**ICU-Gossis Research Paper**

**Data Collection -** We have got our Gossis eICU dataset from Physionet, MIT Laboratory for Computational Physiology Institute for Medical Engineering and Science, MIT, E25-50577 Massachusetts Ave, Cambridge, MA 02139. They gave us GOSSIS eICU the eICU-CRD subset of the Global Open Source Severity of Illness Score (GOSSIS-1) dataset

GOSSIS-1 is a modern, free, open-source in-hospital mortality prediction algorithm for critical care patients, achieving excellent discrimination and calibration across three countries (Australia, New Zealand and the USA). GOSSIS-1 was developed on two large datasets of critical care patients. This project contains the USA subset of patients derived from the eICU Collaborative Research Database (eICU-CRD). The dataset, which we call GOSSIS-1-eICU, consists of 131,051 unique patients from 204 hospitals from ICU admissions discharged in 2014-15. The code to create the dataset from eICU-CRD and generate GOSSIS-1 predictions are also available.

## Background

GOSSIS-1 [1] was developed as the first version of a series of global open-source severity of illness scores by the GOSSIS consortium [2]. The consortium aims create a database of critical care datasets from ICUs around the globe and to use these datasets to develop a family of open-source scoring systems for assessing the severity of illness of critical care patients internationally. GOSSIS-1 was developed using data from two well-known datasets consisting of data from Australia and New Zealand via the ANZICS-APD dataset [3], and the USA via the eICU-CRD dataset [4]. The GOSSIS-1 model achieved high discrimination and calibration in all countries and relevant subsets [1]. This project contains the USA subset of data that was used to train the GOSSIS-1 model. The data originates from eICU-CRD and we call this dataset, GOSSIS-1-eICU.

## Methods

The GOSSIS-1-eICU data were extracted from the eICU-CRD database [4]. The eICU-CRD is a relational database consisting of about 200,000 ICU admissions from over 200 hospitals throughout the USA. Importantly, the GOSSIS-1-eICU data consists of critical care admissions from 2014-15, where the length of ICU stay was >6 hours. Data, including physiologic and vital signs were collected from the first 24 hours of the ICU stay. Readmissions to the ICU, patients <16 years old, and those with a missing outcome or with no heart rate recorded were excluded.

This dataset contains total 131,051 unique ICU patients data out of which total **123734** patients are got normal or discharged and **7317 patients got Mortality.**

This dataset has total 216 columns (Features)

['patientunitstayid', 'data\_source', 'encounter\_id', 'hospital\_id', 'patient\_id', 'age', 'bmi', 'country', 'elective\_surgery', 'ethnicity', 'gender', 'height', 'hospital\_admit\_source', 'hospital\_bed\_size', 'hospital\_bed\_size\_numeric', 'hospital\_death', 'hospital\_disch\_location', 'hospital\_los\_days', 'hospital\_type', 'icu\_admit\_source', 'icu\_admit\_type', 'icu\_death', 'icu\_disch\_location', 'icu\_id', 'icu\_los\_days', 'icu\_stay\_type', 'icu\_type', 'pre\_icu\_los\_days', 'pregnant', 'readmission\_status', 'smoking\_status', 'teaching\_hospital', 'weight', 'albumin\_apache', 'apache\_2\_diagnosis', 'apache\_3j\_diagnosis', 'apache\_post\_operative', 'arf\_apache', 'bilirubin\_apache', 'bun\_apache', 'creatinine\_apache', 'fio2\_apache', 'gcs\_eyes\_apache', 'gcs\_motor\_apache', 'gcs\_unable\_apache', 'gcs\_verbal\_apache', 'glucose\_apache', 'heart\_rate\_apache', 'hematocrit\_apache', 'intubated\_apache', 'map\_apache', 'paco2\_apache', 'paco2\_for\_ph\_apache', 'pao2\_apache', 'ph\_apache', 'resprate\_apache', 'sodium\_apache', 'temp\_apache', 'urineoutput\_apache', 'ventilated\_apache', 'wbc\_apache', 'd1\_diasbp\_invasive\_max', 'd1\_diasbp\_invasive\_min', 'd1\_diasbp\_max', 'd1\_diasbp\_min', 'd1\_diasbp\_noninvasive\_max', 'd1\_diasbp\_noninvasive\_min', 'd1\_heartrate\_max', 'd1\_heartrate\_min', 'd1\_mbp\_invasive\_max', 'd1\_mbp\_invasive\_min', 'd1\_mbp\_max', 'd1\_mbp\_min', 'd1\_mbp\_noninvasive\_max', 'd1\_mbp\_noninvasive\_min', 'd1\_padias\_invasive\_max', 'd1\_padias\_invasive\_min', 'd1\_pamean\_invasive\_max', 'd1\_pamean\_invasive\_min', 'd1\_pasys\_invasive\_max', 'd1\_pasys\_invasive\_min', 'd1\_resprate\_max', 'd1\_resprate\_min', 'd1\_spo2\_max', 'd1\_spo2\_min', 'd1\_sysbp\_invasive\_max', 'd1\_sysbp\_invasive\_min', 'd1\_sysbp\_max', 'd1\_sysbp\_min', 'd1\_sysbp\_noninvasive\_max', 'd1\_sysbp\_noninvasive\_min', 'd1\_temp\_max', 'd1\_temp\_min', 'h1\_diasbp\_invasive\_max', 'h1\_diasbp\_invasive\_min', 'h1\_diasbp\_max', 'h1\_diasbp\_min', 'h1\_diasbp\_noninvasive\_max', 'h1\_diasbp\_noninvasive\_min', 'h1\_heartrate\_max', 'h1\_heartrate\_min', 'h1\_mbp\_invasive\_max', 'h1\_mbp\_invasive\_min', 'h1\_mbp\_max', 'h1\_mbp\_min', 'h1\_mbp\_noninvasive\_max', 'h1\_mbp\_noninvasive\_min', 'h1\_padias\_invasive\_max', 'h1\_padias\_invasive\_min', 'h1\_pamean\_invasive\_max', 'h1\_pamean\_invasive\_min', 'h1\_pasys\_invasive\_max', 'h1\_pasys\_invasive\_min', 'h1\_resprate\_max', 'h1\_resprate\_min', 'h1\_spo2\_max', 'h1\_spo2\_min', 'h1\_sysbp\_invasive\_max', 'h1\_sysbp\_invasive\_min', 'h1\_sysbp\_max', 'h1\_sysbp\_min', 'h1\_sysbp\_noninvasive\_max', 'h1\_sysbp\_noninvasive\_min', 'h1\_temp\_max', 'h1\_temp\_min', 'd1\_albumin\_max', 'd1\_albumin\_min', 'd1\_bilirubin\_max', 'd1\_bilirubin\_min', 'd1\_bun\_max', 'd1\_bun\_min', 'd1\_calcium\_max', 'd1\_calcium\_min', 'd1\_creatinine\_max', 'd1\_creatinine\_min', 'd1\_glucose\_max', 'd1\_glucose\_min', 'd1\_hco3\_max', 'd1\_hco3\_min', 'd1\_hemaglobin\_max', 'd1\_hemaglobin\_min', 'd1\_hematocrit\_max', 'd1\_hematocrit\_min', 'd1\_inr\_max', 'd1\_inr\_min', 'd1\_lactate\_max', 'd1\_lactate\_min', 'd1\_platelets\_max', 'd1\_platelets\_min', 'd1\_potassium\_max', 'd1\_potassium\_min', 'd1\_sodium\_max', 'd1\_sodium\_min', 'd1\_wbc\_max', 'd1\_wbc\_min', 'h1\_albumin\_max', 'h1\_albumin\_min', 'h1\_bilirubin\_max', 'h1\_bilirubin\_min', 'h1\_bun\_max', 'h1\_bun\_min', 'h1\_calcium\_max', 'h1\_calcium\_min', 'h1\_creatinine\_max', 'h1\_creatinine\_min', 'h1\_glucose\_max', 'h1\_glucose\_min', 'h1\_hco3\_max', 'h1\_hco3\_min', 'h1\_hemaglobin\_max', 'h1\_hemaglobin\_min', 'h1\_hematocrit\_max', 'h1\_hematocrit\_min', 'h1\_inr\_max', 'h1\_inr\_min', 'h1\_lactate\_max', 'h1\_lactate\_min', 'h1\_platelets\_max', 'h1\_platelets\_min', 'h1\_potassium\_max', 'h1\_potassium\_min', 'h1\_sodium\_max', 'h1\_sodium\_min', 'h1\_wbc\_max', 'h1\_wbc\_min', 'd1\_arterial\_pco2\_max', 'd1\_arterial\_pco2\_min', 'd1\_arterial\_ph\_max', 'd1\_arterial\_ph\_min', 'd1\_arterial\_po2\_max', 'd1\_arterial\_po2\_min', 'd1\_pao2fio2ratio\_max', 'd1\_pao2fio2ratio\_min', 'h1\_arterial\_pco2\_max', 'h1\_arterial\_pco2\_min', 'h1\_arterial\_ph\_max', 'h1\_arterial\_ph\_min', 'h1\_arterial\_po2\_max', 'h1\_arterial\_po2\_min', 'h1\_pao2fio2ratio\_max', 'h1\_pao2fio2ratio\_min', 'apache\_3j\_hospital\_death\_prob', 'apache\_3j\_score', 'apache\_4a\_hospital\_death\_prob', 'apache\_4a\_icu\_death\_prob', 'apsiii', 'aids', 'cirrhosis', 'diabetes\_mellitus', 'hepatic\_failure', 'immunosuppression', 'leukemia', 'lymphoma', 'solid\_tumor\_with\_metastasis', 'apache\_3j\_bodysystem', 'apache\_2\_bodysystem']

