Patient Survival Prediction

Capstone Group 9

Bhupendra Mewada
Madhavkumar
Tumu Adithya
Ajinkya Dandgavhal

Project Mentor

Ms. Vibha Santhanam

Problem Definition

Business Problem :

- To develop a model which can predict whether a patient will survive or not based on certain health markers which are recorded when the patient is hospitalized.

Importance of the Business Problem :

- Survival Prediction after hospitalization is important for doctors and the patient or their family members.
- It helps doctors to understand the degree of severity and accordingly plan the medical treatment.
- Helps the medical staff to prioritize patients based on fatality of medical condition.
- Patients and their families can get adequate time to make necessary arrangements.
- Leads to timely prevention and treatment and worse treatment decisions (over-treatment or late palliative care) can be avoided.

Current Practices :

- Survival ability predicted using APACHE-II scoring system (Acute Physiology and Chronic Health Evaluation II). Developed in 1981.
- Severity-of-disease classification system. Higher scores - higher risk of death.
- Applied within 24 hrs of admission of a patient to an ICU. Integer score from 0 to 71.
- Score calculated from 12 physiological parameters like body temp, blood pH, heart rate, arterial pressure, serum sodium, creatinine, WBC count, age etc.
- Also, prediction is made by specialist doctors based on their medical experience.

Value Addition :

- Major gap which was observed during the topic survey phase is that there are many studies regarding survival prediction for certain specific diseases like cancer, sepsis, brain tumor, but very few studies which include many diseases together.
- We wish to bridge this gap by including as many diseases which can adversely affect a patient's survival. These diseases are namely liver cirrhosis, cardiovascular, respiratory failure, trauma, sepsis, metabolic, neurologic and gastrointestinal diseases.
- Also, using the model, time required to predict the survival status is very less as compared to that of APACHE-II Scoring System (24 hrs).

Approach & Dataset Considered

Approach :

- To study the health markers which are recorded once a patient is hospitalized.
- Understand how each health marker can potentially indicate the severity of the patient's condition.
- Study how different markers interact with each other and their effect on the survival ability of the patient using EDA, Statistical Analysis, Visualizations.
- Identify most important markers and build a model using these markers.

Dataset Considered :

- Project domain - Healthcare Analytics | Dataset Source - Kaggle.com | Dataset Link : https://www.kaggle.com/mitishaagarwal/patient
- No. of records : 91,713 | No. of columns : 85 | No. of numeric columns : 63 | No. of categorical columns : 22

0	encounter_id	29	d1_diasbp_max	57	h1_mbp_noninvasive_max
1	patient_id	30	d1_diasbp_min	58	h1_mbp_noninvasive_min
2	hospital_id	31	d1_diasbp_noninvasive_max	59	h1_resprate_max
3	age	32	d1_diasbp_noninvasive_min	60	h1_resprate_min
4	bmi	33	d1_heartrate_max	61	h1_spo2_max
5	elective_surgery	34	d1_heartrate_min	62	h1_spo2_min
6	ethnicity	35	d1_mbp_max	63	h1_sysbp_max
7	gender	36	d1_mbp_min	64	h1_sysbp_min
8	height	37	d1_mbp_noninvasive_max	65	h1_sysbp_noninvasive_max
9	icu_admit_source	38	d1_mbp_noninvasive_min	66	h1_sysbp_noninvasive_min
10	icu_id	39	d1_resprate_max	67	d1_glucose_max
11	icu_stay_type	40	d1_resprate_min	68	d1_glucose_min
12	icu_type	41	d1_spo2_max	69	d1_potassium_max
13	pre_icu_los_days	42	d1_spo2_min	70	d1_potassium_min
14	weight	43	d1_sysbp_max	71	apache_4a_hospital_death_prob
15	apache_2_diagnosis	44	d1_sysbp_min	72	apache_4a_icu_death_prob
16	apache_3j_diagnosis	45	d1_sysbp_noninvasive_max	73	aids
17	apache_post_operative	46	d1_sysbp_noninvasive_min	74	cirrhosis
18	arf_apache	47	d1_temp_max	75	diabetes_mellitus
19	gcs_eyes_apache	48	d1_temp_min	76	hepatic_failure
20	gcs_motor_apache	49	h1_diasbp_max	77	immunosuppression
21	gcs_unable_apache	50	h1_diasbp_min	78	leukemia
22	gcs_verbal_apache	51	h1_diasbp_noninvasive_max	79	lymphoma
23	heart_rate_apache	52	h1_diasbp_noninvasive_min	80	solid_tumor_with_metastasis
24	intubated_apache	53	h1_heartrate_max	81	apache_3j_bodysystem
25	map_apache	54	h1_heartrate_min	82	apache_2_bodysystem
26	resprate_apache	55	h1_mbp_max	83	Unnamed: 83
27	temp_apache	56	h1_mbp_min	84	hospital_death
28	ventilated_apache				

