Predicting Heart Disease Risk

Fabian Roulin, Léa Goffinet, Samuel Mouny
EPFL, Switzerland

Abstract—**Cardiovascular diseases (CVD) are an increasing global health issue, making early detection and prevention crucial. This project utilizes fundamental machine learning techniques to evaluate an individual's risk of developing myocardial infarction and coronary heart disease (MICHD) based on lifestyle factors. By using data from the Behavioral Risk Factor Surveillance System (BRFSS), our goal is to develop a predictive model for MICHD. This tool could be vital for early intervention.**

I. Introduction

Cardiovascular diseases (CVD) are a leading cause of mortality worldwide [**?**], making early detection and prevention crucial. This project utilizes fundamental machine learning (ML) techniques to evaluate an individual's risk of developing myocardial infarction and coronary heart disease (MICHD) based on lifestyle factors.

Using data from the Behavioral Risk Factor Surveillance System (BRFSS) [1], our goal is to develop a predictive model for MICHD. Such a tool could be vital for early intervention and reducing the global burden of CVD.

II. Models and Methods

A. Exploratory Data Analysis and Feature Processing

An initial step in designing an ML model is examining the available data. The dataset comprises approximately 330,000 samples with 74 features, including both numerical and categorical data. During data exploration, we observed a significant class imbalance, with about 90% healthy and 10% MICHD samples.

To handle the complex nature of the features, we created a JSON metadata dictionary based on the BRFSS codebook [1]. This metadata includes feature types (categorical or numerical), subtypes (e.g., discrete, nominal, binary), encoding methods, and invalid values treated as NaN. Features irrelevant to prediction, such as interview-related data, were flagged for exclusion.

Feature Selection and Processing: We initially considered removing features with a high percentage of missing values (over 90%) and those flagged for exclusion. We also analyzed feature correlations to identify and remove redundant features. Pearson correlation and mutual information were used to assess the relevance of features to the target variable.

Despite these efforts, feature selection did not significantly improve model performance over using the full dataset. Therefore, we proceeded with all remaining features after cleaning. Missing numerical values were imputed with the mean, and categorical values were imputed with the mode. Nominal features were one-hot encoded, and ordinal features were appropriately mapped if needed.

B. Goal of the Model

While maximizing prediction accuracy is a common objective, it is not sufficient in the context of imbalanced datasets. With a 90/10 class imbalance, a model predicting all samples as healthy would achieve 90% accuracy but fail to identify sick individuals.

To address this, we focused on maximizing the F1-score, which balances precision and recall, providing a better measure of a model's effectiveness on imbalanced datasets. The F1-score is defined as:

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

C. Baseline Study

We implemented and evaluated several machine learning methods to predict MICHD, including least squares regression, logistic regression, gradient descent methods, and regularized logistic regression. Each method was tested with default parameters (e.g., learning rate $\gamma = 0.01$, maximum iterations max_iters $= 1000$), and performance was compared using accuracy and F1-score. As shown in Table I, logistic regression yielded the best results, making it the focus of our further analysis.

TABLE I: Baseline Model Performance

Method	**Accuracy**	**F1-score**
Least Squares Regression	0.347	0.210
Logistic Regression	0.912	0.03
Gradient Descent Methods	0.303	0.201
Regularized Logistic Regression	0.912	0.00

D. Logistic Regression

Based on our results and other tests performed, we chose the logistic regression as our model. Logistic regression models the probability that a sample belongs to the positive class using the sigmoid function $\sigma(t) = \frac{1}{1+e^{-t}}$. To handle the class imbalance, we incorporated a class weight ω_i to penalize misclassification of the minority class more heavily. The weighted loss function we optimized is:

$$L(w) = -\frac{1}{N}\sum_{i=1}^{N} \omega_i \left[y_i \log \sigma(x_i^\top w) + (1 - y_i) \log \left(1 - \sigma(x_i^\top w)\right)\right] \tag{2}$$

where

$$\omega_i = \begin{cases} \omega_c & \text{if } y_i = 1 \\ 1 & \text{if } y_i = 0 \end{cases} \tag{3}$$

This adjustment helps the model pay more attention to the minority class during training by increasing the penalty for misclassifying positive samples.

