Finding improved predictive models with Generalized Boosted Models on Hungarian Myocardial Infarction Registry

Péter Piros, Rita Fleiner, Levente Kovács

Óbuda University, Budapest, Hungary

Contents

1. **Background**
   - Hungarian Myocardial Infarction Registry (HUMIR)
   - Myocardial Infarction
2. **Research**
   - Authors former publications on the topic
   - Dataset, Prediction targets
   - Generalized Boosted Models
   - Comparison with Random Forest model
   - Results & Conclusions

Hungarian Myocardial Infarction Registry (HUMIR)

The **Hungarian Myocardial Infarction Registry (HUMIR)** focuses directly on myocardial events and treatments.

- In 2014: the Hungarian government made it mandatory for hospitals to participate in the project in the whole country and report all cases to the registry [1].

- Hospitals, cases, patients: Until 1st of Oct 2020, the 93 participating hospitals reported 127,249 cases in 116,029 patients [2].

- Purpose of the registry: To audit the quality of care of patients with acute myocardial infarction and provide a database for scientific research.

Myocardial Infarction

Significance: Cardiovascular disease (CVD) continues be one of the most serious health problem of mankind. 10 leading causes accounted for 74.1% of all registered deaths in US - and this causes in 2016 were the same as in 2015. In this Top 10 list, heart disease can be found in the 1st position [3]

Incidence – in Hungary, Budapest [4]: out of 10.000:
- Man: 28,63
- Woman: 16,21

Two main types:

- STEMI: there is a pattern known as ST-elevation on the EKG („ST elevation myocardial infarction")
- NSTEMI: there is elevation of the blood markers suggesting heart damage, but no ST elevation seen on the EKG

Authors former publications on the topic

An Overview of Myocardial Infarction Registries and Results from the Hungarian Myocardial Infarction Registry [1]:

History and early results of HUMIR. Conclusion: Only a few such registers exist in Europe.

Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry [5]:

Conclusions: The difference between the predictive power of our neural network and logistic regression models were not significant, but decision tree was not able to achieve such a performance.

Comparing the predictive power of decision tree models with different tuning approaches on Hungarian Myocardial Infarction Registry [6]:

Conclusions: On the investigated dataset, repeated cross validation slightly outperformed cross validation and both had significantly better results than models trained with bootstrap method.

Authors former publications on the topic

Random Forest-based predictive modelling on Hungarian Myocardial Infarction Registry [7]:

Conclusions:

- Random Forest (RF) models clearly outperformed our previously reported decision tree (DT) models: improvement of 5.5% and 7.3% (30-day models, training and validation); 8.1% and 9.2% (1-year models, training and validation)
- The most important features in RF models - 30-day mortality: Age, Cardiogenic shock, Smoking, Hyperlipidaemia and Level of creatinine. 1-year mortality: 4 more features reached the same level of importance: Hyperlipidaemia, Heart failure, Peripheral artery disease and Percutan coronary intervention
- The RF models represent a stable learner. Standard deviations: 0.0047 (30-day) and 0.0036 (1-year) on the validation datasets.