**1- First Experiment Model (Full ICU Model) -**

**Data Cleaning and Preprocessing -** In this model we have trained our model on full ICU dataset consist of 131,051 unique ICU patients data out of which total **123734** patients are got normal or discharged and **7317 patients got Mortality.** This data contains

**Min ICU Length Of Stay days - 4 Hours**

**Max ICU Length Of Stay days - 506 Days**

So after getting data we drop some features which have null data and some feature which are causing overfitting our model like - Hospital death, 'apache\_3j\_hospital\_death\_prob', 'apache\_4a\_hospital\_death\_prob'

So we basically drop total 37 feature for training over model and these feature are:

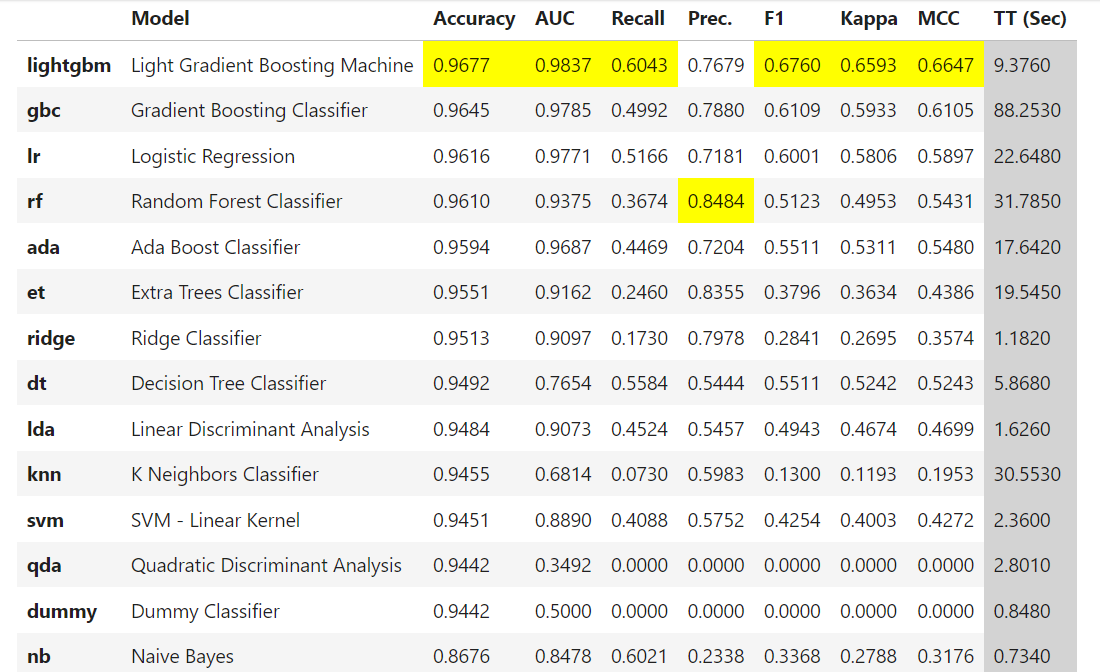
**variables\_to\_drop**= ["patientunitstayid", "data\_source",'smoking\_status', "encounter\_id", "hospital\_id", "patient\_id", "country", "hospital\_admit\_source", "hospital\_bed\_size", "hospital\_bed\_size\_numeric", "hospital\_disch\_location", "hospital\_type", "icu\_admit\_type", "icu\_disch\_location", "icu\_id","pregnant", "d1\_padias\_invasive\_max", "d1\_padias\_invasive\_min", "d1\_pamean\_invasive\_max", "d1\_pamean\_invasive\_min", "d1\_pasys\_invasive\_max", "d1\_pasys\_invasive\_min", "h1\_padias\_invasive\_max", "h1\_padias\_invasive\_min", "h1\_pamean\_invasive\_max", "h1\_pamean\_invasive\_min", "h1\_pasys\_invasive\_max", "h1\_pasys\_invasive\_min", "apache\_3j\_hospital\_death\_prob","hospital\_death",'icu\_stay\_type', 'icu\_type', 'apache\_3j\_bodysystem','apache\_2\_bodysystem','apache\_3j\_hospital\_death\_prob', 'apache\_4a\_hospital\_death\_prob', 'apache\_4a\_icu\_death\_prob','ethnicity']

