


HEALTHCARE ASSISTANT

Submitted by

Remya Mavila

Samridhi Agrawal

Ashwini Rudrawar

Deleted: 

Deleted: 

| | |
|--|-------------------------------------|
| Abstract | Error! Bookmark not defined. |
| 1.Introduction | 5 |
| 2.Related Work | 7 |
| 3.DataSet Overview | 7 |
| 3.1 What is Mimic 3? | 8 |
| 3.2 Why is it important? | 8 |
| 3.3 Methods to access Mimic III | 9 |
| 3.4 Overview of the MIMIC-III data | 9 |
| 4.Length of Stay | 10 |
| 4.1 Data Analysis and Model Evaluation | 10 |
| 4.2 Exploratory Data Analysis and Visualization: | 10 |
| 4.2.1 ADMISSIONS | 11 |
| LOS | 11 |
| ETHNICITY | 12 |
| RELIGION | 13 |
| ADMISSION_TYPE | 14 |
| INSURANCE | 15 |
| MARITAL_STATUS | 15 |
| 4.2.2 PATIENTS | 16 |
| GENDER | 17 |
| AGE | 17 |
| 4.2.3 DIAGNOSES_ICD | 18 |
| 4.2.4 ICUSTAY | 20 |
| 4.3 Model Building and Evaluation | 21 |
| 4.3.1 Experiments and Results | 22 |
| 5. Chances of Readmission | 22 |
| 5.1 Data Analysis and Model Evaluation | 22 |
| 5.2 Exploratory Data Analysis and Visualization | 23 |
| 5.2.1 ADMISSIONS | 23 |
| 5.2.2 PATIENTS | 29 |

| | |
|-------------------------------|-------------------------------------|
| | 3 |
| 5.2.3 DRGCODES | 30 |
| 5.2.4 LABEVENTS | Error! Bookmark not defined. |
| 5.2.5 D_LABITEMS | |
| Pre-processing | |
| Socio_Economic Bias Check | 36 |
| For Gender | 36 |
| For Religion | 37 |
| For Ethnicity | 33 |
| Model Building and Evaluation | 38 |
| Experiments and Results | 38 |
| Train Test Split Method | Error! Bookmark not defined. |
| Cross Validation | 38 |
| | 37 |
| Future Work | 39 |
| References | 41 |

Abstract

The purpose of this study is to improve the current health care system using machine learning by developing ML models that can predict the length of hospital stay and risk of readmission in advance, based on electronic health records. The study is focussed on using machine learning techniques to predict patient length of stay at the time of admission and suggest the risk of readmission of a patient within 30 days of discharge using MIMIC-III clinical dataset. The main purpose of this thesis is to build models to predict the variables that can be used in a hospital prediction software by applying machine learning techniques for regression and classification models we learned. The metrics used to define the performance for length of stay is root-mean-square error (RMSE). The metrics used to evaluate the chances of Readmission are precision, recall, roc_auc. We were able to achieve good results for both Length of Stay and Readmission models. The results obtained are better than many existing models.

1. Introduction

The requirements for application of Machine Learning techniques within the Healthcare domain is rapidly expanding with improvement of modern computing systems. For example, modern ICUs provide continuous monitoring of critically ill patients susceptible to many complications and mortality, which require a high staff-to-patient ratio and generate a sheer volume of data. For clinicians, the real-time interpretation of these data and decision-making is a challenging task. Machine Learning, powered by increasing availability of healthcare data, can be used in such areas of healthcare for applications ranging from early detection of high-risk events to outcome prediction. In the perspective of a patient and hospital in general, there are multiple unknowns to the patient and hospital like patient length of stay, chances of readmission etc. Supervised machine learning is a good fit for optimizing these hospital procedures.

The healthcare sector is facing ever increasing challenges like economical challenges, lack of expertise, staff and hospital beds. In order to face these challenges, hospitals need to be equipped with administrative planning tools that can allow them to allocate the available resources in an efficient manner. There are various methods identified, applied and restructured with new ways of patient interaction that have been tested by using for example mobile apps. Therefore, a new area of interest that could bring new tools to the hospital's administrative toolbox, namely machine learning (ML). Machine learning has been on the rise in several fields for the last decade following improved computational power, availability of data and improved algorithms. It has excelled in tasks such as image segmentation and classification, machine translation and recommender systems.

Predictive analytics has become an important tool in the healthcare field since modern machine learning (ML) methods can use large amounts of available data to predict individual outcomes for patients. For example, ML predictions can help healthcare providers determine likelihoods of disease, aid in diagnosis, recommend treatment, and predict future wellness. So, in this project we will focus on two most important factors: length-of-stay of patients at the time of admission and risk of readmission of patients within 30days of discharge. Our project is entirely based on the MIMIC-III clinical database. MIMIC is an openly available dataset developed by the MIT Lab for Computational Physiology, comprising de-identified health data associated with approximately 40,000 critical care patients. It includes demographics, vital signs, laboratory tests and medications details.

Patient length of stay is most commonly defined as the total hospitalization time, i.e. from admission to discharge. LoS predictions can be used in many different ways and serve as a very valuable method for resource planning. Not only could it provide an overview of future bed capacity, but it could also be used as a precautionary warning that extra measures should be taken given that a patient's LoS might be longer than usual, such as social planning or extra medical attention. Additionally, certain ML methods can potentially provide valuable insights into what features affect LoS and hence be used as a way to evaluate procedures in order to more efficiently treat patients and reduce unnecessary workload. Reduction in the number of inpatient days results in decreased risk of infection and medication side effects, improvement in the quality of treatment, and increased hospital profit with more efficient bed management.

There has been a study on U.S. hospital stays that cost the health system at least \$377.5 billion per year and recent Medicare legislation standardizes payments for procedures performed, regardless of the number of days a patient spends in the hospital. This incentivizes hospitals to identify patients of high LOS risk at the time of admission. Once identified, patients with high LOS risk can have their treatment plan optimized to minimize LOS and lower the chance of getting a hospital-acquired condition such as staph infection. Another benefit is that prior knowledge of LOS can aid in logistics such as room and bed allocation planning.

The second part of the project will focus on hospital readmission. Hospital readmission is costly for hospitals and is associated with worse outcomes for patients. Many readmissions are avoidable if patients receive further care during their initial admission. Therefore, a predictive model that can indicate whether a patient is likely to be readmitted is very valuable. A study has been done on hospital readmission using CNN(convolutional neural network)[7] which also takes unstructured data like diagnosis notes into account. The hospitals are given a score which is used by many other hospitals to score the readmission risk which is considered as an important factor to decide.

We can learn how data processing techniques and machine learning models can help us to build better health care systems. We can also learn how these models help us to better visualize, summarize and classify this data to observe similarities and differences in each patient's recovery path.

One of the limitations of our project is that it is not tested on real time data. MIMIC -III is a structured clinical database but it is still being maintained and not as real time. The results may differ after deployment on actual data. Like every other model the quality of the model is dependent on the quality of data used to train the model and there could be some unknown or unidentified features correlation between the features that could induce a significant impact on our model.

2. Related Work

Previous studies have examined effective management of LOS. Majority of these involved subjects stratified by condition or admitting unit, for example, patients admitted to specialized departments, such as psychiatric wards or the intensive care unit (ICU) or for patients with hip fractures or undergoing coronary artery surgery. And there are models predicting Length of stay for patients admitted with a specific diagnosis, such as heart failure or pulmonary disease.

Among the numerous works aiming to provide decision-making tools for ICU clinicians at discharge time, two in particular caught our attention in terms of performance and similarity of setting to our own. One of the previous researches proposed an advanced neural network for 30-day ICU readmission prediction (LSTM-CNN based model) achieving an Area Under Curve of the Receiver Operating Characteristic (AUROC) metric of 0.791 on MIMIC-III, using chart events 48h time series, diagnostic ICD-9 codes embeddings, and demographic information of the patients. The authors claim to offer higher sensitivity (0.742) compared to existing solutions, regardless of the specificity trade-off. There is no mention of precision nor F1-score. Another research trained a simpler and more interpretable gradient boosting model (XGBoost) for predicting risk of ICU bounceback and readmission at a variety of time points using MIMIC-III, achieving AUROC of 0.76 and 0.75, F1-score of 0.20 and 0.34, for 72h and 30-days ICU readmission respectively. They use chart events, time series, ICD-9 codes indicators, as well as admission, demographic and length-of-stay information of the patients.

3. Data

In order to discuss health data analytics and the role it plays in the health care sector, we must first understand the data that is being collected and analyzed. There is data being collected on the processes and procedures of the business side of healthcare, but there is also an enormous amount of health data being gathered, stored and analyzed. Health data is any data relating to

the health of an individual patient or collective population. This information is gathered from a series of health information systems (HIS) and other technological tools utilized by health care professionals, insurance companies and government organizations.

There are a variety of tools and systems used to collect, store, share and analyze health data gathered through various means. These tools include:

- Electronic Health Records (EHRs)
- Personal Health Records (PHRs)
- Electronic Prescription Services (E-prescribing)
- Patient Portals
- Master Patient Indexes (MPI)
- Health-Related Smart Phone Apps and more

These data sets are so complex that traditional processing software and storage options cannot be used.

What is Mimic 3?

MIMIC-III ('Medical Information Mart for Intensive Care') is a large, single-center database comprising information relating to patients admitted to critical care units at a large tertiary care hospital. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more. The database supports applications including academic and industrial research, quality improvement initiatives, and higher education coursework.

Why is it important?

The MIMIC-III critical care database is notable for the following reasons:

- It is the one of few freely accessible critical care database of its kind;
- The dataset spans more than a decade, with detailed information about individual patient care.
- Analysis is unrestricted once a data use agreement is accepted, enabling clinical research and education around the world.

It was de-identified in accordance with Health Insurance Portability and Accountability Act (HIPAA) standards using structured data cleansing and date shifting. Protected health information was removed from free text fields, such as diagnostic reports and physician notes,

using a rigorously evaluated de-identification system based on extensive dictionary look-ups and pattern-matching with regular expressions.