Patient record

Our patient record consists of the following fields. These

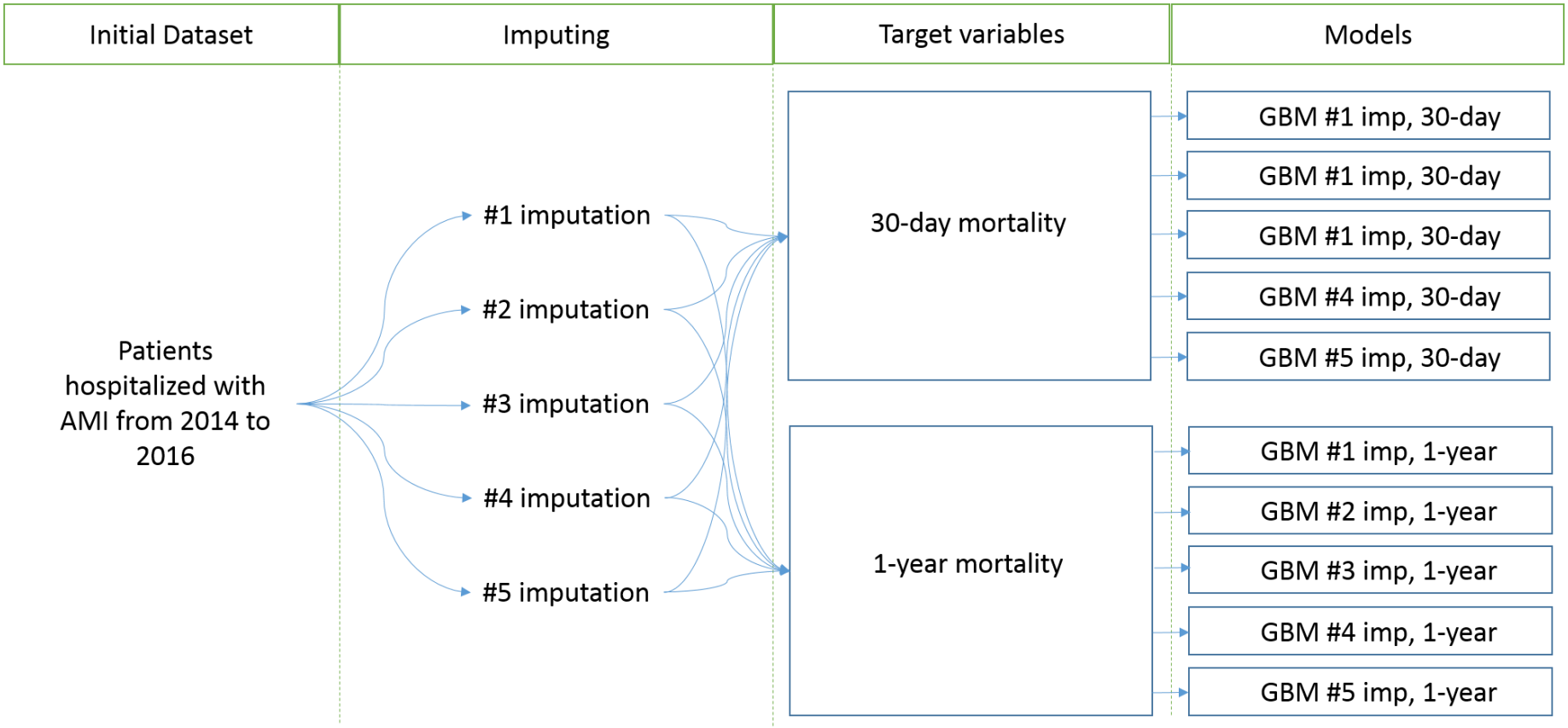
23 variables can be categorised into 3 homogene groups:

Group 1: General information about the patient (Event ID, Patient ID, If the patient alives, Date of death, Gender, Date of birth, ZIP code)

Group 2: Previously reported diseases (Myocardialis infarctus, Hearth failure, Hypertonia, Stroke, Diabetes,Peripheral vascular disease, Hyperlipidaemia, Smoking)

