Predicting Health Care Facility Stay Duration: A Machine Learning Approach

Georgios Kleitou
Department of Science and Engineering
Southampton Solent University
Southampton, United Kingdom
kleitougeorgioswork@gmail.com

Jarutas Andritsch
Department of Science and Engineering
Southampton Solent University
Southampton, United Kingdom
jarutas.andritsch@solent.ac.uk

***Abstract*— The COVID 19 pandemic revealed shortcomings in healthcare, particularly concerning bed occupancy and resource allocation. During the Delta variant wave, it was highlighted how much improvement is needed in management strategies. One promising solution is the prediction of inpatient Length of Stay. Accurate predictions can enhance efficiency, reduce infection risks, lower mortality rates and decrease bed occupancy. This research proposes a predictive model using Random Forest Regression to accurately forecast hospital length of stay, aiming to enhance resource management and patient care. We utilized a 2010 inpatient dataset from the New York Department of Health and conducted thorough data preprocessing, including cleaning, handling missing values, and numerical encoding of categorical variables for regression. Additionally, we experimented with three database variations: one with targeted and frequency encoding, another using synthetic minority oversampling technique for handling imbalances, and a third applying synthetic minority oversampling technique for regression with gaussian noise for continuous variables. Each database was tested with and without scaling using four different scalers. The objective was to achieve a mean absolute error below the industry standard of 6.5, prioritizing unbiased metrics. Our results indicate that the final model achieved a 2.93 mean absolute error on the normal database, demonstrating its effectiveness in predicting length of stay. The study underlined the potential of machine learning in accurately predicting the Length of Stay in hospitals and the possibility of a more accurate model of the industry standard. Further advancements could be made to the models with more balanced datasets and a user-friendly interface for hospital staff usage.**

Keywords*— *length of stay, machine learning, health care, Random Forest, Regression

I. INTRODUCTION

The COVID-19 pandemic indicated various shortcomings in healthcare on a global scale. Reports of overcrowding, mishandling of space and resource allocation resulted in a lack of oxygen and beds [1]. When the Delta variant followed, the situation escalated affecting over 78 countries [2]. In an effort to examine the situation the Cyber Security and Infrastructure Agency developed a predictive model based on the data they possessed which was 75% of bed occupancy resulted in 12000 excess deaths, then the prediction showcased that in the duration of 2 weeks that could spiral in a 100% bed occupancy and 80000 excess deaths [3]. Researchers have discovered that prolonged length of state (PLoS) had tremendous impact on bed occupancy. The studies made revealed that targeting extended length of stay (LoS) can optimize a health care facility's resource distribution [4]. Thus, tackling LoS is essential to handling healthcare facility efficiency. LoS represents the duration spent by an inpatient in a healthcare facility [5]. Successful supervision of LoS can improve resource distribution, decrease infection risks, lower mortality rates, declined bed occupancy and grow the profits of the facility [6]. The National Health Service (NHS) has been investigating hospital stay since 2018 aiming to improve on leadership, communication, support and research [7]. This experiment was done to uncover prediction of an inpatients LoS with precision based on various variables of the patient and regression models.

II. LITERATURE REVIEW

A. Models Used in prior research

A noteworthy study in Saudi Arabia aimed to predict the likelihood of COVID-19 patients being transferred to the ICU and forecast their LoS. After preprocessing the dataset, the researchers developed various predictive models. The models utilized were: Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost) and an ensemble of these models. The better predictive model was the RF [8]. The study focused on classification but most of these models have a regression counterpart. A separate study used data from the Emergency Department (ED) of Sant'Orsola Malpighi university hospital in Bologna, Italy to evaluate different predictive models for LoS and PLoS to assist health care decision making. The researchers used a total of 8 models which included the Least Absolute Shrinkage and Selection Operator (LASSO), Elastic-Net (EN), Random Forest Regressor (RFR), GB, AdaBoost, Support Vector Machine (SVM), XGBoost and K-Nearest-Neighbor (KNN). The study used regression models to measure the shorter LoS and classification for PLoS. The GB excelled in classification while the XGBoost excelled in regression with a prediction error of 6-7 days [9]. These two projects also accentuated the importance of data preprocessing before commencing any kind of experimentation with the predictive models.

