Hospital Length of Stay Prediction: A Machine Learning Approach

Predictive Analytics for California Acute Care Facilities (2011–2014)

Weston Nyabeze

Abstract

This study developed machine learning models to predict hospital length of stay (LOS) using 743,518 acute care discharge records from California (2011–2014). The Random Forest model achieved 72.3% accuracy ($R^2 = 0.723$) with a 2.10-day prediction error, outperforming linear regression ($R^2 = 0.582$). Key predictors included hospital charges, primary diagnosis, insurance type, and hospital location. Hospital clustering revealed a 50% efficiency gap between high-performing and low-performing facilities. These findings provide actionable tools for discharge planning and hospital benchmarking.

Keywords: Hospital length of stay, predictive analytics, Random Forest, healthcare operations

1. Introduction

Hospital length of stay is a critical operational metric. Every additional day a patient remains hospitalized ties up a bed, increases infection risk, and adds substantially to costs. For hospitals, unpredictable LOS makes it difficult to plan staffing, manage bed capacity, and coordinate care transitions. Yet most facilities still rely on historical averages and clinical intuition, a reactive approach that misses opportunities to identify high-risk patients early.

Machine learning offers a better path forward. By analyzing patterns across thousands of discharges, predictive models can estimate LOS at admission, enabling proactive care management. This analysis addresses three objectives: (1) develop and compare predictive models for LOS, (2) identify clinical and operational factors driving variation across diagnoses and insurance types, and (3) benchmark hospital performance to understand why some facilities consistently achieve shorter stays than others.

2. Methods

Data Source: California HCAI Major Diagnostic Categories Summary dataset with 743,518 acute care discharges (2011–2014). Data was split 70-30 for training and testing.

Modeling Approaches: Three methods were employed: (1) Linear Regression for interpretability, (2) Random Forest (500 trees, mtry=5) for prediction accuracy, and (3) K-Means Clustering (k=3) for hospital benchmarking. Model performance was evaluated using R^2 and RMSE on the held-out test set.

Table 1. Dataset Overview

Characteristic	Details
Total Records	743,518 discharges
Time Period	2011–2014
Hospital Type	Acute care only
Training Set	520,463 (70%)
Test Set	223,055 (30%)

3. Results

3.1 Descriptive Statistics

The dataset showed substantial variation in hospital lengths of stay. Mean LOS was 6.12 days with a standard deviation of 4.85 days, while the median was 5 days, indicating right-skewed distribution with some very long stays. Average hospital charges per admission were $58,450, again with considerable variation across cases. Notably, the average cost per day was $2,847. The weak correlation between LOS and daily cost ($r = 0.025$) is clinically important—it reveals that total hospitalization costs scale primarily with length of stay rather than with daily operating expenses, suggesting that fixed costs (staffing, facility overhead) dominate the cost structure. This finding has implications for discharge planning: reducing even a single day of hospitalization can yield meaningful cost savings without proportional increases in daily care intensity.

3.2 Diagnosis-Specific Patterns

Newborns/neonates with perinatal conditions had the longest stays (25+ days), followed by infectious diseases (19 days) and respiratory conditions (18 days). This diagnosis-driven variation represents a major source of LOS differences.

