Machine Learning: Project 1

Benoit Mathey-Doret
Henri Roniger
Simon Deconihout

***Abstract*—This work focus on the classification of sample in order to predict the risk of Cardiovascular diseases. This has been achieved using data processing and augmentation, followed by a ridge gradient descent. This method permit to obtains a F1-score on the test samples of $F1 = 0.434$.**

I. Introduction

The aim of this paper is to predict the risk of Cardiovascular Diseases (CVD) using machine learning tools. Since CVDs are among the leading causes of mortality worldwide, it's a major challenge to be able to predict risks effectively and accurately so that patients can be treated on time. The considerable number of factors that contribute to the development of CVDs, and the variability in these factors between individuals, makes the prediction of CVDs very challenging. For this purpose, the *BRFSS: Behavioral Risk Factor Surveillance System* dataset [1] will be analyses only troughout a data science pipeline.

II. Exploratory Data Analysis

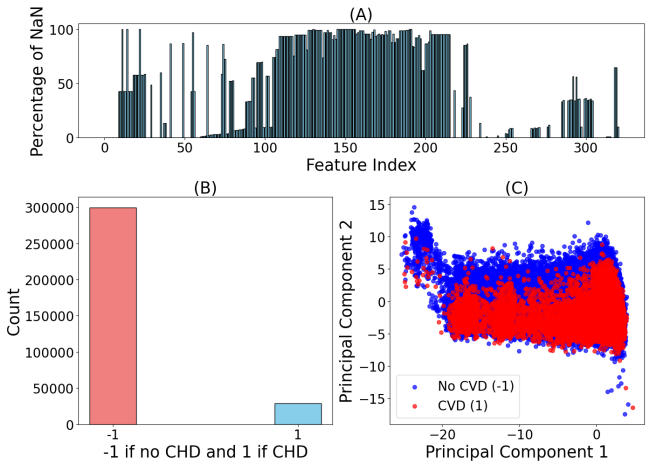

Figure 1. **A**: Percentage of Missing Values per Feature in the train dataset, **B**: Distribution of CVDs in the train dataset, **C**: 2D PCA Projection of Data

The first step is to study the structure of the Data [2], [3]. The BRFSS dataset collects data in the United States via telephone calls to American households. The received dataset includes 328,135 individuals who participated in the telephone survey, representing the same number of samples. They were asked 321 questions, which corresponds to the number of available features. The variable *MICHD* access if the patient have ever had CVD (1) or not (-1) is stored in the y dataset and is widely unbalanced (1 figure B). When a person did not respond to a question, a NaN (not a number) fills the feature's cell. Replacing the large number of NaNs in the dataset was one of the initial challenges (1 figure A). Concerning the two principal components (1 figure C), the two classes appear to overlap significantly, indicating that it does not provide a clear separation between the classes. Additional components or other feature engineering techniques might be necessary for better class separation.

III. Methods and Models

To achieve best model performance at predict CHDs, two aspect where examined and optimized.

A. Data Processing

As it is implied in the previous section, the dataset need to be process in order to get the most out of it. Our data processing goes trough the folowing steps [4]:

Missing values
: By studying in depth how the questionnaire was constructed, we noticed that for each question, a specific number was designated if the person chose not to answer. This number consisted only of 9s (9, 99, 999, etc.), allowing us to replace the NaNs to retain more samples without losing information on these samples.

Features processing
: First by inspecting the features, we drop *IDATE, SEQNO, PSU* because their shape is irrelevant in our data set. Then *WEIGHT2* and *HEIGHT3* where respectively convert to pounds and inshes. Normalisation where coonduct in order to to improve model convergence and balance the weight of each features. The formula used to normalize is the following: $X_{\text{standardized}} = \frac{X-\mu}{\sigma}$. PCA where then perform in order to reduce the dimensionality of our dataset.

Sample processing
: To handle the data unbalanced II, we first try to artificially augment the number CVD in order to achieve the same ratio. For this purpose, CVD data have been duplicated N times, where N is an integer optimized in the code (see below). This

dataset is referred as **D1**. An other approach is to reduce the data to have less -1, with a ratio of N times more -1 than 1 (thus, not the same N as in D1, but similar idea). This second method is then preferred for the gradient descend because of the smaller matrices size and a best computational speed.

B. Model selection

For the regression function, many possibilities where considered. In particular, regularized logistic regression, regularized least square (ridge), and Support Vector Machine (SVM) where tested after the data processing explained in sec.III-A combined with data augmentation/reduction. Various N had been tested and the best for each method is written in Tab.I. Different regularization terms where tested on a range of $\lambda \in [10^{-8}, 10^{0}]$. The same process was used for γ, for regression that use gradient descent. To have robust hyperparemeters, k-folds were also implemented (usually with 4,5, or 6 folds). This process allows us to find the optimal parameters for each regression function. Then, our model is tested on an intact test set, which was put apart at the beginning of the code and not used for training or in the k-folds. Finally, as the predict does not return a vector with -1 and $+1$ but with real value, one need to use $np.sign(y_{pred} - p)$ and optimize with a penality p. Indeed, if there is too few cardiac crisis predicted (because of data unbalance), then a small $y_{\text{pred},i} = -0.1$ can be still considered as a $+1$ label. The accuracy and F1 score are used as an evaluation metric, before submiting our results.

