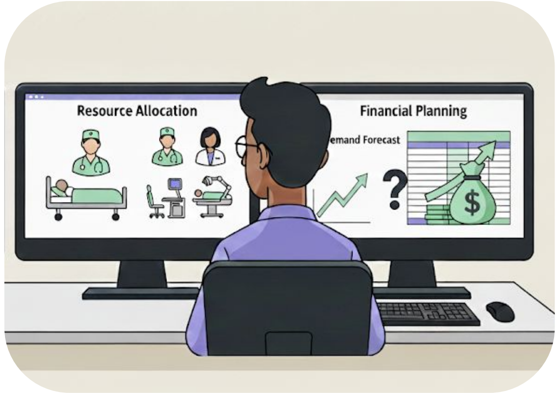
Predicting Hospital Length of Stay

By: Allison, Anna, Parisha, and Samuel

Table of Contents

01 Introduction + Problem Statement



Problem Statement

Hospitals face pressure to provide high quality care while staying operationally efficient. A major driver of both patient outcomes and financial performance is **length of stay (LOS).**

Using the **SPARCS inpatient dataset**, we built a regression model to predict LOS for each patient.

Better LOS Predictions can Help Hospitals:

1. Forecast Demand More Effectively
2. Allocate Staff and Resources
3. Improve Budgeting and Financial Planning

02 Data Prepping + Cleaning Process

Data Preparation & Cleaning

Examining Target Variable

Checked all unique values in **Length of Stay** to understand distribution

Converted Length of Stay into an **integer** for analysis

Handled Data Types & Missing Values

Dropped columns not applicable to analysis

Removed non-numeric values like “120+” from Length of Stay

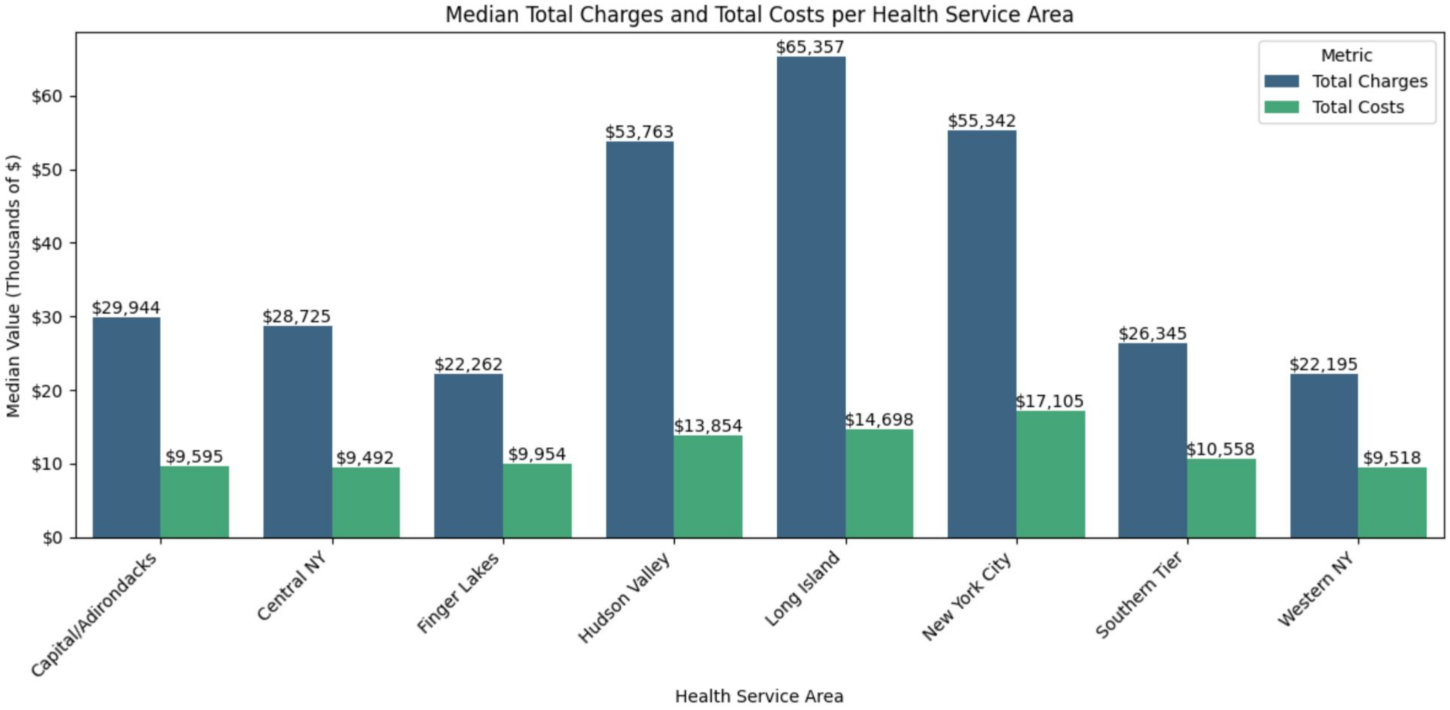
Converted variables **object → category**