EDA and Statistical Analysis

greatlearning

Pre-processing :

- Dataset shape before null value treatment : (91713, 85)
- Dataset shape after null value treatment : (72454, 80)
- Dataset shape after checking for multi-collinearity : (72454, 58)
- Redundant Columns : 5

Multi-Collinearity :

- We have dropped columns having correlation coefficient greater than 0.8. In total, we have dropped 22 columns. Dataset shape after checking for multi-collinearity : (72454, 58)

Outlier Analysis :

- Outliers were present in the dataset for many columns. We have used the capping technique to handle the outliers. The outliers above (Q3 + 1.5*IQR) were capped at (Q3 + 1.5*IQR) and those below (Q1 - 1.5*IQR) were capped at (Q1 - 1.5*IQR).

Balancing the Dataset :

Degree of Imbalance Proportion of Minority Class :

- Mild : 20 - 40 % of dataset
- Moderate : 1 - 20 % of dataset
- Extreme : < 1 % of dataset

With respect to the above metric, the imbalance in our dataset falls in the category of moderate imbalance. To handle this imbalance, we have experimented with multiple approaches including SMOTE, SMOTE + Tomek T- Links, Under and Over Sampling, Balanced Class Weights approach.

EDA and Statistical Analysis

greatlearning

Feature Engineering :

- Scaling the data - We have scaled the numeric columns in the dataset using StandardScaler() function from sklearn library.
- Dummy Encoding - Also, we have encoded the categorical columns using One Hot Encoding.

Statistical Analysis :

The conclusions of the analysis are as below :

Chi-Sq test for independence of attributes :

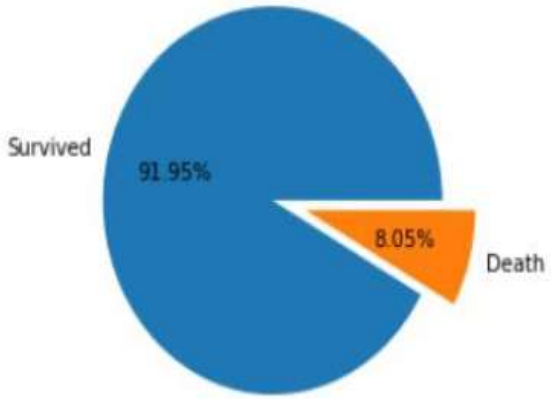
- The attributes, gender and hospital death are independent. Both males and females are at an equal risk of surviving or not surviving any particular medical condition.
- Ethnicity and hospital death are dependent.
- ICU Admit Source and hospital death are dependent.
- ICU type and hospital death are dependent.
- Apache_3j_bodysystem and hospital death are dependent.
- Ventilated_or_not and hospital death are dependent.

Z-Test for equality of means :

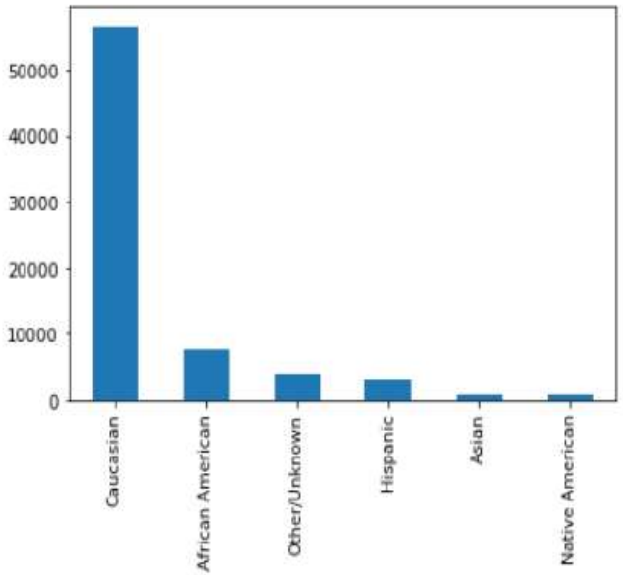
- The cases resulting in death are having a higher value of pre-icu loss-days.

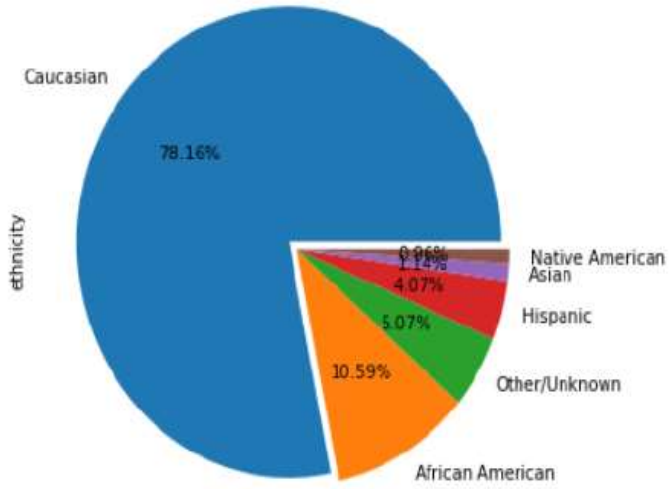
Data Exploration (EDA) :

