USC Viterbi
School of Engineering

90-Day Mortality Prediction for ICU Trauma Patients

Group 1
Sanath Sridhar, Rithik Mandipati, Tian Cui, Jie Gu, ZiMin Zhu

OUTLINE

1. Introduction
2. Dataset
3. Literature review
4. Methodology
5. Result
6. Conclusion

USC Viterbi
School of Engineering

INTRODUCTION

INTRODUCTION

Objective: The aim of this study was to develop and validate a machine learning-based predictive models that predicts 90-day mortality in ICU trauma patients.

Trauma patients admitted to ICU face life threatening situation and predicting their mortality is very important to health care professionals in evaluating a patient's treatment and also allocating the resources for the hospital during dire situations.

Our study is based on a retrospective cohort study of a critical care database, Medical Information Mart for Intensive Care IV. This database includes demographic, vital signs, laboratory tests, fluid balance, and life state chart events; Document International Classification of Diseases and Ninth Revision (ICD-9) Code

Machine learning (ML) “learns” models from past data to predict future data.

In this study, we will be training our data on respective models namely XGBoost, logistic regression, random forest, AdaBoost, multilayer perceptron neural networks (MLP), support vector machine (SVM), Light GBM, KNN, and Gaussian Naive Bayes (GNB) machine learning models. The best performing model metrics will be used for further post processing techniques.

Challenges in data processing.

Heterogeneous Data: The MIMIC-III dataset contains diverse data types, including vital signs, lab results, medications, and clinical notes. Integrating and making sense of this heterogeneity is a significant challenge.

Imbalanced Data: Mortality events are relatively rare compared to patient survival. Imbalanced datasets can lead to biased models that prioritize the majority class.

Techniques like over sampling , class weighting will be used to tackle this issue.

Feature Engineering: Identifying the most relevant features or variables for predicting mortality is essential. This is done by using recursive feature elimination (RFE) algorithm based on Shapely additive explanation(SHAP).

USC Viterbi School of Engineering

LITERATURE REVIEW

LITERATURE REVIEW

Study	Method
S. Yang et al (2023)	Models: XGBoost, logistic regression, random forest, AdaBoost, MLP, GNB, SVM, KNN, and LightGBM The AUCs of XGBoost was1.00 The accuracy of XGBoost was 0.991
J. Li et al (2023)	The AUROC of XGBoost and logistic regression were 0.83 and 0.88 at different prediction windows
K. Pang et al (2022)	The AUCs of XGBoost, logistic regression, SVM, and decision tree were 0.918, 0.872, 0.872 and 0.852
Y. Deng et al (2022)	The AUC of recurrent neural network was 0.87

- S. Yang, L. Cao, Y. Zhou, and C. Hu, “A retrospective cohort study: Predicting 90-day mortality for ICU trauma patients with a machine learning algorithm using XGBoost using MIMIC-III database,” Journal of Multidisciplinary Healthcare, vol. Volume 16, pp. 2625–2640, Sep. 2023. doi:10.2147/jmdh.s416943
- K. Pang, L. Li, W. Ouyang, X. Liu, and Y. Tang, “Establishment of ICU Mortality Risk Prediction Models with Machine Learning Algorithm Using MIMIC-IV Database,” Diagnostics, vol. 12, no. 5, p. 1068, Apr. 2022, doi: 10.3390/diagnostics12051068.
- J. Li, F. Xi, W. Yu, C. Sun, and X. Wang, “Real-time prediction of sepsis in critical trauma patients: Machine Learning–Based Modeling Study,” JMIR Formative Research, vol. 7, 2023. doi:10.2196/42452
- Y. Deng et al., “Explainable time-series deep learning models for the prediction of mortality, prolonged length of stay and 30-day readmission in intensive care patients,” Frontiers in Medicine, vol. 9, 2022. doi:10.3389/fmed.2022.933037

USC Viterbi
School of Engineering

Dataset

DATA EXTRACTION

We have 5556 patients in original dataset and more than 10,000,000 records.

Conditions:

1. Get the last admission for every patient
2. Get trauma patients
3. Get patients who stay in icu more than 1 days

Now we have 2107 patients.