Finalized a Clean Dataset

Numeric Variables (LOS, Total Cost / Charges)

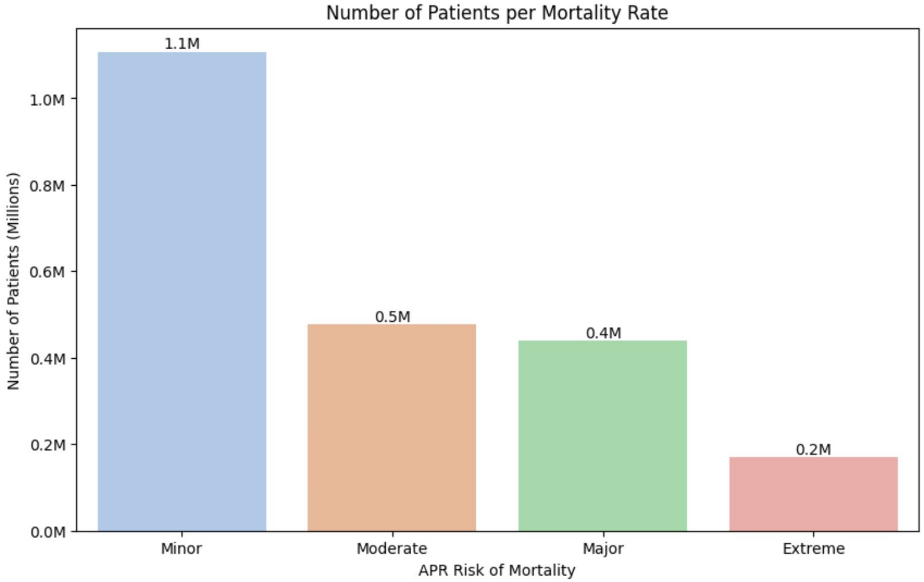
Categorical Variables (Risk of Mortality, Age Group / Gender)

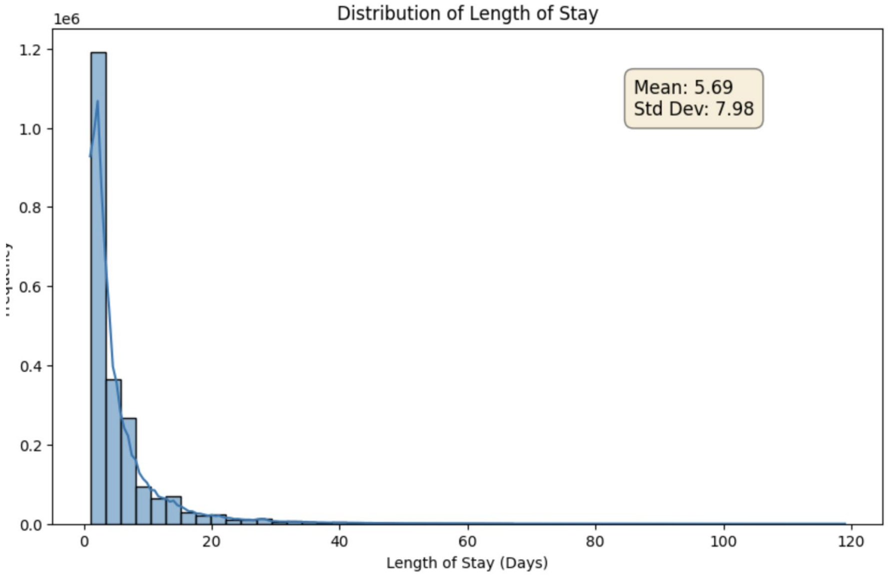
EDA: Preliminary Visualizations



Charges: The amount the hospital billed for the inpatient stay
Costs: The hospital's estimated cost to provide the care

EDA: Preliminary Visualizations





03 Choosing + Tuning Best Models

Initial Model Comparison

Model	MAE	RMSE	R2	Tuning
Linear Regression	3.238	6.473	0.439	Yes
Decision Tree	3.640	7.204	0.305	Yes
Linear SVR	2.938	6.805	0.380	No
Random Forest	4.139	7.717	0.202	No
Light Gradient Boosting	2.994	6.218	0.482	Yes

1. Linear Regression

We began with a **linear regression**, which tries to draw the best straight line by choosing coefficients for each feature.

