Machine Learning Team Project

Group 5

Professor Ram Gopalan

[53:716:545:90 MSBA\_MACHINE LRNG APPLIC](https://rutgers.instructure.com/courses/303763)

October 114, 2024

ROBERT GORMAN

DOMINICK FILAR

CATHARINE TARQUINIO

**Part A: Telco Churn Analysis**

Part A-1: Accuracy

* *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)*

|  |  |  |
| --- | --- | --- |
| **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 %** |  |  |

Part A-2 Decision Tree Cross-validation

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

Part A-3 Logistic Regression Cross-validation

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

Part A-4 Cost-Benefit Matrix

*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)*

Part A-5 Churning Segments

*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?*

**Part B: Simmons Dataset**

* *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.*

Part B-1 Regression Coefficients

* *What are the coefficents (BETAs) for the logistic*  
  *regression model? Answer as below:*

|  |  |
| --- | --- |
| **LR coefficients** | Value |
| BETA0 (or constant term) |  |
| BETA1 (coeff. For X1 ) |  |
| BETA2 (coeff. For X2) |  |
|  |  |
|  |  |

Part B-2 Predictions

* *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)*

Jack: Spends $2000 annually (X1 = 2) and has a Simmons card (X2 = 1).

Jill: Spends $4000 annually (X1 = 4) but does not have a Simmons card (X2 = 0).

|  |  |
| --- | --- |
| **Customer** | **Probability of Response** |
| Jack |  |
| Jill |  |

Part B-3 Large Rollout

* *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).*