Methods to access Mimic III

This data is collected from patients who were admitted to Beth Israel Deaconess Medical Center in Boston, Massachusetts from 2001 to 2012.

Below are the steps to access this data

- complete CITI training course
- create a PhysioNet account
- request access to MIMIC III
- accessing MIMIC III

The MIMIC-III database is also available on two major cloud platforms: Google Cloud Platform (GCP) and Amazon Web Services (AWS). To access the data on the cloud, simply add the relevant cloud identifier to your PhysioNet profile.

Overview of the MIMIC-III data

MIMIC is a relational database containing tables of data relating to patients who stayed within the intensive care units at Beth Israel Deaconess Medical Center. A table is a data storage structure which is similar to a spreadsheet: each column contains consistent information (e.g., patient identifiers), and each row contains an instantiation of that information (e.g. a row could contain the integer 340 in the patient identifier column which would imply that the row's patient identifier is 340).

The tables are linked by identifiers which usually have the suffix "ID". For example HADM_ID refers to a unique hospital admission and SUBJECT_ID refers to a unique patient. One exception is ROW_ID, which is simply a row identifier unique to that table.

Tables prefixed with "D_" are dictionaries and provide definitions for identifiers. For example, every row of OUTPUTEVENTS is associated with a single ITEM_ID which represents the concept measured, but it does not contain the actual name of the drug. By joining OUTPUTEVENTS and D_ITEMS on ITEMID, it is possible to identify what concept a given ITEM ID represents.

The Tables are divided into 4 types

1) The following tables are used to define and track patient stays: ADMISSIONS, CALLOUT, ICU STAYS, PATIENTS, SERVICES, TRANSFERS

2) The following tables contain data collected in the critical care unit: CAREGIVERS, CHARTEVENTS, DATETIMEEVENTS, INPUTEVENTS_CV, INPUTEVENTS_MV, NOTEEVENTS, OUTPUTEVENTS, PROCEDUREEVENTS_MV

3) The following tables contain data collected in the hospital record system: CPTEVENTS, DIAGNOSES_ICD, DRG CODES, LABEVENTS, MICROBIOLOGYEVENTS, PRESCRIPTIONS, PROCEDURES_ICD

4) The following tables are dictionaries: D_CPT, D_ICD_DIAGNOSES, D_ICD_PROCEDURES, D_ITEMS, D_LABITEMS

4. Length of Stay

4.1 Data Analysis and Model Evaluation

The model building to predict the Length of stay has been done using the MIMIC III dataset. This section explains some of the preprocessing methods followed, filling missing data and some exploratory analysis. Problem understanding and model evaluation is also presented.

From the initial analysis conducted on the MIMIC III dataset, it has been identified that the most relevant features that can be selected in or to predict the Length of Stay of are distributed mainly among the four tables listed below. Hence the

The tables used for the Length of Stay Prediction using MIMIC III dataset are,

1. Admission
2. Patients
3. ICUStay
4. Diagnoses_ICD

4.2 Exploratory Data Analysis and Visualization:

Under this section, the process of data exploration and different visualization methods used to identify the underlying relationships in the data has been presented.

4.2.1 ADMISSIONS

From further analysis on MIMIC III data to explore the target variable, it has been figured out that the Admission table has the important features to extract the LOS in days along with many other contributing features. The admission table columns are listed below.

```
RangeIndex: 58976 entries, 0 to 58975
Data columns (total 19 columns):
#   Column                Non-Null Count  Dtype
---  -
0   ROW_ID                 58976 non-null  int64
1   SUBJECT_ID             58976 non-null  int64
2   HADM_ID                58976 non-null  int64
3   ADMITTIME              58976 non-null  object
4   DISCHTIME              58976 non-null  object
5   DEATHTIME              5854 non-null   object
6   ADMISSION_TYPE         58976 non-null  object
7   ADMISSION_LOCATION     58976 non-null  object
8   DISCHARGE_LOCATION     58976 non-null  object
9   INSURANCE              58976 non-null  object
10  LANGUAGE               33644 non-null  object
11  RELIGION                58518 non-null  object
12  MARITAL_STATUS         48848 non-null  object
13  ETHNICITY               58976 non-null  object
14  EDREGTIME               30877 non-null  object
15  EDOUTTIME               30877 non-null  object
16  DIAGNOSIS              58951 non-null  object
17  HOSPITAL_EXPIRE_FLAG   58976 non-null  int64
18  HAS_CHARTEVENTS_DATA   58976 non-null  int64
dtypes: int64(5), object(14)
memory usage: 8.5+ MB
```

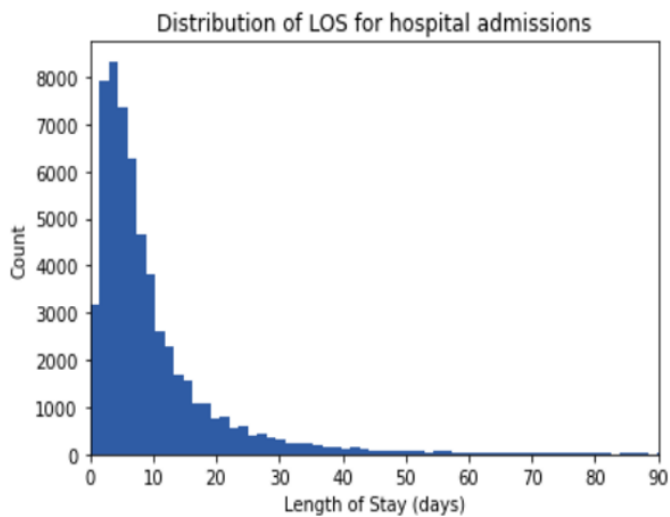
| | ROW_ID | SUBJECT_ID | HADM_ID | ADMITTIME | DISCHTIME | DEATHTIME | ADMISSION_TYPE | ADMISSION_LOCATION | DISCHARGE_LOCATION | INSURANC |
|---|--------|------------|---------|---------------------|---------------------|-----------|----------------|---------------------------------|------------------------------|----------|
| 0 | 21 | 22 | 165315 | 4/9/2196 12:26 | 4/10/2196 15:54 | NaN | EMERGENCY | EMERGENCY ROOM ADMIT | DISC-TRAN CANCER/CHLDRN H | Privat |
| 1 | 22 | 23 | 152223 | 9/3/2153 7:15 | 9/8/2153 19:10 | NaN | ELECTIVE | PHYS REFERRAL/NORMAL DELI | HOME HEALTH CARE | Medicar |
| 2 | 23 | 23 | 124321 | 10/18/2157 19:34 | 10/25/2157 14:00 | NaN | EMERGENCY | TRANSFER FROM HOSP/EXTRAM | HOME HEALTH CARE | Medicar |
| 3 | 24 | 24 | 161859 | 6/6/2139 16:14 | 6/9/2139 12:48 | NaN | EMERGENCY | TRANSFER FROM HOSP/EXTRAM | HOME | Privat |
| 4 | 25 | 25 | 129635 | 11/2/2160 2:06 | 11/5/2160 14:55 | NaN | EMERGENCY | EMERGENCY ROOM ADMIT | HOME | Privat |

LOS

The length of Stay is calculated in number of days using the ADMITTIME and DISCHTIME by subtracting ADMITTIME from DISCHTIME and dividing by 24*60*60 (number of seconds in a day)

$$\text{LOS} = (\text{DISCHTIME} - \text{ADMITTIME}) / 24*60*60$$

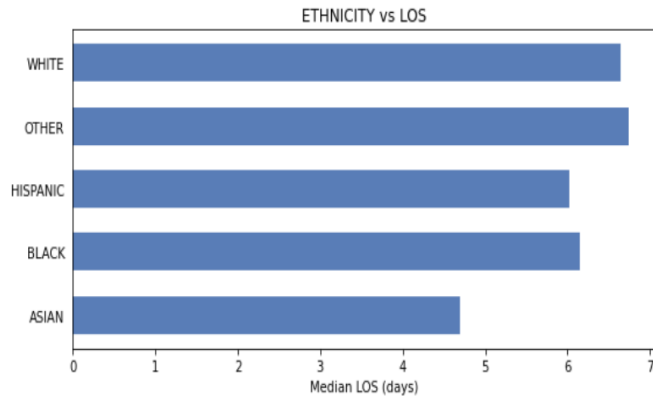
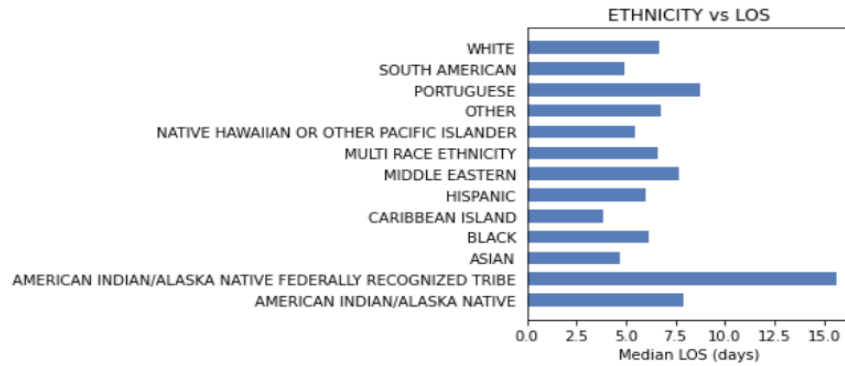
For the following analysis, length of stay is kept as the primary variable along the y-axis of the plots I create since it is the predictor variable for this project. First, the distribution of length of stay is visualized.



The above distribution shows that even though the hospital stay ranges for a few months, the distribution is very skewed, with most of the patients in the data having lengths of stay between ~1 to 7 days or most of the patients staying for less than 10 days in the hospital.

