447 Project

Bryce Anderson, Zoe Cruz

June 2024

Abstract

Cirrhosis is a chronic liver disease that significantly impacts liver function and can lead to liver failure. This study leverages data from the Mayo Clinic trial on primary biliary cirrhosis (PBC) conducted between 1974 and 1984. The goal is to develop a classification model to accurately predict the stage of cirrhosis based on 16 predictor variables,both categorical and quantitative. Using Random Forest and Gradient Boosting methods, we assessed the classification performance and identified the most significant predictors. The Random Forest model achieved an overall accuracy of 61.45 percent, while the Gradient Boosting model suffered from overfitting, achieving a best accuracy of 50.06 percent. The analysis highlighted the importance of variables such as albumin and hepatomegaly, while sex was found to be less influential. We underscores the complexity of predicting cirrhosis stages and the need for further investigation with larger datasets to enhance model performance and reliability.

1 Introduction

Cirrhosis is a chronic liver disease that occurs when healthy liver cells are replaced by scar tissue, also known as fibrosis. This scarring prevents the liver from functioning properly and can eventually lead to liver failure. We are going to be analyzing a dataset from information collected from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984 of patients with cirrhosis from stages 1-4. Our goal is to create a classification model that can accurately predict the stage of cirrhosis someone has based on a number of predictor variables and identify which variables are most influential to the models. It is an important classification problem, because we could identify which predictors are clear indicators of cirrhosis, and ultimately narrow down the amount of testing needed to classify someone with cirrhosis. We have 16 predictor variables. There are 6 categorical variables and 10 quantitative variables. 5 of the categorical variables are related to other conditions directly related to cirrhosis for example, hepatomegaly (enlarged liver). These are all binary variables coded as Y/N. The other 10 quantitative variables have to do with various metrics that are important to liver health. We will be using Random Forest and Gradient Boosting methods to create classification tree models to analyze the data.

2 Methods

2.1 Decision Tree

We are going to be using tree based methods to capture both the categorical and numeric variables in our data. Firstly, we will build a simple decision tree to visualize the data using the rpart library in R. We have adjusted the complexity parameter to 0.02 to prune back the tree and make it more interpretable as seen in Figure 1. From the first look of things, it looks like this is going to be a tough classification problem as there is a lot of variability in the data.

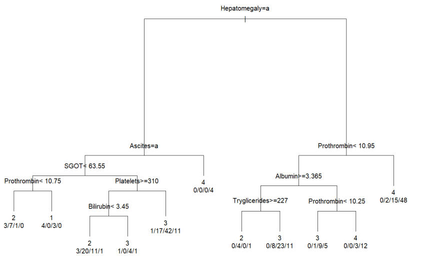
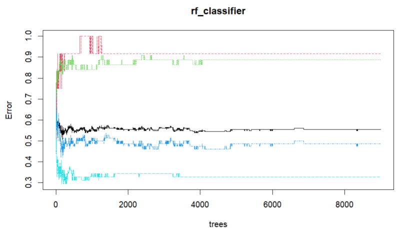
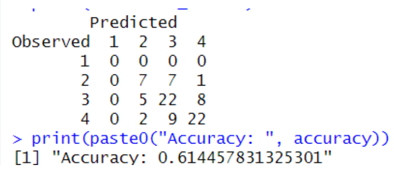


