**CAP 5768: Homework 3**

**Due on Canvas by Sunday, April 9, 2024, at 11:59pm**

**Place name here:**

**Preliminary instructions**

All analyses must be performed in Python using the packages that we discussed in class. Fill in all your solutions in the appropriate spaces provided in this Word document, and then upload a PDF copy of your solutions to Canvas. **Only PDF copies will be graded**.

**Brief overview of the assignment**

In this assignment, you will analyze the **college** dataset that is available with this assignment in Canvas and under the datasets module. This dataset has information on 18 features for 777 US colleges obtained from the 1995 issue of US News and World Reports. The columns in the dataset are:

Name Description

**Private** A factor with levels **No** and **Yes** including private or public university

**Apps** Number of applications received

**Accept** Number of applications accepted

**Enroll** Number of new students enrolled

**Top10perc** Percent new students from top 10% of high school class

**Top25perc** Percent new students from top 25% of high school class

**F.Undergrad** Number of full-time undergraduates

**P.Undergrad** Number of part-time undergraduates

**Outstate** Out-of-state tuition

**Room.Board** Room and board costs

**Books** Estimated book costs

**Personal** Estimated personal spending

**PhD** Percent of faculty with a Ph.D.

**Terminal** Percent of faculty with a terminal degree

**S.F.Ratio** Student/faculty ratio

**perc.alumni** Percent alumni who donate

**Expend** Instructional expenditure per student

**Grad.Rate** Graduation rate

**Questions and problems**

**1. [6%]** Recode the binary feature **Private** with values **0** and **1** in place of **No** and **Yes** and store it in a new data frame called **College\_recoded**.

**Provide the code below:**

**2. [10%]** Fit a multiple linear regression model to predict private school status with the 17 other features. Which feature is most important in this model, and what evidence tells you that?

**Provide the code below:**

**Provide the answer to the question below:**

**3. [5%]** Fit a simple linear regression model to predict private school status based on the most important feature from Question 2. Is this feature still important in this model, and what evidence tells you that?

**Provide the code below:**

**Provide the answer to the question below:**

**4. [10%]** Visualize the simple linear regression model from Question 3 using a scatter plot and the fitted linear model.

**Provide the code below:**

**Provide the figure below:**

**5. [10%]** Make predictions of the classes for the training dataset using your simple linear regression model from Question 3, and add these predictions to the data frame **College\_recoded** that you created in Question 1. Create a confusion matrix and estimate classification accuracy for the training dataset.

**Provide the code below:**

**Provide the confusion matrix below:**

**Provide the accuracy estimate below:**

**6. [12%]** Perform the same operations as in Question 5, except use the multiple linear regression model from Question 2. Has the classifier improved in performance on the training data compared to the results from Question 5? Explain why you conclude this and provide a reason as to why this model did or did not improve upon the training error from Question 5.

**Provide the code below:**

**Provide the confusion matrix below:**

**Provide the accuracy estimate below:**

**Provide answers to the questions below:**

**7. [10%]** Fit a multiple logistic regression model to predict private school status with the 17 other features. Which feature is most important in this model, and what evidence tells you that?

**Provide the code below:**

**Provide the answer to the question below:**

**8. [5%]** Fit a simple logistic regression model to predict private school status based on the most important feature from Question 7. Is this feature still important in this model, and what evidence tells you that?

**Provide the code below:**

**Provide the answer to the question below:**

**9. [10%]** Visualize the simple logistic regression model from Question 8 using a scatter plot and the fitted logistic model.

**Provide the code below:**

**Provide the figure below:**

**10. [10%]** Make predictions of the classes for the training dataset using your simple logistic regression model from Question 8 and add these predictions to the data frame **College\_recoded** that you created in Question 1. Create a confusion matrix and estimate classification accuracy for the training dataset.

Provide the code below:

Provide the confusion matrix below:

Provide the accuracy estimate below:

**11. [12%]** Perform the same operations as in Question 10, except use the multiple logistic regression model from Question 7. Has the classifier improved in training accuracy compared to the results of the multiple linear regression model from Question 6? Explain why you conclude this and provide a reason as to why this model did or did not improve upon the training error from Question 6.

**Provide the code below:**

**Provide the confusion matrix below:**

**Provide the accuracy estimate below:**

**Provide answers to the questions below:**