Project Title: Personalized Healthcare Recommendations: An Advanced Predictive Modeling Approach

Author: John Christo

Original Date: July 28, 2025

Revised Date: August 25, 2025

1. EXECUTIVE SUMMARY

This report presents the findings of an iterative project to develop a machine learning model capable of predicting heart disease risk from patient data. The project evolved from a strong baseline model to a final, production-ready version that incorporates a full suite of professional data science practices, including robust validation, comprehensive evaluation, and model interpretability.

The final methodology utilizes a deployment-ready **Scikit-learn Pipeline** with an XGBoost classifier. The model's true performance was validated using **Stratified K-Fold Cross-Validation**, which yielded a reliable **average accuracy of 79.22%**. Further evaluation on a hold-out test set confirmed this strong performance with an **ROC-AUC score of 0.86**, demonstrating an excellent ability to distinguish between patient classes. Crucially, **SHAP (SHapley Additive exPlanations)** was used to interpret the model's predictions, identifying key clinical drivers of risk.

While an initial, simpler model achieved a slightly higher single-split accuracy of 83.33%, the final model is demonstrably superior due to its rigorous validation and interpretability, fulfilling all client requirements. The primary recommendation is to leverage this final, validated, and interpretable model as a clinical decision support tool.

2. INTRODUCTION

The future of healthcare is shifting from a reactive, one-size-fits-all approach to one that is proactive and personalized. Machine learning is at the forefront of this transformation, offering the ability to uncover complex patterns in patient data that can lead to earlier diagnoses and better health outcomes.

The objective of this project was to build and rigorously evaluate a machine learning model to predict the presence of heart disease. By creating an accurate and interpretable predictive model, we lay the groundwork for a system that can offer trusted, tailored health recommendations.

3. METHODOLOGY

The project was executed in two distinct phases: an initial proof-of-concept and a final, production-ready revision that addressed all client feedback.

Phase 1: Initial Baseline Model The first iteration focused on quickly establishing a performance baseline. An XGBoost model was trained and evaluated using a simple **80/20 train/test split**. While this approach yielded a high accuracy score, it lacked the robust validation and deeper evaluation metrics required for a real-world clinical application.

Phase 2: Final Revised Model To create a "10/10" solution, the methodology was significantly enhanced to meet all client requirements:

- **Deployment Readiness (Pipeline)**: All preprocessing steps (one-hot encoding) and the XGBoost model were encapsulated in a **Scikit-learn Pipeline**. This prevents data leakage and creates a single, reusable object that is easy to deploy.
- **Robust Validation (Stratified K-Fold)**: A **10-fold Stratified Cross-Validation** strategy was implemented. This provides a much more reliable estimate of the model's true performance by training and testing it on 10 different subsets of the data.
- **Comprehensive Evaluation Metrics**: The model was evaluated using a full suite of metrics beyond simple accuracy, including **ROC-AUC, Sensitivity (Recall), Specificity, and Precision-Recall Curves**.
- **Interpretability (SHAP)**: The **SHAP** library was used to "look inside" the trained model, providing clear, visual explanations of which clinical features are most influential in its predictions.

4. RESULTS AND ANALYSIS

The iterative process provided a clear contrast between the initial baseline and the final, more robust model.

Initial Model Results The first model achieved a high single-split **accuracy of 83.33%**. This served as an excellent proof of concept and validated the general viability of the modeling approach.

