Cardiovascular Diseases Detection and Prevention

Neila Benlamri, Bipasha Goyal, Sofia Sannino

CS-433 Project 1: Girl Power Team

Abstract

This report presents our work on designing robust data processing and regression machine learning techniques to predict cardiovascular diseases (CVD) and identify the key factors influencing its development.

1. Introduction

Cardiovascular diseases (CVD) are a leading global health emergency, particularly in the context of an aging population. Thus, developing effective detection and prevention systems is essential. In this report, we compare the performance of linear and logistic regression models trained on the Behavioral Risk Factor Surveillance System (BRFSS) dataset.

2. Data Processing

In large and complex datasets the design of a structured data processing pipeline is critical for model stability and performance. The raw data was processed through the following steps.

2.1. Initial Cleaning & Feature Selection

Redundant columns (identifiers, date/time stamps, administrative info, derived/weight variables) were dropped, as they carry no predictive power and can introduce noise.

2.2. Missing Value Handling

Features missing more than 40% of values were removed entirely. For remaining missing values, imputation was performed using simple statistical measures, chosen to minimize distortion of the feature distributions. In particular, we classified and divided features as numerical and categorical, replacing missing values with their median and mode respectively.

2.3. Data Augmentation

Since our dataset is heavily unbalanced, we implemented data augmentation to avoid overfitting. In particular, we added small Gaussian noise to numerical features and added these noisy clone samples to the original dataset.

2.4. One-hot encoding

In order to properly train models on categorical features without introducing artificial order within categories, we implemented a one-hot encoding algorithm. We chose one-hot encoding to preserve interpretability and stability, with a reasonable increase in computational cost.

2.5. Correlation analysis

We performed Pearson correlation analysis on numerical features and dropped features with a Pearson's correlation coefficient higher than 0.90 in absolute value, in order to reduce multicollinearity and simplify features space, avoiding redundant information.

2.6. Winsorization Outlier Analysis

Winsorization statistical technique was used on numerical features for detecting and mitigating the influence of extreme values that could skew the models performance. We applied Winsorization above the 1st percentile and below the 99th percentile.

2.7. Standardization

All numerical features were standardized (zero mean, unit variance, $\mu = 0, \sigma = 1$), to ensure that all features are on the same numerical scale.

2.8. Principal Component Analysis (PCA)

We performed PCA on the standardized numerical features to further reduce dimensionality and remove redundancies.

2.9. Polynomial Basis Expansion (PBE)

To allow the inherently linear models to capture non-linear decision boundaries, features were expanded using a polynomial basis up to a predefined degree D. We chose $D = 2$ trying to maximize models' representation capacity without reaching an unsustainable computational cost.

2.10. Outcome

After step 2.2, 75 features were retained from the dataset, describing clinically relevant signals spanning demographics (age/race), access to care (insurance, personal doctor, cost barriers), diagnosed conditions (e.g., diabetes, hypertension, stroke), health status/limitations, behaviors (smoking, alcohol, diet, activity), and anthropometrics. The resulting feature set is compact, reproducible across train/test via saved artifacts, and aligned with cardiovascular risk factors expected to matter for heart-attack prediction. Clinically, the biggest signal is expected from age, cardiometabolic history, smoking status, and activity indicators, with height/weight providing BMI-like risk. Variables such as state and interview language are less directly causal but can capture regional and contextual differences that sometimes improve calibration.

3. Training and Model Evaluation

We describe below our machine learning workflow steps and choices.

3.1. Machine Learning Models and Metrics

- **Datasets** We trained models and then compared results using different compatible data processing pipelines to find the most suitable: *Dataset 1*, applied every step above sequentially, except PBE and data augmentation; *Dataset 2*, applied every step above sequentially, except PBE; *Dataset 3*, applied every step above sequentially, except data augmentation; *Dataset 4*, applied every step above sequentially, except data augmentation and one-hot encoding. We report here the results obtained on Dataset 1: similar trends were observed on the other datasets, though with lower performance.

- **Models** We compared and analyzed the following linear and logistic regression models : linear regression using *least squares*, linear regression minimizing mean squared error through *stochastic gradient descent*, *ridge regression* and *L2-regularized logistic regression* minimizing logistic loss through Adam algorithm. We implemented Adam to reach the optimum faster with respect to other algorithms such as stochatisc gradient discent (SGD).
- **Metrics** To compare models, we evaluated their performance using the following metrics: Area Under the ROC Curve (AUC), F1-Score (F1) and accuracy. We used AUC to discriminate between models and hyperparameter choices, since it is insensitive to class imbalance and evaluates how well the model ranks positive samples above negatives independently of the threshold chosen to classify 1 or 0. Once we found the optimal models and hyperparameters, we determined the discrimination threshold that maximized F1. F1 quantifies the balance between precision and recall, making it an appropriate metric for evaluating disease prediction models, where correctly identifying positive cases is critical.
- **Dealing with unbalanced data** Since our dataset is deeply unbalanced, we trained linear regression models with *oversampling* on the dataset and introduced *positive ratio class weights* in logistic regression to increase positive samples impact on the loss.
- Notation : λ regularization parameter, γ step-size in SGD or Adam, α positive class ratio parameter.

