DA 350 Advanced Methods in Data Analytics

Take Home Final Exam

May 07, 2018

Name (Print Neatly): _____

Rules of the Take Home Exam:

- You may download and begin the exam anytime between Monday, May 7 and Friday, May 11.
- Once you have downloaded the exam, it is due on Notebowl by midnight on the fifth day. For example, if you begin Monday, May 7 it is due Friday May 11 at midnight, and if you begin on Friday May 11 it is due Tuesday May 15 at midnight.
- I will check the time you downloaded the files from Notebowl to ensure it is turned in by the appropriate deadline. No extensions will be considered.
- You ARE permitted any outside resources, including the textbook, prepared R scripts, and google searches
- You ARE permitted to read statistics or coding questions on sites such as Stack Overflow
- You are NOT permitted to post your own questions to sites such as Stack Overflow
- You are NOT permitted to work with anyone else or discuss the exam with others
- If you have any questions, please contact the instructor for clarification

You are obligated to comply with the Denison Code of Academic Integrity. You are not allowed to receive or give aid on this examination, and you are subject to follow the rules written above.

Please write the following Honor Pledge at the bottom of your exam:
"I have neither given nor received aid on this examination",
and digitally sign by writing your name beneath.

The 'ProductSalesTraining.csv' file on Notebowl contains historical data on the sales of 1,579 new products for the company you work for. Your task is to use this data to generate predictions of how future products will sell to help guide manufacturing and marketing. The two outcomes of interest are 'Sales', the number of product sales within the first month of release, and 'Longevity', a binary variable indicating whether or not the product continues to be actively demanded 12 months later. The full data dictionary is on the following page. 'ProductSalesTest.csv' contains the test set of 674 products you will generate predictions on. Your tasks are the following:

(1) Use unsupervised methods to describe the relationships between the predictor variables. What patterns and trends do you find in the predictors and types of products?
(2) Use the relevant predictor variables to create predictions of the product sales for the test set. Provide your best estimate of the performance your method will have on the test set. Describe any insights you may discover about the relationship between the predictors and the response from your model.
(3) Use the relevant predictor variables to create predictions of whether the products in the test set will continue to be demanded for 12 months. Provide your best estimate of the performance your method will have on the test set. Describe any insights you may discover about the relationship between the predictors and the response from your model.

You should consider the methods we've learned and practiced over the semester. You should carefully consider how best to handle each variable, and some you may not use at all. It is at your discretion how to handle outliers. Explain your justification for the final models you chose, considering the pros and cons of each type of possible approach. You will turn in 4 items:

- ◇ Your R code with the methods you tried. It should be organized and commented enough that a reader can easily understand what the purpose of each section is. You should include the code for methods that you explored, even if it did not contribute to your final answers.
- ◇ A written document explaining your approach, a justification of why you chose those approach, your findings (including the performance of all methods you tried), and your best estimate of how your predictions will perform on the test set.
- ◇ A .csv named 'Regression predictions' of your predictions for (2) on the test set.
- ◇ A .csv named 'Classification predictions' of your predictions for (3) above on the test set. It should include two columns: one named 'Class Predictions' that are 0/1 predictions, and one named 'Probability Predictions' which are probabilities.

Your grade will derive from:
20 pts - Explaining your approach and what steps you took to arrive at your final answers.
10 pts - Justifying the use of your final models.
20 pts - The thoroughness and correctness of your description of predictor relationships in task (1).
30 pts - How close your test set predictions are in tasks (2) and (3) (compared to the best achievable predictions). The criteria for correctness will be Root Mean Squared Error for the regression problem, accuracy for the class predictions, and cross entropy for the probability predictions.
20 pts - Your estimate of your predictive performance on the test set in tasks (2) and (3).

Data Dictionary

The data columns consist of the following. Predictors are labeled (Q) or (C) to denote quantitative or categorical, accordingly. Unless otherwise noted, categorical predictors are 0/1.

- Sales - the continuous response. How many sales of the product are made in the first month.
- Longevity - the binary response. Whether a product lasts on the market 12 months or longer.
- Similar Sales (Q) - the revenue made in similar product sales.
- Prior demand (Q) - the demand for similar products in the prior month.
- Advertising budget (Q) - the money spent advertising the product.
- Similar Products (Q) - the number of similar products currently on the market from your company.
- Competitor Products (Q) - the number of similar products offered by competitors.
- Ads (Q) - the number of ads to promote the product.
- Competitor Ads (Q) - the number of ads run by competitors for similar products.
- Magazine (Q) - the number of magazine articles featuring your product.
- Awards (Q) - the number of consumer awards the product won.
- Innovative (C) - whether the product can be considered innovative.
- Product necessity (C) - whether or not the product can be considered a necessity.
- Type (C) - the type of product. Levels are Bath, Kitchen, Office, Toy.
- Cheap (C) - whether the product can be considered cheap.
- Disposable (C) - whether the product is disposable.
- Tested (C) - whether the product was tested with a focus group.
- Competitor Launch(C) - whether or not competitors launched a new similar product.
- Warranty(C) - whether the product has a warranty.
- Styles (C) - whether the product is available in multiple styles.