B. Feature Importance

When the choice for the variables to consider in the models was to be done academic literature was researched to include relevant information from the patients. A study done in France by the Assistance Publique university hospital in 2022 investigated features that were associated with PLoS and evaluating those features through machine learning models to predict LoS [10]. The study focused on binary classification and employed models like Neural Networks, decision trees, and Random Forest. The GB was crowned as the best performing model. The study also concluded in features that had an impact on PLoS to be the ED admission age, neurological issues, discharge destination, mental health conditions and heart disease [10].

C. Metrics

Almost every study on predictive models operates with metrics to evaluate model performance. A leading example of measuring performance on regression model instances would be the research done at Duke University in 2019 [11]. The researchers gathered 3 years of elective surgery data along with their personal information to evaluate performance of patients undergoing such surgeries under the COVID-19 pandemic. Their trials included three models: LASSO, RF and MLP. To measure the loss of their models they used Mean Absolute Error (MAE), Mean Squared Error (MSE) and Mean Relative Error (MRE). Although the outcome was determined with a two-stage classification/regression model the loss was measured by metrics of loss that measure regression model loss. What was discovered during the measurement is that MAE was a primary candidate since the MSE was prone to large errors when dealing with outliers and MRE was sensitive to smaller values making the MAE to be more balanced than the other two [11].

D. Literature Gap

The literature review led to the idea of the possibility of precise predictions in healthcare facility inpatients. While the research has been done on the same subject no idea tried to push predictions on the precision front using regression. The only thing close that could be discovered were two stage models. The idea of discovering models that could predict closer than the normal classification group of days could really help with the measurement of resource and staff allocation.

III. Methodology

The project relied heavily on the identification of a highly efficient predictive model. This followed worked similar to the project of Gupta, S et al [12], following practices like data manipulation, data cleaning, model comparison relied on metrics and finally model application and investigation of the most effective models [12].

A. Data Selection

The New York Department of Health (NYDH) has anonymous inpatient data available throughout the years on their website which can be accessed and downloaded by anyone [13]. As hospitals rarely distribute such data the decision was made to gather what was given publicly. The year that was chosen was 2010 due to the reason of no major out breaks of any viruses or infections according to the CDC at that time in the United States [14]. By choosing a calm year more common ailments could be used for the models to predict rather than having outliers that could spike or provide outliers to the data.

B. Data Description and Data Dropping

Due to the age of the dataset the data included some codes from the old and out of use International Classification of Disease 9th Revision Clinical Modification (ICD9CM). This system provides many codes for procedures, diseases and diagnosis. What was kept from the ICD9CM was the Clinical Classification Software (CCS) due to its sorting and categorizing [14]. Additionally, the All-Patient Refined codes were kept because it included 2 separate categories, the Diagnosis Related Group (DRG) and the Major Diagnostic Categories (MDC). The DRG was dropped due to its financial inclusion while the MDC is purely diagnostics for patients [15]. If those columns were kept, then that would potentially have created multicollinearity which would have led to relationships developing due to the repetition of groupings with high correlations [16]. Due to the multicollinearity the model would overfit and result in poor generalization abilities [17]. To overcome this obstacle such columns were dropped including some columns that were irrelevant to the scope, down at Table I there is a reason behind every variable dropped.

TABLE I. Data drop

Data Drop	Reason
Facility ID, CCS & APR Descriptions Health Service Area Zip Code	These were removed due to already existing variables that existed as it would cause relationships if not removed
APR DRG	APR MDC for simplification
Licensing Variables	Irrelevant
Costs/Charges	Cost could be a factor for leaving in the U.S but patients that left against medical advice were removed from the dataset to counter that. This was also done to make the study have universal applications

C. Data Mapping

The data had to be transformed into numerals to be used in regression models. Various encoding methods were employed to map out the whole dataset from categorical to numerical. Binary Encoding was used to label variables that could be interpreted with 0 or 1 like Gender or Medical/Surgical Description (Medical or Surgical) [18]. Base n encoding was used that uses various numerals to arbitrary represent variable and used when binary encoding cannot capture the dimensionality, like Race and Type of Admission. Lastly label/ordinal encoding that functions the same as base n encoding but the number represents an ordinal value that represents a relationship with the succeeding number for example age group would be one [19].

