Identifying Predictors of Myocardial Infarction Complications with Interpretable Machine Learning

Yasona Neocleous
August 2022

Myocardial Infarction (MI, commonly known as heart attack) is a serious medical emergency in which the supply of blood to the heart is suddenly blocked, usually by a blood clot. Acute MI is associated with high mortality in the first year after it and its high incidence —especially in urban populations, due to lifestyle factors— makes it a leading cause of death globally. The course of the disease in patients with MI can vary considerably; MI can occur with or without complications and said complications can worsen the long-term prognosis or not. Even experienced specialists cannot always foresee the development of these complications. This makes, predicting MI patient survival an important task in order to timely carry out the necessary preventive measures. Therefore, throughout this project shallow machine learning techniques have been used to provide some explainability and improve the understanding of predictors of MI patient survival. A decision tree classifier (DTC) was used to differentiate between survivors and non-survivors based on clinical & demographic features, treatments administered and confirmed complications. A peak F1 score of 0.8478 was obtained when using a decision tree classifier, while when using a less explainable ensemble model (histogram gradient boosting, HXGB) a peak F1 score of 0.8222 was achieved.

The machine learning models were trained and tested using the UCI Myocardial Infarction Complications Database. At first glance the dataset had various issues. First of all, there were large amounts of missing data from the dataset. To overcome this, two methods were used. Firstly, a copy of the dataset was created using Pandas, on which columns having 15% of their values missing were removed. We shall call the original dataset the *full set* and the copy with features with a high degree of missingness removed as the *reduced set*. The remaining missing values were replaced by a constant value of -1 (all features are binary with values 0 or 1, so -1 is a valid indicator of missingness). This allowed the decision tree classifier to still make decisions based on the null value. The datasets were also severely imbalanced with survivors being the overwhelming majority class. This led to the models having very high accuracies because they consistently predicted the majority class. However, this resulted in a low precision due to all the false positives. To resolve this, the Synthetic Minority Oversampling technique known as SMOTE was used [1]. Using SMOTE, artificial datapoints from the minority class (non-survivors) were generated based on the existing data. This allowed the training data sets to be balanced, thus removing the problems caused by having an overwhelming majority class.

Shallow learning techniques employed using Scikit-learn were selected over deep learning techniques. This is because shallow learning techniques tend to perform better and converge faster when creating models based around tabular data [2]. As one of the main goals of the study was to provide a model with high explainability the decision tree classifier was used to differentiate between the survivors and non-survivors. This allowed the model to be printed in a tree format which facilitates visual analysis. This eased the analysis of what predictors and complications played the greatest role in determining if a patient survived. Through hyperparameter tuning a maximum tree depth of 7 was chosen. This was because it still provided a high F1 score while allowing for an easy-to-explain model. Once the decision tree classifier had been analysed an HXGB model was implemented along with permutation feature importance to potentially improve the binary classification abilities of the model. The eli5 library was used to provide deeper levels of interpretability to the HXGB model as unlike the decision tree, displaying the HXGB model is much more laborious and difficult. The models are then evaluated using the F1 score metric. The F1 score captures the balance between recall and precision

and doesn't rely on accuracy. This was optimal as the testing dataset will still retain the large survivor bias leading to high accuracies being meaningless without inspection.

Comparing the *full* and *reduced* data sets we can see that 20 columns of data had been removed. SMOTE had to be applied individually and only to the training sets to avoid the models learning on data they would later be tested on. This guaranteed that the class imbalance in the test data would remain unchanged, for evaluation purposes and that all evaluation data is real data. After multiple sets of cross validation training and evaluating, the model trained on the *reduced* data set performed better. At the optimised max depth of 7 the *reduced* model achieved an F1 score of 0.7865, while the model containing the full data set achieved an F1 score of 0.7473, making the *reduced* model score 5.3% better.

