Heart Failure: predicting hospital re-admission after 6 months

Statistical Learning for Healthcare Data (056867) – A.Y. 2022/2023

Prof. M. Ferrario and Prof. A. M. Paganoni

Teo Bucci*, Giulia Montani* and Alice Traversa†
*M.Sc. Mathematical Engineering †M.Sc. Biomedical Engineering, Politecnico di Milano
Email: {teo.bucci, giulia.montani, alice.traversa}@mail.polimi.it
Code available at: github.com/teobucci/slhd

1. Introduction

Heart Failure (HF) is a prevalent condition with high re-admission rates: the literature[1] states that the number of HF cases worldwide almost doubled from 33.5 million in 1990 to 64.3 million in 2017.

Studies[2] also suggest that half of the patients diagnosed with HF will be re-admitted once within a year and 20% will be re-admitted twice or more. The focus of this analysis is on predicting the re-admission within 6 months for HF patients.

However, our emphasis extends beyond prediction performance; we prioritize the development of an interpretable model that healthcare professionals can easily understand. By examining patient characteristics and clinical variables, we aim to identify key predictors and provide valuable insights for targeted interventions and personalized care plans.

2. Materials and methods

The dataset was collected from 2008 patients with heart failure who were admitted to a hospital in Sichuan, China between 2016 and 2019, 773 of which were readmitted within 6 months.

A total of 168 variables were provided, which included basic patient characteristics such as age, sex, height, weight, occupation, admission department, visit times, etc. Clinical characteristics such as respiratory rate, systolic blood pressure, diastolic blood pressure, hemoglobin, red blood cells, D-dimer, etc. It also included comorbidities such as diabetes, dementia, liver disease, etc.

Variables were of both types: categorical (binary and multi-class) and numerical (continuous and integer).

The majority of patients had age in the 59-89 range, namely a total of 1675 patients accounting for 86.07% of the total. Regarding sex, 58.22% were females and 41.78% were males.

73.54% suffered a whole HF, while 23.84% suffered a Left HF and 2.62% a Right HF.

The other dataset provided contained drugs information: which drugs each patient took.

The problem is framed as a binary classification problem, one for patients that were readmitted to the hospital within 6 months, and another for those who were not.

2.1. Data cleaning

To begin with, we merged the information of the drugs to the main dataset, but since there were 18 drugs, we grouped them into four categories: Diuretics, Vasodilatory, Inhibitor, Increase Force of Heart Contraction (IFHC).

We removed the 57 dead patients and all information on re-hospitalizations prior to 6 months. We also identified and removed 5 patients with inconsistent information: their `DestinationDischarge` was `Died` but their outcome during hospitalization wasn't `Dead`.

1. N. L. Bragazzi et al. "Burden of heart failure and underlying causes in 195 countries and territories from 1990 to 2017". In: *European Journal of Preventive Cardiology* 28.15 (Feb. 2021), pp. 1682–1690.

2. A. Groenewegen et al. "Epidemiology of heart failure". In: *European Journal of Heart Failure* 22.8 (June 2020), pp. 1342–1356.

2.1.1. Missing values. In healthcare datasets, missing values are often a big concern. The first step was to inspect the situation:

- among the numerical variables, 105 contained NaNs;
- among the categorical variables, 1 contained NaNs, which was `occupation` 1.34% of missing, which we imputed with the most frequent.

All 14 variables with over 60% missing were discarded, while for those with 50% to 60% missing we computed whether there was any variable in the sub-50% missing part of the dataset with more than 0.80 correlation. If that was the case, we discarded them: 9 variables were removed.

Finally, among the remaining ones in the 50% to 60% missing range we computed the correlation matrix, clustered with hierarchical clustering, with the aim of seeing a clearer block structure, see Figure 1. With this techniques we discarded 3 more variables.

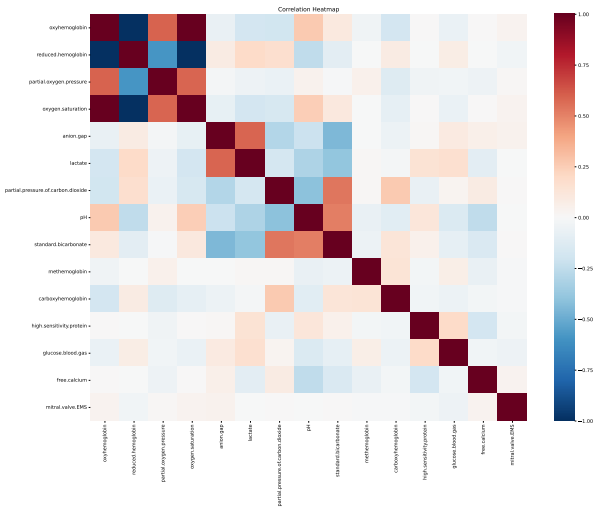

Figure 1. Correlation matrix of variables with 50%-60% missing values.

