**Predicting Inpatient Length of Stay in Hospitals**

Lidiia Tronina

City University of New York School of Professional Studies

DATA 698 - Capstone

November 30, 2022

**Table of Contents**

Introduction 3

Literature Review 4

Hypothesis/Research question 7

Data and Variables 8

Statistical Methods 9

Data Preparation 10

Model Building 13

Results 13

Discussion 14

Conclusion 15

References 16

## Introduction

Hospital stays in the U.S. cost insurance companies at least $377.5 billion per year and recent Medicare legislation standardizes payments for the procedure performed, regardless of the number of days a patient spends in the hospital [3]. This motivates hospitals to evaluate patients carefully at the time of admission and determine whether they are at risk of having a prolonged stay. Once identified, patients with high length of stay (LOS) risk can have their treatment plan optimized to minimize LOS and lower the chance of contracting an additional illness while hospitalized. Another benefit is that prior knowledge of LOS can aid in logistics, such as room and bed allocation planning. From an operational perspective, LOS is a meaningful indicator of costs, resource consumption and patient flows.

The average LOS for a patient in the U.S. is 4.5 days [7]. Many hospitals are aware that LOS is an important metric to keep track of, and that it’s better to strive for shorter LOS where possible.

LOS also has a significant impact on patient experience. Patients don’t want to stay at the hospital longer than necessary. Though their hospital stay sets them up for a successful recovery, they are often eager to complete their recovery at home in a comfortable environment.

Longer lengths of stay also negatively impact hospitals. They increase costs and are often linked to inefficiency—indicating that processes may need to be revisited. Likewise, LOS directly impacts bed management, which lowers turnover and decreases revenue. It also means that hospitals may not be able to meet the needs of their patients; when a patient is kept in a bed longer than necessary, it may mean that the bed is not available for another patient who needs it more.

Increasing healthcare costs and availability of a large volume of medical data motivate the search for ways to increase the care efficiency. The LOS of hospitalized patients is considered as an important factor for healthcare policy planning as it has a major impact on technical and financial resources as well as facilities occupation. It also indicates the quality of hospital service. The objective of this work is to build a model that can assist the physicians to advice on the inpatient hospital LOS using predictive modelling.

While length of hospital stay generally depends on many factors, the author seeks to quantify the precise relationship and form insights from the most important available variables at time of admission that translate to action, and explore whether this model can help find opportunities for optimization. The goal of this research is to create a model that predicts the LOS for each patient at time of admission.

## Literature Review

Length of stay prediction has received substantial interest in academic literature. Much historical work has sought to better understand the factors that predict the length of hospital stays and used age, gender, diagnosis, and emergency room activity.

Chrusciel and Girardon [2] offer a LOS prediction using two random forest models. The first included unstructured text extracted from electronic health records. The second model was primarily based on structured data in the form of diagnoses coded and triage codes. Variables common to both models were: age, gender, zip/postal code, LOS in the Emergency Department (ED), recent visit flags, assigned patient ward after the ED stay and short-term ED activity. The authors decided to exclude short hospital stays less than 2 days. LOS was considered as a categorical variable to maximize the power of the model, however the authors agree that to improve research in this area, future studies should consider a study of LOS as a continuous variable. Models were trained on 80% of data and performance was evaluated for accuracy on the remaining 20% test data. The model using unstructured data had a 75% accuracy compared to 74% for the model containing structured data. The two models produced a similar prediction in 86% of cases. In a secondary analysis restricted to intensive care patients, the accuracy of both models was also similar (76% vs 75%). However their model evaluated patients who were admitted through an Emergency room only and had a limited number of cases - 5,006.

Tien [6] expanded the dataset and included the publicly available “NY Hospital Inpatient Discharges” dataset from New York State Government health data website. This dataset contains 2.3+ million rows of patient data, including information such as patient demographics, diagnosis, type of admission, severity of illness, and type of insurance. In this study LOS was also treated as a classification problem, instead of regression, and was grouped into the following bins: 1–5 days, 6–10 days, 11–20 days, 21–30 days, 31–50 days, and 50–120+ days*.* The author was able to predict a patient’s LOS only using data available from the moment they entered the hospital and were diagnosed with an accuracy of 70% using Boosted Decision Trees. APR DRG Codes and APR Severity of Illness Codes were the two most important features in predicting a patient’s length of stay. In addition, the payment typology Medicare subgroup has a relatively high importance in predicting LOS in comparison to the other payment typology groups. Even though this study is similar to ours, it only utilizes publicly available data for one year - 2015. The author mentioned the COVID-19 pandemic in his study, but the data is not going to reflect that. It also won’t get a year over year progression.

