Predicting Myocardial Infarction and Coronary Heart Disease Risk

Amene Gafsi, Rosa Mayila, Elias Mir
Department of Computer Science, EPFL, Switzerland

***Abstract*—This paper presents a classification model that combines several pre-processing tools and feature engineering for a robust heart attacks prediction based on the the Behavioral Risk Factor Surveillance System (BRFSS) dataset. The primary focus is tuning hyperparameters for ridge and logistic regression, assessing model performance on both undersampled and original datasets, and optimizing the decision threshold for maximal F1 score.**

I. Introduction

Prediction is critical in many fields, from healthcare diagnostics to financial fraud detection, where clear, reliable predictions influence important decisions. Selecting an appropriate model and preprocessing strategy significantly affects performance, especially when working with large datasets prone to missing values and class imbalances.

In this project, we develop a regularized logistic regression model to classify binary outcomes, focusing on techniques that improve prediction accuracy and interpretability.

Our approach is tested on a dataset containing demographics and health-related entries, evaluating model performance using accuracy and F1-score as benchmarks. This study demonstrates the advantages of combining regularized logistic regression with effective preprocessing to achieve robust performance in real-world binary classification.[1].

II. Data Preprocessing

The dataset, obtained from the Behavioral Risk Factor Surveillance System (BRFSS), consists of 328,135 samples representing U.S. patients, with 321 distinct features ranging from personal identifiers, such as phone numbers, to behavioral information, such as alcohol and tobacco consumption. Each patient is assigned one of two labels: (1) indicating that the patient suffers from heart disease, and (-1) indicating the absence of the condition.

A. Missing values handling

The dataset contains numerous missing values. Handling them was necessary in order to improve the model and avoid computational issues. The approach was to replace missing values with the median of the column they belong to. Unlike the mean, the median is less sensitive to extreme values which is more suitable for this dataset.

[1]Detailed description of the dataset can be found at https://www.cdc.gov/brfss/annual_data/annual_2015.html.

B. Class imbalance handling

Due to the imbalanced distribution between the -1 and 1 labels (representing "no heart attack" and "heart attack," respectively), our process manages to balance the dataset by undersampling the majority class. It randomly removes a portion of records from the dominant class. **Importantly, the original, full dataset is retained separately for testing and evaluation purposes to ensure a reliable measure of model performance.**

C. Normalization

The dataset's features are on different scales, which could cause certain features to dominate the model's learning process. To ensure that all features contribute equally, Z-score normalization is applied to standardize the features. [2]

D. Polynomial feature expansion

To address underfitting observed in preliminary testing, we employ polynomial feature expansion with degree 2 without interaction between samples to avoid computational cost. That way, we are able to significantly increase the model complexity and its performance.

III. Model training

During our project, we encounter several computational challenges when training linear models, particularly using methods like least squares and ridge regression. One primary issue is **matrix singularity**, which poses significant problems for calculating model parameters. Matrix singularity in linear models, especially with the least squares method, arises when the matrix $X^T X$ (where X is the feature matrix) is non-invertible. This situation is common in high-dimensional data or datasets with multicollinearity, where the features lack linear independence.

A. Challenges in Training with Least Squares and Ridge Regression

Due to these computational issues, we can only successfully train our model on the preprocessed data using ridge regression. Ridge regression mitigates the singularity issue by adding a regularization term to the objective function, stabilizing the optimization process. However, while ridge regression improves model stability, interpreting the results remains challenging.

[2]Data codebook shows how features are coded : https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf

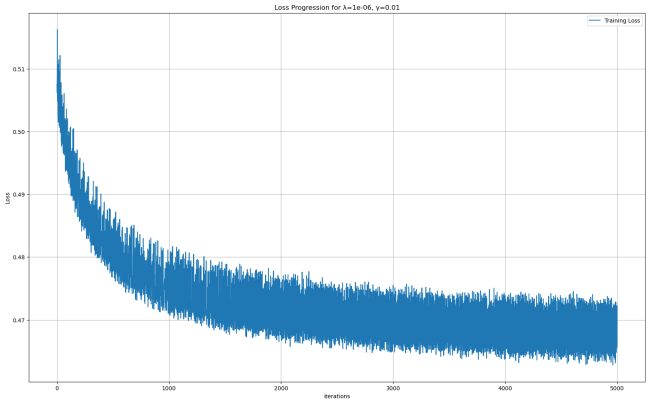

Figure 1: Regularized logistic regression loss evolution during training.[3]

