Patient Survival Prediction

Capstone Final Report

Mentored by
Animesh Tiwari

Group 1
Aakriti Gulati
Ciana S Raj
Md Sharib Zeya
Misha Rajagobi
Sheetal Yadav

Problem Statement

The aim is to build a model that helps to predict whether the patient will survive based on their physiology factors that is collected at the time of patient's admission. By building the model, it is to help healthcare providers prioritize their patients as well as improve survival outcomes since population is increasing at a faster rate along with increase in comorbidities. The source of the dataset used to help build the model is from Kaggle. The dataset contains 91713 records and 186 attributes.

About the Dataset

The dataset has 186 attributes which is split up into 9 categories. From these attributes, we are considering hospital_death as our target variable. There are 170 float, 8 int, and 8 object data types.

- Identifier
- Demographic
- APACHE covariate
- APACHE comorbidity
- APACHE grouping
- APACHE prediction
- Labs
- Labs blood gas
- Vitals

Attributes

Identifier

The attributes under the identifier category are to uniquely identify the records.

Variable Name	Data Type	Description
encounter_id	integer	Unique identifier associated with a patient unit stay
hospital_id	integer	Unique identifier associated with a hospital
patient_id	integer	Unique identifier associated with a patient
icu_id	integer	A unique identifier for the unit to which the patient was admitted

Demographic

The demographic attributes tell us the details of patient, as well as information about which unit the patient is being admitted to and whether it was a scheduled surgery.

Variable Name	Data Type	Description
hospital_death	binary	Whether the patient died during this hospitalization
age	numeric	The age of the patient on unit admission
bmi	string	The body mass index of the person on unit admission
elective_surgery	binary	Whether the patient was admitted to the hospital for an elective surgical operation
ethnicity	string	The common national or cultural tradition which the person belongs to
gender	string	The genotypical sex of the patient
height	numeric	The height of the person on unit admission
hospital_admit_source	string	The location of the patient prior to being admitted to the hospital
icu_admit_source	string	The location of the patient prior to being admitted to the unit
icu_admit_type	string	The type of unit admission for the patient
icu_stay_type	string	
icu_type	string	A classification which indicates the type of care the unit is capable of providing
pre_icu_los_days	numeric	The length of stay of the patient between hospital admission and unit admission
readmission_status	binary	Whether the current unit stay is the second (or greater) stay at an ICU within the same hospitalization
weight	numeric	The weight (body mass) of the person on unit admission

APACHE covariate

The following attributes contains readings that were taken at the time of admission. It contains the diagnosis as well.

Variable Name	Data Type	Description
albumin_apache	numeric	The albumin concentration measured during the first 24 hours which results in the highest APACHE III score

apache_2_diagnosis	string	The APACHE II diagnosis for the ICU admission
apache_3j_diagnosis	string	The APACHE III-J sub-diagnosis code which best describes the reason for the ICU admission
apache_post_operative	binary	The APACHE operative status; 1 for post-operative, 0 for non-operative
arf_apache	binary	Whether the patient had acute renal failure during the first 24 hours of their unit stay, defined as a 24 hour urine output <410ml, creatinine >=133 micromol/L and no chronic dialysis
bilirubin_apache	numeric	The bilirubin concentration measured during the first 24 hours which results in the highest APACHE III score
bun_apache	numeric	The blood urea nitrogen concentration measured during the first 24 hours which results in the highest APACHE III score
creatinine_apache	numeric	The creatinine concentration measured during the first 24 hours which results in the highest APACHE III score
fio2_apache	numeric	The fraction of inspired oxygen from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation
gcs_eyes_apache	integer	The eye opening component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
gcs_motor_apache	integer	The motor component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
gcs_unable_apache	binary	Whether the Glasgow Coma Scale was unable to be assessed due to patient sedation
gcs_verbal_apache	integer	The verbal component of the Glasgow Coma Scale measured during the first 24 hours which results in the highest APACHE III score
glucose_apache	numeric	The glucose concentration measured during the first 24 hours which results in the highest APACHE III score
heart_rate_apache	numeric	The heart rate measured during the first 24 hours which results in the highest APACHE III score
hematocrit_apache	numeric	The hematocrit measured during the first 24 hours which results in the highest APACHE III score
intubated_apache	binary	Whether the patient was intubated at the time of the highest scoring arterial blood gas used in the oxygenation score
map_apache	numeric	The mean arterial pressure measured during the first 24 hours which results in the highest APACHE III score
paco2_apache	numeric	The partial pressure of carbon dioxide from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation
paco2_for_ph_apache	numeric	The partial pressure of carbon dioxide from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for acid-base disturbance
pao2_apache	numeric	The partial pressure of oxygen from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for oxygenation
ph_apache	numeric	The pH from the arterial blood gas taken during the first 24 hours of unit admission which produces the highest APACHE III score for acid-base disturbance
resprate_apache	numeric	The respiratory rate measured during the first 24 hours which results in the highest APACHE III score
sodium_apache	numeric	The sodium concentration measured during the first 24 hours which results in the highest APACHE III score

temp_apache	numeric	The temperature measured during the first 24 hours which results in the highest APACHE III score
urineoutput_apache	numeric	The total urine output for the first 24 hours
ventilated_apache	binary	Whether the patient was invasively ventilated at the time of the highest scoring arterial blood gas using the oxygenation scoring algorithm, including any mode of positive pressure ventilation delivered through a circuit attached to an endo-tracheal tube or tracheostomy
wbc_apache	numeric	The white blood cell count measured during the first 24 hours which results in the highest APACHE III score

APACHE Comorbidity

The comorbidities attributes contains indication of whether the patient has previously or additionally diagnosed with the following diseases.

Variable Name	Data Type	Description
aids	binary	Whether the patient has a definitive diagnosis of acquired immune deficiency syndrome (AIDS) (not HIV positive alone)
cirrhosis	binary	Whether the patient has a history of heavy alcohol use with portal hypertension and varices, other causes of cirrhosis with evidence of portal hypertension and varices, or biopsy proven cirrhosis. This comorbidity does not apply to patients with a functioning liver transplant.
diabetes_mellitus	binary	Whether the patient has been diagnosed with diabetes, either juvenile or adult onset, which requires medication.
hepatic_failure	binary	Whether the patient has cirrhosis and additional complications including jaundice and ascites, upper GI bleeding, hepatic encephalopathy, or coma.
immunosuppression	binary	Whether the patient has their immune system suppressed within six months prior to ICU admission for any of the following reasons; radiation therapy, chemotherapy, use of non-cytotoxic immunosuppressive drugs, high dose steroids (at least 0.3 mg/kg/day of methylprednisolone or equivalent for at least 6 months).
leukemia	binary	Whether the patient has been diagnosed with acute or chronic myelogenous leukemia, acute or chronic lymphocytic leukemia, or multiple myeloma.
lymphoma	binary	Whether the patient has been diagnosed with non-Hodgkin lymphoma.
solid_tumor_with_metastasis	binary	Whether the patient has been diagnosed with any solid tumor carcinoma (including malignant melanoma) which has evidence of metastasis.

APACHE grouping

The variables in APACHE grouping are all categorical variables. They describe the diagnosis for the patient.

Variable Name	Data Type	Description
apache_3j_bodysystem	string	Admission diagnosis group for APACHE III
apache_2_bodysystem	string	Admission diagnosis group for APACHE II

APACHE prediction

The death probability is calculated on the basis of the APACHE covariates.

Variable Name	Data Type	Description
apache_4a_hospital_death_prob	numeric	The APACHE IVa probabilistic prediction of in-hospital mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis.
apache_4a_icu_death_prob	numeric	The APACHE IVa probabilistic prediction of in ICU mortality for the patient which utilizes the APACHE III score and other covariates, including diagnosis

Labs

The variables containing information on highest and lowest concentration of certain tests at the first hour(h1) and first 24 hours(d1).

Variable Name	Data Type	Description
d1_albumin_max	numeric	The lowest albumin concentration of the patient in their serum during the first 24 hours of their unit stay
d1_albumin_min	numeric	The lowest albumin concentration of the patient in their serum during the first 24 hours of their unit stay
d1_bilirubin_max	numeric	The highest bilirubin concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_bilirubin_min	numeric	The lowest bilirubin concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_bun_max	numeric	The highest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_bun_min	numeric	The lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_calcium_max	numeric	The highest calcium concentration of the patient in their serum during the first 24 hours of their unit stay
d1_calcium_min	numeric	The lowest calcium concentration of the patient in their serum during the first 24 hours of their unit stay
d1_creatinine_max	numeric	The highest creatinine concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_creatinine_min	numeric	The lowest creatinine concentration of the patient in their serum or plasma during the first 24 hours of their unit stay

d1_glucose_max	numeric	The highest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_glucose_min	numeric	The lowest glucose concentration of the patient in their serum or plasma during the first 24 hours of their unit stay
d1_hco3_max	numeric	The highest bicarbonate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_hco3_min	numeric	The lowest bicarbonate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_hemaglobin_max	numeric	The highest hemoglobin concentration for the patient during the first 24 hours of their unit stay
d1_hemaglobin_min	numeric	The lowest hemoglobin concentration for the patient during the first 24 hours of their unit stay
d1_hematocrit_max	numeric	The highest volume proportion of red blood cells in a patient's blood during the first 24 hours of their unit stay, expressed as a fraction
d1_hematocrit_min	numeric	The lowest volume proportion of red blood cells in a patient's blood during the first 24 hours of their unit stay, expressed as a fraction
d1_inr_max	numeric	The highest international normalized ratio for the patient during the first 24 hours of their unit stay
d1_inr_min	numeric	The lowest international normalized ratio for the patient during the first 24 hours of their unit stay
d1_lactate_max	numeric	The highest lactate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_lactate_min	numeric	The lowest lactate concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_platelets_max	numeric	The highest platelet count for the patient during the first 24 hours of their unit stay
d1_platelets_min	numeric	The lowest platelet count for the patient during the first 24 hours of their unit stay
d1_potassium_max	numeric	The highest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_potassium_min	numeric	The lowest potassium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_sodium_max	numeric	The highest sodium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_sodium_min	numeric	The lowest sodium concentration for the patient in their serum or plasma during the first 24 hours of their unit stay
d1_wbc_max	numeric	The highest white blood cell count for the patient during the first 24 hours of their unit stay
d1_wbc_min	numeric	The lowest white blood cell count for the patient during the first 24 hours of their unit stay
h1_albumin_max	numeric	The lowest albumin concentration of the patient in their serum during the first hour of their unit stay
h1_albumin_min	numeric	The lowest albumin concentration of the patient in their serum during the first hour of their unit stay
h1_bilirubin_max	numeric	The highest bilirubin concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_bilirubin_min	numeric	The lowest bilirubin concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_bun_max	numeric	The highest blood urea nitrogen concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_bun_min	numeric	The lowest blood urea nitrogen concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_calcium_max	numeric	The highest calcium concentration of the patient in their serum during the first hour of their unit stay

h1_calcium_min	numeric	The lowest calcium concentration of the patient in their serum during the first hour of their unit stay
h1_creatinine_max	numeric	The highest creatinine concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_creatinine_min	numeric	The lowest creatinine concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_glucose_max	numeric	The highest glucose concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_glucose_min	numeric	The lowest glucose concentration of the patient in their serum or plasma during the first hour of their unit stay
h1_hco3_max	numeric	The highest bicarbonate concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_hco3_min	numeric	The lowest bicarbonate concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_hemaglobin_max	numeric	The highest hemoglobin concentration for the patient during the first hour of their unit stay
h1_hemaglobin_min	numeric	The lowest hemoglobin concentration for the patient during the first hour of their unit stay
h1_hematocrit_max	numeric	The highest volume proportion of red blood cells in a patient's blood during the first hour of their unit stay, expressed as a fraction
h1_hematocrit_min	numeric	The lowest volume proportion of red blood cells in a patient's blood during the first hour of their unit stay, expressed as a fraction
h1_inr_max	numeric	The highest international normalized ratio for the patient during the first hour of their unit stay
h1_inr_min	numeric	The lowest international normalized ratio for the patient during the first hour of their unit stay
h1_lactate_max	numeric	The highest lactate concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_lactate_min	numeric	The lowest lactate concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_platelets_max	numeric	The highest platelet count for the patient during the first hour of their unit stay
h1_platelets_min	numeric	The lowest platelet count for the patient during the first hour of their unit stay
h1_potassium_max	numeric	The highest potassium concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_potassium_min	numeric	The lowest potassium concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_sodium_max	numeric	The highest sodium concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_sodium_min	numeric	The lowest sodium concentration for the patient in their serum or plasma during the first hour of their unit stay
h1_wbc_max	numeric	The highest white blood cell count for the patient during the first hour of their unit stay
h1_wbc_min	numeric	The lowest white blood cell count for the patient during the first hour of their unit stay

