ÓBUDAI EGYETEM
ÓBUDA UNIVERSITY

DOKTORI ÉRTEKEZÉS TÉZISFÜZETE

PETER PIROS

Improving Mortality Prediction with Machine Learning Models

Supervisors:
Prof. Dr. Levente Kovács
Dr. Rita Fleiner

DOCTORAL SCHOOL OF APPLIED INFORMATICS AND APPLIED MATHEMATICS

Budapest, November 26, 2023

Contents

Acknowledgments

I would like to express my gratitude to *Levente Kovács*, who, in addition to all his other duties, always assisted me in professional and administrative issues both under my master's and doctoral studies. His problem-solving ability always made possible to define the strategic plans and fulfill them. I am thankful for *Rita Fleiner* who helped me on a day-to-day basis. Her well-structured thinking and hardworking behaviour teached me a lot and made possible to publish the dissertation.

I am a great admirer of *András Jánosi*, who made a real, unfiltered dataset from the Hungarian Myocardial Infarction Registry (HUMIR) available for the current researches. His career with the heart attack registry is fascinating and I consider it an honor to be involved in his momentous work. I thank to *Tamás Ferenci* who was always ready to give feedbacks and advices on several questions regarding statistics and the dataset.

Thank you very much for the support of *Epam Hungary*, who supported my doctoral research, as an industrial partner.

1 Background of the Research

1.1 Importance of myocardial registries and mortality prediction

In *Heart Disease and Stroke Statistics*, American Heart Association annually reports, that approximately every 40 seconds, an American will have an myocardial infarction (MI) - they did the same in the recent statistics titled *2022 Update* [1].

In the area with numbers like these, mortality prediction can and should play a very important role in the hand of physicians: with validated models, it becomes possible to select patients with high-risk of death and use this information in the process of treatment. Using new, real-life datasets to extract hidden information can lead to more effective treatment and prevention. As US surgeon Dr. Ernest Amory Codman suggested: ”Every hospital should follow every patient it treats long enough to determine whether or not the treatment was successful and to inquire ‘if not, why not?’ with a view to preventing similar failures in future.” [2]

Reliable, high-quality datasets are mandatory to build and train any type of predictive model. For this reason, first, I reviewed the ongoing European myocardial projects then focused on the Hungarian register. The Hungarian Myocardial Infarction Register (HUMIR) project was introduced in 2010. In the recent years, around 15,000 new patients got registered per year and until December 2022, the 94 participating hospitals reported 157,724 cases in 142,439 patients.

In all my related publications and theses I used a dataset from HUMIR to predict the mortality of patients hospitalized with acute myocardial infarction.

1.2 Acute myocardial infarction

My theses are built around *acute myocardial infarction*. As [3] defines, it is ”a condition that happens because of a lack of blood flow to one’s heart muscle. The lack of blood flow can occur because of many different factors but is usually related to a blockage in one or more heart’s arteries. Without blood flow, the affected heart muscle will begin to die. If blood flow

isn't restored quickly, a heart attack can cause permanent heart damage and death." The most common symptom is chest pain or discomfort which may travel into the shoulder, arm, back, neck or jaw [4].

Acute myocardial infarction (commonly called a heart attack) remains a leading cause of morbidity and mortality worldwide, despite substantial improvements in prognosis over the past decade [5].

The statistical characteristics of MI also speaks volumes: as *Heart Disease and Stroke Statistics* reports [1], the estimated annual incidence of MI is 605,000 new attacks and 200,000 recurrent attacks in the US. The overall prevalence for MI is 3.1% in US adults (>19 years of age). Males have a higher prevalence of MI than females for all age groups except 20 to 39 years of age. MI prevalence is 4.3% for males and 2.1% for females.

1.3 Regression and Machine learning solutions

Logistic regression is the most commonly adopted and trusted model in the field of mortality prediction. From the view of medicine, it is import to know if there is a competitive opponent for the mostly used and trusted regression. Several studies work with regression - in general, it represents the ”classical” statistical approach with fast computational time and high accuracy. Next to regression, researchers try to use machine learning-based solutions to build predictive models to reach higher accuracy. In this subsection, I list a few attempts and their results.

