Berkeley
UNIVERSITY OF CALIFORNIA

Master of Engineering
Industrial Engineering and Operations Research

Final Report

Applications in Data Analysis

Student :
Emma Scharfmann

Professor :
Paul Grigas
GSI :
Hyungki Im
Hong-Seok Choe

February 4, 2022

Contents

1 Introduction

Our project focuses on predicting health outcomes of patients with Alcohol Use Disorder (AUD), who are admitted in Intensive Care Unit (ICU)

2 Motivation

Alcohol is one of the most commonly abused addictive substances in the general population, that causes millions of deaths, and is legal to buy and consume in most of the countries. People with AUD can't stop or control their alcohol consumption, and AUD is considered as a brain disorder. Patients with AUD have higher risks of comorbidities, serious illness and death, and this is why an early intervention and prognosis for severity of illness is highly important to improve treatment and survival rates of AUD patients.

We want to focus on two different predictions. The first one is the prediction of people's survival, when they are admitted to the Intensive Care Unit (ICU). The main idea is trying to strengthen medical care for those less likely to survive, and help them survive. But it can be used to increase doctor's knowledge about the chances of survival of those patients, and so better inform and support families and relatives in this difficult process. The second prediction is the length of stay in the hospital, as some patients can stay for a short period of time, and others can stay longer, depending on their diseases, and what has to be done. This would help hospital staff better plan the space requirements, the staff needed, how many people they could admit in ICU... As medical and hospital expenses are really expensive in the US, it would also help make better predictions of the medical costs when a person is admitted in an ICU.

3 Data

3.1 Data source

The data used in our report is entirely extracted from the Medical Information Mart for Intensive Care (MIMIC-III), a freely-available database comprising deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center. This database was co-developed by the Laboratory for Computing Physiology at MIT and Philips Healthcare, containing 26 tables storing data such as vital signs, demographics, and survival for more than 40,000 ICU patients from years ranging from 2001 to 2012. This database is widely used in academic discourse for training models to predict medical outcomes for patients.

3.2 Data Extraction and Cleaning

A database connection using Postgres-SQL was necessary to extract data from MIMIC. We identified AUD patients using ICU-9 codes that indicate alcohol abuse. The paper "Red Blood Cell Distribution Width as a Predictor of 28-Day Mortality in Critically Ill Patients With Alcohol Use Disorder" creates a linear regression model predicting mortality of AUD patients. The features used in this model provided insight into what medical indicators are important factors for health outcomes of AUD patients. As a result of this research, 40 features were chosen including: demographics, comorbidities, lab measurements, sequential organ assessment (SOFA) score and simplified acute physiology (SAPS II) score. Patients were excluded if they were under 18 years old, had been to the ICU previously, or stayed in the ICU for less than 24 hours. The final dataset contained 3435 rows when null values were excluded.

4 Data Analysis

4.1 EDA

To understand underlying patterns in our data and describe it better, we have performed a univariate analysis of all the independent variables.

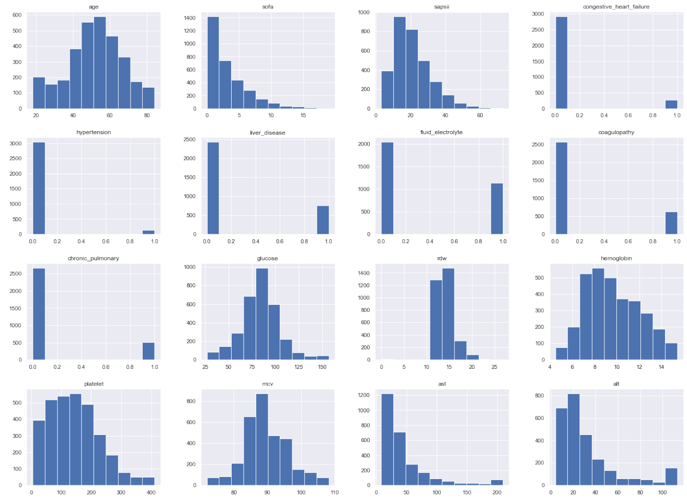


Figure 1: Final Univariate Plots

While performing this, our first observation was that a couple of our variables are heavily skewed and many of them have a lot of outliers. Hence for outlier treatment, we have adopted a modified version of 'winsorisation' (Imputing extreme variables to certain percentile limits). We have clipped the outliers to reasonable values based on intuition to ensure that underlying information is not lost but outliers are taken care of. We have also removed all the variables which have a considerable amount of nulls present from our datasets to ensure better performance.