E. Hyperparameter Optimization

To enhance model performance, we performed hyperparameter tuning using stratified 5-fold cross-validation. The hyperparameters adjusted include:

- γ: Learning rate of the gradient descent.
- λ: Regularization coefficient (we found that setting $\lambda = 0$ was optimal).
- ω_c: Class weight for the minority class.
- T: Decision threshold for classification.
- *Patience*: Number of iterations with no improvement before early stopping.

We implemented early stopping to prevent overfitting by monitoring the validation loss and stopping training when no improvement was observed for a specified number of iterations (*Patience*).

III. Results

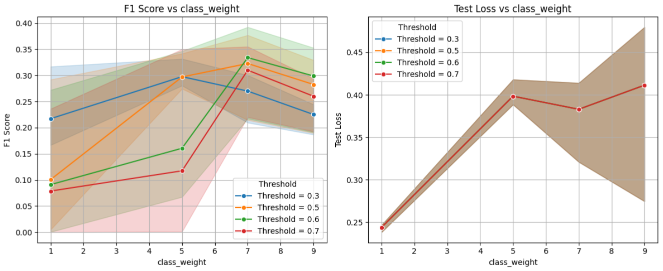

Fig. 1: F1 score and Loss vs class weight

We conducted sensitivity analyses for all hyperparameters and compared the performances of the optimized model and the baseline model. The tuned hyperparameters are shown in Table II.

TABLE II: Optimized Hyperparameters

Parameter	**Symbol**	**Value**
Learning Rate	γ	0.01
Regularization Coefficient	λ	0.001
Class Weight	ω_c	7
Threshold	T	0.6

A. Impact of Learning Rate γ

The learning rate significantly affects the convergence of the gradient descent algorithm. A too high a learning rate can cause divergence, while too low a learning rate slows convergence. We found that a learning rate of $\gamma = 0.01$ provided the best trade-off between convergence speed and stability, resulting in a higher F1-score.

B. Regularization Coefficient λ

Regularization helps prevent overfitting by penalizing large weights. We tested various values of λ (including 0, 0.001, 0.1, 1, and 10) and chose the best one. Consequently, we set $\lambda = 0.001$ in our final model.

C. Class Weight ω_c

Addressing class imbalance is crucial for our model. We adjusted the class weight ω_c to give more importance to the minority class. As depicted in Figure 1, increasing ω_c improved the F1-score up to a point. We achieved the best F1-score with $\omega_c = 7$, beyond which performance plateaued or declined.

D. Threshold T

We experimented with different decision thresholds for classification. Adjusting the threshold can affect the trade-off between precision and recall. We found that a threshold of $T = 0.6$ maximized the F1-score on the validation set.

E. Optimized Model vs. Baseline Model

Overall, these methods significantly improved the performance of our system, as shown in Table III. The baseline model is a logistic regression with default hyperparameters, no class weighting, and default threshold ($T = 0.5$).

TABLE III: Comparison of Baseline and Optimized Models

Metric	**Baseline Model**	**Optimized Model**
F1-score	0.03	0.432
Accuracy	0.912	0.859

As shown in Table III, the optimized model improved precision and recall, indicating better identification of sick individuals. The slight decrease in accuracy is acceptable since our focus is on correctly identifying the minority class.

IV. Discussion

Our approach effectively addresses the class imbalance issue through class weighting and hyperparameter optimization. The substantial improvement in F1-score demonstrates the importance of tuning model parameters and adjusting for data imbalances. However, the model may still misclassify some positive cases, and further improvements could involve more sophisticated algorithms or additional data preprocessing.

V. Summary

We developed a logistic regression model to predict the risk of MICHD using the BRFSS dataset. By performing extensive hyperparameter tuning and addressing class imbalance through class weighting, we significantly improved the model's F1-score. Our optimized model effectively balances precision and recall, making it more practical for early detection of CVD.

Future work could explore more advanced models, such as ensemble methods or neural networks, and incorporate external data sources to further enhance prediction accuracy.

REFERENCES

[1] Centers for Disease Control and Prevention, *2015 Behavioral Risk Factor Surveillance System (BRFSS) Codebook*, Centers for Disease Control and Prevention, Atlanta, Georgia, USA, 2015, accessed: 2024-11-01. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html