PREDICTING PATIENT MORTALITY AFTER COVID-19 HOSPITALIZATION

SARAH WARDA

THE PROBLEM AND IMPORTANCE

- During the height of the pandemic in 2021, hundreds of thousands of patients died from Covid-19
- Although medical professionals could narrow down risk factors, some patients without these risk factors died after hospitalization unexpectedly
- Hospitals suffered shortages in PPE, hospital beds, medicine, staff, and various medical supplies
- Hospitals attempted to prioritize patients with the greatest likelihood of survival

- ***What risk factors and diagnostic features can be used to classify patient outcomes as a result of Covid-19, and can we use this model to predict patient death?***

ETL : DATA SOURCE

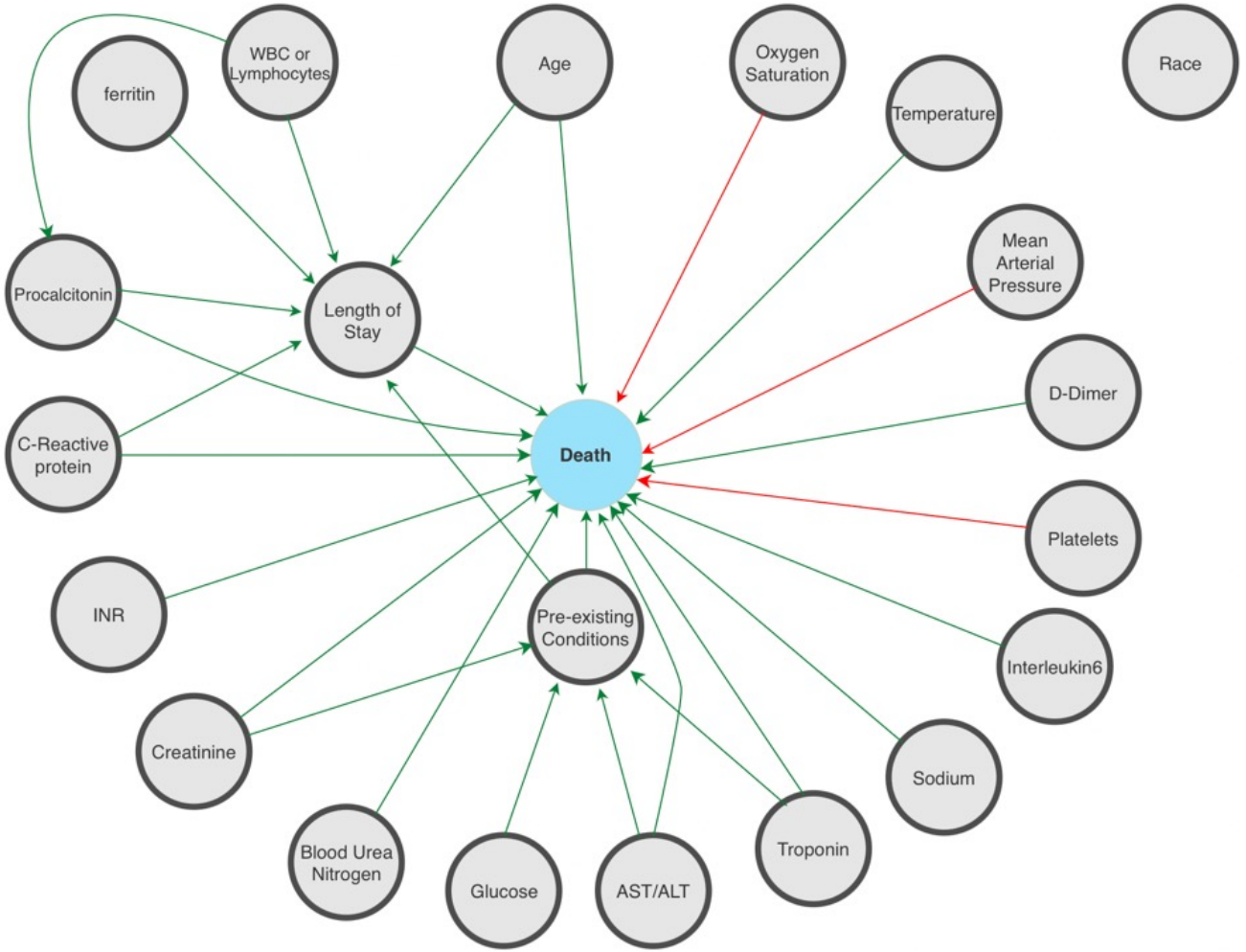
- Kaggle Link or fig-share link from publication:
  - Kaggle Link
  - Publication (2021) Data Link
- The original, anonymized data set contained 85 variables and 4,711 patients/rows
- 43 of the variables were untransformed, raw data
- The data consisted of allowable patient characteristics, pre-existing or current conditions, and blood panel results

