Machine Learning Strategies for Coronary Heart Disease Classification in the Face of Data Imbalance

Wesley Monteith, Camille Challier & Gonçalo Braga
EPFL CS-433, 30/10/2023

***Abstract*—This study deals with the classification of Coronary Heart Disease (CHD) positive individuals based on the *Behavioral Risk Factor Surveillance System* (BRFSS) data set [1]. Heart disease being the leading cause of death in developed countries the need to find complementary diagnostic tools is real. We propose to leverage Machine Learning methods to tackle this problem. The methods explored here are particularly suited for such a problem since they take into account the natural class imbalance of this type of data.**

I. Introduction

Heart disease is the leading cause of death in the United States [2] and Europe [3]. In our dataset consisting of 328135 responses, 28975 reported to suffer from this disease. From a machine learning perspective, the frequency of our event class (ie. having the disease) is $\approx 8.83\%$ which can be considered as a challenge, as the event and non-event classes are heavily imbalanced. This poses the need for innovative approaches that can classify reliably despite this imbalance. This study explores several such methods and highlights the significance of addressing class imbalance to improve CHD classification.

II. Exploratory Data Analysis

After a systematic review of our data, the survey codebook [4] allowed us to split the features into discrete and continuous versus categorical, and to build a glossary for the latter. The feature distributions were plotted against the target feature to determine their relevance (panels a, b and d in 1). We thematically grouped the variables and looked for correlation among ones of the same topic (panel c). Taking into account their meaning and degree of correlation we could then discard or not one of them, helping reduce dimensionality.

III. Data pre-processing

Some features of the data set can be deemed useless for training a model: "Record Identification" variables (participant contact, interview details,...), redundant variables or multi-frequency variables (day^{-1}, $year^{-1}$). Fortunately, many of these features were condensed into "calculated variables" by the surveyors, so only the latter were kept.

We addressed the issue of multiple non-answer choices ("Don't Know", "Refused", etc) by merging them into a single "missing value" answer. Variables that were constant across

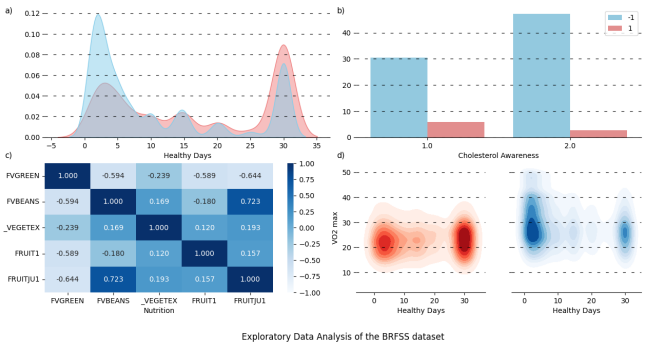

Figure 1. Different preprocessing vizualisations. a) Health-Related Quality of Life distribution in individuals with (1) and without the disease (-1) (discrete feature). b) Disparities in identified high cholesterol levels (categorical). c) Correlation among nutrition-related variables. d) Correlation of $VO2_{max}$ and general health in disease vs. non-disease groups.

all individuals, along with ones that had too many missing values (threshold of 95%), were not considered useful and removed. To mitigate multicollinearity, only one among proportional, linearly correlated variables was retained.

For discrete and continuous columns, missing values were replaced with either the median or the mean of the category. Ultimately standardisation was applied to each of them.

For categorical features we decided to use One-Hot Encoding: each feature's category is mapped to a binary column, which prevents assignment of magnitudes and unintended ordinal relationships to categorical variables. One downside to this encoding is that it greatly increases the dimensionality of the data. This was attenuated by dropping one binary column per categorical variable, as it can be deducted from the other categories.

IV. Implemented Methods

1) Regressors: Some regression methods were implemented: Least Squares, which computes analytically the weights that minimize the MSE; Gradient Descent (GD) and its Stochastic counterpart (SGD) which minimize the loss using its gradient (MSE or MAE); regularized regressors, Lasso, Ridge and ElasticNet which penalize the weights by adding resp. the L_1, L_2 or a combined norm. These methods generate continuous predictions, which we adapt for classification by taking their signs.

2) Classifiers: Implemented classification models include

Logistic regression (LR) and its regularized derivatives (L_1 & L_2 norm). These aim at optimizing the log-likelihood objective thus maximizing the overall accuracy, with GD or Newton's method.

Our data being highly imbalanced, LR will tend to favour the non-event class. To tackle this we tried both down-sampling the non-event class and up-sampling the event class to obtain perfectly balanced data before training a LR (resp. DLR and ULR).

Hyper-parameter tuning for DLR or ULR is challenging since doing plain Cross Validation (CV) on the up- or down-sampled data does not give an accurate approximation of the test score as the train data is balanced while the test data is imbalanced. Therefore to search hyper-parameters more effectively for ULR, we implemented a Stratified Upsampled CV (SUCV), which splits the folds while preserving the imbalance of the data and then up-samples the event class to a 1:1 ratio for the training folds, but leaves the testing fold imbalanced.