Labs Blood Gas

Variable Name	Data Type	Description
d1_arterial_pco2_max	numeric	The highest arterial partial pressure of carbon dioxide for the patient during the first 24 hours of their unit stay
d1_arterial_pco2_min	numeric	The lowest arterial partial pressure of carbon dioxide for the patient during the first 24 hours of their unit stay
d1_arterial_ph_max	numeric	The highest arterial pH for the patient during the first 24 hours of their unit stay
d1_arterial_ph_min	numeric	The lowest arterial pH for the patient during the first 24 hours of their unit stay
d1_arterial_po2_max	numeric	The highest arterial partial pressure of oxygen for the patient during the first 24 hours of their unit stay
d1_arterial_po2_min	numeric	The lowest arterial partial pressure of oxygen for the patient during the first 24 hours of their unit stay
d1_pao2fio2ratio_max	numeric	The highest fraction of inspired oxygen for the patient during the first 24 hours of their unit stay
d1_pao2fio2ratio_min	numeric	The lowest fraction of inspired oxygen for the patient during the first 24 hours of their unit stay
h1_arterial_pco2_max	numeric	The highest arterial partial pressure of carbon dioxide for the patient during the first hour of their unit stay
h1_arterial_pco2_min	numeric	The lowest arterial partial pressure of carbon dioxide for the patient during the first hour of their unit stay
h1_arterial_ph_max	numeric	The highest arterial pH for the patient during the first hour of their unit stay
h1_arterial_ph_min	numeric	The lowest arterial pH for the patient during the first hour of their unit stay
h1_arterial_po2_max	numeric	The highest arterial partial pressure of oxygen for the patient during the first hour of their unit stay
h1_arterial_po2_min	numeric	The lowest arterial partial pressure of oxygen for the patient during the first hour of their unit stay
h1_pao2fio2ratio_max	numeric	The highest fraction of inspired oxygen for the patient during the first hour of their unit stay
h1_pao2fio2ratio_min	numeric	The lowest fraction of inspired oxygen for the patient during the first hour of their unit stay

Vitals

Variable Name	Data Type	Description
d1_diasbp_invasive_max	numeric	The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_diasbp_invasive_min	numeric	The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_diasbp_max	numeric	The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_diasbp_min	numeric	The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_diasbp_noninvasive_max	numeric	The patient's highest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_diasbp_noninvasive_min	numeric	The patient's lowest diastolic blood pressure during the first 24 hours of their unit stay, non-invasively measured

d1_heartrate_max	numeric	The patient's highest heart rate during the first 24 hours of their unit stay
d1_heartrate_min	numeric	The patient's lowest heart rate during the first 24 hours of their unit stay
d1_mbp_invasive_max	numeric	The patient's highest mean blood pressure during the first 24 hours of their unit stay, invasively measured
d1_mbp_invasive_min	numeric	The patient's lowest mean blood pressure during the first 24 hours of their unit stay, invasively measured
d1_mbp_max	numeric	The patient's highest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_mbp_min	numeric	The patient's lowest mean blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_mbp_noninvasive_max	numeric	The patient's highest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_mbp_noninvasive_min	numeric	The patient's lowest mean blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_resprate_max	numeric	The patient's highest respiratory rate during the first 24 hours of their unit stay
d1_resprate_min	numeric	The patient's lowest respiratory rate during the first 24 hours of their unit stay
d1_spo2_max	numeric	The patient's highest peripheral oxygen saturation during the first 24 hours of their unit stay
d1_spo2_min	numeric	The patient's lowest peripheral oxygen saturation during the first 24 hours of their unit stay
d1_sysbp_invasive_max	numeric	The patient's highest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_sysbp_invasive_min	numeric	The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, invasively measured
d1_sysbp_max	numeric	The patient's highest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_sysbp_min	numeric	The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, either non-invasively or invasively measured
d1_sysbp_noninvasive_max	numeric	The patient's highest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_sysbp_noninvasive_min	numeric	The patient's lowest systolic blood pressure during the first 24 hours of their unit stay, non-invasively measured
d1_temp_max	numeric	The patient's highest core temperature during the first 24 hours of their unit stay, invasively measured
d1_temp_min	numeric	The patient's lowest core temperature during the first 24 hours of their unit stay
h1_diasbp_invasive_max	numeric	The patient's highest diastolic blood pressure during the first hour of their unit stay, invasively measured
h1_diasbp_invasive_min	numeric	The patient's lowest diastolic blood pressure during the first hour of their unit stay, invasively measured
h1_diasbp_max	numeric	The patient's highest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_diasbp_min	numeric	The patient's lowest diastolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_diasbp_noninvasive_max	numeric	The patient's highest diastolic blood pressure during the first hour of their unit stay, non-invasively measured
h1_diasbp_noninvasive_min	numeric	The patient's lowest diastolic blood pressure during the first hour of their unit stay, non-invasively measured
h1_heartrate_max	numeric	The patient's highest heart rate during the first hour of their unit stay

h1_heartrate_min	numeric	The patient's lowest heart rate during the first hour of their unit stay
h1_mbp_invasive_max	numeric	The patient's highest mean blood pressure during the first hour of their unit stay, invasively measured
h1_mbp_invasive_min	numeric	The patient's lowest mean blood pressure during the first hour of their unit stay, invasively measured
h1_mbp_max	numeric	The patient's highest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_mbp_min	numeric	The patient's lowest mean blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_mbp_noninvasive_max	numeric	The patient's highest mean blood pressure during the first hour of their unit stay, non-invasively measured
h1_mbp_noninvasive_min	numeric	The patient's lowest mean blood pressure during the first hour of their unit stay, non-invasively measured
h1_resprate_max	numeric	The patient's highest respiratory rate during the first hour of their unit stay
h1_resprate_min	numeric	The patient's lowest respiratory rate during the first hour of their unit stay
h1_spo2_max	numeric	The patient's highest peripheral oxygen saturation during the first hour of their unit stay
h1_spo2_min	numeric	The patient's lowest peripheral oxygen saturation during the first hour of their unit stay
h1_sysbp_invasive_max	numeric	The patient's highest systolic blood pressure during the first hour of their unit stay, invasively measured
h1_sysbp_invasive_min	numeric	The patient's lowest systolic blood pressure during the first hour of their unit stay, invasively measured
h1_sysbp_max	numeric	The patient's highest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_sysbp_min	numeric	The patient's lowest systolic blood pressure during the first hour of their unit stay, either non-invasively or invasively measured
h1_sysbp_noninvasive_max	numeric	The patient's highest systolic blood pressure during the first hour of their unit stay, non-invasively measured
h1_sysbp_noninvasive_min	numeric	The patient's lowest systolic blood pressure during the first hour of their unit stay, non-invasively measured
h1_temp_max	numeric	The patient's highest core temperature during the first hour of their unit stay, invasively measured
h1_temp_min	numeric	The patient's lowest core temperature during the first hour of their unit stay

Missing Values

The first step after checking out the dataset, is to find the missing values. In this step, we check if there any attributes that can be dropped if it contains more than 70-80% missing data.

Attributes	Total missing values	Percentage of missing values
h1_bilirubin_max	84619	92.27 %
h1_bilirubin_min	84619	92.27 %
h1_lactate_max	84369	91.99 %
h1_lactate_min	84369	91.99 %
h1_albumin_max	83824	91.4 %
h1_albumin_min	83824	91.4 %

h1_pao2fio2ratio_max	80195	87.44 %
h1_pao2fio2ratio_min	80195	87.44 %
h1_arterial_ph_max	76424	83.33 %
h1_arterial_ph_min	76424	83.33 %
h1_hco3_max	76094	82.97 %
h1_hco3_min	76094	82.97 %
h1_arterial_pco2_max	75959	82.82 %
h1_arterial_pco2_min	75959	82.82 %
h1_wbc_max	75953	82.82 %
h1_wbc_min	75953	82.82 %
h1_arterial_po2_max	75945	82.81 %
h1_arterial_po2_min	75945	82.81 %
h1_calcium_max	75863	82.72 %
h1_calcium_min	75863	82.72 %
h1_platelets_max	75673	82.51 %
h1_platelets_min	75673	82.51 %
h1_bun_max	75091	81.88 %
h1_bun_min	75091	81.88 %
h1_creatinine_max	74957	81.73 %
h1_creatinine_min	74957	81.73 %
h1_diasbp_invasive_max	74928	81.7 %
h1_diasbp_invasive_min	74928	81.7 %
h1_sysbp_invasive_max	74915	81.68 %
h1_sysbp_invasive_min	74915	81.68 %
h1_mbp_invasive_max	74844	81.61 %
h1_mbp_invasive_min	74844	81.61 %
h1_hematocrit_max	73420	80.05 %
h1_hematocrit_min	73420	80.05 %
h1_hemaglobin_max	73123	79.73 %
h1_hemaglobin_min	73123	79.73 %
h1_sodium_max	72617	79.18 %
h1_sodium_min	72617	79.18 %
h1_potassium_max	72102	78.62 %
h1_potassium_min	72102	78.62 %
fio2_apache	70868	77.27 %
paco2_apache	70868	77.27 %
paco2_for_ph_apache	70868	77.27 %
pao2_apache	70868	77.27 %
ph_apache	70868	77.27 %
d1_lactate_max	68396	74.58 %
d1_lactate_min	68396	74.58 %
d1_diasbp_invasive_max	67984	74.13 %
d1_diasbp_invasive_min	67984	74.13 %
d1_sysbp_invasive_max	67959	74.1 %
d1_sysbp_invasive_min	67959	74.1 %
d1_mbp_invasive_max	67777	73.9 %

d1_mbp_invasive_min	67777	73.9 %
d1_pao2fio2ratio_max	66008	71.97 %
d1_pao2fio2ratio_min	66008	71.97 %
d1_arterial_ph_max	60123	65.56 %
d1_arterial_ph_min	60123	65.56 %
d1_arterial_pco2_max	59271	64.63 %
d1_arterial_pco2_min	59271	64.63 %
d1_arterial_po2_max	59262	64.62 %
d1_arterial_po2_min	59262	64.62 %
bilirubin_apache	58134	63.39 %
d1_inr_max	57941	63.18 %
d1_inr_min	57941	63.18 %
h1_inr_max	57941	63.18 %
h1_inr_min	57941	63.18 %
albumin_apache	54379	59.29 %
d1_bilirubin_max	53673	58.52 %
d1_bilirubin_min	53673	58.52 %
h1_glucose_max	52614	57.37 %
h1_glucose_min	52614	57.37 %
d1_albumin_max	49096	53.53 %
d1_albumin_min	49096	53.53 %
urineoutput_apache	48998	53.43 %
wbc_apache	22012	24 %
h1_temp_max	21732	23.7 %
h1_temp_min	21732	23.7 %
hospital_admit_source	21409	23.34 %
hematocrit_apache	19878	21.67 %
bun_apache	19262	21 %
creatinine_apache	18853	20.56 %
sodium_apache	18600	20.28 %
d1_hco3_max	15071	16.43 %
d1_hco3_min	15071	16.43 %
d1_platelets_max	13444	14.66 %
d1_platelets_min	13444	14.66 %
d1_wbc_max	13174	14.36 %
d1_wbc_min	13174	14.36 %
d1_calcium_max	13069	14.25 %
d1_calcium_min	13069	14.25 %
d1_hemaglobin_max	12147	13.24 %
d1_hemaglobin_min	12147	13.24 %
d1_hematocrit_max	11654	12.71 %
d1_hematocrit_min	11654	12.71 %
glucose_apache	11036	12.03 %
d1_bun_max	10514	11.46 %
d1_bun_min	10514	11.46 %
d1_sodium_max	10195	11.12 %

