Predicting ICU Patient Mortality

Based on non-physiological features:
Statistical and Machine Learning Approaches

Christopher Forbath
ID 501168902

Ryerson University

Introduction

Predicting ICU Patient Mortality based on non-physiological features: Statistical and Machine Learning Approaches

Christopher Forbath

ID 501168902

December 5, 2022

Objectives

- Develop sparse statistical and machine learning models to predict probability of ICU Patient Mortality based on non-physiological features (i.e. not dependent on laboratory analysis or acquisition by sensors)
- Evaluate and compare models' performance
- Compare with established physiologically-based APACHE IV statistical model
- Identify issues, limitations and opportunities for further exploration and analysis.

Research Questions

1. How well can various models based on non-physiological features predict the probability of in-hospital mortality?
2. Of those evaluated in this study, how do the statistical and machine learning models perform relative to each other and with an established physiological-based one?
3. How do the models perform given the imbalanced dataset, where the minority class accounts for only 8.4% of the observations? How do the models perform using oversampled data to address the imbalance?

Applications and Ethical Considerations

The use of statistical and machine learning modeling to predict the probability of patient mortality is well established and has several applications, including:

- Criteria for enrollment in clinical studies
- Assessing relative risk of patients to determine level of monitoring and possible intervention required
- Evaluating individual hospital / ICU aggregate performance relative to broadly expected outcomes

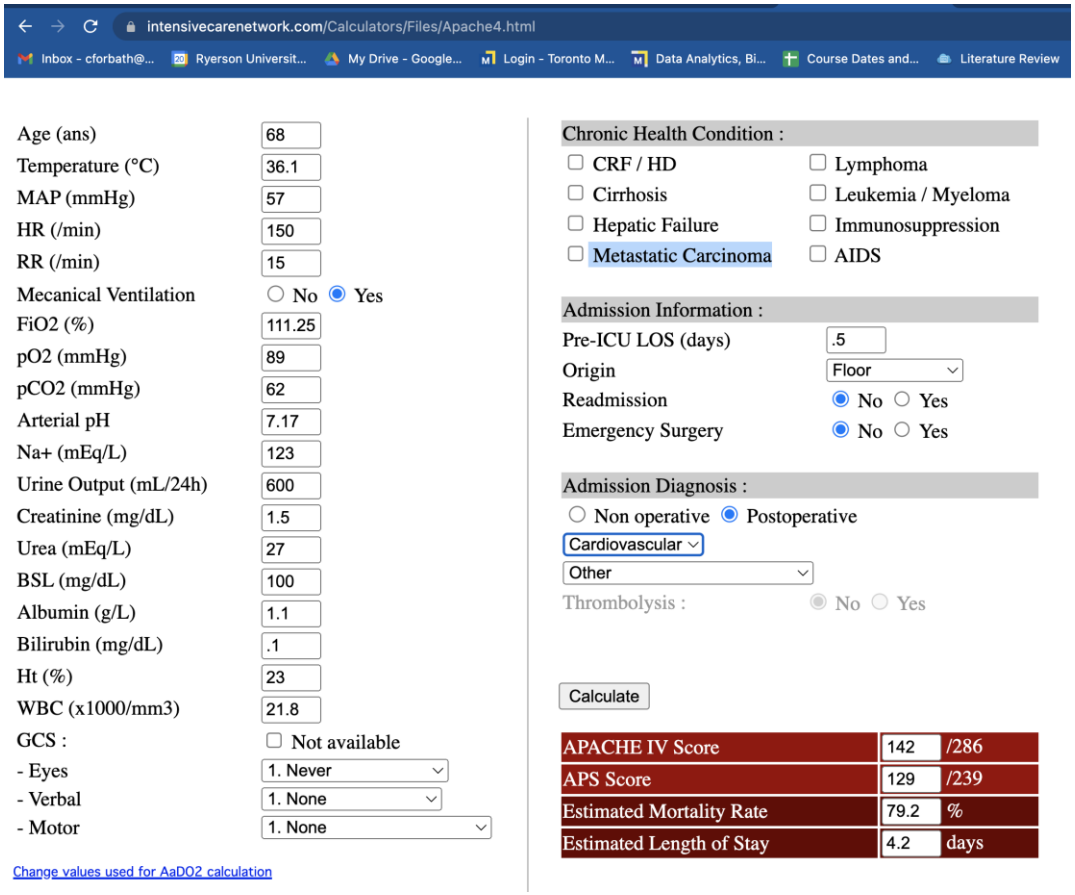
These can all be subject to potential ethical pitfalls if insufficient consideration is given to unexpected consequences and/or potential misuse or abuse. There are also many subjective, cultural and highly emotive factors that come into play in the treatment of the critically ill.

