DeathPredict Report

Name:	Pierre-alexandre BARIL, Morgan VATERKOWSKI, Quentin LEFORT
Research Centre:	EPITA
Research Project Title:	DeathPredict
Primary Supervisor:	Antoine Liutkus

1 Abstract

The predictors of in-hospital mortality for intensive care units (ICU)-admitted HF (Heart Failure) patients remain poorly characterized. This is why we aim to predict if a patient will die from heart failure(HF) while they are in intensive care unit (ICU). For that we use the patient's different constant like 'age', 'diabetes', 'calcium in blood' and many others (48 of them).

2 Introduction

We aim to predict if a patient will die from heart failure(HF) while they are in intensive care unit (ICU). For that we use the patient's different constant like 'age', 'diabetes', 'calcium in blood' and many others (48 of them).

The repository is on github[1] and the code was developped with the help of kaggle[2].

3 Data

3.1 Data Source and collection

The MIMIC-III database (version 1.4, 2016) is a publicly available critical care database containing de-identified data on 46,520 patients and 58,976 admissions to the ICU of the Beth Israel Deaconess Medical Center, Boston, USA, between 1 June, 2001 and 31 October, 2012. These data include comprehensive information, such as demographics, admitting notes, International Classification of Diseases-9th revision (ICD-9) diagnoses, laboratory tests, medications, procedures, fluid balance, discharge summaries, vital sign measurements undertaken at the bedside, caregivers notes, radiology reports, and survival data12. After successful completion of the National Institutes of Health Protecting Human Research Participants web-based training course, we obtained approval to extract data from MIMIC-III for research purposes (Certification Number: 28860101).

3.2 Final Version

The data we used is a kaggle dataset[3] . The data come under an Excel (csv) file containing 51 columns for each of the 1177 patients.

4 Implementation

We decided to code under the Python language. Moreover, we used the machine learning package: sci-kit learn.

4.1 Other Implementations

This study[4] already worked on this dataset and came with this conclusion:
"Patients meeting the inclusion criteria were identified from the MIMIC-III database and randomly divided into derivation and validation groups. Independent risk factors for in-hospital mortality were screened using XGBoost and LASSO regression models in the derivation sample. Multivariable logistic regression analysis was used to build prediction models. Discrimination, calibration, and clinical usefulness of the predicting model were assessed using the C-index, calibration plot, and decision curve analysis. After pairwise comparison, the best performing model was chosen to build a nomogram according to the regression coefficients."

Prediction model of in-hospital mortality in intensive care unit patients with heart failure: machine learning-based, retrospective analysis of the MIMIC-III database

5 Experiences

5.1 Data cleaning

The first problem : Replacing the NaN value, what we did first was to replace NaN value with the average of the column.
However, this is not ideal. What we did was to use the K-NN method, Nearest Neighbor with k = 5.

5.2 Data Split

We split the data in two different set, the training set (70 percent) and the test set (30 percent) for that we used the "train_test_split" function which can be found in the "sklearn.model_selection" library

5.3 MLP

The first idea is to use a Multi-Layer Perceptron. We used a standart scaler to normalise all the dataset to feed integers to the perceptron.

However, this implementation gives poor results : 40 percent accuracy. Antoine advised to use trees, therefore leading to the next experience.

5.4 Trees

1. Decision Tree
2. Random Forest

5.5 AutoML

We used Pycaret module[5]. to train all models on our data in order to compare them all.

Figure 1: AutoML model training results

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
ridge	Ridge Classifier	0.90	0.00	0.08	0.55	0.14	0.12	0.19	0.01
lr	Logistic Regression	0.89	0.75	0.13	0.32	0.17	0.14	0.17	0.60
rf	Random Forest Classifier	0.89	0.78	0.01	0.10	0.02	0.02	0.03	0.35
et	Extra Trees Classifier	0.89	0.78	0.00	0.00	0.00	0.00	0.00	0.29
xgboost	Extreme Gradient Boosting	0.89	0.77	0.09	0.36	0.14	0.11	0.14	14.36
lightgbm	Light Gradient Boosting Machine	0.89	0.77	0.13	0.43	0.19	0.16	0.19	0.25
catboost	CatBoost Classifier	0.89	0.78	0.03	0.20	0.06	0.05	0.08	6.92
dummy	Dummy Classifier	0.89	0.50	0.00	0.00	0.00	0.00	0.00	0.01
knn	K Neighbors Classifier	0.88	0.62	0.06	0.30	0.10	0.06	0.09	0.07
ada	Ada Boost Classifier	0.88	0.71	0.21	0.35	0.25	0.20	0.21	0.11
gbc	Gradient Boosting Classifier	0.88	0.74	0.11	0.31	0.15	0.11	0.13	0.33
lda	Linear Discriminant Analysis	0.88	0.77	0.20	0.34	0.24	0.19	0.21	0.01
qda	Quadratic Discriminant Analysis	0.85	0.54	0.06	0.11	0.04	0.01	0.02	0.01
dt	Decision Tree Classifier	0.83	0.56	0.22	0.23	0.22	0.13	0.13	0.02
nb	Naive Bayes	0.82	0.75	0.43	0.31	0.35	0.25	0.26	0.01
svm	SVM - Linear Kernel	0.81	0.00	0.21	0.11	0.13	0.06	0.07	0.01

We can find again the scores we had previouvsly with the trees. Ridge Classifier has the best accuracy score, however the Naive Bayes has better Recall and F1-Score.

5.6 Naive Bayes

6 Discussion & Conclusion

The concrete proof that the accuracy score is not enough to choose the best model was demonstrated here. Indeed, the model with the highest accuracy score is "Ridge Classifier". It is best at predicting the survival outcome, because of the imbalanced data. However, it fails at predicting death outcome, which is about 13 percent of the total data. For that , we have to look at the F1-score.

Interestingly enough, the Naive Bayes model is the best for predicting death and survival, which is equivalent for the highest F1-score.

Notes

1. See https://github.com/PierreAlexandreDev/DeathPredict/ for more information.
2. See https://www.kaggle.com/pierrealexandre78/deathpredict for more information about the main code and history.
3. See https://www.kaggle.com/saurabhshahane/in-hospital-mortality-prediction
4. See https://datadryad.org/stash/dataset/doi:10.5061/dryad.0p2ngf1zd
5. Pycaret Module: https://www.kaggle.com/sonalisingh1411/pycaret-automl-heart-failure-prediction