UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

BUDT758T

DATA MINING AND PREDICTIVE ANALYTICS

Homework 3

- Please submit on Canvas.
- Your submission should consist of either:
  i) This document (with answers filled in in the appropriate places) and a separate file with your R code.
  ii) An R Markdown file (I will post an example)
- Please ensure that answers are appropriately numbered and clearly legible.

Quantitative analysis of credit

The recent slowdown of the US economy was due in a sizeable degree to the process of extending credit to people who defaulted on their loans (typically mortgages for their houses) as they were not able to repay them. Combined with decreasing real estate prices, many of the institutions that extended the loans ended up owning property that has decreased in value, and therefore lost significant amount of money.

In the spreadsheet Credit_Dataset.csv, you will find data pertaining to 1000 personal loan accounts at a bank. The Excel spreadsheet Credit_Dictionary.xlsx contains a description of what the various variables mean.

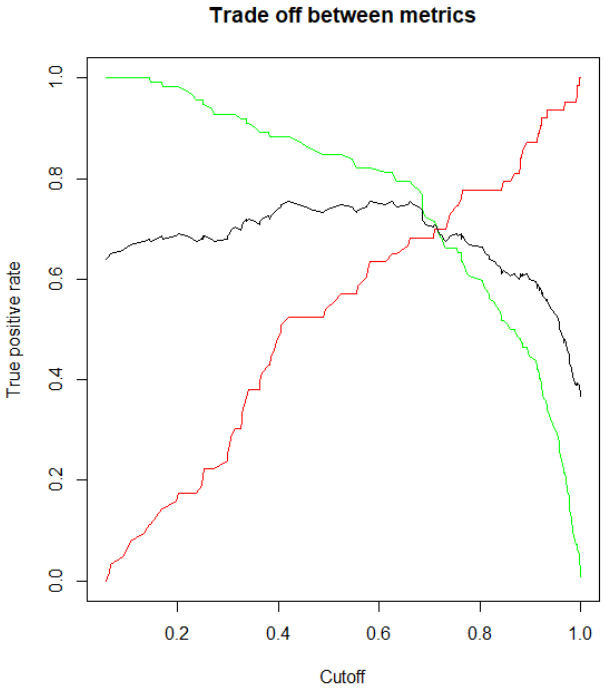
When a new applicant applies for credit, as a part of the application, the company collects information which is available in the form of Variables 2 to 21. The company then decides an amount to be credited (the variable CREDIT_EXTENDED.) For these 1000 accounts, we also have information on how profitable each account turned out to be (variable PROFIT). A negative value indicates a net loss. This typically happens when the debtor defaults on his/her payments.

The goal in this case is to investigate how one can use this data to better manage the bank's credit extension program. **Specifically, our goal is to develop a classification regression model to classify a new account as "profitable" or "not profitable".**

1. Data Preparation:
    a. Read the data set in R.
    b. The goal is to use classification methods to predict whether or not a new credit account will be profitable (not default). Create a new categorical variable to use as the dependent variable in the model (call it PROFITABLE, which is 1 if the account is not a net loss and 0 if the account is a net loss).
    c. Create factor variables from CHK_ACCT, SAV_ACCT, HISTORY, JOB, and TYPE.
    d. Set the seed to 12345 and randomly partition the data into 30% testing data and the remaining 70% data.
    e. Next, randomly partition the remaining data into 75% training data and 25% validation data.
    f. Carefully inspect the information in the data dictionary. Which variable(s) should you **not** use to classify account profitability?

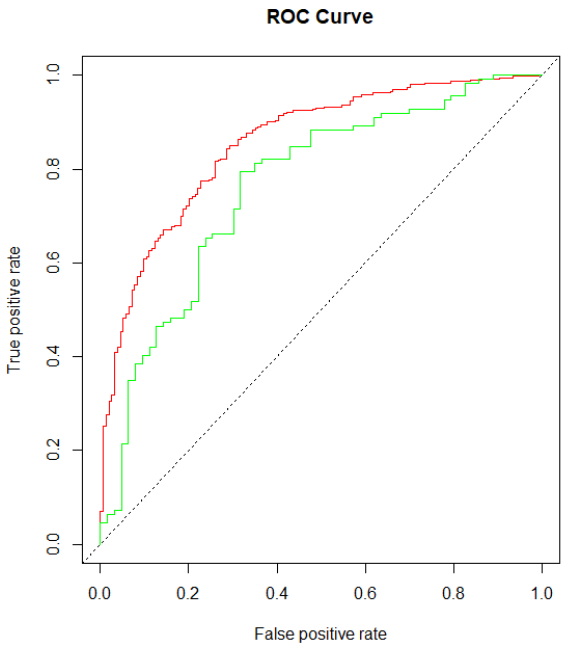
The first variable 'observation no.' has no value here. The last variable is what we are trying to predict. So, PROFIT should not be used to classify account probability.