B. Regularized logistic regression

After evaluating the performance of various models, we choose to proceed with the regularized Logistic Regression model, as it demonstrates the best balance between accuracy and probabilistic interpretability for our dataset.

C. Optimization and Hyper-parameter Tuning

Three optimization techniques are used to train and fine-tune these models hyperparamers. Grid search is used to tune Ridge regression and logistic regression parameters. In the other hand, a 5-fold cross-validation with 150 iterations for the gradient descent is used to optimize regularized logistics regression parameters.

The optimal parameters are : $\lambda = 1e^{-5}$ for Ridge regression, $\gamma = 0.02$ for logistic regression and the tuple ($\lambda = 1e^{-6}$, $\gamma = 0.01$) for regularized logistic regression.

We also adjust the prediction threshold. Tuning the prediction threshold has been extremely powerful in this project.

IV. Results

After completing the preprocessing and hyperparameter tuning steps, we train the models to optimize the weight vector w that best fits the data. Table I, II, and III show the results obtained for each mentioned models while (Figure 1) shows the loss evolution during training.

For all the models, tuning the threshold has a strong impact on the F1 score, especially on the original(imbalanced) data. This indicates threshold tuning efficiency to improve the positive class detection. Each model has its own sensitivity to threshold changes which reflects how the preprocess steps influence the optimal threshold choice. Lower threshold tend to be more beneficial when model struggle with detecting positives cases.

Accuracy remains high on original data but does not correspond to F1 score particularly for ridge regression. For this model, the high accurracy and low F1 score on original data strongly suggests that the accurracy result is due to the data imbalance. Conversly processed data accuracy is lower than F1 score which shows that preprocessing enhances model's ability to detect true positives and true negatives.

Ridge regression performs poorly on the original data but shows improvement with threshold tuning, indicating that ridge regression might benefit from data balancing techniques to optimize its performance in real-world applications.I

	Accuracy	F1 score	F1 score tuned	Threshold
processed data	0.54	0.69	0.80	0.60
original data	0.91	0.17	0.4293	0.66

Table I: Ridge regression results

	Accuracy	F1 score	F1 score tuned	Threshold
processed data	0.74	0.75	0.76	0.05
original data	0.77	0.34	0.38	0.98

Table II: Logistic regression results

	Accuracy	F1 score	F1 score tuned	Threshold
processed data	0.72	0.78	0.79	0.40
original data	0.87	0.37	0.42	0.72

Table III: Regularized logistic regression results

When evaluated on the test set, our model achieves an F1 score of 0.436 and an accuracy of 0.877. While the accuracy is high, the low F1 score suggests a higher-than-desired rate of false negatives or false positives. In healthcare, the F1 score is crucial as it reflects the model's ability to correctly identify unhealthy individuals.This underscores that the choice of evaluation metric is domain-specific.III

V. Discussion

The model we implemented is robust against outliers and extreme values. However, it may lack complexity as it underfits slightly, achieving 72% accuracy on the training set but 87% accuracy on the test set on AIcrowd. To improve performance, we could consider adding interaction terms between features or increasing the degree of polynomial expansion; however this would also increase the computational cost.

VI. Summary

This project aims to predict heart attack risk using healthcare data. We preprocess the data to handle missing values, standardized features, and address class imbalance. After evaluating several models, we focus on regularized Logistic Regression, optimizing it through feature engineering and hyperparameter tuning. Our model achieves competitive accuracy and F1 scores on the AIcrowd platform, demonstrating the effectiveness of machine learning in healthcare risk prediction.

[3]The oscillation is due to a large learning step $\gamma = 0.01$, which we maintained for faster convergence.