University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

"Predicting length of stay (LOS) and categorization for healthcare management"

SHAMLI BAJAD
2211489

Supervisor: **Dr. Na,You**

November 24, 2023
Colchester

Contents

Abstract

The hospital Length of Stay (LOS) across many categories such as short, medium, long, very long and prolonged stays can be predicted by a machine learning algorithm described in this paper. The project aims to enhance care for patients, operational efficiency, and resource allocation in the hospitals while addressing a major difficulty associated with healthcare administration. Hospital data is carefully examined for the study after which data preparation and Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the dataset. On each balanced and unbalanced datasets three machine learning models; Ordinal Logistic Regression, Decision Tree Classifier and XGBoost were used and evaluated. To determine the effectiveness of each model performed, comparative analyses were performed including ROC-AUC curve evaluations. The XGBoost model performed more effectively accurately predicting LOS in a variety of categories, particularly after the hyperparameter adjustment. Through establishing an effective model for LOS prediction improving patient care planning and optimizing hospital resource management this study improves healthcare analytics. Future investigations require to delve into supplementary data sources and advanced machine learning methodologies in order to improve the estimated precision and practicality of the model within the health care sector.

Keywords: Hospital Length of Stay (LOS),Machine Learning Algorithm,Synthetic Minority Over-sampling Technique (SMOTE),Ordinal Logistic Regression,Decision Tree Classifier,XGBoost,ROC-AUC

List of Figures

List of Tables

CHAPTER 1

Introduction

The effective utilization of resources in healthcare organizations is a major problem within the ever evolving health sector.One of the main issues within this context is LOS which extends beyond management practices and reflects the standard of hospital operations in terms of the health of patients. The quality of treatment is a additional aspect of LOS that significantly impacts the number of patients treated and hospital expenditure control.Traditionally, generic statistical methods have utilized for predicting LOS.Although their wide accessibility, these strategies sometimes lack to give the particular circumstances the attention they deserve.

Length of stay is determined by factors like medical diagnosis, treatment methods, socio backgrounds and health backgrounds. Inadequate consideration of such variables may result in illfated strategies that will be characterised by improper use of resources and poor patient care planning especially when considering a demographic like the elderly that usually has multiple medical conditions prolonged admissions and complicated care needs.

Predicting LOS in the rapidly developing field of medical science is also challenging. Advances in medical science and changes in treatment approaches often reinterpret the timeframes for recovering and the needs of patients when they are discharged. Therefore for improved healthcare services a flexible and thorough LOS estimation approach is essential.

In this context, the use of machine learning and data analytics becomes a revolutionary remedy. These enable one to discover hidden trends and findings from large and heterogeneous sets of patient data that traditional statistics cannot pick. The study explores various models including Decision Trees, XGBoost, and Ordinal Logistic Regression having distinct strengths towards LOS prediction and administration. With techniques such as Synthetic Minority Over sampling Technique (SMOTE) to counteract the class balance they refine the model to be accurate and representative of various age groups, ethnicities and health histories.

This study aims at using big data and machine intelligence for strategic alteration in the LOS forecasting process. It is instrumental in supporting the broader paradigm switch of the healthcare system towards a patient oriented model with an individualized care and resources dispensation. This study aims for improving resource utilization and personalised care provision through prediction of short stay, medium stay and long stay. The purpose in doing so is to provide medical staff members with an all encompassing tool that they can use to classify stay lengths for managing patient care making optimal use of resources and reducing costs.

CHAPTER 2

Literature Review

All health care systems have long been interested in learning more about length of hospital stay (LOS). The length of a patient's hospital stay (LOS) is a greater issue for clinical and surgical specialists. Accurate prediction of patient length of stay (LOS) is important as health care systems change, and benefits range from better surgery to better patient care Researchers around the world have addressed this issue in recent years, each highlighting different aspects of this multifaceted problem. [1]

In a 2013 study titled "Analysing Hospital Length of Stay Using Electronic Health Records: A Statistical Data Mining Approach," researchers Hynyong Baek, Minsu Cho, Seok Kim, Hee Hwang, Minseok Song, and Sooyoung Yoo examined electronic resources the health impact is under-reported (Length of stay (LOS) in EHR-using hospitals) data. The study used advanced statistical analysis and data mining techniques to provide a nuanced understanding of LOS dynamics. Studies found a mean LOS of 7 days that varied widely across sectors, such as rehabilitation medicine and neuropsychiatry, which notably had a median delay in which nearly 55% of patients discharged within 4 days, and it shows the variation in patient stay. Factors such as brain tumors were found to be significant contributors to LOS prolongation. Furthermore, the study showed that chronic inpatients, representing 3% of the study population, had high rates of surgery and antibiotic use.In addition, the investigators observed that various modes of transfers impacted LOS, and transferred individuals experienced prolonged LOS. Models for predicting length of stay (LOS) and hospitalization were developed using frequency of transfer, frequency of surgery and frequency of diagnosis as predictors. Even though these studies investigated some other variables, they depended entirely on a single clinic's data without full consideration of broader demographic and environmental factors that may affect LOS. [2]

In the paper "Predicting Hospital Patients Length of Stay: A Federal Study Method" research presents length of hospital stay (LOS) . Machine learning method that models multiple decentralised organizations can be trained while keeping their data private. This new approach is especially important in the healthcare industry where patient privacy is paramount. Researchers use three types of classical machine learning regression model: linear, lasso and ridge regression each instances of data from an individual hospital clients was trained reflecting decentralization of the ideal training approach. The uniqueness of the method lies in the use of integrated learning, which allows learning samples from multiple sources where true patient information is need to be shared, thus preserving privacy and confidentiality. client 1 with 20,353 patient's information had many errors predictions. Additionally, a smaller data set (client 7 with 2,007 patient's records) was exposed, and the linear regression model was shown to be outstanding too, as it lowered the mean absolute error (MAE) with a decrease in patient population when using two clients from 2225 to 1389.In spite of its original approach and important results, this research has some shortcomings. Using correlation between variables could help improve strong nonlinear relations in the data. Furthermore, variations in the size and caliber of clients dataset may affect the generalised ability of results. [3]

In "Prediction and Analysis of Length of Stay Based on Nonlinear Weighted XGBoost Algorithm in Hospital" by Yong Chen. Several machine learning models were implemented and compared, including Naive Bayes, Decision Tree, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Standard XGBoost, and Nonlinear Weighted XGBoost without cross-validation, parameter tuning : Procedure 79.12 percent careful along with designed Naive Bayes accuracy emerged while the decision tree reached a slightly higher accuracy of 82.47% and the Support Vector Machine (SVM) emerged as one of the best performers at 86.61% accuracy Metric types such as F1-score are equal This highlights the power of machine learning models optimized to outperform standard algorithms in determining complex tasks such as duration of residence. [4]

Jan Crucial, Francois Girardon, Lucien Roquet, David Laplanche, Antoine Duclos, and Stephane Sanchez, authors of the research article " The prediction of hospital length of stay using unstructured data," presented raw data from implemented electronic health records (EHRs) in a large French-language clinic. With a focus on length of stay (LOS) the researchers used machine learning, specifically random forest samples to analyse the data. Models were developed to classify hospital stays as "short" or "long" based on a median of seven days. The modelling system used two factors: one model used only structured data (codes and CCMU classification codes), while the other used the Unified Medical Language System (UMLS) to treat these unstructured data this was addressed to eliminate the relevant medical perspective. The findings of the study were surprising. In terms of the fact that the model incorporating raw data showed slightly higher accuracy accuracy (75.0%) compared to the model using only structured data (74.1%), both models agreed so in their prediction in 86.6% of the cases. In a second study focused on intensive care patients, the unstructured data model again showed slightly better performance. However, its limitations, the time frame of study was limited to specific months in 2019. This relatively short period provides a wealth of information, but additionally raises questions approximately the generalise ability and scale of the findings. [5]

In the study entitled "Network evaluation and machine learning for predicting the length of stay of elderly patients with chronic diseases" shen, wang, qiu entered in this essential subject matter.The researchers set out on a undertaking to seamlessly integrate machine learning and network analysis. Knowing the existence of lacking statistics at the point of admission(POA), a unique method for predicting LOS for elderly patients with chronic illness was developed. To determine the predictive power of these features, the team used machine learning models, each of which has incorporated different inputs into these models along with extreme gradient boosting algorithms and deep neural networks.The findings of the observe were encouraging and informative.This study used big data on 2.5 million hospital transferred records from urban area of Chengdu, China, during the period of 2015 to 2019. In summary, it was discovered that XGBoost beat other models and produced a very good resolution (R2=0.375) for Baseline, History, MN ,PSN features combined. Using such strategy jumped of 18.7% point higher than the baseline history materials. Research also emphasized that historical network attributes could be used as the main predicting factor for getting the highest modelling precision. [6]

The study titled "Predictors of in-hospital length of stay among cardiac patients: Studies carried out with a machine learning approach by Tahani A. Daghistani et al. The purpose of this research was to create a machine learning model, aimed at anticipating the In-Hospital LOS for those who have been admitted into adult cardiology services. The research team used electronic medical records to retrieve non-concurrent patient visit data retrospectively. Patients were categorized into three groups based on their LOS: for example, less than 3 days, between 3 and 5 days, and over 5 days. Information gain algorithm was used to identify appropriate attributes for this model. The prediction model was based on four different machine learning techniques including random forest (RF).Random Forest model had better performance when compared with other models. Sensitivity was 0.80, while accuracy was also 0.80. Therefore, an AUROC was 0.94. Generalisation was limited by potential biases due to retrospective, data heterogeneity, and specific healthcare settings. The studies further point out a necessity of perpetual improvement of ML models including issues related to data quality and explaining the models. [7]

The study conducted by Mustafa Ataman, MD, and Sariyer, PhD, introduces a notable advancement in the field of predicting waiting and treatment times in emergency departments (EDs) through the application of ordinal logistic regression models. This methodological innovation plays a pivotal role in enhancing our understanding of the dynamics within EDs and how various factors influence patient experiences.One of the fundamental strengths of ordinal logistic regression models is their ability to classify patients into ordered categories based on waiting and treatment times. Traditional regression models might struggle with capturing the nuanced variations in time-related quality indicators within an ED. Ordinal logistic regression, on the other hand, categorizes patients into meaningful groups, such as short wait times, moderate wait times, and long wait times, allowing for a more precise analysis of the data.By adopting ordinal logistic regression, the study provides a structured and systematic framework for assessing the impact of multiple factors on ED operations. This structured approach is essential for identifying the key drivers of delays and variations in treatment times.One of the notable advantages of using ordinal logistic regression models is their ability to improve the accuracy of predictions. Unlike binary logistic regression, which focuses on two outcomes (e.g., admission or discharge), ordinal logistic regression considers multiple ordered categories. This not only allows for more precise predictions of waiting and treatment times but also enables healthcare institutions to allocate resources more efficiently. [8]

In total, applying machine learning methodological framework towards healthcare and specifically in predicting patients expected stay duration does encourage us. In this way many other areas of research and improvement leading to increased levels of preciseness by limiting the downfalls and usage in numerous health care areas.

CHAPTER 3

Methodology

3.1 Data Description

The project used detailed information on hospital patient discharges from New York State in 2015. Obtained from the SPARCS, this dataset contains discharged records, and provides public access for research and analysis Description, clinical research Includes a wide range of data groups including, treatments received, and clinical specialties. dataset contain total 1045306 records and 34 variables. [9]

1. Demographic Information The data is comprised of the basics columns like age, gender, race, and ethnicity, plus part of the zip code (the first three digits) that gives the region's outlook.

2.Hospital and admission information This also includes information about the clinic's site and how to access the clinic.It comprises health care, hospital district and mode of admission (emergency admission, elective admission, etc.).The dataset also includes patient characteristics such as a diagnosis or reason for discharge.