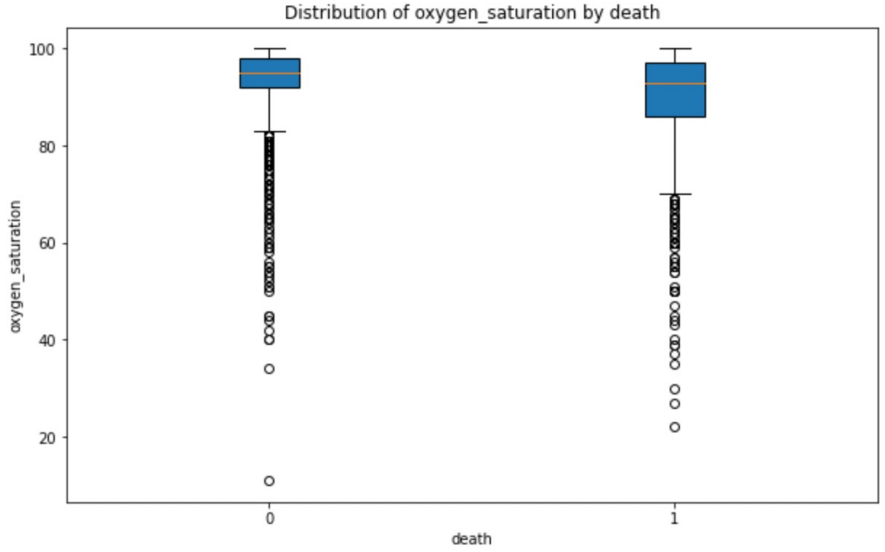
ETL : DATA STORAGE AND LOADING

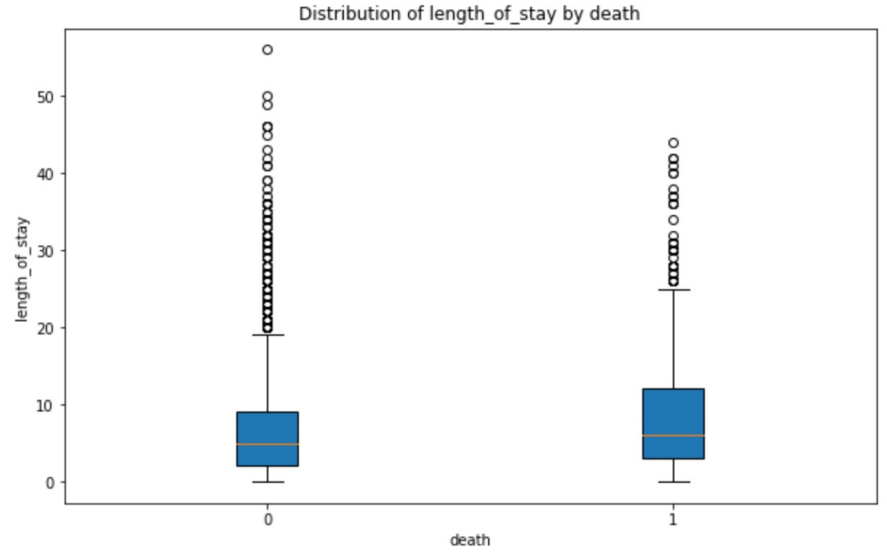
- 43 variables and all 4,711 instances were extracted from the excel file via the .py file and stored in a sqlite database
- The SQL database includes two tables:
  - **blood**: 21 variables
  - **patient**: 22 variables
- Queried database for necessary columns from both tables with a join statement for EDA

EDA : CAUSAL LOOP DIAGRAM AND DOMAIN KNOWLEDGE

- Many of the variables are expected to have a direct influence on *death*
- Length of stay and all pre-existing conditions are influenced by many other variables
- Normal ranges for all blood panel results were researched and cited for reference during EDA
- Most patients had abnormal levels in blood panels, but patients who died had greater abnormalities

