Project Summary

Batch details	PGP-DSE Apr’23 Gurgaon
Team members	Rishi Tanwar, Neeraj Meena, Ayush Singh Verma, Sarthak Barnwal, Harsh Tyagi
Domain of Project	Forecasting Hospital Stay Durations: A Predictive Modeling Approach for Enhanced Patient Care
Proposed project title	Healthcare Management
Group Number	Group-6
Team Leader	Sarthak Barnwal
Mentor Name	Ms. Prachi Tare

Date: 26-December-2023

PRACHI TARE

Signature of Mentor

SARTHAK BARNWAL

Signature of Team Leader

Table of Contents

Project Details

OVERVIEW

The healthcare management project aims to optimize healthcare delivery by implementing an integrated system. This initiative focuses on enhancing patient care through streamlined processes such as efficient appointment scheduling, secure patient record management, and automated billing. Leveraging advanced technology, the project promotes seamless communication among healthcare providers, ensuring a holistic and patient-centric approach. Additionally, it addresses data interoperability challenges to foster collaboration among different healthcare entities. The incorporation of artificial intelligence enables predictive analytics, facilitating resource optimization and reduced wait times. Compliant with healthcare regulations, the project prioritizes data security to maintain patient confidentiality. Ultimately, the healthcare management solution promises to elevate operational efficiency, improve patient outcomes, and contribute to the overall advancement of healthcare services.

BUSINESS PROBLEM STATEMENT:

"Forecasting Hospital Stay Durations: A Predictive Modeling Approach for Enhanced Patient Care"
Recent Covid-19 pandemic has raised alarms over one of the most overlooked areas to focus: healthcare management. while healthcare management has various use cases for using data science, patient length of stay is one cd predict if one wants to improve the efficiency of the healthcare management in a hospital, this parameter helps hospitals to identify patients of high loss risk at the time of admission and prior knowledge of loss can aid in logistics such as room and bed allocation planning and task to accurately predict the length of stay for each patient on case-by-case basis so that the hospitals in a professional and optimal number

What would we achieve by this project ?

Healthcare facilities, including hospitals, often face the challenge of efficiently managing their resources to provide optimal patient care. One critical aspect of this challenge is predicting the duration of a patient's hospital stay. Accurately estimating how long a patient is likely to remain in the hospital can significantly impact resource allocation and operational efficiency.

Resource Optimization:

Predicting hospital stay durations enables healthcare facilities to optimize the allocation of resources such as hospital beds, medical staff, and equipment. By understanding how long a patient is expected to stay, hospitals can strategically plan for the required resources, preventing underutilization or overcommitment.

Staffing Levels:

Staffing levels play a crucial role in delivering quality healthcare. Knowing the expected duration of a patient's stay allows hospitals to adjust staffing levels accordingly. For example, if a surge in patient admissions with shorter expected stays is predicted, the hospital can optimize its nursing and support staff schedules.

Operational Efficiency:

Efficient operations contribute to improved patient experiences and outcomes. Predictive models for hospital stay duration can be integrated into the overall hospital management system to streamline processes. This includes discharge planning, scheduling follow-up appointments, and coordinating post-discharge care.

Bed Utilization:

Hospital bed availability is often a bottleneck in healthcare systems. Predicting how long a patient will occupy a bed facilitates better bed management. This knowledge can be instrumental in reducing wait times for incoming patients, minimizing congestion in emergency departments, and ensuring timely admission for those in need.

Patient-Centric Care:

Accurate predictions contribute to patient-centric care by allowing healthcare providers to communicate more effectively with patients and their families. When patients are informed about the expected duration of their hospital stay, it helps manage expectations and plan for post-discharge activities and support.

Financial Impact:

Efficient resource allocation and reduced length of stay can have positive financial implications for healthcare institutions. Predictive models that contribute to shorter hospital stays can lead to cost savings and increased overall financial sustainability.

In summary, developing a predictive model for hospital stay duration addresses a critical business problem by enhancing resource management, improving operational efficiency, and ultimately delivering better patient care. The integration of such a model into healthcare systems can lead to more informed decision-making, positively impacting both the quality of care and the financial health of the institution.

What are the limitation of this model ?

Our model is not a generalized model but a business oriented, rreealistic and specific model -meaning neeeds to be calibrated and updated with time and business.

Data Quality and Availability:

The accuracy of the predictive model heavily relies on the quality and availability of historical patient data. If the data used for training the model is incomplete, outdated, or contains errors, it can negatively impact the model's performance.

Complexity of Healthcare Variables:

Healthcare is a complex field with various influencing factors. Predicting hospital stay duration involves considering a multitude of variables, such as the patient's medical history, severity of illness, and response to treatment. Some factors may be difficult to quantify or predict accurately.

