HEALTHCARE ANALYTICS PROJECT REPORT

LENGTH OF-STAY PREDICTION AT TIME OF ADMISSION

PREPARED BY:

Siddhartha Patra

CONTENTS

INTRODUCTION

Predictive analytics is growing day-by-day as an important tool in the healthcare industry because of its abilities to predict the individual outcomes of patients using machine learning (ML) techniques on vast volumes of data. Hospitals & Clinics can use Predictive analytics to better manage their time & resources and ultimately provide better services to their patients. For this project, we have chosen to focus on hospital length-of-stay (LOS), which is a very crucial aspect of determining the wait times and beds availability in the hospitals, especially in ICU (Intensive Care Unit). LOS is defined as the period of time, expressed in days, between hospital admission time and discharge time.

PROJECT BACKGROUND

Business Problem

According to a recent study from the Fraser Institute (*Comparing Performance of Universal Health Care Countries, 2019*), Canada has a relatively short supply of doctors and hospital beds—and the longest wait times, despite spending more on health care than most other developed countries with universal coverage.

This problem is growing bigger, as Canada's senior population grows in number with each passing day, seeking more medical care. As suggested *by Canadian Institute of Health Information* (*CIHI*), Over the next 20 years, Canada's seniors' population — those age 65 and older — is expected to grow by 68%.

Even from our personal experiences, we have witnessed our own friends and family members facing long wait times in the direst of situations, especially in emergency wards in Canada. This was one of our biggest concern and motivation to choose this topic for our research.

Why Length of Stay (LOS)?

We want to tackle the problem of long wait time in hospitals at time of admission. One of the ways to deal with this issue was to estimate the Length-of-stay (LOS) of patients during the time of admission. This is due to the following reasons:

Improve Communication: An accurate prediction of a patient's remaining LOS would be a useful guide for expectation management and would improve communication between doctors and their patients.

Improve Resource Allocation: One of the primary goals of hospital administration is to allocate resources as efficiently as possible. They seek to improve the quality of healthcare services for patients (for example, on-time diagnosis and shorter wait times) while reducing costs.

Improve Service Quality: Precise LOS prediction of patients is critical for efficient use of ICU resources such as staff, ventilators, and other medical devices.

Project Goal

Exploratory Data Analysis: Through the project we initially plan to study the impact of factors, such as, patient age, gender, race, ethnicity, marital status, diagnosis category etc. on a patient's LOS. We want to observe the minimum, maximum and median LOS of patients across these factors and determine the overall median LOS.

Predictive Analytics: We will be further utilizing these factors to develop a predictive ML model (using a regressor) to predict the LOS of future patients, that the hospitals can utilize to optimize their medical resources and better manage bed availability in the emergency wards.

Evaluation Criteria

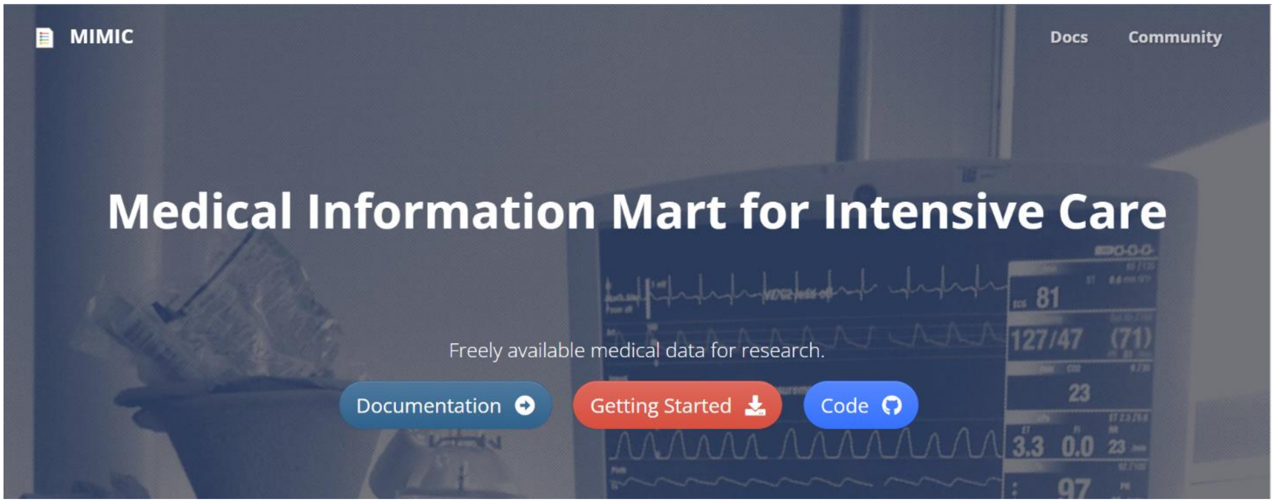
The goal of this project is to predict the Length-of-stay (LOS) of patients at the time of admission. Hence, the measure of success of our project will be the accuracy of ML Model to predict the LOS of patients based on their demographic and health diagnosis features.

Furthermore, the project's success is also dependent on the methods and ways we can prescribe to hospital management to manage their resources better and minimize patients' wait times, once the LOS of patients is predicted at the time of admission.

Project Resources

For this project, we will be using the MIMIC-III ((Medical Information Mart for Intensive Care III) Clinical dataset, which is a large, freely-available database comprising deidentified health-related data associated with over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

This is a relational database consisting of 26 tables, that strongly simulates real life databases that are used in modern day hospitals.

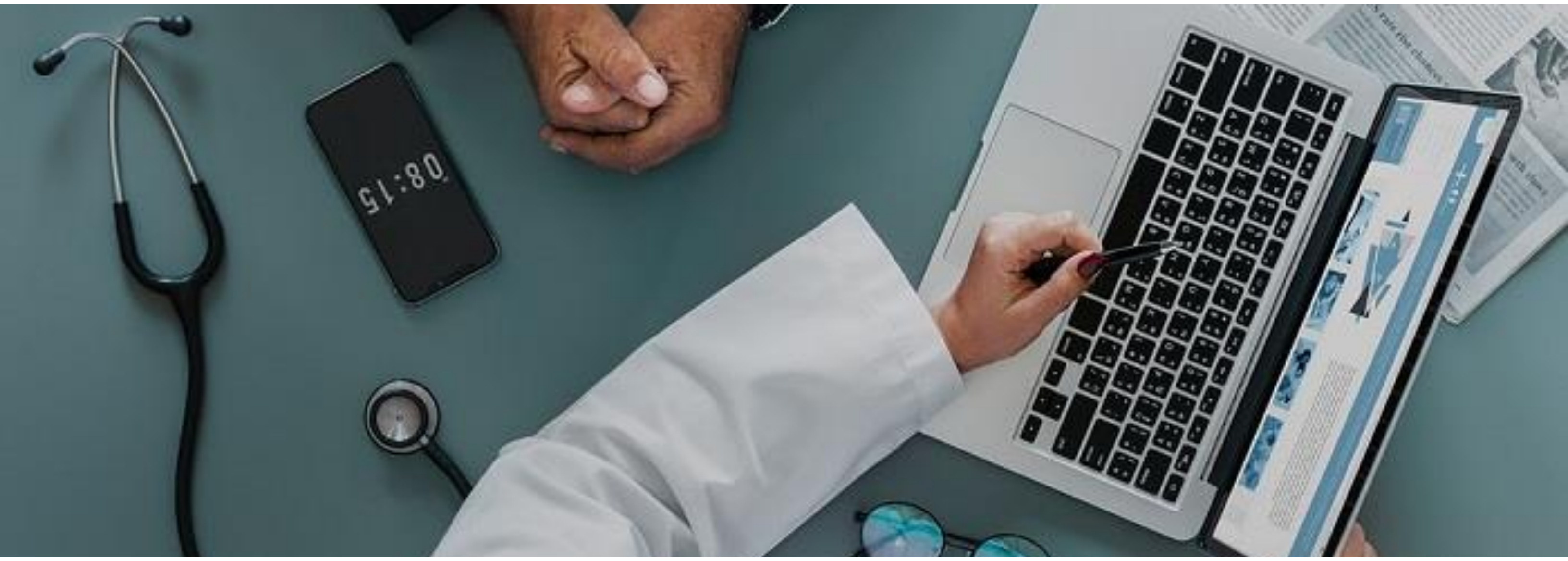


Medical Information Mart for Intensive Care

Freely available medical data for research.