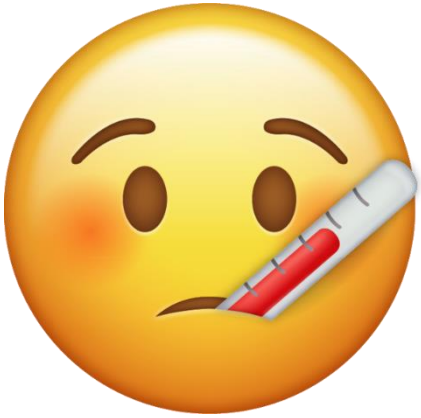
Sample ethical issues: Should a patient mortality probability prediction be factored into resource allocation ($, staffing, ICU bed, etc.)? In organ donations? Determining who gets into experimental clinical trials?

The dataset used in this study was already fully anonymized re: patient and hospital identity.

APACHE IV: most widely used model in USA

Logistic Regression based primarily on physiological features (39 total)





APACHE IV performance

Performance Measure	Metric	APACHE VERSION	
		IVa	III Hc
Discrimination	AUROC	0.880	0.868
Calibration	Hosmer-Lemeshow chi-square	16.8 (p = .08)	635.4 (p < .001)
SMR (Standard Mortality Rate)	Observed hospital mortality / Predicted hospital mortality	0.997 (p = .76)	0.799 (p < .001)

Zimmerman, J. E., Kramer, A. A., McNair, D. S., & Malila, F. M. (2006). Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for today's critically ill patients. Critical care medicine, 34(5), 1297–1310. https://doi.org/10.1097/01.CCM.0000215112.84523.F0

Subgroup	AUROC
Coronary artery bypass graft surgery	0.75
Sepsis	0.79
Gastrointestinal bleeding	0.81
Ventilated	0.83

