Healthcare Case Studies
Project Description

Instructions

In this project, you will create a linear regression model for predicting total post-index cost and a classifier for predicting the pdc flag for diabetic patients. For the classifier, you must examine the performance of 3 types of models, namely Logistic Regression, Naive Bayes, and KNN.

The first step in creating the models is variable selection. As you select your variables, keep in mind that smaller models are easier to interpret and less prone to overfitting.

Utilize scatter and box-plots, statistics tests, and Stepwise, Ridge, and Lasso among other techniques for variable selection. You should also compare the performance of different models on the validation sets and assess the final performance by measuring the out-of-sample error.

For the linear regression model, you will need to perform residual analysis to check model assumptions.

Demo

Your group will do a demo (in powerpoint, ioslides, etc.) on Saturday where you'll share your methodology and initial results with rest of your peers, and answer instructor's questions. Please see schedule on Canvas for your presentation time.

Deliverables

You will turn in the following on Canvas:

1. Completed set of slides with a summary of methodology and results (15-20 sides). You need to report the performance of your models and justification of what classifier you have chosen as your final model.

2. A function in R or Python where I can test your analysis on my test file. The values for RSS and confusion matrix must be clearly displayed.

3. Your source code (markdown/Python) and the html file.