Executive_Summary_Modeling

Eleanor Colligan

3/7/2022

Background

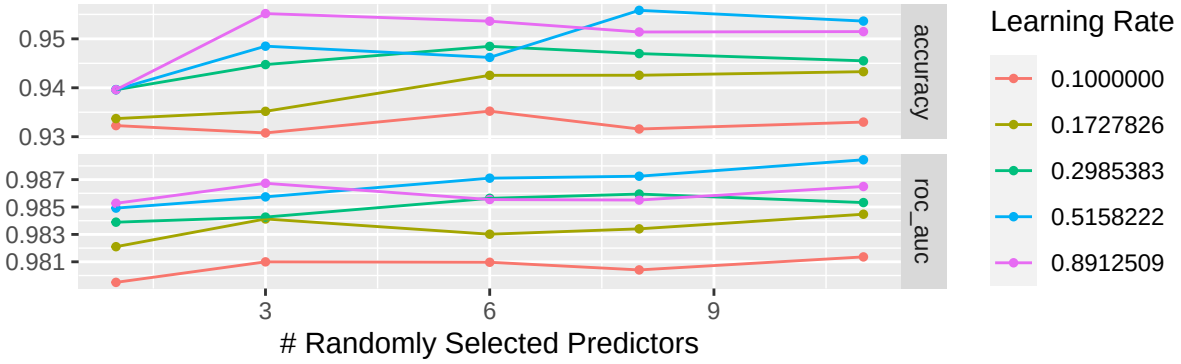
As a woman, I find woman's health very interesting. Furthermore, due to the fact we live in a historically (and currently) man-focused world, there is a huge gap in women's health research and medicine compared to men's health. One example of this is birth control, and how reliable forms of birth control rest on women to take and deal with troubling (and) at times very serious) side effects. Furthermore, cancer is an illness that virtually touches everyone; it seems that everyone I know has a family member or friend that has been diagnosed with cancer. Breast cancer, and this data set as a result, which is at the intersection of woman's health and cancer, combines both pressing issues. The data set I chose is from the UCI Machine Learning Respository, and can be found at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

The research question I sought to answer was: How accurately can I predict the diagnosis of a tumor based off of various physical measurements of the tumor?

Modeling

During class, I feel like we did a lot more practice with regression than with classification, or perhaps classification is just less intuitive to me given I've learned about regression in other statistics classes. Either way, I wanted to do my project on classification so I could understand it better. I tried four algorithms: K-Nearest-Neighbor, Logistic Regression, Random Forest, and Boosted Tree.

My most successful model was a boosted model, so I decided to try to further tune the parameters manually min_n, mtry, and learn_rate to see if I could get a better accuracy and ROC AUC score.



I used min_n as 2 since in my other models I saw that min_n= 2 had much higher scores for both accuracy and ROC AUC. I then set the learning rate to a range of (-1, -.05) and my mtry to a range of (1,11) since I had 11 predictors. From looking at both my metrics graph, which is shown above, and the *select_best()* function from the *tune* package, I decided to use an mtry of 11, a learning rate of .5158222, and a min_n of 2. I ended up getting an **accuracy of 96.5% and an ROC AUC of 98.77%**. So, to answer my research

question: yes, I can accurately predict the diagnosis of a tumor based off of various physical measurements of the tumor.