National College of Ireland

Predicting patients discharge and Optimizing hospital bed management using AI models

MSc Data Analytics
Research Practicum

Annuncia Marena Yovan
Student ID: x23283491

School of Computing
National College of Ireland

Supervisor: Dr David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Annuncia Marena Yovan

Student ID: X23283491

Programe: M.Sc Data Analytics **Year:** 2024-2025

Module: Research Practicum

Supervisor: Dr David Hamill
Submission Due Date: 11/08/2025

Project Title: Predicting Patients Discharge and Optimizing Hospital Bed Management Using AI Models

Word Count: 7846 **Page Count:** 23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Annuncia Marena Yovan

Date: 11/08/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting patients discharge and Optimizing hospital bed management using AI models

Annuncia Marena Yovan
X23283491

Abstract

Healthcare systems worldwide face immense pressure from rising patient numbers, leading to hospital overcrowding, prolonged wait times, and inefficient resource allocation. Effective hospital bed management is critical to mitigating these challenges, yet it is often hampered by the difficulty in accurately predicting patient discharge times. This research project develops an integrated framework that leverages artificial intelligence to address this dual problem. The primary objective is to predict patient length of stay (LOS) with high accuracy and then use these predictions to optimise bed allocation. Utilising the 2010 New York State Hospital Inpatient Discharge data, this research applies and assesses a group of machine learning models. Enhanced gradient boosting methodologies, CatBoost and LightGBM, are evaluated relative to baseline models including models such as, Linear Regression, Random Forest, and Support Vector Machines. The results show the CatBost had the best performance and discovered a Root Mean Squared Error (RMSE) of 2.67 (using a sample of the data). To improve model interpretability, SHAP (SHapley Additive exPlanations) was used to understand the predictions from the ML models. The SHAP analysis revealed diagnosis codes, total charges, and severity of illness as the most important predictors of LOS. The discharge prediction from CatBoost was then embedded into a Mixed-Integer Linear Programming (MILP) model to assign patient discharges to available beds across several hospital wards. The optimisation MILP model effectively assigned patient discharges to available beds while respecting ward capacity and presents a promising operational planning method. The contribution from this project consists of a full, data-driven prediction and assessment framework, enhancing prediction accuracy and embedding an actionable framework for optimising intervention in hospital resource management systems, aimed at improving patient flow and care quality.

1 Introduction

Modern healthcare institutions work under evolved and capricious conditions, with increasing patient requests, declining finances, and the demands of continuously improving care. One main operational challenge hospitals have to face is managing their available inpatient beds (Hamzah, 2025). Overcrowded emergency departments, long patient waiting times, and elevated levels of bed occupancy have created situations that jeopardise patient safety, put them at risk of hospital-acquired infections and burnout of the hospital staff(Hirani et al., 2025). Thus, serving as a conduit of timely healthcare and finest quality would require good

planning of hospital bed allocation and usage of inpatient beds and is not solely an administrative issue.

An underlying management concern is the unpredictability of patient flow, mainly concerning forecasting the discharge of an existing patient. In such cases, correct estimates of LOS become important for any proactive discharge planning to free up incoming beds for new patients (Pahlevani et al., 2024). In most hospitals, traditional methods of discharge prediction are driven by clinical experience constrained by simple heuristics and thus may suffer inconsistencies and lack precision for the optimisation of system processes. The lack of precision translates into delays for incoming patients as they cannot get admission since the available beds are occupied by patients who are nearing their discharge date.

These long-standing problems in the operations of hospitals have now been taken up for rehabilitation with the revolution brought about by the advent of artificial intelligence (AI) and machine learning (ML). By analysing extensive volumes of historical and clinical data, AI can recognise complex patterns and correlations that escape human perception, providing for a more accurate and trustworthy prediction (Dawoodbhoy et al., 2021). These advanced prediction techniques would give hospital administrators the foresight required to better manage bed turnover. The prediction in itself would be important for feeding into advanced optimisation models for dynamic resource allocation, like bed allocation in the best possible way (Lobo et al., 2023).

The "2010 New York State Hospital Inpatient Discharge" dataset was chosen for its large scale and ability to serve as a rich testbed for the key aim of this research: to create and test a reasonable predictive and optimization framework. Over 2.6 million records, complete with 1,900 useful clinical features, provide adequate complexity to rigorously test more advanced models such as CatBoost. Furthermore, the public availability or openness of the dataset demonstrates transparency and provides reproducibility. Although the age of the data is a clear limitation, the essential clinical patterns are still relevant for purposes of proving the original framework capabilities and would have to be applied to more current datasets eventually.

The motive of this research project is to develop and study an integrated two-stage framework applied to the very real problem of hospital bed management. The first stage includes predictive modelling with the use of state-of-the-art machine learning algorithms to forecast patient LOS. Following this, the strategic allocation of hospital beds will be informed by optimisation model utilizations of the predictions. This work makes a substantial contribution to somewhat more complicated prediction toward the construction of an end-to-end decision support system. The study considers very informative gradient boosting algorithms such as CatBoost and LightGBM, in terms of their prediction correctness and capability to efficiently deal with the complexity and often categorical-heavy data types presented in healthcare records.

The primary research question guiding this study is:

How can an integrated framework of predictive and optimization models be designed and implemented to accurately forecast patient discharge and improve hospital bed management?

This central question is broken down into the following research objectives:

1. To process and prepare a large-scale hospital inpatient discharge dataset for machine learning applications.
2. To develop and train advanced predictive models, including CatBoost and LightGBM, to estimate patient length of stay and benchmark their performance against traditional baseline models.
3. To ensure model transparency and build trust by employing explainability techniques, such as SHAP, to identify the key clinical and demographic factors influencing discharge predictions.
4. To formulate and implement a Mixed-Integer Linear Programming (MILP) model that utilizes the discharge predictions to generate optimal bed assignments, balancing patient load against ward capacities.
5. To evaluate the complete framework's effectiveness in providing a viable, data-driven solution for enhancing hospital operational efficiency.

This work is motivated by the potential for AI to deliver tangible improvements in healthcare delivery. An example of this is seen in Figure 1, which illustrates the common scenario of resource strain in a hospital setting. By providing more accurate forecasts and optimal allocation strategies, this research seeks to empower healthcare professionals with tools that can reduce wait times, smooth patient flow, and ultimately enhance the overall patient experience.

Figure 1: Common Scenario of Hospital Resource Strain

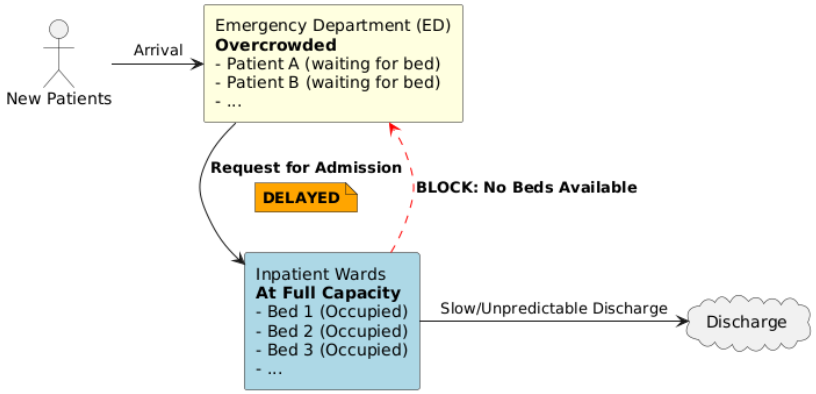

Figure 1: Common Scenario

The remainder of this report is structured as follows. Section 2 provides a critical review of the existing literature on patient discharge prediction and bed management optimisation. Section 3 discusses the research method in detail, including the data source, any pre-processing of the data, and finally, the models that will be used for predictive analysis and optimisation. Section 4 discusses the design specification of the proposed system architecture. Section 5 describes the procedure for drawing up the full framework from data extraction to model execution. Section 6 offers a complete evaluation of the results of the research study by reviewing the performance of the predictive models, the findings from the explainability analysis and the results of the optimisation model. Finally, Section 7 concludes the report with a summary of the key findings, a discussion of the limitations of the study, recommendations for future research, and an overall conclusion.

