Boosted Tree Ensembles for Predicting Postsurgical ICU Mortality

Garrick Aden-Buie, Yun Chen, Rashad Kayal, Gina Romero, Hui Yang

Dept. of Industrial and Management Sciences Engineering
College of Engineering
University of South Florida, Tampa, FL

INFORMS Annual Meeting 2013, Minneapolis, MN

Outline

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING

Outline

Trends in Critical Care in US

- Critical care beds increased by 6.5% (2000-2005)
  - Despite 12.2% decrease in hospitals with critical care and 4.2% reduction overall in hospital beds
- Constrained ICU capacity
- High quality care: *safe, effective, equitable patient-centered, timely and efficient* (IOM)

Halpern, Neil A, and Stephen M Pastores, 2010
"Critical Care Medicine in the United States 2000-2005"

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING

Acuity Scores in ICUs

- Existing acuity scores
  - APACHE
  - SAPS
  - MPM
  - SOFA
- Aim to compensate for population differences to objectively compare practices across ICUs
- Need for patient-specific prognostic models

Objective

- To develop a data-driven, patient-specific prognostic model to predict in-hospital death in post-surgical ICU patients.
- To support effective, efficient use of critical care resources

Overview

- We created and evaluated a gradient boosted trees model using routine patient data recorded during the first 48 hours of an ICU visit.
  - Uses heterogeneous, routinely-collected data
  - Requires minimal preprocessing
  - Effectively addresses sampling and missing information issues
  - Accurately predicts in-hospital mortality

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING

Outline

MIMIC II Clinical Data

- Physiologic signals and vital signs from patient monitoring and hospital information systems
- PhysioNet Computing in Cardiology 2012 Challenge
- 12,000 patients divided into 3 sets of 4,000
  - Set A: Training
  - Set B: Validation
  - Set C: Testing
- Inclusion criteria
  - Age $\geq$ 16 years
  - Initial ICU stay $\geq$ 48hrs

http://www.physionet.org/challenge/2012/

MIMIC II Clinical Data

- Physiologic signals and vital signs from patient monitoring and hospital information systems
- PhysioNet Computing in Cardiology 2012 Challenge
- 12,000 patients divided into 3 sets of 4,000
  - Set A: Training
  - Set B: Testing
- Inclusion criteria
  - Age $\geq$ 16 years
  - Initial ICU stay $\geq$ 48hrs

http://www.physionet.org/challenge/2012/

Input Variables

- Up to 41 variables recorded per patient
  - 5 general descriptors
  - 36 time series variables

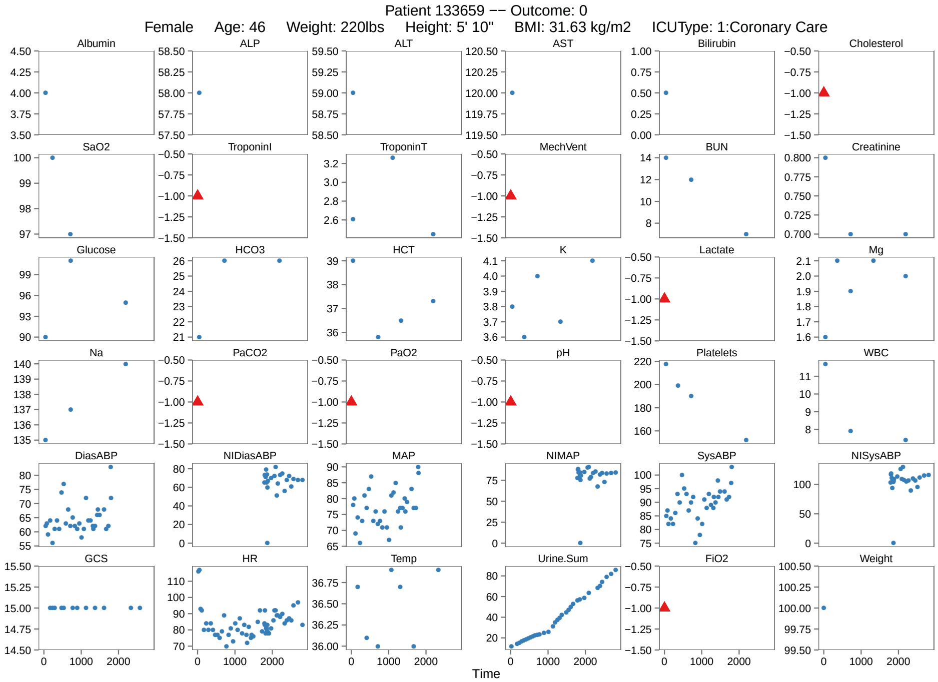
General Descriptors

Variable	Mean	S.D.
Age	64.5 *yrs*	17.1
Height	169.5 *cm*	17.1
Weight	81.2 *kg*	23.8
Gender	*Male*:	56.1%
	Female:	43.8%
ICU Type	*Medical*:	35.8%
	Surgical:	28.4%
	Cardiac surgery:	21.1%
	Coronary:	21.1%
In-Hospital Death		*13.85%*

