Datathon 4, Date:13th Nov, 2023 **Group: unspecified (99) Team members:** Benjamin (Benny) Zhang

Most Important Features in Classifying a Person's Mortality Risk Prior to admission into the ICU

Introduction

Accurate prediction of mortality is of critical importance in clinical medicine. Early and precise mortality risk assessments allow for the effective allocation of limited resources and ensures hospitals are working optimally in saving and bettering lives. In a face-paced environment, every saved second counts. While several scoring systems and evaluation programs have been used to help predict patient mortality rate using data from the first 24 hours of ICU admittance, less is known about the predictive accuracy using data provided prior to this point. Gradient Boosting is a powerful machine learning algorithm that is particularly suitable for handling structured data and is used both in classification and regression tasks. Thus, the goal of our study is to investigate whether it is possible to use information available prior to formal ICU admittance to accurately predict patient mortality rate using tree-based classification models.

Data Engineering Process

Using the mortality dataset provided from MIT's GOSSIS community initiative, over 91,000 ICU patient outcomes were observed and a subset of features were selected based on their availability prior to ICU admission. Numerical and categorical variables were specified from these selected features and processed. All continuous variables such as age, bmi, days spent in the hospital prior to icu admission (pre_icu_los_days), and weight were inputted as numerical variables into our tree-building pipelines. Conversely, binary variables containing information on gender, icu readmission status, aids, cirrhosis, diabetes mellitus, hepatic failure, immunosuppression, leukemia, lymphoma and presence of solid tumor with metastasis were inputted as categorical variables. Multinomial categorical variables such as the location in the hospital they were staying at prior to ICU admission (hospital_admit_source), the type of ICU they were admitted to (icu_type), and the ethnicity of the individual (ethnicity), however, needed to be processed differently. For each, dummy variables were created to represent each possible response category, and through this method, these categorical features were imputed for analysis. Observations that had missing data in any of these variables were removed prior to data analysis.

Analysis:

To determine the viability of this data in predicting icu mortality rate, a simple decision tree and a XG-boosted decision tree-based classifier were constructed using the prespecified features. For both models, to observe for overfitting, observations were first split into a training (70%) dataset to train the model and a test (30%) dataset to observe its performance. To find the optimal hyperparameters to fit the both models, variations in the hyperparameter values were investigated using a Grid Search with a 5-fold cross validation method. After fitting the models with optimal parameters, the models were run on the test data to observe prediction accuracy. Additionally, to determine the most significant features relevant for mortality status classification, the SHapley Additive exPlanations (SHAP) value will be analyzed to observe the overall impact of specific variables on the model output.

Findings:

The simple decision tree and the XG-boost models were initialized and hyper parameters tuning were conducted using grid Search. Optimal hyperparameter settings for the decision tree model were found for maximum depth (5), maximum features ('sqrt'), the minimum samples required at a leaf node (15), and splitter function ('best'). Optimal values for the XG-boost model were found for number of trees (n_estimators = 100), maximum tree depth (max_depth = 5), the minimum samples required at a leaf node (min_samples_leaf = 500), and learning rate (learning_rate = 0.1). After rerunning the models, with these parameters, the final decision tree model and XG-boost model performed at 80.39% and 91.67%

accuracy, respectively, on the test dataset (**Figure 1**). Upon observing the simple decision tree, the first few splits are based on hospital admission source, weight, bmi and ethnicity. Additionally in our XG-boost decision trees, splits based on BMI and age play are often the first split amongst the XG-boost decision trees. Additionally, from our SHAP values, it would appear that categorical variables such as ethnicity, especially when african american (Ethnicity_AA) or hispanic (Ethnicity_Hisp) and being admitted into the ICU from the Step-Down Unit (HAS_SDU) or the recovery room (HAS_RR) play an important part in predicting ICU mortality rate (**Figure 2**).

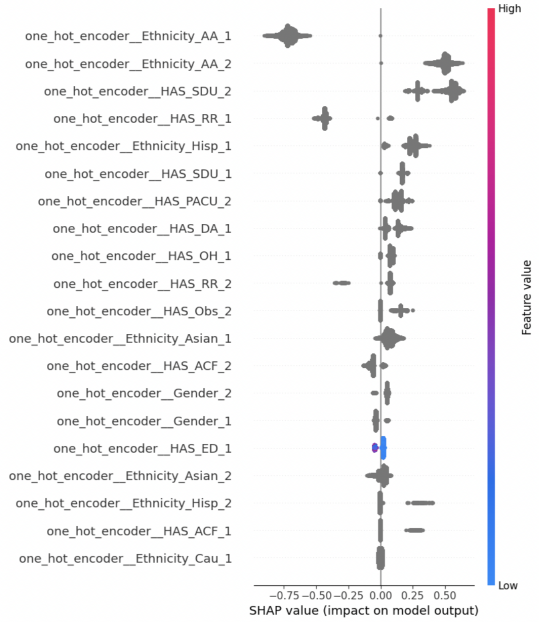
Figure 1. XG_Boost model performance

Regular decision tree	**XG-Boost**
Accuracy: 80.40%	Accuracy: 91.80%
Precision: 11.22%	Precision: 60%
Recall: 19.56%	Recall: 0.06%

Figure 1. Decision tree vs XG-boosted model performance on test datasets.