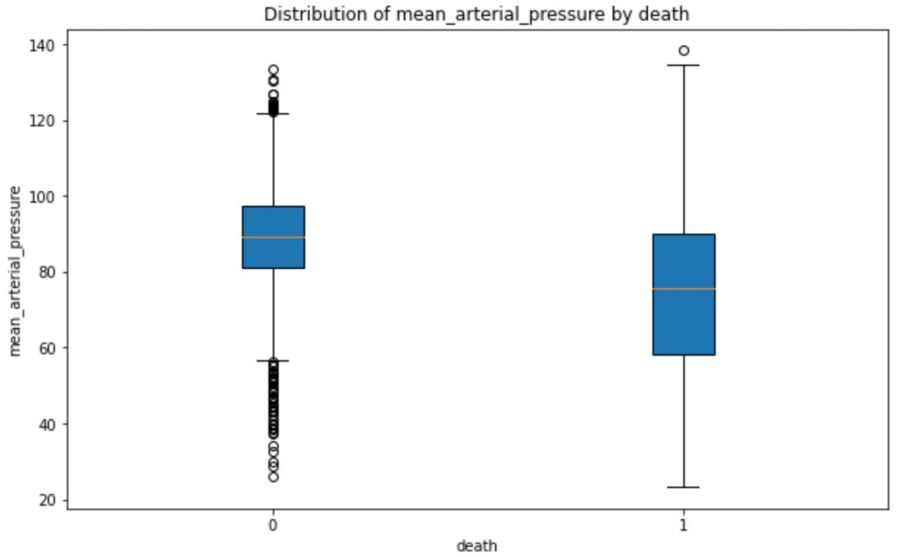


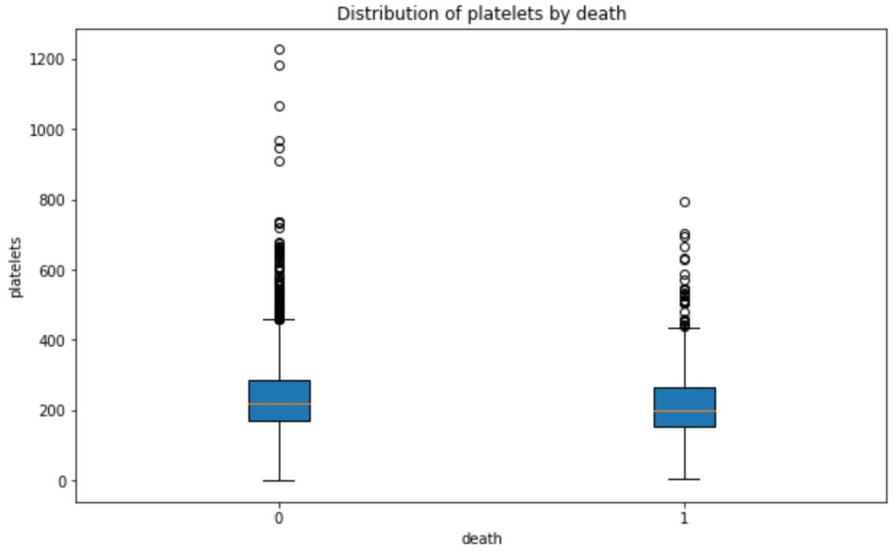


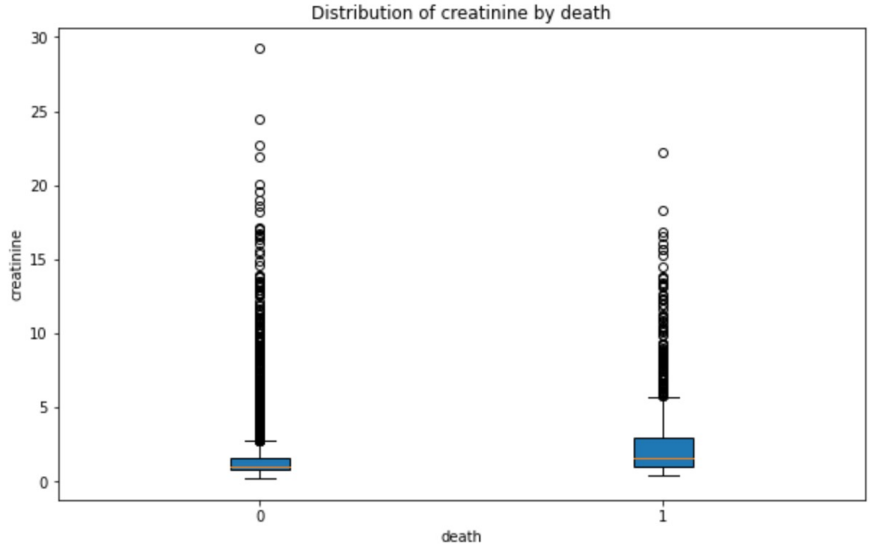


EDA : SIGNIFICANT RELATIONSHIPS

- Some significant patient characteristics

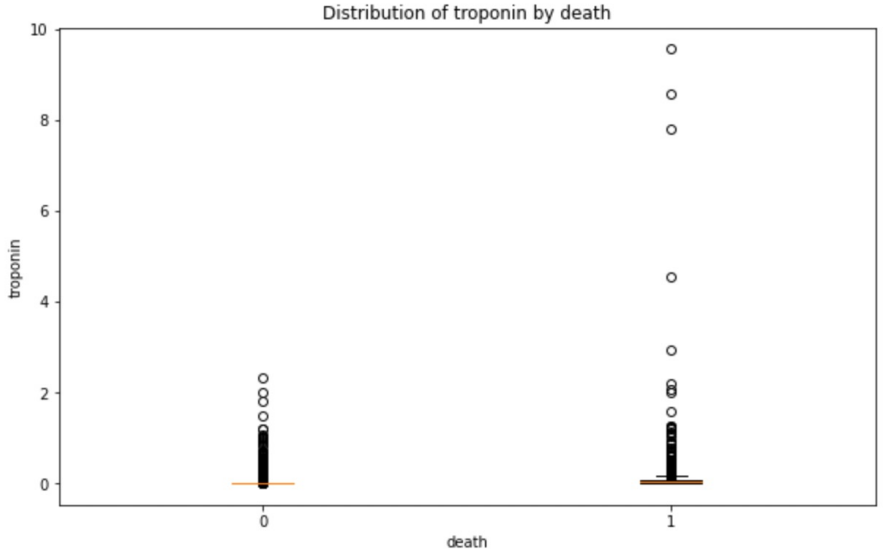






EDA : SIGNIFICANT RELATIONSHIPS

- Some significant blood panel results



MODELING

- Null Model:
  - $P(death = 1) = 0.24$
  - $error\ rate = 0.76$
- We will use Logistic regression, since our target variable is categorical
- Using the results from EDA, we narrowed down 43 variables to 25 variables:
  - 16 numerical variables
  - 9 categorical variables

	feature	r	rho
0	age	0.2906	0.2998
1	oxygen_saturation	-0.2188	-0.1771
2	temp_F	0.0184	0.0154
3	mean_arterial_pressure	-0.3715	-0.3048
4	D_dimer	0.1777	0.1245
5	platelets	-0.0803	-0.0883
6	creatinine	0.1385	0.2431
7	sodium	0.1205	0.0832
8	aspartate_aminotransferase	0.0952	0.1654
9	alanine_aminotransferase	0.0555	0.0198
10	white_blood_cell	0.0635	0.0940
11	interleukin6	0.0407	0.0608
12	C_reactive_protein	0.2173	0.1782
13	procalcitonin	0.1716	0.1837
14	troponin	0.1139	0.2337
15	length_of_stay	0.1337	0.1499

MODELING : "ALL-IN MODEL"

- The error rate is 17.73% and the pseudo R^2 is 0.29, which is *okay*, but maybe we can do better.
- model seems to classify negative cases relatively efficiently but has an issue with false negatives.