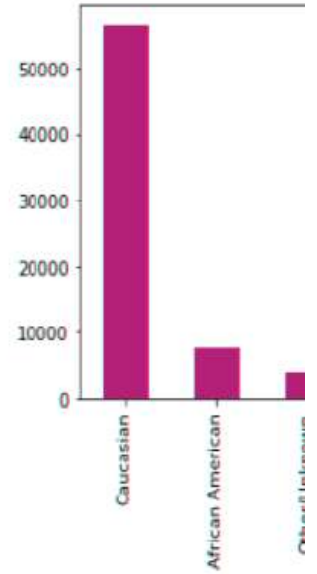
The findings from the dataset are summarized below :



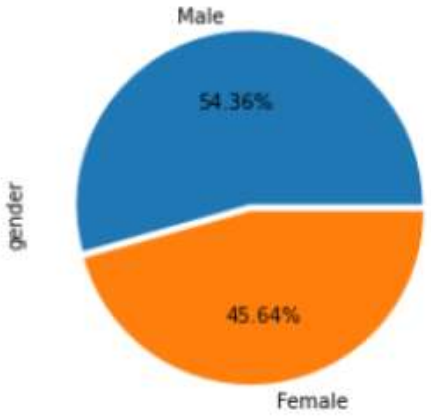
From the pie chart, we can infer that 92 % of the patients have survived their medical condition and 8.05 % patients have died.



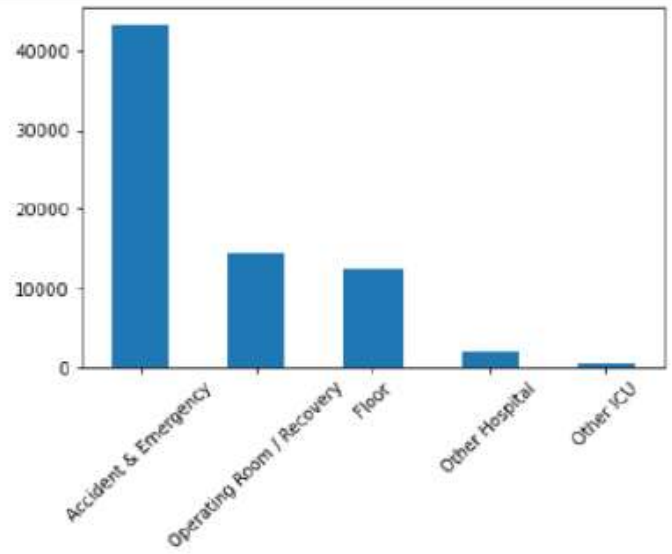




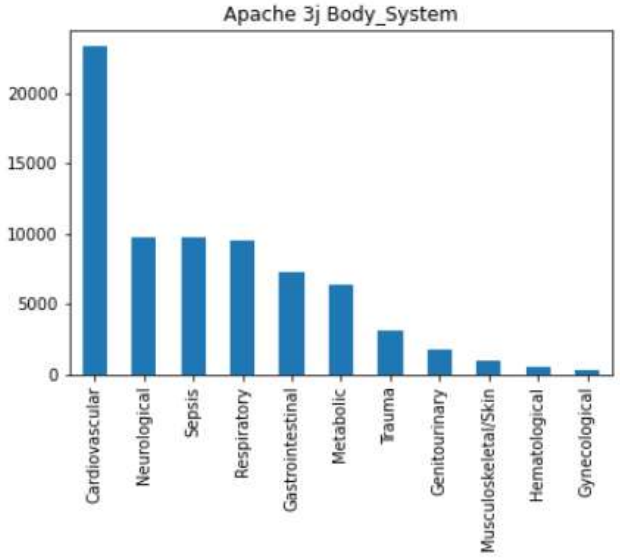
1. Majority of the patients i.e. 78 % belong to Caucasian Ethnicity
2. African American patients account for 10.6 % of the patients

