Exploring the Prediction of ICU Patient Readmission Using Machine Learning

Rishabh Sharad Pomaje[1], Rutanshu Jhaveri[2], Shruthi Shekar[3]

{[1]rishabhp, [2]rutanshu, [3]scshekar}@stanford.edu

[1]Department of Electrical Engineering, [2]Institution of Computational and Mathematical Engineering, [3]Department of Biomedical Data Science

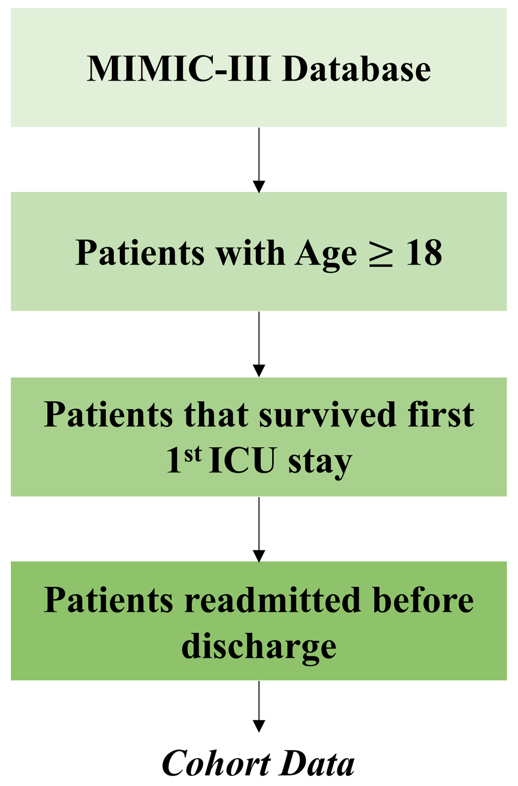
Stanford University

Motivation, Problem, and Overview

- In the United States, unplanned readmissions cost the Centers for Medicare and Medicaid Services (CMS) an estimated $17-26 billion annually [1], which simultaneously straining limited hospital capacity.
- Early identification of high risk patients enables targeted interventions.
- An interpretable Machine Learning (ML) system can assist medical experts in clinical decision-making.
- We explore and evaluate different ML models and algorithms to predict the risk of all-cause 30-day readmission while considering the data from a patients first ICU stay.
- Our experiments evaluate models of varying complexity, from baseline logistic regression to gradient-boosted decision trees (XGBoost, CatBoost) and Deep Learning models (TabNet). Among these, we found that XGBoost achieves the strongest predictive performance

Data and Features

- We derive our data from the MIMIC-III (PhysioNet) database [2].



