Physionet 2020 - Predicting patient survival in ICU

Khushbu Desai • Sruthi Gopalakumaran • Priyanka Sharma

Mentors: Prof. Patricia & Tom Brander

Overview of Competition : Physionet 2020

- Kaggle Competition: **Women in Data Science** Datathon (WIDS 2020)
- Timeline: January 30th - February 24th **(3 weeks)**
- Dataset: **130,000** hospital Intensive Care Unit **(ICU) visits** from patients in **one-year** timeframe
    - From: Argentina, Australia, New Zealand, Sri Lanka, Brazil, and more than **200 hospitals in the United States**
- Model: to predict **patient survival** in the **ICU** using data **of first 24hrs**
    - hospital_death **= 0 (survived)** and **1 (death)**
    - Training set: 186 variables and 91,713 samples
- Submissions: are **evaluated** on the **AUC** curve **between** the **predicted mortality** and **hospital_death**
    - Graded using only 50% of testing data (final scores were different)
    - 2 final submissions for judging

WOMEN IN DATA SCIENCE
@STANFORD UNIVERSITY

Exploratory Data Analysis

Exploratory Data Analysis

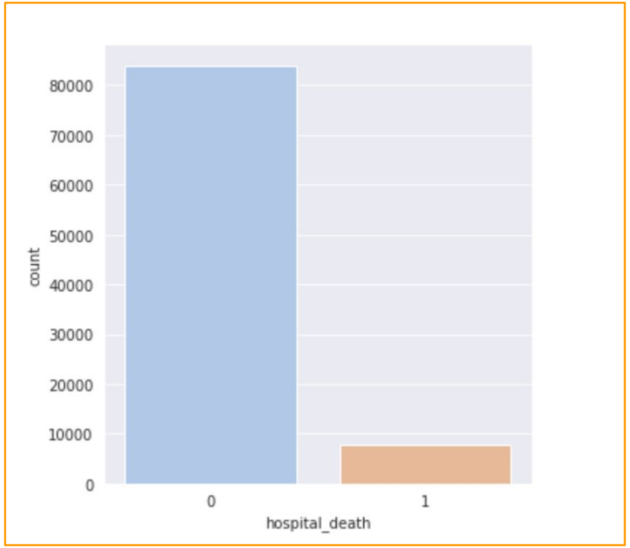
- Understanding which are **important** out of 186 variables:
  - Between **similar or same tests**
  - **H1 vs D1** for each lab test
  - **Apache** scores
  - **Researching** lab test

Snapshot of Variables

Apache Variables	
hematocrit_apache	Gcs_unable_apache
intubated_apache	Gcs_verbal_apache
map_apache	Glucose_apache
paco2_apache	Heart_rate_apache
paco2_for_ph_apache	Wbc_apache
pao2_apache	Apache_post_operative
ph_apache	Arf_apache
resprate_apache	Bun_apache
sodium_apache	Creatinine_apache
temp_apache	Fio2_apache
urineoutput_apache	Gcs_eyes_apache
Apache_2_diagnosis	Gcs_motor_apache
Apache_3j_diagnosis	Ph_apache
Apache_post_operative	Ventilated_apache

Blood Pressure Variables	
d1_diasbp_invasive_max	h1_diasbp_invasive_min
d1_diasbp_invasive_min	h1_diasbp_max
d1_diasbp_max	h1_diasbp_min
d1_diasbp_min	h1_diasbp_noninvasive_ma
d1_diasbp_noninvasive_max	h1_diasbp_noninvasive_min
d1_diasbp_noninvasive_min	h1_mbp_invasive_max
d1_mbp_invasive_max	h1_mbp_invasive_min
d1_mbp_invasive_min	h1_mbp_max
d1_mbp_max	h1_mbp_min
d1_mbp_min	h1_mbp_noninvasive_max
d1_mbp_noninvasive_max	h1_mbp_noninvasive_min
D1_mbp_noninvasive_min	h1_sysbp_invasive_max
d1_sysbp_invasive_max	h1_sysbp_invasive_min
D1_sysbp_invasive_min	h1_sysbp_max
d1_sysbp_max	h1_sysbp_min
d1_sysbp_min	h1_sysbp_noninvasive_max
D1_sysbp_noninvasive_max	h1_sysbp_noninvasive_min
D1_sysbp_noninvasive_min	h1_diasbp_invasive_max
D1_sysbp_noninvasive_min	
h1_diasbp_invasive_max	

