Machine Learning Project 1 - Generating Good Predictions on Heart Attacks

Guilherme Costa Ferreira, Mehdi Amor, Rodrigo Anjos
Machine Learning Course, EPFL, Switzerland

***Abstract*—Using data from the 2015 BRFSS Survey, this project employed Regularized Logistic Regression and Ridge Regression to predict heart attack risk. After employing an adequate preprocessing pipeline and finding the best hyper-parameters, Ridge Regression showed the best performance, with an F1 Score of 0.423, compared to 0.418 for Logistic Regression, after being tested in an aicrowd arena designed for this challenge.**

I. Introduction

In this project, a dataset of over 300,000 individuals with more than 300 features was used to predict heart attacks. Our approach began with an analysis of the dataset to understand its structure before selecting models. Initial inspection revealed several challenges: numerous NaN entries across features (figure 1A), varying types of features (categorical, ordinal), values with differing meanings depending on range, and a strong skew towards healthy individuals, typical of health datasets (figure 1B).

Given the high dimensionality, certain models, like mean-squared error gradient descent (MSE GD) and least squares, were expected to be computationally expensive, while MSE SGD could reduce costs but with noisier convergence. The models' specifications are explained in the II-B section [1].

In health datasets, accuracy alone is insufficient. Here, the priority is correctly identifying individuals with a disease or high risk of health problem, rather than maximizing general accuracy. A model predicting all patients as healthy, for instance, would yield high accuracy, yet be ineffective. Thus, emphasis was placed on the F1 score to balance precision and recall, minimizing false negatives.

II. Models and Methods

A. Exploratory Data Analysis

The preprocessing steps are as follows: to start, columns related to survey details rather than lifestyle factors (e.g. cell phone type) are removed. Next, one of each pair of highly correlated features is eliminated to prevent multicollinearity, which can negatively impact performance, and decrease data dimensionality while maintaining the information. Values with meanings that don't provide useful information, answers such as "Refused" or "Don't know", are replaced with NaN, and values in the incorrect order are adapted to be in the right place - for instance, oftentimes the value "8" relates to a meaning lower than the one from value "5". Following this, columns with more than a certain threshold of missing values are deleted. For remaining columns with missing values, NaNs are replaced with random values generated randomly, with the probability of each value equal to its prevalence on the non-missing values in each column. Categorical features, detected by having less than a threshold of unique values, are expanded into binary columns to avoid ordinal assumptions, and then all features are standardized to ensure consistent scaling across the dataset. The last step balances data skewness, randomly removing rows with negative samples (-1 in the target variable y) based on a specified probability (skew chance).

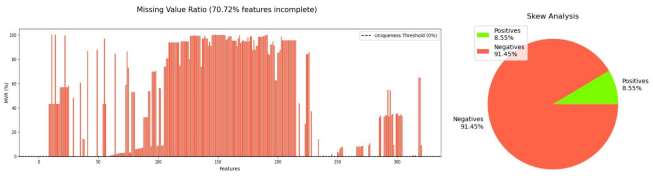

Figure 1. A (left) - Plot of all features' NaN ratios. If we removed every feature that had a missing value, we would delete 70.72% of the dataset. B (right) - Training set output skew. Most of the data is from healthy patients (91.45%) versus only 8.55% from sick patients.

B. Models

Two models were evaluated—Ridge and Logistic regressions—due to their respective strengths suited to this dataset. Ridge regression efficiently handles high-dimensional data through a closed-form solution, making it faster and less computationally intensive. However, ridge regression predicts continuous values and may be more sensitive to outliers due to squared-error minimization. Logistic regression, while slower due to iterative optimisation, is naturally suited for binary classification, directly modeling probabilities, and is less impacted by outliers. Testing both allows for the determination of the most effective model for this prediction task.

The best values for each hyperparameter: threshold for removing features with high correlation (R2), with many missing values (MV), for detecting categorical features with low unique values (CAT), skewness adjustment (S), regularization lambda (L), learning rate gamma (G) were found by grid search, using k-fold cross validation. The models were evaluated using key metrics, including F1 score and accuracy, computed on the validation set for each fold, while the mean and standard deviation of these metrics were also recorded to evaluate the model's stability.