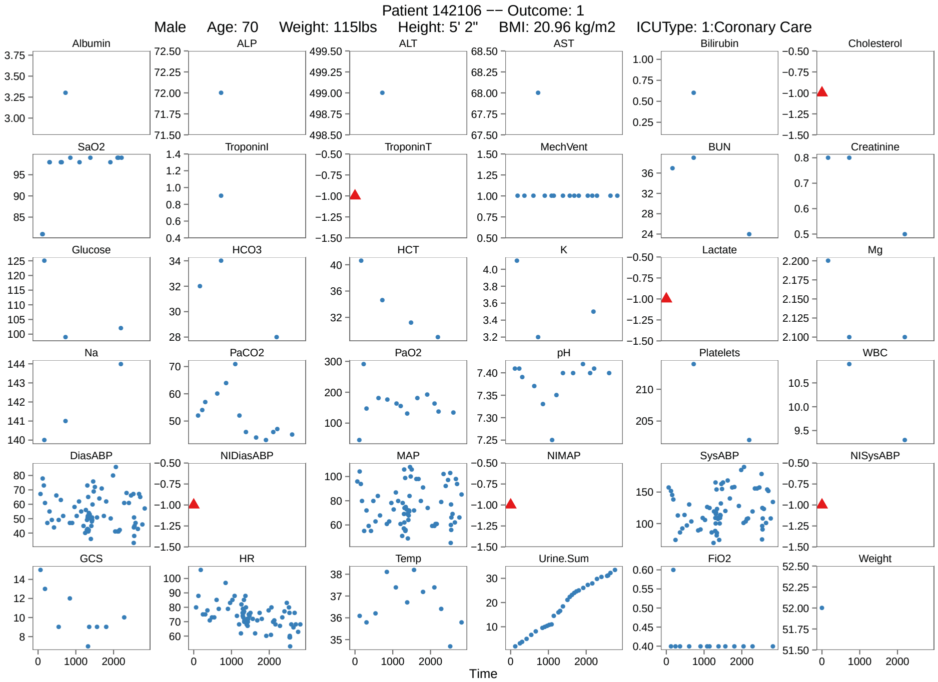
Time Series Variables

36 variables describing

- Arterial Blood Gasses
- Cardiac Biomarkers
- Blood Count
- Consciousness
- Hepatic Function
- Overall Condition
- Renal Function
- Serum Electrolytes
- Ventilation Support
- Vital Signs



Patient 133659 -- Outcome: 0
Female Age: 46 Weight: 220lbs Height: 5' 10" BMI: 31.63 kg/m2 ICUType: 1:Coronary Care
Albumin
ALP
ALT
AST
Bilirubin
Cholesterol
SaO2
TroponinI
TroponinT
MechVent
BUN
Creatinine
Glucose
HCO3
HCT
K
Lactate
Mg
Na
PaCO2
PaO2
pH
Platelets
WBC
DiasABP
NIDiasABP
MAP
NIMAP
SysABP
NISysABP
GCS
HR
Temp
Urine.Sum
FiO2
Weight
Time


Patient 142106 -- Outcome: 1
Male Age: 70 Weight: 115lbs Height: 5' 2" BMI: 20.96 kg/m2 ICUType: 1:Coronary Care
Albumin
ALP
ALT
AST
Bilirubin
Cholesterol
SaO2
TroponinI
TroponinT
MechVent
BUN
Creatinine
Glucose
HCO3
HCT
K
Lactate
Mg
Na
PaCO2
PaO2
pH
Platelets
WBC
DiasABP
NIDiasABP
MAP
NIMAP
SysABP
NISysABP
GCS
HR
Temp
Urine.Sum
FiO2
Weight
Time