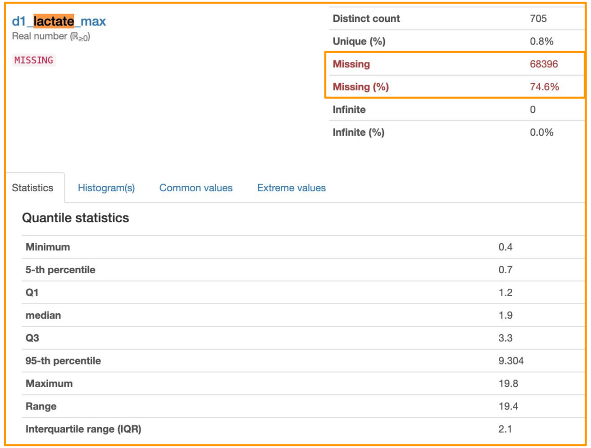
Imbalanced



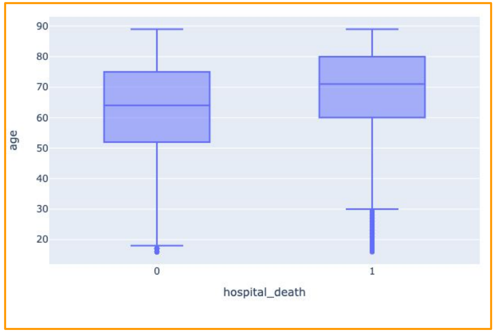
Exploratory Data Analysis cont.

- **Importance plots**
  - (RF vs. XGB)
- Using **pandas**:
  - Pandas profiling
  - Importance tables (RF vs. XGB)
  - Correlation matrix
- Using **Google Colab**
  - Bar and boxplot distributions for each variable
- Using **Clinical Knowledge**
  - Max versus Min value for labs
  - Expected range of values
  - Right censored data

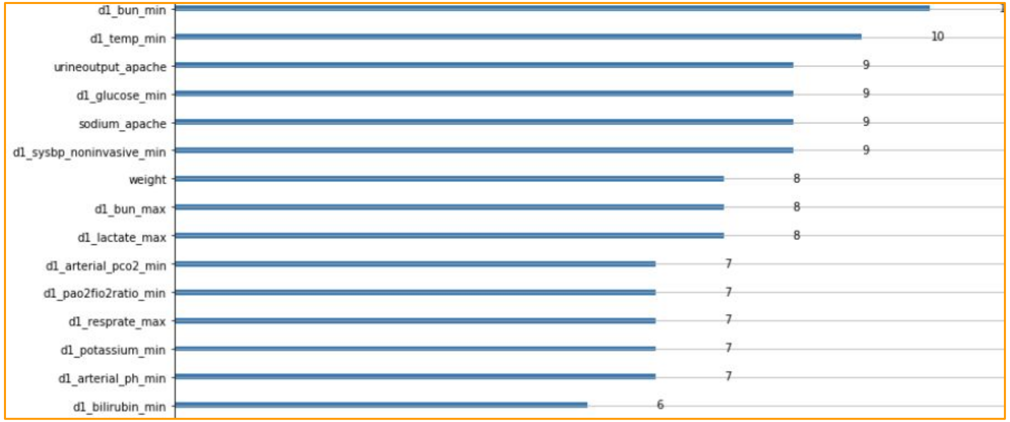
Pandas profiling



Google Colab



XGB Importance plot



Feature Selection And Feature Engineering

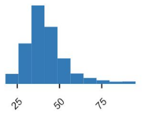
Feature Selection

Paco2_for_ph_apache: 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**

Paco2_apache: 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_apache
Real number ($\mathbb{R}_{\geq 0}$)
MISSING