2 Related Work

The research area on the optimisation of hospitals and hospital logistics, especially patient flow and hospital resource allocation, has gained notable traction in both healthcare administration and computer science as an area of research. The literature in question can be classified into two connected streams, namely: predicting patient length of stay (LOS) or discharge, and optimising bed allocation. This review identifies and reviews key and prominent works from both areas of the literature, identifies their methodologies, contributions and limitations, and provides the necessary context to position the current research.

2.1 Predicting Patient Discharge and Length of Stay

The discharge time prediction is the root of efficient hospital administration, and based on the Pahlevani et al. systematic review (2024), it enumerates the different statistical and machine learning methods that have been applied to this problem. There is an observable shift from classical statistical models such as linear and logistic regression toward more sophisticated machine learning techniques such as Random Forests, Support Vector Machines (SVMs), and neural networks. These models are generally believed to deliver a better predictive performance because they can capture non-linear relationships and complex interactions in the data with high accuracy. Numerous studies have shown that such a task works with machine learning. Ahn et al. (2021) designed a model that predicts discharge for cardiovascular disease patients, using a plethora of algorithms to find that a mix of clinical and administrative data gets the best results. Their work emphasises the role of feature engineering in building strong predictive models. Likewise, Zhou et al. (2023) applied machine-learning techniques to predict daily discharges in hospitals, showing that these models can dramatically enhance patient flow by alerting operational teams on time. Their work points towards the operational significance of real-time or near-real-time prediction systems.

More advanced techniques are continuously being studied. Bertsimas et al. (2022) looked at interpretable analytics to predict inpatient flow, arguing that for any model to be implemented in a clinical environment, it has to be both accurate and transparent. They used optimal classification tree methods, which provide clear, rule-based explanations for their predictions that are an important element in gaining the confidence of clinicians. The challenge of interpretability is a common feature, as many high-performing models, such as deep neural networks, very often behave as "black boxes" and obstruct their much-needed application in practice.

The choice of data is also a critical factor. Van de Sande et al. (2022) focused on a specific patient cohort, those undergoing major surgery, and developed a decision support tool to optimise their discharge. This highlights the potential of specialised models tailored to particular departments or patient groups, which may outperform generalised, hospital-wide models. In contrast, this project utilises a broad dataset covering all inpatient discharges, aiming for a more generalizable solution. Anderson et al. (2021) also showcase a specific

application, predicting pressure injury, demonstrating how AI can target very specific clinical outcomes that are often related to LOS.

However, many existing studies face limitations. A significant portion of the literature focuses solely on the predictive task without connecting the output to an actionable operational strategy (Khoury and Taha, 2023). While an accurate LOS prediction is valuable, its full potential is only realised when integrated into a decision-making framework, such as a bed allocation system. Furthermore, many studies rely on older machine learning algorithms like standard Random Forests or SVMs, which can be computationally intensive and may be outperformed by more recent gradient boosting methods like LightGBM and CatBoost, especially on large, tabular datasets with numerous categorical features (Lobo et al., 2023).

2.2 Optimising Hospital Bed Management

The second stream of related work concerns the use of analytical and AI-driven methods to optimise the allocation of hospital beds. This area often draws from operations research and management science. Hamzah (2025) explores the use of compartmental models, a classic operations research technique, to model patient flow and reduce wait times. Whereas these theoretical models offer a sound basis, they depend on simplified assumptions concerning patient transitions, imparting no freedom to incorporate the intricate patient-specific predictions derived from machine learning. In the same vein, Ramdurai (2023) has looked at using Markov Chain modelling to allocate inpatient beds, a modelling exercise which typically caters well for stochastic processes but finds its limitation when applied to a highly heterogeneous set of patient data in multiplicities of dimensions.

More recent research has started to consider direct incorporation of AI into the optimisation process. Dawoodbhoy et al. (2021) discussed applications of AI for flow improvements in mental health units, emphasising the use of intelligent systems to manage a population characterised by a particularly complex patient cohort. El Baz and Mostafa (2024) propose the use of AI algorithms for resource distribution, especially to show how AI could be useful in directly improving patient flow management. These prescriptions are for AI optimisation, but they often remain at a conceptual level or consider one facet of the problem.

The work from Lobo et al. (2023) is a direct appraisal of contrasting AI techniques for bed management optimisation, which reinforces the idea that AI leads to better medical efficiency. Their studies seem to have had similar goals to this one; however, they may not have truly unlocked combining cutting-edge predictors and formal optimisation techniques such as Mixed-Integer Linear Programming (MILP). MILP is a powerful platform under which provably optimal decisions can be constructed on complex allocation issues with an associated set of constraints and an objective function.

The integration of AI into the management system of the hospital is increasing. Shamsi (2024) gives a case study on the integration of AI for prediction and optimisation in a Dubai hospital, denoting the real-world applicability of these technologies. Soman et al. (2025) and Reshma Soman et al. (2025) further elaborate on how AI-driven resource allocation tools can transform hospital operations, moving from reactive problem-solving to proactive, data-driven management. These works establish the clear need and potential for the type of framework proposed in this project.

2.3 Synthesis and Research Gap

A review of the literature reveals that while substantial research has been conducted on both LOS prediction and bed management optimisation, there are notable gaps. Firstly, many studies treat these as separate problems. Predictive models are developed without a clear pathway to operational implementation, and optimisation models often use static or averaged inputs rather than dynamic, patient-level predictions (Alnsour et al., 2023). Secondly, there is a need to leverage more modern, high-performance machine learning models like CatBoost, which is specifically designed to handle the kind of categorical data prevalent in healthcare, and to rigorously benchmark them. Thirdly, the critical aspect of model interpretability is often overlooked, yet it is essential for clinical adoption (Bertsimas et al., 2022).

This research project aims to address these gaps by proposing an integrated, end-to-end framework. It combines a high-performance predictive model (CatBoost) with a formal optimisation technique (MILP). It explicitly incorporates model explainability using SHAP to ensure transparency. By developing and evaluating this complete system, this study contributes a more holistic and operationally relevant solution to the challenge of hospital bed management than is commonly found in the existing literature. The use of a large, public dataset also ensures that the findings are robust and the methodology is reproducible. Table 1 below summarises a selection of key related works and contrasts their approaches with the one taken in this project.