Then we moved ahead to performing a bivariate analysis on our dataset to understand the relationship between our independent variables and our target variables. For this we plotted scatter plots and a correlation heatmap. We have used the Spearman correlation method instead of Pearson as medical data usually contains a lot of outliers
Please find below a heatmap of our dataset

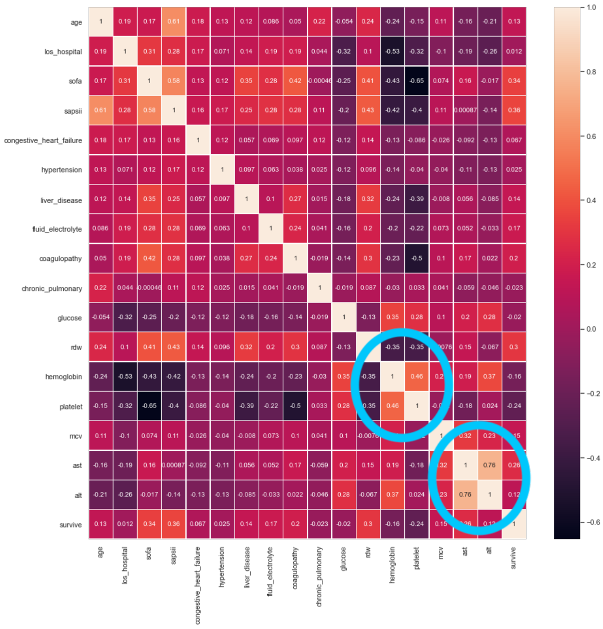


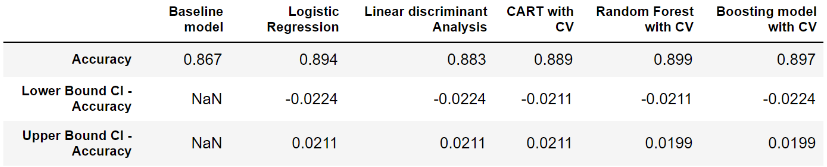
Figure 2: Heatmaps

We can see that the pair ('ast','alt') has a excessively high correlation while the pairs ('platelet','hemoglobin') and ('platelet','sofa') have moderate correlation.
We have removed the variable 'alt' from our classification models and 'ast' from our regression model. Removal has been done based on their predictive performances, with respect to our target variables. The variables SOFA and SAPSII have been derived from the physiological conditions of the patient, which can explain this high correlation as well as their high correlation with other variables. Therefore, we have removed the variable SOFA from all our models.

Here onwards, our dependent variables don't have any strong predictors as the correlation map shows. Now, we are going to apply different predictive methods (linear methods, decision tree, boosting and random forest methods) on the clean data to predict the survival of patients and the length of stay of patient.

4.2 Predicting the survival of a patient

The first prediction that we have realized is whether a patient would survive or not. For this classification problem, we have created the dependent variable *Survive_or_not* from the column *Death_time* which has missing values only if the patient survives. Then we have dropped the columns *los_hospital* and *los_icu* which indicates the length of stay of a patient in hospital and ICU respectively. Indeed, in reality, these data are available only once the patient dies or survived but not when he/she arrives at the hospital. Before making any prediction, we have split the database in order to have the training (70% of the database) and test set. Then, we have built different classification models such as baseline, logistic regression, LDA, decision tree, random forest and gradient boosting. The goal was to find the best model that would maximize the accuracy. Therefore, we have used cross validation to find some parameters as CCP_α for CART, maximum features m for random forest and the number of estimators for boosting. Finally, we have implemented boostrap and confidence interval (95%) to find the variation of the accuracy on the test set of each model. All in all, we have obtained the following results:



	Baseline model	Logistic Regression	Linear discriminant Analysis	CART with CV	Random Forest with CV	Boosting model with CV
Accuracy	0.867	0.894	0.883	0.889	0.899	0.897
Lower Bound CI - Accuracy	NaN	-0.0224	-0.0224	-0.0211	-0.0211	-0.0224
Upper Bound CI - Accuracy	NaN	0.0211	0.0211	0.0211	0.0199	0.0199

Figure 3: Results for various classification models

All these models seem to work well on the test set since the accuracy is very high and is higher than the baseline model (which has a high accuracy too). Moreover, the boostrap shows that the accuracy has a low variation reflecting the robustness of the models. If we look to the highest accuracy with the lowest variation, we can say that the gradient boosting model with cross validation is the best one. The AUC of the boosted model to be around **0.71** indicating good performance. We also have gotten the feature importance tabulated and seen that the features (SAPS II,ast and platelet) have the highest influence over the survival of the patient.

