**Data Science for Social Scientists**

Psyc 546, Spring 2023

Week 12 – In-Class Assignment

**Due Date**: April 13th (by 11:59 PM)

**Reminder**: See the assigned readings, resources on Canvas, and the lecture slides for a tutorial on how to use R to perform the various functions included in the in-class assignment below. **Once completed, you should submit a completed version of this document and your final R script file to the Week 12 – In-Class Assignment – Submission Portal on Canvas**.

Your submitted R script file should contain code to answer the questions below. Please use comments (e.g., #Question 1) to label the code for each question.

1. Q1 uses the **binary.csv** from Canvas. This data file contains whether students were admitted to a graduate program or not (admit), their undergraduate GPA (gpa), their GRE score (gre), and the rank of their undergraduate institution (rank). First, perform a logistic regression model predicting graduate admittance with the other three variables as predictors. Then, use the fitted model to provide predicted admittance for all the participants (rounded to 0 and 1). Finally, use confusionMatrix() to calculate the following values: [2 points overall]:
2. The accuracy of the model: 0.705
3. The sensitivity of the model: 0.7208
4. The specificity of the model: 0.5918
5. The kappa of the model: 0.1856
6. Using the logistic regression model from Q1, produce a ROC curve figure. Make sure that the area-under-the-curve (AUC) is included. Paste the figure below. [2 points]

Chart

Description automatically generated

1. Q3 will use the holdout method of cross validation with the model from Q1. First, set the seed of the R environment to 257. Then, split the binary data file into 70%/30%. Have the 70% applied to an object called binary\_train; have the 30% applied to an object called binary\_test.

Re-run the logistic regression model from Q1 but on the training data. Then, use the fitted model to provide predicted admittance for the participants in the testing data (rounded to 0 and 1). Finally, use confusionMatrix() to calculate the following values on the test data: [2 points overall]:

1. The accuracy of the model: 0.7062
2. The sensitivity of the model: 0.7463
3. The specificity of the model: 0.5000
4. The kappa of the model: 0.1858
5. Q4 will use repeated k-fold cross validation with the model from Q1. First, set the seed of the R environment to 717. Then, use the trainControl() function to choose repeated k-fold CV with 5 folds and 5 repetitions. Finally, use the train() function to run the model from Q1. Report following values from the results of the CV: [2 points overall]:
6. The accuracy: 0.6964856
7. The kappa: 0.1619478
8. Q5 will use repeated k-fold cross validation with a predictive/regression model (still using the binary data file). First, set the seed of the R environment to 500. Then, use the trainControl() function to choose repeated k-fold CV with 10 folds and 5 repetitions. Finally, use the train() function to run a model with GRE as the DV and GPA and Undergrad Institution Rank as two predictors. Report following values from the results of the CV: [2 points overall]:
9. Root Mean Squared Error: 106.2056
10. R2: 0.1689974