Intro to Data Science

Prof. Dalessandro

DS-GA-1001

Fall 2016

Instructions for Term Project

For your term project, you will use PYTHON (and any additional necessary tools) to build a model for a problem of interest. The data can be from a problem at your current job, something of interest to the school, data acquired from the web, etc. You will design the data mining task, build the models, and describe your results. You also will research existing solutions to the problem, if any have been proposed or documented. 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. Your project will demonstrate appropriate understanding of the data mining process, including: problem formulation, data prep/understanding, modeling and an appropriate discussion on implementation.

In writing up/presenting your research, think of yourselves as analysts employed by or retained by a company (large or small), by a non-profit with a certain objective, or by a funding source (e.g., a VC firm or incubator), who wants to understand the state of the art for using data mining for the task in question. Review what has been done to date on your problem. Take as an example data mining for on-line advertising, which we will discuss in class. A VC firm considering funding on-line ad networks or ad-tech startups would need to understand the state of the art in using data mining for targeting on-line advertising, when considering an idea for applying data mining. 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). Purely conceptual projects should be as comprehensive as possible in analyzing the business problem and potential data mining solution.

You should use the "data mining 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 and the section leader 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 us 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.

IMPORTANT ADVICE: Reach out to me regularly and frequently to make sure I can help if things seem to get stuck.

It is possible that you realize somewhere down the road that the data is not supporting what you want to do. There is a fine line between being able to anticipate this (I will let you know if I suspect issues) and things that do not work out just because. You are not being judged on the performance of the model and in particular NOT being able to predict very well is OK. Not finding evidence for a hypothesis is fine too!

Finally, you are free to take certain liberties both with the data and the business setup. Be creative! You can pretend to have less data than you actually got and you can invent a business problem, but you have to create a convincing case for your problem and solution.

Deliverable #1: On **October 19th** you will submit your choice of team and initial ideas for projects. Teams will comprise of 3-4 students. Remember when choosing a team, search for diversity. Look for a good mixture of technical understanding as well as business understanding. Initial ideas can simply be a few sentences about what you are thinking you might do. When submitting, please email me directly (briand@gmail.com: Subject 'Term Project 2016 Deliverable 1'). In one email, give the members of the teams (name and netid), a team name, and the ideas.

Deliverable #2: On **October 26nd** you should have by now at least one lead for a dataset. Otherwise you will have major problems with the proposal for next week. Just tell me in 2 sentences the state of your hunt for datasets. An email submission is fine.

Deliverable #3: On **November 2nd** you will present me with a **proposal** for your project. This should give as much detail as possible on your ideas, so that I can give you feedback. At this point you should have your data. Include in your proposal your ideas about: What is the exact business problem? What is the use scenario? What precisely is the supervised data mining problem? What is a data instance? What might be the target variable? What features would be useful? How exactly would it add business value? Etc. This should be about 2 paragraphs and email submission is preferred (5%)

Deliverable #4: On **November 18th** you will present me with a status report, including preliminary modeling results or issues that you are facing in developing your project. This is your last chance to change course in case something is not working out. This should be 1-2 pages and emailed with an attachment (Word or PDF) (15%).

Deliverable #5: On **December 9th**, your final write-up should include the information detailed on the next/back of this 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. Be as precise/specific as you can. The write-up should be about 10 double-spaced pages, 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. The report should include an appendix describing the contributions of each team member. (80%)

You will get the most out of the project if you interact with me during the development of your ideas. Talk to me especially before choosing one of the business problems we cover in class (see the syllabus). And please feel free to come talk to me about your ideas as often as you'd like. ***Please do not choose stock/index prediction or market forecasting (talk to me).***

Your write-up should include all of the following elements:

Business Understanding

- Identify and motivate the business problem that you are addressing.
- How (precisely) will a data mining solution address the business problem?

Data Understanding

.

- Identify and describe the data (and data sources) that will support data mining to address the business problem. Include those aspects of the data that we routinely talk about in class and/or in the homework.

Data Preparation

- Specify how these data are integrated to produce the format required for data mining.
- Give a clear and precise definition of the target variable.
- Make a summary of any feature engineering that should be performed, which may include binning, non-linear transformations and domain knowledge based feature extraction.

Modeling & Evaluation

- Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
- Identify an appropriate baseline model and report its performance.
- Describe an evaluation framework you will use to improve upon the baseline.
- Perform an analysis of possible algorithms and use the data science experimental framework to choose an optimal candidate.
- Demonstrate how you were able to improve upon the baseline and document the process of doing so.
- Discuss why and how this model should "solve" the business problem (i.e., improve along some dimension of interest to the firm).
- Discuss the type of evaluation metric that should be used to choose the best algorithm. How does this metric relate to the business problem?

Deployment

- Discuss how the result of the data mining will be deployed.
- Discuss how it should be monitored and evaluated in an actual production system.
- 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.