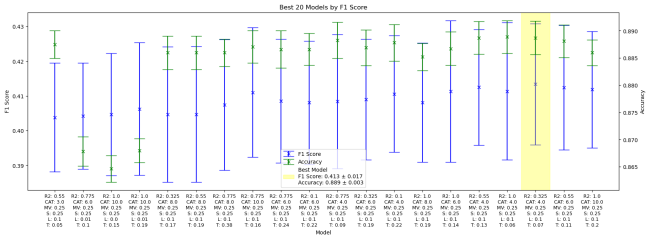

Figure 2. The twenty best hyperparameter combinations with ridge regression models, according to mean F1 score. In the second vertical axis, the accuracy is also depicted, however, it was not used to pick these models. The error bars are referent to standard error: $\sigma/\sqrt{k}$. Fitting Duration in seconds is also depicted as T

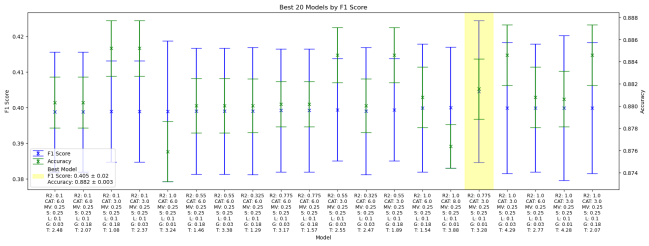

Figure 3. The twenty best logistic regression models according to mean F1 score. The labels are the same as in Figure 2 with the addition of G for gamma

Execution time for each parameter combination was also logged to account for computational efficiency. We started with Ridge Regression, as it was the quickest to run, since it has a closed form, and then ran again the grid search for Logistic Regression, varying only the hyperparameters that took different values within the 20 best models in the search with Ridge Regression, to optimise time. We used cross-validation with $k = 5$ to get the F1 Scores and Accuracy estimations. All the possible hyperparameters values that we used are in table I. They are defined in a way that the ones that were bounded, like MV, R2 and S, had the possible values either uniformly or logaritmically distributed, while the rest were defined through experimentation. CAT range was chosen with computation time in mind-thresholds bigger than 10 would add too many and too big categorical features.

III. Results

In the grid search with Ridge Regression, we obtained the results in figure 2.

We can verify that some hyperparameters didn't vary much in the top 20 models. The most consistent hyperparameters were skew chance (S = 0.25), lambda (L = 0.1) and missing value rate threshold (MV = 0.25). The rest varied more.

The results of the second grid search are in figure 3.

It is evident that the performance does not vary too much, implying that both models would perform similarly. Nonetheless, we run both in the arena and obtained the F1 scores: 0.423 (Ridge) and 0.418 (Logistic).

R2	MV	CAT	S	L	G
0.1	0	3	0	0	0.001
0.325	0.25	4	0.062	0.01	0.005
0.55	0.5	6	0.25	0.1	0.031
0.775	0.75	8	0.562	1	0.178
1	1	10	1	10	1

Table I
HYPERPARAMETERS

IV. Discussion

Our results show that Ridge regression worked better than logistic regression, although both of them are far from great in a real application, where both the f1-score and accuracy would sit between around 0.8-0.9 [2], implying our data might not be linearly separable, being this the greatset weakness of our approach. For future works, other types of models, especially non-linear models, such as random forests or neural networks could be tested. If not, at least polynomial feature expansion could be done, to improve the performance of linear models.

Ridge regression performing practically as well as the logistic regression implies that its disadvantages were nullified. The skew issue was fixed in the preprocessing, but nothing was made about outliers. This was not an issue due to the high dimensionality of the dataset, distances between all points were generally high, reducing the effect of outliers.

We consider one of the strengths of our approach to be our preprocessing, which is shown to be relevant since hyperparameters such as the skew chance and missing value ratio were always the same or similar in the best models in grid search, showing if we hadn't done it, the models would always have turned out worse.

A last observation we can make is that we saw the results from the arena were slightly better than the mean of our local results, which could be attributed to the variance of the dataset, since they even still fall within the error.

V. Summary

In this project, we explored the predictive capabilities of machine learning for heart attack risk assessment, using two primary models—Regularized Logistic Regression and Ridge Regression—following extensive data preprocessing. Both models performed similarly in terms of F1 score, though Ridge Regression showed slightly superior results.

However, our models' moderate performance suggests potential limitations in linear separability within the dataset, leading us to suggest that future research should explore feature processing such as polynomial feature expansion, or non-linear models such as random forests or neural networks, which could better capture complex relationships in health data.

REFERENCES

[1] M. J. . N. Flammarion, "Machine learning course lectures," 2024, https://github.com/epfml/ML_course/tree/main/lectures.

[2] M. Rizwan, S. Arshad, H. Aijaz, R. A. Khan, and M. Z. U. Haque, "Heart attack prediction using machine learning approach," in *2022 Third International Conference on Latest trends in Electrical Engineering and Computing Technologies (INTELLECT)*, 2022, pp. 1–8.