Homework #2 − Real-life Problem Proposal: Heart Failure Analysis and Prediction

Zhehao (Jonas) Lu

Master of Mechanical Engineering (MSME)

zhehaolu2022@u.northwestern.edu

1   Introduction – Fundamentals of My Proposal

Heart failure is a chronic, progressive condition in which the heart muscle is unable to pump enough blood to meet the body's needs for blood and oxygen. It is commonly agreed that certain factors increase chances of heart failure and thus cause a death event, such as blood pressure, smoke, diabetes and so on. Diagnosing the possibility of death event due to heart failure is essential to make predictions and select reasonable treatment plans.

In this project, I will propose a prediction model regarding heart failure events leveraging machine learning (ML) techniques. A variety of factors will be taken into consideration to evaluate the correlation between these factors and heart failure. A widely-accepted dataset will be applied and further processed using the "six basic data science concepts" discussed in the course. Different kinds of linear and non-linear regression models will be applied and compared.

2   Implementation – Six Steps to Achieve my Proposal

2.1   Multimodal Data Generation & Collection

The dataset is collected from Kaggle website (Please refer to: https://www.kaggle.com/gopalj/heart-fail-analysis-with-ensemble-modeling/data). The dataset contains 13 columns. The upshot of the collected data is a Boolean type parameter use either 1 or 0 to show whether or not a death event happens.

2.2   Extraction of Mechanistic Features

Some initial pre-processing methods are applied to the raw dataset that is collected. Invalid data, such as those containing missing columns should be removed. Data should be re-arranged and separated according to the Boolean value of death event.

2.3   Knowledge-driven Dimensional Reduction

In this step, the number of variables should be reduced under consideration. The correlation between the possibility of death event with respect to any single variable is obtained and observed. If they do not yield a good R-squared value, then the variable can be ignored, otherwise it should attract more attention.

2.4   Reduced Order Surrogate Models

Reduced order surrogate model is applied to reduce the complexity of the problem based on physical principles, but there seems to be few existing mathematical principles regarding this heart failure prediction model, so that this is a purely data-driven problem, similar to the example of diamond price prediction, which may not include advanced order reduction methods due to the absence of existing principles.

2.5   Neural Networks for Regression & Classification

A variety of linear and non-linear regression models will be applied and compared with regard to the accuracy and efficiency of the prediction, which may include, but are not limited to, SVM, Logistic Regression and K-Neighbors Classifier methods.

2.6   System & Design

With all of the previous work completed and carefully checked, we can draw a conclusion on which regression model demonstrates its accuracy with regard to this case. After the validation of prediction accuracy, the model can make predictions on whether a death event may occur when new patients come with all the essential information given.