Similarly, a recently proposed solution by Zhang et al [5] is to use a penalized log-likelihood function defined as

$$LL(\omega, \lambda) = \sum_{i=1}^{N} \lambda_i y_i log(\sigma(\omega^T x_i)) + (1-y_i) log(1-\sigma(\omega^T x_i))$$

Here misclassification is penalized only for the event observations, giving more importance to the under-represented class. A GD model can be derived for both the weights and the penalization term $\lambda \in \mathbb{R}^N$ to create an Imbalanced LR (ILR).

Using multiple Lasso LRs (LLR) to make an Ensemble (LLRE) as proposed by Wang et al[6], one can combine different clusters of non-event class with the event class and train LLRs on these balanced subsets. Hyper-parameters were tuned to maximize F1 score of the test data containing every other cluster plus all event data points. For this method to work, we implemented a simple KMeans clustering algorithm, and applied it to the pre-processed data.

Hyper-parameters of all other regressors, were tuned using an implemented grid search paired with a 3-fold CV.

V. Results

In Fig. 2 we observe that continuous regressors tend to have a high accuracy but a low F1 score. LRs on the other hand have higher F1 scores and reasonable accuracy.

Overall estimated and true accuracy are close, exceptions being ElasticNet (Δ=0.2$\pm$0.006), Mini-Batch GD (Δ=-0.203$\pm$0.029) and Upsampled LR (Δ=0.129$\pm$0.001). The same is true for F1, exceptions being Mini-Batch (Δ=-0.103$\pm$0.014), Up- (Δ=-0.256$\pm$0.002) and Down- (Δ=-0.399$\pm$0.001) sampled LR. SUCVLR delivers the best performance ($Accuracy_{true}$=0.871, $F1_{true}$=0.437), ILR the 2^{nd} best ($Accuracy_{true}$=0.876, $F1_{true}$=0.422) and L2 LR the worst ($F1_{true}$=$Accuracy_{true}$=0). The best F1 score is upsampled LR ($F1_{true}$=0.437) and the best accuracy is held by LS, polynomial Ridge and Ridge ($Accuracy_{true}$=0.915).

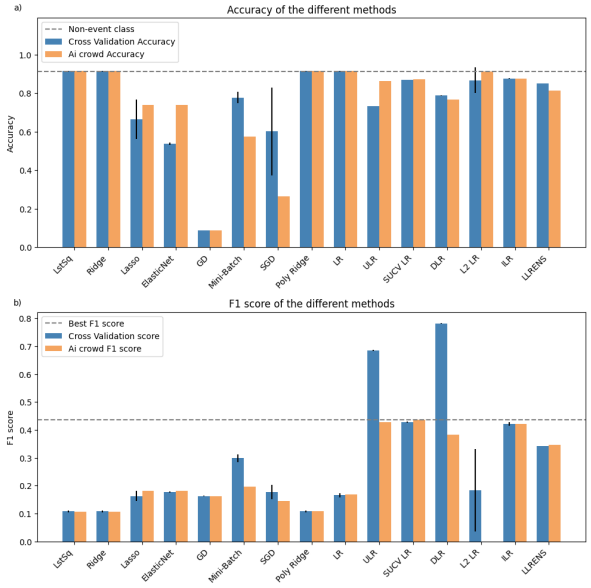

Figure 2. Comparison of estimated Accuracy (a) and F1 score (b) for the different implemented methods (with standard deviations) versus truth values (AiCrowd). ULR (γ=0.3, I=100) & ILR (γ=0.4 α=0.06) are our best models. All other hyperparameters can be found in our README.

VI. Discussion

Continuous regressors achieve high accuracy but low F1 scores, indicating a tendency to over-predict the non-event class as illustrated by the dotted line (Fig 2.a). In contrast LRs demonstrate higher F1 scores along with reasonable accuracy, making them better suited for correctly identifying true positive CHD cases.

For ULR, the substantial differences between the estimate and the truth value of the test F1 score, using CV, can be effectively corrected by using SUCV. We used it for hyper-parameter tuning, and this lead to our best model for the test F1 score. One downside with ULR is its computational complexity. By up-sampling the data we obtain a very large train feature matrix, which takes a lot of RAM and slows down the overall training time.

On the other hand, the ILR is more computationally efficient as it already considers the imbalance in its log likely-hood function. Plus, this model has some merit as it yielded the second best test F1 score and a sound accuracy.

Unfortunatly LLRENS did not prove itself as fruit-full as ULR or ILR.

VII. Conclusion

CHD occurance classification was performed through different regression models. We observed that most regressors tend to favor the dominant/non-event class and thus classify poorly (data imbalance problem). By leveraging appropriate models (ULR, SUCV-LR, ILR and LLRENS) and hyper-parameter tuning one can obtain a higher number of true positives.

REFERENCES

[1] "CDC - 2015 BRFSS Survey Data and Documentation," Apr. 2022. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html

[2] CDC, "Heart Disease Facts | cdc.gov," May 2023. [Online]. Available: https://www.cdc.gov/heartdisease/facts.htm

[3] "Causes of death statistics." [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Causes_of_death_statistics

[4] "2015 BRFSS Codebook." [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html#:~:text=phone%20data%20set.-,2015%20BRFSS%20Codebook,-pdf%20icon

[5] "Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function - PMC." [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9542776/

[6] "Large Unbalanced Credit Scoring Using Lasso-Logistic Regression Ensemble | PLOS ONE." [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117844