Cummings [3] also works on the same goal - to create a model that predicts the LOS for each patient at time of admission. However, he decided to work on LOS as a continuous variable and make it a regression, as a primary prediction technique. The author used MIT’s MIMIC-III database. It's an openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with 58,967 admission events. It includes demographics, type of admission, insurance category, vital signs, laboratory tests, medications, and more. There are a few extra data preparation steps taken in this research that are worth mentioning. First, the author excludes admissions that resulted in death, since they jeopardize an accurate length of stay prediction. Also, the author created age groups. Fig. 1 highlights the MIMIC groups of newborns and >89 year olds, where there is an increasing number of admissions in the age groups from 20 to 80 years old. Because of the discrete-like distribution of data on the extremes of age, the author decided to convert all ages into the categories of newborn, young adult, middle age, and senior.

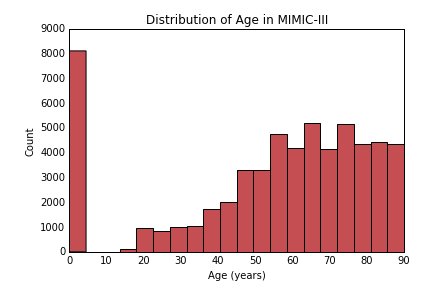


Fig. 1 Distribution of Age in MIMIC -III dataset [3]

Cummings also attempts to solve another challenge in terms of feature engineering - a great variety of Diagnosis codes. There were 6,984 unique codes used in the MIMIC dataset and 651,047 ICD-9 diagnoses given to patients since most were diagnosed with more than one condition. The author sorts all the unique codes per admission into these categories as shown in Fig. 2.

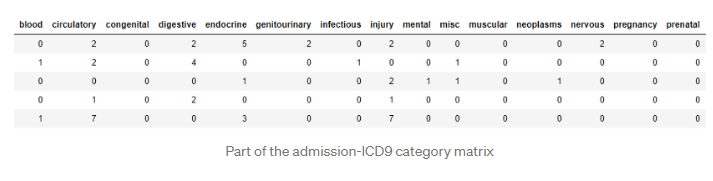


Fig. 2 Part of the admissions ICD9 category matrix in Cummings Study [3]

The highest R2 score of 0.39 was from the Gradient Boosting Regressor model. The author finds that diagnoses related to prenatal issues have the highest feature importance coefficient followed by respiratory and injury. In fact, in the 20 top features, only emergency admission type, gender, and Medicaid insurance showed any importance outside of diagnosis groups.

Another study [4] that also treats LOS as a continuous variable has an R2 score of 0.26. However, it was found that the accuracy of the classification model predicting long term admissions (>30 days) is 0.9732.

While LOS has been previously modeled, there has not been any analysis on year over year changes, especially since COVID, and its relationship to change in LOS. This study will use 8 years (2015-2022) of data with >900,000 admissions.

While there is a high accuracy in predicting LOS as a categorical variable, there is room for improvement for prediction as a continuous variable.

## Hypothesis/Research question

The hypothesis is that the length of stay predictions can be made based on patient demographics, diagnosis, condition severity, insurance, and type of admission. This analysis focuses on predicting length of stay in the hospital at time of admission as a continuous variable, hence, the more accurate predictions compare to other researchers in this area. The author also seeks to extend the body of research by focusing on improving the feature selection and engineering and improving the accuracy of LOS predictions as a continuous variable.

Ultimately the author seeks to answer whether the length of stay in hospital can be predicted at time of admission and gain knowledge on what contributes to longer stays in the hospital. Factors to be considered are: whether there is a negative dynamic in any specific hospital over a number of years; whether the type of insurance a patient has affects LOS; and whether there is a difference in LOS before and after the COVID-19 pandemic.

## Data and Variables

The data for this study will be collected from encounter data provided by Mount Sinai Hospital from their 7 New York hospitals. Mount Sinai is one of the oldest and largest teaching hospitals in the country, with thousands of admissions per year that can inform our research. The admissions data were collected continuously from 2015 to the present (2022). It’s estimated that there were over 1M admissions during that period of time. All admissions have been de-identified according to HIPPA privacy rules.