3.Medical information However, the highest priority in this set of information is the clinical data.This includes a diagnostic group code ,APR DRG Code, categorizing patients according to similar medical conditions as well as severity of diseases, mortality risk classification and some important variables of the interest as well as duration of recorded sleep, which is one of its vital determinants of the hospital.

4.Financial and operational information The hospital has documented financial data like the total charges and total expenses. The other information contained in the dataset is payment types highlighting the main form of payment made e.g., Medicare or private insurance.

5. Target variable: length of stay An important aspect of the project objective is length of stay, which is initially recorded as a continuous variable. It has an important application in the assessment of health service utilization and hospital resource utilization. For research purposes, these changes were converted into categories, such as short-term, medium-term, and long-term, which made them easier to use in predictive modelling methods. The sets of data give broad perspective on patient admission, taking into accounts socio-demographic characteristics, clinical features, clinical environment, as well as financial components. The main target is to forecast length of stay in hospital aiming to improve health care administration.

3.2 Data Pre-processing

3.2.1 Data Cleaning Process

First, it is important to note that data cleansing should be done before discussing peculiarities of the data analysis in detail. Nevertheless, cleaning might more reasonably be seen as the initial stage which impacts the integrity and reliability of every consequent exploration. The process of cleaning data is essential for several reasons: data quality means ensuring that data are properly coded and consistent; this improves model performance, deals with missing data, and removes redundancy enlarge valuable sample size. [10]

1. Initial Data Assessment

First dataset was evaluated the original data that contained 34 columns. The columns incorporated different items, for instance, demographics, clinical data, as well as financial attributes of the hospital's discharge.

2. Handling Missing Values

Missing data imputation can help enhance the performance of prediction models in circumstances where missing data hides useful information. [11]

In this regard, a decision was taken to exclude 'Payment Typology 2', 'Payment Typology 3', and 'Zip Code 3 digits' columns containing huge misses. Likewise, any rows having any of those missing values were deleted from the dataset.

3. Duplicate Entries

Data duplication’s needed to be identified and addressed for a quality analysis’s process to happen.Some of these duplication had to be eliminated in order to curb overlaps that might have led to distortion of data and biased results thereby guaranteeing the authenticity of the findings’.After this data cleaning exercise, the dataset was significantly transformed.

3.2.2 Data Encoding

The first basic process of data analysis in particular, whenever conducting advanced statistical analysis or constructing a machine learning algorithm is encoding categorical variables. This will consider an analysis on the purpose, functioning and results of coding in this part of the discussion.

Encoding took place primarily for analyzing it through numerous statistical test For the tests involving independent categorical variables, correct mathematical computation depends on the numeric value. Using this procedure, scikit-learn Label encoder converted each categorical variable into a unique integer set. [12] Predictive modeling constitutes one of the key components of this project; these mathematical models comprising mostly numbers are also primarily computer-based. The data was categorized and converted into its numerical form, which made it usable by many different types of machine learning algorithms. This ensured the algorithms decoded the data correctly to get accurate results.

Special emphasis was placed on the preservation of the distinctiveness of each category during coding. Label encoding assigned unique integer values to categories that corresponded to the qualities of the data. Consequently this critical transformation was necessary for maintaining the truthfulness and integrity of the information during analysis processes.

Encode also organized or categorized data by converting the text data into numbers. The simplification of the dataset helped analysts work more quickly with the aid of various data analysis techniques in order to obtain optimal results using minimum resources. [13] In this regard the Label Encoding is a mechanism that allots each of the above-mentioned categories with an integer value. [14]

3.3 Feature Engineering

There were some essential procedures which transformed ′Length of Stay′ field in order to classify it properly within data set. The initial contents in the field were both numbers and non-numbers. Therefore, each entry, including ′120+′ representing one-year stay or more, was encoded numerically only. For instance, ′120+′ was normalized to the numeric value of 120. Such transformation allowed for a more accurate and standard rating system.

The categories in the ′Stay Category′ column were encoded. This is because these categories form a natural order and hence the OrdinalEncoder was used for coding. The categories are numbered by increasing numbers that identify them in ascending order of their position with the help of OrdinalEncoder. For example, ′Short stay′ may be coded by zero while ′Medium stay′ is indicated by one. Another column, ′Stay_Cat_Ordinal_Data′, encoded it.

3.4 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) played an important role in project. In particular, it allowed to analyse the data in depth, gaining a better understanding of variables, their categories and classification This preliminary analysis was important as it provided the necessary foundation for subsequent research steps all. Once got a better understanding of the data, statistical analysis was performed on the numeric data.

3.4.1 statistical analysis

This step is the most important in terms of making quantitative sense of the data, especially for important variables such as ′length of stay′ as its a target variable.General statistics like mean, median, standard deviation, range was observed.

The metrics were essential in revealing the underlying trends and spread of the central tendencies of the dataset thus giving me an overview of the dataset′s inherent characteristics. The analysis helped point out anything unique or remarkable about the statistical variables that may influence features selection and the formation of samples at subsequent stages.Additionally, found meaningful facts column "Length of Stay".

Here's a detailed breakdown discovered: Statistical analyses are made more powerful and trustworthy through the utilisation of large databases comprising of 1,041,939 patient's medical records. Besides, this large sample size adds to the validity of the paper and greatly reinforces the significance as well as reliability of the final conclusions. However, the main target was the length of stay that provided a detailed interpretation about the most significant part in that data set.

Average Stay (Mean = 5.39 Days): This mean value served as the lowest common denominator understanding how long the average hospital is stay in days. A central point around which other observations could have been contextualized, for instance, an average of about 5.39 days per patient. It observed that on average the patients were in hospital for slightly less than one week. It showed a great variability in the length of patient stay with high standard deviation estimated at 7.32 days. This indicates that the patients present different kinds of needs and therapies which can vary as per the duration from short term to long duration. Therefore, this variation should be considered in the development of a model, because these different experiences were manifested in that data set. Besides, the first hospitalisation period took only a day while the last one lasted for as long as 119 days.

The great variance in hospital stay pointed out from mild to severe cases. A working model must reflect such diversity adequately so that it can handle this patient diversity. Distribution Insights (25th Percentile = 2 Days, Median = 3 Days, 75th Percentile = 6 Days): Quartiles gave a clear indication of a positively skewed distribution. Many cases only spent two days on the study, with roughly quarter of it lasting for 2 days or even less. The skewness proved important during my modelling approach because it implied existence of outliers and care needed when dealing with the data. Overall, all these findings from this analysis are having huge range of stay duration in hospital. Most of the patients tent to have shorter stay compared to other categories. Understanding this distribution was vital for the next stages of my project, particularly in selecting and tuning the predictive models.

3.4.2 correlation analysis

This later prepared the project for correlation analysis, an important step in several ways. The broad statistical analysis having been completed, this phase was especially vital in perfecting insights. In addition, correlation analysis was used for further findings relationships between different variables.

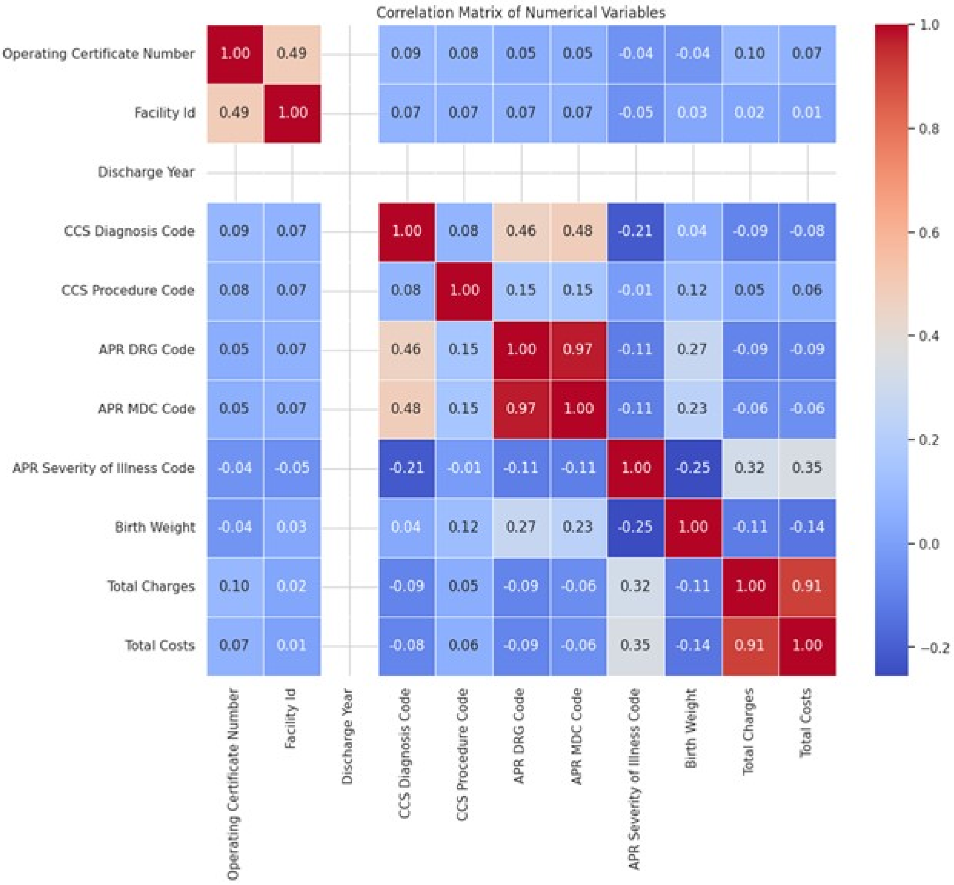

Figure 3.1: Correlation matrix

In correlation analysis It was found the high correlated variables that explain very peculiarly to understand an obstacle of given dataset. Particularly, 'Total Charges' and 'Total Costs' showed a high positive correlation coefficient of 0.91 reflecting the close linearity between financial characteristics of hospital stays. This had been a predicted result given its conformity to the commonly accepted notion in most cases which claims that expensive items are usually costly across the board.

This was seen using 'APR DRG Code' and 'APR MDC Code' which had a correlation score of 0.97. It showed how close these two classification systems are related with each other when it comes to healthcare billing and administration. Additionally a weak positive connection of 0.49 between 'Operating Certificate Number' and 'Facility Id' implied that some operating certificate are designated for particular facilities. These indicated a structured scheme of certifying in the healthcare sector.

The correlation of 'APR Severity of Illness Code' with 'Total Charges' (0.29) and 'Total Costs' (0.31) was particularly insightful. It indicated that cases deemed more severe tend to incur higher charges and costs providing a direct link between the clinical aspect of patient care and its financial implications. Additionally, 'Birth Weight' showed interesting correlations. These included "APR DRG Code" with a relation coefficient of 0.27 and 'APR MDC Code' which indicated a relation coefficient of 0.23 which were probably due birth related diagnoses and proceedings. [15]

However, it went in a reverse relation to 'APR Severity of Illness Code' (-0.25) and 'Total Costs' (-0.11) meaning that most of such cases tend to be less serious and costly. Lastly the 'Discharge Year' variable showed NaN correlations due to it being a constant across the dataset, rendering it non contributory in the correlation analysis.

These correlations offered a comprehensive view of how different aspects of hospital stays, such as facility characteristics, medical classifications severity of illness and financial aspects are interconnected. This understanding was crucial in informing the predictive modeling phase of my project, especially in identifying key factors that could significantly influence the length of hospital stay.

3.4.3 categorical variable analysis

A critical step was doing an analysis of the categorical variable because it allowed me to see relationships that exist between non numerical data because that often gives very important information.These categorical variables e.g. gender, race, ethnicity and type of admission are usually not obvious but they have significant effects on patients outcomes and hospital process.

