IEOR 4650: Business Analytics
Spring 2019 Professor Adam Elmachtoub

Assignment 3: Classification Methods

Due date: March 1, 11:59am

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_3.pdf` and `Obama_Barack_3.R`. Your submissions must be based on your own original work. Late submissions will be penalized at 10% per hour.

1. In this question, you are asked to analyze consumer data to help design a future promotion campaign for the MM brand. We will use the data from `OrangeJuice.csv`, which includes observations on customer orange juice purchases. The first variable `Purchase` is the brand of orange juice the consumer previously purchased, which is either the brand MM or CH. The other variables are as following:

   - WeekofPurchase - Week of purchase
   - StoreID - Store ID
   - PriceCH/PriceMM - Price charged for CH/MM
   - DiscCH/DiscMM - Discount offered for CH/MM
   - SpecialCH/ SpecialMM - Indicator of special on CH/MM
   - LoyalCH - A proxy for customer brand loyalty for CH
   - SalePriceCH/SalePriceMM - Sale price for CH/MM
   - PriceDiff - Sale price of MM less sale price of CH

   (a) Load the data from `OrangeJuice.csv` and split your sample into training (50%), validation (25%), and test (25%) data. Use the command set.seed(1337) to set the randomizer's seed. Print the summary of the training data. Which variables are qualitative and require special treatment?

   (b) Fit a logistic regression to predict `Purchase` using all the covariates over the training data. Print the estimated coefficients and interpret them.

   (c) Now let's fit logistic regression with a LASSO penalty. We shall try values of λ from $10^{[-3:3]}$. Do cross-validation on the *training (50%) data* and report the best λ. What is the final model on the training data, using the best λ?

(d) Fit an LDA classifier on the training data to predict `Purchase`. What is the classification error on the training data?

(e) Use cross-validation on the training data to find the best k-Nearest Neighbors algorithm (find the best k). What value of k is best, and what is the classification error on the training data?

(f) Choose the best model from the 4 previous parts based on the validation data. Specifically, choose the model that has the lowest classification error on the validation data. Which method performed the best?

(g) Refit your best model from the previous question on the combined training and validation data. Assess the final model on the test data.

(h) You are running a promotion aimed to convince customers to sample the MM orange juice. Your campaign wishes to target customers who bought CH orange juice and give them a coupon for MM. Your marketing team tells you that a coupon handed to a customer who bought CH will help convert the customer and will generate a \$3.50 profit, however a coupon that is handed to a customer who already bought MM will be a waste and will generate a loss of \$0.50. Using one of the previous models, you need to decide who receives coupons. For the chosen method, find the optimal threshold to convert the probability to a decision. What is the the best attainable payoff on the test data?