Predicting 30-day ICU readmissions from the MIMIC-III database

I. Definition

Project Overview

This project describes the development of a model for predicting whether a patient discharged from an intensive care unit (ICU) is likely to be readmitted within 30 days. To do this, we will use the MIMIC-III database which contains details records from ~60,000 ICU admissions for ~40,000 patients over a period of 10 years. Using these records, we will find patients in the database that were readmitted within 30 days of discharge, and use their records to train a generalized classifier to predict readmission. The project also includes parameter optimization and in depth performance analysis for the classifier

Problem Statement

Problem Statement - Heart failure is a very common ailment leading to fatalities if not
attended to promptly. Even for the patients who get proper treatment, hospital re admissions result in a significant risk of death and a financial burden for patients, their
families as well as the already overburdened healthcare systems. Prediction of at-risk patients for readmission allows for targeted interventions that reduce morbidity and mortality.

1. Develop a machine learning model with the end objective to predict readmission of heart-failure patients within 30-days of discharge from the hospital.
2. We have provided a subset of tables at the below one-Drive link for this problem. Veersa Hackathon Submission 2024
3. Participants will further decide which tables they will use to solve the problem.
4. Link to mimic-III table descriptions. - https://mimic.mit.edu/docs/iii/tables/
5. Before actual prediction, detailed data analysis is expected to support the model.
6. Below are the list of Diagnosis codes (Icd9_codes) representing heart-failure:
('39891','40201','40211','40291','40401','40403','40411','40413','40491','40493','4280','4281','42820','42821','42822','42823','42830','42831','42832','42833','42840','42841','428 42','42843','4289')

The solution outline will be as follows:

1. **Preliminary research:** Consult the relevant literature to assemble a list of relevant features to use.
2. **Data wrangling:** Use SQL to obtain the relevant subset of data from the MIMIC-III database.
3. **Exploratory data analysis:** Produce relevant statistics and visualizations to characterize the data

4. Data preprocessing: Prepare the dataset for application of a machine learning model – clean missing or invalid values, calculate the time between subsequent
admissions to produce label dataframe.
5. ML model implementation and refinement: Define model metrics, implement relevant code and perform hyperparameter optimization. Once optimized values are found, estimate model performance with cross-validation. After extracting our features of interest (see table 1) for all admission records or interest
(i.e. admissions to the MICU1), and cleaning up the extracted dataset, we will create the
label for each admission which will indicate whether that discharge resulted in readmission within 30 days. We will then feed the clean dataframe and calculated labels into a gradient-boosted decision tree classifier, perform hyperparameter optimization and
analyze the model's predictive capability.

Metrics
Our problem can be described as a binary classification problem on an unbalanced dataset (only ~12% of the observations fall into the "positive" label category). For this reason, we cannot rely on a simplistic metric such as prediction accuracy, and must look at more robust metrics.
A standard approach in such cases is to look at the precision, sensitivity, specificity, F1 score and the Receiver Operating Characteristic curve, all defined in the following:
• **Precision** is defined as:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

• **Recall**, also known as **Sensitivity** or **true positive rate (TPR)**, is defined as:

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

• **Specificity** is defined as:

$$Specificity = \frac{True\ Negative}{True\ Negative + False\ Positive}$$

• **F1 score**, the harmonic mean of precision and recall:

