Assignment 3

(To be done in teams. **Submit a pdf file (with cover sheet) to Quercus by 11:59 PM March 31**.)

This assignment is based on the Relay Revisited case at the end of Chapter 9 of the VFW book (see Class 7 Module in Quercus). There are two data sets associated with the case: RetailRelay(C)TestData.xlsx and RetailRelay(C)TrainingData.xlsx. Please read the case before doing the assignment.

In this assignment you will use two modeling strategies, RFM and logistic regression, to predict customer retention (the variable "retained") and evaluate their relative predictive efficacy via lift and gain analysis.

Part I: Calibrating models using the training data

1. Estimate a logistic regression model with "retained" as the dependent variable and the following explanatory variables: firstorder, lastorder, esent, eopenrate, eclickrate, avgorder, ordfreq, paperless, refill, and doorstep. Report the coefficient estimates and identify which variables have statistically significant effects. (10 points)
2. Compute and report the marginal effects of each independent variable using the average-marginal-effect method. (10 points) Which of the marginal effects seem "economically important"? (5 points) What are the managerial implications of your findings? (5 points)
3. Run a RFM model using "lastorder," "ordfreq," and "avgorder" with deciles and calculate the average retention rate for each composite index.

Part II: Validate the two models on the test data

4. Check the predictive performance of the logistic regression on the test data using a 0.5 classification threshold. Report the confusion matrix, accuracy, precision, and sensitivity. (10 points)
5. Assign each customer in the test data to a decile based on his/her predicted probability of retention based on the estimated logistic regression.
6. Report the number of customers, the number of actual retained customers, and the actual average retention rate in each of the deciles created in the previous question. (10 points)
7. Assign each customer in the test data an RFM composite index and predict her retention rate based on the RFM analysis you did on the training data in Q3. Sort customers into deciles based on his/her predicted probability of retention.
8. Report the number of actual retained customers, and the actual average retention rate for each of the deciles created in the previous question. (10 points)
9. Assess the predictive performance of RFM in the test data by developing a classification table similar to one developed above for the logistic regression. Report the confusion matrix, accuracy, precision, and sensitivity. (10 points)

Part III: Lift and Cumulative Lift in the test data

10. Use the computations in Q6 and Q8 to create a table showing the lift and cumulative lift for each decile, for both logistic regression results and RFM results. You may want to use Excel for these calculations.
11. Use the computations in Q6 and Q8 to create a table showing the gains and cumulative gains for each decile, for both logistic regression results and R(FM) results. You may want to use Excel for these calculations. Report a composite table showing the computations in Q10 and Q11.
12. Create a chart plotting cumulative gains against cumulative customers for the logistic regression segments, the RFM segments, and random segmentation. Does the logistic regression do a better job predicting retention than RFM? (20 points) Explain why it does or does not do so. (10 points)