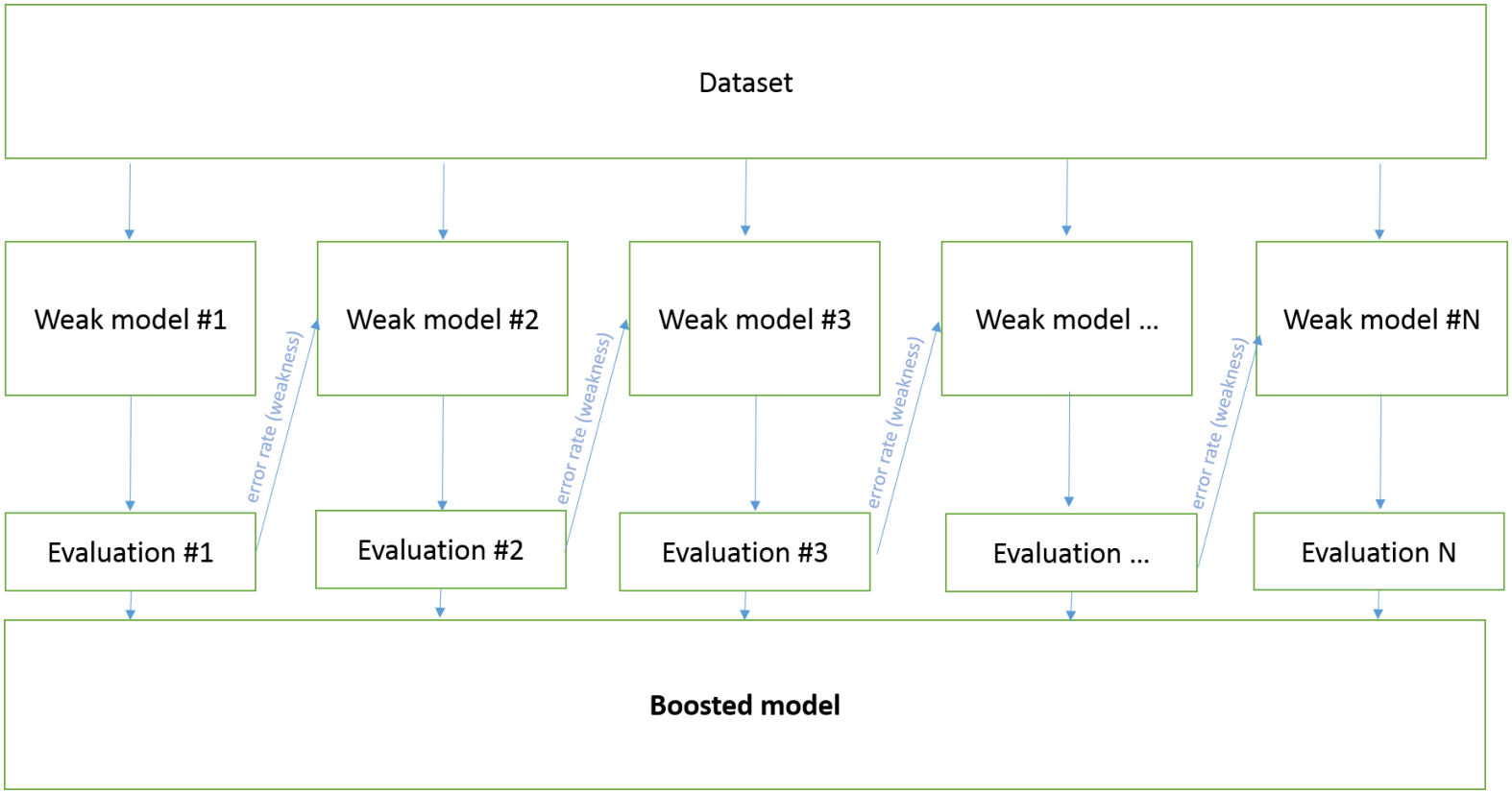
Group 3: Information about the pre- and in-hospital treatment (Prehospitalis reanimatio, Cardiogenic shock, Percutaneous Coronary Intervention, Level of creatinine, Diagnosis, Treatment ID, Date of admission, Creatinine)

Data and model structure of the current research



Target variables: 30-day and 1-year mortality. **Missing data:** multiple imputation, Fully Conditional Specification and Bayesian linear regression, 5 imputations.

Generalized Boosted Models



Boosting is the process of iteratively adding basis functions in a greedy fashion so that each additional basis function further reduces the loss function.

