Data Science
COMP5122M

Business understanding

Roy Ruddle

Parts of this lecture are based on the book "Data Science for Business" by Foster Provost and Tom Fawcett, 2013. Slides and figures from the book are used with the authors' permission.

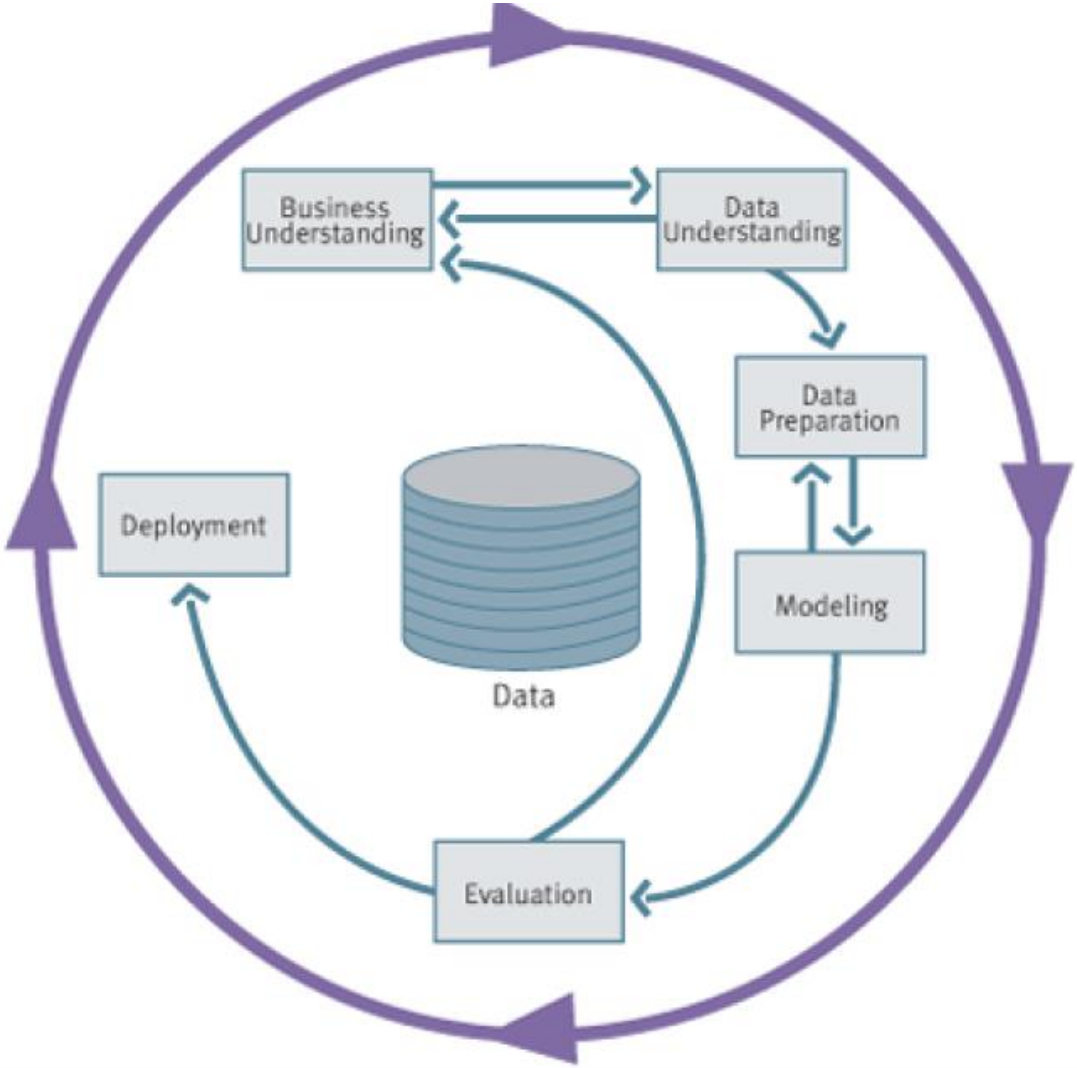
Private study

- See Minerva Announcements for up-to-date info
- Private study for this lecture
  - Data Science for Business:
    - Chapters 1, 2 & 11
    - Appendix A (Proposal Review Guide)

What will you learn?

- The CRISP data mining process
- What we mean by "decision analytic thinking" and why it's important
- Some simple methods that will help you understand business problems
- Appreciate the long time that may be required

CRISP data mining process



Cross-Industry Standard Process for Data Mining (Shearer, 2000).