D. Exploratory Data Analysis

The backbone for good quality data will always have to go through Exploratory Data Analysis (EDA). The EDA performed provided an oversight of the dataset as well as informed about the short comings of it. The dataset was unfortunately imbalanced, and an imbalanced dataset could have led to poor performance in the prediction models [20]. Viewing the quartiles provided the mean LoS was 5.47 days and the standard deviation 7.63 days. The dataset in quartiles goes as follows: Q1 = 2 days or lower, Q2 = 3 days or lower (median), Q3 = 6 days or lower and that is with minimum of 1 day and maximum of 119 days. Moreover, measuring the correlation coefficient was used to calculate standard deviation, mean and product of the variables [21]. Nothing really stood out at first glance, but more were discovered later model feature importance.

E. Metrics

Given the experimentation with various models, metrics were a necessity to evaluate the models. Four models were used to lower the collection of available models that were assembled through literature. The metrics were the MSE, MRE, MAE and the R squared which was used to evaluate the variance of independent variables and observe if the data

were accustomed to the model. Having a lower R squared suggested worse fitting models, but the R squared is less sensitive to extreme errors and outliers and could potentially ignore overfitting or be unable to handle negative space of the coefficient. Nevertheless, it was implemented as an additional metric of measure to provide supplementary information on model quality [22]. This was done to avoid bias by being focused on only one specific metrics.

F. Experiment Methodology

This research was a large-scale experiment for model performance discovery on precise LoS. The experiment comprised of 4 dataset variations, 4 scalers and 1 additional time with no scalers and 6 machine learning models for a total of 120 model trials.

1) Database Variations

Due to the imbalance contained in the data on the LoS the dataset experimented with the normal dataset that was mapped for regression, a database using Synthetic Minority Oversampling Technique (SMOTE) to oversample all kinds of days for LoS to a cap of 20000 instances each. Based on that, a trial on Synthetic Minority Oversampling Technique for Regression with Gaussian Noise (SMOGN) was also attempted to aim for increasing the data diversity on continuous variables and an encoded database which was the original regression mapped database but encoded even further with frequency encoding on non-continuous variables and impact encoding for ordinal variables to uncover any extra relationships for the models.

2) Scalers

For scalers a trial with no scalers was used and an additional of 4 other scalers were used for each model. The models included were the Standard Scaler, MinMax Scaler, Robust Scaler and MaxAbsolute Scaler. Every scaler has their own unique use and trait. However, all of them were attempted in the trials to appraise models with each of them.

3) Models

The assortment of models that were used included an ElasticNet was used to uncover shortcomings located in ridge regression and LASSO the ElasticNet combines both penalties of shrinking some coefficients to zero for variable selection while reducing magnitude of all coefficients [23]. The Polynomial on the other hand was used to explore non-linear relationships that could have developed between variables. XGBoost regressor was used in the place of Gradient Boosting regressor to provide speed, accuracy as well as preventing overfitting [24]. Following examples from the literature reviewed the MLP was put to the test due to its data splitting, early stop, learning rate and batch size features. The KNN was also included for the prediction capabilities on continuous variables by averaging responses to the k-nearest neighbors [25]. The RFR was employed due to its single measure of variable importance making it important for research purposes [26]. Furthermore, hyper parameters were used but due to their time consumption and expensive computational resources even with automated methods such are Grid Search were only used at the last few promising models [27]. Segmetation of the database was also made to observe model performance in more concentrated areas of the database. The segments were split into 1-7 days due to most of the data being there, 8-23 days as per the research of the NHS and 24+ days for the rest of the sparse data.

IV. RESULTS

After model trials commenced it was obvious that the feature importance relied heavily on Facility ID so the decision was made to be removed as viewed on Table I. Even though Facility ID was a logical feature the decision was taken to be removed as the model should rely heavily on inpatients so it can have an application for every hospital wanting to use it. After this minor adjustments it was a showdown between the XGBoost with the normal database and the RFR being compatible with most databases. Even though the SMOTE database provided better metrics the synthetic samples made no sense. After hyper tuning and database segmentation the RFR proved to be more consistent in bigger ranges of LoS but the XGBoost proved minorly better in lower dates where the majority of data were. The metrics can be observed on Table II and the comparison with actual LoS are on Table III.