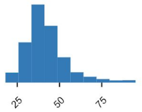
Distinct count	705	Mean	42.18323819
Unique (%)	0.8%	Minimum	18
Missing	70868	Maximum	95
Missing (%)	77.3%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	716.6 KiB



Toggle details

paco2_for_ph_apache
Real number ($\mathbb{R}_{\geq 0}$)
MISSING

Distinct count	705	Mean	42.18323819
Unique (%)	0.8%	Minimum	18
Missing	70868	Maximum	95
Missing (%)	77.3%	Zeros	0
Infinite	0	Zeros (%)	0.0%
Infinite (%)	0.0%	Memory size	716.6 KiB



Toggle details

Feature Selection

Readmission had only a single value for the entire dataset

readmission_status
Boolean

CONST

Distinct count	1
Unique (%)	< 0.1%
Missing	0
Missing (%)	0.0%
Memory size	716.6 KiB

Impossible Values

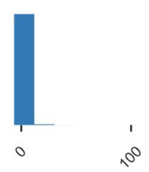
Pre ICU Length of Stay has a minimum of -24.9 days.

pre_icu_los_days
Real number (ℝ)

ZEROS

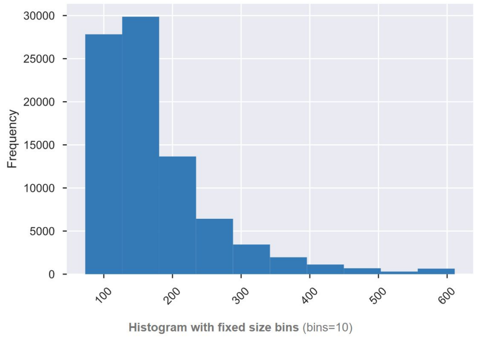
Distinct count	9757
Unique (%)	10.6%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	0.8357660507
Minimum	-24.94722222
Maximum	159.0909722
Zeros	3711
Zeros (%)	4.0%
Memory size	716.6 KiB



Extreme Values: Are they impossible?

Glucose level that is considered to be normal is 90 - 140 mg/dL. This dataset had a maximum of 611 and minimum of 33



Value	Count	Frequency (%)
611	423	0.5%
610	1	< 0.1%
609	2	< 0.1%
608	3	< 0.1%
607	3	< 0.1%

Feature Engineering

- Label Encoding
- Comorbidity Column
- Creation of a new Column for each lab test

Label Encoding

Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form.

ICU STAY TYPE	LABEL ENCODED ICU STAY TYPE
Med-Surg ICU	0
Med-Surg ICU	0
CCU-CTICU	1
Med-Surg ICU	0
Neuro ICU	2

Comorbidity Column

New Column for each test

BS	BT	BU
d1_temp_max	d1_temp_min	temp_new
39.9	37.2	2
36.3	35.1	2
37	37	1
38	34.8	2
37.2	NA	1
NA	NA	0