We then attempted to optimize our linear regression with:

- **Ridge regression:** shrinks all coefficients, keeps every feature
- **Lasso regression:** shrinks coefficients, sets some to zero, picks important features

Model	MAE	RMSE	R2
Linear Regression	3.238	6.473	0.439
Ridge Regression	3.237	6.472	0.439
Lasso Regression	3.634	7.227	0.297

2. Decision Tree

Next, we did a **decision tree**, a model that splits the data into branches based on feature values to make predictions. Can capture **nonlinear relationships**, but are prone to **overfitting.**

We then attempted to optimize our decision tree with:

- **RandomizedSearchCV:** tunes max depth, min samples split, and min samples leaf to improve performance

Model	**MAE**	**RMSE**	**R2**
Decision Tree	3.640	7.204	0.305
Randomized Search	3.222	6.676	0.403

Decision Tree Nodes

3. Linear SVR

- XXX

Model	MAE	RMSE	R2

4. Light Gradient Boost

Lastly, we used **Light Gradient Boosting** which is an ensemble model that builds many small decision trees and learns patterns by boosting mistakes from previous trees.

We tested several LightGBM variations:

Model	MAE	RMSE	R2
Base Model	2.994	6.218	0.482
Tuned Model	2.884	6.061	0.505
SFS with Pre-filtering (10% sample dataset)	4.359	68.390	0.082

Feature Importance

1. Clinical Severity (Strongest Predictors)

- APR **Severity of Illness** and **Risk of Mortality** categories appear repeatedly at the top.
- These measures clearly show that **patient acuity is the primary driver of LOS**.

2. Discharge Disposition

- "Home/Self Care," "Skilled Nursing Facility," "Home with Health Services," and "Expired" all rank highly.
- Indicates that **patients with complex discharge needs tend to stay longer**.

3. Demographics

- Gender (F), older **Age Groups (50–69, 70+)**, and multiple **Race/Ethnicity** indicators contribute meaningful signal.
- Suggests **population differences in care patterns**.

4. Healthcare System & Payer Factors

- Hospital location (New York County) and **Medicare/Medicaid** coverage show notable importance.
- Reflects **system-level and socioeconomic influences** on LOS

04 Model Stacking Process

Our Stacking Process

Linear Regression (Ridge) — Starts the stack with a simple model to establish baseline predictive ability

Decision Tree — Adds structure by learning rule based patterns and interactions

Light Gradient Boost — Builds on the earlier models by correcting their errors using many boosted trees

Final Model — Integrates all model predictions to deliver the most accurate and stable output

Implementing the model

Current Situation

The best estimation for LOS is 3 days (median) or 5.8 days (mean)

Differences between patients conditions are overlooked

The best estimation for cost per patient is $17.1K (median) or $31.4k (mean)

Staffing is static and changes reactively instead of proactively

With Our Model

Model provides individualized predictions tailored to each patient based on symptoms

