Data Mining Case Description Deliverable

Case Description:

The goal of the project is to define an analytics strategy for predicting patient survival from the first 24hrs in a hospital's intensive care unit.

The dataset is provided by the 2020 Global Women in Data Science (WiDS) Conference hosted on Kaggle. It is originally from MIT's GOSSIS community, with privacy certification from the Harvard Privacy Lab. The data contents are from more than 130,000 hospital Intensive Care Unit visits from patients all over the world, spanning a one-year timeframe. It has information on demographics, acute physiology and chronic health evaluation (APACHE), vital sign measures and lab tests for over 91,713 encounters. The data is presented in 186 columns, with one column indicating whether the patient died during the first 24 hours of hospitalization.

We will split the labeled dataset into training and testing data to develop our predictive model. K-fold cross-validation may be used to validate how the result will generalize to unseen data. We are given an unlabeled dataset that will also be used to test our model. Since such data will not be used to train our model but test only, we are more confident that our model performance will be a good predictor. The goal of our model is to produce a high probability of ranking a random positive example more highly than a random negative example.

In addition to our evaluation against commonly used metrics like AUC, we would like to explore gender bias in our models. The motivation for this is that clinical trials are often unbalanced with a higher proportion of males. This leads to models that work very well on males, but fail to generalize when more females are added to the data. We plan to evaluate our first model's probabilities on demographics to see if there is any bias. If we do note a bias, we would like to create two separate models: one for males and one for females. These models would then be evaluated to see if there is strong predictive power when bias is taken into account.

Case Motivation:

We hope that our analysis will inform the hospital staff so that they can take initiative within the first 24 hours of patient admission. Intensive Care Units are often understaffed, patients in this unit often need the most attention, and most staff does not remain constant over the course of 24 hours. With all these changes, it is extremely difficult for staff to keep track of all the information they are receiving. Therefore, a predictive model like ours could inform staff of the

most important features that need to be monitored when a patient is admitted to the Intensive Care Unit.

Team Members:

Our team includes Meghna Diwan, Viviana Hernandez, Vamika Venkatesan, Jessica Wang, and Jenny Zhen. We all hope to equally contribute to the success of the project. To achieve this, we have started cross-collaborating through Github and Google Drive. Going forward, we hope to leverage both business and technical acumen of all the members.