Project 1 report

Yurshevich Daniil, Debouvry Marine, Pollet Arthur

***Abstract*—In this research project, we conducted an in-depth analysis of a substantial dataset derived from interviews with over 300,000 patients in the United States, with the primary objective of predicting the risk heart diseases. After extensive preprocessing, logistic regression, random forests and decision tree models were implemented.**

I. Introduction

In this project, a dataset based on interviews of more than 300 000 patients in the US was used. The interviews were conducted in order to predict the possibility of heart attacks. The goal of the project was to apply machine learning concepts and algorithms in order to predict the risk of an individual to develop a cardiovascular disease in the future.

II. Models and Methods

A. Pre-processing of data

The provided dataset was composed of an input dataset of around 300 features.They were treated using the following steps, in the following order:

1) **Feature transformations:** According to the Codebook Report provided with the dataset, certain numerical values are associated to answer types in the dataset. In particular, the numbers 7, 77, or 777 were associated to the answer "Don't know/Not Sure", and 9, 99, or 999 to "Refuse to answer". These types of answers were considered equivalent to a missing answer and were thus transformed into "NaN" values. Similarly the numbers 8, 88, or 888 had been associated to "None" answers, when the answer was supposed to be numerical, and were changed to 0. Many features contained these answer types and were treated using a function that works for features in general. The features that didn't follow the same scheme were manually isolated and modified differently.
   Other feature transformations included unit changes, for example from months to days. The data set's final features were "calculated variables". These features were calculated according to the previous features. If all was correct, this therefore ended up in features with correlated values, correlated features were eventually removed according to point 5.
   Finally, features that directly stood out in the Codebook as useless were removed for simplicity. These features were deemed irrelevant to the problem. These manual changes though were for a minority of features.
2) **Standardization:** Data was standardized ignoring the NaN values for each feature.
3) **Removal of features with insufficient data:** A threshold value α was fixed. If more than $\alpha \cdot N$ data points were missing for that feature, the whole feature was removed. A point was said missing when the value was "NaN".
4) **Replacement of NaN values to mean of each feature:** All remaining NaN values were replaced by the mean of the feature. This yielded similar or better results than replacing with the median of the feature or to zero.
5) **Removal of correlated features:** The correlation matrix was calculated in order to detect features that correlated. If it was the case, only one of the concerned columns was kept. Features were considered correlated when the correlation coefficient between them exceeded 0.7.
6) **Removal of features with low correlation to y:** For every feature, the Pearson correlation coefficient r with the output y was computed and a minimal r_{min} was fixed. If $r < r_{min}$, the feature was removed.

The values for α, σ_{min}, and r_{min} were chosen in order to achieve a tradeoff between a too large loss of information and a too large computational cost. Different values were tested in order to achieve better a model.

Attempts were made to implement feature expansion to give more weight to "better" features. To do so, the values of the features with highest correlation to the output variable were raised to a chosen degree. The effectiveness of this technique is commented later in Part III.

Following the previously described steps, a resampling technique was applied to the dataset with the goal of achieving a balance between the number of data points labeled as '1' and '0' before starting the model training process. This was done to avoid an overwhelming prediction of 'O's of our model.

Subsequently, the remaining '0' labeled data points, were allocated for use in the cross-validation procedure, with a specific emphasis on ensuring stratification.

To obtain an accurate assessment of model performance during testing, the actual proportion of '1' labeled data points was preserved, about 10%. This approach was taken to ensure that the testing accuracy closely mirrors real-world conditions.

B. Models

In this study, a selection of conventional machine learning techniques were employed to address the problem at hand. In addition to the logistic regression method, which was introduced in the course, the performance of decision trees and random forests using predefined hyperparameters was also implemented. If you are unfamiliar with these models, you can refer to the following resources for detailed explanations: [1] and [2].

