COVID 19 Machine Learning

The COVID-19 pandemic changed many assumptions about health care and allocation of resources in the medical community. The widespread nature of the infections caused individuals and organizations to modify their historic approaches to work and life in general. One the most important issues that arose was the allocation of limited health care resources at hospitals to patients who were the most at risk for severe illness and/or death. Among the population of individuals admitted to hospitals, these at-risk individuals can be identified as those who were subsequently moved to the intensive care unit (“ICU”) due to the degradation of their condition. Having a reasonable way to predict which patients, based upon various demographic and medical measures, are most likely to eventually need ICU level care can not only allow hospitals to more quickly identify those individuals and allocate resources to their care but also identify the individuals who are less likely to degrade and thus be more appropriate for discharge to home care in order to more efficiently use the limited staff and bed resources available. To these ends, the development and implementation of a machine learning model is desirable, achievable and appropriate.

In developing the model, use of the anonymized data from the Sírio-Libanês hospital in São Paulo Brazil was used. The data as obtained was previously cleaned and scaled by column (attribute/feature) using the Min Max Scaler to fit between -1 and 1. The data set consists of 1,925 records (rows) each of which have 231 attributes. These attributes consist of 225 float, 4 integer and 2 categorical data types. To develop the machine learning model, the 2 attributes of this type (Age Percentile, Window) were more closely examined to determine that neither of the columns had any missing values and they were then subsequently encoded to discrete values. Examination of the 4 integer type attributes indicated no missing data and evaluation of the 225 float attributes indicated widespread missing data from every column. Since it is reasonable to assume that a patient who does not have a measurement recorded in a time window is clinically stable, potentially presenting vital signs and blood labs similar to neighboring windows, missing values were filled using the next or previous entry via first forward filling and then back filling of the data. Since it has been advised not to use the records where the ICU attribute is present, a new attribute (“ICU\_EVER”) was created that indicated by patient whether the ICU attribute was present in any record window of that patient. This allowed the dropping of all records where the ICU attribute was present while retaining all other records for each patient and window.

The cleaned and prepared data is now used to train both a XGBoost and CatBoost classifier using a 20% test sample size and a random state of 42 and defining the target value as the ICU\_EVER attribute. The results revealed that the XGBoost classifier provided the best overall results with an accuracy score of 89.36% vs. 88.30% for the CatBoost Classifier.

Based on these results, we would recommend the implementation of the XGBoost Classifier machine learning model. The model results indicate that this model will accurately predict patients who will not end up in the ICU with an accuracy of 91% and patients who will end up in the ICU with an accuracy of 87% with an overall accuracy score of 89.36%.