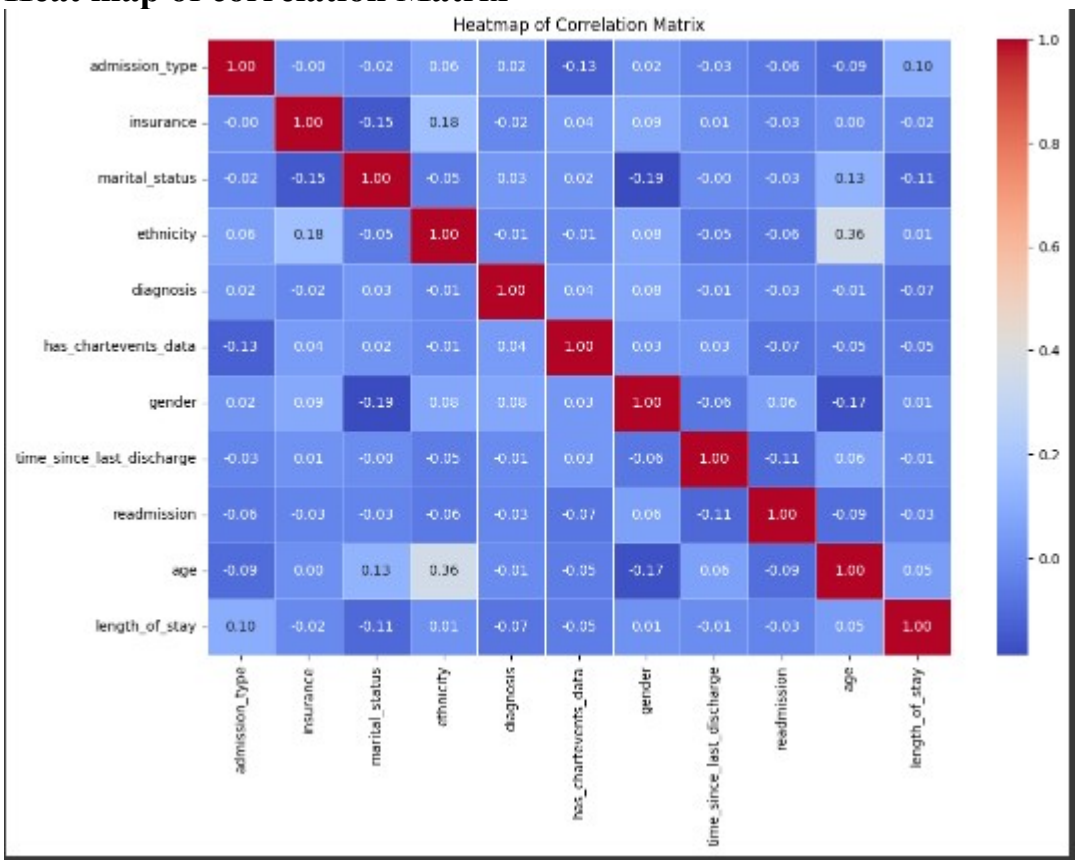
$$F1 = \frac{2 * (True\ Positive)}{(True\ Positive + False\ Negative) + (True\ Positive + False\ Positive)}$$

<u>Pseudo-code for ML:</u>
X contains all the independent columns.

Y contains only target column.
First split the data into train and test datasets
X_train,X_test,Y_train,Y_test = split(X,Y)
Train the model using the training data
model.fit(X_train,Y_train)
Predict on the testing data
predictions = model.predict(X_test)
Calculate the metric i.e. MAE for regression/Accuracy for classification
metric(Y_test,predictions)

Data Visualisations

Heat map of correlation Matrix



Bibliography

1. MIMIC-III, a freely accessible critical care database. Johnson AEW, Pollard TJ, Shen L,
Lehman L, Feng M, Ghassemi M, Moody B, Szolovits P, Celi LA, and Mark RG.
Scientific Data (2016). DOI: 10.1038/sdata.2016.35. Available from: http://www.nature.com/articles/sdata201635
2. RB Alfonso, Feature Extraction and Selection for Prediction of ICU Patient's
Readmission Using Artificial Neural Networks, Master's thesis, 2013.

3. Nguyen et al., Predicting All-Cause Readmissions Using Electronic Health Record
Data From the Entire Hospitalization: Model Development and Comparison,
Journal of Hospital Medicine, 2016.
4. Kareliusson et al., Risk prediction of ICU readmission in a mixed surgical and
medical population. Journal of Intensive Care, 2015.