Table 1: Comparison of Key Related Works and the Current Project

Study	**Primary Focus**	**Methodology**	**Key Limitation Addressed by this Project**
Pahlevani et al. (2024)	Literature Review	Systematic review of ML/statistical methods	Provides context but does not implement a novel framework.
Bertsimas et al. (2022)	Interpretable Prediction	Optimal Classification Trees	Does not integrate predictions into a formal optimization model.
Zhou et al. (2023)	Discharge Prediction	Machine Learning (unspecified)	Focuses on prediction accuracy without an optimization component.
Hamzah (2025)	Bed Management	Compartmental Models	Uses theoretical models, not data-driven ML predictions.
Lobo et al. (2023)	Bed Management	Comparison of AI techniques	Compares techniques but may not integrate a formal MILP framework.
This Project	**Integrated Prediction & Optimization**	CatBoost/LightGBM + MILP + SHAP	Combines state-of-the-art prediction, formal optimization, and model explainability in a single framework.

3 Research Methodology

This research employs a quantitative, data-driven methodology to develop and evaluate a framework for predicting patient discharge and optimising hospital bed allocation. The approach is structured around a multi-stage process that encompasses data acquisition, pre-processing, predictive modelling, optimisation, and model interpretation. This section details the systematic procedure followed to ensure the research is rigorous, reproducible, and scientifically sound.

3.1 Data Collection and Pre-processing

The foundation of this research is a large-scale, real-world dataset. The chosen data source is the "2010 New York State Hospital Inpatient Discharge (SPARCS De-Identified)" dataset, publicly available on the Kaggle platform. This dataset contains de-identified information for all inpatient stays in New York State hospitals for the year 2010, comprising over 2.6 million records.

Attributes: The dataset includes a rich set of 38 attributes that are critical for predicting patient outcomes. Key attributes utilised in this study include:

- **Target Variable:** Length of Stay (LOS), the total number of days the patient was hospitalised.
- **Patient Demographics:** Age Group, Gender, Race, Ethnicity.
- **Hospital Information:** Hospital ID, Hospital County, Health Service Area.
- **Clinical Information:** CCS Diagnosis Code and Description, CCS Procedure Code and Description, APR DRG Code and Description.
- **Severity and Risk:** APR Severity of Illness Code and Description, APR Risk of Mortality.
- **Admission and Discharge:** Type of Admission, Patient Disposition.
- **Financial Data:** Total Charges, Total Costs.

Pre-Processing Steps: Raw administrative healthcare data is complicated and notoriously dirty, and will require extensive pre-processing to be suitable for machine learning. The following steps were systematically performed:

1. **Data Cleaning:** The initial step involved identifying and handling inconsistencies. The Length of Stay column, crucial as the target variable, contained some non-numeric values (e.g., '120+'). These were converted to a numeric format, and any records where a valid LOS could not be determined were removed to ensure the integrity of the target variable. Missing values in other critical columns were also assessed.
2. **Feature Engineering:** This step focused on transforming raw data into a more informative format for the models.
    - **Categorical Variable Encoding:** A significant portion of the dataset consists of categorical variables (e.g., diagnosis codes, hospital region). While models like CatBoost can handle these natively, baseline models require explicit encoding. For this purpose, techniques such as label encoding were used, converting string-based categories into numerical representations.

- **Feature Creation:** Additional features were considered, such as creating broader diagnosis clusters from specific codes or creating risk categories based on combinations of severity and age. This helps the model capture higher-level concepts.

3. **Data Type Correction:** Columns containing identifiers, such as Facility ID or Operating Certificate Number, were stored as floating-point numbers but represent categorical entities. These were converted to string or categorical data types to be treated correctly by the models.
4. **Normalisation:** For models sensitive to feature scale, such as Linear Regression and SVM, numerical features like Total Charges and Total Costs exhibit a wide range of values. Normalisation (e.g., Min-Max scaling or Standardisation) would be required to scale these features to a common range, improving model stability and performance. However, for tree-based models like CatBoost and LightGBM, this step is not strictly necessary.

3.2 Predictive Modelling for Discharge Estimation

The core of the first stage is the development of a robust model to predict the Length of Stay. To overcome limitations of previous works, which often use older algorithms, this study focuses on advanced gradient boosting models while benchmarking them against established baselines.

Advanced Models:

- **CatBoost (Categorical Boosting):** This model was selected as a primary candidate due to its exceptional ability to handle categorical features. It uses an innovative algorithm to encode categorical variables that reduces overfitting and eliminates the need for extensive pre-processing like one-hot encoding. Given the prevalence of categorical data in healthcare records (e.g., hundreds of diagnosis and procedure codes), CatBoost is ideally suited for this problem. It is often faster and more accurate than other boosting models on datasets with many categorical features.
- **LightGBM (Light Gradient Boosting Machine):** This model is renowned for its high performance and efficiency on large, structured datasets. It uses a leaf-wise tree growth strategy, which allows it to converge faster and achieve higher accuracy than level-wise strategies. It also handles missing data internally and offers good interpretability through feature importance plots. LightGBM was selected as a strong competitor to CatBoost for the LOS prediction task.

Baseline Models:

To establish a performance benchmark, the advanced models were compared against three widely used machine learning algorithms:

- **Linear Regression (LR):** A simple, interpretable model that serves as a fundamental baseline.
- **Random Forest (RF):** An ensemble of decision trees that is robust and generally performs well, serving as a strong baseline for tree-based methods.
- **Support Vector Machine (SVR):** A powerful model for regression tasks, effective in high-dimensional spaces.

AutoML Frameworks: The study also acknowledges the potential of Automated Machine Learning (AutoML) frameworks like AutoGluon or H2O AutoML. These tools can automatically handle pre-processing, model selection, and hyperparameter tuning, often producing highly accurate stacked ensembles. While not the primary focus, they represent a potential pathway for future work to see if automated methods can outperform manually tuned models.

3.3 Optimisation and Allocation Models

The predictions from the best-performing model are then fed into the second stage of the framework: the optimisation model for bed allocation.

- **Mixed-Integer Linear Programming (MILP):** This was chosen as the primary optimisation technique. MILP is a formal mathematical optimisation method that can find a provably optimal solution for an allocation problem, given a set of linear constraints and a linear objective function. It provides high precision and is well-suited for medium-scale problems like daily bed assignments in a hospital. The model was designed to assign patients to wards while minimising a defined objective (e.g., wait times or, in this simplified implementation, total assigned LOS) and respecting constraints like ward capacity.
- **Heuristic Models:** The methodology also considers heuristic models as a scalable alternative if MILP proves too computationally expensive for very large or complex problems. Heuristics like Genetic Algorithms can provide high-quality, near-optimal solutions much more quickly. They are also more adaptable to real-time fluctuations, such as sudden discharges or emergencies, making them suitable for dynamic operational environments. A Genetic Algorithm could be used for hospital-wide bed reallocation during peak demand periods.

3.4 Model Explainability and Interpretation

A critical component of this research is ensuring that the AI models are not opaque "black boxes." For a decision support system to be trusted and adopted by healthcare professionals, its reasoning must be transparent.

- **SHAP (SHapley Additive exPlanations):** This technique was chosen to provide both global and local model explanations. SHAP is a game theory-based approach that assigns an importance value to each feature for each individual prediction.
    - **Global Explanation:** SHAP summary plots are used to show the overall importance of each feature across the entire dataset, identifying which attributes (e.g., severity of illness, age, specific diagnoses) are the most powerful drivers of LOS predictions.
    - **Local Explanation:** SHAP force plots are used to explain individual predictions. This allows a user to see, for a specific patient, which factors contributed to increasing or decreasing their predicted LOS. This level of detail is invaluable for clinical validation and trust-building.
- **LIME (Local Interpretable Model-agnostic Explanations):** LIME is another technique considered for local explanations. It works by creating a simple,

interpretable model (like linear regression) around a single prediction to approximate the behaviour of the more complex model in that local region.