The examination of these factors was essential for revealing patterns, preferences, and disparities which might help determinate the duration of hospital stays and general patient care. This phase in the project was very important in determining which elements to choose as well as model adjustments. This improved the accuracy of the developed classifier towards portraying the complex nature of heath care data and thus increasing the reliability and generalization in the research findings.

1.Age Group:

Most of the short duration hospital admissions were recorded in the youngest of patients who are 0-17 years old.The fact that health problems in these people require

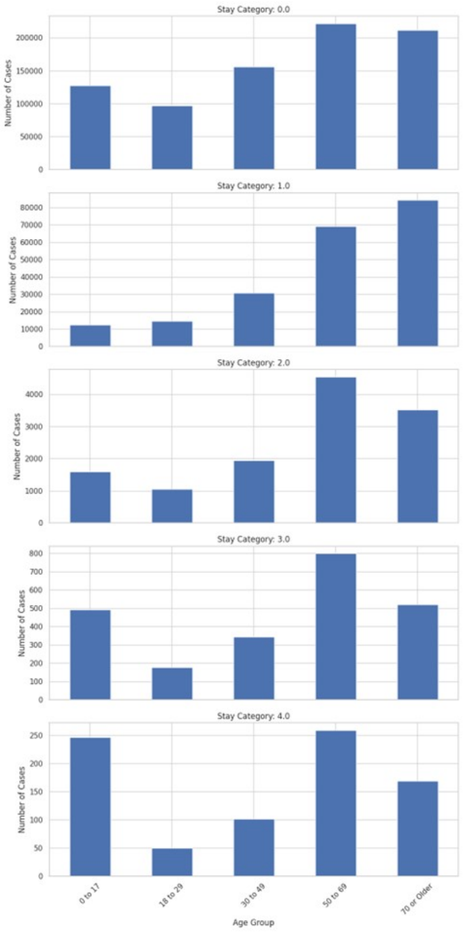

Figure 3.2: Age and stay category analysis

shorter hospital period implies mildness, and easy treatment of their ailments in general.Additionally, there were few episodes of long term hospitalization that lasted up to one month or more for children younger than 2 years old.

The results were also comparable with those found in the youngest children aged 18 to 29 years.Most of these cases were short stays implying such a demo-graphical cohort did not have serious health problems which required lengthy hospitalization.

Looking at these middle-aged groups, with the age range between 30 and 69years, another trend occurred. As evident an increase of prolonged hospital admissions especially at 50-69 years grouping. It should be expected, considering that these are probably the first acute and more short-term problems which are less dependent on old age. However, the bulk of cases involved short stays, pointing at an equally broad spectrum of health problems, but with different degrees of severity. The greatest variations were discovered among seniors, those seventy years old and above.

As expected for a population with chronic problems common in older people, this was the oldest group that had the highest proportion of longer stays. Nevertheless, they remained high rates of short stay, suggesting that healthcare services were required in different forms within this population.

2.Gender:

Investigating the impact of gender on hospital stay lengths. This analysis aimed to discover any significant differences in hospitalization patterns between male and female patients which could be indicative of gender specific health trends or healthcare needs.

Stay Category	Short Stay	Medium Stay	Long Stay	Very Long Stay	Extended Stay
Gender					
F	465,030	105,771	5,795	1,058	385
M	348,765	105,690	6,856	1,276	441

Table 3.1: No. of cases by Gender and stay category

1. Short Stays: Females: For instance, female patients had more occurrences of short term admissions as shown in the dataset of about 465030 cases. Such prevalence indicates that although mostly women are admitted for inpatient services the length of stay is usually shorter than mens.

Males: Besides male patients also had many short stays which amounted to 348765. Such situations mean that short term hospitalization is a frequent event in which both sexes participate more often by females.

2. Medium Stays: Females: Female patients had a total of 105771 cases of medium-term hospitalization.

Males: The number of men with medium length admissions stood at 105690, which means that gender was not necessarily a factor.

Longer Stays: Females: The data showed that females had slightly higher numbers in longer stay categories: 5,795 in 'Long Stay', 1,058 in 'Very Long Stay' and 385 in 'Extended Stay'. Males: In comparison males recorded 6,856 long stays, 1276 very long stays and 441 extended stays.

However, these findings provide a more complex depiction of gender specific hospitalizations. Generally short stays prevail among both females and males but females tend to have a slight percentage of longer stays. That may indicate differences in health statuses, health requirements and health seeking patterns between males and females. Such an understanding is vital in relation to health planning and policy formulation, because it emphasizes on the need for gender specific healthcare demands and supplies. The reason behind females might have a higher number of short stays might involve issues like childbirth related admissions that generally tend to be shorter.

3.Type of Admission

The "Type of Admission" aspect in hospital data serves as critical indicator for understanding the reasons behind hospitalizations encompassing a range of scenarios from unplanned admissions to various other cases. This analysis not only highlights the dynamics of patient inflow but also aids in optimizing hospital operations and enhancing patient care management.

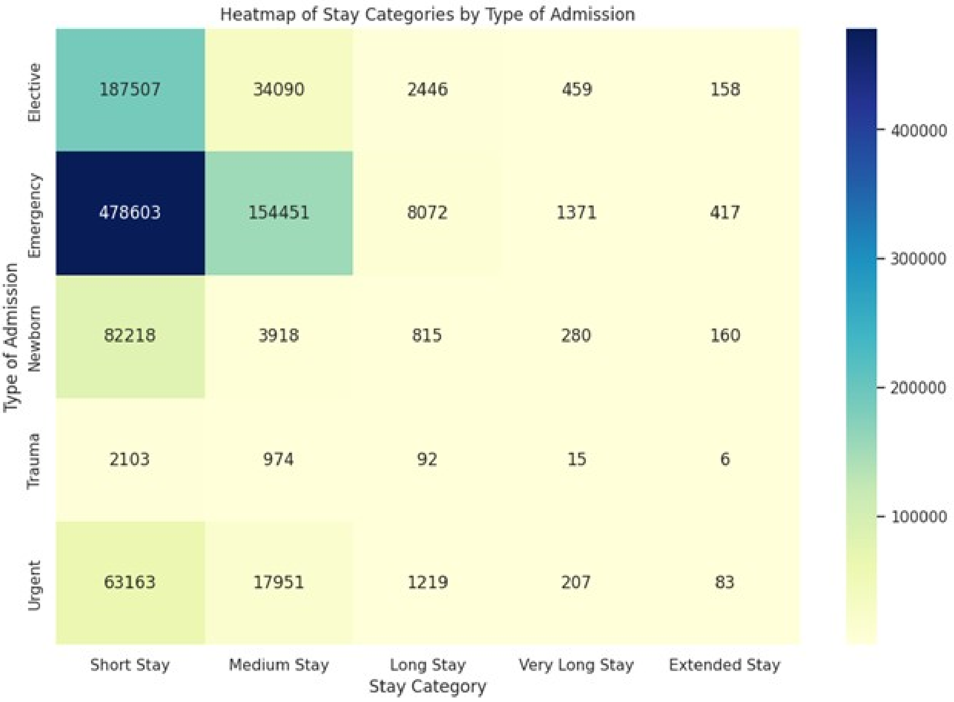

Figure 3.3: stay category by type of admission

Intricate trends in length of stay by type of admission were observed when analysing the type of admission variable in dataset. Here is the key findings: During short-stay cases in elective admissions, 187507 cases were recorded as short stay cases. Lastly there were many cases of medium stays(34,090). Longer ones were however rarely recorded. These are expected because most of the elective procedures require shorter recovery periods.

On the other hand, this was not true for Emergency Admissions. Due to these admissions there was a large number of short stays (478603 cases) and medium stays (154451 cases) but longer stays were noticed as well. This variance indicates that emergency case could vary, including cases that would require prolonged stay in hospitals.

Regarding newborn admissions most were short stays 82218 cases, probably mirroring typical birth circumstances. In contrast, the percentage distribution of long and very long stays was relatively higher than the other admissions type possibly due to complications during labor and or problems relating to prematurity.

4. APR (All Patient Refined) Severity of Illness Code

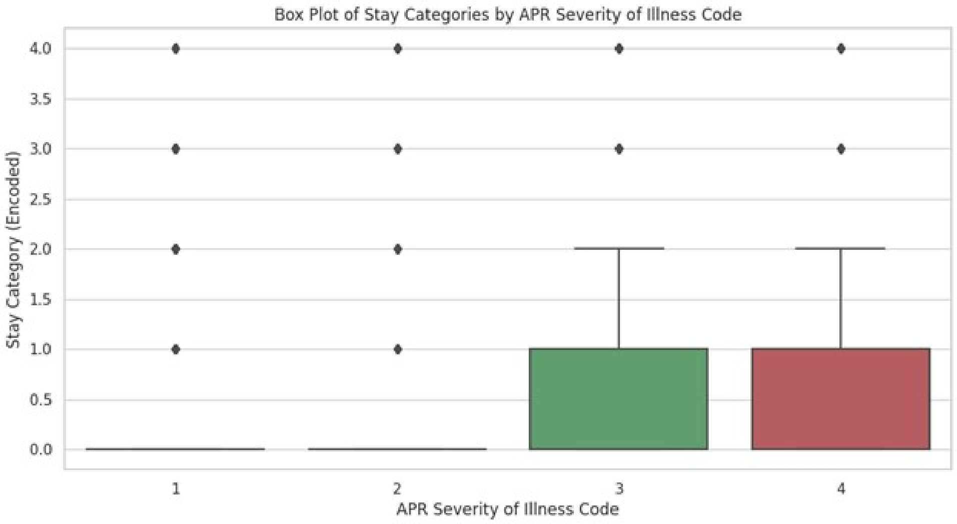

Figure 3.4: Box plot of stay category by APR Severity of Illness Code

The impact of medical severity on the stay categories in was examined through a retrospective comparative study. APR Severity of Illness Code places patients according to how severe and complicated their conditions are. The analysis gave out major trends congruent to expectations concerning patient maintenance and hospitalization lengths. Key findings from each severity level are summarized as follows:

Severity Level 1 (Least Severe): For the second level that had 328,831 cases, there was a propensity for short stays having an average of 0.077 with a small standard deviation of 0.281. Nonetheless, some relatively longer stay cases were observed, which attested the highest score of 4.0.

Severity Level 2: The third category had 403, 191 cases and a longer mean stay duration at 0.183. The mean value was 0.420, which indicated a higher variation across the stay length distribution but the median still remained zero, implying almost exclusively short stays.

Severity Level 3 A level of 243,571 gave an even higher mean stay length of 0.387. This category had a mean standard deviation of 0.539, which indicated that there were some shorter as well as longer stay within the group.

Severity Level 4 (Most Severe): With an average stay duration of 0.818, this group recorded the highest number 65478 cases. This category had a large standard deviation of 0.707 and a median value of 1.0 which mostly represented long stays for patients under study.

5. race and stay category

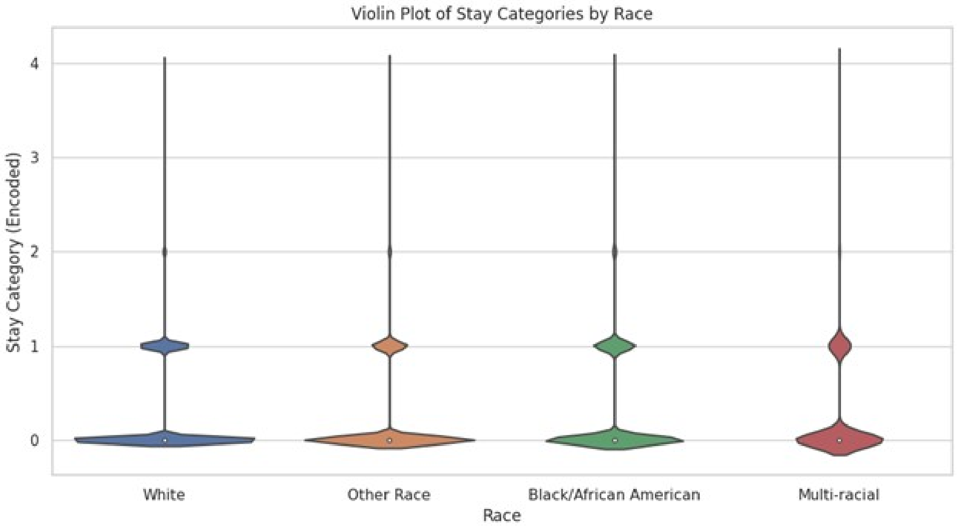

Figure 3.5: stay category by APR Severity of Illness Code

There were patterns relating to race and type of stay for various racial groups during the health experience analyses within the dataset. Key observations include:

Black/African-American Patients: Majority of them did not have prolonged periods in hospital after admission while some had extended stay duration.

