Pre-Read Report for Heart Health Risk Intervention Program

Presented by: Parag Subhash Padekar
Prepared for: Healthy Hearts Inc.

Executive Summary

This report examines the pressing issue of heart health risks among patients in the United States, where heart disease remains the leading cause of death. Alarmingly, nearly half of the population may be at risk without their knowledge, underscoring the urgent need for predictive health interventions. By utilizing demographic and medical data, we evaluated three classifiers K-Nearest Neighbours (KNN), Random Forest, and Logistic Regression to determine their effectiveness in accurately predicting heart disease risk levels and to assess the potential return on investment for Healthy Hearts Inc. The Random Forest classifier stood out as the most beneficial, projecting a net profit of $22,400 across 283 patients, which translates to $128 per patient. This method, with further enhancements in data quality, offers not only improved predictive accuracy but also actionable insights that could significantly enhance patient outcomes and business growth.

1. Original Request and Proposed Approach

Healthy Hearts Inc. requested the development of a predictive model to classify patients into five heart health risk categories based on demographic and medical data. The goal was to implement a proactive intervention strategy that identifies individuals at elevated risk for heart disease, allowing for timely care while ensuring profitability.

Proposed Approach:

- **Data Collection:** Gathered demographic and medical parameters, including age, cholesterol levels, peak heart rate, and blood pressure, from a sample of 283 patients.
- **Cost Model:** Each accurate prediction is estimated to yield a net profit of $500, with an average profit of $128 per patient across the dataset.

2. Dataset Overview

The dataset comprises various demographic and medical attributes categorized into five health risk classes: "No Risk," "Slight Risk," "Moderate Risk," "High Risk," and "Extreme Risk." It is important to note that the data is imbalanced, with a predominance of the "No Risk" class, which affects the model's accuracy for the other classes.

3. Data Analysis and Classifier Selection

Following a thorough exploratory analysis, we chose K-Nearest Neighbours, Random Forest, and Logistic Regression classifiers based on their unique strengths in classification tasks. Data preprocessing steps included:

- **Data Balancing:** Addressed class imbalance through resampling techniques.
- **Data Normalization:** Ensured consistent scaling across continuous variables for optimal model performance.
- **Feature Importance Analysis:** Employed Random Forest to identify critical predictors.

4. Classifier Analysis and Results

Random Forest was identified as the top-performing model, striking a balance between accuracy and net benefit. It exhibited consistent precision in the “No Risk” category and yielded higher net benefits compared to the other classifiers.

Evaluation Metrics:

- **Accuracy:** Average of 63%
- **Net Benefit:** Highest at $22,400
- **Strengths:** High precision in the predominant “No Risk” class and interpretable feature importance.
- **Limitations:** Lower accuracy for minority classes, indicating room for improvement with more balanced data.

Classifier Comparison:

- **Random Forest:** Strong accuracy and net benefit; best suited for the current dataset.
- **KNN:** Moderate accuracy with slightly lower net benefit.
- **Logistic Regression:** Low accuracy due to complexity and class imbalance issues.

5. Recommendations and Future Steps

Based on its performance, we recommend implementing the Random Forest model for initial deployment. To enhance predictive capabilities:

- **Enhanced Data Collection:** Increase sample sizes in minority classes, particularly for "High" and "Extreme Risk."

- **Periodic Model Evaluation:** Deploy the Random Forest model with regular assessments to adapt to real-time data changes.
- **Extended Feature Set:** Consider incorporating additional health metrics, such as lifestyle choices and family history, to improve predictive power.

Conclusion and Next Steps

The proposed model provides Healthy Hearts Inc. with a practical tool for early detection of heart disease risk, focusing on both patient care and profitability. Future research should delve into advanced feature extraction techniques and real-time model calibration to boost accuracy and extend benefits to higher-risk groups. Thank you for the opportunity to present these findings and potential next steps.