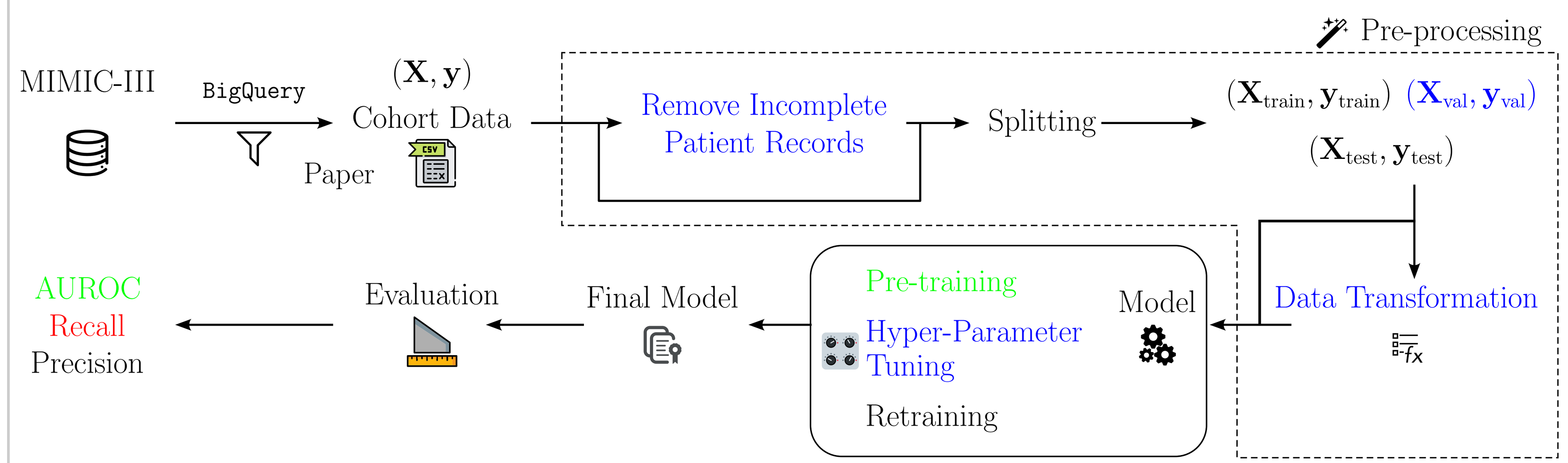
	Train Set	Validation Set	Test Set	Total
No. of Entries	14939	6403	9147	30489
No. of Readmissions	1605	688	983	3276
(% Readmissions)	~10.74	~10.74	~10.74	~10.74

Summary Statistics of the Cohort Data (top) and features considered for every patient (bottom).

Anion-Gap	Bicarbonate	Calcium
Chloride	Creatinine	Glucose
Hematocrit	Hemoglobin	MCHC
MCV (Mean Corpuscular Volume)	Magnesium	PT (Prothrombin Time)
Phosphate	Potassium	RDW (Red Cell Distribution Width)
RBC Count	Sodium	Urea-Nitrogen
WBC Count	Age (years)	ICU-time (hrs)

- A feature vector of a patient consists of the mean, min, max, and standard deviation of each lab parameter and the age and time spent in the ICU during first visit.
- Lab test results are usually available in 4-12hrs, supporting early readmission prediction.
- **Issues:**
  1. Missing Data
     - Solution: Filtering records that are incomplete
     - Problem: Losing valuable information.
     - Solution: Imputation techniques that use exisiting data to estimate missing values.
  2. Imbalanced Data: Readmission rate was 10.74%. Can balancing techniques improve the performance?
     - Minority Oversampling.
     - Majority Undersampling and loss upweighting.
- Possible Pitfalls: Data Leakage (Split first), Disturbed Priors (Stratification).

Pipeline



Models, Experiments, & Results

Family	Model	Precision		Recall		AUROC
		Macro	Weighted	Macro	Weighted	
Linear Classifier (Baseline)	LogReg (Base)	0.54	0.78	0.50	0.86	**0.670 ± 0.025**
	LogReg (Balanced)	0.56	0.82	0.62	0.66	**0.668 ± 0.025**
	LogReg (Imputation)	0.67	0.85	0.51	0.89	**0.709 ± 0.017**
GBDT	XGBoost (Base)	0.60	0.83	0.52	0.89	**0.687 ± 0.017**
	XGBoost (HyperOpt)	0.63	0.85	0.60	0.87	**0.745 ± 0.015**
	XGBoost (Imputed)	0.64	0.85	0.59	0.88	**0.746 ± 0.015**
	CatBoost (Base)	0.75	0.86	0.51	0.89	**0.726 ± 0.016**
	CatBoost (HyperOpt)	0.76	0.86	0.51	0.89	**0.737 ± 0.016**
Neural Networks	TabNet (Base)	0.55	0.83	0.51	0.89	**0.648 ± 0.018**
	TabNet (Pretrained)	0.56	0.85	0.64	0.67	**0.680 ± 0.018**

Logistic Regression (LogReg)

- LogReg models the probability of readmission:

$$P(y=1 \mid \mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b), \quad \sigma(z) = \frac{1}{1+e^{-z}}$$

- It finds the optimal parameters w and b by minimizing loss-function, typically log-likelihood.
- Experiments:

1. Base - Using only complete entries.
2. Balanced - Using balancing technique: Undersampling + Loss upweighting.
3. Imputation - Filling in missing data using median strategy.