Basic info

Variable Name	Type	Comments
admission_age	INTEGER	age of admission
gender	STRING	gender description
race	STRING	race description
admission_type	STRING	admission_type include 9 types
icu_length	DATETIME	length in icu
newest_admittime	DATETIME	last admit date for a patient
los_hospital	DATETIME	Length in hospital
dod (response variable)	DATE	patient death date
mortality(response variable) (added in the data clean part)	Boolean	If this patient death

Laboratory test

Variable Name	Type	Comments/ admission
avg_rdw	FLOAT	Red cell distribution width
avg_calcium	FLOAT	Calcium
avg_sodium	FLOAT	Sodium
avg_inr	FLOAT	international normalized ratio
avg_glucose	FLOAT	Glucose
avg_chloride	FLOAT	Chloride
avg_creatinine	FLOAT	Creatinine
avg_bicarbonate	FLOAT	Bicarbonate
avg_aniongap	FLOAT	Anion Gap

Vital signs

Variable Name	Type	Comments/ admission
avg_heart_rate	FLOAT	heart rate
avg_spo2	FLOAT	Pulse Oxygen Saturation (SpO2): This is frequently represented as SpO2.
avg_temperature	NUMERIC	Body Temperature
avg_resp_rate	FLOAT	Respiratory Rate (RR): Commonly abbreviated as RR.
avg_dbp	FLOAT	Diastolic Blood Pressure (DBP): Often abbreviated as DBP. When blood pressure is provided as a reading (e.g., 120/80 mmHg), the first number represents the systolic pressure and the second represents the diastolic pressure.
avg_sbp	FLOAT	Systolic Blood Pressure (SBP): Commonly abbreviated as SBP.

Scores

Variable Name	Type	Comments/admission
avg_sapsii	INTEGER	The SAPS II score ranges from 0 to 163, with higher scores indicating a higher risk of mortality. The score is calculated using 17 variables which include: Age Type of admission (e.g., scheduled surgical, unscheduled surgical, medical) Three neurological variables (Glasgow Coma Score) Twelve physiological variables (e.g., heart rate, systolic blood pressure, temperature, pH, PaO2, sodium, potassium, creatinine, urea, total bilirubin, white blood cell count, and bicarbonate levels) Chronic diseases (e.g., metastatic cancer, hematologic malignancy, AIDS) It's important to note:
avg_gcs_score	INTEGER	The Glasgow Coma Scale (GCS) is a neurological scale that aims to give a reliable, objective way of recording the conscious state of a person.
avg_SOFA score	INTEGER	The Sequential Organ Failure Assessment (SOFA) Score is used to track a person's status during the stay in an Intensive Care Unit (ICU) to determine the extent of a person's organ function or rate of failure.

DATA EXTRACTION

Variables(29):

Category(8): subject_id,gender,stay_id,admission_type,newest_admittime, admission_age,race, dod(if is died, response variables),mortality

