Prediction of Heart Attacks from Lifestyle Factors

CS-433 Machine Learning Project 1

Mohamed Sami Ghrab, Yelizaveta Kulynych

***Abstract*—Cardiovascular Diseases (CVDs) are a leading cause of death globally. This project explores the use of machine learning to predict the risk of heart attack (MICHD) based on personal and lifestyle factors from the Behavioral Risk Factor Surveillance System (BRFSS) dataset. This report details our methodology, including data preprocessing, feature engineering, and model implementation. We present a comparative analysis of the models' performance and discuss key findings from our exploratory data analysis.**

I. Introduction

Cardiovascular diseases, particularly heart attacks, pose a significant public health challenge . Early detection and prevention are key to mitigating this issue. Machine learning models offer a promising approach by identifying high-risk individuals based on observable data, such as lifestyle habits and clinical measurements.

The goal of this project is to build and evaluate models for this binary classification task. We use a large dataset from the BRFSS, containing over 300,000 individual entries.

II. Data Preprocessing

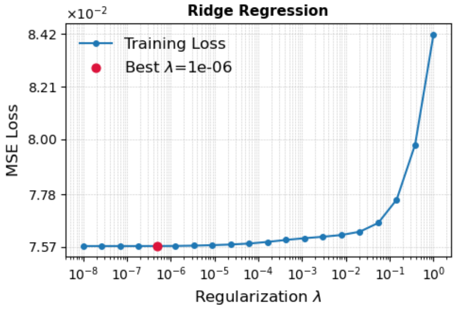
The BRFSS dataset includes 328,135 U.S. patients with 641 features, covering personal and behavioral factors such as alcohol and tobacco use. Each sample is labeled 1 for heart disease or -1 otherwise.

- **Undersampling:** To address the severe class imbalance, we created a balanced training set by randomly undersampling the majority class (label -1) to match the count of the minority class (label 1). Models were trained on this balanced data.
- **Missing Value Imputation:** Missing values were replaced with `NaN`. We then computed the *median* of each feature from the *original* full training set and used this value to impute all `NaN`s in the balanced training, imbalanced validation, and test sets.
- **Standardization:** All features were standardized (z-score) to have zero mean and unit variance. The mean and standard deviation were calculated from the *balanced* training set and then applied to all sets.
- **Feature Expansion:** We expanded the feature space using a 2nd-degree polynomial, concatenating the original standardized features with their squared values. This allows our linear models to capture non-linear relationships.
- **Bias Term:** A bias (intercept) column of ones was added to the feature matrices.

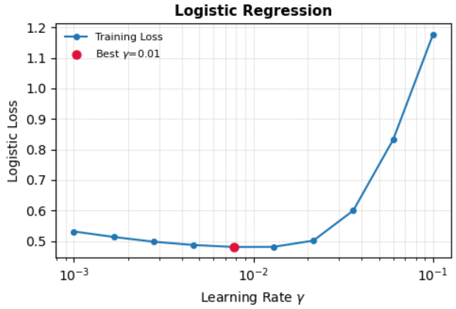
III. Machine Learning Models

We implemented and evaluated the following models on the preprocessed data. The target labels y were converted from $\{-1, 1\}$ to $\{0, 1\}$ for the logistic models.

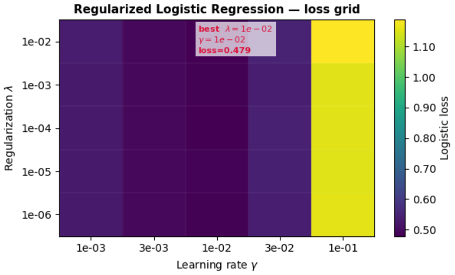
1) Ridge Regression : This model is a regularized version of least squares using an L_2 penalty. We first tuned the regularization hyperparameter λ by searching the best value that minimized the MSE loss. We found the optimal value to be 10^{-6}, which we then used to train the final model via the normal equations.