Generalized Boosted Models

Adaptive Boosting (AdaBoost): weakness of each learner is the set of misclassified data points. Solution: adding increased weights to these points (while decreasing the weight of well-classified items) so that the next weak learner will pay extra attention to putting it to the right class.

Gradient Boosting: instead of adding sample weights and tuning them based on the success of classification, it compares the difference between the predicted and the real value coming from the dataset.

Software and hardware environment

Software:

- R was used as an open-source software environment and language for statistical computing and graphics.

- The implementation of R's generalized boosted modeling framework closely follows Friedman's Gradient Boosting Machine [8].

Hardware (same as with RF models):

- Usual hardware environment (Intel Core i3 processor, 12 GB memory) was not suitable for GBM modelling on this size of dataset.

- Applied environment: Amazon AWS service with 48 vCPU, 168 ECU, 192 GB memory (m5.12xlarge configuration)

- Average training times: below 5 minutes

Results & Conclusions

Resulted ROC AUC values of GBM models for each imputations in case of 30-day mortality as target variable:

Model Nr.	*Training set*	*Validation set*
#1	0.847	0.841
#2	0.848	0.838
#3	0.847	0.837
#4	0.844	0.844
#5	0.849	0.835

The average for 30-day mortality is 0.847 for training set and 0.839 for validation set.

Results & Conclusions

Resulted ROC AUC values of GBM models for each imputations in case of 1-year mortality as target variable.

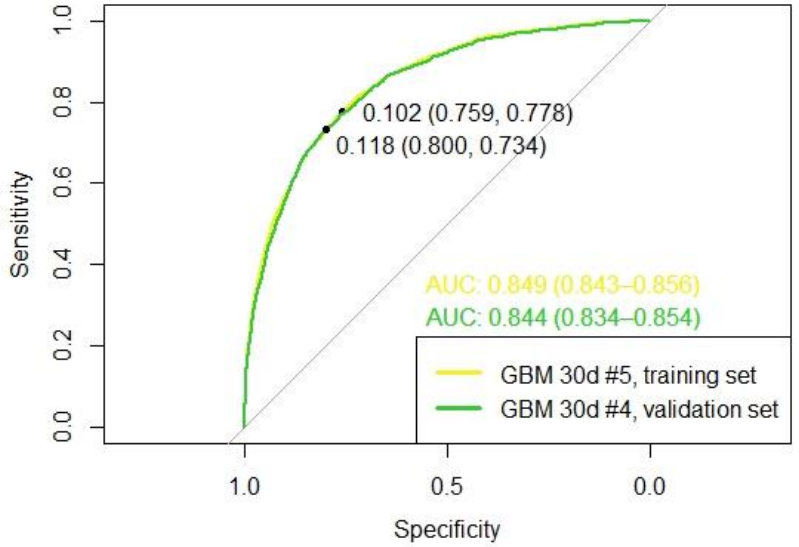
Model Nr.	*Training set*	*Validation set*
#1	0.829	0.817
#2	0.829	0.820
#3	0.826	0.825
#4	0.829	0.818
#5	0.825	0.825

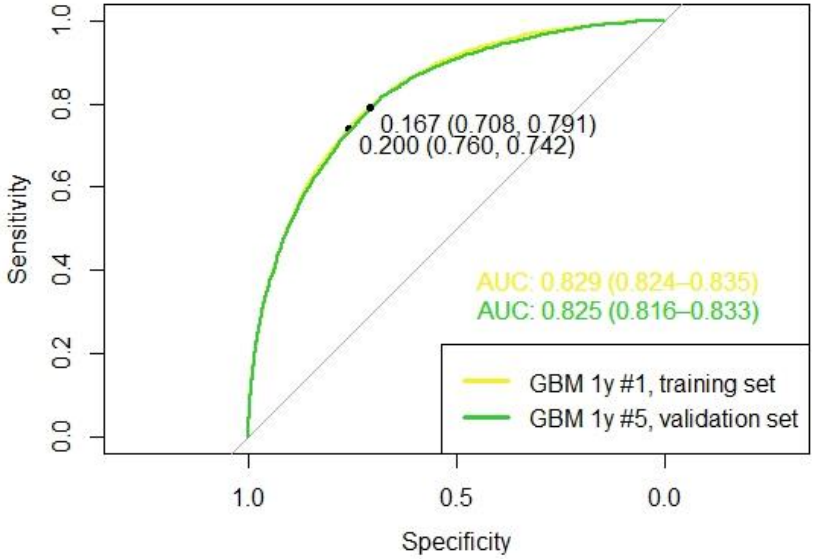
This means an average of 0.828 on the training set and 0.821 on the validation set.