Figure 1: Decision tree with cp=0.02

2.2 Random Forest

Next, we will use build a Random Forest model using 70 percent of the data for training. We tried values between 50 and 80 percent for the training data, however it had little effect on the results. We kept the mtry parameter at 4 as it was approximately equal to the square root of all the variables in our data by recommendation of several other papers. The OOB estimate of error rate was 55.44 percent. Not a very convincing to start. In fact, the RF model had a 0.9166667 error rate when classifying a stage 1 observation. This does make sense as these were the smallest sample sizes in the data set with only 21 observations having Stage 1. To try and make these results slightly less horrendous, we will tried to tune the model using the tuneRF function in the RF library to find the optimal mtry value. We adjusted the step factor as needed, however it was only able to reduce the OOB error to 50.36 percent with an mtry = 1, which was not desirable. We can calculate the overall accuracy of this model by looking at the confusion matrix from the validation set and comparing the actual values to the prediction ones. If we just take the sum diagonals and divide it by the sum of the entire matrix we get an overall accuracy of 61.45 percent. To analyze the variable importance, we can take a look at the mean decrease in the Gini-index and the mean decrease in accuracy. We can interpret the variables on these graphs as ones that have higher values are of more important, as when we take them out of the model, the accuracy/Gini-index decreases more. From Figure 3, we see that it would be difficult to decide which variables to perhaps remove from the model. The only variables that I would consider removing would be Edema and Drug as they are both very low in the graphs on both. It is important to note that the predictor Drug is a variable in which someone either received a drug or received a placebo. This is not important to our research question and thus should have been removed.





```
          Predicted
Observed  1  2  3  4
       1  0  0  0  0
       2  0  7  7  1
       3  0  5 22  8
       4  0  2  9 22
> print(paste0("Accuracy: ", accuracy))
[1] "Accuracy: 0.614457831325301"
```

Figure 2: Random Forest Model Metrics

2.2.1 Boosting Model

Now we will implement a gradient boosting model that will learn sequentially. The goal of this to reduce both bias and variance and get a more accurate model. When we fit the boosting model, it suffered from substantial over fitting. We tried reducing our interaction depth to remedy this but it only slightly improved

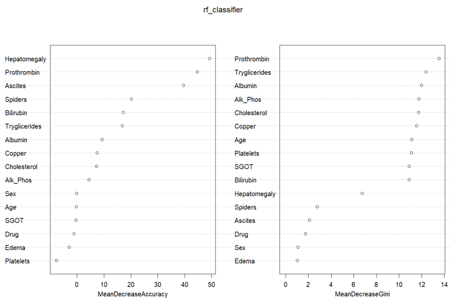
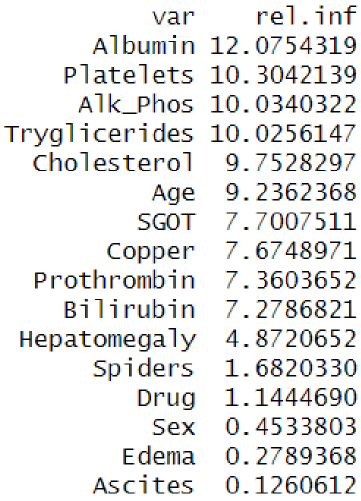


Figure 3: Variable Importance Graph



```
               var    rel.inf
           Albumin 12.0754319
         Platelets 10.3042139
          Alk_Phos 10.0340322
     Tryglicerides 10.0256147
       Cholesterol  9.7528297
               Age  9.2362368
              SGOT  7.7007511
            Copper  7.6748971
       Prothrombin  7.3603652
         Bilirubin  7.2786821
      Hepatomegaly  4.8720652
           Spiders  1.6820330
              Drug  1.1444690
               Sex  0.4533803
             Edema  0.2789368
           Ascites  0.1260612
```

Figure 4: Variable Importance from Boosting Method

the model. We tried slower learning by adjusting the interaction.depth, adjusting the number of trees, number of folds and even the size of the training data. The best accuracy we got was 50.06 percent, a decrease of nearly 10 percent from the RF model. I think the data we had was just prone to over fitting in the first place and we were unfortunately unable to come up with a remedy for this problem. However, we were able to gain some more insight into the variable importance. Interestingly, hepatomegaly was near the bottom indicating it was not very important to the boosting model.

