Predicting Cardiovascular Disease from Clinical and Lifestyle Factors

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

***Abstract*—Cardiovascular diseases are the globally most prevalent cause of death. Predicting them from clinical and lifestyle information could support prevention, diagnosis, and treatment. For this task, we assess machine learning models on 328,135 surveyed adults. We find weighted least squares most effective, relying on known risk factors and demonstrate concerning model unfairness for women, young and black people. We also replicate the hispanic paradox in our model. Code available under https://github.com/CS-433/project-1-ed4ml.**

I. Introduction

Cardiovascular diseases (CVDs) are the leading cause of death and a major contributor to the reduction in life expectancy worldwide [1]. About 79.6% of premature deaths from CVD are attributed to modifiable risk factors related to lifestyle. Empirically assessing the risk of developing CVD could support prevention, diagnosis, and early 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, fairness 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 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. For the best performing model, we estimate the performance with respect to race, sex and age, to evaluate model fairness. For explainability, we compute permutation feature invariance (PFI) that estimates which features decrease the F1 score the most when randomly permuted. For the prediction on the test set on AICrowd, we use the entire training set for the inner loop of the cross-validation.

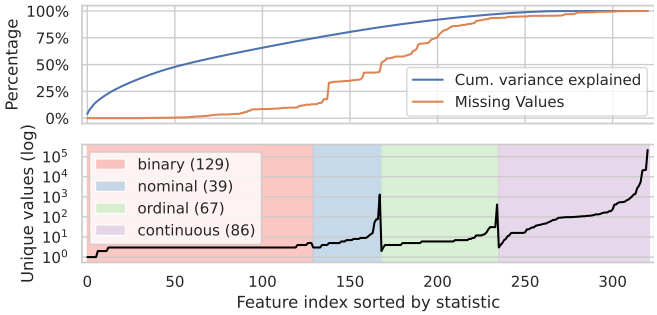

Figure 1. The dataset's features in terms of cumulative explained variance of principal components, missing and number of unique values, and scale.

IV. Preprocessing

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

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	**43.9**
Logistic Reg.	42.4±0.4	47.6±0.7	85.9±0.2	**43.9**
Linear SVM	35.0±0.2	**50.7**±0.2	82.7±0.1	35.6
Decision Tree	40.5±0.5	47.7±1.1	84.5±0.1	41.5
kNN	38.4±1.1	43.8±2.2	83.2±0.4	39.7

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

Factor	Setting	Δ F1 (%)
Missing codes → N/A	No replacement	-2.0
One hot encoding	No encoding	-0.1
Squaring features	No squaring	-0.3
Tuning decision threshold / weighting samples	No threshold tuning	-5.2
	No weighted loss	-0.8
	Neither	-31.9

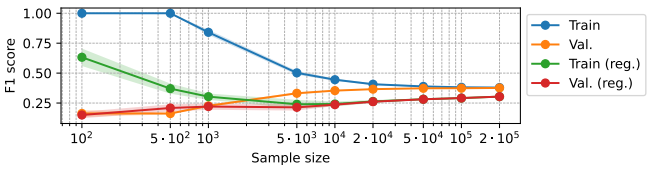

Figure 2. WOLS training and validation performance (generalization gap) with increasing data w/ and w/o regularization (CI across 5 random seeds)

Table III
WOLS PERFORMANCE (%±STD) BY DEMOGRAPHIC.

Race	F1-score (%)
White	43.2±0.5
Black	39.1±1.3
Hispanic	38.4±0.9
Other	44.3±2.0
Multiracial	49.9±4.3
Unknown	43.8±4.8

Sex	F1-score (%)
Male	45.9±0.3
Female	38.9±0.4

Age	F1-score (%)
Under 65	36.6±0.5
Above 65	45.3±0.6
Unknown	30.5±4.8

V. MODELS AND ESTIMATION

Weighted *ordinary least squares* (WOLS) is our baseline due to its closed-form solution; 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 convexity of the loss and dataset size. We test linear *support vector machines* (SVM) aiming for better generalization in high-dimensional spaces and insensitivity to class imbalance. For gradient-based methods, we use the validation set 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 nonlinear models, *k-nearest neighbors* (kNN) and *Decision Trees*. For all models, we list the tuned hyperparameters and tested configurations in our `README.md`.

VI. RESULTS

Table I shows the estimated performance of our models. Unregularized WOLS wins most metrics, but F2, where SVM is on par. 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. The Decision Tree cannot confirm the hypothesis that non-linear models outperform linear models.

1) Modeling 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 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 WOLS model across demographic subgroups. The model performs better for white, multiracial, and other racial groups than for Black and Hispanic individuals, likely due to lower representation in the training data and distinct underlying risk factors (see next section). The model performs better for older adults, likely because their data include both risk factors and CVD outcomes measured during the same survey, making the relationships clearer to the model. Although the training data included more women, the model performs better for men, suggesting an unclear interaction between structural sexism and healthcare [4].

3) Explainability: The PFI across five random seeds ranks maximum oxygen consumption(-), drinking(+), physical activity(-), asthma(+) and smoking(+) as the top risk(+)/preventive(-) factors for CVD in the WOLS model. These have been long known and studied [5], [6], [7], [8]. Interestingly, "being Hispanic" appears as an important risk factor despite a lower CVD prevalence (6.1%), reflecting the "Hispanic paradox," where traditional risk factors do not fully account for observed outcomes [9].

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. The dataset size is large enough to not regularize model complexity. White, old, men are favored by the predictive performance. The model uses exercise, smoking, drinking as classical predictive factors and exhibits the hispanic paradox.

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

[4] E. C. Dore, S. Shrivastava, and P. Homan, "Structural Sexism and Preventive Health Care Use in the United States," *Journal of Health and Social Behavior*, vol. 65, no. 1, pp. 2–19, Mar. 2024, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.1177/00221465231194043

[5] R. A. Bruce, F. Kusumi, and D. Hosmer, "Maximal oxygen intake and nomographic assessment of functional aerobic impairment in cardiovascular disease," *American Heart Journal*, vol. 85, no. 4, pp. 546–562, Apr. 1973. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0002870373905024

[6] J. B. Lakier, "Smoking and cardiovascular disease," *The American Journal of Medicine*, vol. 93, no. 1, Supplement 1, pp. S8–S12, Jul. 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/000293439290620Q

[7] M. Xu, J. Xu, and X. Yang, "Asthma and risk of cardiovascular disease or all-cause mortality: a meta-analysis," *Annals of Saudi Medicine*, vol. 37, no. 2, pp. 99–105, Mar. 2017. [Online]. Available: http://www.annsaudimed.net/doi/10.5144/0256-4947.2017.99

[8] P. E. Ronksley, S. E. Brien, B. J. Turner, K. J. Mukamal, and W. A. Ghali, "Association of alcohol consumption with selected cardiovascular disease outcomes: a systematic review and meta-analysis," *BMJ*, vol. 342, no. feb22 1, pp. d671–d671, Feb. 2011. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.d671

[9] S. Gomez, V. Blumer, and F. Rodriguez, "Unique Cardiovascular Disease Risk Factors in Hispanic Individuals," *Current Cardiovascular Risk Reports*, vol. 16, no. 7, pp. 53–61, Jul. 2022. [Online]. Available: https://link.springer.com/10.1007/s12170-022-00692-0