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 information, 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 data set, whether individuals will suffer from MICHD or not.**

I. Introduction

The goal of the classification 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. We ultimately kept the most efficient, accurate and reliable model, Regularized Logistic Regression.

We pre-processed the data set 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 data set, 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 assessed the absence of duplicates among the samples and shuffled the data set for randomization. We then replaced missing information in each column with the median of the feature's values, which is robust to outliers and avoids creating new categories among the categorical features.

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 data set, notably the ones with information that seemed useless for the learning algorithm (date of sample's collection, Id, ...). We also removed features with a null standard deviation, having only one common value and thus not useful to distinguish points among the data set. Following this, we used a correlation matrix with the Pearson product moment correlation to remove high correlated features that provided redundant information.

B. Feature expansion

After cleaning the data, we added one column of "1" for the bias. Then we expanded the numerical features of the data set, to improve the model performance. We tried both a polynomial and cosine expansion, leading us to different results and accuracy.

We chose cosine expansion, as it showed better results. Moreover, this type of expansion has a lower computational complexity compared to polynomial expansion of degrees higher than 2.

C. Data splitting

We split the data in training and validation samples, with a 80:20 proportion (80% for training, 20% for validation). We inspected the occurrence 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 data set was very unbalanced.

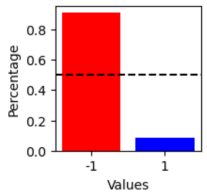

Fig. 1: Frequency distribution of -1 and 1 outputs (-1 : negative MICHD prediction; 1 : positive MICHD prediction)e

We tested our model on the unbalanced data set using k-fold cross validation.

In parallel, to prevent a bias towards negative MICHD prediction, we tried splitting the training samples in multiple smaller balanced subsets. Each of these subsets contained the same set of samples with output 1 but different sets of -1 output samples. In other words, this method allowed us to under-sample the majority class without loss of information. However, to test the model in conditions corresponding to the initial data, we kept the validation data set unbalanced.

Testing both options led us to the conclusion that the training with balanced data sets was more effective for the predictions.

III. Models and methods

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

A. Testing models

To tackle the 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 include the threshold for prediction making, and hyperparameters of some of the algorithms. The training was made on the balanced subsets, while the validation was based on the unbalanced data set, ensuring real condition results. We based our choice of model on F1 score and accuracy. The results of the tests are shown in Fig. 2.

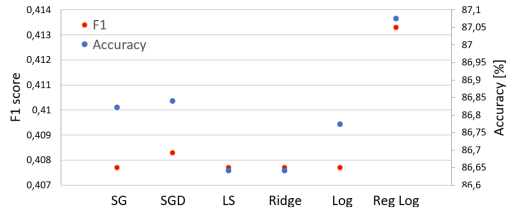

Fig. 2: Accuracy and F1 score of the six learning algorithms after best parameters grid search

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. Regularized Logistic Regression showed the best accuracy and F1 score. Moreover, this type of algorithm is suitable for binary classification problems with complex data sets.

As discussed above, we performed grid search to find the best hyperparameters λ and γ for this model. This was achieved by minimizing the loss, with training on balanced subsets and cross validation on an unbalanced subset, representative of the initial data set. The results of the grid search are shown in Fig. 3.

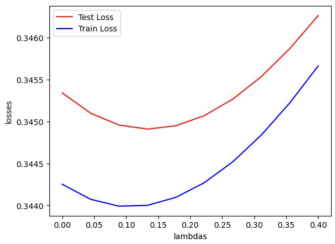

Fig. 3: Loss for the Regularized Logistic Regression algorithm depending on λ, for the best γ found

C. Prediction making

To generate a prediction for our testing data set, 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 choose a threshold that was not 0.5, as the data were 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 validation data set.

IV. Results

Among the six models, the model that predicted most accurately the diagnosis of MICHD was the Least Squares algorithm followed by Regularized Logistic Regression, trained on balanced subsets. The best parameters found by grid search were λ = 0.133, γ = 0.022, and threshold = 0.586. Applying the model on the provided testing set gave us a F1 score of 0.413 and accuracy of 86.9%.

Discussion

An accuracy of 86.9% is a good result and can be an indicator of the effectiveness of the model. However, as the data set is very unbalanced, the F1 score is more relevant to judge the quality of the algorithm. The obtained F1 score is good, but could have probably been improved, notably with a better pre-processing of the data. We used the median to replace all the missing values, but it could have been interesting to use a prediction model to replace them. Moreover, treating those missing values differently for categorical and numerical features could also have been an improvement.

A better identification and handling of the outliers could 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.

In addition, the balance of the data set and how we approach the problem could have been improved. Our subset method allowed us to work with balanced data but still led to an undersampled data set, since we always used the same set of 1 output samples. Using weights on the minority class, ensuring that the training model pays more attention to it, could be an efficient way to offset the unbalance of the data set.

Finally, the method of determination of the threshold based on the training data set is questionable. 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 effectiveness.