2.1.2. Outlier analysis. An overall histogram of all numerical variables made very clear that there were many outliers in the data.

The approach we followed was to identify them and judge one by one whether or not the values were physiologically possible or not, by checking the literature. In that case, instead of discarding the patient completely, we set the value to NaN to impute it later. During the approach we tried to be as conservative as possible to retain as much information as possible, but simultaneously give to the models high quality data that would enable good separation.

Outliers were identified by calculating the sample Z-score of each variable:

$$Z\text{-score} = \frac{X - \bar{X}}{s}$$

where $\bar{X}$ is the sample mean and s is the sample standard deviation.

A threshold of 3 or above was applied to detect them, depending on the specific variable.

After this procedure, there were still three variables that showed possible outliers: `eosinophil.count`, `high.sensitivity.troponin` and `glutamic.pyruvic.transaminase`. However, these variables were marked by many models as important, so we decided to keep them anyway without changes. We suspect there may have been discordances in inserting the values, i.e. for some patients with some unit of measure, for other patients with another unit. We strongly recommend double checking these variables.

2.1.3. Removing low variance variables. Among the categorical variables, we removed 16 variables with more than 95% dominance in values counts, while for the numerical variables we removed 2 constant ones.

2.1.4. Correlation analysis. Using clinical knowledge, we inspected some groups of variables we thought could be quite correlated, then we proceeded with discarding variables with more than 0.85 correlation with the others, ultimately removing 12 more variables.

Up to this point, 62 variables had been discarded.

2.2. Modeling

The dataset was split according to an 85:15 train-test ratio, preserving the target proportions, and part of the training was designated to validation set.

2.2.1. Preprocessing categorical data. Categorical variables were encoded using one-hot encoding.

2.2.2. Preprocessing numerical data. Numerical variables were imputed using a KNN imputer with 5 neighbours and then standardized for letting the training converge.

2.2.3. Class imbalance. The dataset was a bit imbalanced, with 39.6% of observations belonging to class 1 (positives). To tackle the problem, whenever the model allowed for it, we passed class weights based on the inverse of percentage of samples over the total.

2.2.4. Performance indexes. The main metric used to compare models was the Area Under Curve (AUC) under the Receiver Operating Characteristic (ROC) curve.

2.3. Feature selection

Ideally we wanted to perform a backward selection. However, since the initial number of features is 131, the process is very computationally intensive (about 20 seconds per feature to be removed) and greedy, so we would risk discarding important variables too soon. To speed up the process we developed the following method:

- Train a Logistic Regression (LR) with strong L^1 penalty to select a set S_{LR} of features.
- Train a Random Forest (RF) and create an S_{RF} set of features made by the top 30 features by impurity decrease importance.
- Create a set $S_{\text{reduced}} = S_{\text{LR}} \cup S_{\text{RF}}$. The reason for this is that the way LR and RF select variables is very different, in fact the intersection $S_{\text{LR}} \cap S_{\text{RF}}$ was very small. In this way we take advantage of both.

We could now perform a backward selection based on this S_{reduced} set of features, and inspecting the evolution of the AUC we selected a suitable number of features $S_1^\star$. Then, since the results are often different, we performed a forward selection from 0 to 10 features on the same model to get the set $S_2^\star$, and finally we took the union of the two as our final features set $S^\star = S_1^\star \cup S_2^\star$ of 13 variables.

For visualization purposes, in Figure 2 we plotted the KDE of some features in $S^\star$ separately with the respect to the target.

It's really clear the difference in distribution of the `creatinine.enzymatic.method`, we expect it to be important in the final model.

2.3.1. Model selection. We trained 6 different classifiers on the unscaled selected features with a 5-fold stratified cross-validation to tune the hyperparameters.

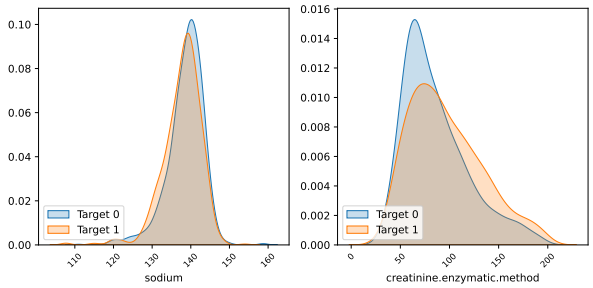

Figure 2. Distribution with the respect to the target of sodium and creatinine enzymatic method.

Besides those, we made an attempt with an SVM, but the training failed to converge except on the scaled data (which makes interpretability harder) and didn't perform well, therefore we don't include it.

