DUKE
FUQUA
SCHOOL OF BUSINESS

Instruction for Term Project

Your team will mine actual data for a problem of interest. These could be data from a problem from your current/previous job, something of interest to the school, data acquired from the web, etc. You will describe a business application in which data mining can make a positive impact. You will design the data mining task, mine the data, 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.)

In writing up/presenting your research, 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 mining for the task in question. Review what has been done to date on your problem. 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 targeting on-line advertising, when considering an idea for applying data mining. Do not worry too much about coming up with a novel idea. (Specifics of the business can many times make standard data mining tools develop new insights.) It is more important to develop the idea well (within the scope of what we've discussed in class).

You should use the "data mining process" to structure your research and write up (see last page). 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 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 would be able to find out if you actually could interact with the client firm.

It is important to realize that a well structure business application will motivate several different data mining tasks. Although I expect you to discuss some of them in the project, due to time constraints and data availability, I do not expect a team to attempt all of them. However, a successful project is expected to conduct (at least) two related data mining tasks:

- data visualization, to effectively explain and communicate facts (e.g., plots, clustering, networks);
- build and evaluate a predictive/causal model (e.g. regression, Lasso, text mining)

Keep in mind that your data must be rich enough to allow for these tasks. Typically the data would allow you to use many of the topics we discussed in class (but unlikely to be applicable to all of them). Finally, make sure that the data is compatible with the use in your application. **(Note: There are no requirements to do the project using a specific software. Feel free to use the tool that works better for each stage of the project. Using multiple tools can potentially help you. R covers all topics but you might be more familiar with other software (like Excel) for specific tasks that you want to use.)**

Schedule for Term Project Deliverables

Deliverable #1: **Day Before Class 4**, you will submit a **proposal** for your project. (**Submission should be made in pdf format following the guidelines in the syllabus.** A single page suffices but no strict page limit. Additional files can be submitted with support material.) This should give as much detail as possible of your ideas, so that I can give you feedback. Include in your proposal your ideas about: What is the exact business problem? What is the use scenario? Which is the data source? What is a data instance/unit? What might be the target variable? What features would be useful? What precisely is the data mining problem? Is it supervised or unsupervised? How exactly would it add business value? Etc.

Deliverable #2: **Day Before Class 8**, you will submit a status report, including preliminary results or issues that you are facing in developing your project. (**If you have doubts about the project talk to me during office hours**, do not make questions on the Project Update and definitely do not wait for the feedback of the Project Update.) At this stage, **I expect the team to have at least cleaned up the data**. Keep in mind that this can be time consuming. You will also **specify the "Industry"** for which the project belongs, e.g. Consulting, Investment Banking, General Management, Corporate Finance, Healthcare, Technology, Marketing, Operations, Energy, Media & Entertainment, Private Equity, Public/Government Policy, Sports. If you feel your project does not fall into any of those, please specify. (**Submission should be made in pdf format following the guidelines in the syllabus.** Two pages suffice but no strict page limit. Additional files can be submitted with support material.)

Deliverable #3: **Two days before Class 12,** you will **submit your final write-up**, which 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 and specific as you can.

The write-up should be up to 10 double-spaced pages.

Any appendices with additional material you would like to submit can be submitted via a separate file. Make the main write-up self-contained (do not expect people to look at the appendices). (**Submission should be made in pdf format following the guidelines in the syllabus.**) 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.

By this date you should also submit your data, your (well documented) code, and slide presentation (ppt, in the standard slide format as presentations will be combined).

Deliverable #4: **During Class 12**, you will present to the class the results of your research. I will give you a time limit for your presentation beforehand, and your presentation will be expected to remain within the time limit. Please keep in mind that this is a very important skill to master: if a VC or a corporate officer/board member tells you she'll give you 5 minutes to present your idea or proposal**, you present your proposal in 5 minutes -- not 7 or 10.** Going over the time by more than one minute will be reflected negatively in your grade, but I'll warn you when you're getting close.

Note: 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. And please feel free to talk to me about your ideas.

Your write-up 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 the information without searching too hard. Be as precise/specific as you can.

Business Understanding (take this seriously)

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

(NB: 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 for the results to deliver highly precise predictions (since that depends on the quality of the data and we would have more time in practice). It's more important that you have the experience of working through a realistic problem definition, outline the whole path to get to the solution, and provide evidence you are approaching the data mining tasks properly.

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 assignments. (If appropriate highlight potential bias, full disclosure always better, and identify potential bias directions.)

Data Preparation

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

(NB: data preparation can be time consuming. Get started early. Talk to the TAs 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. Provide good measures of the performance of predictive models. How should a business case be developed to project expected improvement? ROI? If this is impossible/very difficult, explain why and identify any 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.