Multiracial Patients: The findings of this study indicated that patients who identified themselves as "multiracial" usually remained for considerably shorter periods.

White Patients: White patients were often in and out of the hospital. However, there were some who had moderate to long-term convalescent stages.

However, these differences suggested that most patients irrespective of ethnicity would probably be released after several days. The average stay length was mostly same across the racial groups but the variation within and heterogeneity between these groups indicate that there were other differences in staying duration's, which imply disparities in hospital stays among races.

6. Total Charges and Stay Category

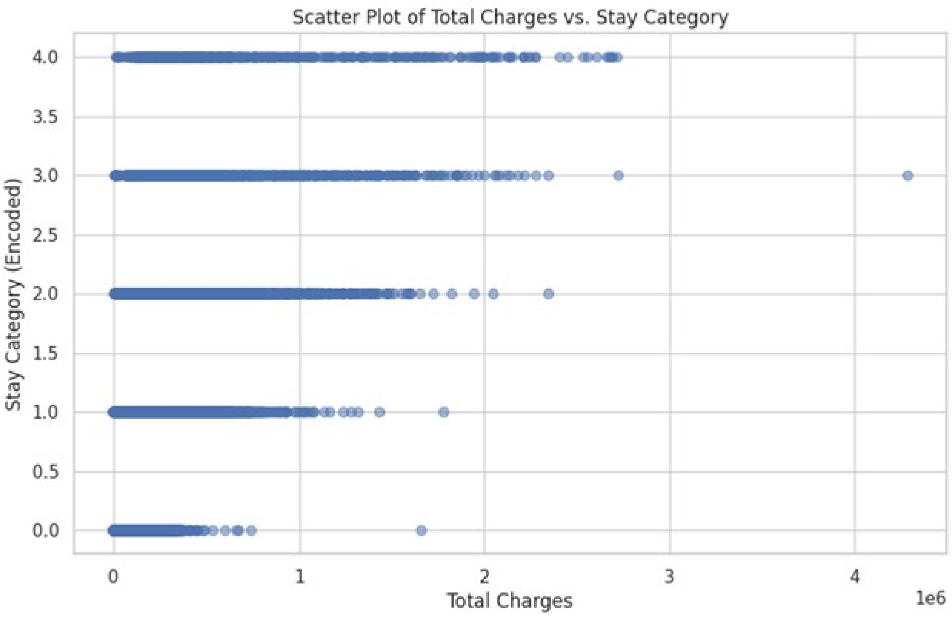

Figure 3.6: total charge vs stay category

The study on hospital data also obtained interesting results related about the causation between "Total Charges"and "Stay Category". The analysis demonstrated a distinct correlation however, one must remember that if a patient stays longer than expected upon admission the average total charge would also have increased in accordance. Therefore, this kind of stay is considered as bit long, although costly because it involves extra and intensive care.

The study showed that numerous charges were charged in relation to a long stay. It refers to various clinical cases and each of these cases comes with its own different management processes, treatment options, or patient's related conditions which also differ. It shows clearly that estimating the health care cost is so complicated due to such diversity.

Right skewness in charges for different forms of stay was a place for these data. In most instances, median charges are lower than average except a few cases where the charges are very high. Such high costs come on average with more complex or customized therapeutic measures. As a result such medical products receive often higher prices for their cost.

For example, there was an excessive fee to cover extended admission in the hospitals. This signifies that various factors such as length of stay are not the largest determiner of total hospital expenses. This involves the condition of illness, complexity of applied treatment, or specific medical device.

Such cost estimation complexities also present challenges to the health care providers who face financial complications on the health systems and patients as well. In summary, good financial planning and provision of good care are prerequisites for sound health care case coordination.

Consequently, the duration of stay in the hospital versus charges associated should no longer be treated as intangible by healthcare providers. They also aid in allocation of resources, patient care planning and management of finances in health care institutions. This underscores that a clear understanding of the dynamics is paramount to provision of quality healthcare, as well as sustenance of the healthcare system.

7. APR Severity of Illness Description in Hospital Admissions

Moderate Severity Dominance: Of these cases, those that are severe but not extreme constitute the largest number of admissions that stand at 403,191 representing 38.73% of the information set. This prevalence implies that most of these hospital-related admissions result from very serious but non-life threatening health conditions. Such patients require extra nursing, which could be done using the traditional hospital's staff.

Substantial Minor Severity Cases: The final category comprises of minor severity cases which account for 31.59%. These include conditions that are not so serious and call for hospitalization but are not meant to lead to a longer stay in the hospital. The

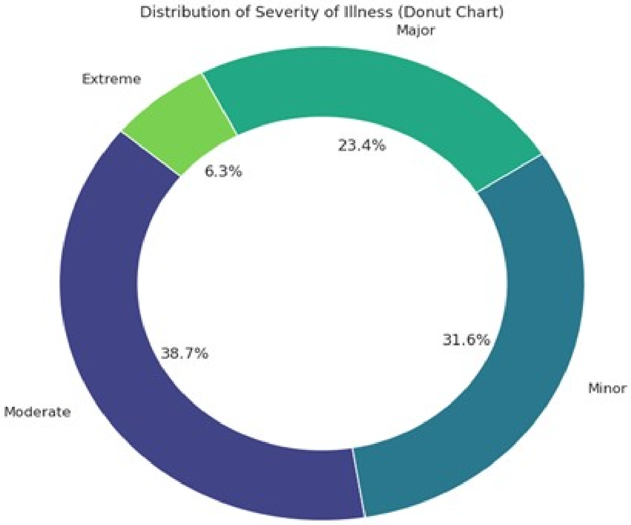

Figure 3.7: APR Severity of Illness Code and Stay Categories

last section looks at the way different kinds and grades of diseases ranging from slight to severe are handled by hospitals simultaneously.

Major Severity Insights: The main severity category has approximately 23.60% of cases, highlighting its importance to hospital data. These often involve circumstances that necessitate expensive in-patient surgical procedures and extended stays in the hospital.

Extreme Severity - The Critical Minority: In this sense, the data set comprises sixty-five thousand, four hundred and seventy eight severe cases which are critical involving just about three one billionth part. Such a status needs special surgical interventions and long-hour rest at the hospital.

8. APR Severity of Illness Description and APR Risk of Mortality

It is important to analyze illness severity across hospital admissions as a way of understanding how hospitals operate. Significant case should be the one that is treated with greater urgency since it is mainly made up of moderate or severe health problems. This forms important input that guides hospitals on how to allocate resources. They also train their staff in preparedness for medical cases. These patterns must be identified and then this will enable hospitals to be efficient in providing care according to patient needs with good and high quality.

Minor Severity and Risk of Mortality:

It is demonstrated through patients belonging to the "Minor" group that these latter

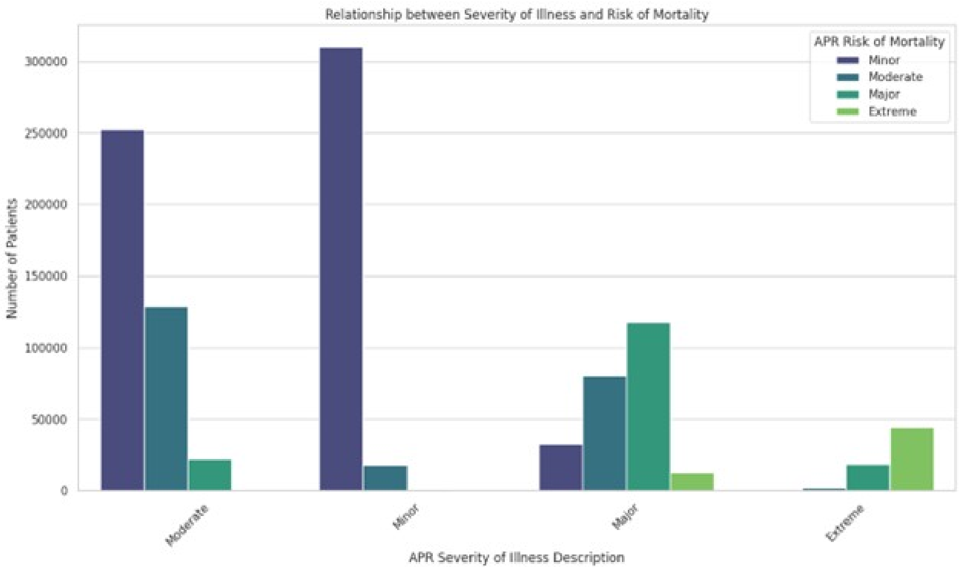

Figure 3.8: relationship between severity illness and Risk of Mortality

individuals represent the "Minor" mortality risk class. This relationship is expected but points out that less serious health problems typically go along with fewer deaths. However, there are some cases even in the moderate major and extreme risk category where disease is very mild but few exceptions take place regardless of the gravity of illness.

Moderate Severity and Its Risks:

Mortality in patients categorized as having moderate severity was inconsistent during patients risk distribution among them. These risks were mostly low risks and then moderately high. This therefore indicates that these moderate health conditions pose several types health conditions. Furthermore some diseases under "major" risks also demonstrate that such illnesses can be aggravated without effective medical care.

Major Severity and Diverse Risks:

A variety of associated mortalities ranged from mild, moderate, and even severe as recorded in the major severity category. The different risk profiles reveal that serious diseases like cancer and cardiac illnesses might not end up having similar results.

Excessively Severe and High Mortality:

The noticeable point was from the "high" severity group and showed strong link to the ′Extreme′ mortality potential. These are highly relevant conditions that indicate this relationship. However, most of the "Major" risk cases also became "extreme" but a very small proportion of "Minor" and "Moderate" risk levels were recorded.

This analysis establishes a clear trend the more deadly a disease is the higher the chance of death. These results underscore importance of appropriate severity ratings in health care with focus on guiding, on control and outcome projections. This shows us the perspective through which the patients of hospitals should be handled, and hence healthcare professionals should take note of this relationship. These interactions can assist in determining the best distribution of the limited medical resources in a manner whereby those with more crucial needs come first. Therefore it is expected that there would be an improvement in health service delivery.

9. APR Medical Surgical Description This analysis seeks to explain why there are variations between medical and surgical cases which eventually translates to high cost of hospital stay length of hospitalization, patient profiles, and types of admissions. The study sheds light on the frequency of various kinds of acute admission and unplanned readmission.