Dynamic Nature of Healthcare:

Healthcare practices and patient conditions can evolve over time. Changes in medical protocols, treatment options, or the emergence of new diseases may render the model less effective over time. Continuous updates and retraining may be necessary to maintain accuracy.

Unforeseen Events and Complications:

Unforeseen events or complications in a patient's condition can significantly affect the accuracy of predictions. A sudden change in a patient's health status, unexpected complications, or external factors can disrupt the predicted hospital stay duration.

Patient-Specific Factors:

Individual patient responses to treatment can vary, and the model might not account for all the nuances of a specific patient's case. Patients may also have external factors, such as social support or home environment, that influence their recovery and discharge timing.

TOPIC SURVEY IN BRIEF

The project focuses on developing a predictive model for hospital stay duration, aiming to enhance resource allocation and operational efficiency in healthcare facilities. By accurately estimating how long a patient is likely to stay, the model addresses challenges such as optimizing resource utilization, adjusting staffing levels, improving bed management, and ultimately delivering better patient-centric care. However, the project is subject to limitations, including the complexity of healthcare variables, data quality issues, ethical considerations, and the dynamic nature of healthcare, which may impact the model's accuracy and generalizability. Continuous monitoring, collaboration with healthcare professionals, and adherence to regulatory requirements are crucial for addressing and mitigating these limitations.

"Forecasting Hospital Stay Duration" is crucial for understanding the landscape of predictive modeling in healthcare. Here's a critical assessment of the key aspects of the survey:

Relevance:

The topic is highly relevant, addressing a significant challenge in healthcare—efficient resource allocation and management of hospital stays.

Comprehensive Overview:

The survey provides a comprehensive overview of the problem, covering key aspects such as resource optimization, operational efficiency, and patient-centric care.

Consideration of Ethical Issues:

Acknowledges ethical concerns related to patient data usage, demonstrating an awareness of the need for ethical considerations in healthcare predictive modeling.

METHODOLOGY TO BE FOLLOWED

CRISP - DM - Cross Industry Standard Process for Data Mining

– Business Understanding

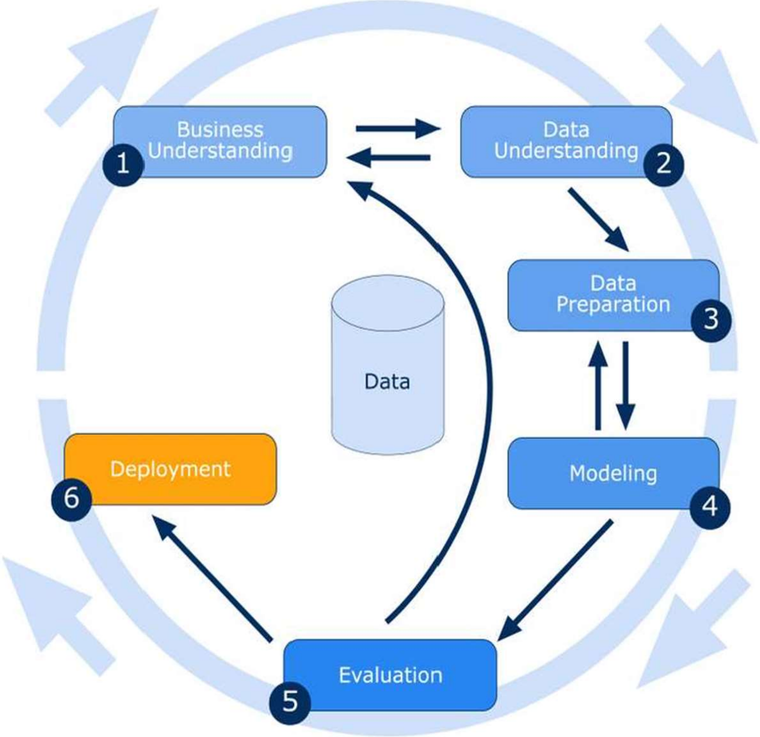
– Data Understanding

– Data Preparation

– Modeling

– Evaluation

– Deployment



Business Understanding

The primary business problem is to optimize resource allocation in healthcare facilities by predicting the duration of a patient's hospital stay. This involves accurately estimating how long each patient is likely to stay in the hospital, allowing for more efficient planning of resources.

PROBLEM STATEMENT - Recent Covid-19 pandemic has raised alarms over one of the most overlooked areas to focus: healthcare management. while healthcare management has various use cases for using data science, patient length of stay is one cd predict if one wants to improve the efficiency of the healthcare management in a hospital, this parameter helps hospitals to identify patients of high loss risk at the time of admission and prior knowledge of loss can aid in logistics such as room and bed allocation planning and task to accurately predict the length of stay for each patient on case-by-case basis so that the hospitals in a professional and optimal number

This Dataset has the required data to train a classification model that will do the delivery time estimation, based on all the features

Data Understanding

Rows: 100000

Variables: 18

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

Target Variable

- We have derived our target, **“STAY”.**
- It describes the stay days by the Patient.