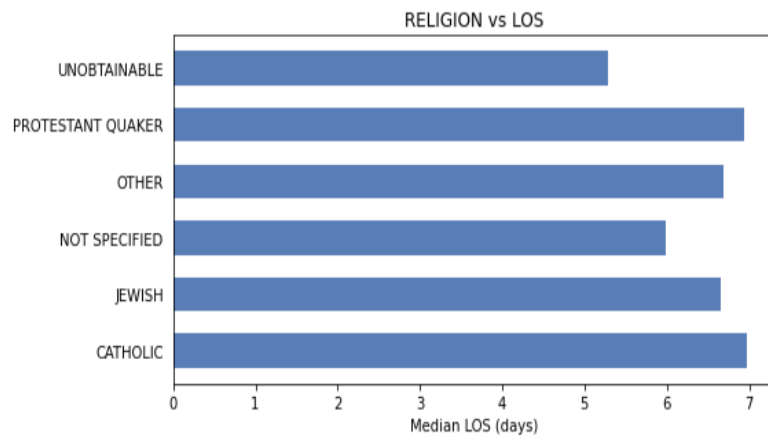
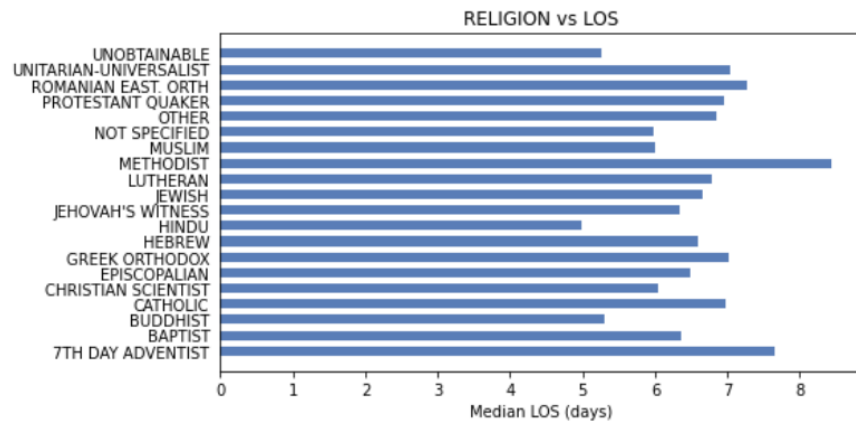
ETHNICITY

Next let's look at the length of stay distribution of various Ethnicities. As in the figure below there were many Ethnicity categories and to get the best result it has been combined and categorized to 5 different categories like Asian, Black, Hispanic, White and Other. The distribution of Ethnicity categories before and after are shown below.



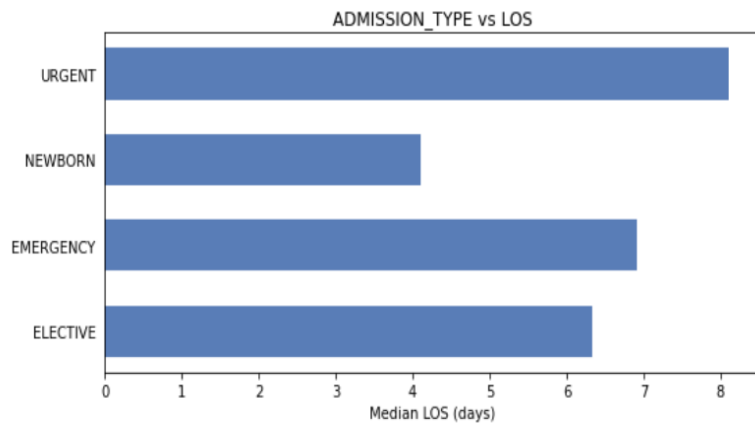
RELIGION

Similarly, the length of stay distribution of different religious groups has been visualized and analyzed to identify the category wise distribution. Based on the number of people belonging to each category, the religious categories are combined into six different groups as in figure below.



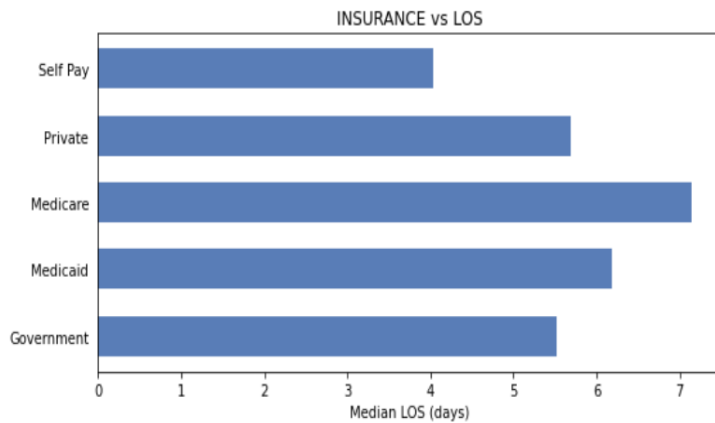
ADMISSION_TYPE

Another feature that can contribute to LOS prediction is Admission Type. There are four different types of admissions present in the dataset which are, Emergency, Newborn, Elective and Urgent. The admission type distribution is as given below.



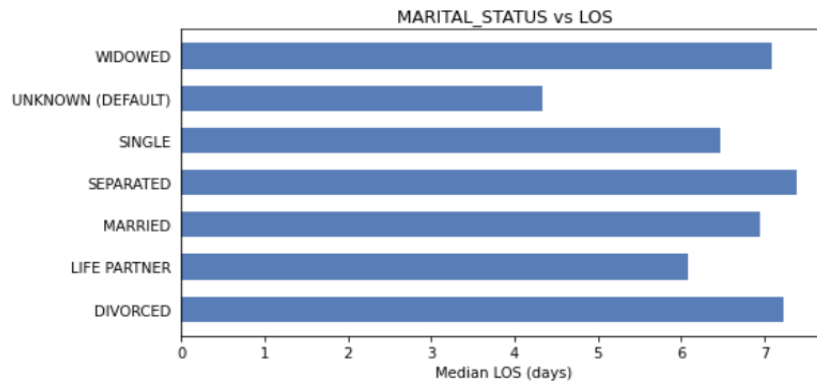
INSURANCE

The Length of Stay distribution has been visualized and analyzed for the multiple insurance types present in the dataset. The diagram below shows Medicare has the most number of days of average length of stay.



MARITAL_STATUS

After filling all the missing values as unknown, the marital status of all the hospital admissions records shows the below listed categories where the distribution shows Median LOS is more in the cases of Separated status.



4.2.2 PATIENTS

The second important table that was used in the LOS prediction is Patients. It has the information regarding the Gender and Date Of Birth of all the patients having SUBJECT_ID as it's Key. The columns of the Patients table are listed below.

```

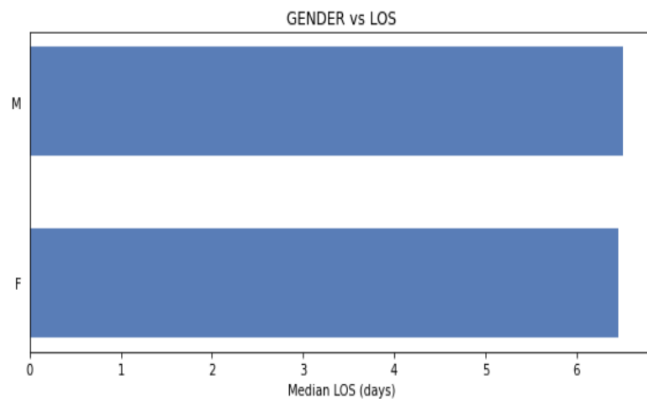
RangeIndex: 46520 entries, 0 to 46519
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   ROW_ID      46520 non-null  int64
1   SUBJECT_ID  46520 non-null  int64
2   GENDER      46520 non-null  object
3   DOB         46520 non-null  object
4   DOD         15759 non-null  object
5   DOD_HOSP    9974 non-null   object
6   DOD_SSN     13378 non-null  object
7   EXPIRE_FLAG 46520 non-null  int64
dtypes: int64(3), object(5)
memory usage: 2.8+ MB

```


| ROW_ID | SUBJECT_ID | GENDER | DOB | | DOD | DOD_HOSP | DOD_SSN | EXPIRE_FLAG |
|--------|------------|--------|-----|---------------------|---------------------|---------------------|---------|-------------|
| 0 | 234 | 249 | F | 2075-03-13 00:00:00 | NaN | NaN | NaN | 0 |
| 1 | 235 | 250 | F | 2164-12-27 00:00:00 | 2188-11-22 00:00:00 | 2188-11-22 00:00:00 | NaN | 1 |
| 2 | 236 | 251 | M | 2090-03-15 00:00:00 | NaN | NaN | NaN | 0 |
| 3 | 237 | 252 | M | 2078-03-06 00:00:00 | NaN | NaN | NaN | 0 |
| 4 | 238 | 253 | F | 2089-11-26 00:00:00 | NaN | NaN | NaN | 0 |

GENDER

The Length of stay Distribution of different patients based on their Gender is given below. Both men and women have almost similar distribution of LOS in the data.



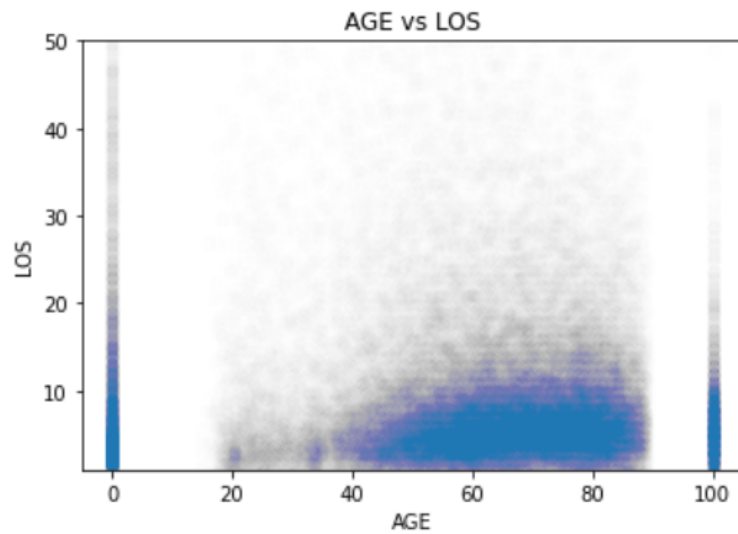
AGE

The DOB information of Patients table can be used to calculate the Age of a Patient. The Age of a patient will be the date of birth subtracted from the first admit time and divided by the number of days in a year.

$$\text{AGE} = (\text{FIRST ADMIT TIME} - \text{DOB}) / 365$$

To extract the first admit time the minimum value of the ADMITTIME column of the Admissions table is used for each SUBJECT_ID and converted to date and subtracted the DOB date from the Patients table to calculate the Age.

