Alex Mendoza  
12/17/2023  
Extra Credit 2  
BDAT 640  
Section: 01W  
Instructor: Chris Shannon  
File Name: ExtraCredit2\_Mendoza\_Alex.docx

Your boss in the HR department wants you to develop a predictive model for employee attrition. You are to develop the following models, assess them and make a formal recommendation to your boss.

You are not allowed to use the function confusionMatrix from the caret package or any equivalent function in any other R library. You must create the confusion matrix as shown in the demo and calculate the required metrics like sensitivity, specificity and accuracy by hand-calculations, as shown in the demo file.

Use set.seed(1) right before all random sampling and functions for training models such as tree, randomForest and svm.

1. Load data from the file employee\_attrition.csv and view it. All categorical variables are properly set in the data set on loading.

A screenshot of a computer code

Description automatically generated

2. Split your data into training and test sets, with two-thirds of your data in the training set and one-third in the test set.

A close-up of a computer code

Description automatically generated

**CONTINUED ON NEXT PAGE**

3. Use logistic regression to develop a model, with Attrition as the target variable and all the rest of the variables as predictors. You can use the short-cut formula to specify the target and all predictors:

A screenshot of a computer code

Description automatically generated

4. Create a confusion matrix using the test data and calculate by hand without using ONLY the sensitivity, specificity and accuracy of the model based on the results of the confusion matrix.

- See Code, hand calculations in comments.

A screenshot of a computer code

Description automatically generated

5. Evaluate the model and make a decision as to whether a synthetic data can improve the model. Explain to your boss why the synthetic data set is needed if it is needed.

- See Code.

6. Use R to create three synthetic data sets, one using random undersampling, the second using random oversampling, and the third using the ROSE algorithm.

A computer code with blue text

Description automatically generated

**CONTINUED ON NEXT PAGE**

7. Pick one of the synthetic data sets to train your models and explain to your boss the advantages and disadvantages of the synthetic data set you picked.

- See Code.

8. Using the synthetic data set, train a logistic regression model using all the predictors to predict Attrition. Use the stepAIC function to select a best model. Set the attribute "direction" to "backward". Create a confusion matrix and calculate by hand the sensitivity, specificity and accuracy and load the results in the following variable, replacing the zeros with your hand-calculations:

- See Code, hand calculations in comments.

A close-up of a text

Description automatically generated

A screenshot of a computer code

Description automatically generated

9. Train a classification tree with the synthetic data set using all the variables to predict Attrition. Use cv.tree to select the optimal tree size, and then use prune.tree to prune your tree. Use the pruned tree to predict the outcome using the test set and create a confusion matrix. From the confusion matrix, calculate the sensitivity, specificity, and accuracy, replacing the zeros with your hand-calculations.

- See Code, hand calculations are in the comments.

A computer code with blue text

Description automatically generated

**CONTINUED ON NEXT PAGE**

A graph of a tree

Description automatically generated

A computer code with numbers and symbols

Description automatically generated

10. Train a random forest model with the synthetic data set and use all variables to predict Attrition. Print out the variable importance using importance() and display the importance using varImpPlot(). Use predict() with the test data to create a set of predictions. Create a confusion matrix and enter your hand-calculations for the sensitivity, specificity and accuracy into the below variable:

- See Code, hand calculations are in the comments.

**CONTINUED ON NEXT PAGE**

A screenshot of a computer

Description automatically generated

A computer screen shot of a code

Description automatically generated

**CONTINUED ON NEXT PAGE**

11. Combine the variables model.logistic, model.classtree and model.randomforest into a table and print it.

A close-up of numbers

Description automatically generated

12. Create an ROC curve featuring the curves of all three models. Give the graphic the following title: "ROC Curves of Attrition Models". Give the model a legend, and in the plot report which model has the maximum Area Under Curve (AUC) and also report the value of the maximum AUC in the plot.

**CONTINUED ON NEXT PAGE**

A screenshot of a computer code

Description automatically generated

A graph of a curve

Description automatically generated with medium confidence

13. Using the table in problem 11 and the ROC curve graphic in problem 12, recommend to your boss which model the department should use and why.

- See Code.