DS 740 Midterm

Paul Isaacson

October 30, 2017

In this study, I looked at the factor response 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 differentiator 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.

With such a classification model, one could determine 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 hypocalcemia. 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 a sign of more serious diseases, such as lung and breast cancer (Mayo Clinic Staff (n.d.)). With early detection, symptoms such as kidney problems or arrhythmia may be avoided.

For classification problems, knn, logistic regression, lda, decision trees, and ensemble methods such as bagging, random forests, and boosting are appropriate. However, because the response variable was not binomial, I discarded logistic 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.

A critical step in fitting a good model is to find the appropriate parameters. Because I only had 5 parameters to work with (Fig. 2), 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 plot of error rate versus K values (Fig. 1).

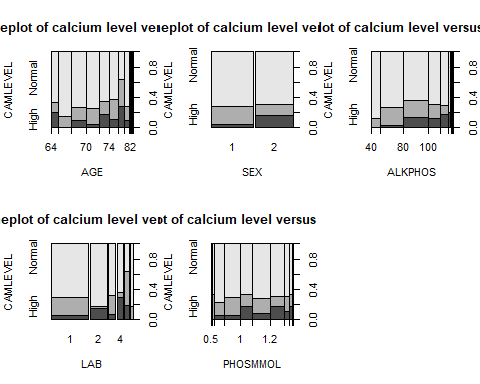
The random forest, the best performing model had 3 predictors (PHOSMMOL, LAB, ALKPHOS). This produced a .

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 and rate classification error based on this.

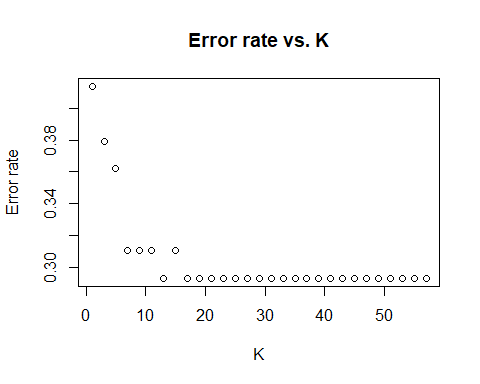
After the optimal two models were found using cross validation, I used double cross validation on the knn model. For the inner cross validation, I tested error rate versus various K values. The optimal model produced .

The models would benefit from additional observations and predictors. Because of the CV values near chance, I would not feel comfortable using this to make decisions until the data was enriched.

# Figure 1



# Figure 2



# References

Mayo Clinic Staff. n.d. “Hypercalcemia.” <https://www.mayoclinic.org/diseases-conditions/hypercalcemia/symptoms-causes/syc-20355523>.