International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

Enhancing Healthcare Operations with Predictive Length of Stay Models

Umamaheswara Reddy Kudumula

Engineer Lead, EDA- Provider, Employer and Financial Analytic Solutions, Anthem, Inc, Georgia, Virginia, United States

Abstract: ***Healthcare costs are a significant concern in the United States, posing a substantial burden on American families and leading many to bankruptcy. Hospitalization stands out as a major contributor to these costs. According to the American Hospital Association, there were over 36 million hospital admissions in 2018, with an average length of stay (LOS) of approximately 5.5 days—an increase of over 19% from previous years. The average daily hospitalization cost is estimated at $2,883, resulting in nearly $13,000 per stay. If hospitalization involves surgery, costs can range from $5,000 to $150,000, with potential for exceedingly high out-of-pocket expenses. Consequently, healthcare costs are a primary reason for bankruptcy filings. The correlation between healthcare costs and the length of hospital stays is clear: longer stays result in higher costs. Excessive hospital days contribute to operational inefficiencies and negatively impact patient outcomes. Therefore, reducing LOS is crucial for benefiting both patients and healthcare providers. This white paper examines the use of predictive data analytics to forecast average length of stay (ALOS). By leveraging these predictions, healthcare systems can develop strategies to enhance management efficiency, optimizing the use of healthcare workers and resources like beds. Ultimately, this approach aims to improve the quality of care, reduce LOS, and lower associated healthcare costs.***

Keywords: Healthcare, Length of Stay, American Hospital Association (AHA), Data Analytics, Average Length of stay (ALOS), Patient Care, Bankruptcy, Hospitalization, Providers, Medicare Severity Diagnosis-Related Group (MS DRG)

1. Introduction

Length of Stay (LOS) measures the duration a patient spends in the hospital from admission to discharge. Average Length of Stay (ALOS) is a key performance indicator in the healthcare industry, assessing efficiency, healthcare costs, and quality of care. Extended stays contribute to inefficiencies, higher out-of-pocket payments, and increased risk of hospital-acquired infections, while shorter stays enhance facility efficiency, patient experience, and outcomes, and reduce costs.

LOS is also vital for provider reimbursement. Providers often receive incentives from insurance companies for reduced LOS. In Medicaid and Medicare programs, hospitals are compensated based on the Medicare Severity Diagnosis-Related Group (MS-DRG) rather than LOS. Accurate LOS predictions are critical for healthcare, allowing providers to better plan resources, improve operational efficiency, patient care, and reduce costs.

This paper explores the use of predictive analytics to forecast LOS for inpatient admissions. By leveraging predictive models, healthcare providers can develop strategies to enhance operational efficiencies, allocate resources effectively, and improve patient care and outcomes.

2. Solution

We will employ predictive data modeling techniques to forecast the Length of Stay (LOS). Predictive data modeling is a statistical technique that utilizes data mining and machine learning to predict future outcomes. Specifically, we will leverage the Random Forest model to predict LOS based on multiple explanatory variables.

Random Forest is a supervised learning algorithm that [1] addresses both classification and regression problems. It constructs multiple decision trees and makes predictions by aggregating the results from each tree through a voting mechanism. In our case, we will use a Random Forest Classifier model to predict LOS. This method enables the development of a model that healthcare providers can use to enhance operational efficiencies.

To build this model, we will use data from the Hospital Length of Stay Dataset by Microsoft, which includes patient demographics and medical information. The process involves the following steps:

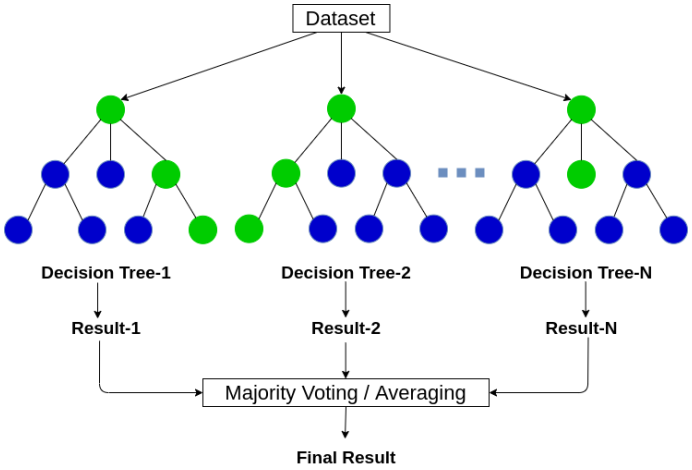

Figure 1: Random Forest Prediction [4]

1) Data Collection: Identify and collect the required data from various sources.
2) Exploratory Data Analysis (EDA): Analyze and explore the datasets, summarizing the main characteristics using statistical graphs or visualizations.
3) Data Splitting and Feature Selection: Split the data into training and testing sets. Select feature variables that impact model performance.
4) Model Fitting and Evaluation: Evaluate the relationship between dependent and independent variables, and fit the model to predict LOS.