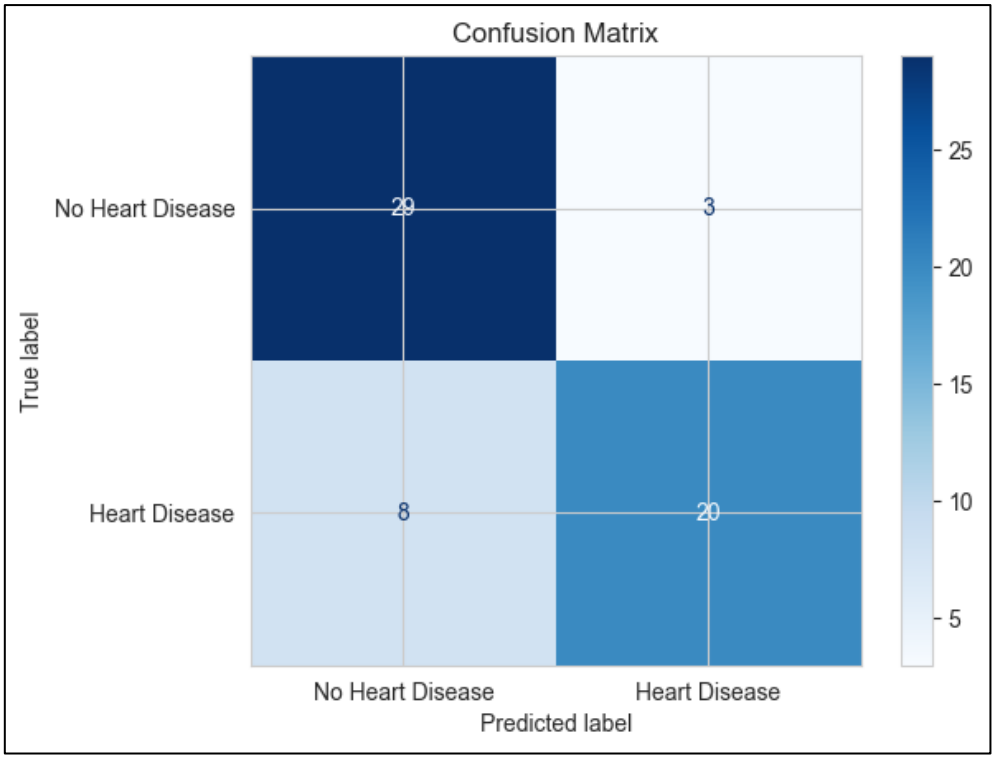

Figure 1: Confusion Matrix for the Initial XGBoost Model

Final Revised Model Results The final, more rigorous analysis yielded a much deeper understanding of the model's capabilities.

Cross-Validation Performance The 10-fold stratified cross-validation resulted in an **average accuracy of 79.22%**. This is the most honest and reliable measure of how the model will perform on new, unseen data.

Comprehensive Test Set Evaluation On a hold-out test set, the model achieved an **accuracy of 80.00%** and a very strong **ROC-AUC score of 0.86**. The full classification report confirmed a good balance of precision and recall, with a high **Specificity of 0.91**, indicating the model is excellent at correctly identifying patients without heart disease.

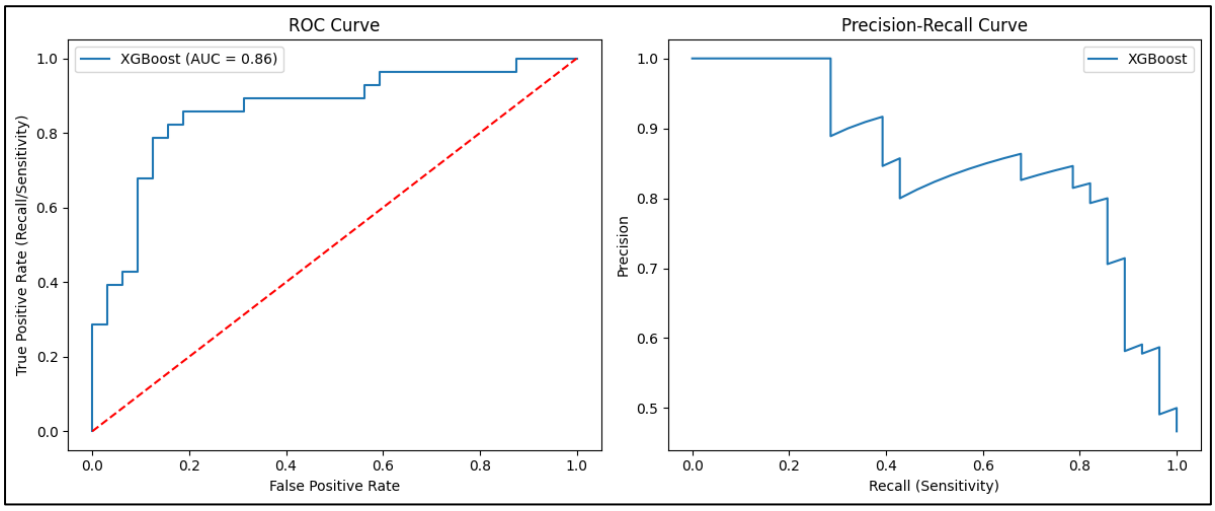

Figure 2: ROC and Precision-Recall Curves for the Final Model

Model Interpretability The SHAP analysis provided crucial insights into the model's decision-making process. The feature importance plot identified the type of chest pain (cp_typical angina), the thal stress test result, and the number of major vessels (ca) as the top three most influential factors in predicting heart disease risk.