Lee et al. [6] developed a logistic regression model based on a dataset of patients diagnosed with heart failure at multiple hospitals in Ontario, Canada. In a derivation set of 2,624 patients, the mortality rates were 8.9% in-hospital, 10.7% at 30 days, and 32.9% at 1 year. While validating the model, the area under the Receiver Operating Characteristics (ROC) curve was 0.80 for 30-day mortality and 0.77 for 1-year mortality.

Based on a dataset of 52,616 patients, Jack et al. [7] developed logistic regression models to predict 30-day and one-year mortality after an AMI. They predicted mortality with an area under the ROC curve of 0.78 for 30-day mortality and 0.79 for one-year mortality. In two independent validation datasets, this model reached 0.77 and 0.78, respectively.

Chin et al. [8] developed (n = 65,668) and validated (n = 16, 336) a logistic regression model to predict the risk of in-hospital mortality of patients with AMI. They reported AUC of 0.85 and 0.84 in the derivation and validation cohorts, respectively.

Clermont et al. [9] compared the performance of logistic regression and artificial neural network (ANN) models while predicting hospital mortality for patients in the intensive care unit. Seven intensive care units with 1,647 admissions were investigated, and finally they found that the two models have similar performance (0.80 and 0.84 as the area under the ROC curve).

Nilsson et al. [10] aimed to develop a method to select

risk variables and predict mortality after cardiac surgery by using artificial neural networks. They also used area under the ROC curve as performance indicator and found that area of artificial neural networks (0.81) was larger than the logistic model's (0.79).

Orr found [11] that implementing a probabilistic neural network model to estimate mortality risk following cardiac surgery is relatively rapid, and it is an alternative to standard statistical approaches. He got 0.72 and 0.81 as ROC AUC for the training and validation sets. The neural network model reached 0.74 on an independent dataset of the following year.

Voss et al. [12] investigated if neural networks improved on the risk estimate of the commonly used logistic regression. They used multi-layer perceptron (MLP) and probabilistic neural networks (PNN) to estimate the risk of MI or acute coronary death. As they reported, the AUC of the MLP was greater than that of the PNN (0.897 versus 0.872), and both exceeded the AUC for LR of 0.840. As a conclusion, the authors declare that use of the MLP to identify high-risk individ-

uals as candidates for drug treatment would allow prevention of 25% of coronary events in middle-aged men.

Austin compared [13] the predictive power of logistic regression with that of regression trees for predicting mortality after hospitalization with an AMI. His study shows that regression trees (0.762 AUC) do not perform as well as logistic regression (0.845 AUC). Author used data on 9 484 patients admitted to hospital with an AMI in Ontario, Canada.

In another study, Austin et al. [14] used ensemble-based methods, including bootstrap aggregation (bagging) of regression trees, random forests, and boosted regression trees. They found that ensemble methods offered substantial improvement in predicting cardiovascular mortality compared to conventional regression trees, but may not lead to clear advantages over conventional logistic regression models.

Convolutional neural network was applied by Acharya et al. [15] to automatize the diagnosis of congestive heart failure using ECG signals. They presented an 11-layer deep convolutional neural network model and out of four different datasets,

one attained the highest accuracy of 98.97%, specificity and sensitivity of 99.01% and 98.87% respectively.

One group of quoted researches uses only regression; other ones compare regression with neural network, but in another field of medicine; some publishes only the results of neural networks or decision trees; but only a few of them investigated the differences on an AMI dataset and, the most important, none of them used the official Hungarian myocardial registry to predict short- and long-term mortality.