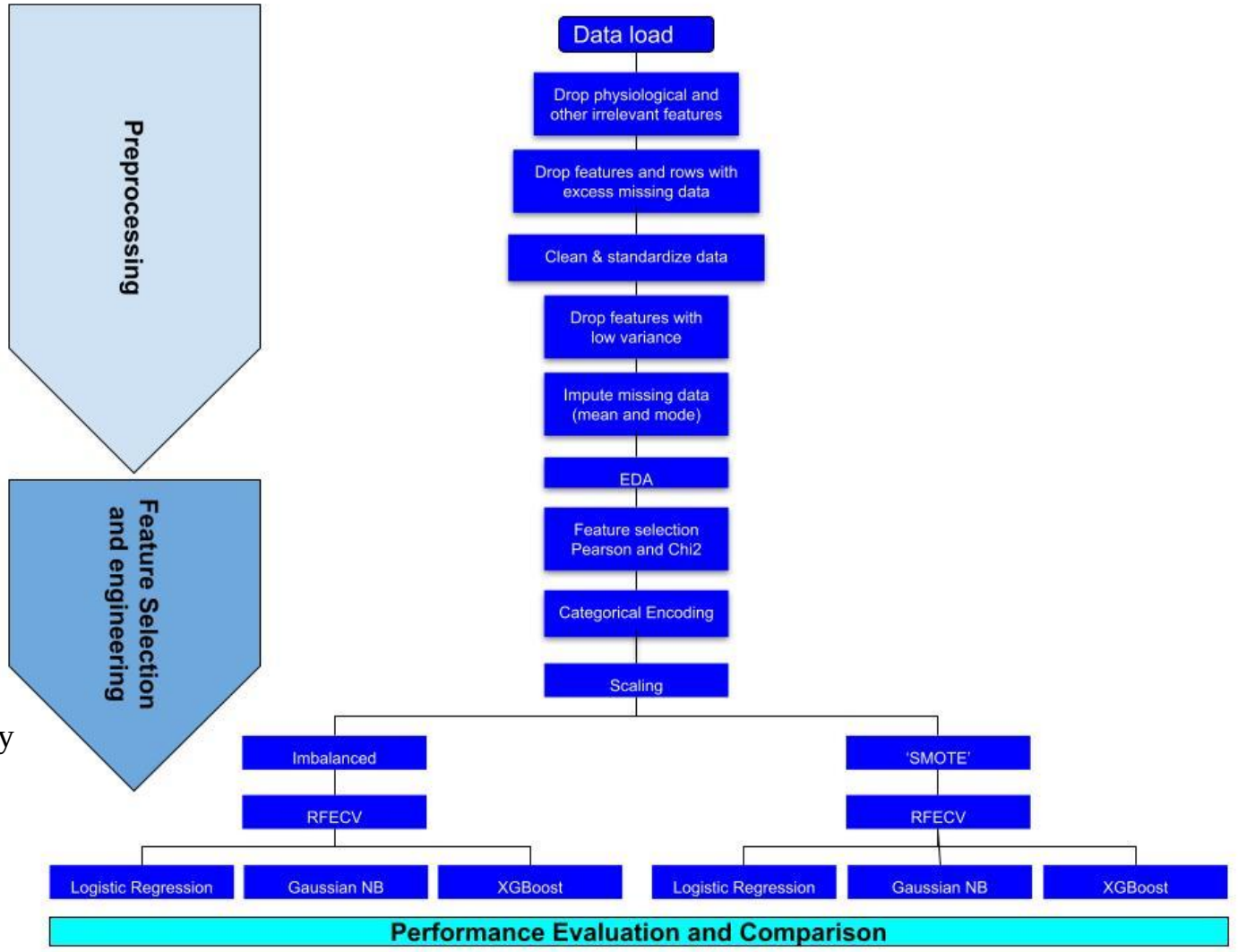
Raschke, R. A., Gerkin, R. D., Ramos, K. S., Fallon, M., & Curry, S. C. (2018). The explained variance and discriminant accuracy of APACHE IVa® severity scoring in specific subgroups of ICU patients. Southwest J Pulm Crit Care, 17, 153-64.

Veith & Steele: sparse “administrative model” AUROC: 0.751

Attribute	Description	Ranking
Admission Type	Urgent, Elective, Newborn or Emergency	1
Insurance	Patient’s insurance provider	2
Admission Location	Previous location of patient	3
Religion	Patient’s religion	4
Marital Status	Patient’s marital status	5
Language	Spoken language of patient	6
Ethnicity	Ethnicity of patient	7

Veith N, Steele R. Machine Learning-based Prediction of ICU Patient Mortality at Time of Admission. Proceedings of the 2nd International Conference on Information System and Data Mining; 2018 Mar; Lakeland, USA. New York: Association for Computing Machinery, 2018: 34-38.

