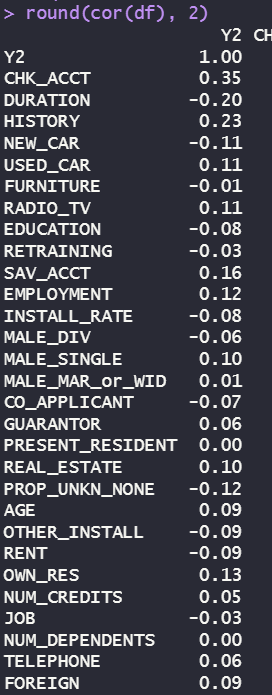
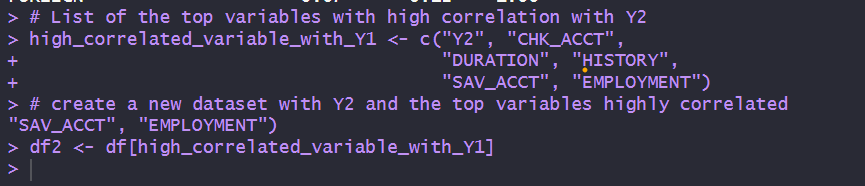
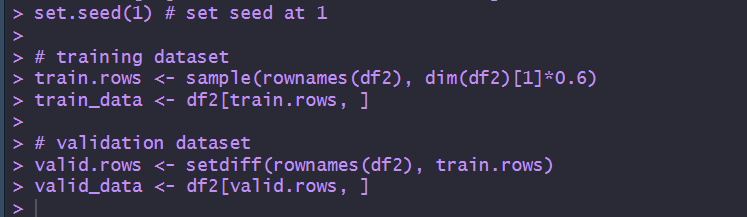
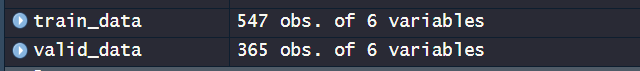
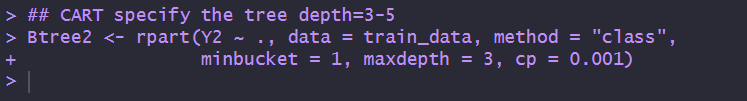
**Classification Tree Model for Y2. Y1 is excluded from this analysis.**

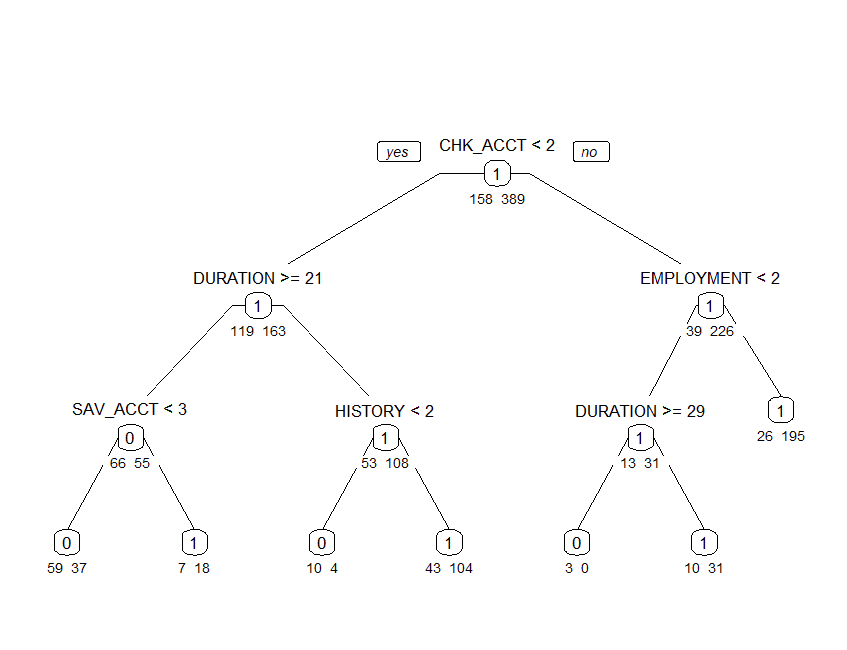
Use your individual data.

1. Explain what data you are using, number of observations, number of variables,  
  
  
I have loaded the German Credit dataset into a data frame called 'df'. This data frame includes **912** observations, with each observation corresponding to a row, and **31** variables, with each variable corresponding to a column.  
  
