Predicting Patient Length of Stay via Machine Learning and Correlation Analysis: A Data-Driven Approach to Optimize Hospital Resource Allocation and Operational Efficiency

Abstract

This study addresses hospital inefficiencies in resource allocation by predicting patient length of stay (LOS) to optimize bed and staff management. Using 318,438 hospitalization records, we will apply machine learning and correlation studies to identify key LOS predictors, including illness severity, age, and admission type. Expected outcomes include a classification model categorizing LOS into 11 intervals and actionable recommendations for clustering high-LOS patients. Results will enable hospitals to proactively allocate resources. The findings emphasize improving operational efficiency and healthcare delivery, though limitations like regional data specificity and model interpretability require further refinement for broader applicability.

Introduction/Background

Hospital management has exposed many inefficiencies during the COVID-19 pandemic, especially in the allocation of beds and medical resources (*Neha Prabhavalkar, 2021*). If hospitals can predict the length of a patient's stay when they are admitted, they can more effectively allocate wards, arrange nursing resources, reduce congestion, and thus improve overall medical efficiency (*Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N., 2022*).

In this context, we hope to help hospitals predict the specific length of stay of patients based on their basic information when they are admitted to the hospital. So our study will focus on the following questions: Which key variables have the greatest impact on LOS? Can machine learning models be used to accurately predict hospital stay? How can the prediction results be used to optimize hospital management? At the same time, we will focus on predicting length of stay using patient demographics, hospital characteristics, and admission information, rather than focusing on specific disease diagnoses and treatment options. The rationale behind this choice is threefold. First, considering diagnoses information will make the model more complex, as integrating detailed clinical information typically requires more detailed data and expertise. Second, focusing on widely available data (e.g., Age, Bed Grade, Deposit) ensures that the resulting model is more generalizable and can be deployed across a variety of healthcare settings without the need for disease-specific modules. Finally, from an operational perspective, many hospitals (especially those with limited resources) may not have complete diagnostic details available immediately upon admission, and therefore a more general hospital length prediction approach would be more practical and scalable.

This can be done through machine learning techniques (*Jain, R., Singh, M., Rao, A.R. et al., 2024*). The results of this study can help hospitals identify high-risk patients in advance, improve bed and staff allocation, and enhance the operational efficiency of the medical system.

Goal/ Objectives

Objective 1 - Identification of all variables that present a positive correlation with patients' length of stay: our first goal is study the correlation between all dataset variables and LOS, extrapolating the relevant ones and giving an explanation to why they present a positive or negative correlation.

Objective 2 - Identification of clusters of patients whose length of stay substantially deviates from the average: after studying all relevant correlations, we will analyze cross-correlations and try to cluster patients whose characteristics make them more likely to stay hospitalized for a prolonged amount of time.

Objective 3 - Creation of a prediction model through machine learning: through our training and test datasets, we aim to create a model that hospitals can use to accurately predict the expected LOS by inputting the relevant values.

Objective 4 - Recommendations to increase the hospital efficiency: once the model is created and ready for use by the hospitals, we will provide recommendations on how to use such a model, and the consequent adjustments that are expected to increase efficiency based on the model results.

Methods

Data description

The dataset contains 318,438 hospitalization records, covering patient information, hospital characteristics, admission type, severity of illness, etc. The target variable Stay is a categorical variable, indicating the length of hospital stay, which is divided into 11 categories. A summary of the variables found in dataset is shown in table 1:

Table 1 | Data Description Table

Column	*Data Type*	*Null Values (Yes/No)*	*Description*	*Sample Size*
Case_id	Integer	0	Case_ID registered in Hospital	318438
Hospital_code	Integer	0	Unique code for the Hospital	318438
Hospital_type_code	String	0	Unique code for the type of Hospital	318438
City_Code_Hospital	Integer	0	City Code of the Hospital	318438
Hospital_region_code	String	0	Region Code of the Hospital	318438
Available Extra Rooms in Hospital	Integer	0	Number of Extra rooms available in the Hospital	318438
Department	String	0	Department overlooking the case	318438
Ward_Type	String	0	Code for the Ward type	318438
Ward_Facility_Code	String	0	Code for the Ward Facility	318438
Bed Grade	Float	113	Condition of Bed in the Ward	318438
Patientid	Integer	0	Unique Patient Id	318438
City_Code_Patient	Float	4532	City Code for the patient	318438
Type of Admission	String	0	Admission Type registered by the Hospital	318438
Severity of Illness	String	0	Severity of the illness recorded at the time of admission	318438
Visitors with Patient	Integer	0	Number of Visitors with the patient	318438
Age	String	0	Age of the patient	318438
Admission_Deposit	Float	0	Deposit at the Admission Time	318438
Stay	String	0	Stay Days by the patient	318438

