	2.729261
ua_epismall	2.175929	1.9465410	2.435398
ua_epinegative	2.041279	1.7353770	2.401661
ua_bacteriamarked	1.706999	1.5106642	1.928692
psychiatric_confusion1	1.683237	1.4529255	1.947886
ua_epiother	1.447856	0.1989411	6.589884
psychiatric_confusionnot_reported	1.426368	1.3267113	1.534411
ua_epimoderate	1.388268	1.2283681	1.570323
ua_bacteriamany	1.379587	1.2676226	1.501389
ua_bacteriamoderate	1.375097	1.2771083	1.480446

1. Interpretability:

- Understanding of Variable Impact: The odds ratios (OR) provide insight into how different clinical variables influence the likelihood of a specific outcome. For instance, variables such as "ua_epinot_reported" (OR = 2.40) and "ua_epismall" (OR = 2.18) significantly increase the likelihood of the event occurring, while variables like "ua_wbcnegative" (OR = 0.12) and "ua_wbcnot_reported" (OR = 0.06) strongly decrease it. The interpretability of these results allows practitioners to understand which features contribute most to the outcome and how they affect predictions.
- Clinical Relevance: Understanding which features drive model predictions is essential for translating statistical findings into clinical practice. The top variables, such as urinary bacteria presence ("ua_bacteriamarked", OR = 1.71) and psychiatric confusion ("psychiatric_confusion1", OR = 1.68), reflect clinically meaningful associations. This ensures that the model aligns with known clinical knowledge and is interpretable within the medical context. By examining odds ratios, clinicians can determine which factors are most predictive and make informed decisions based on those factors.
- Easy Human Prediction: The odds ratios for key predictors (e.g., "age" OR = 1.01) indicate that small increases in variables like age slightly raise the likelihood of the outcome, making the relationship intuitive for clinicians. Since this pattern aligns with clinical expectations, it enhances trust in the model. The straightforward, linear relationship also makes it easier for humans to interpret and predict outcomes.

2. Explainability:

- Clarifying Model Decisions: The odds ratios offer a clear explanation of how each variable contributes to the predicted probability. For example, a unit increase in the variable "ua_epinot_reported" results in a 2.40 times higher chance of the event, providing an easy-to-understand explanation for clinicians and stakeholders. This enhances the overall explainability of the model as clinicians can trace back predictions to individual features.

- Highlighting the Most Impactful Variables: By focusing on the top 10 variables with the highest odds ratios, the model's behavior is made transparent. The significant impact of variables such as "ua_bacteriamarked" (OR = 1.71) and "psychiatric_confusion1" (OR = 1.68) makes it clear which clinical factors are most influential in driving predictions, making the model easier to communicate to non-technical users.
- Post-Hoc Explanation via Feature Analysis: The feature importance table can be used as a post-hoc explanation tool, where the contributions of each variable to the final decision are presented in understandable terms. This is especially useful when explaining the model to clinicians or healthcare professionals who may not be familiar with machine learning techniques. The table breaks down each variable's impact, helping bridge the gap between the model's complex internal workings and human understanding.

3. Practical Application of Interpretability and Explainability:

- Informed Decision-Making: The odds ratios, along with their confidence intervals, equip healthcare providers with actionable insights. For instance, knowing that "ua_bacteriamarked" has a strong positive association with the event outcome allows practitioners to prioritize this feature when making diagnostic or treatment decisions, ensuring that decisions are based on the most relevant factors.
- Model Refinement: The interpretability of the odds ratios also provides feedback for further refinement. Features with very low odds ratios, such as "ua_wbcnegative" (OR = 0.12), might be reconsidered for exclusion in the model or further investigation, improving model efficiency.
- Transparency and Trust: By being transparent about the relationships between features and predictions, the model builds trust with users. Clinicians are more likely to adopt and rely on the model if they understand how and why certain factors are influencing the outcome, thus enhancing both the practical use and the credibility of the system.

4.2 SHAP Analysis

1. SHAP Summary Plot - Top 20 Features

- Interpretability: This plot visualizes the impact of the top 20 most influential features in the model's predictions and uses color bars to represent the distribution of SHAP values for each feature, where color indicates feature value (ranging from low to high). The length of each bar along the x-axis represents the spread of SHAP values, showing the variability in how much a feature influences predictions. Wider bars indicate features with a larger impact on model predictions, meaning their influence varies significantly across observations.

- Explainability: The summary plot highlights the features the model relies on most for prediction, with the color gradient in the bars indicating whether higher or lower feature values drive the prediction positively or negatively. Some features exhibit a symmetric spread of SHAP values, meaning their effect on predictions can be positive or negative depending on their value and interactions with other features. Features with the widest bars, such as "abx," have the greatest impact on predictions, even if they appear lower on the plot. In contrast, features that are crowded in the middle with narrower bars contribute little to the model's output.