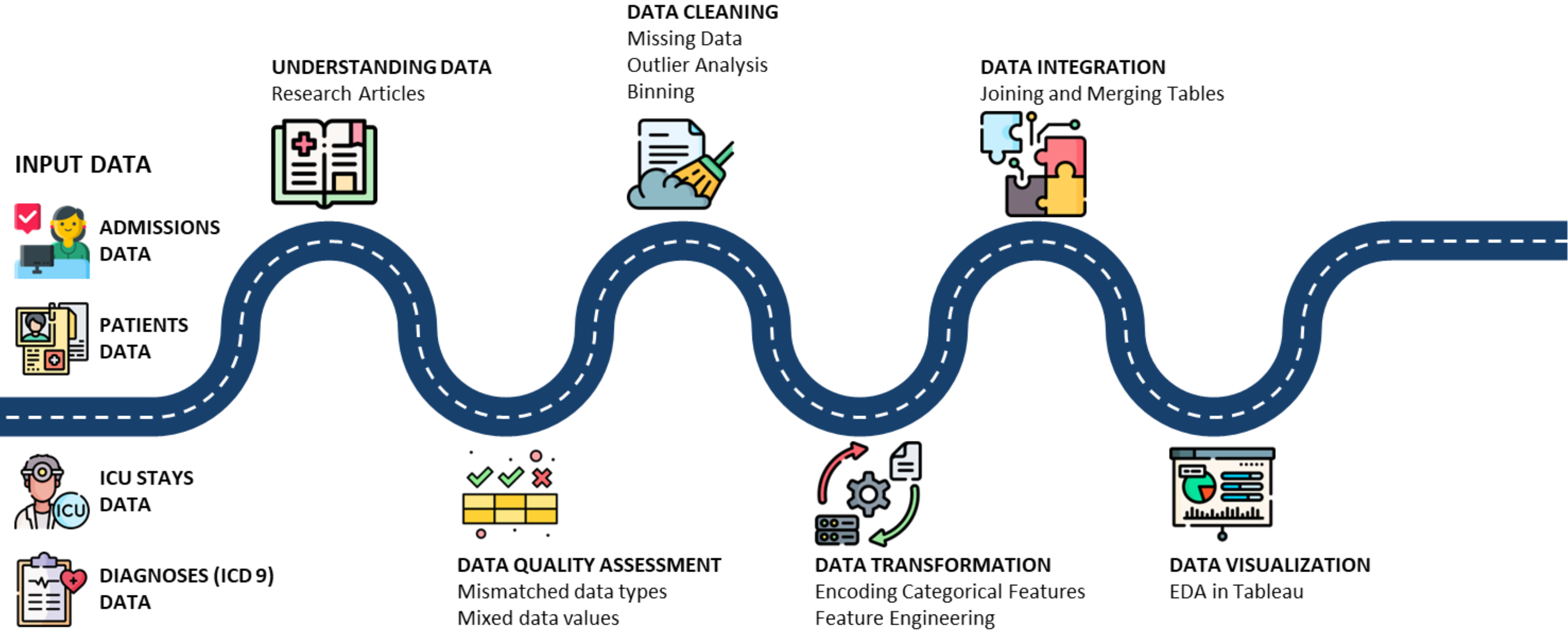
Following methods and tools will be used to execute this project:

- Data Access for MIMIC-III (through completion of required trainings and becoming a credentialed PhysioNet user)
- Data Cleaning (MS Excel, SQL, Python)
- Exploratory Data Analysis (Python)
- Data Visualization (Python - Matplotlib & Seaborn, Tableau)
- Build, Train and Test ML Models (Python)
- Model Refinement for better accuracy (Python)



DATA PREPROCESSING

After going through the 24 tables present in the MIMIC-III dataset, we decided to focus on 4 tables for the scope of our analysis. These 4 input tables were Admissions, Patients, ICU Stays & Diagnosis (ICD 9) tables. The roadmap for data preprocessing was as follows:



INPUT DATA

ADMISSIONS DATA

PATIENTS DATA

ICU STAYS DATA

DIAGNOSES (ICD 9) DATA

UNDERSTANDING DATA
Research Articles

DATA QUALITY ASSESSMENT
Mismatched data types
Mixed data values

DATA CLEANING
Missing Data
Outlier Analysis
Binning

DATA TRANSFORMATION
Encoding Categorical Features
Feature Engineering

DATA INTEGRATION
Joining and Merging Tables

DATA VISUALIZATION
EDA in Tableau

Understanding Data

We started with initial research on the LOS problem and studied different publicly available research-based articles on the MIMIC-III Dataset. This was done to develop an understanding of the dataset and study how to approach this kind of problems.

We also found that 6984 unique ICD9 codes from the DIAGNOSES (ICD 9) Table. Research was done on the ICD9 codes, and it was found that the first 3 digits of the ICD9 Codes could be arranged into the super categories. This helped us significantly to generalize diagnosis categories, which further led to better performance of our ML models.

Data Quality Assessment

Mismatch Data Types: Features with Date-time values such as Admission Time, Discharge Time, Death Time, and Date of Birth, were changed to the correct date type data to calculate features such as Age and Length of Stay. Numerical Features were also converted from object to integer/float data types.

Mixed Data Values: Data values that hold similar or same meaning were replaced with one uniform value. Eg: 'SEPERATED' and 'DIVORCED' were clubbed together as 'DIVORCED'; 'UNKNOWN' and 'UNOBTAINABLE' were clubbed as 'UNKNOWN'; Empty characters were replaced with NaNs.

Data Cleaning

Missing Data: With observation, we found that most feature variables with missing values were less than 1% of the total number of records. Hence, it was decided to drop those records with missing values. In some cases, categorical variables like Marital Status, missing values were replaced with values like 'UNKNOWN'.

Outlier Analysis: Outliers were found in numerical features like Age and Length of Stay. For the scope of our analysis, we decided to limit the values of age to 100 years and Length of Stay to 30 days. Any records with values above that threshold limit were dropped from our dataset.

Binning: Binning was used to generalize categorical features with large number of categories. Example Ethnicity column had 41 unique ethnic groups but many of them has the same roots such as 'ASIAN – THAI', 'ASIAN - KOREAN', 'ASIAN - JAPANESE' etc. all 41 ethnicity groups were clubbed under 5 main ethnic groups, i.e., WHITE, BLACK/AFRICAN AMERICAN, HISPANIC/LATINO, ASIAN and OTHER/UNKNOWN. Binning was also used to bin numerical features like Age and Los to categorical features, like Age was binned into Child, Youth, Adult and Seniors based on different age ranges or groups.

Data Transformation

Feature Engineering: Features like Admission Time, Discharge Time was used to calculate Length of Stay (Our Target Variable). Similarly, the patient's first (or the earliest) Admission Time and Date of Birth was used to calculate their relative Age.

