Question 2: Part 2

One of the hardest things about making the decision to screen a patient or not is understanding which patient factors may put them at a higher risk for breast cancer. In order to find what factors could potentially be more correlated with cancer, and those that our sample of doctors used in their recall decisions, we partitioned the data into training and testing splits and ran a random forest regression focused on recursive feature elimination in order to find the best subset of features for both our recall variable and recall observations, and on our cancer variable, in order to find what variables might more accurately predict that a patient is at higher risk for cancer.

The reason for running this recursive feature elimination on the two separate outcomes is to see if there is a chance that doctors are focusing on the wrong, or lower correlation/accurate, variables when making their recall decisions. Additionally, due to the small sample size, and very low probability of both recall and cancer occurring, the recursive feature elimination randomly samples with replacement from this original data set in order to artificially increase our sample size so that we can attempt to find a more accurate prediction model. During this sampling process, it randomly selects different variables to use and holds on to the highest performing models before outputting the best predictive variables to be used in order to train a model to accurately predict the outcome variable based on cross validation. It is our hope that by doing this we can help doctors be more confident in their recall decision, and so that patients know whether they fall under a higher risk category or not.

We first began by removing all radiologists from the dataset. By doing so we are hoping to remove any impact that persona bias or conservatism regarding recall decisions based on each individual doctor from affecting the regression. Next, using recall as the out outcome variable, we found that doctors tend to base their recall decision most commonly on the following variables: cancer, premenopause, postmenoNoHT, and age groups for those between 50-59 and 70 plus. We then included these variables in another, more finely tuned, random forest, neural net, and generalized linear model, in order to test these variables predictive power under different models and situations.

The neural net and random forest models came back with the most accurate and significant of the three regressions. Both had an accuracy rating of roughly 86.02% and a Kappa rating of 16.97%, meaning that they are relatively useful and accurate for out of sample prediction. The GLM model returned a statistically significant accuracy rating of 85.21% and a 12.89% kappa rating. Each model came back with a high true positive rate and low false positive rate, and therefore we can assume that all of these variables are indeed good predictors of what doctors tend to look for when making their recall decision based on cross validation.

Next we followed the same procedure and used cancer as the outcome variable. Using the recursive feature elimination, our model narrowed the important variables down to recall, density4, age 60-69, pre-menopause, and density3. This time the neural net returned a statistically significant accuracy rate of 96%, however it also had a kappa value of 0% and thus has no true out of sample predictive power. This is likely due to the very small number of cancer outcomes in the data set, and thus it did not obtain enough data to give back any truly useful information. The GLM model however returned a 96.15% accuracy rate that is statistically significant with a kappa rating of 12.59%, as well as a low false positive and a high positive prediction value. Therefore we can assume that the neural net model was likely heavily over fitted and therefore a simpler model is likely to give better results.

Last, we used the recall predictors in a GLM model with cancer as the outcome variable in order to see if it could potentially give more accurate results than those from the original regression. This model did have a higher accuracy rating, however it has a kappa rating of 0% and thus is not useful for out of sample prediction. Although we are confident that doctors are not completely inaccurate in their decision for recalling, the reason such a low out of sample prediction rate occurred is likely due to the very small number of cancer outcomes in our sample size.

Using the data that we have been given, and advanced statistical regression techniques, we believe that the models and variables that we have provided may be of some use for doctors who are concerned with whether they are paying attention to the right factors regarding their patients potential for testing positive for cancer during a mammogram screening, and whether they should make a recall decision. We still believe that it is worthwhile to take the variables that we have found through our regression into consideration, and hope that doing so will lead to less unnecessary recalls in the future, and more accurate cancer detection rates.