Improving Patients' Length of Stay Prediction Using Clinical and Demographics Features Enrichment

Hamzah Osop[1], Basem Suleiman[2,3], Muhammad Johan Alibasa[4], Drew Wrigley[2], Alexandra Helsham[2], and Anne Asmaro[2]

[1] Nanyang Technological University, Singapore
hamzah.osop@ntu.edu.sg
[2] The University of Sydney, Australia
basem.suleiman@sydney.edu.au
{dwri5016,ahel9827,aasm3350}@uni.sydney.edu.au
[3] The University of New South Wales, Australia
basem.suleiman@unsw.edu.au
[4] Telkom University, Indonesia
alibasa@telkomuniversity.ac.id

Abstract. Predicting patients' length of stay (LOS) is crucial for efficient scheduling of treatment and strategic future planning, in turn reduce hospitalisation costs. However, this is a complex problem requiring careful selection of optimal set of essential factors that significantly impact the accuracy and performance of LOS prediction. Using an inpatient dataset of 285k of records from 14 general care hospitals in Vermont, USA from 2013-2017, we presented our novel approach to incorporate features to improve the accuracy of LOS prediction. Our empirical experiment and analysis showed considerable improvement in LOS prediction with an XGBoost model RMSE score of 6.98 and R2 score of 38.24%. Based on several experiments, we provided empirical analysis of the importance of different feature sets and its impact on predicting patients' LOS.

Keywords: machine learning · length of stay · electronic health records.

1 Introduction

The global cost of healthcare is rising faster than any increase in provisions for its funding. Thus, with an aging population, there is an added pressure to reduce the costs associated with patient treatment in hospitals [6]. Patients' length of stay (LOS) and hospital readmissions are factors that make up the true cost of hospitalisation [10]. LOS remains one of the biggest drivers of costs and a determinant in patients' lives saved within healthcare. Predicting the LOS allows for more effective and efficient planning within the hospital. It also improves the scheduling of elective surgeries, while also supporting the long-term strategic planning of the hospital [11]. A prime example of the significance of

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_9

LOS, besides the availability of health equipment, has been COVID-19 which unprecedentedly tested the effectiveness of healthcare systems across the globe.

The goal of this paper is to predict LOS using machine learning methodologies to support hospitals in LOS management. We enrich our data by contributing to novel feature extraction techniques utilising International Classification of Disease (ICD) 9 and ICD10 codes, and admission information. We provide an empirical evaluation by training and testing five machine learning models to evaluate and to present the best performing model, to identify opportunities for future research in this domain.

In this paper, we present a preliminary study on a methodological approach for predicting patients' LOS based on the Vermont dataset, a real-world hospitalisation dataset. Predicting LOS is very challenging as it requires employing an optimal set of diverse features that are often not available in many datasets [12]. Hence, we present our empirical approach for enriching the Vermont dataset by incorporating features including Charlson Comorbidity Index (CCI), rank procedure severity, categorise procedures, patient median income and median LOS of the prior year. Besides these features, the Vermont dataset includes a large number of other features which might be necessary for prediction LOS [3]. Thus, we also present a machine learning method to identify the optimal and essential features that are crucial for predicting LOS. We construct several regression models to predict LOS using the enriched and selected features of Vermont dataset. We present our empirical experiments and show the best-performing regression model that significantly outperforms other benchmarks in terms of RMSE and R2 score. Our empirical results further highlight the importance of three categories of features and their impact on the accuracy of LOS prediction.

2 Related Work

There are a number of factors that influence and contribute to a patient's LOS and in turn, the ability to predict LOS. Buttigieg et al [3] summarised factors that impacted LOS from 46 research papers, with relationships identified between emergency department crowding, early patient transfer to specialists, date and time of admission, access to early imaging and patient income to name a few. The relationship between patient demographics and LOS is prevalent in prior studies. Particularly, age was identified as one of the main drivers, explaining for 87% of LOS variability [10]. This can also be seen when predicting patient discharge [1], LOS and readmission after colorectal resection [8] with age being one of the most predictive features of LOS. Further, patients' payment type has also emerged as a predictor of LOS, for example, in predicting LOS amongst cardiac patients [7]. The socio-economic position of a patient also influences LOS, with low income associated with inadequate housing increasing LOS by 24% [3].