We also calculated some statistics about number variables, presented in image 1:

Image 1 | Descriptive Statistics

	Available Extra Rooms in Hospital	Bed Grade	Visitors with Patient	Admission_Deposit
count	318438.000000	318325.000000	318438.000000	318438.000000
mean	3.197627	2.625807	3.284099	4880.749392
std	1.168171	0.873146	1.764061	1086.776254
min	0.000000	1.000000	0.000000	1800.000000
25%	2.000000	2.000000	2.000000	4186.000000
50%	3.000000	3.000000	3.000000	4741.000000
75%	4.000000	3.000000	4.000000	5409.000000
max	24.000000	4.000000	32.000000	11008.000000

This is a very important step so that we can better understand the numerical characteristics of each variable and facilitate subsequent analysis.

In this study, the dependent variable is "Stay", while the independent variables have been selected based on expected influence factors on LOS (*Mieke Deschepper, Chloë De Smedt, Kirsten Colpaert, 2025*). Several dummy variables have been created for the categories “Department, Type of Admission, and Severity of Illness”. The independent variables are: Radiotherapy (dummy variable), Anesthesia (dummy variable), Gynecology (dummy variable), TB & Chest Disease (dummy variable), Surgery (dummy variable), Bed Grade, Trauma (dummy variable), Urgent (dummy variable), Emergency (dummy variable), Minor (dummy variable), Moderate (dummy variable), Extreme (dummy variable), Visitors with Patient, Randomized Age, Admission Deposit.

The regression model will take the form of:

$$\text{Stay} = \beta_0 + \text{Radiotherapy} \times \beta_1 + \text{Anesthesia} \times \beta_2 + \text{Gynecology} \times \beta_3 + \text{TB \& Chest Disease} \times \beta_4 + \text{Surgery} \times \beta_5 + \text{Bed Grade} \times \beta_6 + \text{Trauma} \times \beta_7 + \text{Urgent} \times \beta_8 + \text{Emergency} \times \beta_9 + \text{Minor} \times \beta_{10} + \text{Moderate} \times \beta_{11} + \text{Extreme} \times \beta_{12} + \text{Visitors with Patient} \times \beta_{13} + \text{Randomized Age} \times \beta_{14} + \text{Admission Deposit} \times \beta_{15}$$

Data Preprocessing

Process missing values, select the median to fill missing values of numerical types, and select the mode to fill missing values of categorical data.

Afterwards, we perform numerical mapping encoding on ordered categorical variables, such as the severity of the disease, and perform dummy variable processing on non-ordered categorical

variables.

Feature Engineering

To reduce the dimensionality of the dataset and address potential collinearity between features, we employ principal component analysis (PCA). Here are the steps:

1. *Data Standardization:* We first normalize each feature to have zero mean and unit variance. This ensures that variables at different scales do not overly influence the principal components.
2. *Covariance Matrix Calculation:* After normalization, we calculate the covariance matrix to capture the linear relationship between features.
3. *Principal Component Extraction:* We perform eigendecomposition (or singular value decomposition, SVD) on the covariance matrix to identify the principal components (PCs)—the directions in which the data vary the most.
4. *Component Selection:* We keep the previous k principal components that together explain a sufficiently large fraction of the total variance (e.g., 90% or 95%). By doing this, we effectively reduce the number of input variables while retaining most of the original information.
5. *Feature Transformation:* Finally, we project the original data onto these selected principal components. The transformed data (now in a low-dimensional space) serves as input for our final modeling phase.

By using PCA, we aimed to capture the most significant variations in the data with fewer dimensions, thereby simplifying the model, reducing the risk of overfitting, and potentially improving the performance of hospital stay prediction.