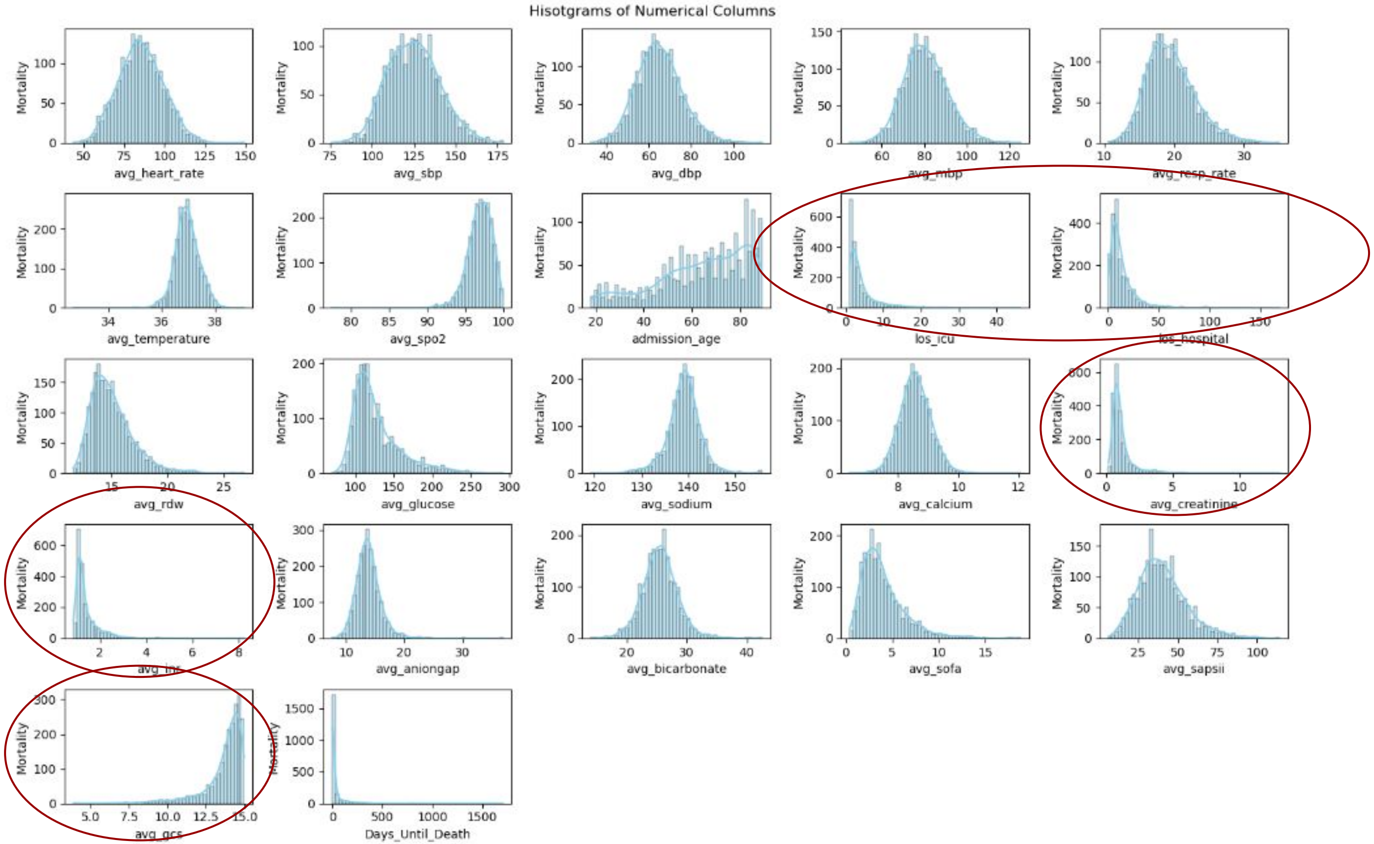
Numeric(21): avg_heart_rate, avg_sbp, avg_dbp,avg_mbp, avg_resp_rate, avg_temperature, avg_spo2,icu_length, los_icu, los_hospital, avg_rdw, avg_glucose, avg_sodium, avg_calcium, avg_creatinine, avg_inr, avg_aniongap, avg_bicarbonate,avg_sapii,avg_sofa,avg_gcs

4 groups: Basic info(11), Laboratory test(9), Vital signs(6), Scores(3)