Methodology Overview



A similar methodology was followed on a 'Compact' balanced version of the dataset (12k records)

Dataset

The “Patient Survival Prediction Dataset” is used

- The initial dataset consists of 91,713 instances (i.e. unique patient ICU stays) and 186 features.
- Completely pre-anonymized: no patient or hospital identifiers
- Patient ages range between 19 and 89
- Data was collected from 147 hospitals
- **Variable of interest is ‘hospital_death’ (0/1), i.e. whether the patient died in hospital**
- Includes the APACHE IV predicted probability of patient dying on hospital, which was removed from the outset.

https://www.kaggle.com/datasets/sadiaanzum/patient-survival-prediction-dataset
THE GLOBAL OPEN-SOURCE SEVERITY OF ILLNESS SCORE (GOSSIS)
https://journals.lww.com/ccmjournal/Citation/2019/01001/33__THE_GLOBAL_OPEN_SOURCE_SEVERITY_OF_ILLNESS.36.aspx

Feature Selection

1. Continuous: Pearson Correlation
2. Ordinal and Categorical: Chi-Square
3. Recursive Feature Elimination (RFECV):
   - There were 39 total features prior to RFECV (including categorical dummies)
   - Random Forest classifier
   - 10-K Cross-Validation
   - scoring: ‘roc_auc’
   - RFECV is run after SMOTE was applied to address the class imbalance

Class Imbalance

1. Minority class (patient died) accounts for 8.4% of observations
2. Three approaches were evaluated and compared:
    1. SMOTE (Synthetic Minority Oversampling Technique): ~78K records each of patient died and did not die (RFECV run after SMOTE process)
    2. As-is dataset – no measures taken to address the imbalance: ~78K patients did not die and ~7.2K patients died
    3. A compact subset of the dataset was extracted, with an initial 12K observations evenly split between patient died and did not die (selected by sorting on a random number generator).

RFECV: Features to keep / discard

After running RFECV on the 3 versions of the dataset:

1. 22/39 features were kept for the imbalanced
2. 36/39 were kept for 'SMOTE'
3. 23/39 were kept for the compact balanced.

Note: these feature counts all include categorical dummies. The initial 39 features are actually only 12 if un-encoded.

Dataset version	RFECV Optimum Features	
	Features to keep	Features to Drop
Imbalanced	**Total features to keep = 22** •age •elective_surgery •pre_icu_los_days •weight •intubated_apache •ventilated_apache •solid_tumor_with_metastasis •immunosuppression •gcs_motor_apache_1.0 •gcs_motor_apache_4.0 •gcs_motor_apache_5.0 •gcs_motor_apache_6.0 •gcs_verbal_apache_1.0 •gcs_verbal_apache_4.0 •gcs_verbal_apache_5.0 •apache_3j_bodysystem_Cardiovascular •apache_3j_bodysystem_Gastrointestinal •apache_3j_bodysystem_Metabolic •apache_3j_bodysystem_Sepsis •apache_2_bodysystem_Cardiovascular •apache_2_bodysystem_Neurologic •apache_2_bodysystem_Respiratory	**Total features to drop = 17** •gcs_motor_apache_2.0 •gcs_motor_apache_3.0 •gcs_verbal_apache_2.0 •gcs_verbal_apache_3.0 •apache_3j_bodysystem_Genitourinary •apache_3j_bodysystem_Gynecological •apache_3j_bodysystem_Hematological •apache_3j_bodysystem_Musculoskeletal/Skin •apache_3j_bodysystem_Neurological •apache_3j_bodysystem_Respiratory •apache_3j_bodysystem_Trauma •apache_2_bodysystem_Gastrointestinal •apache_2_bodysystem_Haematologic •apache_2_bodysystem_Metabolic •apache_2_bodysystem_Renal/Genitourinary •apache_2_bodysystem_Trauma •apache_2_bodysystem_Undefined diagnoses
SMOTE	**Total features to keep = 36** •age •elective_surgery •pre_icu_los_days •weight •intubated_apache •ventilated_apache •solid_tumor_with_metastasis •immunosuppression •gcs_motor_apache_1.0 •gcs_motor_apache_2.0 •gcs_motor_apache_3.0 •gcs_motor_apache_4.0 •gcs_motor_apache_5.0 •gcs_motor_apache_6.0 •gcs_verbal_apache_1.0 •gcs_verbal_apache_2.0 •gcs_verbal_apache_3.0 •gcs_verbal_apache_4.0 •gcs_verbal_apache_5.0 •apache_3j_bodysystem_Cardiovascular •apache_3j_bodysystem_Gastrointestinal •apache_3j_bodysystem_Genitourinary •apache_3j_bodysystem_Metabolic •apache_3j_bodysystem_Neurological •apache_3j_bodysystem_Respiratory •apache_3j_bodysystem_Sepsis •apache_3j_bodysystem_Trauma •apache_2_bodysystem_Cardiovascular •apache_2_bodysystem_Gastrointestinal •apache_2_bodysystem_Haematologic •apache_2_bodysystem_Metabolic •apache_2_bodysystem_Neurologic •apache_2_bodysystem_Renal/Genitourinary	**Total features to drop = 3** •apache_3j_bodysystem_Gynecological •apache_3j_bodysystem_Hematological •apache_3j_bodysystem_Musculoskeletal/Skin

