A Beginners Guide to Data Science

Data Science and Machine Learning can come across as complex, and many businesses are unsure about how they can make use of the two.

What is Data Science?

Although Data Science has received a lot of attention in the media, it is still a relatively recent term and almost always mentioned alongside Machine Learning, AI or Big Data. Many people find themselves asking what exactly Data Science is, and what it is good for.

Data Science is a specialism using methods that are distinct from familiar ones in software development, data analysis and statistics, whilst at the same time borrowing heavily from all of these fields. As you might have guessed, those methods include machine learning.

So what is Machine Learning*?

On the surface, it sounds self-explanatory: a machine that can learn. But what is so new and powerful about machine learning?

Machine Learning is an umbrella term for all kinds of learning algorithms, ranging from the simple line-of-best-fit Linear Regression, to the more advanced e.g. Neural Networks, Deep Learning and Reinforcement Learning. Non-machine learning algorithms are explicitly programmed to solve a task by the programmer. Any algorithm that can learn from historical data how to solve the task is classed as machine learning; think of it as learning the underlying patterns in the data to solve the task at hand.

The Data Science Process

What kinds of activities can we expect to see in a Data Science process? What kinds of tools would be used? How long does a Data Science project take?

Below I will explore the main features of this process and the tools that they typically use. It is worth mentioning that there are many tools out there for partially automating these steps.

What can Machine Learning do?

Machine Learning is successful at a growing number of tasks. Here are some useful subcategories of problem types:

- Reading, interpreting and creating text.
- Interpreting photos and video.
- Forecasting (specifically focused on data where there is a time dimension, this is an area where machine learning has been used for some time and continues to make an impact).
- Optimisation deserves a special mention, as although it isn't a machine learning method (as it isn't learning from historical data), it is a very powerful toolbox that is proven in many industries, and often competes with machine learning methods.

Step 1) Understand the Business Problem

Tools used: At this sensitive stage of the project, a data scientist will need to use their interpersonal skills as well as their research of the problem to be effective.

Step 2) Data Collection

Tools used: Collaborating well with people experienced in a problem the data scientist may be new to, getting up-to-speed quickly with complex data and most likely writing SQL to query that data.

Step 3) Exploratory Data Analysis

Tools used: Statistics, Python, Excel Tableau, or any kind of visualisation software they are comfortable with.

Step 4) Data Cleaning

Tools used: Usually include SQL and Python.

Step 5) Feature Engineering

Tools used: Python.

Step 6) Model Development

Tools used: Python, scikit-learn, TensorFlow, Pytorch, Spacy, NetworkX, to name just a small sample.

Step 7) Evaluation

Tools used: Python aplus whichever package was used for model development.

Step 8) Trial

Tools used: Back to collaborating with stakeholders and the wider business.

That is the data science process in a nutshell. Each business will have slightly different steps, and if you'd like to discuss any of these steps in greater detail, or discuss Data Science more generally, please drop me an e-mail at alexander.tarroni@cgi.com.

*Machine Learning is often called by its more handsome name, Artificial Intelligence. This tends to evoke mostly negative sci-fi images of Hal from Space Odyssey: 2001 or more recently Ava from Ex Machina. I think this terminology is a bit misleading from a business perspective as these are examples of what AI researchers call Artificial General Intelligence, i.e. intelligence that eclipses human skill and understanding in all areas and doesn't currently exist. Although this is a genuine project in AI research and one that has genuine existential risks, in practice we are interested in 'Narrow Artificial Intelligence', which is either nearly as good as or better than humans at a specific task. This has the benefit of actually existing and is proving to be a powerful tool in many sectors. The risks associated with narrow AI are more immediate, such as biased models and behaviour change.

About CGI

Founded in 1976, CGI is among the largest IT and business consulting services firms in the world.

We are insights-driven and outcomes-based to help accelerate returns on your investments.

Across hundreds of locations worldwide, we provide comprehensive, scalable and sustainable IT and business consulting services that are informed globally and delivered locally.

For more information
Visit www.cgi.com/uk/en-gb/emerging-technologies
Email us at enquiry.uk@cgi.com