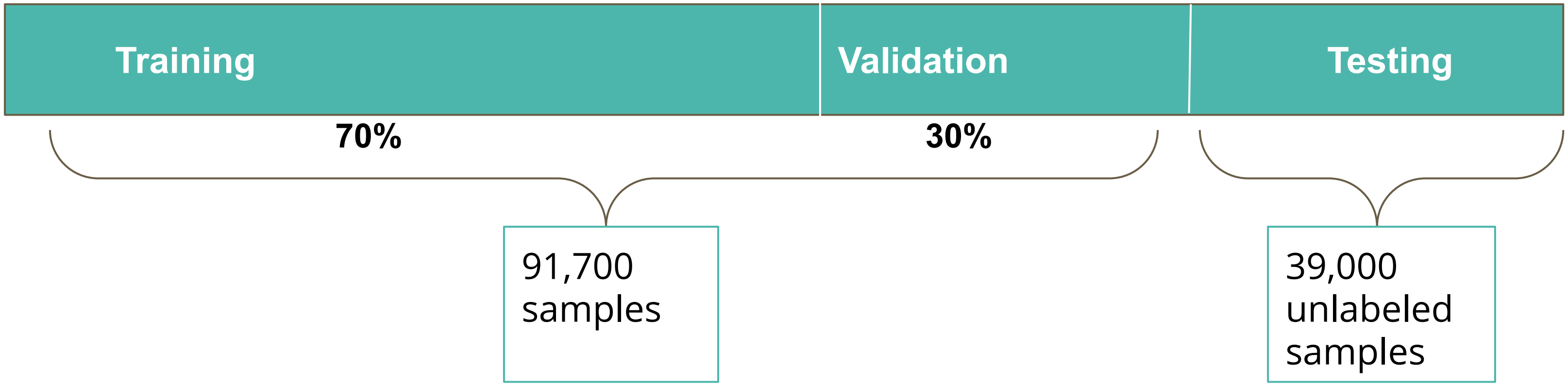
Imputation Technique

The dataset contained 45 columns with more than 75% data missing. We handled the missing values by:

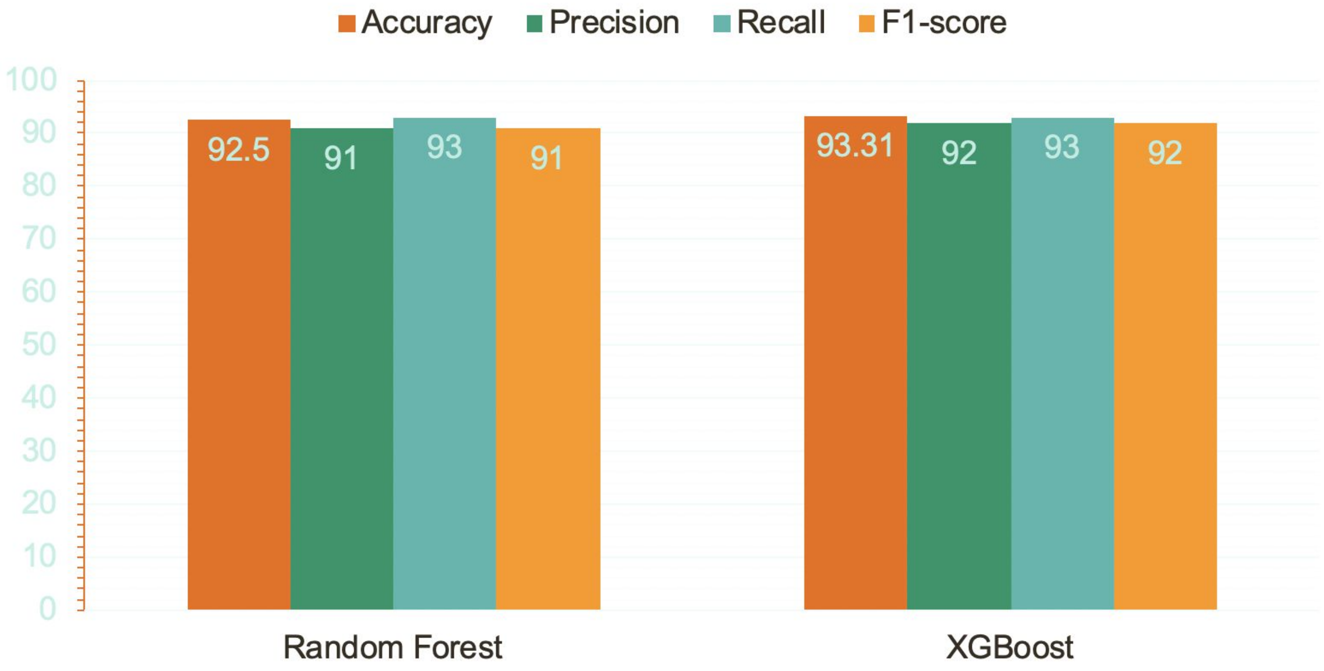
- Imputing by Mean/ Median
- Filled in NAs with “Normal/Typical” values observed in patients

