Machine Learning Project 1

Valentine Delevaux, Axel Croonenbroek, Julie Le Tallec *EPFL, October 2023*

***Abstract*—Heart diseases are the top one leading cause of death in the United States of America. However, many factors such as lifestyle and medical background can highly influence heart condition and thus, based on these informations, one can anticipate the likelihood of an individual having coronary heart diseases (MICHD). Based on health-related data provided by the Behavioral Risk Factor Surveillance System (BRFSS), we developed a machine learning algorithm that aims to predict for a new dataset, whether individuals will suffer from MICHD or not.**

I. Introduction

The goal of the algorithm is to predict the likelihood of having a coronary heart disease or not. We implemented six different algorithms to attempt to reach the best classification prediction possible. We ultimately kept the most efficient, accurate and reliable model, Regularized Logistic Regression.

We pre-processed the dataset via data cleaning, feature expansion, data splitting and searched for the best parameters in order to obtain the most accurate prediction.

II. Data pre-processing

A. Data cleaning

With a first glance at the dataset, we observed that it is composed of 437 514 samples, with 321 features. 328 135 samples were used for training, and the rest for the final evaluation of the model. Visualization of the features allowed us to notice that some of them were categorical and other numerical. We started the cleaning with very basic but necessary steps, by assessing the absence of duplicates among the sample and shuffling the dataset for randomization. We then replaced missing information with the median of the features's values. This avoids creating new categories among the categorical features, like using the mean would have done.

We normalized the data and thereby, gave an equal weightage to all features, while handling potential outliers and improving the model performances.

In addition, we removed features from the dataset, notably the ones with informations that seemed useless for the learning algorithm (date of sample's collection, Id, ...), or with a null standard deviation. Columns with null standard deviation have only one common value, therefore they are not useful to distinguish points among the dataset. Following this, we used a correlation matrix with the Pearson product moment correlation to remove the features that had a very high correlation coefficient, since they would provide us the same kind of information.

B. Feature expansion

After cleaning the data, we expanded the numerical features of the dataset, to improve the model performance. We tried both a polynomial and cosine expansion, leading us to different results and accuracy.

We observed better results with the cosine expansion. Moreover, this type of expansion has a lower computational complexity compared to polynomial expansion of degrees higher than 2.

C. Data splitting

We splitted our data between training and testing data, with a 80:20 proportion (80% for training, 20% for testing). We inspected the occurence frequency of the 2 possible outputs (1 corresponding to a positive prediction for MICHD and -1 a negative one) (Fig.1) and noticed that the proportion of 1 was considerably smaller, meaning that our dataset was very unbalanced.

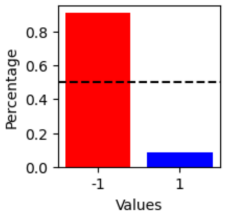

Fig. 1: Frequency distribution of -1 and 1 outputs (-1 : negative MICHD prediction; 1 : positive MICHD prediction)

We tested our model on the unbalanced dataset using cross validation.

In parallel, to prevent a bias towards negative MICHD prediction, we splitted our training dataset in multiple balanced subsets. Each of these subsets contains the same set of samples with output 1 but a different set of -1 output samples. In other words, this method allowed us to undersample the majoritary class without loss of information. However, to test the models in conditions corresponding to the initial dataset, we kept the testing samples unbalanced.

Testing both options led us to the conclusion that the training with balanced dataset was more efficient for the predictions.

III. Models and methods

We tested our different learning algorithms on the balanced data subsets, in order to choose to most efficient model.

A. Testing models

To tackle our classification problem, we implemented six different algorithms:

- Linear regression using gradient descent (GD)
- Linear regression using stochastic gradient descent (SGD)
- Least squares regression using normal equations (LS)
- Ridge regression using normal equations (Ridge)
- Logistic regression using gradient descent (Log)
- Regularized logistic regression using gradient descent (Reg Log)

In order to test the algorithms in their respective optimal conditions, we searched for their best parameters, using grid search. The best parameters correspond to the threshold for prediction making, and hyperparameters of some of the algorithms. The training was made on the balanced subsets, while the testing was based on the unbalanced dataset, ensuring real condition results. We based our choice of model on F1 score and accuracy. The results of the tests are shown in 2.