Results & Conclusions

1.) There is no significant difference exists between the predictive power of models trained on different imputations.

2.) Next to our RF models, the GBM models also represent a stable learner: the standard deviation for the 30-day models are 0.0019 and 0.0035 (training and validation, respectively). These numbers are 0.0019 and 0.0038 for the 1-year models (training and validation, respectively).





Results & Conclusions

3.) Most important variables of our GBM models:

30-day mortality	1-year mortality
Cardiogenic shock	Age
Age	Cardiogenic shock
Level of creatinine	Level of creatinine
Percutan coronary intervention	Percutan coronary intervention
Prehospitalis reanimation	Hearth failure
Diagnosis	Prehospitalis reanimatio

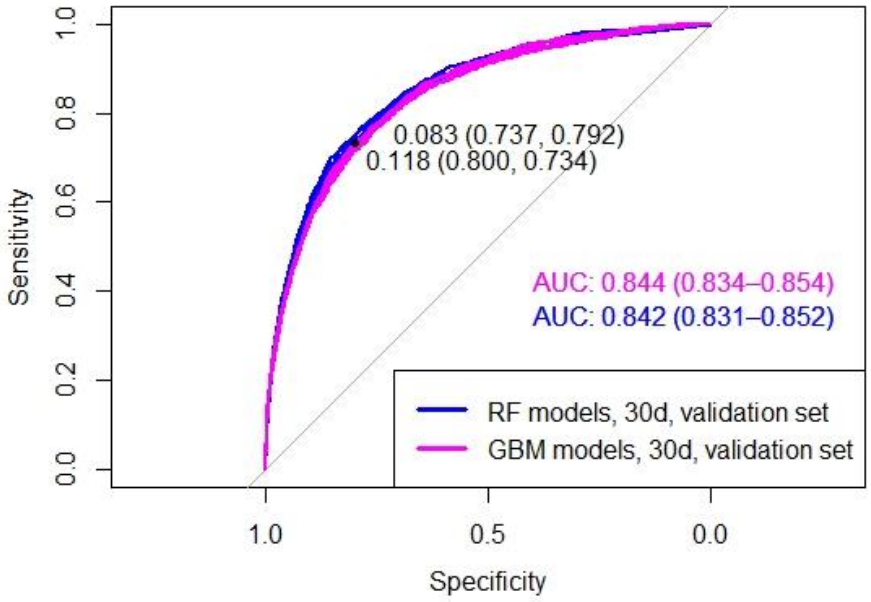
Results & Conclusions

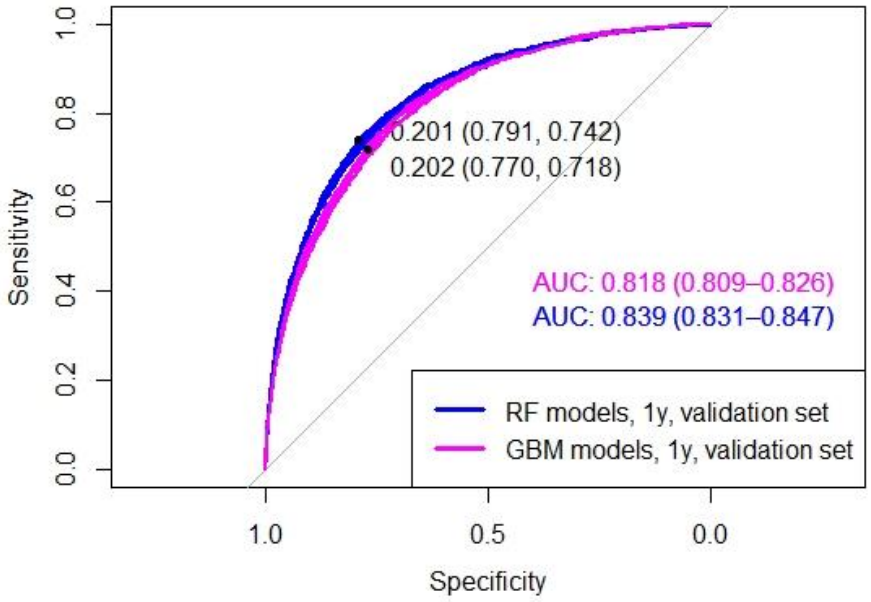
Comparing with our former results (RF models):

- The difference between our RF and GBM models are between 0.5% and 0.9%, except in the case of 1-year model on the validation dataset: it’s 1.7% compared to the RF results. A slight advantage of RF's performance power is revealed.
- There are common factors in the list of five most important fields, namely Age, Cardiogenic shock and Level of creatinine for both 30-day and 1-year models.
- There are few factors appearing only in one of the models:
  - RF: Smoking and Hyperlipidaemia for both 30-day and 1-year models;
  - GBM: Percutan coronary intervention, Prehospitalis reanimatio for 30-day models and Prehospitalis reanimatio and Hearth failure for 1-year models.

Results & Conclusions

Examples – differences between our RF and GBM models on validation sets:





References

[1] Piros P, Fleiner R, Ferenci T, Andreka P, Fujita H, Ofner P, Kovacs L,

