Evaluating Machine Learning Models for Heart Disease Diagnosis

Zhengyang Fei

CHL5230: Applied Machine Learning for Health Data

Dalla Lana School of Public Health

UNIVERSITY OF TORONTO

March 16, 2025

Contents

Section 1: Introduction

Heart disease remains one of the leading causes of mortality worldwide. Studies have shown that early and accurate diagnosis of heart disease is crucial for timely intervention and improved patient outcomes. With the increasing availability of data, machine learning techniques have emerged as valuable tools for enhancing diagnostic accuracy by identifying patterns within complex datasets.

In this study, we focus on the classification of heart disease using a subset of the Cleveland Heart Disease dataset from the University of California, Irvine Machine Learning Repository. We will apply three different classification methods: logistic regression with regularization, random forests, and gradient boosting machines, and then evaluate and compare their predictive performance.

The primary objective of this analysis is twofold: (a) to assess the classification performance of the three selected models using appropriate evaluation metrics, and (b) to determine the most effective model for classifying future observations. Model performance will be evaluated using a validation set approach. Given the clinical significance of false negatives in heart disease diagnosis, special consideration will be given to strategies for handling imbalanced misclassification costs.

Section 2: Methodology

2.1 Data Processing

The Cleveland Heart Disease dataset used in this study [1] comprises 14 attributes, including 13 predictor variables and one class variable that denotes the presence or absence of heart disease. The dataset was examined for missing values, and a complete case analysis was performed. Categorical variables including sex, cp, fbs, restecg, exang, slope, ca, and thal were encoded as factors, as specified in the dataset documentation. The outcome variable, num, was converted into a binary format for classification.

Variable Name	**Role**	**Type**	**Description**
age	Feature	Integer	Age
sex	Feature	Categorical	Sex
cp	Feature	Categorical	Chest pain type
trestbps	Feature	Integer	Resting blood pressure (on admission)
chol	Feature	Integer	Serum cholesterol
fbs	Feature	Categorical	Fasting blood sugar > 120 mg/dl
restecg	Feature	Categorical	Resting electrocardiographic results
thalach	Feature	Integer	Maximum heart rate achieved
exang	Feature	Categorical	Exercise-induced angina
oldpeak	Feature	Integer	ST depression induced by exercise relative to rest
slope	Feature	Categorical	Slope of the peak exercise ST segment
ca	Feature	Integer	Number of major vessels (0-3) colored by fluoroscopy
thal	Feature	Categorical	Thalassemia type
num	Target	Integer	Diagnosis of heart disease

Table 1: Summary of Variables in the Cleveland Heart Disease Dataset

2.2 Model Selection and Training

Three classification models were employed in this study:

- Logistic Regression with Lasso (L1) Regularization.
- Random Forest
- Gradient Boosting Machines

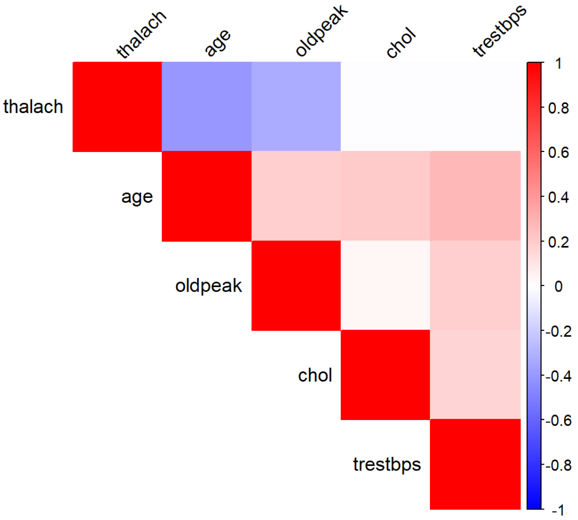

Figure 1: Correlation plot: None of the variables are very correlated (ie. cor value < 0.7)

For logistic regression, the Lasso (L1) regularization approach was selected for its ability to perform feature selection by shrinking some coefficients to zero. Another reason is that we might want to exclude variables which have high correlation. The correlation matrix is given in figure 1. The model matrix was created using one-hot encoding for categorical variables, ensuring compatibility with the regression framework. The optimal regularization parameter (lambda) was determined using 100-fold cross-validation within the training set. This process is repeated for each loop, hence giving us 5 lambda values, of which we take the mean of them. The best model was then used to make probability predictions on the test set, and class labels were assigned based on a threshold of 0.5.

