Intro to Data Science

Prof. Dalessandro

DS-GA-1001

Fall 2020

Instructions for Term Project

For your term project, you will use Python, and any additional necessary tools (i.e., Spark) to build a supervised learning 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 science 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 science process, including: problem formulation, data prep/understanding, modeling and an appropriate discussion on implementation and ethical considerations.

In writing up your research, think of yourselves as a data scientist 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 machine learning for the task in question. Review what has been done to date on your problem. Take as an example machine learning for on-line advertising, which will be discussed 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, but going beyond what we learn in class is always welcome).

You should use the "data science process" as well as agile development to structure your research and write-up. After framing the problem and exploring the data, this means identifying a suitable baseline model and testing that as a starting point. Your research plan should be hypothesis driven and will seek out ways to improve upon your baseline. This process, and the progress along the way, should be well documented as part of the write-up. And note – it is the design, execution and presentation of this process that I care about, more so than final model performance.

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.

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.

<u>Pre - deliverable:</u> By **11:59 PM EST October 8th** you should have 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 skills as well as business acumen. If by this time you do not have a team, please reach out to me. Also, because of the size of class and constraints on grading, I require each team to have at least 3 and at most 4.

Deliverable #1 (5% of grade): By **11:59 EST October 22nd** 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. For your submission, email me directly at briand@gmail.com, with the subject: "Term Project 2020 Deliverable 1." In the email, I need the name and netid of each team member, a team name, and the above proposal. **I also want the proposal to be included in the body of the email**, and not as an attachment (it is easier to give feedback on a line by line basis this way).

Check in: By **November 12th** you should have the data, you should have read it in and explored it, and should have a baseline model. I generally expect you to stick to your proposal, but you can change course by this date if you find that the problem or data is simply intractable. I don't require that you send me an update, but if you want to change course please send me a check-in note so that you can get my feedback.

Final Write-Up (95% of grade): By **7:00 AM EST December 3rd:** your final write-up should include the information detailed on the later in this document, in approximately the order given. Your write-up need not have corresponding sections (though to keep your structure easy to manage, I recommend it), 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 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.

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 machine learning solution address the business problem?
- What is the size / magnitude of the opportunity?
- What is the current state of the art in solving the 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. (note – do not show EDA plots for the sake of showing them. Do the EDA, and report out on any interesting observations that motivate the decisions you make in how you handle or model the data).

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 and show the distribution 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.