Methodolgy

Univariate Analysis

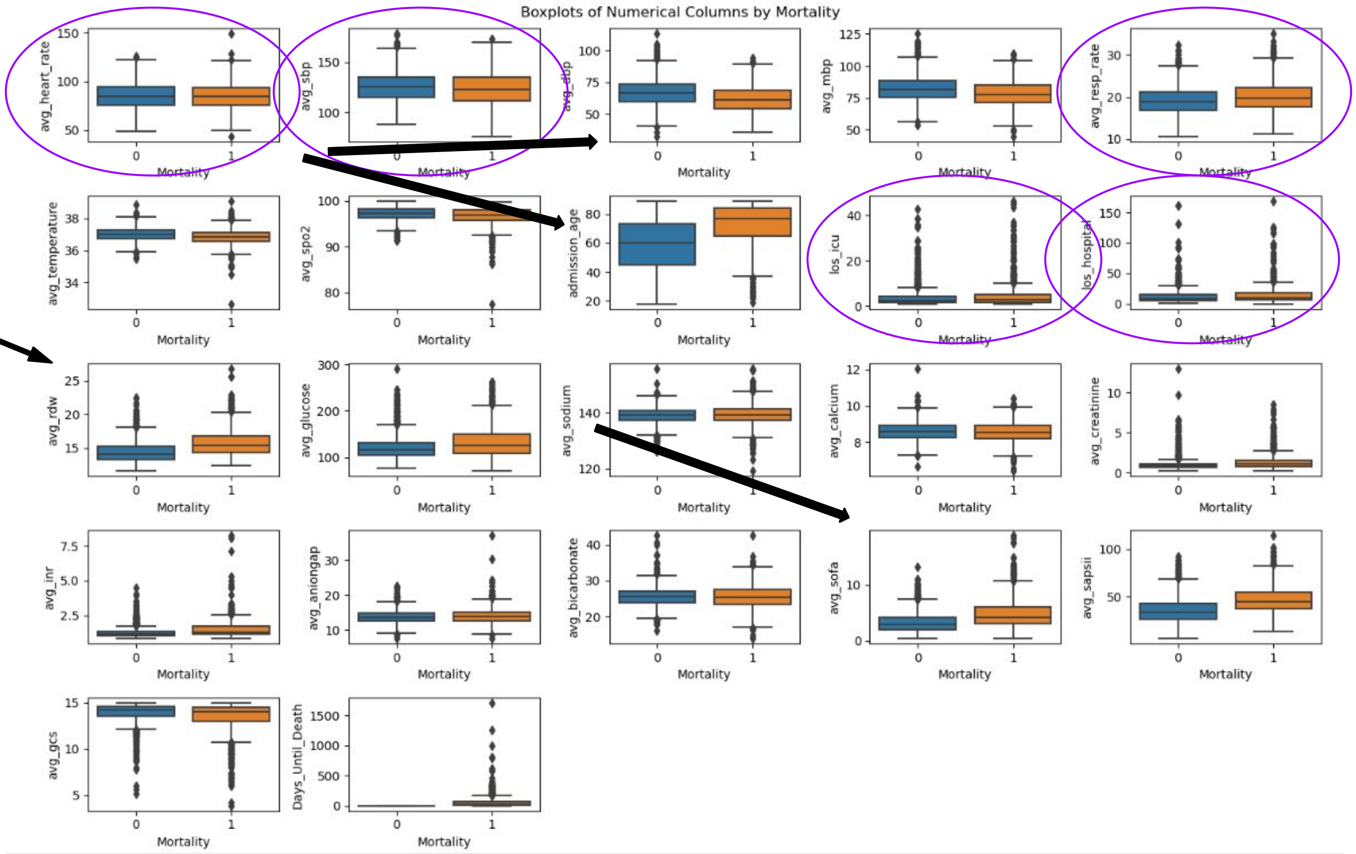
Most columns have high skew. And 6 of them (||>2): avg_gcs, avg_inr, avg_cr
tinine, los_hospital, los_i



Bivariate Analysis

Admission_age, avg_sofa, avg_rdw,avg_dbp seem to have high correlation with mortality

avg_Heart_rate, avg_sbp, avg_resp_rate, Los_icu, Hospital_length, avg_bicarbonate seems to have low correlation with mortality



Z-score Analysis

- Conducting z-tests between feature distributions of patients who survived and those who didn't survive to reveal correlations with mortality

- 19 out of 23 numerical columns have p-value<0.01 indicating significant that there exists significant statistical difference between the distributions of both classes

	Variable Name	Z Statistic	P Value	P Value < 0.01
0	avg_heart_rate	-0.868992	3.848517e-01	False
1	avg_sbp	-3.573991	3.515815e-04	True
2	avg_dbp	-11.375453	5.541436e-30	True
3	avg_mbp	-9.216277	3.075940e-20	True
4	avg_resp_rate	6.013919	1.810907e-09	True
5	avg_temperature	-9.304624	1.344672e-20	True
6	avg_spo2	-5.674608	1.390068e-08	True
7	admission_age	19.478590	1.668118e-84	True
8	icu_length	3.544757	3.929751e-04	True
9	los_icu	3.545249	3.922427e-04	True
10	los_hospital	4.172844	3.008209e-05	True
11	avg_rdw	16.841214	1.217145e-63	True
12	avg_glucose	7.765760	8.115699e-15	True
13	avg_sodium	2.568677	1.020877e-02	False
14	avg_calcium	-1.346789	1.780481e-01	False
15	avg_creatinine	9.205775	3.392133e-20	True
16	avg_inr	10.411594	2.195146e-25	True
17	avg_aniongap	3.456709	5.468153e-04	True
18	avg_bicarbonate	-0.288077	7.732875e-01	False
19	avg_sofa	15.142067	8.548448e-52	True
20	avg_sapsii	17.849829	2.899231e-71	True
21	avg_gcs	-7.551119	4.315335e-14	True
22	Days_Until_Death	19.444393	3.250998e-84	True