TABLE II. COMPARISON OF MODEL EVALUATION METRICS

Metrics	**Models**					
	ELASTIC	***POLYNOMIAL***	***XGBOOST***	***MLP***	***KNN***	***RFR***
MAE	3.55	3.41	2.80	2.98	3.11	**2.93**
MSE	49.66	46.80	34.29	37.50	41.78	37.86
MRE	11.23 %	104.09 %	75.71 %	81.59 %	79.90 %	75.79 %
R^2	0.15	0.20	0.41	0.36	0.28	0.35

TABLE III. COMPARISON OF ACTUAL LENGTH OF STAY AND PREDICTED LENGTH OF STAY

Actual LoS	**XGBOOST Predictions**	**RFR Predictions**
2	1.78	**1.94**
3	2.79	**2.51**
3	3.66	**3.44**
15	12.73	**14.95**
16	12.43	**14.39**
100	88.17	**95.09**
100	66.55	**90.99**

V. DISCUSSION

The experiment trialed the possibility of an accurate LoS prediction for healthcare facility inpatients. Even with an imbalanced dataset and limited computational resources promising results were yielded, this also proved a roadblock for further researcher. More models were in the list for trials but due to time they were not such as gradient boosting and support vector machine. As for the database an attempt was made to access the MIMIC-IV but to no avail. For future work finding a more diverse dataset and adding a GUI is certainly a must.

VI. CONCLUSION

The possibility of predicting length of stay through machine learning algorithms is a feasible feature and seems positively promising to help health care facilities with their distribution of resources and staff allocation. With the predicting abilities of the Random Forest Regressor the

experiment was proven to be a great leap forward compared to its classification counterparts who narrow down predictions to discrete classes, regression provided a more precise prediction of LoS. While there is the concern for the prediction radius it falls off, this research indicates that it is still considerably closer to classification studies that bundle days in classes.

REFERENCES

[1] E. Mahase, "Covid-19: Hospitals in crisis as ambulances queue and staff are asked to cancel leave," *BMJ*, pp. m4980–m4980, Dec. 2020, doi: https://doi.org/10.1136/bmj.m4980.

[2] A. M, Tareq, T. B. Emran, Kuldeep Dhama, M. Dhawan, and T. E. Tallei, "Impact of SARS-CoV-2 delta variant (B.1.617.2) in surging second wave of COVID-19 and efficacy of vaccines in tackling the ongoing pandemic," Human Vaccines & Immunotherapeutics, vol. 17, no. 11, pp. 4126–4127, Sep. 2021, doi: https://doi.org/10.1080/21645515.2021.1963601. .1971312. [PMCID: PMC8425453].

[3] G. French et al., "Impact of hospital strain on excess deaths during the COVID-19 pandemic—United States, july 2020–july 2021," American Journal of Transplantation, vol. 22, no. 2, pp. 654–657, Feb. 2022, doi: https://doi.org/10.1111/ajt.16645.

[4] M. P. Quinn, A. E. Courtney, D. G. Fogarty, D. O'Reilly, C. Cardwell, and P. T. McNamee, "Influence of prolonged hospitalization on overall bed occupancy: a five-year single-centre study," QJM, vol. 100, no. 9, pp. 561–566, Jun. 2007, doi: https://doi.org/10.1093/qjmed/hcm064.

[5] Md. M. Rahman, D. Kundu, S. A. Suha, U. R. Siddiqi, and S. K. Dey, "Hospital patients' length of stay prediction: A federated learning approach," *Journal of King Saud University - Computer and Information Sciences*, Jul. 2022, doi: https://doi.org/10.1016/j.jksuci.2022.07.006.

[6] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, "Analysis of length of hospital stay using electronic health records: A statistical and data mining approach," PLoS ONE, vol. 13, no. 4, pp. e0195901–e0195901, Apr. 2018, doi: https://doi.org/10.1371/journal.pone.0195901.

[7] NHS, "NHS England» Reducing length of stay," England.nhs.uk, 2018. https://www.england.nhs.uk/urgent-emergency-care/reducing-length-of-stay/ (accessed Sep. 28, 2024).

[8] D. A. Alabbad et al., "Machine learning model for predicting the length of stay in the intensive care unit for Covid-19 patients in the eastern province of Saudi Arabia," Informatics in Medicine Unlocked, vol. 30, pp. 100937–100937, Jan. 2022, doi: https://doi.org/10.1016/j.imu.2022.100937.

[9] A. J. Zeleke, P. Palumbo, P. Tubertini, R. Miglio, and L. Chiari, "Machine learning-based prediction of hospital prolonged length of stay admission at emergency department: a Gradient Boosting algorithm analysis," Frontiers in Artificial Intelligence, vol. 6, Jul. 2023, doi: https://doi.org/10.3389/frai.2023.1179226.

