Instruction for DSBI Group Project

Think of yourselves as analysts employed by or retained by a company (large or small) or by a funding source (e.g., a VC firm or incubator), who wants to understand the state of the art for using data science for the task in question.

You will use *R* and *Tableau* to analyze and visualize data for a business problem of interest. These could be data from a problem from your current job, something of interest to the school, data acquired from the web, etc. You will design the data science task, analyze the data, and describe your results. Your own data and results need not be on par with actual industry results; the goal is for you to get as realistic a hands-on experience as possible, given the constraints of what you have learned. Don't worry too much about coming up with a novel idea. It is more important to develop the idea well (within the scope of what we've discussed in class). You must choose a classification problem (will talk about this in week2).

You should use the "data science process" to structure your research and write-up. Keep in mind that it may be ineffective simply to proceed linearly through the steps, and this may need to be reflected in your analysis. You should interact with me from the preparation of your initial ideas through your write-up, as a consulting group would interact with a firm or funding source in preparing a research report. Use your imagination, prior experience, or ask me to help to fill in any gaps between the material available and what you would be able to find out if you actually could interact with the client firm.

Deliverable #1: In Week 3, you will submit a **proposal** for your project. This should give as much detail as possible your ideas, so that I can give you feedback. Include in your proposal your ideas about: What is the exact business problem? How will the result be used? What is a data instance (i.e. what does each data point represent? a customer? a county? a product?) What might be the target variable (i.e. the variable of interest to be predicted)? What features would be useful? How exactly would it add business value? **A maximum one page Word document.**

Deliverable #2: In Week 6, you will submit a **status report**, including preliminary results (e.g. summary statistics, some visualization of data) or issues that you are facing in developing your project.

Deliverable #3: In Week 8, you will submit your **final report**, which should include the information detailed in the next page, in approximately the order given. Your write-up need not have corresponding sections or bullet points, but I should be able to find the information without searching too hard. The write-up should be **maximum 10 double-spaced pages Word document**, plus any appendices you would like to include. Use external sources where appropriate, and provide clear citations and bibliography. All group members should contribute to the analysis and write-up.

Presentation: In Week 8, you will present the results of your research to the class. Each team will be given 15-20 minutes for the presentation (depending on the class size) including Q&A. Going over the time will be negatively reflected in your grade.

Business Understanding – 15 points

- Identify, define, and motivate the business problem that you are addressing.
- How (precisely) will a data science solution address the business problem?
- Review what has been done to date on your problem.

Data Understanding & visualization – 20 points
- Identify and describe the data (and data sources) to address the business problem.
- Specify how these data are integrated to produce the format required for data science.
- Do other interesting visualizations
- Discover interesting relationships or phenomena

Modeling – 15 points

- Specify the type of model(s) built and/or patterns mined.
- Discuss choices for data science algorithm: what are alternatives, and what are the pros and cons?
- Interpret the results. What are the patterns discovered through data science? Do the results make sense?

Evaluation – 15 points

- Assess the performance of your model. How good is your model?
- Discuss how the result of the data science is/should be evaluated. How should a business case be developed to project expected improvement?
- If this is impossible/very difficult, explain why and identify any viable alternatives.

Deployment – 15 points

- Discuss how the result of the data science will be deployed.
- Discuss any issues the firm should be aware of regarding deployment.
- Are there important ethical considerations?
- Identify the risks associated with your proposed plan and how you would mitigate them.

Presentation/Communication – 20 points

- Present to the class the results of your research
- Clearly understand your own project contents and process, as well as limitation
- This portion also considers the overall quality of writing of the final report.

Each team member will submit a peer evaluation form separately. If all team members contribute to the project equally, they will share the same grade for the project. Otherwise, the peer evaluation results will affect your final score for the project.
All members of a team must have some face time during presentation.