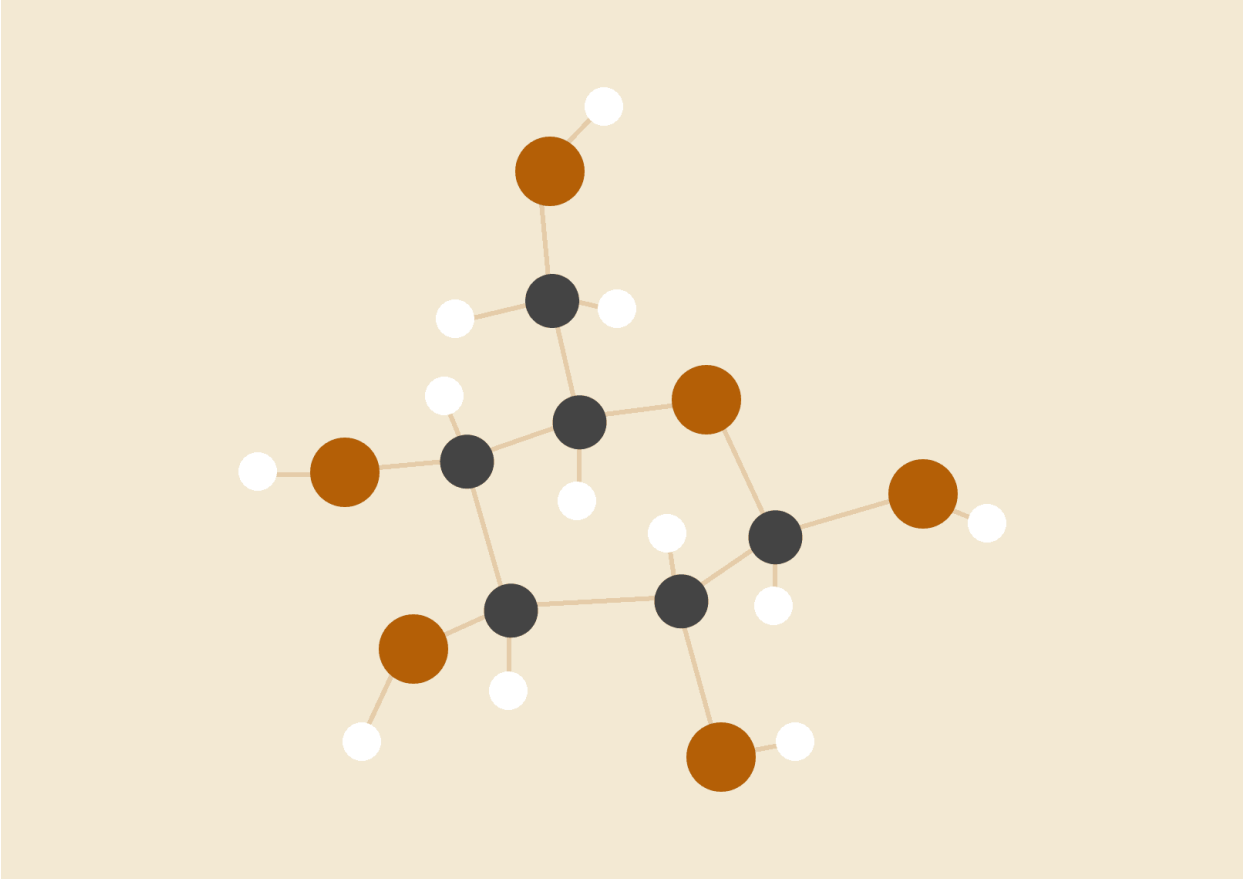
AML Midterm Report

Hospital Length of Stay Prediction



MILIND SONI - 1810110131

SANSKAR TEWATIA - 1810110215

Group - 14

11.03.2021

Applied Machine Learning EED363 (Spring 2021)

INTRODUCTION

> “The goal of this project is to create a model that predicts the length-of-stay for each patient at time of admission.”

Every time a patient is admitted to a hospital, the first question that comes to the mind is the number of days the patient will be admitted. This information is of utmost importance to the hospital, because they need to allocate resources accordingly, and plan future admittances based on the number of free beds. If there is one thing we learnt from the Pandemic, it’s the fact that there can be situations where a huge chunk of the population needs to be hospitalized simultaneously. In such cases, there is a need to predict the hospital capacity weeks in advance, so as to provide proper care to patients and refer new patients to another hospital in case there are no beds free. There can be significant variation of LOS across various facilities and across disease conditions and specialties even within the same healthcare system. Advanced LOS prediction at the time of admission can greatly enhance the quality of care as well as operational workload efficiency and help with accurate planning for discharges, resulting in lowering of various other quality measures such as readmissions.

In this project we will demonstrate how to build a model predicting the length of stay using the following steps

- Data Exploration
- Data Encoding
- Feature Generation
- Building training/ Validation/ test samples
- Model selection
- Model Evaluation

HYPOTHESIS

Using various machine learning models, we will try to make predictions on the number of days a patient is going to be admitted to a hospital. These models will take into account various types of information about the patient and their medical history like Unique Patient ID, date of admission, date of discharge, gender, facility ID, past history of medical conditions like asthma, depression, malnutrition, pneumonia, etc.

