ETH zürich

363-1098-00L Business Analytics

Project Report

Authors:	*Student Number:*
Belén Cantos Bernal	18-941-732
Enxhi Gjini	18-948-075
Joan Puig Sallés	18-950-238

May 28, 2019

Contents

1 Introduction

Measuring and predicting patient health is the main issue in critical care research. One of the most important outcomes for admission in the Intensive Care Unit (ICU) is the risk of patients' mortality. Predicting this indicator could help practitioners to efficiently assess the severity of illness and effectively allocate resources for treatments and interventions.

This project aims to focus on predicting patients' death probability using different machine learning algorithms and compare their performance. To achieve this goal, we have chosen the Medical Information Mart for Intensive Care III (MIMIC III) [1]. This publicly available database contains health-related information from over 50,000 patients, admitted to ICUs at the Beth Israel Deaconess Medical Center in Boston, Massachusetts.

2 Methods

This section starts with a description of the chosen dataset and the methods used for data retrieval and processing. Next, a description of the exploratory analysis conducted is given. The last part is devoted to predictive modeling. The programming languages used to develop this project were R, for the data retrieval and processing and exploratory analysis part, and Python for the predictive modeling. The reason for choosing Python for the second part of this project was because of the SKLearn package. Thanks to the clean and intuitive API it offers, the excellent documentation and online support, it is very straightforward to develop and implement different machine learning algorithms using this Python package.

2.1 Data Retrieval and Processing

2.1.1 Dataset Description

MIMIC-III (Medical Information Mart for Intensive Care) is a large, publicly available database comprising information relating to over 58,000 patients (38,645 adults and 7,875 neonates), admitted to ICUs at the Beth Israel Deaconess Medical Center in Boston, Massachusetts. The data spans from June 2001 to October 2012 and is de-identified, for confidentiality reasons. It includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more.

2.1.2 Dataset Preprocessing

Following the approach by Johnson et al. [2], we first identified all adult patients in the dataset by filtering out those whose age was under 15. Second, for each patient, we use their first admission for subsequent analysis. Admissions with no charted observations, or an incomplete recording of ICU admission and discharge, were removed.

There are a total of 26 tables in the MIMIC-III relational database. For benchmarking purposes and as suggested by Purushotham et al [3], we extracted data for the selected cohort from the *outputevents*, *labevents* and *prescriptions* tables. Due to the size of *chartevents*, we did not extract features from this table. These tables include the 19 features used for the prediction of the risk of mortality. 14 of these features correspond to the calculation of the SAPS-II score, used to measure the severity of disease for adults [4], namely, the type of diagnosis, ICD9 code, oxygen levels, systolic blood pressure, urine output, serum urea nitrogen level, white blood cells count, serum bicarbonate level, sodium level, potassium level, bilirubin level, age, gender, and admission type. The rest of the features correspond to socioeconomic factors, such as admission location, type of insurance, marital status, and ethnicity.

After extracting the features, we performed a last cleaning of the data. First, for each set of features, outliers were removed, as most of them corresponded to data entry mistakes. Finally, to fill in missing values, we performed a mean imputation to all variables that were 0 or with a *Na*.

2.2 Exploratory Analysis

The data consists of a dependent variable (death), and 19 predictors, or independent variables. Among the predictors, 8 are categorical and the rest are continuous. As for the categorical ones, which are summarised in Table 1, they offer some insights about the data: (i) the patients that died are, on average, about 8 years older than the ones that did not; (ii) a significant majority of males dying over women, around 12 % more; (iii) regarding insurance contract, privately insured patients that did not die, is 14 percentage points higher than for the ones that did die; and (iv) concerning the origin of patients, we observe that the majority of the patients that died, did not plan the admission but came either from the emergency room or from inside the hospital.

