Patient Survival Prediction

03/16/2022

Prepare for Data Mining Principles

By: Naibo(Ray) Hu, Weijia(Joyce) Wang, Wen Zhang, Jingwen Nan, Moxuan(Polly) Zheng, Rujue Du

Agenda

01 Business Problem

02 Data Profile & Quality

03 Data Processing

04 Exploratory Data Analysis (EDA)

05 Feature Engineering

06 Model Building & Evaluation

07 Conclusion

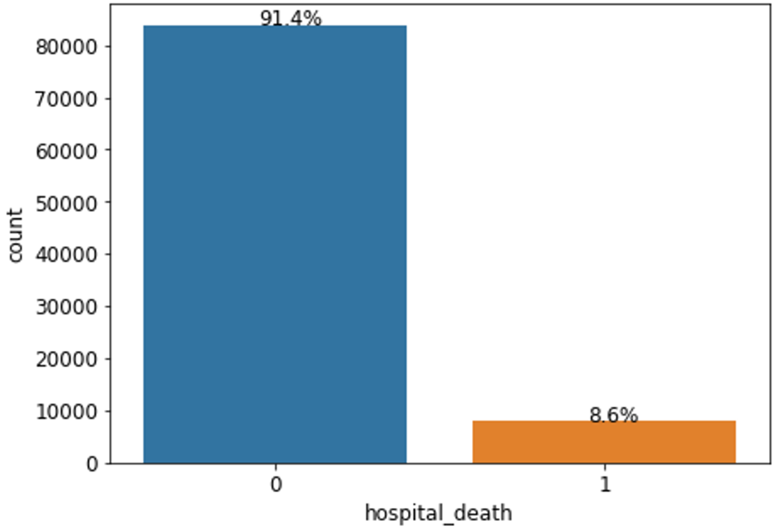
08 Lessons learned

Business Use Case

The predictors of in-hospital mortality for admitted patients remain poorly characterized. **We aim to develop and validate a prediction model for all-cause in-hospital mortality among admitted patients, and identify factors that are most relevant to the ICU survival rate.**

We hope that our analysis can provide clinical research and public health professionals with valuable insights, save lives by increasing patients' survival rates, and ultimately bring a positive impact to the community.

Data Profile



Potential Problem: imbalanced response variable

91.4% of hospital_death is 0, and 8.6% of hospital_death is 1

Name	Patient Survival Dataset
Dimensions	85 variables 91, 713 rows
Data Type	Semi-structured
Data Size	31.4 MB
Description	The dataset includes predictors for admitted patients in hospital
Source	Kaggle

Data Quality

Completeness

Any missing values?

Out of 85 variables, **75 of them have missing values.** There are 288,046 missing values in total.

Validity

Does data match the rules?

All fields are checked and formatted to be the appropriate data type in our database.

Uniqueness

Are there duplicate values?

Each row in the dataset is unique.

Consistency

Consistent across various data stores?

Dataset is stored in and sourced from Google Drive, so it is consistent for all users.

Timeliness

Does data represent reality from required point in time?

Dataset was created and uploaded in 2021, so the data is up-to-date.

Accuracy

Degree to which data represents reality

Dataset contains patients' various types of health data, which is objective.

Handling Missing Values

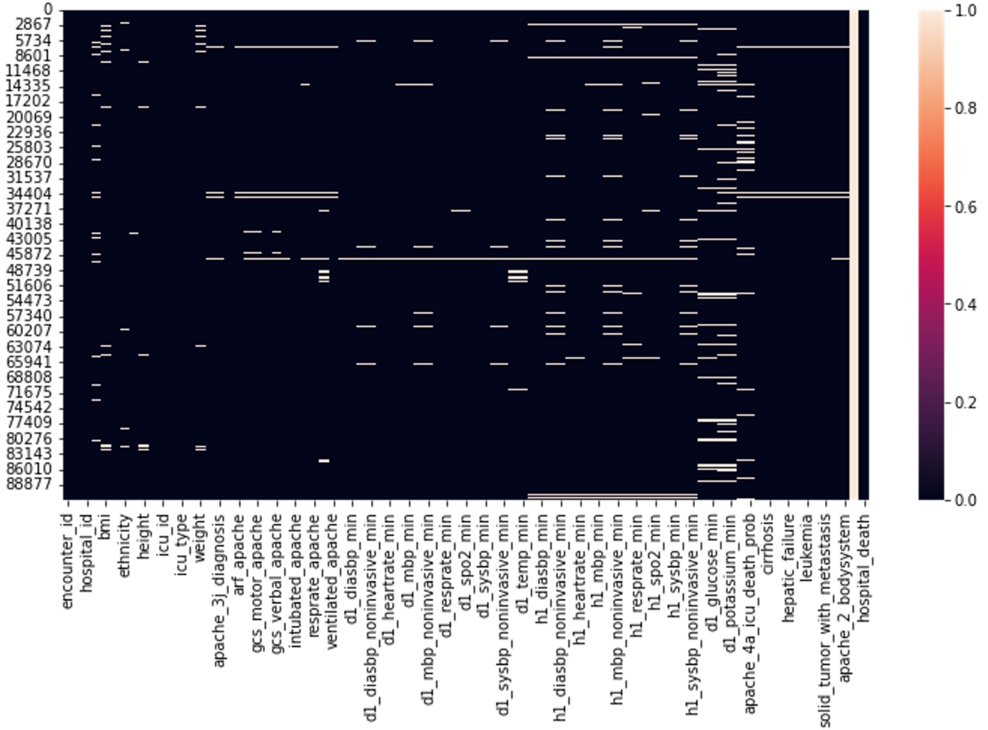
- For missing values in numeric variables, we use **interpolation method.**

- There are 25 missing values in “gender” column, and we replaced them with **the most frequent value, “M.”**