Model building

We selected multiple machine learning models for comparison and finally selected the best model for prediction: First, we selected the logistic regression benchmark model to evaluate the model performance. It provides basic interpretability, but may not capture complex nonlinear relationships. Then we used a random forest for prediction. It is suitable for high-dimensional data and can capture nonlinear relationships. Finally, we used XGBoost, which has strong generalization ability and is suitable for class imbalanced data. It can automatically learn the interaction between features and improve prediction accuracy. After completing the model building, we calculated the accuracy, F1 score, and confusion matrix to evaluate the model.

Correlation Analysis

In addition to building a model that can accurately predict, we also performed correlation analysis. For categorical variables, we used Dummy Variables to transform them and used R Studio to calculate their correlation with LOS.

Expected outcomes

1. More severe illnesses / injury types, and specific departments are expected to generate relevant correlations. The patient's age is also expected to be positively correlated to the LOS. Analysis will be conducted on the initial admission deposit to explore a potential correlation.

2. Clusters are expected to be composed of patients suffering from genetic diseases (starting at a relatively young age), severe diseases that require urgent attention, and older patients with high initial deposits due to the intensity of future care needed.

3. The model will be fed the train dataset, and then analyzed with a test dataset. It is expected to predict the length of stay of a patient based on all relevant variables, placing each observation in the corresponding 10-days ordinal categorical binning group.

4. Based on our results, we are expecting to be able to recommend a management model that avoids bottlenecks and inefficient allocation of patients in a hospital. It will involve clustering similar LOS patients in same or adjacent rooms, possibly by category of treatment needed, and using the least amount of resources for the most amount of patients at the same time, without forgoing the same treatment quality level.

Tables analysis

Through an early analysis of our data, we have explored the distribution of some interest variables, reported in the box-plots of table (3), in relation to our dependent variable length-of-stay. From the table, it is possible to notice some initial associations between specific categories of patients (like admissions to surgery department, or extreme severity of illness) and a relatively higher expected length of stay. In table (4) instead, a multi-function relation of admission deposit grouped by Stay intervals is shown. In this table, mean, minimum, maximum,

and standard deviation of admission deposit is calculated for each binning group, represented by Stay intervals (0-10 days, 11-20 days, etc.). We intend to use this table to explore whether admission deposit could result as a key variable.

Conclusions

Our goal is therefore to study which variables are relevant to the prediction of LOS, and to build a model able to do so. We will establish which variables influence patients' length of stay, and whether these variables highlight clusters of patients which most contribute to differences in the length of stay.

Highlight the Main Expected Results

We expect to find relevant correlations between, but potentially not only, length of stay and:

- Severity of illness
- Type of admission
- Department
- Age

Among these results, we expect: at least one of the departments to outstand in relation to length of hospitalization following, or during, the treatment; a higher average length of stay for patients who are admitted for highly concerning diseases, or particularly traumatic injuries; a positive correlation between age and length of stay. Any other potential results will be analyzed and included in the paper if relevant to our research.

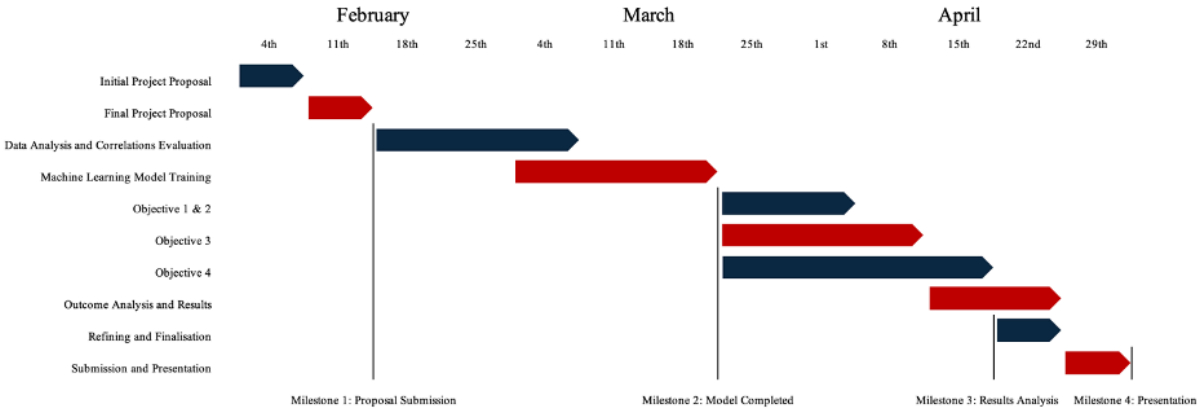
We believe these expected outcomes to be of critical importance, as being able to identify and categorize future patients with a high expected LOS in clusters of interest allows hospitals to optimize their treatment and improve the overall efficiency of their healthcare management. Knowing which department has the relatively higher length of stay, for which type of admission, or predicting it based on the admission deposit, removes substantial uncertainty and allows more accurate predictions.