Also in the conclusion of bivariate part

Data Cleaning

- Drop columns stay_id, subject_id
- Log transform avg_inr, avg_creatinine, avg_calcium, icu_length, avg_glucose, avg_aniongap, los_hospital, los_icu, avg_gcs, exponential transform avg_sofa score
- Clip los_hospital, los_icu at 0.99 quantile to reduce skew
- One hot encode gender, race, admission type
- Min Max scaling of numeric variables

USC Viterbi School of Engineering

Result

Models Used

- Models Chosen: Decision Tree, Logistic_regression ,AdaBoost, Random Forest Classifier, XGBoost Classifier, SVM, KNN, Naive Bayes, LightGBM
- Metrics of Evaluation: Validation Accuracy, Precision ,Recall, F1 Score, AUC-ROC

	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	decision_tree	0.663391	0.601227	0.576471	0.588589	0.651104
1	logistic_regression	0.759214	0.725000	0.682353	0.703030	0.748349
2	adaboost	0.712531	0.658683	0.647059	0.652819	0.703276
3	random_forest	0.771499	0.736196	0.705882	0.720721	0.762224
4	xgboost	0.773956	0.737805	0.711765	0.724551	0.765165
5	svm	0.759214	0.716867	0.700000	0.708333	0.750844
6	knn	0.705160	0.678571	0.558824	0.612903	0.684475
7	naive_bayes	0.479115	0.443850	0.976471	0.610294	0.549417
8	lightgbm	0.751843	0.719745	0.664706	0.691131	0.739526