2. Use the training data to train a logistic regression model to predict your newly created categorical dependent variable PROFITABLE, using AGE, DURATION, RENT, TELEPHONE, FOREIGN, and the factors you created from CHK_ACCT, SAV_ACCT, HISTORY, JOB, and TYPE. Install and load the ROCR package (if you haven't done so already).
    a. Plot the accuracy, sensitivity, and specificity against all cutoff values (using the ROCR package) **for the validation data.** What is that maximum accuracy value? At what value of the cutoff is the accuracy maximized?



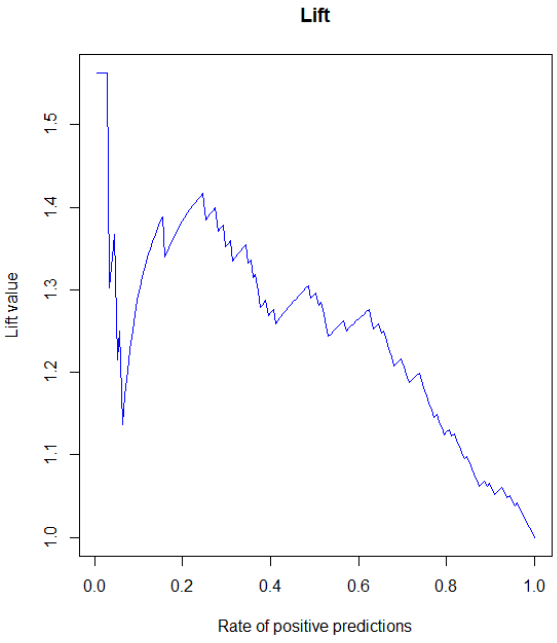
Maximum Accuracy Value is 0.754 and this occurs at a cutoff of 0.66

b. Plot the ROC curves for both the training and validation data on the same chart. What do you observe from this graph? Is there anything unexpected?



The area under the curve AUC for the validation data is lower than that of the training data. This is expected as, for every positive you get (or sale you make) you get a lot of negatives (wasted mailings or ads) in our validation data compared to our training data.

c. Plot the lift curve for the validation data. What is the maximum lift we can achieve? If we have the budget to approve 20% of the loans, what lift would we expect? If we would like to approve as many loans as possible while still achieving a lift of 1.3, how many instances will be classified as positives?

