Predicting Patient Length of Stay (LOS) for Efficient Hospital Management

Mohammad Shahnewaz Morshed

10 March 2024

Abstract

A statistical framework was developed to predict hospital Length of Stay (LOS) at admission, a critical, discrete, and non-negative outcome in healthcare management. To address the right-skewed distribution and overdispersion characteristic of LOS data, tailored count-based models were employed to account for greater variability than standard approaches permit. Additionally, mixture models were integrated to capture the frequent occurrence of short stays, such as same-day or one-day admissions, ensuring accurate differentiation from longer hospitalizations. These methodological enhancements significantly improved prediction accuracy and provided well-calibrated forecasts, supporting robust hospital resource planning.

1 Problem Formulation

Length of Stay (LOS) is the number of calendar days between admission and discharge. Statistically, LOS is a non-negative, discrete random variable $Y \in \{0, 1, 2, \dots\}$. Empirically, LOS exhibits right skewness and overdispersion ($\mathrm{Var}(Y) > \mathbb{E}[Y]$) where the variance exceeds the mean. Ignoring these properties leads to inefficient estimates and biased predictions, which in turn generates misallocation of hospital resources.

2 Methodology

2.1 Statistical Characterization of LOS Predictors

The modeling approach was initially formulated using Poisson regression, which is the canonical model for count outcomes. In Poisson regression, the conditional mean of the outcome Y_i given covariates X_i is modeled as

$$\mathbb{E}[Y_i|X_i] = \mu_i, \quad \log(\mu_i) = X_i^\top \beta,$$

with the assumption of equidispersion, i.e.,

$$\mathrm{Var}(Y_i|X_i) = \mu_i.$$

This model is considered canonical because it directly leverages the Poisson distribution's properties, using a logarithmic link function to ensure non-negative predictions and providing interpretable coefficients as incidence rate ratios. Its simplicity and well-established maximum likelihood estimation make it the standard starting point for modeling count data in statistical practice.

An empirical analysis of the LOS variable was conducted, revealing that the sample variance substantially exceeded the sample mean, indicating overdispersion. Under these conditions, Poisson regression leads to underestimated standard errors and inflated Type I error rates.

Consequently, the model was extended to the Negative Binomial (NB) specification, which introduces a dispersion parameter $\alpha > 0$ and relaxes the equidispersion assumption by allowing

$$\mathrm{Var}(Y_i|X_i) = \mu_i + \alpha\mu_i^2.$$

This parameterization accommodates greater variability in LOS and provides more reliable inference when patient heterogeneity produces higher variance than the Poisson model permits.

Further examination of the distribution was performed, indicating a higher frequency of very short stays (including same-day or near-zero discharges) than predicted by either Poisson or NB models. In such cases, standard count models fail to capture the probability mass at zero or near zero. A Zero-Inflated model was therefore adopted, which assumes that the data are generated by a mixture of two processes: one that produces structural zeros with probability π, and another that follows a count distribution (Poisson or NB) with probability $1 - \pi$. The probability mass function of the Zero-Inflated Poisson (ZIP) model is given by

$$\Pr(Y_i = 0) = \pi + (1-\pi)e^{-\mu_i}, \quad \Pr(Y_i = y) = (1-\pi)\frac{e^{-\mu_i}\mu_i^y}{y!}, \quad y > 0.$$

This specification enables the excess occurrence of short stays to be captured without distorting the overall mean–variance relationship of the count distribution.

Concurrently, the predictor space, which included high-dimensional categorical variables (e.g., physician identifiers, ward codes) with potentially weak or redundant effects, was addressed through penalized regression methods. The Least Absolute Shrinkage and Selection Operator (LASSO) estimator was applied, minimizing

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i^\top \beta)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where the penalty term $\lambda \sum |\beta_j|$ enforces sparsity. This property forces many coefficients toward zero, thereby selecting only the most influential predictors and reducing variance from overfitting. The Elastic Net was also employed, combining LASSO's sparsity (ℓ_1) with Ridge's grouping and stability (ℓ_2) properties, making it particularly suitable when predictors are correlated.

Model interpretability was improved through the application of Shapley value decomposition, derived from cooperative game theory. For each patient i, the predicted LOS $\hat{y}_i$ was decomposed as

$$\hat{y}_i = \phi_0 + \sum_{j=1}^{p} \phi_{ij},$$

where ϕ_{ij} denotes the average marginal contribution of predictor j to the prediction for observation i, calculated across all possible orderings of feature inclusion. This method was selected for its provision of a rigorous, distribution-based measure of predictor importance, ensuring consistency across diverse model classes and delivering locally accurate attributions at the individual prediction level. Unlike regression coefficients, which depend on linearity assumptions, Shapley value decomposition maintains additive and comparable contributions, even in nonlinear models.

2.2 Development and Validation of Predictive ML Models: Capturing Nonlinearities and Interactions

To effectively model the complex, nonlinear relationships and interactions among predictors influencing hospital Length of Stay (LOS), tree-based ensemble methods, including Random

Forest (RF), Gradient Boosting Machine (GBM), and Extreme Gradient Boosting (XGBoost), were employed. These models were trained and validated using k-fold Grouped Cross-Validation (CV) with grouping by patientid to prevent information leakage across multiple admissions of the same patient, ensuring robust and unbiased performance estimates. Model performance was evaluated using standard regression metrics, namely Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2), to assess prediction accuracy and goodness-of-fit. Additionally, to explore classification performance, LOS was optionally discretized into categorical bins, enabling the assessment of Area Under the Receiver Operating Characteristic Curve (AUROC) and F1-score for predictive discrimination and balance between precision and recall. Calibration diagnostics, such as reliability curves, were incorporated to verify the alignment between predicted and observed LOS values, ensuring the reliability of probabilistic forecasts for operational use.

2.3 Operational Insights and Policy Implications

Predictions were translated into actionable insights to optimize hospital resource allocation. Average Marginal Effects (AMEs), Partial Dependence, and SHAP interaction effects were computed to quantify the impact of operational levers, such as staffing levels and bed capacity, on Length of Stay (LOS). AMEs were used to estimate the average change in LOS associated with a unit change in a predictor, providing a clear measure for resource planning. Partial Dependence analyses were conducted to illustrate how individual predictors influence LOS across their range, offering insights into their marginal effects. SHAP interaction effects were analyzed to capture the combined impacts of predictor pairs, enhancing the interpretability of complex relationships. Additionally, scenario analyses were performed to simulate changes in LOS under hypothetical resource allocation scenarios, such as increased staffing or bed availability. These analyses, while informative, were associational in nature and require integration with causal inference methods to establish definitive causal relationships.