Data Science Case Study

Background:

We have data regarding patients with COVID-19 and a hospital has asked us to predict the need for mechanical ventilation among patients. The goal is to reduce patient mortality by identifying high risk patients based on the available data and devoting the limited ventilators to them.

Data:

baselines.csv – Contains patient level data regarding patient characteristics.

Ed_before_order_set column specifies if the patient was admitted before the new series of tests were introduced for a better understanding of COVID-19. For any other column which has 0/1 binary values, **1** corresponds to **yes** and **0** corresponds to **no. Checked** corresponds to **yes** and **Unchecked** corresponds to **no.**

labs and vitals.csv – Contains time series biometric information for patients.

Value column indicates the time series value associated with the vital specified in the **name** column. The **subject** column indicates the patient ID (MRN) and **time_stamp** column gives information about the time when the value was captured

Exercise:

1. Perform data cleaning and feature engineering
2. Explore the data and perform appropriation data visualization to extract insights
3. Use statistical methods to make inferences
4. Select ML classifiers to predict the "event" (need for mechanical ventilation among patients). Note: Check assumptions and provide reasoning for this model selection
5. Tune hyperparameter to improve any chosen metric (Eg: misclassification error/ accuracy/ precision/ recall) and discuss the results.

Please include explanations wherever necessary. Please use python to perform analysis and predictions. Please finally share the code along with your understandings in a powerpoint by mail

Please try to complete this on/before 7 days from the date of email