

Data Science for Business

Lecture #1

Course Overview

Prof. Alan L. Montgomery

The University of Hong Kong & Carnegie Mellon University, Tepper School of Business

email: alanmontgomery@cmu.edu

All Rights Reserved, © 2021 Alan Montgomery

Do not distribute, post, or reproduce without Alan Montgomery's Permission

Course Information



Course Objectives

Understand Data Science

- Data mining is a process of using data to solve problems. The focus of our problems are business and e-commerce problems like retaining customers. Data scientists combine statistics, computer science, and business domain knowledge to solve problems empirically.

Introduce and apply a set of data mining tools

- Techniques include cluster analysis, regression, decision trees, and logistic regression. Our focus is on the application and interpretation of the techniques – not their implementation.
- Use R Studio as a tool for data mining.

Apply these techniques to specific case studies to solve business and e-commerce problems

- Some problems include market segmentation, predicting customer profitability and retention, analyzing text data from keyword search and social media.



What you will learn

Purposes and practice of data science

Highlights of major data mining algorithms

- Visual Analysis
- Clustering
- Text mining
- Decision trees
- Linear and Logistic Regression

Illustrations of using data mining to solve business and e-commerce problems

- Use of data mining using RStudio

Point out limitations of data, as well as ethical and legal issues in data mining



Prerequisites

Mathematics knowledge

- Models rely on algebraic knowledge.
- Optimization of most techniques requires some calculus, but we will not cover this
- Helpful if you know something about linear regression and statistics. However, I assume that you are not familiar with these techniques

Software

- We will use R for data analysis, and will not need any custom programming. R is a full featured statistical and programming environment for manipulating data, to perform data analysis, and display graphics and results.
- I presume you are familiar with Windows OS and interface, and may use Excel for some examples.

Your backgrounds are diverse. My objective is to aim towards the middle of the class. Additional readings available for those that want a challenge.



Course Outline

Understanding Data and the Data Mining Process (1, May 23, Sunday)

Exploring and Visualizing Data

- Exploring and Visualizing Data: Cluster Analysis (2, 10Jun, Thursday)
- Market Segmentation of Consumers, **Case Study: Ford Ka** (3, 17Jun, Thursday)
- Working with Unstructured Datasets (4, 24Jun, Thursday)

Predictive Modeling

- Predictive Modeling (5, 8Jul, Thursday)
- Data Mining Techniques for Prediction, (6, 11Jul, Sunday)
- Overfitting and Evaluating Models **Case Study: Freemium** (7, 15Jul, Thursday)

Business Applications

- Pro-Active Churn Management, (8, 18Jul, Sunday)
- Data Based Decision Making, **Case Study: Cell2Cell I** (9, 22Jul, Thursday)
- Business Strategy for employing Data Science, **Case Study: Cell2Cell II** (10, 29Jul, Thursday)

Final Exam (Aug)



Course Expectation

Exercises (60% written assignment + 5% presentation)

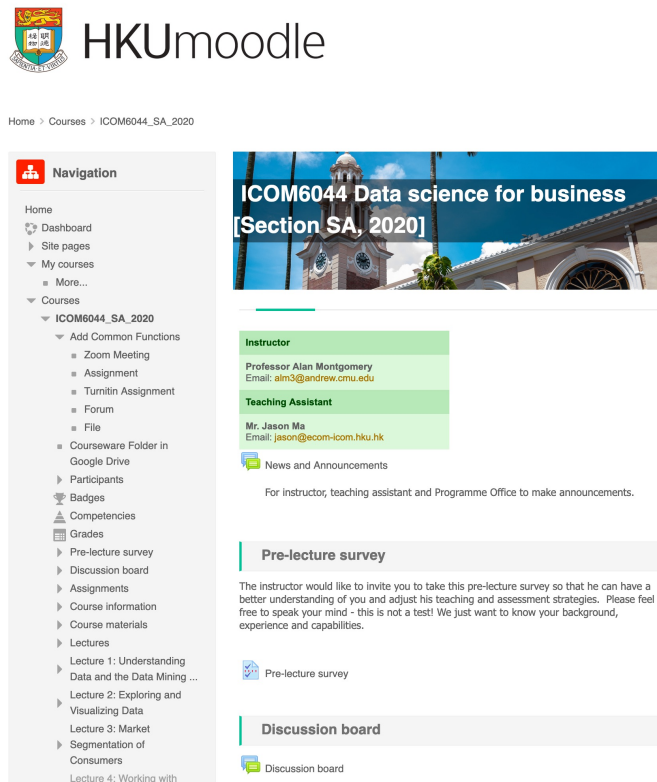
- Meant to facilitate learning
- 4 cases (individual or group of 4), but you may omit one as optional. If you do all then I will take your 3 best scores.
- No late assignments accepted since we will discuss the solutions in class
- I will ask you keep the same teams for all assignments.
- Given format I will ask presentations to be recorded to make it easy to play during class and discuss. I may not be able to have everyone present. I will let you know the problem that your team is to present.