Overall, this research will aim to show the causes of an expected prolonged length of stay, potentially linked to a more severe illness, urgent admissions, higher deposits, and long-term therapy departments. Subsequently, a model will be provided to anticipate, through such correlations, the LOS of future patients. Finally, recommendations will be presented as to how to incorporate such information into hospitals’ management model to optimize patient treatment and resource allocation efficiency.

Limitations

The study's discretization of the target variable "Stay" into 11 categories facilitates classification but may lead to information loss and uneven category distribution, impacting prediction accuracy. Additionally, machine learning models may struggle with generalization due to evolving medical environments, requiring frequent retraining. The model captures associations but may not fully reflect causal relationships, limiting interpretability. Moreover, as the dataset is region-specific, differences in medical policies, insurance, and hospital management may reduce its applicability across hospitals. Future improvements could include advanced modeling techniques (e.g., deep learning), enhanced feature selection with more medical variables, and improved interpretability through rule-based models to optimize medical resource management and hospital operations.

Timeline of the Project



References

Almeida G., Brito Correia F., Borges A.R., Bernardino J. (2024), Hospital Length-of-Stay Prediction Using Machine Learning Algorithms—A Literature Review, https://www.mdpi.com/2076-3417/14/22/10523

Chrusciel J., Girardon F., Roquette L., Laplanche D., Duclos A., Sanchez S. (2021), The prediction of hospital length of stay using unstructured data, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01722-4

Deschepper M., De Smedt C., Colpaert K. (2025), Literature-based approach to predict continuous hospital length of stay in adult acute care patients using admission variables: A single university center experience, https://www.sciencedirect.com/science/article/pii/S1386505624003411

Dexur (2025), Understanding & Predicting Length of Stay (LOS) using Machine Learning, https://dexur.com/a/ml-research-los/6/

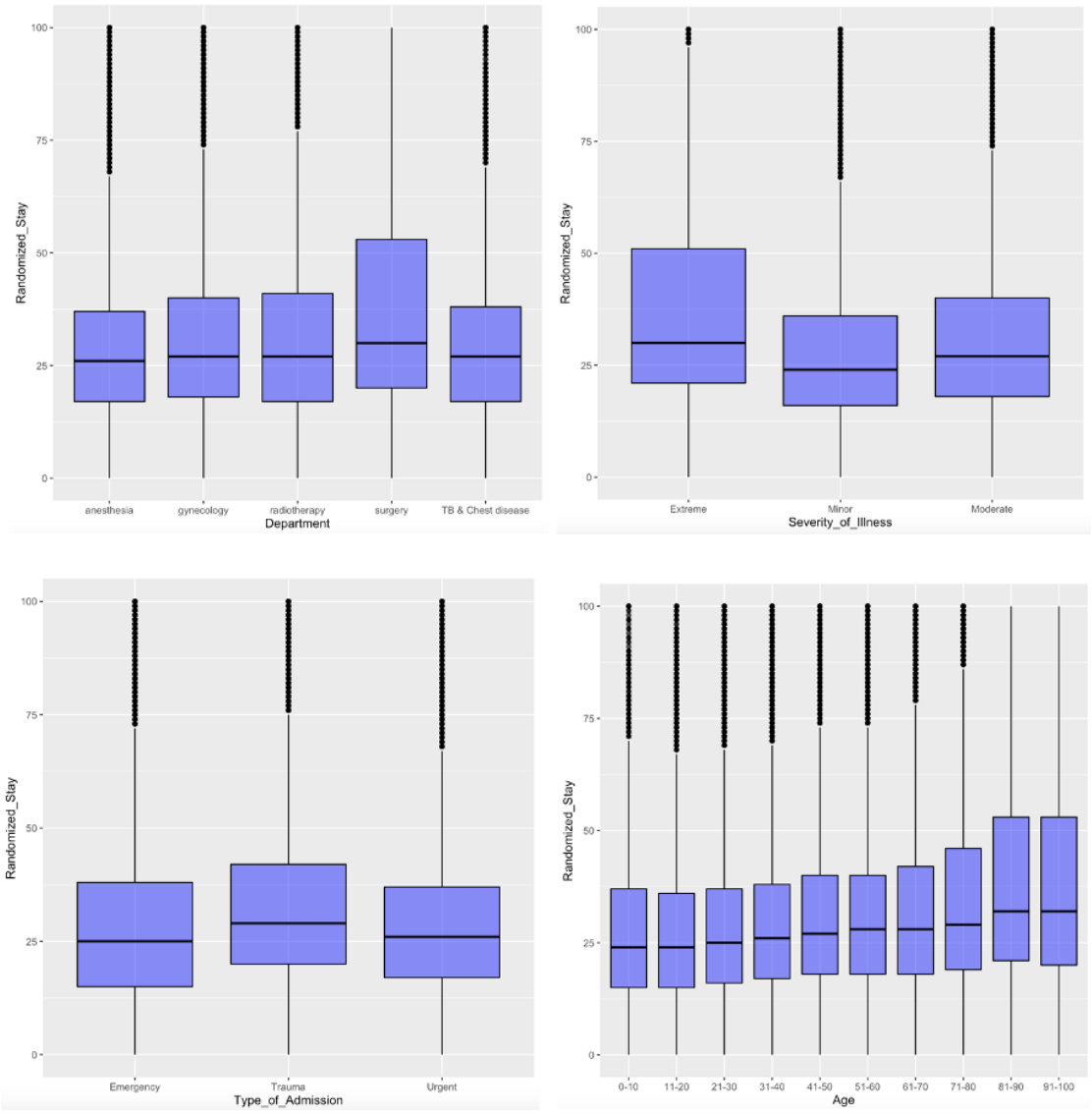
Jain R., Singh M, Ravishankar Rao A., Garg R. (2024), Predicting hospital length of stay using machine learning on a large open health dataset, https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-024-11238-y

