NEW YORK UNIVERSITY STERN
Leonard N. Stern School of Business

DATA MINING FOR BIZ ANALYTICS
Spring 2019

Instructions for term project

Prof. George Valkanas

For your term project: (see the syllabus for the due dates for the deliverables)

The goal is for you to get as realistic a hands-on experience as possible of Data Mining projects and what they involve. To that end, you need to address all aspects of the "data mining process" and your term project should reflect that.

Focusing on a domain and problem of your interest, you will design the data mining task that solves the problem, mine the data as proposed and describe your results. All of that information (domain, problem, data) may be related to your current or previous job(s), something of interest to the school or to an (objectively) large audience, or something that you find interesting and could support a business model. Data mining projects with the public's interest in mind ("social good" projects) are very welcome too.

In writing up/presenting your research, think of yourselves as data scientists employed by or retained by a company or organization (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 mining for the task in question. This means that as part of your work, you will also research existing solutions to the problem, if any have been proposed or documented, including ones that do not rely on data mining. Consider as an example predictive analytics for on-line advertising: 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 targeted 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). Also, your own data and (data mining) results need not be on par with actual industry results. However, your state-of-the-art review should be as comprehensive as possible (re-review the above paragraph for details).

It's a good idea to use the "data mining process" to structure your research and write-up. Keep in mind that it may be ineffective to just linearly proceed through the steps, and this may need to be reflected in your analysis. You should interact with me and the course assistants 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 could find out if you actually could interact with the client firm.

Deliverable #1: By **Feb 10th** you will submit your choices for teams and initial ideas for projects. Teams will comprise 3-4 students. Initial ideas can simply be a few sentences about what you are thinking you might do. It is a good idea, though not a requirement, to form teams where the members have complementary skills (e.g., ideation, communication, technical, etc), but other factors may also come into play (e.g., past project success, etc). You are free to choose your teammates among classmates from your section, but should do so responsibly: you will be working together for the entire semester. If you need help finding a team let us know.

Deliverable #2: By **March 10th** you will present me with a **proposal** for your project. This should include as much detail as possible, so that I can give you feedback. The more detail you provide, the better feedback I can give. For example, your proposal should, at least, address the following: What is the exact (business) problem? What is the use scenario? What precisely is the data mining problem? Is it supervised or unsupervised? What might be the target variable (if supervised)? What features would be useful to solve the problem? Where would the data come from in the real setting? How exactly would the "technical" solution solve the business problem? Etc.

Deliverable #3: By **Fri April 14th** you will present me with a status report ("Project Update"), including preliminary results and any issues that you are facing in developing your project.

Deliverable #4: As your final deliverable (**TBA**), you will submit the full write-up that should include the information detailed below, in *approximately* the order given. Your write-up need not have corresponding sections or bullet points, but I should be able to find all of the information rather easily. Be as precise/specific as you can. The write-up should be about 20 double-spaced pages, plus any appendices you would like to include. Use external sources where appropriate, and provide

clear citations and bibliography. In addition to the write-up and as part of that final deliverable, you will also submit your technical work (code, data) that supports your results. Finally, all group members should contribute to the analysis and write-up. The report should include an appendix describing the contributions of each team member.

You will get the most out of the project if you interact with me and the CA during the development of your ideas. Talk to me especially before choosing one of the business problems we cover in class. And please feel free to 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 project write-up should include the information detailed below, in approximately the order given. Your write-up need not have the corresponding sections or bullet points, but I should be able to find all of the information rather easily. Be as precise/specific as you can.

Business Understanding (take this seriously)

- Identify, define, and motivate the business problem that you are addressing. Discuss existing approaches, if any.
- How (precisely) will a data mining solution address the business problem?

(Note: I'd like to see a good definition/motivation of the business problem and a precise statement of how a data mining solution will address the problem. It's not so important that the hands-on results match perfectly. It's more important that you have the experience of working through a realistic problem definition.)

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 homeworks.

Data Preparation

- Specify how these data are integrated to produce the format required for data mining.

(Note: data preparation can be time consuming! Get started early. Talk to the CAs or Prof if you need advice.)

Modeling

- Specify the type of model(s) built and/or patterns mined.
- Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
- Discuss why and how this model should "solve" the business problem (i.e., improve along some dimension of interest to the firm).

Evaluation

- Discuss how the result of the data mining is/should be evaluated. How should a business case be developed to project expected improvement? ROI? If this is impossible/very difficult, explain why and identify viable alternatives.

Deployment

- Discuss how the result of the data mining 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.
