Final Project

**Data Science II - Statistical Learning**

**Instructions**:

* The report should be a Word, pdf or html document and should be uploaded to canvas.
* You are required to submit your R code - either by kniting an R markdown into one of the above format or in an Appendix.
* Do not present unnecessary R output.  Length of the report should not exceed 6 pages. Detailed explanation of your steps of analysis is required and points will be deducted for inadequate reasoning.

**Total points: 100**

**Background**:  The data set is a simulated data set. It represents the data on COVID+ patients (i.e. patients diagnosed with COVID) who were hospitalized in the New York Presbyterian hospital (Weill Cornell and Lower Manhattan campus). Some COVID+ patients suffer from acute respiratory distress and have acute difficulties in breathing. Such critically ill patients then require artificial respiratory support through an invasive mechanical ventilator, a process known as intubation. Since there was a surge in COVID+ cases in New York City during the period of March, the hospital authorities were concerned that there might be a shortage in availability of mechanical ventilators. The authorities wanted to predict the need for intubation within 5 days of hospitalization for each patient in order to better manage their resources. Your goal is to build a predictive model that will predict the outcome of intubation within 5 days of hospitalization.

**Data**: There are two data files "baseline.csv" and "labs and vitals.csv".

* The baseline.csv file contains patient level information (one row per patient) on baseline clinical measures that were determined at the time of admission to the hospital. The measures include demographic variables, relevant clinical history and certain diagnostic tests.
* The labs and vitals file contain data on vitals signs of patients (e.g. blood pressure, heart rate, SpO2 or blood oxygen levels, etc). The vital signs are measured multiple times for each patient during the course of hospitalization. There number of times it is measured per individual also varies with each individual. The vitals data is in a long format i.e. there are multiple rows for a patient.
* A variable dictionary is also available to explain the variables.

**Goal**:

1. **Feature engineering**: Extract features from the longitudinal vital signs measure so that you have one value for patient.

*Hints*: You can explore various aspects of the distribution of each vital sign e.g. mean, standard deviation, skewness, kurtosis, trends over time etc to create featuers).

*Note*: You can and should create multiple features per vital sign but you need to make sure there is one value per patient for each feature.

2. **Predictive modeling**: Using all baseline variables and features extracted from labs and vitals, your goal is to build a predictive model predicting the binary event of intubation within 5 days of hospitalization. You can use various learning methods for prediction and you need to compare the prediction error of the models. You need to provide detailed explanations of the steps you have taken to tune your models, estimate and compare prediction error. Finally, you have to choose a final model that you will present to the authorities and explain the prediction model the best you can to the authorities so that they understand factors are important for prediction.

[Previous](https://medcornell.instructure.com/courses/1732894/modules/items/24703007)[Next](https://medcornell.instructure.com/courses/1732894/modules/items/24715751)