"THE GOOD", "THE BAD" and "THE UGLY" for Cardiovascular Diseases Prediction on the Behavioral Risk Factor Surveillance System

Antonio Mari[1], Matteo Santelmo[1], and Jakhongir Saydaliev[1]

[1]Department of Computer Science, EPFL, Switzerland

Abstract—**According to World Health Organization, Cardiovascular Diseases (CVD), such as heart attacks, are becoming one of the leading causes of death globally. The Behavioral Risk Factor Surveillance System (BRFSS) provides a dataset containing information on health-related risks and behaviors of the U.S. residents. We develop machine learning models to help with the early detection of CVDs, experimenting with different pre-processing strategies for the dataset and gaining insights from analyzing the importance of the features. Despite similar performance for different strategies, we found out that a careful pre-processing highlights the importance of meaningful features, opening the way to proceed with deeper analysis together with experts in the field.**

I. INTRODUCTION

The rise of new technologies such as machine learning algorithms can help with the early detection and prevention of developing CVDs. The BRFSS is an ongoing surveillance system designed to measure behavioral risk factors for the noninstitutionalized adult population (aged 18 years of age and older) residing in the United States and the dataset provided is derived by landline telephone- and cellular telephone-based surveys. Here, interviewers collect data from a randomly selected adult in a household.

Adults are classified as having coronary heart disease (MICHD) if they reported having been told by a provider they had MICHD, a heart attack or angina. In terms of early detection and prevention of these disease, it is our job to build a model able to estimate the likelihood of developing MICHD given a certain clinical and lifestyle situation.

II. MODELS AND METHODS

In this section, we describe the engineering choices for our ML system[1], providing insights into the dataset (section II-A), illustrating our pre-processing strategies (section II-B), and reporting how we performed model selection (section II-C).

A. Dataset Overview

The dataset collects the answers of 328135 respondents in 322 features, semantically grouped in different sections, such as demographics, health conditions, alcohol consumption or exercise. Based on their possible values, features can be categorized into

- binary (yes-no questions, $\approx 34\%$ of the total),
- categorical (no meaningful answers order $\approx 18\%$),
- numerical (quantitative or meaningful order $\approx 48\%$).

Moreover, since the condition to be detected for our task is not that common (less than 9% of positive answers), the dataset exhibits a strongly unbalanced target variable, so the accuracy is not a good metric for evaluating the model performance. Precision and recall rather give us more useful insights, in particular, the recall is a well suited metric to measure the ability of correctly identifying an individual who is actually at risk, which is fundamental for the prevention of life-threatening conditions. Putting these two metrics together, we compute the F_1 score.

B. Pre-processing

Many features within the same category often share the encoding of the answers but, due to the heterogeneity of the questions, an harmonization of values is still needed and addressed by the pre-processing.

Furthermore, some features are completely irrelevant for our task, e.g. the phone-number confirmation given at the beginning of the survey, and/or they have an high proportion of missing values. To prevent these variables from affecting our results, only a subset of 122 features has been selected.

We define the **informed pre-processing** as following.

1) For each selected feature, all invalid values are mapped to NaN.
2) The answers are mapped to meaningful values. E.g., for the "ASERVIST" feature, values between 1 and 87 constitute the number of times of attending an emergency room because of asthma, while 88 indicates no such episodes. Thus, we remap 88 to 0.
3) Missing values (NaN) are filled. Based on the semantic of the questions and their answers, NaNs are replaced with either specific values that suit the question (such as 0 or 1), or the mean/median of the remaining values.
4) Finally, values are normalized using min-max normalization for boolean variables and z-scores for the others.

[1] https://github.com/epfml/ml-project-1-sarcastic-gradient-descent.git

Note that the informed pre-processing treats features based on their encoding and semantic relevance for our task.[2]

We also define the **default pre-processing** strategy as a simpler alternative. It uses the whole raw dataset, mapping all missing data to -1 (which is never a valid value) and applies min-max normalization without taking into account the encoding of answers.