CRISP as a framework

- **Business understanding**
  - **Think about the problem to be solved and about the use scenario**
- **Data understanding**
  - **Strengths & limitations of your data; historical data; data sources; cost of data; does data match the problem?**
- **Data preparation**
  - **Data quality; converting & transforming data; data linkage; data leaks**
- **Modelling**
  - **Unsupervised & supervised tasks**
    - **Classification & probability estimation; Regression; Similarity matching; Clustering; Co-occurrence grouping; Profiling; Link prediction; Data reduction; Causal modelling**
- **Evaluation**

Example – Signet Bank

Provost & Fawcett Ch. 1

- 1990s
  - Became possible to model profitability, not just probability of default
    - > 100% profit came from small proportion of customers
  - But Signet didn't have the data they needed
- Strategy
  - Invest in acquiring the data & modelling
    - Experiment with loan terms for customers
    - Charge-off losses (unpaid balances) doubled
- Success
  - 1994 spun off credit card operations into Capital One
    - Now 4th largest credit card issuer, and 8th largest bank, in USA

Decision analytic thinking

- How can we design (or 'engineer') a solution to our business problem?
- Charity asks you to help fund-raise
  - Three people tell you the campaign's aim
    - "We want as many donors as possible"
    - "We want to maximise the donations we receive"
    - "We want the campaign to be as profitable as possible"
  - Are the aims different?
    - In terms of measuring success?
    - The data required to model the campaign?
  - Which aim is best?

What is the problem?

- **Stakeholder**
  - **Has a vague idea of the problem**
  - **No idea how to tackle it (apart from silver bullets)**
- **You**
  - **Know little about their business**
  - **Don't understand the terminology**
  - **Have never seen any data**
- **In these circumstances, a successful data scientist will succeed!**

> "If you can't explain it simply, you don't understand it well enough"
> – Albert Einstein

Methods for solving the problem

- **Discussion**
  - **Preferably face-to-face (especially the 1st time)**
- **Read publications**
  - **Reports, presentations, etc.**
- **Questionnaire**
  - **Request written answers**
- **Run a stakeholder workshop**
- **Obtain some example data**

Example #1: A small business

- Pre-proposal (2 months)
  - Face-to-face discussion (build understanding & develop relationship)
  - Email/phone follow-up
- Proposal (2 months)
- Exploratory analysis (6 weeks)
  - Face-to-face discussion (1-2 times/week)
  - Publications (presentation)
  - Example data
  - Summary statistics & data visualization
- Final report (4 weeks)

Example #1 (contd.)

- **Quick wins**
  - **Improve calibration protocol for sensors**
- **Medium-term benefits**
  - **Workflow for data analysis**
  - **Data cleaning methods, to identify & characterise abnormal events**

Example #2: QuantiCode (2016-2020)

- **Aim**
  - **Develop novel data mining and visualization tools and techniques**
  - **10 person-years of research**
- **Partners**
  - **Leeds City Council, Sainsbury's, NHS Digital, Leeds Informatics Board, Bradford Institute for Health Research, Consumerdata, and aql**
- **Funded by the EPSRC (EP/N013980/1)**
  - **£980k**
- **Based in LIDA**

QuantiCode timeline

- Research project proposal (Months 1-11)
- Recruitment (Months 11-17)
- Scoping the problem (Months 15-21)
- Research (Months 18-60)
  - Data sharing agreement (Months (20-23)
  - Ethical approval (Months 21-23)
  - Received example data (Months 24, 30 & 40)
  - First tool released (Month 46)

QuantiCode: Scoping the problem (1)

- **Questionnaire (see Minerva)**
  - **"Please write brief answers to the following questions (a few sentences is enough, and incomplete answers are fine …"**
  - **Background information about 7 scenarios from 5 organisations**
    - **What is the goal of the analysis?**
    - **Who is involved and what are their roles?**
    - **What data is involved?**
    - **What analysis steps are involved?**
    - **What do you already know?**
    - **What would you like to do, but cannot do today?**

QuantiCode: Scoping the problem (2)

- **Discuss data analysis scenarios**
  - **7 face-to-face meetings and 3 phone discussions**
  - **Involving 16 people from 4 external organisations**
- **What did we learn?**
  - **What's important to each organisation**
    - **Track the journey of people, to predict future care needs**
    - **Data quality (that's why the business is successful)**
    - **Want to improve data quality (little consistency checking)**
    - **Need to better understand customers (especially convenience stores)**
  - **Good quotes**
    - **"We're spectacularly poor at using our own data"**
    - **"We don't understand our own data"**
    - **"That would be cool – a tool to visualize quality rather than reading a statistician's report"**

QuantiCode: Scoping the problem (3)

- Workshops
  - Describe the stakeholders' problem
  - Then invite them to correct you!
  - Clarifies misunderstandings
  - Agree mutually beneficial priorities
- Organisation A (2 stakeholders)
  - Sample data sent in advance
  - Created analysis storyboard to stimulate discussion
- Organisation B (6 stakeholders)
  - Breadth of stakeholders helped generalise our ideas

Summary

- Understand the business problems and build rapport with stakeholders
  - Discussions (especially face-to-face)
  - Workshops
- Excite the stakeholders and the potential
- Show them you're interested in their problem
  - Not just the technical intricacies of the solution
- Make the stakeholders' life easier
  - Their effort will be paid back many times over