IA|BE Data Science Certificate

Module 1 on Foundations of machine learning in actuarial sciences
Knowing me, knowing you - data science meets insurance

Katrien Antonio
LRisk - KU Leuven and ASE - University of Amsterdam

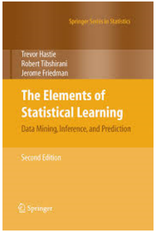
October 5, 2021

https://katrienantonio.github.io

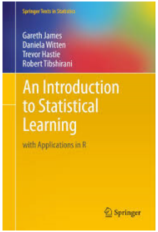
Analytics: what's in a name?

(Data) analytics or data science or data mining or predictive modeling or . . .

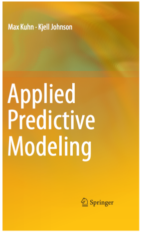
. . . refers to a vast set of tools for understanding data.



Hastie, Tibshirani & Friedman



James et al.

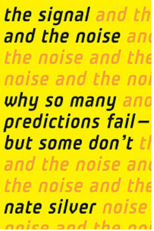


Kuhn & Johnson

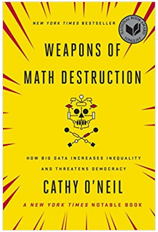
Analytics: what's in a name?

(Data) analytics or data science or data mining or predictive modeling or . . .

. . . refers to a vast set of tools for understanding data.



Nate Silver



Cathy O'Neil

Hannah Fry

Analytics: supervised learning

Determine the structural function f ('the Signal') such that outcome or target Y can be written as

$$Y \quad = \quad f(x_1, \ldots, x_p) + \epsilon,$$

with features $x_1, \ldots, x_p$ and error term ϵ ('the Noise').

- **Main questions:**
  - What are the relevant features x_i to be included (with predictive power)?
  - What does the structural function f look like?
  - How can features x_i be constructed from (continuous) data?

Analytics: unsupervised learning

Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be the feature vector. Assume we have n (possibly noisy) observations of such a feature vector:

$$\mathcal{F} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).$$

There is **NO** target or outcome Y!

- **Main questions:**
  - Find patterns and differences in these features $\mathcal{F}$ ('pattern recognition').
  - Reduce the dimension of $\mathcal{F}$ to a small set of useful features ('dimension reduction').

(Provocative) Statement Nr. 1

Doing data science - Straight talk from the frontline

What is the eyebrow-raising about big data and data science?

The hype is crazy.

Getting past the hype?

There might be some meat in the data science sandwich. Data science, as it's practiced, is a blend of Red-Bull-fueled hacking and espresso-inspired statistics.

Quote from *Doing data science - Straight talk from the frontline*, by Rachel Schutt and Cathy O'Neil, 2013.

What data scientists really do

- Technical tasks of a data scientist:
  - identifying models, including selecting/building appropriate features
  - training the models and testing their performance
  - interpreting the results and re-evaluating model selection
  - visualization of data and findings.
- Technical skills of a data scientist:
  - programming (e.g. R or Python), including standard packages for machine learning and visualization.
  - proficient knowledge of machine learning techniques and how they differ from each other.

Insurance analytics

- The actuary plays a central role in data analysis and predictive modeling:
  - insurance pricing and product development
  - reserving and accounting
  - risk management and Nat Cat modeling
  - marketing
  - claims handling.
- "All these actuarial fields go through massive, data driven changes."

  (quoting prof. Mario Wüthrich, ETH Zurich)

(Provocative) Statement Nr. 2

Aviva's CEO

> From a skills perspective, Wilson is aware of the need to reskill employees to navigate this digital era; for example, retraining actuaries to become data scientists. ‘I'm desperate for that skill set but universities don't train people in it. I'm willing to pay more for a data scientist than an actuary,‘ he reveals.

Quote from *Reviving Aviva: Exclusive interview with Mark Wilson*, published on May 29, 2018.

☺ IA|BE Data Science Certificate! ☺

Actuaries of the 5th kind

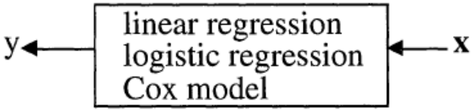
(Picture taken from Data Science Strategy, Working party of the Swiss Association of Actuaries, August 2018.)

Insurance analytics: ‘the two cultures’

- Read the Breiman (2001, Stat Science) paper on *Statistical modeling: the two cultures*.
- The problem:
  - data generated by a black box
  - vector of input variables $\boldsymbol{x}$ go in
  - vector of response variables $\boldsymbol{y}$ come out.

