ACST891

Data Analytics Tools for Finance and Insurance

S2 Evening 2017

Dept of Applied Finance and Actuarial Studies

Contents

Disclaimer

Macquarie University has taken all reasonable measures to ensure the information in this publication is accurate and up-to-date. However, the information may change or become out-dated as a result of change in University policies, procedures or rules. The University reserves the right to make changes to any information in this publication without notice. Users of this publication are advised to check the website version of this publication [or the relevant faculty or department] before acting on any information in this publication.

General Information

Unit convenor and teaching staff
Unit convenor and lecturer
Sachi Purcal
sachi.purcal@mq.edu.au
Contact via Email
E4A615
Tuesdays 1400–1600 during teaching weeks

Lecturer
Neil Fraser
neil.fraser@mq.edu.au
Contact via Email or mobile (+61 408 419 691)
Refer to iLearn
Refer to iLearn

Credit points
4

Prerequisites
ACST890

Corequisites

Co-badged status

Unit description
The world of `Big Data' is rapidly evolving in finance and insurance, with new technologies emerging while existing technologies mature. Hadoop is the first high-performance commercial computing platform that works at scale and is affordable at scale. This unit focuses on the Hadoop platform and the Hadoop ecosystem of tools. These technologies are at the core of the `Big Data' phenomenon, and they facilitate scalable management and processing of vast quantities of data. Students who complete this unit will understand the architecture of Hadoop clusters. Using Hadoop and related `Big Data' technologies such as MapReduce, Hive, Impala and Pig, they will develop analytics to devise solutions to the types of problems challenging finance and insurance today. Students undertaking this unit are expected to simultaneously enrol and complete the Cloudera course on Apache Hadoop, with the aim of obtaining the resulting Cloudera professional credentials.

Important Academic Dates

Information about important academic dates including deadlines for withdrawing from units are

available at https://www.mq.edu.au/study/calendar-of-dates

Learning Outcomes

On successful completion of this unit, you will be able to:

Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.

Assemble statistical and machine learning techniques to tackle data science problems.

Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.

Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.

Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

General Assessment Information

For all assessments the following apply.

- Assessment criteria for all assessment tasks will be provided on the unit iLearn site.
- All individual assessment results will be made available under Grades on the website.
- It is the responsibility of students to view their marks for each within-session assessment on iLearn within 20 working days of posting. If there are any discrepancies, students must contact the unit convenor immediately. Failure to do so will mean that queries received after the release of final results regarding assessment marks (not including the final exam mark) will not be addressed.
- In the case where a disruption to studies application is approved, the student may be offered an alternative assessment or may receive a mark based on the percentage mark achieved by the student in one or more other assessment tasks, at the unit convenor's discretion.

Assessment Tasks

Name	Weighting	Hurdle	Due
Online Quiz	0%	No	23 August
Assignment (group component)	10%	No	17/10/17

Name	Weighting	Hurdle	Due
Assignment (individual part)	40%	No	17/10/17
Final examination	50%	No	University Examination Period

Online Quiz

Due: **23 August**
Weighting: **0%**

The online quiz will cover the first three weeks' material. The quiz is due on Wednesday 23 August (Week 04) at 11.30 p.m. (2330) to be submitted online via the iLearn site.

Please use the quiz an an indicator of whether you are progressing satisfactorily in the unit. If you are having difficulties, please see the Unit Convenor and consider withdrawing before the census date on Friday of Week 04.

On successful completion you will be able to:

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.
- Assemble statistical and machine learning techniques to tackle data science problems.
- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.
- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.

Assignment (group component)

Due: **17/10/17**
Weighting: **10%**

The assignment will consist of two parts: a group component and an individual component.

The group component will consist of external analysis based on big data techniques. The group component should be about 1000–2000 words (12pt font size with 1.5 spacing). It must be submitted (as a *readable* PDF file—it is students' responsibility to check this) via iLearn.

You will be a member of a syndicate group that selects or builds data sets from publicly available data that can be used to formulate a data science strategy for a company. The comprehensive analysis will utilise knowledge and skills developed during ACST890 and ACST891.

No extensions will be granted. Students who have not submitted the task prior to the deadline will be awarded a mark of zero for the task, except for cases in which an application for disruption to studies is made and approved.

On successful completion you will be able to:

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.
- Assemble statistical and machine learning techniques to tackle data science problems.
- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.
- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.
- Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

Assignment (individual part)

Due: **17/10/17**
Weighting: **40%**

The assignment will consist of two parts: a group component and an individual component.