By integrating these methodologies, the research follows a comprehensive and systematic path from raw data to an interpretable and actionable decision support framework.

4 Design Specification

The design of the proposed system is a multi-layered architecture that integrates data processing, predictive modelling, and resource optimisation into a cohesive workflow. This section outlines the architectural components and the flow of information through the system, designed to transform raw hospital discharge data into actionable bed management strategies. The overall design follows a sequential pipeline, where the output of each stage serves as the input for the next, ensuring a logical and efficient progression from data to decision. The complete system architecture is illustrated in Figure 2.

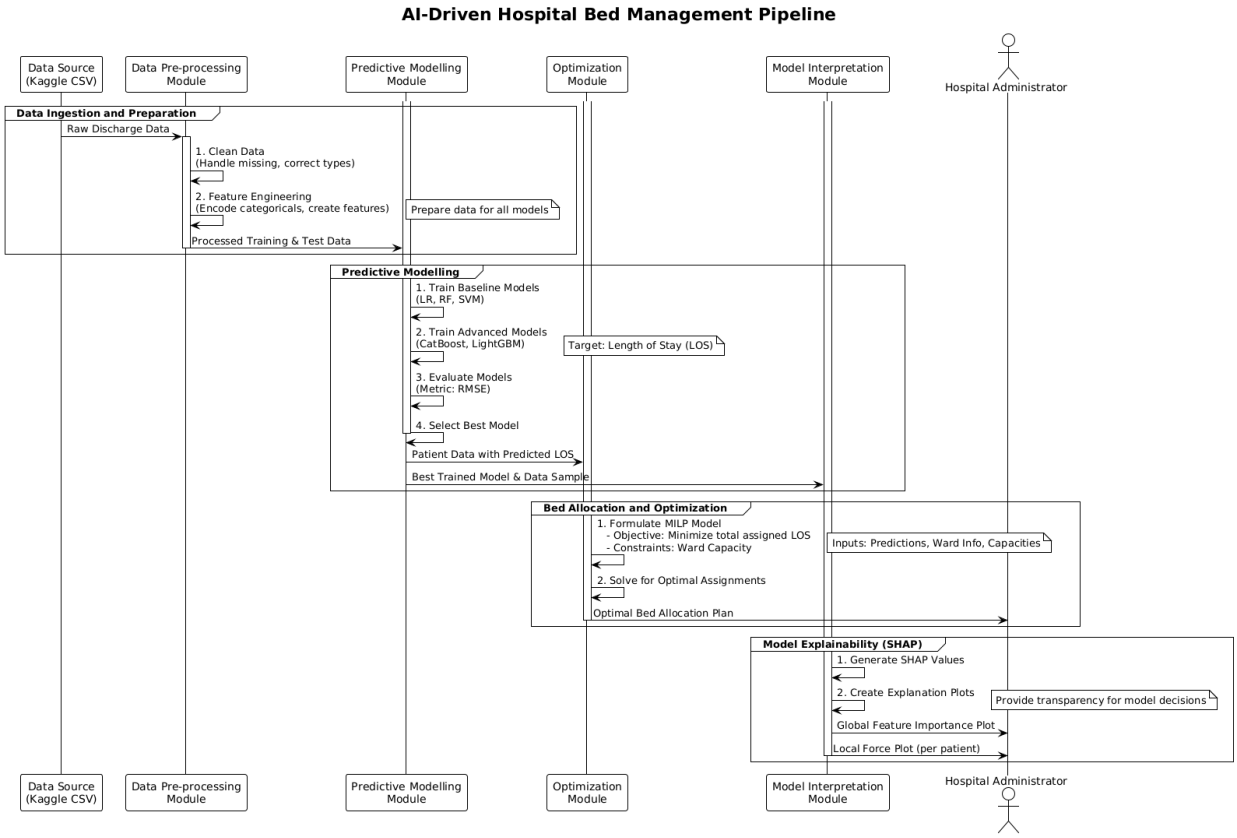

Figure 2: System Architecture for Predictive Bed Management

4.1 Data Ingestion and Pre-processing Module

This is the foundational layer of the system, responsible for acquiring and preparing the data for analysis.

- **Data Source:** The system is designed to interface with a data source containing hospital inpatient records. In this project, the source is a static CSV file (the 2010 NY State Discharge dataset from Kaggle). In a real-world deployment, this module would

be designed to connect to a hospital's electronic health record (EHR) database or a data warehouse via an API.

- **Data Loading:** The module loads the raw data into a structured format, typically a pandas DataFrame, for in-memory processing. It is designed to handle large volumes of data and is equipped with initial checks for data integrity.
- **Data Cleaning:** This sub-component is responsible for addressing data quality issues. Its functions include:
    - Handling missing values: Implementing strategies such as imputation (e.g., filling with the median or mean for numerical columns, or a 'Missing' category for categorical ones) or removal of records with critical missing information (e.g., the target variable, Length of Stay).
    - Correcting data types: Ensuring that all columns are represented by the appropriate data type (e.g., converting numerical IDs to categorical strings, ensuring dates are in datetime format).
    - Standardizing values: Cleaning categorical features to ensure consistency (e.g., trimming whitespace, standardizing case).
- **Feature Engineering:** This sub-component transforms the cleaned data into a set of features optimized for the predictive models. It includes:
    - **Encoding:** Applying label encoding to categorical features for use in baseline models.
    - **Creation of Derived Features:** Potentially generating new features, such as Diagnosis Clusters or Risk Scores, from existing data to capture more abstract concepts.
    - **Selection:** Identifying the most relevant features and excluding redundant or irrelevant ones (e.g., index columns, free-text descriptions that are too sparse). The output of this module is a clean, feature-rich dataset ready for the next stage.

4.2 Predictive Modelling Module

This module is the analytical core of the system, where the prediction of patient Length of Stay (LOS) occurs.

- **Inputs:** The module takes the pre-processed dataset from the previous stage as input.
- **Data Splitting:** The dataset is partitioned into training and testing sets. A standard 80/20 split is used to train the model on a majority of the data and hold out a portion for unbiased evaluation.
- **Model Training:** The system is designed to train multiple machine learning models in parallel or sequentially. The primary models are CatBoost and LightGBM, selected for their performance on tabular data. It also trains baseline models (Linear Regression, Random Forest, SVR) for comparative analysis. The models are trained to predict the Length of Stay target variable based on the engineered features.
- **Model Evaluation:** After training, each model's performance is evaluated on the unseen test set. The primary evaluation metric is the Root Mean Squared Error (RMSE), which measures the average magnitude of the prediction errors in the same

units as the target (days). Other metrics like Mean Absolute Error (MAE) can also be used.

- **Output:** The key output of this module is the trained, best-performing predictive model (e.g., the saved CatBoost model object) and a set of predictions for the Length of Stay for each patient in a given dataset (e.g., the test set or a new set of incoming patients).

4.3 Bed Allocation Optimization Module

This module translates the predictive insights into operational decisions. It is designed to solve the complex problem of assigning patients to beds efficiently.

