Econ 608 Data Mining and Big Data 2022-2023

COURSE AIMS & OBJECTIVES, KEY SKILLS AND LEARNING OUTCOMES

Course Aims & Objectives: The purpose of this course is to provide students with an introduction to data mining and how to best handle big data. Most applied research rely on data that is only getting larger and more complicated. Researchers have to first clean, manage, and mine the data to better understand any existing relationships and guide analysis. This course is aimed towards research and public policy applications.

Key Skills: By the end of this course, students should have some knowledge and understanding of:

- Different data structures
- Coding dos and donts
- Looping
- Data visualization
- Data (text) mining
- Machine learning

Desired Outcomes: By the end of this course, students should be able to:

- Understand big data basics (data storage, quality, etc.)
- Handle all sorts of data properly
- Data movement and manipulation
- Identify inconsistencies and problems with the data
- Present complex data in graphs and plots that are easy to follow
- Apply basic machine learning algorithms
- Work effectively both individually and within a team environment.

COURSE STRUCTURE

Econ 608 is a 10 credits course and therefore students are expected to input approximately 100 hours of study into the course. The total number of contact hours on Econ 608 is 15 hours. This leaves 85 hours for private study. Course Delivery comes in the form of Lectures with 15 hours delivered over the first 3 weeks of the term (10 hours of lectures and 5 hours of tutorials). There will be optional clinics on the last day of the course.

During your private study you should strike a balance between reading the course material (which is the primary source of information) and the recommended textbooks, thinking critically about how these fit in to the body of knowledge on the subject and about how our level of knowledge can be improved, performing exercises, completing

coursework and revising for examinations. You can expect to perform well on this course only if you work consistently through the year.

COURSE CONVENOR

Professor George Naufal

LECTURERS CONTACT INFORMATION (Including Office Hours)

Email: gnaufal@gmail.com

Available by appointment (please email to arrange a convenient time)

COURSEWORK ASSESSMENT

The final mark for the course will depend on a written exam. Timetable for details of time and venues will be communicated via Moodle and by Timberlake well in advance.

The CWA mark will be calculated as 100% coursework. The coursework will be assigned at the beginning of the module.

Coursework must be submitted electronically through the Moodle site for this course: https://mle.lancs.ac.uk/course. Login using your regular Lancaster University access details. This opens a page headed MLE: My home.

The format of the submission is as follows.

- The submitted file must be in pdf format with the following name

  stud#_studname_cw_**cw#.pdf**

  where: **stud#** is your student number, **studname** is your name in the format *surname_name*, **cw#** is either 1 or 2 according to the piece of coursework submitted. Eg a student with student number 111 would submit a file named *111_surname_firstname_cw_1.pdf*.

- Maximum file size is 2MB: figures resolution must be adjusted accordingly.

Note that your work will be screened using software designed to detect plagiarism.

Do not rely upon someone else to submit your coursework.
*Word counts are inclusive of all material submitted apart from the Bibliography.

FEEDBACK ON COURSEWORK:

The coursework will be marked and returned to students within 4 weeks of the submission deadline. Feedback will consist of marker's notes appended to the pdf of your coursework.

MARKING CRITERIA AND PENALTIES

Marking criteria can be found in the Economics Undergraduate Handbook and the general course information paper. An electronic copy of this can be found via the Current Student page of the university website then follow the Academic Regulations link
https://gap.lancs.ac.uk/ASQ/QAE/MARP/Documents/UG-Assess-Regs.pdf

FINAL MARK INFORMATION

This course is assessed 100% by means of coursework. The final mark is the average of the marks obtained in the two pieces of coursework.

COURSE TEXT AND RECOMMENDED READING

Main texts

The main recommended textbook is: M. N. Mitchell (2020) Data Management Using Stata: A Practical Handbook, Second Edition, Stata Press.

Students will also find the following texts useful as further reading.

- N. C. Cox, H. J. Newton (2014) One Hundred Nineteen Stata Tips, Third Edition, Stata Press
- J. P. Hoffmann (2017) Principles of Data Management and Presentation, University of California Press
- M. N. Mitchell (2012) A Visual Guide to Stata Graphics, Third Edition, Stata Press
- N. J. Cos (2014) Speaking Stata Graphics, Stata Press
- M. J. Zaki and W. Meira (2020) Data Mining and Machine Learning, Second Edition, Cambridge University Press

Note Copies of the lecture slides will be made available on the course web pages. You **MUST** print off the notes for each lecture **prior to** attending. Solutions to exercises, and some additional material associated with these lectures and course announcements will also be placed on this website.

COURSE OUTLINE/LECTURE SCHEDULE

Day 1: **Introduction to Big Data**

- Types of data (data structure, size, etc.)
- What makes data labelled as big data
- Data storage and quality
- Inputting and outputting data
- Merging, appending, and producing data

Day 2: **Coding with Big Data**

- The art of coding
- Organizing programming files
- Doing repetitive things
- Looping

Day 3: **Data Visualization**

- Simple graphs and plots
- Use loops to draw graphs
- How to prep data for complicated graphs

Day 4: **Machine Learning**

- Introduction to machine learning
- Benefits of machine learning
- Basic machine learning alghorithms

Day 5: **Machine Learing - continued**

- LASSO
- Random Forest
- Unsupersived and supervised regressions