Introduction to Data Science
Prof. Spirling
DS-GA-1001
Spring 2017

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 something connected with your course of study, something you obtain from a data service at the university, something you 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, by a non-profit with a certain objective, or a government organization, 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 type of problem. 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 weve discussed in class).

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. 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.

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 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 perfectly fine. Not finding evidence for a hypothesis is fine too!

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

Deliverables and Dates

Done

1. March 10: submit your choice of team and the basic outline of an idea you would like to work on (this can be one sentence!). Teams will comprise 3 students. It would be helpful if you choose a team that's diverse: people with different types of skills often make great partners. When submitting the names, email our TA (using his official NYU account) with the subject '[IDS] Team Project names'. In one email, give the members of the teams (name and netid),

a team name, and the idea. If you fail to turn in a team at this point, you will be put in a random team of (up to) three people.

Done

2. March 31: submit your proposal for the project. This should be approximately two paragraphs. It should answer the following questions: what is the (business or research) problem? Is it supervised or unsupervised? What is the data, and where will you get it? What is an instance in the data? What is the target variable (if there is one)? What are the attributes of interest?

3. April 21: Status report. This is one page, and includes any preliminary data or modeling results. It also includes any issues you are facing in your project. This is your last chance to change course in case something is not working out.

4. May 9: The final write-up is due. It 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 we should be able to find the information without searching too hard. Be as precise/specific as you can. The write-up should be no longer than 8 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.

Final Write-Up Structure

Your write-up should generally include the following elements. It doesn't need to cover everything in exactly the way laid out here, but this should help you structure your efforts.

Business Understanding

- Identify and motivate the business/industry/government 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 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 (if appropriate)
- 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 and Evaluation

- Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
- If you are undertaking a supervised problem, then you should try to...
  - 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 central problem (i.e., improve along some dimension of interest to the organization).
- 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 organization should be aware of regarding deployment.
- Are there any important ethical considerations?
- Identify any risks associated with your proposed plan and how you would mitigate them.