PROJECT

On

HEALTHCARE ANALYTICS

By Cedric Mutoni

TABLE OF CONTENT

Introduction

Healthcare organizations face growing pressure to improve patient care outcomes and achieve better care. However, this challenge also presents an opportunity for organizations to significantly enhance the quality of care by unlocking more value and insights from their data. Health care analytics refers to the use of quantitative and qualitative techniques to explore trends and patterns in data. Although healthcare management uses various performance metrics, such as patient outcomes and efficiency, the patient's length of stay (LOS) is a vital indicator. By accurately predicting LOS, hospitals can optimize treatment plans to reduce LOS, minimize the spread of infections among patients, staff, and visitors, and improve the overall quality of care.

Project Goal

The primary objective of this project is to develop a predictive model that can accurately estimate the length of stay for each patient in a hospital setting. By achieving this goal, hospitals can improve their resource allocation and enhance their overall functionality. This model will be developed by analyzing patient data, including demographic information, medical history, diagnosis, treatment plans, and other relevant data. The ultimate aim of this project is to provide hospitals with a tool to identify patients who may require a longer stay, allowing them to allocate resources effectively and optimize patient care. The successful implementation of this model will lead to improved patient outcomes, increased hospital efficiency, and reduced costs.

Hypothesis Generation

Before conducting any data exploration or analysis, it is crucial to understand the problem in detail by generating hypotheses regarding the factors that may impact the outcomes of the Length of Stay. These variables can be classified into two levels: Patient-Level and Hospital-Level.

At the Patient-Level, the following factors may impact the Length of Stay:

• Type of Admission - Patients admitted in Urgent, Emergency, and Trauma levels have varying lengths of stay. Urgent care patients may have shorter stays while Trauma patients require longer stays for monitoring until they are qualified to be discharged.
• Severity of Illness - Patients classified as Minor, Moderate, and Extreme have varying lengths of stay. Patients classified as Minor will have shorter stays while patients classified as Extreme may stay for a longer period.
• Visitors with Patient - Patients with more visitors may require longer stays in the hospital.
• Age - Infants and older patients typically require longer recovery time, so they may have a longer Length of Stay than younger patients.
• Admission Deposit - Patients who are required to deposit a high amount of money at the time of admission may have severe conditions and require longer stays.

At the Hospital-Level, the following factors may impact the Length of Stay:
• Ward Type - Patients allocated in ICU may require longer stays than those in the general ward as their conditions are more severe.
• Department - Patients under surgery may have longer stays than those in gynecology as their recovery time is longer.

Understanding and exploring these factors can help in developing an accurate predictive model to estimate the Length of Stay and optimize patient care and resource allocation in hospitals.

Data Exploration:

4.1 Overview of Data

The training dataset for this project comprises 318,438 observations, and it includes 17 variables that can be used to predict the patient's Length of Stay. A detailed description of all the variables in the dataset is presented in Table 4.1. By analyzing this data, we can identify patterns and trends that may impact the Length of Stay and help to develop a predictive model to estimate it accurately.

Table 4.1: Dataset overview

Variables	**Description**
case_id	Case_Id register in Hospital
Hospital_code	Unique code for the Hospital
Hospital_type_code	Unique code for the type of Hospital
City_Code_Hospital	City Code of the Hospital
Hospital_region_code	Region Code of the Hospital
Available Extra Rooms in Hospital	Number of Extra rooms available in the Hospital
Department	Department overlooking the case
Ward_Type	Code for the Ward type
Ward_Facility_Code	Code for the Ward Facility
Bed Grade	Condition of Bed in the Ward
patientid	Unique Patient Id
City_Code_Patient	City Code for the patient
Type of Admission	Admission Type registered by the Hospital
Severity of Illness	Severity of the illness recorded at the time of admission
Visitors with Patient	Number of Visitors with the patient
Age	Age of the patient
Admission_Deposit	Deposit at the time of Admission
Stay	Patient Length of Stay

Data Cleaning and Preparation:

4.2 Data Cleaning and Preparation:

The dataset used in this project contains missing values in the "City_code_patient" and "Bed Grade" variables, which can affect the accuracy of the predictive model. To address this issue, we employed the "mode" imputation technique to replace the missing values before feeding the data into the algorithm.

