Ds 740 midterm

Paul Isaacson

October 30, 2017

# PART 1

# Discuss why you selected the response variable (factor versus quantitative).

In this study, I looked at the factor reponse of CAMLEVEL for the data set calcium. This data set came from the study by Boyd, Delost, and Holcomb (1998) entitled 'Calcium, inorganic phosphorus and alkaline phosphatase levels in elderly patients. The original intent of the study was to determine if gender and lab was a significant deferentiator in levels of calcium, inorganic phosphorus and alkaline phosphorus. I, on the other hand, took a slightly different approach. I believe it to be important to predict which of the factors will lead to classification in either the high calcium (hypercalcemia) or low calcium (hypocalcemia). For this goal, a categorical response variable is needed.

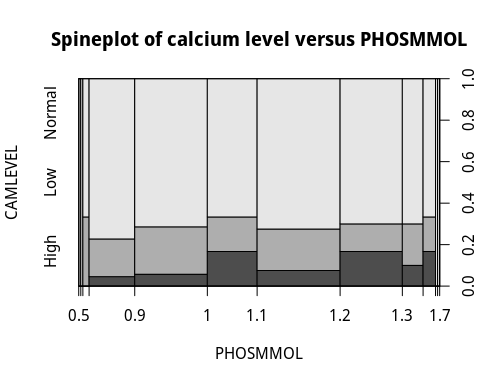
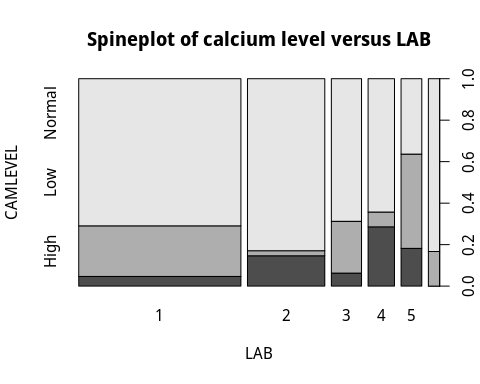
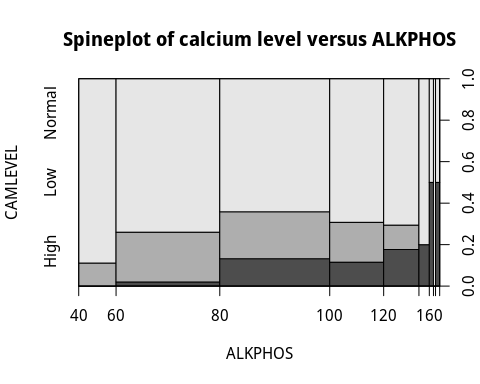
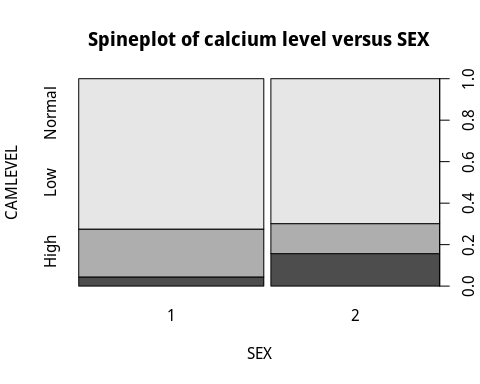
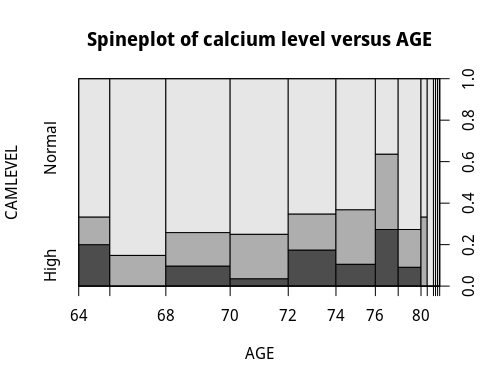
# Describe a possible practical purpose for the data analysis and an audience who might be interested in this purpose.

With such a classification model, one could make a determination if an elderly (over 65) individual was at risk for either conditions. Further tests, such as routine blood tests, could be targeted for individuals who are classified as hypocalcemic. However, this is not an overwhelming benefit as such routine blood tests are performed regularly on this population. Perhaps more importantly, a classification of hypercalcemia could be the a sign of more serious disesease, such as lung and breast cancer (cite 1) With early detection, symptoms such as kidney problems or arryhthmia may be avoided.

# PART 2

# Identify two possible methods (from two different lessons) for predicting the response and provide justification for why these methods are appropriate for the data.

For classification problems, knn, logistic regression, lda, decision trees, and ensemble methds such as bagging, random forests, and boosting are approprate. However, because the response variable was not binomial, I discarded logisitic regression as a choice. The two models selected for analysis was knn and random forest. I selected knn because it is simple and non-parametric. I was also interested in exploring optimal K values. I selected random forest from the similar decision tree methods because I suspected only a small subset of the variables would be more informative than the others.



# Use cross-validation techniques to select between the two methods (and among any parameters needed for those methods).

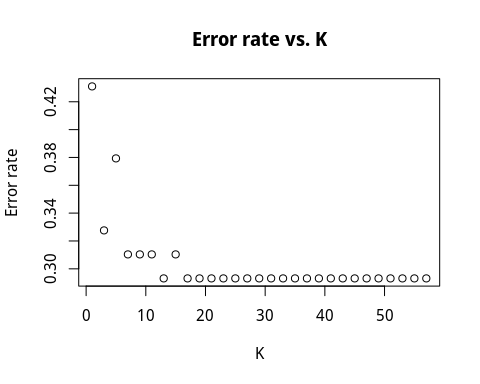
done

# Add an outer level of cross-validation to further assess the predictive ability of the model selected.

not done

# Fit the final selected model to the data and summarize and discuss the outcome of the fit. Include estimates of any model parameters.

A critical step in fitting a good model is to find the appropriate parameters. Because I only had 5 parameters to work with, I spend much time performing cross validations on all the permutations of variables. For knn, the best performing model had 2 predictors (AGE, LAB) for knn. This produced a . Additionally, the k was set to 7 based on the following plot



The random forest, the best performing model had 3 predictors (PHOSMMOL, LAB, ALKPHOS). This produced a .

In some regards, both models did little better than classifying everyone as 'Normal'. This approach would produce an error of 0.2873563. However, overall error is perhaps not the best measure for this study. In may be more acceptable to have a higher overall error rate if the model was better at identifying 'Low' and 'High' classifications. Further studies should explore this distinction.

cite 1https://www.mayoclinic.org/diseases-conditions/hypercalcemia/symptoms-causes/syc-20355523