By utilizing predictive modeling techniques and the Random Forest algorithm, we aim to create a robust model that healthcare providers can use to improve operational efficiencies and patient care.

3. Data Collection

This data collection contains demographics and medical information of the patient. We can find several variables from the data set in the (.csv) file; some are independent, and there is only one target dependent variable. Below is the information on the dataset attributes [3]

- eid: Unique ID of the hospital admission.
- vdate: Visit date.
- rcount: Number of readmissions within the last 180 days.
- gender: Gender of the patient - M or F.
- dialysisrenalendstage: Flag for renal disease during the encounter.
- asthma: Flag for asthma during the encounter.
- irondef: Flag for iron deficiency during the encounter.
- pneum: Flag for pneumonia during the encounter.
- substancedependence: Flag for substance dependence during the the encounter.
- psychologicaldisordermajor: Flag for the major psychological disorder during the encounter.
- depress: Flag for depression during the encounter.
- psychother: Flag for other psychological disorders during the encounter.
- fibrosisandother: Flag for fibrosis during the encounter.
- malnutrition: Flag for malnutrituion during the encounter.
- hemo: Flag for blood disorder during the encounter.
- hematocrit: Average hematocrit value during encounter (g/dL).
- neutrophils: Average neutrophils value during encounter (cells/microL).
- sodium: Average sodium value during encounter (mmol/L).
- glucose: Average sodium value during encounter (mmol/L).
- bloodureanitro: Average blood urea nitrogen value during the encounter (mg/dL).
- creatinine: Average creatinine value during encounter (mg/dL).
- bmi: Average BMI during the encounter (kg/m2).
- pulse: Average pulse during the encounter (beats/m).
- respiration: Average respiration during the encounter (breaths/m).
- secondarydiagnosisnonicd9: Flag for whether a non ICD 9 formatted diagnosis was coded as a secondary diagnosis.
- discharged: Date of discharge.
- facid: Facility ID at which the encounter occurred.
- length of Stay: Length of Stay for the [2] encounter.
- losgroup: Discrete values representing groups. 1-4 days as 1; 5-8 days as 2; 9-12 days as 3; 13-16 days as 4 ; >17 days as 5

4. Exploratory Data Analysis

Import the required Python libraries. Below are some of the libraries we would be using in our model.

- Pandas: Software Library used for data analysis.
- NumPy: The library is used to work with large multidimensional arrays.
- Sklearn: Machine Learning Library featuring various algorithms.
- Matplotlib: Library used for creating visualizations.
- Seaborn: The library is based on Matplotlib and is used to create advanced visualizations.
- Spicy: The library is used for statistical and probabilistic analysis.
- Sklearn.model_selection: Library used to split the dataset into test and train datasets
- Sklearn.ensemble: The library includes two averaging algorithms based on the random decision trees; random forest algorithm and extra tree methods.

Import the Length of Stay data set in csv format, to the data frame using Panda's library.

```
In [1]: import numpy as np
        import pandas as pd

In [2]: LOS_dataset=pd.read_csv("..../LengthOfStay.csv")
```

Exploratory data analysis is a critical step in [3] predictive model building. This involves using various functions to identify missing values, default values, outliers, errors, inconsistencies, and inaccuracies. We will use various statistical summaries to gain insights into the data. Below are some.

- Describe (): Generate descriptive statistics for the dataset.
- Isnull(): Verify if there are any null values in the data set.
- shape: Returns tuple value
- head (): returns the top 5 records in the data set
- value_counts(): Returns the count of unique values in the dataset

```
In [4]: LOS_dataset.shape
Out[4]: (100000, 29)

In [5]: LOS_dataset.head()
```

Out[5]:

	eid	vdate	rcount	gender	dialysisrenalendstage	asthma	irondef	pneum	substancedependence	psychologicaldisordermajor	...	bloodureanitro	c
0	1	8/29/2012	0	F	0	0	0	0	0	0	...	12.0	
1	2	5/26/2012	5+	F	0	0	0	0	0	0	...	8.0	
2	3	9/22/2012	1	F	0	0	0	0	0	0	...	12.0	
3	4	8/9/2012	0	F	0	0	0	0	0	0	...	12.0	
4	5	12/20/2012	0	F	0	0	0	1	0	1	...	11.5	