MIMIC III data values are encoded for patients who are more than 89 years old. The age values above 100 are aggregated to 100. The below given scatter plot shows the distribution of LOS for ages of admitted Patients ranging from 0 to 100.



4.2.3 DIAGNOSES_ICD

The diagnoses_ICD table has all the details of each diagnosis for hospital admissions encoded with ICD codes which can be used to predict the Length of stay for the admission. The columns are given below.

```

RangeIndex: 651047 entries, 0 to 651046
Data columns (total 5 columns):
#   Column      Non-Null Count  Dtype
---  -
0   ROW_ID      651047 non-null  int64
1   SUBJECT_ID  651047 non-null  int64
2   HADM_ID     651047 non-null  int64
3   SEQ_NUM     651000 non-null  float64
4   ICD9_CODE   651000 non-null  object
dtypes: float64(1), int64(3), object(1)
memory usage: 24.8+ MB

```

| ROW_ID | SUBJECT_ID | HADM_ID | SEQ_NUM | ICD9_CODE |
|--------|------------|---------|---------|-----------|
| 0 | 1297 | 109 | 172335 | 1.0 40301 |
| 1 | 1298 | 109 | 172335 | 2.0 486 |
| 2 | 1299 | 109 | 172335 | 3.0 58281 |
| 3 | 1300 | 109 | 172335 | 4.0 5855 |
| 4 | 1301 | 109 | 172335 | 5.0 4254 |

The first three digits of the ICD code indicate the main category of the diagnoses. Hence using the ICD Code category Ranges the ICD9_CODE can be classified into 18 Categories. The International Classification of Diagnoses codes and corresponding Categories are given below.

International Statistical Classification of Diseases and Related Health Problems

- 001–139: infectious and parasitic diseases
- 140–239: neoplasms
- 240–279: endocrine, nutritional, and metabolic diseases, and immunity disorders
- 280–289: diseases of the blood and blood-forming organs
- 290–319: mental disorders
- 320–389: diseases of the nervous system and sense organs
- 390–459: diseases of the circulatory system
- 460–519: diseases of the respiratory system
- 520–579: diseases of the digestive system
- 580–629: diseases of the genitourinary system
- 630–679: complications of pregnancy, childbirth, and the puerperium
- 680–709: diseases of the skin and subcutaneous tissue
- 710–739: diseases of the musculoskeletal system and connective tissue
- 740–759: congenital anomalies
- 760–779: certain conditions originating in the perinatal period
- 780–799: symptoms, signs, and ill-defined conditions
- 800–999: injury and poisoning
- E and V codes: external causes of injury and supplemental classification

ICD-9 Categories

0: 'infectious', 1: 'neoplasms', 2: 'endocrine', 3: 'blood', 4: 'mental', 5: 'nervous', 6: 'circulatory', 7: 'respiratory', 8: 'digestive', 9: 'genitourinary', 10: 'pregnancy', 11: 'skin', 12: 'muscular', 13: 'congenital', 14: 'prenatal', 15: 'misc', 16: 'injury', 17: 'misc'

| | SUBJECT_ID | HADM_ID | LOS | GENDER | blood | circulatory | congenital | digestive | endocrine | genitourinary | ... |
|--------------|--------------|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|-----|
| count | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | 58878.000000 | ... |
| mean | 33761.791382 | 149866.149886 | 10.151266 | 0.558613 | 0.395988 | 2.379089 | 0.070587 | 0.654880 | 1.217178 | 0.556592 | ... |
| std | 28092.613275 | 28882.995648 | 12.459774 | 0.496557 | 0.678072 | 2.278877 | 0.343045 | 1.163365 | 1.354162 | 0.872232 | ... |
| min | 2.000000 | 100001.000000 | 0.001389 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... |
| 25% | 11999.250000 | 124942.750000 | 3.755556 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | ... |
| 50% | 24141.000000 | 149887.000000 | 6.489583 | 1.000000 | 0.000000 | 2.000000 | 0.000000 | 0.000000 | 1.000000 | 0.000000 | ... |
| 75% | 53862.750000 | 174958.000000 | 11.805556 | 1.000000 | 1.000000 | 4.000000 | 0.000000 | 1.000000 | 2.000000 | 1.000000 | ... |
| max | 99999.000000 | 199999.000000 | 294.660417 | 1.000000 | 7.000000 | 17.000000 | 11.000000 | 11.000000 | 12.000000 | 7.000000 | ... |

4.2.4 ICUSTAY

The ICUSTAY table contains the information regarding the icustay for the hospital admissions which is an important feature that can determine the Length of Stay. The table columns are as listed below from which the first care unit and the length of stay can be more useful for the model.

RangeIndex: 61532 entries, 0 to 61531

Data columns (total 12 columns):

| # | Column | Non-Null Count | Dtype |
|----|----------------|----------------|---------|
| 0 | ROW_ID | 61532 non-null | int64 |
| 1 | SUBJECT_ID | 61532 non-null | int64 |
| 2 | HADM_ID | 61532 non-null | int64 |
| 3 | ICUSTAY_ID | 61532 non-null | int64 |
| 4 | DBSOURCE | 61532 non-null | object |
| 5 | FIRST_CAREUNIT | 61532 non-null | object |
| 6 | LAST_CAREUNIT | 61532 non-null | object |
| 7 | FIRST_WARDID | 61532 non-null | int64 |
| 8 | LAST_WARDID | 61532 non-null | int64 |
| 9 | INTIME | 61532 non-null | object |
| 10 | OUTTIME | 61522 non-null | object |
| 11 | LOS | 61522 non-null | float64 |

dtypes: float64(1), int64(6), object(5)

memory usage: 5.6+ MB

| | SUBJECT_ID | HADM_ID | ICUSTAY_ID | FIRST_CAREUNIT | LOS |
|----------|------------|---------|------------|----------------|--------|
| 0 | 268 | 110404 | 280836 | MICU | 3.2490 |
| 1 | 269 | 106296 | 206613 | MICU | 3.2788 |
| 2 | 270 | 188028 | 220345 | CCU | 2.8939 |
| 3 | 271 | 173727 | 249196 | MICU | 2.0600 |
| 4 | 272 | 164716 | 210407 | CCU | 1.6202 |

The First care unit categories are aggregated to ICU and NICU where the various other categories other than the NICU can be classified as ICU for simplification. The final structure of ICUSTAY data will look like below.

| | SUBJECT_ID | HADM_ID | ICUSTAY_ID | FIRST_CAREUNIT | LOS |
|---|------------|---------|------------|----------------|--------|
| 0 | 268 | 110404 | 280836 | ICU | 3.2490 |
| 1 | 269 | 106296 | 206613 | ICU | 3.2788 |
| 2 | 270 | 188028 | 220345 | ICU | 2.8939 |
| 3 | 271 | 173727 | 249196 | ICU | 2.0600 |
| 4 | 272 | 164716 | 210407 | ICU | 1.6202 |

4.3 Model Building and Evaluation

After combining all the various processed data together and final verification of non-null values or any unwanted data, the data has been standardized and normalized and made ready for using the algorithm. The final data has all the categorical columns along with the main attributes. In order to check the model performance built a model and calculated the r2 score which was .374.

Algorithm Used:

The below algorithms were experimented during model creation,

1. Linear Regression
2. Random Forest Regressor
3. KNN Regression
4. Gradient Boosting Regressor
5. SGD Regression

Evaluation: The model has been evaluated using Kfold cross validation with K =10 and the performance metrics such as MSE, RMSE, MAE and R2 Score are calculated. The performance of the model using 10 fold cross validation with various algorithms are as below.

4.3.1 Experiments and Results

| Results/ Model | R2- SCORE | MAE | MSE | RMSE |
|----------------------|-------------------------|--------------------------|---------------------------|--------------------------|
| Linear Regression | 0.360723251073 1572 | 0.019684999337 371663 | 0.001100453934 8199005 | 0.033173090522 589245 |
| Random Forest | 0.344540531563 5881 | 0.019320084907 308688 | 0.001128310942 58733 | 0.033590216218 057675 |
| KNN Regression | 0.212346912193 54593 | 0.020146242192 356844 | 0.001355869646 7605277 | 0.036822135282 470074 |
| Gradient Boosting | 0.433261445368 66024 | 0.017837393360 655686 | 0.000975586353 6491177 | 0.031234374597 513086 |
| SGD Regression | 0.309826442388 6207 | 0.019894266511 280696 | 0.001188067935 2988754 | 0.034468355102 61455 |

Since the performance was the highest when using GradientBoostingRegressor, the model is created using GradientBoostingRegressor. Feature selection is performed using SelectKBest with K=62..