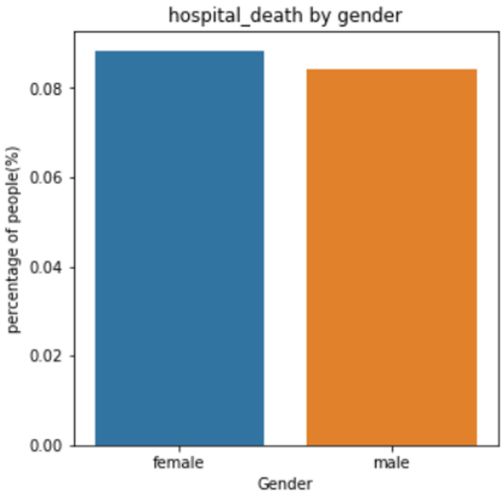
Value counts in gender column

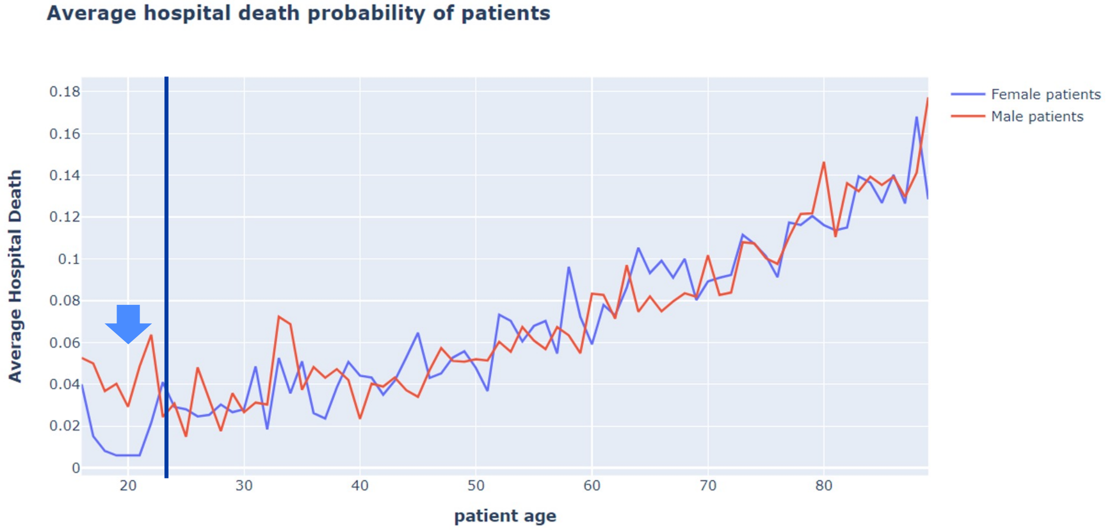
```
M     49469
F     42219
Name: gender, dtype: int64
```

Visualizing missing values with heatmap



Hospital death is not related to gender





- **8.44% of male and 8.84% female** died in hospital, which are roughly equal.
- The mortality rate for male is much higher than that of female around **age 15-25**; this is probably due to male are more likely to engage in risky activities than female.
- The **mortality rates are similar for both men and women** after age 25.
- As patients get older, hospital death rate increases.

Apache score is not as predictive as expected, as low apache scores associate with high mortality rate

Insights:

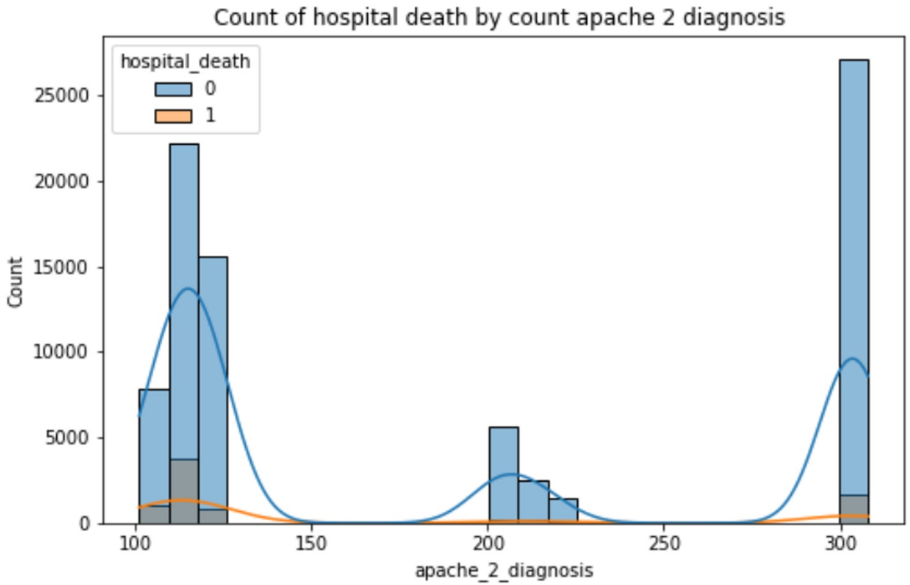
- Most patients are **distributed around extreme apache scores**.

- **Extremely high or low values** indicate relative **higher probabilities of deaths** compared to mortality rate around score of 200, which is counterintuitive.

Intuition:

- Patients with lower apache II scores might not be taken care of properly.

- Since the apache II score only has 75% of accuracy, it is likely that the scores underestimate the severity of patients' disease.

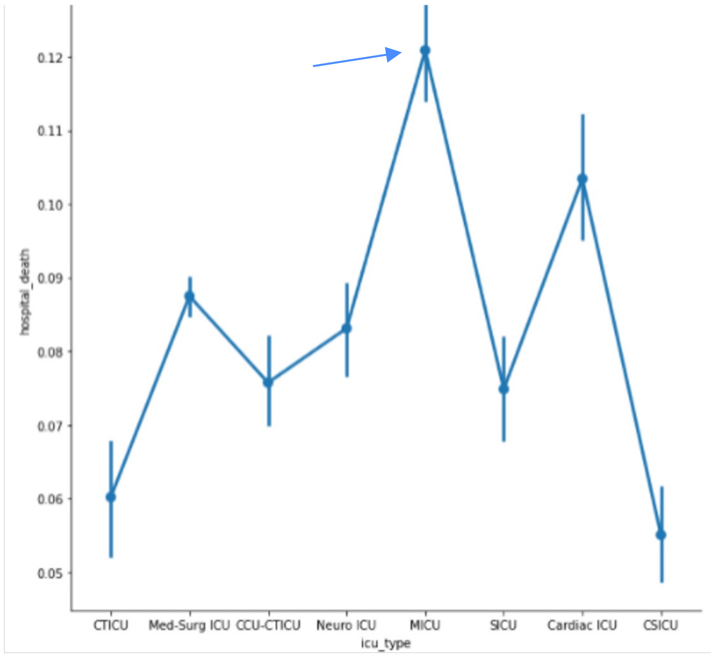


Explanation of Apache_2_diagnosis:

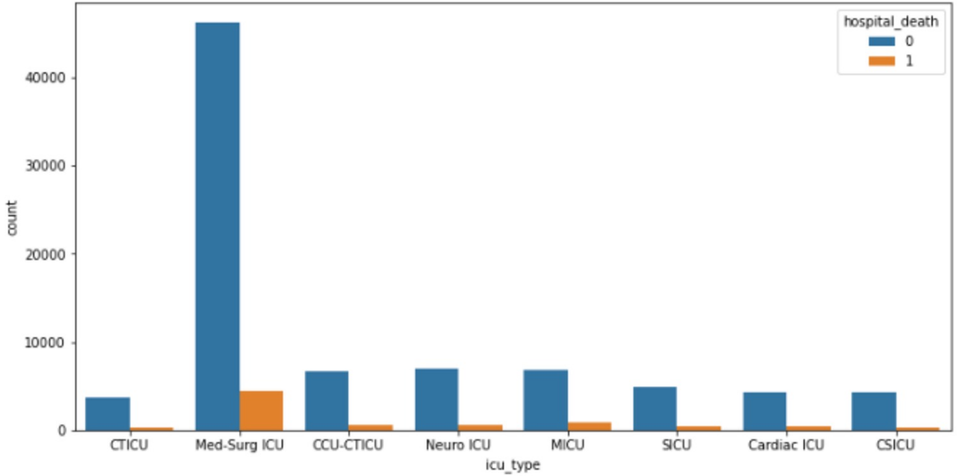
- ICU scoring systems
- Higher scores correspond to more severe disease and a higher risk of death

ICU types are associated with patients' mortality rate

Death rate by ICU type



Number of patients by ICU type



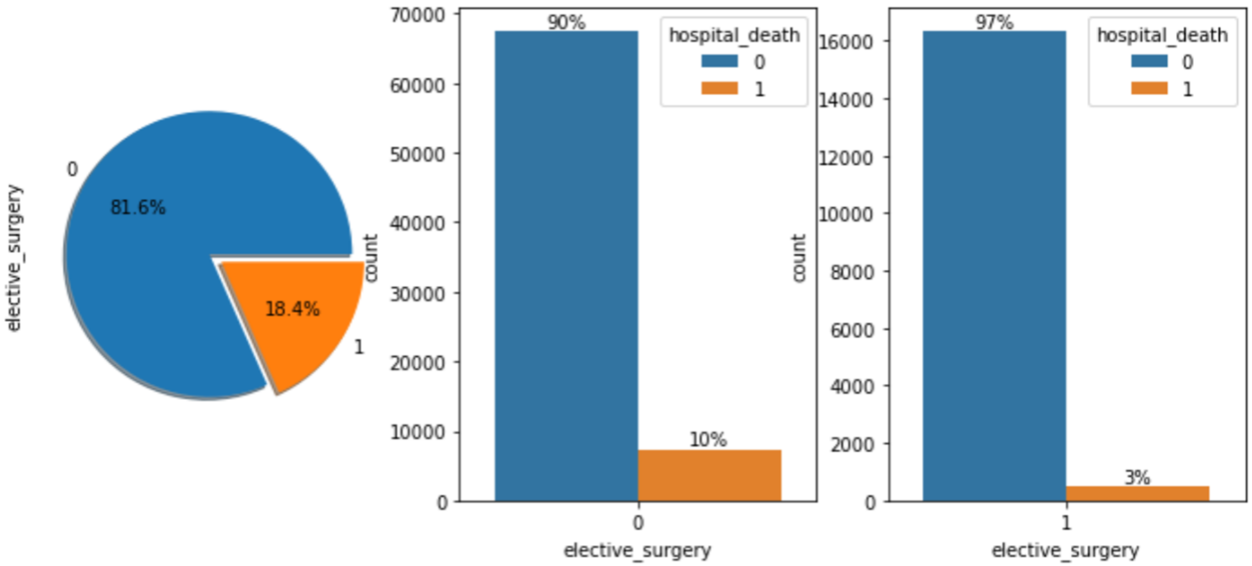
Most patients are assigned to Med-Surg ICU. **However, MICU**(medical intensive care unit) **has the highest death rate** among all ICU types. This unit is a medical specialty that deals with seriously or critically ill patients who have, are at risk of, or are recovering from conditions that may be life-threatening.

Receiving Elective Surgery is likely to reduce mortality rate

Insights:

The percentage of patients who were admitted for elective surgery was as low as **18.4%**

However, as we find from the bar charts, **patients who have received elective surgery had a lower mortality rate** than those who did not.



Intuition:

Even for mild symptoms, giving treatments earlier is likely to increase patients' survival rate.

Elective_surgery:

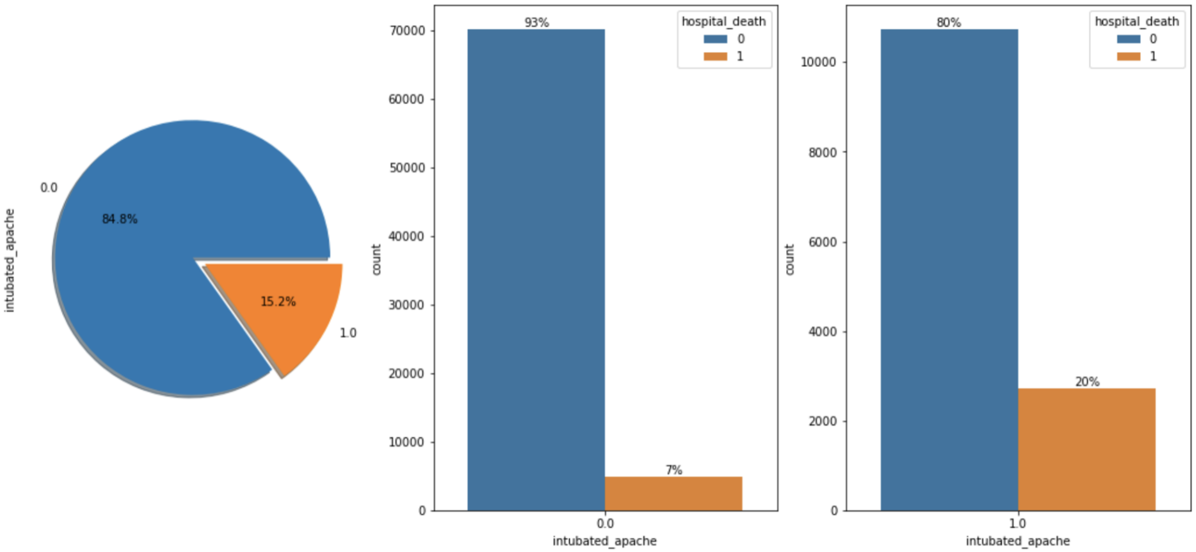
- Whether the surgery can be scheduled in advance.
- It may be a surgery you choose to have for a better quality of life, but not for a life-threatening condition.

Intubated Apache is associated with patients' mortality rate

Insights:

15% of the patients were intubated during the treatment, which means that they may have experienced respiratory failure or shock.

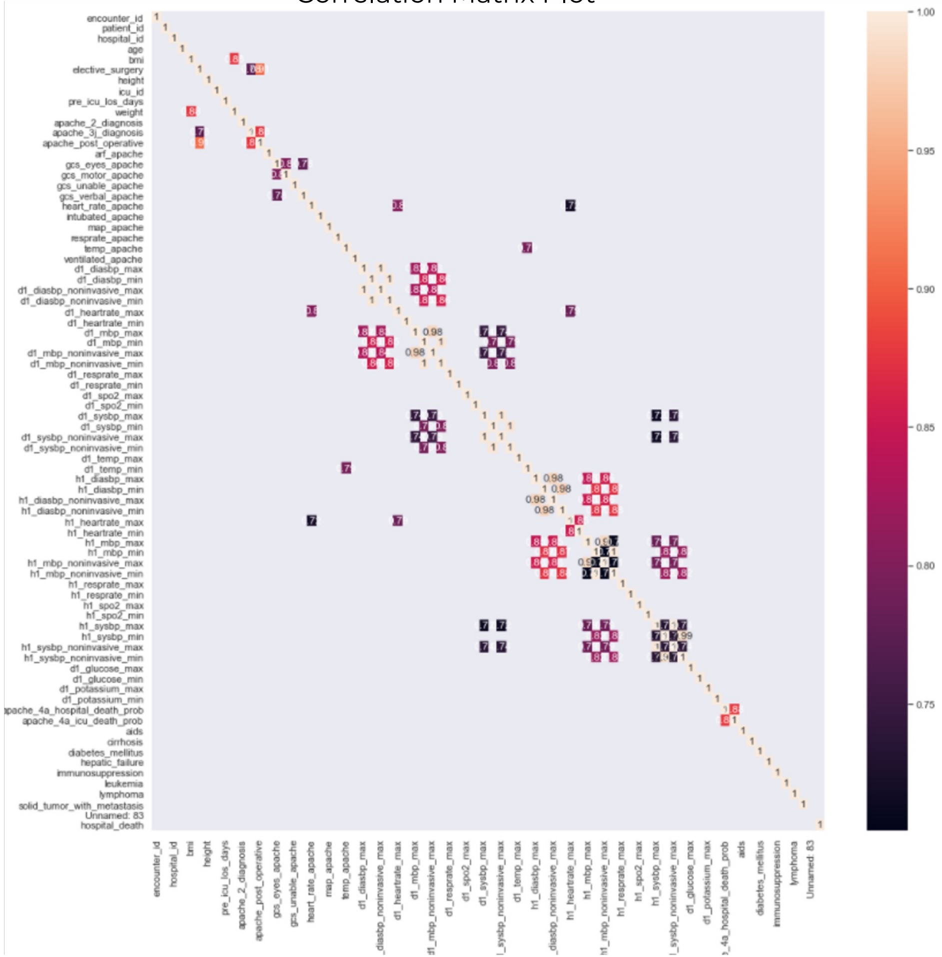
Our histograms show that **the mortality rate is 20% for patients with intubated treatment**, which is greater than the mortality rate of patients who were not intubated.



Intubated Apache:

- Whether the patient was intubated at the time of the highest scoring arterial blood gas used in the oxygenation score

Correlation Matrix Plot



There is multicollinearity among numeric variables

Variables are considered to have **collinearity** if they have **correlation values greater than 0.7.**

There are **65 pairs of variables** having correlation values greater than 0.7.

Thus, we decided to **use PCA** for dimension reduction and further reduce collinearity.

Encode categorical variables and drop unnecessary columns

Dummy Coding

gender

One Hot Encoding

ethnicity
icu_admit_source
icu_stay_type
icu_type
apache_3j_bodysystem
apache_2_bodysystem

Drop columns

Drop “encounter_id” and “patient_id”