Ryerson University

Classifiers (versions of models)

1. Logistic Regression
2. Gaussian Naïve Bayes
3. XGBoost

Common Elements

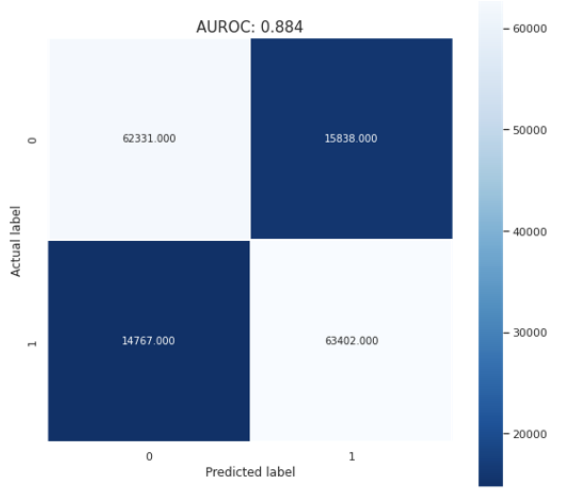
For all 3 classifiers and across all the versions of the dataset, the following were applied:

1. 10-Fold Cross Validation
2. AUROC was specified as the scoring metric

Logistic Regression

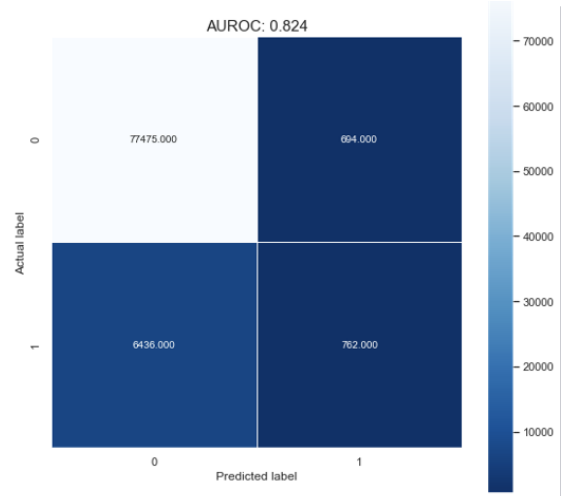
Logistic Regression was selected as it is also used in APACHE IV.