IV. Result

Reg. function	**Ridge (D1)**	**Ridge**	**Logistic**	**SVM**	**Ridge poly**
λ_{opt}	$5.2 \cdot 10^{-8}$	$9.2 \cdot 10^{-7}$	$2.8 \cdot 10^{-4}$	$7.5 \cdot 10^{-4}$	$7.2 \cdot 10^{-5}$
γ_{opt}	-	-	0.3	0.5	-
k **folds used**	4	4	4	4	5
N_{opt}	4	3	-	-	4
p_{opt}	-0.017	-0.068	-0.174	-0.156	-0.217
F1	0.421	0.418	0.417	0.412	0.434
Accuracy	0.868	0.871	0.864	0.865	0.871

Table I
Optimal parameters (λ, γ, k, N and penalty p) and performance metrics for Ridge, Logistic, and SVM models. D1 is for the augmented datasets, all the other use reduces dataset with the same $N_{opt} = 3$

The best score obtained, in Tab.I, is $F1 = 0.434$, using the preprocessing III-A, and ridge regression with a polynomial feature expansion $\phi : x \to (x, x^2)$. We did not use PCA for this regression function because it did not improve our prediction. In addition, since it does not contain any optimization, we do not need to train for different learning rate γ. As consequences, we save some computational time which allows us to train with high dimensional data for this approach. Concerning the sample processing, data reduction was used. To try this method with data augmentation would have been good, but because of lack of time (and lack of AI crowd submissions) it was not implemented.

V. Discussion

The fact that ridge regression better performed than other regression technique in our case can be explained by many factors. First, it is guaranteed to find a global optimum because the objective function is convex. The regularization term reduce the model complexity and mitigate over fitting on the train data. These characteristics make Ridge regression a good choice to handle high dimensional dataset such a BRFSS.
The feature augmentation $\phi : x \to (x, x^2)$ might also help to classify feature with hight absolute values from one with low values in the dataset. Indeed, it is very unlikely that the data are linearily separable; but since N is too large, Kernel methods could not have been implemented.

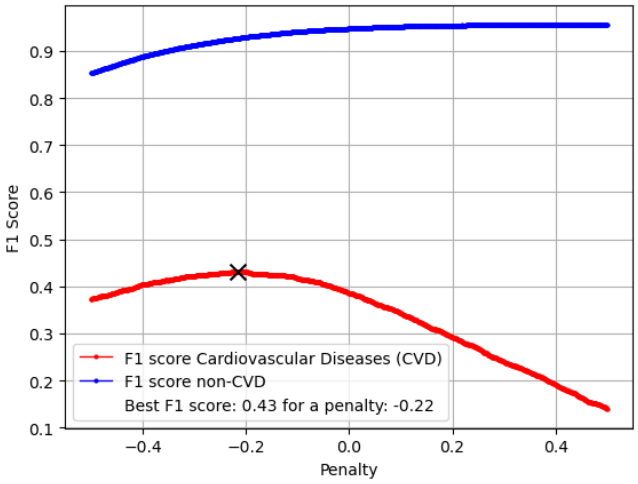

Figure 2. Impact of penalty term on F1 score for ridge regression with polynomial features, with a data reduction such that $number(-1) = N \cdot number(1)$, with $N = 4$

During the search of the best F1-score, it has been observed the impact of the penalty term optimization, also known as threshold, on the prediction quality. This parameter can increased the F1-score by 50% when using the same model, as we can see on Fig.2. As the dataset is unbalanced, meaning that the humans (samples) with CVD represent less than 10% of the population (dataset), the model tends to underestimate the real number of CVD on the prediction, the penalty is here to adjust the level of CVD prediction. The data augmentation/reduction have a similar role as the penalty, by adding more samples with CVD, or reducing samples with not-CVD, it will restablish the right amount of $+1$ (people with CVD risk) labels. That is why when data augmentation is used the penalty will tend next to zero, and when to much augmentation is make/or too much reducing non-CVD , penalty become positive.

REFERENCES

[1] U. S. Government, "Brfss survey data," 2015.

[2] M. Brbic, "Cs-401: Applied data analysis," 2024.

[3] /, "Exploratory data analysis," *Wikipedia, The Free Encyclopedia*, last edited on 12 September 2024.

[4] J. Brownlee, "Discover feature engineering, how to engineer features and how to get good at it," *Machine Learning Mastery*, 2020.