Predicting Hospital Length of Stay from Imbalanced Data

Miguel Fuentes, Pinlin [Calvin] Xu, Lisa Liu | {migufuen, pinlinxu, lisacliu}@stanford.edu

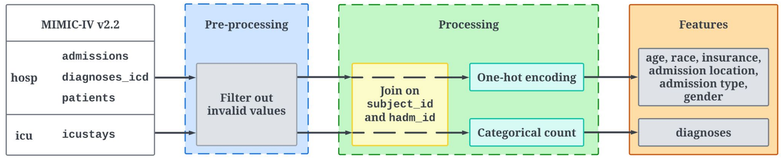
Overview

This project strives to predict the **length of stay (LoS)** of admitted patients from **electronic health record (EHR)** data, a crucial factor for efficiently planning healthcare capacities and resources[2].

We collected raw data from **MIMIC-IV**[1] (Medical Information Mart for Intensive Care). After addressing **highly-imbalanced** data, we trained a custom **classification-regression** pipeline to predict the LoS of the respective patient.

Data & Feature Engineering

We created our dataset from 30 GB of MIMIC-IV data by joining tables by patient admissions and cleaning up invalid entries, outliers, and data unavailable at admission. We set up 10-fold cross validation with 249772 training examples, 30836 validation examples, and 27753 test examples. The LoS range is limited to between 1 and 15 days, calculated from the admission and discharge times.



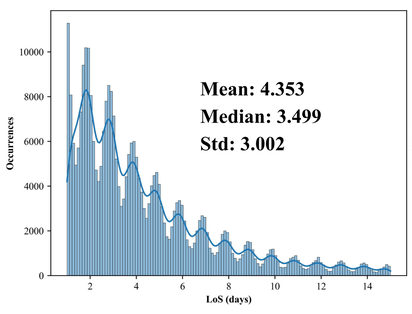
The features selected for the regression task consist of categorical data (**race, gender, insurance type, admission location**, etc.) in one-hot encoding, desensitized numerical values such as **age** grouped by Medical Subject Headings (MeSH) definitions, and individual patient **diagnoses** converted to 19 categories based on the International Statistical Classification of Diseases and Related Health Problems (ICD) standard, for a total of 53 columns.

Baselines

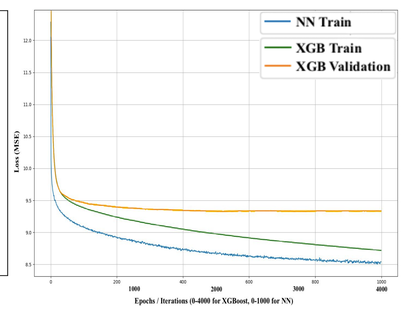
We evaluated the regression performance of 5 baseline algorithms on the dataset. Poor accuracy and overfitting were theorized to be due to the very imbalanced/skewed distribution of LoS in the data.

Baseline	OLS (with Ridge)	LASSO	ElasticNet	SGD Regressor	NN	XGBoost
Validation	3.1416	3.1692	3.1544	3.1499	3.0868	3.04795

Length of Stay Distribution (Histogram)



XGBoost & NN Learning Curves



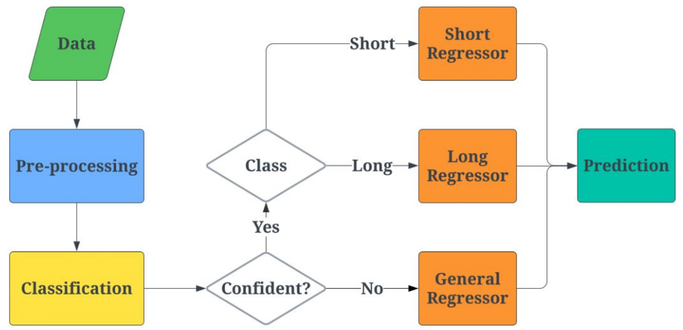
Methods & Experiments

We use a modified **SMOTE (Synthetic Minority Oversampling Technique)** to **balance the data**[3]. Instead of simply oversampling the minority class (long stays), SMOTE generates new data from the feature space. For a given data in the minority class, we select at random two of the five closest neighbors and create a new data from their convex combination given by $x_{\text{new}} = tx_1 + (1-t)x_2$ for a random $t \in (0,1)$. This procedure is repeated until the dataset is balanced, with which we were able to train a better classifier.

