**Individual Assignment 3: k-NN and Classification Trees Applied to Credit Extensions**

This assignment continues the analysis of the credit data, exploring whether we can improve on our earlier analysis that utilized linear and logistic regression. Please refer to the earlier assignments for the data description, and repeat if needed the data preparation steps, using the credit.csv data:

1. Create a new categorical variable that indicates whether or not a new credit extension will result in a positive NPV.
2. The knn function needs all variables to be numerical, therefore create dummy variables for all categorical variables. You may want to take advantage of the R script provided (or utilize Excel).
3. Split the data into two parts, setting the seed to 1, 70% training and 30% validation.

Please answer all questions. Supply supporting documentation and show calculations as needed. **Please submit a single well-formatted PDF or Word file.** The instructor should not need to go searching for your answers! In addition, feel free to include code and screenshots as an Appendix – this will not be graded, but will help the instructor give you feedback, if your models differ significantly from the solutions.

**k-NN**

*Classify customers as profitable/not profitable with k-NN. As the second part of the assignment focuses on a bias check you may want to carefully think about which variables to include in your model.*

1. Run the k-NN algorithm for **classification**, testing multiple values of *k* (increase k until you no longer observe an improvement in the accuracy on the validation data).
2. Plot the accuracy of the validation sample. Include the plot as an Exhibit.

Chart, scatter chart

Description automatically generated

1. What is the best value of *k*?  
   16, accuracy = 0.68333
2. Briefly explain why the % Error is zero for the training sample when k=1, but not for the validation sample.

When K = 1, you'll choose the closest training data to your test data. Since your test data is in the training dataset, it'll choose itself as the closest and never make mistake. For this reason, the training error will be zero when K = 1, irrespective of the dataset.

1. Obtain predictions for the best k on the validation data.
2. Include a confusion matrix as an Exhibit.  
   Text

   Description automatically generated
3. What is the sensitivity?

0.12500

1. What is the specificity?

0.96429

1. How do these values compare to your Logistic Regression model from Week 2 (or feel free to use the solutions). **Briefly**comment.

My knn model has a slightly less accuracy but high true negative rate (specificity) and low true positive rate(sensitivity), whereas my logistic regression model from week 2 has higher accuracy but medium true negative rate (specificity) and medium true positive rate (sensitivity).

**Classification trees**

*Classify customers as profitable/not profitable with a classification tree*

1. Run the Classification Tree algorithm using all the relevant independent variables (excluding as before Credit Extended, Obs, etc. ) with the profitable/not profitable as the output variable. Use the xerror approach to prune back the tree (or if you prefer, clearly describe your approach).

Original tree

Diagram

Description automatically generated  
  
Include an Exhibit of the “prune log” or the Complexity Parameter table (in R you can use printcp()). Identify in the Exhibit which tree you chose.

Original tree cp table

Text

Description automatically generated

Min xerror prune tree

Text

Description automatically generated Text

Description automatically generated

Diagram

Description automatically generated

1. Select either the min error tree or the best pruned tree for scoring.
   1. Include the classification confusion matrix for the validation sample and a figure of your selected tree as Exhibits.

A screenshot of a computer

Description automatically generated with medium confidence Text

Description automatically generated

Min xerror prune tree from validation data

Diagram

Description automatically generated

1. Analyze the output.
   1. How many decision nodes are in your selected tree?

5

* 1. What is the error rate for the validation data in your selected tree?

Error = 0.67368

Xerror = 0.90526

Xstd = 0.082446

* 1. Would you expect the rate to be different between the training and validation samples?

Yes, given that the training data has greater sample set for prediction and that the tree was pruned to be less complex with a smaller nsplit (not overfitting).

* 1. Which applicants for credit will get rejected by the model (using the best pruned tree)? (Describe the type of customers using the English language.)

Using min. xerror prune tree from validation data, **customers** with duration of credit greater than or equal to 44 months, and **customer** with duration of credit smaller than 44 months, employment history smaller than 7 years, no guarantor, and requested credit amount smaller than 2318 will be rejected by the model.

1. Using the model for decision making.
2. Consider a 27-year-old domestic student that has $100 in her checking account but no savings account. The student has one existing credits, which has so far been paid back duly. The credit duration is 12 months. The applicant has been renting her current place for less than 12 months, does not own any real estate, just started graduate school (the present employment variable is set to 1 and nature of job to 2). The applicant has no dependents and no guarantor. The applicant wants to buy a used car and has requested $4,500 in credit, and therefore the installment rate is quite high, or 2.25%. However, the applicant does not have other installment plan credits. Finally, the applicant has a phone in her name.

How would the your tree classify the student?

CHK\_ACCT <2

DURATION <44

EMPLOYMENT <3

GUARANTOR = 0

AMOUNT\_REQUESTED > 2318

Using min. xerror prune tree from validation data, the tree will classify the student as profitable.

**Method comparisons**

1. You have now run three different classification algorithms on this data; logistic regression, classification tree and k-NN. Compare their performance in two ways. First using statistical measures (that you deem appropriate) and second using their possible impact on the credit extension process. Feel free to take advantage of the solutions to Individual Assignment 2 or your own previous work.

Text

Description automatically generatedKNN using best K of 16 on validation data. The possible impact would be variance in performance given k figure and inability to create complex results. In this specific dataset, there will be high false positive rate.

A screenshot of a computer

Description automatically generated with medium confidencemin. xerror prune tree from validation data. The possible impact would be variance in performance given predefined classes and level of complexity. In this specific dataset, the model will be highly inaccurate.

Text

Description automatically generated Regression model with 0.5 cutoff point from IA\_2. The possible impact would be variance in performance given the quality of datasets (numerical vs categorical) and # of variables. In this specific dataset, the model has a mediocre performance.