ML Project 1

Giuditta Del Sarto, Lorenzo Ferri, Emanuele Rimoldi

I. Introduction

In a world where Cardiovascular Diseases (CVDs) are a serious concern for people's health, it would be interesting to use ML algorithms to predict the development of these diseases. The aim of this project is to build a model able to the presence of CVDs from clinical and lifestyle features of a patient. The dataset consists of 322 features retrieved from 328 135 individuals, and the task is a binary classification problem.

II. Data cleaning and exploratory analysis

To deal with missing data in the dataset, we started by discarding all features having more than 10% of missing values. Then, we kept only the features that were not redundant between each other and that were related to health or lifestyle, based on the provided description [1]. If there were duplicate columns, one with missing values and the other with imputed values, we only keep the latter. If there were duplicate columns, one with numerical values and the other with the same values grouped in a range, we only kept the former. If the same value had two different labels, we grouped such labels into one.

To address the remaining missing values, we standardized the dataset and, *for each missing value* in row i, we computed the L^2 norm of all rows while excluding the columns for which row i has missing values. Then, we select the $1\,000$ most similar rows to row i according to this norm, among those that do not have a missing value in the column whose value has to be imputed. Then, we replace the missing value by the median (for categorical variables) or the mean (for continuous variables) of such selected rows, ensuring that the imputed values are contextually relevant.

The resulting dataset retains all the original $328\,135$ rows and 63 columns, and the main challenge is that the patients with CVDs are a small percentage of the sample: only 0.088% ($28\,975$) of all patients have CVDs, while the remaining 0.912% ($299\,160$) does not.

III. Models and methods

We compared different regression methods:

- Unpenalized methods:
  - Linear regression with loss function Mean Square Error, optimized by full gradient descent (GD), batch stochastic gradient descent with batch size 1000 (SGD), or normal equations.
  - Logistic regression with logistic loss and full gradient descent.

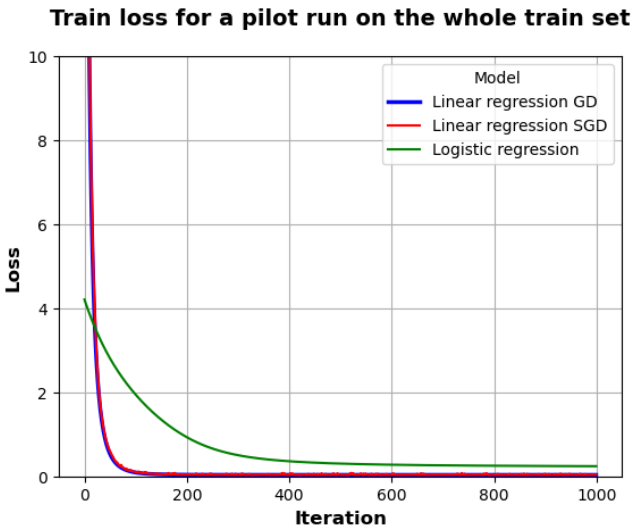

Figure 1. Loss function convergence for a pilot run.

- Penalized methods:
  - Ridge regression;
  - Logistic regression with ridge penalization.

To deal with imbalance in the response variable, we perform downsampling of the the majority class, i.e. we uniformly randomly select only a percentage p of all the patients without CVDs, while we retain all patients with CVDs. The hyperparameter p will be selected via cross-validation.

The regularization parameter λ, when used, will also be tuned by cross-validation. Each cross-validation will be performed over 4 folds, providing a good balance between variance and computational efficiency, each one retaining the same response variable proportion as the original dataset.

For iterative methods, we use $1\,000$ as the maximum number of iterations and 0.05 as learning rate, since a pilot run on the whole train set with these parameters shows that the loss function converges for all models (Figure 1). As a consequence, when training these models on a smaller portion of the train set or when adding a penalization term, the loss will converge as well.

All these operations will be performed on the standardized dataset, while also adding a column of 1s to avoid bias.