d1_sodium_min	10195	11.12 %
d1_creatinine_max	10169	11.09 %
d1_creatinine_min	10169	11.09 %
d1_potassium_max	9585	10.45 %
d1_potassium_min	9585	10.45 %
h1_mbp_noninvasive_max	9084	9.9 %
h1_mbp_noninvasive_min	9084	9.9 %
apache_4a_hospital_death_prob	7947	8.67 %
apache_4a_icu_death_prob	7947	8.67 %
h1_diasbp_noninvasive_max	7350	8.01 %
h1_diasbp_noninvasive_min	7350	8.01 %
h1_sysbp_noninvasive_max	7341	8 %
h1_sysbp_noninvasive_min	7341	8 %
d1_glucose_max	5807	6.33 %
d1_glucose_min	5807	6.33 %
h1_mbp_max	4639	5.06 %
h1_mbp_min	4639	5.06 %
h1_resprate_max	4357	4.75 %
h1_resprate_min	4357	4.75 %
age	4228	4.61 %
h1_spo2_max	4185	4.56 %
h1_spo2_min	4185	4.56 %
temp_apache	4108	4.48 %
h1_diasbp_max	3619	3.95 %
h1_diasbp_min	3619	3.95 %
h1_sysbp_max	3611	3.94 %
h1_sysbp_min	3611	3.94 %
bmi	3429	3.74 %
h1_heartrate_max	2790	3.04 %
h1_heartrate_min	2790	3.04 %
weight	2720	2.97 %
d1_temp_max	2324	2.53 %
d1_temp_min	2324	2.53 %
gcs_eyes_apache	1901	2.07 %
gcs_motor_apache	1901	2.07 %
gcs_verbal_apache	1901	2.07 %
apache_2_diagnosis	1662	1.81 %
apache_3j_bodysystem	1662	1.81 %
apache_2_bodysystem	1662	1.81 %
d1_mbp_noninvasive_max	1479	1.61 %
d1_mbp_noninvasive_min	1479	1.61 %
ethnicity	1395	1.52 %
height	1334	1.45 %
resprate_apache	1234	1.35 %
apache_3j_diagnosis	1101	1.2 %
d1_diasbp_noninvasive_max	1040	1.13 %

d1_diasbp_noninvasive_min	1040	1.13 %
gcs_unable_apache	1037	1.13 %
d1_sysbp_noninvasive_max	1027	1.12 %
d1_sysbp_noninvasive_min	1027	1.12 %
map_apache	994	1.08 %
heart_rate_apache	878	0.96 %
arf_apache	715	0.78 %
intubated_apache	715	0.78 %
ventilated_apache	715	0.78 %
aids	715	0.78 %
cirrhosis	715	0.78 %
diabetes_mellitus	715	0.78 %
hepatic_failure	715	0.78 %
immunosuppression	715	0.78 %
leukemia	715	0.78 %
lymphoma	715	0.78 %
solid_tumor_with_metastasis	715	0.78 %
d1_resprate_max	385	0.42 %
d1_resprate_min	385	0.42 %
d1_spo2_max	333	0.36 %
d1_spo2_min	333	0.36 %
d1_mbp_max	220	0.24 %
d1_mbp_min	220	0.24 %
d1_diasbp_max	165	0.18 %
d1_diasbp_min	165	0.18 %
d1_sysbp_max	159	0.17 %
d1_sysbp_min	159	0.17 %
d1_heartrate_max	145	0.16 %
d1_heartrate_min	145	0.16 %
icu_admit_source	112	0.12 %
gender	25	0.03 %

Inference

The labs and vitals attributes such as h1_bilirubin, h1_lactate, h1_albumin, h1_calcium, h1_creatine, h1_sodium, d1_lactate, d1_arterial, fio2, paco2 and few more contains more than 70% missing data. A few demographic and APACHE attributes contain missing values as well, but they are essential for model building and are in the range of 4.6-0.03% and 64-0.7% respectively. Hence, these attributes will be processed in the later stages.

Data Preprocessing

Missing Value Imputation

Attributes that contain more than 70% missing values

The number of attributes that had more than 70% of missing values is 55. As mentioned, the attributes belonged to the labs, vitals and apache covariate groups. Since a large amount of data is missing, the attributes are removed from the dataset.

Attributes that contain less than 70% missing values

The missing values are handled by imputing mean/mode depending on their types using sklearn SimpleImputer library.

For categorical attributes, the missing value was replaced with most frequently occurring value, whereas for numerical attributes, depending on the skewness of the attribute, mean or median was added.

Irrelevant attributes

By researching more into the attributes, we had decided to remove few attributes as they do not contribute to the survival rate of a patient. The attributes were mainly identifiers, apache covariate, labs and lab blood gas variables, such as albumin_apache, apache_post_operative, bilirubin_apache, encounter_id, fio2_apache, hospital_id, icu_id, paco2_apache, paco2_for_ph_apache, pao2_apache, patient_id, ph_apache, pre_icu_los_days, readmission_status and urineoutput_apache

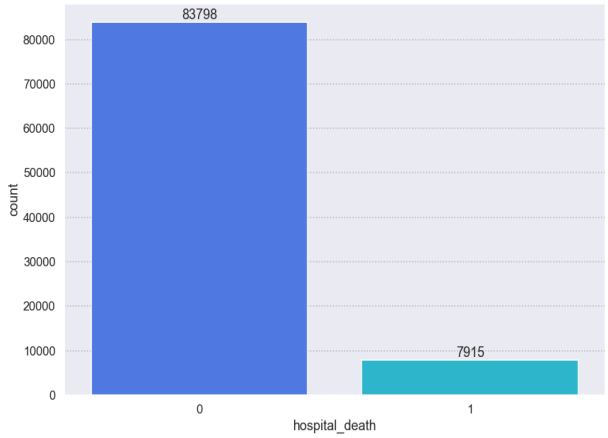
Outlier Detection and Treatment

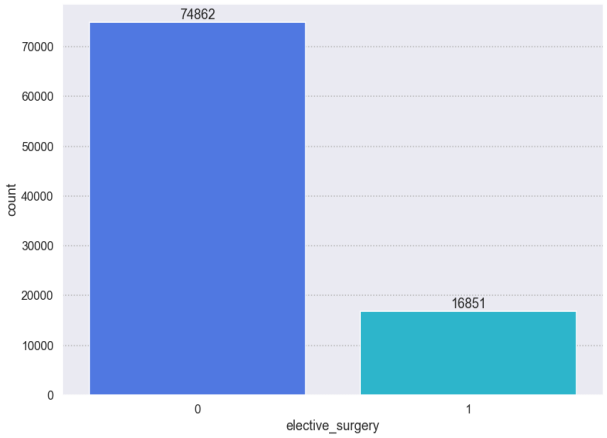
The outliers were detected using the IQR method and boxplots were used for visualization.

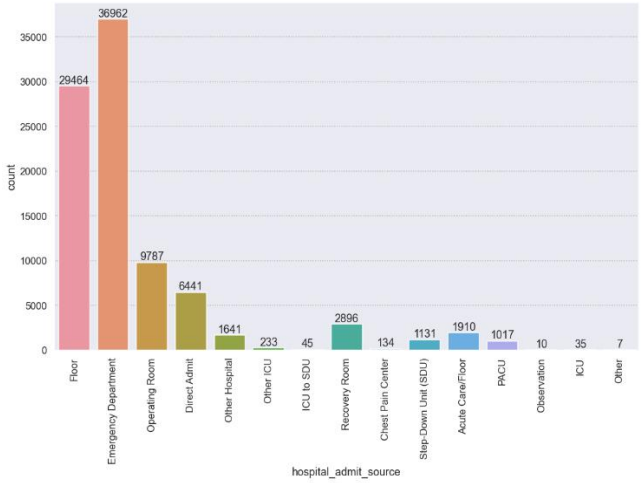
For the treatment, we had initially used IQR and Z scale methods were applied to treat the outliers depending on their skewness. On further research, we used PowerTransform to treat the outliers as it stabilizes the variance or in other words reduces the skewness. By doing so, it also does feature scaling of the data.