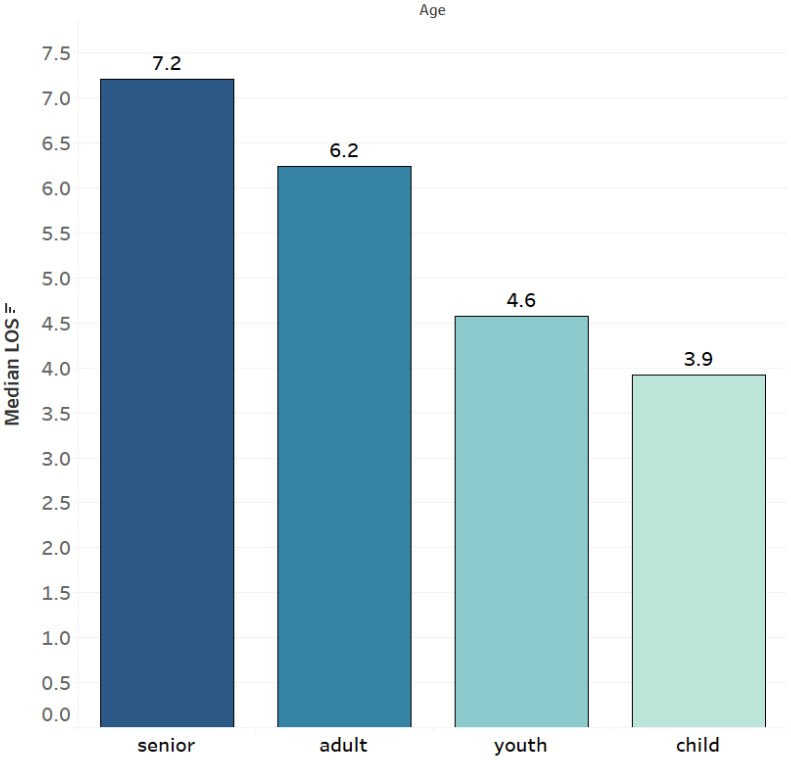
Encoding Categorical Features: Finally, all categorical features were converted into numerical features using One-Hot Encoding, so that they can fitted into the ML model.

Data Integration

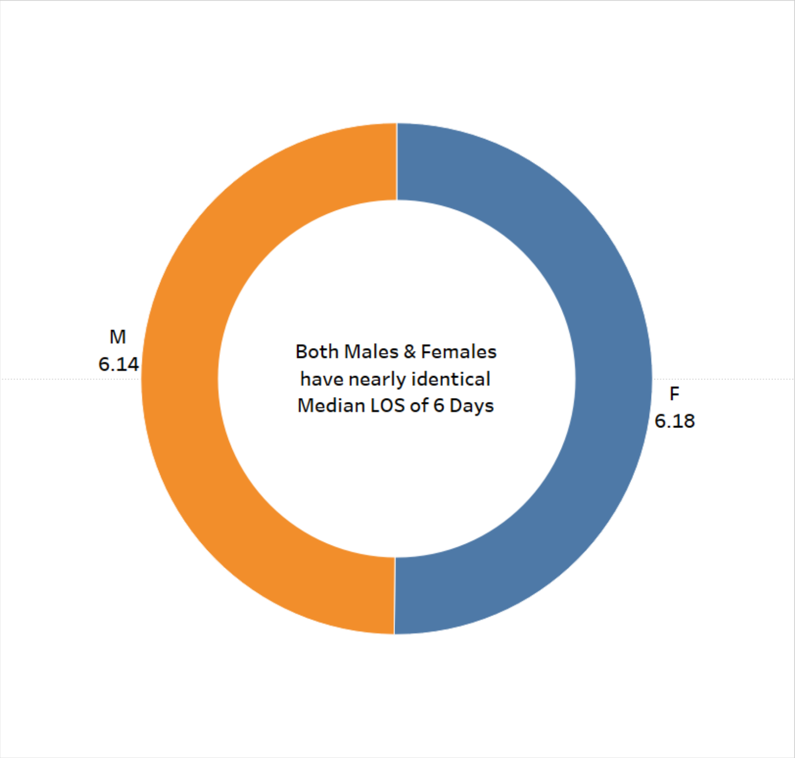
Joins & Merges: Lastly data from all the tables (Patients, ICU Stays and Diagnosis - ICD 9) were merged with the Admissions table using primarily inner joins and left joins on either Subject IDs or Hospital Admit ID.

