Introduction

- Project Name: Using Machine Learning Models to Predict the Length of Stay in a Hospital Setting
- Course Name: Machine Learning-I
- Student Name: Maria Baloch (23231), Nafeesa Yousuf Murad (23661)

Health Care Domain

- Proper prediction of Length Of Stay (LOS) has become increasingly important these years. The LOS prediction provides better services, managing hospital resources and controls their costs.
- Healthcare institutions, academic researchers and industry organizations in various areas are working in coordination to improve the quality of care and the management of healthcare systems.
- The Length Of Stay (LOS) is considered as one of the basic indicators to evaluate the performance of care services and care quality which explain the growing interest of predicting LOS in hospitals these past years.
- The LOS represents the interval time between the admission of the patient and his discharge. Estimating the LOS at the admission time provides an approximation of the patient's discharge date involving an appropriate planning of care activities. As a result, expecting the acute value of the LOS at the time of patients' admission is useful to highlight a planning strategy for the hospital's logistics.
- There are 17 predictors for predicting Length of stay of each patient.
- Predictive analytics (PA) is a new trending approach in the field of healthcare that uses machine learning to build a prediction model using supervised learning algorithms.
- The aim of this healthcare Predictive analytics project is to develop and evaluate a model to predict postoperative length of stay (PLoS) for cost-estimation open-heart surgery patients using supervised machine learning.

Problems:

Identifying factors that constitute LOS models was the first task to carry out in the project of LOS prediction. Awareness of factors and elements that determine LOS was a challenge in every data science project. lhc_done_at_thi, myocardial_infarction, last_hematocrit, last_wbc_count, initial_icu_hours, ejection_fraction, patient_age, weight, diastolic are used to predict LOS in a hospital. To analyze the huge amount of medical data, several methods of data mining were used. Preceding studies have used multiple supervised learning techniques to create a prediction model of LOS based on factors impacting this variable.

Pipeline:

1- Manual Deletion of identifiers, other cardiac procedures, valve procedures, mortality variables, complexity, follow-up note variables, date variables.
2- Merging variables that belong to one type.
3- LOS is calculated by subtracting date of discharge variable from date of admission.
4- Correlation and Random Forest feature selection technique was conducted for important Feature Selection
5- Algorithms used: Random Forest, Decision Tree, SVM, Gradient Boosting, AdaBoost, Voting regression, polynomial regression.
6- Choosing the best Algorithm on the basis of RMSE.

Methodology:

I have chosen RMSE(Root Mean Square Mean) to have a fair comparison between the models. The algorithm giving the lowest value of RMSE is the winner algorithm.

Best Algorithm:

Which combination is best? (based on test data performance)

Ridge and Lasso Regression with Random Forest Feature Selection

Solution:

The best results were provided by Ridge and Lasso Regression with Random Forest Feature Selection. Our model gives an RMSE of 3.7 on test data.

- Suppose your model is deployed in industry and generates predictions on live test data. What if the distribution of the original data changes, e.g., the customers' behavior pattern changes? Then your model will start to give the wrong predictions. How will you update your model in this situation?

  I will work more on outliers of my dataset, regularization, exploring more models for better results.