2 Directions and Goals of the Research

Nowadays, most of the countries have their own mortality and disease statistics based on International Classification of Diseases – however, these statistics never contain clinical informations, for example results of former examinations, comorbidity or smoking behavior of the patients. Several databases store information about patients and diseases, but only a few system exists that focuses directly on myocardial events and

treatments. My first aim was to collect these systems and to take advantage of the opportunities offered by the Hungarian register with developing the solution which turns the registers' raw data into an input data of machine learning algorithms.

Then, in my researches I developed several machine-learning models based on Decision Tree (DT), Neural Networks (NN), Logistic Regression (LR), Random Forest (RF), Generalized Boosted Model (GBM) and Ensembled algorithms to predict 30-day and 1-year mortality on the same, real-world, unfiltered dataset originated from HUMIR. The main question I was facing was if there is a competitive opponent for the mostly used and trusted regression in the world of machine learning algorithms.

The idea behind the approach that I was working on the same dataset in all the connected researches is the following: I was trying to establish an order in the list of different modelling techniques by keeping the dataset fixed and trying to maximize the prediction capability of each of our models.

3 Materials and Methods of Investigation

3.1 Methodology and Methods

In the area of data mining, a well-known methodology called *CRISP-DM* exists which summarizes the main stages and questions of a given project, hence, serves as a base for the whole process. This process framework was used in all of the researches.

In the studies, I applied area under the Receiver Operating Characteristics curve, or simply ROC AUC as a single-number measure for evaluating performance of a learning algorithm. Huang et al. [16] proved that AUC is – in general – a better measure than accuracy. Bradley [17] found that ROC AUC has several desirable properties compared to accuracy.

Decision tree is a graphical model that uses a tree-like structure of classifying examples. Starting with the full dataset, in each step, the algorithm splits the source set into two sub-

sets based on a feature and a corresponding value. This operation gets repeated in a recursive manner, until a node has all the same values of the target variable, or another stopping criteria fires.

Artificial neural networks are based on a collection of connected neurons, where the connections can transmit a signal from one neuron to the other. This theory is inspired by biological neural networks. In my study, a feed-forward neural network with a single hidden layer was used.

The basic idea of Random Forest algorithm is building many small, weak, less-correlated trees in parallel. The RF algorithm works as follows: first, we select a bootstrap sample. Second, on each bootstrap sample and on each node, a decision tree-based learning method gets performed, but with only a randomly selected, small subset of features.

Boosting represents the idea that the final model could be more powerful if we continuously add weak models (e.g. decision trees) to our system, each compensating the weaknesses of its predecessors.

Ensembled modelling as a strategy based on the idea that if we combine the predictive performance of different classifiers, it can produce a stronger learner. Boosting itself also belongs to ensembled approach, but here *Stacking* was also applied: when different types of first-levels learners are used, then we try to exploit the common predictive power of them in an upper level.

To deal with missing data multiple imputation using Fully Conditional Specification (FCS) and Bayesian linear regression was applied with 5 imputations and 5 iterations, leaving the final, prepared dataset size at $n = 47{,}391$.

3.2 Software and Hardware environments

Statistical analysis, data preparation, modelling and visualisation were all principally performed under the R statistical language [18], version 3.6.1.

The applied Machine Learning methods are listed in pair with the two different hardware environments and configurations I used during the investigation:

1. *Normal configuration* with the following resources: Intel Core i3 processor (i3-4030U CPU 1.90 GHz), 12 GB memory, no SSD. This environment was used in the early publications connected with Logistic regression, Decision tree and Neural network.

2. *Extended configuration* is a cloud-based architecture powered by *Amazon Web Services* with *EC2* instances[1]. It had the following parameters: 16 vCPU, 70 ECU, 64 GB memory (*m5.4xlarge* configuration). This environment was applied with the latest researches and models like Random forest, General Boosted Model and Ensembled.

4 New Scientific Results

Thesis group 1: In a methodological approach, I have discussed and analyzed the data preparation of artificial intel-

[1] Amazon EC2: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html

ligence algorithms on the dataset of the Hungarian National Myocardial Infarction Register.