Final Exam (35%)

- Based upon required readings and lectures
- Comprised of short answer questions and data analysis
- To be Scheduled for August



Course Materials



HKU Moodle

Home > Courses > ICOM6044_SA_2020

Navigation

- Home
- Dashboard
- Site pages
- My courses
 - More...
- Courses
 - ICOM6044_SA_2020
 - Add Common Functions
 - Zoom Meeting
 - Assignment
 - Turnitin Assignment
 - Forum
 - File
 - Courseware Folder in Google Drive
 - Participants
 - Badges
 - Competencies
 - Grades
 - Pre-lecture survey
 - Discussion board
 - Assignments
 - Course information
 - Course materials
 - Lectures
 - Lecture 1: Understanding Data and the Data Mining ...
 - Lecture 2: Exploring and Visualizing Data
 - Lecture 3: Market Segmentation of Consumers
 - Lecture 4: Working with

ICOM6044 Data science for business [Section SA, 2020]

Instructor
Professor Alan Montgomery
Email: alm3@andrew.cmu.edu

Teaching Assistant
Mr. Jason Ma
Email: jason@ecom-icom.hku.hk

News and Announcements
For instructor, teaching assistant and Programme Office to make announcements.

Pre-lecture survey
The instructor would like to invite you to take this pre-lecture survey so that he can have a better understanding of you and adjust his teaching and assessment strategies. Please feel free to speak your mind - this is not a test! We just want to know your background, experience and capabilities.

Discussion board

All course materials (syllabus, schedule, assignments, project details, announcements, links, etc.) are posted online

Make sure you check the site regularly

You must read the assigned readings before coming to each lecture.



Software

We will use Rstudio

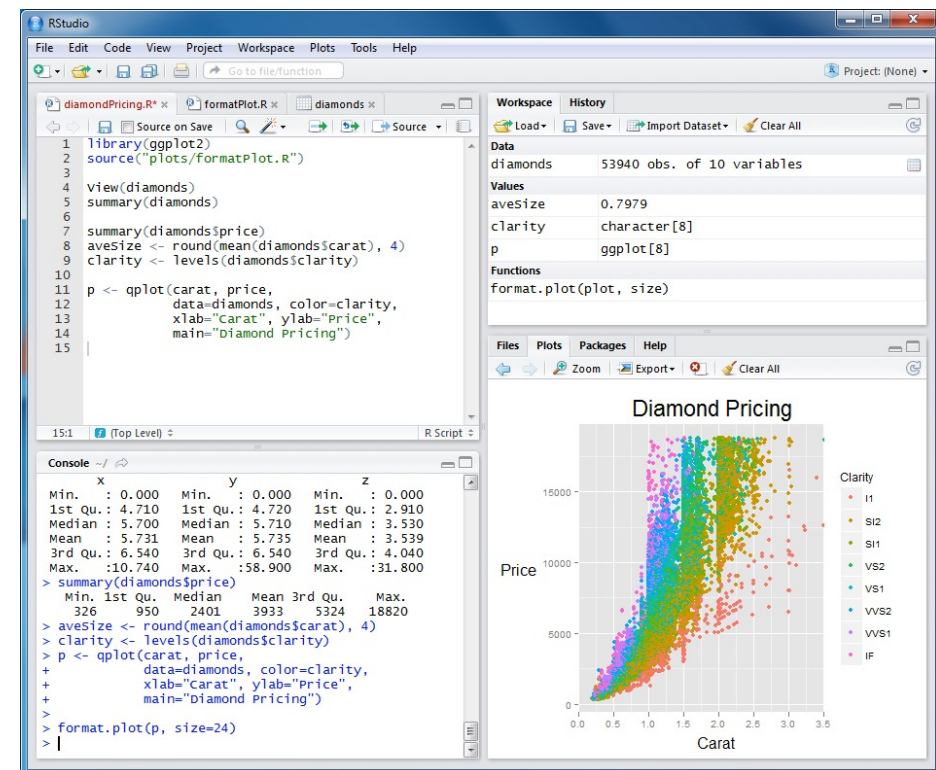
- Instructions on downloading and installation are on website (see www.rstudio.com)

