CIS-STA 3920 Project Assignment

Submit project on the designated Portal. Do not wait for the last minute to submit the project as it could happen that the Baruch server will freeze up and your work will then be late; 10 points per day late.

**Overall Guidance**

Showcase your ability to produce a well-crafted document, your capacity to learn how to do something challenging on your own, and your ability to engage with the reader and to explain what is happening. If asked for a letter of recommendation, I would hope I could without hesitation *include excerpts from your project as concrete evidence of your abilities and your integrity*.

* Submit a *single* Microsoft Word file as the project. Accompany that with a single R workspace file that you used to do your work.
* Write in the first person.
* Share some of your (a) thought-processes, (b) miscues, (c) work-arounds, and (d) insights. For style guide, review my lecture notes.
* Demonstrate engagement with the *concepts* and *approaches* taken in the Lecture Notes.
* Do not share your project work with anyone else; that is cheating.
* Do not plagiarize; do not plagiarize me.
* Number the questions in the same way as shown: A1, A2,...
* Answer each question at its numbered position. I will not look elsewhere.
* Number your pages.
* Do not show the text of my questions. Instead, *fold the question into the answer*. For example, instead of repeating my second question, you could say, "A.3. I am now going to work on plotting the classification space. The term *classification space* means..."
* Use a cover page with an interesting title and picture – the goal is to encourage the reader to turn to turn to the second page rather than giving up at a glance. Do not plagiarize the picture – either make your own image or *cite the source*.
* On the cover page, include your name, the course number and date.
* On the second page, place a well-formatted *Table of Contents*.
* For your filename, put your *last name* first.
* When referencing sources, use quotation marks and use footnotes to show the source. Do likewise when citing the source of text or images taken from the Web, books or articles.
* I require z**ero plagiarism**.

TO START: First, select a new classification methodology to learn in your project. Choose from either of the following chapters in the ISLR text:

(a) Chapter 8, pages 303-331, covers SVM (“support vector machines”);

(b) Chapter 9, pages 337-368, covers CART (“classification and regression trees”).

After making your choice, get introduced to it by walking through the lab example provided in the ISLR text. Show that work in an Appendix, including any graphs. To actually get the data used in the ISLR examples, you will likely need to download an R package called *ISLR*; it contains the data sets used in the text.

**Part A** Apply your chosen methodology to your stock risk data, as follows.

A1. [25 points] Work through your stock risk data at the level of detail used in ISLR. In Chapter 9, that means getting through boosting. In Chapter 8, that means getting through support vector machines. Show how well your new method forecast risk levels.

A2. [25 points] Select an aspect of the methodology that you find challenging and engage with it by reading and researching about it. For example, in Chap. 8, the Gini index (p. 312) is crucial to classification trees but is barely mentioned. What does it do, and how? Likewise, in Chap. 9, how is the parameter C (p. 346) controlled and what does it do? Demonstrate that you worked diligently on this problem: the goal is to engage in with the issue, not to produce miracles of comprehension or a plagiarism dump!

A3. [25 points] Create a classification space plot using your chosen methodology. Be sure to explain how you went about this. Create the plot using the same techniques that we did in our plots for other methods; that is, work it out yourself, step by step. If your method comes with a pre-canned classification space routine, you can show as an alternative method, but that will not by itself address this question. Again, remember that the primary goal is improve your learning skills by working diligently on this yourself: that demonstration is more important than getting your plot to work!

**Part B** [25 points] Select between 8 and 12 variables from the Cardiac data set, which is the NIH data set I used in LN8. Create a new y-variable by mapping one of the continuous variables to 0-1. Do that by mapping values less than or equal to the median to 0 and the rest to 1. Then, do a comparative study of the performance of knn, naive Bayes, logistic regression (with bestglm) and your method from above.