SMOTE



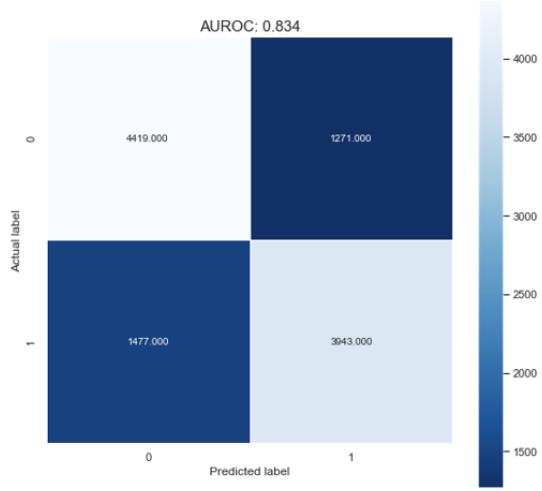
	precision	recall	f1-score	support
0	0.81	0.80	0.80	78169
1	0.80	0.81	0.81	78169
accuracy			0.80	156338
macro avg	0.80	0.80	0.80	156338
weighted avg	0.80	0.80	0.80	156338

Imbalanced



	precision	recall	f1-score	support
0	0.92	0.99	0.96	78169
1	0.52	0.11	0.18	7198
accuracy			0.92	85367
macro avg	0.72	0.55	0.57	85367
weighted avg	0.89	0.92	0.89	85367

Compact Balanced

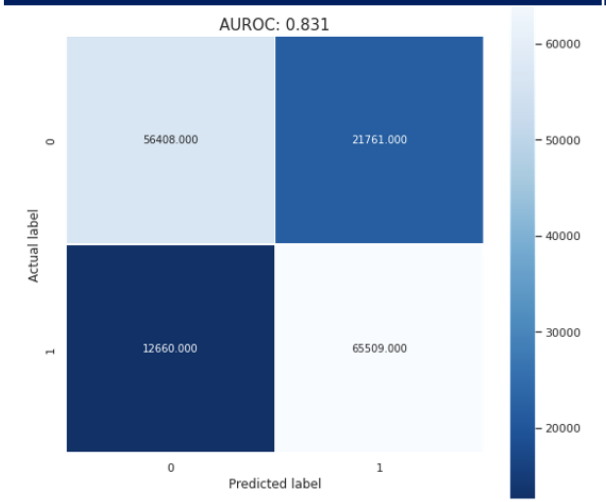


	precision	recall	f1-score	support
0	0.75	0.78	0.76	5690
1	0.76	0.73	0.74	5420
accuracy			0.75	11110
macro avg	0.75	0.75	0.75	11110
weighted avg	0.75	0.75	0.75	11110

Gaussian Naïve Bayes

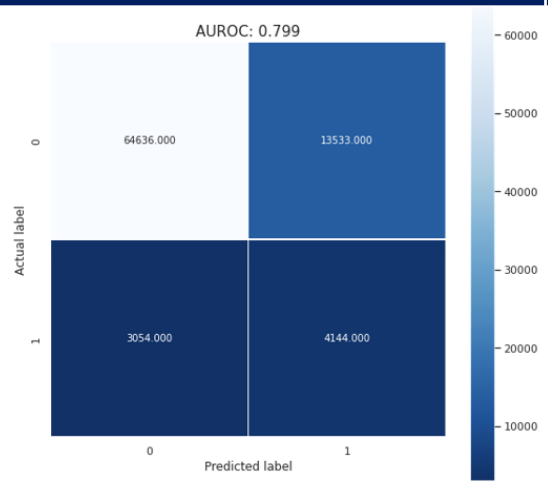
Gaussian Naïve Bayes was selected as an alternative statistical approach

