STAT5009 Decision Methods & Predictive Analytics

Take-Home Project

Semester 1, 2023

Objective

This take-home project is one of three assessments (along with Tests 1 & 2) in this unit. It is worth 60% of the overall mark.

The main objective is to allow the participants in the unit to demonstrate their grasp of the fundamentals and practical use of statistical and machine learning methods for prediction/classification that have been discussed during the unit, **and those that participants will research on their own**.

The project should focus on the analysis of a substantive dataset that participants may obtain from online sources, or from their place of work.

In both lectures and tutorials, we have analyzed data and fitted predictive models as if the steps to do so were clear, well-laid out, and led invariably to a 'correct' answer. Reality, however, is messier. There is not a linear path from problem and data to solution, and one of the pedagogical objectives of the project is to allow participants to get some sense of that.

Participants should work (with some exception) in **teams of 4 people**. Analysis and reporting are to be carried out in *R/RStudio* using R Markdown.

Assessment

Task	Mark	Due
Proposal (2–3 pages)	10%	TW7, Friday 21 April
Peer Review of Proposal	5%	TW10, Wednesday 10 May
Written Report	35%	Study Week, Wednesday 31 May
Oral Presentation	10%	TW11&12, Lecture A schedule will be available on By Teaching Week 10

Details of each of these assessments are shown below, and rubrics and other reference material will be available on Blackboard.

Data

Participants may wish to use data from their own workplaces, as long as confidentiality requirements do not prevent them from writing a report that will be read by the lecturer nor from speaking about their analysis to the other participants in the unit.

There are many public sources of data available, including open data websites such as OpenDataSoft. Also available on BB as a separate document are data sources compiled by the Data Analytics Practice Committee of the Institute of actuaries Australia.

The idea is to find a sufficiently complex dataset to allow you to demonstrate your familiarity with the methods studied in the unit **and those that we have not**. There may be several response variables for which prediction/classification methods have to be used. In addition, you will find yourself more motivated if you select a dataset from a field that is of interest to you.

Project Proposal

The project proposal is a short (2-3 page) Word document produced using R Markdown that contains:

1. Title
2. Data & Analyses
    a. Objective: What do you plan on predicting/classifying and why?
    b. Where do the data come from? Have these data been analyzed before?
    c. Describe context and variables and their types; show some plots/tables
    d. What analyses do you propose to carry out?
    e. How will you evaluate the predictive models?

Your proposal will be marked by one of your classmates.

Peer Review of Proposal

You will be provided with a rubric and some general guidelines to help you evaluate a proposal written by one of your classmates.

Project Report

The project report should be written as a formal technical report. It can be written wholly in R Markdown and then converted to Word, or some combination of R

Markdown for technical appendices and Word for the main body. There is no prescribed structure, but it should contain the following elements:

1. Problem Statement and Background
    - What is the problem you are trying to solve? Where do the data come from? Include background material as appropriate.
2. Methods
    - What are the methods you used for exploratory analysis and for prediction/classification? Provide background information on methods that we did not cover in the unit.
    - What hyper-parameter choices did you make and why?
    - What data cleaning/wrangling did you have to do before analysis?
    - Include methods that didn't work as well as those that did.
3. Results
    - Provide a detailed description of your results. What are the performance measures you used to assess predictive/classification accuracy?
    - If the data have been analyzed before, how well did your methods perform compared to those that others used?
    - Use informative and interesting visualizations for EDA and for displaying your results.
4. Conclusions and Lessons Learned
    - What would you have done differently? What other methods could you have used?

Depending on the complexity of the problem you have decided to tackle, the main body of the report will be 8-10 pages long, including important plots and tables, excluding references. The appendix should contain the R Markdown file and the resulting output from your data wrangling, exploratory data analysis, and quantitative analysis. **If you use any external resources such as books or websites – and you are encouraged to do so! – please make sure that you cite them appropriately.**

A rubric will be provided on BB to guide you as you write the report. If you are working in a team, please provide a breakdown of the effort of each member, and what each individual worked on.

You will be required to add the following statement to your report:

1. This assignment is my/our own original work, except where I/we have appropriately cited the original source (appropriate citation of original work will vary from discipline to discipline).

2. This assignment has not previously been submitted in any form for this or any other unit, degree or diploma at any university or other institute of tertiary education.

3. I/we acknowledge that it is my responsibility to check that the file I/we have submitted is: a) readable, b) the correct file and c) fully complete.

Oral Presentation

The last two lecture slots will be devoted to a light oral presentation of your work. Each presentation will be 3 minutes long, depending on the number of presentations. And a 10 - 15 minute long full presentation must be recorded and submitted through the Blackboard. A rubric will be made available on Blackboard.

Resources

The course textbook and supporting materials should be your starting point for help on exploratory data analysis and predictive methods. There is plenty of online help on using *R*. For example, the website stackoverflow has a subsection devoted to *R* that's very useful. A searchable archive of the *R* help list may be found. And, of course, you are welcome to contact the teaching team for guidance.

Good luck.