DATA SCIENCE: Machine Learning TEAM PROJECT

This programming assignment contains 2 independent parts, creatively called Part A and Part B ☺)

PART A (telco churn analysis): (10 points)

For this part of the project, you will analyze the TelCo CHURN data set. Divide the entire data set into training and test sets (the test set should be 25% of the original data set).

PART A Deliverable: Apply 10-fold cross-validation to build TWO distinct models to predict customer CHURN. The two techniques are a) Decision Trees and b) Logistic Regression. Use the best combination of predictor variables for this purpose (ok to use the variables in the Lab notebook for Decision trees). There are 5 parts to PART A of your submission:

A.1: For both Decision Trees and Logistic Regression, report the accuracy for the 10 folds. Also, compute the AVERAGE accuracy across the 10 folds as well as the STANDARD DEVIATION of accuracy across the 10 folds. Which technique (LR or Trees) has a higher average accuracy? (2 points)

A.2: Accurate Jupyter notebook pdf in Appendix A.2 of your Decision Tree cross validation code (2 points)

A.3: Accurate Jupyter notebook pdf in Appendix A.3 of your Logistic Regression cross validation code (2 points)

A.4: Consider the 4 cells (p, Y), (p, N), (n, Y) and (n, N) (see chapter on confusion matrix from Provost book). For each of these cells come up with a BENEFIT/COST for every customer that falls into the cell. There is no right or wrong answer here, but this has NOTHING to do with parts A.1,A.2,A.3 above. This is based upon a BUSINESS understanding of the costs/benefits of misclassification. State your rationale for the numbers you provide (2 points).

A.5: Look carefully at ALL the predicted "CHURN/LEAVE" node-leafs of your decision tree. As a business manager, describe each churning segment in words. Recommend ONE choice of CHURN segment where you will focus your resources to reduce churn. Why did you pick this one segment from all the available alternatives? (2 points)

TOTAL for PART A: 10 points.

Hints for writing a good report for Part A:

The report for PART A does not have to be long (no more than 5 pages), but should be super-clear. Imagine you are presenting the report to senior management. Here are some suggestions:

Part A.1: Describe the numbers below in a table:

Cross-validation Fold	Decision Tree	Logistic Regression
Fold 1		
Fold 2		
Fold 3		
Fold 4		
Fold 5		
Fold 6		
Fold 7		
Fold 8		
Fold 9		
Fold 10		
Average Error %		
Std. Dev. Error %		

See IRIS_PRACTICE_CROSSVAL Jupyter notebook in Module 6 for how to create a 10-fold cross-validation (that notebook shows you a 5-fold).

Next place Part A.4 (not parts A.2, A.3) :

	p	n
Y	$ Benefit	$ cost
N	$ cost	$ benefit

Write a few sentences supporting the $benefit/cost numbers. How did you come up with these numbers?

Next place Part A.5 (not parts A.2, A.3) :

	RULES for identifying	Description in words
Churn segment 1…		
Churn segment 2…		

The Rules are as stated by the decision tree. WHICH CHURN SEGMENT DO YOU RECOMMEND FOCUSING ON? WHY?

Finally, in appendices for PART A, place Jupyter notebooks Parts A.2 and A.3.

Appendix A.2: Decision Tree cross validation notebook: construct this notebook by combining ideas from Churn_Telco and Iris_practice_crossval notebooks.

Appendix A.3: Logistic regression cross validation notebook: construct this notebook by combining ideas from Churn_Telco, WBCD and Iris_practice_crossval notebooks.

NOTE: The data set used for logistic regression is STILL the Telco churn dataset.

TEAM PROJECT: Part B (5 points). (Recommended length for PART B: 2 pages).

Part B: Use the Simmons data set in module 10. See the Excel file titled Simmons-data-raw in Module 10. Watch the video that explains the contents of the file. The data set uses two predictors X1 = Annual spend on a similar credit card and X2 = Presence/Absence of the Simmons loyalty card to PREDICT Y = Will customer use coupon or not? For Part 2, build a logistic regression model to predict Y = coupon usage from X1 and X2 and then answer the following questions.

PartB-1 (2 points): What are the coefficents (BETAs) for the logistic regression model? Answer as below:

LR coefficents	Value
BETA0 (or constant term)	
BETA1 (coeff. For X1)	
BETA2 (coeff. For X2)	

PartB-2 (2 points): Use the model above to compare TWO customers Jack and Jill. Jack spends $2000 annually (note: X1 for Jack = 2) and HAS the Simmons card (X2 = 1). Jill spends $4000 annually (X1 = 4) and does NOT have the Simmons card (X2 = 0). Who is more likely to use the coupon? (Hint: A complete answer must evaluate their probabilities for response).

	Probability of Response
Jack	
Jill	

XXXX is more likely to respond because…

PartB-3 (1 point): If you were to ROLL OUT the logistic regression model to PREDICT coupon usage for a LARGE database of customers, what CUTOFF probability will you choose? (Hint: No right or wrong answer here, but a concept such as a CONFUSION MATRIX may help make your call for cutoff probability).

TOTAL POINTS FOR PART B: 5 points.

Please submit ONE single word or pdf file (that collates your answers to both PARTS A and B) under ASSIGNMENTS, by the due date stipulated in the learning management system. The time stamp provided by the system will be the official assignment submission time. Late assignments will not be accepted by the LMS. The total for parts A and B should not exceed 7 pages including tables/figures.

State all contributing team member names INSIDE the assignment (at top) clearly. Good luck!