- Findings: "Patid" and "abx" have the largest range of SHAP values, indicating that their impact on predictions varies widely across individuals. This suggests strong interaction effects and a significant contribution to model uncertainty. In contrast, "ua_wbc" (White Blood Cell Count) is also highly influential but has a narrower SHAP value range, meaning its effect on predictions is more consistent. Higher "ua_wbc" values generally increase risk, while lower values decrease it. Some features exhibit mixed effects, suggesting possible non-linear relationships or interactions. Given their impact, the most important features (particularly "patid" and "abx") should be further examined to understand their clinical or predictive significance and how they drive variability in predictions.

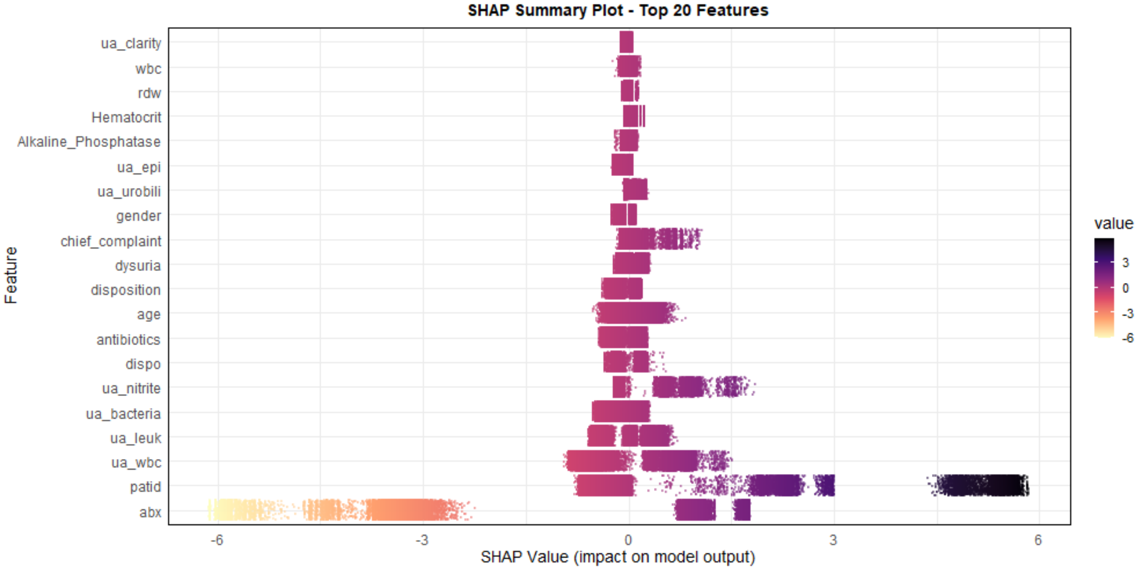

Figure 10: SHAP Summary Plot - Top 20 Features

2. SHAP Feature Importance - Top 20 Features

- Interpretability: This bar plot ranks the top 20 features based on their mean absolute SHAP values, reflecting their overall contribution to model predictions. Unlike the summary plot, this visualization does not show directionality but instead provides an aggregate view of feature importance. The higher a feature appears in the ranking, the greater its average impact on model output, making it crucial for prediction consistency.

- Explainability: The feature importance plot allows for direct comparison of the relative importance of different variables in the model. Unlike regression coefficients, SHAP values account for complex interactions, ensuring that features with significant but non-linear effects are properly weighted. This visualization answers the question: *"Which features contribute the most to model predictions?"* and helps prioritize variables for further analysis.

- Findings: The results align with the SHAP summary plot. "abx" ranks as the most important feature, having the highest mean absolute SHAP value, meaning it has the largest average impact on model output magnitude. "patid" follows as the second most critical feature, contributing substantially to prediction variability. "ua_wbc" (White Blood Cell Count) is the third most influential, with its effect being more consistent across observations. The ranking of these features provides a data-driven foundation for prioritizing variables in future modeling, interpretability studies, and clinical evaluations.