Prabhavalkar N. (2021), AV : Healthcare Analytics II, https://www.kaggle.com/datasets/nehaprabhavalkar/av-healthcare-analytics-ii

Stone K, Zwiggelaar R, Jones P, Mac Parthaláin N., 2022, A systematic review of the prediction of hospital length of stay: Towards a unified framework, https://pmc.ncbi.nlm.nih.gov/articles/PMC9931263/#:~:text=Hospital%20length%20of%20stay%20of,costs%20and%20increase%20service%20efficiency.

(3)

Box-plots of interest variables (X-axis) in relation to length of stay (Y-axis)



(Y-axis of the four box-plots: Randomized Stay)

(Legend of Y-axis: number of days of Stay in the hospital after admission)

(X-axis of the four box plots, in order top-left, top-right, bottom-left, bottom-right: Department; Severity of Illness; Type of Admission; Age)

(Legend of X-axis: 1st box-plot - department names; 2nd box-plot - intensity of illness; 3rd box-plot: type of admission; 4th box-plot: age intervals, binned in tens)

The size of axis titles is small due to the software used (Radiant). Time constraints didn't allow us to find a solution to increase the size, but we are already working on it. The size will be increased by the next submission.

(4)

Multi-function table of {admission deposit} grouped by {Stay} intervals

		Function				
Stay	variable	n_obs	mean	min	max	sd
All	All	All	All	All	All	All
0-10	Admission_Deposit	23,604	4,615.215	1,801	9,673	1,059.092
11-20	Admission_Deposit	78,139	4,931.125	1,832	10,419	1,001.676
21-30	Admission_Deposit	87,491	5,025.310	1,807	10,729	1,016.774
31-40	Admission_Deposit	55,159	4,871.071	1,820	11,008	1,091.839
41-50	Admission_Deposit	11,743	4,888.819	1,825	11,008	1,143.829
51-60	Admission_Deposit	35,018	4,748.784	1,831	10,771	1,166.789
61-70	Admission_Deposit	2,744	4,845.449	1,809	10,254	1,301.187
71-80	Admission_Deposit	10,254	4,709.845	1,833	10,842	1,209.603
81-90	Admission_Deposit	4,838	4,590.645	1,827	10,729	1,322.041
91-100	Admission_Deposit	2,765	4,715.539	1,805	10,506	1,282.309

(n_obs: number of observations in the interval; mean: mean of the admission deposit for the interval; min: minimum; max: maximum; sd: standard deviation of the interval)