RStudio is meant to be an easier interface to the R Statistical Package.

- Very powerful, and we will only scratch the surface of its use.
- Takes a lot of time to learning, so my examples are going to be limited.

Most popular data mining package (open source), SAS and SPSS are popular commercial packages.

Or an alternative is a general programming language like python with data science libraries.



Why R?

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation.

Features

- Object oriented
- Free, Open source, Widely used
- CRAN Community support (4800+ packages)
- Cross-platform (Linux, Windows, Mac)
- Works with other tools (odbc, sql, ...)
- Easy to extend (web scraping, ...)
- Graphics

Detractions

- Learning curve
- Command driven (but has GUI front-ends like RStudio and Rattle)
- Documentation can be terse, especially for non-statistician and non-programmer
- Not all packages are robust
- No one to complain to if it doesn't work (except for S-Plus)
- Memory bound (unless you use GraphLab's sFrame)



Learning R

We will be using RStudio to do data mining. This can handle many small to medium sized problems, but large problems (e.g. enterprise level) often require specialized software.

Recommendation:

- watch the demonstrations
- read the handouts
- immediately do on own.

Welcome to follow along, but you will get lost at times: don't get frustrated or feel stupid. Work in groups, and spend time exploring.

Will have further information to help provide assistance for those interested



Learning R

My focus is on using data mining tools to solve problems, but we need to have a tool to do this which is R

R is a full-featured data analysis environment. Really three parts: scripting, packages, solving problems.

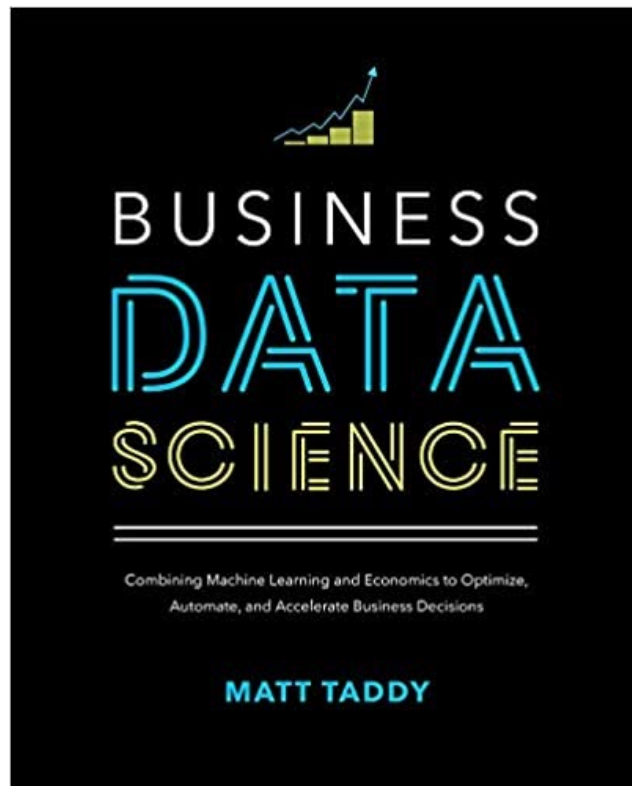
- To become expert takes more time than we have, so I take a “guided” learning approach. I post scripts that allow you to complete (or nearly complete) the exercise. If you delve more deeply into the scripts then you can become proficient – but you will have to be willing to put additional effort.