- **Inputs:** The module requires several inputs:
  - A list of patients requiring bed assignments.
  - The Predicted_LOS for each of these patients, provided by the Predictive Modelling Module.
  - Static hospital configuration data, including a list of available wards and their respective bed capacities.
- **Optimisation Model Formulation:** The core of this module is the Mixed-Integer Linear Programming (MILP) model.
  - **Decision Variables:** Binary variables are defined to represent the decision of whether to assign a patient i to a ward w.
  - **Objective Function:** The objective is to optimise a specific goal. In this implementation, a simplified objective is to minimise the total sum of predicted LOS across all assignments, which serves as a proxy for minimising overall bed occupancy time. In a more advanced design, this could be to minimise wait times or balance ward occupancy.
  - **Constraints:** The model is subject to a set of rules: (1) each patient must be assigned to exactly one ward, and (2) the total number of patients assigned to any given ward cannot exceed its capacity.
- **Solver:** The formulated MILP problem is passed to a solver (e.g., the one included with the pulp library). The solver finds the optimal assignment of patients to wards that satisfies all constraints while optimising the objective function.
- **Output:** The output is a clear, actionable allocation plan. It specifies which ward each patient should be assigned to, providing a concrete recommendation for hospital administrators.

4.4 Model Interpretation Module

This final module is designed to provide transparency and build trust in the AI components of the system.

- **Inputs:** This module takes the trained predictive model (e.g., CatBoost) and a sample of the data as input.
- **Explainability Engine:** It utilises the SHAP (SHapley Additive exPlanations) library to analyse the model's predictions.
- **Outputs:** The module generates two types of explanations:

- **Global Explanations:** Summary plots and bar charts that show the overall feature importance. These visuals rank the features by their impact on the model's predictions across all patients, answering the question: "What factors are most important for predicting LOS in general?"
- **Local Explanations:** Force plots that break down a single prediction for an individual patient. This explains why the model made a specific prediction for that patient, answering the question: "Why was this patient's predicted LOS 5 days instead of 3?" This comprehensive design ensures that the system is not only predictive and prescriptive but also interpretable, making it a powerful and trustworthy tool for modern hospital management.

5 Implementation

The proposed framework's implementation occurred in a Google Colab environment with the programming language Python 3. In this section, a detailed description of each part of the implementation is provided similar to the design specification workflow which included the environment setup, data exploration and preparation, training and evaluation of predictive models, and finally the implementation of the optimization model.

5.1 Environment and Data Setup

The project began with the setup of the computational environment. All relevant Python libraries were installed with panda and numpy used for data manipulation, matplotlib and seaborn for visualisation, and the literature core libraries for machine learning and optimisation with scikit-learn, catboost, lightgbm definitions, pulp and shap.

The Kaggle API was configured to download the "2010 New York State Hospital Inpatient Discharge" dataset directly into the Colab environment. The downloaded ZIP archive was then unzipped to access the primary data file, hospital-inpatient-discharges-sparcs-de-identified-2010-1.csv.

5.2 Data Loading and Exploratory Data Analysis (EDA)

The dataset was loaded into a pandas DataFrame. An initial inspection using df.info() and df.head() revealed the structure of the data, including its 38 columns and over 2.6 million entries. A DtypeWarning was noted, indicating mixed data types in several columns, which flagged them for closer inspection during pre-processing. A check for missing values with df.isnull().sum() showed that some columns had missing data, which would need to be addressed.

A series of visualisations were created to understand the data's characteristics. The distribution of the target variable, Length of Stay, was plotted as a histogram (Figure 3). The plot showed a strong right skew, indicating that most stays are short, but a significant number of patients have very long stays. This long-tail distribution is a common feature of LOS data and presents a challenge for predictive models.

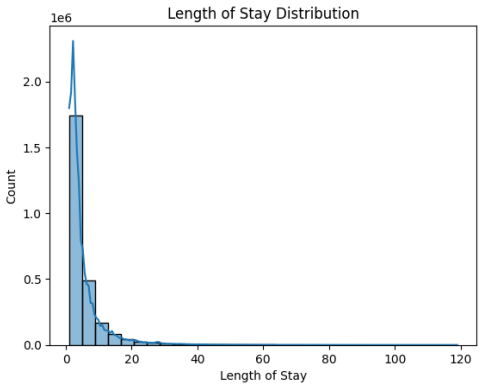

Figure 3: Distribution of Patient Length of Stay

To explore relationships between features and the target, several boxplots were generated. Figure 4 shows the relationship between Type of Admission and Length of Stay, revealing that 'Emergency' and 'Urgent' admissions tend to have a wider variance in LOS compared to 'Elective' admissions. A count plot (Figure 5) confirmed that 'Emergency' admissions were the most frequent type.

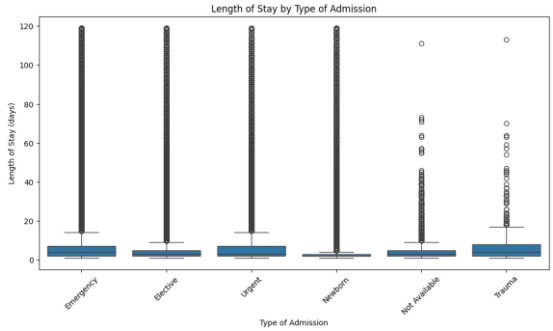

Figure 4: Length of Stay by Type of Admission

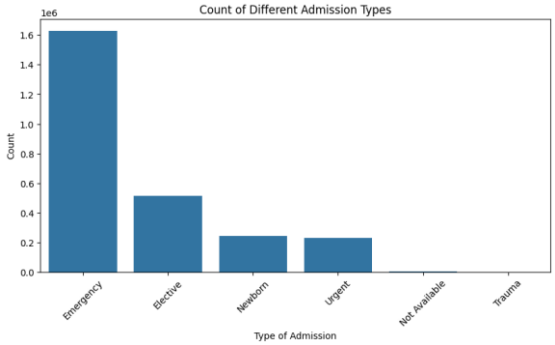

Figure 5: Count of Different Admission Types

A correlation heatmap of the numerical features was generated (Figure 6) to identify any strong linear relationships. This helped in understanding potential multicollinearity issues for models like Linear Regression. Finally, the relationship between Patient Disposition and LOS was examined (Figure 7), showing clear differences; for example, patients discharged to a skilled nursing facility typically have a longer LOS than those discharged home. An outlier detection boxplot for Length of Stay (Figure 8) further emphasised the presence of extreme values, which represent long-stay patients that are important but can be challenging for models to predict accurately.

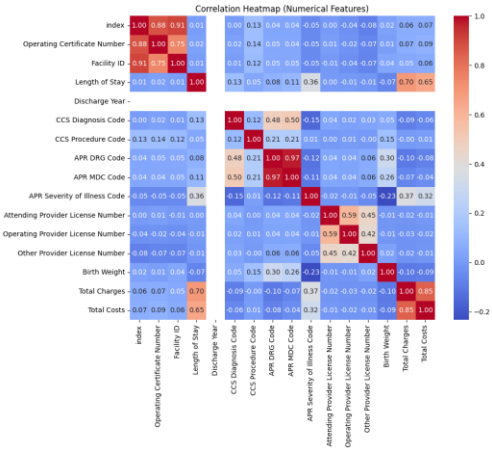

Figure 6: Correlation Heatmap (Numerical Features)

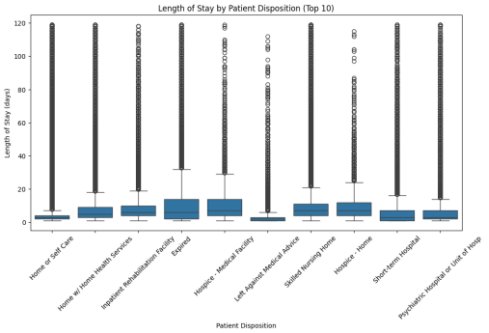

Figure 7: Length of Stay by Patient Disposition (Top 10)