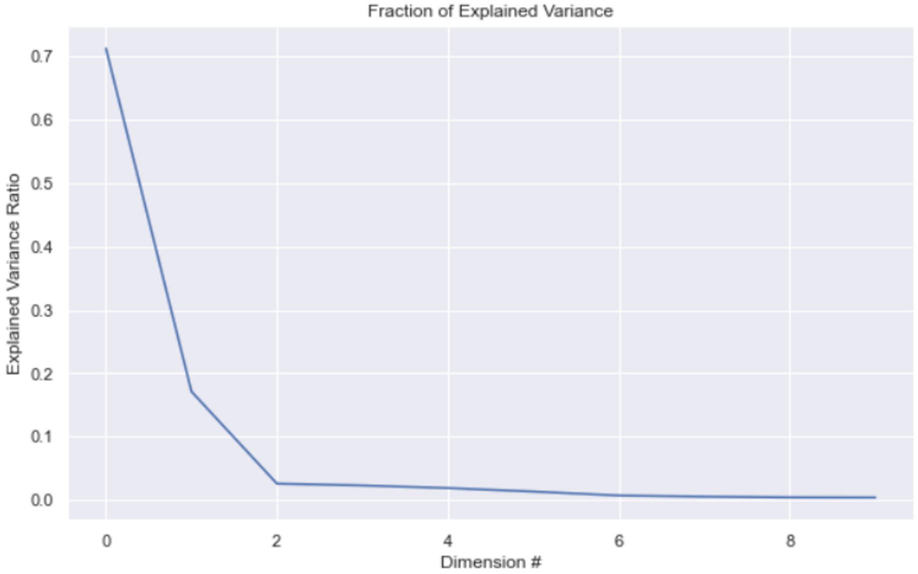
121 variables
88,589 rows

Principal Component Analysis (PCA)

Since we have 85 variables, and many of them have multicollinearity issue, we decided to **eliminate noise** through PCA.

From the PCA variance plot, we find that the **first two components can explain roughly 85%** of response variable. Thus, we decide to use first two components (0,1).

Based on PCA component plots, we end up selecting **32 significant variables** (shown in Appendix)



Utilize VIF to check multicollinearity for selected 32 predictors

	feature	VIF
0	age	1.201421
1	apache_3j_bodysystem_Trauma	1.238756
2	d1_sysbp_noninvasive_min	1.612316
3	apache_2_bodysystem_Neurologic	1.686407
4	apache_3j_bodysystem_Genitourinary	1.129171
5	d1_spo2_min	1.116406
6	d1_mbp_max	1.486617
7	d1_temp_max	1.095657
8	d1_glucose_max	1.320305
9	h1_sysbp_noninvasive_max	1.958513
10	icu_stay_type_readmit	22.011346
11	icu_admit_source_Other ICU	1.015118
12	h1_resprate_max	1.137979
13	apache_3j_bodysystem_Cardiovascular	2.137271
14	solid_tumor_with_metastasis	1.009992
15	apache_2_bodysystem_Haematologic	1.037935
16	ethnicity_Asian	1.113664
17	icu_stay_type_admit	3254.250352
18	icu_stay_type_transfer	183.568043
19	apache_2_bodysystem_Gastrointestinal	1.463269
20	ethnicity_Caucasian	1.965351
21	d1_heartrate_min	1.905610
22	d1_potassium_max	1.092980
23	d1_glucose_min	1.230969
24	h1_mbp_min	1.997788
25	h1_heartrate_min	2.008103
26	apache_3j_bodysystem_Gynecological	1.025808
27	lymphoma	1.002761
28	ethnicity_African American	1.876258
29	apache_2_bodysystem_Metabolic	1.572000
30	icu_type_SICU	1.021476
31	apache_2_bodysystem_Cardiovascular	2.580895

"icu_stay_type_readmit," "icu_stay_type_admit," and "icu_stay_type_transfer" have VIF values greater than 5.

We decided to drop the three columns, and we ended up having **29 predictors**

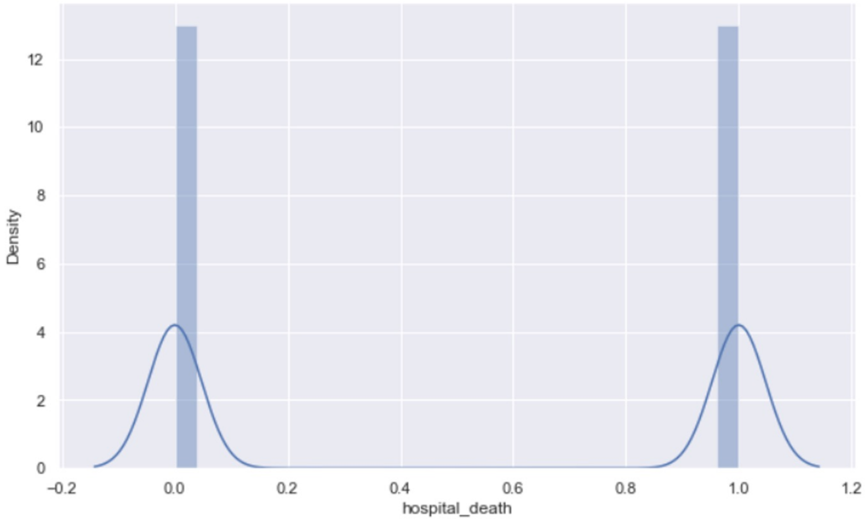
Utilized SMOTE Resampling method to oversample minority target variable

We split data into **80% train and 20% test**, and we applied SMOTE to **train data**.

91.4% of hospital_death is 0, and 8.6% of hospital_death is 1 in training dataset