3. Results

3.1. Model comparison

Table 1 shows the performance of the different models on the test set after the hyperparameter tuning. Even if it's not the most performing model, we choose the LR as final model for its interpretability and simplicity, at the cost of just one percentage point in performance with respect to the RF.

Figure 3 shows the ROC curves of all the models.

TABLE 1. COMPARISON OF PERFORMANCE.

Model	AUC
RandomForestClassifier	0.6769
LogisticRegression	0.6702
GaussianNB	0.6452
DecisionTreeClassifier	0.5943
KNeighborsClassifier	0.5681
MLPClassifier	0.5028

3.2. Model analysis

Given a vector of variables $\boldsymbol{X} = X_1, \ldots, X_p$ the LR model assumes the following relationship between the variables and a prediction score $p(\boldsymbol{X})$, where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown real coefficients to be numerically estimated.

$$\begin{aligned}\text{logit}(p(\boldsymbol{X})) = \log(\text{odds}(p(\boldsymbol{X}))) &= \log\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right) \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\end{aligned}$$

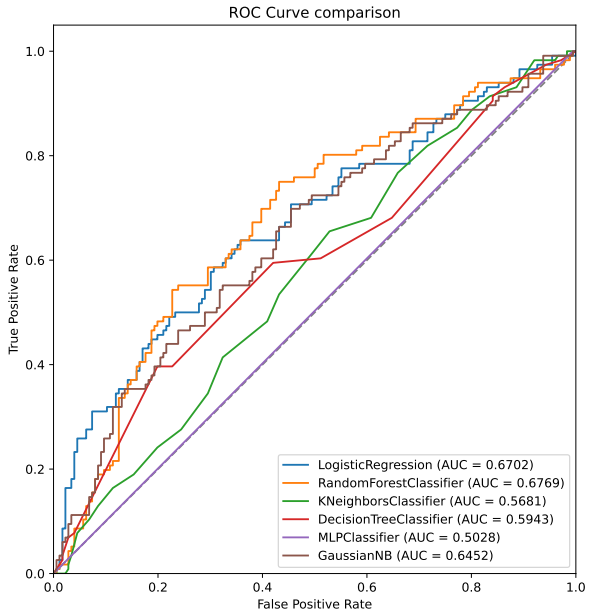

Figure 3. ROC curves comparison.

TABLE 2. LOGISTIC REGRESSION COEFFICIENTS.

feature	beta	exp_beta
occupation_farmer	-0.6973	0.4979
glutamic.pyruvic.transaminase_log	-0.1305	0.8776
D.dimer	-0.0911	0.9129
partial.pressure.of.carbon.dioxide	-0.0261	0.9743
sodium	-0.0049	0.9951
basophil.ratio	0.0055	1.0055
creatinine.enzymatic.method	0.0066	1.0066
dischargeDay	0.0294	1.0298
eosinophil.ratio	0.0486	1.0498
NYHA.cardiac.function.classification_IV	0.3599	1.4332
diabetes_True	0.4049	1.4991
international.normalized.ratio	0.4074	1.5029
type.of.heart.failure_Both	0.5502	1.7336

We can give an interpretation to the coefficients: if $\mathbf{1}_i$ is a vector of all zeros, and with a 1 in i-th position:

$$\mathrm{OR}_i = \frac{\mathrm{odds}(\boldsymbol{X}+\mathbf{1}_i)}{\mathrm{odds}(\boldsymbol{X})} = \frac{\left(\frac{p(\boldsymbol{X}+\mathbf{1}_i)}{1-p(\boldsymbol{X}+\mathbf{1}_i)}\right)}{\left(\frac{p(\boldsymbol{X})}{1-p(\boldsymbol{X})}\right)}$$
$$= \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_i(X_i+1)+\cdots+\beta_p X_p}}{e^{\beta_0+\beta_1 X_1+\cdots+\beta_i X_i+\cdots+\beta_p X_p}} = e^{\beta_i}$$

The odds multiply by e^{β_i} for every 1-unit increase in X_i.

In Table 2 we report the β_i coefficients for the final model, together with e^{β_i}.

TABLE 3. CLASSIFICATION REPORT.

	precision	recall	f1-score	support
False	0.7244	0.6420	0.6807	176
True	0.5368	0.6293	0.5794	116
macro avg	0.6306	0.6357	0.6300	292
weighted avg	0.6498	0.6370	0.6405	292

3.3. Performance evaluation

The confusion matrix produced by the LR obtained with the default 0.5 threshold of Scikit-learn can be seen in Figure 4, while Table 3 contains all the classification metrics. The Accuracy is 0.6370.