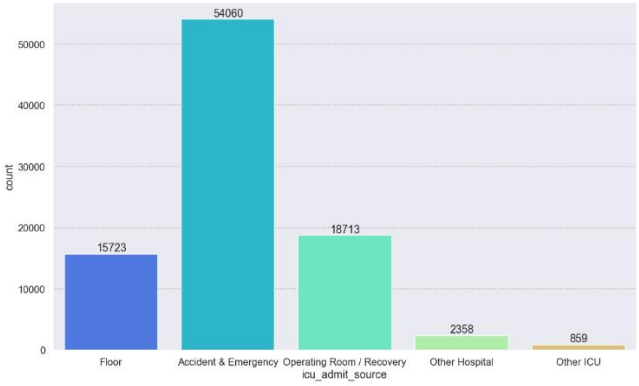
Analysis of Dataset

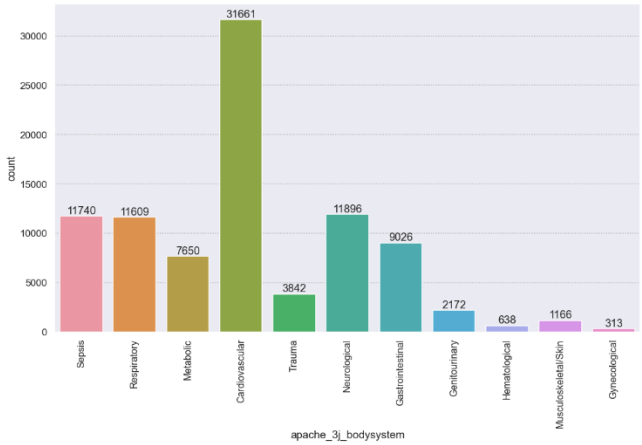
Univariate Analysis

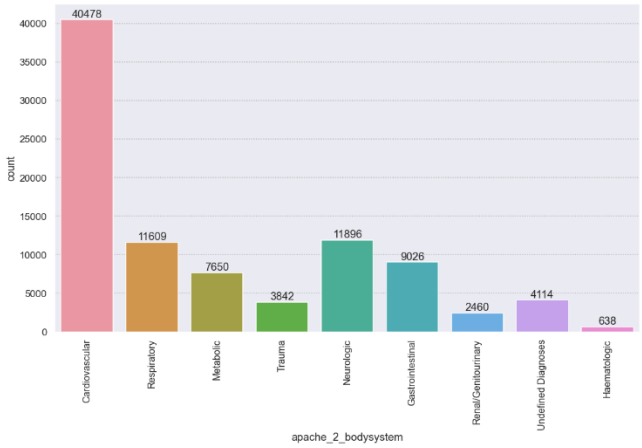


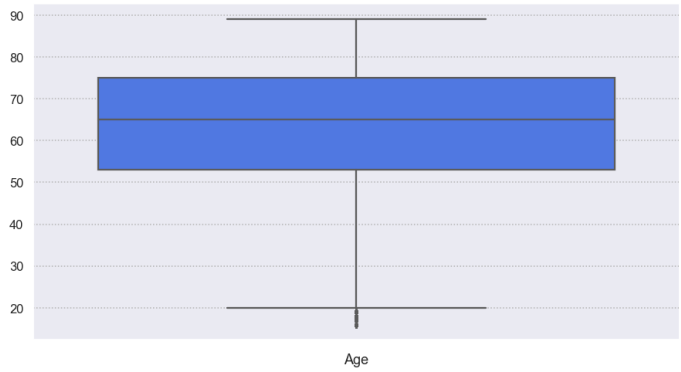


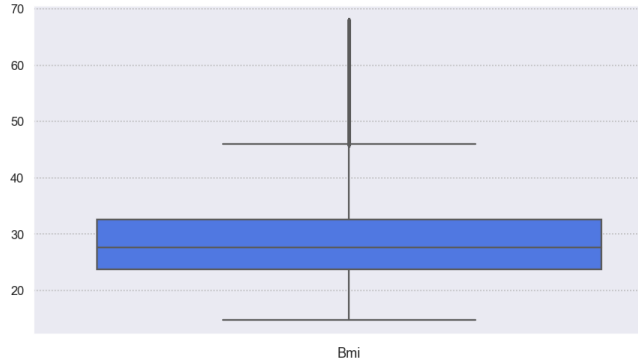


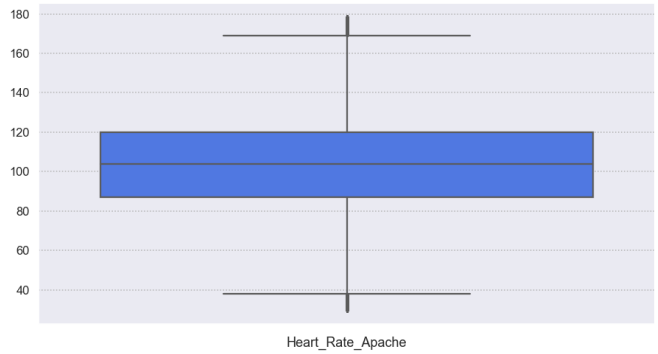










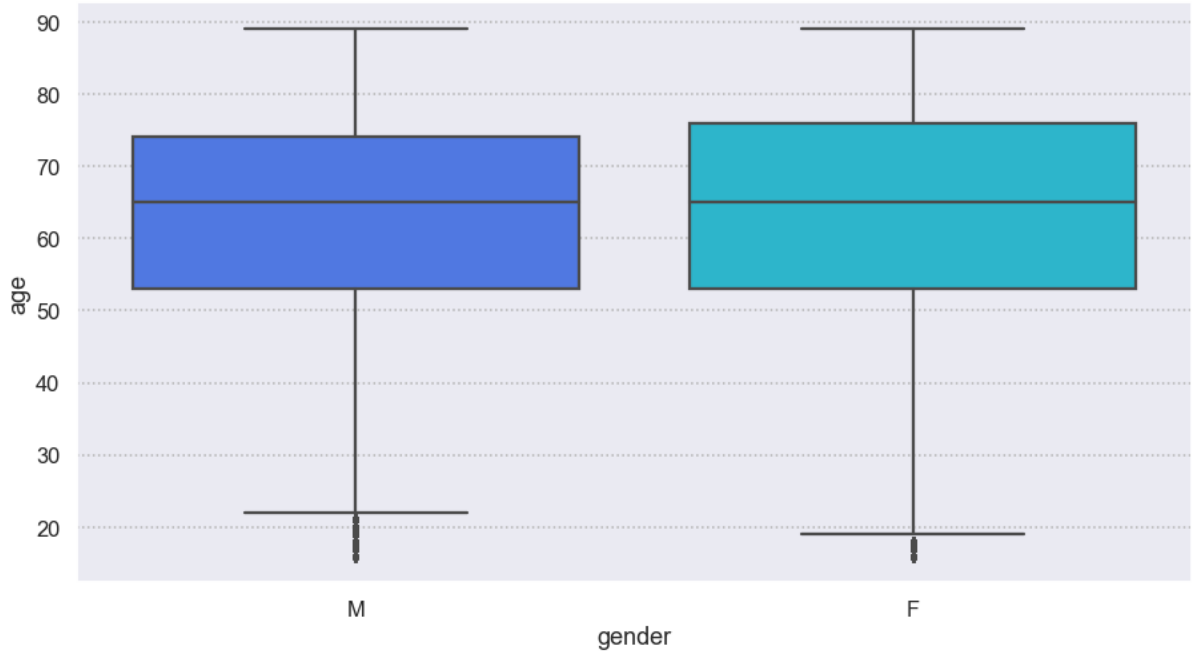


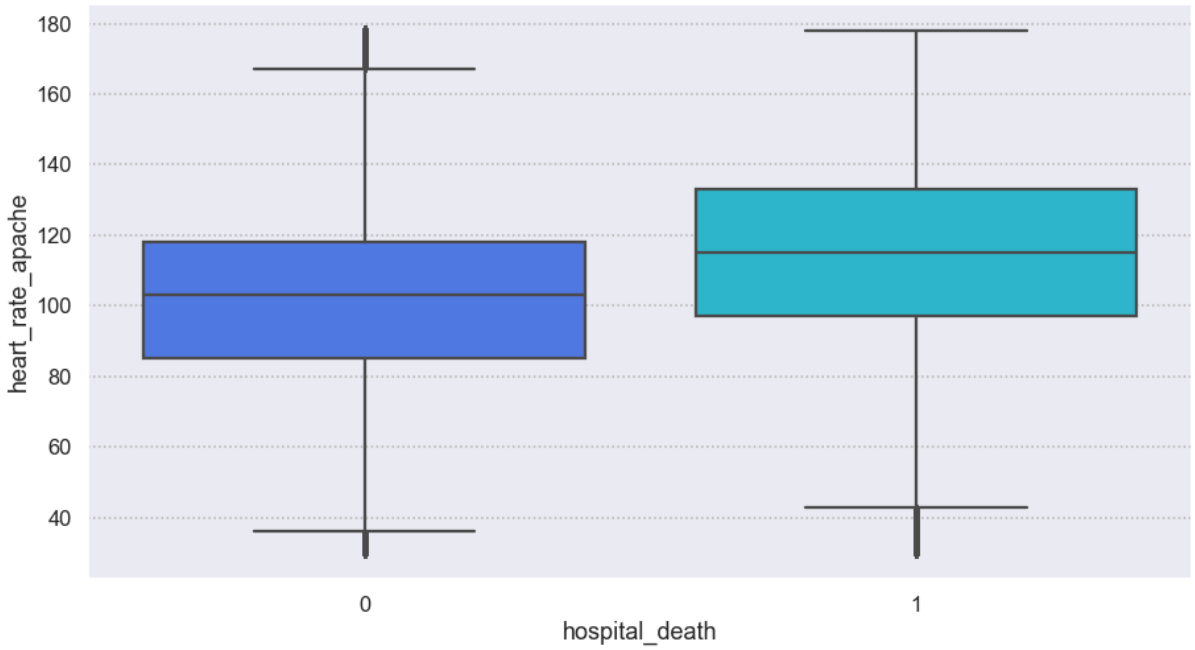
Inferences

- The target variable hospital death has an imbalanced proportion as there are more survivals than deaths.
- From the ethnicity graph it is seen that patients that were admitted were Caucasians in majority. And the least admitted ones are native Americans.
- There isn’t much of a difference between male and female patients.
- In hospital admit source, Emergency Room and Floor are maximum.
- Admission is the most popular one in ICU stay type.
- The unit that had a greater number of patients admitted is Med-Surg.
- Almost all of them have significant number of cases except for Gynecology and Hematology
- The range of age had been admitted ranges from 16 to 90. The age group that is highly likely to be admitted is in the range 50-85, in which the age 85 is the highest number of patients admitted.
- In terms of bmi, it ranges from 15-70. Patients with bmi in the range 20-35 are high in number, especially with bmi 30.
- For height, it ranges from 137 to 195, where the patients with height range 160-180 are higher in number in admissions.
- For weight, patients with weight in the range of 62-100 are more in number.