Deep Learning approach

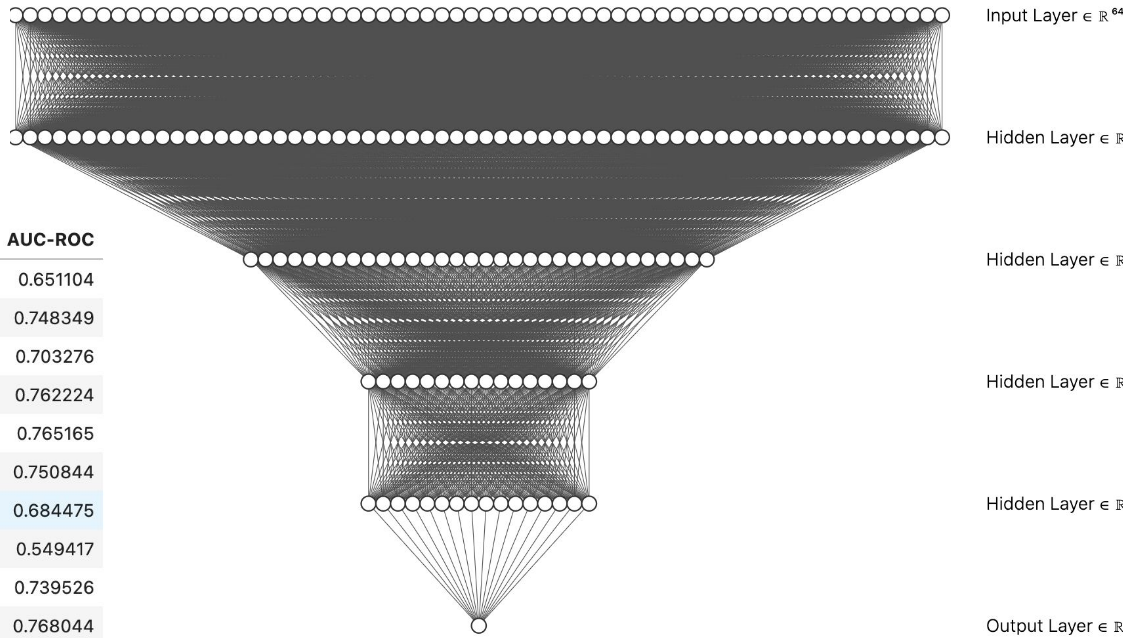
Loss Function : Cross Entropy Loss

Optimizer:Adam

Learning Rate : 0.01

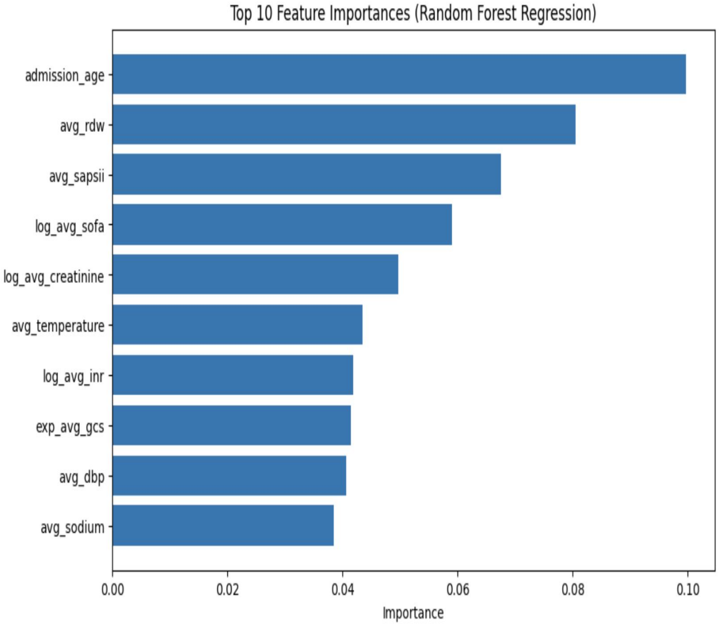
Hidden Layers: 4

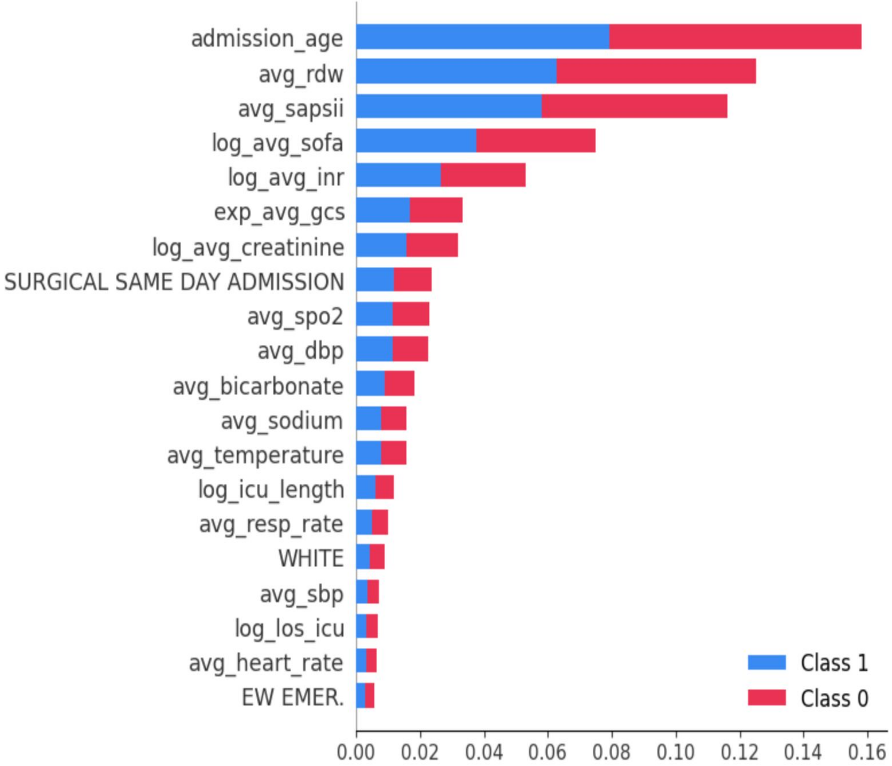
	Model	Accuracy	Precision	Recall	F1 Score	AUC-ROC
0	decision_tree	0.663391	0.601227	0.576471	0.588589	0.651104
1	logistic_regression	0.759214	0.725000	0.682353	0.703030	0.748349
2	adaboost	0.712531	0.658683	0.647059	0.652819	0.703276
3	random_forest	0.771499	0.736196	0.705882	0.720721	0.762224
4	xgboost	0.773956	0.737805	0.711765	0.724551	0.765165
5	svm	0.759214	0.716867	0.700000	0.708333	0.750844
6	knn	0.705160	0.678571	0.558824	0.612903	0.684475
7	naive_bayes	0.479115	0.443850	0.976471	0.610294	0.549417
8	lightgbm	0.751843	0.719745	0.664706	0.691131	0.739526
9	Neural Network	0.771499	0.717514	0.747059	0.731988	0.768044