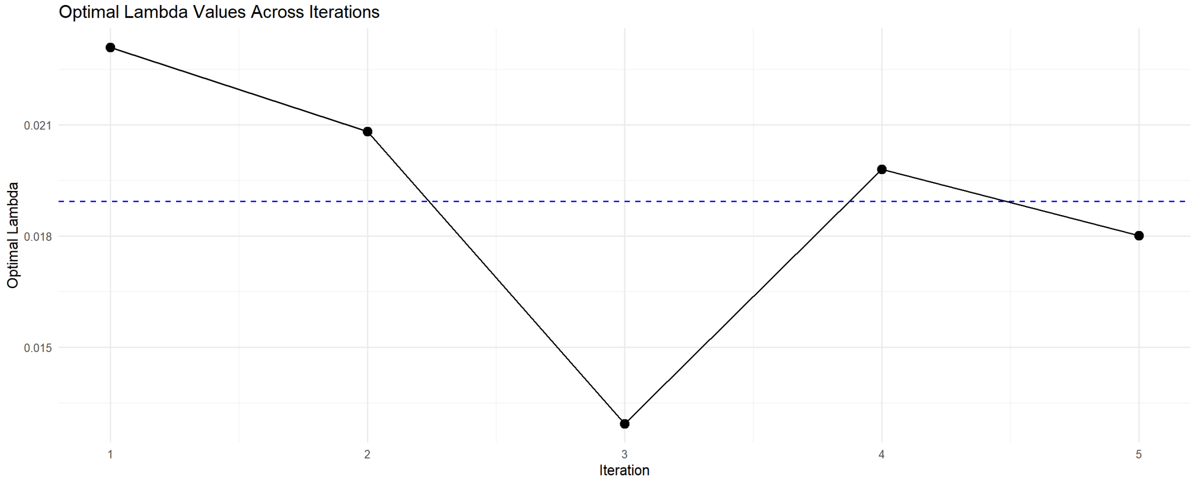

Figure 2: Optimal Lambda Values Across 5 Iterations of 5-CV Lasso Regression

Random forest classification was performed using an ensemble of 500 decision trees. To optimize the model, the mtry parameter (number of features randomly selected at each split) was tuned using grid search over the values 3, 5, 7. These values were chosen because they are close to the square root of 13 (the number of predictor variables) and 13/2, which are rules commonly used for selecting mtry in random forest models. Cross validation with 100 folds was applied within the training set to determine the best-performing model. Predictions were generated using the trained model, and probability scores were used to calculate performance metrics.

Note that several hyperparameters were not tuned and were set to their default values to balance computational efficiency and model stability, as well as because they were not covered in the lecture. For instance, the number of trees (ntree) was fixed at 500 to reduce variance while maintaining reasonable runtime. The minimum number of samples required in terminal nodes (nodesize) and the maximum number of terminal nodes per tree (maxnodes) were left at their defaults, allowing trees to grow without predefined restrictions. Ideally, tuning these parameters could further reduce the model's standard deviation and improve stability by better controlling tree depth and overall complexity.

The gradient boosting model was trained using 5000 boosting iterations, with a fixed shrinkage factor of 0.01 and an interaction depth of 3 to control tree complexity and reduce overfitting. 100-fold cross-validation was applied during training to determine the optimal number of trees, which was selected using gbm.perf(). Predictions were generated using the optimized tree count, and class probabilities were used to evaluate model performance.

