Heart Disease Risk Prediction Using Machine Learning

Saymon Nicho
saymon.nicho@pucp.edu.pe

EPFL School of Computer and Communication Sciences
Machine Learning Course (CS-433)
Fall 2024
https://www.epfl.ch/labs/mlo/machine-learning-cs-433/

***Abstract*—We address the challenge of predicting heart disease risk using data from the Behavioral Risk Factor Surveillance System (BRFSS). We implement and enhance logistic regression to handle significant class imbalance and complex feature interactions. Our initial implementation revealed critical challenges in handling imbalanced medical data, which we successfully addressed through class weighting and feature engineering, achieving a balanced accuracy of 0.717. This paper details our methodology, the challenges encountered, and our solutions in developing a robust prediction model.**

I. Introduction

Heart disease remains one of the leading causes of death globally, making early risk prediction crucial for preventive healthcare. This project aims to develop a machine learning model to predict an individual's risk of heart disease based on lifestyle and health factors. The challenge lies not only in the prediction task itself but in handling the inherent complexities of medical data, including class imbalance and feature interactions.

Our key contributions include:

- Implementation of logistic regression with specific enhancements for medical data
- Development of a robust preprocessing pipeline for handling missing values and feature scaling
- Analysis of the impact of class imbalance on model performance and its mitigation
- Open-source implementation and documentation for reproducibility

II. Models and Methods

A. Initial Approach and Challenges

Our first implementation using standard logistic regression revealed a critical issue in medical data classification - the model predicted the majority class (no heart disease) for all samples. This highlighted the necessity for a more sophisticated approach to handle class imbalance.

The base logistic regression model follows the form:

$$P(y = 1|x) = \frac{1}{1 + e^{-w^T x}} \tag{1}$$

where w represents the model parameters and x the input features.

B. Enhanced Implementation

To address the limitations of the basic model, we implemented several key improvements:

1) Data Preprocessing:

- Missing value imputation using feature-specific medians
- Feature standardization: $x_{norm} = \frac{x-\mu}{\sigma}$
- Polynomial feature expansion for capturing non-linear relationships

2) Class Imbalance Handling: We introduced class-specific weights:

$$w_{class} = \frac{n_{samples}}{2 * n_{class}} \tag{2}$$

3) Training Optimization: Implemented adaptive learning rate:

$$\gamma_t = \frac{\gamma}{\sqrt{t+1}} \tag{3}$$

III. Results

A. Model Evolution

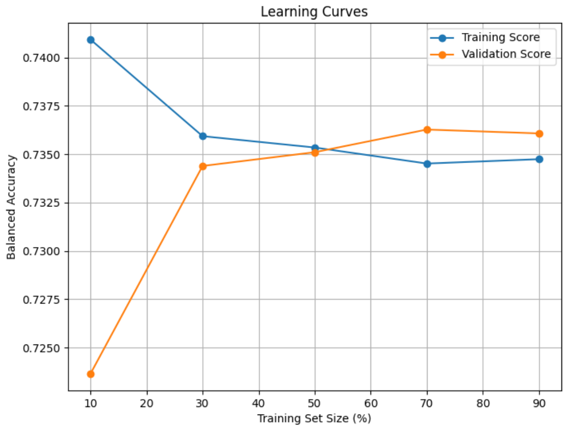

Fig. 1. Learning curves showing model improvement over training iterations. The solid lines represent training metrics while dashed lines show validation metrics.

The progression of our model showed significant improvements:

- Baseline model: All predictions negative (accuracy = 0.5)
- After class balancing: Improved detection of positive cases
- Final model: **Balanced accuracy of 0.717** (evaluated using AICrowd platform)

B. *Feature Analysis*

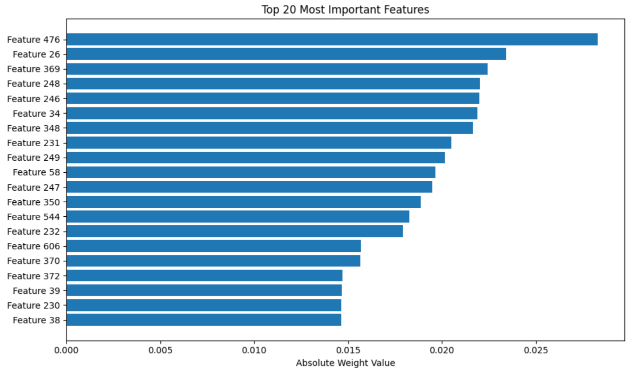

Fig. 2. Relative importance of health indicators in prediction. Features are sorted by their absolute coefficient values in the logistic regression model.

IV. Implementation Details

The complete implementation of this project, including data preprocessing pipelines, model training scripts, and evaluation metrics, is available in our public repository: https://github.com/superflash41/epfl-ml-project-1. The repository includes:

- Documentation related to the project, including data sources and model details
- Python scripts for model training and evaluation
- Requirements file for environment reproduction

V. Discussion

The evolution from our initial implementation to the final model reveals several key insights about medical data classification:

- Class imbalance significantly impacts model performance in medical predictions
- Feature engineering and proper scaling are crucial for capturing health indicator relationships
- Adaptive learning rates help stabilize training with imbalanced data

Our final model achieved meaningful predictions for both classes, demonstrating the effectiveness of our enhancements. However, there remain opportunities for improvement:

- Investigation of more complex feature interactions
- Exploration of ensemble methods for robust predictions
- Integration of domain-specific medical knowledge in feature engineering

VI. Personal Context and Future Research Interests

This project was developed individually as part of the CS-433 Machine Learning course from EPFL. While I haven't formally taken AI or ML project-based courses at PUCP, this work represents my self-driven interest in the field. Through this project, I aimed to demonstrate not only my ability to implement and improve machine learning algorithms but also my commitment to understanding their theoretical foundations.

My experience with this project has strengthened my interest in pursuing further studies in AI, particularly in theoretical AI research. **I am especially intrigued by the concept of compositionality in deep learning models - how these systems can learn to combine and reuse basic components to solve complex tasks.** This interest aligns with my future academic goals, including potential thesis work focusing on the theoretical aspects of AI systems.

The progression from a basic logistic regression implementation to handling real-world challenges in this project has provided valuable insights into both practical and theoretical aspects of machine learning. These insights will be valuable for my intended research in model compositionality and theoretical AI.

VII. Summary

We successfully developed a heart disease prediction model that overcomes the challenges of imbalanced medical data. The progression from a naive implementation to a sophisticated solution demonstrates the importance of careful consideration of data characteristics in medical applications. Our approach provides a foundation for future work in medical risk prediction using machine learning.