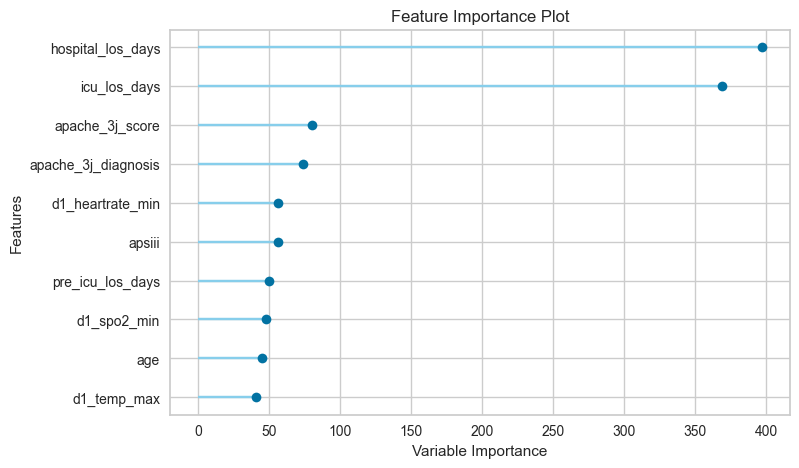
So We trained our model on total 179 features.

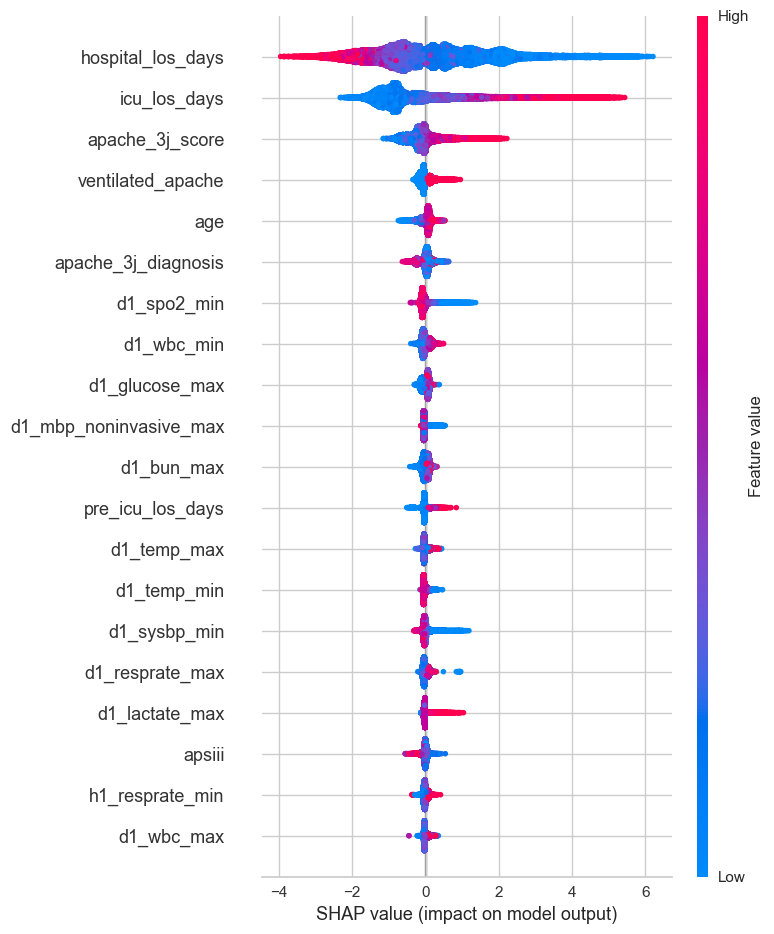
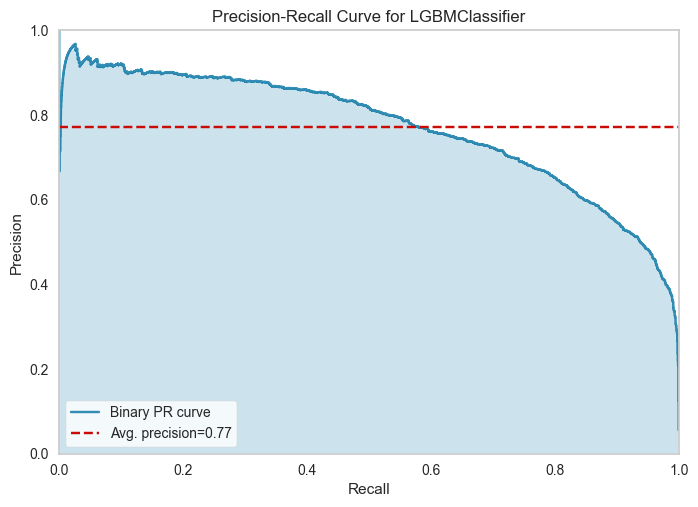
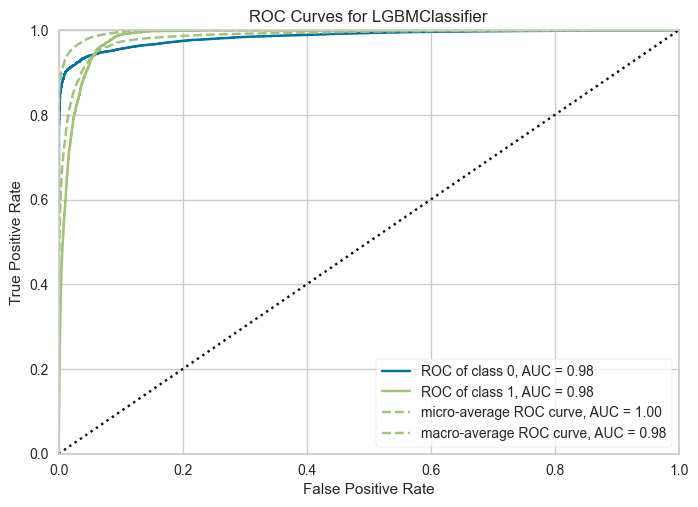
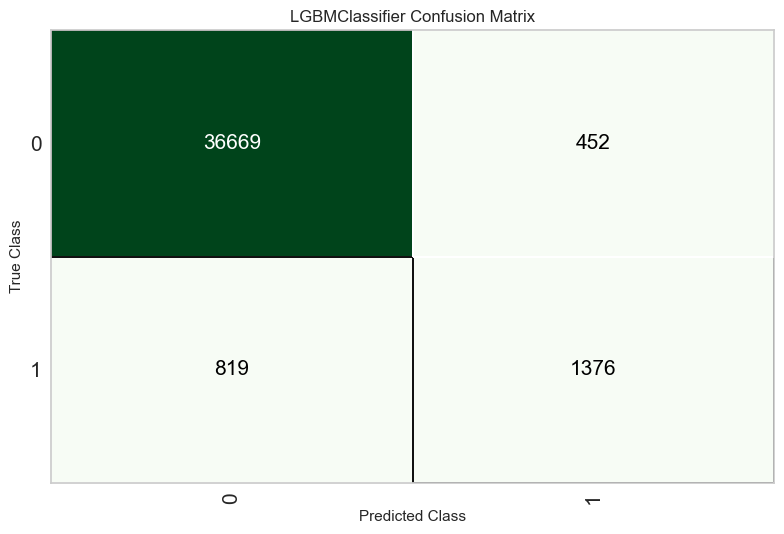
In these features we have some numerical feature and some are categorical features so we have applied onhot encoding in the categorical features. Then we filled null values of the numerical features with mean values of the features and for categorical features we filled with mode of the features. We have also used Label Binarizer technique to convert binary features(True,False/ Yes,No) into numericals(0 and 1).

