**Predicting Hospital Length of Stay (LOS) at Admission**

**Business Problem**

Hospitals are reimbursed a fixed lump-sum amount based on diagnosis-related groups (DRGs), regardless of how long patients stay.

* **If patients stay longer than expected → hospital profitability decreases.**
* **If patients are discharged too early → risk of readmissions and worse patient outcomes.**

👉 Goal: **Classifying (LOS) at admission** to determine if the stay is **short (2 days or less)** or **long (more than 2 days)** using only information available at the time of admission.

**Dataset Description**

Each row represents **a single patient admission** to a New York hospital.

* **Target column**:
  + length\_of\_stay: number of days admitted (numeric, capped/log-transformed for modeling).
* **Admission-time features (used for modeling):**
  + age\_group: patient age bucket (e.g., *0–17*, *18–29*, *30–49*, *50–69*, *70+*).
  + gender: patient gender.
  + race: self-reported race category.
  + birth\_weight: newborn birth weight (grams).
  + apr\_severity\_of\_illness: categorical severity indicator (*Minor, Moderate, Major, Extreme*).
  + apr\_risk\_of\_mortality: categorical risk indicator (*Minor, Moderate, Major, Extreme*).
  + apr\_mdc\_description: medical condition description
  + apr\_drg\_description: refined Diagnosis Related Group
  + ccsr\_diagnosis\_description: primary diagnosis (categorical).
  + ccsr\_procedure\_description: primary procedure (categorical, can be empty for medical admissions).
  + payment\_typology\_1/2/3: insurance/payment source (e.g., *Medicare, Medicaid, Private*).
  + birth\_weight\_missing: Derived variable as birth weight is often missing for adults, then birth\_weight\_missing=1 encodes patients who are not newborns.
* **Excluded features (not available at admission, hence dropped):**
  + patient\_disposition: discharge outcome (home, rehab, expired, etc.).
  + total\_charges, total\_costs: final hospital bill.
  + discharge\_year: discharge timestamp.
  + hospital\_service\_area, hospital\_county, operating\_certificate\_number, permanent\_facility\_id, zip\_code\_3\_digits : To generalize the model
  + ccsr\_diagnosis\_code, ccsr\_procedure\_code, apr\_drg\_code, apr\_mdc\_code: Duplicate with descriptions

This ensures the model only uses **information available at admission**, avoiding data leakage.

**Exploratory Data Analysis (EDA)**

* **Target distribution (LOS)**: highly right-skewed with extreme outliers (>30 days).
* **Class imbalance**: majority of patients stay ≤5 days.
* **Key categorical drivers**: severity of illness, risk of mortality, procedure codes.
* **Numeric drivers**: birth weight, age groups.

A graph of a number of patients

AI-generated content may be incorrect.

A graph of a bar graph

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

**Modeling Approach**

**Baseline Models**

* **Logistic Regression**: poor fit, could not capture non-linearity.
* **KNN Classifier**: too sensitive to noise.
* **SVR**: computationally heavy, didn’t converge.

**Advanced Models**

* **Deep Learning (Keras MLP)**: underperformed due to tabular nature of data.
* **Stacking Regressors**: did not improve over single models.

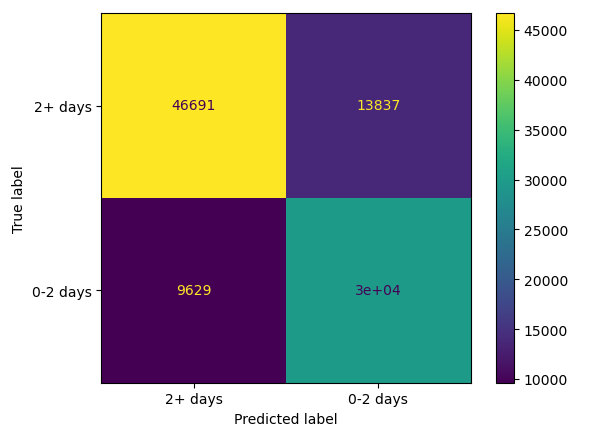
**Best Model: Tuned XGBoost**

* Applied **log-transform** on LOS target.
* Hyperparameters tuned via **Bayesian Optimization**.

*Results (Test Set):*

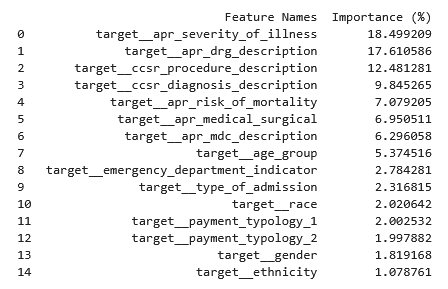
* Accuracy: 0.7756
* Classification Report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Label | Precision | Recall | F1-score | Support |
| 0-2 days | 0.69 | 0.77 | 0.73 | 60528 |
| 2+ days | 0.84 | 0.78 | 0.81 | 39437 |



**Model Interpretability (SHAP)**

* **Top Predictors:**
  + Severity of illness
  + DRG description
  + Procedure descriptions
  + Diagnosis groupings
  + Risk of mortality



A graph with blue and pink dots

AI-generated content may be incorrect.

**Limitations**

* **Data limitations**:
  + No data for the total number of beds and the number of available beds at admission
  + No lab values, vitals, or comorbidities at admission.
  + Some discharge-only variables (costs, disposition) had to be excluded.
* **Target skew**: long-stay patients (>20 days) are rare and hard to predict.
* **Single-center bias**: model may not generalize to other regions.

**Future Improvements**

* Add **lab results, vitals, comorbidity indices** at admission.
* Expand to a **multi-class classification.**
* Add **number of beds** and **the number of available beds** at admission for each hospital
* Use **time-series data** (day-by-day patient progression).
* Try **survival models (Cox, DeepSurv)** instead of plain classification.