2.3 Performance Evaluation

To evaluate model performance, the dataset was split into training and validation sets using stratified random sampling. The classification model were assessed based on the following key evaluation metrics:

- Accuracy - Provides an overall measure of correctness by computing the proportion of correctly classified instances. This metric was chosen because of it is simple and intuitive. A draw back is that it may not be reliable for imbalanced datasets, where a model could achieve high accuracy by predicting the majority class most of the time.

- AUC-ROC (Area Under the Receiver Operating Characteristic Curve) - Measures the model's ability to distinguish between classes across different classification thresholds. It is particularly useful in this case as it evaluates the trade-off between sensitivity and specificity, helping to assess model discrimination power independently of a specific threshold.

- Sensitivity and Specificity – Sensitivity (true positive rate) is prioritized in this study to minimize false negatives, as failing to diagnose a patient with heart disease could have serious consequences. Specificity (true negative rate) ensures the model does not falsely classify healthy individuals as having heart disease. These measures are crucial in clinical applications where the cost of false negatives is high.

- Brier Score - Evaluates the calibration of predicted probabilities by computing the mean squared difference between predicted probabilities and actual outcomes. Unlike AUC-ROC, which measures discrimination, the Brier Score assesses how well-calibrated the probability predictions are.

To account for randomness in data splitting, the validation process was repeated five times with different random splits, and the average performance of each model was recorded. A summary table was created to display the mean and standard deviation of each evaluation metric across five iterations.

Section 3: Results

3.1 Comparison of classification performance

The results of the five-fold validation are summarized below in Table 1. The results of the five-fold validation experiment indicate that all three models demonstrated strong classification performance, with slight differences in their strengths and weaknesses. Random Forest achieved the highest AUC score (0.9218 ± 0.0315), marginally outperforming Gradient Boosting (0.9203 ± 0.0307) and Logistic Regression (0.9201 ± 0.0274). This suggests that all three models are effective at distinguishing between positive and negative cases, with only minor differences in discrimination ability.

In terms of overall accuracy, Logistic Regression achieved the highest mean accuracy (0.8539 ± 0.0346), indicating that it made the fewest classification errors. However, accuracy alone does not fully capture the importance of correctly identifying positive cases, particularly in a medical context where false negatives carry significant consequences. When examining sensitivity, which measures the model's ability to correctly identify patients with heart disease, Gradient Boosting performed the best (0.8917 ± 0.0697). This suggests that Gradient Boosting is the most effective model for minimizing false negatives, making it a strong candidate for clinical applications where failing to diagnose heart disease could lead to serious health risks.

Specificity, on the other hand, measures the ability of the model to correctly classify healthy individuals. Logistic Regression had the highest specificity (0.8146 ± 0.0475), meaning it had the lowest false positive rate. This is beneficial in reducing unnecessary follow-up tests or treatments for individuals who do not actually have heart disease. However, the trade-off between sensitivity and specificity must be carefully considered.

Another important consideration is probability calibration, which was assessed using the Brier Score. A lower Brier Score indicates that a model's predicted probabilities closely reflect actual outcomes. Logistic Regression (0.1144 ± 0.0230) and Gradient Boosting (0.1156 ± 0.0218) both demonstrated strong probability calibration, whereas Random Forest had a significantly higher Brier Score (1.1427 ± 0.1015). This suggests that although Random Forest performs well in classification tasks, its probability estimates are poorly calibrated and may be overconfident or unreliable for clinical decision-making.

