

Group 19 – Project Proposal for Predicting a Patient’s Length of Stay

Project Host: NHS Wrightington, Wigan and Leigh (WWL)

Point of Contact: Thomas Ingram (thomas.ingram@wwl.nhs.uk)

A comparative study of different Machine Learning models to predict Patient’s Length of stay.

In this research, we aim to build and propose a machine learning framework for predicting a patient's length of stay based on different data factors. We will be evaluating different frameworks against each other and propose the best one.

Muhammad Maaz Bin Adnan

Department of Computing and Communications, Lancaster University, m.m.adnan@lancaster.ac.uk

Divya

Department of Computing and Communications, Lancaster University, d.divya@lancaster.ac.uk

Eyimofe Nathan Ope-Faniran

Department of Computing and Communications, Lancaster University, e.ope-faniran@lancaster.ac.uk.

Hafil Abdul Kadhar

Department of Computing and Communications, Lancaster University, h.abdulkadhar@lancaster.ac.uk.

Kumara Guru Kumar

Department of Computing and Communications, Lancaster University, kumarkg@lancaster.ac.uk.

With varying demand but a fixed capacity of beds, bed managers at NHS Wrightington, Wigan and Leigh (WWL) have the difficult job of making critical decisions in facilitating the effective use of acute hospital beds, often relying on their own experience, initiative, supplementary data, and forecasts during this process. In recent times, tools taking advantage of artificial intelligence and other statistical methods have been developed with the objective of assisting bed managers in making these decisions. One such method is predicting an individual patient's length of stay – the period from admittance to a ward to discharge from the hospital. Predicting a patient's length of stay can conceivably be completed by considering a range of patient information, including their demographics, diagnoses, and comorbidities.

CCS CONCEPTS • Machine Learning • Artificial Intelligence • Health Informatics

1 INTRODUCTION

A patient's Hospital length of stay (LOS) is defined as the time and number of days an inpatient is admitted to the hospital and is using the bed. The LOS is defined for a single admission event, that is, how many consecutive days has the patient been in the hospital. This can vary from anytime between 0 to 50+ days and depends on a number of factors that will be discussed further.

Determining a patient's LOS will help the hospitals assign better resources and establish appropriate healthcare planning by providing medical facilities and personnel accordingly. This not only helps manage the resources effectively but also takes the strain off healthcare workers without overworking them. Healthcare systems are becoming more cost-conscious, and having concrete data along with forecasting systems in place can help reduce the associated cost and improve overall patient care. Similarly, it may also help understand the underlying patterns in the data, especially which wards and hospital care units experience the most requirement for beds, which can significantly boost the evaluation of operational functions of a hospital. [1]

Alternatively, from a patient's perspective, shortening their LOS is an ideal scenario since it mitigates the dangers of them remaining in care for longer than required. These include falls, hospital-acquired infections and medication errors, all of which could in turn threaten a patient's life and lead to a prolonged LOS, straining the hospital's resources even more. [1]

Many methods have been developed for determining the LOS, and while some literature focuses on determining LOS through a pre-defined disease group, treatment, or medication [2], it involves many complicated factors, including but not limited to their characteristics i.e., their ethnicity, pre-existing complications, or general discharge planning.

2 AIMS AND OBJECTIVES

With this project, we not only aim to come up with a system to manage a patient's discharge and provide accurate predictions for better planning but also to understand which variables have the most impact on the length of the stay and whether explainable models could be built to help understand the predictions it makes. We aim to review existing methods and, similarly, build a framework that can be applied to the specific hospitals in question, and what data would be needed to create a model that could be used universally. All of this is essential for the hospital's resources and the patient's well-being.

3 PROPOSED APPROACH

3.1 Data Understanding

Our project is in collaboration with the NHS in Wigan, Wrightington and Leigh. This section aims to provide a basic understanding of the data. The data consists of **101** columns and approximately **41,000** rows.

The target column is **spell_episode_los**, which is the duration of time a patient has been in the hospital. It is a calculated column that is essentially a subtraction of the discharge date and the admission date columns.

The data consists of 56 numerical columns and 45 string columns.

It also consists of a high number of missing values, especially for certain columns where the probability of a true value is very low, for example, the **covid_19_diagnosis_flag**.

3.2 Data Preparation

This section details the end-to-end data preparation plan for predicting patient Length of Stay (LOS) at WWL. Since the data is diverse, one of the main goals will be to encode categorical variables for regression models. Next, the data will be cleaned, with particular focus on columns containing a high proportion of missing values. This may involve simple mean imputation or building regression models to estimate missing entries, depending on the data distribution and model requirements. We will also normalise variables with high variance to ensure consistent scaling across predictors.

In addition, feature selection will be carried out to reduce dimensionality and eliminate redundant or irrelevant variables. An example is the variables called '**site_national_code**', '**site_description**', and '**site_local_code**'; all three represent the same thing and are simply extensions of one another, so we can only keep '**site_national_code**'. Outliers, especially in clinical measurements, will be identified using statistical methods such as the interquartile range (IQR) and either treated or removed based on how they affect the model. Finally, continuous variables can be standardised using z-score or min–max normalization, while categorical variables will be carefully encoded to maintain consistency across training and testing sets.

3.3 Models

N = 41846 records

Mean LOS = 1.76 days

Median LOS = 0 days

SD LOS = 5.28

Max LOS = 123 days

There are over forty thousand patient entries in our dataset, with the average patient LOS being 1.76 days. This distribution is heavily right-skewed, meaning standard regression models (like Linear Regression) would struggle because they assume a roughly normal target variable. We'll use models designed for skewed outcomes.

The prediction of hospital length of stay (LOS) will be approached as a regression problem. Given the highly skewed nature of the LOS variable, a combination of generalised linear models (GLMs) and tree-based ensemble methods will be developed and compared.

