

## Assignment 4 – Machine Learning Models

### Submission Guidelines

Submit your answers as a PDF file. You do not need to be overly concerned with formatting for this deliverable. Just ensure your responses are clear, readable, and address the full elements of each question (you can use screenshots of your results). Please put your name in the document and also in the filename

Background Information: US book publishing industry is over \$28 billion annually. The Sierra Book Club was established in 1986 for the purpose of selling **specialty books** through direct marketing. SBC is strictly a distributor and does not publish any of the books it sells. In anticipation of using database marketing, SBC made a strategic decision right from the start to build and maintain a detailed database about its members containing all the relevant information about them. Readers fill out an insert and return it to SBC, which then enters the data into its database. The company currently has a database of over a million readers and sends out a mailing about once a month.

SBC is exploring whether to use predictive modeling approaches to improve the efficacy of its direct mail program. For a recent mailing, the company selected a random sample of customers from its database and mailed a brochure for the book **The Art History of Florence**. SBC then developed a database to calibrate a response model to identify the factors that influenced these purchases. For purposes of analysis, we will use a subset of the database available to SBC. The data for this exercise appears in the file “*Targeting.xlsx*”. The sheet called “Data” contains data on 1600 customers, 400 of whom purchased the book, and 1200 that did not. There are 2300 observations for holdout prediction. Customer IDs starting with “T” are part of the training sample, while those starting with “H” are part of the holdout sample.

Here is a description of the variables used for the analysis:

**Choice:** Whether the customer purchased **The Art History of Florence**. 1 corresponds to a purchase and 0 corresponds to a non-purchase.

**Gender:** 0 = Female and 1 = Male

**Amount purchased:** Total money spent on SBC books

**Frequency:** Purchase frequency

**Last\_purchase** (recency of purchase): Months since last purchase

**First\_purchase:** Months since first purchase

**P\_Child:** Number of children’s books purchased

**P\_Youth:** Number of youth books purchased

**P\_Cook:** Number of cookbooks purchased

**P\_DIY:** Number of do-it-yourself books purchased.

**P\_Art:** Number of art books purchased

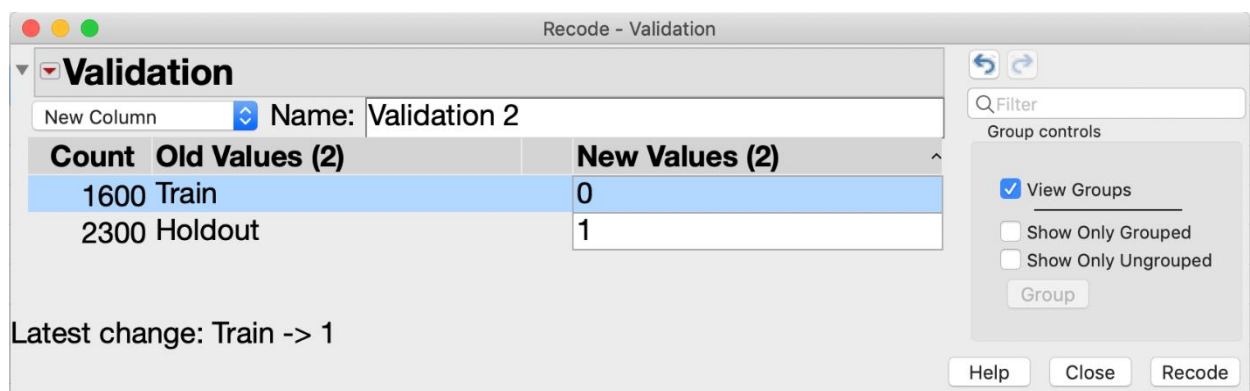
**Objective:** Sierra Book Club is considering a similar nationwide mail campaign for 1,000,000 customers. The allocated cost of the promotional mailing is \$0.65/addressee (including postage). The book costs Sierra \$15 to purchase and mail to a buyer. The company allocates overhead to each book at 45% of cost (i.e., 45% of \$15). The selling price of the book is \$31.95. Assume that the firm has decided to mail to only 10% of the 1,000,000 customers. (Note: All of this information is contained in the spreadsheet “Financial Calculations”).

*Question 1: Machine Learning (8 points)*

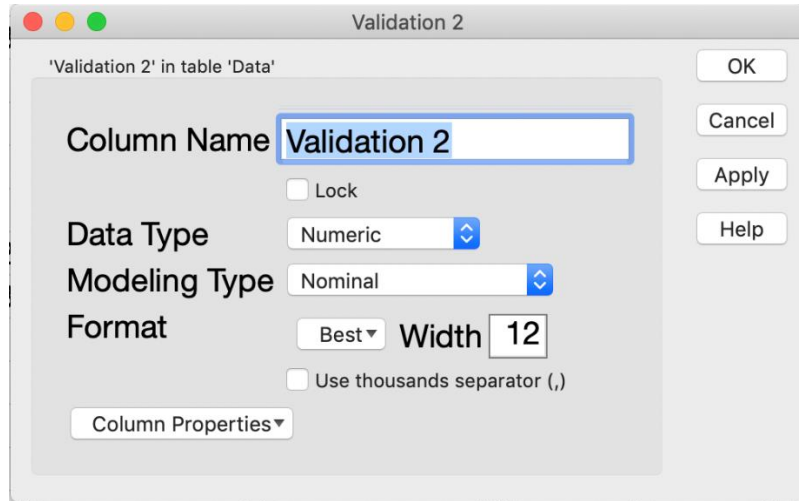
How much more profit would you expect the company to generate if they used a targeting model as compared to sending the mail offer to 10% of the entire list chosen at random?

Approach:

- 1) What % of customers bought (choice=1) in the “Train” and “Holdout” data?
  - a. Train \_\_\_\_\_
  - b. Holdout \_\_\_\_\_
  
- 2) Note: When running machine learning models (e.g., logistic regression; bootstrap forest, boosted trees, neural network models), the underlying Validation column has to be a “numeric” variable instead of a “character” variable. To convert the current Validation variable to a numeric one, do the following:
  - i. Click the Validation column (any cell).
  - ii. Click “Cols > Recode” and set the values according to the screenshot below. Click “Recode”.



- iii. Click the “Validation 2” column in the “Columns” pane on the left, right click and select “Column Info”
- iv. Set the “Data Type” to “Numeric” as shown below.



- v. Use “Validation 2” in all of your machine learning models
- 3) Run a Logistic Regression with Choice as the dependent variable and all other variables as predictors. Note that you have to either (1) Use “Validation 2” column in running the model, or (2) if you don’t have the latest JMP, “hide & exclude” the holdout sample before running your model. Share these results.
  - 4) Save “Probability formula” from your model.
  - 5) Create a subset for the **Holdout Sample**. Use the column Prob[1] to create deciles (10 groups) for the Probability. You can use the JMP Add-in “Interactive Binning”. Show your decile results in a bar chart and specifically, what is the response rate (percentage of Choice=1) in the Top Decile of the holdout sample?
    - a. Response in Top Decile\_\_\_\_\_

*Question 2: Financial Calculations (4 points)*

- 6) Note the numbers in 5 above and then go to Excel sheet “Financial Calculations”. Instructions are provided there (you will be filling in the shaded cells). What is the gain from Targeting?

Gain From Targeting\_\_\_\_\_

*Question 3: Model Comparison (8 points)*

Run the following models to predict “Choice” with all other variables as Predictors. Compare the model performance.

- a. Bootstrap Forest
- b. Boosted Trees
- c. Neural Network
- d. Elastic Net