Outline

Motivation

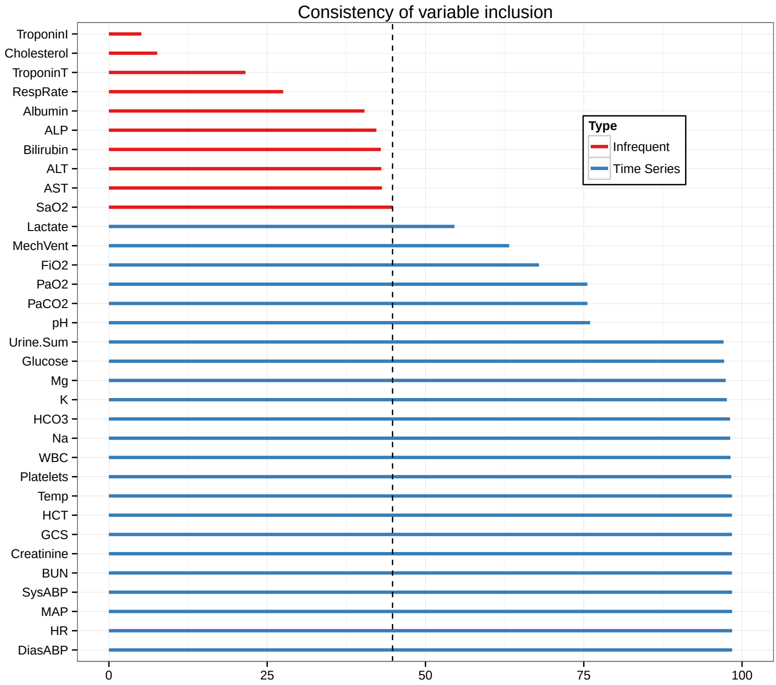
MIMIC II Clinical Data

Methods

Results

Preprocessing Overview

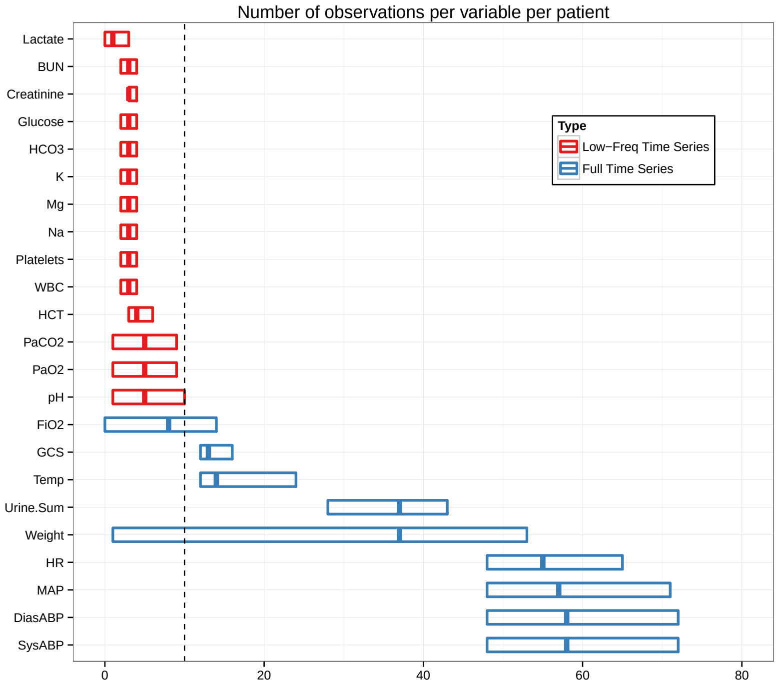
- Correct implausible values
- Categorize variables by
  1. Consistency of inclusion
  2. Number of observations when recorded
- Missing information
- Feature extraction
- Feature selection


Consistency of variable inclusion
TroponinI
Cholesterol
TroponinT
RespRate
Albumin
ALP
Bilirubin
ALT
AST
SaO2
Lactate
MechVent
FiO2
PaO2
PaCO2
pH
Urine.Sum
Glucose
Mg
K
HCO3
Na
WBC
Platelets
Temp
HCT
GCS
Creatinine
BUN
SysABP
MAP
HR
DiasABP
0
25
50
75
100
Type
Infrequent
Time Series

Infrequently Included Variables

- *Infrequently included variables*
  - Are included in $\leq$ 45% training set patients
- Transformed to a categorical variable:
  - 0 = Not recorded
  - 1 = Recorded & within normal range
  - 2 = Recorded & abnormal
- Significant portion of missing minimal information