Lot's of resources to learn R, check out moodle. The right one depends upon you:

- Programmer: <http://tryr.codeschool.com>
- General: <https://www.datacamp.com/courses/free-introduction-to-r>



Required Textbook



Business Data Science: Combining Machine Learning and Economics to Optimize, Automate and Accelerate Business Decisions (1st Ed)

by Matt Taddy

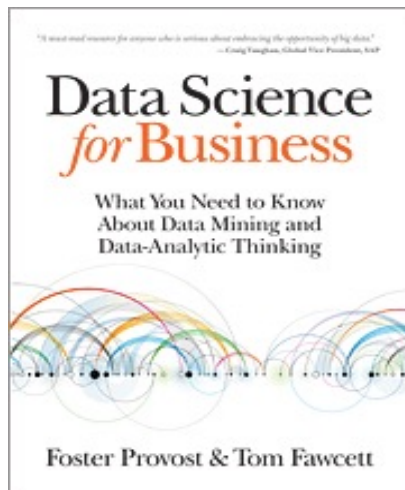
Good coverage of the topics that we cover, but his emphasis will be different than what is covered in class. It highlights the important concepts but does not have full depth or background. Please see optional books for additional background content.

Optional, Supplemental Book

For those that need more conceptual help

Data Science for Business

By Foster Provost & Tom Fawcett



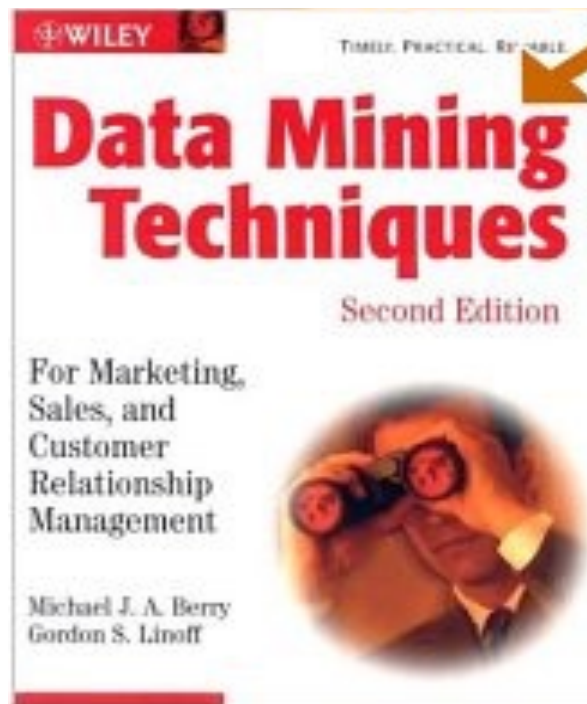
Introduction to Data Science : Using the R Language for Statistical Computing and Graphics

By Jeffrey M. Stanton



Optional, Supplemental Book

For those that want more business examples



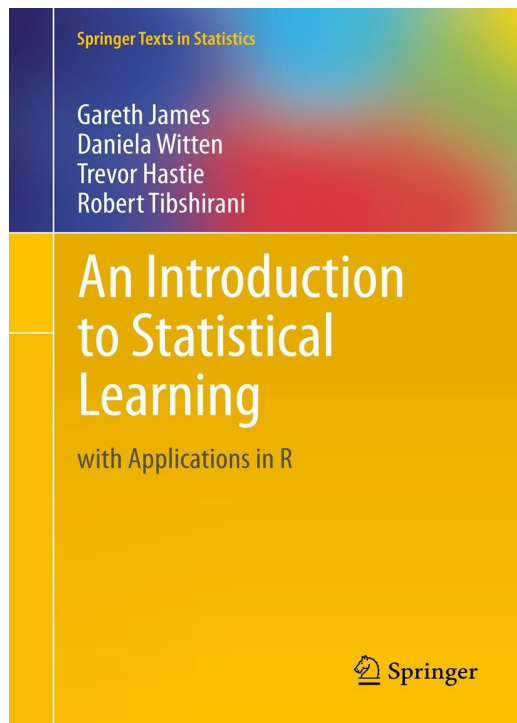
Data Mining Techniques (2nd Ed)

by Michael Berry and Gordon Linoff
(Amazon, etc.)