	Actual 1	Actual 0
Predicted 1	432	169
Predicted 0	571	3003

Metrics	Result
Accuracy	0.822754
Error Rate	0.177246
Sensitivity	0.430708
False Negative	0.569292
Specificity	0.946721
False Positive	0.0532787
Precision	0.718802
F1	0.538653

		95% BCI		
Coefficients	Mean	Lo	Hi	P(y=1)
β_0	0.00001	-0.00034	0.00023	0.50000
age β_1	0.05061	0.04580	0.05789	0.01265
myocardial_infarction β_2	-0.00004	-0.00451	0.00221	-0.00001
peripheral_vascular_disease β_3	-0.00412	-0.02529	-0.00187	-0.00103
congestive_heart_disease β_4	0.00035	-0.00316	0.00516	0.00009
chronic_obstructive_pulmonary_disease β_5	0.00034	-0.00161	0.00577	0.00009
all_central_nervous_system_disease β_6	0.00210	0.00055	0.01623	0.00053
diabetes_mellitus_simple β_7	-0.00023	-0.00817	0.00216	-0.00006
renal_disease β_8	0.00196	0.00011	0.01464	0.00049
stroke β_9	0.00070	0.00024	0.00630	0.00017
seizure β_{10}	0.00019	-0.00028	0.00300	0.00005
oxygen_saturation β_{11}	-0.03848	-0.05455	-0.03292	-0.00962
temp_F β_{12}	0.00629	-0.00843	0.02799	0.00157
mean_arterial_pressure β_{13}	-0.04463	-0.05095	-0.03854	-0.01116
D_dimer β_{14}	0.01366	-0.00419	0.02928	0.00341
platelets β_{15}	-0.00193	-0.00288	-0.00121	-0.00048
creatinine β_{16}	0.01915	0.00858	0.07258	0.00479
sodium β_{17}	0.01286	0.00053	0.02290	0.00322
aspartate_aminotransferase β_{18}	0.00067	0.00012	0.00299	0.00017
alanine_aminotransferase β_{19}	0.00053	-0.00349	0.00102	0.00013
white_blood_cell β_{20}	0.00180	-0.00853	0.01418	0.00045
interleukin6 β_{21}	0.00001	-0.00002	0.00069	0.00000
C_reactive_protein β_{22}	0.01898	0.00945	0.02663	0.00474
procalcitonin β_{23}	0.03190	0.01874	0.04783	0.00798
troponin β_{24}	0.00288	0.00097	0.02023	0.00072
length_of_stay β_{25}	0.03426	0.02170	0.04632	0.00856

Metrics	Mean	Lo	Hi
Error (%)	17.72455	15.96240	18.73737
Efron's R^2	0.29478	0.26928	0.33443

MODELING : FINAL MODEL

- Additional transformations, besides the ones included in the final model, did not improve the model
- Other interaction terms did not improve the model
- Compared to the all-in model, we can identify more true positives, but its not perfect
- Some coefficient Interpretations (detailed description can be found in linear model notebook):
  - For every one-year increase in a patients age, the log odds of death increases by 0.0467. The coefficient sign is expected and is strongly supported by the 95% BCI. Using the divide by four rule, the probability of death, or P(death = 1), increases by 1.17 percentage points for each year increase.
  - For every 1% increase in the creatinine level, when the patient does not have renal disease, the log odds of death increases by 0.2919. Using the divide by four rule, the probability of death, or P(death = 1), increases by 7.30 percentage points for every 1% increase in creatinine level, when the patient does not have renal disease.
  - For every one unit increase in the AST_ALT ratio, the log odds of death increases by 0.2542. Using the divide by four rule, the probability of death, or P(death = 1), increases by 6.35 percentage points for every one unit increase in the AST_ALT ratio.

	Actual 1	Actual 0
Predicted 1	462	166
Predicted 0	541	3006

Metrics	Result
Accuracy	0.830659
Error Rate	0.169341
Sensitivity	0.460618
False Negative	0.539382
Specificity	0.947667
False Positive	0.0523329
Precision	0.735669
F1	0.566524