**Model Building -** We have tried multiple models and our best-performing model is Light Gradient boosting model. We have done train-test split of our dataset in which we split our data into 80% for Training and 20% for Testing purpose.

We have also used grid search cv for finding best parameters and we find these parameters ('num\_leaves': 50, 'learning\_rate': 0.1, 'n\_estimators': 200) are performing best.

****

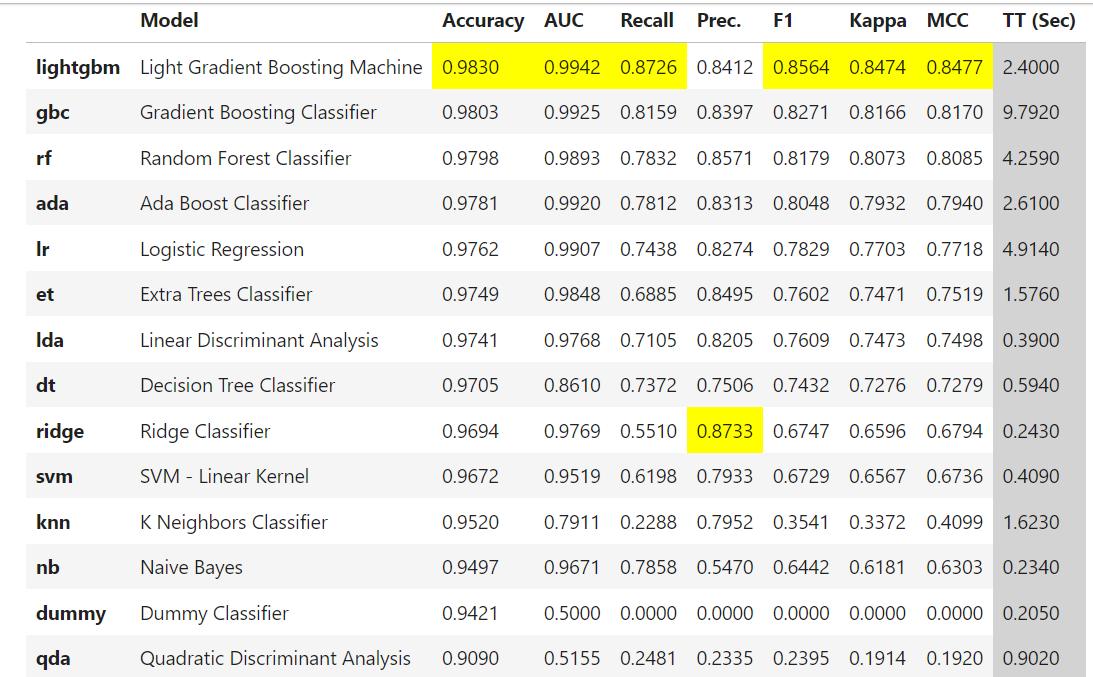
****

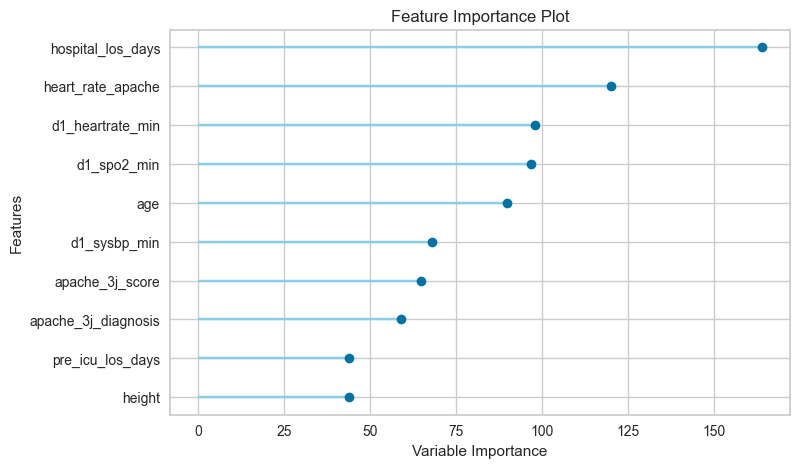
****

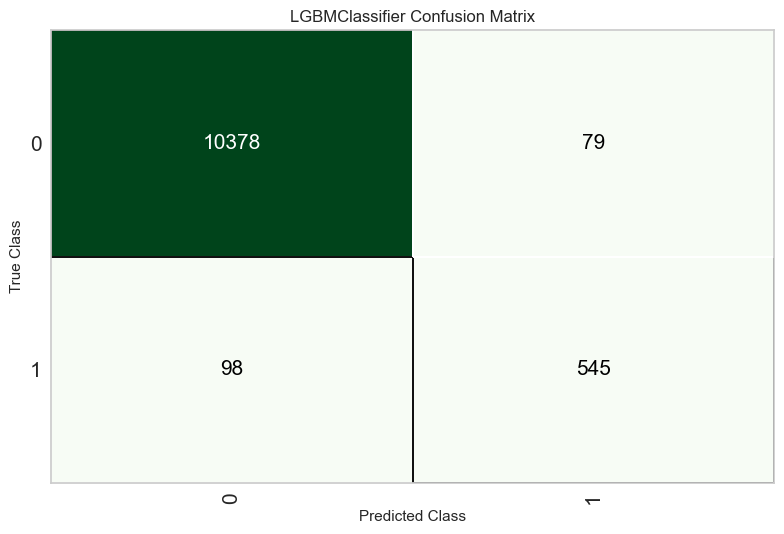
**2- Second Experiment (First 24 hour Full ICU model) -**

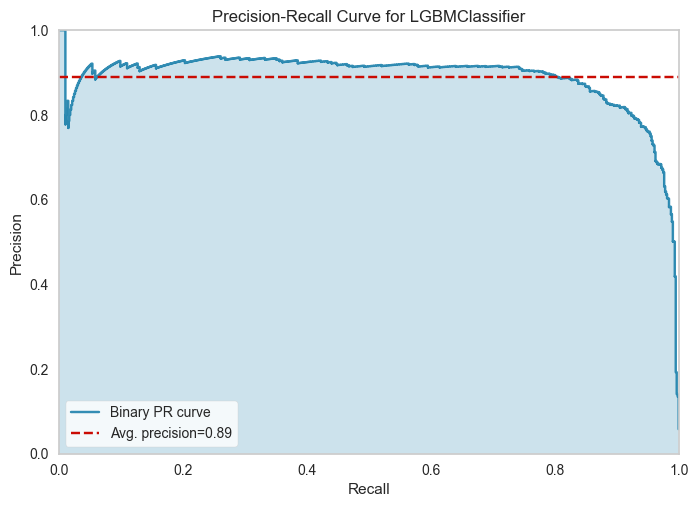
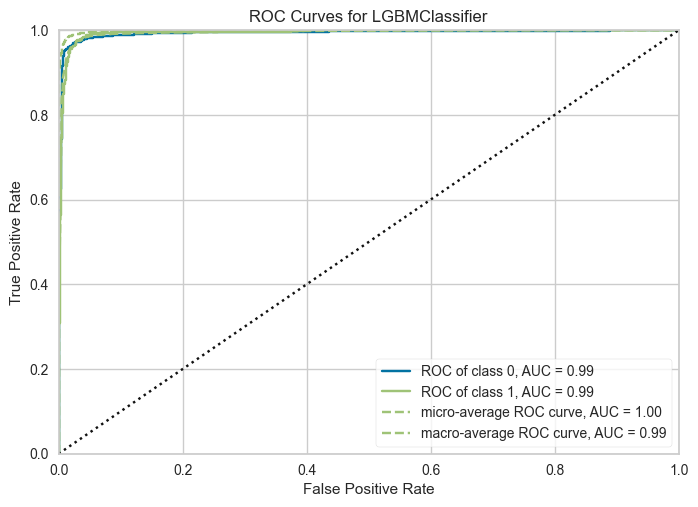
In this model we have taken first 24 hours of ICU patient data and predicting the outcomes of ICU mortality.

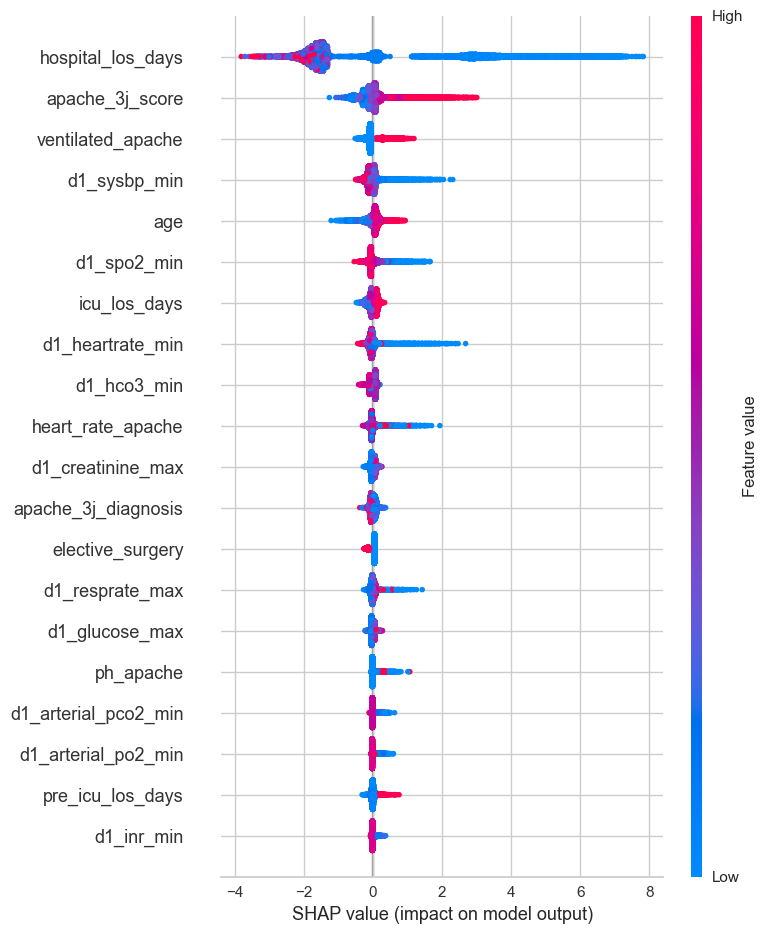
We have trained our model total on 37000 ICU patients and 179 features and classified into two category 0(Normal) and 1 (ICU Death). The counts of each categories are for normal = 34858 and ICU death = 2142