Thesis 1.1

At the international level, I have took a look at the official registries that collect cardiovascular data; within this, I have highlighted the uniqueness of the Hungarian National Myocardial Infarction Register and gave a ”literature recipe” for the use of artificial intelligence methods.

Thesis 1.2

For the dataset of the Hungarian Myocardial Infarction Registry, I have developed a data preparation procedure, with which the raw data became suitable for participating in the implementation of the machine learning process as input for the predictive models.

Relevant own publications pertaining to this thesis group:

[P-1] [P-2] [P-3]

Thesis 2:

Thesis 2

I have developed machine learning models for predicting 30-day and 1-year mortality that met and in some cases exceeded the predictive capabilities of regression.

Relevant own publications pertaining to this thesis:

[P-3] [P-4] [P-5] [P-6]

Thesis 3:

Thesis 3

I have showed that in the case of decision trees, there are minimal differences between the resampling methods used to determine the tuning parameters; and these differences disappear with a larger data set ($n > 15000$).

Relevant own publication pertaining to this thesis:

[P-7]

5 Practical Applicability of the Results

For demonstrating a possible clinical application of the research results, I have developed a web-based application where visitors can check the prediction capability of the given model.

Through the application physicians can enter the patient data, then calculate the predicted possibilities. In the background, in the process of prediction, the application is communicating with the original (R-environment-based) modelling infrastructure to use originally developed models to predict 30-days and 1-year mortality as outcomes.

The application is fully developed and working as it is described in this section. It uses the previously published General Boosted Model to predict the 30-day and 1-year mortality for field values added on the User interface layer.

Although the current version has some limitations (which

are listed in the dissertation, together with their possible solutions) mainly coming from the software and hardware architecture - the application demonstrates a possible clinical application of the research results present in the current dissertation, and it is ready to use by physicians.

6 Bibliography

References

[1] Connie W Tsao et al. “Heart disease and stroke statistics—2022 update: a report from the American Heart Association”. In: *Circulation* 145.8 (2022), e153–e639.

[2] M Pearson. “Lessons from the management of acute myocardial infarction”. In: *Heart* 91.suppl 2 (2005), pp. ii28–ii30.

[3] *Cleveland Clinic website*. Accessed 2023.01.23 14:25. URL: `https://my.clevelandclinic.org/health/`

diseases/16818-heart-attack-myocardial-infarction.

[4] Lung National Heart, Blood Institute, et al. “What are the signs and symptoms of coronary heart disease”. In: *Bethesda: National Heart, Lung, and Blood Institute* (2016).

[5] Grant W Reed, Jeffrey E Rossi, and Christopher P Cannon. “Acute myocardial infarction”. In: *The Lancet* 389.10065 (2017), pp. 197–210.

[6] Douglas S Lee et al. “Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model”. In: *Jama* 290.19 (2003), pp. 2581–2587.

[7] Jack V Tu et al. “Development and validation of the Ontario acute myocardial infarction mortality prediction rules”. In: *Journal of the American College of Cardiology* 37.4 (2001), pp. 992–997.

[8] Chee Tang Chin et al. “Risk adjustment for in-hospital mortality of contemporary patients with acute myocardial infarction: The Acute Coronary Treatment and Intervention Outcomes Network (ACTION) Registry®–Get With The Guidelines (GWTG)™ acute myocardial infarction mortality model and risk score”. In: *American heart journal* 161.1 (2011), pp. 113–122.

[9] Gilles Clermont et al. “Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models”. In: *Critical care medicine* 29.2 (2001), pp. 291–296.

[10] Johan Nilsson et al. “Risk factor identification and mortality prediction in cardiac surgery using artificial neural networks”. In: *The Journal of thoracic and cardiovascular surgery* 132.1 (2006), pp. 12–19.

[11] Richard K Orr. “Use of a probabilistic neural network to estimate the risk of mortality after cardiac

surgery". In: *Medical Decision Making* 17.2 (1997), pp. 178–185.