5. Chances of Readmission

5.1 Data Analysis and Model Evaluation

The model building to predict binary classification of Chances of Readmission has been done using the MIMIC III dataset. This section explains some of the preprocessing methods followed, filling missing data and some exploratory analysis. Problem understanding and model evaluation is also presented. From the initial analysis conducted on the MIMIC III dataset, it has been identified that the most relevant features that can be selected in or to predict the Chances of Readmission are distributed mainly among the five tables listed below.

| Tables Used |
|-------------|
| ADMISSIONS |
| PATIENTS |
| DRGCODES |
| LABEVENTS |
| D_LABITEMS |

5.2 Exploratory Data Analysis and Visualization

5.2.1 ADMISSIONS

Each row of this table contains a unique HADM_ID, which represents a single patient's admission to the hospital. HADM_ID ranges from 1000000 - 1999999. It is possible for this table to have duplicate SUBJECT_ID, indicating that a single patient had multiple admissions to the hospital. The ADMISSIONS table is linked to the PATIENTS table using SUBJECT_ID as Foreign Key. ADMISSIONS table also contains admit time and discharge time with this we can create CHANCE_OF_READMISSION.

| | SUBJECT_ID | HADM_ID | ADMITTIME | DISCHTIME | DEATHTIME | ADMISSION_TYPE | ADMISSION_LOCATION | DISCHARGE_LOCATION | INSURANCE |
|---|------------|---------|------------------------|------------------------|-----------|----------------|---------------------------------|------------------------------|-----------|
| 0 | 22 | 165315 | 2196-04-09 12:26:00 | 2196-04-10 15:54:00 | NaN | EMERGENCY | EMERGENCY ROOM ADMIT | DISC-TRAN CANCER/CHLDRN H | Private |
| 1 | 23 | 152223 | 2153-09-03 07:15:00 | 2153-09-08 19:10:00 | NaN | ELECTIVE | PHYS REFERRAL/NORMAL DELI | HOME HEALTH CARE | Medicare |
| 2 | 23 | 124321 | 2157-10-18 19:34:00 | 2157-10-25 14:00:00 | NaN | EMERGENCY | TRANSFER FROM HOSP/EXTRAM | HOME HEALTH CARE | Medicare |
| 3 | 24 | 161859 | 2139-06-06 16:14:00 | 2139-06-09 12:48:00 | NaN | EMERGENCY | TRANSFER FROM HOSP/EXTRAM | HOME | Private |
| 4 | 25 | 129635 | 2160-11-02 02:06:00 | 2160-11-05 14:55:00 | NaN | EMERGENCY | EMERGENCY ROOM ADMIT | HOME | Private |

```

df_admission['LOS'].describe()
[16] ✓ 0.8s Pyth
... count    58976.000000
   mean      10.133916
   std       12.456682
   min       -0.945139
   25%        3.743750
   50%        6.467814
   75%       11.795139
   max       294.660417
   Name: LOS, dtype: float64

```

CHANCE OF READMISSION

A new Column has been created for Next admit time by grouping using Subject_Id and the Chances of Readmission is calculated using the below Steps.

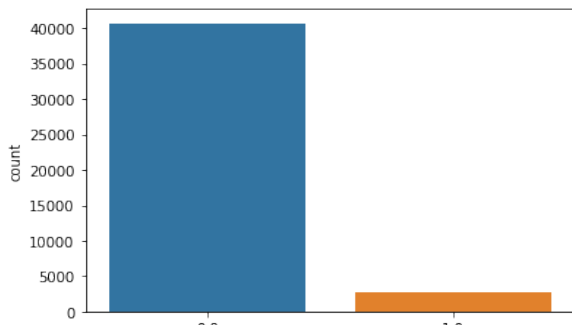
Step 1. *DAYS_TO_NEXT_ADMIT= NEXT_ADMITTIME- DISCHTIME*

Step 2. *IF DAYS_TO_NEXT_ADMIT <= 30*

THEN CHANCE_OF_READMISSION]= 1

Step 3. *FILLNA 0*

Below is the distribution of the target variable Chance of Readmission

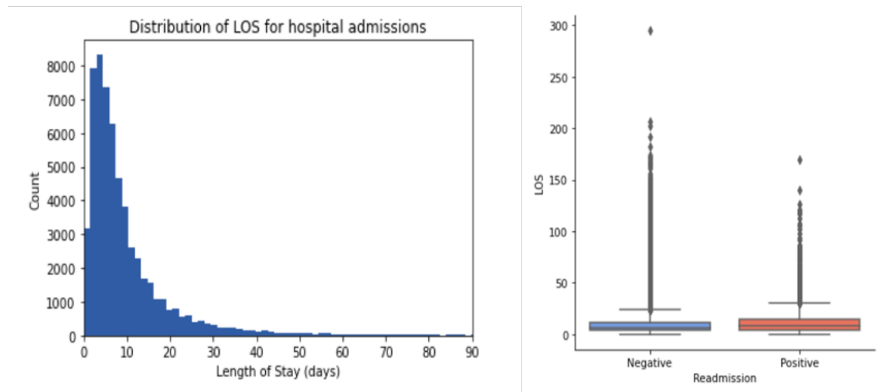


LOS

The length of Stay is calculated in number of days using the ADMITTIME and DISCHTIME by subtracting ADMITTIME from DISCHTIME and dividing by 24*60*60 (number of seconds in a day)

$$LOS = (DISCHTIME - ADMITTIME) / 24*60*60$$

For the following analysis, length of stay is kept as the primary variable along the y-axis of the plots I create since it is the predictor variable for this project. the distribution of length of stay is visualized.



The negative LOS columns were removed after calculating the LOS. Performed more processing on other categorical features such as admission Type. Removed newborn and elective type admissions and combine emergency, urgent admission type, merge categories with less number of samples need to be performed.

Below are some of the figures and plots with the processed features.

```

... WHITE 30252
BLACK/AFRICAN AMERICAN 4224
UNKNOWN/NOT SPECIFIED 3535
HISPANIC OR LATINO 1169
OTHER 961
UNABLE TO OBTAIN 697
ASIAN 695
PATIENT DECLINED TO ANSWER 296
HISPANIC/LATINO - PUERTO RICAN 204
ASIAN - CHINESE 189
BLACK/CAPE VERDEAN 148
WHITE - RUSSIAN 145
BLACK/HAITIAN 93
MULTI RACE ETHNICITY 89
ASIAN - ASIAN INDIAN 66
HISPANIC/LATINO - DOMINICAN 66
PORTUGUESE 50
WHITE - OTHER EUROPEAN 47
WHITE - BRAZILIAN 46
ASIAN - VIETNAMESE 38
BLACK/AFRICAN 37
HISPANIC/LATINO - GUATEMALAN 35
MIDDLE EASTERN 35
AMERICAN INDIAN/ALASKA NATIVE 22
ASIAN - FILIPINO 22

show more (open the raw output data in a text editor) ...

CARIBBEAN ISLAND 7
ASIAN - JAPANESE 5
HISPANIC/LATINO - HONDURAN 4
ASIAN - THAI 3
AMERICAN INDIAN/ALASKA NATIVE FEDERALLY RECOGNIZED TRIBE 3
Name: ETHNICITY, dtype: int64

WHITE 30507
OTHER 5715
BLACK 4502
HISPANIC 1541
ASIAN 1057
Name: ETHNICITY, dtype: int64

```

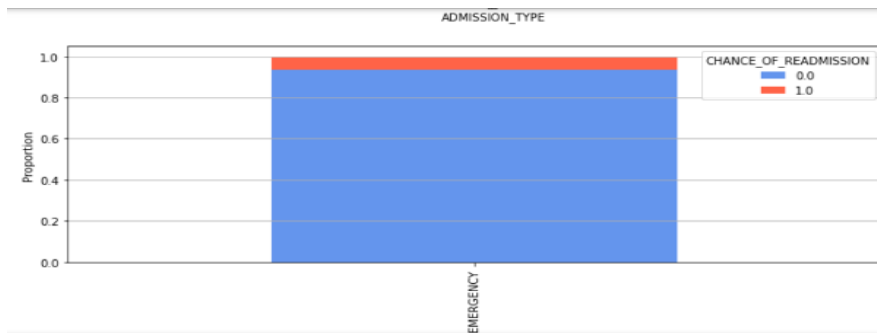
Reducing number of 21 religion type columns to 6 columns

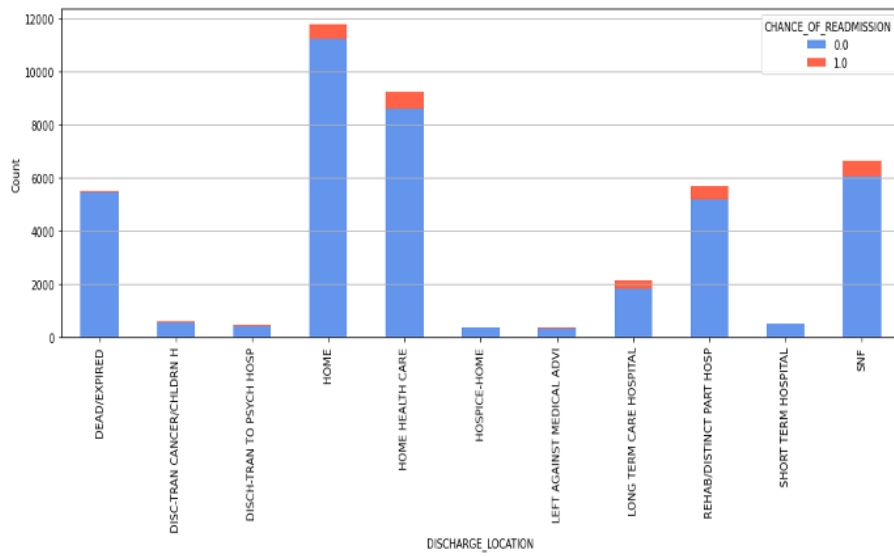
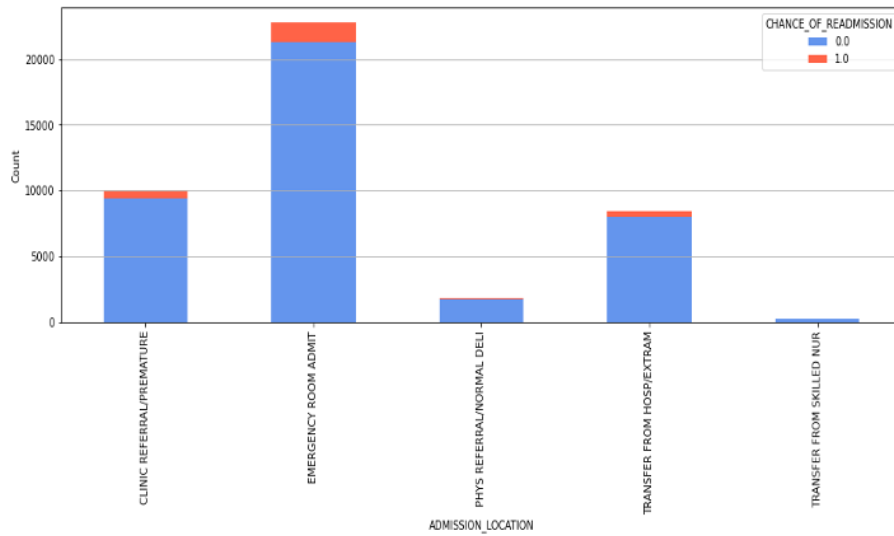
| | |
|------------------------|-------|
| CATHOLIC | 15275 |
| NOT SPECIFIED | 8700 |
| PROTESTANT QUAKER | 5620 |
| UNOBTAINABLE | 5037 |
| JEWISH | 4241 |
| OTHER | 2087 |
| EPISCOPALIAN | 573 |
| NaN | 425 |
| GREEK ORTHODOX | 364 |
| CHRISTIAN SCIENTIST | 260 |
| BUDDHIST | 174 |
| MUSLIM | 161 |
| JEHOVAH'S WITNESS | 110 |
| UNITARIAN-UNIVERSALIST | 80 |
| 7TH DAY ADVENTIST | 62 |
| ROMANIAN EAST. ORTH | 57 |
| HINDU | 52 |
| BAPTIST | 23 |
| HEBREW | 15 |
| METHODIST | 5 |
| LUTHERAN | 1 |