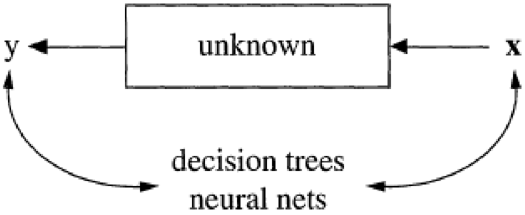
Insurance analytics: ‘the two cultures’

- Read the Breiman (2001, Stat Science) paper on *Statistical modeling: the two cultures*.
- Data modeling culture
  - assume stochastic data model, estimate parameter values
  - validate with goodness-of-fit tests and residual inspection.



Insurance analytics: 'the two cultures'

- Read the Breiman (2001, Stat Science) paper on *Statistical modeling: the two cultures*.
- Algorithmic modeling culture
  - inside of the box is complex and unknown
  - find algorithm $f(\boldsymbol{x})$ to predict $\boldsymbol{y}$
  - measure by predictive accuracy.



Insurance analytics: ‘the two cultures’

	Statistical Learning	Machine Learning
origin	statistics	computer science
$f(X)$	model	algorithm
emphasis	interpretability, precision and uncertainty	large scale applicability, prediction accuracy
jargon	parameters, estimation	weights, learning
CI	uncertainty of parameters	no notion of uncertainty
assumptions	explicit a priori assumption	no prior assumption, learn from the data

(Taken from Why a mathematician, statistician and machine learner solve the same problem differently.)

Actuarial learning: (some) challenges

- Past/Present

  Risk classification in competitive markets using (standard) regression models ($\sim$ GLMs) for frequency and severity.

- Ongoing

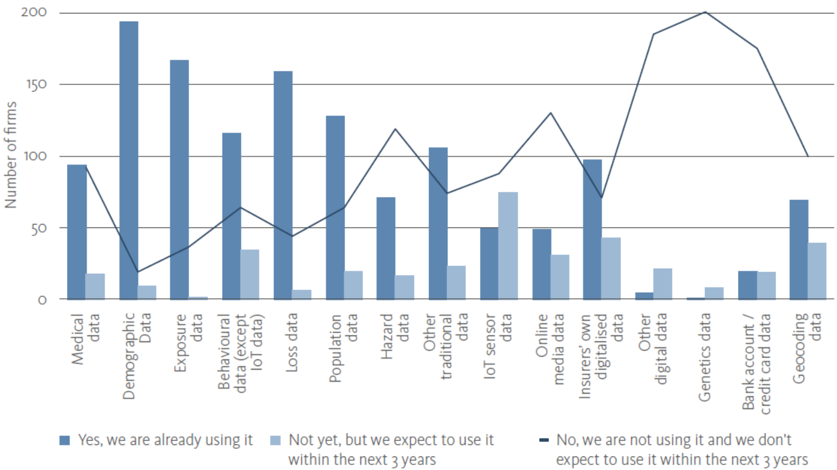
  From statistical learning to machine learning with shallow but also deep learning techniques.

  New data sources (structured, but also unstructured).

- Challenges?
  - keep model explainable to clients, regulators, ICT
  - !!be aware of specific features of insurance data!!

Actuarial learning: (some) challenges

Figure 3 – Usage of different types of data



Source: EIOPA BDA thematic review

Actuarial learning: (some) challenges

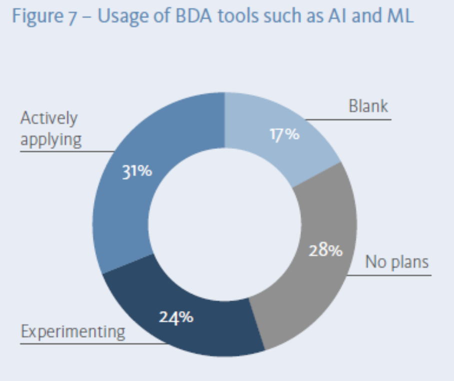
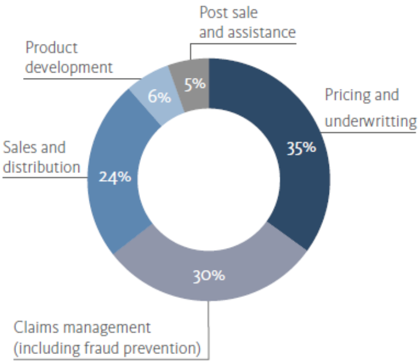