	Overall	Dead at the hospital	Alive at the hospital
General			
#admissions	38569	4336	34233
Age	63.7	70.4	62.8
Gender(female)	16730 (38%)	2044 (47%)	14686 (43%)
Length of stay	4 days	5.8 days	3.8 days
Insurance			
Medicare	20339 (55.3%)	2899 (69.5%)	17500 (53.5%)
Private	13289 (36.1%)	990 (23.7%)	12299 (37.6%)
Medicaid	3174 (8.6%)	285 (6.8%)	2889 (8.8%)
Origin			
Emergency / Urgent	32300 (83.7%)	4181 (96.4%)	28119 (82.1%)
Planned admission	6269 (16.3%)	155 (3.6%)	6114 (17.9 %)

Table 1: Baseline characteristics and in-hospital mortality measures from our cleaned dataset

For the numerical predictors, which mainly contain lab tests and fluids into/out patients, we have plotted a correlation matrix (see Figure 1). A priori, we cannot reach any conclusion from this outcome, as there are just low correlations between the variables and the death label, and some irrelevant correlations between the variables.

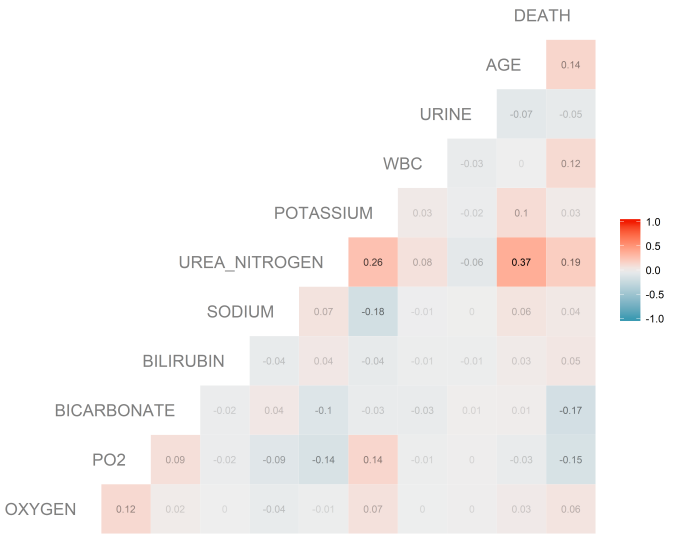

Figure 1: Correlation matrix of the numerical predictors

2.3 Predictive Modeling

Predicting mortality is one of the main issues of interest in hospital admission. We have formulated mortality as a binary classification task, where 1 indicates the event of death for a patient. The goal is to predict whether a patient will or will not die, based on the 19 clinical and socioeconomic features (independent variables).

Seven different machine learning algorithms have been used to predict the risk of mortality: four linear models (Logistic Regression, Lasso Regression, Elastic Net, Bayesian Ridge) and three ensemble methods (Gradient Boosting, Bagging, and Random Forest). These models have been implemented using Scikit-learn, a free software machine learning library for the Python programming language.

While linear models are more straightforward to interpret, they often perform worse on many prediction tasks compared to less interpretable models like gradient boosting. When predicting the risk of mortality, the imperative for supporting explanations for predictions is of great value, as the risk score may drive critical care decisions [5]. For this reason, we should find a compromise between performance and interpretability. That is, a model that performs well but is also easy for practitioners to understand. Besides, since the cost of misclassifying a patient with a high risk of mortality as healthy (false negative) is much higher than the cost of reverse error (false positive), we have used recall as the main predictive performance metric.

One of the main issues we had to pay special attention to was that some of our features, which were used as predictors, were not numerical values but categories. To solve this issue we used label encoding for converting the categorical labels into numeric ones.

Another issue we encountered was that the data was heavily unbalanced. As Table 1 indicates, out of the 38,569 patients, only 12% died during their stay at the hospital. To balance the data, we increased the number of dead patients in the sample using a Synthetic Minority Over-sampling Technique [6].

3 Results

This section presents the key findings and summarizes the results.

Table 2 summarizes accuracy, recall and F1 score for each of the algorithms applied. As we need to minimize the risk of misclassifying ill patients as healthy ones rather than maximizing the overall accuracy, we have mainly focused on recall and F1 score as a metric of performance instead of accuracy.

The comparison among the seven models can be best realized by means of the Receiver Operating Characteristics (ROC) curve (see Figure 2). This tells us how well each model distinguishes between classes. The ROC curve shows the True Positive Rate (TPR) over the False Positive Rate (FPR) for different configurations of the model. The model chooses a pair - TPR, FPR - to maximize the area under the curve, which approximately corresponds to the accuracy of the model. This figure shows that some of the models perform better than others in the whole span of the curve.