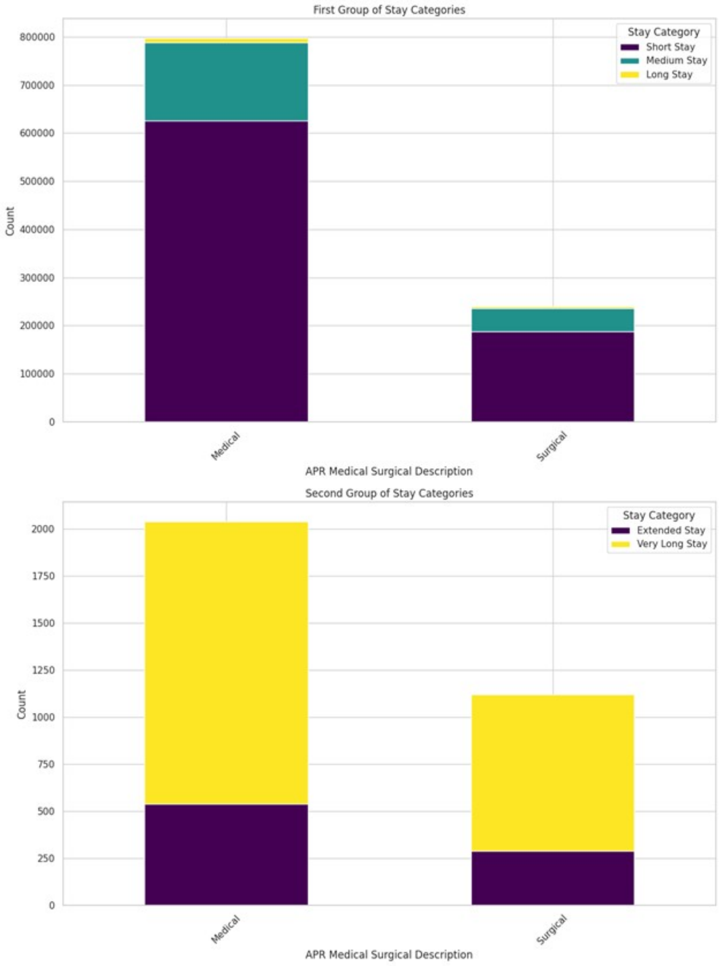

Figure 3.9: Medical and Surgical analysis

Findings on Medical Cases

As shown by the case showing, hospitals admission typically has many cases indicating existence of some health issues among individuals with varying period of stay. For example, the medical cases that occurred among Short stay, Medium stay, and long stay were 625808, 162975, and 8688 respectively. Surprisingly enough the number of medical cases falls into the Extended Stay and Very Long Stay groups For example majority of hospital admissions are caused by multiple diseases where the length of stay at hospital varies between short and long duration.

Insights on Surgical Procedures

Despite the smaller volume of the surgical than that of the medical cases, the latter is considerable in the hospital records. It had 187,991 short and day/short case surgical episodes; 48486 medium stay and 3963 for long stays. Medical cases that indicate an increase in the number of short-stay surgical admission is one such area that has witnessed significant reduction in such patients for surgical admissions that lead to an extended or very long stay. This sequence means the operations may happen in several lengths of stay within hospital setting at a time.

Comparative Analysis and Implications In most cases though, this means that patients have to do a stay of medium or long term at hospital. Nonetheless, this is vital information to hospital managers as well as health care providers. This strategy involves the proper deployment of available human resources in the hospital and fortification of the health care delivery manpower strength. In one study, researchers underlined that such a good health care system should work out ways on how to address many diseases to ensure success. Diseases should be managed more appropriately.

This can also be done in the case of medical and surgical time in finding out if at all there exists recurrent pattern within the medical and surgical case in an effort to meet their demand for limited resources. It has a great significance whereby hospitals need to have important facilities that meet patient's preferences.

3.5 Methodology Execution flow

The study adopted a specific procedure in the selection of models by assessing each model under varied situations with a view to finding out the best one. The approach

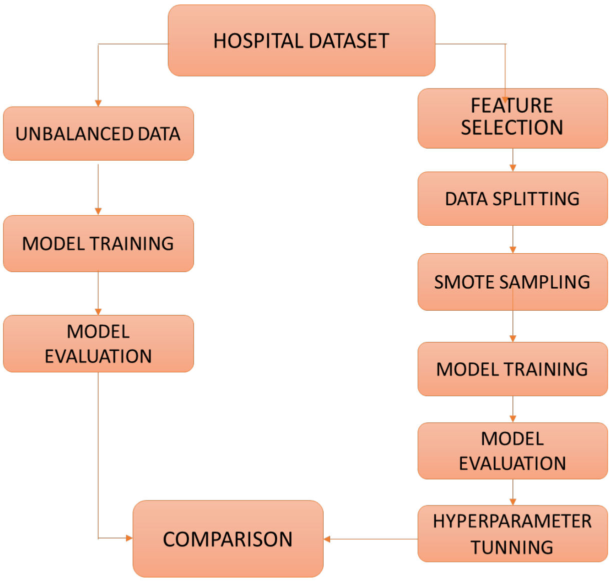

Figure 3.10: Execution Flow

used was systematic in order to undertake a complete examination on the specified model as well as fine-tune it for maximum functionality. The following steps outline the methodology:

1.Initial Model Trials with Unbalanced Data:

Several prediction models applied on the skewed dataset. In this way it assisted in creating an initial benchmark for various models performances at the original data that were unaltered. Nevertheless uneven data can cause bias results but the main aim in this part is just an initial observation of how well the models behave.

2. Feature Selection with SMOTE(Balanced data):

Balanced data followed by feature selection. To be able to achieve valid predictive models in this process, extra features were not needed and hence they had to be removed. Feature subset evaluation reduces over learning, heightens precision and improves efficiency in learning. Furthermore the SMOTE method coupled with the feature selection approach ensured that only key factors were employed during the learning of the models.

3. Hyperparameter Tuning of the Best-Performing Model:

Model among those which demonstrated excellence in preceding stages underwent hyperparameter tuning. Various aspects including varying multiple model parameters in a bid to attain better performance were incorporated. Hyperparameter tuning is part of model optimization and offers a great enhancement in the predictive capacity of a model making it responsive to the data set.

This systematic technique allowed appropriate selection of model that best suited the employed data so as to make accurate and credible forecasts.

3.6 Model training for unbalanced data

Ordinal Logistic Regression:

The use of the Ordinal Logistic Regression for this study was driven by the properties of the data set as well as the type of predictor involved. Hospital stay time is categorical ordinal variable (short, medium, and long). The appropriateness of this type is evidenced by the fact that ordinal logistic regression is designed for these outcomes. [16]

Decision Tree classifier:

Among these algorithms, the decision tree algorithm is unique because of its simple yet effective style of data analysis. Its works by splitting the data according to feature values and building a tree like structure for making decisions. Such a process proves very effective in handling complicated, non linear relationships between the data. In this particular research there were several medical as well as demographic features where the capabilities of the decision tree proved extremely important.

XGBoost classifier:

XGBoost, known for its efficiency and scalability is adept at handling complex datasets. Its design is particularly suited for large datasets with a mix of numerical and categorical variables as often found in healthcare analytics. This algorithm stands out for its ability to intuitively manage diverse data types thus offering a robust approach to analysing hospital inpatient data.

3.7 Model training for balanced data

3.7.1 Feature Selection

feature selection for a dataset with both categorical and numerical features to determine which features have a significant relationship with the target variable.

Initial Feature Assessment and Elimination:

The initial phase involved a thorough evaluation of the dataset to identify and remove variables that functioned as unique identifiers or possessed minimal predictive value. Specifically, this referred to the elimination of variables that constituted noise sources such as ′Health Service Area′ and ′Hospital County′, ′Operating Certificate Number′, ′Facility Name′, ′Abortion Edit Indicator′, ′Discharge Year′, ′Race′.

Application of Random Forest Classifier:

A random forest classifier known for toughness when working with numerous attributes, then was used to assess the importance of every surviving feature in the data set. Following training this model systematically ranked the features and highlighted their individual impacts toward predicting length of stay in hospital. [17]

Dimensionality Reduction:

Finally, some of the least important features were removed to reduce the dimensionality of the data. The first involved pruning of the dataset to 23 relevant features out of an initial larger set which facilitated more focused and relevant model building process during the second stage.

3.7.2 SMOTE: Balancing Data

In the exploratory data analysis phase special attention was focused on the distribution of the target variables, ′Stay_cat_ordinal_data′ and ′Stay Category′. This examination revealed a significant imbalance in the distribution of stay lengths.

To address this the Synthetic Minority Over-sampling Technique (SMOTE) was chosen. SMOTE capability to generate synthetic examples in the minority classes aimed to create a more evenly distributed class representation thus mitigating the limitations of simple oversampling techniques. [18]

Prior to the application of SMOTE the dataset underwent a data splitting process. This ensured that the balancing technique was applied only to the training data maintaining the integrity and distribution of the test data for accurate model evaluation. The application of SMOTE on the training data was designed to replicate the complex patterns present in the minority classes thereby preventing the model from overfitting to the majority class.

SMOTE created synthetic examples in the underrepresented classes (′Medium Stay′, ′Long Stay′, ′Very Long Stay′, and ′Extended Stay′) by interpolating between existing examples. This increased the number of instances in these categories without merely replicating the existing data thus maintaining diversity in the dataset. The newly synthesized samples balanced the class distribution ensuring that each category of hospital stay length was adequately represented. This prevented the models from being biased towards the 'Short Stay' category.

3.7.3 Ordinal Logistic Regression

The first model employed was Ordinal Logistic Regression chosen for its suitability in handling the ordinal nature of the target variable. The balanced dataset enhanced this models ability to differentiate between various stay length categories(short, long, medium) effectively. [19]

The basic equation of ordinal logistic regression is represented as:

$$\log\left(\frac{P(Y > j)}{P(Y \leq j)}\right) = \alpha_j - \beta X \tag{3.1}$$

[20]

Where:

- Y is the ordinal target variable (such as hospital stay length, categorized into ordered groups like short, medium, and long stays).
- j represents each category of the target variable, except for the last one (since it's the reference category).
- $P(Y > j)$ is the probability of the target variable being in a category higher than j.
- $P(Y \leq j)$ is the probability of the target variable being in category j or lower.
- α_j is the threshold or intercept for category j.
- β denotes the coefficients for the predictor variables.
- X is the matrix of predictor variables.

The model's coefficients:

β denotes the degree that an associated value of a certain person's outcome shifts to a specific category. For example, positive coefficients may suggest that some diseases or higher grades of sickness are related with prolonged hospitalization. Thresholds α_j define transitions between different stay lengths, shedding light on the probability distribution across categories.

Applying the predictor to a dataset with used model provided probability estimates to each of length of stay categories. That hierarchical character was also reflected by the threshold and coefficient which were ordered log odds for each of these classes.

They showed how certain factors affected the length of stay in a facility. High values of the positive coefficients meant that the longer the hospitalization was the higher the likelihoods. To understand the distribution of lengths of stay in the hospital data through various points with different probabilities the thresholds of the model will divide the sample into one or the other category.

Lastly Ordinal Logistic Regression aided in examining the ordered categorical data of the 'Length of Stay' parameter. Even though the ordinal nature of such models is highly significant in terms of providing a more detailed picture of the effect exerted by different impacting elements on the LOS such models were not considered while doing such study.

3.7.4 Decision Tree Classifier

According to the Matan Marudi,Irad Ben-GalORCID researchers [21] Ordinal methods are good because they take advantage of information that shows a natural order in class values for ordinal classification problems. Moreover the proposed ordinal decision tree based approaches demonstrated competitive results as compared with existing order methods. Decision trees are highly flexible because they can handle both numeric and categorical data. It played an important role as some of the mixed data types were present within this data set population.

Initially the algorithm identifies an attribute that best fits the data points with respect to the duration they stayed in the hospital or ICU. For example it may regard the 'APR Severity of Illness Code' or 'total cost' as a major predictor and employ this to split the dataset for the first time. This leads to subbranches each branch of which is constituted by segments with common characteristics as regards to the chosen trait.

Since data comprises of categorical and numerical components, the tree splits are designed uniquely for each case. On the other hand categorical data is subdivided in accordance with its separate categories while numerical data could be broken down using certain limits.

Splitting continues and depends among other things upon whether max depth per tree or min samples per node. Each tree has a node or leaves that indicates the forecasted length of stay outcome. The model searches for a leaf in the tree growth from the root and this is after following the specified decision paths of certain feature characteristics on the incoming patient record. Finally the last prediction concerning the patient length of stay is deduced from the leading category among training samples in the leaf node attained.