The following fields will be taken from the original dataset:

|  |  |
| --- | --- |
| Facility | Identifies the hospital where the patient was admitted |
| Year | Year when patient was admitted to inpatient care |
| Month | Month when patient was admitted to inpatient care |
| Age | Patient’s Age |
| Race | Patient's Race |
| Gender | Patient's Gender |
| Type of insurance | Groupings of like payers into high-level categories, i.e. Commercial, Medicare HMO, Medicare, Medicaid etc. |
| Type of Admission | Distinguishes whether the admission was emergent, elective, urgent, traumatic, or newborn. |
| DRG | MS DRG (Medicare Severity Diagnosis Related Group) Code. Groups patients into categories based on clinical and resource consumption similarities. |
| DRG Severity | There are 4 possible options - 1: Minor, 2: Moderate, 3: Major and 4: Extreme. |
| Diagnosis | Identifies the diagnosis that led to the patient being admitted for inpatient treatment. |
| Readmission Flag | Qualifying inpatient encounters within 30-days prior to the current admission will have a value of Y. |
| Risk of Mortality Score | Premier-defined expected mortality for the admission. |
| Risk of Mortality Categorical | Categorical variable that measures relative risk of death during admission within a given DRG. There are 4 possible options - 1: Minor, 2: Moderate, 3: Major and 4: Extreme. |

Furthermore, the following parameters will be added to the dataset:

* Age Group

About 10% of all inpatient admissions in the Mount Sinai data are newborns with age ‘0’. It makes sense to group age by category.

* COVID-19 diagnosis flag.
* Post COVID admission flag. Patient admitted after 03/01/2020
* Emergency Department Admission Indicator

It is also important to exclude admissions that resulted in death since they can affect the accuracy of the model.

## Statistical Methods

Since the length of stay is the count variable, this will be a regression problem. The three specific methods are described below at a high level.

The Linear Regression (LR) is the baseline model in this research. The objective of ordinary least squares regression is to find the plane that minimizes the sum-squared error between the observed and predicted responses. In order to avoid including correlated predictors and exclude predictors with low significance, the level backward elimination technique will be performed.

A popular alternative to standard Linear regression is the use of penalized models, such as lasso and ridge regression. This extremely fast method should work well for a dataset this size.

Another method that is widely used for count outcomes is Poisson regression. It is a generalized linear model form of regression analysis used to model count data and contingency tables. It should be more suitable for our case because the response is the number of days.

There is a high computation cost during runtime if sample size is large for non-linear models. K-Nearest Neighbors (KNN) is the only non-linear model that will be used for this dataset. KNN is a non-parametric model, whereas LR is a parametric model. KNN mainly involves two hyperparameters, K value & distance function. If the training data is much larger than the number of features, KNN is better than support vector machines.

## Data Preparation

There are no missing values in the data set. However, there were some discrepancies, such as invalid DRG codes or age. To prepare the data, encounters with ages less than 0 or larger than 113 were excluded, and encounters with DRGs such as 'XXX', '999', '998', '989', '988', and '987' were also eliminated. The DRG field had 786 distinct DRGs even after exclusions. After some investigation, a service group that is based on DRG was used. It groups 786 DRGs into 34 groups.

Due to a high number of admissions that have age 0, which is a newborn encounter, and increasing number of admissions moving from the 20s toward the 80s, age was grouped into the following categories: Newborns (age 0), Kids (1-12), Teens (13-18), Youth (19-24), Young Adult (25-34), Middle Age (35-54), Boomers I (55-64), Boomers II (65-74), Seniors (75 <).

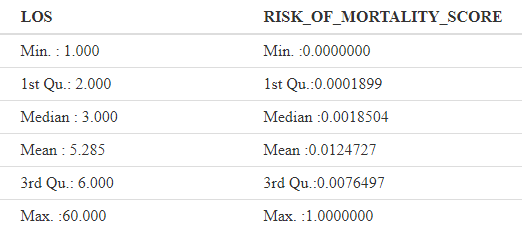
There are also some unidentified values in fields such as gender and race. Encounters with unknown genders were excluded, and unknown race was updated to 'Other'.

