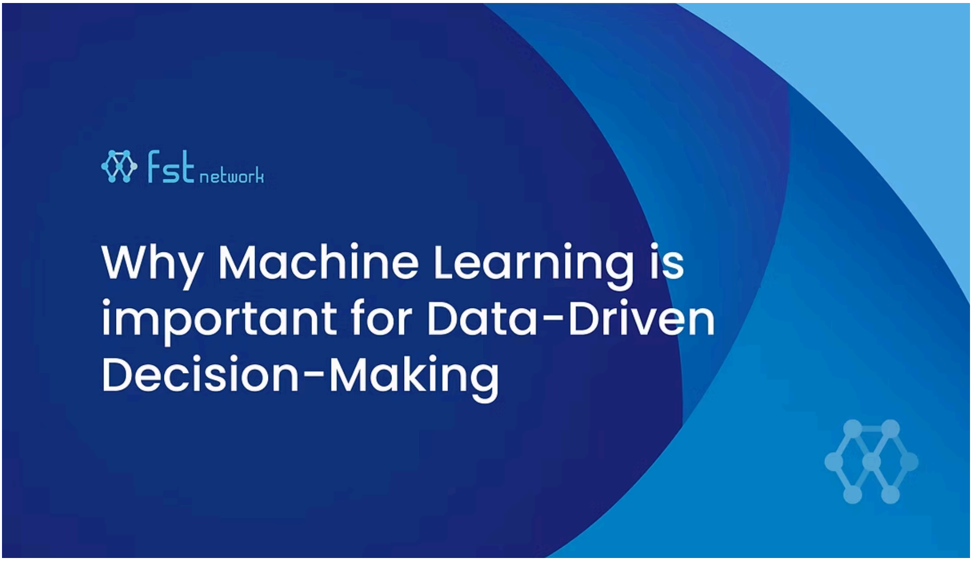
Dec 9, 2022 · 7 min read

Why Machine Learning is Important for Data-Driven Decision-Making

- the New Norm for Smarter and Faster Data Insights


fst network
Why Machine Learning is important for Data-Driven Decision-Making

In Gartner's Hype Cycle for Data and Analytics Programs and Practices (2022), we can see many of the data trends share one aspect: artificial intelligence (AI). In the world of software technologies, AI is in fact the synonyms of machine learning (ML).

But what exactly is *machine learning*? A ever fancy term, sounds more like digital *alchemy* than anything else. It is mentioned almost everywhere, not just in the data industry - yet how it works and its relationship with data is often not adequately explained.

The importance of applying machine learning on data is, in fact, no longer a novelty. So in today's article, we will have a little dive into this new-yet-not-new topic.

What is Machine Learning (ML)?

> Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems.
>
> – Machine learning, explained (MIT Management Sloan School, 2021)

In the simplest way to explain possible, machine learning is to **have a machine learn to identify patterns in data**.

Or, more precisely, we "train" a set of machine learning algorithms to recognise *the difference between instances of data*, for example, individual sales records or policy application data. Is a instance of data belongs to group A, group B, or one of the eight other groups? Can the model predicts new results with enough accuracy without human review?

For a lot of tasks, ML are overkill - a simple rule-based algorithm can be just as good. Humans can also learn much, much faster than machines. However there are key exceptions:

1. Machines works 24/7 while humans can get emotional, tired or sick.
2. Some patterns are too complicated or too time-consuming for humans to learn, sometimes even for professionals.

In other words, *we outsource pattern-learning to machines so we can do something else instead*, preferably something only humans know how. The knowledge or experience of pattern is hence stored in the model itself. We can even automate the model training and testing process, than use continuous deployment (CD) to deploy the model as a software service. You then input an instance of data and it will tell you the most possible answer.

The quest of utilising ML in business is already underway. Data lakes from 2010s (and data lakehouses in recent years) are built exactly for machine learning purposes, so that you to dump large amount of potential training data (big data) rapidly into a unstructured data storage. We then extract relevant data with ETL pipelines for the models.

ML-Assisted Decision Making

The root of machine learning is actually *mathematical statistics*. But whereas statistics is about finding patterns in existing data (*data analytics*), ML is about predicting new data with these knowledge using far more advanced math models.