Metric	Logistic Regression	Random Forest	Gradient Boosting
Mean AUC (SD)	0.9201(0.0274)	0.9218(0.0315)	0.9203(0.0307)
Mean Accuracy (SD)	0.8539(0.0346)	0.8404(0.0547)	0.8404(0.0408)
Mean Sensitivity (SD)	0.8875(0.0700)	0.8667(0.1048)	0.8917(0.0697)
Mean Specificity (SD)	0.8146(0.0475)	0.8098(0.0631)	0.7805(0.0751)
Mean Brier Score (SD)	0.1144(0.0230)	1.1427(0.1015)	0.1156(0.0218)

Table 2: Performance Metrics for Different Models

The boxplot comparisons of the evaluation metrics further illustrate the distribution of AUC, accuracy, sensitivity, and specificity across five runs for each model. The AUC distributions confirm that all three models have comparable performance in distinguishing between classes, though Random Forest shows slightly higher variability. Logistic Regression has the most consistent accuracy, reinforcing its reliability in classification. Sensitivity results highlight that Gradient Boosting is consistently higher, aligning with its effectiveness in detecting positive cases. Conversely, specificity results confirm that Logistic Regression maintains the best performance in correctly identifying negative cases. Random forest shows the higher brier scores, suggesting poorer calibration.