SMOTE



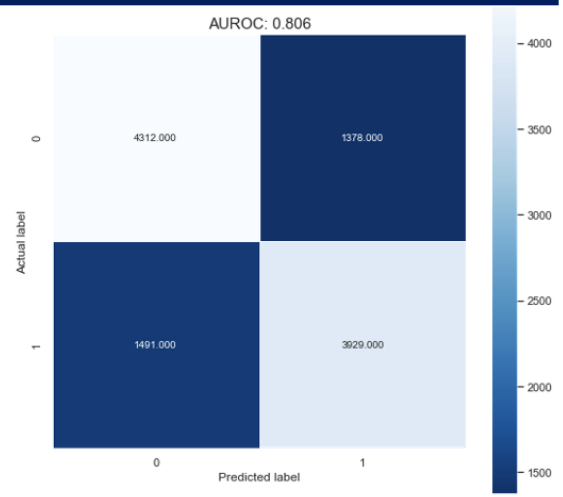
	precision	recall	f1-score	support
0	0.82	0.72	0.77	78169
1	0.75	0.84	0.79	78169
accuracy			0.78	156338
macro avg	0.78	0.78	0.78	156338
weighted avg	0.78	0.78	0.78	156338

Imbalanced



	precision	recall	f1-score	support
0	0.95	0.83	0.89	78169
1	0.23	0.58	0.33	7198
accuracy			0.81	85367
macro avg	0.59	0.70	0.61	85367
weighted avg	0.89	0.81	0.84	85367

Compact Balanced



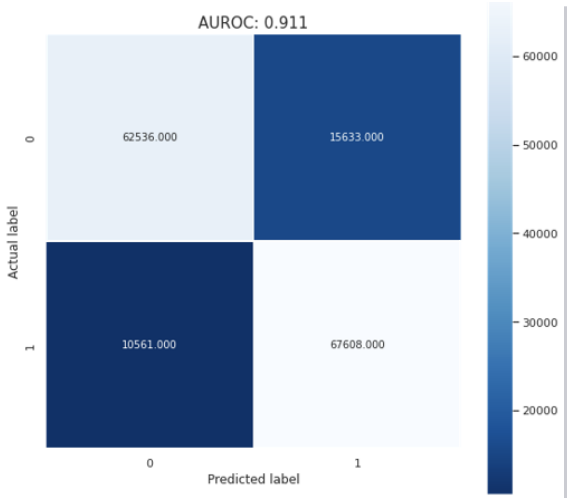
	precision	recall	f1-score	support
0	0.74	0.76	0.75	5690
1	0.74	0.72	0.73	5420
accuracy			0.74	11110
macro avg	0.74	0.74	0.74	11110
weighted avg	0.74	0.74	0.74	11110

Ryerson University

XGBoost

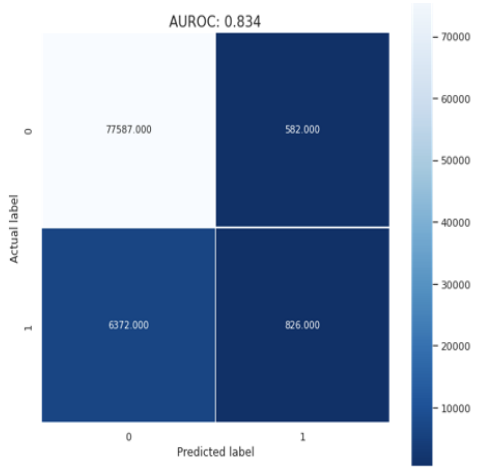
XGBoost was selected in order to compare a machine learning model with the statistical counterparts

SMOTE



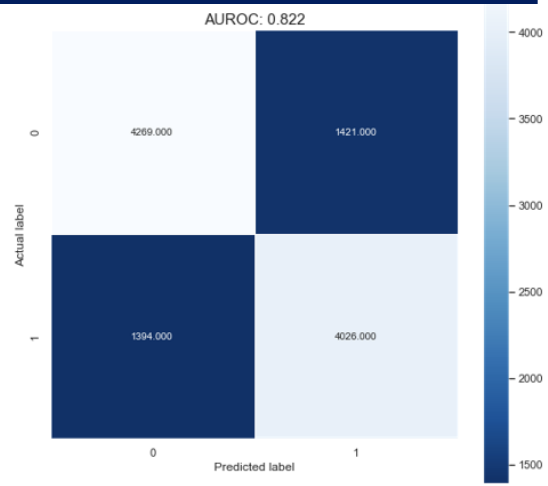
	precision	recall	f1-score	support
0	0.86	0.80	0.83	78169
1	0.81	0.86	0.84	78169
accuracy			0.83	156338
macro avg	0.83	0.83	0.83	156338
weighted avg	0.83	0.83	0.83	156338