Permutation importance using the eli5 package was then applied to the models in order to calculate which features were most important to the model when making classifications [3]. Permutation importance works by calculating the decrease in a model score when the values of a single feature are randomly shuffled across all datapoints [3]. The physical decision trees for both models were also outputted. From both these methods the most important features could be analysed and assessed. The feature with the greatest importance for both models was NA_R_3_n, which represents the "Use of opioid drugs in the ICU in the third day of the hospital period". This is clearly a medical treatment and provides close to zero value in determining predictors of complications as the treatment has been applied in response to a complication and reoccurring chest pain. Therefore, to improve the model's ability to predict based on complications and other clinical & demographic features, all medical treatments were removed from the dataset and a new model (*`DTC - no treatments'*) was ~~created and~~ trained.

Once the treatments had been removed the F1 score went down to 0.7800 which is only 0.0065 (0.8%) worse than the *reduced set* based model and 0.0327 (4.3%) better than the *full set* based model. The 0.0065 loss is deemed acceptable as the model is now able to make predictions on only clinical & demographic features and complications and to a relatively similar standard. This has facilitated a greater level of explainability as the identification of clinical & demographic features and complications are now much easier using Figure 1. Each branch in Figure 1 can be followed and each predictor can be analysed to identify how this impacts survivability.

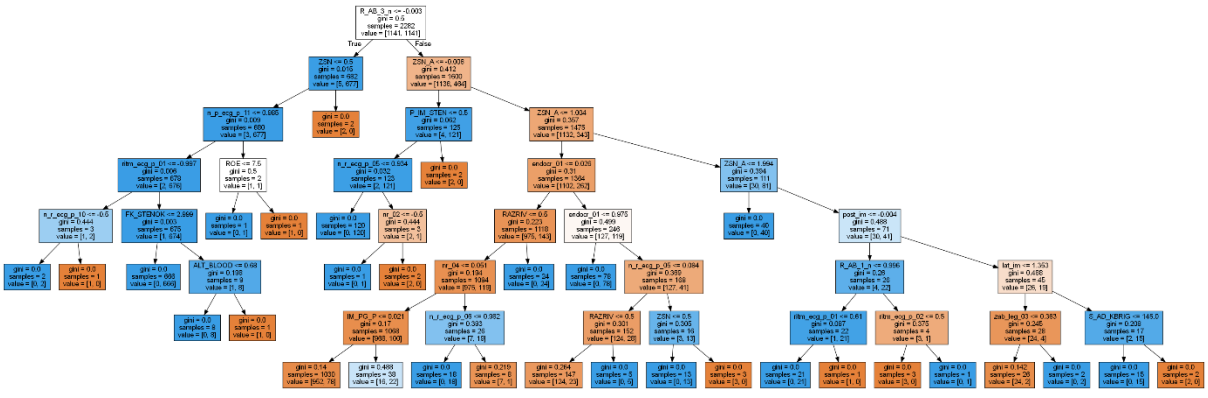

Figure 1 shows the plotted decision tree using the no treatment data set model (*`DTC - no treatments'*) for the training data. Bluer shades indicate a higher 'probability' of survival while a more orange shade indicates a high probability of non-survival.

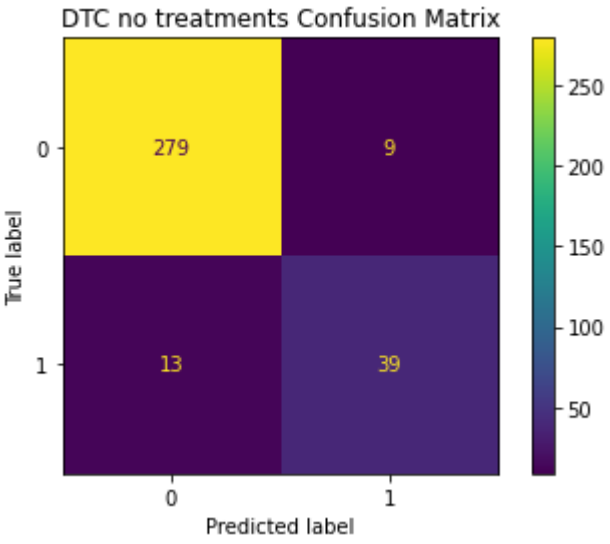

Figure 2 indicates the confusion matrix for the *`DTC - no treatments'* model. The label 0 indicates a patient survived while 1 indicated they were a non-survivor.