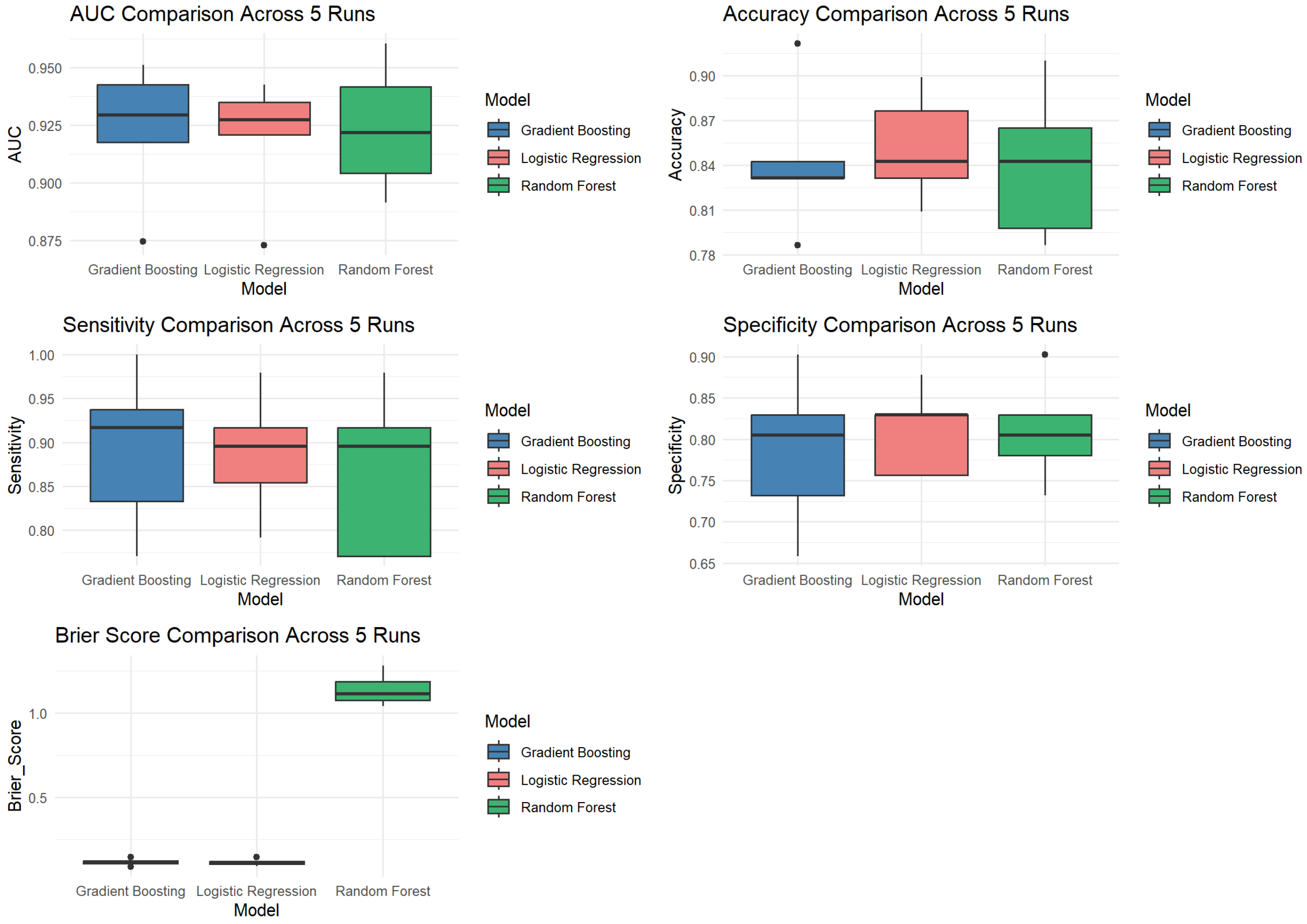

Figure 3: Barplots of AUC, Accuracy, Sensitivity, and Specificity for the three models across 5 runs

A closer examination of the results from each run reveals that no single model consistently outperforms the others across all metrics, but distinct patterns emerge. All three models perform comparably in terms of AUC, though Random Forest exhibits greater fluctuation across iterations. Logistic Regression demonstrates the highest consistency in classification accuracy and excels in specificity, making it the most reliable for minimizing false positives. In contrast, Gradient Boosting is the best choice for minimizing false negatives, as it maintains high sensitivity across all five iterations. Finally, Logistic Regression and Gradient Boosting produce more reliable probability estimates compared to Random Forest, which struggles with calibration. These findings suggest that the best model depends on the specific clinical priority: whether it is minimizing false negatives or optimizing overall classification reliability.

Iteration	**Model**	**AUC**	**Accuracy**	**Sensitivity**	**Specificity**	**Brier Score**
Gradient Boosting						
1	Gradient Boosting	0.9446	0.8427	0.9792	0.6829	0.1070
2	Gradient Boosting	0.9268	0.8202	0.8333	0.8049	0.1166
3	Gradient Boosting	0.9101	0.8202	0.8958	0.7317	0.1238
4	Gradient Boosting	0.9477	0.9101	0.9375	0.8780	0.0857
5	Gradient Boosting	0.8725	0.8089	0.8125	0.8049	0.1451
Logistic Regression						
1	Logistic Regression	0.9426	0.8764	0.9792	0.7561	0.1027
2	Logistic Regression	0.9273	0.8427	0.8542	0.8293	0.1106
3	Logistic Regression	0.9228	0.8427	0.8958	0.7805	0.1177
4	Logistic Regression	0.9346	0.8989	0.9167	0.8780	0.0899
5	Logistic Regression	0.8730	0.8089	0.7917	0.8293	0.1513
Random Forest						
1	Random Forest	0.9583	0.8764	1.0000	0.7317	1.2911
2	Random Forest	0.8951	0.7865	0.7708	0.8049	1.0709
3	Random Forest	0.9207	0.8427	0.8958	0.7805	1.1904
4	Random Forest	0.9482	0.9101	0.8958	0.9024	1.0356
5	Random Forest	0.8869	0.7865	0.7500	0.8293	1.1253

Table 3: Model Performance Over 5 Iterations

3.2 Feature Selection

We now analyze feature importance for both the random forest and gradient boosting models. The result for both models are displayed in figure 4.

Notably, both models identify thal, ca, and cp as the top three most important features, though their ranking differs. Additionally, variables such as oldpeak, thalach, age, chol, and trestbps consistently hold significant importance across both models. In contrast, fbs, restecg, slope, sex, and exang are deemed less influential. Based on these findings, we will proceed by retaining thal, ca, cp, thalach, age, oldpeak, chol, and trestbps as predictors while excluding sex, fbs, restecg, and slope from the model.