			95% BCI		
Coefficients		Mean	Lo	Hi	P(y=1)
	β_0	0.0033	-0.0099	0.0132	0.5008
age	β_1	0.0467	0.0397	0.0543	0.0117
congestive_heart_disease	β_2	0.0066	-0.0640	0.0770	0.0017
chronic_obstructive_pulmonary_disease	β_3	0.0213	-0.0333	0.0685	0.0053
all_central_nervous_system_disease	β_4	0.1011	0.0112	0.1852	0.0253
renal_disease	β_5	0.1113	0.0190	0.2658	0.0278
stroke	β_6	0.0497	-0.0002	0.1005	0.0124
oxygen_saturation	β_7	-0.0399	-0.0494	-0.0317	-0.0100
mean_arterial_pressure	β_8	-0.0424	-0.0493	-0.0372	-0.0106
D_dimer	β_9	0.0104	-0.0040	0.0257	0.0026
platelets	β_{10}	-0.0017	-0.0026	-0.0009	-0.0004
creatinine_log	β_{11}	0.2919	0.1515	0.3604	0.0730
sodium	β_{12}	0.0147	0.0097	0.0216	0.0037
AST_ALT	β_{13}	0.2542	0.1909	0.3851	0.0635
C_reactive_protein	β_{14}	0.0206	0.0123	0.0274	0.0052
procalcitonin	β_{15}	0.0225	0.0106	0.0380	0.0056
troponin_sq	β_{16}	0.2972	0.0704	0.4691	0.0743
length_of_stay	β_{17}	0.0357	0.0226	0.0465	0.0089
creatinine_log:renal_disease	β_{18}	0.1719	0.0548	0.3180	0.0430

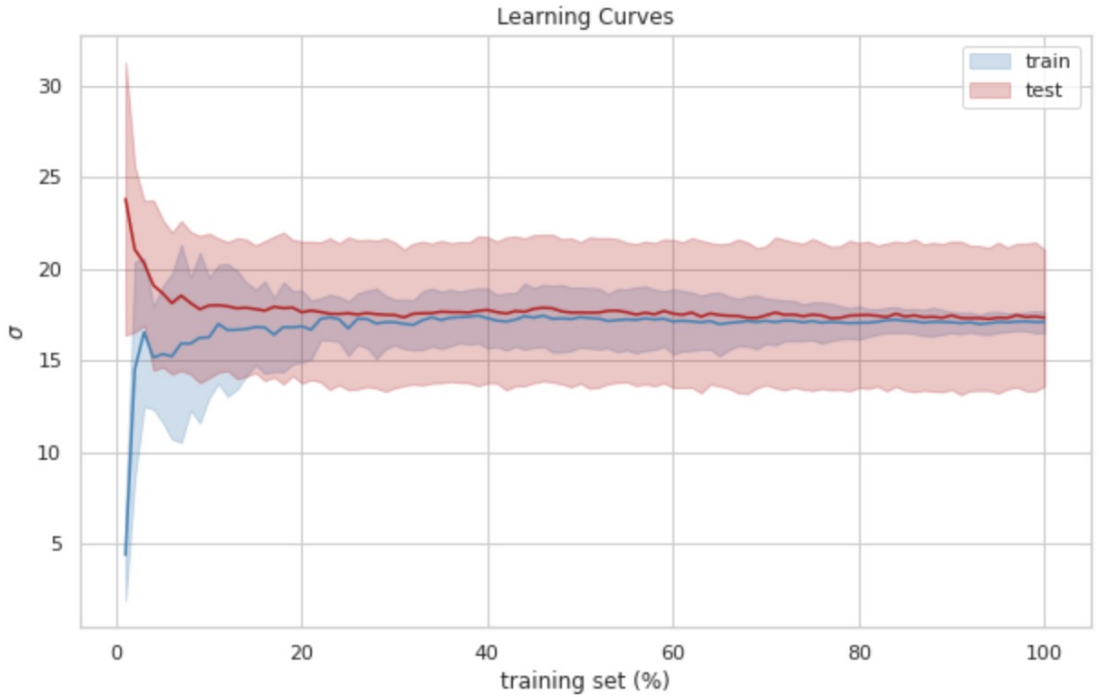
Metrics	Mean	Lo	Hi
Error (%)	16.9341	15.8802	18.3804
Efron's R^2	0.3098	0.2764	0.3462

