Report: Machine Learning Project 1

Hanqi Lu (326798), Zhiyan Liao (395790), N'Zian Cédric Koffi (345346)
Neuro-X Section, EPFL, Lausanne, Switzerland

***Abstract*—Cardiovascular Diseases (CVD) have been a major health threat for decades, yet early predictions and prevention strategies are still under development. Here, we utilize classical machine learning algorithms to predict the risk of developing CVD based on lifestyle factors.**

I. Introduction

Cardiovascular Diseases (CVD) account for about 32% of global deaths [1]. Personal factors—including race, diet, and sleep—contribute to the elevated risk of CVD. Understanding these correlations aids in early prediction and prevention.

We apply five classical machine learning algorithms to analyze the 2015 BRFSS Survey Data on CVD [2]. Our goal is to predict whether respondents are classified as having coronary heart disease (MICHD) or not.

We outline our data preprocessing and model training methodology in Section II, present the model training results in Section III, and conclude with a brief discussion of our findings in Section IV.

II. Methodology

In this section, we explain the data preprocessing and model training procedures.

A. Data Preprocessing

1) Downsampling: The MICHD samples constitute about 8% of the dataset, indicating significant class imbalance. To mitigate bias, we downsample the dataset so that MICHD samples represent 20% of the population.

2) Filling Missing Values: To address the high proportion of missing values, we remove columns with over 15% missing values. We fill remaining missing values based on feature types: continuous features are filled with the mean, while categorical features are filled with the mode.

3) Normalization: We apply different normalization strategies based on feature types. Continuous features undergo z-score normalization, while categorical features use one-hot encoding to treat each category distinctly.

B. Model Training

We use five algorithms: gradient descent (GD), stochastic gradient descent (SGD), ridge regression, logistic regression, and regularized logistic regression. A grid search method finds the optimal step size γ and λ for each model.

1) Cross Validation: We use 4-fold cross-validation on the training dataset and compute the following metrics:

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy measures the proportion of correctly predicted instances (both true positives TP and true negatives TN) out of all instances.

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures the number of TP relative to the total positive predictions (TP and false positive FP), indicating the reliability of positive predictions.

F1-score

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F1-score shows the model's performance in predicting the positive class while accounting for both FP and FN. A high value indicates a good balance between precision and recall.

We then choose the model with the highest F1-score and relative good performance on other metrics to predict the test dataset.

III. Results

A. Model training

Through grid search method, we use following parameter settings for different methods:

a. GD & SGD: $\gamma = 0.01$, max_iter $= 10000$;
b. Ridge regression: $\lambda = 0.01$;
c. Logistic regression: $\gamma = 0.5$, max_iter $= 1000$;
d. Regularized logistic regression: $\gamma = 0.5$, $\lambda = 0.1$, max_iter $= 1000$.

To compute the gammas for the different approaches, we decided to set an acceptable number of iterations, then to choose the gamma that would produce the smallest error in this set number of iteration. If gamma is too high, the model may diverge or produce higher error. If gamma is too small, converge is slow and the error will be high after a set number of iterations. Thus, the gamma selected should be a nice

