

Bram van Es

January 28, 2020

Abstract

Short description of results/methods

1 General approach

We have the following dataset:

- ECG: 646 columns with aggregated data from ElectroCardioGram's,
- CELLDYN: 40 columns
- Phenotypical: age, gender, BMI
- Clinical: historical clinically relevant information (hypertension, hyperchol, diabetes, etc.)
- heartscore: ..
- troponine: both the initial measurement and the slope obtained after a second measurement

We first handled the NaN's in the dataset; we have 5 columns with NaN's for 8% of the samples, 4 of them have a specific clinical meaning, the other is BMI. The 4 clinical features are related to the P-R interval of the ECG and the absence of these specific features can signal atrium fibrillation. For this reason we map all non-NaN's in these columns to 1 and all NaN's to 0. For the BMI we simply take the median given it's low variance in the dataset, and given the fact that we are more interested in the impact of adding CELLDYN data.

We expanded the CELLDYN dataset using pairwise feature combinations, specifically we applied $A/(A+B+\epsilon)$ and $\sqrt{A \cdot B}$, where A and B are different celldyn features.

We considered several transformations (quantile/minmax/standard) but the non-transformed features seemed to perform the best. We also considered removing the collinearity but it did not positively affect the predictive accuracy of the models. We prune the data by including only minimally informative features, which was based on multiple statistic distance metrics such as Anova, Kolmogorov-Smirnov, the 1st and 2nd Wasserstein distance.

We define the following feature sets:

1. age, gender, BMI, ECG, (CELLDYN)
2. age, gender, BMI, ECG, initial troponine, (CELLDYN)
3. age, gender, BMI, ECG, initial troponine, history, (CELLDYN)

4. age, gender, BMI, ECG, initial troponine, history, heartscore, (CELLDYN)
5. age, gender, BMI, ECG, initial troponine, history, heartscore, troponine slope, (CELLDYN)

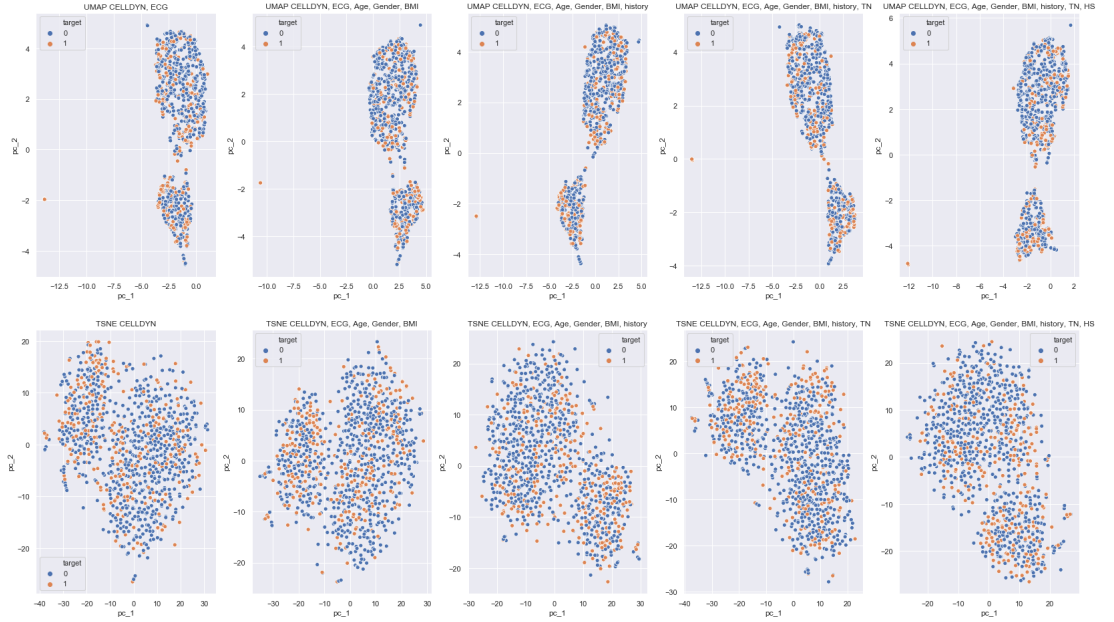
where for each feature set we have a with/without CELLDYN variant.

For each of these feature set we apply 10-fold cross-validation on an ensemble of machine learning models using a calibrated voting classifier. The models used are: XGboost, LightGBM, ExtraTrees, RandomForest and μ -SVM. For comparison we use a second ensemble of a convolutional network and a fully connected neural network. For the application of the neural network we did apply dimension reduction and feature transformation, namely *PCA* and a quantile transformation.