EXPLORATORY DATA ANALYSIS

Comparision of Median LOS across Age Groups

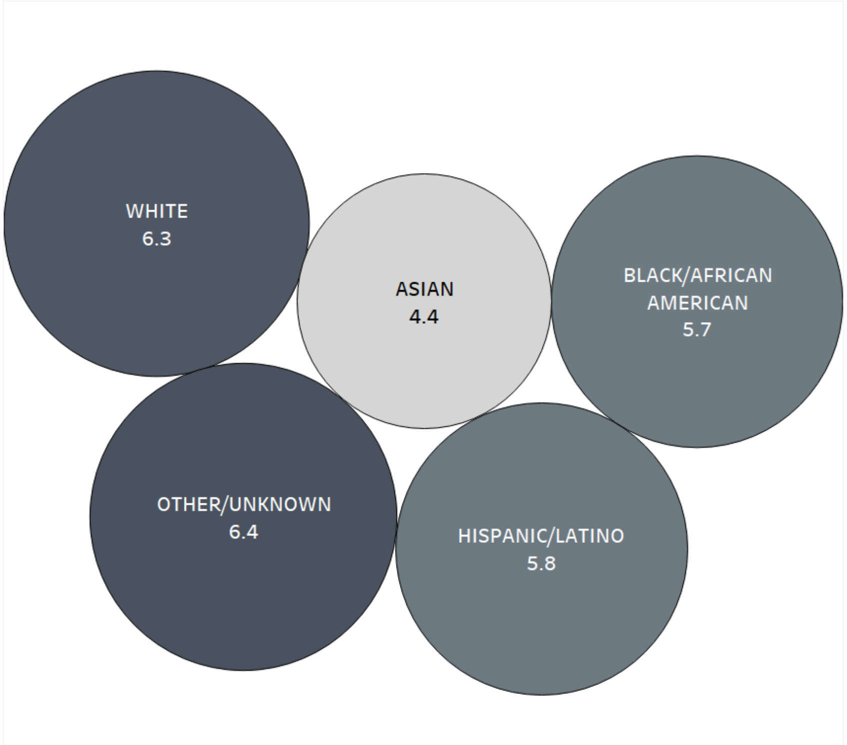


Comparision of Median LOS across Gender

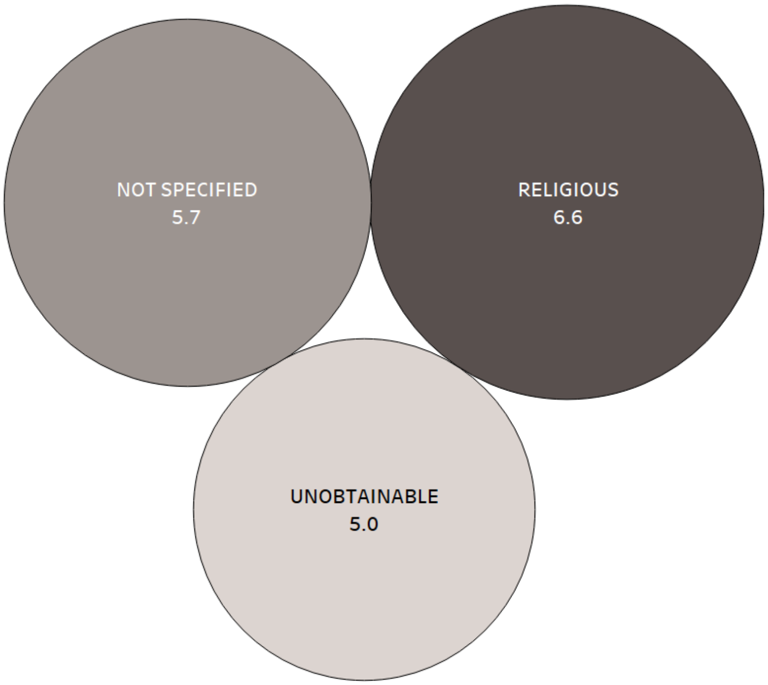


Both Males & Females have nearly identical Median LOS of 6 Days

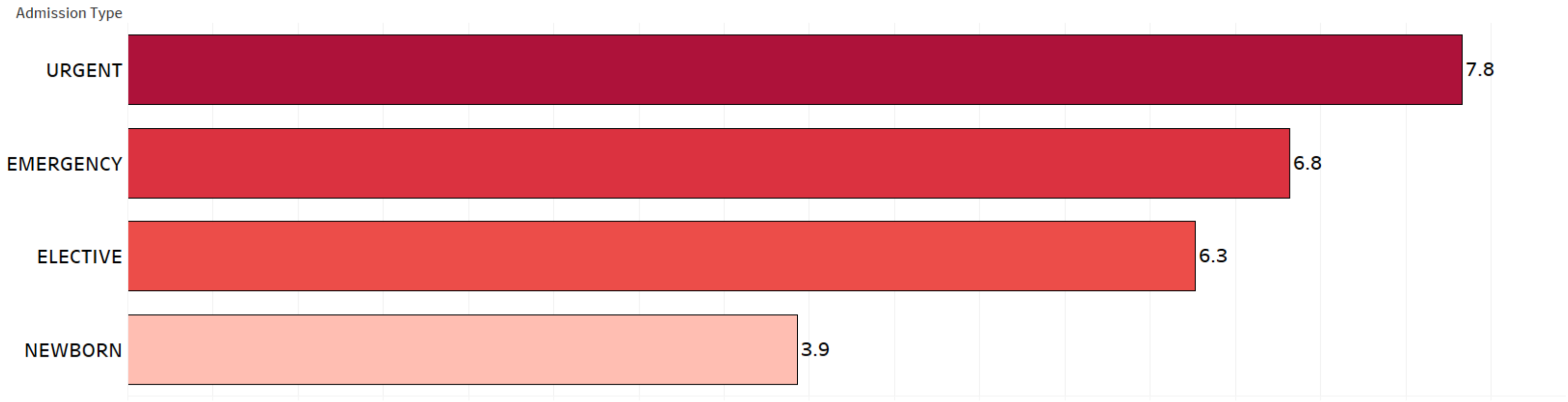
Comparision of Median LOS across Ethnicity



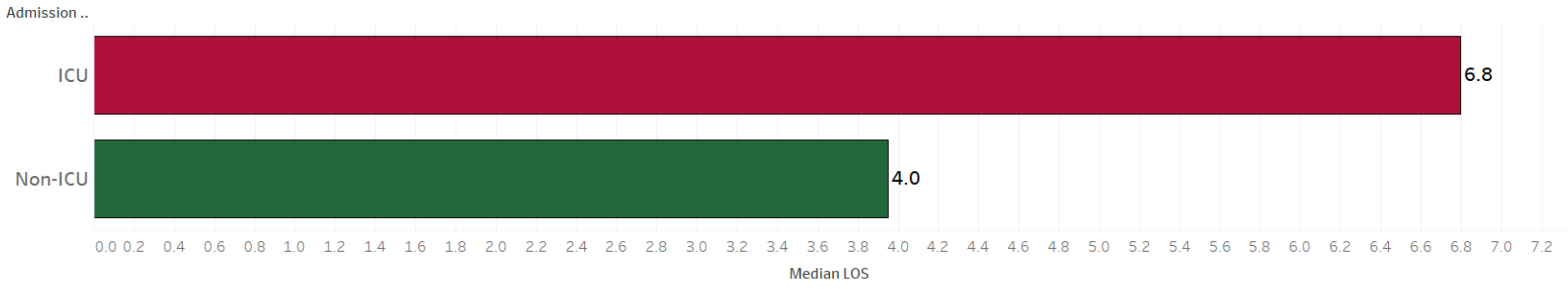
Comparision of Median LOS across Religious Belief



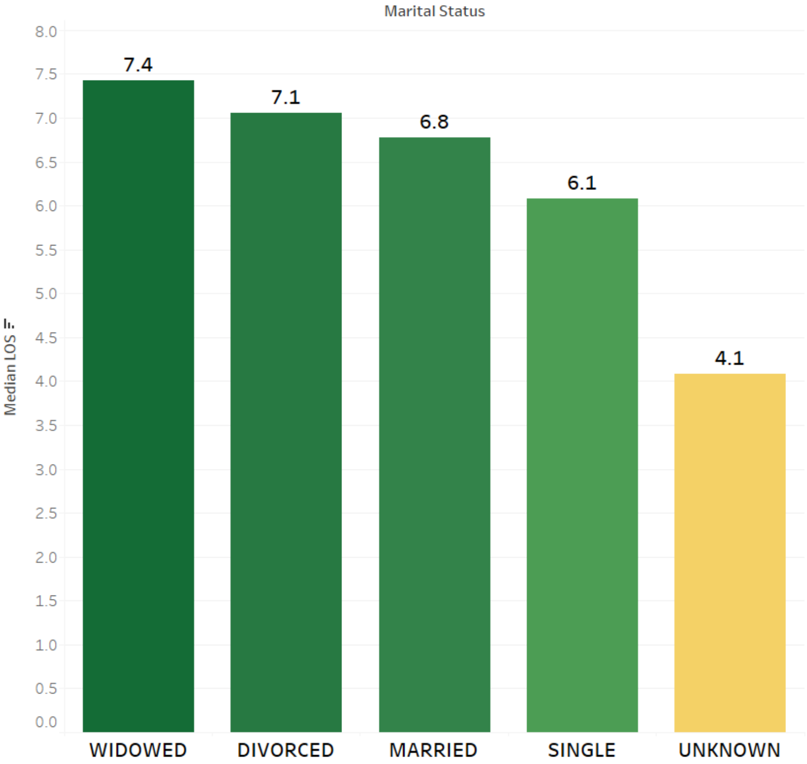
Comparision of Median LOS across Admission Types