Modeling

Split of our data



Comparison for base model



* Random Forest + SMOTE performance < XGBoost performance

Approach

Algorithm

- XGBoost

Data Pre-processing

- Imputing
- Label Encoding

Different Models

- Model 1 - Feature engineering - EDA
- Model 2 - Top 70 features using Importance plots
- Model 3 - Top 30 features using Importance plot

Improving our model

Model 1 (Added derived variables)

Imputing	•Imputing with mean/median •Imputing with normal values
Feature Engineering	•Adding a new variable – combination of multiple variable •Total 106 variables (old + new)
Parameter Tuning	• Grid search CV • Class weight (class imbalance) • Other parameters tuned were related to tree complexity

Model 2 (Top 70) & Model 3 (Top 30)

Imputing	•Imputing with mean/median
Feature engineering	•Model 2 - Top 70 features - XGBoost+RF importance plots •Model 3 - Top 30 features - XGBoost Importance plot
Parameter Tuning	• Grid Search CV • Class weight (class imbalance) • Other parameters tuned related to tree complexity

Winning Team
AUC =0.9149

Model Evaluation Metrics

Evaluation metrics on validation data:

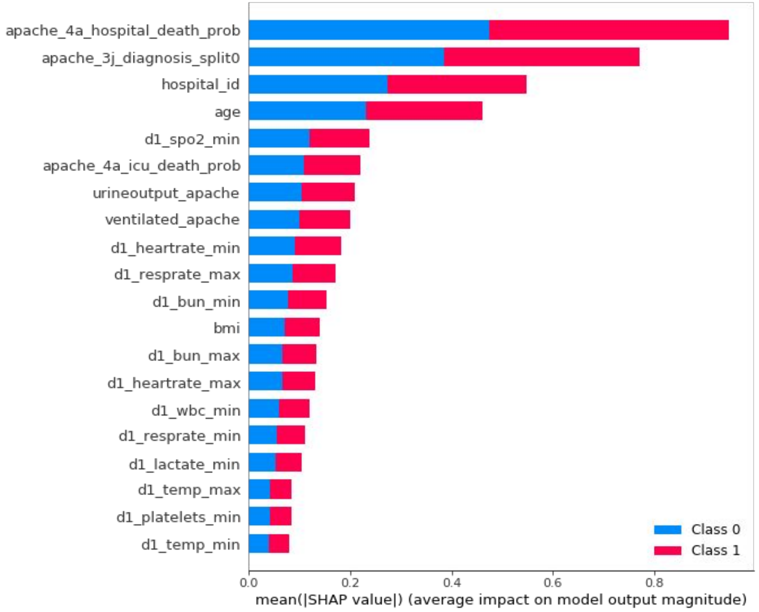
Model / Metrics	Model 1	Model 2	Model 3
Accuracy	0.9327	0.9345	0.9303
Precision	0.92	0.93	0.92
Recall	0.93	0.93	0.93
F1-score	0.92	0.92	0.92

AUC for Test Data:

1. **Model 2**
   Top 70
   AUC = 0.9045
2. **Model 1**
   New derived variables
   AUC = 0.90382
3. **Model 3**
   Top 30
   AUC = 0.89355

Interesting approaches used by Top 3 winners

- Importance plots based on class distribution



- Library (Missingno) : to combine similar features based on missing data
- LGBM or Catboost
- Bayesian Optimization for hyper parameter Tuning
- Adversarial Validation to further drop variables in the model

Thank You

Questions?