Predicting Cardiovascular Disease Risk Using Machine Learning

Aolaritei Valentin, De Laurentis Alberto, Le Bras Martin Louis
Machine Learning (CS-433) Project-1, EPFL, Switzerland

***Abstract*—Cardiovascular diseases (CVD), including heart attacks, are among the leading global causes of death, according to the World Health Organization. Utilizing the Behavioral Risk Factor Surveillance System (BRFSS) dataset, which captures health-related risks and behaviors of U.S. residents, we developed machine learning models for the early detection of CVDs. Our study involved implementing various pre-processing techniques and analyzing feature importance. Careful pre-processing underscored the significance of meaningful features, paving the way for further analysis in collaboration with field experts.**

I. Introduction

Cardiovascular Diseases (CVD), such as heart attacks and strokes, have been identified by the World Health Organization as a leading cause of death worldwide, with prevalence increasing as populations live longer. Early detection and prevention of CVD have become critical public health goals, and machine learning offers innovative methods to aid these efforts. This project aims to leverage machine learning techniques to assess the likelihood of individuals developing CVD, based on lifestyle and clinical risk factors.

II. Models and Methods

In this section, we describe the engineering choices for our ML system providing insights into the dataset (II-A), illustrating the pre-processing strategy (II-B), reporting how feature engineering and model selection was performed (II-C and II-D) and finally how and which hyperparameters were tuned (II-E.

A. Dataset Overview

Using data from the Behavioral Risk Factor Surveillance System (BRFSS), we analyzed records where respondents were classified as having coronary heart disease (MICHD) if they reported diagnoses of MICHD, heart attack, or angina pectoris. The dataset comprised 328,135 respondents in 322 features, categorizable in:

- binary (yes-no questions, $\approx$ 34% of the total)
- categorical (no meaningful answers order $\approx$ 18%)
- numerical (quantitative or meaningful order $\approx$ 48%)

Moreover, the dataset exhibited a strongly unbalanced target variable (8.83 % of positive cases), highlighting F_1 score would be a better metric to evaluate model performance rather than accuracy.

B. Pre-processing

Initially, the answers were mapped to meaningful values. E.g, for the "MENTHLTH" feature, values between 1 and 30 constituted the number of past days of bad mental health, while 88 indicated no such episodes. Thus, we remapped 88 to 0.

The pre-processing followed these steps:

1) Features with a missing value rate exceeding 25% were removed, because of a substantial proportion of missing values in the dataset (44.75% of total entries).
2) To refine the analysis, certain uninformative features (e.g. telephone number) and redundant interacting features were excluded to mitigate the risk of overfitting.
3) Cross-correlation was computed for the subset of features that successfully cleared our initial filtration criteria, guiding us to drastically reduce the selected features to achieve a smaller subset of low-correlated features.
4) This subset only contained 36 features divided into non-interacting features and *calculated variables* of the remaining interacting terms. (See *Figure 1*).
   It was remarkable to process differently some specific features, especially the ones concerning old people (e.g. not removing sparse features) and close related diseases (e.g diabetes).
5) To reduce the proportion of missing values and to enhance the dataset's resilience against potential outliers we decided to substitute them with the median for numerical features and with the mode for categorical one.

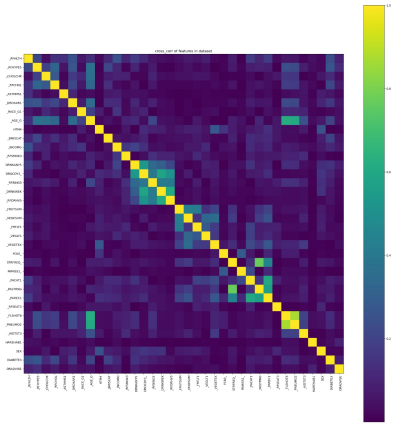

Figure 1: **HeatMap Correlation of Final Subset Features**

C. Feature Engineering

Categorical variables were encoded as $k-1$ dummy variables, where k is the number of categories, with one category omitted as the baseline to prevent collinearity with the constant term.
For numerical features, the following foresights were taken.