3 Conclusion

To conclude, we were quite unsuccessful in creating an effective classification model for this data set. However, we did gain insight into some of the predictors of late stage cirrhosis. Based upon our variable importance graphs from the models, I think we can say that cirrhosis is quite independent of a person's sex. On the other hand, we could clearly see that hepatomegaly and albumin was a clear influence of a later stage cirrhosis. This makes sense hepatomegaly means that the person's liver is enlarged and much research has been done on increase levels of albumin and liver diseases. I think we were limited in the number of observations of our data. As previously stated, we only had 20 observations of people with stage 1 cirrhosis. In addition, I think the overall variability of the predictors was concerning to begin with and perhaps we should have narrowed down the predictors a bit more. More analysis could be done on this data such as comparing the groups of people who received a drug treatment versus the placebo to see if it benefited the person.

Works Cited

Data Set: fedesoriano. (August 2021). Cirrhosis Prediction Dataset. Retrieved [May, 2024] from https://www.kaggle.com/fedesoriano/cirrhosis-prediction-dataset. https://www.geeksforgeeks.org/gradient-boosting-vs-random-forest/ https://medium.com/swlh/gradient-boosting-trees-for-classification-a-beginners-guide-596b594a14ea

Appendix

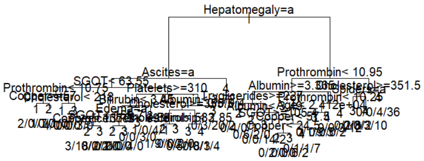


Figure 5: cp=0.01

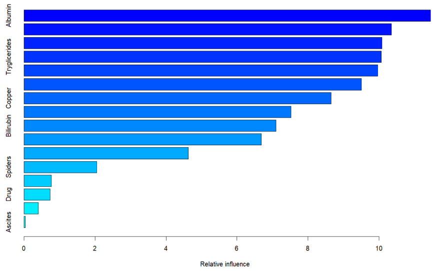


Figure 6: Boosting variable importance Graph

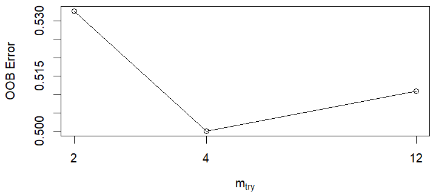


Figure 7: RF Tuning Parameters

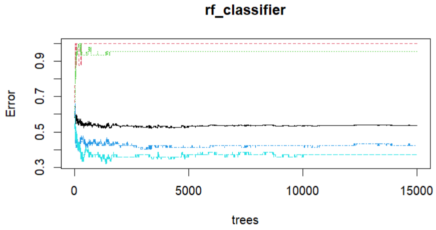


Figure 8: A different RF model with different parameters

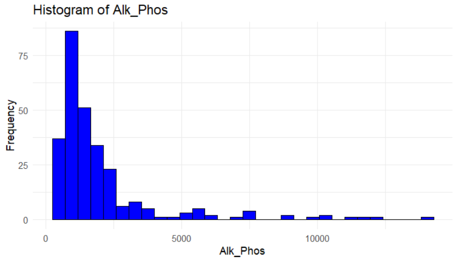


Figure 9: Albumin Histogram

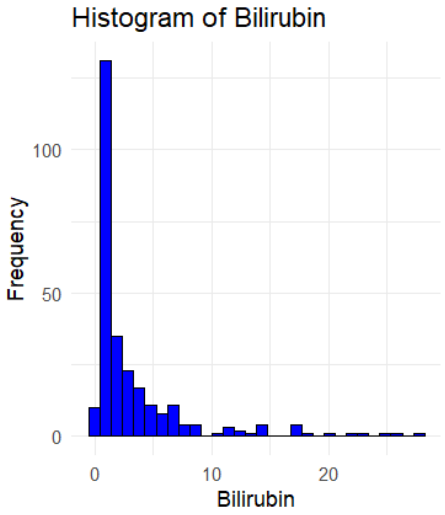


Figure 10: Histogram of Bilirubin