Implementing Machine Learning Models and Predicting the Risk of Developing Cardiovascular Diseases

Madeleine Hueber
School of Computer & Communication Sciences, EPFL, Switzerland

Adrian Martinez Lopez
School of Computer & Communication Sciences, EPFL, Switzerland

Duru Bektas
College of Management of Technology, EPFL, Switzerland

***Abstract*—This project aims to use machine learning techniques to predict the risk of developing cardiovascular diseases (CVDs), using the data from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. The study discusses about the challenges of data preprocessing, such as feature engineering and data cleaning and modelling. Since the dataset presented imbalances, undersampling techniques have been applied to enhance detection over the minority class. For inference, a ridge logistic regression model is applied. The model achieves a promising result for CVD risk prediction with an F1 score of 0.4275 and an accuracy of 0.878.**

I. Introduction

Cardiovascular diseases (CVDs) are one of the main reasons of death, responsible for an estimated 17.9 million deaths in 2019, which makes for 32% of total deaths [1]. More than 75% of deaths caused by CVDs occur in countries with low or middle income, where there are limited access for healthcare, which also underlines the relevance of decisive tools for detecting and preventing the CVDs [2].

The aim of this project is to use machine learning techniques to estimate the tendency of a person to develop CVDs, using the BRFSS survey done across US [3]. With the model, the main goal is to contribute to the detection of CVDs which can be beneficial for global health.

II. Models and method

A. Data preprocessing

Data processing presented a significant challenge in this project, due to the size and complexity of the dataset, which included over 300 features and 300,000 samples.

Merging Landline and Cellphone Features: The BRFSS employs two distinct types of interviews based on the phone type—landline or cellphone. Consequently, certain variables have separate versions for each interview type, containing missing values when they do not correspond to the respective phone type. To address these issues, their values were merged into a single feature containing the information of both.

Abnormal Dataset Values: The *BRFSS 2015 Codebook Report* presented many anomalies and discrepancies among all of its features, according to a preliminary analysis. Abnormal values, like 77 or 99, were present in some variables and indicated missing data for various reasons. Furthermore, these values were not consistent between features.

These values have been dealt with by performing text analysis over the report to identify the abnormal value labels described with keywords such as *Don't know*, *Refused*, *Missing*, and so forth. They have been replaced by NANs or 0s depending on the context of their descriptions.

Feature Type Detection: The type of a feature determines how it should be treated and how its information is conveyed. This process is relatively straightforward: binary features are identified by counting the number of unique values, while continuous features are identified by the presence of non-integer values.

The remaining attributes fall into one of three categories: numerically discrete, ordinal, or categorical. A manual evaluation has been performed on this set of features, as the descriptions of their value labels play a significant role in determining their types

Treatment of NAN values: The dataset also contained a considerable number of NAN values. These can be mainly attributed to unasked questions due to previous answers or the state of residence. However, due to a shortage of time, the causes of NAN values have not been fully explored. In order to deal with them, binary and categorical features have been filled with the mode of their train data, continuous features with the mean and discrete numerical and ordinal ones with the discrete mean.

Feature engineering: The feature engineering process was executed in a relatively straightforward manner, with binary encoding (less columns than OHE) applied to categorical features and standardisation of numerical data. Additionally, the logistic regression bias was incorporated into the dataset.

B. Feature selection

After preprocessing the data, we obtained a dataset containing over 300 features, among which we identified both redundancy and irrelevance. Therefore, feature selection was necessary to keep only those features with a meaningful impact on the outcome, in order to improve our model.

To do so, a correlation matrix was generated (Fig. 1) to examine the relationships between features. It showed that pairs of features with significantly high correlation coefficients did exist, although they were a minority.

A pair (i, j) of features was defined as redundant if their correlation coefficient was markedly high ($|c_(i, j)| > 0.9$). In such instances, only the most correlated feature with the target labels was retained in the dataset.

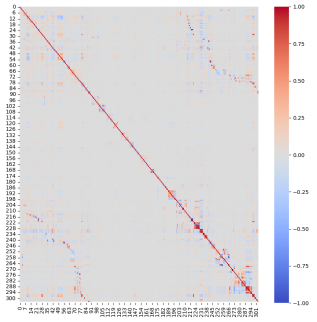


Figure 1: Correlation matrix between the features of the training dataset

C. Undersampling

Another noteworthy finding is the fact that the dataset had a ratio of 1:10 between their targets, making it **imbalanced**. The first attempts to train a model on this dataset produced a severely overfitted model that consistently predicted that patients were healthy.