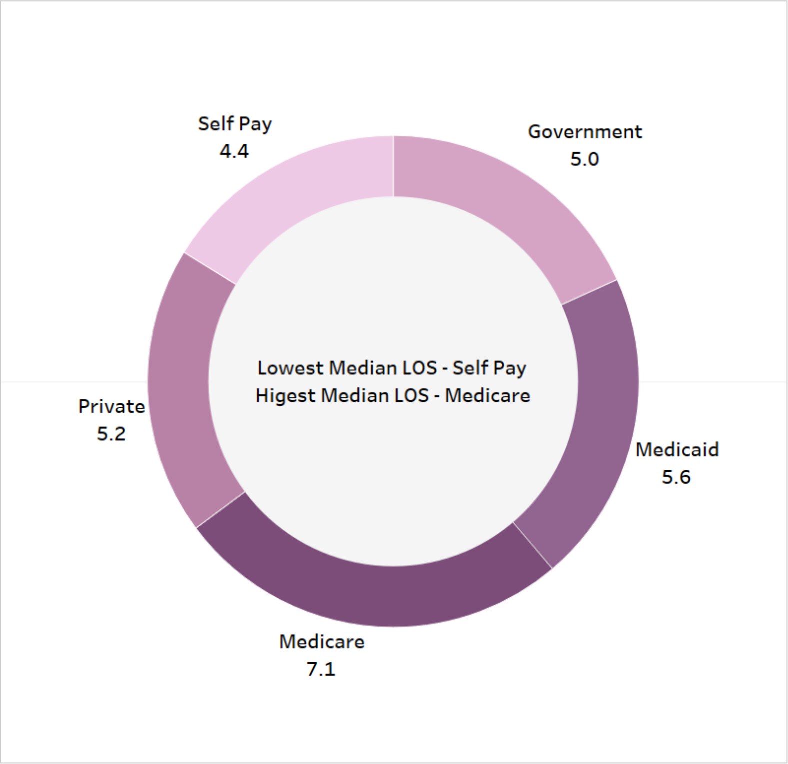
Comparision of Median LOS across Admission Ward



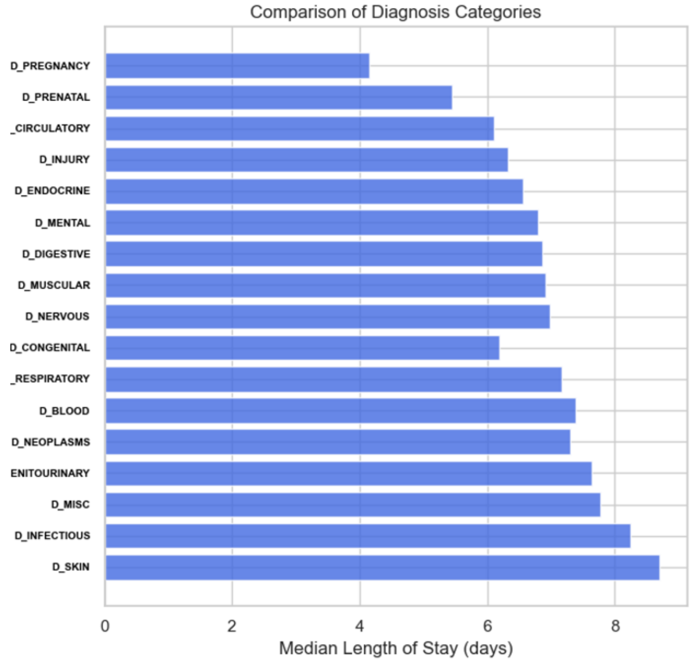
Comparision of Median LOS across Marital Status



Comparision of Median LOS across Insurance



Comparision of Median LOS across Diagnosis Type



ML MODEL IMPLEMENTATION

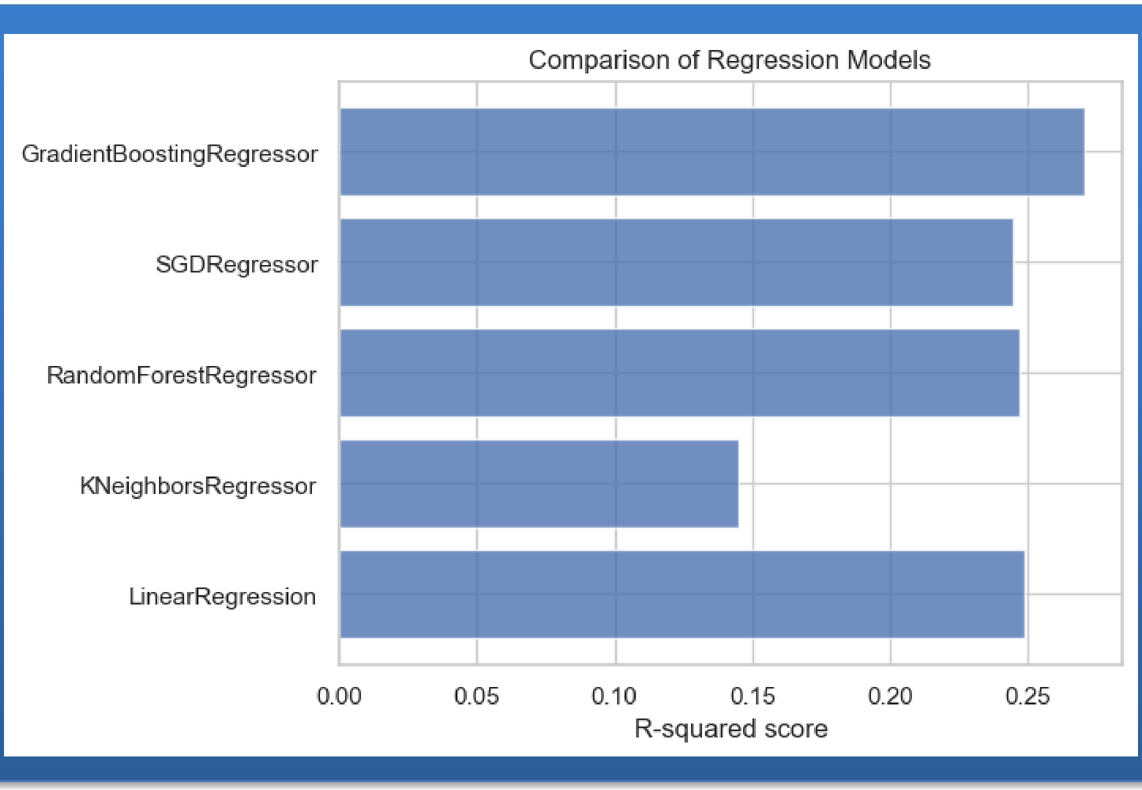
We used the cleaned dataset for implementing Machine Learning Models to predict the Length of Stay (LOS) for the patients. The dataset was divided into 80% training set and 20% testing set. Since LOS is a continuous numerical feature, forecasting or predicting LOS becomes a problem of regression.

