Statistical Methods for Discrete Response, Time Series, and Panel Data (W271): Lab 2

W271 Instructional Team

October 8, 2017

Instructions:

- **Due Date: 10/23/2017**
- Submission:
  - Submit your own assignment via ISVC
  - Submit 2 files:
    1. A pdf file including the summary, the details of your analysis, and all the R codes used to produce the analysis. Please do not suppress the codes in your pdf file.
    2. R markdown file used to produce the pdf file
  - Each group only needs to submit one set of files
  - Use the following file naming convensation
    - SectionNumber_hw01_FirstNameLastNameFirstInitial.fileExtension
    - For example, if you are in Section 1 and have two students named John Smith and Jane Doe, you should name your file the following
      - Section1_hw01_JohnS_JaneD.Rmd
      - Section1_hw01_JohnS_JaneD.pdf
  - Although it sounds obvious, please write the name of each members of your group on page 1 of your report.
  - This lab can be completed in a group of up to 3 people. Each group only needs to make one submission. Although you can work by yourself, we encourage you to work in a group.
  - When working in a group, we encourage student not to use the "division-of-labor" approach to complete the lab. That is, do not divide the lab by having Student 1 completed questions 1 - 3, Student 2 completed questions 4 - 6, etc. Asking your teammates to do the questions for you is asking them take away your own opportunity to learn.
- Other general guidelines:
  - If you use R libraries and/or functions to conduct hypothesis tests not covered in this course, you will have to explain why the functions you use are appropriate for the hypothesis you are asked to test. Lacking explanations will result in a score of zero for the corresponding question.
  - Thoroughly analyze the given dataset. Detect any anomalies, including missing values, potential of top and/or bottom code, etc, in each of the variables.
  - Your report needs to include a comprehensive Exploratory Data Analysis (EDA) analysis, which includes both graphical and tabular analysis, as taught in this course. Output-dump (that is, graphs and tables that don't come with explanations) will result in a very low, if not zero, score.

  - Your analysis needs to be accompanied by detailed narrative. Remember, make sure your that when your audience (in this case, the professors and your classmates) can easily understand your your main conclusion and follow your the logic of your analysis. Note that just printing a bunch of graphs and model results, which we call "output dump", will likely receive a very low score.
  - Your rationale of any (EDA and modeling) decisions made in your modeling needs to be explained and supported with empirical evidence. Remember to use the insights generated from your EDA step to guide your modeling step, as we discussed in live sessions.
  - All the steps to arrive at your final model need to be shown and explained very clearly.
- Students are expected to act with regards to UC Berkeley Academic Integrity.

Description of the Business Problem, the Data, and Your Tasks

The file *lab2data.csv* summarizes a sample of the contributions received a private university. Information in each record in the sample includes graduating class (Class.Year), gender, marital status, major of studies when the alumnus attending the university (Major), whether or not the alumnus has attended any university events hosted by the Alumni organization between year 2012 and 2015 (AttendenceEvent), and the contribution in each of the years between 2012 and 2016 (FY12Giving, FY13Giving, etc). This is a carefully constructed sample, including only alumni who graduated from the institution and not the former students who spent time at the institution without graduating. Alumni not contributing have the entry "0" in the related column.

For a university foundation, it is very important to know who is contributing, because those information allows the foundation to target their fund-raising resources to those alumni who are likely to donate in the future.

In this lab, your group, as a team of data scientists working for the university foundation, are tasked to utilize the given information to predict who are likely to donate in the future. The data, *lab2data,csv*, contains recent historical information. You will need to build a model to predict the most recent (i.e. fiscal year 2016) contribution "category" using techniques covered in lecture 1 - 5.

The variable of interest is *FY16Giving*, which is a numeric variable. However, I'd like you to create another variable, named *FY16GivingCat*, representing various categories of contribution in 2016. The categories are $[0, 1)$, $[1, 100)$, $[100, 250)$, $[250, 500)$, $[500, 200000)$. Note that we specifically want to separate out those who did not contribute and put them in the $[0, 1)$ bin.

Even though I said "build a model", you are more than likely to experiment various model specifications as well as techniques. Some may consider using multinomial logistic regression, even though the categories are clearly ordered.

As in any data science project, start your project with examination of the data and then exploratory analysis. These analyses will help the administration of the university foundation to understand the sample (before you present any model results to them). In fact, your report should consider the following sections:

- Section 1: An introduction to the project, which should include a concise summary of the key results as well as techniques you used in your final model.
- Section 2: Data examination and EDA. This section should statr with a summary of the key insights you learn from examining the data and conducting the EDA. Since there will be a page limit (see below), select your graphical and tabular results carefully and accompany each one with narrative. **DO NOT USE OUTPUT DUMP!**
- Section 3: Statistical Modeling. Start the section summarizing the key results - what variables, if any, are the key predictors of the year 2016 contribution? What are the key techniques you have experimented? What method did you use in your final model? How did you choose the final model? What model performance criteria did you use to choose the final model? What statistical infernece did you perform? Explain them. Comment on statistical significance vs. economic significance.

- Section 4: Final Remarks. After examining the data and using the data to build a predictive model, what are your departing thoughts? What are the strengths and weaknesses in your analysis? Should the administration trust your result? Are there subsample in your sample that your model did a bad job in predicting their contribution behavior? If so, why? Are there other "things", a wish list, that you think can be used to improve your model? If so, what are they? Perhaps you can make a suggestion to the administration to collect those information in the future.