Classification for prediction of heart attacks

Arthur Chansel, Marianne Scoglio & Gilles de Waha
Class Project 1, Machine Learning, EPFL

I. Introduction

Cardiovascular disease (CVD) is a major U.S. health concern. Hopefully, Behavioral Risk Factor Surveillance System (BRFSS) offers us a large health survey dataset to conduct an analysis of CVD risk factors. It includes many features describing U.S. residents' health-related risk behaviors, chronic health conditions, and use of preventive services. This project employs various machine learning methods to develop accurate predictive models. By doing so, we aim to enhance our understanding of CVD and contribute to more effective prevention and management strategies.

II. Exploratory data analysis and feature processing

Data handling is a crucial step in the conception of a model. The dataset of the health survey is quite large, it contains many NaN values and different types of features such as continuous and categorical. Missing data can introduce bias and affect the performance of the model, thus we removed the features with more than 50% of NaN values. In order to deal with the missing data of the remaining features we tried two different approaches: we imputed them with the median [1](simple and quite robust) or we created new features with them later through one hot encoding (in that way they do not have any effect on the other categories). Feature selection is important in order to choose a subset of the most relevant ones while discarding less important and redundant features. We calculated the correlation [2]between features and the predictions and we kept the most correlated ones (adding more was not improving our model). It is a simple and quick way to do it. Since the selected features seemed important to us and allowed the model to reach a good performance, we estimated that this method was enough. We applied one hot encoding [3]on the categorical features to ensure that the model would make a distinction between the different categories. Later we standardized them and the continuous ones. The last step of our data processing consisted in PCA [4], which reduces the dimensionality of the dataset while retaining most of the essential information and filtering out noise and redundancy.

III. Models and Methods

Here we will describe the different models and methods we tried to implement. The following methods are implemented for output labels $y \in \{0, 1\}$. The data provided contains $y \in \{-1, 1\}$. Therefore, we replaced all -1 by 0, and then we reversely substituted labels in the predictions to be consistent with the data.

A. Gradient Descent

The gradient descent is an optimization technique, to find the minimum value of the loss function. It is defined by:

$$w_{t+1} = w_t - \gamma \nabla L(w_t) \tag{1}$$

B. Logistic regression

The logistic regression is a binary classification technique that models the probability of an outcome belonging to a specific class. It's an extension of linear regression, but instead of directly predicting values, it uses the logistic function to map the model's output to a probability range between 0 and 1. It is also more robust to extreme values.

The logistic function is defined as:

$$\sigma(\mu) = \frac{e^{\mu}}{1 + e^{\mu}} \tag{2}$$

The loss uses the negative log of max likelihood estimation (MLE). For $y \in \{0, 1\}$ the loss and its gradient are defined as follows [5]:

$$L_A(w) = \frac{1}{N} \sum_{n=1}^{N} -y_n x_n^T w + \log\left(1 + e^{x_n^T w}\right) \tag{3}$$

$$\nabla L_A(w) = \frac{1}{N} X^T (\sigma(Xw) - y) \tag{4}$$

C. Regularized logistic regression

The regularized logistic regression is a similar method, in which we add a penalty to the loss to control the weights, and avoid overfitting. Then we want to find w which minimizes the loss and the penalty term together. The regularization parameter λ can be fine tuned, to get best results.

Here we use the Ridge regression regularizer, defined by:

$$\Omega(w) = \lambda ||w||_2^2 \tag{5}$$

Then the loss, and its gradient are:

$$L_B(w) = L_A(w) + \lambda ||w||_2^2 \tag{6}$$

$$\nabla L_B(w) = \nabla L_A(w) + 2\lambda w \tag{7}$$

D. Weighted regularized logistic regression

In logistic regression, the standard loss function can be modified to include class weights θ_0 and θ_1 to address the problem of imbalanced datasets [6]. The class weight for the minority class is set to a higher value, making errors on the minority class more costly. The new loss and its gradient are:

$$\begin{aligned} L_C(w) = \Omega(w) + \frac{1}{N}\sum_{n=1}^{N}\Big[&\theta_1 y_n \log\left(\sigma(x_n^T w)\right) \\ &+ \theta_0(1-y_n)\log\left(1-\sigma(x_n^T w)\right)\Big] \end{aligned} \tag{8}$$

$$\nabla L_C(w) = \frac{1}{N}X^T(\theta_0(1-y)\sigma(Xw) - \theta_1 y(1-\sigma(Xw))) \tag{9}$$

We choose the weights with the inverse class frequency:

$$\theta_i = \frac{\text{samples}}{2\cdot\text{samples}(\text{class}_i)} \tag{10}$$

E. Cross-validation and optimization

Cross-validation, a technique ensuring fair data splitting, involves systematically rotating subsets. Data is divided into k random groups, with one as the validation set and the remaining k-1 forming the training set. This process is repeated k times, and losses are averaged. We used 5-fold cross-validation to tune hyperparameters λ and γ.