Weight	Feature
0.3243 ± 0.0747	R_AB_3_n
0.1628 ± 0.0302	RAZRIV
0.1213 ± 0.0258	ZSN_A
0.0785 ± 0.0575	ZSN
0.0352 ± 0.0103	ritm_ecg_p_01
0.0239 ± 0.0130	post_im
0.0224 ± 0.0101	zab_leg_03
0.0159 ± 0.0142	nr_04
0.0056 ± 0.0097	n_r_ecg_p_05
0.0049 ± 0.0121	R_AB_1_n
0.0025 ± 0.0099	ALT_BLOOD
0.0006 ± 0.0174	lat_im
0 ± 0.0000	SVT_POST
… 86 more …	

Figure 3 shows the permutation importance of each feature, their mean weight and uncertainty (standard deviation across 10 shufflings) for the *`DTC - no treatments'* model. The greater the value the more impact randomised shuffling of the feature in question had on the model performance.

Looking at Figure 3 we can see that only 12 features make a contribution to the model. With the other 87 features having zero weight and being functionally useless from a machine learning point of view. The top 5 features seen in Figure 3 represent, respectively: “Relapse of the pain in the third day of the hospital period”, “Myocardial rupture”, “Presence of chronic Heart failure (HF) in the anamnesis”, “Chronic heart failure”, “ECG rhythm at the time of admission to hospital – sinus (with a heart rate 60-90)”. From Figure 1 and Figure 3 we can see that R_AB_3_n is the most predictive factor of survival.

In order to further the machine learning discussion and explore the possibility of maximising model performance, two new datasets and models were created. The new datasets were made by trimming the “no treatment” dataset down to the top 10 (‘*DTC top 10’*) and top 5 (‘*DTC top 5*’) permutation importance features. Training models on the two new datasets resulted in an improved F1 score. The DTC top 10 and top 5 models scored 0.8298 and 0.8478, respectively. When comparing to the DTC no treatments model this is a 6.3% and 8.7% improvement. Looking at Figure 4 we can see that most of the improvement is afforded due to the increased precision. Therefore, the new models are now predicting far fewer false positives, but in terms of recall have not improved. In terms of the medical

application this does not particularly improve the results, as predicting false negatives are far more dangerous than false positives. Therefore, the *DTC top 10* and *DTC top 5* models are no better than the original *`DTC - no treatments'* model. However, from the perspective of ML the *DTC top 10* and *DTC top 5 are superior.* In a real world setting where the models would be deployed, it would be beneficial to use the no treatments model over the top 10 or top 5 due to the increased feature value and therefore more complex decision tree.

