Transforming Healthcare Efficiency With Length Of Stay Predictions

Bharath Srinivasaiah
Engineer Lead Sr, Elevance Health Inc
Richmond, Virginia, USA.
Email: bharathsrinivasaiah@gmail.com

ABSTRACT:

Healthcare Costs are one of the biggest problems in the United States of America. It has become a burden on American families and one of the primary reasons for them to go bankrupt. Hospitalization is one of the top contributors to healthcare costs. Based on the latest data shared by the American Hospital Association, AHA, there have been more than 36 million hospital admissions in the year 2018[1]. The average Length of Stay in the United States is around 5.5 days, which has seen an increase of more than 19% from previous years [1]. The average per day hospitalization costs are estimated to be around $2,883, with 4.5 average lengths of Stay – totaling up to nearly 13,000 per Stay [3]. If the hospitalization involves surgery, the average cost could range between $5,000 to $150,000, and the out-of-pocket payment could go exponentially high. This is why Healthcare costs are among the top reasons for filing bankruptcy.

Healthcare costs are directly correlated to the Length of stays in hospitals. The longer the Length of Stay, the higher the healthcare costs. Unnecessary hospital days could lead to operational inefficiency and negatively impact patient outcomes. Hence, it is critical to reduce the Length of stays in the hospital, which will benefit both patients' and healthcare providers' systems. In this white paper, we will explore utilizing predictive data analytics in predicting the Average Length of Stay, ALOS, which will help the healthcare provider systems develop strategies to improve the efficiencies in management, not limited to healthcare workers but also resources like beds. This will help improve delivery quality and patient care and reduce the Length of Stay and 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)

I INTRODUCTION:

Length of Stay is a quality metric defined as the duration the patient spends in the hospital from the time of admission to the time of discharge. Average Length of Stay is often used as a key performance indicator in the Hospital industry to assess the efficiency, healthcare costs, and quality of care. A more extended stay contributes to inefficiencies in the provider facilities, higher out-of-pocket payments, and a higher risk of hospital-acquired infections. On the contrary, shorter lengths of Stay can contribute to improved facility efficiencies, patient experience and outcome, and reduced healthcare costs.

Besides being a quality metric, Length of Stay is vital in provider reimbursement. Healthcare providers often receive incentives from health insurance companies with reduced lengths of Stay. In Medicaid and Medicare programs, hospitals are paid based on the Medicare Severity Diagnosis-Related Group (MS DRG), not the Length of Stay. Predicting the Length of Stay is critical for the healthcare industry. By accurately predicting these metrics, the healthcare provider system can plan for the resources, thus improving operational efficiency, patient care, and outcomes and reducing healthcare costs. In this paper,

we explore predictive analytics and use a model to predict the Length of Stay for inpatient admission. Healthcare providers can utilize this predicted data to develop strategies to improve operations efficiencies.

II SOLUTION:

We will use predictive data modeling techniques to predict the value of Length of Stay. Predictive data model is a statistical technique that uses data mining and machine learning [2] to predict the future or outcomes. In the predictive model, we will leverage the random forest model to predict the Length of Stay based on multiple explanatory variables.

The Random Forest is a supervised learning algorithm. It can be broadly used to solve two problems: classification and regression. It uses many decision trees to make predictions. It takes predictions from each decision tree using the sample data and selects the best solution by voting. In our case, we will use a Random Forest Classifier model to predict the value for Length of Stay. By utilizing this method, we can develop a model that healthcare providers can use to improve operational efficiencies.