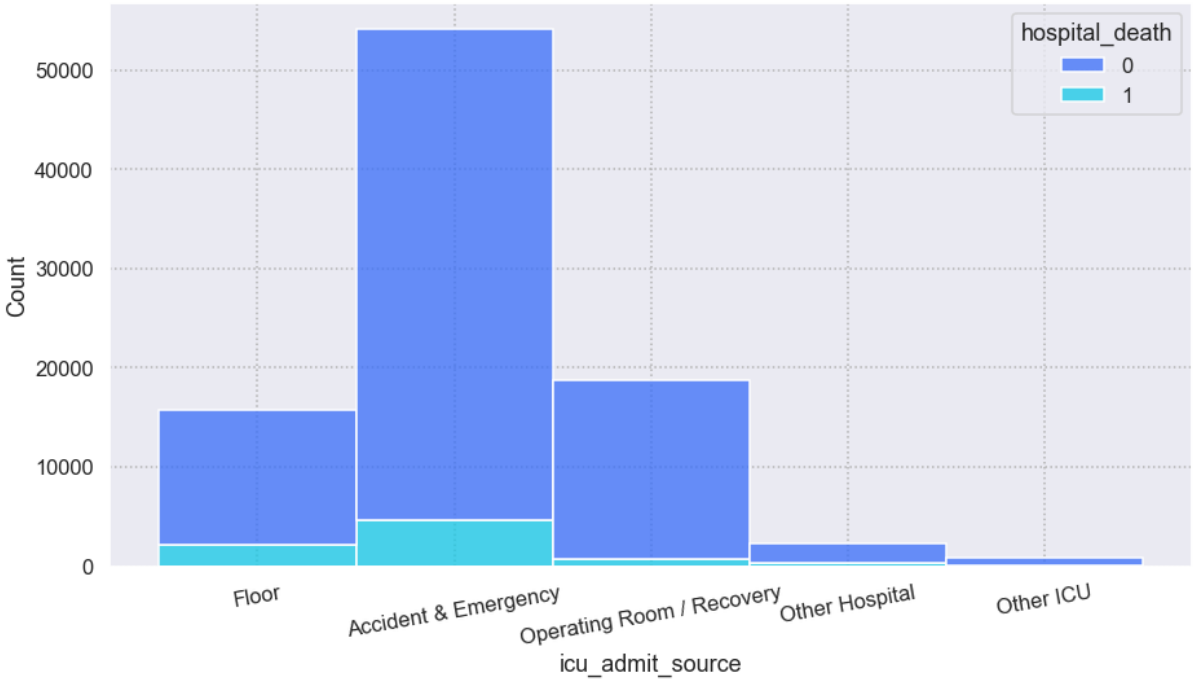
Bivariate Analysis

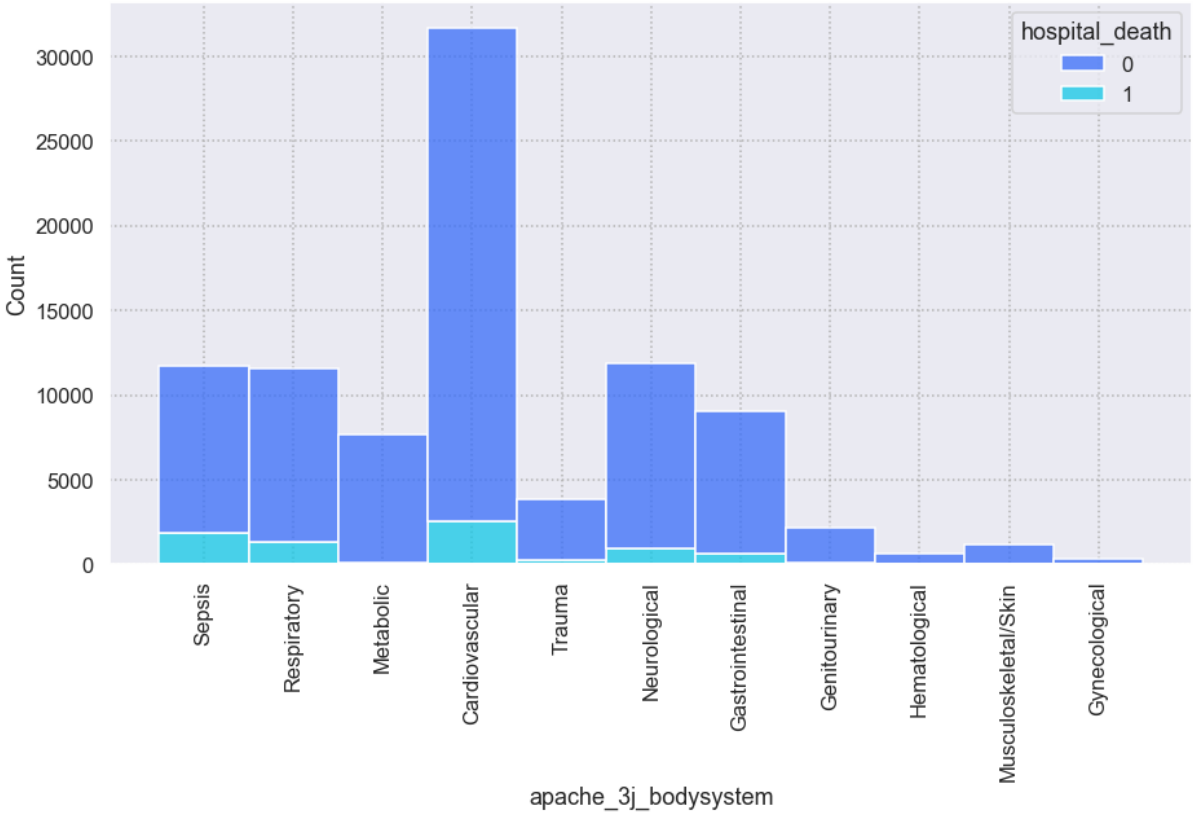
Numerical vs Categorical

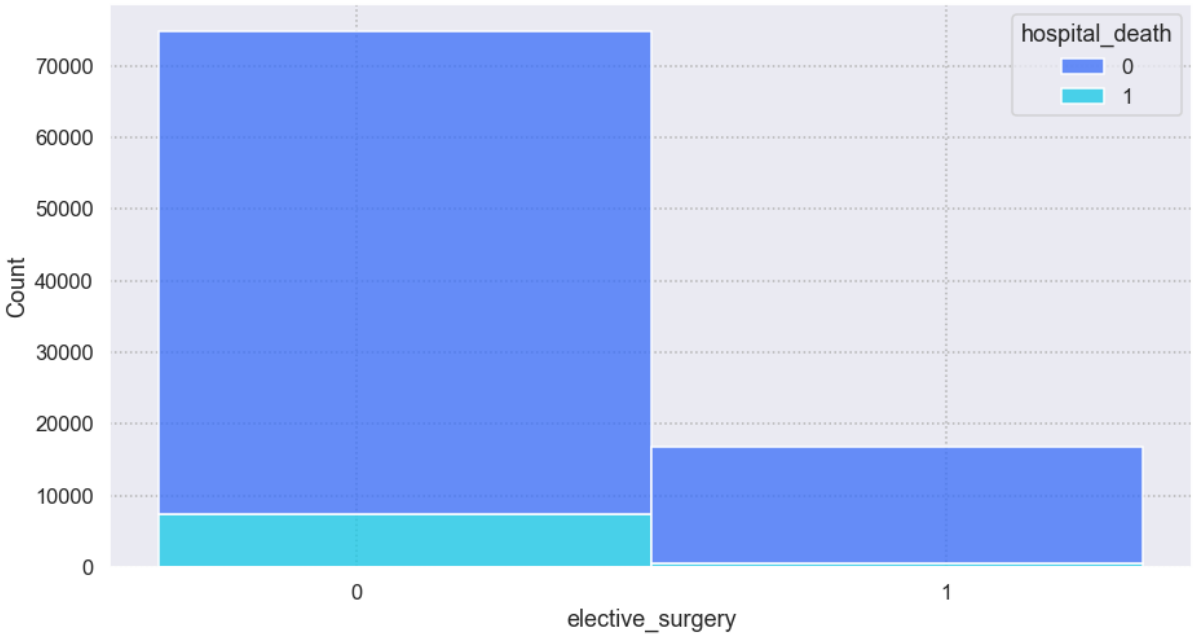


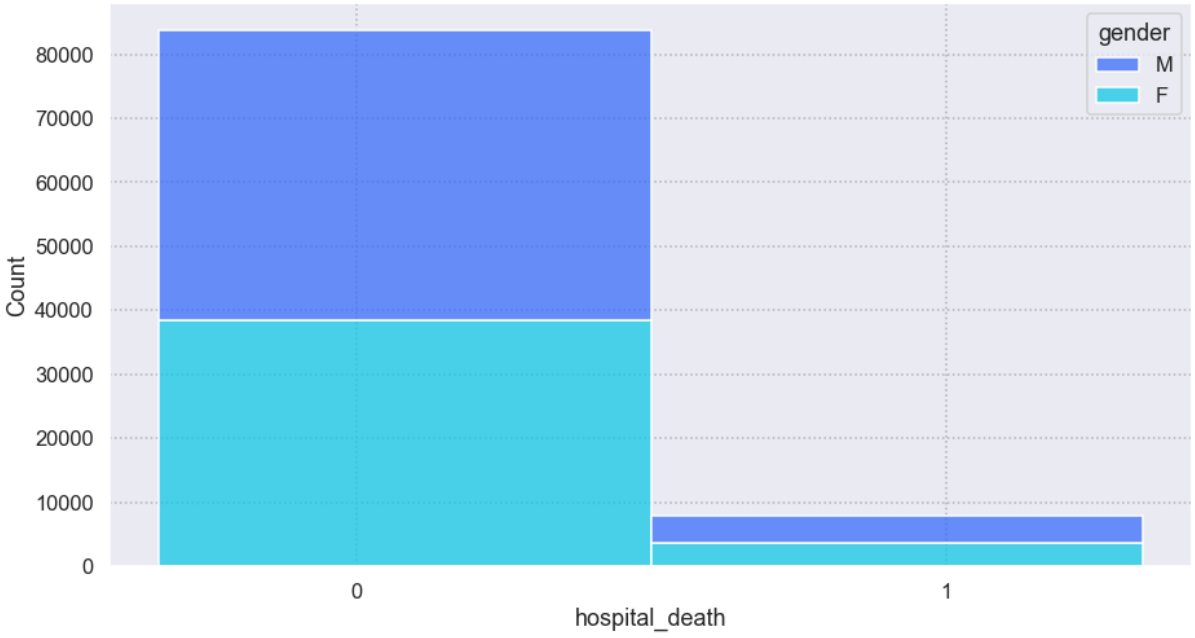


Categorical vs Categorical

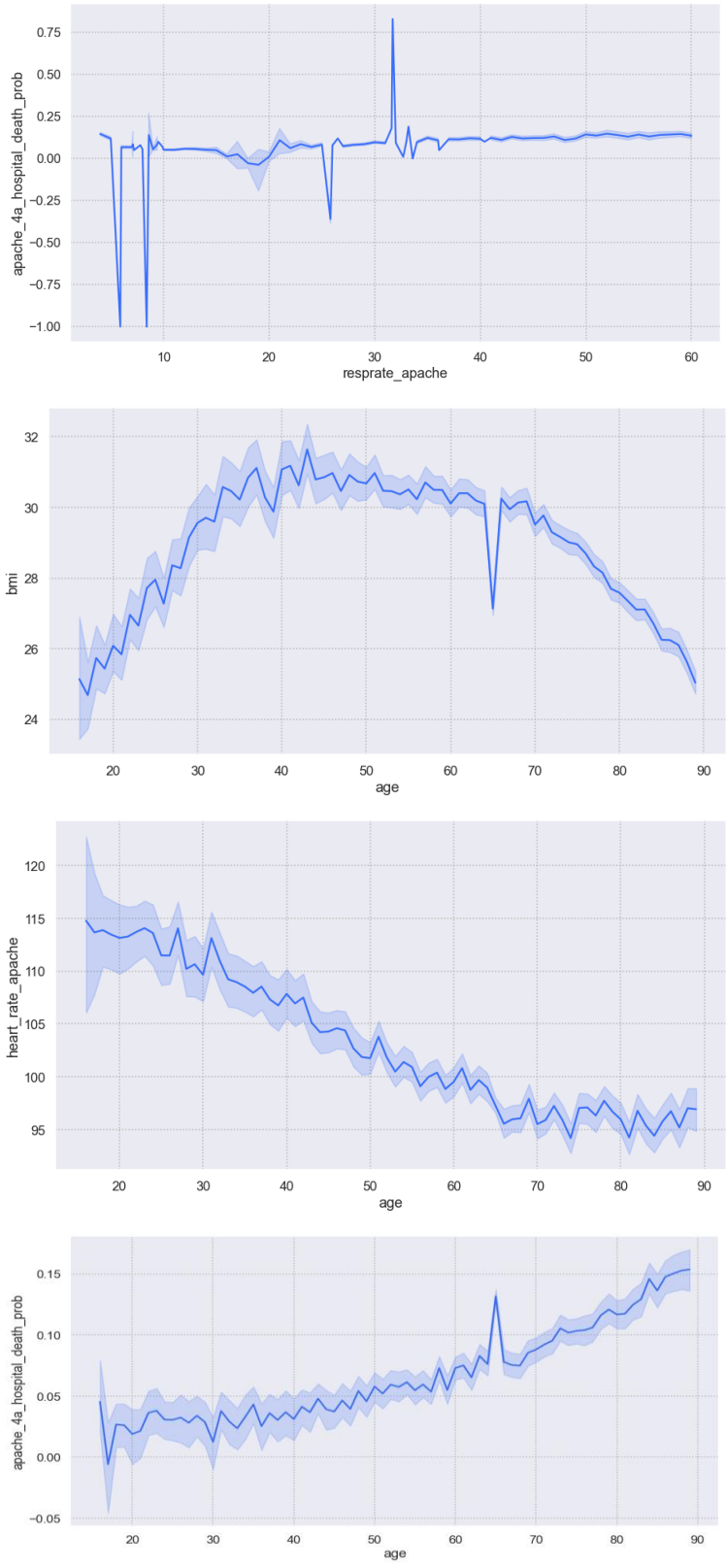





hospital_death
0
1
70000
60000
50000
40000
30000
20000
10000
0
Count
0
1
elective_surgery


gender
M
F
80000
70000
60000
50000
40000
30000
20000
10000
0
Count
0
1
hospital_death