IV. Model comparison outline

Firstly, we randomly split our data in a train set (95% of the data) that we use to compare the different models and

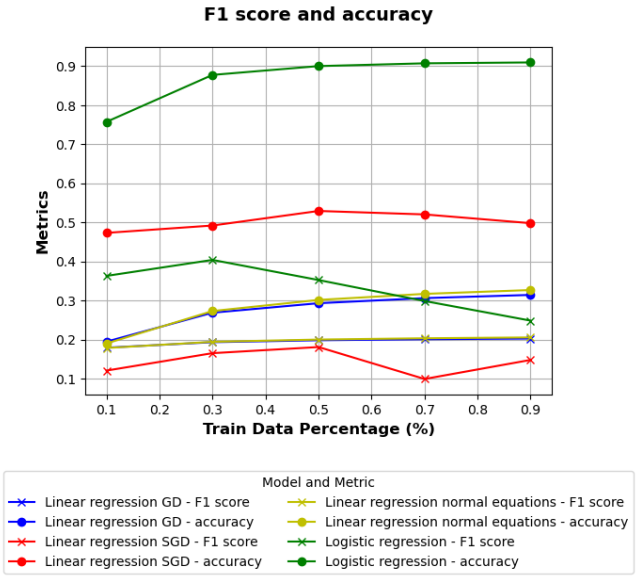

Figure 2. F1 score and accuracy comparison for unpenalized models.

to perform cross-validation, and a final validation set (5% of the data) that will be used to make the final choice for the best model. The proportion of the response values in the two sets is the same as the original dataset.

The model comparison will take into account the following metrics:

- Accuracy, measuring how many response variables are predicted correctly;
- F1 score, a metric accounting for both false positives and false negatives.

Maximizing these metrics via cross-validation and adding a regularization when performing penalized regression ensures that we avoid overfitting the data. When computing accuracy and F1 score, we consider 0.5 as the threshold to predict positive or negative, to distinguish the two classes in a balanced way.

The ranges of hyperparameters that we consider are:

- Proportion of observation from the majority class to keep p: `[0.1, 0.3, 0.5, 0.7, 0.9]`;
- Regularization parameter λ: `[0.001, 0.01, 0.1, 1]`.

V. Model discussion

We first compare the unpenalized methods, while performing cross-validation for p (Figure 2).

Accuracy and F1 scores are consistently better for the logistic regression model for any value of p, justifying the choice of this model as more preferable than the others. For this model, the best value for the hyperparameter p is 0.3, and this confirms that downsampling the majority class to train the model is an effective way to handle class imbalance.

These models could be further improved by incorporating ridge regularization, in order to reduce overfitting and handle

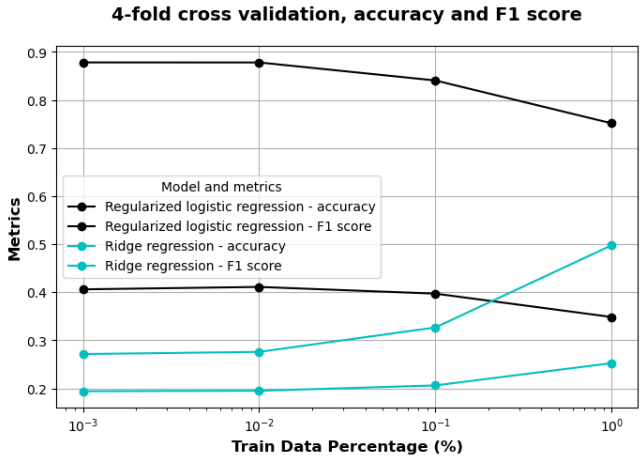

Figure 3. F1 score and accuracy comparison for penalized models.

the large number of features. In particular, we consider ridge regression for $p = 0.9$, since this was the best value of p for the F1 score of linear regression models, and logistic regression with ridge regularization for the optimal value $p = 0.3$.

Penalized logistic regression displays better metrics than ridge regression for any value of λ, confirming the fact that logistic regression is a preferable method for this kind of task than linear regression (Figure 3).

Then, for the final choice we compare the best logistic regression model ($p = 0.3$) with the best penalized logistic regression model ($p = 0.3$, $\lambda = 0.01$) on the test set (Table V).

	F1 score	Accuracy
Unpenalized logistic regression	0.404	0.868
Penalized logistic regression	0.411	0.866

The performance appears to be quite similar, although slightly better for the penalized model, at least for the F1 score.

VI. Conclusion

To tackle the binary problem of predicting the presence of CVDs in the case of unbalanced data, we proposed to keep only a percentage p of the majority class and then we investigated linear regression models, that displayed a suboptimal performance for any value of p and regardless of penalization, unpenalized logistic regression, and penalized logistic regression.

Altogether, logistic regression models are superior than linear regression models, while adding a penalization term slightly improves some metrics and can be deemed as preferable due to its greater generalization power.

Hence, we propose as final model a penalized logistic regression with $p = 0.3$ and $\lambda = 0.01$.

REFERENCES

[1] Centers for Disease Control and Prevention, "Behavioral Risk Factor Surveillance System 2015 Codebook Report: Land-Line and Cell-Phone Data," Centers for Disease Control and Prevention, Atlanta, GA, Tech. Rep., August 23 2016.