Shapley Values and Feature Importances

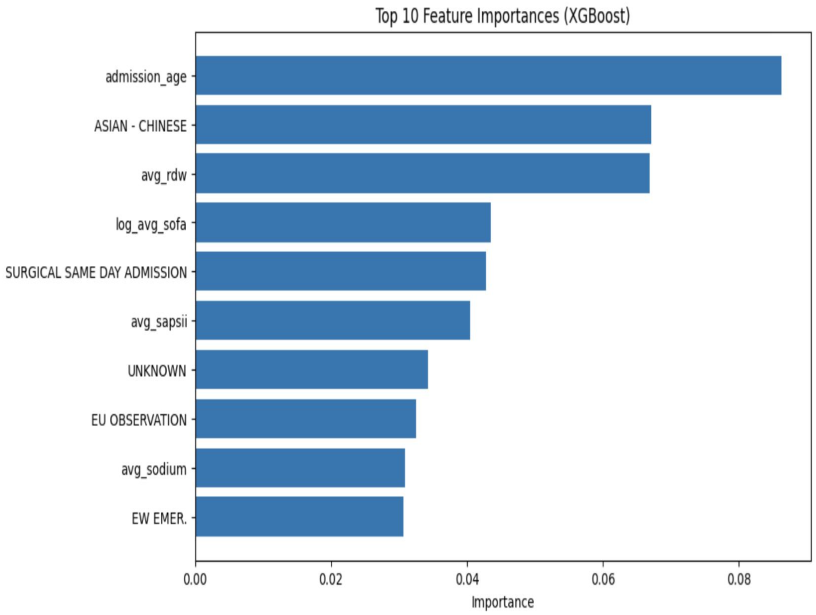
Shapley Values and Top 10 features of Random Forest Model

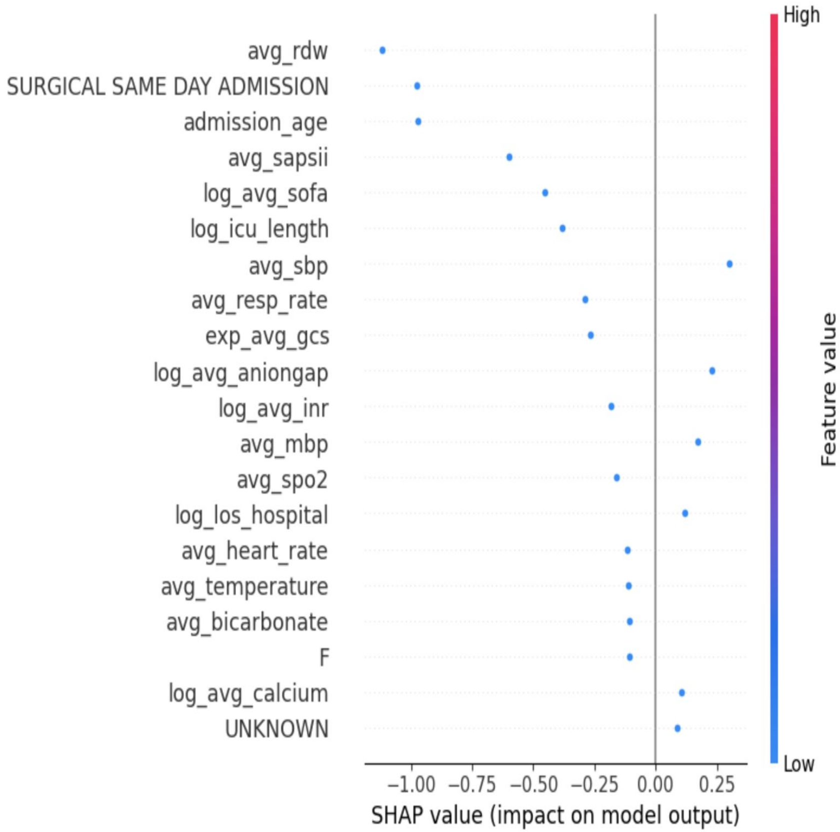




Shapley Values and Feature Importances

Shapley Values and Top 10 features of XGBoost Model





Conclusion

- The accuracy was observed highest in the models randomforest , xgboost and neural networks with accuracy of around 77%.

- The AUC and F1 score metric was observed highest in the the Neural networks with values 76.8% and 73.2% respectively.

- The AUC and F1 values of XGboost are also comparable with the NN with values 76.5 % and 72.4% respectively

Thank You !