Past studies have used machine learning to predict LOS from inpatient data. Daghistani et al [7] used four machine learning algorithms, including Random Forest (RF), Bayesian Network (BN), Support Vector Machine (SVM) and Artificial Neural Networks (ANN). They found that RF algorithm performed best on

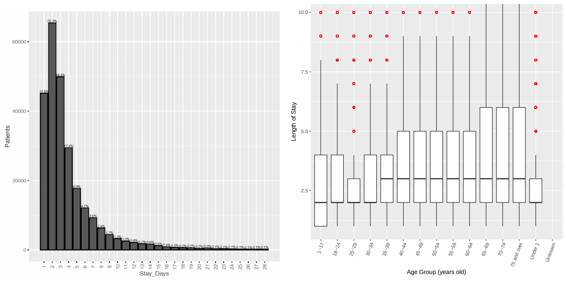
DOI: 10.1007/978-3-031-36021-3_9

70 features derived from demographic information, cardiovascular risk factors, vital signs on admission and admission criteria. Further developments on LOS prediction include extending it to predict the time of patient discharge. Barnes et al [1] conducted a study to classify if a patient will be discharged at 2 pm or midnight that day, using data available at 7 am daily. While this use case is a slight variation of LOS prediction, it highlights the range of applications and value of predicting LOS. The key feature used were the elapsed length of stay, observation status, age, reason for visit and day of the week. However, these past studies did not include various rich features such as comorbidity measures, and other feature extraction techniques as they mostly used demographic and hospital admission information.

3 Methodology

3.1 Dataset and Exploratory Analysis of Vermont Dataset

We utilised the Vermont Hospital Dataset, consisting of inpatient discharge data (260k records), outpatient procedures and services data (8.3M records), and emergency department data (1.3M records) collected across Vermont's 14 general care hospitals using the Vermont Uniform Hospital Discharge Data System (VUHDDS). The datasetspanned between 2013 to 2017, included attributes such as (i) diagnostic discharge data, (ii) patient socio-demographic characteristics, (iii) ICD9 and ICD10 codes for diagnosis and reasons for admission, (iv) patient treatment and services provided, (v) length of hospital stay, and (vi) financial data such as billing and charges.



(a) LOS by day (b) LOS by Age Band

Fig. 1. Distribution of Length of Stay (LOS)

Initial analysis of the dataset revealed a strong skew in the number of stay days, with about 93% of LOS being ten days or less (refer to Figure 1(a)). Given the age banding and a disproportion towards low stay days, an analysis of patients' age group with LOS showed a general association of older patients with a longer period of stay as shown in Figure 1(b). The primary diagnostic group (ccsdx) mapped patients' diagnosis to a grouped category of diagnoses, was analysed with the distribution of LOS. The analysis of the top 15 ccsdx codes showed a significant amount of variance between these categories, suggesting a strong relationship between diagnosis and patient LOS.

3.2 Data Processing and Feature Engineering

To accommodate missing values, data variables were approximated, turned to factors or removed completely. The dataset was anonymised and without features like medical history, readmission visits, medication, or clinical data such as laboratory testing, the dataset was further enriched using secondary data sources. The ICD9 and 10 codes consisted of unique string values and were difficult to consume by machine learning models. Hence, the ICD codes were mapped to (1) replace 'NA' values, (2) map ICD9 to ICD10 codes, (3) calculate the Charlson Comorbidity Index, (4) rank procedure severity and, (5) categorise procedures, making it more meaningful.

Handling missing values. There were several data fields that had high counts of 'NA' values: the diagnosis codes fields (DX1-DX20) and procedure codes fields (PX1-PC20). ICD mappings were used in place of the missing values.

Aligning ICD9 and ICD10 Data. The dataset contained a mix of ICD9 and ICD10 codes depending on the year the codes were used. To create consistency in the use of ICD codes, all ICD9 codes were mapped to ICD10, based on mappings provided by the Bureau of Economic Research (NBER).

Calculating Charlson Comorbidity Index (CCI). The CCI predicts the 10-year mortality of a patient with a range of comorbid procedures [5]. The CCI calculation was performed using the ICD library in R, where each ICD9/10 codes correspond to a certain score.

ICD Procedure Classes. Two conversions were performed on the procedure data fields. First, procedures were classified according to the first character in the ICD codes which specified the classes of procedures performed. This classified each ICD code into categories with the corresponding nature of procedures. The second conversion categorised the procedures into four classes of Minor Diagnostic, Major Diagnostic, Minor Therapeutic and Major Therapeutic. The categorisation was determined based upon whether a procedure was (1) diagnostic or therapeutic, and (2) performed in an operating room.

Median Income. Patients with lower socioeconomic status have been shown to stay in hospitals longer as their external environment is not adequate for proper care requirements. Given the details of the 14 hospitals in the dataset, the zip codes of those hospitals were extracted. The zip codes were then mapped to publicly available median income data sources [4].