Initial modelling will involve Poisson Regression and Negative Binomial Regression, both of which are well-suited to count or duration-type outcomes [3]. These models provide interpretable coefficients that can quantify how demographic and clinical features influence the expected LOS.

To capture nonlinear relationships and complex feature interactions, Random Forest Regression will be employed as a non-parametric alternative [4]. If performance improvements are sought, Gradient Boosting techniques such as XGBoost may also be explored, given their established success in healthcare prediction tasks [5].

Model selection will be guided by a combination of predictive performance metrics—including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R²—as well as interpretability considerations relevant to clinical decision-making.

3.4 Evaluation

The performance and dependability of the models created to forecast patient length of stay (LOS) will be the focus of the evaluation phase. Metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R²) will be used to gauge how accurate the model is. To guarantee generalisability and avoid overfitting, a train–test split and k-fold cross-validation will be used. To determine the best strategy, the performance of machine learning models (Random Forest, XGBoost) and statistical models (Poisson, Negative Binomial Regression) will be compared. To ensure both predictive accuracy and clinical interpretability, feature importance and SHAP analysis will also be performed to interpret model predictions and identify the factors that have the greatest impact on LOS.

4 PROPOSED TIMELINE

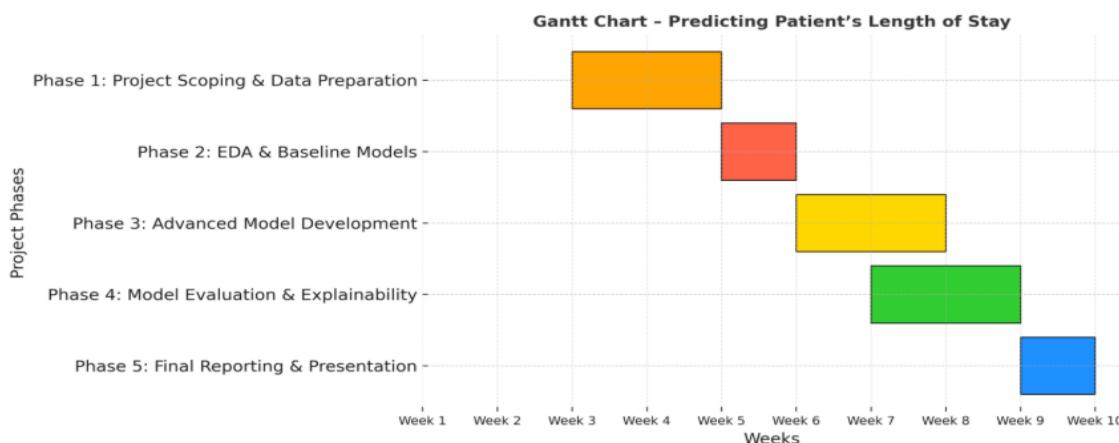


Fig 4.1 – Gantt Chart for Project timeline

5 TEAM SETUP

Group 19: Predicting Patient Length of Stay

Role	Team Members	Responsibilities
Data Analysts	Kumara Guru Kumar [kumarkg@lancaster.ac.uk], Hafil Abdul Kadhar [h.abdulkadhar@lancaster.ac.uk]	Conduct initial data exploration, statistical analysis, and visualisation to identify key Length of Stay (LOS) drivers. Support feature selection and provide analytical insights for model refinement.
Data Engineers	Muhammad Maaz Bin Adnan [m.m.adnan@lancaster.ac.uk], Eyimofe Nathan Ope-Faniran [e.ope-faniran@lancaster.ac.uk]	Manage data collection, cleaning, and transformation. Develop reproducible preprocessing pipelines, handle missing values, and ensure data integrity across modelling stages.
Data Scientists	Divya[d.divya@lancaster.ac.uk] Eyimofe Nathan Ope-Faniran [e.ope-faniran@lancaster.ac.uk], Muhammad Maaz Bin Adnan[m.m.adnan@lancaster.ac.uk]	Build and compare baseline (Poisson, Negative Binomial) and advanced (Random Forest, XGBoost) models. Perform model tuning, evaluation, and interpretability analysis using SHAP.
Project Manager	Divya[d.divya@lancaster.ac.uk.]	Oversee project planning, coordination, and documentation. Ensure objectives, timelines, and ethical data practices are maintained throughout all phases.
Presenters	Eyimofe Nathan Ope-Faniran [e.ope-faniran@lancaster.ac.uk], Muhammad Maaz Bin Adnan[m.m.adnan@lancaster.ac.uk], Divya[d.divya@lancaster.ac.uk].	Prepare and deliver the final presentation. Summarise findings, visualise results, and communicate outcomes clearly to stakeholders.

REFERENCES

- [1] K. Stone, R. Zwiggelaar, P. Jones, and N. Mac Parthaláin. 2022. A systematic review of the prediction of hospital length of stay: Towards a unified framework. *PLOS Digital Health* 1, 4 (2022), e0000017. <https://doi.org/10.1371/journal.pdig.0000017>
- [2] S. Shea, R. V. Sideli, W. DuMouchel, G. Pulver, R. R. Arons, and P. D. Clayton. 1995. *Computer-generated Informational Messages Directed to Physicians: Effect on Length of Hospital Stay*. Journal of the American Medical Informatics Association 2, 1 (Jan. 1995), 58-64. <https://doi.org/10.1136/jamia.1995.95202549>
- [3] Breiman, L. Random Forests. *Machine Learning* 45, 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
- [4] Hilbe, J. M. (2011). *Negative Binomial Regression* (2nd ed.). Cambridge: Cambridge University Press. <https://doi.org/10.1017/CBO9780511973420>

- [5] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *The New England journal of medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMra1814259>