The University of Texas at Dallas

School of Management

BUAN 6383/MIS 6386 **Modeling for**
Syam Menon **Business Analytics**

Homework 03

Objective

- Learn to build, evaluate, and select classification models

Instructions

- **Due Date: See Syllabus**
  eLearning will stop accepting submissions after the due date, and late submissions will not be accepted
- **Submit one report per group via eLearning as a Microsoft Word document**
  - The report should be named **hw03-group*xx*.docx**
    (for example, group 05 should name the report hw03-group05.docx)
  - Clearly identify your group number and all group members on the cover page
  - A professional quality report is expected – messy or hard-to-read reports will be penalized
- **Submit all the code you have developed as a jupyter notebook**
  - The file should be named **hw03-group*xx*.ipynb**
  - (for example, group 05 should name the notebook hw03-group05.ipynb)
  - If you prefer, you can submit separate jupyter notebooks for each question If you choose to do so, the files should be named **hw03-group*xx*-p*yy*q*zz*.ipynb** (for example, group 05 should name the notebook
  - for Part 2, question 3 hw03-group05-p02q03.ipynb)
  - Clearly identify which question each part of the code is for, and what it is supposed to do
  - Clear, detailed comments are required; I should be able to run the codes you submit
- **This homework counts for 70 points**

Data Sets

- `dmtrain.csv`
- `dmtest.csv`

Predicting Customer Response

A direct marketing firm mails catalogs to its customer base of about 5 million households. Customers respond either by ordering items from the catalog, or do not respond. The firm distinguishes itself by mailing expensive catalogs, and – while the response rates to the firm's mailers are higher than the industry average (30% vs 22%) – they incur considerable printing and mailing costs. They are trying to improve their performance by identifying and targeting profitable customers, i.e., customers who are likely to respond (and order items that would justify the printing and mailing costs). They are particularly interested in *lapsing customers* (customers who made their last purchase 13 to 24 months ago).

A preliminary study shows that customers seem to make their buying decision in two phases – they decide whether or not to respond first and, if they decide to respond, make a follow-up decision on what to order. `dmtrain.csv` contains information about 2,000 customers from the last mailing campaign. Everyone included has made at least one purchase from the firm in the past. The variables involved are:

Variables	Description
`id`	customer ID
`n24`	number of orders in the last 24 months
`rev24`	total order amount ($) in the last 24 months
`revlast`	amount of last order ($)
`elpsdm`	time elapsed since last order (months)
`ordfreq`	order frequency over the last 24 months (1, 2, 3 → actual number of orders; 4 → 4 or more orders)
`ordcat`	order amount category (1 → $0.01–$1.99, 2 → $2.00–$2.99, 3 → $3.00–$4.99, 4 → $5.00–$9.99, 5 → $10.00–$14.99, 6 → $15.00–$24.99, 7 → over $25.00)
`response`	1 → customer responded, 0 → no response

The broad objective is to classify customers who are likely to respond to mailers, and customers who are not (i.e., `response`) is the dependent variable.

1. Read in the data and review the non-binary variables to see if any are skewed and need to be transformed. If so, transform them and drop the non-transformed versions of the variables. Make sure that you do not include the customer identifier `id` in your calculations. Explain what you found, what transformations you applied, and why.
2. Generate a decision tree on the entire dataset, without any limitations on the depth of the tree. Use entropy as the metric. What is the depth of the tree that is generated? Provide a plot of the tree.
3. We will focus on decision trees first, and try to identify the best decision tree classifier by pruning the tree at different depths. Use 10-fold cross validation and identify the best tree depth (again, using accuracy as the metric), by trying as many possible depths as you deem necessary. Provide your reasoning for using the values of tree-depth that you tried. Based on your results, what depth do you recommend? What is the accuracy associated with this tree depth? If you had to select the three best values of tree-depth, what would they be?
4. Next, we will consider random forests. Develop a random forest classifier with 100 trees, using the three best values of tree-depth you identified in the previous question. Provide all relevant results. Which tree-depth results in the best random forest classifier? How does it perform relative to the best decision tree?
5. Repeat this experiment with 50 trees. Provide all relevant results. Does your recommendation change?
6. We will now consider *k*-nearest neighbor models. Use 10-fold cross validation and identify the best value of *k*, by trying as many values of *k* as necessary. Keep in mind that very large values of *k* can affect speed, and that 5 is the default – try at least values from 5 to 10 (you can try more if you wish). Provide all relevant results. What value of *k* do you recommend? What is the accuracy associated with this value of *k*? If you had to select the three best values of *k*, what would they be?
7. Develop a logistic regression model using 10-fold cross validation. What is the associated accuracy?
8. Develop a logistic regression model on the *entire* training dataset. Provide the output. What is the model developed?
9. Using the four best models identified in each category (decision tree, random forest, *k*-nearest neighbor, logistic regression), perform an evaluation with 10-fold cross-validation. Your results should be similar (but not necessarily identical) to the results you have already obtained for these models. Across these four models, which one would you recommend, and why?
10. Use the entire dataset to develop a final version of the recommended model for testing. Provide all details of the model (and the tree if the recommended model is the decision tree). What is that accuracy of this model on the training dataset?
11. Read the file `dmtest.csv` and make predictions (using the final model) on which customers are likely to respond, and which are not. The predictions have to be 0 or 1 – if the model you selected naturally gives a probability score,

use 0.5 as the threshold to determine whether your prediction will be 0 or 1. For example, if you use a logistic regression model that gives you a probability estimate of 0.51, the prediction should be 1, and if it gives a probability estimate of 0.49, the prediction should be 0.

Create a file **group*xx*dmtest.csv** that adds a column named "prediction" to the original variables in `dmtest.csv` and submit it with your report. I will assess the quality of your predictions based on the actual values (which are not included in `dmtest.csv`).

12. If you were to focus on the "lapsing customers" (customers who made their last purchase 13 to 24 months ago), do you expect your model to be different? For the selected model, compare the quality of predictions for these customers relative to predictions for the others on records in the training set. Discuss your findings.