Figure 8 – Usage of BDA tools such as ML and AI across the value chain



Source: EIOPA BDA thematic review

Use cases

Figure 9 – BDA uses cases

Use Case	Output
Churn models	Use of ML churn models for the prediction of consumer's propensity to shop around at the renewal stage, which can be useful for pricing and underwriting (e.g. for price optimisation in combinaiton with a demand price-elasticity analysis) or for servicing the customer (e.g. "Next Best Action" approach)
Chatbot	Enable "human like" conversations with consumers by analysing customer unstructured data via text or voice with the use of natural language processing and other ML algorithms
Sentiment Analysis	Evaluate the sentiment in feedback provided by consumers to transform it into usable information to help improve customer satisfaction and engagement
Electronic document management	Robotic process automation (RPA) – Deep learning networks used for automatic classification of incoming documents of unstructured data (e.g. emails, claims statements), routing them to the correct department
Claims management	Optical character recognition (OCR) - Deep learning networks used to extract information from scanned documents such as images from damaged cars to estimate repair costs
Fraud prevention	Analysis of fraudulent claim patterns based on FNOL data provided by the consumer
Product development	Use of ML and graph database in predictive modeling for the identificaiton of disease development patterns
Pricing and underwriting	BDA tools used in motor and health insurance for processing large quantities of data from different sources, often on a real-time basis (e.g. quote manipulation), using a wide array of statistical techniques

Source: EIOPA BDA thematic review

Big data and insurance: changing role of data

- Aggregated personal data; personal information directly collected from policyholders.
- Data from third-party data sources.
- IoT and the Digital Society produce continuously large amounts of real-time data.
  - online behaviour
  - sensors built into appliances.

Big data and insurance: ethical and societal concerns

Three categories of concerns

- privacy and data protection
- individualisation of insurance
- implications for competition.

These are not new, though become more prominent in the big data era!

More on this will be covered in Module 3 of the IA|BE Data Science Certificate!

Big data and insurance

Read:

Big data and insurance. Implications for innovation, competition and privacy by the Geneva Association (2018).

Big data analytics in motor and health insurance: a thematic review by EIOPA (2019).

(Provocative) Statement Nr. 3

The mindset of the actuary - course ambition

> The narrative must be that actuaries are entering the data science world not entirely to compete but also to bring the element of the actuarial profession where we build integrity and transparency into any work that we do, and how documentation of that is possible.

Quote from What data science means for the future of the actuarial profession, British Actuarial Journal, June 2018.

Check-list for an insurance data science project

A take home message:

- Is it (technically) possible?
- Is it allowed (by regulation)?
- Should we do it (cfr. reputation of the company)?

Outlook

Common themes in my lab's research lines:

- open the black box (as much as possible) and document
- fill methodological gaps that arise when working with insurance data
- analyze real life data.

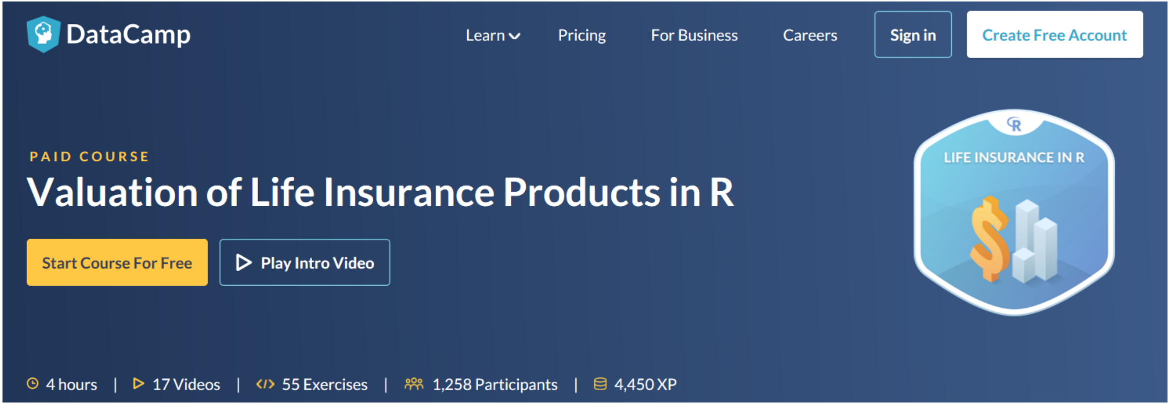
More information

For more information, please visit:

LRisk website, www.lrisk.be

https://katrienantonio.github.io

Thanks to



Online course with DataCamp on Valuation of Life Insurance Products in R

designed by Katrien Antonio & Roel Verbelen

http://www.datacamp.com/courses/2333