Gradient Boosted Decision Trees

- GBDTs can capture more complex distributions and feature relations than linear models.
- XGBoost [3] uses an ensemble of decision trees. The output of an ensemble is given as:

$$\hat{y} = \sum_{k=1}^{K} f_k(\mathbf{x}), \quad f_k \in \mathcal{F}$$

- Minimizes regularized-convex loss:

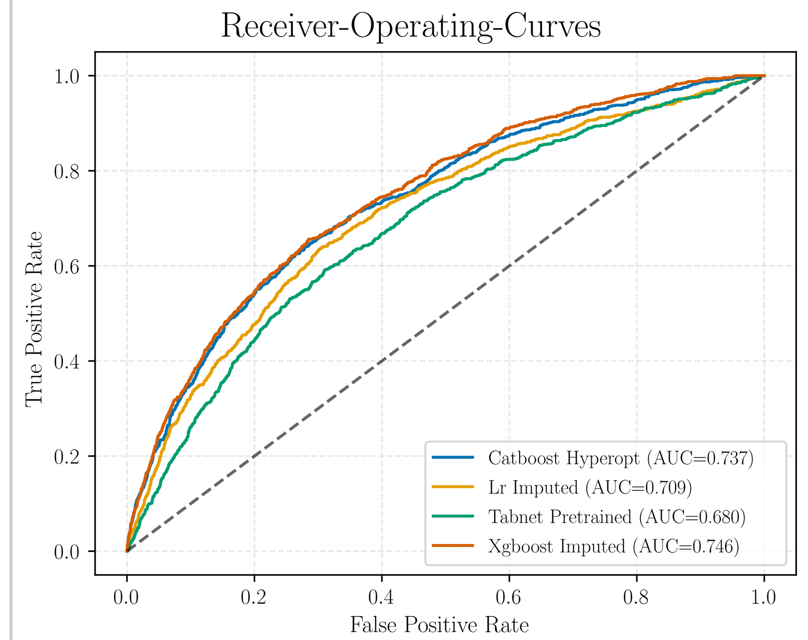
$$\mathcal{L} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \left(\gamma T_k + \tfrac{1}{2}\lambda \|\mathbf{w}_k\|^2\right)$$