Numerical vs Numerical

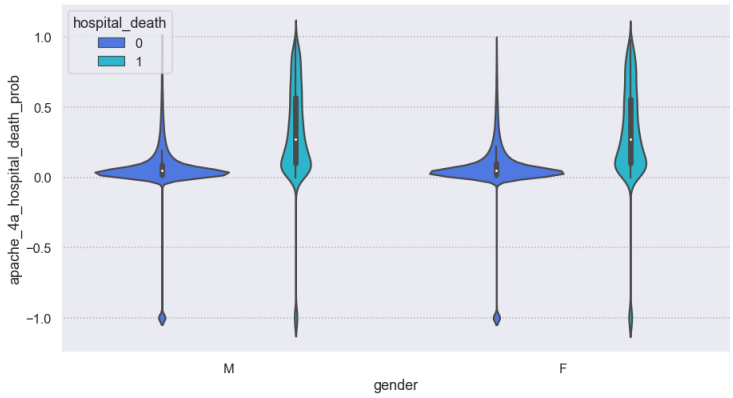


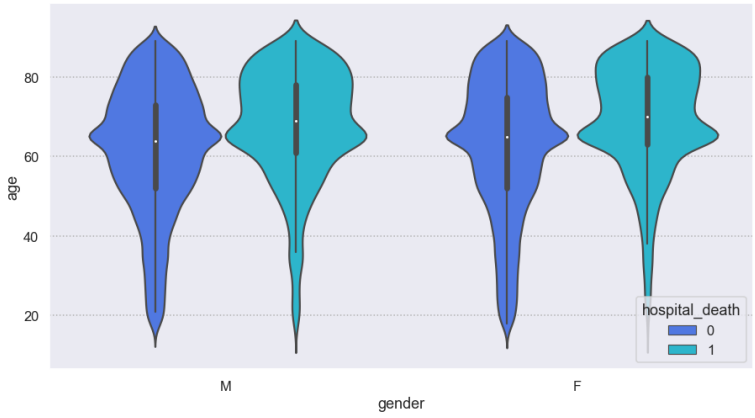
Inferences

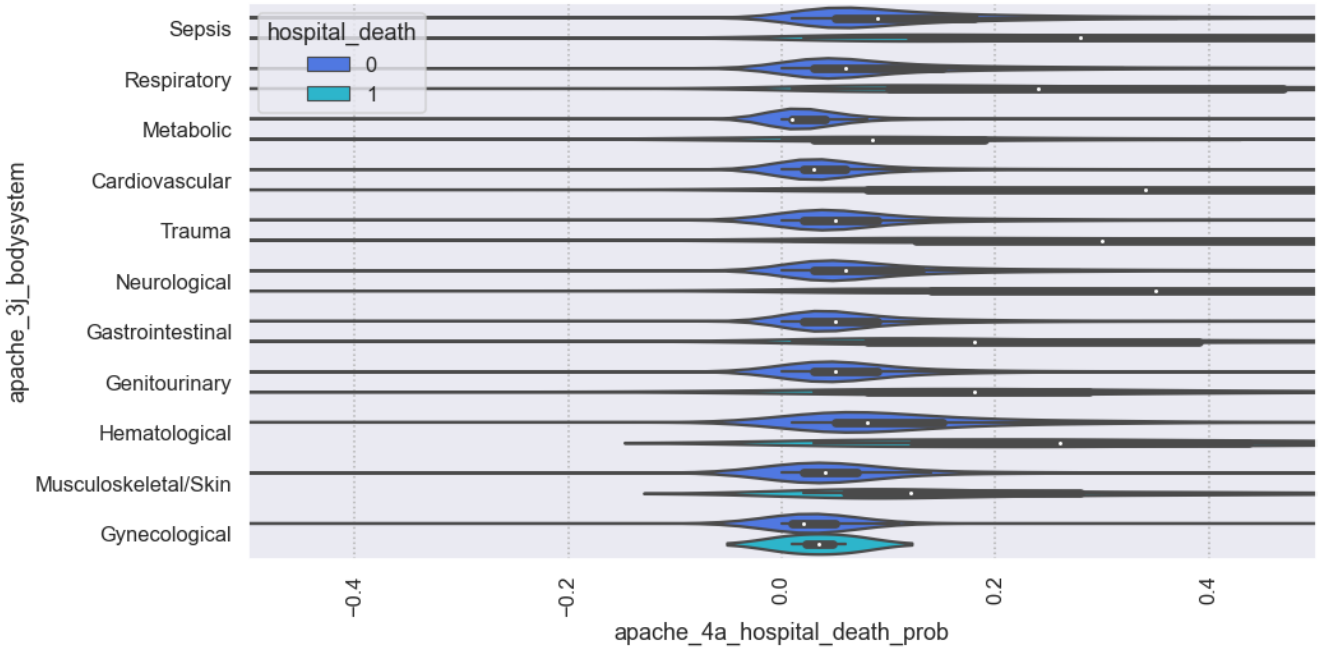
- There’s no difference in the age groups for both male and female patients

- For patients with higher heart rate have a lower chance of survival
- It is seen Accident and Emergency cases have higher chances on both survival and death
- Cardiovascular bodysystem has the similar conclusion as above.
- As age increases, the apache death probability also increases.

Multivariate Analysis



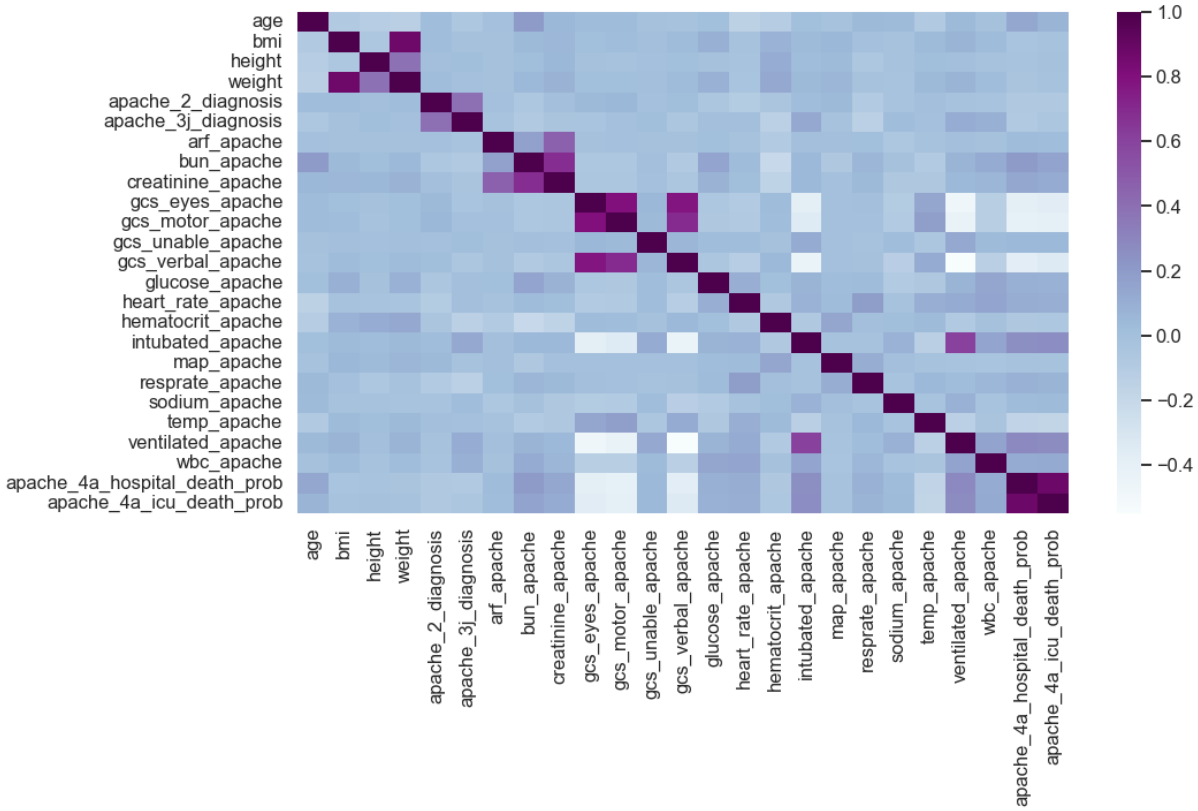




Inferences

- The apache death probability is a bit higher for Female when it comes patients not surviving.
- The common age for both genders for possible death is the same.
- The death probability is higher in the case of Cardiovascular bodysystem.

Multicollinearity



Inferences

There are only a handful of variables that show some collinearity.

Hypothesis testing

Chi-Square Test

The chi-square test is used to check whether the variables are dependent on one another. The categorical variables are compared with the target variable, hospital death. From the table, we see that icu_admit_source and hospital_death are independent of one another, and hospital_death is dependent on the other variables.

Mann-Whitney U Test

The first step that was taken was to check if the data is normally distributes using the Shapiro Wilk test. From the test, we see that all the variables did not have a normal distribution, hence we used the Mann-Whitney test if there are differences between two independent outcomes, i.e. it checks whether the mean of a variables varies for different outcomes of hospital_death. From the results of the test, it is seen that all the variables had differences for different outcomes of hospital_death.

Feature Engineering

Feature Encoding

The function get_dummies from pandas is used to create indicator/dummy attributes except for the target variable hospital_death.

Feature Selection

The variation inflation factor method was used to determine the features involved in multicollinearity. These features were removed from the dataset.

Feature Scaling

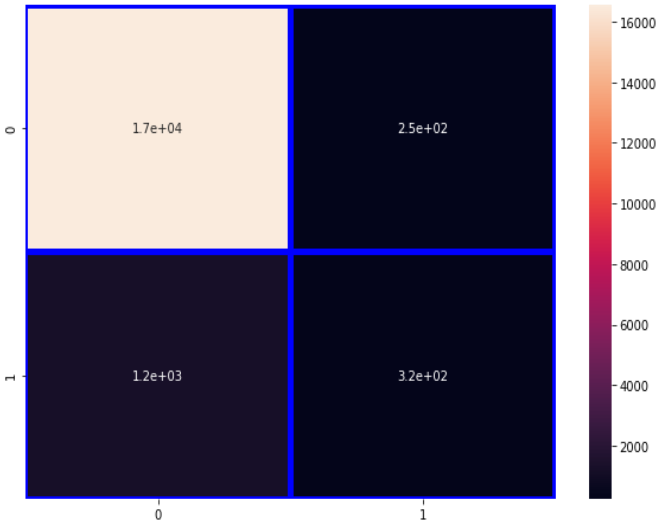
In the outlier treatment, we used Power Transform to reduce the skewness by scaling the data.

Model Building

We first began trying variety of base models and comparing the results to help choose a model giving a good recall score. In the following models, we will see the confusion matrix, classification report and the ROC curve. The data has been split into train and test sets.

Logistic Regression

The model building first begins with logistic regression model. After generating the confusion matrix, we calculated the optimal threshold value using the Youden's Index. After getting the optimal threshold value of 0.51, we generated the following reports once more.

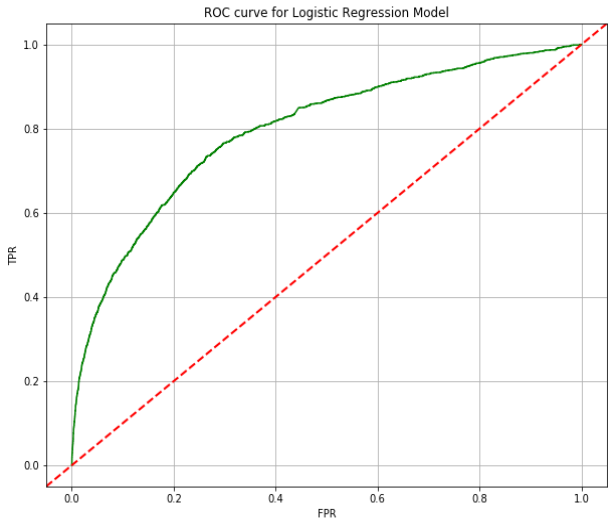


From the confusion matrix, it predicted that 17000 patients will survive, and 320 patients will not survive. However, it incorrectly predicted that 250 patients will not survive and 1200 will survive.

Logit model classification report:

---

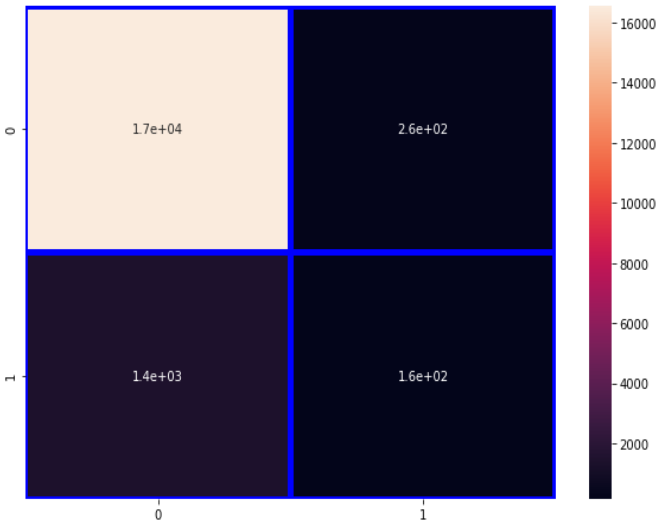
	precision	recall	f1-score	support
0	0.93	0.99	0.96	16796
1	0.56	0.21	0.30	1547
accuracy			0.92	18343
macro avg	0.75	0.60	0.63	18343
weighted avg	0.90	0.92	0.90	18343



Inferences:

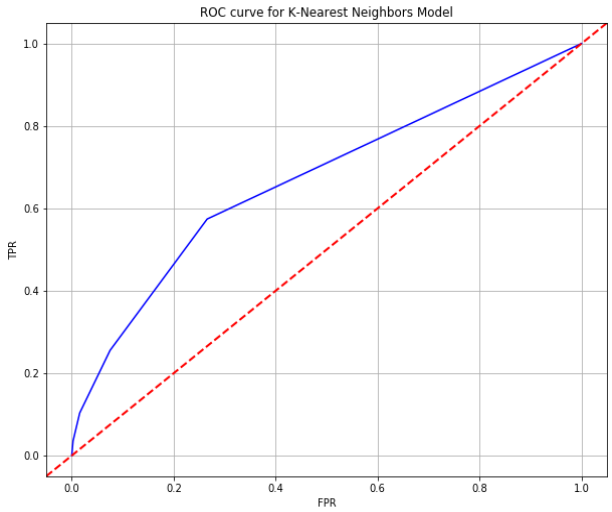
- Cross Entropy for the Logistic Regression Model is 2.77
- ROC AUC Score for the Logistic Regression Model is 79.55
- The Model Accuracy for the Logistic Regression Model is coming out to be around 92%.
- f1 weighted average for the Logistic Regression Model is around 90%.
- Specificity : 98%
- Sensitivity : 20.74%