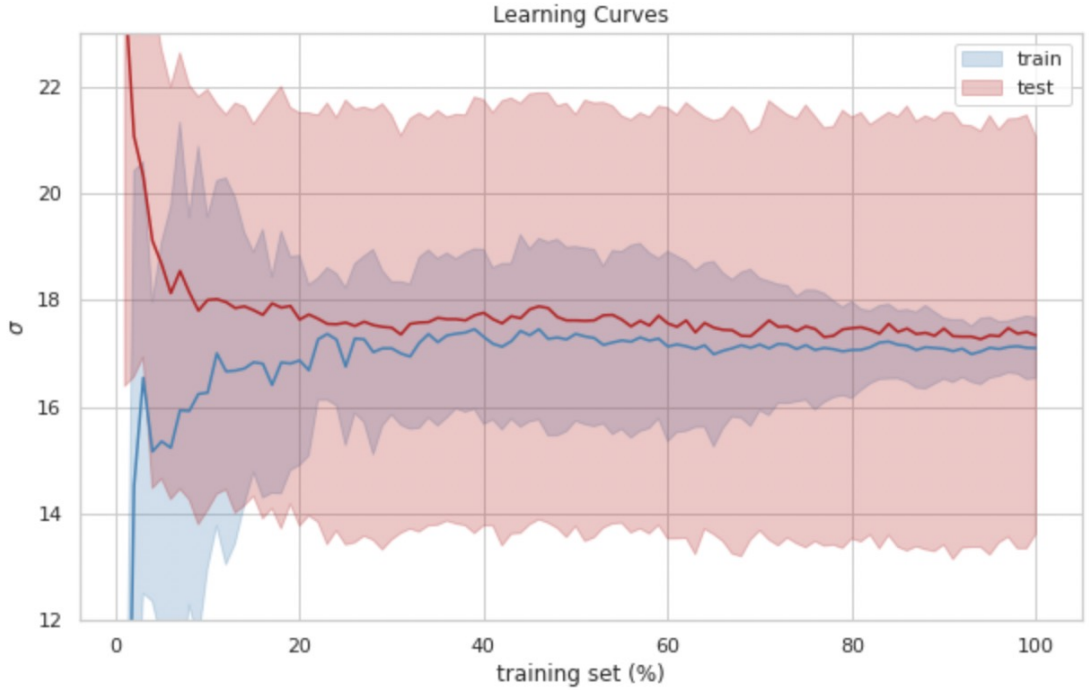
EVALUATING THE MODEL : CROSS VALIDATION

- Three rounds of 10-Fold Cross Validation
- Estimated credible bounds for metrics, *given the data*:
  - 95% CI for error rate: 14.36, 19.19
  - 95% CI for Efron's R^2 : 0.20, 0.40
- Average Performance of the model, *given the data*:
  - estimated mean error rate of 17.23%
  - estimated mean Efron's R^2 of 0.31
  - 95% CI for *mean* error rate: 16.70, 17.67
  - 95% CI for *mean* Efron's R^2 : 0.29, 0.32

EVALUATING THE MODEL : LEARNING CURVES

- Curves are well converged, so we are not dealing with a high variance scenario
- if we refer back to the null model, the error rate was 76%. In comparison to the null model, we are not dealing with a high bias scenario.
- For our *application*, this amount of error seems tolerable





USING THE LOGISTIC REGRESSION MODEL: PREDICTION #1

- Let's say we have a 64 year old patient who has COPD and the following physical and blood panel results:
  - OSat: 91%
  - MAP: 80 mm Hg
  - D-Dimer: 3.5 u/mL
  - platelets: 210 platelets/uL
  - creatinine: 1.9 mg/dL
  - sodium: 140 mEq/L
  - AST/ALT: 1.5
  - c-reactive protein: 8 mg/dL
  - procalcitonin: 1.3 ng/mL
  - troponin: 0.06 ng/mL
  - length of stay: 14 days

- Result:
  - Patient outcome: 0.0
  - P(death = 0): 0.74
  - P(death = 1): 0.26

USING THE LOGISTIC REGRESSION MODEL: PREDICTION #2

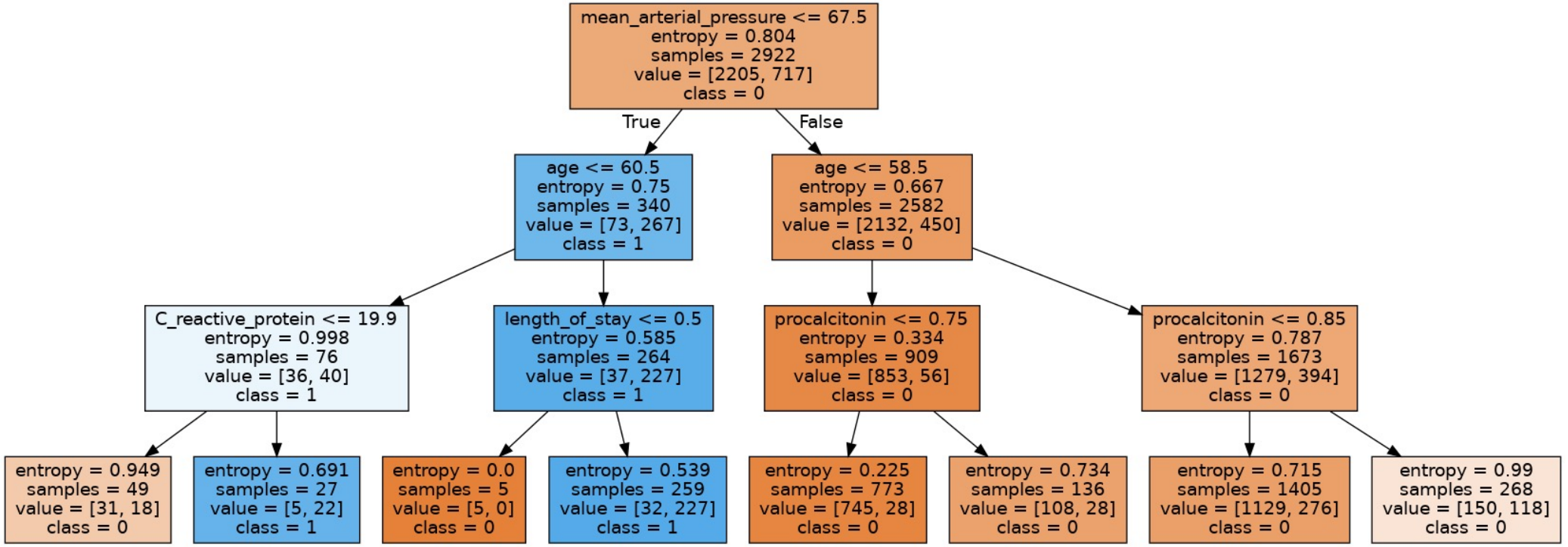
- Let's say we have a 40 year old patient who has renal disease and the following physical and blood panel results:
  - OSat: 82%
  - MAP: 60 mm Hg
  - D-Dimer: 5.1 u/mL
  - platelets: 170 platelets/uL
  - creatinine: 4.3 mg/dL
  - sodium: 150 mEq/L
  - AST/ALT: 1.8
  - c-reactive protein: 17 mg/dL
  - procalcitonin: 3.1 ng/mL
  - troponin: 1.8 ng/mL
  - length of stay: 8 days

