Predicting Cardiovascular Diseases from Clinical and Lifestyle Data

Abhinand Shibu, Dominik Glandorf, F. Betül Güres
EPFL, Switzerland

***Abstract*—Cardiovascular diseases are the globally most prevalent cause of death. Predicting it from clinical and lifestyle information could support its prevention, diagnosis, and early treatment. We develop and assess a predictive model using data from 328,135 surveyed adults. We find weighted least squares most effective and demonstrate concerning model unfairness.**

I. Introduction

Cardiovascular diseases (CVDs) are the globally leading cause for deaths and loss of life years [1]. 79.6% of lost life years are attributed to modifiable risk factors such as lifestyle. Empirically assessing the risk of developing CVD could support prevention, early diagnosis, and treatment. Supervised machine learning has the potential to model the complex relationship between risk factors and CVD [2]. In this paper, we report our work on optimizing ML models for CVD prediction, evaluating modeling and model estimation decisions and explaining predictions.

II. Dataset

The data originates from telephone surveys by the Behavioral Risk Factor Surveillance System. The prediction target is whether a person has ever had a myocardial infarct or coronary heart disease. The data has 328,135 cases in the training split, 109,379 in the test set without public labels. The target is imbalanced, with only 8.8% of positive cases in the training data. The 321 features mostly concern lifestyle, clinical diagnostics and treatments. Figure 1 summarizes the features. 74.5% of the features have any, 30.8% have more than 90% missing values. A Principal Component Analysis (PCA) of mean-imputed, standardized data shows that 56 components can explain 50% of the total feature variance and that 19 components do not linearly explain any variance. Lastly, 40.1% of the features are binary and 100 are recalculated from other variables ("_" in their name).

III. Evaluation

As prediction performance *metrics*, we use F1, F2 and AUC-ROC. The F1 score reflects both the classifier's sensitivity and specificity, but depends on a decision threshold. Therefore, we also report AUC-ROC as a threshold-summarizing metric. We leave out accuracy as it is a misleading metric due to the class imbalance. We share the belief that, in preventive medicine, false positives are less severe than false negatives[3], which is why we also report

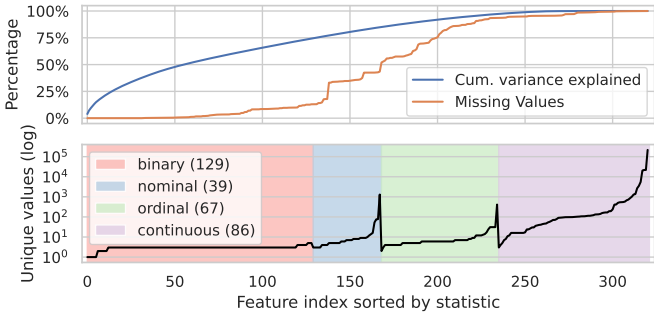

Figure 1. The dataset's features in terms of cumulative explained variance of principal components, missing and number of unique values, and scale.

the F2 score that prefers recall over precision:

$$F_2 = (1 + 2^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(2^2 \cdot \text{Precision}) + \text{Recall}}$$

To estimate model *generalization* to unseen data, we use 5-fold Cross Validation, which ensures using the entire dataset. Hyperparameter tuning is done within the respective training split, the validation set only serves for estimating the test error. We use a seed for reproducibility of the random split. We estimate the performance metrics for the best performing model on three demographic attributes, race, sex and age, to evaluate model fairness. For the final prediction on the test set and submission to AICrowd, we use the entire training set for the inner loop of the cross-validation.

IV. Preprocessing

From the documentation, we manually extract the feature-specific numeric codes for missingness (e.g., 77 or 99900) and classify features' scales into binary, nominal, ordinal, and continuous. We remove uninformative features without variance but add a constant column as a bias term. We also remove clear duplicates (e.g., height in meters/inches). As models require complete data, we impute missing values using mean imputation. To model the effects of nominal variables, we use a one-hot encoding of these features. To stabilize model estimation, we standardize all features by the training data's mean and standard deviation. To enable linear models to model non-linear relationships, we square ordinal and continuous features.

V. Models and Estimation

As a linear baseline, we use weighted *ordinary least squares* (WOLS) due to its closed-form solution and despite

Table I
ESTIMATED PERFORMANCE METRICS (%±STD) OF ASSESSED MODELS.