Overall, we obtained relatively low recall and F1 score values for all the models. The Elastic Net algorithm scored the maximum recall: 0.696. As for the overall performance, Gradient Boosting outputs the best F1 score at 0.416. However, even though it managed to get a very high accuracy (0.856), its recall was still very low, about 50%, which means that it wrongly classifies as healthy, half of the patients at high risk of mortality. Hence, this algorithm is not the most appropriate to be used in hospital settings.

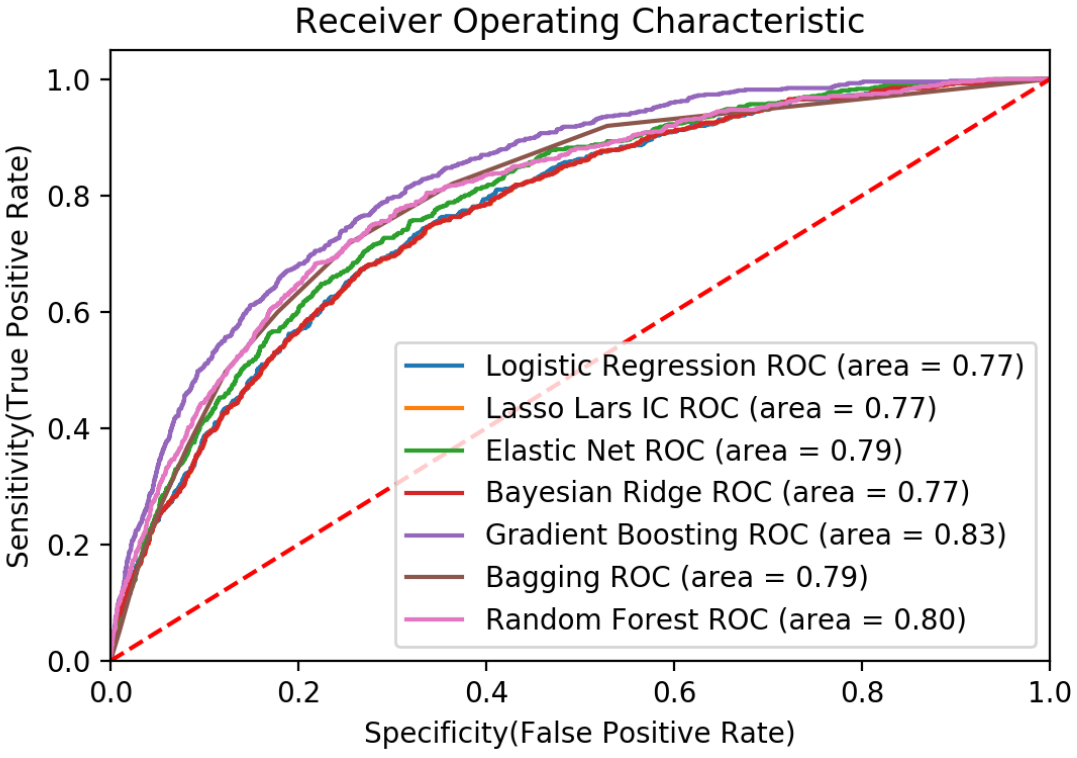

Figure 2: Receiver Operating Characteristic (ROC) curve for the differents models we tested

Algorithm	**Accuracy**	**Recall**	**F1 Score**
Logistic Regression	0.748	0.648	0.345
Lasso Lars IC	0.742	0.652	0.341
Elastic Net	0.719	**0.696**	0.337
Bayesian Ridge	0.742	0.652	0.342
Gradient Boosting	0.856	0.499	**0.416**
Bagging	**0.867**	0.351	0.351
Random Forest	0.751	0.66	0.352

Table 2: Summary of the Accuracy, Recall and F1 scores of the models.

4 Discussion

To achieve the goal of our project, that is to predict the risk of patients' mortality accurately, we have used various machine learning algorithms on the MIMIC III medical database.