****

****

****

****

****

**3- 3rd Experiment (First 24 hours and top 25 features)-**

In this model we have dropped 191 features and only trained on top 25 features of ICU data to check performance and how our model can be affected. We have trained our model total on 37000 ICU patients and only top 25 features and classified into two category 0(Normal) and 1 (ICU Death). The counts of each categories are for normal = 34858 and ICU death = 2142

We have dropped all these features

variables\_to\_drop= ["patientunitstayid", "data\_source",'smoking\_status', "encounter\_id", "hospital\_id", "patient\_id", "country", "hospital\_admit\_source", "hospital\_bed\_size",

"hospital\_bed\_size\_numeric", "hospital\_disch\_location", "hospital\_type",

"icu\_admit\_type", "icu\_disch\_location", "icu\_id","pregnant", "d1\_padias\_invasive\_max",

"d1\_padias\_invasive\_min", "d1\_pamean\_invasive\_max", "d1\_pamean\_invasive\_min",

"d1\_pasys\_invasive\_max", "d1\_pasys\_invasive\_min", "h1\_padias\_invasive\_max",

"h1\_padias\_invasive\_min", "h1\_pamean\_invasive\_max", "h1\_pamean\_invasive\_min",

"h1\_pasys\_invasive\_max", "h1\_pasys\_invasive\_min", "apache\_3j\_hospital\_death\_prob",

"hospital\_death",'icu\_stay\_type', 'icu\_type', 'apache\_3j\_bodysystem','apache\_2\_bodysystem','apache\_3j\_hospital\_death\_prob', 'apache\_4a\_hospital\_death\_prob',