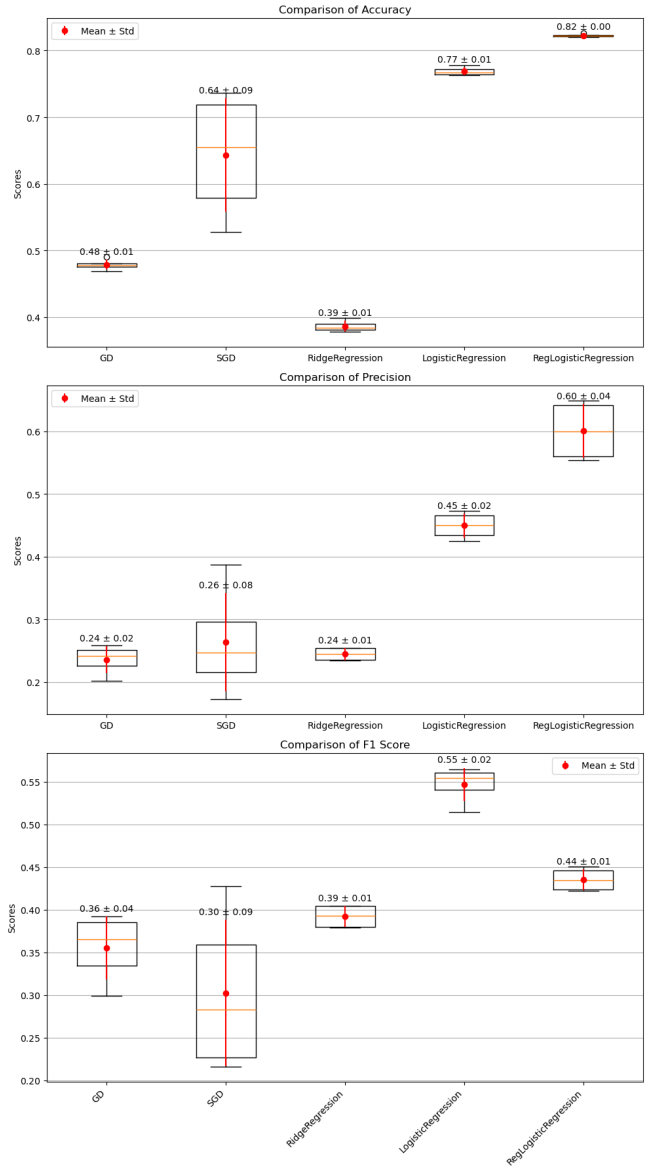

Figure 1. Comparison of accuracy (top), precision (middle) and F1-score (bottom) of different models on training datasets. The values labeled on each box represent the mean ± standard deviation.

trade-off. A more complete approach would involve sweeping between the orders of magnitude of lambdas, then fine-tuning its value to get close to divergence while maintaining a small error. The results of our approach may be a bit inconsistent. It turned out that the computing time was small enough for us to use $\gamma = 0.01 which has sure sconversion$.

We then compute accuracy, precision and F1-score to systematically characterize models' ability under the cross-validation settings. Figure 1 compares these metrics across models and the results suggest that logistic regression has the best performance among all, with the highest F1-score and relatively high accuracy and precision.

B. Prediction

Because of the outstanding metric performance of logistic regression, we decide to use this model to predict the label of test dataset. The final prediction results are included in the submission.csv, and the submission on https://www.aicrowd.com generates a F1-score of 0.364 and accuracy of 0.791 (ID 275066).

IV. Discussion

A. Data Preprocessing

In this project, we handle missing values with different strategies, considering the encoding density and type of the feature: we remove features that encode too little information (for example IDs and phone numbers, that are irrelevant) and features with too many missing values. Then we fill the remaining empty values for continuous and categorical variables separately.

Given that this dataset is very unbalanced with majority of samples being negative labels, we downsample the training data to increase the proportion of MICHD labels to avoid bias learning. Other strategies such as data augmentation, penalizing misclassifications in the minority class may also work in this situation.

We do not perform stringent feature selection for training model. However, it would be interesting to explore the impact of selecting the most relevant features (through computing Cramér's V and correlation coefficient, etc.) or use the decomposition methods (PCA, ICA, etc.) to train better models.

B. Model Training

We use five classical machine learning methods and the results suggest that the prediction model based on logistic regression has the best performance. To find the best γ and λ that could reach faster convergence rate, we perform grid search. Although straightforward, this strategy is computationally inefficient. Since the loss functions in our setting are all L-Lipschitz smooth (L is the Lipschitz constant), using the optimal learning rate $\gamma = \frac{1}{L}$ may significantly reduce the burden of computation.

Finally, our model's F1-score and accuracy on the test set indicate that there is room for improvement. Exploring more complex models and refining our training strategy are likely to yield better results.

References

[1] World Health Organization, "Cardiovascular Diseases (CVDs) Fact Sheet," https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), accessed: 2023-11-01.

[2] Centers for Disease Control and Prevention. (2015) Behavioral Risk Factor Surveillance System: 2015 BRFSS Data. Accessed: 2024-11-01. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html