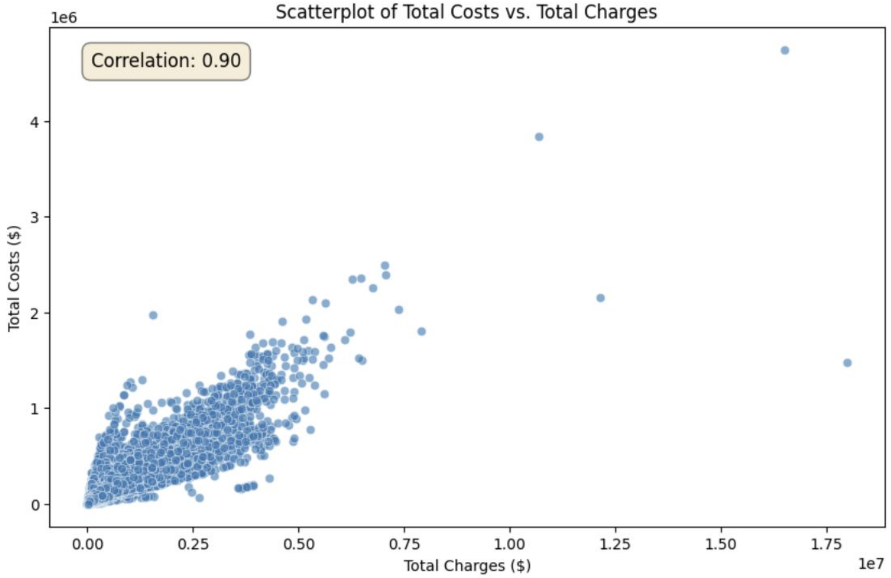
The prediction's error is 2.8 days (MAE) and for atypical patients is 6.1 days (RMSE)

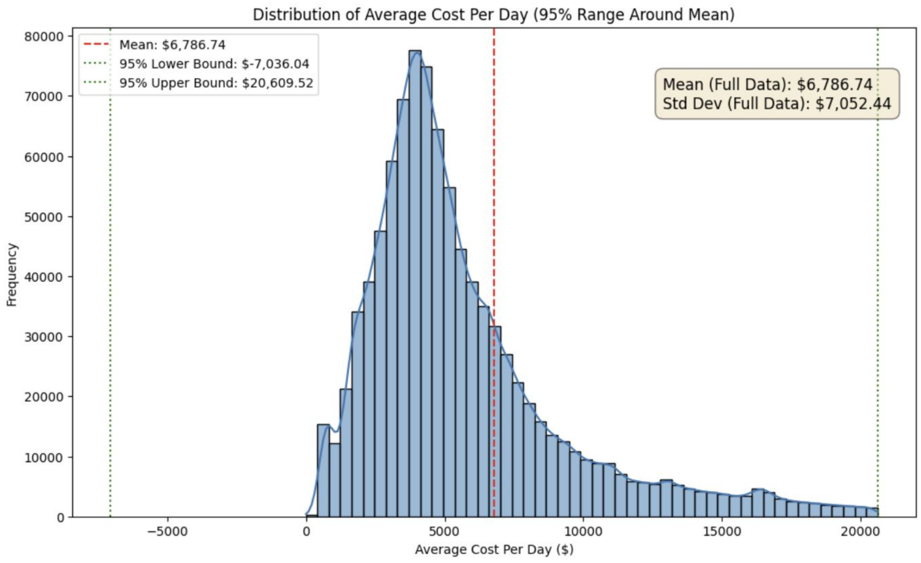
There is more certainty on the estimated cost per patient

Staffing plans can now be adjusted according to the occupation expectatives on a weekly or semi-weekly basis

05 Our Findings + Business Insights

Our Findings





Business Insights

Acuity

Clinical Severity Drives LOS More Than Demographics

Reimbursement

Payer Mix Influences Stay Duration and Utilization

Discharge Planning

Discharge Destination Significantly Impacts Predictions

Triage

Emergency Admissions Need Different Resource Planning

06 Recommendations + Limitations

Recommendations Benefits from LOS Prediction

Capacity and Bed Planning

By correctly predicting patient's length of stay...

1. Health facilities can anticipate discharge dates
2. Balance and Manage ER admissions
3. Determine bed turnover

Staffing Optimization

By estimating patient volume, health facilities can...

1. Plan workloads more effectively
2. Maintain balanced nurse-to-patient ratios
3. Reduce overtime and agency staffing costs

Financial Forecasting & Cost Management

By accurately predicting length of stay, hospitals can...

1. Improve budgeting and cost visibility
2. Reduce unreimbursed or avoidable hospital days
3. Inform capacity and investment decisions

Project Limitations

1. **Large dataset (~6 million rows)**
   - Some tuning took ~55 minutes
   - To reduce computation time, took **10% sample** for feature selection & **CV folds** = 3 during hyperparameter tuning

2. **Limited Clinical Detail**
   - Dataset mainly includes **demographics/clinical codes**
   - **No lab results/vitals,** restricting the model's ability to capture full patient complexity

3. **LOS Distribution Highly Skewed**
   - Many short stays, few very long stays
   - **Right-skewed LOS data** makes extreme cases harder to predict accurately

THANK YOU !

Educational Icons

Medical Icons