K-Nearest Neighbor Classification Model



K-Nearest Neighbors Model classification report:

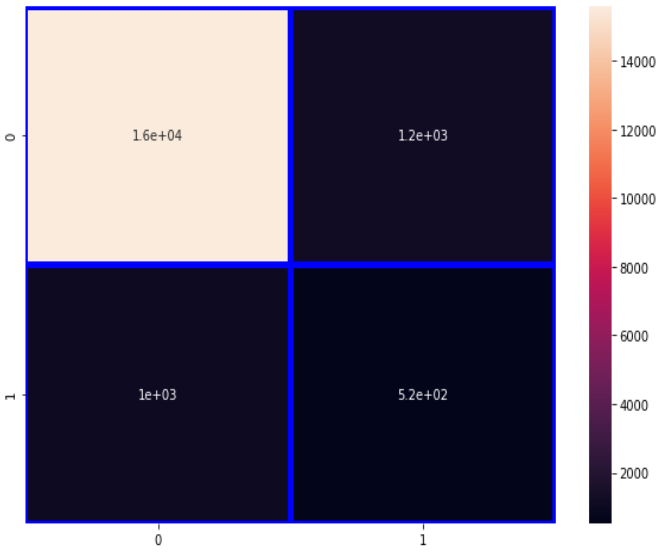
	precision	recall	f1-score	support
0	0.92	0.98	0.95	16796
1	0.38	0.10	0.16	1547
accuracy			0.91	18343
macro avg	0.65	0.54	0.56	18343
weighted avg	0.88	0.91	0.89	18343



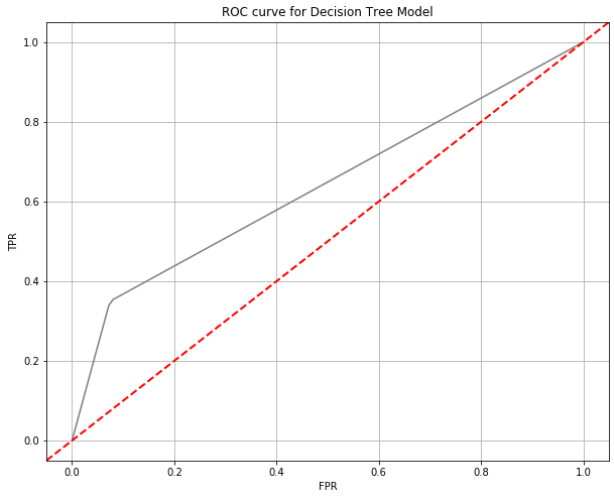
Inferences:

- Cross Entropy for K-Nearest Neighbours Model is 3.10
- ROC AUC Score for the K-Nearest Neighbours Model is 66.84
- The Model Accuracy for the K-Nearest Neighbours Model is coming out to be around 91%.
- f1 weighted avergae for the K-Nearest Neighbours Model is around 89%.
- Specificity : 98.44%
- Sensitivity : 10%

Decision Tree Classification Model



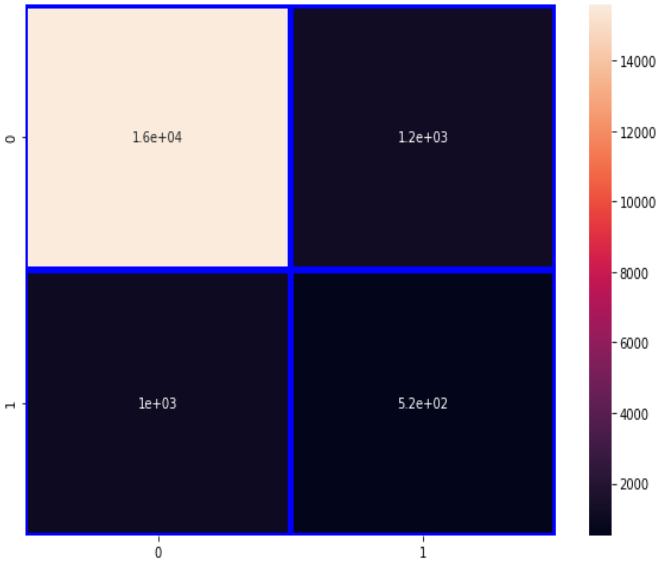
```
Decision Tree Model classification report:
---------------------------------------

              precision    recall  f1-score   support

           0       0.94      0.93      0.93     16796
           1       0.30      0.34      0.32      1547

    accuracy                           0.88     18343
   macro avg       0.62      0.63      0.63     18343
weighted avg       0.88      0.88      0.88     18343
```



Inferences

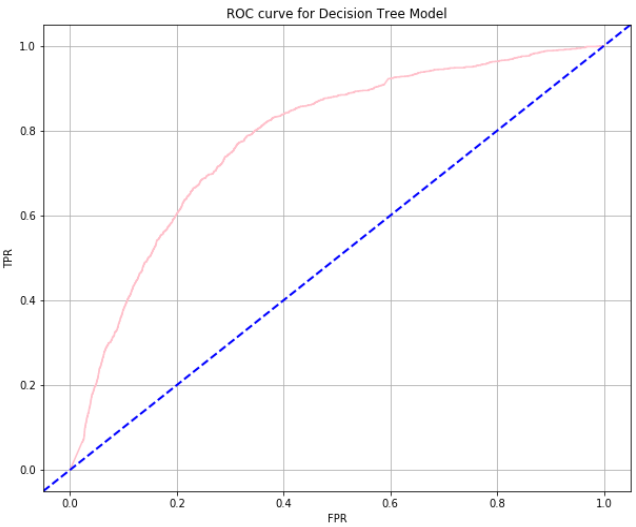
- The Cross entropy for the Decision Tree model is 4.19
- ROC AUC Score for the Decision Tree Model is 63.89
- The Model Accuracy for the Decision Tree Model is coming out to be around 88%.
- f1 weighted avg for the Decision Tree Model is around 88%.
- Specificity : 92%
- Sensitivity : 34%

Naïve Bayes Classification Model



Naive Bayes Model classification report for test data:

	precision	recall	f1-score	support
0	0.95	0.87	0.90	16796
1	0.24	0.47	0.32	1547
accuracy			0.83	18343
macro avg	0.59	0.67	0.61	18343
weighted avg	0.89	0.83	0.86	18343



Base Model Comparison

	Overall Accuracy Score	Accuracy for train data	Accuracy for test data	Specificity	Sensitivity	f1 score weighted avg	ROC AUC Score	Cross Entropy	Bias Error	Variance Error
Logistic Regresion Model	92%	91.66%	91.96%	98.52%	20.74%	90%	79.55	2.77	8.37	42.05
K-Nearest Neighbors Model	91%	93.01%	90.28%	98.44%	10.34%	91%	66.86	3.10	9.12	31.38
Descision Tree Model	88%	99.90%	87.86%	88%	34.19%	88%	63.89	4.19	12.35	42.54
Naive Bayes Model	83%	83.47%	83.24%	92.80%	34%	86%	77.95	5.70	16.51	66.10

Inferences

Overall Accuracy Score:

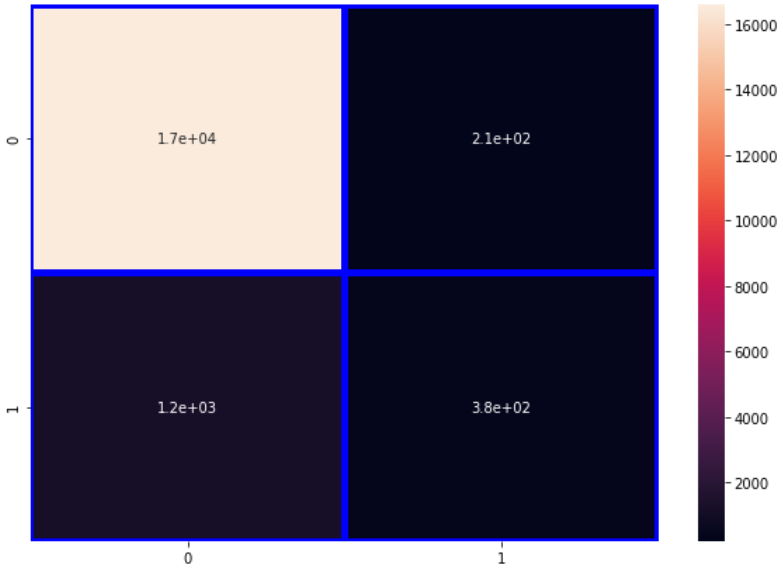
1. Logistic Regression Model has the highest overall accuracy of about 92%.
2. Decision Tree Model yields the lowest overall model accuracy of about 88%.

Overfitting/Underfitting:

1. All the classification models exhibit overfitting of the trained data with respect to the test data.
2. The model accuracy for train data and test data for both Logistic Regression Model and K-Nearest Neighbors Model has very less overfitting.
3. As observed, the model accuracy for train data and test data for the Decision Tree Model has a considerably high difference in accuracies which can be considered a high overfitting condition in comparison to other models.

Model Optimization

Random Forest

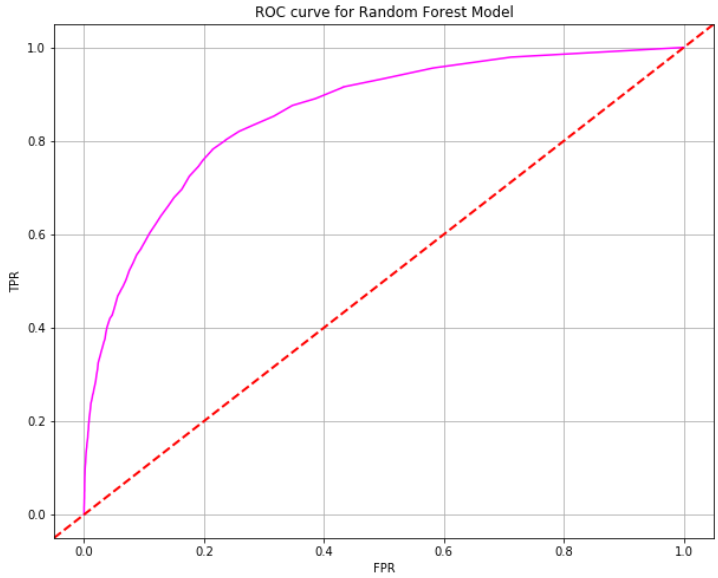


Random Forest Model classification report for train data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	67002
1	1.00	0.99	0.99	6368
accuracy			1.00	73370
macro avg	1.00	0.99	1.00	73370
weighted avg	1.00	1.00	1.00	73370

Random Forest Model classification report for test data:

	precision	recall	f1-score	support
0	0.93	0.99	0.96	16796
1	0.64	0.24	0.35	1547
accuracy			0.92	18343
macro avg	0.79	0.62	0.66	18343
weighted avg	0.91	0.92	0.91	18343



Hyperparameter tunning

A good choice of hyperparameters can really make a model succeed in meeting desired metric value or on the contrary it can lead to a unending cycle of continuous training and optimization. Hence, we have used Grid Search Cross Validation technique for Hyperparameter tunning the models where the Cross Validation method considered is 10-Fold Cross Validation.