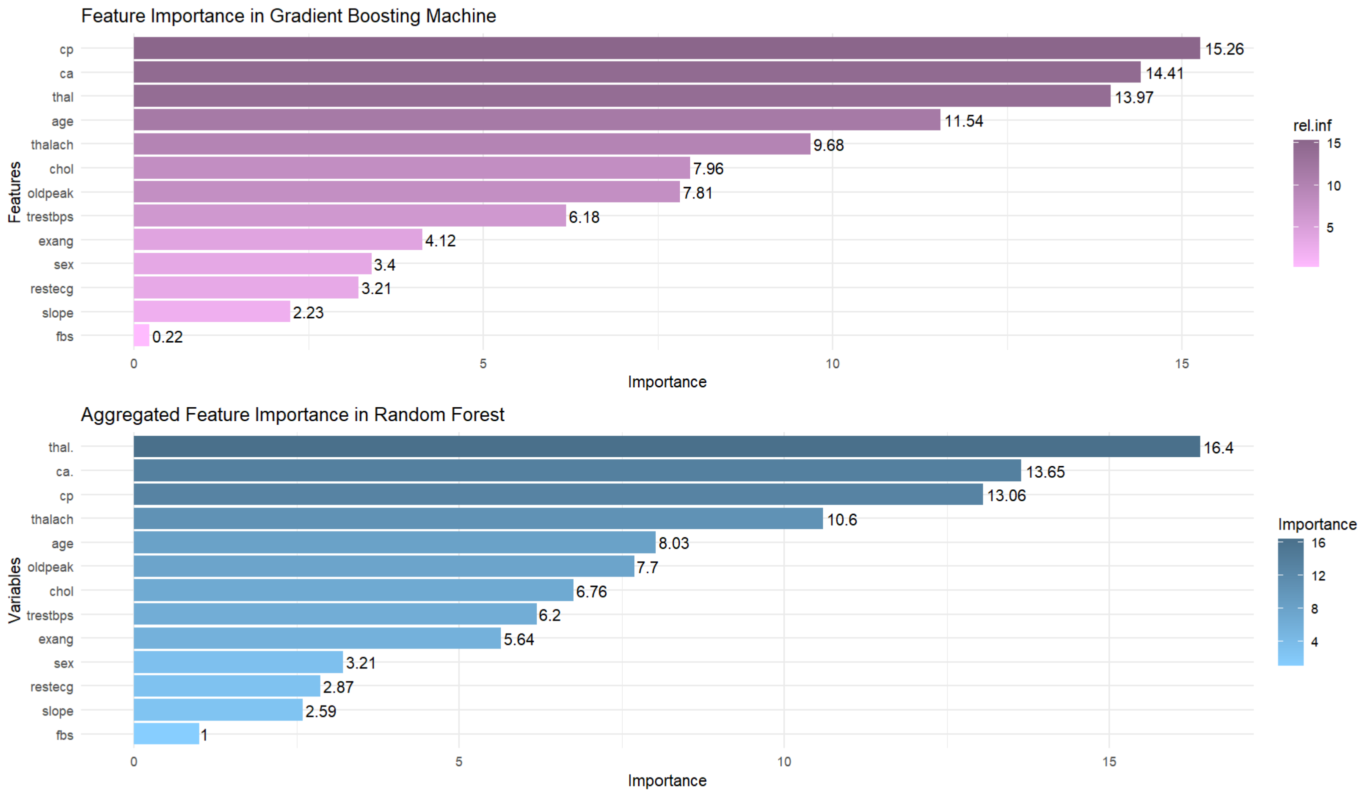

Figure 4: Feature Importance for Gradient Boost Model

3.3 Model for Future Classification

Based on the above analysis of model performance across five iterations, gradient boosting emerges as the most suitable model for future classification tasks. It is particularly suitable in a clinical setting where correctly identifying heart diseases cases is most important.

The conclusion is drawn from the following observations:

- Highest mean sensitivity (0.8917): it is most effective at correctly identifying patients with heart disease. Given the serious consequences of false negatives in medical diagnosis, a model with higher sensitivity is preferred

- High mean AUC (0.9203): while random forest has slightly higher AUC, its poor calibration (high Brier score) makes it less reliable for decision making

- Balanced tradeoff between sensitivity and specificity: While logistic regression has the highest specificity (0.8146), this came as a tradeoff for a worse sensitivity. Gradient boosting offers a slightly more reasonable balance.