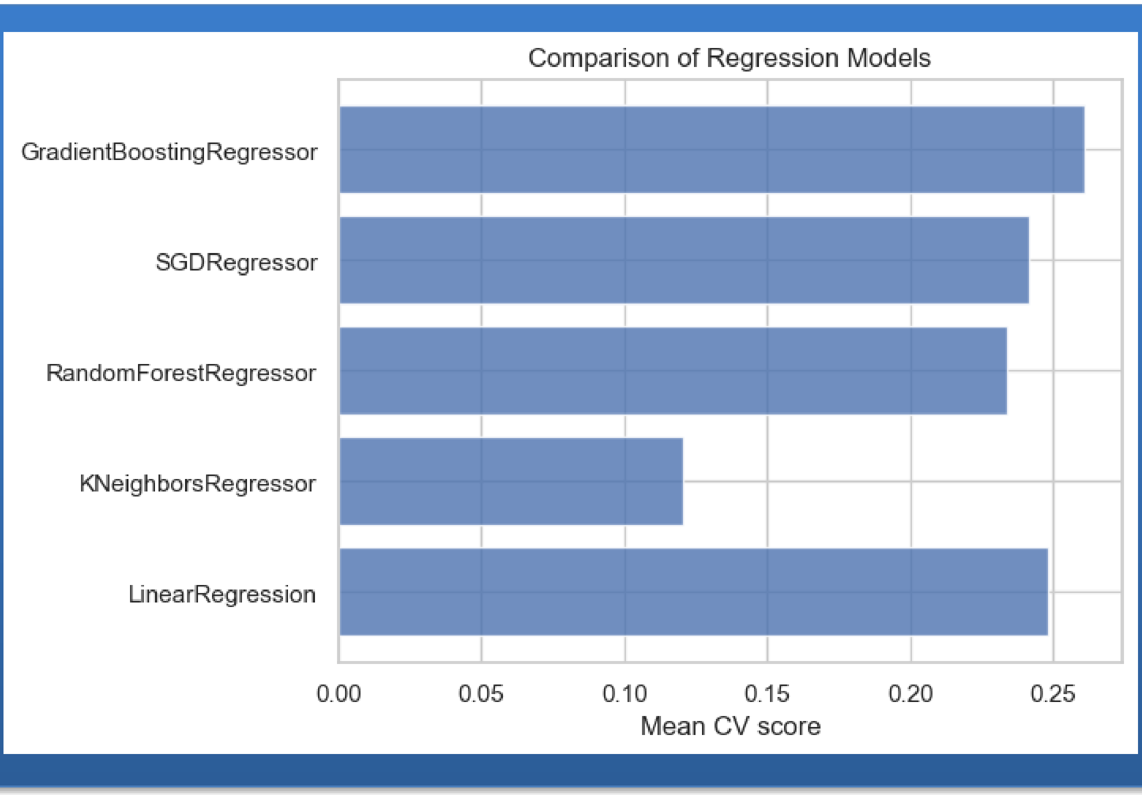
Approach 1: Regression

For the purpose of our analysis, we used some of the most widely used Regressors i.e., Linear Regressor, K Neighbors Regressor, Random Forest Regressor, SGD Regressor and Gradient Boosting Regressor. We also used 5-fold cross validation and the accuracy for our model was tested on R-squared score and Mean Cross Validation (CV) score.

GradientBoosting Regressor performed the best among all the Regressors, with R-squared score of around 27% and Mean Cross Validation (CV) score of around 26%. Further, by hyperparameter tuning the accuracy increased by 1%. But overall, the accuracy of our regression model was still far below our expectation and was not enough to implement any real-life solutions.

Comparison of accuracy scores across all regression models:





Approach 2: Classification

The problem at hand: Our regression model has a terrible accuracy even after data cleaning (which dealt with missing values as well as outliers). Hyperparameter tuning barely increased the accuracy by 1%. The underlying reasons for this problem is that we are trying to estimate LOS, which is continuous numerical variable, using all categorical variables converted into numerical features using One-Hot Encoding. Regression models perform much better when the input independent variables are also continuous numerical in nature.

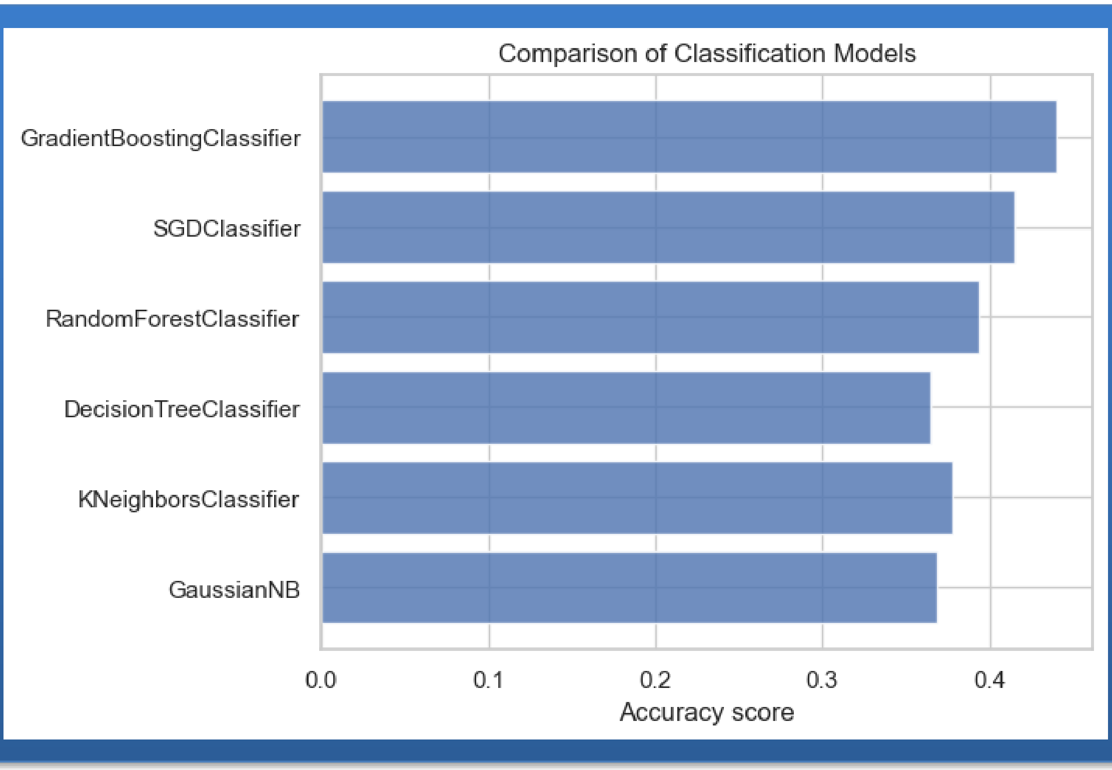
Solution: We can convert LOS into a categorical variable through binning LOS in various range of values, such as, Low (0 to 3 days), Medium (4 to 6 days), High (7 to 10 days) and Very High (More than 10 days). This way instead of predicting the precise LOS, we can change this regression problem into a classification problem of determining the correct bin of LOS. Thus, we don't have to rely on numerical features to tackle this problem.

1. Multiclass Classification

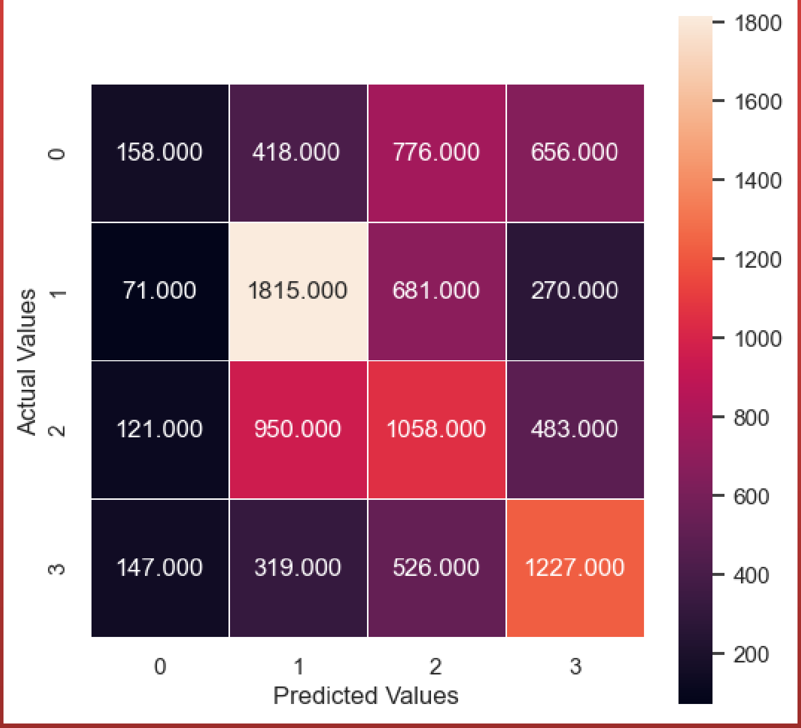
After binning our target variable LOS into 4 categories (or class), we used some of the most widely used Classifiers i.e., Gaussian NB Classifier, K Neighbors Classifier, Decision Tree Classifier, Random Forest Classifier, SGD Classifier, Gradient Boosting Classifier, and the accuracy for our model was tested based on the Confusion Matrix.