Logistic Regression

```
Sensitivity :  26.63
Specificity :  98.06

Classification report of DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_split=8,
                       splitter='random') for resampled train data :
-------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.94      0.98      0.96     67002
           1       0.63      0.31      0.42      6368

    accuracy                           0.92     73370
   macro avg       0.78      0.65      0.69     73370
weighted avg       0.91      0.92      0.91     73370
```

```
Classification report of DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_split=8,
                     splitter='random') for resampled test data :
-------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.94      0.98      0.96     16796
           1       0.56      0.27      0.36      1547

    accuracy                           0.92     18343
   macro avg       0.75      0.62      0.66     18343
weighted avg       0.90      0.92      0.91     18343
```

Random Forest Model

```
Sensitivity :  24.76
Specificity :  98.82

Classification report of RandomForestClassifier() for resampled train data :
-------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     67002
           1       1.00      0.99      0.99      6368

    accuracy                           1.00     73370
   macro avg       1.00      0.99      1.00     73370
weighted avg       1.00      1.00      1.00     73370
```

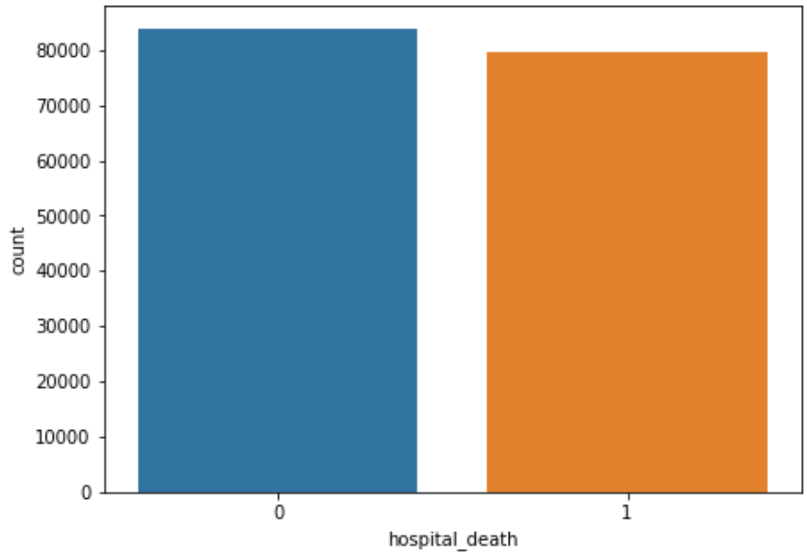
```
Classification report of RandomForestClassifier() for resampled test data :
-------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.93      0.99      0.96     16796
           1       0.66      0.25      0.36      1547

    accuracy                           0.93     18343
   macro avg       0.80      0.62      0.66     18343
weighted avg       0.91      0.93      0.91     18343
```

Imbalanced Target

As seen in univariate analysis, our target variable is imbalanced. Even though it makes sense that there are high survivals than deaths, to improve our model, we use SMOTE analysis it correct the imbalanced data.

SMOTE Analysis

The minority class, i.e., the patients that have not survived, gets oversampled. After doing so, the ratio between survival and death became 51:48.



We performed the same steps of model building, and compared the scores once again.

Logistic Regression

```
LogisticRegression(max_iter=200) Model classification report after resampling for train data:
-------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.77      0.81      0.79     66957
           1       0.79      0.75      0.77     63767

    accuracy                           0.78    130724
   macro avg       0.78      0.78      0.78    130724
weighted avg       0.78      0.78      0.78    130724
```

```
LogisticRegression(max_iter=200) Model classification report after resampling for test data:
-------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.77      0.81      0.79     16841
           1       0.79      0.75      0.77     15841

    accuracy                           0.78     32682
   macro avg       0.78      0.78      0.78     32682
weighted avg       0.78      0.78      0.78     32682


Sensitivity :  0.7452181049176189
Specificity :  0.8115907606436672

AUC score of the LogisticRegression(max_iter=200) Model for resampled test data :  0.8524666031564991

The Cross Entropy score of LogisticRegression(max_iter=200) Model :  7.618645620021943
```

Decision Tree

```
DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_split=8,
                       splitter='random') Model classification report after resampling for train data:
-----------------------------------------------------------------------
```

	precision	recall	f1-score	support
0	0.83	0.83	0.83	66957
1	0.82	0.82	0.82	63767
accuracy			0.83	130724
macro avg	0.83	0.83	0.83	130724
weighted avg	0.83	0.83	0.83	130724

```
DecisionTreeClassifier(criterion='entropy', max_depth=10, min_samples_split=8,
                       splitter='random') Model classification report after resampling for test data:
-----------------------------------------------------------------------
```

	precision	recall	f1-score	support
0	0.82	0.82	0.82	16841
1	0.81	0.81	0.81	15841
accuracy			0.82	32682
macro avg	0.82	0.82	0.82	32682
weighted avg	0.82	0.82	0.82	32682

Random Forest

```
RandomForestClassifier() Model classification report after resampling for train data:
-----------------------------------------------------------------------
```

	precision	recall	f1-score	support
0	1.00	0.99	1.00	66957
1	0.99	1.00	1.00	63767
accuracy			1.00	130724
macro avg	1.00	1.00	1.00	130724
weighted avg	1.00	1.00	1.00	130724

```
RandomForestClassifier() Model classification report after resampling for test data:
-----------------------------------------------------------------------
```

	precision	recall	f1-score	support
0	0.95	0.92	0.94	16841
1	0.92	0.95	0.93	15841
accuracy			0.93	32682
macro avg	0.93	0.93	0.93	32682
weighted avg	0.93	0.93	0.93	32682

Inference

From the comparisons above, it is seen that precision and recall has improved for Random Forest compared to the one built with the original data,

Feature Extraction

The important features were determine with the help of Recursive Feature Elimination using the Random forest model. The following features were selected

- map_apache
- apache_4a_icu_death_prob
- apache_2_diagnosis
- apache_3j_diagnosis
- apache_4a_hospital_death_prob
- bun_apache
- creatinine_apache
- wbc_apache
- gcs_verbal_apache
- glucose_apache
- diabetes_mellitus
- resprate_apache

Inference

- The Sensitivity and Specificity of both Logistic Regression model and Decision Tree model has increased significantly from the base models resulting in an improved weighted-avgerage of Precision and Recall rates.
- The Decision Tree Model has a better overall performance than the Logistic Model for the new dataset obtained after recursive feature elimination and resampling the data.
- The Random Forest Model has not shown any major improvement from the previous model but has the best overall performance as compared to the Logistic Regression and Decision Tree model.
- Random forest has been considered.

Bagging and Boosting Algorithms

AdaBoost Model

```
Classification report of AdaBoostClassifier() for resampled train data :
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.87      0.84      0.86     66957
           1       0.84      0.87      0.85     63767

    accuracy                           0.85    130724
   macro avg       0.85      0.86      0.85    130724
weighted avg       0.86      0.85      0.85    130724
```

```
Classification report of AdaBoostClassifier() for resampled test data :
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.87      0.84      0.85     16841
           1       0.84      0.86      0.85     15841

    accuracy                           0.85     32682
   macro avg       0.85      0.85      0.85     32682
weighted avg       0.85      0.85      0.85     32682
```

```
AUC Score for AdaBoostClassifier() for resampled test data :
0.9277744034192948

The Cross Entropy score of AdaBoostClassifier() Model :  5.093906234177455

Mean Score :  0.8552140562069261
Bias error :  14.478594379307385
Variance error :  0.27860125804568026
```

GradientBoost Model

```
Classification report of GradientBoostingClassifier() for resampled train data :
--------------------------------------------------------------------------------
              precision    recall  f1-score   support

           0       0.90      0.87      0.88     66957
           1       0.87      0.90      0.88     63767

    accuracy                           0.88    130724
   macro avg       0.88      0.88      0.88    130724
weighted avg       0.88      0.88      0.88    130724
```

```
Classification report of GradientBoostingClassifier() for resampled test data :
--------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.90      0.87      0.89     16841
           1       0.87      0.90      0.88     15841

    accuracy                           0.88     32682
   macro avg       0.88      0.88      0.88     32682
weighted avg       0.88      0.88      0.88     32682


AUC Score for GradientBoostingClassifier() for resampled test data :
0.9510644196706552

The Cross Entropy score of GradientBoostingClassifier() Model :  4.000092741186987

Mean Score :  0.8840075112298973
Bias error :  11.599248877010272
Variance error :  0.3036704502786237
```

XGBoost Model

```
Classification report of XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, eval_metric='logloss',
              gamma=0, gpu_id=-1, importance_type='gain',
              interaction_constraints='', learning_rate=0.300000012,
              max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=100, n_jobs=0,
              num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=1, subsample=1, tree_method='exact',
              validate_parameters=1, verbosity=None) for resampled train data :
--------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.95      0.97      0.96     66957
           1       0.97      0.94      0.96     63767

    accuracy                           0.96    130724
   macro avg       0.96      0.96      0.96    130724
weighted avg       0.96      0.96      0.96    130724
```

```
Classification report of XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, eval_metric='logloss',
              gamma=0, gpu_id=-1, importance_type='gain',
              interaction_constraints='', learning_rate=0.300000012,
              max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
              monotone_constraints='()', n_estimators=100, n_jobs=0,
              num_parallel_tree=1, random_state=0, reg_alpha=0, reg_lambda=1,
              scale_pos_weight=1, subsample=1, tree_method='exact',
              validate_parameters=1, verbosity=None) for resampled test data :
--------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.94      0.96      0.95     16841
           1       0.96      0.93      0.95     15841

    accuracy                           0.95     32682
   macro avg       0.95      0.95      0.95     32682
weighted avg       0.95      0.95      0.95     32682
```

Model Comparison

	Overall Accuracy	Accuracy for train data	Accuracy for test data	Specificity	Sensitivity	f1 score weighted avg	ROC AUC Score	Bias Error	Variance Error	Cross Entropy
Random Forest Model	95%	99.60%	93.41%	92.08%	94.82%	93%	0.98	6.76	0.27	2.27
AdaBoost Model	86%	86%	86%	84.82%	86.29%	86%	0.93	14.15	0.24	4.99
GradientBoost Model	89%	89%	88%	87.06%	89.65%	89%	0.95	11.46	0.30	4.03
XGBoost Model	96%	95%	95%	96.24%	93.31%	96%	0.98	5.17	0.22	1.78

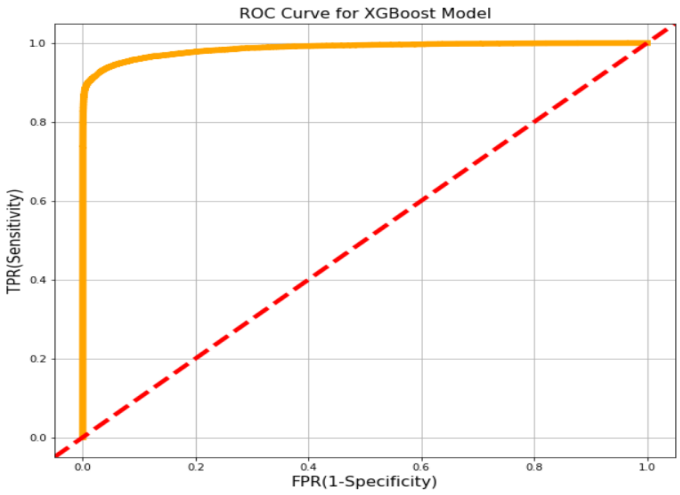
Based on the comparison we can conclude that:

- XGBoost Model and Random Forest Model have the approximately overall accuracy of 95% on test data.
- XGBoost has less overfitting of train data as compared to Random Forest Model.
- XGBoost has a slightly lesser Bias error and variance error than Random Forest Model.
- f1-score weighted for XGBoost Model: 96%
- AUC score for XGBoost Model: 0.98

Final Model

From the comparisons, we have decided to use XGBoost Model. The following are it evaluation metrics

```
Sensitivity :  93.3
Specificity :  96.16

Classification report of XGBoost Model for resampled train data :
--------------------------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.95      0.97      0.96     66957
           1       0.97      0.94      0.96     63767

    accuracy                           0.96    130724
   macro avg       0.96      0.96      0.96    130724
weighted avg       0.96      0.96      0.96    130724


Classification report of XGBoost Model for resampled test data :
----------------------------------------------------------------

              precision    recall  f1-score   support

           0       0.94      0.96      0.95     16841
           1       0.96      0.93      0.95     15841

    accuracy                           0.95     32682
   macro avg       0.95      0.95      0.95     32682
weighted avg       0.95      0.95      0.95     32682
```



```
AUC Score for XGBoost Model for resampled test data :
0.9861564517690254

The Cross Entropy score of XGBoost Model :  1.8050531614357996

Mean Score :  0.9481044004053688
Bias error :  5.18955995946312
Variance error :  0.1870198716707829
```

Conclusion

We have decided on XGBoost Model as our final model. The important features gives some idea on what factors to look at more closely when determining the survival rate of the patient. Some of the features are diabetes, respirate, glucose, Apache diagnosis. The limitation is that there it is still not very precision in determining the survival rate but out of all the models, XGBoost showed the best result.