5 rows × 29 columns

```
In [6]: LOS_dataset.describe()
```

Out[6]:

dium	glucose	bloodureanitro	creatinine	bmi	pulse	respiration	secondarydiagnosisnonicd9	lengthofstay	losgroup
0000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.00000	100000.000000
1397	141.963384	14.097185	1.099350	29.805759	73.444720	6.493768	2.123310	4.00103	1.423550
9669	29.992996	12.952454	0.200262	2.003769	11.644555	0.568473	2.050641	2.36031	0.572659
2632	-1.005927	1.000000	0.219770	21.992683	21.000000	0.200000	0.000000	1.00000	1.000000
1062	121.682383	11.000000	0.964720	28.454235	66.000000	6.500000	1.000000	2.00000	1.000000
7151	142.088545	12.000000	1.098764	29.807516	73.000000	6.500000	1.000000	4.00000	1.000000
2885	162.180996	14.000000	1.234867	31.156885	81.000000	6.500000	3.000000	6.00000	2.000000
7283	271.444277	682.500000	2.035202	38.935293	130.000000	10.000000	10.000000	17.00000	5.000000

```
In [7]: LOS_dataset['losgroup'].value_counts()
Out[7]: 1    61694
        2    34393
        3     3781
        4      128
        5        4
        Name: losgroup, dtype: int64
```

A. Splitting Data and Feature Selection:

We will be creating the target and independent variables. The target variable will have a discrete value representing the group. We will be dropping the Length of Stay, losgroup, discharged, dvate, eid, facid, gender, rcount columns as they do not correlate much with the Length of Stay.

```
In [9]: y=LOS_dataset['losgroup']

In [83]: X=LOS_dataset.drop(["lengthofstay","losgroup", "discharged", "vdate", "eid","rcount","gender","facid"], axis=1)
```

The next step in building the model is splitting the data into training and testing sets.

```
In [103]: from sklearn.model_selection import train_test_split
          X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 10)
          X_train.shape
Out[103]: (80000, 21)
```

Splitting the data sets into test and train datasets will help assess the model's performance. Train data set is used to train the model, while test data sets are used to evaluate the model.

B. *Model Fitting & Evaluation:*

Random forests are more accurate than decision trees as they reduce overfitting by combining multiple decision trees. This makes Random Forest models less sensitive to outliers and improves predictive performance. We will use the `RandomForestClassifier` to build and train our model. The process includes data collection, exploratory data analysis, data preprocessing, and splitting data into training and testing sets. By leveraging this method, we aim to create a robust

predictive model for Length of Stay, enhancing operational efficiencies and patient care.

```
In [107]: from sklearn.ensemble import RandomForestClassifier
          rfc = RandomForestClassifier(n_estimators=10, random_state=50)
          rfc.fit(X_train, y_train)
          print("Training Score: ", rfc.score(X_train, y_train))
          print("Testing Score: ", rfc.score(X_test, y_test))

          Training Score:  0.9770375
          Testing Score:  0.6651
```

The next step of the process is to fit the model and validate it against the test dataset. Also, check the accuracy of the model.

```
In [108]: y_pred = rfc.predict(X_test)
          y_pred

Out[108]: array([1, 1, 1, ..., 1, 1, 1], dtype=int64)

In [109]: from sklearn.metrics import accuracy_score

          accuracy_score(y_test, y_pred)

Out[109]: 0.6651
```

This model predicts the Length of Stay with discrete values with an accuracy of 66.51%.

We would increase the number of decision trees in random forests to improve accuracy. We will test the model with 100, 200, and 500 decision trees and verify the accuracy.

First run with n_estimators set to 100, which is 100 decision trees.

```
In [110]: from sklearn.ensemble import RandomForestClassifier
          rfc = RandomForestClassifier(n_estimators=100, random_state=50)
          rfc.fit(X_train, y_train)
          y_pred = rfc.predict(X_test)
          accuracy_score(y_test, y_pred)

Out[110]: 0.6903
```

First run with n_estimators set to 200, which is 200 decision trees.

```
In [111]: from sklearn.ensemble import RandomForestClassifier
          rfc = RandomForestClassifier(n_estimators=200, random_state=50)
          rfc.fit(X_train, y_train)
          y_pred = rfc.predict(X_test)
          accuracy_score(y_test, y_pred)

Out[111]: 0.69185
```

First run with n_estimators set to 500, which is 500 decision trees.

```
In [112]: from sklearn.ensemble import RandomForestClassifier
          rfc = RandomForestClassifier(n_estimators=500, random_state=50)
          rfc.fit(X_train, y_train)
          y_pred = rfc.predict(X_test)
          accuracy_score(y_test, y_pred)

Out[112]: 0.6932
```

The accuracy of Random Forest models increases with the number of decision trees. Our Random Forest model can predict the Length of Stay with nearly 70% accuracy. Evaluating the model's performance is done through a classification report, which includes precision, recall, F1 score, and support score. Here are key metrics:

- False Positive: The prediction is positive, but the actual case is negative.
- False Negative: The prediction is negative, but the actual case is positive.
- True Positive: Both the prediction and the actual case are positive.
- True Negative: Both the prediction and the actual case are negative.