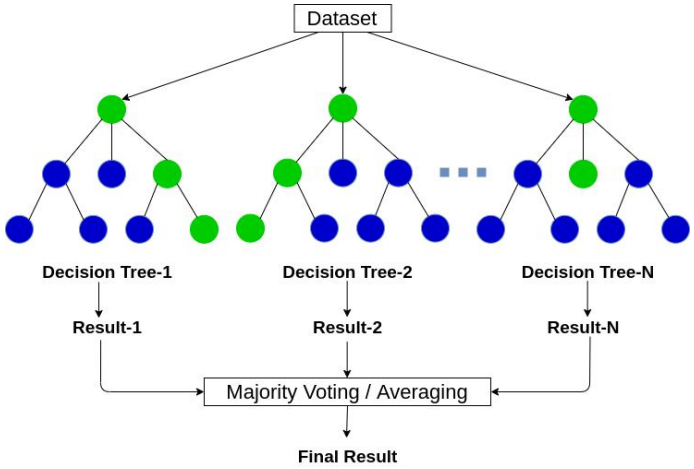

Figure 1 Random Forest Prediction [4]

To build this model, we will use the data from the Hospital Length of Stay Dataset Microsoft [3], which consists of patient demographics and medical information. Below are the steps we will be using in the process

- Data Collection: Identify and collect the required data from various sources
- Exploratory Data Analysis: Analyze and explore the datasets by summarizing the main characteristics using statistical graphs or visualizations.
- Splitting Data and Feature Selection: Split the data into two, one used for training the model and the other set used to test the model. Select the feature variables that would impact the model performance.
- Model Fitting & Evaluation: Evaluate the relationship between the dependent and independent variables.

A. *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 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

B. 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 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
```

C. 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.

D. ***Model Fitting & Evaluation:***

Random forests are more accurate than the decision tree as they reduce overfitting by combining multiple decision trees. Hence, Random Forest models are less sensitive to outliers and provide better performance with the prediction. We will use the RandomForestClassifier to build the Random Forest model and train them.

```
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 also increases with the increase in the number of decision trees. Finally, this Random Forest Model can predict the Length of Stay in discrete values with an accuracy of close to 70%.
A classification report is a way to evaluate the model's performance. It consists of the model's precision, recall, f1, and support scores. Below are the four ways to check whether a model's prediction is accurate or wrong.

- False Positive: The case was negative, but the prediction was positive
- False Negative: The case was positive, but the prediction was negative
- True Positive: Both case and prediction were positive
- True Negative: Both case and prediction were negative

Precision: It provides % the number of cases where predictions were correct.
Recall: It provides the % number of positive cases found.
F1 Score: It provides the % of predictions that were correct.
Support Score: It is the actual number of occurrences of cases in the given dataset.

III. APPLICATIONS OF THE SOLUTION IN VARIOUS ORGANIZATIONAL PROCESSES:

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

A. Assess Credit Risks in Finance Industry

Credit Risk Modelling plays a critical role in financial institutions' risk management. In the United States, bankruptcy filings are increasing steeply. There are more than 400,000 bankruptcy filing cases in the year 2021 [5]. Random Forest Models can be used to build credit risk models, which provide lenders with information, helping them to make informed decisions by lowering the risks and increasing profitability. Based on the outcome, banks can optimize lending strategies involving loan terms, interest rates, and other pricing structures.

B. Reduce Customer Churn in Ecommerce

Typically, in any business industry, acquiring a new customer cost more than retaining an existing customer , which is more accurate in the case of e-commerce. According to the latest reports, companies lose over hundred billion annually due to avoidable customer churn [6]. Also, based on the report shared by Forrester, acquiring a new customer would cost five times more than retaining an existing customer, which adds to the company's loss [7]. A random forest model can help build a predictive model to predict customer churn. The prediction can be used by the business to come up with strategies to retain the customer and minimize losses.

C. Hospitality Industry – Forecast Hotel Demands.

Forecasting hotel demands is critical for the hotel industry in this digital age. This would help the hotels plan to optimize resources and occupancy, regulate prices, and increase revenue. Hotels would incur significant losses due to overstaffing and lower prices without this data. Random Forest models can be used to predict hotel demands, which will help hotels develop strategies to run promotions, adjust prices, and adjust inventory and staff, which will significantly help increase profits.