The most common ML application is so-called **supervised learning** - we first provide a training dataset which already contains *answers*. The answers are called *labels* (target values associated with an instance of data), which are usually manually added by humans or extracted from metadata. Thus the ML model would be able to learn the relationship between variables and labels, and use this to “guess“ (or *infer*) unknown labels for new data instances.

There are, in fact, two types of prediction in supervised learning:

1. **Classification** (predict a “class” or a category, which is a discrete value). For example: predicting if a purchase with credit card is an act of fraud or if an insurance policy applicant is likely to breach the contract.
2. **Regression** (predict a continuous value). For example: predicting sales based on historical sales data or service demand based on time of the day and holidays.

(*Deep Learning* (DL) is a subset of machine learning itself, which use multiple-layered neural network models to learn extremely complicated patterns like images and speech. Some advanced models can even draw pictures and write articles. The way they work are similar to human brain, but essencially they are still a form of math algorithms.)

ML models is not just about letting you sneak a peek into the near future. They can also let you know something might be going to go wrong. For instance, you get a sales forecast that is lower than expected - why? You can act sooner to fix the problem before the real impact arrives.

Or, after training the model, you find it can only make predictions with mediocre accuracy. Were we not using the right variables? Or we were simply asking the wrong questions and need to inspect the problem from another angle?

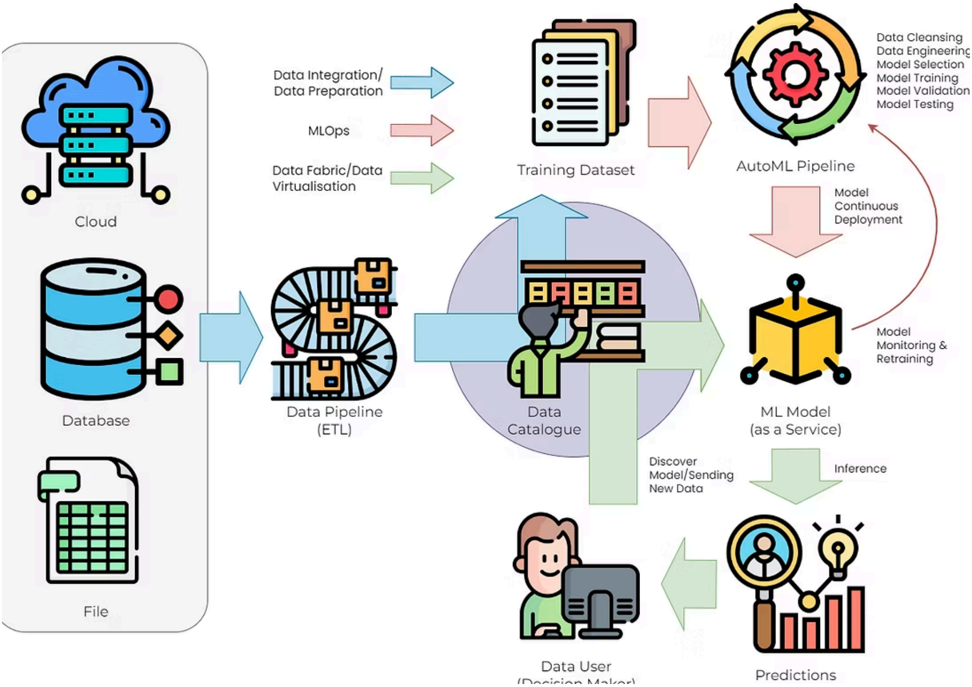
> *“I checked it very thoroughly,” said the computer, “and that quite definitely is the answer. I think the problem, to be quite honest with you, is that you've never actually known what the question is.”*
>
> *“But it was the Great Question! The Ultimate Question of Life, the Universe and Everything,” howled Loonquawl.*
>
> *“Yes,” said Deep Thought with the air of one who suffers fools gladly, “but what actually* is *it?”*
>
> *A slow stupefied silence crept over the men as they stared at the computer and then at each other.*
>
> *“Well, you know, it's just Everything ... Everything ...” offered Phouchg weakly.*
>
> *”Exactly!” said Deep Thought. “So once you know what the question actually is, you'll know what the answer means.”*
>
> – *The Hitchhiker’s Guide to the Galaxy* (Douglas Adams, 1970), while Deep Thought tried to explain why the answer of The Ultimate Question of Life, the Universe and Everything is “42”, the most famous joke in the book series

ML Life-Cycle Automation: AutoML and MLOps

Understandably, ML model can be immensely valuable for assisting business decision-making. However, the traditional way to train, calibration and test models is a lengthy process and still require a lot of human supervision (by experts like data scientists).