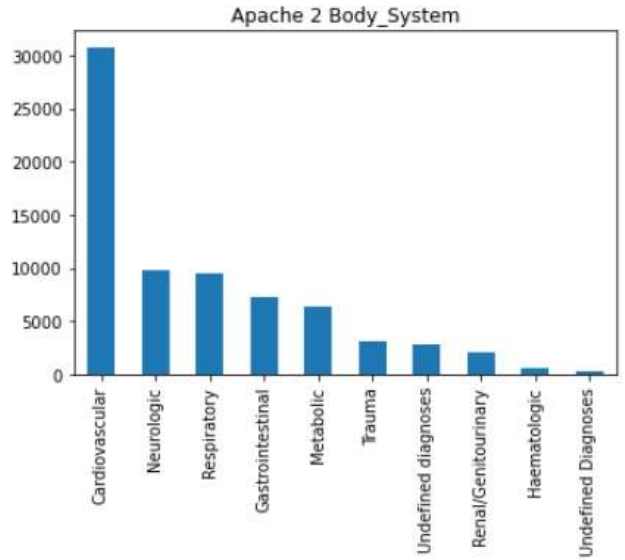


Males are 54 % and Females are 46 % of the total patient population



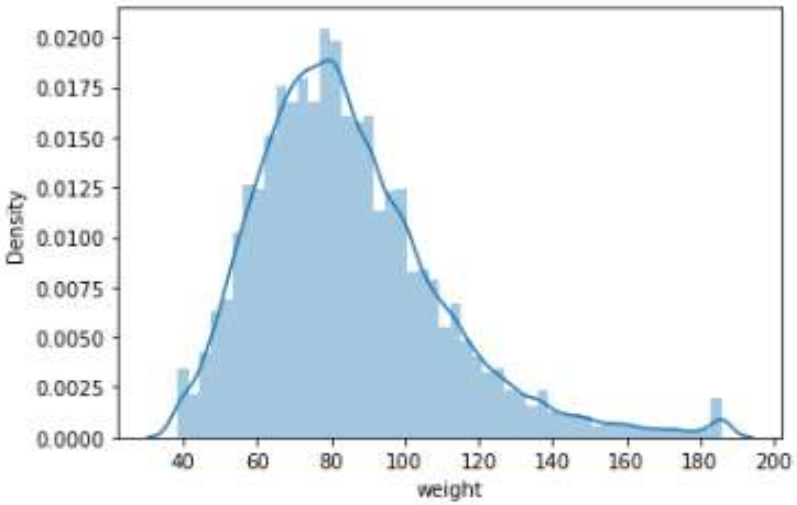
Majority of the cases are of Accident & Emergency

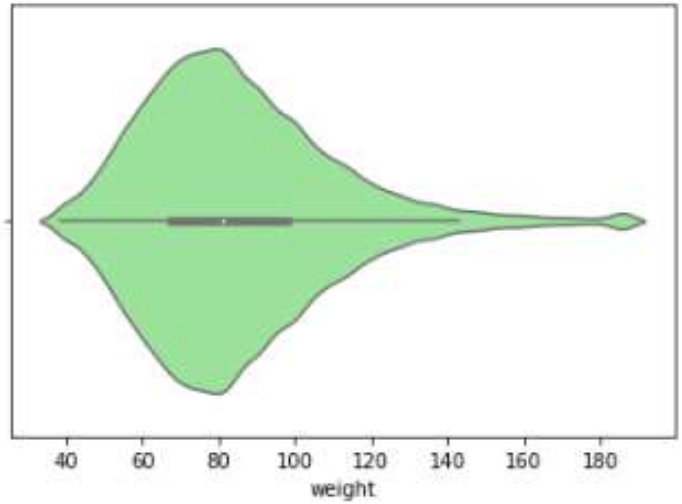




There are more cases of cardiovascular diseases as compared to other medical conditions

