AY2021 Sem 2 BC2407 Computer Based Assessment Homework

In-Hospital Mortality of ICU Patients with Heart Failure

Introduction

Hospital admits critically ill patients into Intensive Care Unit (ICU) for close monitoring and medical intervention. In general, ICU patients have higher risk of death compared to non-ICU patients. A real dataset from a hospital is publicly available (see data01.csv). The data contained ICU patients with heart failure and 51 variables. The dataset and summary statistics were explained in Li F., et al (2021) research paper[1]. Patients who died are coded outcome = 1 in the dataset. They proposed a new risk scoring model Nomogram to assess mortality risk.

The American Heart Association Get With the Guidelines – Heart Failure (GWTG) is an established risk model popularly used to assess the risk of mortality among heart failure patients at hospital. This simple model was explained in Peterson P.N. (2009).

In this assignment, you will analyzed the dataset and evaluate the predictive performance of Logistic Regression, Random Forest, GWTG and Nomogram.

Part A: Data Exploration and Preparation (20%)

1. Data Preparation Part 1:

    a. Import the csv dataset and name it **data1** and ensure that all categorical data are treated as categories instead of integers, numeric or text string characters. Show your code.

    b. Li F, et al (2021) used the terms Derivation group and Validation group. What is the purpose of the two groups and how is this reflected in the dataset?

    c. There are missing values in the dataset. Show a table of missing value counts that shows all those variables that has missing values, it's data type (numeric, integer, factor, character, etc.) and it's missing value count. The structure of your table should look like this:

	Variable.Name	Data.Type	NA.Count
1	NA	NA	NA
2	NA	NA	NA
3	NA	NA	NA
4	NA	NA	NA

    d. Explore data1. Produce charts, tables or/and statistics to explain 3 interesting findings.

[1] Two research papers are provided for your reading. You do not need to know XGBoost, LASSO or medical terms to complete this assignment.

2. Data Preparation Part 2:

    a. Create a copy of data1 and named it data2. In data2, replace all missing values by the median if the variable is continuous or the mode if the variable is categorical. Check that data2 has no missing values. Show the code used to check for missing value count and the output. Henceforth, all analysis will be performed on data2 unless mentioned otherwise.

    b. Produce a trainset from data2 using group = 1 and named it trainset. Remove group and ID from the trainset and show the proportion of cases who died vs alive.

    c. Produce a testset from data2 using group = 2 and named it testset. Remove group and ID from the testset and show the proportion of cases who died vs alive.

Part B: Analytics and Insight (50%)

Set seed as 22 before running each technique that requires randomization.

In all models, set the threshold as 50%. i.e. model predicts patient will die if P(death) > 50%.

3. Briefly explain (in bullet points) how you would compute the GWTG predicted outcome on the dataset.

4. Briefly explain (in bullet points) how you would compute the Nomogram predicted outcome on the dataset.

5. Show the table of trainset errors (false positive[2] rate, false negative rate, overall error) from Logistic Regression with Backward Elimination[3], Random Forest with default settings, GWTG and Nomogram. Your table structure should look like this:

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	NA	NA	NA
2	Random Forest	NA	NA	NA
3	GWTG	NA	NA	NA
4	Nomogram	NA	NA	NA

    Briefly state your findings.

[2] Positive event is defined as the model predicts patient will die.

[3] Refer to Chew C.H. (2021) AI, Analytics and Data Science Vol. 1 Chap 7, or BC2407 session 2 or BC2406 unit 7. If you cannot execute Backward Elimination, you can use the standard logistic regression with all variables or significant variables only but revise the model name in the table of errors accordingly.

6. Show the table of testset errors (false positive rate, false negative rate, overall error) from Logistic Regression with Backward Elimination, Random Forest with default settings, GWTG and Nomogram. Your table structure should look like the above table too. Briefly state your findings.

7. The testset errors might be skewed to one side as the trainset is unbalanced. Balance the trainset by sampling from the majority[4] to obtain 50-50 distribution of alive vs death in the trainset. Show the table of testset errors (false positive rate, false negative rate, overall error) from Logistic Regression with Backward Elimination and Random Forest with default settings. Your table structure should look like the below. Briefly state your findings.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	NA	NA	NA
2	Random Forest	NA	NA	NA

8. Li F., et al (2021) used the top 20 variable importance from XGBoost as predictors into logistic regression. Extract the top 20 variable importance (permutation approach) from Random Forest trained on balanced trainset and fit them as predictors into logistic regression. Append the testset results and show the table. Your testset errors table structure should look like the below. Is this model superior than stand-alone logistic regression (with backward elimination) or Random Forest? Briefly state your findings.

	Model	FPR	FNR	Err
1	Logistic Reg (BE)	NA	NA	NA
2	Random Forest	NA	NA	NA
3	RF VarImpt into Logistic Reg	NA	NA	NA

Part C: Conclusions (30%)

9. A hospital in Singapore is thinking of using a risk scoring system to assess ICU patient mortality. What is your recommendation?

10. Suggest other ideas that may improve the accuracy of the model.

[4] Refer to Appendix A: Sample Rcode for Sampling the Majority to Create Balanced Trainset, for an example. You may modify and use this Rcode or use another code. There are many ways to code.

Appendix A

Sample Rcode for Sampling the Majority to Create Balanced Trainset

Note: You may use other code. There are many ways to code.

Sources:

- Chew C.H. (2021) AI, Analytics and Data Science, Vol. 1, Chap 8 (CART), Cengage.
- RScript default CART.R from BC2406 Analytics I CART topic.

```
# Random sample from majority class Default = No and combine with Default =
Yes to form new trainset -----
majority <- trainset[Default == "No"]

minority <- trainset[Default == "Yes"]

# Randomly sample the row numbers to be in trainset. Same sample size as
minority cases.
chosen <- sample(seq(1:nrow(majority)), size = nrow(minority))

# Subset the original trainset based on randomly chosen row numbers.
majority.chosen <- majority[chosen]

# Combine two data tables by appending the rows
trainset.bal <- rbind(majority.chosen, minority)
summary(trainset.bal)
## Check trainset is balanced.
```