- Result:
  - Patient outcome: 1.0
  - P(death = 0): 0.24
  - P(death = 1): 0.76

USING THE LOGISTIC REGRESSION MODEL: PREDICTION #3

- Let's say we have a 40 year old patient with no pre-existing conditions and the following physical and blood panel results
  - OSat: 82%
  - MAP: 57 mm Hg
  - D-Dimer: 5.0 u/mL
  - platelets: 200 platelets/uL
  - creatinine: 1.5 mg/dL
  - sodium: 130 mEq/L
  - AST/ALT: 0.5
  - c-reactive protein: 17 mg/dL
  - procalcitonin: 3.0 ng/mL
  - troponin: 2 ng/mL
  - length of stay: 8 days

- Result:
  - Patient outcome: 0.0
  - P(death = 0): 0.53
  - P(death = 1): 0.47

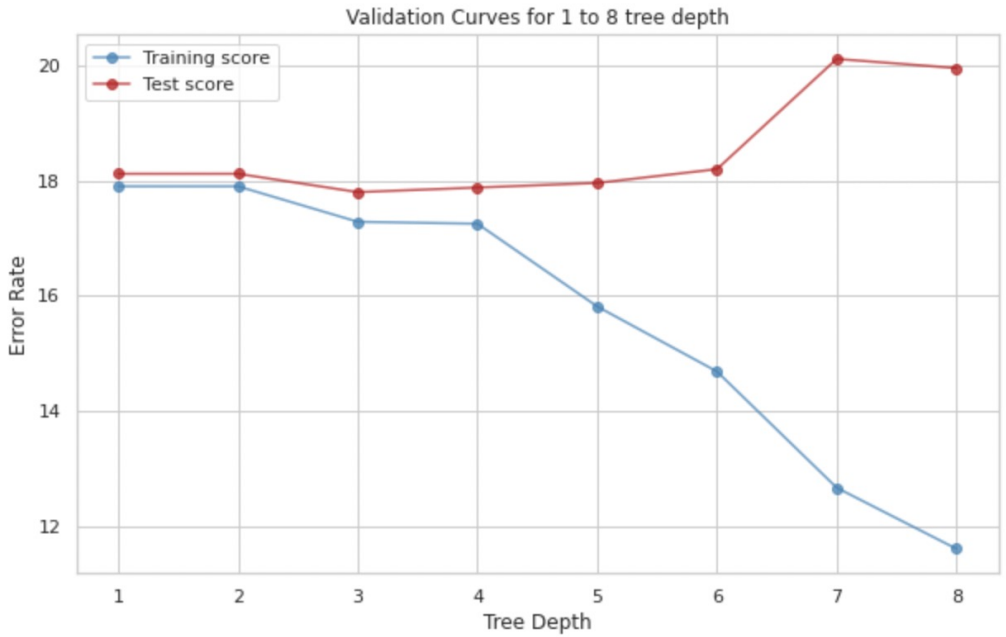


DECISION TREE CLASSIFIER

- To start, we implemented a decision tree classifier with a max depth of 3
  - Accuracy: 82.20
  - Error rate: 17.80

