P8160 - Breast Cancer Data: To lasso or to not lasso

Group ??

Amy Pitts, Hun Lee, Jimmy Kelliher,
Tucker Morgan, Waveley Qiu

2022-04-01

Abstract

we will write an abstract eventually.

1. Introduction

1.1. Overview

1.2. Objectives

2. Methods

2.1. Data Cleaning and Exploratory Analysis

The data set of interest contains 569 rows and 33 columns all related to breast tissue images. Each entry of the table represent an individual patient. Of primary interest is the column containing information about patient diagnosis of cancer taking on values either malignant or benign. One column contains information about patient ID which will be removed from our dataset. The other 30 columns correspond to numerical information about the breast imaging.

In a quick exploration of the data it has been found that there is a high correlation between many of the variables. This is seen in **Figure #**. The heat map correlation plot there are a lot of varaible with dark blue coloring representing a high correlation. High correlations can cause major issues with the methods we are going to use so understanding why there is high correlation in our data is important.

To explore the correlation more **Figure #** shows the highest 25 correlation in our data. We see that the highest correlation is between radius mean and perimeter mean with a correlation values of 0.998. In the graph there are 21 combinations of variables that achieve a correlation over 0.90. This is cause for major concern.

The variables {`area_mean`, `area_worst`, `perimeter_mean`, `perimeter_worst`, `radius_mean`, `radius_worst`} are all mutually correlated. This grouping represents the top six, eigth correlation in **Figure #**. Mathematically, if we consider the equivalence classes of variables that are highly correlated, these six variables would belong to the same equivalence class. To identify the best proxy for this grouping we look at the highest mean correlation which turns out to be `radius_worst`. The next grouping of correlated variables is {`radius_sd`, `perimeter_se`, `area_se`}. The best represenitive will be `radius_se`. Next grouping is {`concavity_mean`,`concavity_worst`, `concave.point_worst`, `concave.point_mean`}, with the best variable being `concave.point_worst`. Finally, {`texture_mean`, `texture_worst`} with `texture_mean` being the variable saved. Each of these equivalence classes represents all the correlations above 0.90. In total there will 10 variables removed.