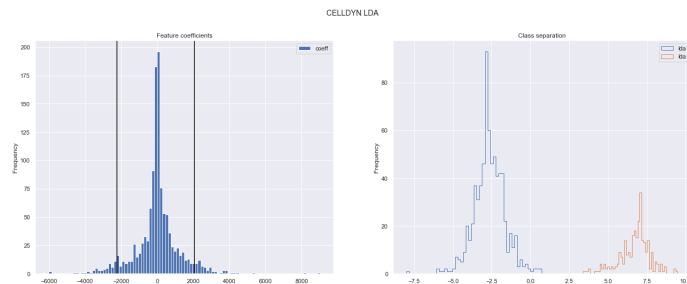
2 Results

The classes are slightly imbalanced; the NSTEMI's are underrepresented with 30%, in the modelling we compensate for this using higher target weights for the NSTEMI's.

2.1 Exploration



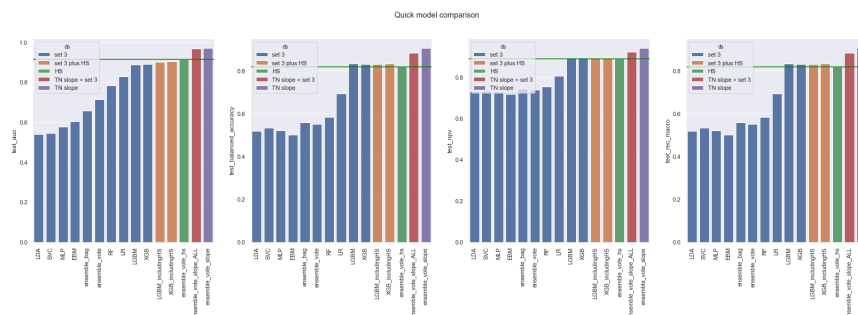
Using unsupervised clustering on UMAP and t-SNE reduced data we see that using CELLDYN data we can split the patients into two clusters, these two clusters persist when we add more data. Using LDA we look at the linear separability of the NSTEMI/non-NSTEMI patients.



The LDA transform suggests that the two classes are linearly separable. To test this we generate an LDA model on a training set and apply it to a test set, which results in poor performance, with the AUC close to 0.50. This suggests that a combination of the following is at play

- the target NSTEMI/non-NSTEMI can only be separated using non-linear methods
- high sensitivity of class separation to outliers (should be taken care of with quantile transformations)
- high variance of variance within dataset; i.e. there are a lot of patient segmentations

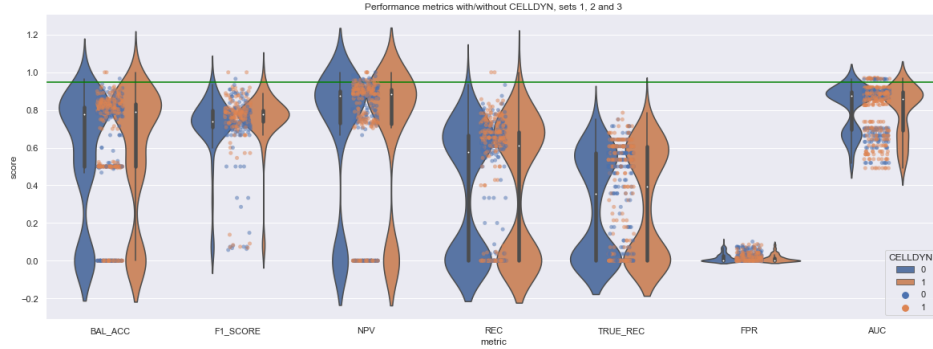
We repeated the LDA using quantile transformed variable and we found a similar result, and the same when we applied synthetic upsampling using SMOTE. Using feature set 3 we test the following classifiers; LDA, LR, RF, μ -SVM, 3 layered MLP, XGB and LGBM, and an ensemble of RF, μ -SVM, LR and MLP.