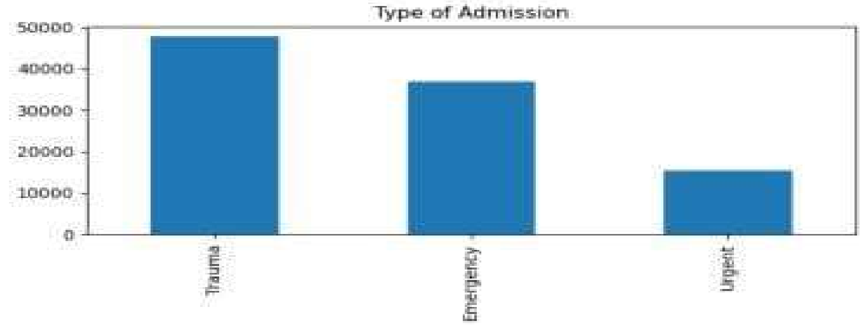
Existing Columns:

- Case_id
- Hospital_code
- Hospital_type_code
- City_code_hospital
- Hospital_code_region
- Available extra Rooms in the Hospital
- Department
- Ward_type
- Ward_Facility_Code
- Bed Grade
- Patientid
- City_code_Patient
- Type of Admission
- Severity of Illness
- Visitors with patient
- Age
- Admission_Deposit
- Stay

Univariate Analysis

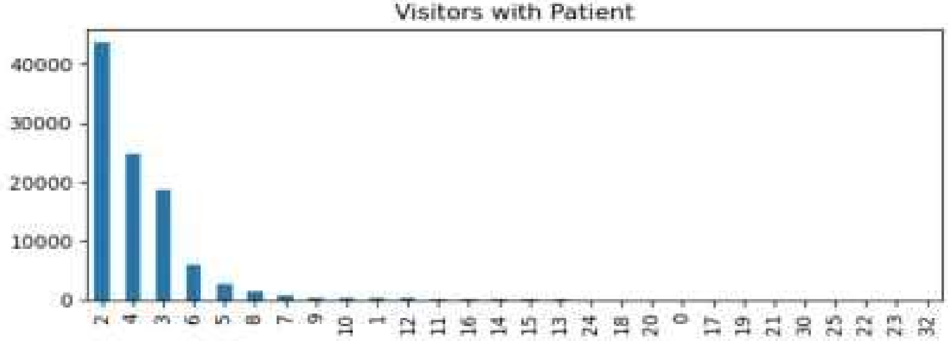
Type of Admission:-

There are three types of Admission, Trauma, emergency and Urgent, as we can see that mostly data lies Trauma.



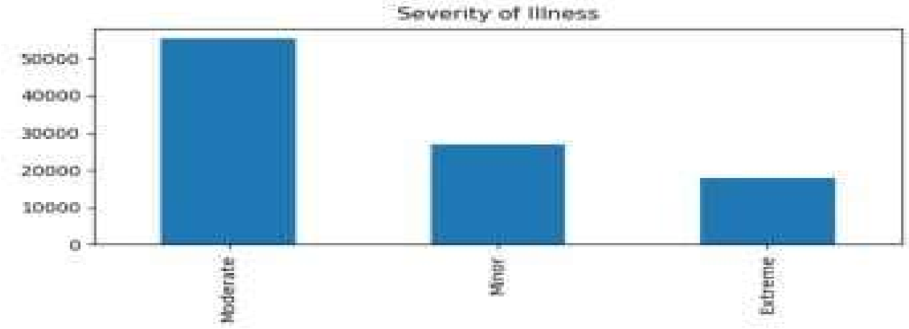
Visitors with Patient –

The data is highly right skewed, and we can see that, as we go after more than 15 number of visitors, the data is very less.



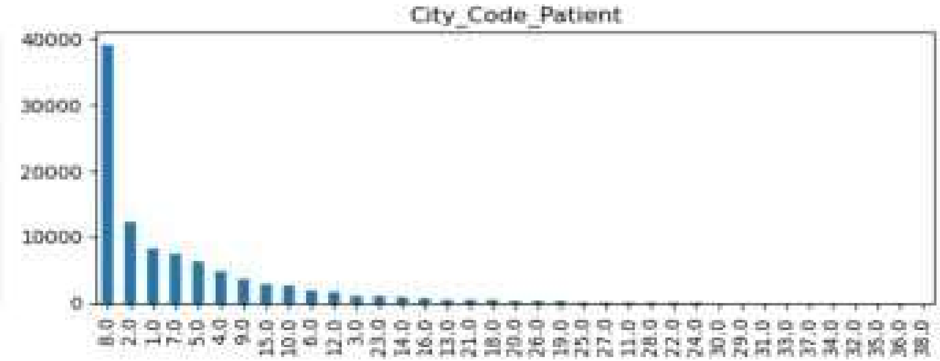
Severity of Illness –

There are three category Moderate,
Minor and Extreme, Moderate category
has the most data.



