What's Data Science Anyway?
Data Expert's Guide to Business

DR. MARC LANGE*, GITANJALI NAIR
*marc.lange@elbformat.de

elbformat

How to read this...

When you know what machine learning pipeline you want to implement, it is usually not the biggest problem to learn what models are involved. Hence, consider this a poster format cheat sheet for analytics cases, where it is less clear. Business users will come with questions and tasks, which need to be massaged into sequences of data sources and machine learning analyses.

Usually you have quite the flexibility to look at the same business question: Take the simple example: *Can you work out which customers made us most revenue?* You now have the flexibility to make that question precise:

- **Make a regression** on revenue data and find the most relevant features. This can be applied for targeting parameters in marketing for example.
- **Classify customers** into high-and-low-value. Depending on the business model that classification can be infeasible or a quick-win for business. This knowledge will transfer into a lot of departments.
- **Cluster customers** in general and see, if the resulting clusters have any clear relation to business value? It will definitely inform what groups of customers the business has, and allow tailored products.

Supervised Learning

TYPICAL QUESTION STYLES:

- Which customers are high-value?
  – in €
  – as subscribers
  – easily retained?
- Which customers are at risk to leave?
- What's the main topic of this article?
- To which support team does this message belong?
- Is there a dog in this picture?

METHODS:

- Regression: Linear, Ridge, ...
- Classification: kNN, SVM,
- Just a spreadsheet?

	A	B	C	D	E	F	G	H
1	age	5	6	7	8	9	10	11
2	5	10	10	10	10	11	11	10
3	10	10	10	11	11	10	10	11
4	15	11	12	11	11	10	11	11
5	20	12	13	12	12	11	11	11
6	25	13	14	13	12	11	10	10
7	30	13	14	14	12	11	11	11
8	35	13	13	13	12	11	10	10
9	40	12	13	13	12	10	11	10
10	45	11	12	11	11	11	10	11
11	50	11	10	11	11	12	11	11
12	55	11	11	12	13	13	13	12
13	60	11	13	14	15	14	14	14
14	65	12	13	15	15	15	15	14
15	70	12	14	15	15	15	16	15
16	75	11	13	14	15	15	15	14
17	80	10	12	14	14	15	14	13
18	85	11	11	12	13	13	13	13
19	90	11	11	10	12	12	12	11

Side Note Business colleagues will occasionally use the word *clustering* to refer to some of the questions above. Don't be mislead, it could be a classification task for example.

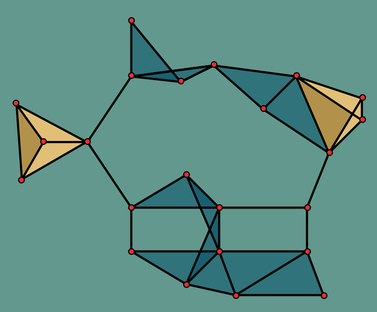
Unsupervised Learning

TYPICAL QUESTION STYLE:

- We have all these server monitor KPIs, can we show the *most surprising* automatically?
- We want to migrate this big stock of articles, can we *structure* them somehow?
- Is this customer behaving *normal* in our shop?

METHODS:

- Clustering: k Means, Spectral, (H)DB SCAN,..
- Anomaly Detection
- Autoencoders
- Topological Data Analysis



Side Note Unsupervised Learning tasks are an effort to understand the *shape* of your data. As such it might not be an answer to a business question itself, but an intermediary result of feature engineering, just a helper for visualisation and the like. Don't underestimate the added value a good communication of these results can have to business knowledge though!

Communication

As data people, communication of the results is a core part of our job. Business results are much shorter than academic results. It's not about defense, but relaying information. Don't be intimidated, if they do find an error, it means you communicated it well. We empower data-driven decisions for people who need us to use our toolbox to empower their decisions.

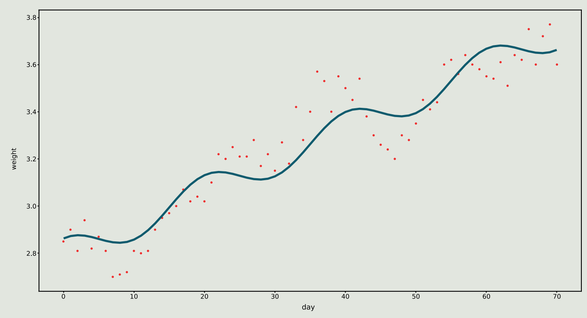
It is one of our skills in permanent training to make an analysis that took half a year into maybe 5 lines of an email to CEO, CTO and other decision makers.

AUDIENCE AND MODES:

- CEOs and similar: usually mails, the shorter, the better
- Product Managers of development teams: a meeting with them or whole team depending on the impact of the result
- Business Units like marketing, sales, HR: possibly manager, usually team
- Customers, colleagues, writers .. ask yourself who could profit from your discoveries, take charge and ownership of your result, ask if they want to know more

MEDIA STYLES:

- Plain text
- Plots, Slides, Spreadsheet (maybe with tunable parameters)
- Tableau, Plotly/Dash, or similar dashboards



Dimension Reduction and Visualisation

When you already have your data prepared, i.e. you know how you make your images, articles, etc. into data that is actually digestible by machine learning, you might still be disoriented what to do next. Dimension reduction and the resulting visualisations help you in the beginning stages, they inform your intuition, and of a few hundred pictures you will probably find the one on the way that helps with communication of your results later.

METHODS TO TRY:

- universal, cheap, brutal: PCA (un/centered), SVD, ..
- universal, expensive, a bit safer: t-SNE, MDS, ..
- occasional: feature elimination with e.g. multilinear regression

We Can Help!

- Do you need a data expert to coach or organise your business or developer team into a data unit, so they have a head start to lift your data competence?
- Do you need someone to assist with your machine learning pipelines for a few months until the full time data scientist starts her magic?
- Do you have a pool of data that no one has adequate time to analyse?
- Do you want to include data driven decisions into your business, but don't know where to start?

WE CAN HELP WITH THAT AND MORE!

Find us at www.elbformat.de
or directly contact
marc.lange@elbformat.de