Relevant Equations:

The Gini Impurity of a dataset D is calculated as:

$$\text{Gini}(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{3.2}$$

[22]

where:

- D is the dataset at the node where the Gini Impurity is being calculated.
- p_i is the proportion of samples that belong to class i at this node.
- n is the number of different classes or labels in the dataset.

Information Gain

The Information Gain $IG(D_p, f)$ for a dataset D_p based on a feature f is defined as:

$$IG(D_p, f) = \text{Entropy}(D_p) - \sum_{j=1}^{m} \frac{|D_j|}{|D_p|} \cdot \text{Entropy}(D_j) \tag{3.3}$$

[22]

where:

- D_p is the dataset at the parent node.
- f is the feature used for the split.
- D_j is the j-th subset of the dataset created after the split.
- $|D_j|$ is the number of samples in the subset D_j.

- $|D_p|$ is the number of samples in the parent dataset D_p.

The Entropy of a dataset D is given by:

$$\text{Entropy}(D) = -\sum_{i=1}^{n} p_i \log_2(p_i) \tag{3.4}$$

[22]

where:

- p_i is the proportion of samples that belong to class i in the dataset D.
- n is the number of different classes or labels in the dataset D.
- m is the number of subsets created from the split.

Interpretation with Dataset:

Gini Impurity:

For instance at a node where there are different lengths of stay patients (short and long) Gini Impurity would be high meaning the different lengths of stay patients are mixed together. Each split intends to cut down on this impurity in the decision tree.

Information Gain:

Entropy change measure of any attribute is estimated by the algorithm whenever data is divided into classes using this trait. Such a example would entail separation through APR severity of illness code, total cost and others with low entropy within the sub clusters translating into narrow distribution on average stay length.

3.7.5 XGBoost Classifier

Finally XGBoost Classifier was applied. However this model, which is known for it is reliability and veracity, performed exceptionally well when a balance was achieved in the data source. XGBOOST building successive trees learning from the mistakes of preceding trees is an intelligent way of developing great predictive models thus highly effective. [23]

Objective Function Setting (objective=′multi:softmax′):

The algorithm is configured with the ′multi: softmax′ objective for multiclass classification. For the purpose of classification this setting is crucial for Xgboost to provide a probability distribution across the different length of stay categories for every patient case.

Classification Categories (numclass=5): This model is customized for five 'numclass' which distinguishes 5 different lengths of hospital stay. The model training process is adjusted to the respective classifications of the dataset which makes all the aspects aligned.

Consistency in Results (random state=42): 'Random state' is set to provide consistent output results in various models executions. It is important to maintain the validity and repeatability of the predictive results in this regard.

XGBoost Equation:

$$\hat{y}_i^{(t+1)} = \hat{y}_i^{(t)} + \eta f_{t+1}(x_i) \tag{3.5}$$

[24]

Where:

- $\hat{y}_i^{(t)}$ is the prediction at iteration t for the i-th instance.
- $\hat{y}_i^{(t+1)}$ is the updated prediction at iteration $t+1$ for the i-th instance.
- η is the learning rate, a factor controlling the step size during the boosting process.
- $f_{t+1}(x_i)$ is the output of the new decision tree at iteration $t+1$ for the i-th instance.

During XGBoosting the hospital stay data model makes an estimated base prediction that is successively tuned upwards. This culminates with the end point of a patient discharge being represented by every new tree that is added to the model addressing each time there is deviation form the earlier assumptions. With respect to prediction, XGBoost shows strong strength due to it is iterative and detail oriented approach that can be associated with different data types without overfitting. This makes it an important asset that is able to handle health data nuances thereby presenting accurate information in healthcare analytics.

The resultant data after balancing and the set of tuned features increased the output for all 3 models (Ordinal Logistic Regression, Decision Tree, and XGBoost). Since these models would now be able to effectively extract the latent trends hidden within the data resulting in improved generalization and overall better results on yet to be seen data. Specifically there were significant developments in terms of class distribution and predictions that were improved with regard to resilience and reliability compared to previous iterations.

3.8 Hyperparameter tunning

Hyperparameter tunning on XGBoost:

The selection of XGBoost for hyperparameter tuning following it is superior performance compared to the other two models Ordinal Logistic Regression and Decision Tree was driven by several compelling reasons. The fact that XGBoost could capture these intricate patterns and relations within complex datasets like with mixture of numerical and categorical variables was facilitated by it is efficiency for handling complex data structures. Regularisation features that helped against overfitting were crucial given this multidimensional complex datasset.

The RandomizedSearchCV approach was utilized to systematically explore and identify the optimal set of hyperparameters. [25] The focus was on parameters such as 'n_estimators', 'learning_rate', 'max_depth', 'gamma', 'subsample', and 'colsample_bytree'.

n_estimators (500): The value of this parameter affects how many tree elements are included into the XGBoost model which contributes greatly to the overall complexity.At 500 trees this balance was achieved offering enough information about sophisticated patterns without overfitting.

learning_rate (0.05): The learning rate determines how fast the model adapts.

A steady though progressing pace of 0.05 was chosen to ensure learning is achieved with precision while at the same time maintaining some level of training efficiency.

max_depth (13): Setting the maximum depth of each tree in the ensemble at 13 made the model robust enough to explore the complex relationships that it was able to detect in the data without overfitting given to the training set.

gamma (0.2): With a regularization parameter gamma=0.2 the model became conservative and thus required loss reduction of order magnitude for any more splits of data.This approach effectively curbed overfitting.

subsample (0.8): By using 80% of the training data for each tree, the subsample parameter enhanced the model's robustness. This sampling strategy fostered a balance between comprehensive learning and overfitting prevention.The adopted sampling strategy helped ensure comprehensive learning without the possibility of overfitting.

colsample_bytree (1.0): This parameter ensured that every feature was utilized at a value of 1.0 for each tree construction process.The predictive accuracy of the model made sense considering that it is based on the full range of available features since the nature of the dataset allowed this.

$$\hat{y}_i(\theta) = \sum_{k=1}^{n} f_k(x_i; \theta) \tag{3.6}$$

[26]

- $\hat{y}_i(\theta)$: Represents the prediction of the XGBoost model for instance i using the set of parameters θ.
- n: The total number of trees in the XGBoost model, corresponding to the 'n_estimators' parameter.
- $f_k(x_i; \theta)$: The output of the k-th tree for instance x_i under the parameters θ. This reflects the contribution of each individual tree in the ensemble to the final prediction.
- k: Index variable representing each tree in the ensemble, ranging from 1 to n.
- x_i: The i-th instance in the dataset for which the prediction is being made.

The optimal parameters of an XGBoost model were determined following a hyperparameter tuning procedure conducted via RandomisedSearch CV.In other words conducted an exhaustive sweep spanning through the defined grid of hyperparameter selections intended for maximization of the model outputs. Number of trees (n_estimators), learning rate (learning_rate), tree depth (max_depth) among other parameters.This was done in order to strike the right compromise for which the model would have neither been underfitted nor overfitted.

Model Evaluation with Optimized Parameters:

The best parameter values found in tuning were used in re-training the XGBoost model which was subsequently applied on the test data set.Evaluation metrics used were Accuracy, Precision, Recall, F1-Score, and Confusion Matrix.This gave an insight about how well the model can be used beyond being able to classify the discharge home versus nursing center stay lengths as correct or incorrect.Performance of the optimized model had to be evaluated to ascertain if it accurately identified classes and balanced the trade-off between precision and recall.

CHAPTER 4

Results

4.1 Findings from Data Pre-processing

The analysis showed that there were 2329 missing data in important columns which included, 'Health care Service Areas', 'Hospital County', 'Operating Certificate Number' and 'Facility ID'. There was also a large percentage of missing values 15,920 for the zip code 3 digit category. A substantial portion of the dataset lacked information in payment related categories with nearly 398071 values in Payment Typology 2 and 761102 values in Payment Typology 3 . also the duplicate found was 4234 after dropping these columns and duplicate values dataset columns became 31 and 1041939 records.

4.1.1 Data encoding

A systematic encoding was undertaken for the 'type of admission' variable. Through a sequential or alphabetic approach each category was identified with an exclusive numeric code. The encoding assigned numeric codes 0 for Emergency 1 Elective, 2 Newborn, 3 Trauma Center, and 4 Urgent. similarly and all the other categorical columns was encoded.

4.1.2 Feature Engineering

'Length of Stay' statistics were clustered according to the number of days the patient spent in hospital. These categories were defined as follows:

Short stay: 0-7 days

Medium stay: 7-30 days

Long stay: 30-60 days

Very long stay: 60-90 days

Extended stay: 90-120 days

These categories were used to create a new column called 'Stay Category'. The breakdown of this category was necessary to transform "length of stay" from a continuous unit to grouped units for classification analysis.Again the column were introduced which was encoded data of this 'saty category' column named Stay_cat_ordinal_data.

summarising this The 'Length of stay' originally recorded staying duration continuously The categories were defined and encoded 0 Short Stay 1 Medium Stay, 2 Long Stay, 3 Very Long Stay and 4 Extended Stay.

4.2 Model Results for unbalanced data

4.2.1 Ordinal Logistic Regression

The evaluation results of the Ordinal Logistic Regression model, applied to the dataset without balancing.

Category	Precision	Recall	F1-score	Support
Short Stay	0.81	1.00	0.89	244185
Medium Stay	0.67	0.06	0.10	63389
Long Stay	0.33	0.26	0.29	3799
Very Long Stay	0.17	0.28	0.21	696
Extended Stay	0.20	0.48	0.28	253
Accuracy	0.80			
Macro Avg	0.44	0.41	0.35	312322
Weighted Avg	0.77	0.80	0.72	312322

Table 4.1: Evaluation Metrics for Ordinal Logistic Regression Model Without Balancing

Analysis of the predictive ability of the Ordinal Logistic Regression model shows that it is more accurate than other predictive variables with an accuracy of 79.51% and excellent class specific measures for the Short Stay category. The model work well except with the long term stay categories such as 'Medium Stay', 'Long Stay', 'Very Long Stay' and 'Extended Stay'. Precision and recall are lower in these cases. This variation in performance across different stay lengths is further elucidated by the macro and weighted averages 'Macro Averages' (Precision = 0.44; Recall = 0.41; F1-Score = 0.36) represent moderate performance in every class while the 'Weighted Averages' (Precision = 0 However the values for 'Support' confirm the imbalanced nature of data set which is a major reason for bias to model towards better prediction of the most common 'Short stay category'.

4.2.2 Decision Tree classifier

Stay Category	**Precision**	**Recall**	**F1-Score**	**Support**
Short Stay	0.95	0.95	0.95	244185
Medium Stay	0.79	0.79	0.79	63389
Long Stay	0.58	0.59	0.59	3799
Very Long Stay	0.46	0.48	0.47	696
Extended Stay	0.51	0.42	0.46	253
Accuracy	0.91			
Macro Avg	0.66	0.64	0.65	312322
Weighted Avg	0.91	0.91	0.91	312322

Table 4.2: Evaluation Metrics for Decision Tree Classifier

The Decision Tree classifier proved relatively accurate at forecasting LOS among various groups. This had an overall accuracy rate of approximately 91.03%, showing the system capabilities at differentiating between number of days. The model had really high scores in precision and recall (0.95) for the 'Short Stay' category producing a very high F1-score of 0.95. It was highly commendable as precision and recall were at 0.79 while the F1 score was of the same magnitude for the 'Medium Stay' category. The predictive capacity as observed for longer stay categories that include 'Long Stay', 'Very Long stay' as well as 'Extended stay' was moderate and involved precision and recalls varying from 0.46 to 0.58 and F1-scores of 0.46 This indicates that medium and longer stays are more effectively identified as opposed to shorter stays, albeit with some scope for further refinement in the less frequent stay groups.