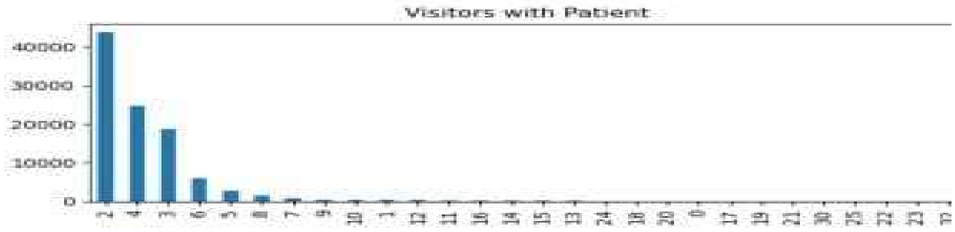
City_Code_Hospital –

It is the city code of hospitals and
city code 1,2 and 3 has the maximum patients or cases.



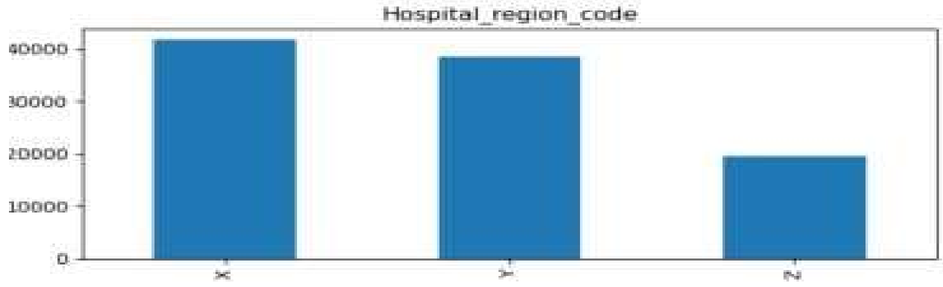
Department –

There are five categories such as gynecology,
anesthesia, radiotherapy, TB & Chest Disease and Surgery,
the major data lies in gynecology.



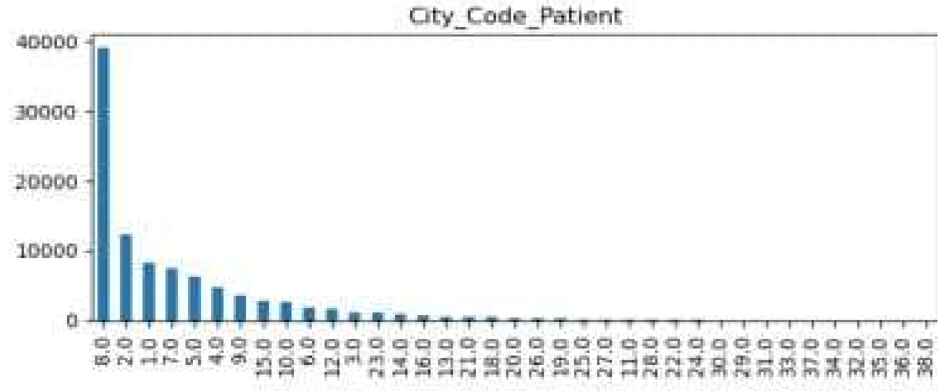
Hospital Region Code –

There are three categories such as X, Y and Z
and we can say that X and Y region are the busy regions.



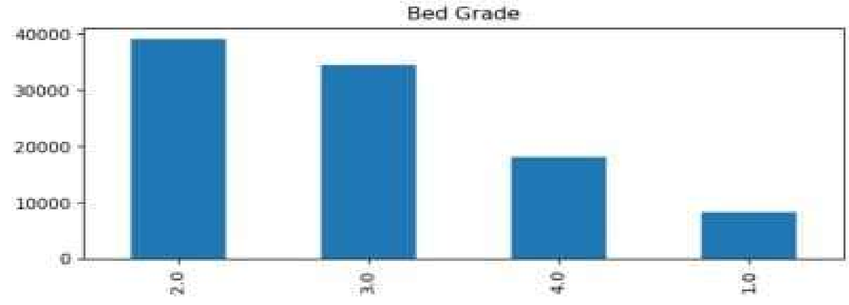
City_Code_Patient –

There are 38 categories in City Code of patient, and we can easily say 8 city code has the major data but we can also see that the patients from the city code of 8 are going to the hospital with the city code 1 and 2, maybe the hospitals lie in those cities
are the good one.



