Introduction to Actuarial Data Science

Actuarial Data Science - Open Learning Resource

Fei Huang, UNSW Sydney

With the rise of AI and big data, every firm has effectively become a data-driven enterprise. Whether it is supermarkets, hairdressers, restaurants, or dentists, businesses of all types are collecting data on customer behaviour and engagement. We are living in a data-driven era, where data has become one of the most valuable assets for businesses. However, possessing data alone does not automatically translate into insights or business value. It is crucial to apply appropriate data science techniques to extract meaningful insights and communicate those findings effectively to stakeholders.

There are many textbooks available today that cover data science, statistical methods, and machine learning techniques. What does this Open Learning Resource (OLR) offer that is different? While many excellent data science textbooks exist, they often focus on individual techniques or models without addressing the entire problem-solving process in a business setting. In contrast, this OLR aims to equip readers with the skills to solve data problems in real-world business environments. It guides readers through the complete process: from asking the right questions and performing exploratory data analysis to modeling, interpreting results, communicating findings, and considering ethical implications.

The aim of this learning material is aligned with the Australian Actuaries Institute's Part II Data Science Principles syllabus.

> "The Data Science Principles aim to extend students' knowledge of modern analytical tools and techniques beyond those introduced in the Foundation Program subjects and to teach students how to apply this knowledge in real-life business settings"
>
> — Actuaries Institute, *Data Science Principles syllabus*

Data science is an interdisciplinary field that covers many areas of knowledge, including but not limited to statistics, machine learning, databases, optimization, algorithms, programming, and domain knowledge in a business setting. This book mainly focuses on applying the data analysis cycle with statistical machine learning techniques to address actuarial applications, referred to as Actuarial Data Science. The techniques and concepts introduced in this book can be applied more broadly to other business problems. Therefore, this book can also be used as a textbook for solving general business data problems.

It is often argued that data science is a discipline rooted in science and engineering, with a vast arsenal of quantitative tools. However, it is important to recognise that data science is also an art (Peng and Matsui 2015). While a wide range of analytical techniques—from linear regression to classification trees and deep learning—have been codified into software packages, the role of the data scientist extends far beyond selecting and running algorithms. Effective data science involves making numerous judgment calls throughout the problem-solving process: choosing appropriate tools for specific tasks, interpreting results, communicating findings clearly to stakeholders, and embedding ethical and regulatory considerations into each decision. These nuanced aspects of data science—judgment, interpretation, and communication—remain, at least for now, beyond the capabilities of machines.

Data Science Lifecycle (DSL)

Actuaries apply the Actuarial Control Cycle (ACC), as shown in Figure 1 for problem solving. Figure 1 is adapted from the Actuaries Institute, based on Bellis et al. (2010).

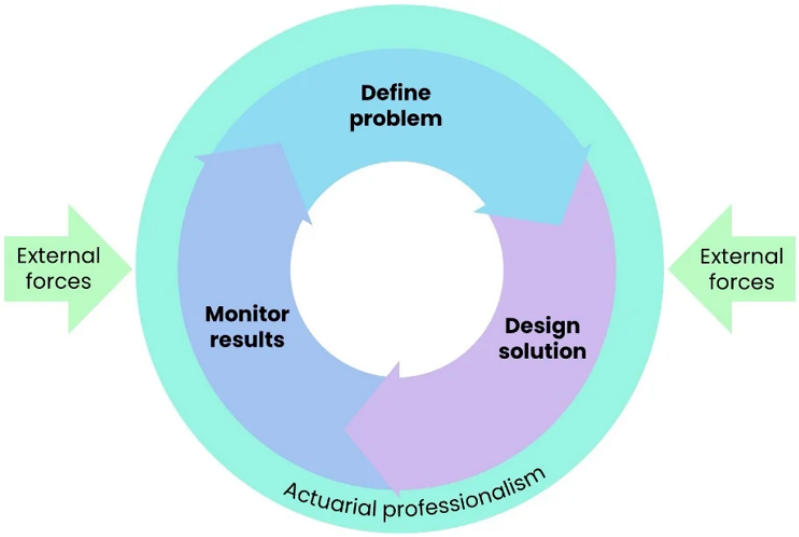

Figure 1: The Actuarial Control Cycle

In a similar spirit, we introduce the Data Science Lifecycle (DSL) as a practical methodology for tackling data-driven problems. While rooted in the principles of the ACC, the DSL reflects the unique, iterative, and often non-linear nature of data analysis. It comprises six key steps and serves as a tailored application of the ACC to the context of data science, as shown in Figure 2.

The six steps of DSL are listed below:

1. Problem statement

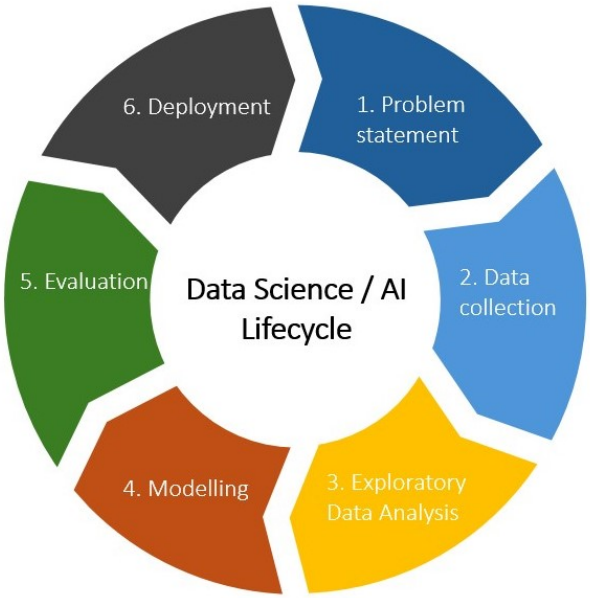

Figure 2: Data Science Lifecycle

2. Data collection
3. Exploratory data analysis
4. Modelling
5. Evaluation
6. Deployment

Throughout the entire process, ethical considerations and professional conduct are essential and must be integrated into every step. Just as importantly, each step of the DSL is grounded in a deep understanding of the business context—without this, data science risks becoming disconnected from real-world impact.

In the chapters that follow, we illustrate how to apply the DSL to address practical business challenges. In addition to the six steps, we also emphasize communication and ethics as core components that must be embedded into the lifecycle.

Roadmap

The chapters are organised as follows:

Chapter 2 covers Step 1, focusing on how to ask the right question and frame it as a data science problem.

Chapter 3 addresses Steps 2 and 3, including data collection and exploratory data analysis.

Chapter 4 presents a range of modelling techniques, explaining when and how to apply them.

Chapter 5 introduces a systematic evaluation toolbox for data-driven decision-making, with a particular focus on applications in the insurance sector.

Chapter 6 explores effective communication strategies for engaging different stakeholders.

Chapter 7 introduces the Ethical Data Science (AI) Lifecycle (EDSL)—a framework for embedding ethical thinking across the entire DSL.

References

Bellis, Clare, Richard Lyon, Stuart A. Klugman, and John Shepherd, eds. 2010. *Understanding Actuarial Management: The Actuarial Control Cycle*. 2nd ed. Sydney, Australia: Institute of Actuaries of Australia; Society of Actuaries.

Peng, Roger D, and Elizabeth Matsui. 2015. *The Art of Data Science: A Guide for Anyone Who Works with Data*. Skybrude Consulting, LLC.