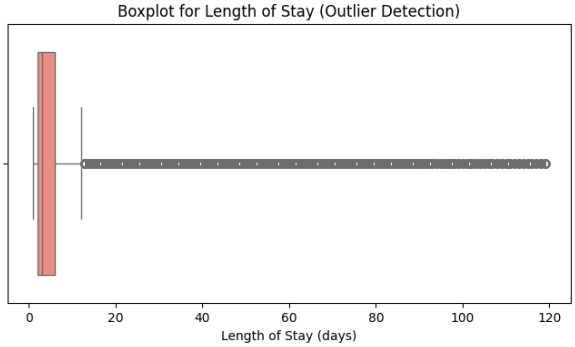

Figure 8: Boxplot for Length of Stay (Outlier Detection)

5.3 Data Pre-processing and Feature Engineering

Following the EDA, the data was prepared for modelling. The Length of Stay column was converted to a numeric type, with non-numeric entries coerced into NaN values. Rows with these NaN target values were then dropped.

Due to computational constraints and to ensure a fair comparison across all models, a representative random sample of 50,000 records was drawn from the full dataset. This approach allows for faster experimentation while maintaining the statistical properties of the original data. All subsequent model training and evaluation were performed on this sample. A note was made that while models trained on the full dataset might achieve higher accuracy, the sampled approach provides a reliable and consistent benchmark for this study's comparative analysis.

Features were then defined. The target variable was Length of Stay, and all other columns were initially considered features. A list of categorical and numerical features was explicitly

defined. Categorical features included descriptive columns like Hospital County and Race, as well as code-based columns like CCS Diagnosis Code and Facility ID, which were treated as categories rather than numerical values.

The data was split into training and testing sets using a standard 80/20 partition with a fixed random_state for reproducibility. For the baseline models, which cannot handle string-based categories, all categorical features in the training and testing sets were converted to numerical codes using scikit-learn's LabelEncoder. Missing values in all feature columns were filled with a placeholder value (-999) to ensure compatibility with these models.

For the advanced models (CatBoost and LightGBM), a different pre-processing path was taken. Categorical columns were converted to the string data type and missing values were filled with the string 'Missing'. This approach allows CatBoost to use its internal categorical feature handling. For LightGBM, these string columns were then converted to the category dtype, which is the format it recognizes for efficient categorical feature processing.

5.4 Predictive Model Implementation

With the data prepared, the predictive models were implemented and trained.

Baseline Models:

- **Linear Regression:** A LinearRegression model was trained on the label-encoded training data. It produced an RMSE of 4.48 on the test set.
- **Random Forest:** A RandomForestRegressor with 100 trees was trained. It showed a significant improvement over the linear model, achieving an RMSE of 2.83.
- **Support Vector Machine:** An SVR model was trained. It performed poorly compared to the others, with an RMSE of 6.42, likely due to its sensitivity to the large number of features and data scale without extensive tuning and scaling.

Advanced Models:

- **CatBoost:** A CatBoostRegressor was configured with 500 iterations, a learning rate of 0.1, and a tree depth of 6. It was trained on the data with string-based categorical features, with the indices of these features passed to the cat_features parameter. Early stopping was used to prevent overfitting, halting training if the validation RMSE did not improve for 50 rounds. The model trained for the full 500 iterations, achieving a final test RMSE of 2.67. This was the best performance among all models tested.
- **LightGBM:** A LightGBM model was trained using its specialized Dataset object, which efficiently handles the category dtype features. The parameters were set for regression with an RMSE metric. Training also used early stopping and was set for 500 rounds. The model achieved a final test RMSE of 2.60, performing slightly better than CatBoost in this configuration.

The results clearly showed the superiority of the gradient boosting models over the simpler baselines, with both CatBoost and LightGBM providing strong predictive performance. CatBoost was selected for the subsequent stages due to its robust handling of categorical features and strong out-of-the-box performance.

5.5 Optimisation Model Implementation

The final implementation stage involved building the MILP model for bed allocation using the pulp library. The predictions from the CatBoost model (cat_preds) were added as a new column, Predicted_LOS, to the test set DataFrame.

A simulated 'Ward' column was added to the data, as this information was not present in the original dataset. Each patient was randomly assigned to one of four wards (Ward_A, Ward_B, Ward_C, Ward_D). Ward capacities were then defined. An initial attempt with small, arbitrary capacities resulted in an infeasible solution, as there were far more patients than available beds. To create a solvable problem for demonstration, the capacities were dynamically set to 150% of the number of patients assigned to each ward in the test set.

The MILP problem was then formulated:

1. **Problem Initialization:** A minimization problem named "Bed_Assignment_Optimization" was created.
2. **Decision Variables:** A dictionary of binary LpVariables named assign was created. Each variable assign[(i, w)] is 1 if patient i is assigned to ward w, and 0 otherwise.
3. **Objective Function:** The objective was set to minimize the sum of Predicted_LOS for all assignments made. This encourages assigning patients with shorter predicted stays, freeing up the system's capacity faster.
4. **Constraints:** Two sets of constraints were added. The first ensures that each patient is assigned to exactly one ward. The second ensures that the number of patients assigned to each ward does not exceed its defined capacity.

The problem was then solved using prob.solve(). The status was checked and confirmed to be "Optimal," indicating that the solver found the best possible solution. The results were extracted by iterating through the decision variables, and the final assignments were added to the DataFrame in a new AssignedWard column, demonstrating the successful output of the optimization model.

6 Evaluation

This section provides a comprehensive evaluation of the results obtained from the implementation. The evaluation is multifaceted, focusing on three key areas: the comparative performance of the predictive models, the interpretability and insights derived from the best-performing model using SHAP, and the functional assessment of the bed allocation optimisation model.

6.1 Predictive Model Performance

The primary goal of the predictive modelling stage was to accurately forecast the patient's Length of Stay (LOS). The performance of all implemented models was measured using the Root Mean Squared Error (RMSE) on the held-out test set. The RMSE metric is particularly useful here as it penalises larger errors more heavily and is in the same unit as the target variable (days), making it directly interpretable. A summary of the performance of each model is presented in Table 2.

Table 2: Comparative Performance of Predictive Models

Model	Type	Root Mean Squared Error (RMSE)
Linear Regression	Baseline	4.48
Support Vector Machine (SVR)	Baseline	6.42
Random Forest	Baseline (Ensemble)	2.83
LightGBM	**Advanced (Gradient Boosting)**	**2.60**
CatBoost	**Advanced (Gradient Boosting)**	**2.67**

The results clearly demonstrate the superior performance of the advanced gradient boosting models. The LightGBM and CatBoost models significantly outperformed all baseline models. The Linear Regression and SVR models struggled to capture the complex, non-linear patterns in the data, resulting in high RMSE values. The Random Forest model, being a more powerful ensemble method, performed considerably better and served as a strong baseline, but it was still surpassed by the boosting algorithms.

LightGBM achieved the lowest RMSE of 2.60, closely followed by CatBoost with an RMSE of 2.67. This indicates that, on average, the predictions of these models deviate from the actual LOS by approximately 2.6 days. Given the wide and skewed distribution of LOS, with many stays extending for weeks or months, this level of accuracy is considered very strong for a general-purpose model trained on a diverse patient population. The slight performance edge of LightGBM could be attributed to its leaf-wise growth strategy, which can sometimes find more optimal splits than CatBoost's level-wise approach. However, both models are exceptionally well-suited for this task. For the remainder of the analysis, CatBoost was used due to its excellent out-of-the-box performance and its robust native handling of categorical features, which simplifies the pipeline.