Median LOS for Prior Year. Given the correlation between LOS and age, a new feature called Median LOS of Prior Year was introduced into the dataset. Based on the patient's age group, the median LOS from the previous year was calculated and labelled in the dataset as median_intage.

Y Variable Log Transformation. Due to the dataset being highly skewed, a log transformation on the LOS data field was performed to convert it towards

DOI: 10.1007/978-3-031-36021-3_9

being normally distributed. This is to ensure that models such as linear regression could deliver the best model performance.

4 Experiments

4.1 Constructing Model to Predict LOS

We selected four regression models to empirically evaluate the LOS prediction, namely Linear Regression, SVM, RF and XGBoost. The dataset was split into training and testing data, where training data contained 2014 to 2016 records, and testing data 2017 records. Altogether, the training data had 157k records and the testing data, 53k records.

We only selected optimal features for the model to achieve balance between model performance and computational effort. The feature selection was performed using RF and Recursive Feature Elimination (RFE). Utilising RF approach allowed for the features information gain to be isolated while taking into consideration the impact of multivariate nature of the problem. RFE was a wrapper type feature selection algorithm, by fitting a model using all features, and then ranking the features by importance. It recursively discarded the least important features and re-fitted the model. This continued until the optimal features were selected for the model. As such, we did not consider the correlation analysis of the data features.

Optimisation of model performance was implemented using Random Search hyperparameter tuning. This method required input for the tuneLength parameter, which defined the total number of parameter combinations to be evaluated. In this case, the parameter was set to ten. This method of tuning was selected over alternates such as Grid Search due to the computational challenges that came with exhaustive search. Studies such as [2] have also shown than Random Search is far more efficient that Grid Search for hyperparameter tuning.

4.2 Results

We based the benchmark for the performance of LOS regression on the study by Liu et al [9], given the similarities to our dataset. However, we were not able to obtain the benchmark dataset to replicate and compare the model result with our dataset. Both our dataset and the benchmark dataset consisted of administrative data, such as demographics and admission diagnosis. The similarities end there with the benchmark study using two additional features of Laboratory Acute Physiology Score (LAPS) and Comorbidity Point Score (COPS) derived from its dataset. Similarly, we adopted the CCI and Prior-Year LOS fea-tures. The benchmark employed a linear regression with an R2 of 0.124 and RMSE of 173.4.

Predicting LOS (Regression). Table 1 summarised the LOS prediction results using different regression models. The best regression model, XGBoost, significantly outperformed the benchmark in both RMSE and R2. Although the underlying datasets were different, the massive lift in model performance suggested the methodologies adopted in this study had the potential to improve the performance of models trained on similar datasets. The XGBoost model could predict the patients' LOS based on information available at admission. However, the model might not be strong enough for implementation, with the RMSE above the median LOS of the population at 3 days. As shown in the table, the SVM model performed far worse compared to the other models. Considering a big number of data that we used, the possible reason is that the SVM model was unable to find a clear decision boundary based on the provided features.

Table 1. Regression model performance for all diagnosis

Model	RMSE	R2
Baseline	8.2	N/A
Linear Regression	7.82	34.33%
SVM	399.7	8.87%
Random Forest	7.15	37.42%
XGBoost	6.98	38.24%
Liu et al. [9]	173.5	12.4%

Feature Importance. There is significant value in understanding the drivers of LOS, including the ability to generalise the insights. Our results showed that the optimal number of features based on collection of dataset are 175 and 18 for all diagnosis and heart & Circulatory diagnosis, respectively. In this study, the data features were categorised into three groups, including features that originated from the base dataset and features from secondary data sources. The key features for the models were summarised as follows:

- **Features originated from base data**: Among features from base data, attributes related to primary diagnosis are most important. Fields such as ccsdx (primary diagnosis) and MDC (major diagnostic category), ccspx (primary medical procedures), are in the top 10 most important features of most models.
- **Features derived from base data**: This study created over 1,000 new features from base data. The prior year median LOS features turn out to be very predictive of LOS, seemingly more predictive than the features they were created from. This reflects prior-year LOS could be a good indicator for future studies.
- **Features derived from secondary sources**: Although adding new features from secondary data sources has improved the model performance, not many features in this category are identified are as the top 10 features for regression or classification models.

5 Discussion

The initial motivation of the feature analysis was to assist hospitals in identifying key improvement areas in LOS management. As most fields in the base data

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_9

were out of control for hospitals, the key features identified might not provide good guidance for hospital LOS management in practice. Future research might need to analyse this dataset together with hospital operational data to provide actionable insights for hospitals. Interestingly, the median income of hospitals was predictive of LOS in the classification model. Typically, the literature used the median income based on patient location, rather than hospital location. Perhaps the socio-economic region that a hospital is based in can be used as a proxy in the absence of patient location.