As most of the variables in the dataset have ordinal data, we used a label encoder to transform them into levels, allowing us to perform further analysis on the data. The number of distinct observations of ordinal data in the dataset is presented in Table 4.2. This process of data cleaning and preparation is necessary to ensure that the data used for analysis is accurate and reliable, allowing us to develop a robust predictive model for Length of Stay.

Table 4.2 Distinct Observation of ordinal data

Variables	**Number of distinct observations**
Hospital_type_code	**7**
Hospital_region_code	**3**
Department	**5**
Ward_Type	**6**
Ward_Facility_Code	**6**
Type of Admission	**3**
Severity of Illiness	**3**
Age	**10**
Stay	**11**

4.3 Feature Engineering

We conducted feature engineering on the cleaned and prepared data. To extract more meaningful insights, we created a new column "count_id_patient" by grouping patientid and case_id. This new variable captures the count of multiple admissions of a patient under different case_id. Additionally, we grouped "Hospital_region_code" and "ward_facility_code" into patientid and case_id, resulting in two more variables: "count_id_patient_hospitalCode" and "count_id_patient_wardfacilityCode". These variables capture the count of multiple admissions in a hospital region and the count of multiple wards allocated to a patient.

To begin the analysis, we split the train data into two parts: the first part containing all the feature variables and the second part containing the target variable ("Stay"). The train data was then preprocessed into train and validation sets. For the Naïve Bayes and XGBoost models, we partitioned the train set with 80% of the data and the validation set with 20% of the data.

Model Building

5.1 Model 1 - Naïve Bayes

The first model used in this project is Naïve Bayes, which is a classification technique based on Bayes theorem. It assumes independence among the variables and is suitable for multilevel classification. The main objective is to predict the Length of Stay for each patient, which is classified into 11 levels. The model uses feature variables, such as the patient's condition and hospital-level information, to calculate the probability of the patient's Length of Stay.

Bayes theorem provides a framework for calculating the probability of a hypothesis (H) given some observed evidence (E). In our case, H represents the prior probability of a patient's Length of Stay, E represents the probability of a feature variable, P(E|H) represents the probability of a patient's Length of Stay given that the features are true, and P(H|E) represents the probability of the features given that the patient's Length of Stay is true.

The model is trained using a Gaussian Naïve Bayes classifier, and the partitioned train data is fed to the model in array format. Then, the trained model is validated using the validation data. After validation, this model provides an accuracy score of 34.55%.

5.2 Model 2 – XGBoost

XGBoost is an ensemble method that sequentially combines a set of classification and regression trees. The principle of ensemble learning is to weigh model outcomes based on the outcomes of the previous instant. XGBoost tries to reduce the misclassification rate by growing trees one after another. The final prediction score of the model is the sum of each individual score.

Before training the XGB Classifier model with the train data, booster parameters must be tuned to prevent overfitting and yield higher accuracy. In this XGBoost model, the following parameters were used for tuning:

• Learning rate = 0.1: the step size shrinkage used to prevent overfitting.

• Max_depth = 4: the maximum depth of the tree that describes the model complexity. Increasing its value results in overfitting.

• N_estimators = 800: the number of gradient boosting trees or rounds. Increasing the number of trees can yield higher accuracy, but the model reaches a point of diminishing returns quickly.

• Objective = ‘multi: softmax’: sets XGBoost to do multiclass classification using the softmax objective because the target variable has 11 levels.

• Reg_alpha = 0.5: the L1 regularization term on weights. Increasing this value makes the model more conservative.

• Reg_lambda = 1.5: the L2 regularization term on weights that is smoother than L1 regularization. Increasing this value makes the model more conservative.

• Min_child_weight = 2: the minimum sum of instance weight needed in a child.

After training and validating, the XGBoost model yielded an accuracy score of 43.04%, an 8.5% improvement over the Naïve Bayes model.

5.3 Model 3 – Neural Networks

The Neural Networks model consists of neurons that take in a real value, multiply it by weight, and pass it through a non-linear activation function. The model records one record at a time and learns by comparing the classification of the record with the actual classification. The errors from the initial classification are fed back into the network and used to modify the network's algorithm for further iterations.