Again, Gradient Boosting Classifier performed the best among all the Classifiers, with accuracy score of around 44%. Further, hyperparameter tuning increased the accuracy by 1%. Overall, the accuracy of our classification model is much better than our regression model but it was still below our expectation and was again not enough to implement any real-life solutions.

Comparison of accuracy scores across all classification models:

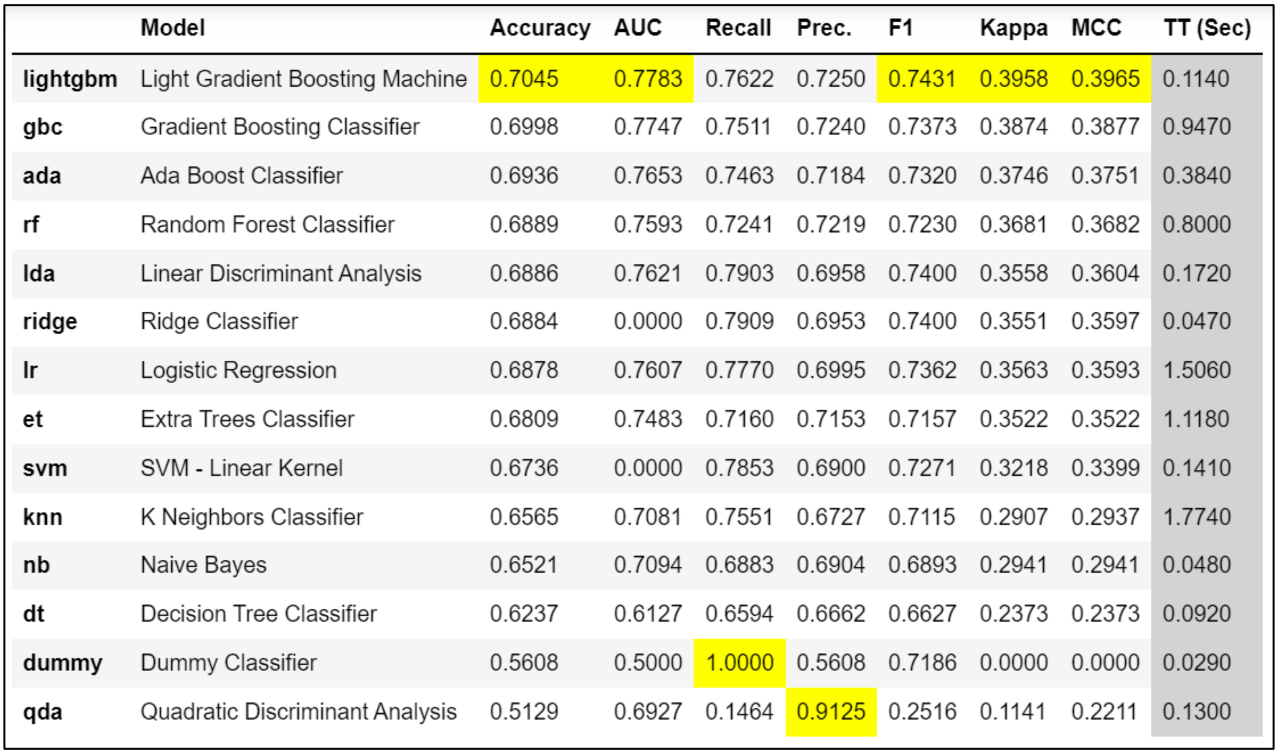


Confusion Matrix of Gradient Boosting Classifier for 4 Class Classification:



2. Binary Classification

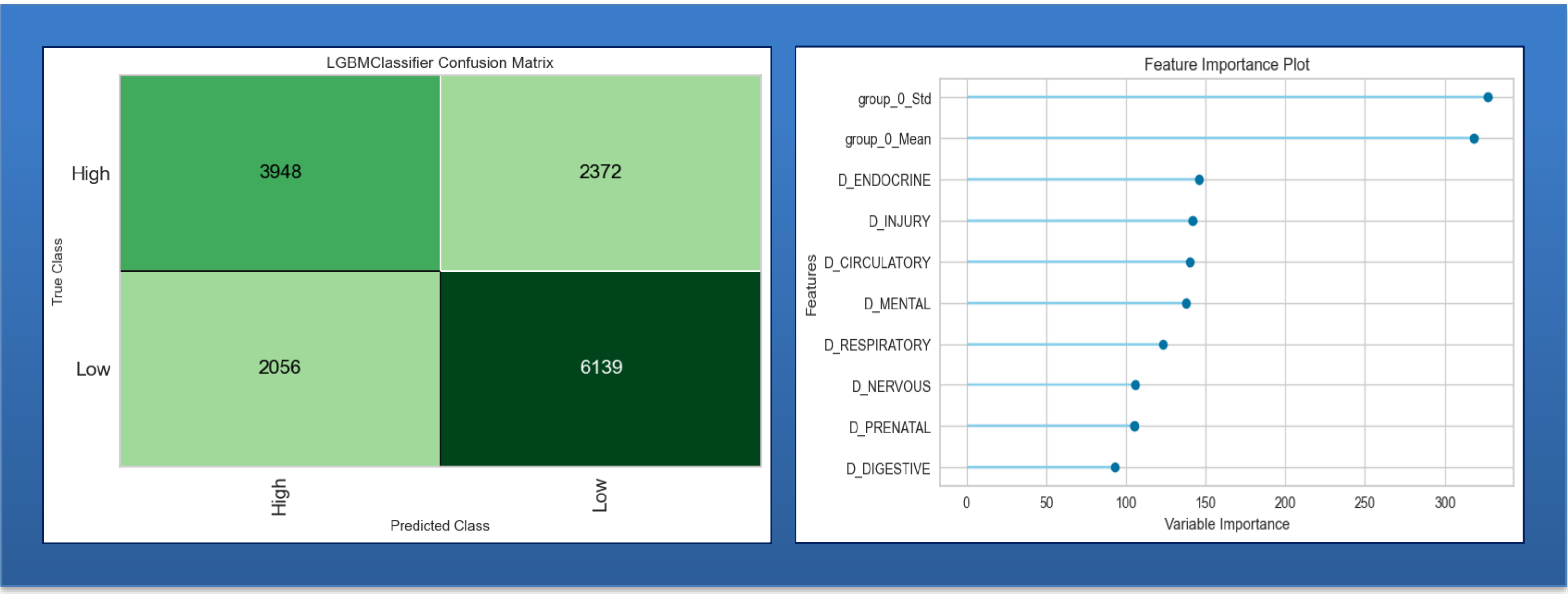
To get a better performing model, we decided to generalize our target variable, LOS, even further into binary class i.e., High and Low. We even decided to step further and test our cleaned data using auto-ML, PyCaret package in python. To our surprise, auto-ML trained and tested our data on 14 classifiers and gave us the model with the best accuracy, F1 & Kappa scores and overall much better Recall & Precision score.