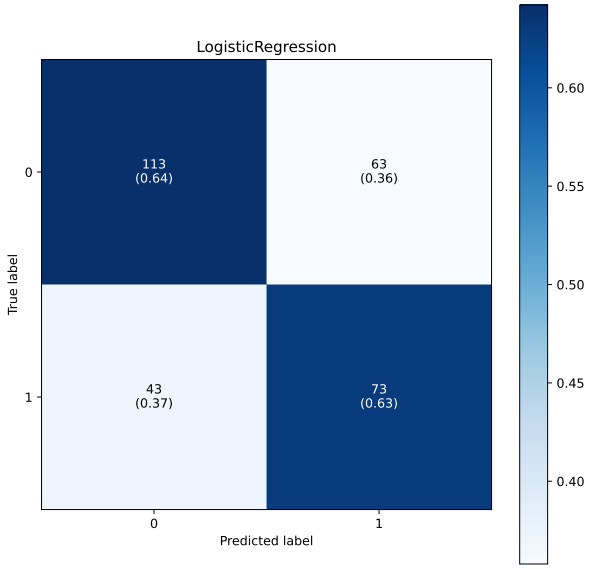

Figure 4. Normalized confusion matrix.

3.4. Web App

Trying to address the needs of clinicians who need user friendly tools to put our statistical results into practice, we developed a web app using the Python library `streamlit`, to easily input values and get predictions from our final model.

The web app is accessible through the following link: https://teobucci-slhd-app-3iahgf.streamlit.app/

4. Discussion and conclusion

The two most performing models are the RF and the LR with comparable AUC, but we chose the latter

one for the sake of interpretability. The GaussianNB shows promising performance, while we clearly discard the neural network approach by MLPClassifier.

Looking at the coefficients of the LR in Table 2 there is a strong effect caused by the presence of diabetes, which accounts for an increase in the odds ratio by 1.5; in fact, it is known that diabetes can be responsible for the development of lesions in arteries that are the basis of major cardiovascular diseases. The presence of a level 4 NYHA accounts for an increase in the odds of 1.43 in favour of a re-admission, an it is indeed symptomatic of severe HF, which causes the patient to be bedridden or have serious limitations in their daily activities. Having suffered from a HF of both Left and Right ventricles also accounts for an increase in the odds of 1.7, the strongest of all.

Higher D-dimer and log-transformed glutamic-pyruvic transaminase were associated with lower risk of re-admission. D-dimer, in healthy conditions, should not be present in the blood, except when a coagulation process is active. High D-dimer values may be a symptom of an increased level of tissue repair following HF. However, according to the literature, a high value is not sufficient to confirm this hypothesis. Strangely, if the patient belonged to the farmer group of occupation, it accounted for a protective odds ratio of about 0.5, but this is likely caused by some external confounder, such as diet and habits of Chinese farmers.

One of the goals was to address drugs importance. None of the 4 categories of drugs identified in subsection 2.1 made it to the final set of features, so we can't affirm that they are more relevant in predicting re-admission rather than biomarkers.

In fact, according to clinical practice,[3] most patients are treated with both diuretics – to counteract fluid retention – and vasodilators – to prevent the formation of clots – making these drugs ineffective in predicting readmission to 6 months. Inhibitors are also poor predictors, as is shown in the literature, because of how variable the effect can be on different patients. However, among them, the one that proved to be most informative was the IFHC that made it to the S_{reduced} set of features.

3. E. Lonn. "Regular review: Drug treatment in heart failure". In: *BMJ* 320.7243 (Apr. 2000), pp. 1188–1192.

4.1. Comparison with previous models

In previous studies[4] about a 30-day prediction framework, an AUC of 0.654 was reached using XGBoost, but the authors benefited from having data about prior hospitalization which were ranked as the most important features. Our model was trained without such data and less samples and provided a slightly better performance. The other important variables are aligned with our findings, like sodium and day of discharge.

4.2. Limitations

Limitations of this analysis are the difficulty in comparing results with other papers, given that different dataset provide very different variables.

This analysis is also quite biased given that the sample is not representative of the global population, as it comes from a city in China.

Our model also lacks data coming from electrocardiography, which would be well suited for the task.

4.3. Further development

We focused on simplicity and interpretability, but for the sake of performance only, one could keep more variables selected by the step-wise methods.

Furthermore, one could explore more advanced classifiers such as XGBoost or AdaBoost, and try an interpretation path through the Shapely value.

It would be interesting to repeat the analysis for the 28-day re-admission and see how the results change; however, this task is not said to be easier given a more prominent imbalance in the data for this class.

4.4. Recommendations

Overall we recommend a better management of missing values and a closer look to units of measure. We also iterate about checking the three variables mentioned in subsubsection 2.1.2.

4. V. Sharma et al. "Predicting 30-Day Readmissions in Patients With Heart Failure Using Administrative Data: A Machine Learning Approach". In: *Journal of Cardiac Failure* 28.5 (May 2022), pp. 710–722.