The implementation of the decision tree algorithm involved recursively partitioning the entire dataset based on specific features and corresponding threshold values. The selection of

the "best" split was determined by minimizing the entropy, calculated as follows:

$$H(D) - \frac{|L|}{|D|}H(L) - \frac{|R|}{|D|}H(R), \tag{1}$$

Here, $L \sqcup R$ represents a partition of the dataset D, and this partition is achieved through the application of a specific feature and its associated threshold. The function H is used to measure the entropy of the dataset and is based on the distribution of class labels within the dataset. In the implementation, two variations of the entropy function were considered, namely, the Gini index and standard entropy. For more in-depth information about these entropy measures, you can refer to the following resource: [3].

In our research, we utilized Random Forests, which were implemented as a collection of Random Subspace Method (RSM) trees. These trees were trained using bootstrap-resampled data, and the prediction on unseen test data samples was obtained by averaging the predictions from all the individual trees.

The RSM mechanism involves a modification in how features are selected for the optimal split during the tree's construction. Instead of considering all available features at each split, the options are narrowed by randomly selecting a subset of features to evaluate.

In summary, Random Forests with the RSM approach provide a robust and less prone-to-overfitting ensemble model by employing random subsets of features during the construction of individual trees. This technique not only enhances the generalization performance but also ensures a more equitable use of feature information.

III. Results

The results of conducted experiments for different models are gathered in the Table I. The results obtained from penalized logistic regression align with the AIcrowd scores, providing additional validation. However, for the remaining models, we only have access to local cross-validation results (a discussion regarding their validity is presented in the section IV in the document). Upon inspecting the table, it becomes apparent that logistic regression outperforms the other models, but the two alternative models demonstrate a slightly higher level of confidence in their predictions when analyzing the probabilities associated with class 1.

Model	F1-score	Accuracy
Penalized logistic regression with $\lambda = 0.1$	0.385	0.893
Decision tree of depth 3 with entropy loss	0.63	0.29
Random forests with 50 RSM trees of depth 2	0.63	0.39

TABLE I
COMPARISON OF DIFFERENT MODEL'S PERFORMANCES.

IV. Discussion

Extensive testing was conducted using various parameters, different models, and dataset preprocessing methods. Unfortunately, none of these efforts yielded an F1-score higher than 0.38. Similarly, feature expansion did not prove itself to be efficient in addressing the problem. It's worth noting that our models not only predict the labels of unseen data but also provide probability vectors for each class. One of the parameters we were fine-tuning is the threshold (denoted as t) at which we predict a label of 1. Interestingly, our experiments revealed that even slight adjustments of this threshold toward 1 result in a decrease in the F1 score. This empirical evidence suggests that our models exhibit a high level of uncertainty in their predictions.Furthermore, an additional piece of evidence supporting this notion is the significant disparity in F1-scores between our local cross-validation and AIcrowd results. On occasion, the variance between the F1-score achieved locally and the score obtained on AIcrowd can be as high as 0.1, however, it is noteworthy that the leading submission on AIcrowd surpasses our local performance by a margin of 0.05. Moreover, upon reviewing the AIcrowd result table, it becomes evident that no one has achieved an F1-score exceeding 0.45. This suggests the possibility that our models may not be capable of effectively addressing the problem at hand. To address this issue, we may need to explore additional data properties or consider using different models for this task.

V. Conclusion

This project aimed to predict the risk of an individual to develop a cardiovascular disease in the future by applying machine learning concepts and algorithms to a real dataset.

Extensive pre-processing, including feature transformation, computation of correlation coefficients, and treatment of "NaN" values. Furthermore, a resampling technique was used to address the problem of class imbalance.

Three models were tested: Logistic regression, Random forest and Decision tree. The Random forests algorithm used a RSM approach to avoid overfitting.

The results indicated that, despite trying many parameter adjustments, no model achieved an F1-score higher than 0.38. The model predictions exhibited high uncertainty. The problem therefore requires further investigation, by considering new data properties, reinforcing feature engineering techniques, or using alternative models. Another idea to improve the model would be to implement Gradient Boosting.

In summary, this project's extensive data preprocessing and modeling efforts revealed the challenges of accurately predicting cardiovascular disease risk, and further research is needed to improve the models' performance in addressing this issue.

References

[1] Wikipedia. Decision trees.
[2] Wikipedia. Random forests.
[3] Gini Index Vs Entropy for Information Gain in Decision Trees.