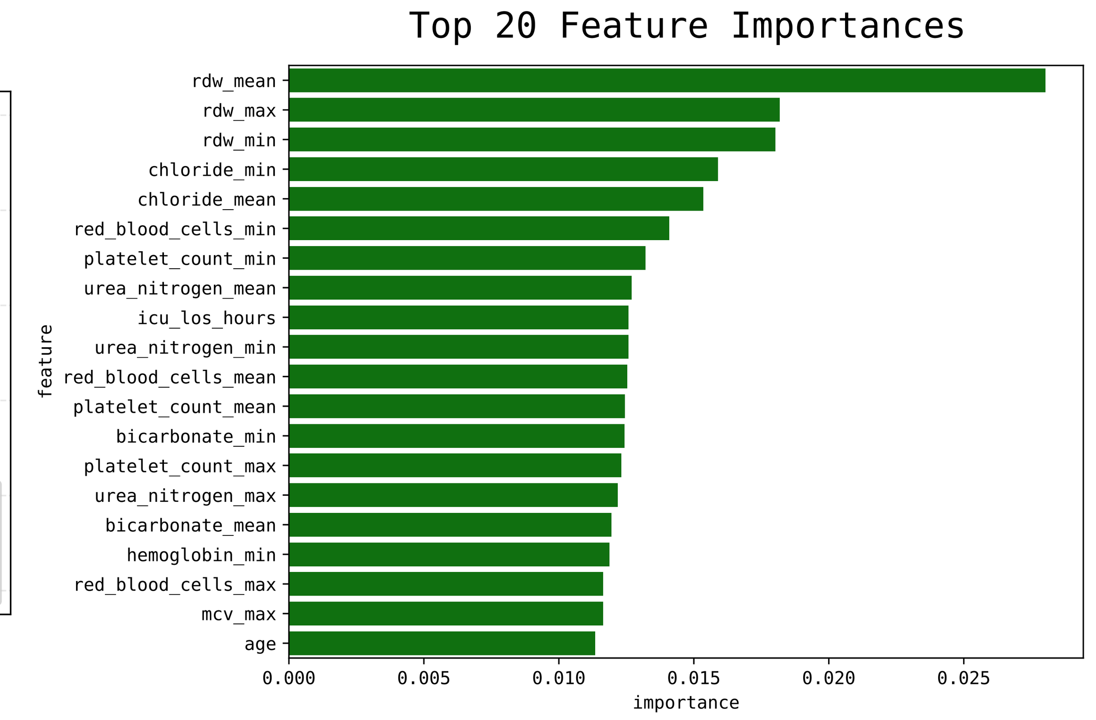
- Experiments:
  1. Base - Using default hyperparameters.
  2. HyperOpt - Tree-Structured Parzen Estimation (TPE) Hyperparameter search.
  3. Imputed - Filling in missing data using median strategy + Hyperparameter search.
- CatBoost [4] provides a fast, efficient alternative that is useful for rapid prototyping.

TabNet [5]

- Neural Network architecture designed specially for Tabular Data.
- Sequential-Attention mechanism dynamically selects which features to consider at each step.
- Unsupervised pre-training allows to learn features and improves performance.

Discussion





- LogReg provides a solid baseline, however suffers from (i) limited expressivity and (ii) inability to handle missing data.
- XGB performs the best due to (i) additional expressivity and (ii) ability to find the optimal split while accounting for missing data.
- Neural Network has the capability to do well but needs more data and more importantly parameter tuning.
- Data Imputation techniques do improve performance and maximize the utility of available data.
- RDW variation is linked to systemic stress and adverse ICU outcomes. Urea/BUN reflects dehydration, catabolic stress, or impaired perfusion.
- Chloride and bicarbonate highlight acid-base and fluid disturbances which relate to mortality.
- Platelet count signals ongoing inflammation or coagulopathy.
- The model learned patients who appear clinically stable yet retain physiological signs of incomplete recovery.

Future Work

- Improvement due to data imputation technique indicate that it is work exploring more sophisticated strategies such as KNN- or Deep Learning based methods.
- Early feature-engineering experiments showed performance gains, suggesting value in augmenting the dataset with additional categorical variables: diagnostic/medication/procedure codes.
- Additional information extracted from clinical notes potentially incorporates domain-expert knowledge.
- Attention-based deep learning architectures like FT-Transformer [6] and TabNet that are designed for use with tabular data warrant further investigation to better leverage both structured and unstructured information.

References

[1] M. Alvarado, B. Lahijanian, Y. Zhang, and M. Lawley, "*Penalty and incentive modeling for hospital readmission reduction*," Operations Research for Health Care, vol. 36, p. 100376, 2023.

[2] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "*Mimic-iii, a freely accessible critical care database*," Scientific data, vol. 3, no. 1, pp. 1–9, 2016.

[3] T. Chen and C. Guestrin, "*Xgboost: A scalable tree boosting system*," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. ACM, Aug. 2016, p. 785–794.

[4] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "*Catboost: unbiased boosting with categorical features*," 2019.

[5] S. O. Arik and T. Pfister, "*Tabnet: Attentive interpretable tabular learning*," 2020.

[6] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "*Revisiting deep learning models for tabular data*," 2023.