[10] F. Jaotombo, V. Pauly, G. Fond, V. Orleans, P. Auquier, B. Ghattas, et al., "Machine-learning prediction for hospital length of stay using a French medico-administrative database," *Journal of Market Access & Health Policy*, vol. 11, no. 1, Nov. 2022, doi: https://doi.org/10.1080/20016689.2022.2149318.

[11] Z. Xu, C. Zhao, C. D. Scales, R. Henao, and B. A. Goldstein, "Predicting in-hospital length of stay: a two-stage modeling approach to account for highly skewed data," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, Apr. 2022, doi: https://doi.org/10.1186/s12911-022-01855-0.

[12] S. Gupta, K. Saluja, A. Goyal, A. Vajpayee, and V. Tiwari, "Comparing the performance of machine learning algorithms using estimated accuracy," Measurement: Sensors, vol. 24, p. 100432, Dec. 2022, doi: https://doi.org/10.1016/j.measen.2022.100432.

[13] New, "Hospital Inpatient Discharges (SPARCS De-Identified): 2010," *Ny.gov*, Sep. 16, 2013. https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/mtfm-rxf4/about_data (accessed Sep. 28, 2024).

[14] "Clinical Classifications Software (CCS) for ICD-9-CM," Ahrq.gov, 2015. https://hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (accessed Sep. 28, 2024).

[15] R. Averill, N. Goldfield, J. Hughes, C. Muldoon, R. Gregg, A. McCullough, et al., "ALL PATIENT REFINED DIAGNOSIS RELATED GROUPS (APR-DRGs) Methodology Overview 3M Health Information Systems," 2023. Available: https://hcup-us.ahrq.gov/db/nation/nis/APR-DRGsV20MethodologyOverviewandBibliography.pdf

[16] K. Kunanbayev, I. Temirbek, and A. Zollanvari, "Complex Encoding," 2021 International Joint Conference on Neural Networks (IJCNN), Jul. 2021, doi: https://doi.org/10.1109/ijcnn52387.2021.9534094.

[17] Jireh Yi-Le Chan et al., "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review," Mathematics, vol. 10, no. 8, pp. 1283–1283, Apr. 2022, doi: https://doi.org/10.3390/math10081283.

[18] D. Shah, Z. Y. Xue, and T. M. Aamodt, "Label Encoding for Regression Networks," *arXiv.org*, 2022. https://arxiv.org/abs/2212.01927 (accessed Sep. 28, 2024).

[19] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," IEEE Access, vol. 9, pp. 114381–114391, Jan. 2021, doi: https://doi.org/10.1109/access.2021.3104357.

[20] [1] C. G. Weng and J. Poon, "A new evaluation measure for imbalanced datasets," in *Proceedings of the 7th Australasian Data Mining Conference (AusDM '08)*, Glenelg, Australia, 2008, pp. 27-32. Available: https://doi.org/10.5555/2449288.2449295.

[21] "11. Correlation and regression | The BMJ," The BMJ | The BMJ: leading general medical journal. Research. Education. Comment, Oct. 28, 2020. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression (accessed Sep. 28, 2024).

[22] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," PeerJ Computer Science, vol. 7, p. e623, Jul. 2021, doi: https://doi.org/10.7717/peerj-cs.623.

[23] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," Journal of the Royal Statistical Society Series B (Statistical Methodology), vol. 67, no. 2, pp. 301–320, Mar. 2005, doi: https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[24] R. Wang, L. Wang, J. Zhang, M. He, and J. Xu, "XGBoost Machine Learning Algorism Performed Better Than Regression Models in Predicting Mortality of Moderate-to-Severe Traumatic Brain Injury," World Neurosurgery, vol. 163, pp. e617–e622, Apr. 2022, doi: https://doi.org/10.1016/j.wneu.2022.04.044.

[25] M. Azadkia, "OPTIMAL CHOICE OF k FOR k-NEAREST NEIGHBOR REGRESSION," 2020. Available: https://arxiv.org/pdf/1909.05495

[26] A. Liaw and M. Wiener, "Classification and Regression by randomForest," Classification and Regression by randomForest, vol. 2, no. 3, pp 18-20, 2002, Available: https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[27] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," Neurocomputing, vol. 415, pp. 295–316, Nov. 2020, doi: https://doi.org/10.1016/j.neucom.2020.07.061.