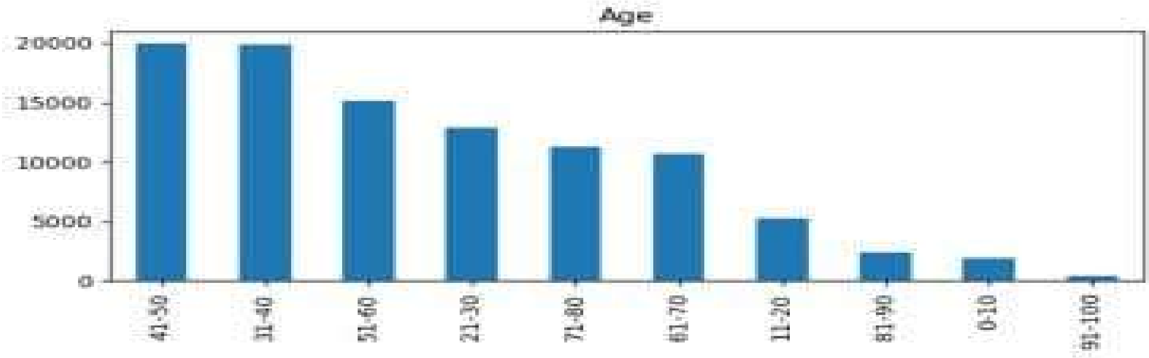
Bed Grade –

There are 4 categories in Bed Grade and we can say that 2 and 3 has the most patients.



Age –

There are 10 brackets in the Age category and we can say that, the age bracket with 41-50 and 31-40 has the most patients.



Bivariate Analysis

Type of Admission vs Stay:

Urgent admissions tend to have shorter stays, while Trauma and Emergency admissions often result in longer stays.

Visitors with Patient vs Stay:

Higher visitor numbers generally correspond to shorter stays, indicating potential support for patients during recovery.

Severity of Illness vs Stay:

Patients with moderate illness severity tend to have varied stays, while extreme severity levels correlate with longer stays.

City_Code_Hospital vs Stay:

Stay durations vary across different hospital locations (City Codes), suggesting potential regional healthcare disparities.

Department vs Stay:

The department overseeing the case influences stay duration, with gynecology cases often having shorter stays.

Hospital Region Code vs Stay:

Stay durations differ based on the region of the hospital, indicating potential regional healthcare management variations.

Hospital_Code vs Stay:

Specific hospitals (Hospital_Code) exhibit varying stay durations, suggesting hospital-specific factors influencing patient stays.

City_Code_Patient vs Stay:

Patient city codes impact stay duration, indicating potential connections between patient location and healthcare outcomes.

Available Extra Rooms in Hospital vs Stay:

The number of available extra rooms in a hospital appears to have minimal impact on patient stay durations.

Ward Facility Code vs Stay:

Different ward facilities exhibit varied stay durations, reflecting the influence of facility-specific factors.

Bed Grade vs Stay:

Bed grades influence stay durations, with certain grades associated with shorter or longer patient stays.

Hospital Type Code vs Stay:

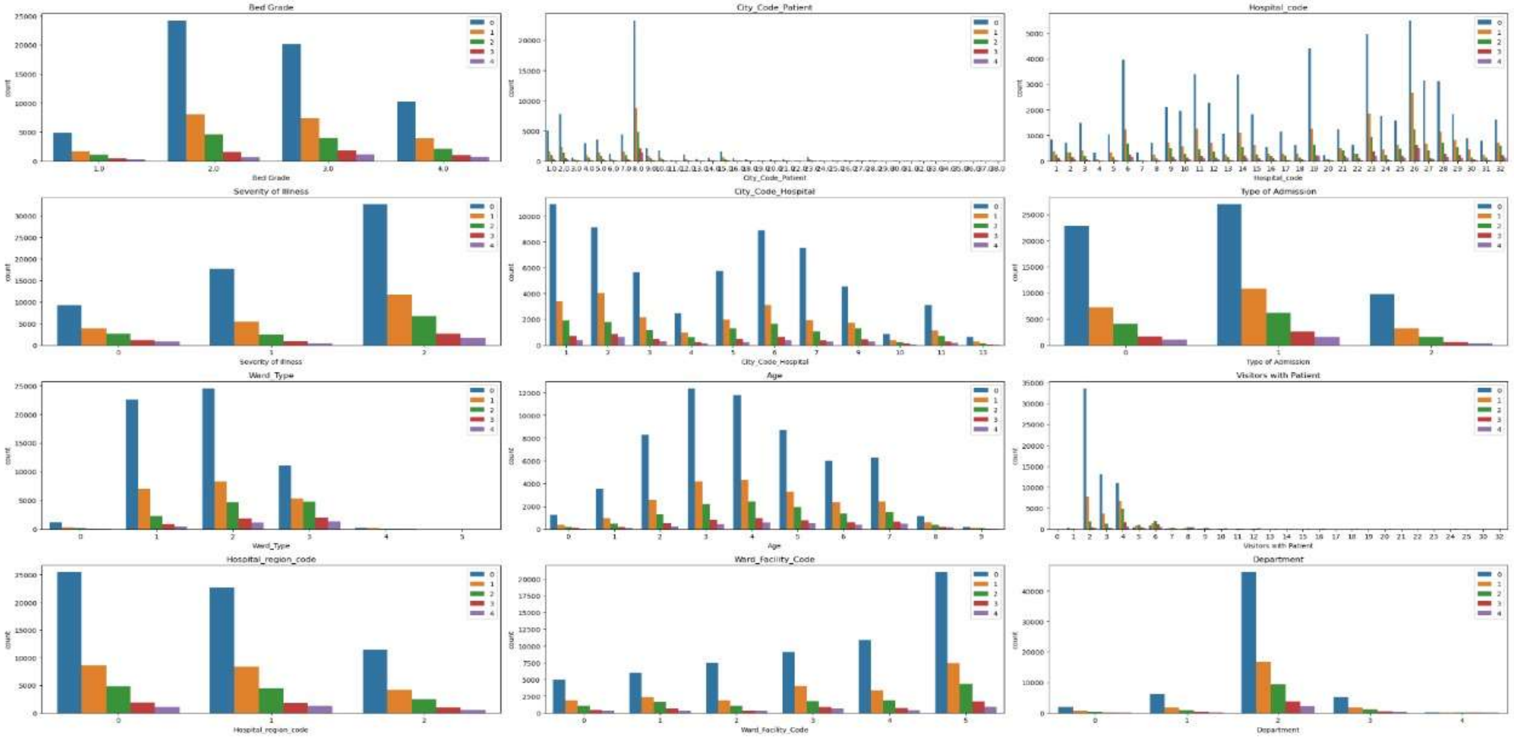
The type of hospital influences stay duration, with specific hospital types correlating with shorter or longer stays.

Age vs Stay:

Patient age impacts stay duration, with distinct patterns observed across age groups.

Ward Type vs Stay:

Different ward types influence stay durations, reflecting the significance of the ward in patient care and recovery.