More on applications of data mining in business. Helpful if you want more background on usage of data mining.

Optional, Supplemental Book

For those that want more technical information



An Introduction to Statistical Learning: With Applications in R

by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani

An introduction to statistical learning methods aimed at masters students in non-mathematical sciences.

The book is available for free:

<http://www-bcf.usc.edu/~gareth/ISL/>



Optional Background Reading

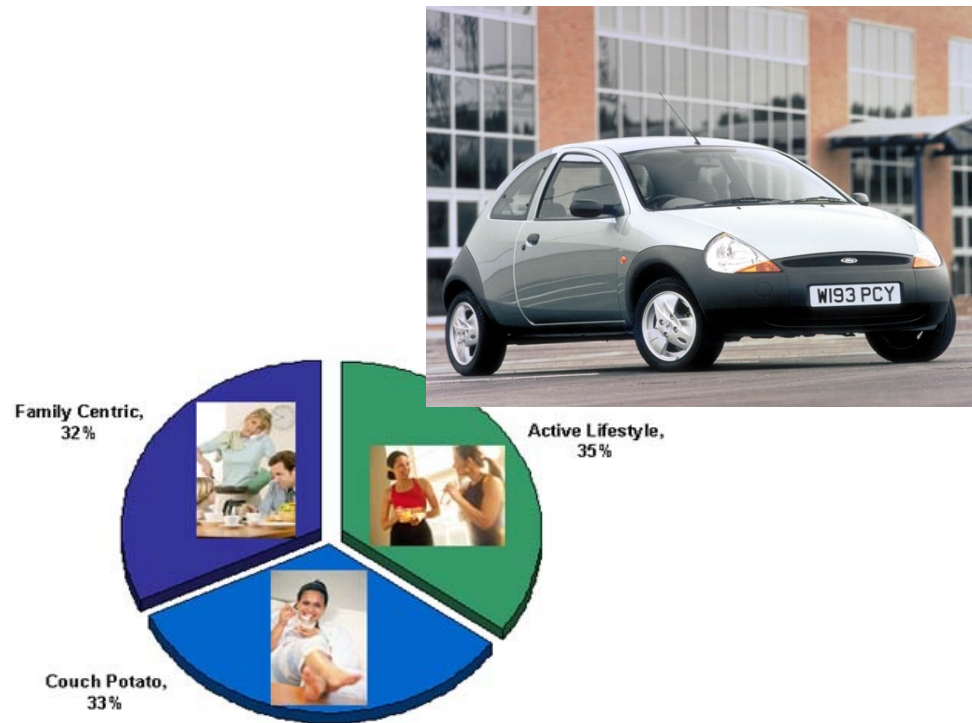
More Technical Materials

Here are some suggestions, if you want more technical material concerning the algorithms (some may require more background in statistics and linear algebra):

- *Handbook of Statistical Analysis and Data Mining Applications* by Nisbet, Elder, and Miner
- *The Elements of Statistical Learning: Data Mining, Inference and Prediction* by Hastie, Tibshirani, and Friedman
- *Data Mining: Practical Machine Learning Tools and Techniques* by Witten and Frank
- *Making Sense of Data*, by Myatt