The individual component will consist of analysis. Your individual contribution to the assignment should be about 2000–3000 words (12pt font with 1.5 spacing). Each member of the syndicate group must clearly identify which element of the group assignment is his or her individual contribution. This can be done by putting your names in brackets next to a section heading and/or in the table of contents (if you use one).

Your individual work must be submitted (as a *readable* PDF file—it is students' responsibility to check this) via iLearn.

No extensions will be granted. There will be a deduction of 10% of the total available marks made from the total awarded mark for each 24 hour period or part thereof that the submission is late (for example, 25 hours late in submission—20% penalty). This penalty does not apply in cases for which an application for disruption to studies has been made and approved. No submissions will be accepted after solutions have been posted.

On successful completion you will be able to:

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.
- Assemble statistical and machine learning techniques to tackle data science problems.
- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.

- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.
- Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

Final examination

Due: **University Examination Period**
Weighting: **50%**

The final examination will be a three-hour written paper with ten minutes reading time, held during the university examination period.

The exam will be open book.

On successful completion you will be able to:

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.
- Assemble statistical and machine learning techniques to tackle data science problems.
- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.
- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.

Delivery and Resources
Textbook

No textbook in envisioned for this course. Readings will be assigned over the semester from a variety of sources.

Technology used and required

We will learn a variety of data science packages over the semester. In addition, you will need to be familiar with document processing software (e.g., WORD) to produce your group assignment.

Unit Schedule

	Sachi Purcal	**Applied data science** • What is data science? • What kind of person does data science? • Corporate data science • International data science • Government data science • Life sciences (bio-informatics and Pharma)	VM setup for Cloudera
	Neil Fraser	**External analysis techniques** • The web as a data warehouse • Data markets • Linked open data • Social profiling • Undertake industry analysis, to assess existing and future industry forces, industry structure and industry attractiveness (big data techniques)	Sourcing a web data set
	Neil Fraser	**Internal analysis techniques** • Identifying strategic data • Identifying operational data • Orchestrating value-chains that cover data activities for analysis • Data lineage and data science life cycle • Analytical frameworks	Google Refine tutorial. Undertake a data resource audit check: quality, cleaning and parsing.

	Neil Fraser	**Big data tools**
		• Visualisation tools
		   ◦ Google Data Studio
		   ◦ Power BI
		   ◦ Python Seaborn
		• Query tools
		   ◦ Google Big Query
		   ◦ Fusion
		   ◦ Impala and Hive
			Visualisation with Google Data Studio

	Neil Fraser	**Big data technologies**
		• Database technologies
		   ◦ Hadoop
		   ◦ NoSQL
		   ◦ RDBMS
		   ◦ Parallel processing
		• Hadoop architecture
		• MapReduce
			Ingest data to VM and query with Impala and Hive

Neil Fraser	**Natural language processing** • Information annotation and extraction ◦ Optical character recognition sentence segmentation ◦ Part-of-speech tagging ◦ Named entity recognition ◦ Gazetteers ◦ Information extraction ◦ Relation and co-reference extraction • Information meaning ◦ Ontology mapping ◦ Topic models ◦ Sentiment analysis • Information tools and techniques ◦ NLTK ◦ GATE	Ingest VODAFAIL data to GATE/ LEXIMANCER /NVIVO
Neil Fraser	**Machine learning** • What is machine learning? • The history of machine learning • What's new in machine learning? • Supervised versus unsupervised learning • Six machine learning tasks 1. Clustering 2. Detecting outliers 3. Affinity analysis 4. Classification 5. Regression analysis 6. Recommendation • Machine learning and MapReduce • Spark	Spark or Sickit or Mahout
Neil Fraser	**Taking data science to production** • Implementing • Scalability and upgrading models • Versioning	Assignment
Neil Fraser	**Business Models in data science** • Recommend new business models of firms built on data science • Critically analyse the structural and cultural requirements for competitive data analysis • Internet commerce, insurance, automobiles, banking, mining	Assignment

	Sachi Purcal	**Moving beyond linearity + Tree-based methods** • Polynomial regression • Step functions • Basis functions • Regression splines • Smoothing splines • Local regression • Generalised additive models (GAMs) • Decision trees	Tutorial problems on this material
	Sachi Purcal	**Tree-based methods + Support Vector Machines** • Bagging, random forests, boosting • Maximal margin classifier • Support vector classifiers • Support vector machines • SVMs with more than two classes • Relationship to logistic regression	Tutorial problems on this material
	Sachi Purcal	**Unsupervised learning** • Challenge of unsupervised learning • Principal components analysis • Clustering methods	Tutorial probelms on this material
	Sachi Purcal	**Revision** • Exam preparation	Mahout (Apache open source machine learning library)