In addition, we will fit the model using the predictors described in the previous section. This process is done in R. Here are the key metrics for the final model.

Metric	Gradient Boosting Score
AUC	0.877
Accuracy	0.809
Sensitivity	0.8125
Specificity	0.8049
Brier Score	0.1406

Table 4: Performance Metrics for the Tuned Gradient Boosting Model

3.4 Partial Dependence Plots

Furthermore, we can look at the partial dependence plots among the continuous variables in Figure 5. Note that we assume independence among the variables.

The plot suggests that Oldpeak and thalach appear to be the strongest predictors. Age and cholesterol level have more complex non linear effect. Lastly, resting blood pressure shows a threshold effect.

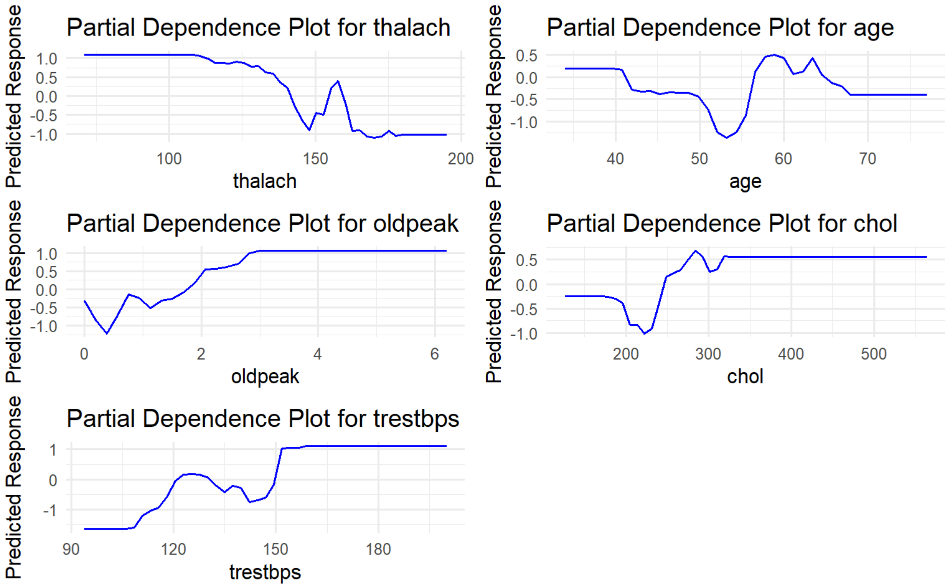

Figure 5: Partial Dependence Plots for Continuous Variables

Section 4: Discussion and Conclusion

This study compared Logistic Regression with Lasso Regularization, Random Forest, and Gradient Boosting Machines (GBM) for heart disease classification. All three models demonstrated strong performance, with Random Forest exhibiting the highest AUC but poor probability calibration, Logistic Regression achieving the highest specificity, and GBM excelling in sensitivity. Given the clinical importance of minimizing false negatives, GBM was selected as the final model due to its superior ability to identify positive cases.

In medical diagnostics, missing a heart disease diagnosis (a false negative) can have severe consequences. To mitigate this risk during model training, several strategies can be employed. One approach is threshold adjustment, where the default probability threshold of 0.5 is lowered to enhance sensitivity, increasing the likelihood of correctly identifying positive cases. Another method is cost-sensitive learning, which assigns a higher penalty to false negatives in the loss function, encouraging the model to prioritize sensitivity over overall accuracy. Additionally, human consultation remains valuable in ensuring accurate and informed decision-making.

References

[1] UCI Machine Learning Repository. *Heart Disease Dataset*. 2024. URL: https://archive.ics.uci.edu/dataset/45/heart+disease.