2) Logistic Regression : For the the logistic regression we first tuned the learning rate γ by searching on a log scale (from 10^{-3} to 10^{-1}) to find the best value for convergence. Using the optimal γ, we then trained the final model using gradient descent for 5000 iterations.



3) Regularized Logistic Regression (GD): This model adds an L_2 penalty to the logistic regression loss to prevent overfitting. We performed a grid search to find the optimal combination of the regularization parameter λ and the learning rate γ. Once the best pair of hyperparameters was identified, we trained the final model using gradient descent for 5000 iterations.



IV. Model Implementation and Results

A. Ridge Regression using Normal Equations

We tuned λ over $\text{logspace}(-6, 0, 20)$ and obtained the lowest loss at $\lambda = 10^{-6}$. The model achieved an accuracy of **78.82%** and an F1-score of **0.7935** on the balanced training set, and an accuracy of **76.22%** with an F1-score of **0.3767** on the imbalanced validation set.

B. Logistic Regression using Gradient Descent

The learning rate γ was tuned over $\text{logspace}(-3, -1, 10)$, yielding the best value $\gamma = 7.74 \times 10^{-3}$. After 5000 iterations, the model reached an accuracy of **78.46%** and F1-score of **0.7882** on the balanced training set, and **76.71%** accuracy with **0.3779** F1-score on the imbalanced validation set.

C. Regularized Logistic Regression using Gradient Descent

A grid search over $\lambda \in [10^{-6}, 10^{-2}]$ and $\gamma \in [10^{-3}, 10^{-1}]$ identified the best parameters as $\lambda = 0.01$ and $\gamma = 0.01$. The model achieved **78.36%** accuracy and **0.7873** F1-score on the balanced data, and **76.46%** accuracy with **0.3754** F1-score on the imbalanced validation set.

After tuning the classification threshold on the validation set, the F1-score improved to **0.4259** at a threshold of **0.73**. This regularized logistic regression model was therefore selected as the final submission model.

D. Experimental Setup

Models were trained on the balanced, preprocessed training set. We evaluated performance on two datasets: (1) the balanced training set itself, and (2) the full, imbalanced original training set (as our validation set). The key metric is the F1-score on the imbalanced validation set, as it reflects real-world performance given the class disparity. Hyperparameters (λ and γ) were tuned to minimize the loss on the balanced training set.

E. Model Performance

This table summarizes the performance of each model. While all models achieve high scores on the balanced data, their performance on the imbalanced validation set reveals their true utility.

On the balanced Data: All three models perform very similarly, achieving high Accuracy and F1 Scores, both in the range of approximately 0.78 to 0.79. Ridge Regression shows a slightly higher F1 Score (0.7934). On the Imbalanced Validation Set : All models experience a sharp drop in their F1 Score, which falls to around 0.37–0.38. The accuracy also drops, but less dramatically (to ∼0.76–0.77)..

TABLE I
Model Performance Comparison

Model	Data	Accuracy	F1 Score
Ridge Regression	Processed	0.7882	0.7934
	Original	0.76221	0.3767
Logistic Regression	Processed	0.7850	0.7872
	Original	0.7711	0.3804
Regularized logistic regression	Processed	0.7836	0.7873
	Original	0.7646	0.3754

F. Prediction Threshold Tuning

Our best model on the imbalanced data was **Regularized Logistic Regression**. Trained on a balanced subset, its default threshold of 0.5 performed poorly on the imbalanced validation set.

To improve the **F1-score** for the minority class, we ran a grid search over thresholds from 0.0 to 1.0. The optimal value, **0.32**, boosted the F1-score from 0.3754 to **0.4259** , highlighting the importance of post-training threshold calibration for imbalanced data.

V. Conclusion

In this project, we successfully implemented a full ML pipeline to predict heart attack risk. The most impactful steps were addressing the severe class imbalance via majority class undersampling and expanding the feature space with 2nd-degree polynomials.

Our results showed that Regularized Logistic Regression, followed by threshold tuning, was the most effective approach. The final F1-score of 0.4259 highlights the complexity of the dataset and the challenge of minority class detection.