'apache\_4a\_icu\_death\_prob', 'ethnicity', 'icu\_admit\_source', 'age',

'Gcs\_eyes\_apache', 'apache\_4a\_hospital\_death\_prob',

'gcs\_motor\_apache',

'apache\_4a\_icu\_death\_prob',

'gcs\_verbal\_apache',

'aids',

'bilirubin\_apache',

'bun\_apache',

'cirrhosis',

'creatinine\_apache',

'diabetes\_mellitus',

'elective\_surgery',

'gcs\_unable\_apache',

'glucose\_apache',

'heart\_rate\_apache',

'hematocrit\_apache',

'height',

'hepatic\_failure',

'h1\_albumin\_max',

'h1\_albumin\_min',

'h1\_arterial\_pco2\_max',

'h1\_arterial\_pco2\_min',

'h1\_arterial\_ph\_max',

'h1\_arterial\_ph\_min',

'h1\_arterial\_po2\_max',

'h1\_arterial\_po2\_min',

'h1\_bilirubin\_max',

'h1\_bilirubin\_min',

'h1\_bun\_max',

'h1\_bun\_min',

'h1\_calcium\_max',

'h1\_calcium\_min',

'h1\_creatinine\_max',

'h1\_creatinine\_min',

'h1\_diasbp\_invasive\_max',

'h1\_diasbp\_invasive\_min',

'h1\_diasbp\_max',

'h1\_diasbp\_min',

'h1\_diasbp\_noninvasive\_max',

'h1\_diasbp\_noninvasive\_min',

'h1\_glucose\_max',

'h1\_glucose\_min',

'h1\_hco3\_max',

'h1\_hco3\_min',

'h1\_hematocrit\_max',

'h1\_hematocrit\_min',

'h1\_hemaglobin\_max',

'h1\_hemaglobin\_min',

'h1\_inr\_max',

'h1\_inr\_min',

'h1\_lactate\_max',

'h1\_lactate\_min',

'h1\_mbp\_invasive\_max',

'h1\_mbp\_invasive\_min',

'h1\_mbp\_max',

'h1\_mbp\_min',

'h1\_mbp\_noninvasive\_max',

'h1\_mbp\_noninvasive\_min',

'h1\_pao2fio2ratio\_max',

'h1\_pao2fio2ratio\_min',

'h1\_platelets\_max',

'h1\_platelets\_min',

'h1\_potassium\_max',

'h1\_potassium\_min',

'h1\_resprate\_max',

'h1\_resprate\_min',

'h1\_sodium\_max',

'h1\_sodium\_min',

'h1\_spo2\_max',

'h1\_spo2\_min',

'h1\_sysbp\_invasive\_max',

'h1\_sysbp\_invasive\_min',

'h1\_sysbp\_max',

'h1\_sysbp\_min',

'h1\_sysbp\_noninvasive\_max',

'h1\_sysbp\_noninvasive\_min',

'h1\_temp\_max',

'h1\_temp\_min',

'h1\_wbc\_max',

'h1\_wbc\_min',