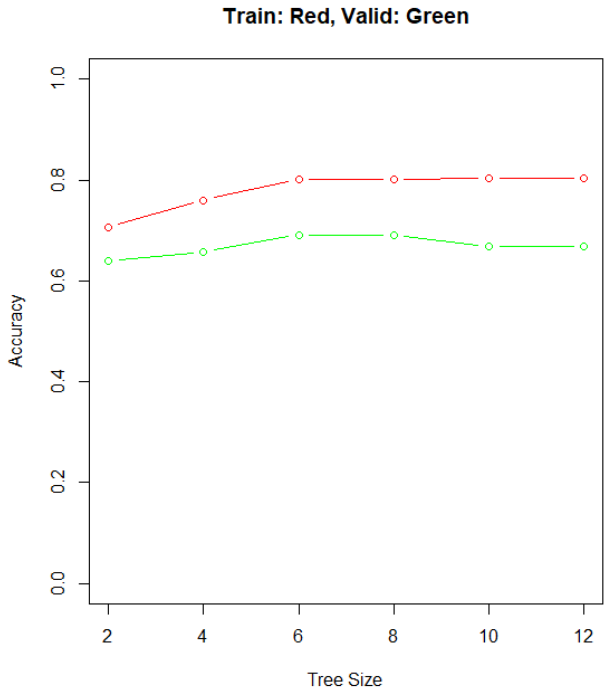


The maximum lift we can achieve is 1.56

If we have the budget to approve 20% of the loans, the Lift is 1.378

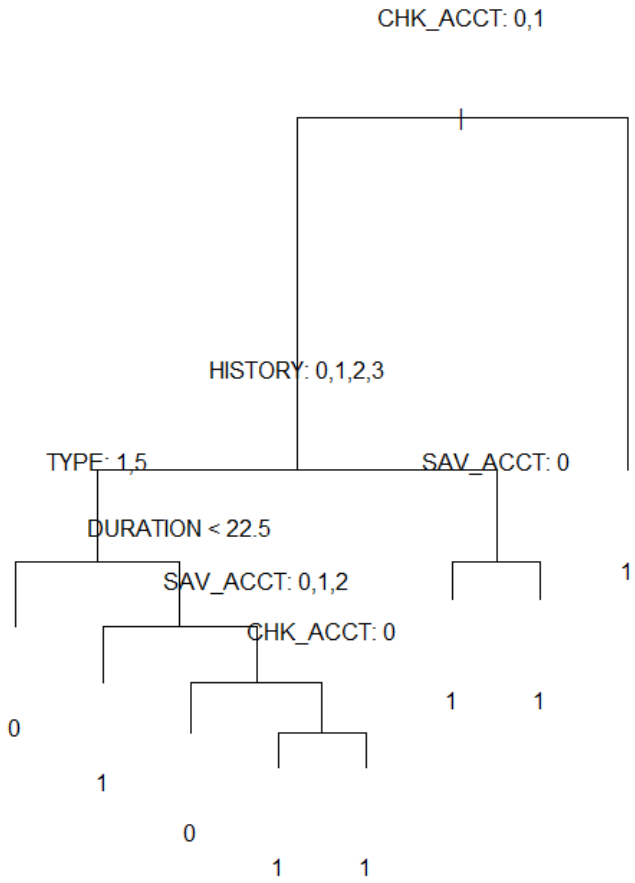
48.57% of instances will be classified positive with a lift of 1.3

3. Run a Classification Tree algorithm to predict PROFITABLE using the training data and the variables you used in your logistic regression model. Experiment with different tree sizes by modifying the number of terminal nodes in the tree. Use 10 values: 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, as well as the full (unpruned) tree.
    a. Use a cutoff of 0.5 to classify accounts and measure the accuracy in the training and validation data for each tree. Plot the tree size versus accuracy in the training and validation data (respectively) and select the best tree size.



The full tree itself has 12 terminal nodes. So, the pruning has been done for 5 trees with terminal nodes: 2, 4, 6, 8 and 10. The best tree is the pruned tree with 8 terminal nodes, as it has the highest accuracy on the validation data. The tree with the 6 terminal nodes has the same accuracy on the validation data, but is rejected because its residual mean deviance is higher than that of the tree with 8 terminal nodes.

    b. Plot the tree that results in the best accuracy in the **validation** data.



c. How many decision nodes (NOT terminal nodes) are in the full tree? How many decision nodes are in tree you plotted in part (b)?

There are 11 decision nodes on the Full Tree.
There are 7 decision nodes on the Best Tree.

d. Compare your tree from part (b) to your logistic regression summary. Do your two models agree on which variables are important to use if you want to predict if a loan will be profitable? Justify your answer.

Logistic Regression Summary:

```
> summary(logmodel)

Call:
glm(formula = PROFITABLE ~ AGE + DURATION + RENT + TELEPHONE +
    FOREIGN + CHK_ACCT + SAV_ACCT + HISTORY + JOB + TYPE, family = "binomial",
    data = credit_train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0668  -0.6551   0.3211   0.6440   2.2835

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.188385   1.179516  -0.160 0.873107
AGE          0.013876   0.011888   1.167 0.243149
DURATION    -0.043529   0.010202  -4.267 1.98e-05 ***
RENT        -0.588325   0.312075  -1.885 0.059403 .
TELEPHONE    0.246283   0.279471   0.881 0.378184
FOREIGN      1.177799   0.809904   1.454 0.145878
CHK_ACCT1    0.209688   0.297723   0.704 0.481242
CHK_ACCT2    1.617758   0.701072   2.308 0.021024 *
CHK_ACCT3    1.722917   0.324705   5.306 1.12e-07 ***
SAV_ACCT1    0.613821   0.405963   1.512 0.130531
SAV_ACCT2    1.102342   0.741336   1.487 0.137023
SAV_ACCT3    3.138442   1.129216   2.779 0.005447 **
SAV_ACCT4    0.873499   0.341425   2.558 0.010516 *
HISTORY1    -0.105693   0.741289  -0.143 0.886622
HISTORY2     0.986171   0.566963   1.739 0.081966 .
HISTORY3     1.277544   0.665854   1.919 0.055028 .
HISTORY4     2.305329   0.616521   3.739 0.000185 ***
JOB1        -0.132117   0.908414  -0.145 0.884366
JOB2        -0.335066   0.883860  -0.379 0.704618
JOB3        -0.812971   0.921007  -0.883 0.377399
TYPE1       -1.163075   0.521793  -2.229 0.025814 *
TYPE2        0.877643   0.644875   1.361 0.173529
TYPE3       -0.465463   0.539525  -0.863 0.388288
TYPE4        0.176654   0.524132   0.337 0.736086
TYPE5       -1.342524   0.738392  -1.818 0.069038 .
TYPE6        0.006256   0.609905   0.010 0.991816
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 635.36  on 524  degrees of freedom
Residual deviance: 442.46  on 499  degrees of freedom
AIC: 494.46

Number of Fisher Scoring iterations: 6
```