Compared to the Ordinal Logistic Regression model the Decision Tree classifier exhibited substantial advancement in it is ability to predict hospital stay lengths. While the Ordinal Logistic Regression model had a overall accuracy of approximately 79.51% the Decision Tree improved this metric significantly to around 91.03%. The Decision Tree also showed more balanced performance across different stay categories particularly in identifying medium to longer stays where it outperformed the previous model in both precision and recall. The macro and weighted average scores for Decision Tree were notably higher indicating more consistent and reliable prediction across all classes as opposed to the Ordinal Logistic Regression model which was more skewed towards predicting the majority class accurately.

4.2.3 XGBoost Classifier

Category	**Precision**	**Recall**	**F1-Score**	**Support**
Short Stay	0.96	0.97	0.97	244185
Medium Stay	0.85	0.85	0.85	63389
Long Stay	0.71	0.64	0.68	3799
Very Long Stay	0.61	0.44	0.51	696
Extended Stay	0.68	0.38	0.49	253
Accuracy	0.94			
Macro Avg	0.76	0.66	0.70	312322
Weighted Avg	0.94	0.94	0.94	312322

Table 4.3: Evaluation Metrics XGBoost Model

Compared to the Ordinal Logistic Regression and Decision Tree models XGBoost stands out with it is significantly higher overall accuracy and balanced performance across different stay lengths. The Ordinal Logistic Regression model while achieving a decent overall accuracy of approximately 79.51% showed a marked deficiency in correctly identifying medium, long, and extended stays. The Decision Tree model improved upon this with an overall accuracy of 91.03% offering better balance across different categories. However it still lagged behind XGBoost, particularly in predicting less frequent stay lengths. XGBoost's superiority is evident in it is higher macro and weighted average scores indicating it is enhanced ability to handle class imbalance and provide more reliable predictions across all categories. This comparative analysis highlights XGBoost effectiveness in dealing with complex, imbalanced datasets, making it a more suitable choice for predicting hospital stay lengths in this study.

4.3 Feature selection

Some of the columns like ′Health Service Area′,′Hospital County′,′Operating Certificate Number′,′Facility Name′,′Abortion Edit Indicator″Discharge Year′,′Race′,′Ethnicity′ dropped as they are unique identifiers. and the data reduced to 25 columns and 1041071 rows.

Insights from Feature Importance Analysis:
This highlighted that among those selected using random forest feature selection method ′total costs′ accounted for 27.74% whereas ′total charges′ amounted to 18.89%. This strongly implied relationship between length of stay in hospital and it is financial components. Another important contributory variables included ′Facility Id′ and ′CCS Diagnosis Code′. It highlighted the impact of health care facilities and diagnostic groups on the length of stay in hospitals. Additionally APR DRG code and severity of illness was a major determining factor suggesting the impact of diagnosis severity and type. However the variables like Emergency Department Indicator 0.012 . Minimal effect was experienced by the model in terms of prediction on birth weight variables such as 0.00 and 0.02 respectively. hence it was dropped.

4.4 SMOTE

The distribution of target variable that is Stay Categories before SMOTE was as follows:

Category	Percentage
Short Stay 0	78.2562%
Medium Stay 1	20.2311%
Long Stay 2	1.2104%
Very Long Stay 3	0.2233%
Extended Stay 4	0.0790%

Table 4.4: Distribution of Hospital Stay Categories

Given this imbalance, it became evident that balancing the dataset was essential.After SMOTE the distribution became:

Category	data Count
Short Stay 0.0	654413
Medium Stay 1.0	654413
Long stay 2.0	654413
Very Long Stay3.0	654413
Extended Stay4.0	654413

Table 4.5: Count of Data Samples per Category After Balancing

Now, the class distribution for each class in the target variable is equal to 654413 making it balanced. By having a equal number of records of each class present in the training data, the model will not be bias towards a specific classification during execution.

4.5 Model Results for Balanced data

4.5.1 Ordinal Logistic Regression Model (Post-Balancing)

The balanced model shows nearly the same work output throughout the different stay lengths. This reduced their general accuracy but enormously boosted the ability of the

Category	**Precision**	**Recall**	**F1-score**	**Support**
0.0	0.88	0.63	0.74	244185
1.0	0.25	0.38	0.30	63389
2.0	0.06	0.57	0.10	3799
3.0	0.17	0.49	0.25	696
4.0	0.22	0.44	0.29	253
Accuracy	0.58			
Macro Avg	0.31	0.50	0.34	312322
Weighted Avg	0.74	0.58	0.64	312322

Table 4.6: Classification Report of Ordinal Logistic Regression Model (Post-Balancing)

model to point out those rare stay lengths (Medium Stay, Long Stay, Very Long and Extended Stays). This can be seen in higher recall values for those categories, indicating a more accommodating model that is appropriate for datasets having different classes of hospital stay durations.

4.5.2 Decision Tree Classifier Results (Post-Balancing)

Category	Precision	Recall	F1-Score	Support
Short Stay (0)	0.95	0.94	0.94	244185
Medium Stay (1)	0.75	0.78	0.77	63389
Long Stay (2)	0.54	0.58	0.56	3799
Very Long Stay (3)	0.44	0.50	0.47	696
Extended Stay (4)	0.43	0.42	0.42	253
Accuracy	0.90			
Macro Avg	0.62	0.64	0.63	312322
Weighted Avg	0.90	0.90	0.90	312322

Table 4.7: Decision Tree Classifier Results (Post-Balancing)

The decision tree classifier gave commendable results with a overall accuracy approximating at 90.03%. This indicates that the precision for balance has fallen slightly compared to the balanced model with about 91.03% accuracy. However the balanced model was more effective in categorising various types of stay.

For example the precision and recall in the 'Medium Stay', 'Long Stay', 'Very Long Stay', and 'Extended Stay' categories witnessed noticeable improvements. This is evident in the macro average precision, recall, and F1-score, which are approximately 0.62, 0.64, and 0.63 respectively. These scores when compared to the unbalanced model macro averages of 0.66, 0.64, and 0.65 suggest slightly more uniform performance across all classes in the balanced model.

4.5.3 XGBoost Classifier Results (Post-Balancing)

Category	Precision	Recall	F1-Score	Support
0.0 (Short Stay)	0.96	0.95	0.96	244185
1.0 (Medium Stay)	0.81	0.82	0.82	63389
2.0 (Long Stay)	0.53	0.73	0.61	3799
3.0 (Very Long Stay)	0.45	0.60	0.51	696
4.0 (Extended Stay)	0.44	0.58	0.50	253
Accuracy		0.9232		
Macro Avg		0.64	0.74	0.68
Weighted Avg		0.93	0.92	0.92

Table 4.8: XGBoost Classifier Results (Post-Balancing)

The performance of the XGBoost classifier on a balanced dataset improved considerably and achieved about 92.32% accuracy. It shows it predicts effectively for any hospital length of stay category. The model had an over 95% precision and recall for predicting short stays whereas it was very robust in the medium stay category. It also achieved notable improvement in identifying rare stay lengths such as long, very long, and extended stays by up to 44% precision and recall of 58%. Macro and weighted average scores show that in general the model predicted more successful outcomes than unsuccessful ones for every class confirming the predictability of the model in a balanced dataset.

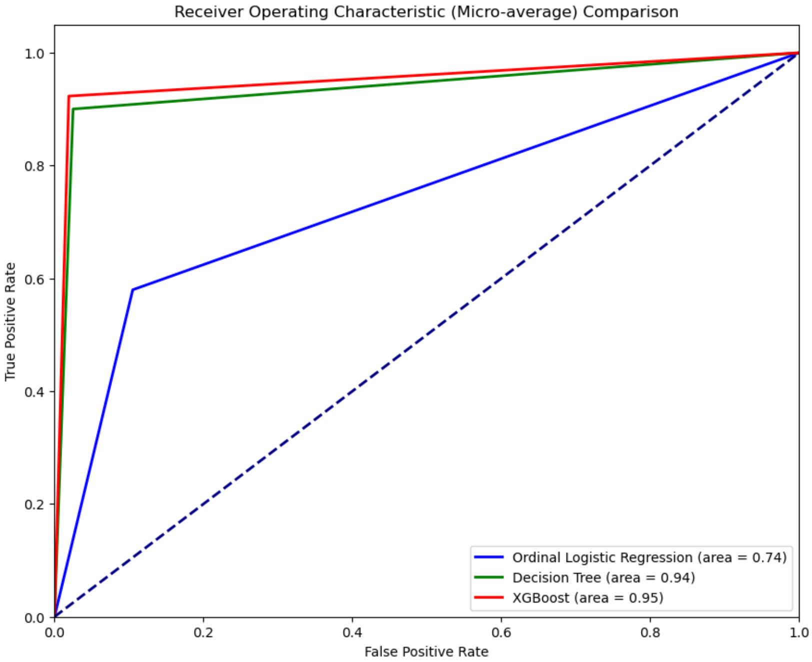

Figure 4.1: Receiver Operating Characteristic (Micro-average) Comparison

4.5.4 ROC Plot

In the evaluation of our multi class classification problem, encompassing five distinct classes, the micro-average technique was used to construct ROC curves for a comprehensive analysis. This approach ensures a balanced representation of model performance across all classes particularly beneficial given the potential class imbalance.

The comparative analysis revealed that the XGBoost model outshines it is counterparts with a micro average AUC of 0.95 indicating a superior predictive power. It is ROC curve closely approaches the ideal top left corner of the ROC space denoting high true positive rate coupled with low false positive rate. These characteristics suggest that XGBoost is exceptionally adept at discerning between the classes affirming it is efficacy and robustness for our classification task.

4.5.5 Hyperparameter tunning on XGBoost

Hyperparameter tunning helped to find out the best parameter for XGBoost those are as follows:

Parameter	**Value**
subsample	0.8
n_estimators	500
max_depth	13
learning_rate	0.05
gamma	0.2
colsample_bytree	1.0
Best F1 score	0.9596

Table 4.9: Best parameters for the XGBoost model

Model performance after hyperparameter tunning:

Class	Precision	Recall	F1-Score	Support
0.0	0.97	0.97	0.97	244,185
1.0	0.86	0.87	0.86	63,389
2.0	0.69	0.69	0.69	3,799
3.0	0.59	0.56	0.58	696
4.0	0.59	0.55	0.57	253
Accuracy		0.94		
Macro Avg	0.74	0.73	0.73	312,322
Weighted Avg	0.94	0.94	0.94	312,322

Table 4.10: XGBoost Classifier Results with hyperparameter tuning

XGBoost model post hyperparameter tuning demonstrates a significant enhancement in predicting hospital stay lengths. With an overall accuracy of 94% the model shows a marked proficiency in identifying short stays (category 0.0) as evidenced by the high precision and recall of 0.97. Notably there is considerable improvement in the precision and recall for medium to long stays (categories 1.0 to 4.0) indicating the model increased sensitivity and specificity in these more challenging categories. The F1 scores across all categories are consistently high reflecting balanced performance between precision and recall. The macro average of 0.74 and weighted average of 0.94 further illustrate the model effectiveness both in terms of equitable treatment across different categories and emphasis on the more represented classes.

