The University of Texas at Dallas

School of Management

BUAN 6383/MIS 6386 — **Modeling for Business Analytics**

Syam Menon

Project II

Objective

- Learn to build advanced customized 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 **project-II-group*xx*.docx**
    (for example, group 1 should name the report project-II-group01.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 **project-II-group*xx*.ipynb**
    (for example, group 1 should name the notebook project-II-group01.ipynb)
  - If you prefer, you can submit separate jupyter notebooks for each question If you choose to do so, the files should be named **project-II-group*xx*-pyyq*zz*.ipynb** (for example, group 1 should name the notebook for Part I, question 1 project-II-group01-p01q01.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 project counts for 120 points**

Data Sets

- `candy.csv`
- `articles.csv`

Part I: Replicating Models from Class

1. Consider the hard candy example from class. The associated data is in the file `candy.csv`. Develop the following models discussed in class using maximum likelihood estimation (MLE):
   (a) the **Poisson model**,
   (b) the **NBD model**,
   (c) the **Zero Inflated NBD model**, and
   (d) **Finite Mixture models** for 2, 3, and 4 segments.

   Report your code and all relevant details, including the estimated values of the parameters for each model and the corresponding log-likelihood values. Please add comments to your code to make it easy to understand.

2. Evaluate the models developed; explain which of them is best, and why. Are there any significant differences among the results from these models? If so, what exactly are these differences? Discuss what you believe could be causing the differences.

3. Based on the 2, 3, and 4-segment finite mixture models, how many packs are the following customers likely to purchase over the next 8 weeks?

   (a) a customer who purchased 5 packs in the past week, and

   (b) a customer who purchased 9 packs in the past week.

Part II: Analysis of New Data

`articles.csv` contains the number of publications by 915 doctoral candidates (`articles`), along with five predictors:

1. `female`: 1 if candidate was female, 0 otherwise
2. `married`: 1 if candidate was married, 0 otherwise
3. `kids`: number of children aged ≤ 5
4. `prestige`: prestige of the candidate's department (higher is better)
5. `mentorpubs`: number of publications by the candidate's mentor over the past 3 years

Your task is to to predict the number of articles as a function of the five independent variables.

1. Estimate all relevant parameters for **Poisson regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. What are the managerial takeaways — which customer characteristics seem to be important?

2. Estimate all relevant parameters for **NBD Regression** using MLE. Report your code, the estimated parameters and the maximum value of the log-likelihood. What are the managerial takeaways — which customer characteristics seem to be important?

3. In this question, you will apply the ideas learned in this course to build a model that you have not seen before – **the Zero Inflated NBD Regression**.

   First, recall that zero inflated models view 0s as coming from 2 sources - (i) from a fraction π who is 0 "by type" (in the context of this problem, these are candidates who will never publish), and (ii) from the remaining fraction $(1 - \pi)$ who are likely to eventually become nonzero (these are candidates who will publish at some point, but have not done so yet). You can assume that the candidates in the latter group are distributed as a negative binomial (making the NBD regression appropriate for them).

   Explain the logic used in developing the model in detail. (**hint**: you do not need anything beyond what you have learned in the class to do this.)

   Report your code, the estimated parameters and the maximum value of the log-likelihood. What are the managerial takeaways — which customer characteristics seem to be important?

4. Evaluate the models developed; explain which of them is best, and why. Are there any significant differences among the results from these models? If so, what exactly are these differences? Discuss what you believe could be causing the differences.

As with project 01, briefly summarize what you learned from project 02. Remember – this is an open-ended question, so please include anything you found worthwhile.