Stellenbosch University Master's Course

Data Science

Course Outline

Lecturers: Nico Katzke & Dr. Dawie van Lill

1 PART I (Katzke)

The aim of the first part of the course is to introduce students to quantitative techniques with a strong emphasis on practical application. In this, the first part of the course, we will be covering topics around building sustainable, replicable and well documented functional coding infrastructures using the freeware statistical package `R`, as well as effectively using Github as a means of securing code in an efficient version controlled environment. This is a specialized and technical course that will enable students to comfortably engage with data, produce replicable projects and become more comfortable with data visualization and -analysis.

This course is strongly recommended for students looking to further their quantitative finance skills by taking the Financial Econometrics course. All class notes and additional information will be loaded onto the course's official website:
https://datsci.nfkatzke.com

This section will count 50% of the final mark for the course.

1.1 Assessment

Students will be graded on their ability to master the concepts discussed in class. The assessment will be conducted in a timed practical exam. Students will be expected to do data analytics, produce visualizations and build a well-structured and documented project using real world data. Submission will be done using Github, with students graded on their coding, reasoning and creative abilities displayed in answering real world questions.

1.2 Topics Covered

1.2.1 Practical 1: Getting started in R

- Deep-dive in the coding language R
- Introduction to Rstudio
- Base R coding overview

1.2.2 Practical 2: Functional Programming

- Philosophical discussion around coding paradigms - focussing on Object Oriented and Functional Coding developments through time.
- Building a robust functional coding environment
- Functional coding in R - motivation and illustration of the power of functional coding
- This will include a bonus practical serving as practice for effectively using modern R verbage.

1.2.3 Practical 3: Tidy Programming

- An introduction to the Tidyverse

- Motivation and illustration of the power of tidy visualization in R

Readings: R4ds, Hadley Wickham (https://r4ds.had.co.nz/)

1.2.4 Practical 4: Data visualization

- Introduction to effective data visualization
- Overview of the power of tidy visualization using ggplot.

1.2.5 Practical 5: Writing Reports in R

- Overview of using Rmarkdown for writing formal reports in Pdf and HTML
- Introduction to using the Texevier package for writing formal academic reports.

For all the sessions, class notes will be made available online and communicated to students ahead of time.

1.3 Assessment

A practical exam will be arranged with students directly. Students are required to arrange access to a laptop / computer that has internet access and R & R-studio installed.

The practical exam date will be set for 16 - 17 June, and will be a 24 hour practical project.

2 Part II (van Lill)

The first part of the course focuses on programming concepts and data exploration. In the second part we will build on this foundation and try to make rigorous conclusions about the data. We will model the data in various ways, with a special focus on machine learning methods. In addition to the modelling component we will also explore what programming and data science looks like in other modern programming languages like Julia and Python. We will end the course with a brief overview of SQL.

Most of the notes will be loaded onto the following Github page:
https://github.com/DawievLill/DataScience-871

However, we will also consult the following page for the initial portion of the lectures:
https://datasciencebox.org/

The project for this section will count 50% of the final mark for the course.

2.1 Topics Covered

2.1.1 Session 1: Modelling data

- The language of models
- Fitting and interpreting models
- Modelling nonlinear relationships
- Models with multiple predictors

2.1.2 Session 2: Model building

- Getting used to machine learning terminology
- Prediction and overfitting
- Feature engineering

2.1.3 Session 3: Model validation

- Cross-validation
- Feature engineering contd.

2.1.4 Session 4: Uncertainty quantification

- Bootstrapping
- Hypothesis testing
- Inference

2.1.5 Session 5 + 6: Machine learning methods

- Shrinkage methods
- Decision trees
- Random forests and gradient boosting

2.1.6 Session 7: Alternative programming languages

- Introduction to Python (and Julia) for data manipulation
- Basic overview of SQL

2.2 Assessment

The assessment for this part of the module takes the form of a semester project. Students are expected to hand in a proposal for a project during the semester, which will have to be approved by the lecturer.

The submission date for the project is the 30th of June at midnight. Details of the project will be provided at the start of the semester.

To pass the module, a final mark of at least 50 percent has to be obtained. To obtain a distinction in this module, a minimum final mark of 75 percent is required