C. *Model Selection*

We trained several **logistic regressors** to predict MICHD. This is a natural choice to model the conditional probabilities of the target classes given the input features. We optimized the logistic loss using the L_2 regularization.

$$\mathcal{L}(w) := \frac{1}{N}\sum_{n=1}^{N} -y_n x_n^\top w + \log\left(1 + e^{x_n^\top w}\right) + \frac{\lambda}{2}\|w\|_2^2$$

Our experiments were performed using three different processed versions of the BRFSS dataset:

1) "The good", containing only our "selected" features, treated with the informed pre-processing
2) "The bad", containing all the original features, treated with the default pre-processing
3) "The ugly", containing all the original features. Here, the "selected" ones are treated with the informed pre-processing and the rest with the default pre-processing

For each processed dataset, we grid-searched over the following hyper-parameters, splitting the dataset intp a training and a validation set, with a proportion of 90-10.

- The regularization coefficient:
$$\lambda \in \{0.00001, 0.0001, 0.001, 0.01, 0.1\}$$
- The step-size for the descent:
$$\gamma \in \{0.01, 0.05, 0.1, 0.5, 1\}$$
- The batch size:
$$b \in \{500, 5000, 10000\}$$

We consistently used mini-batch gradient descent, training every model for 5000 epochs and keeping the parameters that achieve the lowest loss on the validation set, which is computed without the regularization term.

Finally, for each learned model we compute the decision threshold looking at the validation set. To do so, we linearly searched the decision threshold that maximizes the F_1 score of our model, iterating over all the values in [0, 1] using a 0.005 step.

III. Results and Discussion

Table I: Best models performance on validation set. The best submission was "The ugly" ($\gamma = 0.5$, $\lambda = 0.0001$, $b = 10000$), achieving $F_1 = 0.442$ on the test set.

Model	Acc.	Prec.	Rec.	F_1	Threshold
"The good"	86.78%	34.99%	**57.96%**	**43.63%**	0.186
"The ugly"	**87.83%**	**36.91%**	53.26%	43.60%	0.206
"The bad"	85.95%	33.07%	57.71%	42.05%	0.181
Less-features	0	0	0	0	0

The validation statistics for the best models that resulted from the grid-searches are reported in table I. Our main observation is that the different pre-processing strategies and the feature selection do not lead to big difference in terms of the models' performance. The best results on the test set have been achieved using "The ugly" pre-processing.

However, we investigated the role of each feature by looking at the weights of the best models for "The ugly" and "The good". The question is: "what are the real benefits of a good pre-processing?"

Pre-processing comparison: Our aim is figuring out which features are associated with bigger absolute weights. This is an indicator of how important each feature is to the model's prediction. We found major differences in the features associated with the biggest weights. We hereby report the top 5 features for both "The good" and "The ugly" best models. The key difference is the presence of confounders (highlighted in red) for "The ugly", which hinder the interpretability of the model.

"The good"

Feature	w	Expl.
_AGE80	0.82	Discretized age of the respondent
HAREHAB1	0.72	Had rehabilitation after hearth attack
GENHLTH	0.43	measures the general health
SEX	-0.42	Sex of the respondent
PA1VIGM_	-0.39	measures weekly physical activity

"The ugly"

Feature	w	Expl.
HAREHAB1	0.61	Had rehabilitation after hearth attack
_FRTRESP	0.60	missing fruit responses
_FRUITEX	0.55	missing/out of range fruit responses
CTELNUM1	0.45	confirmation correctness phone number
_RFHYPE5	0.44	High blood pressure

IV. Summary

Despite similar performance for the different pre-processing strategies, "The good" reaches the best recall and enables us interrogate about the impact of a feature associated to a heavier weight. Such difference, with respect to "The ugly", really eases further analyses on how such detected factors can be related to our target. With that being said, we support the effort made for a careful pre-processing, to open up the possibility of further investigating cardiovascular diseases.

[2] "/data_preparation/features_info.py" on github repository for details