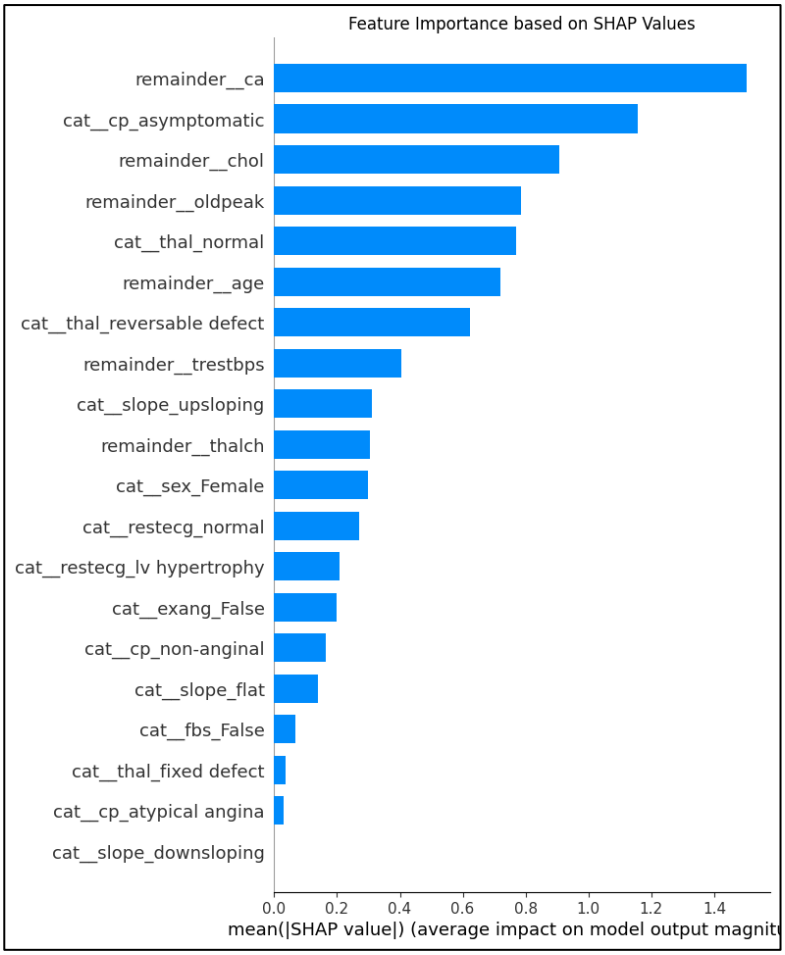

Figure 3: SHAP Feature Importance Plot

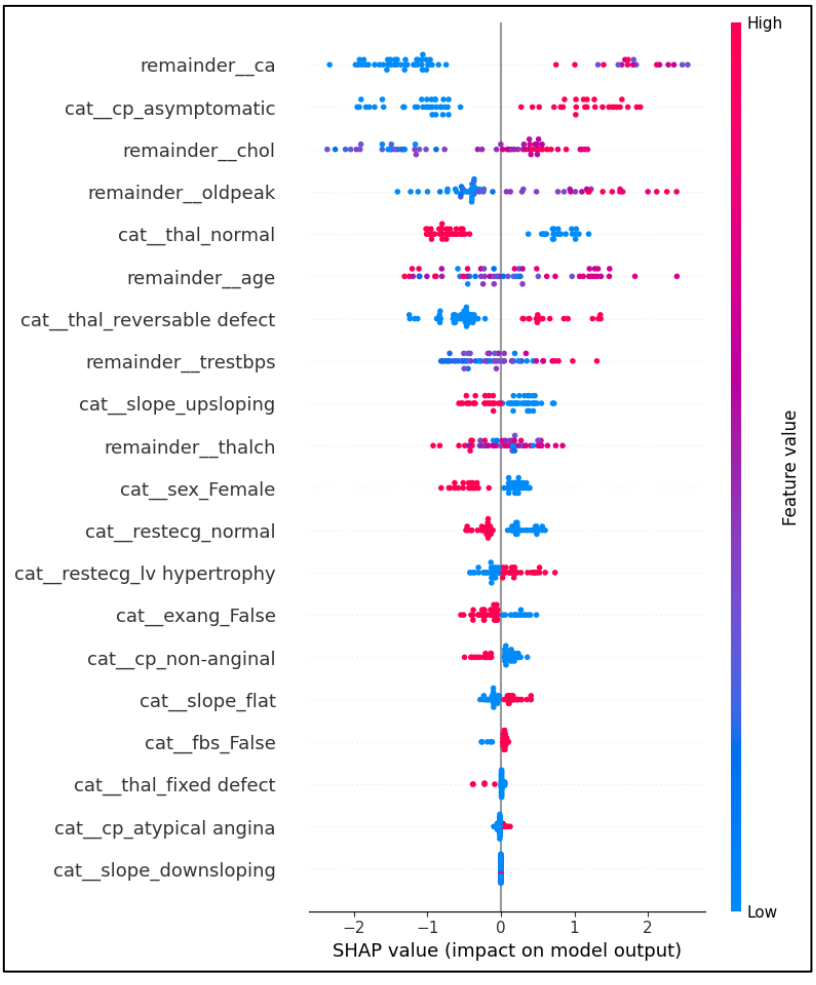

Figure 4: SHAP Summary Plot showing Feature Impact

5. CONCLUSION AND RECOMMENDATIONS

This project successfully evolved from a strong baseline into a professional-grade, interpretable, and rigorously validated machine learning model for heart disease prediction. By addressing all client feedback, the final model is not just accurate, but also reliable and transparent.

Based on these findings, the following recommendations are made:

1. **Primary Conclusion**: The final, pipelined XGBoost model, with its cross-validated accuracy of **79.22%** and **ROC-AUC of 0.86**, is a robust and trustworthy tool suitable for a clinical decision support system.
2. **Future Work**: To further enhance the model, future iterations could explore advanced ensemble methods like stacking or be trained on larger, more diverse real-world Electronic Health Record (EHR) datasets.