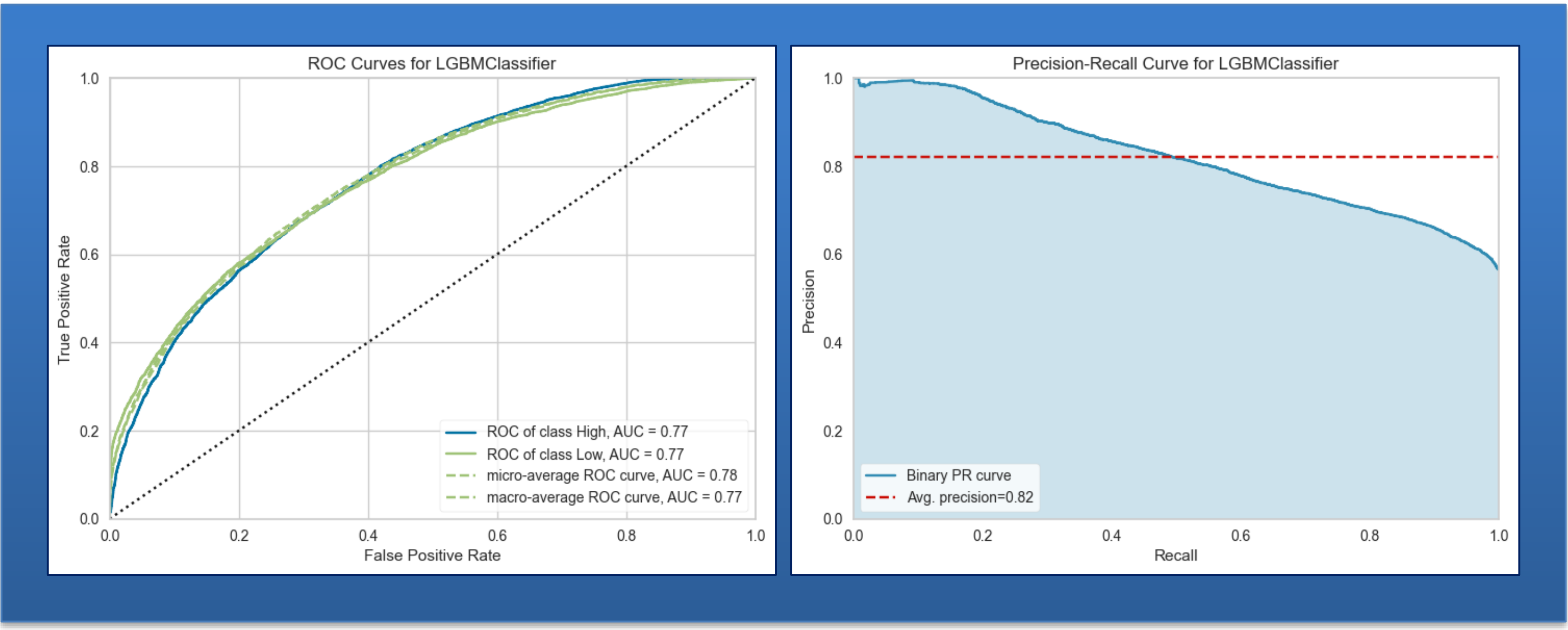
	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
lightgbm	Light Gradient Boosting Machine	0.7045	0.7783	0.7622	0.7250	0.7431	0.3958	0.3965	0.1140
gbc	Gradient Boosting Classifier	0.6998	0.7747	0.7511	0.7240	0.7373	0.3874	0.3877	0.9470
ada	Ada Boost Classifier	0.6936	0.7653	0.7463	0.7184	0.7320	0.3746	0.3751	0.3840
rf	Random Forest Classifier	0.6889	0.7593	0.7241	0.7219	0.7230	0.3681	0.3682	0.8000
lda	Linear Discriminant Analysis	0.6886	0.7621	0.7903	0.6958	0.7400	0.3558	0.3604	0.1720
ridge	Ridge Classifier	0.6884	0.0000	0.7909	0.6953	0.7400	0.3551	0.3597	0.0470
lr	Logistic Regression	0.6878	0.7607	0.7770	0.6995	0.7362	0.3563	0.3593	1.5060
et	Extra Trees Classifier	0.6809	0.7483	0.7160	0.7153	0.7157	0.3522	0.3522	1.1180
svm	SVM - Linear Kernel	0.6736	0.0000	0.7853	0.6900	0.7271	0.3218	0.3399	0.1410
knn	K Neighbors Classifier	0.6565	0.7081	0.7551	0.6727	0.7115	0.2907	0.2937	1.7740
nb	Naive Bayes	0.6521	0.7094	0.6883	0.6904	0.6893	0.2941	0.2941	0.0480
dt	Decision Tree Classifier	0.6237	0.6127	0.6594	0.6662	0.6627	0.2373	0.2373	0.0920
dummy	Dummy Classifier	0.5608	0.5000	1.0000	0.5608	0.7186	0.0000	0.0000	0.0290
qda	Quadratic Discriminant Analysis	0.5129	0.6927	0.1464	0.9125	0.2516	0.1141	0.2211	0.1300

Light Gradient Boosting Machine performed the best among all the Classifiers, with accuracy score of around 70%. Overall, the accuracy of our binary classification model is much better than our multiclass classification model. There is still a lot of room for improvement as the analysis can be further extended on datasets that were not used within our scope of analysis. But for now, this is good enough to implement real-life solutions.

Confusion Matrix and (Top 10) Feature Importance Chart for LGBM Classifier



ROC Curve and Precision-Recall Curve for LGBM Classifier



WAY FORWARD

Now that we have established a working prediction model that successfully classifies patient Length of Stay (LOS) as 'High' (More than 6 days) or 'Low' (6 days or below), we prescribe the following methods that the hospitals or clinics can follow to better their manage time & resources and ensure availability of hospital beds:

Maintain buffer zones for number of beds: Hospitals can maintain a certain buffer zones for a certain number of beds made available for only for low LOS patients. This number can vary hospitals to hospital depending on the traffic of low LOS patients that the hospitals receive and the total number of beds they have. Having a buffer for beds especially for Low LOS patients ensures that not all (or majority) of beds gets occupied by High LOS patients and the whole medical infrastructure does not gets clogged up with non-vacant beds.

Improve communication with patients: Length of Stay can give a rough estimate of a patient's discharge time. This can help doctors to better communicate with patients during the time of admission about the availability of any present/future beds in their medical unit and also lets them refer patients to other hospitals with better bed availability at the moment.

Using Mortality Rate with LOS to give preference to patients in emergency: Patients with High Mortality (such as the ones that come in emergency) can be given higher preference for bed allotment even with higher LOS, as it could be a matter of life and death. The buffer beds can also be utilized at the time to admit such patients.

CONCLUSION

To conclude, with this project we were able to observe the median Length of Stay (LOS) across various patient characteristics such as their age, gender, ethnicity, diagnosis types etc. and then later successfully develop a machine learning model to predict and classify patient LOS using those characteristic features. In the end, we able to constructively recommend hospitals in various ways on how to better manage their time and resources using patient LOS prediction and improve the quality of their services.

REFERENCES

1. MIMIC-III Clinical Database
2. Wikipedia: List of ICD-9 codes
3. Charles, D., King, J., Patel, V. & Furukawa, M. Adoption of Electronic Health record Systems among U.S. Non-federal Acute Care Hospitals. ONC Data Brief No. 9, 1–9 (2013).
4. Collins, F. S. & Tabak, L. A. NIH plans to enhance reproducibility. Nature 505, 612–613 (2014).
5. Waiting Your Turn: Wait Times for Health Care in Canada, 2021 (The Fraser Institute)
6. Comparing Performance of Universal Health Care Countries, 2019 (The Fraser Institute)
7. Canada's seniors population outlook (Canadian Institute of Health Information)

Special thanks to all the people who made and released these great resources for free:

- Word report template by Used to Tech
- Photographs by Unsplash
- Icons by Flaticon