'heart\_rate\_apache',

'height',

'immunosuppression',

'intubated\_apache',

'leukemia',

'lymphoma',

'paco2\_apache',

'paco2\_for\_ph\_apache',

'pao2\_apache',

'pre\_icu\_los\_days',

'readmission\_status',

'resprate\_apache',

'sodium\_apache',

'solid\_tumor\_with\_metastasis',

'teaching\_hospital',

'urineoutput\_apache',

'weight',

'd1\_albumin\_max',

'd1\_albumin\_min',

'd1\_arterial\_pco2\_max',

'd1\_arterial\_pco2\_min',

'd1\_arterial\_po2\_max',

'd1\_arterial\_po2\_min',

'd1\_bilirubin\_max',

'd1\_bilirubin\_min',

'd1\_bun\_max',

'd1\_bun\_min',

'd1\_calcium\_max',

'd1\_calcium\_min',

'd1\_creatinine\_max',

'd1\_creatinine\_min',

'd1\_diasbp\_invasive\_max',

'd1\_diasbp\_invasive\_min',

'd1\_diasbp\_max',

'd1\_diasbp\_min',

'd1\_diasbp\_noninvasive\_max',

'd1\_diasbp\_noninvasive\_min',

'd1\_glucose\_max',

'd1\_glucose\_min',

'd1\_hco3\_max',

'd1\_hco3\_min',

'd1\_hematocrit\_max',

'd1\_hematocrit\_min',

'd1\_hemaglobin\_max',

'd1\_hemaglobin\_min',

'd1\_inr\_max',

'd1\_inr\_min',

'd1\_lactate\_max',

'd1\_lactate\_min',

'd1\_mbp\_invasive\_max',

'd1\_mbp\_invasive\_min',

'd1\_mbp\_max',

'd1\_mbp\_min',

'd1\_mbp\_noninvasive\_max',

'd1\_mbp\_noninvasive\_min',

'd1\_pao2fio2ratio\_max',

'd1\_pao2fio2ratio\_min',

'd1\_platelets\_max',

'd1\_platelets\_min',

'd1\_potassium\_max',

'd1\_potassium\_min','d1\_resprate\_max','d1\_resprate\_min',

'd1\_sodium\_max',

'd1\_sodium\_min',

'd1\_spo2\_max',

'd1\_spo2\_min',

'd1\_sysbp\_invasive\_max',

'd1\_sysbp\_max',

'd1\_sysbp\_min',

'd1\_sysbp\_noninvasive\_max',

'd1\_sysbp\_noninvasive\_min',

'd1\_temp\_max',

'd1\_temp\_min',

'd1\_wbc\_max', 'd1\_wbc\_min',

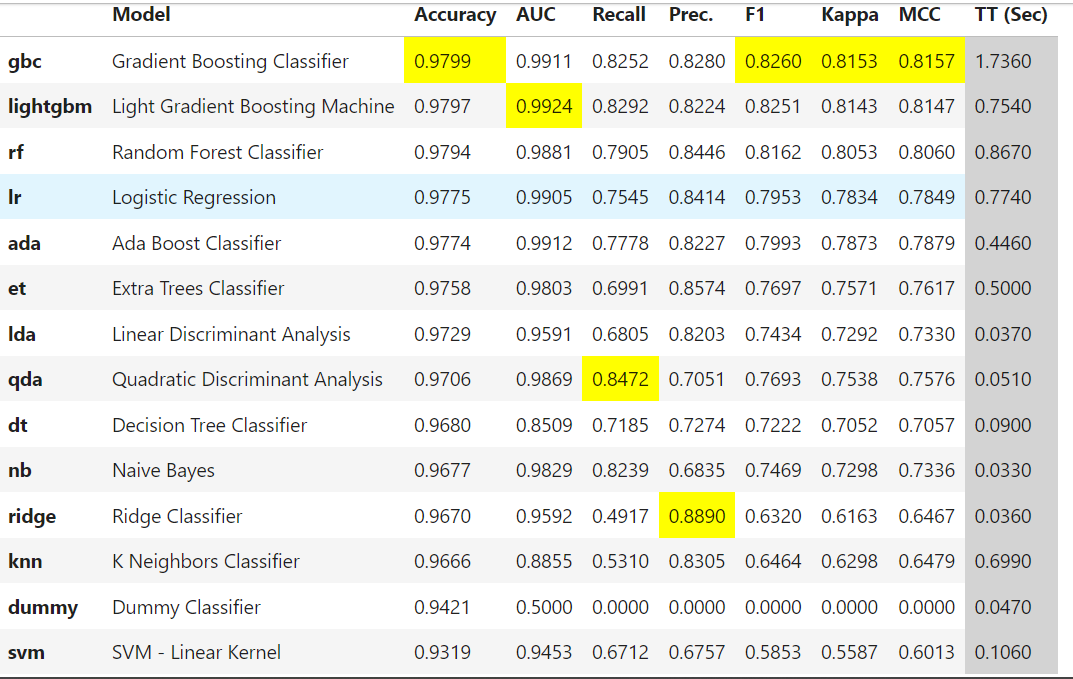
'gcs\_eyes\_apache', 'gcs\_motor\_apache',

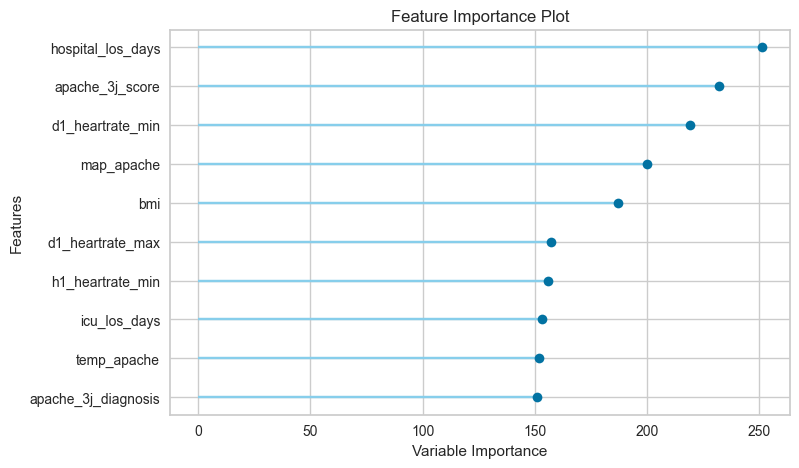
'gcs\_verbal\_apache',

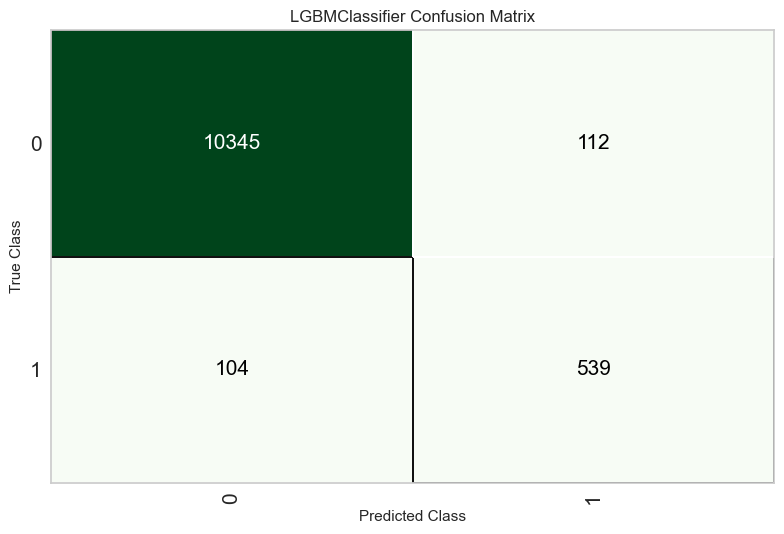
]