Among the seven machine algorithms applied (Logistic Regression, Lasso Regression, Elastic Net, Bayesian Ridge, Gradient Boosting, Bagging, and Random Forest), three stand out: Bagging, Gradient Boosting, and Elastic Net.

Bagging and Gradient Boosting are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one. Even though these models output the best accuracy and F1 scores, they fail to score a high recall. What is more, both become very difficult for medical staff to interpret.

This contrasts with the Elastic Net, a linear regression model with combined L1 and L2 regularization. It shows the highest recall and interpretability, which is prime for the goal we set for this project: finding an interpretable machine learning model for doctors, with immediate effects on patients' wellbeing. In addition to being the most interpretable model, the Elastic Net algorithm offers doctors very valuable information using the weight coefficients. These coefficients help practitioners understand how predictions were made (see Prediction Example in the Appendix). This explains why this model is commonly applied in medical settings. Furthermore, as all other models, Elastic Net also outputs the prediction probability for each assessment.

In the Appendix, we illustrate a case scenario of our mortality prediction model. Once a new patient is admitted to a hospital, and the relevant feature values are gathered, the doctor can use them to predict the mortality rate employing our model. The result does not only indicate whether the patient is at high risk of mortality, but it also supports

doctors with the probability of such prediction.

Regarding the model applications, there are some limitations we need to consider: (i) the features used to predict mortality can only be obtained once a patient has been admitted to a hospital; (ii) the model is restricted to adults (>15 years); and (iii) the model has been trained on a single hospital, this could affect its predictive performance and lead to biased results, making it potentially not suitable for other hospitals.

As for the model prediction performance, we are confident that the results obtained in this project are likely to be improved if we (i) increased the number of independent variables; (ii) used more advanced techniques for data imputation; (iii) improved label imbalance; and most important, (iv) increase the data set with a representative sample of hospitals. This would result in a more reliable and valid prediction model for mortality.

We are confident to have fulfilled the primary goal of this project. Not necessarily about having come up with the most reliable and accurate mortality prediction model, but to have contributed in having shed new light to a fast-growing and new research line in medical care. This synergy between data analytics and medicine will bring significant advances in risk assessment, medical diagnosis, and therapy strategies.

References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[2] Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Reproducibility in critical care: a mortality prediction case study. In *Machine Learning for Healthcare Conference*, pages 361–376, 2017.

[3] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmark of deep learning models on large healthcare mimic datasets. *arXiv preprint arXiv:1710.08531*, 2017.

[4] Jean-Roger Le Gall, Stanley Lemeshow, and Fabienne Saulnier. A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963, 1993.

[5] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 559–560. ACM, 2018.

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Appendix

Prediction Example for a new Admission

```
##NEW ADMISSION

# Please, fill in the following data:

# Patient Information:
Age = 65
Gender = 1 #Female
Admission_Type = 2 #Emergency
Admissin_Location = 3 #EMERGENCY ROOM ADMISSION
Insurance = 4 #Private
Religion = 0 #Not known
Marital_Status = 0 #Not known
Ethnicity = 3 #Asian
Diagnosis = 984 #AORTIC RUPTURE
ICD9_code = 7 #ICD-9 codes 390 459: diseases of the circulatory system

# ICU machines / Lab related data
Oxygen = 100
PO2 = 155.421
Bicarbonate = 20.3333
Bilirubin = 0.3
Sodium = 136.867
Urea_Nitrogen = 19.8667
Potassium = 4.70333
WBC = 7.13914
Urine = 3739.39
```

```
model_Elnet = ElasticNet(alpha = 0.01, l1_ratio = 0.1)
model_Elnet.fit(x_train_res, y_train_res)
y_patientX = model_Elnet.predict(PatientX)

if y_patientX > 0.5:
    y_pred_px = "POSITIVE"
    prob = 100*y_patientX
else:
    y_pred_px = "NEGATIVE"
    prob = 100 - 100*y_patientX

print("The newly addmited patient has been classified as %s" % y_pred_px +
"for the death prediction. With a probability of %f %%" %prob)
```

The newly admitted patient has been classified as NEGATIVE for the death prediction. With a probability of 93.700752 %