Distribution of variable - Weight

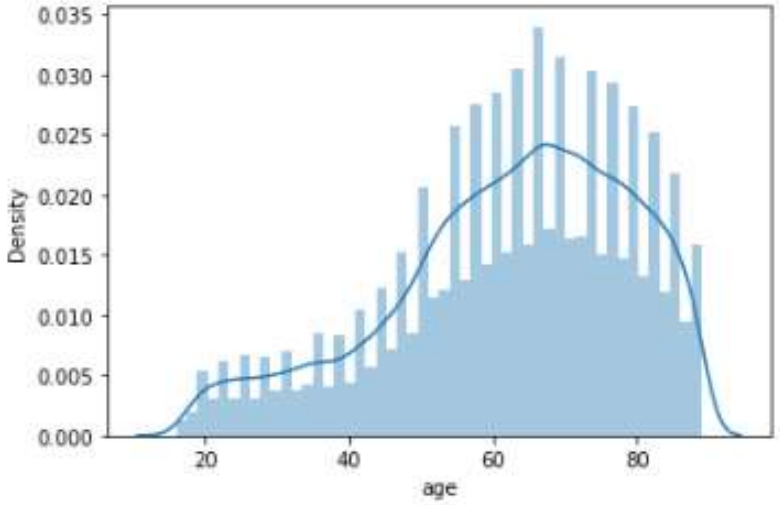


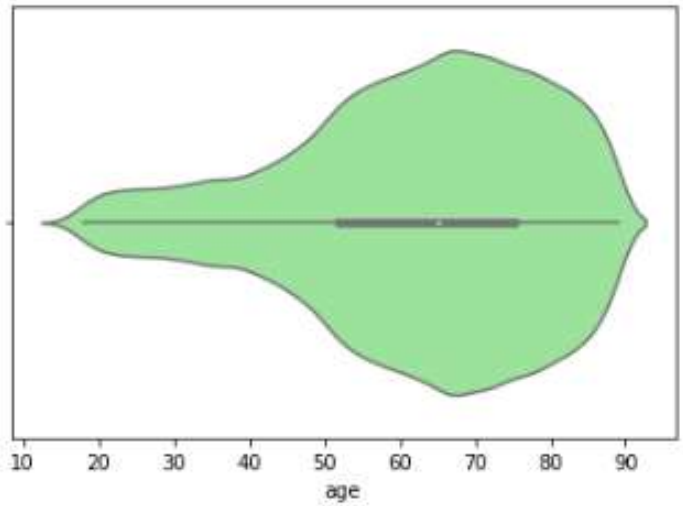



```
Skewness of Weight :  1.066220554588028
Mode of variable Weight :  84.55268653214341
Median of variable Weight :  81.0
```

The distribution of variable weight is positively skewed.
Majority of the patients belong to the weight category of 60 - 100 kg.

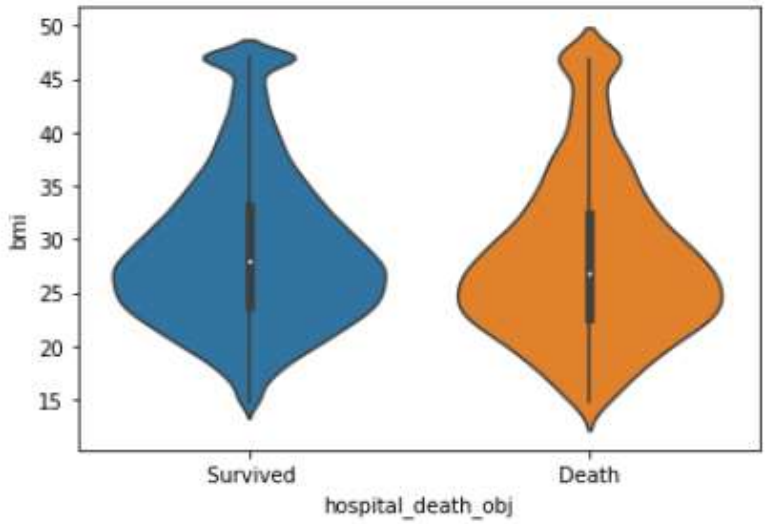
Distribution of variable - Age

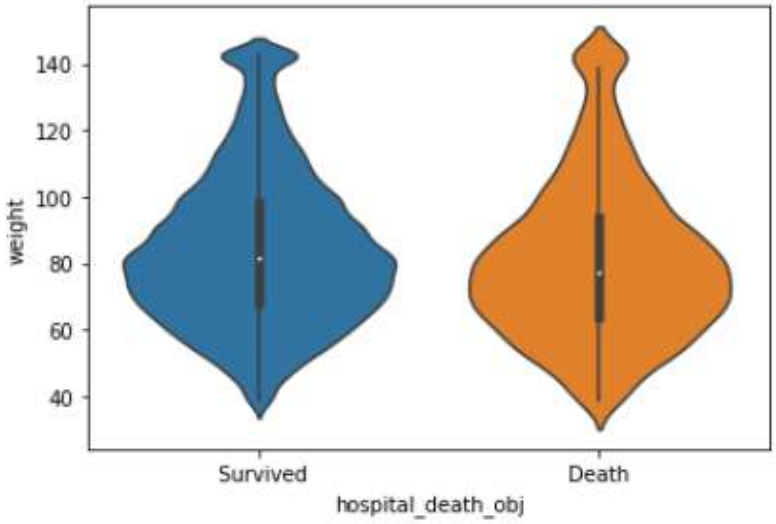





```
Skewness of Age :  -0.6270827791347964
Mode of variable Age :  67.0 yrs
Median of variable Age :  65.0 yrs
```

