Sachin Mohandas – BU ID: U79542832

MET CS555 A1 Term Project

1. Description

Find a dataset for a research problem of interest, here are some good websites for this

Kaggle Data Science Competitions: http://kaggle.com
UCI Machine Learning repository: https://archive.ics.uci.edu/datasets
A list of public data: https://www.teymourian.de/public-data-sets-for-data-analytic-projects/

Describe a research scenario and specify a research question based on data analytic methods that we learned in our class, for example methods like, *one and two sample means, t-test, correlation tests, simple and multiple linear regression, ANOVA and ANCOVA, one and two-Sample Tests for Proportions and logistic regression*.

Clean up your data and reduce it to no more than 2000 observations if your data set is large.

2. Research Scenario Description (no more than 200 words)

Describe your research scenario in no more than 200 words. This is a general description of the use case. Similar to our class examples, we first describe the overall scenario and then we specify a specific research question based on it.

When patients are admitted to the ICU of a hospital, they are subjected to a litany of tests from which substantial amounts of important medical data is obtained. The MIMIC-III database stores such clinical data pertaining to patients admitted to the Beth Israel Deaconess Medical Center here in Boston in an integrated, deidentified, and comprehensive format. This ICU dataset also contains an 'outcome' column indicating the state of mortality of the patient at the terminus of their admission. Given the large ensemble of features for each patient, it becomes shrewd to wonder if the outcome of a patient can be predicted with any accuracy by way of modeling the trends in their respective data. Such advancements could greatly expedite and streamline the triaging process that hurdles many hospital emergency rooms and ICU's. Additionally, the construction and implementation of such a predictive algorithm could significantly improve the odds of survival in patients that are deemed to be at greater risk of facing mortality.

3. Describe the data set (no more than 200 words)

Briefly describe the data set. Describe each variable of the data set that you plan to use in your analysis. Describe any data cleaning you have performed. If possible, provide a link to the main data set source.

We have opted to use the MIMIC-III dataset for ICU patients. This particular dataset contains 1177 observations and is quite tidy and thus requires minimal cleaning and preprocessing before it is ready for use. There are some columns with a number of NA's, and we dealt with these by simply imputing them as the mean of their respective column.

After cleaning, we first make use of all columns to obtain an overall predictive power for the entire ensemble of

features. Then, we select all non-categorical columns ('age', 'outcome', 'BMI', 'heart rate', 'Systolic blood pressure', 'Diastolic blood pressure', 'Respiratory rate', 'temperature', 'SP O2', 'Urine output', 'hematocrit', 'RBC', 'MCH', 'MCHC', 'MCV', 'RDW', 'Leucocyte', 'Platelets', 'Neutrophils', 'Basophils', 'Lymphocyte', 'PT', 'INR', 'NT-proBNP', 'Creatine kinase', 'Creatinine', 'Urea nitrogen', 'glucose', 'Blood potassium', 'Blood sodium', 'Blood calcium', 'Chloride', 'Anion gap', 'Magnesium ion', 'PH', 'Bicarbonate', 'Lactic acid', and 'PCO2') and continue our comparative analysis using all possible groups of three that could be made from them.

I obtained the dataset from Kaggle:

https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction?resource=download

4. Research Question (no more than 100 words)

Describe briefly in one or two sentences the main research question. This is similar to the last sentence of our class examples.

We endeavor to devise an accurate predictive tool that can be reliably implemented in a high traffic hospital ER/ICU environment. We first test the model's efficacy using the entire 49 feature ensemble and find its accuracy. Following this, we analyze the individual predictive power of much smaller ensembles, each containing a different combination of the initial 49 features and determine if any such combination is more accurate than the base model. This could provide valuable insight into patient cases for which only a truncated amount of feature data is available. Lastly, we investigate some of the attributes of those more accurate ensembles to potentially glean statistical reasons for which they may have performed better.

5. State your conclusions and discuss any limitations.

State the conclusion so that a non-statistician can understand. Discuss any potential limitations of your analysis. For example, are you suspicious that the assumptions of your test may not hold? Do you feel the analysis may have limitations for any other reasons?

Using a Logistic Regression machine learning algorithm with a 50/50 training/testing split, we first made 'outcome' predictions by considering the ensemble of features in the dataset containing all columns. This is undoubtedly the most comprehensive approach as it takes into account all pertinent information regarding the patient's health. We found that such a classifier has an accuracy of 85.7% when predicting patient mortality in the test set which indicates relatively high reliability. We then selected all 37 non-categorical columns from the dataset and considered individually the accuracies of all 7770 unique combinations of 3 therein. We found that not only are there a number of such ensembles having higher accuracy than the base model, but the base model's accuracy is actually among the lowest of these combinations (albeit by a relatively slim margin), being outperformed by 7049 of them. The accuracy of the best performing model in this context is 87.8% belonging to the ensemble containing 'Respiratory rate', 'Bicarbonate', and 'Lactic acid'. Upon investigating the model coefficients and odds ratios of the 50 most accurate models, we found that at least one of 'Respiratory rate', 'Lactic acid', or 'Anion gap' is present in almost all 50 models, with 'Lactic acid' and 'Anion gap' being among

the most important features (by coefficient magnitude) having weights in the ranges of 0.14-0.25 and 0.30-0.50, respectively. It should also be noted that, although it was not outstandingly present in this top 50 list, 'Blood potassium' also had a remarkably high feature importance, with coefficient magnitudes ranging from 0.58-1.2.

In addition to this, we conducted some exploratory data analysis in an attempt to determine what potential characteristics of the ensembles could result in higher accuracies. We first calculated the collinearities of each of the 7770 ensembles (by examining their confusion matrices) and found their summary statistics. We found that the mean collinearity among the ensembles was 0.00520 with a standard deviation of 0.151. We then compared these statistics to those belonging to the collinearities of the top 50 ensembles, which we found to have a mean collinearity of 0.0312 with a standard deviation of 0.0490. By inspection, we drew the interim assertion that there was no significant difference between these means, and a quick t-test confirmed this conclusion, having a p-value of 0.455; significantly higher than the standard 0.05 alpha level. We also inspected the pair plots of the top 10 ensembles to ascertain whether there were any trends in the distributions of the most influential predictors, however we were unable to discern any notable such trends aside from what would be manifestly apparent in a medical emergency environment (along the lines of low respiratory rate and high systolic blood pressure leading to higher chance of death).

Although our exploratory data analysis was relatively inconclusive, we believe this research endeavor was fruitful in its determination of a rudimentary predictive tool for determining an ICU patient's risk of death. Our initial model which incorporated all features was able to provide us with an accuracy which would be considered high in many contexts, and the accuracies of our top ensembles was better still.

However, in such a context as a hospital ER or ICU, one could easily argue, from a moral or medical perspective, that an accuracy of just 87.8% provides likely insufficient grounds on which to base crucial such triage decisions as order of admittance or urgency of care. The insufficiency of these grounds is further underlined by factors that lie outside the realm of this study, such as the often-prevalent medical volatility and complexity of patient cases, the very broad spectrum of circumstances that patients may enter the ER under, and the clinical interpretability of the model. The last of these hindrances is perhaps the most significant, as it forces medical professionals to stake faith in an effective black box algorithm without necessarily understanding the intricacies of how it arrived at its conclusion.

Despite these limitations, we believe such further refinement as assessing additional classifying algorithms, fine tuning the parameters therein, and considering additional features and observations can lead to the development of a highly impactful industry standard.