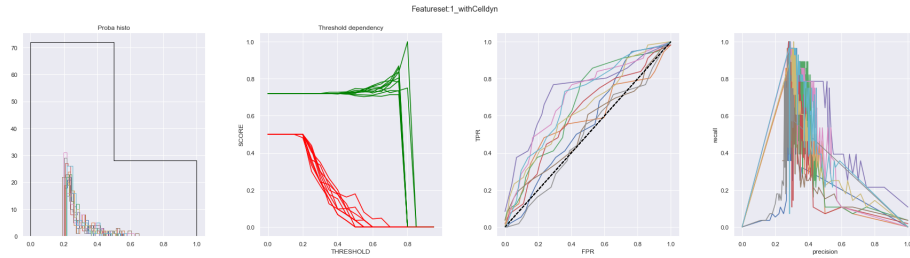
In terms of AUC/F1/precision and NPV the XGB and LGBM models on featureset 3 are an improvement over heartscore or on par. When we combine the troponine data (including the slope) with the other available data we do not see an improvement over just the troponine data. In terms of feature importances: the top-100 features is populated by

2.2 Cross validation

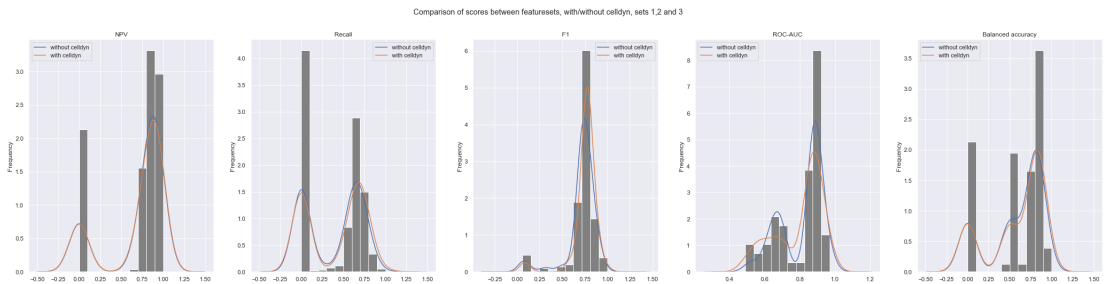
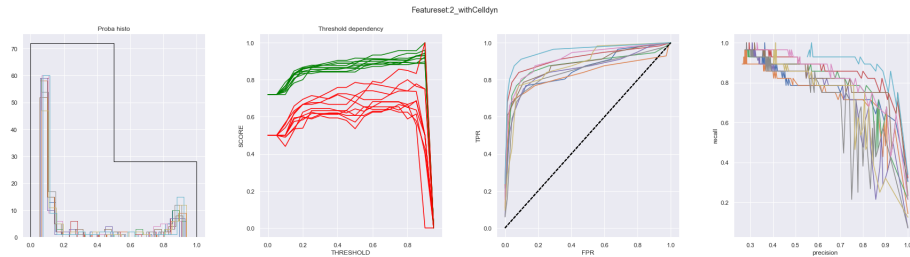
Continuing with the model testing we run a 10-fold cross-validation with an ensemble of XGB, LGBM, RF, ET and μ -SVM.



For the first model ensemble we see a slight improvement in NPV/F1 scores when we add CELLDYN to the feature sets, however this is not significant according to the Kolmogorov-Smirnov test ($p = 0.1$), and for the second model ensemble we see a slight deterioration (with $p < 0.01$). Independently we see that a featureset with age, gender, BMI and CELLDYN has some predictive value.



Adding troponine significantly improves the result and brings the accuracy on par with feature sets 3 and 4 that include the tailor made heart score and the historical clinical information.



We also see that for both model ensembles there are one or two folds in the CV that produce a trivial model with all zeros, supporting the idea that the variance within the population is very large, leading to non-informative patient selections. To overcome this we might benefit from oversampling the minority class and by tweaking the target weights.

We conclude that

- separating NSTEMI from non-NSTEMI patients is a highly non-linear problem
- CELLDYN clusters the heart patients in two distinct groups, however they are not predictive for the target
- CELLDYN can improve the ability to detect NSTEMI's but is not sufficiently accurate as a standalone predictor (with the used set of CELLDYN features)
- using readily available data at patient-intake an accurate prediction can be made regarding NSTEMI's

3 Suggestions

There are things we can do to improve the results;

- look for specific patients groups for which CELLDYN can provide significantly improved predictive accuracy, for this we need more data
- add more CELLDYN features
- parameter optimisation of the models, including the target weights to improve the NPV
- inter-dataset feature enrichment
- remove outlying samples
- rigorously prune the ECG/CELLDYN data
- apply synthetic over-sampling
- mine clinical reports for more sample features
- different thresholds NSTEMI and non-NSTEMI's predictions
- apply AutoML libraries such as AutoGluon, AutoKeras and TPOT