Janosi A, Fujita H, Selamat A, Omatu S (szerk.) An overview of myocardial infarction registries and results from the hungarian myocardial infarction registry. FRONTIERS IN ARTIFICIAL INTELLIGENCE AND APPLICATIONS 297: pp. 312-320. (2017)

[2] Source: https://ir.kardio.hu/, Last access: 1st of Oct, 2020

[3] Benjamin, E. J., Muntner, P., & Bittencourt, M. S. (2019). Heart disease and stroke statistics-2019 update: a report from the American Heart Association. Circulation, 139(10), e56-e528.

[4] Janosi, A., Ofner, P., Branyickine Geczy, G., & Polgar, P. (2013). Incidence of myocardial infarction in Hungary. Population study in five districts of Budapest and Szabolcs-Szatmar-Bereg county. Orvosi hetilap, 154(28), 1106-1110.

[5] Piros, P., Ferenci, T., Fleiner, R., Andréka, P., Fujita, H., Főző, L., ... & Jánosi, A. (2019). Comparing machine learning and regression models for mortality prediction based on the Hungarian Myocardial Infarction Registry. *Knowledge-Based Systems*, *179*, 1-7.

[6] Péter, Piros ; Rita, Fleiner ; Tamás, Ferenci ; Levente, Kovács ; András, Jánosi: Comparing the predictive power of decision tree models with different tuning approaches on Hungarian Myocardial Infarction Registry. SACI 2019 : IEEE 13th International Symposium on Applied Computational Intelligence and Informatics : PROCEEDINGS. Temesvár, Románia : IEEE, (2019) pp. 326-331., 6 p.

References

[7] Piros, P., Fleiner, R., & Kovács, L. (2020, June). Random Forest-based predictive modelling on Hungarian Myocardial Infarction Registry. In *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)* (pp. 525-530). IEEE.

[8] J.H. Friedman (2001). "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics 29(5):1189-1232.

Thank your for your attention!