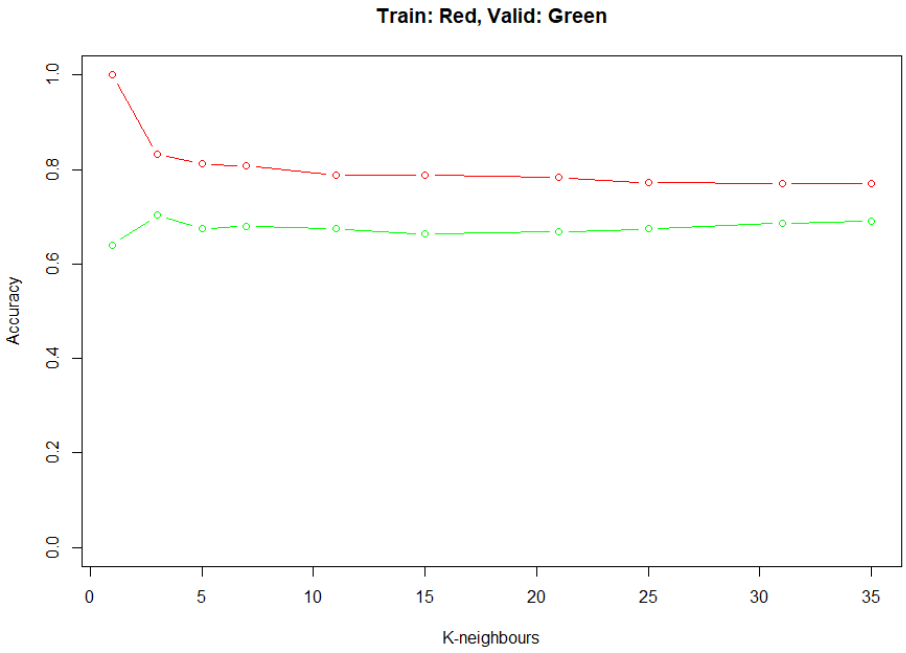
Tree Summary:

```
> summary(credit_tree_8)

Classification tree:
snip.tree(tree = credit_tree, nodes = c(18L, 3L))
Variables actually used in tree construction:
[1] "CHK_ACCT" "HISTORY"  "TYPE"     "DURATION" "SAV_ACCT"
Number of terminal nodes:  8
Residual mean deviance:  0.9223 = 476.8 / 517
Misclassification error rate: 0.1981 = 104 / 525
> plot(credit_tree_8)
```

Yes, they agree to an extent. The tree model uses only the features CHK_ACCT, HISTORY, TYPE, DURATION, SAV_ACCT.
Most of these features on the logistic regression model also have a very low p-value and hence are significant.

4. Run the kNN algorithm for classification on the training data using the following variables: AGE, DURATION, RENT, TELEPHONE, FOREIGN, CHK_ACCT, SAV_ACCT, HISTORY, JOB, and TYPE (note: **DO NOT** use the factor variables that you created in 1c). Try ten values of k: 1, 3, 5, 7, 11, 15, 21, 25, 31, and 35.
    a. Using the output, plot the accuracy for each value of k on both the training data and validation data.



    b. What is the best value of k? How do you know?

The best value of K = 3, as this has the highest accuracy on the validation set of `0.70`

c. Briefly explain why the accuracy is 100% for the training sample when k=1, but not for the validation sample.

When k=1 in the training sample, the knn model overfits and correctly classifies all data instances/points in our training data set.

5. Consider your logistic model, your best tree (the tree you plotted in 3(b)), and your best kNN model (the k value you picked in 4(b)). Calculate the accuracy of all three models on the testing data. Of all the classifiers you've trained in this assignment, which one should you ultimately use for prediction of PROFITABLE? Explain why.

Logistic Regression model accuracy at our optimal cut-off 0.66 is 0.68

Tree model accuracy with Tree Size = 8 is 0.71

Knn model accuracy with K=3 is 0.73

Knn model has the highest accuracy. But, if we care more about inference or interpretability, we should use tree model.

But in this business context, we care about whether an account is PROFITABLE or NOT PROFITABLE, to help us determine whether credit should be extended or not. So, we should use a Knn model here.

6. Did you work with anyone on this assignment? If so, include their names here:

No.