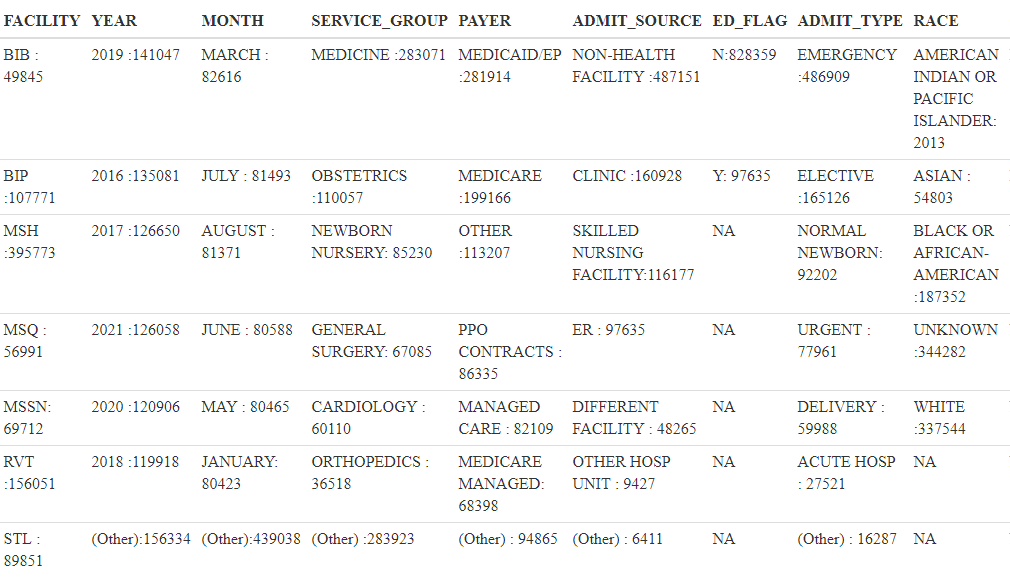
Additionally, three flags were created. The first flag indicates if admission took place during COVID; that is, the patient was admitted after 03/01/2020. The second flag indicates that the patient has COVID based on the diagnosis code. The Emergency Department Admission flag was created based on the admission source; it highlights transfers from the Emergency room.

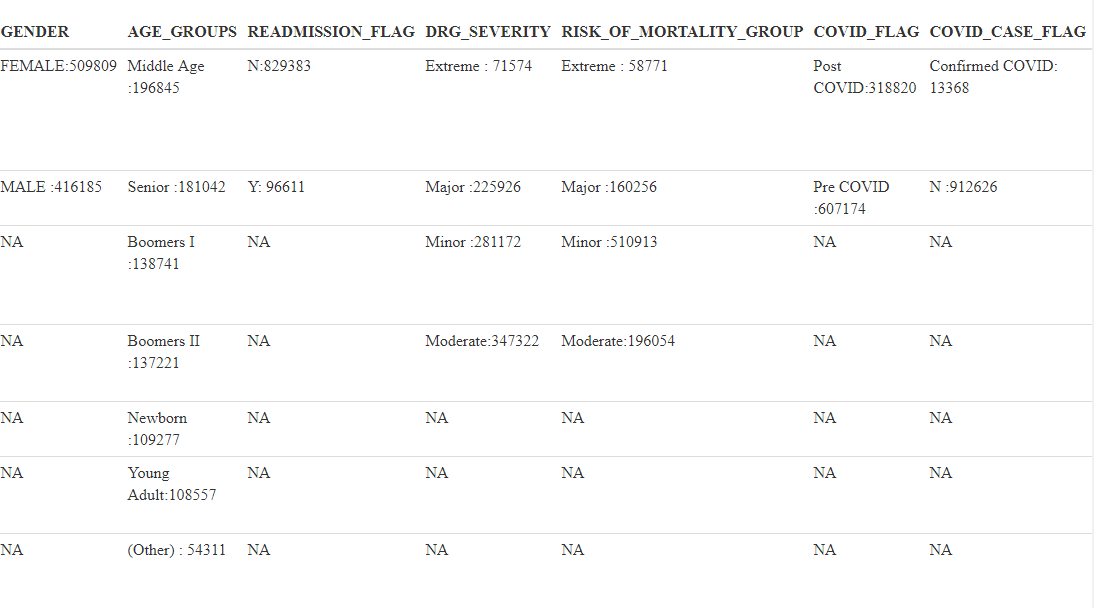
Furthermore, some exclusions have to be made due to data that could affect model accuracy. Admissions that resulted in death and discharges to hospice were excluded since they can affect the length of stay. Admissions with more than 60 days’ stay were also excluded. These are very rare cases that might affect model performance due to unusual lengths of stay.

Data for the KNN model had to be transformed into a numeric format. Categorical predictors were converted into separate columns with numerical data. Predictors with high collinearity (higher than 0.5) were excluded.

The final data set included information from 914,645 hospital admissions. Once the data was in a model-ready format, it was split into a training data set and a testing data set with a train-test split of 70% training and 30% testing.







## Model Building

The following models were trained and then validated on the test data, and their accuracy scores computed:

* Linear Regression with backward elimination
* Penalized Regression
* Poisson Regression
* K-nearest Neighbor Regression

## Results

The accuracy scores of each model are summarized in the next table. They were evaluated using The Mean Squared Error, Mean absolute error, Root Mean Squared Error, and R-Squared.

* The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. It measures the average of the residuals in the dataset.
* Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.
* Root Mean Squared Error is the square root of Mean Squared error. It measures the standard deviation of residuals.
* The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model.