[12] Reinhard Voss et al. "Prediction of risk of coronary events in middle-aged men in the Prospective Cardiovascular Münster Study (PROCAM) using neural networks". In: *International journal of epidemiology* 31.6 (2002), pp. 1253–1262.

[13] Peter C Austin. "A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality". In: *Statistics in medicine* 26.15 (2007), pp. 2937–2957.

[14] Peter C Austin et al. "Regression trees for predicting mortality in patients with cardiovascular disease: what improvement is achieved by using ensemble-based methods?" In: *Biometrical journal* 54.5 (2012), pp. 657–673.

[15] U Rajendra Acharya et al. “Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals”. In: *Applied Intelligence* 49 (2019), pp. 16–27.

[16] Jin Huang, Jingjing Lu, and Charles X Ling. “Comparing naive Bayes, decision trees, and SVM with AUC and accuracy”. In: *Third IEEE International Conference on Data Mining*. IEEE. 2003, pp. 553–556.

[17] Andrew P Bradley. “The use of the area under the ROC curve in the evaluation of machine learning algorithms”. In: *Pattern recognition* 30.7 (1997), pp. 1145–1159.

[18] R R Core Team et al. “R: A language and environment for statistical computing”. In: (2017). URL: `https://www.R-project.org`.

Own Publications Pertaining to Theses

[P-1] Péter Piros et al. “An overview of myocardial infarction registries and results from the Hungarian Myocardial Infarction Registry”. In: *IOS Press* 297 (2017), p. 312. DOI: 10.3233/978-1-61499-800-6-312.

[P-2] László Beinrohr et al. “Anatomy of a Data Science Software Toolkit That Uses Machine Learning to Aid Bench-to-Bedside Medical Research—With Essential Concepts of Data Mining and Analysis Explained”. In: *Applied Sciences* 11.24 (2021), p. 12135. DOI: 10.3390/app112412135.

[P-3] Péter Piros et al. “Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry”. In: *Knowledge-Based Systems* 179 (2019), pp. 1–7. DOI: 10.1016/j.knosys.2019.04.027.

[P-4] Péter Piros, Rita Fleiner, and Levente Kovács. “Random Forest-based predictive modelling on Hungar-

ian Myocardial Infarction Registry". In: *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. IEEE. 2020, pp. 525–530. DOI: 10.1109/SoSE50414.2020.9130476.

[P-5] Péter Piros, Rita Fleiner, and Levente Kovács. "Finding improved predictive models with Generalized Boosted Models on Hungarian Myocardial Infarction Registry". In: *2020 IEEE 20th International Symposium on Computational Intelligence and Informatics (CINTI)*. IEEE. 2020, pp. 179–184. DOI: 10.1109/CINTI51262.2020.9305814.

[P-6] Péter Piros et al. "Further evolution of mortality prediction with ensemble-based models on Hungarian Myocardial Infarction Registry". In: *Acta Polytechnica Hungarica* 20.4 (2023). DOI: 10.12700/APH.20.4.2023.4.7.

[P-7] Péter Piros et al. "Comparing the predictive power of decision tree models with different tuning approaches on Hungarian Myocardial Infarction Registry". In:

2019 IEEE 13th International Symposium on Applied Computational Intelligence and Informatics (SACI). IEEE. 2019, pp. 326–331. DOI: 10.1109/SACI46893.2019.9111525.

Own Publications Not Pertaining to Theses

[Px-1] Péter Piros, Rita Fleiner, and Levente Kovács. “Linked data generation for courses and events at Óbuda University”. In: *2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE. 2017, pp. 000253–000258. DOI: 10.1109/SAMI.2017.7880313.

[Px-2] Rita Fleiner, Barnabás Szász, and Péter Piros. “Indoor navigation Linked Data at Obuda University”. In: *2016 IEEE 11th International Symposium on Applied Computational Intelligence and Informatics (SACI)*.

IEEE. 2016, pp. 25–30. DOI: 10.1109/SACI.2016.7507377.