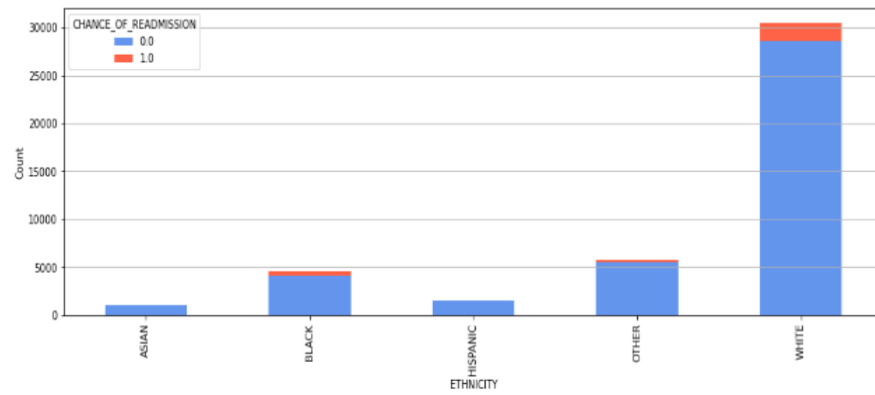
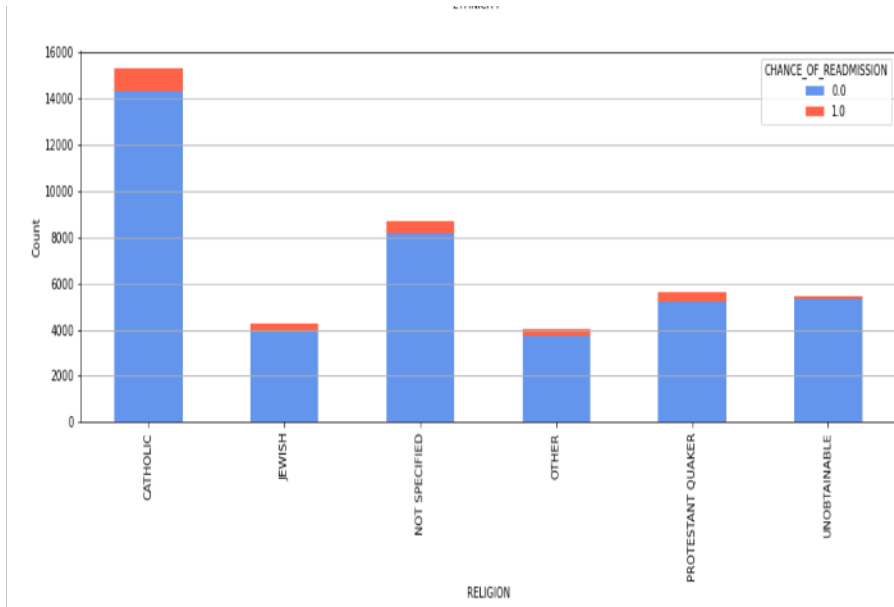
Name: RELIGION, dtype: int64

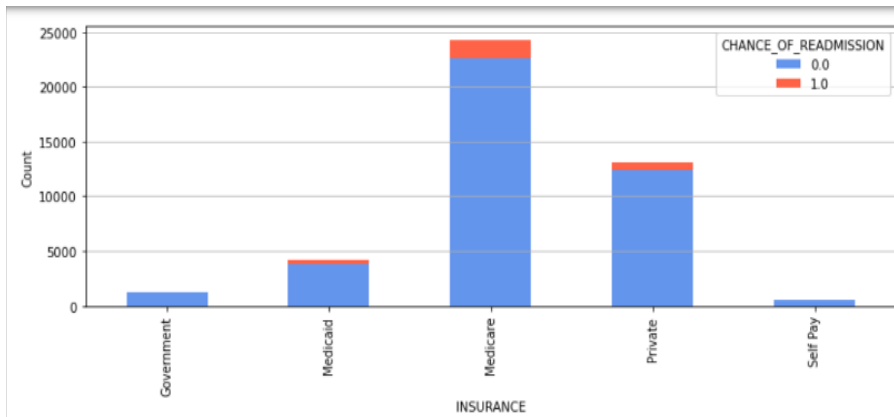
| | |
|-------------------|-------|
| CATHOLIC | 15275 |
| NOT SPECIFIED | 8700 |
| PROTESTANT QUAKER | 5620 |
| UNOBTAINABLE | 5462 |
| JEWISH | 4241 |
| OTHER | 4024 |

Name: RELIGION, dtype: int64









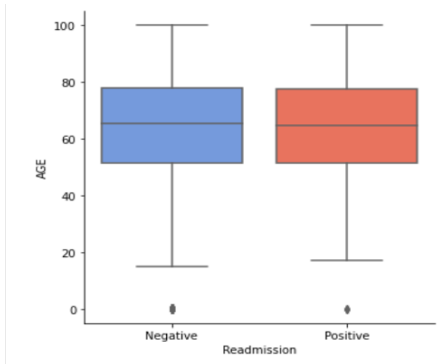
5.2.2 PATIENTS

The Patients table has the patient related data such as Gender, DOB etc. Calculated the age of patients using date of birth and first admit time and classify patients 5 categories like young_child, youth, adult, middle_adult, senior_adult, old age based on their age. Below are the calculation steps and visualization of Age Vs Readmission.

$$AGE = (FIRST\ ADMIT\ TIME - DOB) / 365$$

To extract the first admit time the minimum value of the ADMITTIME column of the Admissions table is used for each SUBJECT_ID and converted to date and subtracted the DOB date from the Patients table to calculate the Age.

MIMIC III data values are encoded for patients who are more than 89 years old. The age values above 100 are aggregated to 100. The below given scatter plot shows the distribution of LOS for ages of admitted Patients ranging from 0 to 100.



Used LabelEncoder to convert string values of M, F of Gender to 1(M) and 0(F). Create dummy columns for Categorical features and merge them into a single dataframe.

5.2.3 DRGCODES

Contains diagnosis related groups (DRG) codes for patient's diagnosis. Number of rows in this table is 125,557. Links to PATIENTS on SUBJECT_ID and ADMISSIONS on HADM_ID.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 125557 entries, 0 to 125556
Data columns (total 8 columns):
#   Column          Non-Null Count  Dtype
---  -
0   ROW_ID          125557 non-null  int64
1   SUBJECT_ID      125557 non-null  int64
2   HADM_ID         125557 non-null  int64
3   DRG_TYPE        125557 non-null  object
4   DRG_CODE        125557 non-null  int64
5   DESCRIPTION     125494 non-null  object
6   DRG_SEVERITY    66634 non-null   float64
7   DRG_MORTALITY   66634 non-null   float64
dtypes: float64(2), int64(4), object(2)
memory usage: 7.7+ MB
```

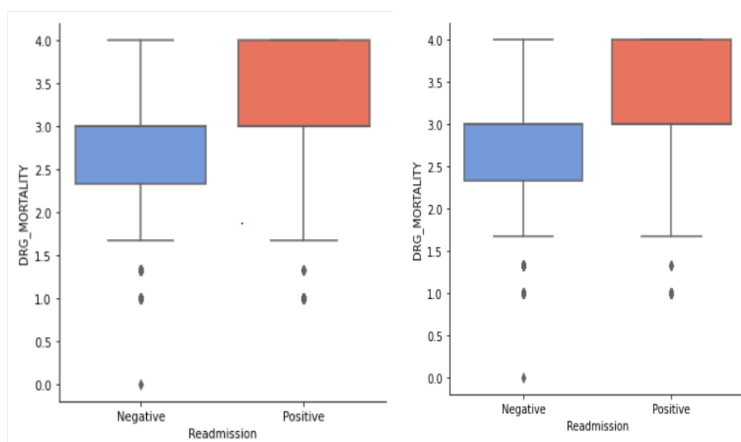
1. SUBJECT_ID is unique to a patient and HADM_ID is unique to a patient's hospital stay.
2. DRG_TYPE: DRG_TYPE provides the type of DRG code in the entry. There are two types of DRG codes in the database which have overlapping ranges but distinct definitions for the codes. The three types of DRG codes in the MIMIC-III database are 'HCFA' (Health Care Financing Administration), 'MS' (Medicare), and 'APR' (All Payers Registry).
3. DRG_CODE: DRG_CODE contains a code which represents the diagnosis billed for by the hospital.

4. **DESCRIPTION:** DESCRIPTION provides a human understandable summary of the meaning of the given DRG code. The description field frequently has acronyms which represent comorbidity levels (comorbid conditions or “CC”). The following table provides a definition for some of these acronyms:

| Acronym | Description |
|------------|--|
| w CC/MCC | with CC or Major CC |
| w MCC | with Major CC |
| w CC | with CC and without Major CC |
| w NonCC | with NonCC and without CC or Major CC |
| w/o MCC | with CC or Non CC and without Major CC |
| w/o CC/MCC | with nonCC and without CC or Major CC |

There are three levels of comorbidities: none, with comorbid conditions, and with major comorbid conditions. These acronyms are primarily used in HCFA/MS DRG codes.

5. **DRG_SEVERITY, DRG_MORTALITY:** DRG_SEVERITY and DRG_MORTALITY are the Severity and Mortality scores of diagnosis ranging from 0-4 and provide additional granularity to DRG codes in the ‘APR’ DRG type. Below are the box plots that shows the distribution of DRG_SEVERITY and DRG_MORTALITY.



After preprocessing steps such as removing unwanted columns, combining descriptions per admission and filling out severity, mortality scores used those features for the model.