Precision: The percentage of correct positive predictions.

Recall: The percentage of actual positive cases correctly identified.
F1 Score: The harmonic mean of precision and recall.
Support Score: The actual number of cases in the dataset.

These metrics help assess the model's accuracy and reliability in predicting Length of Stay.

5. Applications of the Solution in Various Organizational Processes

Random Forest, a powerful predictive data analytics model, has broad applications across various organizations. Below are some key use cases:

a) Assess Credit Risks in the Finance Industry

Credit Risk Modeling is vital for financial institutions' risk management. With over 400,000 bankruptcy filings in 2021, Random Forest models can help build credit risk models, providing lenders with insights to make informed decisions, lower risks, and increase profitability. Banks can optimize lending strategies involving loan terms, interest rates, and pricing structures based on these predictions.

b) Reduce Customer Churn in E-commerce

Acquiring a new customer costs significantly more than retaining an existing one, especially in e-commerce. Companies lose over $100 billion annually due to avoidable customer churn. According to Forrester, acquiring a new customer costs five times more than retaining one. Random Forest models can predict customer churn, enabling businesses to develop retention strategies and minimize losses.

c) Forecast Hotel Demands in the Hospitality Industry

Forecasting hotel demands is crucial for optimizing resources, occupancy, and revenue. Without accurate demand predictions, hotels risk overstaffing and lower prices, leading to significant losses. Random Forest models can predict hotel demands, helping hotels develop strategies to run promotions, adjust prices, and manage inventory and staff, thereby increasing profits.

6. Benefits of the Solution

Implementing predictive Length of Stay (LOS) models offers several benefits to the healthcare industry globally:

a) Optimize Resource Utilization:

LOS predictions enable hospitals to efficiently plan for staffing, manage bed capacity, and optimize patient flow. This data helps in scheduling medical staff, prioritizing patients with longer predicted stays, and allocating resources like X-ray machines, lab equipment, surgical instruments, medications, and administrative staff. Improved resource utilization enhances revenue and quality of care.

b) Enhance Patient Outcomes:

Predicting LOS helps hospitals identify patients at high risk for complications and longer stays, allowing for preventive measures. It aids in patient flow management, reducing wait times, and significantly impacting patient outcomes. Additionally, it supports discharge planning, ensuring smooth transitions to outpatient facilities, other provider facilities, or home care.

c) Improve Patient Satisfaction:

Accurate LOS predictions contribute to better patient outcomes and satisfaction by prioritizing care for high-risk patients, reducing wait times, and ensuring efficient discharge processes. This holistic approach leads to improved patient experiences and overall satisfaction with healthcare services.

d) Reduce Healthcare Costs:

Predicting LOS helps hospitals prioritize patients with higher complications, reducing extended stays and associated costs. Even a one-day reduction in LOS can save patients approximately $13,000 in healthcare expenses. Hospitals also benefit from reduced resource and staffing costs, leading to overall healthcare cost savings.

e) Prevent Hospital-Acquired Infections:

Extended LOS increases the risk of hospital-acquired infections, negatively impacting patient outcomes and further extending stays. According to the CDC, one in 31 US patients and one in 43 nursing home residents [4] contract healthcare-associated infections. Predicting LOS helps hospitals mitigate these risks, improving patient safety and outcomes.

7. Conclusion

Predictive data analytics can significantly enhance healthcare efficiency and reduce costs. By accurately predicting Length of Stay (LOS), healthcare providers can optimize operational efficiency, lower expenses, improve patient care quality, and boost patient satisfaction. This white paper presents a technical perspective on the critical role of data analytics in addressing [5] healthcare system challenges, emphasizing the importance of the LOS metric. It offers guidance on implementing data-driven solutions to transform healthcare delivery, showcasing the potential for predictive analytics to revolutionize healthcare systems worldwide.

References

[1] Ayub, S. (2021). Neutron-Induced Nuclear Cross-Sections Study for Plasma Facing Materials via Machine Learning: Molybdenum Isotopes. Applied Sciences, 11(16), 7359.
[2] Admissions Requirements – ARFC. https://aidsresource.org/dcp-p-2/admissions-requirements/
[3] Data Analysis with Python: Techniques and Libraries for Effective Data Exploration | Shriyash Shukla. https://shukla.com/posts/DAWP/
[4] CDC: HAIs Continued to Increase in 2021 | Ultra Clean. https://ultracleansystems.com/cdc-hais-continued-to-increase-in-2021/
[5] CDC: HAIs Continued to Increase in 2021 | Ultra Clean. https://ultracleansystems.com/cdc-hais-continued-to-increase-in-2021/