Figure 2. Graph of SHAP values for XG-boost data, the more extreme the SHAP value, the greater impact it has on determining hospital death across all the trees.

Figure 2. Graph of SHAP values



<u>Conclusions:</u>

Results from our model are surprising, our simple decision tree had predicted an accuracy percentage of 80.40% for hospital survivability and this accuracy can even rise to 91.80% with the XG-boost approach. However, our recall values for either model are exceptionally low and only a few true icu mortalities were actually predicted as deaths in either. Thus, our models provide little discriminatory value to mortality. However, our models have been informative in revealing important and non-specific trends associated with icu mortality. These include age, bmi, ethnicity and hospital admit source. Associations of between age and bmi, seem biological in nature, with higher mortality associated with higher age and bmi[1,2]. Conversely, ICU mortality rates associated with ethnicity are likely indicative of increased severity upon ICU admission and poorer access to healthcare. While differences in mortality based on hospital admission source is likely a reflection of the threshold of the severity of a patient's conditions to warrant having to be transferred into the ICU. Thus for ICU physicians, knowledge of these factors would provide the most generalized information about the patient's ICU mortality risks. However, these are not the only relevant factors. Knowledge of these factors should be supplemented with knowledge of other specific conditions such as aids and liver cirrhosis status when relevant to treating the patient. Our model findings only demonstrate generic trends that are consistently useful in determining ICU mortality. Thus future studies should seek to fit a regression-based approach to best measure the influence of rarer conditions on icu mortality or attempt deep-learning approaches to attempt to create a model with better discriminatory values.

<u>Team and individual contributions:</u> ***<u>Benjamin Zhang</u>***

Datathon 4, Date:13th Nov, 2023 **Group: unspecified (99) Team members:** Benjamin (Benny) Zhang

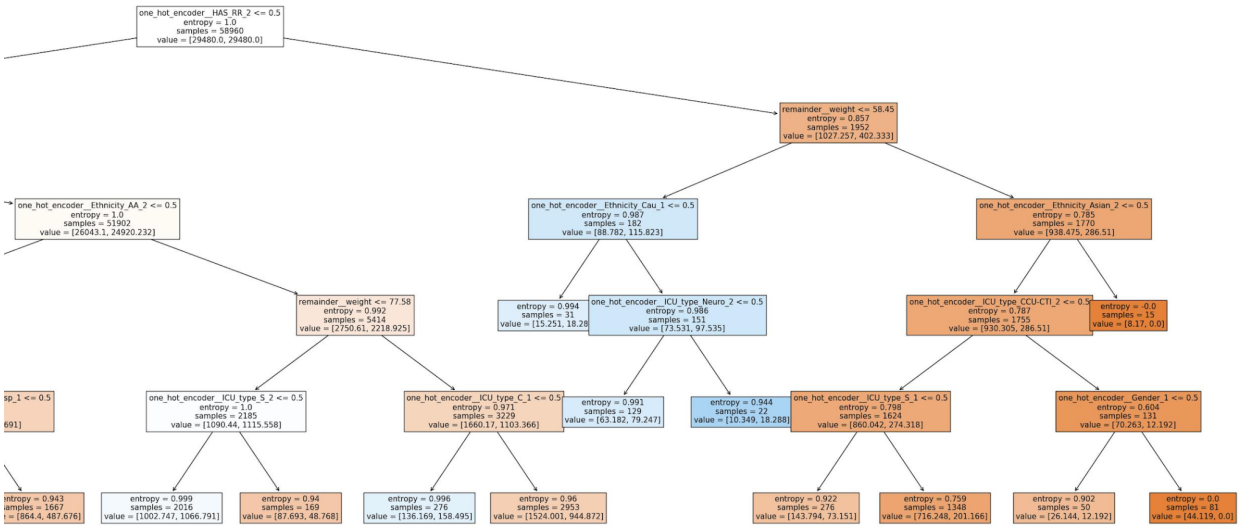
References:

1. Yu, W., Ash, A. S., Levinsky, N. G., & Moskowitz, M. A. (2000). Intensive care unit use and mortality in the elderly. *Journal of general internal medicine*, *15*(2), 97–102. https://doi.org/10.1046/j.1525-1497.2000.02349.x
2. Sanaie, S., Hosseini, M. S., Karrubi, F., Iranpour, A., & Mahmoodpoor, A. (2020). Impact of Body Mass Index on the Mortality of Critically Ill Patients Admitted to the Intensive Care Unit: An Observational Study. *Anesthesiology and pain medicine*, *11*(1), e108561. https://doi.org/10.5812/aapm.108561

Github repository Link:
https://github.com/BennyZhangUofT/Datathon4

Appendix:

A. Tree Structure for Simple Decision tree (Right Side)



B. Tree Structure for Simple Decision Tree (Left side)