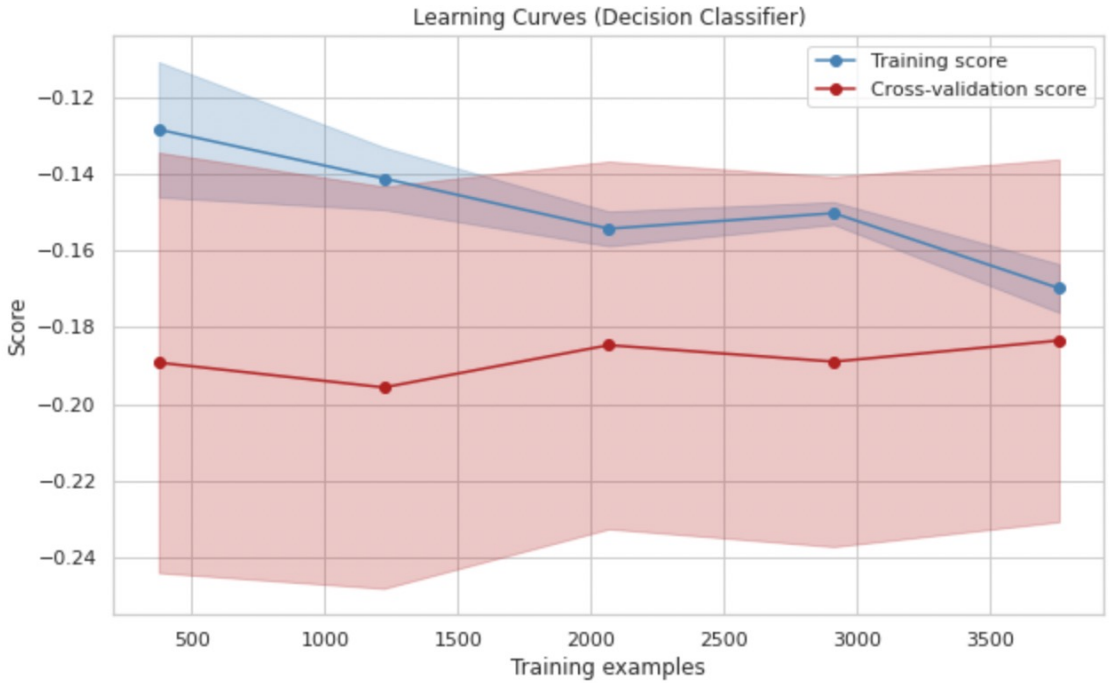
DECISION TREE CLASSIFIER : VALIDATION CURVES

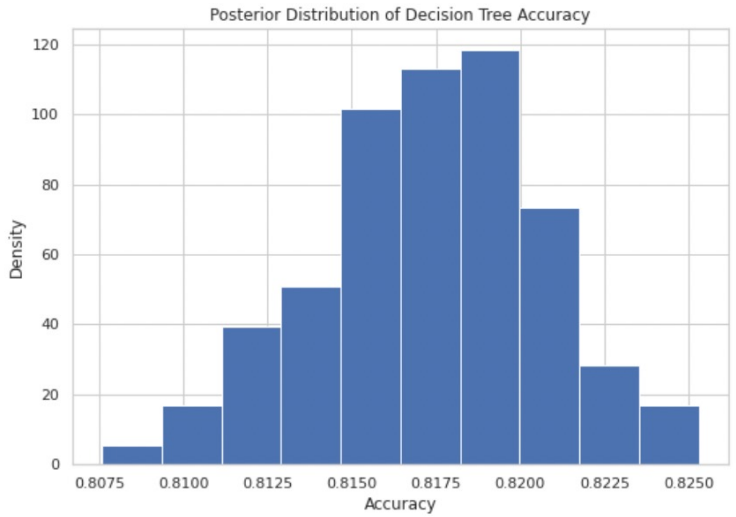
- Validation curve implemented to confirm optimal max depth
- Based on results, max depth of 3 optimizes the bias/variance trade off
- Greater depths, especially over 4, overfit the model

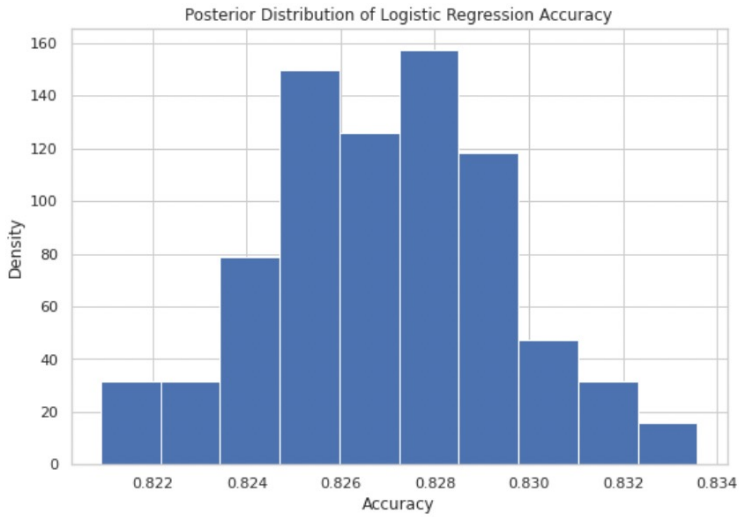


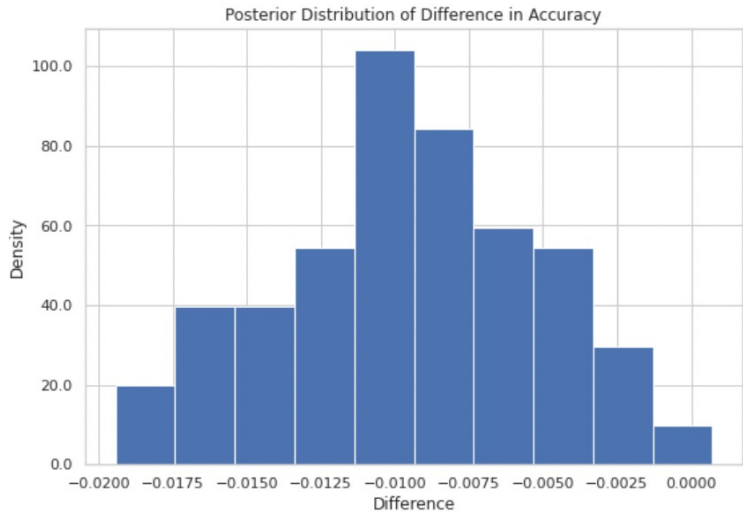
DECISION TREE CLASSIFIER : LEARNING CURVES

- Model may suffer from high variance and high bias
  - we still have a gap between the curves, although it seems to start converging right at the very end.
  - Error higher than we achieved with logistic regression model
- we may need to modify the features used in the model, re-evaluate the number of features, and consider more data.









CROSS VALIDATION AND COMPARISON OF THE MODELS

- Decision tree mean cross validation score: 0.82
- Given the data, unlikely that decision tree model performs better:
  - P(DT > LGR) 0.01
  - P(LGR > DT) 0.99

Thank you! ☺