Hospital _death	Before SMOTE	After SMOTE
0	64,737	64,737
1	6,134	64,737

Distribution of target variable in train data after SMOTE



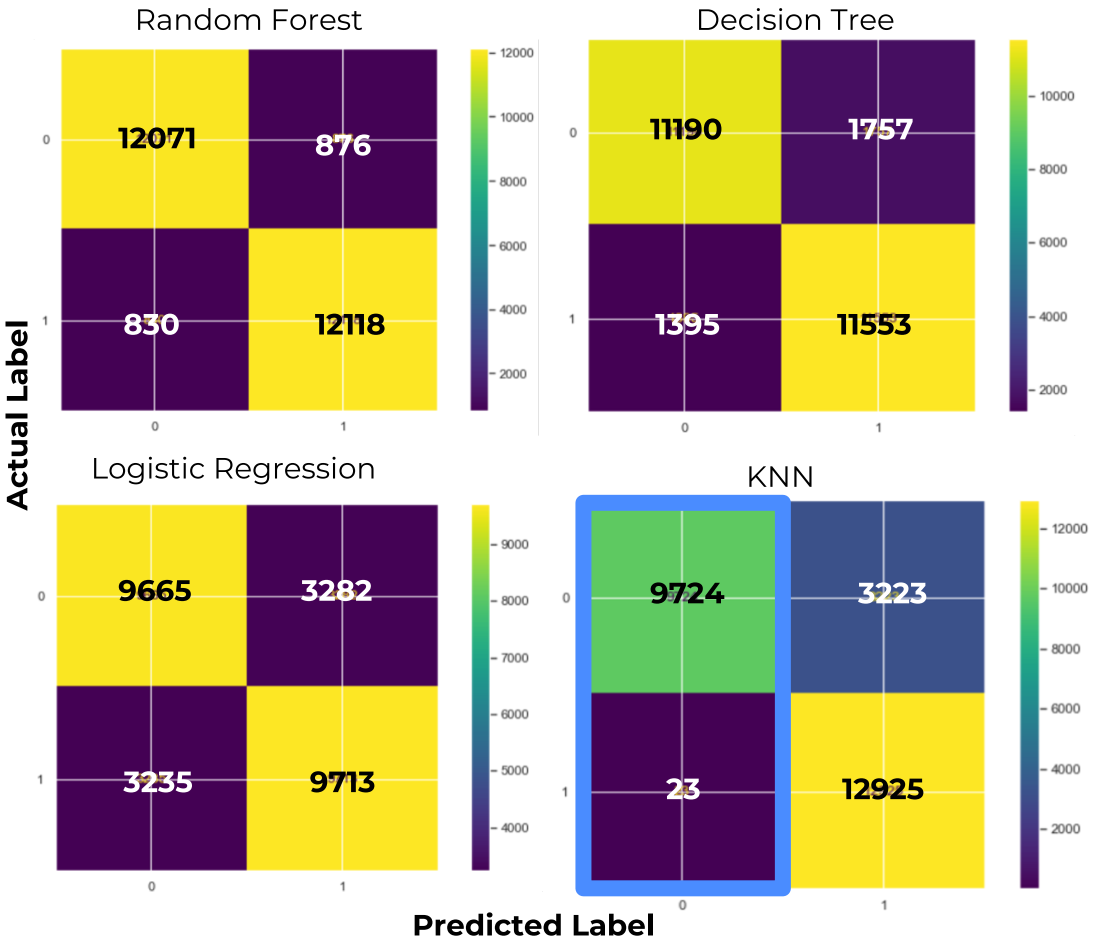
Random Forest is the optimal model

We applied **10-fold cross validation** to the train dataset. Then, we made predictions with the four models.

- **Random Forest** performed the best among the four models
- No model has overfitting problem, as training accuracies roughly equal CV accuracies
- Non-linear models are computationally complex and thus have higher accuracy values than the linear model

Patient Survival Prediction Result

	Training Accuracy	CV Accuracy	F1 Score	Precision Score	Recall Score
Logistic Regression	0.748	0.744	0.748	0.747	0.750
Decision Tree	0.878	0.878	0.878	0.868	0.892
Random Forest	0.934	0.936	0.934	0.933	0.936
KNN	0.875	0.880	0.875	0.800	0.998

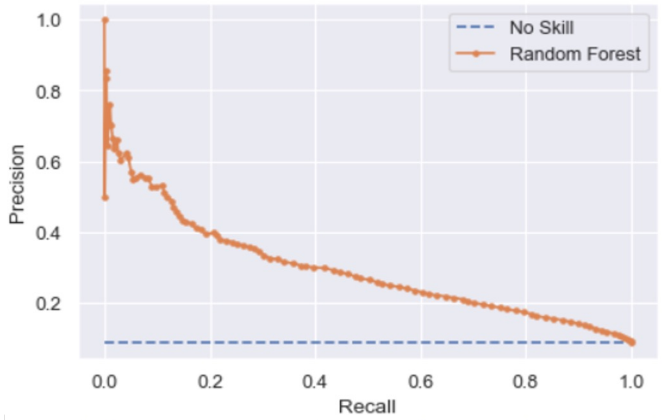


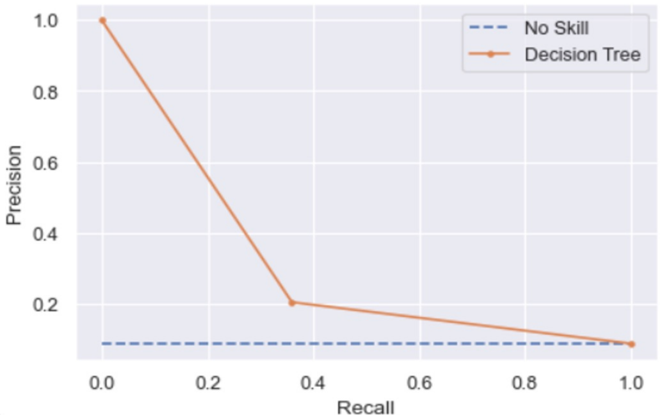
KNN generates the highest recall score (lowest false negative ratio)

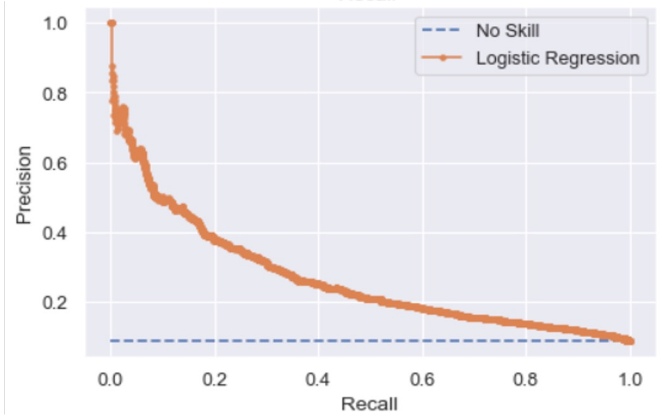
Only **0.2% patients** predicted to survive will actually die.

