Final Project : Statistical Learning and Analytics

The main objective of the project is to apply data mining or machine learning algorithms to analyze a business problem. Students are expected to use methods reviewed in class and present different scenarios with their recommendations.

Project description:

Build a decision support system in your current research/interest area using any technique reviewed in this course that you consider appropriate for the problem studied. If you propose a different methodology, this should be approved by the instructor.

You should prepare this project as if you were consultants or 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. Don't worry too much about coming up with a novel idea. It is more important to develop the idea well.

You should write a program that implements the solution proposed using Python, Java using Weka, C++ or R. I recommend to use Python as it combines the strength of a programming language with powerful machine learning and data analysis libraries.

Make sure to show what novel work you have done for this course, you can't just submit previous work. Each team should have 3 persons.

Project Proposal (only 1 group representative should submit the project proposal). The project proposal should be typed, not hand written, and should include:

1. Problem: What is the exact business problem? What is the use scenario? What precisely is the data mining problem? Is it supervised or unsupervised?
2. Solution: what is the solution proposed? (high level description), which forecasting algorithms do you think are appropriate for this problem domain and why?
3. Programming language.

4. Performance evaluation: How would one test the performance of the algorithm to be used?
5. Data: What is the data to be used? What might be the target variable? What features would be useful?
6. Impact: How exactly would it add business value?
7. Names of all team members on the proposal.

Project Report (only 1 group representative should submit the project)

The final report must include the following sections of the “data mining process” reviewed in class:

Problem or Business Understanding:

- Identify, define, 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 homeworks.

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)

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 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.

Be as precise/specific as you can. The write-up should be about 10 (max. 15) 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.

1. The following paper is a case study that includes the main sections of the CRISP-DM methodology: Richard Adderley and Peter Musgrove, Data mining case study: modeling the behavior of offenders who commit serious sexual assaults, KDD 2001 (in course website).

2. Academic: PhD students should prepare an academic paper that counts as the final project for this course. The above paper is a good example of a potential academic paper, however this should be oriented to conferences such as "Innovative Applications of Artificial Intelligence Conference" or “International Conference on Information Systems.” The paper should be based on a theoretical or applied exploration of one of the methods studied in this course or any other data analysis method approved by the instructor.
3. The following paper could be useful for the preparation of the report as it describes several statistical tests used to compare various learning algorithms: Dietterich, T. G., (1998). Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Computation, 10 (7) 1895-1924. Postscript preprint. (Revised December 30, 1997).

Files that you must include in your final submission:

1. Project report
2. Source code
3. Readme file explaining how to operate your system
4. Output file showing some results of your program.

Project Demo

Each project will have one mandatory demonstration/discussion at the completion of the project (last days of this course). This will involve a poster presentation (3-5 minutes) of your project to the class followed by a discussion of your work and a demonstration of the project. The purpose of this presentation is two-fold:
1) it allows you to highlight portions of your project that may not be adequately shown in your write-up and
2) it allows the instructor to understand how deeply you understand the material and your own work.

Appendix:

CRISP-DM:
Description of data mining and CRISP-DM methodology (pgs. 272-276 and 280-289)
"The CRISP-DM model" by C. Shearer. (pgs. 13-22)
CRISP-DM User guide. This is an IBM guide prepared for its SPSS software and CRISP-DM. Only look at CRISP-DM's conceptual aspects as we are not using SPSS for this course.

You can also look at the following links for ideas or tutorials that could be useful for your project:

http://www-2.cs.cmu.edu/afs/cs.cmu.edu/project/ai-repository/ai/areas/0.html This is a repository about different AI software packages.

http://satirist.org/learn-game/links/tutorial.html Tutorials for machine learning methods (not all links work).

http://www1.cs.columbia.edu/~jebara/4771/PROJECT.htm This link has good project and format suggestions for machine learning papers.

Datasets and resources linked through Kdnuggets.com:
• Government, Federal, State, City, Local and public data sites and portals.
• Data APIs, Hubs, Marketplaces, Platforms, Portals, and Search Engines.
• Free Public Datasets.

Competitions: the team may also participate in a competition such as kaggle as long as the deadline is near or after the project’s deadline.