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 254750, 64)        1344

 dense_1 (Dense)             (None, 254750, 128)       8320

 dense_2 (Dense)             (None, 254750, 256)       33024

 dense_3 (Dense)             (None, 254750, 512)       131584

 dense_4 (Dense)             (None, 254750, 512)       262656

 dense_5 (Dense)             (None, 254750, 11)        5643

=================================================================
Total params: 442,571
Trainable params: 442,571
Non-trainable params: 0
_________________________________________________________________
```

The model has six dense layers, and the final output layer has an activation function "SoftMax" as each patient must be classified in one of the 11 levels in the Stay variable. The number of neurons in each layer is increased to increase the hypothetical space of the model and learn more patterns from the data. The model has a total of 442,571 trainable parameters.

Before training the model, the data were scaled, converted into a sparse matrix, and divided into 80% as the train set and 20% as the test set. The model was compiled using "categorical_crossentropy" as a loss function and "SGD" as an optimizer argument. Initially, the model was trained with 20 epochs, but it was observed that the model was overfitting from the 4th epoch. So, the model was retrained by setting epochs to 4.

Finally, the model was evaluated using the test set, yielding an accuracy score of 42.05%. Although Neural Networks are considered to perform better than other models, due to the smaller dataset, it was not able to learn more accurately than the XGBoost model.

Prediction and Results

As per the results of the project, the Naïve Bayes model has a higher chance of misclassifying patients. This model shows a bias towards predicting a length of stay of 21-30 days, with 72,206 patients being classified under this category.

Table 6.1 Number of observations classified into different levels of length of stays from all model

Length of Stay	**Predicted Observations from Naïve Bayes**	**Predicted Observations from XGBoost**	**Predicted Observations from Neural Network**
0-10 Days	2598	4373	4517
11-20 Days	26827	39337	35982
21-30 Days	72206	58261	61911
31-40 Days	15639	12100	8678
41-50 Days	469	61	26
51-60 Days	13651	19217	21709
61-70 Days	92	16	1
71-80 Days	955	302	248
81-90 Days	296	1099	1165
91-100 Days	2	78	21
More than 100 Days	4322	2213	2799

However, the XGBoost and Neural Network models show similar predictions for the length of stay of patients. The similarity in predictions can be observed in the first five cases in Table 6.2. In Table 6.1, we can see that the predictions made by these models have only marginal similarity.

Table 6.2 – Predicted Length of Stay for first five cases from different models

case_id	Length of Stay predicted from Naïve Bayes	Length of Stay predicted from XGBoost	Length of Stay predicted from Neural Networks
318439	21-30	0-10	0-10
318440	51-60	51-60	51-60
318441	21-30	21-30	21-30
318442	21-30	21-30	21-30
318443	31-40	51-60	51-60

Upon examining these predictions, it was observed that a majority of patients had a hospital stay of 21-30 days, and only a few patients had a stay of 61-70 days. The distribution of length of stay shows that 13% of the patients were discharged within 20 days, and only 1% of patients stayed in the hospital for more than 60 days.

Future Insights

As the volume of quality data in healthcare increases, it can help healthcare professionals to allocate resources efficiently, leading to smart staffing and personnel management. With a comprehensive analysis of health outcomes among individuals in various demographic groups, healthcare professionals can identify factors that prevent individuals from seeking treatment and provide effective preventive care for advanced risk and disease management. Hospitals can effectively decrease hospital admissions by analyzing the duration of patient visits, drug types, and conditions, among other insights.

Clinical Decision Support (CDS) in hospitals can also provide real-time alerting for physicians by analyzing patient evidence and delivering recommendations to health professionals when they make prescriptive choices. However, physicians prefer patients to stay away from hospitals to avoid unnecessary in-house procedures.

In addition, smart wearables can help enhance patient engagement by tracking their heart rates, sleeping habits, and other trackable data to identify potential health risks. All this information can be correlated to provide a more comprehensive understanding of patient health.

Conclusion

In conclusion, this project aimed to analyze various patient-level and hospital-level variables that correlate with Length of Stay (LOS) in hospitals. The ability to predict LOS at the time of admission can help hospitals to manage their resources and patients more effectively. By identifying factors that are associated with LOS, hospitals can predict and manage the number of days patients stay, which can lead to better resource management and the development of new treatment plans. By reducing LOS and effectively utilizing hospital resources, the overall national medical expenses can be minimized. Additionally, this project provides insights into the potential use of advanced technologies such as smart staffing and personnel management, advanced risk and disease management, real-time alerting, and enhanced patient engagement to improve healthcare outcomes in the future.