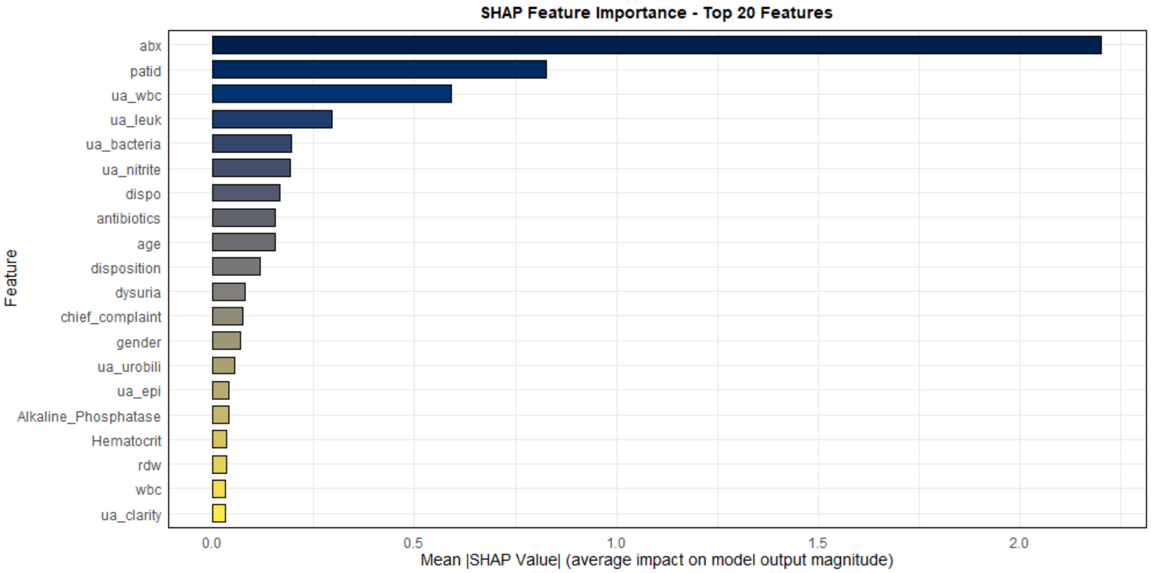

Figure 11: SHAP Feature Importance - Top 20 Features

3. SHAP Dependence Plot for White Blood Cell Count (ua_wbc)

- Interpretability: This dependence plot visualizes the relationship between White Blood Cell Count (ua_wbc) and its SHAP value, illustrating how changes in ua_wbc influence model predictions. Each point represents an individual instance, and the trend line (loess smoothing) captures the overall pattern of feature impact. The spread of points at each ua_wbc level indicates the presence of interactions with other features.

- Explainability: The plot reveals that the effect of ua_wbc on predictions is non-linear. From ua_wbc values of approximately 1 to 4, the smoothed trend line shows a downward pattern, suggesting that lower ua_wbc levels increase the model's predicted risk. Between

4 and 6, the trend stabilizes, forming a nearly horizontal line with a slight upward curvature at the end, indicating minimal but slightly increasing influence on predictions at higher ua_wbc values. The variation in SHAP values at the same ua_wbc level suggests interactions with other variables, influencing the magnitude of its effect.

- Findings: The non-linear relationship between ua_wbc and SHAP values highlights a threshold effect, where lower ua_wbc levels (1–4) have a stronger negative impact on predictions, while values above 4 exhibit a more stable influence. This pattern suggests that ua_wbc's predictive role may depend on interactions with other features, warranting further investigation into its clinical or modeling significance.

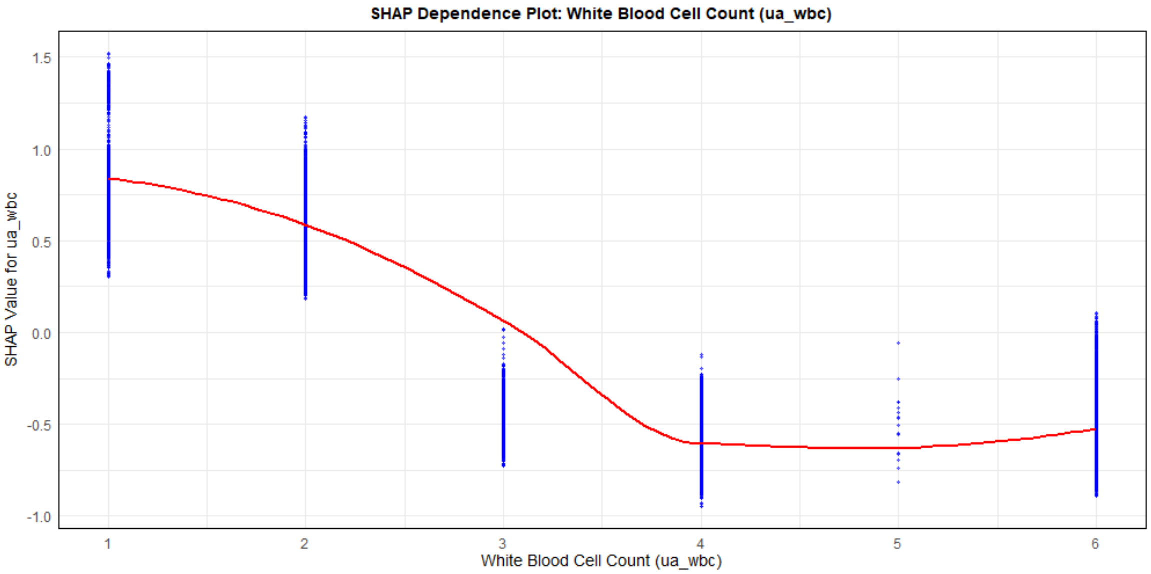

Figure 12: SHAP Dependence Plot - White Blood Cell Count