Since the objective was to develop a model specially for detecting the minority class, undersampling techniques were applied to the majority class, reducing it by a factor of 4. The performance of the model improved significantly after applying this change, making it the most significant discovery in our analysis.

D. Regression model

Once the dataset was preprocessed, a ridge logistic regression model was implemented with Gradient Descent in order to do perform inference. This model is defined by a weight vector $w \in R^D$ determined during the training phase and performs inference with $y = 1_{\sigma(x^T w)}$ where σ the sigmoid function.

To train the model, we used the training data $(x_n, y_n)_{n=1}^N$, along with the regularized logistic loss defined as follows :

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

with λ the regularization parameter. At each step of the training the weight w is then updated :

$$w^{t+1} = w^t - \gamma \nabla L(w^t)$$

with γ the learning rate for our model.

To prevent overfitting, we split our training dataset into a training set (80%) and a validation set (20%). We then trained our model on the training set for up to 1000 steps, stopping early if the difference between validation loss and training loss fell below a threshold ($1e - 6$). Finally, we evaluated over the validation set plus the samples removed with undersampling, to increase the local testing set.

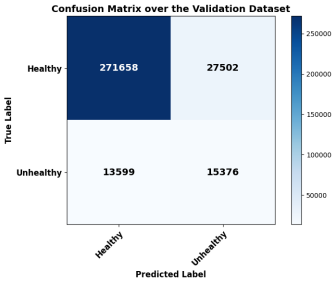
III. Results

The performance of the model depends on the optimization of three hyperparameters: γ, the learning rate, λ, the regularization pattern and ϕ, the undersampling ratio of the majority class. We have performed a grid search of these parameters: for each combination we compute a 5-fold cross-validation, undersample with ϕ and retrieve a mean F1 score.

As shown in figure 2, our optimal hyperparameters are ϕ = 0.25, γ = 0.2 and λ = 0.01 with a mean F1 score of 0.4275. Some combinations have been omitted for the sake of clarity. These results are quite similar to the ones obtained in the AICrowd Server final submission, which obtained 0.439 F1 score.

ϕ	γ	λ	F1 Score
0.1	0.1	0.1	0.3723
0.1	0.1	0.05	0.3747
0.1	0.1	0.01	0.3788
0.15	0.15	0.1	0.4055
0.15	0.15	0.05	0.4076
0.15	0.15	0.01	0.4107
0.2	0.2	0.1	0.4145
0.2	0.2	0.05	0.4197
0.2	0.2	0.01	0.4250
0.25	0.2	0.1	0.4090
0.25	0.2	0.05	0.4178
0.25	0.2	0.01	0.4275
0.3	0.2	0.1	0.3929

(a) Mean f1-score of different cross-validation executions



(b) Confusion Matrix of the resulting model

Figure 2: Cross-validation results and confusion matrix of the best model

IV. Discussion

While there is yet a lot of work that could be done to improve the performance of the model, such as performing a more thorough data processing or leveraging more sophisticated learning algorithms, the results are pretty satisfactory and offer a promising foundation when we measure the improvement in performance respect to the first submission.

Taking into account that we are tackling an imbalanced dataset, it is not an easy task for the model to detect the underrepresented class, which also turns out to be the most important.

We believe that the displayed strategies have proven to be impactful since they have had a significant impact on the performance of the model.

V. Summary

This research presents the effectiveness of a machine learning approach for predicting CVD risk in an early stage from the data of the BRFSS survey done in 2015. This article overviews different data processing and undersampling techniques to improve the quality of the predictions.

A promising F1 score of 0.4275 was obtained even with a simple ridge logistic regression model, demonstrating the influence that data processing has in resolving these types of issues. Despite the faced challenges, this study demonstrates that machine learning methods can be useful in the early prediction of CVDs to save lives.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases (cvds) fact sheet," 2023, accessed: 2024-10-31. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[2] ——, "Global health estimates 2021: Life expectancy and leading causes of death and disability. methods and data sources," World Health Organization, Tech. Rep., 2021, accessed: 2024-10-31. [Online]. Available: https://cdn.who.int/media/docs/default-source/gho-documents/global-health-estimates/ghe2021_cod_methods.pdf?sfvrsn=dca346b7_1

[3] Centers for Disease Control and Prevention, "Behavioral risk factor surveillance system (brfss) 2015 annual data," 2015, accessed: 2024-10-31. [Online]. Available: https://www.cdc.gov/brfss/annual_data/annual_2015.html