```
Int64Index: 58890 entries, 0 to 58889
Data columns (total 4 columns):
#   Column          Non-Null Count  Dtype
---  -
0   HADM_ID         58890 non-null  int64
1   DESCRIPTION     58890 non-null  object
2   DRG_SEVERITY    58890 non-null  float64
3   DRG_MORTALITY   58890 non-null  float64
dtypes: float64(2), int64(1), object(1)
memory usage: 2.2+ MB
```

We used Natural language ToolKit, NLTK to process the description column of the DRGCODES table. It provides us with various text processing libraries. Using snowball stemmer we stemmed the description. Below is a sample output.


```

0          diabet w cc diabet diabet
1  peptic ulcer gastriti peptic ulcer gastriti gi...
2          chronic obstruct pulmonari diseas
3  major small larg bowel procedur w cc w major g...
4  coronari bypass wo cardiac cath or percutan ca...
Name: DESCRIPTION, dtype: object

```

Using TfidfVectorizer created word arrays from the stemmed description by excluding any word with more than 50% occurrence and used top 100 words from the remaining.

Combined severity, mortality and description features into the main dataframe

| | acut | age | ami | bowel | bwt | bypass | card | cardiac | cardiothorac | cardiovascular | ... | term | tracheostomi | tract | trauma | unrel | valv |
|---|------|-----|-----|----------|-----|----------|------|----------|--------------|----------------|-----|------|--------------|-------|--------|-------|------|
| 0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | 0.0 | 0.0 | 0.0 | 0.513495 | 0.0 | 0.000000 | 0.0 | 0.000000 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.429431 | 0.0 | 0.418934 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Removed all the columns with missing values more than 1/5th of its length and imputed the remaining with median for all the new columns added to the dataframe.

5.2.4. LABEVENTS

Contains all laboratory measurements for a given patient, including out patient data. Number of rows is 27,854,055. Links to PATIENTS on SUBJECT_ID and ADMISSIONS on HADM_ID and D_LABITEMS on ITEMID. The LABEVENTS data contains information regarding laboratory based measurements.

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27854055 entries, 0 to 27854054
Data columns (total 9 columns):
#   Column      Dtype
---  ---
0   ROW_ID      int64
1   SUBJECT_ID  int64
2   HADM_ID     float64
3   ITEMID      int64
4   CHARTTIME   object
5   VALUE       object
6   VALUENUM    float64
7   VALUEUOM    object
8   FLAG        object
dtypes: float64(2), int64(3), object(4)
memory usage: 1.9+ GB

```

5.2.5 D_LABITEMS

Definition table for all laboratory measurements. Number of rows is 753. Links to: LABEVENTS on ITEMID.

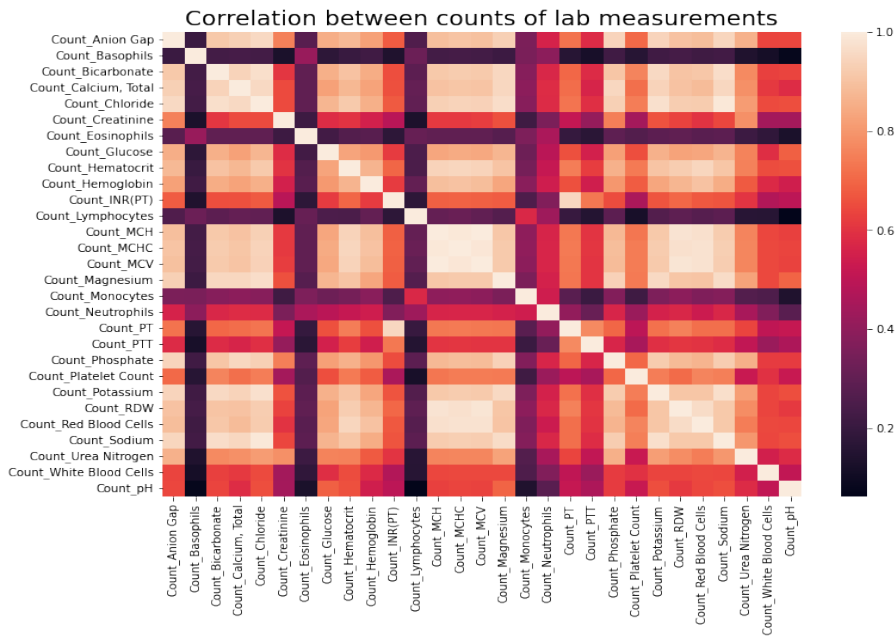
D_LABITEMS contains definitions for all ITEMID associated with lab measurements in the MIMIC database. All data in LABEVENTS link to the D_LABITEMS table. Each unique LABEL in the hospital database was assigned an ITEMID in this table, and the use of this ITEMID facilitates efficient storage and querying of the data. Note that lab items are kept separate while most definitions are contained in the D_ITEMS table, and there were good reasons to keep the lab items separate.