Policies and Procedures

Macquarie University policies and procedures are accessible from Policy Central. Students should be aware of the following policies in particular with regard to Learning and Teaching:

Academic Honesty Policy http://mq.edu.au/policy/docs/academic_honesty/policy.html

Assessment Policy http://mq.edu.au/policy/docs/assessment/policy_2016.html

Grade Appeal Policy http://mq.edu.au/policy/docs/gradeappeal/policy.html

Complaint Management Procedure for Students and Members of the Public http://www.mq.edu.au/policy/docs/complaint_management/procedure.html

Disruption to Studies Policy (in effect until Dec 4th, 2017): http://www.mq.edu.au/policy/docs/disruption_studies/policy.html

Special Consideration Policy (in effect from Dec 4th, 2017): https://staff.mq.edu.au/work/strategy-planning-and-governance/university-policies-and-procedures/policies/special-consideration

In addition, a number of other policies can be found in the Learning and Teaching Category of Policy Central.

Student Code of Conduct

Macquarie University students have a responsibility to be familiar with the Student Code of

Conduct: https://students.mq.edu.au/support/student_conduct/

Results

Results shown in *iLearn*, or released directly by your Unit Convenor, are not confirmed as they are subject to final approval by the University. Once approved, final results will be sent to your student email address and will be made available in eStudent. For more information visit ask.mq.edu.au.

Supplementary exams

Information regarding supplementary exams, including dates, is available at:

http://www.businessandeconomics.mq.edu.au/current_students/undergraduate/how_do_i/disruption_to_studies

Student Support

Macquarie University provides a range of support services for students. For details, visit http://students.mq.edu.au/support/

Learning Skills

Learning Skills (mq.edu.au/learningskills) provides academic writing resources and study strategies to improve your marks and take control of your study.

- Workshops
- StudyWise
- Academic Integrity Module for Students
- Ask a Learning Adviser

Student Services and Support

Students with a disability are encouraged to contact the Disability Service who can provide appropriate help with any issues that arise during their studies.

Student Enquiries

For all student enquiries, visit Student Connect at ask.mq.edu.au

IT Help

For help with University computer systems and technology, visit http://www.mq.edu.au/about_us/offices_and_units/information_technology/help/.

When using the University's IT, you must adhere to the Acceptable Use of IT Resources Policy. The policy applies to all who connect to the MQ network including students.

Graduate Capabilities

PG - Discipline Knowledge and Skills

Our postgraduates will be able to demonstrate a significantly enhanced depth and breadth of knowledge, scholarly understanding, and specific subject content knowledge in their chosen fields.

This graduate capability is supported by:

Learning outcomes

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.
- Assemble statistical and machine learning techniques to tackle data science problems.
- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.
- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/more/training/certification.html.
- Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

Assessment tasks

- Online Quiz
- Assignment (group component)
- Assignment (individual part)
- Final examination

PG - Critical, Analytical and Integrative Thinking

Our postgraduates will be capable of utilising and reflecting on prior knowledge and experience, of applying higher level critical thinking skills, and of integrating and synthesising learning and knowledge from a range of sources and environments. A characteristic of this form of thinking is the generation of new, professionally oriented knowledge through personal or group-based critique of practice and theory.

This graduate capability is supported by:

Learning outcomes

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various

computing tools.

- Assemble statistical and machine learning techniques to tackle data science problems.

- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.

- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/ more/training/certification.html.

- Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

Assessment tasks

- Online Quiz

- Assignment (group component)

- Assignment (individual part)

- Final examination

PG - Research and Problem Solving Capability

Our postgraduates will be capable of systematic enquiry; able to use research skills to create new knowledge that can be applied to real world issues, or contribute to a field of study or practice to enhance society. They will be capable of creative questioning, problem finding and problem solving.

This graduate capability is supported by:

Learning outcomes

- Apply and assess different methods and techniques to formulate data analytic and visualisation solutions to finance and insurance `Big Data' problems using various computing tools.

- Assemble statistical and machine learning techniques to tackle data science problems.

- Reflect on knowledge learned of theory and business practices for future learning and ongoing professional development.

- Complete one of the Cloudera certification qualifications: https://www.cloudera.com/ more/training/certification.html.

- Engage effectively with others to critically examine different viewpoints and work productively in a group by coordinating activities, allocating tasks and synthesizing different material and viewpoints.

Assessment tasks

- Online Quiz

- Assignment (group component)

- Assignment (individual part)

- Final examination

Research and Practice

This unit uses research by Macquarie University researchers, as well as from other Australian and international researchers (references are given in the unit notes).

You are also required to source and use Australian and international research as part of the assignment in this unit.