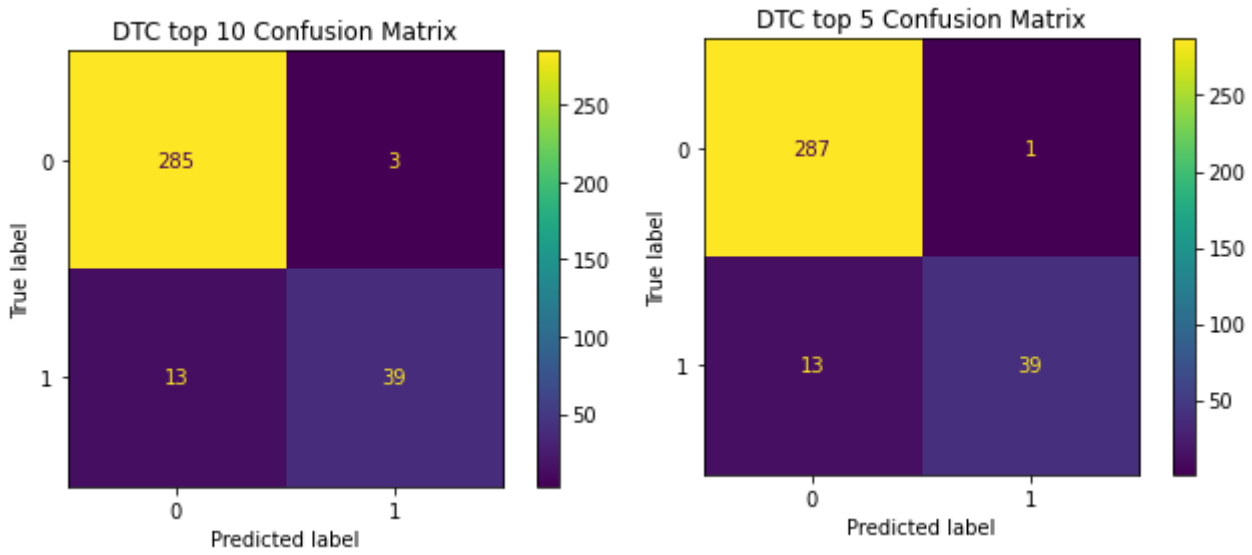

Figure 4 comparing the confusion matrix of a decision tree classifier trained on the top 10 permutation importance features and a model trained using the top 5 permutation importance features. The label 0 indicates a patient survived while 1 indicated they were a non-survivor.

As the main features contributing to the *DTC top 10* and *DTC top 5* models are complications they can't be immediately assessed when a patient arrives at a hospital. Therefore, a new model using the *DTC - no treatments* training data set, but with complications was trained. The new model (*'DTC available data'*) only achieved an F1 score of 0.7059. However, considering the context that this model can be used to assess patients as they enter the hospital and not after potentially irreversibly dangerous complications the lower F1 is deemed acceptable.

Further investigation into alternative machine learning methods were also conducted. Bagged decision tree models, XGB, HXGB and stacking methods were all analysed. HXGB achieved the greatest F1 score out of the ensemble methods with a value of 0.8222. From Table 1 it can be seen that the HXGB models only outperformed the DTC models when trained using the *full, reduced and no – treatments.* Therefore, if the task was to output a probability of a patient surviving given all the available information the HXGB model would be more suitable due to the 10.0% increased F1 score when compared to the *full* dataset model and 5.4% improvement when compared to *DTC - no treatments*. However, as the model provides little to no visual explainability compared to the DTC models making the optimum model that fits the brief of the task the DTC models.

Side by side comparison of the DTC models and the HXGB models

Dataset Trained on	DTC Models F1 Score	HXGB Models F1 Score
Full	0.7473	0.8222
Reduced	0.7865	0.7955
No treatments	0.7800	0.8222
Top 10	0.8298	0.8222
Top 5	0.8478	0.8211
Available Data	0.7059	0.7032

Table 1 comparing the F1 scores of all the DTC models and all the HXGB models trained. The left-hand column represents the dataset that each model was trained on.

In conclusion the *DTC top 5* model outputted the greatest F1 score of 0.8478. As the model is a decision tree classifier it also allowed for a high level of explainability because the tree could be plotted and analysed along side information from permutation importance. However, in a real life setting the top 5 model may be highly suboptimal as it only takes in 5 out 124 features with the main feature (R_AB_3_n) only being accessible 3 days after admission to hospital. This would make real life predictions very difficult using this model. This is because the feature the model depends on most is not available until 3 days into admission and by this point it may be too late to alter treatment or provide extra care. Therefore, the *DTC no treatments* model which achieved an F1 score of 0.7800 is more suitable, as it takes into account all complications and predictors of complications and not just the most predictive features. If a prediction was to be made at the point of entry into the hospital, the *DTC available data* model which achieved an F1 score of 0.7059 is the most suitable as it only takes into account clinical and demographic features. The DTC no - treatment model is more suitable than the HXGB model due to its much greater explainability as the DTC tree can be easily visualized, unlike the HXGB model. This fulfils the task of creating a model to help better understand complications and clinical & demographic features and act as a tool to aid specialists in deducing the correct measures apply to each patient.

References

[1] Alberto Fern´andez et al. “SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary”. In: Journal of artificial intelligence research 61 (2018), pp. 863–905.

[2] Angelo Cannarile et al. “Comparing Deep Learning and Shallow Learning Techniques for API Calls Malware Prediction: A Study”. In: Applied Sciences 12.3 (2022), p. 1645.

[3] Andŕe Altmann et al. “Permutation importance: a corrected feature importance measure”. In: Bioinformatics 26.10 (2010), pp. 1340–1347