6.2 Model Explainability and Interpretation with SHAP

A key component of this research was to ensure that the predictive model was not an uninterpretable "black box." Using the SHAP library on the trained CatBoost model provided deep insights into how it makes predictions.

Global Feature Importance:

The SHAP summary plot (Figure 9) provides a global view of feature importance, ranking features by their overall impact on the model's output. Each point on the plot represents a single patient for a given feature, with the color indicating the feature's value (red for high, blue for low) and its position on the x-axis showing its impact on the LOS prediction.

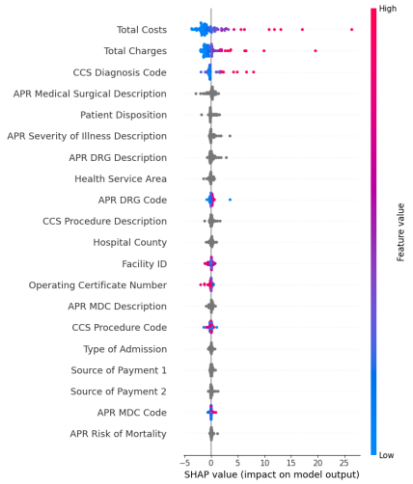

Figure 9: SHAP Summary Plot for Global Feature Importance

From Figure 9, several key insights emerge. The most influential feature is APR DRG Description, which represents the diagnosis-related group a patient is assigned to. This aligns with clinical intuition, as the primary diagnosis is the strongest determinant of the treatment course and recovery time. Other top features include Total Charges, APR Severity of Illness Description, and APR MDC Description (Major Diagnostic Category). The plot shows that high values of Total Charges and APR Severity of Illness (red points) consistently push the predicted LOS higher (positive SHAP values), which is logical.

The SHAP bar plot (Figure 10) provides a more straightforward view of overall feature importance by averaging the absolute SHAP values for each feature. This plot confirms the findings from the summary plot, reinforcing that clinical classification codes and severity markers are the dominant factors in the model's predictions.

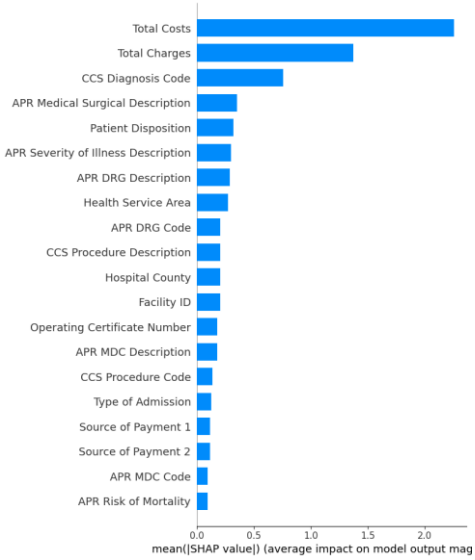

Figure 10: SHAP Bar Plot of Mean Absolute Feature Importance

Local Prediction Explanation:

Beyond global trends, SHAP allows for the explanation of individual predictions. The force plot in Figure 11 deconstructs the prediction for a single patient from the test set.

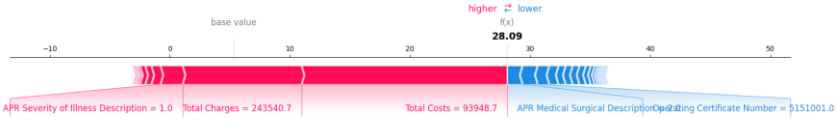

Figure 11: SHAP Force Plot for a Single Patient Prediction

The force plot shows a base value, which is the average predicted LOS across the dataset. Features shown in red are "pushing" the prediction higher, while features in blue are "pushing" it lower. For the patient in Figure 11, their specific APR DRG Description and CCS Diagnosis Description are the main factors increasing their predicted LOS. Conversely, their age group and other factors are pulling the prediction down. The final prediction is the result of these competing forces. This level of local interpretability is invaluable in a clinical setting, as it would allow a physician to understand and validate the model's reasoning for a specific patient, thereby building trust and facilitating adoption.

6.3 Optimisation Model Evaluation

The final stage of the evaluation assesses the functionality of the MILP-based bed allocation model. The model's primary task was to assign all patients from the test set to an appropriate ward without exceeding the capacity of any ward. The model successfully achieved this, finding an "Optimal" solution, which means it generated a valid allocation plan that satisfied all defined constraints.

Figure 12 provides a visual comparison of the number of patients assigned to each ward versus the pre-defined capacity of that ward.

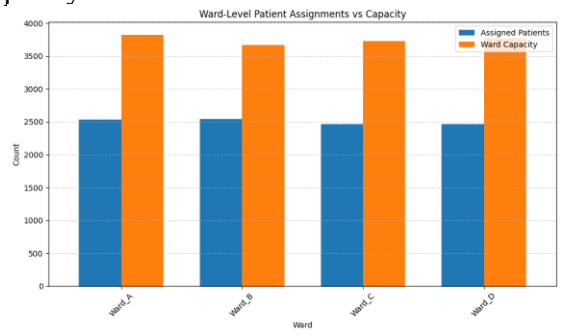

Figure 12: Ward-Level Patient Assignments vs Capacity

The bar chart confirms that the optimisation model worked as designed. For each ward, the number of assigned patients is less than or equal to the total capacity. The model distributed patients across the available wards, effectively solving the allocation problem. It is important to note that the objective function was simplified to minimise the total predicted LOS of assigned patients. This led to an interesting behaviour where patients with shorter predicted stays were preferentially assigned, although the constraints forced a distribution across all wards. In a real-world scenario, the objective function could be made more sophisticated to balance ward occupancy rates, minimise patient transfers, or incorporate patient-ward compatibility (e.g., assigning a cardiac patient to a cardiology ward).

6.4 Discussion of Findings

The evaluation confirms the success of the integrated framework. The predictive component, driven by CatBoost/LightGBM, is capable of forecasting patient LOS with a high degree of accuracy, significantly outperforming traditional methods. The SHAP analysis demystified the model, demonstrating that its predictions are driven by clinically relevant factors, which is a critical step towards practical implementation.

The optimisation component, while implemented with a simplified objective, proved to be a viable method for translating predictions into actionable bed assignment plans. The successful execution of the MILP model shows the potential of this approach for automating and optimising complex logistical decisions in a hospital setting.

A key aspect of the implementation was the use of a data sample. As noted, this was a necessary step due to computational limitations. While this provides a fair basis for model comparison, it is a limitation of the study. Scaling the training to the full 2.6 million record dataset would likely yield even better predictive accuracy, though it would require more significant computational resources. The simulated nature of the 'Ward' data and the simplified capacity planning are also limitations, but they serve to effectively demonstrate the mechanics and potential of the optimisation module. The overall framework is robust, and the results strongly support the hypothesis that an integrated AI system can provide a powerful solution for improving hospital bed management.

7 Conclusion and Future Work

This research project set out to address the critical challenge of hospital bed management by developing an integrated framework combining advanced predictive analytics and mathematical optimisation. The study successfully designed, implemented, and evaluated a system capable of accurately forecasting patient length of stay (LOS) and using these predictions to generate optimal bed allocation plans.