Model	F1-score	F2-score	AUROC	F1 AICrowd
WOLS	42.8±0.3	50.1±0.6	86.0±0.1	43.7
Logistic Reg.	42.4±0.4	47.6±0.7	85.9±0.2	
SVM	37.3±0.5	48.3±0.5	83.3±0.1	
kNN	36.7±0.9	46.6±1.4	82.5±0.6	
Decision Tree				
Neural Network				

Table II
IMPACT OF PREPROCESSING AND MODELING DECISIONS (WOLS).

Factor	Setting	Δ F1 (%)
Missing codes → N/A	No replacement	-1.6
One hot encoding	No encoding	+0.1
Squaring features	No squaring	+0.1
Tuning decision threshold / weighting samples	Decision threshold only	-4.5
	Weighted loss only	-0.5
	Neither	-27.9

its suboptimal quadratic loss function; we weight samples inversely to their class frequency, maintaining a convex loss (affine transformation) and tune the decision threshold after applying the Sigmoid function as the predictions are not probabilities. We also use regularized *logistic regression* (LR) with a similarly weighted logistic loss. We estimate the weights using full-batch gradient descent due to the small dataset and the convexity of the loss. We also use linear *support vector machines* (SVM) aiming for better generalization in high-dimensional spaces and insensitivity to class imbalance. For gradient-based methods, we use a 20% heldout validation set to decide for early stopping when the validation loss does not decrease anymore to prevent overfitting. We suspect nonlinear and hierarchical interactions between the features and therefore implement a nonlinear model, *k-nearest neighbors* (kNN). For the non-parametric kNN, we tune the number of neighbors k on the validation heldout.

VI. RESULTS

Table I shows the estimated performance metrics of our models. WOLS outperforms the other approaches. The kNN might suffer from the curse of dimensionality but works surprisingly well, confirming the observation from the PCA that the features may cluster in a subspace.

1) Ablations: Table II summarizes the impact of preprocessing and modeling design choices. Most importantly, either the loss should be weighted or the decision threshold tuned to counteract the class imbalance. Careful missing code implementation seems also relevant. Automatic feature engineering (one hot encoding and squared features) appear irrelevant.

Figure 2 demontrates how $> 50,000$ datapoints are sufficient to close the generalization gap even without (strong)

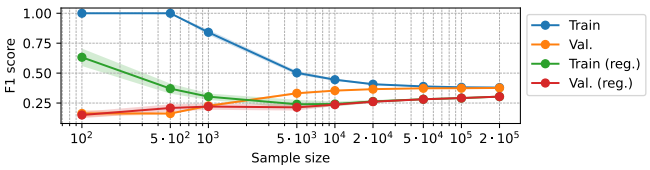

Figure 2. WOLS training and validation performance (generalization gap) with increasing data w/ and w/o regularization (CI across 5 random seeds)

Table III
ESTIMATED PERFORMANCE (%±STD) BY PROTECTED ATTRIBUTE.

Race	F1-score (%)
White	42.2±0.6
Black	38.6±2.6
Other	36.3±4.6
Multiracial	42.5±6.2
Hispanic	47.8±6.8
Unknown	41.5±4.0

Sex	F1-score (%)
Male	44.6±1.3
Female	38.3±0.6

Age	F1-score (%)
Under 65	36.4±1.1
Above 65	43.9±1.1
Unknown	28.9±12.6

L2 model complexity regularization, which closes the gap with less data at the cost of overall validation performance.

2) Fairness: Table III compares the performance of the logistic regression on demographic subgroups, revealing that the model performs better for whites, multiracial and hispanic than for blacks and other races, potentially due to lower representation in the training data. It also works better for males and elderly. This propels historical biases in medical research having the male body studied more rigorously .

VII. CONCLUSION

In this work, we have shown that weighted least squares can predict CVD moderately well (42.8% F1) from clinical and lifestyle data, when carefully counteracting class imbalance with a weighted loss and tuning the decision threshold.

REFERENCES

[1] Global Burden of Cardiovascular Diseases and Risks 2023 Collaborators, "Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023," *JACC*, 2025, publisher: American College of Cardiology Foundation. [Online]. Available: https://www.jacc.org/doi/10.1016/j.jacc.2025.08.015

[2] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 281, Dec. 2019. [Online]. Available: https://doi.org/10.1186/s12911-019-1004-8

[3] K. Ikemura, E. Bellin, Y. Yagi, H. Billett, M. Saada, K. Simone, L. Stahl, J. Szymanski, D. Y. Goldstein, and M. R. Gil, "Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study," *Journal of Medical Internet Research*, vol. 23, no. 2, p. e23458, Feb. 2021, company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical

Internet Research Label: Journal of Medical Internet Research
Publisher: JMIR Publications Inc., Toronto, Canada. [Online].
Available: https://www.jmir.org/2021/2/e23458