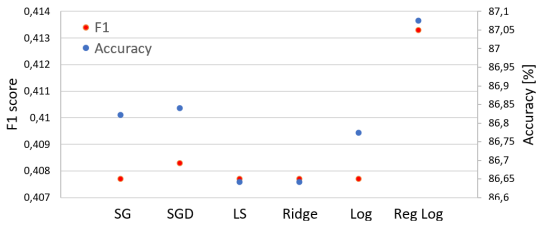

Fig. 2: Accuracy and F1 score of the six learning algorithms after best parameters grid search

B. Regularized logistic regression combined with least squares

According to the testing results, we chose to base our model on a combination of two learning algorithms. First, using Least squares regression with normal equations, allowed to find initial weights. These initial weights were then used in the Regularized Logistic Regression with gradient descent, to obtain the final weights for prediction. Indeed, Regularized Logistic Regression showed the best accuracy and F1 score. Moreover, this type of algorithms is suitable for binary classification problems with complex datasets.

As discussed above, we performed grid search to find the best hyperparameters lambda and gamma for this model by minimizing the loss, with training on balanced subsets and testing on an unbalanced subset. The results of the grid search are shown in 3.

C. Prediction making

To generate a prediction for our testing dataset, we used the sigmoid function with the obtained weights. This gave us the probability (value between 0 and 1) of each output to be classified as 1 or -1. In this situation, it was useful to chose a threshold that is not 0.5, as the data was very unbalanced and could thus lead to bias towards predicting a negative MICHD diagnosis. We performed grid search to find the best threshold, to classify the data, using the unbalanced test dataset.

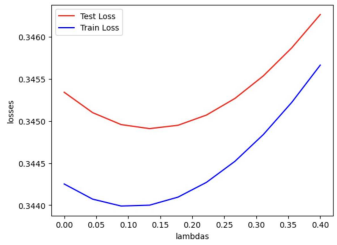

Fig. 3: Loss for the Regularized Logistic Regression algorithm depending on lambda, for the best gamma found

IV. Results

Among the six models, the model that predicted the most accurately the diagnosis of MICHD was the Least Squares algorithm followed by Regularized Logistic Regression, trained on balanced subsets. Applying the model on the provided testing set gave us a F1 score of 0.412 and accuracy of 86.9%.

Discussion

A final accuracy of 86.9% is a good result and can be an indicator of the efficiency of the model. However, as the dataset is very unbalanced, the F1 score is more relevant to judge the quality of the algorithm. The F1 score obtained with our model is good, but could have probably been improved notably with a better pre-processing of the data. We used the median to replace the missing values, but it could have been interesting to use a prediction model (a Random Tree Classifier for the categorical data for example) to replace the missing value. Moreover, treating those values differently for categorical and numerical features could also be an improvement point.

A better identification and handling of the outliers could also lead to a better model. The normalization of the data helps us minimizing the effects of the outliers, but it is not the most effective way to deal with them.

The balance of the dataset and how we approach the problem could also be improved. Our subset method allowed us to work with balanced data but still led to undersampling of the dataset, since we always use the same set of 1 output samples. Using weights on the minority class, so the training model pays more attention to it, can be an efficient way to offset the unbalance of the dataset.

Finally, the determination of the threshold based on the training dataset is questionnable. It worked well in our case, as the testing data seem representative of the data the model was trained on. However, this could be less suitable in the situation where the training and testing sets are very different.

Summary

In this project we achieved to create a model that predicts whether an individual has a risk of suffering from MICHD, based on lifestyle and medical background information. Nevertheless, we found valuable insights for enhancing the predictive model, which we could implement in order to achieve the highest level of efficiency.