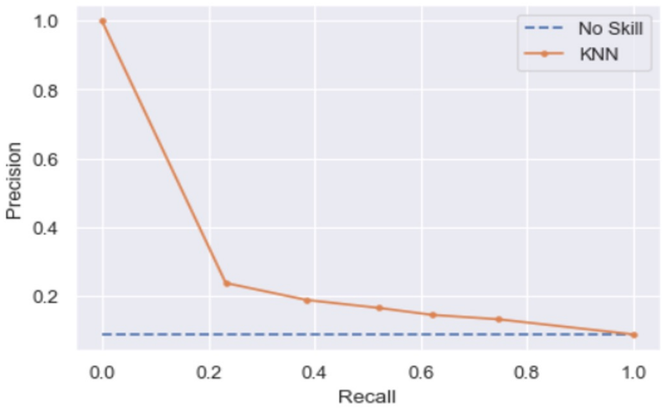
Models	False Negative Ratio
Random Forest	0.064
Decision Tree	0.108
Logistic Regression	0.250
KNN	0.002

There is no ideal trade-off point between precision and recall









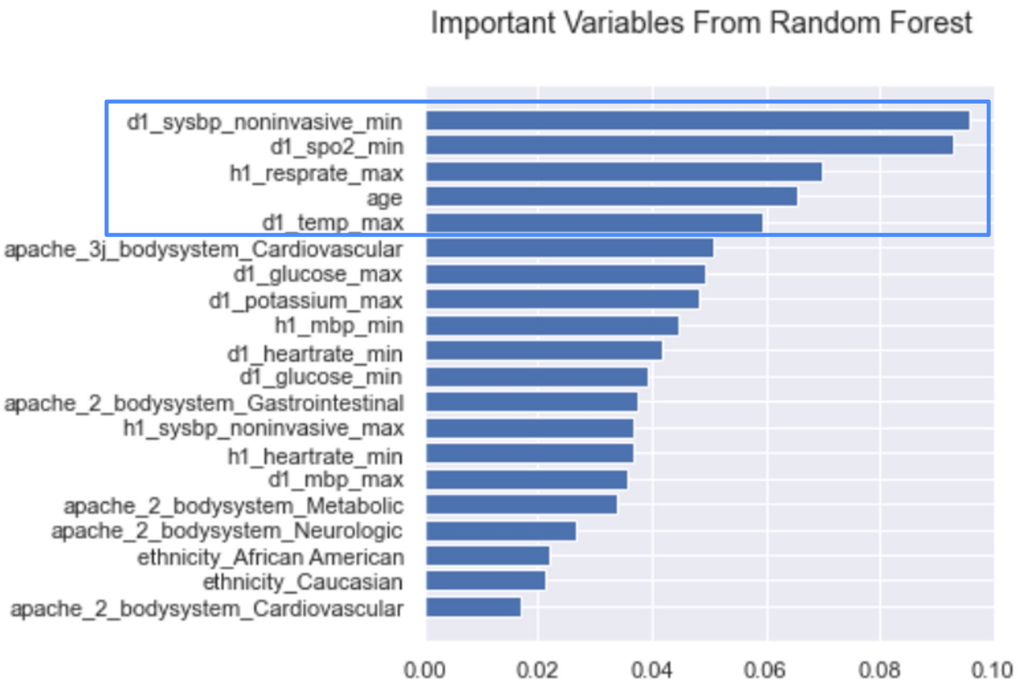
The four curves are **far away** from the upper right corner.

Thus, the trade-off between precision and recall is **not ideal**.

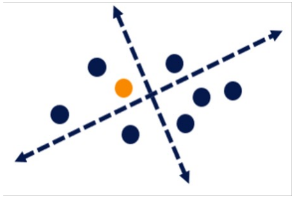
We identified top 20 significant variables from the random forest model

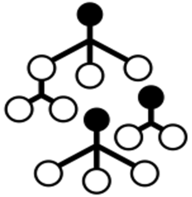
The **top 5 important variables** are "d1_sysbp_noninvasive_min": , "d1_spo2_min", "h1_resprate_max", "age", "d1_temp_max", and they together explain **38.3%** of the response variable.

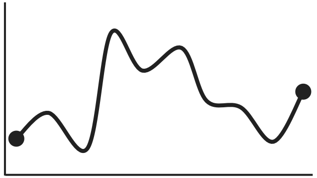
These 5 variables take into consideration of **blood pressure, oxygen saturation, respiratory rate, age, and temperature,** all of which are vital to patients survival chance.



Conclusion







PCA has successfully **reduced 85 variables to 32 variables**, yet does not impact model performance.

Random forest is the optimal model with **93.16% prediction accuracy.**

KNN has **high recall score**, which is **0.998.**

Nonlinear Models, such as random forest and decision tree, **perform better** than linear models, as they are more tolerant to complex data and are not affected by multicollinearity.

Lessons Learned

01

Some **trends identified from EDA do not align with feature importances from random forest model**.

For example, from EDA, we thought that ICU types and elective surgery are strongly associated with patient mortality rate. However, the ICU type is not in the list of top 20 important variables generated from the random forest model.

This result shows that **physiological indicators such as blood pressure, body temperature directly impact patients mortality.**

02

We impute numeric missing values with linear interpolation. However, this method is not precise when we do not have linear data. In the future, we can **use machine learning algorithms to predict missing values.**

03

Since logistic regression does not perform well with post PCA data. It is possible that we do not select enough variables. Thus, we can consider **applying lasso, ridge, and elastic net regularization techniques on pre PCA data.**