As the laboratory data is acquired from the hospital database, the data is consistent across all years in the database. Consequently, there is usually only one ITEMID associated with each concept in the database. Furthermore, the data contains information collected in departments outside the ICU. This includes both wards within the hospital and clinics outside the hospital. Most concepts in this table have been mapped to LOINC codes, an openly available ontology which provides a rich amount of information about the laboratory measurement including reference ranges, common units of measurement and other further detail regarding the measurement.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 753 entries, 0 to 752
Data columns (total 6 columns):
#   Column      Non-Null Count  Dtype
---  -
0   ROW_ID      753 non-null   int64
1   ITEMID      753 non-null   int64
2   LABEL       753 non-null   object
3   FLUID       753 non-null   object
4   CATEGORY    753 non-null   object
5   LOINC_CODE  585 non-null   object
dtypes: int64(2), object(4)
memory usage: 35.4+ KB
```

| | ROW_ID | ITEMID | LABEL | FLUID | CATEGORY | LOINC_CODE |
|---|--------|--------|----------------------------|---------------------------|------------|------------|
| 0 | 546 | 51346 | Blasts | Cerebrospinal Fluid (CSF) | Hematology | 26447-3 |
| 1 | 547 | 51347 | Eosinophils | Cerebrospinal Fluid (CSF) | Hematology | 26451-5 |
| 2 | 548 | 51348 | Hematocrit, CSF | Cerebrospinal Fluid (CSF) | Hematology | 30398-2 |
| 3 | 549 | 51349 | Hypersegmented Neutrophils | Cerebrospinal Fluid (CSF) | Hematology | 26506-6 |
| 4 | 550 | 51350 | Immunophenotyping | Cerebrospinal Fluid (CSF) | Hematology | NaN |

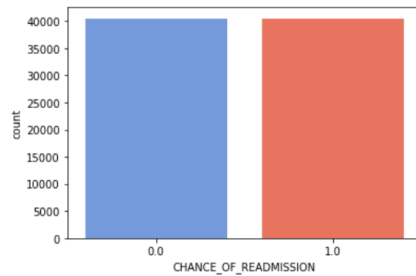
Combined all the Mean Var and Counts for each Labevent and used for the model. Below is the Correlation matrix that shows the correlation of the Counts which was showing the highest correlation among other features.



5.2.6 Pre-processing

As part of preprocessing, imputed any missing values with mean and made sure there is no missing value in the dataframe.

The data was severely imbalanced. Hence before building the model needed to correct the data imbalance. To correct the imbalance performed over sampling using SMOTE(). The distribution of Chance of Readmission after oversampling is given below.



The final dataframe has 237 columns. Feature selection has been performed and selected 200 columns by using Best K Select.

5.2.7 Socio Economic Bias Check

Using FairMLHealth library Fairness and bias analysis has been performed with the entire dataframe. Features like Gender, Ethnicity and Religion were used for the analysis and the results are given below.

Gender

| | | GENDER |
|---------------------|---------------------------------------|---------|
| Metric | Measure | |
| Group Fairness | AUC Difference | 0.0171 |
| | Balanced Accuracy Difference | 0.0305 |
| | Balanced Accuracy Ratio | 1.0322 |
| | Disparate Impact Ratio | 1.7415 |
| | Equal Odds Difference | 0.0629 |
| | Equal Odds Ratio | 1.4360 |
| | Positive Predictive Parity Difference | 0.0040 |
| | Positive Predictive Parity Ratio | 1.0040 |
| Individual Fairness | Statistical Parity Difference | 0.2519 |
| | Between-Group Gen. Entropy Error | 0.0000 |
| | Consistency Score | 0.7331 |
| Model Performance | Accuracy | 1.0000 |
| | F1-Score | 1.0000 |
| | FPR | 0.0000 |
| | Mean CHANCE_OF_READMISSION | 0.4792 |
| | Precision | 1.0000 |
| Data Metrics | TPR | 1.0000 |
| | Prevalence of Privileged Class (%) | 45.0000 |

| | | RELGN_CATHOLIC | RELGN_JEWISH | RELGN_NOT SPECIFIED | RELGN_PROTESTANT QUAKER | RELGN_UNOBTAINABLE |
|---------------------|---------------------------------------|----------------|--------------|---------------------|-------------------------|--------------------|
| Metric | Measure | | | | | |
| Group Fairness | AUC Difference | 0.0422 | 0.0531 | 0.0475 | 0.0635 | 0.1067 |
| | Balanced Accuracy Difference | 0.0629 | 0.1069 | 0.0785 | 0.1118 | 0.2177 |
| | Balanced Accuracy Ratio | 1.0688 | 1.1240 | 1.0880 | 1.1303 | 1.2903 |
| | Disparate Impact Ratio | 2.0936 | 3.0205 | 3.5806 | 3.3481 | 18.2880 |
| | Equal Odds Difference | 0.1276 | 0.2167 | 0.1582 | 0.2135 | 0.4415 |
| | Equal Odds Ratio | 1.4628 | 2.0974 | 1.3097 | 0.2761 | 0.0000 |
| | Positive Predictive Parity Difference | 0.0059 | 0.0071 | 0.0187 | 0.0714 | -0.0054 |
| | Positive Predictive Parity Ratio | 1.0060 | 1.0072 | 1.0192 | 1.0772 | 0.9946 |
| | Statistical Parity Difference | 0.2870 | 0.3340 | 0.3795 | 0.3563 | 0.4841 |
| | Between-Group Gen. Entropy Error | 0.0001 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Individual Fairness | Consistency Score | 0.7331 | 0.7331 | 0.7331 | 0.7331 | 0.7331 |
| | Accuracy | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | F1-Score | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | FPR | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Mean CHANCE_OF_READMISSION | 0.4792 | 0.4792 | 0.4792 | 0.4792 | 0.4792 |
| Model Performance | Precision | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | TPR | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | Prevalence of Privileged Class (%) | 24.0000 | 6.0000 | 12.0000 | 8.0000 | 7.0000 |

Ethnicity

| | | ETHN_BLACK | ETHN_ASIAN | ETHN_HISPANIC | ETHN_OTHER | ETHN_WHITE |
|---------------------|---------------------------------------|------------|------------|---------------|------------|------------|
| Metric | Measure | | | | | |
| Group Fairness | AUC Difference | 0.0436 | 0.1635 | 0.1827 | 0.0677 | 0.0199 |
| | Balanced Accuracy Difference | 0.0721 | 0.3518 | 0.2775 | 0.2095 | 0.0282 |
| | Balanced Accuracy Ratio | 1.0803 | 1.5717 | 1.4019 | 1.2763 | 1.0296 |
| | Disparate Impact Ratio | 1.8667 | 33.6418 | 8.3631 | 19.6164 | 1.5682 |
| | Equal Odds Difference | 0.1372 | 0.7088 | 0.5404 | 0.4227 | 0.0590 |
| | Equal Odds Ratio | 0.3808 | 4.0714 | 2.3510 | 3.0004 | 1.6087 |
| | Positive Predictive Parity Difference | 0.0248 | -0.0051 | 0.2896 | 0.0618 | 0.0021 |
| | Positive Predictive Parity Ratio | 1.0255 | 0.9949 | 1.4103 | 1.0662 | 1.0021 |
| | Statistical Parity Difference | 0.2302 | 0.4708 | 0.4287 | 0.4873 | 0.2227 |
| | Between-Group Gen. Entropy Error | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| Individual Fairness | Consistency Score | 0.7309 | 0.7309 | 0.7309 | 0.7309 | 0.7309 |
| | Accuracy | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | F1-Score | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | FPR | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| | Mean CHANCE_OF_READMISSION | 0.4792 | 0.4792 | 0.4792 | 0.4792 | 0.4792 |
| Model Performance | Precision | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | TPR | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| | Prevalence of Privileged Class (%) | 7.0000 | 1.0000 | 2.0000 | 7.0000 | 61.0000 |

5.3 Model Building and Evaluation

We build the model using train test split and 10 fold cross validation using various models like Decision Tree, Random Forest, Balanced Random Forest, AdaBoost and XGB. Results are given below.

Algorithm Used:

XGBoost

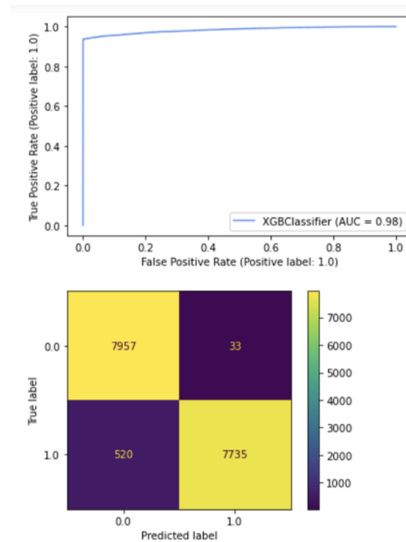
Training and Evaluation Method:

Train and Split

Results:

XGB Classifier

accuracy: 0.965958756540474
f1 score: 0.965958756540474
precision: 0.967204673104069
recall: 0.9664388556561845

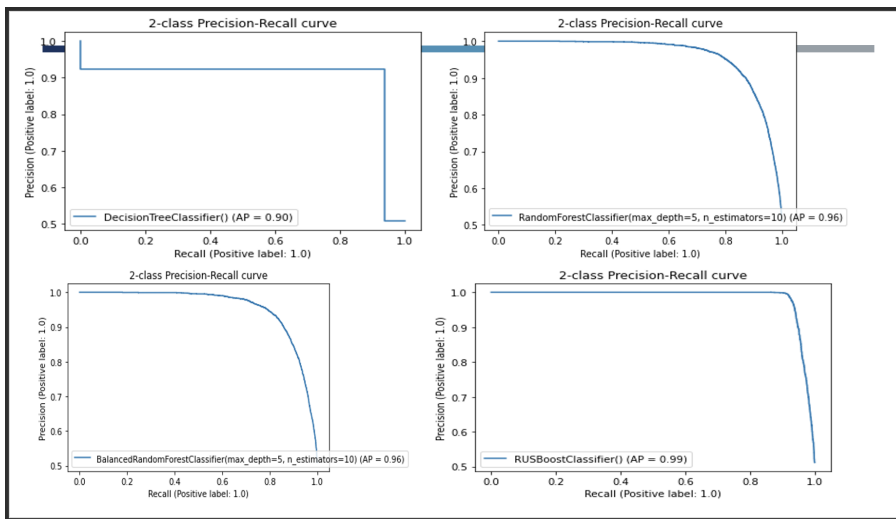


5.3.1 Experiments and Results

Cross Validation (K=10)

| Results/ Model | Accuracy | Precision | Recall | F1 | AUC |
|-------------------|----------|-----------|--------|-------|-------|
| Decision Tree | 0.814 | 0.767 | 0.935 | 0.838 | 0.813 |
| Random Forest | 0.840 | 0.828 | 0.894 | 0.847 | 0.915 |

| | | | | | |
|------------------------|-------|-------|-------|-------|-------|
| Balanced Random Forest | 0.824 | 0.821 | 0.895 | 0.851 | 0.909 |
| Ada Boost | 0.853 | 0.857 | 0.930 | 0.884 | 0.816 |
| XGB | 0.630 | 0.612 | 0.934 | 0.706 | 0.917 |



6. Future Work

We calculated binary classification for chances of readmission for all diseases in the dataset. This classification would help to assess hospital resources, approximate cost of stay for patients and their caretakers. We can extend this work to calculate resources required like medicines, hospital staff, food on a monthly basis for the hospital and calculate effort or cost of stay in advance for patients with chances of Readmission. We can also calculate the percentage of chances of readmission. This would help doctors and patients incorporate various risk factors in post hospital care. We could also build different models to calculate the chance of readmission for different diseases for more accuracy and efficiency.

It has also been observed from current data, patients transferred from the intensive care unit to the wards who are later readmitted to the intensive care unit have increased length of stay, healthcare expenditure, and mortality compared with those who are never readmitted. Improving risk stratification for patients transferred to the wards could have important benefits for critically ill hospitalized patients. A machine learning algorithm built using Patient characteristics, nursing assessments, International Classification of Diseases, Ninth Revision codes from prior admissions, medications, intensive care unit interventions, diagnostic tests, vital signs, and laboratory results and combined with SWIFT (Stability and Workload Index for Transfer score) and MEWS (Modified EarlyWarning Score) is more precise in predicting chances of readmission. Implementation of this approach could target patients who may benefit from additional time in the intensive care unit or more frequent monitoring after transfer to the hospital ward. All these above things have already been achieved in the research community. As part of future work, we could study and implement current research and build better datasets and models for industrial use which are currently needed and more suitable in the real world.

7. Conclusion

The Health Care Assistant project has analysed the possibility of predicting the length of stay for patients and the chances of readmission within 30 days of discharge and created Machine learning Models using the MIMIC III dataset for training. The Length of Stay prediction problem was formulated as a regression type whereas the Chances of Readmission is a classification problem. The results showed that it was possible to achieve a balanced accuracy of more than 90% on the testing set for the Chance of readmission and 0.0009 MSE for the Length of Stay prediction. The learnings related to ICD9 codes and classification have been used to predict the LOS. For the readmission prediction the severely imbalanced data was over samples and used for the prediction which was able to provide good results in terms of metrics used. Gradient Boosting algorithms such as GradientBoosting Regressor and XGBoosting algorithms gave the best results for the LOS and Readmission predictions respectively. The Random Forest algorithm was also performing well in both the cases and almost equally good. These models provided good accuracy, reasonable training and testing speed even with two million rows of LABEVENTS table and processing of description, which would allow for a light-weight implementation using limited resources.

8. Acknowledgements

Dr. Muhammad Aurangzeb Ahmad (Professor)

9. References

1. <https://online.shrs.pitt.edu/blog/data-analytics-in-health-care/>
2. <https://www.frontiersin.org/articles/10.3389/frai.2020.561802/full>
3. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5898738/>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/>
5. <http://www.diva-portal.org/smash/get/diva2:1338294/FULLTEXT01.pdf>
6. <https://cs.brown.edu/research/pubs/theses/ugrad/2020/baruah.prakrit.pdf>
7. https://www.researchgate.net/publication/331487868_An_Exploration_of_Data_Mining_with_Analysis_for_a_Healthcare_System
8. https://en.wikipedia.org/wiki/List_of_ICD-9_codes
9. <https://www.nejm.org/doi/full/10.1056/NEJMsa1702321>
10. <https://ieeexplore.ieee.org/abstract/document/8513181>