The lower value of MAE, MSE, and RMSE implies that Linear regression has the highest accuracy when compared to the other models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Name | R-squared | MSPE | MAE | RMSE |
| Linear Regression | 0.35 | 24.44409 | 2.821081 | 4.944096 |
| Penalized Regression | 0.35 | 24.44666 | 2.819887 | 4.944357 |
| Poisson Regression | - | 48.56436 | 3.853134 | 6.968813 |
| K-nearest Neighbor Regression | - | 28.43573 | 2.963692 | 5.332517 |

## Discussion

Unexpectedly, the best-performing model on the test dataset was the Linear regression model with stepwise regression. Penalized regression had similar results. Poisson regression underperformance might be due to high outcome counts. Reduction of predictors and transformation of the dataset might negatively affect the performance of the KNN model. KNN was also the slowest model as far as computation. The nature of the data (most of the variables are categorical) and the size didn’t allow us to use other techniques.

## Conclusion

There are many factors that contribute to patients’ length of stay in the hospital. The Mount Sinai dataset offered a very large dataset of patients but not necessarily a model-ready format data. Even though the choice of models was limited, we were able to build a model with R2, similar to previous researchers in this area. Using machine learning there was a high accuracy of predictions for LOS as a categorical value. However, there are challenges for predicting exact length of stay, which was seen in high error measurements on the test dataset.

However, using this model we can gain knowledge on factors contributing to longer stays in hospital. There was a negative dynamic in specific hospitals - Brooklyn and South Nassau. Another interesting observation was that Patients with Medicare stay longer than Self-Paying patients, which is logical but still lends context to the data. The COVID era didn’t affect LOS, but if a patient had been diagnosed with COVID, they spent an average of three additional days in the hospital.

As data in the healthcare sector improves, there is a hope for higher prediction accuracy of these models. The development of machine learning in the healthcare sector would be beneficial not only to hospitals, but also to patients and the economy in general.

## References

[1] Barnes, S., Hamrock, E., Toerper, M., Siddiqui, S., & Levin, S. (2015). Real-time prediction of inpatient length of stay for discharge prioritization. *Journal of the American Medical Informatics Association*, *23*(e1), e2–e10. https://doi.org/10.1093/jamia/ocv106

[2] Chrusciel, J., Girardon, F., Roquette, L., Laplanche, D., Duclos, A., & Sanchez, S. (2021). The prediction of hospital length of stay using unstructured data. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01722-4

[3] Cummings, D. (2020, May 11). *Predicting hospital length-of-stay at time of admission*. Medium. https://towardsdatascience.com/predicting-hospital-length-of-stay-at-time-of-admission-55dfdfe69598

[4] Baek, H., Cho, M., Kim, S., Hwang, H., Song, M., & Yoo, S. (2018). Analysis of length of hospital stay using electronic health records: A statistical and data mining approach. *PLOS ONE*, *13*(4), e0195901. https://doi.org/10.1371/journal.pone.0195901

[5] Varghese, D. (2019, May 10). *Comparative study on Classic Machine learning Algorithms*. Medium. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms-24f9ff6ab222

[6] Tien, V. (2020, June 4). *Predicting Inpatient Length of Stay at Hospitals Using Python + Big Data*. Medium. https://towardsdatascience.com/predicting-inpatient-length-of-stay-at-hospitals-using-python-big-data-304e79d8c008

[7] Clarkson, K. (n.d.). *Why Reducing Length of Stay Is Critical for Patient & Hospital Wellbeing*. Www.pulsara.com. https://www.pulsara.com/blog/every-minute-counts-how-increased-length-of-stay-impacts-patients-and-hospitals

[8] Varghese, D. (2018, December 11). *Comparative Study on Classic Machine learning Algorithms, Part-2*. Medium. https://medium.com/@dannymvarghese/comparative-study-on-classic-machine-learning-algorithms-part-2-5ab58b683ec0

GENERAL COMMENTS

Lidiia,

Excellent and concise exploratory study on factors influencing hospital patient length of stay.

My only nit is your in-text reference citations. You have adopted the numbering citation AMA style for the body text, but you’ve listed the references in non-sequential order (first 3, then 7, then 2…).

So if you intent to submit this work, you should fix these. Here’s a guide that might help: <https://owl.purdue.edu/owl/research_and_citation/ama_style/>

Given the quality of the study, I’ve approved your final draft and posted it in the “Final Approved Final Drafts” folder in the “Feedback on Drafts” section on the course site.

Congrats, and I look forward to seeing your recorded presentation.

A