3.2. K-Fold Cross-Validation Between Linear Models

We performed an initial *5-fold cross-validation* scheme on oversampled data to evaluate linear regression models and initialize a rough hyperparameters tuning, ensuring an unbiased estimate. The model achieving the highest AUC was selected for further analysis and hyperparameter tuning. Although the difference is small, Ridge Regression achieves the highest AUC (difference lower than $1e^{-5}$). Moreover, we obtained $\lambda_{Ridge} = 1e^{-4}$.

Table 1: Initial Linear Model Performances on Dataset 1.

Model	AUC	F1	Acc. (%)
Least Squares	0.85778	0.37065	75.70
Ridge Regression	0.85800	0.37081	75.70
MSE (SGD)	0.85800	0.37081	75.70

3.3. Hyperparameters Tuning

A hyperparameter tuning pipeline was applied to the best-performing linear regression model and to the regularized logistic regression model trained with Adam. We followed *successive halving* strategy: computational resources were allocated progressively, refining the tuning around the most promising configurations (with respect to AUC), while discarding poor ones early. To evaluate metrics performance, we used again a k-fold cross validation, using oversampled dataset for the linear model. In particular, we started with $\lambda \in \{e^{-5}, ..., e^{2}\}$, $\gamma \in \{e^{-5}, ..., e^{2}\}$ Adam step-size, $\alpha \in \{0.25, 0.5, 1, .., 5\}$, we trained with randomly chosen tuples $(\lambda, \gamma, \alpha)$ with less folds (3) and iterations (for Adam), and then focused on the most favorable combinations, increasing folds and iterations.

3.4. Finding The Best Threshold

Once hyperparameters were defined, we determined the optimal classification threshold that maximized the F1-score. To this end, a *5-fold cross-validation with different random seeds* was performed, training both the best-found linear model and Adam logistic regression model to identify the best threshold for each.

3.5. Final K-Cross Validation

A final *K-cross validation with five folders* was implemented to train the two chosen models with the right hyperparameters and find the optimal weights for each.

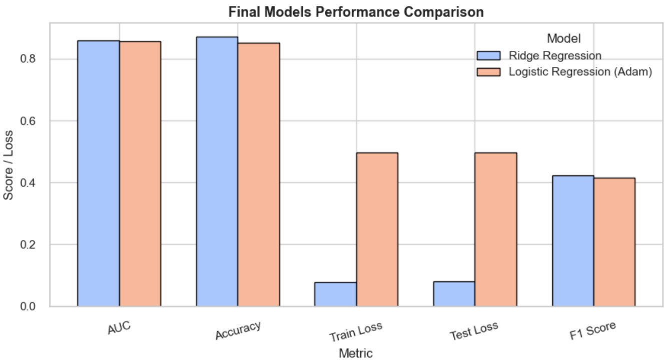

Figure 1: Final K-Cross Validation Results on Dataset 1

4. Validation

Before training, the dataset was randomly divided into a 20% validation set and an 80% training set, on which the steps described in the previous section were performed. Finally, the optimal weights from the linear and logistic models were applied to the validation set to compute performance metrics and identify the final optimal model for our objectives. We chose the best model with respect to F1-Score.

5. Conclusion

We report in the following table 2 the models performances in validation, with hyperparameters and optimal threshold t^* found as described in section 3.3 and 3.4. The Adam regularized logistic regression model used $\alpha = 1.0$, $\gamma = 0.010$. **Thus, the best model is ridge regression on Dataset 1** with hyperparameters and threshold described above. Applying this model, we reached an F1-Score of 0.434 and an accuracy of 0.875 on AIcrowd. The project successfully established a comprehensive machine learning framework and future work involving external libraries, more advanced ML models, and deeper hyperparameter tuning could further enhance overall performance.

Table 2: Final Performance on Validation Set Dataset 1.

Model	Dataset	λ	t^*	AUC	F1	Acc. (%)
Ridge Regression	Dataset 1	4×10^{-4}	0.68	0.858	0.424	87.12
Adam Reg. Log. Reg.	Dataset 1	1×10^{-3}	0.6996	0.8566	0.4187	85.43

References

[1] Centers for Disease Control and Prevention (CDC), "Behavioral risk factor surveillance system (brfss) 2015 annual data," https://www.cdc.gov/brfss/annual_data/annual_2015.html, 2015.

[2] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[3] C. C. Olisah, L. Smith, and M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and machine learning perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260722001596

[4] A. Reifman and J. Keyton, "Winsorize," in *Encyclopedia of Research Design*, N. J. Salkind, Ed. Thousand Oaks, CA: SAGE Publications, Inc., 2010, pp. 1637–1638. [Online]. Available: https://doi.org/10.4135/9781412961288.n502