F. Metrics

Using the F1 score [7] for asset predictions is crucial with imbalanced data. Unlike accuracy, the F1 score considers both precision and recall, providing a more accurate assessment, especially for rare or critical events.

IV. Results and Discussion

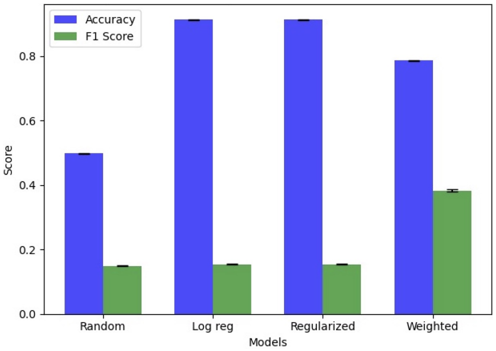

Fig. 1. Comparison of Accuracy and F1 Score for Different Models

We conducted a step-by-step analysis of our classification model with the objective of maximising the F1 score. The first two models used the simple logistic regression function and the regularized one, two fundamental approaches to binary classification. Other functions as MSE and Least Squares are not suitable for this task.

Tuning the λ of the penalty term using 5-fold cross validation, the performance does not increase much since there was not overfitting with the simpler model. However, with $\lambda = 10^{-7}$, the F1 score increases slightly. Although logistic regression is a widely used technique, our dataset exhibited a significant class imbalance and this loss function was insufficient to overcome this challenge. An unbalanced dataset introduces a bias towards predicting the majority class and leading to poor generalization on the minority class. In our case, we have a distribution of 10%-90%. An easy and quick method to unbias the predictions is to introduce class weights. In our case it is important to have good predictions for the class 1 since we want to know which patients have an enhanced susceptibility to heart attack. The class weights allows to make errors on the minority class more costly and to make the model focus on it [6]. Since now we have way more predictions for the class 1 than before, the accuracy drops a bit (more FP), but the F1 score increases significantly 1(especially for the better recall). Since the size of our final dataset is not very large and the model does not take long to run, GD is a better choice than SGD. GD processes the entire dataset in each iteration providing more accurate gradient estimates and leading to a stable and precise convergence [8].

The best predictions on the test test were obtained with the weights used for the best F1 score on a validation set (20% of the initial training set). This weights give a validation F1 equal to 0.410 and the test F1 is 0.426. Figure 2 shows that the the score does not converge, but oscillates. This is probably due to the fact that $\gamma = 1$. However, this value allows to reach better performances.

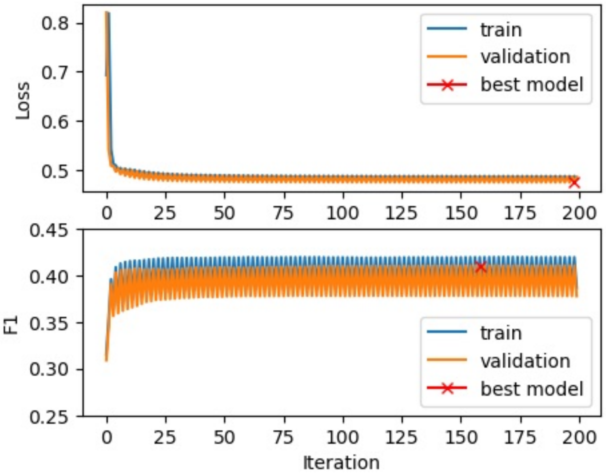

Fig. 2. The loss values and F1 scores for both training and validation sets. The best results based on the lower loss (0.475) and on the higher F1 score (0.410) are showed. The weights at iteration 162 are used to calculate the prediction over the test set.

V. Conclusion

This project provided a practical overview of methods seen in class. We explored data handling techniques and machine learning models based on logistic regression to get the best results with wise metrics. Retrospectively, we could have try other methods for features selection. We used correlation which seemed a good trade-off between reliability and computation time. However, comparing our results with other algorithms would have been valuable. We hopefully designed a good model to predict CVD risk based on health factors.

References

[1] A. Amballa, “Feature engineering part-1 mean/ median imputation.” *Medium*, 2020.

[2] V. R, “Feature selection — correlation and p-value,” *Medium*, 2018.

[3] Lekhana˙Ganji, “One hot encoding in machine learning,” *GeeksForGeeks*, 2023.

[4] Mayureshrpalav, “Principal component analysis(feature extraction technique),” *Medium*, 2020.

[5] S. Thavanani, “The derivative of cost function for logistic regression,” *Medium*, 2019.

[6] R. Abhinav, “Improving class imbalance with class weights in machine learning,” *Medium*, 2023.

[7] R. Kundu, “F1 score in machine learning: Intro calculation,” *V7labs*, 2022.

[8] E. Irby, “Gradient descent vs stochastic gd vs mini-batch sgd,” *Medium*, 2021.