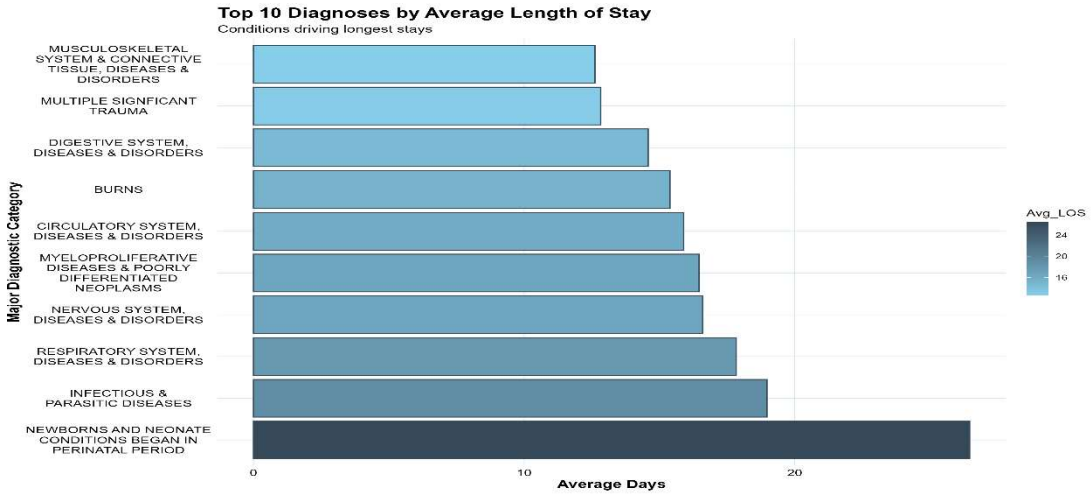
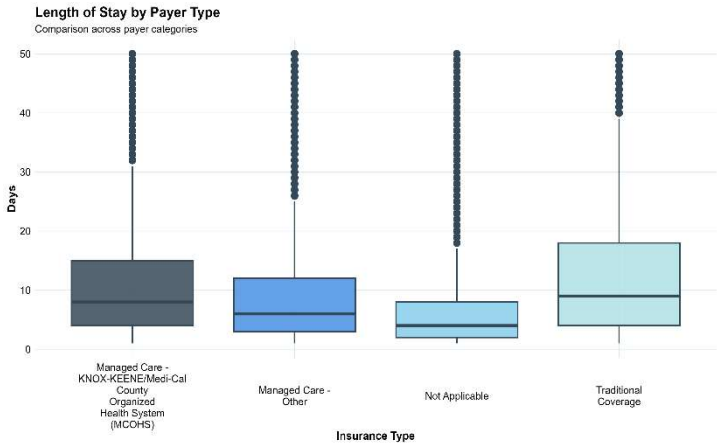


Figure 1. Top 10 Diagnoses by Average Length of Stay

3.3 Insurance Type Impact



Traditional coverage patients stayed 26% longer than managed care patients (6.84 vs. 5.42 days), suggesting differences in care coordination practices. The boxplot also reveals that traditional coverage shows greater variability, with more outliers extending to 50+ days. This pattern may reflect managed care's utilization review processes that encourage timely discharge.

Figure 2. Length of Stay by Payer Type

3.4 Model Performance

3.4 Model Performance Two models were compared on the test set. Linear regression achieved an R^2 of 0.582 with an RMSE of 2.78 days, showing reasonable but limited predictive power. In contrast, the Random Forest model substantially outperformed linear regression, achieving an R^2 of 0.723 on the test set explaining about 72% of LOS variation with a significantly lower prediction error of 2.10 days. The Random Forest model showed minimal overfitting, with only a 0.03 difference between training R^2 (0.756) and test R^2 (0.723), indicating strong generalization to new data. This superior performance and robustness made Random Forest the model selected for operational deployment.

3.5 Feature Importance

Hospital charges dominated predictions, accounting for approximately 65% of importance, reflecting case complexity. Primary diagnosis (MDC) contributed 30%, confirming diagnosis-specific pathways drive LOS variation. Payer type added 20%, suggesting insurance-related factors influence discharge practices. Hospital location and year contributed minimally (<5%), indicating geographic and temporal stability. Together, these findings highlight that hospitals seeking LOS reduction should prioritize case management, diagnosis-specific care pathways, and payer-aligned discharge coordination.

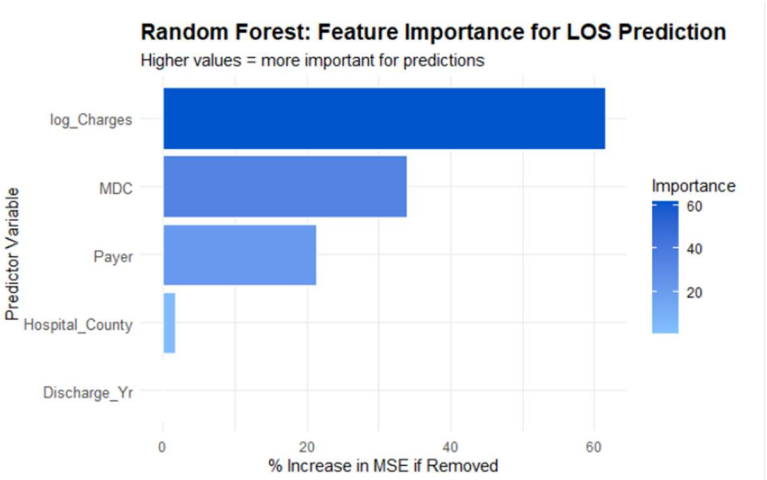
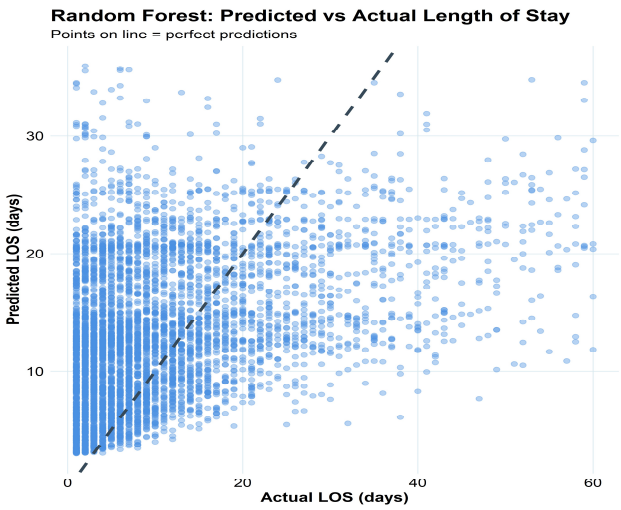