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING


Number of observations per variable per patient
Lactate
BUN
Creatinine
Glucose
HCO3
K
Mg
Na
Platelets
WBC
HCT
PaCO2
PaO2
pH
FiO2
GCS
Temp
Urine.Sum
Weight
HR
MAP
DiasABP
SysABP
0
20
40
60
80
Type
Low−Freq Time Series
Full Time Series

Time Series Variables

- *Low-frequency time series*
  - $<$ 10 observations for $\geq$ 75% training set patients
- *Full time series*
  - Variables not meeting the above criteria
- If no observation recorded for a variable:
  - Impute from normal distribution representing gender-specific normal physiologic values

Feature Extraction

- Low-frequency time series
  1. Mean
- Full time series
  1. Mean, Median
  2. Min, Max
  3. First/Last Observation
  4. Trend over 0–24, 24–48, and 0–48 hours
     - Requires 5, 5, 10 observations

Feature Selection by mRMR

- **mRMR**: Minimum Redundancy, Maximum Relevancy
- *Redundancy*: mutual information between two features
- *Relevancy*: mutual information between features and outcome
- Heuristic: scores and ranks features
- One feature per category selected

Peng, H, Fulmi Long, and C Ding, 2005. "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy."

Boosted Tree Ensembles

- A weak learner can be *boosted* by aggregating the predictions of an ensemble of weak learners
- Boost accuracy and retain benefits of weak learner
- Decision stumps
  - Natural handling of heterogeneous data
  - Non-linear
  - Minimal preprocessing

Schapire, Robert E. "The strength of weak learnability." Machine learning 5, no. 2 (1990)

UNIVERSITY OF SOUTH FLORIDA
COLLEGE OF ENGINEERING

Gradient Boosted Trees

- Given a feature vector, $\mathbf{x} = (x_1, x_2, \ldots, x_i)$, and outcome labels $Y = \{0, 1\}$
- Build a function $g(\mathbf{x}) \colon \mathbf{x} \to y \in Y$
- $g(\mathbf{x}) = \log\left(\frac{p(\mathbf{x})}{1-p(\mathbf{x})}\right)$

Gradient Boosted Trees Algorithm

- Initialize $g_0(\mathbf{x}) =$ basline log-odds of in-hospital death
- Each step: find an $h(\mathbf{x})$ to add to collection $g_m(\mathbf{x})$:
  - Select a random subsample of training data, $\tilde{N}$
  - Search for a decision stump $h(\mathbf{x})$ that best improves fit of $g_m(\mathbf{x}) + h(\mathbf{x})$ on $\tilde{N}$
    - Best fit is determined by maximized Bernoulli log-likelihood
  - $g_{m+1}(\mathbf{x}) \leftarrow g_m(\mathbf{x}) + \lambda h(\mathbf{x})$
- Parameters selected by 10-fold cross validation

Generalized Boosted Regression Models, Greg Ridgeway
R package version 2.0-8 – http://cran.R-project.org/package=gbm

USF UNIVERSITY OF SOUTH FLORIDA COLLEGE OF ENGINEERING

Outline

PhysioNet Scoring

Optimize Precision-Recall curve: $\min(Se, PPV)$

Sensitivity	Positive Predictivity
$Se = \dfrac{\text{TP}}{\text{TP + FN}}$	$PPV = \dfrac{\text{TP}}{\text{TP + FP}}$

Performance on Sets A & B

Set A

Score	0.481
Threshold	0.568
Score at thresh	**0.453**
Sensitivity	0.795
Specificity	0.767
AUC	0.848

Average across 10 folds

Set B

Se	0.532
PPV	0.496
Final Score	**0.496**

Performance Comparison

Method	Score
Random Classifier	0.15
SAPS-I	0.32
Fuzzy Rule Based System	0.36
Cascaded AdaBoost	0.38
Time Series Motifs	0.50
Gradient Boosted Trees	**0.50**
Logistic Regression & Hidden Markov Model	0.50
2-Layer Neural Network	0.51
Bayesian Ensemble	0.53

Summary

- We developed a boosted tree ensemble model for prediction of in-hospital mortality of ICU patients, using patient data collected over the first 48 hours of ICU stay.
- Effectively uses routinely-collected ICU patient data
- Addresses ICU needs in clinical planning
- Future Work:
  - Extend our model to provide and update predictions *during* the 48 hour period

Acknowledgements

US National Science Foundation
CMMI-1266331, IOS-1146882

University of South Florida
Internal Research Award (Grant No. 76734)

Thank you

Questions?