Market Segmentation for Ford Ka

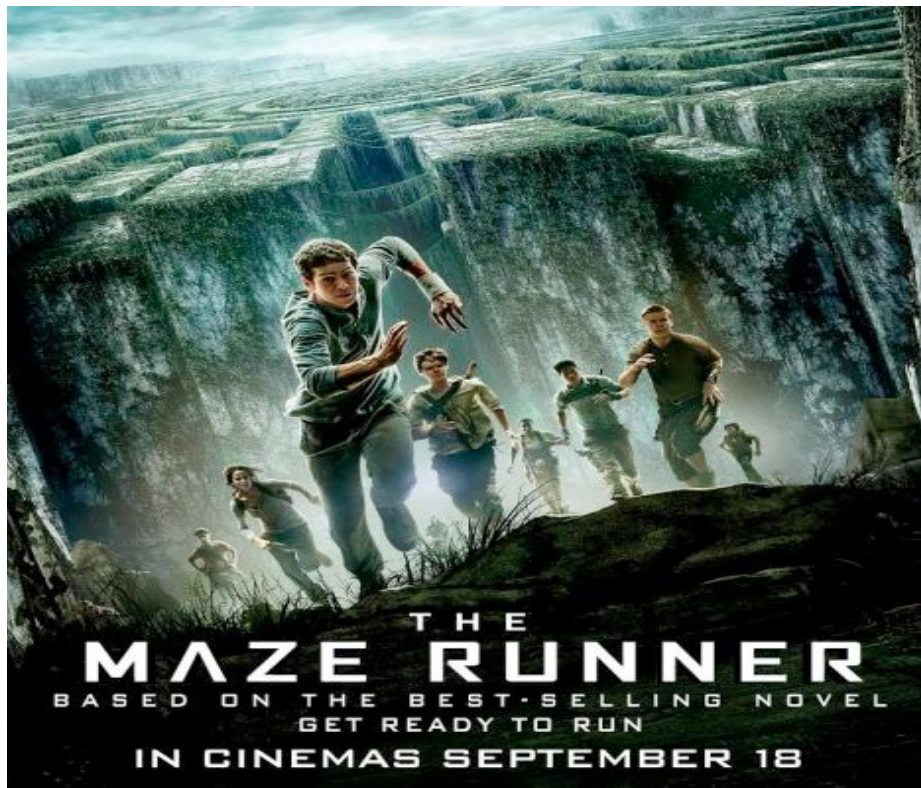


Cluster analysis identifies groupings that are similar so that large numbers of observations can be categorized into smaller, representative groupings. This is frequently used to identify market segments.

Contrast psychographics versus demographics for segmentation.

Movie Scheduling

What week to release?



Often we do not have numeric information, but we have textual information. For example, this movie can be described as: “based on novel”, “escape”, “teenager”, “death”, “human experiment”, ...

How can we use a topic model (or probabilistic cluster model) to compare movies? Let's use these comparisons to decide the week of release by looking for weeks with less competition

'freemium' business models

How do we get people from free to fee?



Oestreicher-Singer & Zalmanson (2013)



Predicting Customer Churn at Cell2Cell

Develop a model for predicting customer churn at “Cell2Cell”, a fictitious wireless telecom company, and use the insights to develop an incentive plan for enticing would-be churners to remain with Cell2Cell



Making lending recommendations

Who to give a loan to?

Banks lend their own money, but Lending Club makes recommendations to Investors on lending to Borrowers

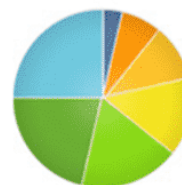
The goal is to be able to predict who will pay back (or default) with as high a precision as possible.

Potentially want to maximize profits, ROI, or satisfaction

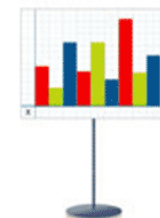
How Lending Club Works



Borrowers apply for loans.
Investors open an account.



Borrowers get funded.
Investors build a portfolio.



Borrowers repay automatically.
Investors earn & reinvest.



Potential Changes

Challenges with format due to COVID

- I did not want to have a 3 hour zoom session, I thought it would be better to break up our contact time into 2 hour zoom + 1 hour video + 30 minutes of Q&A

I try to avoid changes but sometimes I make *minor* changes in the schedule

- Move content when I fall behind
- Add content/examples if concepts are not clear

Technology Concerns

- What happens if our zoom fails (?)



Questions?

Pre-lecture survey

About the syllabus?

About learning objectives?

About R?

About homework?

Anything else?