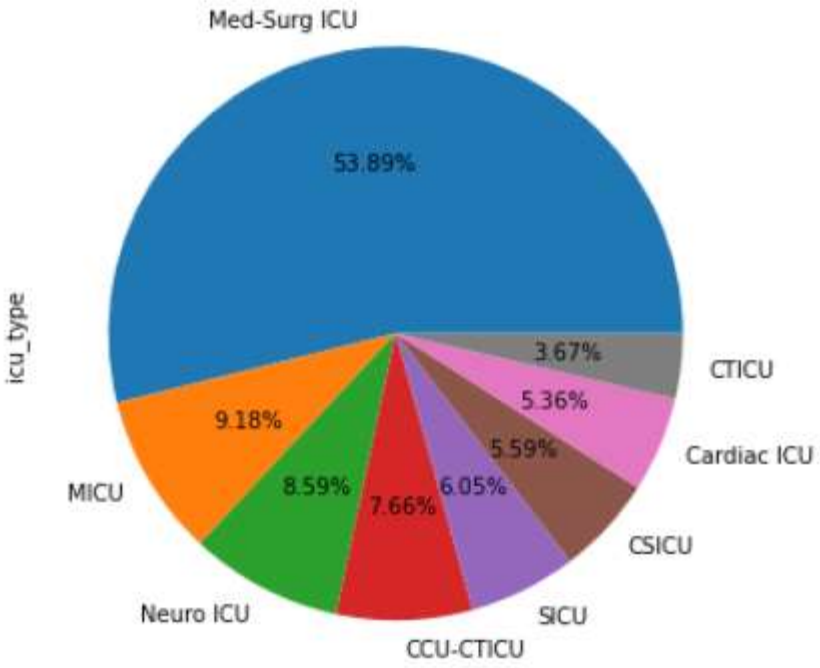
The distribution of variable age is negatively skewed.
Majority of the patients belong to the age group of 55 - 85 yrs.





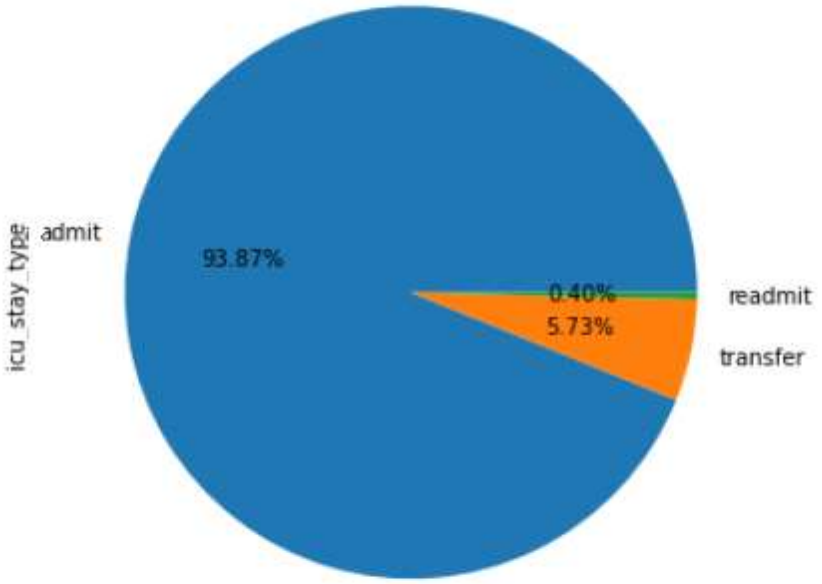
BMI, Weight and Height variables are not having any significant effect on the survival ability of the patient

ICU Type



Nearly 54 % of the patients have been admitted to the Med-Surg ICU.

ICU Stay - Type



Nearly 94 % of the cases are of admit as compared to 6 % of transfer and only 0.4 % of readmit.

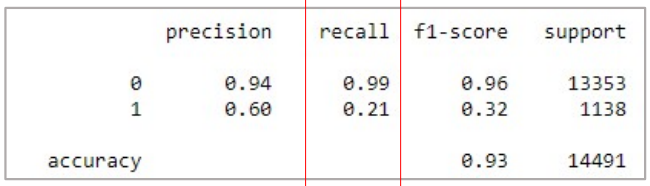
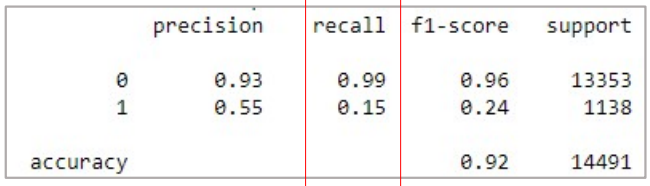
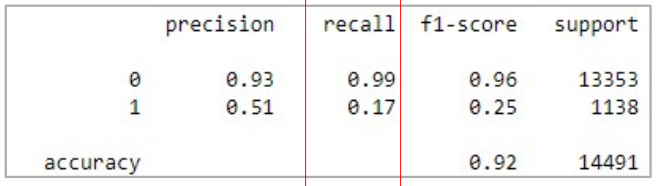
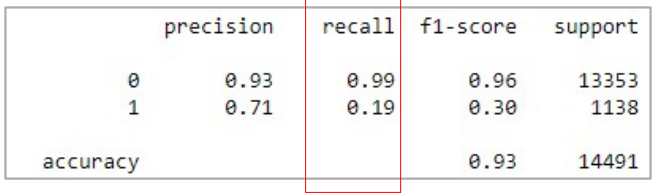
Basic Model Building Steps

greatlearning

Algorithms considered :

As our problem is a Classification problem, we have experimented with Classification Algorithms which include KNN, Decision Tree, Random Forest and Boosting Techniques.