**What is Y2, and what percent of Y2 is =1**.  
Y2 is the binary dependent variable in a logistic regression model, indicating an individual's credit rating: '0' for bad and '1' for good.  
  
  
  
Dropped Y1  
  


**2. (From HW#6). Create a new dataset with Y2 and the top variables with high correlation with Y2 (positive or negative).**  
  
  
  
  


**3. Partition the data in 60% training and 40% validation (use seed).**  
  
  


**4. Build CART model for Y2 using the training data . Use 3 to 5 tree levels.**  
  
  
  
  
**Comment on the result:** This code constructs a tree with a minimum bucket size of 1, a maximum depth of 3, and uses a complexity parameter of 0.001 to regulate the tree's size and prevent overfitting.

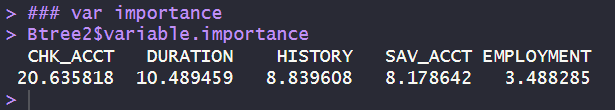
**5. Plot the CART tree (must be readable).**  
  
  
  
**Comment on the tree:** This decision tree has three levels, indicating the depth of questions from top to bottom. It branches out based on conditions starting from the top variable, CHK\_ACCT, which is the root node of the tree. The tree splits into 7 end nodes, or 'leaves,' where each leaf represents a final decision point or outcome.

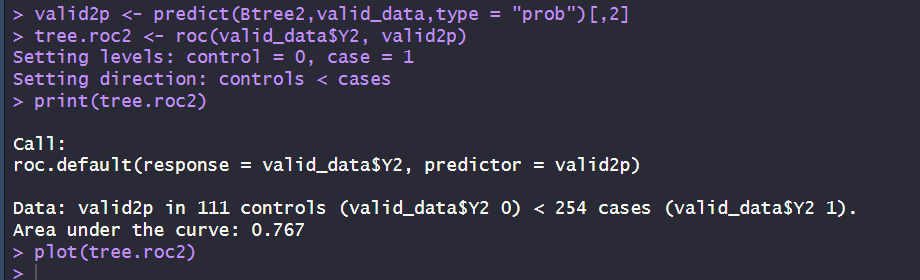
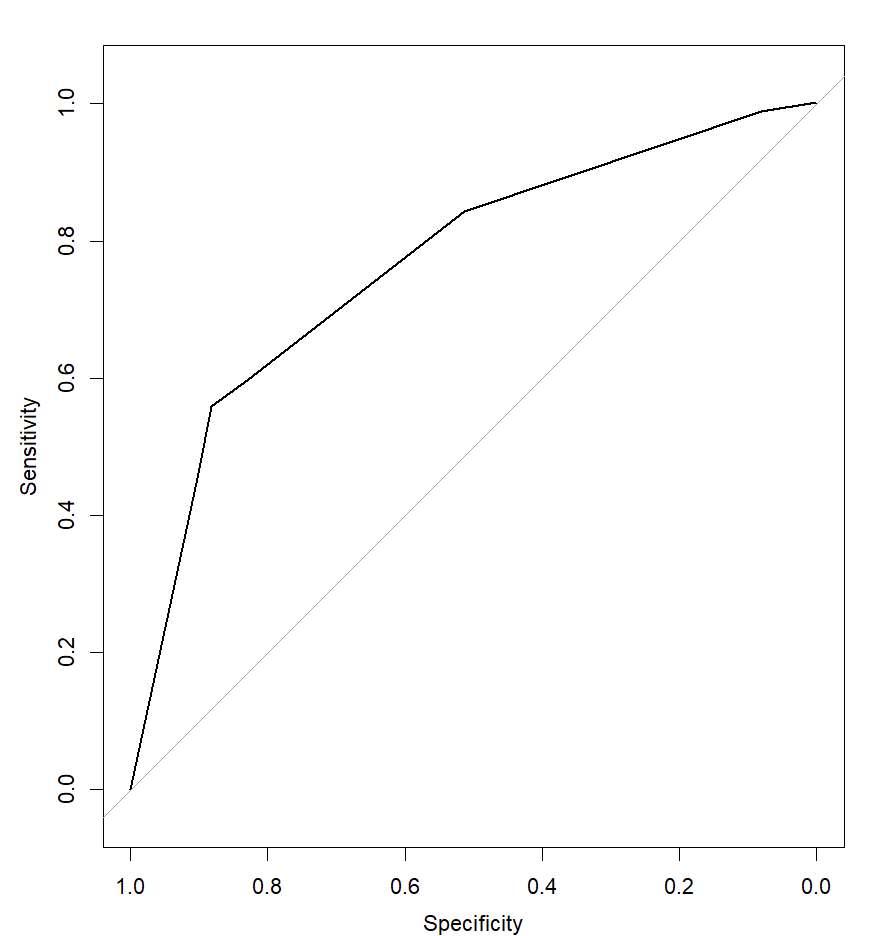
**6. Pick the end-node with the lowest proportion (lowest % Y2 = 1) and write down the path to this node in plain English.**