We designed a classification-regression pipeline to increase regression performance especially for long (> 4 days) stays. We tried four models (**Logistic Regression, XGBoost, Ensemble, Neural Network**) and selected the best-performing classifier based on balanced accuracy and F1 score. For each model, we use the binary logistic (cross-entropy) loss function

$$L(y,p) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)],$$

where p contains the predicted probabilities and y contains the labels $\{0,1\}$ representing short and long stays. LoS is predicted using one of three regressors based on the classifier's predicted probabilities (confidence) as shown below:



Using the ensemble classifier, we compared the performance of three regressors (**SGD, ElasticNet, XGBoost**) each trained with objective function

$$J(w) = \frac{1}{N}\sum_{i=1}^{N}(y_i - w^T x_i)^2 + \alpha \|w\|_2^2 + \beta \|w\|_1,$$

where w is the weights to be learned with $\alpha = \beta = 0$ for SGD and XGBoost scoring while $\alpha = \beta = 0.5$ for ElasticNet, which combines Ridge and Lasso regularization.

Discussion

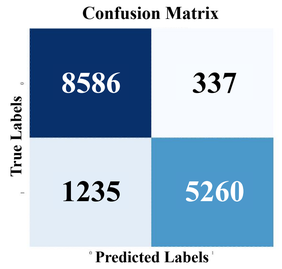
- We found that the Ensemble model is the best classifier with our data-balancing method and combining it with XGBoost gives the best regression result.
- This is expected as Ensemble can enhance prediction accuracy at the cost of additional computational complexity, though still relatively efficient.
- Our multi-stage pipeline approach reduced regression error effectively compared to baselines. However, this may still not be accurate enough in the medical setting, especially for longer stays.
- Predicting long LoS (> 4 days) involves modeling complex stochastic processes and requires larger dataset and more sophisticated models.

Results

Classification (Best Fold)

Model	Accuracy	Balanced Accuracy	Precision	Recall	F1 Score
Logistic	0.71	0.70	0.75	0.74	0.74
XGBoost	**0.72**	**0.71**	0.74	0.80	**0.77**
Ensemble	**0.72**	**0.71**	0.74	**0.81**	**0.77**
NN	**0.72**	**0.71**	**0.76**	0.77	0.76

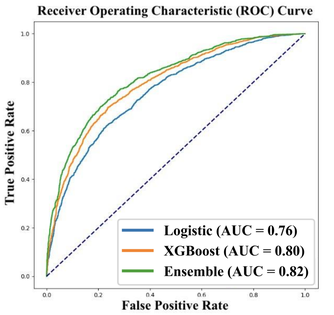
Confusion Matrix



The **Ensemble** model of random forest and XGBoost performs best in four of the five metrics above and has the **highest AUC** as shown. We also evaluate the model on the **fairness metrics** of distribution parity (PR), equalized opportunity (TPR), and equalized odds (TPR and FPR) for gender and race.

Groups	TPR	FPR	TNR	FNR	PR
Female	0.577	0.186	0.814	0.423	0.344
Male	0.645	0.222	0.778	0.355	0.412
Asian	0.510	0.161	0.839	0.490	0.310
Black/African-American	0.587	0.171	0.829	0.413	0.340
Hispanic/Latino	0.534	0.161	0.839	0.466	0.295
White	0.618	0.214	0.786	0.382	0.387
Other/Unknown	0.675	0.214	0.786	0.325	0.431

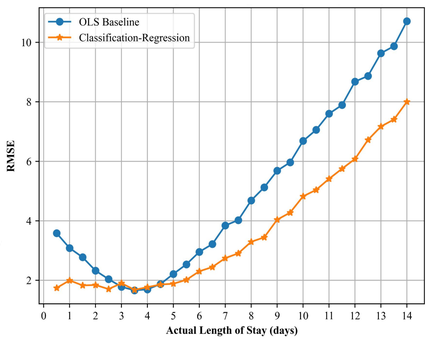
Receiver Operating Characteristic (ROC) Curve



Regression (Best Fold)

Regressor	RMSE	MAE	R^2
SGD	2.53 (2.33)	1.84 (1.68)	**0.29** (0.39)
ElasticNet	2.56 (2.29)	1.87 (1.68)	0.27 (0.41)
XGBoost	**2.52** (2.28)	**1.83** (1.65)	**0.29** (0.42)

Prediction RMSE by Actual LoS (every 0.5 days)



The best regressor we found, in combination with the Ensemble model for the classification stage, was **XGBoost** with better metrics across the board. RMSE by LoS now monotonically increases and is significantly better for longer stays.

Future Research

- We can consider more complex deep learning models and deal with natural language data using transformers and pretrained language models.
- Under this framework, we can make interim predictions for patients already admitted for certain days with additional time-series data.

References

[1] Alistair Johnson et al. 2023a. MIMIC-IV (version 2.2). PhysioNet.
[2] Hsiu Wu et al. 2020. Hospital capacities and shortages of healthcare resources among US hospitals during the coronavirus disease 2019 (COVID-19) pandemic. National Healthcare Safety Network.
[3] N.V. Chawla et al. 2002. Smote: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321-357.