3251 TERM PROJECT

The Project

The 3251 Term Project is your opportunity to showcase what you've learned in the course. You have a choice of two alternative activities.

Choose one of the following for your project:

- Make a hypothesis about a correlation in a dataset and test the hypothesis using a statistical inference technique (such as the t-test)
- Build a predictive model using one of the techniques covered in the course i.e. ordinary least squares regression or Naïve Bayes

Data

The data you use for your analysis should be real (i.e. not randomly generated). The following are links to a variety of interesting datasets. The dataset you use does not need to be from this list. You can use any source as long as it is open data or you have written approval to use it.

https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/
http://www1.toronto.ca/wps/portal/contentonly?vgnextoid=9e56e03bb8d1e310VgnVCM10000071d60f89RCRD
https://www.ontario.ca/search/data-catalogue?sort=asc
http://open.canada.ca/en/open-data
http://blog.yhat.com/posts/7-funny-datasets.html
http://koaning.io/fun-datasets.html

Topic Approval

Prior instructor approval of your choice of project is recommended but not required. Please submit your choice of project to the instructor via email for approval if you would like feedback on its scope and feasibility.

Requirements & Due Dates

You must submit two documents, a report and a presentation. The report is due the second-last class of the term and must be submitted on paper unless you are unable to make it to class. The presentation is due the day before you present, which may be the third-last or second-last class depending on scheduling. (The last class is the exam and your marked paper will be returned to you then).

Report

The report should be organized into the following four sections and should be of a quality that is suitable for presentation to senior management:

1. **Objective**: What are you setting out to prove or predict? What is your rationale for there being a correlation in the data that you're looking to confirm and/or exploit?
2. **Data Preparation**: What was your data source (e.g. web scraping, corporate data, a standard machine learning data set, open data, etc.)? How good was the data quality? What did you need to do to procure it? What tools or code did you need to use to prepare it for analysis? What challenges did you face?
3. **Analysis or Model**: If you're conducting an inference test explain the analysis you performed clearly and include well-labelled diagrams to make your points. If you chose to do a predictive model, explain the model, how you trained and tested it, and how well it works. How did you confirm that the data met the requirements for the test or modeling technique to be valid?
4. **Conclusions**: Did you prove/disprove your hypothesis or create a useful model? What did you learn about your data set?

The report should be about 8-10 pages in length. If there is additional information e.g. code, samples of data, etc. this should be in an appendix.

Presentation

The presentation should be five PowerPoint slides long (excluding title slide), one for each of the four report sections plus one for additional background or diagrams at your discretion. You will only have five minutes to present, which will go by quickly, so please rehearse and time your presentation in advance.

The presentation is due via email the day before you present so we won't lose time in class (for example, while USB drivers install).

Marking Scheme

Marks will be allocated as follows for a total out of 30:

- Suitability as a report and presentation to management – 10 marks
    - Spelling
    - Explanation of use of technical terms
    - Formatting
    - Easy to follow
- Statistical correctness – 10 marks
    - Use of appropriate techniques
    - Checking that the data meets any assumptions the model or test requires

    - Correct interpretation
- Plausibility of conclusions – 5 marks
- Novelty – 5 marks
    - Is this an interesting and different analysis?