Confusion Matrix for XGBoost:

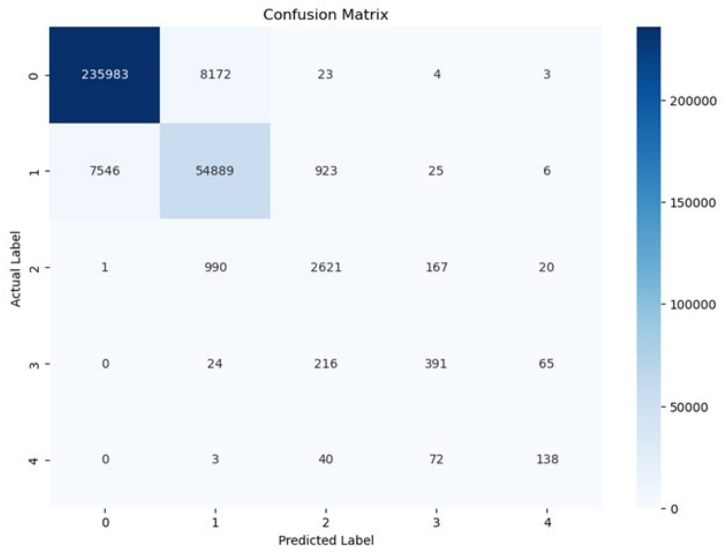

Figure 4.2: Confusion Matrix for XGBoost n

A confusion matrix shows that the XGBoost model had a large number of true positive instances for class 0, which constituted 235983 as compared to negligent proportions of false positive and false negative observations. Class 1 has a high classification accuracy of 54889 true positive but it is notable that class 0 has a misclassification rate. When shifting to Class 2 accuracy becomes poor with a high occurrence of instances classified as classes 1 and 3391 true positives indicate poor class 3 prediction while distinguishing between these two classes proves to be a challenge for the model. The least accurately predicted class is that is class 4 which has 138 true positives, suffers from major misclassification especially from class 2 and class 3 respectively. Darker colors in the matrix color gradient imply that the model is very well predicting on class 0 and class 1 whereas lighter colors indicate that the model needs improvement to reduce misclassification of class 2 to class 4.

4.5.6 Comparative analysis of all model

Approach/Model	Overall Accuracy	Macro Avg Precision	Macro Avg Recall	Macro Avg F1-score	Weighted Avg Precision	Weighted Avg Recall	Weighted Avg F1-score
Without Balancing							
Ordinal Logistic Regression	0.80	0.44	0.41	0.36	0.77	0.80	0.72
Decision Tree	0.91	0.66	0.64	0.65	0.91	0.91	0.91
XGBoost	0.94	0.76	0.66	0.70	0.94	0.94	0.94
With SMOTE Balancing							
Ordinal Logistic Regression (Unbalanced)	0.58	0.31	0.50	0.34	0.74	0.58	0.64
Decision Tree (Unbalanced)	0.90	0.62	0.64	0.63	0.90	0.90	0.90
XGBoost (Unbalanced)	0.92	0.64	0.74	0.68	0.93	0.92	0.92
XGBoost (Hyperparameter Tuning)	**0.94**	**0.74**	**0.73**	**0.73**	**0.94**	**0.94**	**0.94**

Table 4.11: Comparative Results analysis

Machine learning models applied to hospital length of stay prediction showed better performance as compared to other models particularly using Xgboost with hyperparameter tuning that led to the highest accuracy and balanced classification metrics both before and after SMOTE balancing. The smote balance improved the macro recall that is the detection rate across minority classes but it reduced slightly the global accuracy for some models such as the ordinal logistic regression. The decision tree remained consistent for balanced and unbalanced datasets. These results stress the efficacy of ensemble approaches as well as the role of hyper-tuning to maximize precision and recall during prediction modelling.

CHAPTER 5

Conclusions and Future scope

This is a project which aimed at use of advanced machine learning techniques to predict hospitals LOS a crucial issue in hospital resources allocation and patient care delivery improvement. Predicting LOS is difficult process because various illnesses and diverse responses to treatment affect patients differently. While traditional statistical methods provide a very good overview, they seldom take into consideration the uniqueness of each individual case.

One unique feature of this study was that the LOS data used were ordinal in nature. LOS as a continuous variable was grouped into discrete classes Short, medium, long, very long, and extended stays. This order classification captures the fact that there is a natural order of length of hospital stay which must be considered when developing fine tuned prediction model.

Therefore, using several machine learning models such as Decision Tree, XGBoost and Ordinal Logistic Regression, we resolved this ordinal data structure. Techniques such as "SMOTE" the Synthetic Minority Over sampling Technique were highly vital for handling the problem of class imbalance since every LOS group should be well accommodated into the dataset.

The models were examined and measured against each other based on the measures of accuracy, precision, recall and F1 score. Through these evaluations the strong points and weak points of these models were brought up with XGBoost turning out to be the best considering that it had been tuned for hyperparametrs particularly in handling the ordinal nature of the LOS data.

This project demonstrates how the application of machine learning improves hospital length of stay prognosis for ordinal data. The implementation of the project that saw it change from an all inclusive approach to a target oriented evidence based process led to improvement on the reliability of LOS estimates as well as making it possible for doctors to design specific strategies of treating patients and for better utilisation of available resources.

Looking ahead,approaches and ideas are an action plan for future enhancements of healthcare analytics with respect to ordinal data. Further studies could concentrate building complex tailor made models developed for ordinal classification that could include neural networks optimal in ordinal outcomes. Real time data analysis for dynamic LOS predictions is necessary to improve timeliness and accuracy to changeable healthcare environment. These ordinal data methods can be used in different aspects including symptom severity and recovery stages that will help improve patient centric healthcare strategies. In addition it is important to investigate ethically and legally about use ordinal data in health care practice so that developments in predictive analytics would comply with regulations and moral codes. By adopting this holistic approach it is expected to boost the productivity in health sector resulting improved specialized care as well as optimal resource utilization.

Bibliography

[1] Garg L., McClean S.I., Barton M., Meenan B.J., and Fullerton K. *Intelligent patient management and resource planning for complex, heterogeneous, and stochastic healthcare systems*
IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans 42, pp. PLoS ONE, 1332-1345. 10.1109/TSMCA.2012.2210211, 2012.

[2] H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, *Analysis of length of hospital stay using electronic health records: A statistical and data mining approach PLoS ONE*, vol. 13, no. 4, 2018. https://doi.org/10.1371/journal.pone.0195901

[3] Md. Mahbubur Rahman, Dipanjali Kundu, Sayma Alam Suha, Umme Raihan Siddiqi, and Samrat Kumar Dey, "Hospital patients length of stay prediction: A federated learning approach" *ResearchGate*, 2022. https://www.researchgate.net/publication/362215890_Hospital_patients'_length_of_stay_prediction_A_federated_learning_approach

[4] Yong Chen, "Prediction and Analysis of Length of Stay Based on Nonlinear Weighted XGBoost Algorithm in Hospital," *Journal of Healthcare Engineering*, vol. 2021, Article ID 4714898, 2021. https://www.researchgate.net/publication/356697336_Prediction_and_Analysis_of_Length_of_Stay_Based_on_Nonlinear_Weighted_XGBoost_Algorithm_in_Hospital

[5] Jan Chrusciel, FranÃ§ois Girardon, Lucien Roquette, David Laplanche, Antoine Duclos, and StÃ©phane Sanchez, "The prediction of hospital length of stay using unstructured data," *BMC Medical Informatics and Decision Making*, vol. 21, no. 1, Dec. 2021. https://www.researchgate.net/publication/

357153452_The_prediction_of_hospital_length_of_stay_using_unstructured_data

[6] Zhixu Hu, Hang Qiu, Liya Wang, and Minghui Shen, “Network analytics and machine learning for predicting length of stay in elderly patients with chronic diseases at point of admission,” *ResearchGate*, 2022. https://www.researchgate.net/publication/359156896_Network_analytics_and_machine_learning_for_predicting_length_of_stay_in_elderly_patients_with_chronic_diseases_at_point_of_admission

[7] Tahani A. Daghistani, Radwa Elshawi, Sherif Sakr, Amjad M. Ahmed, Abdullah Al-Thwayee, and Mouaz H. Al-Mallah, “Predictors of in-hospital length of stay among cardiac patients: A machine learning approach,” *ResearchGate*, 2019. https://www.researchgate.net/publication/330503464_Predictors_of_in_hospital_length_of_stay_among_cardiac_patients_A_machine_learning_approach

[8] Mustafa Gokalp Ataman, Gorkem, “Predicting waiting and treatment times in emergency departments using ordinal logistic regression models,” *ResearchGate*, 2021. https://www.researchgate.net/publication/350061594_Predicting_waiting_and_treatment_times_in_emergency_departments_using_ordinal_logistic_regression_models

[9] https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/82xm-y6g8

[10] Natarajan, K., Li, J., Koronios, A. (2010). Data mining techniques for data cleaning. In: Kiritsis, D., Emmanouilidis, C., Koronios, A., Mathew, J. (eds) Engineering Asset Lifecycle Management. Springer, London. https://doi.org/10.1007/978-0-85729-320-6_91

[11] Jason Poulos and Rafael Valle (2018) Missing Data Imputation for Supervised Learning, Applied Artificial Intelligence, *PLoS ONE*, 32:2, 186-196, DOI: 10.1080/08839514.2018.1448143

[12] https://machinelearningknowledge.ai/categorical-data-encoding-with-sklearn-labelencoder-and-onehotencoder/

[13] OLLE ANDERSSON. "Predicting Patient Length Of Stay at Time of Admission Using Machine Learning." http://www.diva-portal.org/smash/get/diva2:1338294/FULLTEXT01.pdf

[14] https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

[15] Chang WF, Yan XY, Ling H, Liu T, Luo AJ. A study of the types and manifestations of physicians' unintended behaviors in the DRG payment system. Front Public Health. *PLoS ONE*, 2023 Jun 27;11:1141981. doi: 10.3389/fpubh.2023.1141981. PMID: 37441652; PMCID: PMC10333571.

[16] Bender R, Grouven U. "Ordinal logistic regression in medical research." J R Coll Physicians Lond *PLoS ONE*, .1997 Sep-Oct;31(5):546-51. PMID: 9429194; PMCID: PMC5420958.

[17] DataThanh-Tung Nguyen,oshua Zhexue Huang, and Thuy Thi Nguyen. "Unbiased Feature Selection in Learning Random Forests forHigh-Dimensional" https://www.researchgate.net/publication/264868560_Unbiased_Feature_Selection_in_Learning_Random_Forests_for_High_Dimensional_Data

[18] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, 2008, pp. *PLoS ONE*, 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[19] Lelisho ME, Wogi AA, Tareke SA. Ordinal Logistic Regression Analysis in Determining Factors Associated with Socioeconomic Status of Household in Tepi Town, Southwest Ethiopia. ScientificWorldJournal. *PLoS ONE*, 2022 Feb 3;2022:2415692. doi: 10.1155/2022/2415692. PMID: 35153626; PMCID: PMC8831068.

[20] https://www.norusis.com/pdf/ASPC_v13.pdf

[21] Matan Marudi, Irad Ben-Gal and Gonen Singer (2022) *"A decision tree-based method for ordinal classification problems" PLoS ONE*, IISE Transactions, DOI: 10.1080/24725854.2022.2081745 https://www.tandfonline.com/action/showCitFormats?doi=10.1080%2F24725854.2022.2081745

[22] https://towardsdatascience.com/decision-trees-explained-entropy-information-gain-gini-index-ccp-pruning-4d78070db36c

[23] Yuval Barak-Corren, Pradip Chaudhari, Jessica Perniciaro, Mark Waltzman, Andrew M. Fine and Ben Y. Reis. "Prediction across healthcare settings: a case study in predicting emergency department disposition." https://www.researchgate.net/publication/357082726_Prediction_across_healthcare_settings_a_case_study_in_predicting_emergency_department_disposition

[24] XGBoost: "A Scalable Tree Boosting System" https://dl.acm.org/doi/pdf/10.1145/2939672.2939785

[25] "hyperparameter tuning in xgboost using randomizedsearchcv" https://jayant017.medium.com/hyperparameter-tuning-in-xgboost-using-randomizedsearchcv-88fcb5b58a73

[26] Tianqi Chen and Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, (2016; 29 Jan. 2016) https://ar5iv.labs.arxiv.org/html/1603.02754