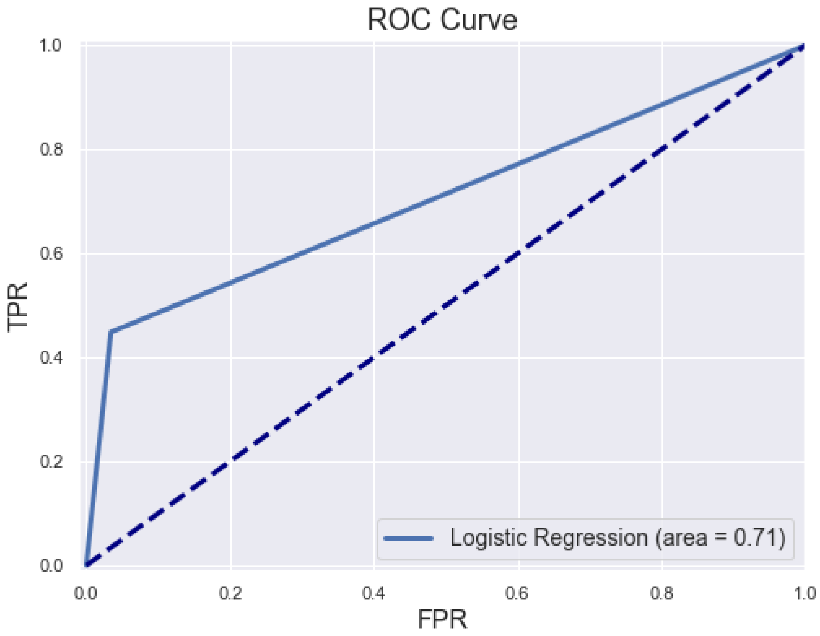
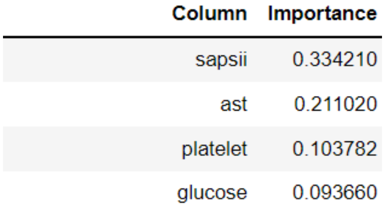


Figure 4: ROC curve

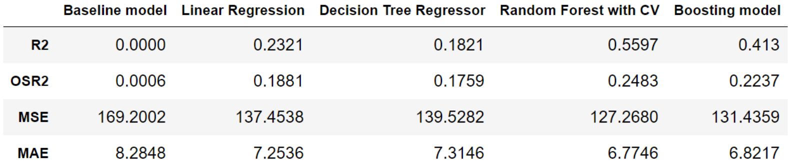


Column	Importance
sapsii	0.334210
ast	0.211020
platelet	0.103782
glucose	0.093660

Figure 5: Important Features

4.3 Predicting the length of stay in hospital

The second data analysis that we have conducted on the database is the prediction of the length of stay in the hospital for an alcoholic patient. Unlike the previous prediction, we had to use the column *los_hospital* since it was our dependent variable. As before, we have split the database to have the training and test sets (70-30%). In this regression problem, we have created various models such as linear regression, decision tree regressor, random forest. For the metrics, we focused on R^2, OSR^2, MAE and MSE. Once more, we have selected some parameters such as CCP_α for CART, maximum features m for random forest and the number of estimators for boosting with cross validation. Finally, we got the following results:



	Baseline model	Linear Regression	Decision Tree Regressor	Random Forest with CV	Boosting model
R2	0.0000	0.2321	0.1821	0.5597	0.413
OSR2	0.0006	0.1881	0.1759	0.2483	0.2237
MSE	169.2002	137.4538	139.5282	127.2680	131.4359
MAE	8.2848	7.2536	7.3146	6.7746	6.8217

Figure 6: Results for different various regression models

Here, the baseline model has an $OSR^2 = 0$, a $MSE = 169$ and a $MAE = 8.28$. The linear regression model,the boosting model and the random forest performs reasonably : (OSR^2, MSE and MAE better than the baseline model). Even if the models don't perform extremely well, the random forest model seems to be the best model we have to predict the length of stay in hospital with the higher OSR^2 and the lower MSE and MAE.

5 Conclusion

To conclude, a boosting model and a random forest model trained on the data we found seem to be the most adapted models to predict the survival and the length of stay of alcoholic patients in hospital. As for the future, we need to find better variables about the patient's medical conditions - for instance - patients' comorbities. A couple of transformations on the numeric variables might also help us get a better fit. We also need to factor in the patients' income status as the stay of hospital usually depends on the affordability of the hospitals as well