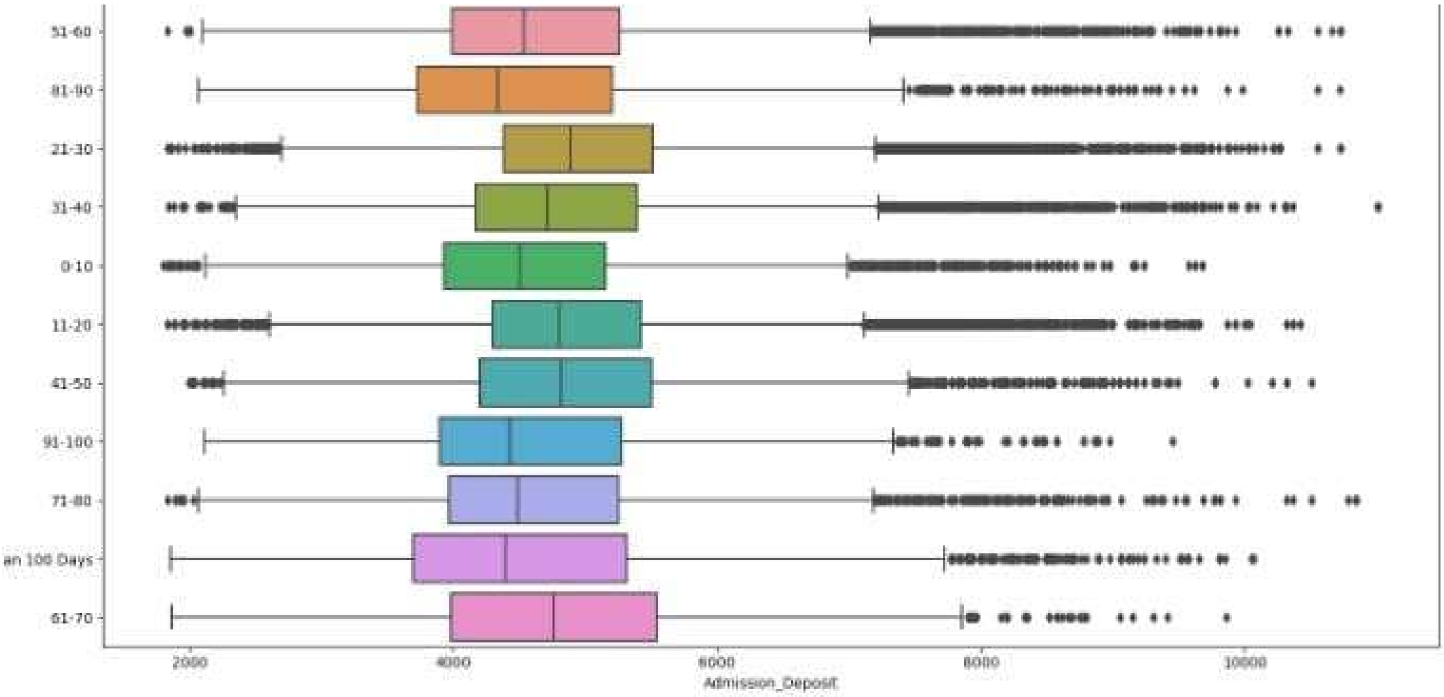
Data Preparation

Missing Values & Outlier Treatment

According to our data, there are are missing values present in variables.

We have treated severe outliers in order to lose least possible rows or information from the data. Hence, we choose selected features only in which extreme outliers were present in Adimission_deposit

Admission_deposit but we can not deal with outliers because these outliers are good distribution of Data and they are related to finance.


51-60
81-90
21-30
31-40
0-10
11-20
41-50
91-100
71-80
an 100 Days
61-70
2000
4000
6000
8000
10000
Admission_Deposit

Data Pre-Processing (Encoding & Scaling)

Encoding the categorical features to numbers such that the model is able to understand and extract valuable information.

□ Label Encoding : “Stay”

□ Performed standard scaling on all numerical variables to bring all the variables having the same scale.

Statistically Significant Features

We performed “chisquare” test on all numerical variables columns for identifying if they contribute towards explaining variation in target variable as the data was not normal and found all variables as significant.

- Hospital_code
- Hospital_Type_Code
- City_Code_Hospital
- Hospital_Region_Code
- Available Extra Rooms in Hospital
- Department
- Ward_Type
- Ward_Facility_Code
- Type of Admission
- Severity of Illness
- Visitors with Patient
- Age

Modeling

We Split the data into train and test sets in the ratio of 80:20.
We tried fitting our data into various models such as Gaussian Naïve Bayes Model, Gradient Boosting, decision tree, random forest and xg boost.

Inferences from Base model (Gaussian Naïve Bayes Model):

We can say that all the features are significant, all features lies below 5% of significance level which are good. Based on extremely low P-value(Admission Deposit P-value 0.0) all the mentioned features are statistically significance and are likely to have a meaningful impact on the outcome variable being studied in the analysis.The probability of ttest for all variables was less than 0.05 which means that all the existing as well as derived columns are also significant as we tested above using statistical tests as well.

Scalling: We have scalled admission deposit using standard scaller.

Before Modelling we did not bin any category apart from stay, as these categories have less data, and binning can disturb the other data, hence it'll decrease the accuracy.

Modelling:

As we have more than 2 catgories in the stay, we can not apply binary models such as logistic regression, bernaulli naïve bayes.

Gaussian Navie Bayes:

In Gaussian Navie Bayes we received a accuracy of 64% , indicating the percentage of correctly predicting instances across all classes .However for classes 1, 2, 3,4 the3 model shows challenges in terms of recall suggesting difficulty in identifying instances of these classes.

While the base model demonstrates the regionable accuracy , there is room for improvement, the next step involves a deeper analysis and refinement of the model to address these challenges.

Decision Tree:-

In Decision tree we received a accuracy of 55% , indicating the percentage of correctly predicting instances across all classes .However for classes 1, 2, 3,4 the3 model shows challenges in terms of recall suggesting difficulty in identifying instances of these classes.While the base model demonstrates that there is decrease in accuracy , need an improvement, the next step involves a deeper analysis and refinement of the model to address these challenges.

Random Forest.:-

In Random Forest we received a accuracy of 66% , indicating the percentage of correctly predicting instances across all classes .However for classes 1, 2, 3,4 the 3 model shows challenges in terms of recall suggesting difficulty in identifying instances of these classes.While the base model demonstrates that there is increase in accuracy as compared to previous models , we are trying to improve the model, the next step involves a deeper analysis and refinement of the model to address these challenges

Gradient Boosting:-

In Gradient Boosting we received a accuracy of 67% , indicating the percentage of correctly predicting instances across all classes .However for classes 1, 2, 3,4 the 3 model shows challenges in terms of recall suggesting difficulty in identifying instances of these classes. While the base model demonstrates that there is increase in accuracy as compared to previous models , we are trying to improve the model, the next step involves a deeper analysis and refinement of the model to address these challenges.