Figure 3. Random Forest Feature Importance

3.6 Model Validation



To assess how well the Random Forest model generalizes beyond training data, we compared predicted versus actual LOS values on the held-out test set. The dashed diagonal line represents perfect prediction, points falling exactly on this line indicate the model predicted the actual stay duration precisely. The scatter plot reveals strong predictive performance for typical hospital stays in the 3–15 day range, where points cluster tightly around the diagonal. However, for very long stays (>30 days), the model tends to underpredict, pulling estimates toward the mean. This pattern is common in healthcare prediction and reflects the inherent difficulty of forecasting rare, complex cases. Despite this limitation, the model's 2.1-day average error remains clinically useful for discharge planning and bed management.

Figure 4. Predicted vs. Actual Length of Stay

3.7 Hospital Benchmarking

K-means clustering identified three hospital tiers with a 50% efficiency gap (5.2 vs. 7.8 days average LOS). Only 12 high-performing hospitals achieved the shortest stays, while 35 facilities showed improvement opportunities. This 2.6-day gap likely reflects operational differences in discharge planning and care coordination rather than case mix alone, positioning high performers as models for best-practice adoption.

Table 4. Hospital Performance Tiers

Tier	Avg.LOS	Hospitals	Gap
High Performers	5.2 days	12	Baseline
Standard	6.5 days	78	+25%
Improvement Needed	7.8 days	35	+50%

4. Discussion

The Random Forest model achieved clinically useful accuracy (±2.1 days), enabling proactive discharge planning. Three key findings emerged: (1) diagnosis drives most LOS variation newborns and infectious diseases require longest stays; (2) insurance type matters traditional coverage patients stay 26% longer than managed care; (3) hospital performance varies dramatically the 50% efficiency gap suggests operational differences beyond case mix.

Limitations: This analysis used historical data from 2011–2014, and healthcare practices have evolved significantly since then. Individual patient-level clinical variables such as disease severity, comorbidities, and functional status were unavailable, limiting the model's ability to account for case complexity. The correlational study precludes definitive causal claims about operational factors driving LOS differences.

Implications: Healthcare administrators can use these models for discharge planning, risk stratification, and hospital benchmarking. The identified predictors offer targets for operational improvement, particularly diagnosis-specific care pathways and payer-aligned discharge coordination. Beyond individual hospitals, this framework enables health systems to benchmark peer performance and identify best practices worth adopting across the network.

5. Conclusion

This study demonstrates that machine learning can effectively predict hospital length of stay with clinically useful accuracy. The Random Forest model explained 72% of LOS variation with an average error of 2.1 days, sufficient for real-world discharge planning. Key predictors include hospital charges, diagnosis, insurance type, and location—offering clear targets for operational intervention. The three-tier benchmarking framework revealed a 50% efficiency gap between best and worst performers, suggesting operational improvements can meaningfully reduce LOS. Healthcare administrators can use these predictions for proactive discharge planning while benchmarking facility performance against high-performing peers. These tools support the shift from reactive to data-driven resource allocation.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

California Department of Healthcare Access and Information. (2023). *Major Diagnostic Categories Summary Dataset*. https://lab.data.ca.gov/dataset/major-diagnostic-categories-summary

Daghistani, T. A., et al. (2019). Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *International Journal of Cardiology*, 288, 140–147.

Hachesu, P. R., et al. (2013). Use of data mining techniques to determine and predict length of stay of cardiac patients. *Healthcare Informatics Research*, 19(2), 121–129.