Business Data Analytics

You will use R to mine actual data for a problem of interest. These could be data from a problem from your current job if you have one or gov data <https://www.gov.uk/search/research-and-statistics> or <https://archive.ics.uci.edu/ml/datasets.php>

**Report (1500 words plus code):** the report involves use of the CRISP-DM to carry out a data analytics project on data of the student’s choice; data modelling is carried out using

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 a 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 yourself as an analyst employed by or retained by a company (large or small) or by a funding source (e.g., a venture capital (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 start-ups 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 (within the scope of what we’ve discussed in class).

You should use the CRISP-DM data mining process to structure your research and report. 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 from the preparation of your initial ideas through to the preparation of your report, as a consultant would interact with a firm or funding source in preparing a research report. Use your imagination, prior experience, or ask for 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.

This assignment will have a phased submission of work, as follows:

Give brief answers. What is the exact business problem? What precisely is the data mining problem? Is it supervised or unsupervised? What data will you be using and where will you obtain it? What is a data instance? What might be the target variable? What features/variables would be useful? How exactly would it add business value? And so on. **Please include a link to the data set you will be using – this is absolutely vital**.

Task 2: Should be about 1500 words, plus any appendices you would like to include. Use external sources where appropriate and provide clear citations and bibliography. You should also submit your **data file** and a **working R script** which I can run against it.

**You should not copy code from a source which is working with the same dataset that you are using**

**Your report should include the information detailed below, in approximately the order given. 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 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 talk about in class and/or in the quizzes. This should include some exploration of the data, such as:

o Summary statistics o Visualisation using graphs/charts

**Data Preparation**

* Specify how these data are integrated and prepared to produce the format required for data mining. *(NB: data preparation can be time consuming. Get started early. Talk to me if you need advice.)*

**Modelling**

* 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? How did you evaluate each of the models you used? Which is the best-performing model?
* 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.

**MARKING CRITERIA**

The submitted and assessed part of this coursework is a report together with R code and data files, rather than an academic essay. Thus, the marking criteria are different from those usually required for an academic essay. Your assignment will be assessed on the criteria shown in the rubric on the next page:

* The percentage given in the leftmost cell of each row shows you the percentage of the final mark available for that criterion
* The % shown in the topmost cell of each column shows you the range of final marks you would achieve if you were awarded marks in this column for all criteria. Your feedback will include a mark for each criterion enabling you to see exactly where you gain/lost marks.