One of the most successful experiments in this study was the creation of prior median LOS features, where new features were created by calculating the median LOS of the prior year for each attribute of the column. This approach applied the methodologies of time series analysis on based data generated over 1000 new features, over 200 of which were kept in the final model. These new features significantly improved the model performance and many of the new features were considered important for many models in the predictive features' analysis. This was a way to make better use of a dataset with time series information, not limited to a similar clinical dataset, for regression and even classification problems. It enabled studies to generate meaningful features without sourcing external data. The only drawback of the approach was the first-year records could not be used in model training and testing, which reduced the data size for model development.

Like many other LOS datasets used in prior studies, the Vermont dataset was highly skewed. This limited the performance of many models which assumed the training data was normally distributed. The Y variable log transformation was a good solution to this problem. Log transformation converted the dataset towards normal distribution and made the dataset better suited for model training. While this approach was simple, it could have been overlooked. This study reaffirmed the effectiveness of the method and would recommend it for consideration in future studies that utilises skewed data. Despite outperforming the benchmark, this prediction might not have been accurate enough to support a decision in practice. It should be highlighted that despite the significant enrichment made in this study, predicting the exact LOS continues to remain a challenge.

6 Conclusion and Future Work

We propose a methodological approach for predicting patients' LOS using a real-world hospital dataset. We empirically enriched the data by incorporating relevant features that contributed to improving LOS prediction. Our empirical experiments showed that our prediction approach outperformed the identified benchmarks that used regression models. Compared to the benchmark research, we introduced an empirical approach that uniquely improved the regression-based LOS prediction. Through the feature selection process, the optimal number of features was selected for each model type. This variance in the number of features highlighted that different models would require different levels of data and different covariates. While our approach seems to only be applicable to

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_9

the Vermont dataset, most electronic health records would contain similar data fields as our dataset. In our future works, replicating the prediction model using similarly typed electronic health records could provide a meaningful comparison of model accuracy. The findings above, therefore, add towards academic studies and medical research. The key features could provide research teams with possible directions in LOS reduction-related research.

Acknowledgements

We would like to thank the State of Vermont for providing us with the Vermont Uniform Hospital Discharge Dataset in agreement with its Public Use policy. We also thank Jirong Liu and Young Rang Choi for helping out with the visual data analysis needed for this work.

References

1. Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., Levin, S.: Real-time prediction of inpatient length of stay for discharge prioritization. Journal of the American Medical Informatics Association : JAMIA **23** (08 2015)
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. J. Mach. Learn. Res. **13**(null), 281–305 (Feb 2012)
3. Buttigieg, S.A., Abela, L., Pace, A.: Variables affecting hospital length of stay: a scoping review. Journal of Health Organization and Management **32** (04 2018)
4. Center, M.P.S.: Zip code characteristics: Mean and median household income (2020), `https://www.psc.isr.umich.edu/dis/census/Features/tract2zip/`
5. Charlson, M.: Charlson comorbidity index (cci) (2020), `https://www.mdcalc.com/charlson-comorbidity-index-cci`
6. Clarke, A.: Why are we trying to reduce length of stay? evaluation of the costs and benefits of reducing time in hospital must start from the objectives that govern change. BMJ Quality & Safety **5**(3), 172–179 (1996)
7. Daghistani, T., El Shawi, R., Sakr, S., Ahmed, A., Thwayee, A., Al-Mallah, M.: Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. International Journal of Cardiology **288** (08 2019)
8. Kelly, M., Sharp, L., Dwane, F., Kelleher, T., Comber, H.: Factors predicting hospital length-of-stay and readmission after colorectal resection. BMC health services research **12**, 77 (03 2012)
9. Liu, V., Kipnis, P., Gould, M.K., Escobar, G.J.: Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. Medical care pp. 739–744 (2010)
10. Masip, J.: The Relationship Between Age & Hospital Length of Stay: A Quantitative Correlational Study. Ph.D. thesis, University of Phoenix (2019)
11. Pendharkar, P., Khurana, H.: Machine learning techniques for predicting hospital length of stay in pennsylvania federal and specialty hospitals. International Journal of Computer Science and Applications, **11**, 45–56 (12 2014)
12. Turgeman, L., May, J., Sciulli, R.: Insights from a machine learning model for predicting the hospital length of stay (los) at the time of admission. Expert Systems with Applications **78** (02 2017)