AutoML (automated machine learning) once again harness the power of math and machine to automatically fine-tune and select most suitable ML models. MLOps (ML+DevOps) push this even further with data continuous integration, model continuous deployment, model scaling as well as performance monitoring, etc. In other words, most of the ML model life cycle can be automated with a pipeline, just like regular software services or data products at large scale. This is crucial for any enterprises who wish to bring machine learning practice into its organisation. When the newer data pattern has changed, the pipeline should be able to detect it and retrain the model.

Some data solutions we’ve discussed in previous articles, like data integration/ETL and data catalogue tools, can be utilised to locate, extract and transfer the right training/testing data for ML models, as well as to send these data to deployed models for prediction. Both the datasets and the models can be regarded as data assets, and the prediction service can be treated as a discoverable data product, which fits well into the design of data mesh.



Explainable AI and Responsible AI

As wonderful as it is, nothing is perfect or without pitfall; so is machine learning.

We've mentioned that machine learning is to have the machine do the learning for us. However, if you don't understand exactly how ML works, a trained model is essentially a *black box*. How are you going to be sure that the model is performing within the right reasons? Can you convince your stakeholders that the ML models are legit and reasonable enough as the basis of your business decisions?

> Responsible AI (sometimes referred to as ethical or trustworthy AI) is a set of principles and normative declarations used to document and regulate how artificial intelligence systems should be developed, deployed, and governed to comply with ethics and laws. In other words, organizations attempting to deploy AI models responsibly first build a framework with pre-defined principles, ethics, and rules to govern AI.
>
> – Responsible AI: Ways to Avoid the Dark Side of AI Use (altexsoft, 2022)

Sometimes the problems are from the human themselves. If you use a *biased* training dataset - for example, a face recognisation dataset which has far less samples of non-caucasian faces and labels - the ML model might inadvertently become a *racist*. And if you make decisions (especially for international business) based on this model's results, they may severely damage the company's revenue even reputation. (Remember Microsoft's disastrous AI chatbot “Tay”?)

So ML models still first needs careful, comprehensive design and review by human experts so that the power of AI can stay trustworthy and comply company policies, privacy laws even general ethics. This can be applied with modern data techniques, like data governance and data stewardship. Once a model is truly ready and green-lighted, it can be managed automatically afterwards.

AI Does Not Necessarily Make Decisions Smarter - But It will Make a Difference

We must bear in mind that the so-called AI people have been talking about are not really *intelligent* by itself - they do not understand the nature of their input and output and do not know how themselves work. We are still very far from the imagined future of rogue robots, and how ML models are used is still fully controlled by humans (or data officers and data scientists alike).

The point of ML is that we can find patterns in data with reliable computing techniques and use them for predictions - data insights learned as a deployable, reusable digital knowledge - so that we can see a little bit into the future or save time and human resources in critical places. An insurance agent can have an application quickly reviewed by a ML model to see how likely it is a potential fraud, instead of requesting a real employee with much longer queuing time. Any extra saved time and money would be indispensable to gain a little more advantage in the modern data race.

So: now you know why machine learning is so important, especially in the data industry. And one of FST Network’s Logic Operating Centre (LOC)'s goal is to support modern data integration, which of course includes machine learning applications. The data needed in one single dataset may be littered across data silos, and it would sure be nice to have a platform to build reusable, reliable, versatile and traceable data pipelines without much hassle.

FST Network aims to revolutionise the data landscape by creating world's first cloud-native and serverless Data Mesh platform. Logic Operating Centre (LOC) is fully capable to productise and eventify your data, with easy-to-deploy data pipelines that seamlessly connects everything across your organisation, bringing in agile, accessible and trustworthy business insights. The data race is constantly changing and fiercer than ever - but we are there to give you a winning edge.

Icon credits: flaticon.com