Figure 1: Correlation Heat Plot of all Covariates

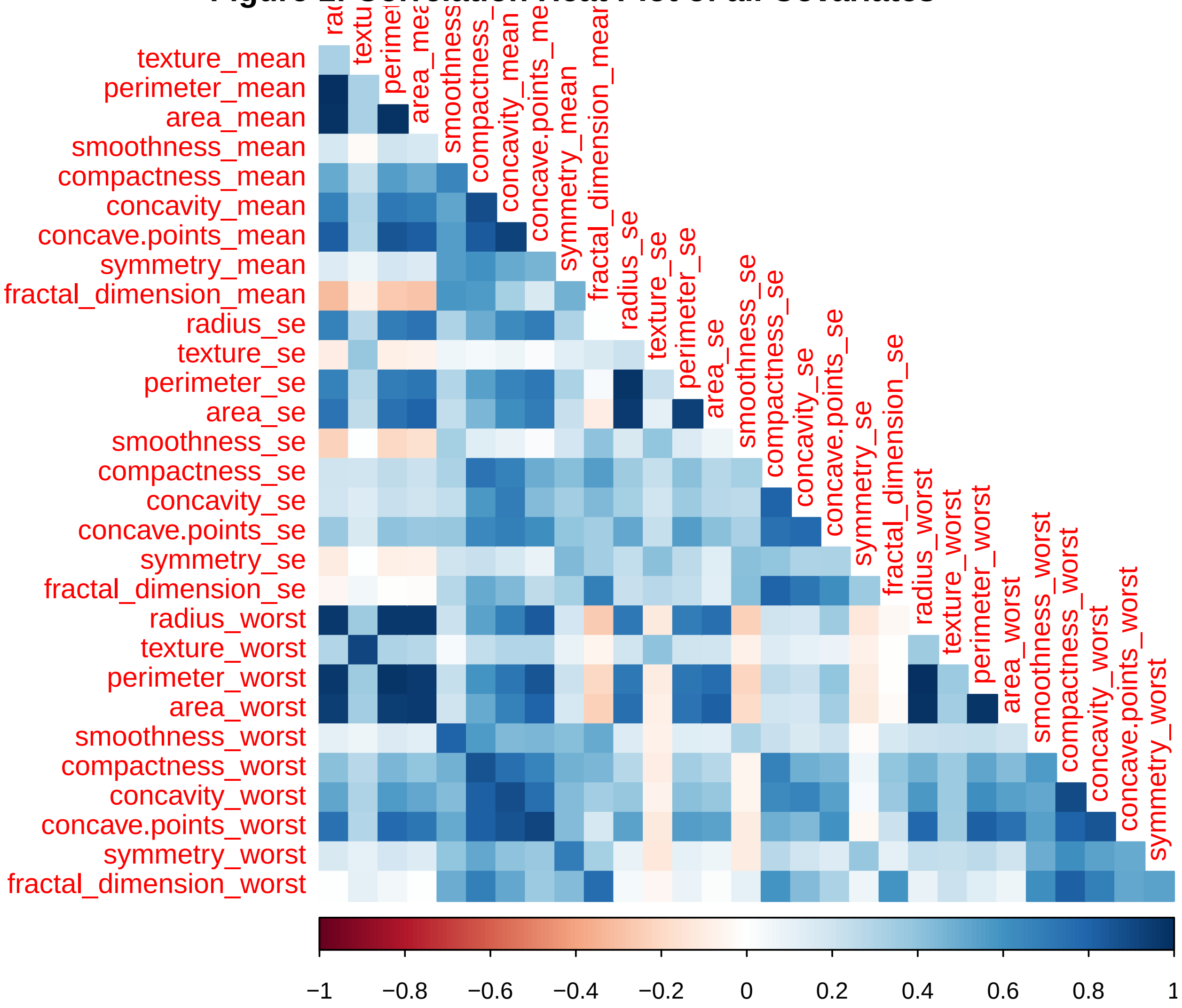
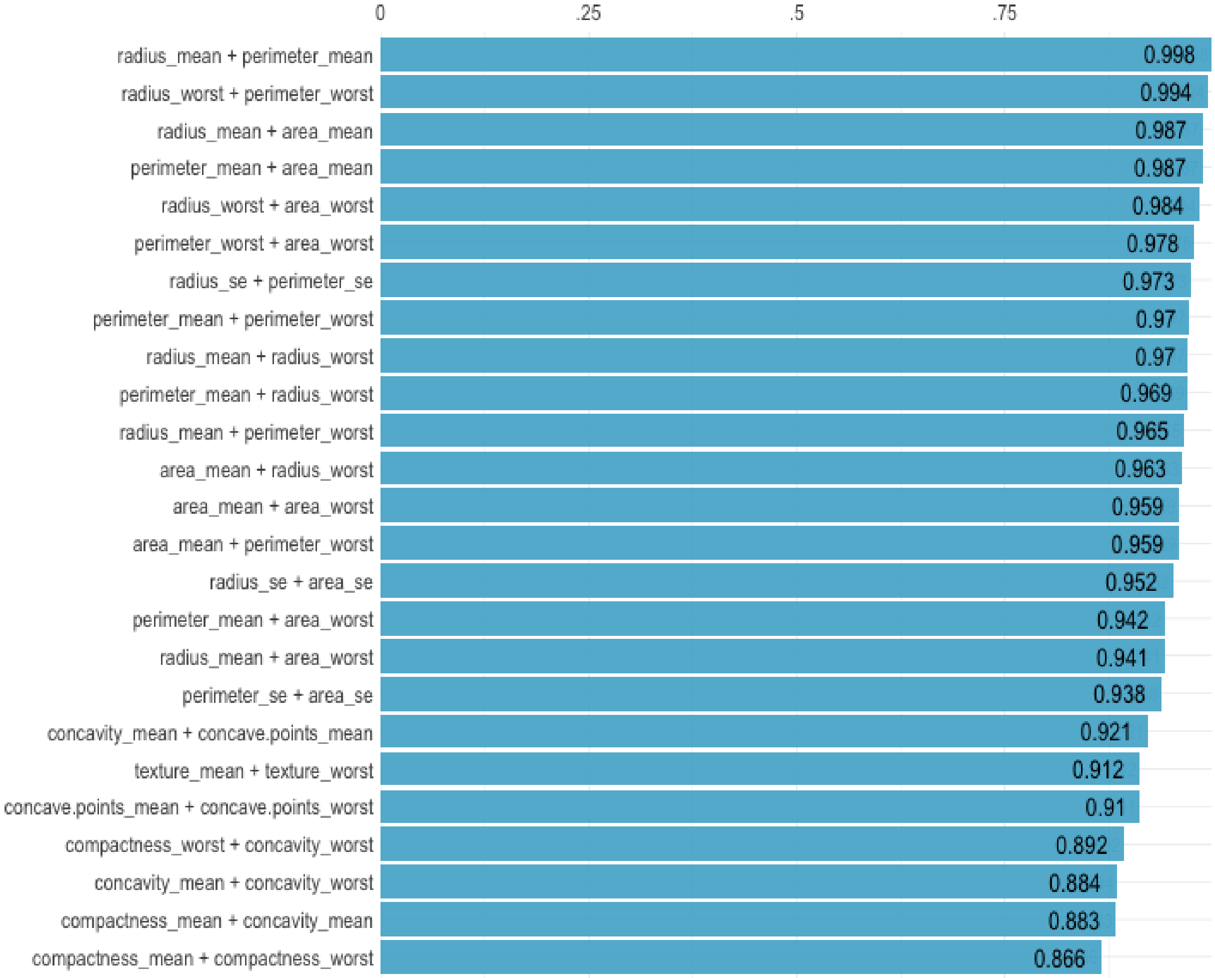


Figure 2: Ranked Cross-Correlations

25 most relevant


radius_mean + perimeter_mean
radius_worst + perimeter_worst
radius_mean + area_mean
perimeter_mean + area_mean
radius_worst + area_worst
perimeter_worst + area_worst
radius_se + perimeter_se
perimeter_mean + perimeter_worst
radius_mean + radius_worst
perimeter_mean + radius_worst
radius_mean + perimeter_worst
area_mean + radius_worst
area_mean + area_worst
area_mean + perimeter_worst
radius_se + area_se
perimeter_mean + area_worst
radius_mean + area_worst
perimeter_se + area_se
concavity_mean + concave.points_mean
texture_mean + texture_worst
concave.points_mean + concave.points_worst
compactness_worst + concavity_worst
concavity_mean + concavity_worst
compactness_mean + concavity_mean
compactness_mean + compactness_worst