IV. BENEFITS OF THE SOLUTION:

This solution offers several benefits to the healthcare industry across the world. Here are the key benefits

A. Optimize resource utilization:

Length of Stay is a critical measure for hospitals. They can use this data in multiple ways, starting with resource utilization. This value can help hospitals plan for staffing needs, manage bed capacity, and plan for the patient flow. They can use this data to plan for medical staffing needs and can prioritize more on the patients who are predicted to stay longer. This data can also be used to plan for other hospital resources like X-ray machines, lab equipment, surgical instruments, medications, and administrative staff. This would help hospitals utilize resources efficiently, which helps increase revenue and the quality of care provided to patients.

B. Enhance Patient Outcomes:

Predicting the Length of Stay is vital in enhancing patient outcomes. This data will help hospitals prioritize the patients with a high risk of developing complications and longer stays and take preventive measures. This metric will also help hospitals plan for the patient flow, which will further help reduce the wait times and, in most cases, can significantly impact patient outcomes. Finally, length of stay can be used to plan discharge activities, like moving to outpatient facility, another provider facility, caregivers, or family, to ensure a smooth transition and continued care.

C. Enhance Patient Satisfaction:

Predicting the Length of Stay is vital in enhancing patient outcomes. This data will help hospitals prioritize the patients with a high risk of developing complications and longer stays and take preventive measures. This metric will also help hospitals plan for the patient flow, which will further help reduce the wait times and, in most cases, can significantly impact patient outcomes. Finally, Length of Stay can be used to plan discharge activities, like moving to an outpatient facility, another provider facility, caregivers, or family, to ensure a smooth transition and continued care.

D. Reduce Healthcare Costs:

Length of Stay is a crucial quality metric that can help reduce healthcare costs from the hospital to the patient. Hospitals can use the predicted Length of Stay and prioritize the patients with higher complications, leading to a more extended stay and reducing it. Even if this can reduce the Length of Stay by one day, it will save the patient close to $13,000 in healthcare costs. This would also help hospitals save on resources and staffing, increasing healthcare cost savings.

E. Prevent Hospital Acquired infections:

Patients with extended Lengths of Stay have a higher probability of developing or contracting hospital-acquired infections. This can significantly impact patient outcomes and further increase the Length of Stay. As per the stats shared by CDC, one in every 31 US patients and 1 in every 43 nursing home residents

contract at least one infection associated to healthcare [8]. Hence, predicting the Length of Stay and its usage by hospitals is critical in the healthcare industry.

V CONCLUSION:

In conclusion, predictive data analytics can significantly improve healthcare efficiency and reduce healthcare costs. By predicting the Length of Stay data, healthcare providers can improve operation efficiency, reduce healthcare costs, improve patient quality care, and enhance patient satisfaction. This white paper provides a technical perspective on the vital role of data analytics in addressing challenges faced by healthcare systems, emphasizing the Length of Stay metric. It includes guidance on data-driven solutions to transform healthcare system delivery.

REFERENCES

[1] Tipton K, Leas BF, Mull NK, et al. Interventions To Decrease Hospital Length of Stay [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2021 Sep. (Technical Brief, No. 40.) Introduction. Available from: https://www.ncbi.nlm.nih.gov/books/NBK574438/

[2]PredictiveModeling:Types,Benefits,andAlgorithms|Rami , Sep 2020- https://www.netsuite.com/portal/resource/articles/financial-management/predictive-modeling.shtml

[3] Predicting Hospital Length of Stay|github, 2016. https://github.com/microsoft/r-server-hospital-length-of-stay/tree/master/Data

[4] Random Forests in Machine Learning: A Detailed Explanation| Saumya, Dec 2020- https://datamahadev.com/random-forests-in-machine-learning-a-detailed-explanation/

[5] September 2021 Quarterly Bankruptcy Filings- https://www.uscourts.gov/statistics-reports/september-2021-quarterly-bankruptcy-filings

[6] Prevent These Top 7 Reasons That Customers Churn, Aug 2021. https://arrows.to/resources/top-7-reasons-customers-churn

[7] Rethinking Customer Loyalty, March 2017- https://www.forrester.com/what-it-means/ep04-rethinking-customer-loyalty/

[8] 2021 National and State Healthcare-Associated Infections Progress Report - https://www.cdc.gov/hai/data/archive/2021-HAI-progress-report.html