Base Models : Test classification results of some base models are summarized below.

Algorithm	Full Model
1. Logistic Regression	
2. KNN	
3. Decision Tree	
4. Random Forest	

	precision	recall	f1-score	support
0	0.94	0.99	0.96	13353
1	0.60	0.21	0.32	1138
accuracy			0.93	14491

	precision	recall	f1-score	support
0	0.93	0.99	0.96	13353
1	0.55	0.15	0.24	1138
accuracy			0.92	14491

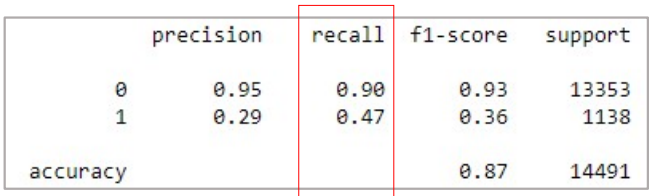
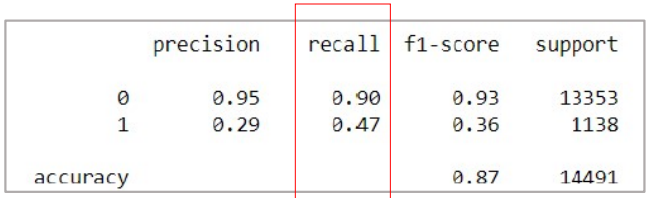
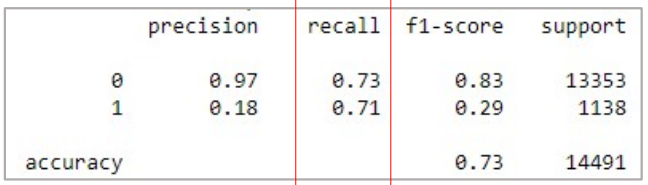
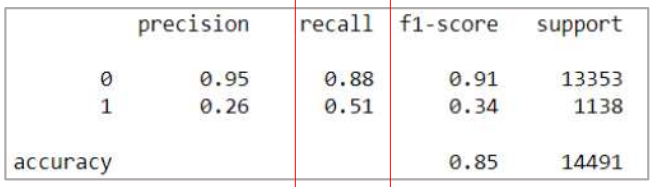
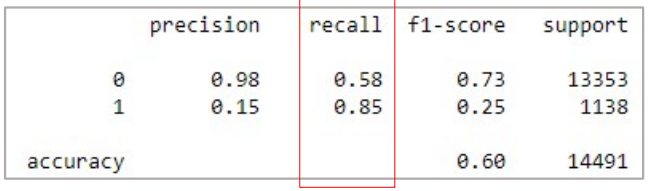
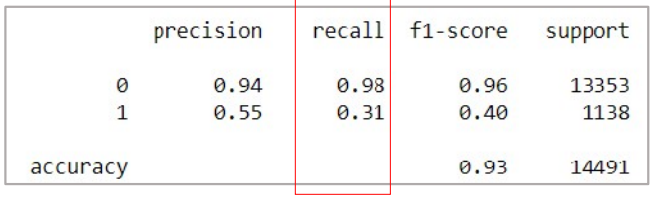
	precision	recall	f1-score	support
0	0.93	0.99	0.96	13353
1	0.51	0.17	0.25	1138
accuracy			0.92	14491

	precision	recall	f1-score	support
0	0.93	0.99	0.96	13353
1	0.71	0.19	0.30	1138
accuracy			0.93	14491

Handling Imbalance

Multiple approaches to handle imbalance :

- SMOTE
- SMOTE + Tomek T-Links

Algorithm	After applying SMOTE and treating Multi-Collinearity	Applying SMOTE + Tomek T-Links
1. Logistic Regression		
2. Ada Boost		
3. XG Boost		

Handling Imbalance

greatlearning

Multiple approaches to handle imbalance :

- Under + Over Sampling
- Balanced Class Weights

Algorithm	25 % Down-Sampling + Up-Sampling	Balanced Class Weights
1. Logistic Regression		
2. Ada Boost		
3. XG Boost		

1. Logistic Regression — 25 % Down-Sampling + Up-Sampling

	precision	recall	f1-score	support
0	0.98	0.79	0.88	13353
1	0.25	0.79	0.38	1138
accuracy			0.79	14491

1. Logistic Regression — Balanced Class Weights

	precision	recall	f1-score	support
0	0.98	0.79	0.88	13353
1	0.25	0.79	0.38	1138
accuracy			0.79	14491

2. Ada Boost — 25 % Down-Sampling + Up-Sampling

	precision	recall	f1-score	support
0	0.95	0.88	0.91	13353
1	0.26	0.51	0.34	1138
accuracy			0.85	14491