DATA

We have chosen the MIMIC III dataset released by MIT. It has a robust amount of information and is described as

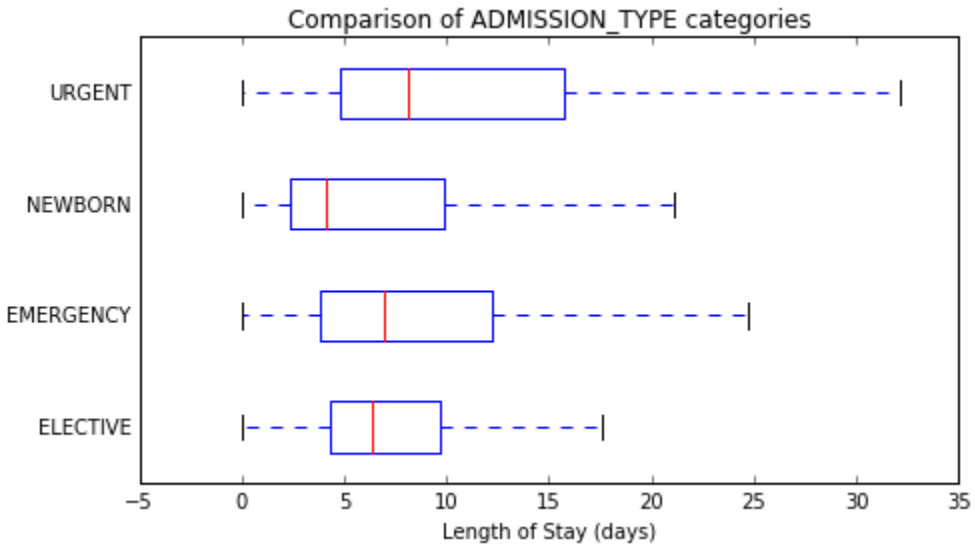
> " An openly available dataset developed by the MIT Lab for Computational Physiology, comprising deidentified health data associated with ~60,000 intensive care unit admissions. It includes demographics, vital signs, laboratory tests, medications, and more."

Link to the dataset:
https://github.com/sdasara95/Machine-Learning-in-Healthcare/tree/master/data

DATA EXPLORATION

We will find out the number of unique observations in the given dataset

LOS	24020
blood	7
circulatory	17
congenital	10
digestive	12
endocrine	12
genitourinary	8
infectious	8
injury	23
mental	12
misc	9
muscular	8
neoplasms	11
nervous	9
pregnancy	14
prenatal	17
respiratory	10
skin	9
GENDER	2
ICU	2
NICU	2
ADM_ELECTIVE	2
ADM_EMERGENCY	2
ADM_NEWBORN	2
ADM_URGENT	2
INS_Government	2
INS_Medicaid	2
INS_Medicare	2
INS_Private	2
INS_Self Pay	2
REL_NOT SPECIFIED	2
REL_RELIGIOUS	2
REL_UNOBTAINABLE	2
ETH_ASIAN	2
ETH_BLACK/AFRICAN AMERICAN	2
ETH_HISPANIC/LATINO	2
ETH_OTHER/UNKNOWN	2
ETH_WHITE	2
AGE_middle_adult	2
AGE_newborn	2
AGE_senior	2
AGE_young_adult	2
MAR_DIVORCED	2
MAR_LIFE PARTNER	2
MAR_MARRIED	2
MAR_SEPARATED	2
MAR_SINGLE	2
MAR_UNKNOWN (DEFAULT)	2
MAR_WIDOWED	2



According to the boxplot we find that the Newborns category has the lowest median LOS. We are able to get an insight into the median of the categories of the dataset using a boxplot.

We can use boxplot to gain statistical insights of different categories.

	LOS
count	51037.000000
mean	10.228510
std	12.461440
min	0.014583
25%	3.857639
50%	6.583333
75%	11.805556
max	294.660417

The statistics related to the Length of stay including the Average length of stay of the people.

After checking for the null values in the dataset we found that there are 0 null values

```
df.isnull().values.sum()
```

```
0
```

we proceed on splitting the dataset into Testing and Training datasets into an 80-20 ratio.

```
Training set has 40829 samples.
Testing set has 10208 samples.
```

We used the scikit-learn python library functions to split the dataset into training and testing data.

DATA ENCODING

Categorical data is handled differently by each machine learning model. However, there is a need to convert all data into numeric format so as to be able to train models on it. There are different methods that can be used to encode such categorical data, however we have encoded all of our data using Label Encoding and OneHot Encoding.

In Label-Encoding, each label is assigned a distinctive integer value based on alphabetical ordering. Hence, for this model 28 features were label encoded using the scikit-learn Label encoder.

One-Hot Encodingcreates additional features based on various categories of values in that specific feature. Every unique value in the category will be added as a new feature column. For example in our dataset we have created a new column called the preconditions column which will be used in the training to improve the score.

Normalisation

Normalization refers to rescaling real-valued numeric attributes into a 0 to 1 range.

Data normalization is used in machine learning to make model training less sensitive to the scale of features. This allows our model to converge to better weights and, in turn, leads to a more accurate model.

FEATURE GENERATION

For the purpose of improving the model accuracy and to make it more robust, we have decided to add a new feature named number of past medical conditions. This is made by creating a new column called preconditions and consists of the count of the preconditions of a given patient.

BASELINE MODEL

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

Using SciKitLearn's library of linear regression, first we trained the model without any kind of data preprocessing or feature generation in order to get the baseline score.

```
LinearRegression done.
```

EVALUATION METRIC

Since this is a regression problem, the metric for evaluation of test set error or validation set error was chosen to be Mean Squared Error, R2 Score .

```
Mean Squared Error
LinearRegression : 101.16879631688285
```

```
R2 Score
LinearRegression : 0.35051452128519556
```

REMAINING WORK

After completion of data preprocessing, we will train various types of machine learning models like KNN, SVM, Decision trees, Gradient Boosting methods, Various Regression models like Polynomial regression, ridge, lasso regression etc. All models will be trained and validated using cross validation which involves dividing the dataset into smaller chunks and performing the training and validation on different pieces of these chunks each separate time. The values of hyperparameters for these models will be selected using extensive testing and using GridsearchCV.