EARIN Midterm Project: Heart Attack Prediction

Krszysztof Kotowski and Juan Manuel Aristizabal Henao

Warsaw University of Technology, Plac Politechniki 1, Warsaw, Poland

1 Project progress

Up to this moment, we have been able to implement so far, almost all solutions of the project and tested them on different combinations of important parameters. These combinations are as follows:

- Logistic Regression
  - C
  - Penalty
  - Solver
  - Max iterations
  - Class weight
- Decision Tree
  - Criterion
  - Max Depth
  - Min Samples Split
  - Min Samples Leaf
  - Class weight
- SVM Classifier
  - C
  - Kernel
  - Gamma
  - Class weight
- Random Forests
  - Number of Estimators
  - Criterion
  - Max Depth
  - Min Samples Split
  - Min Samples Leaf
  - Class weight
- Gaussian Naive Bayes
  - Var smoothing

All these models were tested using cross validation 4:1.

Table 1. Current best results from the experiments.

Name	Best accuracy	Best precision	Best recall
Logistic regression	0,641756988	0,356334842	0,506369427
Decision tree	0,641186537	0,407185629	0,751592357
SVM	0,641756988	0,5	0,535031847
Random forest	0,64346834	1	0,324840764
Gaussian NB	0,641756988	0	0
Random	0,494010268	0,355307263	0,506369427
Constant 0	0,641756988	0	0
Constant 1	0,358243012	0,358243012	1

2 Intermediate results

The results obtained after training the models using different combinations of the before mentioned parameters are presented below:

From these results (from Table 1) it can be seen that the accuracy of the models is around 64%, which is not a totally bad result but it still has more space to improve. The most underperforming method implemented was the Gaussian NB, with a precision and recall of 0%, indicating that the parameters need additional tuning.

In terms of precision, the Random Forest model outperformed the rest with a precision of 100% in some cases, but overall speaking from the experiments, the average precision of the models was between 30% and 40%, which is still suboptimal, and raises the need to both tune the parameters more and/or add more data to the dataset (e.g. oversampling).

In terms of recall, the best performing model was the Decision Tree with around 75%, which aids in identifying the set of parameters that make the model perfom overall well in the predictions. The other models had a recall of around 50% or less, which shows that the models are struggling to understand the patterns of people with an actual high risk of heart attack.

One significant feature that we have learned from this is that the class weight of the Logistic Regression needs to be set to balanced, otherwise the predictions were always 0.

3 Finishing Plans

Noticing how the models are suboptimal in their predictions, there is a need to find more suitable parameters for the models, as well as tuning the existing ones to better fit the data. In addition, it can be noted that the data imbalance on the predicted feature (Heart Attack Risk) may be also influencing this suboptimality in the prediction of people with actual high heart attack risk, in this case implemmenting and testing different oversampling methods may aid in finding a way to reduce the influence of this factor.