

HEART ATTACKS PREDICTION USING MACHINE LEARNING

Levin HERTRICH
levin.hertrich@epfl.ch

Eugénie CYROT
eugenie.cyrot@epfl.ch

Gabriel MARIVAL
gabriel.marival@epfl.ch

Abstract—Cardiovascular diseases are a leading cause of death worldwide. This report describes a machine learning-based approach for predicting heart attacks using data from the Behavioral Risk Factor Surveillance System (BRFSS). After extensive data processing and model selection, an $F1$ -score of 0.4201 and an accuracy of 0.8633 were achieved.

I. INTRODUCTION

Cardiovascular diseases, particularly heart attack, are a leading cause of death worldwide [1]. This project aims to predict heart attacks based on various health indicators obtained by the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS is an ongoing health surveillance system for adults in the United States. For this study, the optimal model for binary classification was selected after processing the medical data.

II. PROCESSING DATA

A. Data cleaning

Some questions from the BRFSS vary from state to state or are conditional upon previous answers [2]; which results in some features containing NaN values. Moreover, some non-NaN values represent missing answers or cases where the respondent was unable to give an answer. Referring to the Codebook [2], we manually identified these values for each feature and replaced them with NaN values. We only did this for the features we manually selected and deemed important, meaning we assume they would impact the risk of a heart attack. For handling missing values, we decided to use the mean value of each feature. Finally, we removed constant columns as they do not provide any useful information.

B. Feature selection

The BRFSS survey contains 321 questions, and thus, many irrelevant questions for our study such as the phone number of the respondent. Therefore, we manually selected 105 features we deemed important based on their description, using the Codebook of the BRFSS [2]. We then performed a Pearson correlation, selecting the most correlated features among the manually selected ones. We tested our dataset with the ridge-regression model at various ratios: $\{0, 0.25, 0.5, 0.75, 1\}$. Selecting 75% of the data yielded the best results, with an $F1$ -score of 0.40685.

C. Re-balancing the data

The dataset is highly imbalanced, in reality most people won't have a heart attack. Around 90% of the individuals are not expected to have one. We thus have chosen to re-balance the dataset by up-sampling the minority class to obtain a 50%-50% balance between the two classes. Up-sampling was performed by duplicating the data-points from the minority class randomly, until reaching a 50-50 balance. A 50-50 balance results in an improvement of the $F1$ -score from 0.213 to 0.413 for a given ratio of 0.75 and threshold of 0.5 for ridge regression model.

III. MACHINE LEARNING MODELS

We implemented different Machine Learning models - linear and logistic regression. All of them have been fitted using optimization by gradient descent, with its associated parameter γ - the learning rate - and a hyper-parameter λ for regularized models.

A. Linear regression models

Linear regressions models involve finding the weights minimizing the following loss function :

$$\mathcal{L}(w) = \frac{1}{2N} \sum_{n=1}^N (y_n - x_n^T w)^2$$

To use linear regression for classification it is necessary to map the obtained results in the target space, which in our case is $\{-1, 1\}$. To find the optimal threshold for this mapping we selected the best value amongst $\{-0.5, -0.25, 0, 0.25, 0.5, 1\}$ according to the $F1$ -score. The maximum $F1$ -score was obtained for a threshold of 0.5.

B. Logistic regression models

Logistic regression models estimate the probability of belonging to a certain class; here, having or not a heart attack. The goal is to minimize the logistic loss :

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

In the case of regularized logistic loss :

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

Model	Step- size γ	Number of iterations N	Regularization parameter λ	F1-score	Accuracy
Linear Regression (GD)	0.1	500	-	0.4201	0.8633
Linear Regression (SGD)	0.01	500	-	0.2830	0.7459
Least squares	-	-	-	0.4199	0.8634
Ridge regression	-	-	0.001	0.4199	0.8631
Logistic regression	0.2	500	-	0.3481	0.7186
Regularized logistic regression	0.1	500	1	0.3904	0.8270

Table I
PERFORMANCE COMPARISON FOR THE DIFFERENT IMPLEMENTED MODELS

Note that we mapped the outputs y_n from $\{-1, 1\}$ to $\{0, 1\}$ for training, and then revert it back from $\{0, 1\}$ to $\{-1, 1\}$ for prediction.

C. Weights initialization

We decided to initialize the weights to constant terms. The code's running time was not a limiting factor, so initializing weights as constants should not affect its convergence, (e.g we can add iterations of gradient descent)

D. Parameters selection

To select the best parameters, we used cross validation with k-fold ($k = 4$ in our case), running 500 iterations per fold. As we did up-sampling, we have many data-points (nearly double); thus, performing many folds and iterations wasn't necessary. We then choose the parameters which give us the best F1-score.

IV. RESULTS

After applying cross-validation, we found that gradient descent, least squares and ridge regression performed the best. Gradient descent showed slightly better performance, thus, it was chosen as our final model.

Least squares and ridge regression perform similarly, which also explains the small optimal regularization weight γ we obtained for ridge regression.

Stochastic gradient descent performed poorly compared to the other models, most likely due to the small step size we selected for the up-sampled data.

It is noteworthy that logistic regression performed worse than linear regression models, despite the fact that they should be well suited for this classification task. Regularized logistic regression outperformed logistic regression, suggesting that the logistic regression tends to overfit in this scenario. This observation is also supported by the relatively high regularization weight $\gamma = 1$ obtained for regularized logistic regression.

V. DISCUSSION

As mentioned above, one of the main issues we encountered was the unbalanced data. Therefore, a naive model that predicts only -1, would have really good accuracy but would not provide any valuable predictions. Moreover, for heart attacks, minimizing false negatives is a primary concern. To

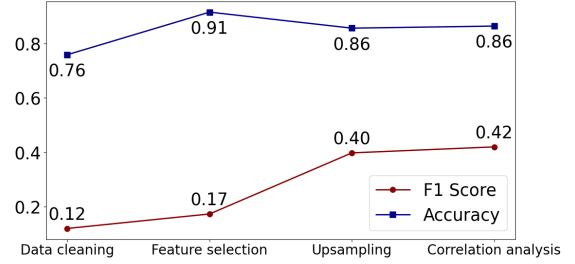


Figure 1. Evolution of F1-score and accuracy

tackle this problem, we focused on maximizing the F1-score and up-sampling the data - in order to artificially increase the proportion of heart attack prone people in our training data. This leads to a model that emphasizes learning from individuals with a disease.

To use the linear regression models effectively for classification it was crucial to identify an optimal threshold to map the results into the target space of $\{-1, 1\}$. As displayed in I with up-sampled data and a threshold of 0.5, gradient descent performed best for this task.

Improvements in model/feature selection could have been done by implementing other algorithms such as PCA (Principal component analysis) or Random Forest. Indeed, those can be really efficient, especially in this case, where feature selection plays a huge role as the data contains a lot of features which are most likely not relevant for the task.

VI. CONCLUSION

This project focused on the prediction of heart attacks based on the BRFSS medical data. To obtain good results, we performed extensive data cleaning, feature selection and data re-balancing by up-sampling the minority class. After searching for the best threshold to map results obtained by linear regression we found that gradient descent, with a threshold of 0.5, performed best for this dataset with a F1-score of 0.4201 and an accuracy of 86.33%.

REFERENCES

- [1] World Health Organization, “The top 10 causes of death,” <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>, 2021, accessed: 2024-10-30.
- [2] Behavioral Risk Factor Surveillance System, “2015 codebook report land-line and cell-phone data,” https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf, 2015, accessed: 2024-10-30.