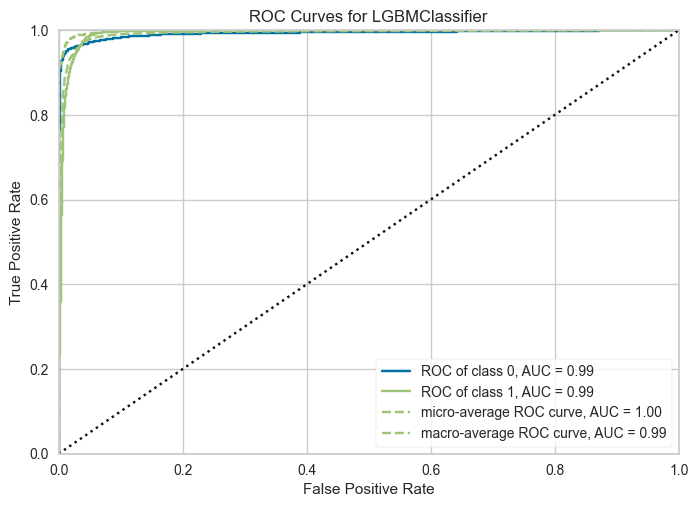
We have taken only these 25 features for training our model =

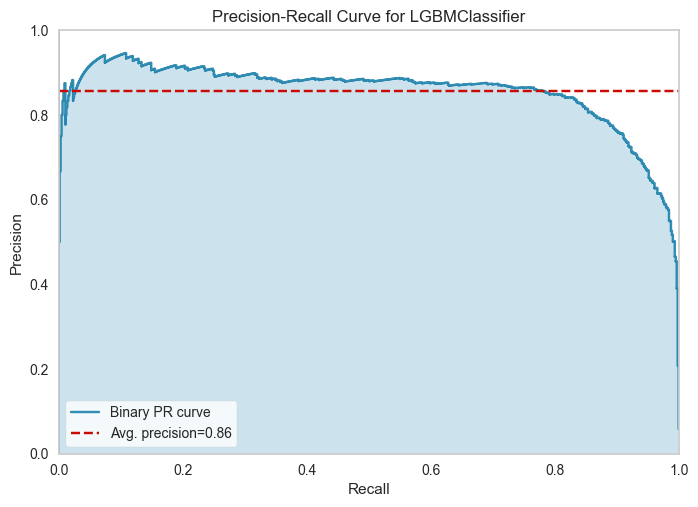
['bmi', 'gender', 'hospital\_los\_days', 'icu\_death', 'icu\_los\_days', 'albumin\_apache', 'apache\_2\_diagnosis', 'apache\_3j\_diagnosis', 'apache\_post\_operative', 'arf\_apache', 'fio2\_apache', 'map\_apache', 'ph\_apache', 'temp\_apache', 'ventilated\_apache', 'wbc\_apache', 'd1\_heartrate\_max', 'd1\_heartrate\_min', 'd1\_sysbp\_invasive\_min', 'h1\_heartrate\_max', 'h1\_heartrate\_min', 'd1\_arterial\_ph\_max', 'd1\_arterial\_ph\_min', 'apache\_3j\_score', 'apsiii']

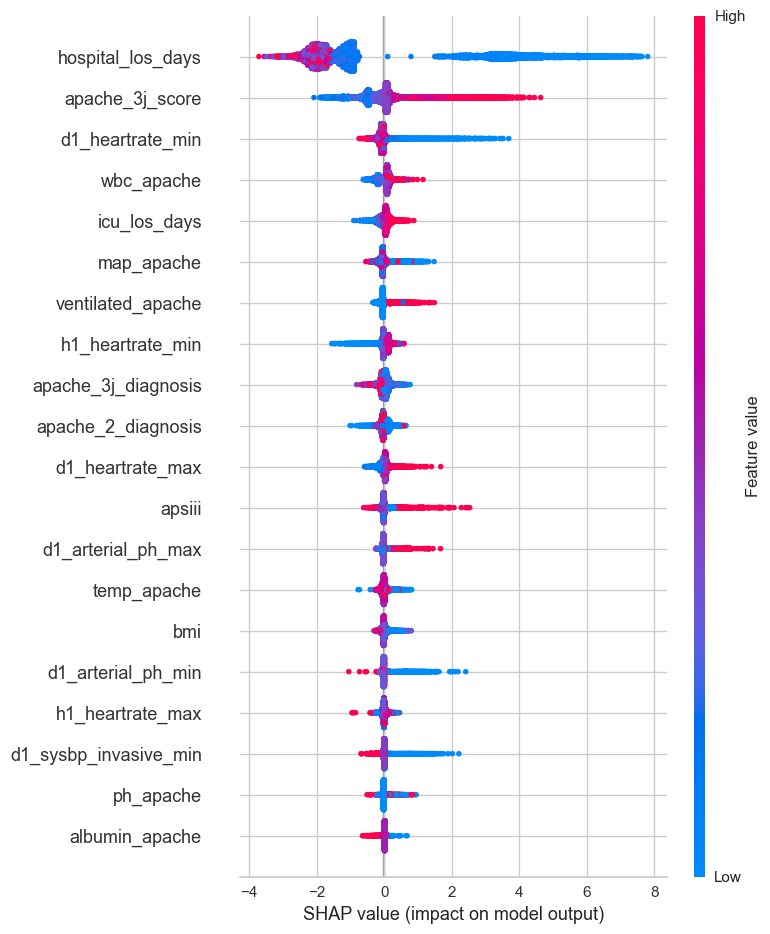
****

****

****

****

****

****