Imbalanced



	precision	recall	f1-score	support
0	0.92	0.99	0.96	78169
1	0.59	0.11	0.19	7198
accuracy			0.92	85367
macro avg	0.76	0.55	0.57	85367
weighted avg	0.90	0.92	0.89	85367

Compact Balanced



	precision	recall	f1-score	support
0	0.75	0.75	0.75	5690
1	0.74	0.74	0.74	5420
accuracy			0.75	11110
macro avg	0.75	0.75	0.75	11110
weighted avg	0.75	0.75	0.75	11110

Common Eler

Comparing model performance across the 3 versions of the dataset

Classifier	Dataset Version														
	SMOTE					Imbalanced					Compact Balanced				
	AUROC	Accuracy	Precision	Recall	f1-score	AUROC	Accuracy	Precision	Recall	f1-score	AUROC	Accuracy	Precision	Recall	f1-score
Logistic Regression	0.884	0.8	0.8	0.8	0.8	0.824	0.92	0.72/0.89	0.55/0.92	0.57/0.89	0.834	0.75	0.75	0.75	0.75
Gaussian NB	0.831	0.78	0.78	0.78	0.78	0.799	0.81	0.59/0.89	0.70/0.81	0.61/0.84	0.806	0.74	0.74	0.74	0.74
XGBoost	0.911	0.83	0.83	0.83	0.83	0.834	0.92	0.76/0.90	0.55/0.92	0.57/0.89	0.822	0.75	0.75	0.75	0.75

Note: macro avg/weighted avg

APACHE IV AUROC: 0.88

Using AUROC as the key metric:

- XGBoost scored highest in both the SMOTE and imbalanced dataset
- Logistic Regression scored slightly higher in the compact dataset*
- Gaussian Naïve Bayes was the lowest performer in all 3 datasets

XGBoost also scored the highest or equally highest in all other metrics across all datasets

Compared with APACHE IV, XGBoost and Logistic Regression perform at or above the same level with the SMOTE dataset.

* Given the much smaller size of the compact dataset, this might not be significant.

Efficiency

Separate versions of the .ipynb files were created for each of the 3 different versions of the dataset.

Component	Version of ipynb for dataset		
	SMOTE	Imbalanced	Compact Balanced
RFECV	3:15:17	1:10:42	0:07:04
Logistic Regression	0:04:27	-	-
Gaussian NB	-	-	-
XGBoost	0:05:50	0:02:01	-
Total duration	**3:32:03**	**1:39:23**	**0:07:57**

h:mm:ss

As seen in the above chart , the recursive feature selection step consumed by far the majority of the processing time in all versions.

The primary environment used was Google Colaboratory, mainly using the sci-kit learn library.

The same .ipynb files were also run locally on Jupyter Notebook on a MacBook Pro.

Interpretation and Caveats

- While we've emphasized not relying on physiological features in our models in contrast with APACHE IV, like APACHE, they all include the Glasgow Coma Scale measures, which possibly accounts for much of the relative similarity in performance (RFECV dropped the middle values of the GCS scores in the imbalanced dataset).
- Our 3 models have not been evaluated for over-fitting. While 10-fold cross validation was used throughout, not enough analysis has been conducted to determine the extent of over-fitting.

Challenges and Lessons Learned

- No domain knowledge
- Integrating and applying new knowledge re: statistics, programming and tools, with abundance of available resources (=information overload)
- Ability to iterate / experiment hindered by long processing times
- Underestimated project management aspects (planning and prioritization based on end-goals)

Q & A

Thank you