MIS 6334 Advanced BA
Project 2: Customer Analytics using Base SAS

Report due (before class starts): 11/27 for Section 002, 11/29 for Section 001

In this group project you will build a customer analytics model to analyze consumer purchase behavior using base SAS. Similar to Project 1, you will **start with a real-world raw dataset, and finish with managerial-relevant knowledge that you discover**. Unlike Project 1 where you can apply standard BA techniques, for this project you will have to build your own analytics model using SAS code. Therefore, this project serves two purposes: it integrates your knowledge learned in the second half of the semester regarding building customized models; and it offers hands-on experience for analytics implementation using base SAS.

Note that SAS codes you wrote in Homework 3 will be useful for this project. Therefore, **please finish your Homework 3 before you start on this project.**

Submit one report per group in eLearning under Assignments -- Project 2. Due time is right before class on the due date. **No late submission is acceptable**. The report should be **a single Microsoft Word document with your group number and all group member names.** The report should address all following project components. Messy reports will receive penalty. Many questions are open-ended, and within-group discussions on these questions are strongly encouraged.

The Dataset, Background Story, and Research Objectives

This project uses dataset "ABA_Project2_data_books.sas7bdat" that is attached to this assignment (zipped for ease of downloading). This dataset records customer purchases at two competing book sellers -- Amazon and BARNESandNOBLE -- in 2007. It also records customer demographics including education, income, race, household size (HHSZ), and more. See the Appendix on the last page for a detailed description of all fields in this dataset.

Suppose you are working for BARNESandNOBLE (BN in short) and would like to understand the factors that affect customer purchasing behavior at BN. In particular, you are interested in the following broad business questions:

- How to build a customized BA model to fit existing data and to make predictions?
- What types of customers tend to purchase a lot of books, and what types tend to make just few purchases? In other words, what consumer characteristics drive the difference?
- Why certain customers prefer Amazon over BN?

Your objective is to leverage the modeling skills you learned from this class to answer the above broad business questions. Below are the detailed instructions.

Part I. Modeling Count Data

1. Process the raw data using SAS to generate a count dataset in a format similar to the raw data in the "khakichinos.com" example. In other words, for each customer, count the number of books she **purchased from BN** in 2007, and keep the demographic variables. Report your code and print the first 10 records of this dataset.

2. For now ignore the demographic information, and run the **NBD Model**. Report your code and the MLE results (including the optimized LL value, all the estimated parameter values, and the according p-values – same requirement for all MLE estimations in this project)**.** (Hint: you will need to create a new dataset similar to the one on slide 5 in the count model lecture.)
3. Based on the NBD Model results, report Reach, Average Frequency and GRPs. Show your calculation.
4. Hereafter we will consider consumer demographic information. Run the **Poisson Regression Model** using the provided customer characteristics. Report your code and the MLE results. Which customer characteristics matter, i.e., what is your managerial takeaway? (Hint: should you "date" in this regression? Why?)
5. For the **NBD Regression Model**, what is the formula for LL? Write it down in your report. Getting this math formula clearly written will help your follow-up coding.
6. Run the **NBD Regression Model** using the provided customer characteristics. Report your code and the MLE results. Which customer characteristics matter, i.e., what is your managerial takeaway?
7. Any noticeable difference regarding the managerial takeaways between Poisson Regression and NBD Regression? If yes, what exactly is the difference? (Optional: Any thought on why the difference?)
8. Does NBD Regression fit the data better than Poisson Regression? (Hint: use the LR test – i.e., likelihood ratio test – on slide 29 in the count model lecture.)

Part II. Improving the Model

Please try to improve the **NBD Regression Model** using the following three methods. First, some hints:

- Note that not all things we try can improve the model – in case of no improvement, concisely write down why you think it didn't work.
- Since we are not using any validation dataset, the correct way to compare models is to use the LR test, which you can easily do manually with the LL values reported by SAS for each model and a Chi-squares table (see, e.g., https://www.medcalc.org/manual/chi-square-table.php, the "0.05" column).
- The following questions regarding model improvement are all open questions. For each question, just give a few tries and report your results and your thoughts.

9. Similar to what you found out in Project 1, not all variables are always useful. Please try feature selection (i.e. selecting only a subset of customer characteristics), and report your findings. (Hint: You can use Enterprise Miner to get some ideas on which variables to keep/remove, or, you can use the built-in variable selection mechanisms in SAS statistical procedures.)
10. You can also construct some variables on your own (e.g. convert date to weekday/weekend, or to holiday/non-holiday, or to seasons, construct percentage of weekend purchases, degree of loyalty to BN etc. -- totally your call and just try 2-3 ideas). Report your code (including the code for constructing the new variables) and the MLE results. Which newly constructed variables matter, i.e., what is your new managerial takeaway?
11. Researchers often try to improve a model by considering interaction effects (e.g., age*income) in the regression. Try 2-3 interaction effects you think are likely. Report your findings.

Part III. Why Certain Customers Prefer Amazon Over BN?

12. Now let's study why certain customers prefer Amazon over BN and vice versa. We will apply the concepts of a choice model – **logistic regression**. For each customer, you need to generate a binary

dependent variable indicating whether a user has made a purchase at BN (denote yes as 1 and 0 otherwise). Then use Proc Logistic to run a logistic regression model, report the results and your takeaways. (Optional: Using the data to answer this question: should you do variable selection?)

Part IV. Summary

13. Summarize what you learned from this project -- it can be key managerial insights you got, BA techniques or SAS skills you learned from this project, new perspective of BA you got by doing hands-on, or anything you feel worthwhile to summarize. Be concise.

Business Analytics is nothing but simply **transforming raw data to knowledge**. I hope this project benefits you in mastering the basic process/steps in building an advanced and customized BA model, and in giving you confidence that you can execute a complete BA initiative using base SAS.

Before submitting your reports through eLearning, please read the announcements at the beginning of this project handout to make sure you abide by all requirements.

Appendix – Dataset "ABA_PROJECT2_DATA_BOOKS"

No.	Variable	Description
1	userid	Unique user id at user computer level
2	education	Lowest – 0, highest – 5. Missing value – 99.
3	region	Four US regions 1 – 4. Missing value – *.
4	hhsz	House hold size.
5	age	Lowest – 1, highest – 11. Missing value – 99.
6	income	Lowest – 1, highest – 7.
7	child	1 – has child, 0 – does not have child
8	race	Four races 1, 2, 3 and 5
9	country	0 – US, 1 – non US.
10	domain	amazon.com or barnesandnoble.com
11	date	YYYYMMDD
12	product	Name of book bought
13	qty	Quantity purchased
14	price	Total price paid

Note about data fields: Many data fields are transformed in order to protect consumer privacy (done by the company providing this data). For example, age is shown from 1 to 11. Such transformation always preserves the order of numbers (e.g., value 8 for age implies a larger true age than value 7 for age).