2. Ada Boost — Balanced Class Weights

	precision	recall	f1-score	support
0	0.93	1.00	0.96	13353
1	0.70	0.11	0.20	1138
accuracy			0.93	14491

3. XG Boost — 25 % Down-Sampling + Up-Sampling

	precision	recall	f1-score	support
0	0.94	0.98	0.96	13353
1	0.55	0.31	0.40	1138
accuracy			0.93	14491

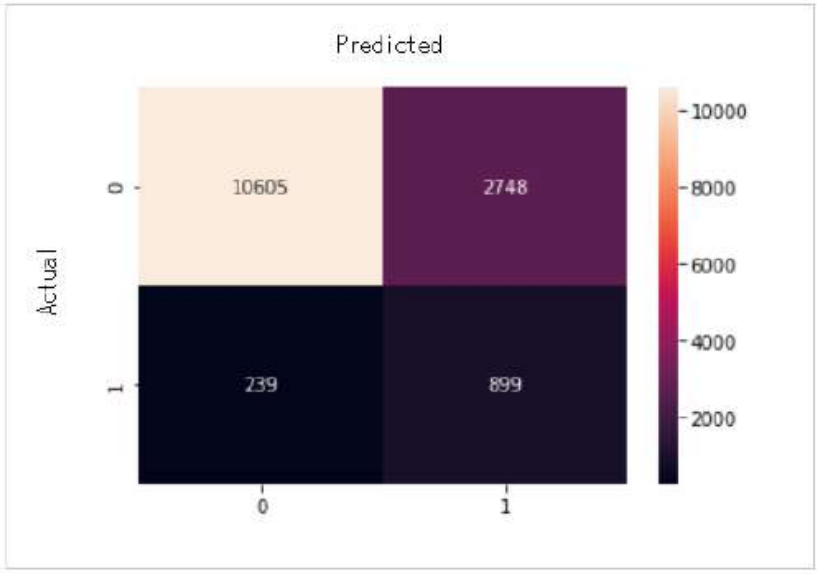
3. XG Boost — Balanced Class Weights

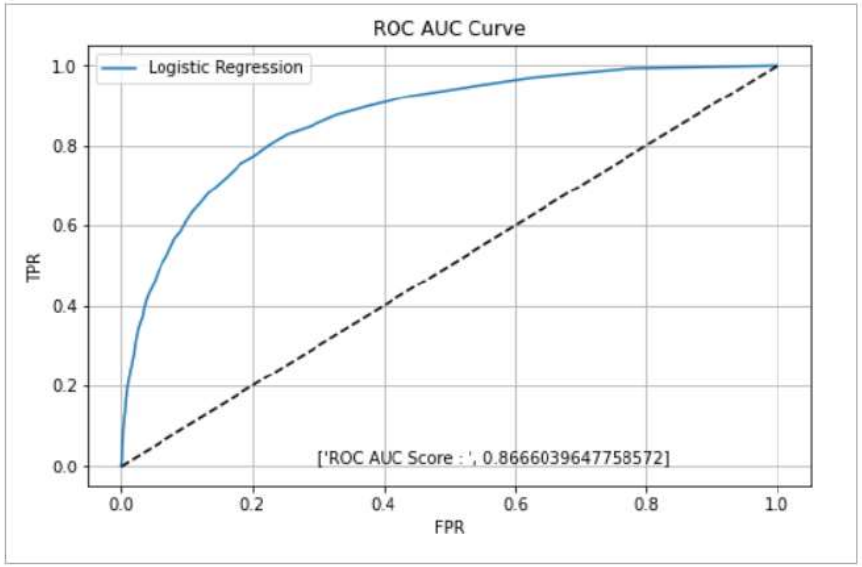
	precision	recall	f1-score	support
0	0.94	0.98	0.96	13353
1	0.59	0.29	0.38	1138
accuracy			0.93	14491

Final Model – Logistic Regression Balanced Class Weights

	precision	recall	f1-score	support
0	0.98	0.79	0.88	13353
1	0.25	0.79	0.38	1138
accuracy			0.79	14491

FPR	TPR	Threshold	Younden_Ind
0.207968	0.792619	0.496324	0.584650





Final Model- Results & Limitations

Results :

- We are considering Recall Score as our main evaluation metric.
- For our final model – Logistic Regression – Balanced Class Weights – Recall Score is 0.79 for both classes.
- High Recall Score means low FNs i.e. low value of Type-II error.
- Type-II error more fatal in our case and our model has effectively reduced it.

Limitations :

- Low precision score i.e. high FPs i.e. high Type-I error.
- Can be improved upon in the future and if addressed rightly, can lead to significant improvement in the reliability of our model for patient survival prediction.

Discussion :

- Model not highly accurate but safe to depend upon and when clubbed with medical expertise of doctors can be helpful.
- Can prove helpful in prioritizing which cases are more severe in terms or survival ability and need immediate attention.

Thank You