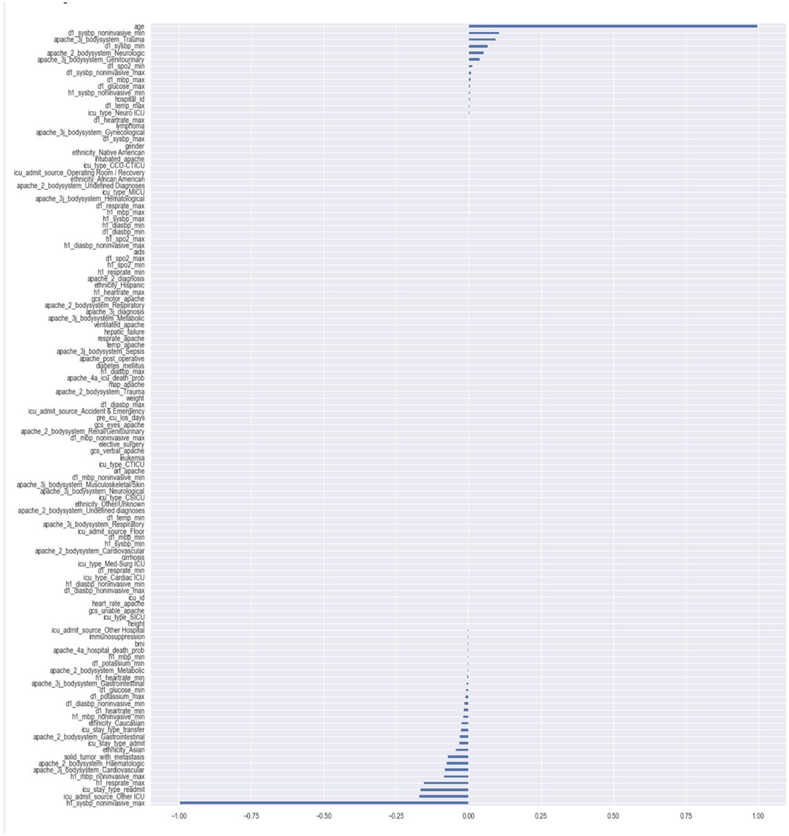
Thanks!

Reference

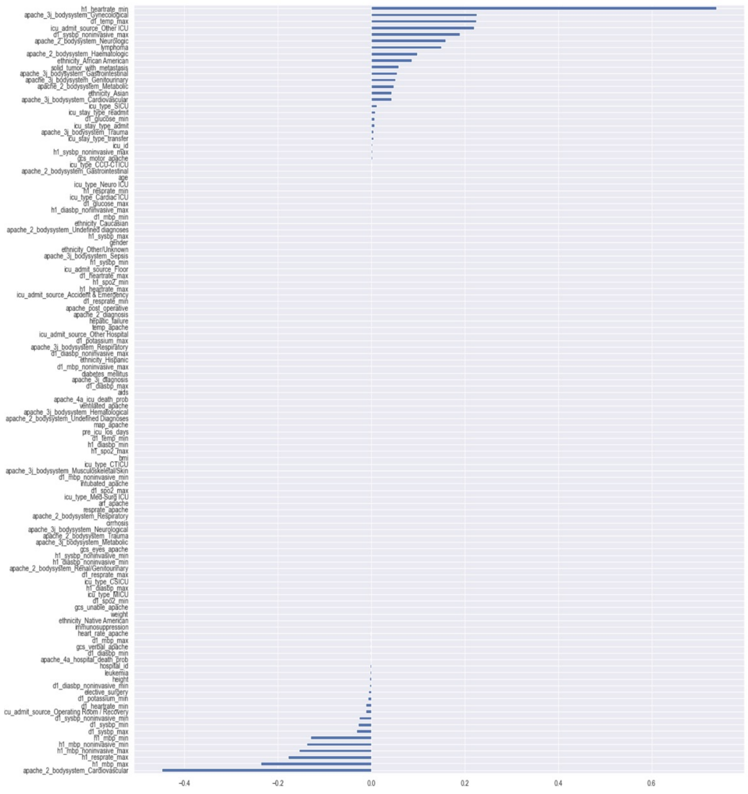
- https://sdsclub.com/how-to-train-and-test-data-like-a-pro/
- https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
- https://www.geeksforgeeks.org/difference-between-pca-vs-t-sne/
- https://www.ibm.com/cloud/learn/random-forest
- https://dhirajkumarblog.medium.com/top-4-advantages-and-disadvantages-of-support-vector-machine-or-svm-a3c06a2b107
- https://nurse.org/articles/hospital-unit-acronyms/#:~:text=MICU%20stands%20for%20medical%20intensive,gastrointestinal%20problems%2C%20and%20blood%20infections.
- https://www.hopkinsmedicine.org/health/treatment-tests-and-therapies/types-of-surgery
- https://www.mountsinai.org/health-library/tests/blood-gases
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3237146/

Appendix

PCA 1 Loadings



PCA 2 Loadings



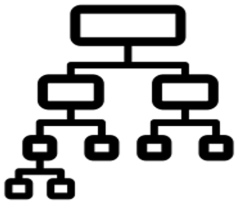
Selected 32 significant variables from PCA components

'age', 'apache_3j_bodysystem_Trauma', 'd1_sysbp_noninvasive_min',
'apache_2_bodysystem_Neurologic', 'apache_3j_bodysystem_Genitourinary',
'd1_spo2_min', 'd1_mbp_max', 'd1_temp_max', 'd1_glucose_max',
'h1_sysbp_noninvasive_max', 'icu_stay_type_readmit', 'icu_admit_source_Other ICU',
'h1_resprate_max', 'apache_3j_bodysystem_Cardiovascular',
'solid_tumor_with_metastasis', 'apache_2_bodysystem_Haematologic', 'ethnicity_Asian',
'icu_stay_type_admit', 'icu_stay_type_transfer',
'apache_2_bodysystem_Gastrointestinal',
'ethnicity_Caucasian', 'd1_heartrate_min', 'd1_potassium_max',
'd1_glucose_min', 'h1_mbp_min', 'h1_heartrate_min',
'apache_3j_bodysystem_Gynecological', 'lymphoma',
'ethnicity_African American', 'apache_2_bodysystem_Metabolic',
'icu_type_SICU', 'apache_2_bodysystem_Cardiovascular'

Models Used

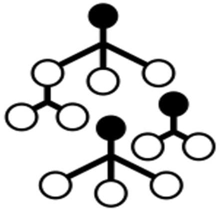
Logistic Regression

A statistical analysis method to predict a binary outcome based on prior observations of dataset



Decision Tree

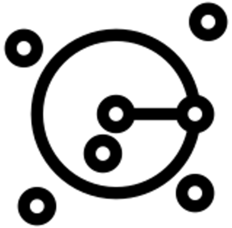
A supervised learning model composed of a set of conditions and leaves organized hierarchically



Random Forest

A classification algorithm consisting of many decisions trees.

Uses bagging and feature randomness to create an uncorrelated forest of trees



KNN

A type of classification where the function is only approximated locally and all computation is deferred until function evaluation.