FAIRLEIGH DICKINSON UNIVERSITY

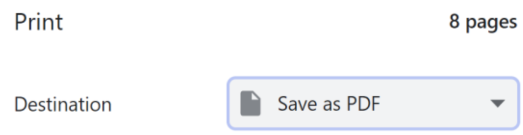
DSCI 6780 Homework 5 Improvement

Due: 5pm Dec 13 Wednesday, 2023 before your Capstone Meeting

Submission Instructions:

Three documents should be submitted to WebCampus.

1. Main Document (will be graded): This word document including all detailed answers/explanations for each question. **Excel and Python outputs, graphs, and plots should be included as part of your answers to support your argument/explanation.**
    a. Readers of your main file should **NOT** need to look into your supporting Python or Excel file to find answers.
    b. Insufficient information in the main document will significantly reduce your grade.
2. Supporting Document 1 (will **NOT** be graded): One single Python **HTML file printed as PDF format** for Part I (After you download the Jupyter Notebook file as html, open the html file, print, choose "Save as PDF").



Print 8 pages

Destination Save as PDF

3. Supporting Document 2 (will **NOT** be graded): One single Excel file for Part II.

PART I: Logistic Regression in Python (65%)

You are working as an external consultant for a hospital to help predict patients who are likely to develop diabetes. You were presented with two datasets: *HW5_diabetes.xlsx* and *HW5_newpatient.xlsx*.

HW5_diabetes.xlsx consists of historical data from 718 patients who have already conducted diabetes diagnosis. The last column "diabetes" indicates if a patient was diagnosed as having diabetes (labelled as 1) or not having diabetes (labelled as 0). The other eight columns are individual information about each patient. *HW5_newpatient.xlsx* is a file of 50 new patients who share the same individual patient information as in *HW5_diabetes.xlsx*. The 50 new patients are to be diagnosed. To most efficiently allocate medical resources, the hospital wants to prioritize the medical care process for those new patients who might have a high probability to develop diabetes.

The task here is to build a model using historical data *HW5_diabetes.xlsx*. Then, use the model built from the historical data to predict the likelihood of developing diabetes for the 50 new patients.

Tasks:

1. Build a logistic regression model using *HW5_diabetes.xlsx*. (**10%**).
   Hint:
   Column 'diabetes' is y and all the other columns are X.
   If Python gives an error message of "Total no. of iterations reached limit", make sure you set max_iter to a larger number in your model.
2. Report confusion matrix of your model. Briefly discuss the meaning of your confusion matrix. (**10%**)
3. Report precision, recall, and accuracy measures. Briefly discuss the meaning of each measure. (**10%**)
4. Plot ROC Curve for your model. Briefly discuss the ROC Curve. (**10%**)
5. Using the model built from step 1 to predict the probability of developing diabetes for the 50 new patients. Report those patients who have a more than 50% probability of developing diabetes. How many of them? (**25%**)
   Hint:
   All columns in HW5_newpatient.xlsx are your new X. We don't have a new y here as we are predicting y.

PART II: Monte Carlo Simulation in Excel (35%)

Higgins Plumbing and Heating maintains a stock of 30-gallon water heaters that it sells to homeowners and installs for them. Owner Jim Higgins likes the idea of having a large inventory on hand to meet any customer demand. However, he also recognizes that it is expensive to do so. He examines water heater sales over the past 50 weeks and notes the following:

Water Heater Sales Per Week	Number of Weeks This Number of Water Heater was Sold
4	6
5	5
6	9
7	12
8	8
9	7
10	3
	50 weeks total data

a) If Higgins maintains a constant inventory of 8 water heaters in any given week, how many times will he stockout during a 20-week simulation? (**10%**)
Hint: There is no correct or wrong answer here because the random number generation will generate different random numbers each time you refresh your simulation. Report one snapshot simulation result in below.

b) Run a Monte Carlo simulation for 52 weeks with 1000 runs and calculate the average number of sales. Report your frequency table in below. What recommendations would you make to Jim? (**25%**)
Hint: One single run is 52 weeks. We need to record this 52-week run result 1000 different times and take the average of these 1000 runs.