XG BOOST:-

In XgBoost we received a accuracy of 68% , indicating the percentage of correctly predicting instances across all classes .However for classes 1, 2, 3,4 the 3 model shows challenges in terms of recall suggesting difficulty in identifying instances of these classes.While the base model demonstrates that there is increase in accuracy as compared to previous models , we are trying to improve the model, the next step involves a deeper analysis and refinement of the model to address these challenges.

Parameters:- “min_child_weight”-1.5; “max_leaves”-0, “max_depth”-6; “lambda”-3; “gamma”-1; “eta”-0.4; “alpha”- 0

Evaluation

Conclusion for the final understanding

Accuracy:

Accuracy is a measure of the overall correctness of the model's predictions. It represents the ratio of correctly predicted instances to the total instances. Higher accuracy values indicate better overall performance.

In this case, the models' accuracies range from approximately 54.7% to 67.8%. The XG Boost model, particularly the hypertuned version, shows the highest accuracy at 67.8%.

Reliability:

Reliability, as mentioned in this context, is not a standard metric used in machine learning. However, it appears to be a measure of the model's consistency or stability.

The reliability values range from approximately 23.9% to 37.5%. Higher reliability values suggest that the model's predictions are more consistent.

Model Comparison:

The XG Boost model, especially the hypertuned version, stands out as having the highest accuracy and reliability among the models presented.

Gradient Boosting, ADA Boost, and Random Forest also demonstrate relatively good accuracy, but their reliability values are slightly lower compared to XG Boost.

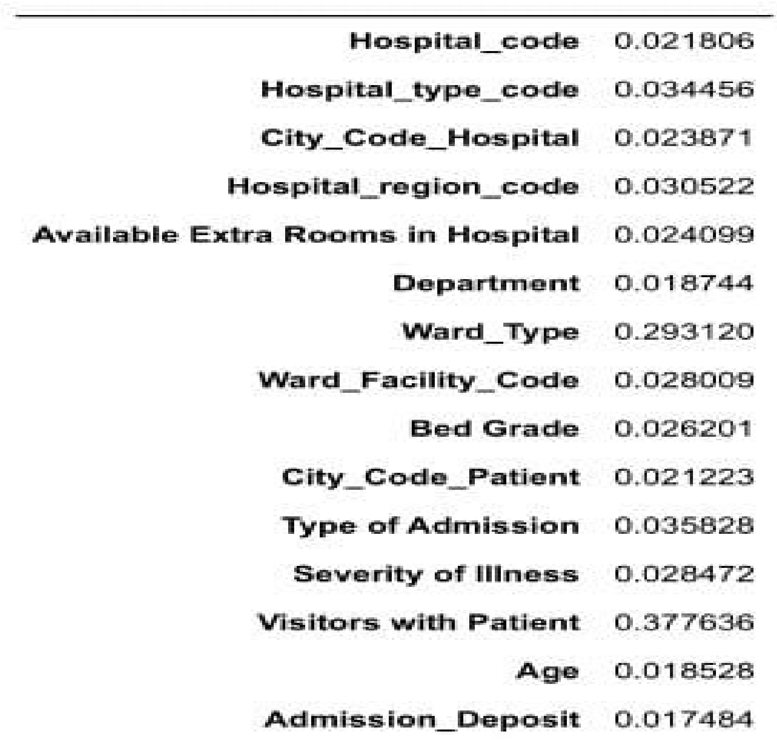
Decision Tree, K-Nearest Neighbor, and Gaussian NB have lower accuracy and reliability compared to the ensemble models.

Considerations:

It's important to note that accuracy alone may not be sufficient for evaluating a model, especially in imbalanced datasets. Other metrics like precision, recall, and F1 score should also be considered, depending on the specific goals of the model.

The reliability measure is less common in standard machine learning evaluation. If this is a custom metric for your specific problem understand how reliability is defined and calculated in your context.

The XG Boost model, particularly the hypertuned version, seems to perform well based on the provided metrics. However, the choice of the best model also depends on the specific requirements and constraints of the problem you are addressing.



Hospital_code	0.021806
Hospital_type_code	0.034456
City_Code_Hospital	0.023871
Hospital_region_code	0.030522
Available Extra Rooms in Hospital	0.024099
Department	0.018744
Ward_Type	0.293120
Ward_Facility_Code	0.028009
Bed Grade	0.026201
City_Code_Patient	0.021223
Type of Admission	0.035828
Severity of Illness	0.028472
Visitors with Patient	0.377636
Age	0.018528
Admission_Deposit	0.017484

Sample Reference for Datasets (to be filled by team and mento)

Original owner of data	Not Disclosed (Delhi Official)
Data set information	Healthcare Management
Previous relevant journals used the data set	XXXXXX—Confedintial --XXXXXX
Citation	
Link to web page	https://www.kaggle.com/datasets/nehaprabhavalkar/av-healthcare-analytics-ii/data