- To preserve the entirety of data rows we decided to winsorize features at 5% level instead of directly dropping outliers datapoints to mitigate their influence.
- To enhance the general efficiency of the algorithms, normalization was applied.
- To improve prediction accuracy, we expanded the feature space by applying a polynomial expansion function, $\psi : [\hat{X}] \to [X, X^2, \ldots, X^d]$, to each numerical column. The polynomial degree d was treated as a hyperparameter for tuning the model.

D. Model Selection

Our model choice fell on several types of logistic regressors, driven by the classification nature of the problem to model the conditional probabilities of the target classes given the input features. We optimized the logistic loss by using Elastic net regularization and its subsets, Ridge ($\alpha = 0$) and Lasso ($\alpha = 1$).

$$\mathcal{L}(w) = \frac{1}{N}\sum_{n=1}^{N}\left(-y_n x_n^\top w + \log\left(1 + e^{x_n^\top w}\right)\right) + \\ + \lambda\left(\frac{1-\alpha}{2}\|w\|_2^2 + \alpha\|w\|_1\right)$$

E. Hyperparameters Tuning

For each model, we grid-searched over several hyperparameters using cross-validation with 5 folds, namely.

- λ: regularization coefficient to reduce overfitting.
- γ: step-size for descent.
- b: batch-size.
- $\bar{t}$: threshold (< 0.5) to reduce the possibility of biased results due to an imbalanced dataset (y_{pred} set equal to 1 if $y_{pred} > \bar{t}$).

To further extend our knowledge of the problem and better address the prediction we also implemented additional parameters to tune.

- f_s: an oversampling ratio to target a ratio of the positive class by resampling.
- *weights*: error weights associated with each feature.
- *cluster*: added feature based on *k-mean* clustering.

All parameters were tuned to reach the highest F_1 score possible.

III. Results and Discussion

The model that achieved the best performance on the AIcrowd test set with an F_1 score and accuracy of 0.405 and 0.859 respectively was the following: Logistic regression ($\gamma = 0.05$, $\bar{t} = 0.46$, $f_s = 0.4$, $cluster = 2$, $d = 1$ other parameters set to 0).

We discovered through several tries that using both *weights* and f_s led to a dysergy which was solved by only using the second hyperparameter. The improvement introduced by f_s showed a plateau at 0.4 threshold, meaning a perfect balance of positive and negative classes was causing overfitting.
One important insight about the real decision boundary was that it was not highly nonlinear. This conclusion was supported by the performance observed during feature expansion, where the best results were achieved with $d = 1$. The performance was slightly worse at $d = 2$ and decreased significantly with higher degrees of expansion.
Our aim was mainly to figure out which features were associated with the highest absolute weights. In this way, we wanted to investigate how important each feature was to the model's prediction. The top 5 features are reported in *Table I*. The features identified in the table are known

Feature	*Weight*	*Explanation*
HAREHAB1	1.29	Outpatient rehab after heart attack
RFHLTH	1.01	Adults with Good or Better health
SEX	-0.92	Sex of the respondent
FC60	-0.67	Estimated Functional Capacity
RFHYPE5	0.66	High blood pressure

Table I: **Top 5 Features by Absolute Weights**

to correlate with heart attack risk due to their association with overall cardiovascular health. It was reasonable that the presence or absence of these features could provide valuable insights into an individual's health status and potential vulnerabilities. Understanding the interplay between these features can enhance risk assessment and inform targeted prevention strategies, ultimately contributing to better health outcomes for individuals at risk of heart attacks.

IV. Conclusion

Despite the possibility of achieving higher performance by using a larger number of features, our model showed how reducing a great pool of features to a way smaller subset can perform the same and be faster in the majority of the cases.
The most important advantage was the access to more insights about the importance of each single feature. Explainability is something really important because it often happens in Machine Learning that models are hard to understand and may be considered black boxes. With that being said, we support the effort of careful pre-processing for this kind of task which may help in further studies on this topic.