The investigation into various predictive models concluded that modern gradient boosting algorithms, specifically LightGBM and CatBoost, offer superior performance compared to traditional baselines like Linear Regression and Random Forest. With an RMSE as low as 2.60 on a representative data sample, these models demonstrated a strong capability to predict patient LOS from complex, high-dimensional inpatient data. The application of SHAP for model explainability was a crucial step, demystifying the "black box" nature of these models. The analysis revealed that predictions were driven by clinically intuitive factors, such as diagnosis-related groups, severity of illness, and total charges, thereby enhancing the trustworthiness and potential for clinical adoption of the system.

The predictive outputs were then successfully integrated into a Mixed-Integer Linear Programming (MILP) model. This optimisation component proved effective in solving the bed assignment problem, allocating patients to wards in a manner that respected all capacity constraints. This demonstrated the viability of translating predictive insights into actionable, operational guidance for hospital administrators. The project has thus successfully answered its primary research question by presenting a cohesive, end-to-end framework that addresses both the predictive and prescriptive aspects of hospital resource management.

7.1 Future Work:

The findings and limitations of this project open up several avenues for meaningful future research.

1. **Utilisation of Contemporary and Real-Time Data:** The most critical next step would be to apply this framework to a more recent dataset. Ideally, this would involve a partnership with a healthcare provider to access real-time or near-real-time data from an EHR system. This would allow for the development of a truly dynamic decision support tool that adapts to current conditions.
2. **Scaling to Full-Scale Datasets:** The models should be scaled to train on complete, multi-million record datasets. This would likely require the use of cloud computing platforms (e.g., AWS, Google Cloud) or distributed computing frameworks (e.g., Spark) to handle the computational load, but would be expected to yield a significant improvement in model performance.
3. **Enhancement of the Optimisation Model:** The MILP model could be significantly enhanced by incorporating more realistic constraints and objectives. This includes

adding patient-ward compatibility rules, considering nurse-to-patient ratios, minimising patient transfers between wards, and creating a multi-objective function that balances competing priorities like wait time, occupancy rate, and operational cost.

In conclusion, this research has laid a strong foundation for the application of an integrated AI framework to one of healthcare's most pressing operational challenges. By continuing to build upon this work, such systems hold the promise of creating more efficient, responsive, and patient-centric hospitals of the future.

References

Ahn, I., Gwon, H., Kang, H., Kim, Y., Seo, H., Choi, H., Cho, H.N., Kim, M., Jun, T.J. and Kim, Y.H., 2021. Machine learning–based hospital discharge prediction for patients with cardiovascular diseases: development and usability study. JMIR Medical Informatics, 9(11), p.e32662.

Alnsour, Y., Johnson, M., Albizri, A. and Harfouche, A.H., 2023. Predicting patient length of stay using artificial intelligence to assist healthcare professionals in resource planning and scheduling decisions. Journal of Global Information Management (JGIM), 31(1), pp.1-14.

Anderson, C., Bekele, Z., Qiu, Y., Tschannen, D. and Dinov, I.D., 2021. Modeling and prediction of pressure injury in hospitalized patients using artificial intelligence. BMC Medical Informatics and Decision Making, 21(1), p.253.

Bertsimas, D., Pauphilet, J., Stevens, J. and Tandon, M., 2022. Predicting inpatient flow at a major hospital using interpretable analytics. Manufacturing & Service Operations Management, 24(6), pp.2809-2824.

Buscarini, L., Romano, P., Cocco, E.S., Damiani, C., Pournajaf, S., Franceschini, M. and Infarinato, F., 2025. Enhancing patient rehabilitation outcomes: artificial intelligence-driven predictive modeling for home discharge in neurological and orthopedic conditions. Journal of NeuroEngineering and Rehabilitation, 22(1), p.117.

Dawoodbhoy, F.M., Delaney, J., Cecula, P., Yu, J., Peacock, I., Tan, J. and Cox, B., 2021. AI in patient flow: applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. Heliyon, 7(5).

El Baz, N. and Mostafa, K., 2024. Applying Artificial Intelligence Algorithms for Optimized Hospital Resource Distribution and Improved Patient Flow Management. Journal of Computational Intelligence for Hybrid Cloud and Edge Computing Networks, 8(9), pp.1-13.

Hamzah, F., 2025. Optimizing Hospital Bed Management Using Compartmental Models: Reducing Wait Times and Overcrowding.

Hirani, R., Podder, D., Stala, O., Mohebpour, R., Tiwari, R.K. and Etienne, M., 2025. Strategies to Reduce Hospital Length of Stay: Evidence and Challenges. Medicina, 61(5), p.922.

Khoury, A. and Taha, R., 2023. An Analytical Study on the Use of Machine Learning for Patient Flow Optimization and Bed Occupancy Forecasting in Urban Hospitals. Journal of Emerging Cloud Technologies and Cross-Platform Integration Paradigms, 7(10), pp.1-13.

Kumar, S., Gupta, S.K., Chaudhary, P. and Kumar, S., 2024, November. Real-Time Prediction of Patient Recovery Time using AI-driven Optimization Techniques in Healthcare. In 2024 4th International Conference on Advancement in Electronics & Communication Engineering (AECE) (pp. 1111-1116). IEEE.

Lobo, A., Barbosa, A., Guimarães, T., Lopes, J., Peixoto, H. and Santos, M.F., 2023, September. Better Medical Efficiency by Means of Hospital Bed Management

Optimization—A Comparison of Artificial Intelligence Techniques. In EPIA Conference on Artificial Intelligence (pp. 260-273). Cham: Springer Nature Switzerland.
Pahlevani, M., Taghavi, M. and Vanberkel, P., 2024. A systematic literature review of predicting patient discharges using statistical methods and machine learning. Health Care Management Science, 27(3), pp.458-478.
Ramdurai, B., 2023. In-Patient Bed Allocation by Using Markov Chain Model.
Reshma Soman, N., Aswathy Prakash, G. and Azza, H., 2025. Optimizing Hospital Operations with AI-Driven Resource Allocation Tools. Transforming Healthcare with Artificial Intelligence: Innovations and Applications, pp.99-110.
Shamsi, M., 2024. Integrating Artificial Intelligence for Prediction and Optimization in Hospital Management Systems (Case study: Iranian Hospital in Dubai). Journal of Business and Future Economy, 1(4), pp.1-9.
Soman, N.R., Prakash, G.A. and Azza, H., 2025. Optimizing Hospital Operations. Transforming Healthcare with Artificial Intelligence: Innovations and Applications, p.99.
Tello, M., Reich, E.S., Puckey, J., Maff, R., Garcia-Arce, A., Bhattacharya, B.S. and Feijoo, F., 2022. Machine learning based forecast for the prediction of inpatient bed demand. BMC medical informatics and decision making, 22(1), p.55.
van de Sande, D., van Genderen, M.E., Verhoef, C., Huiskens, J., Gommers, D., van Unen, E., Schasfoort, R.A., Schepers, J., van Bommel, J. and Grünhagen, D.J., 2022. Optimizing discharge after major surgery using an artificial intelligence–based decision support tool (DESIRE): An external validation study. Surgery, 172(2), pp.663-669.
Zhou, J., Brent, A.J., Clifton, D.A., Walker, A.S. and Eyre, D.W., 2023. Improving patient flow through hospitals with machine learning based discharge prediction. medRxiv, pp.2023-05.