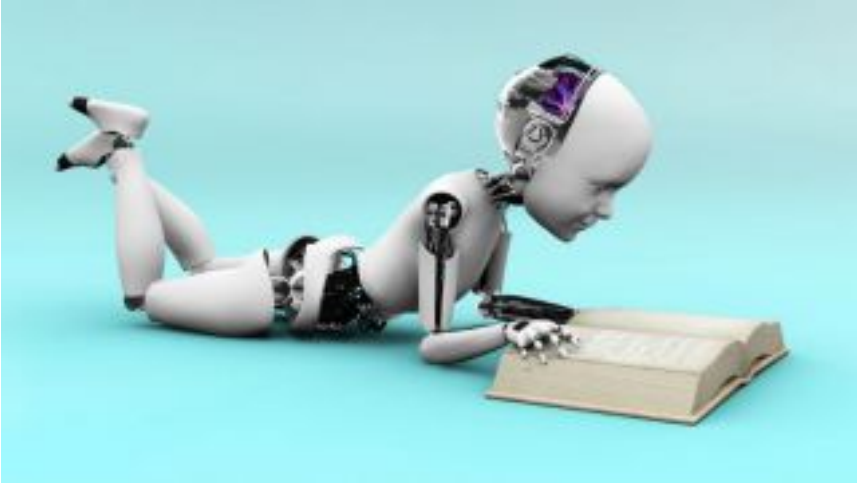
ITProPortal

Machine learning models for mere mortals

By Rafael Zubairov

There is a wide variety of tools that bring this advanced technology to mere mortals.



In our daily lives we are all faced with the need to make decisions. We usually make personal decisions based on personal experience and information. We draw from our friends' experience as well as Internet and other external sources. The data needed to solve a particular challenge usually fits in our head and is structured in sketches in our mind, a notebook or a special document such as mind map. In business, the volume of available data increases in multiples and we are tasked with data collection, analysis and the sheer impossibility of not only holding all the information in our head but even structuring it in a single document. Even before we can attempt to structure the data, we must deal with diverse internal and external data sources and formats.

Decades ago the first computers started to appear with basic spreadsheet management software to help with data analysis and decision making in large companies. Highly specialised software developed for solving a specific narrow range of challenges was giving better results at a faster speed albeit its narrow focus was also limiting its application. Moreover, there were not enough programmers. While the job required not only programming skills, one also needed an understanding of the business domain specific to the task at hand. The combination of these two skills was not common then and is scarce to this day. Bringing together a team that combined these skills was a possibility, but to align their interests, find a compromise and to create a process that would bring immediate results was difficult.

Today we have free access to open source, closed-source, and cloud-based software products that help analysts with their daily tasks. These solutions combine simple functions and modern,

intuitive interfaces for data ingestion and cleansing and model building functionality. Consequently, a wider range of business people are empowered to verify hypothesis and build models of medium complexity without the need to hire software developers.

The vast majority of these solutions were developed in the last few years, enabling businesses to analyse its own data using internal resources. This does not reduce the value that IT departments and outside vendors may bring, especially in the area of data analysis and modelling development scenarios.

It's no magic

We intentionally mentioned data collection and cleansing earlier in this article. Many companies have internal systems for data collection and processing. Data is usually collected from all sources and saved in its raw and cleansed form. However, oftentimes cleansed data is insufficient or not readily available as investigators may find themselves in the tedious time buffer while data is being prepared and delivered to the data warehouse. Deadlines and importance of investigation may warrant the use of raw uncleansed data. Looking into what data science systems offer for preparing data, we can see simple methods of configuring the process of uploading the data from all available sources, whether a database, non - relational database, a file in a shared drive, as well as preliminary data processing. The latter may mean a variety of processes, including filling out empty data slots, and merging and joining of tables. Many of these functionalities can be found in RapidMiner, H2O, DataRobot.
Unfortunately, the majority (up to 80 per cent) of time in data analysis is consumed by data collection and cleansing. Processes that follow are not less important, but less strenuous - feature engineering, model selection and fine tuning are more intellectual, less labour intensive and automated to some level in products like DataRobot, or libraries under AutoML umbrella. When we are choosing the utility for further data analysis, we must take into account not only its functionality but other qualities, including deployment, teamwork, methods of configuring different levels of data access.
It is assumed that the data is valuable to the organisation and internal policies direct the use of exclusively local data centres. In this case convenient tools may be RapidMiner, WEKA and H2O, or DataRobot, albeit it will require a different type of subscription.

Choosing the right software tool is important. Besides flexibility, and intuitive, attractive user interface (UI), factors to consider are the availability of on-premises installation option, capability to run agile data science and share results within the team. For instance, RapidMiner, WEKA and H2O could be easily deployed both in cloud and on-premises scenarios, while tools such as DataRobot would require a different licensing option. Alpine data would provide the team with excellent research-sharing capabilities. On the other hand interactive notebooks like Apache Zeppelin and Spark Notebook can provide higher flexibility, but may require more efforts at integration and deployment stage.
Creating machine learning and data science models in most of the tools is fairly simple. For example, H2O provides an interactive workbook with explanations of each model, ad-hoc suggestions and thorough documentation. Each step of data processing and modelling is represented as a step. RapidMiner on the other hand allows one to create a data flow graph where steps like data processing and model calculation are represented as boxes with configurable properties relative to the action.

We looked at an example of process modelling and training in our recent *article*, illustrating how easy it is to use RapidMiner tool to analyse colds and flu data to predict disease outbreaks.

To conclude we'd like to emphasise that machine learning is not in the realm of magic exclusively accessible to and configurable by data scientists and software developers. There is a wide variety of tools that bring this advanced technology to mere mortals.

Rafael Zubairov, Senior Architect, DataArt
Image source: Shutterstock/Sarah Holmlund

Original article can be found here: https://www.itproportal.com/features/machine-learning-models-for-mere-mortals/