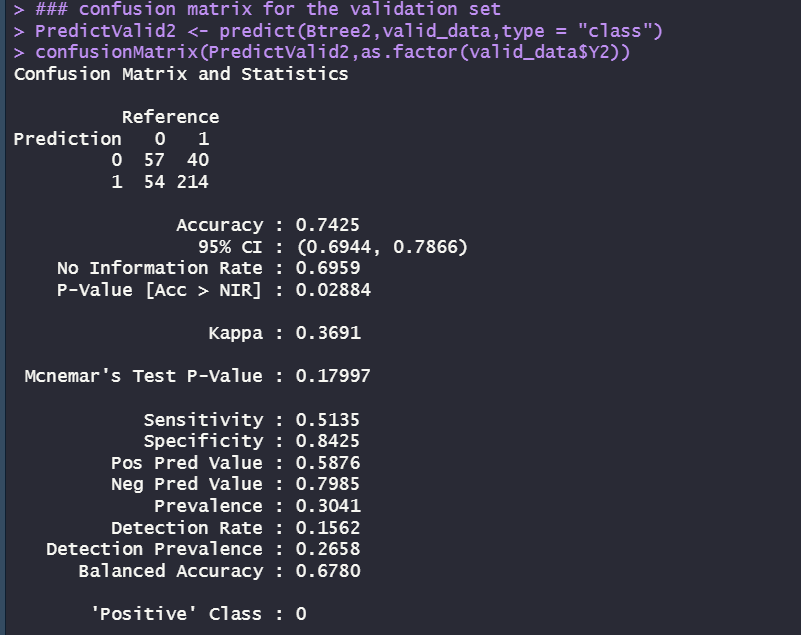
1. Start at the top node labeled “CHK\_ACCT.”
2. If "CHK\_ACCT" is less than 2, move to the left.
3. If "DURATION" is 21 or more, move to the left again.
4. If "SAV\_ACCT" is less than 3, you arrive at the end-node that has the lowest proportion of '1's, indicating a 38.54% chance of obtaining a loan.

**7. Pick the end-node with the highest proportion (highest % Y2 = 1) and write down the path to this node in plain English.**

1. Start at the top node labeled "CHK\_ACCT."
2. If "CHK\_ACCT" is greater than 2, proceed to the right node.
3. If "DURATION" exceeds 21, move to the right once more.
4. You will then arrive at the terminal node, which has the highest percentage of '1's, indicating an 88.24% chance of securing a loan.

**8. Show the variable importance measure**   
  
  
**Comment on the results:** The results show the variable importance scores from a decision tree model. The "CHK\_ACCT" variable has the highest score, indicating it is the most important predictor in the model. "DURATION" follows as the second most significant predictor, with "HISTORY" third, "SAV\_ACCT" fourth, and "EMPLOYMENT" as the least important of the variables listed. These scores help determine which features most influence the model's decisions.

**9. Use the validation data to plot the ROC and compute the Area Under ROC (AUROC).**  
  
  
   
**Comment on the result:** The model's predictability, based on the AUC of 0.767, is generally considered good as it is significantly higher than the threshold of 0.5.

**10. Use the validation data to create the confusion matrix, overall accuracy, sensitivity, and specificity.**   
  
  
 **Comment on the result:** The model achieves an overall accuracy of 74.25%, indicating it correctly predicts three-quarters of the outcomes. However, its sensitivity is relatively low at 51.35%, suggesting it identifies just over half of the actual positive cases. On the other hand, the model is quite specific, with a specificity of 84.25%, meaning it is adept at recognizing true negatives. These metrics suggest the model is more reliable at confirming negatives than detecting positives.