H2O.ai

WHITE PAPER

PRACTICAL GUIDE TO MACHINE LEARNING

MACHINE LEARNING: AN OVERVIEW

You may have heard how companies like Google and Facebook use machine learning to drive cars, recognize human speech and classify images. Very cool, you think. But how does that relate to **my** business?

Consider how these companies use machine learning today:

- A payments processing company detects fraud hidden among more than a billion transactions, and it does so in real time, reducing losses by $1 million per month;
- An auto insurer links losses from insurance claims to detailed geospatial data, enabling it to accurately predict the business impact of severe weather events;
- Working with data produced by vehicle telematics, a manufacturer uncovers patterns in operational metrics and uses them to drive proactive maintenance.

Two themes unify these success stories:

- Each application depends on large scale data sets, in a variety of formats, and at high velocity;
- In each case, machine learning uncovers new insights and drives value.

The technical foundations of machine learning are more than fifty years old, but until recently few people outside of academia were aware of its capabilities. Machine learning requires a lot of computing power; early adopters simply lacked the infrastructure to make it cost-effective.

Several converging trends contribute to the recent surge of interest and activity:

- Moore's Law radically reduces computing costs, and massive computing power is widely available at minimal cost;
- New and innovative algorithms provide faster results;
- With experience, data scientists have accumulated theory and practical guidance to drive value.

Above all, the tsunami of data creates analytic problems that simply cannot be solved with conventional statistics. Necessity is the mother of invention: old methods of analysis no longer work in today's business environment.

MACHINE LEARNING TECHNIQUES

There are hundreds of different machine learning algorithms; a recent paper benchmarked more than 150 algorithms for classification alone. This overview covers the key techniques that data scientists use to drive value today.

Data scientists distinguish between techniques for supervised and unsupervised learning. **Supervised** learning techniques require prior knowledge of an outcome. For example, if we work with historical data from a marketing campaign, we can classify each impression by whether or not the prospect responded, or we can determine how much they spent. Supervised techniques provide powerful tools for prediction and classification.

Frequently, however, we do not know the "ultimate" outcome of an event. For example, in some cases of fraud, we may not know that a transaction is fraudulent until long after the event; in this case, rather than attempting to predict which transactions are frauds, we might want to use machine learning to identify transactions that are unusual, and flag these for further investigation. We use **unsupervised** learning when we do not have prior knowledge about a specific outcome, but still want to extract useful insights from the data.

The most widely used supervised learning techniques include:

- GENERALIZED LINEAR MODELS (GLM): an advanced form of linear regression that supports different probability distributions and link functions, enabling the analyst to model the data more effectively. Enhanced with a grid search, GLM is a hybrid of classical statistics and the most advanced machine learning.
- DECISION TREES: a supervised learning method that learns a set of rules that split a population into progressively smaller segments that are homogeneous with respect to the target variable.
- RANDOM FORESTS: a popular ensemble learning method that trains many decision trees, then averages across the trees to develop a prediction. This averaging process produces a more generalizable solution, and filters out random noise in the data.
- GRADIENT BOOSTING MACHINE (GBM): a method that produces a prediction model by training a sequence of decision trees, where successive trees adjust for prediction errors in previous trees.
- DEEP LEARNING: an approach that models high-level patterns in data as complex multi-layered networks. Because it is the most general way to model a problem, Deep Learning has the potential to solve the most challenging problems in machine learning.

Key techniques for unsupervised learning include:

- CLUSTERING: this technique groups objects into segments, or clusters, that are similar to one another on many metrics. Customer segmentation is an example of clustering in action. There are many different clustering algorithms; the most widely used is k-means.
- ANOMALY DETECTION: in fields like security and fraud, it is not possible to exhaustively investigate every transaction; we need to systematically flag the most unusual transactions. Deep Learning, a technique discussed previously under supervised learning, can also be used for anomaly detection.

- **DIMENSION REDUCTION:** as organizations capture more data, the number of possible predictors (or features) available for prediction expands rapidly. Simply identifying what data provides information value for a particular problem is a significant task. Principal Components Analysis (PCA) evaluates a set of raw features and reduces them to indices that are independent of one another.

While some machine learning techniques tend to consistently outperform others, it is rarely possible to say in advance which one will work best for a particular problem. Hence, most data scientists prefer to try many techniques and choose the best model. For this reason, high performance is essential, because it enables the data scientist to try more options and build the best possible model.

HOW TO GET STARTED

If you are interested in machine learning and wondering how to apply it in your organization, there are some concrete steps you can take.

IDENTIFY A BUSINESS PROBLEM. Identify opportunities in your business where improved predictions will have a compelling impact, in the form of increased revenues, reduced costs or some other key business driver. Possible examples include (but are not limited to): detecting and preventing fraud; detecting security risks and threats; measuring credit and default risk; and other high-impact problems. If you can't find problems like this in your business, you're not looking hard enough; every business has opportunities to improve.

CONSULT WITH YOUR ANALYTICS TEAM. You may be surprised to learn that your analysts already use machine learning; if so, that's great. If not, ask: why not? Most analysts are excited about machine learning, and actively seek out business cases where the techniques can drive value. There may be short-term barriers, however, such as a shortage of personnel, lack of software or lack of support from the IT organization. Work with your analysts to diagnose and resolve these barriers.

If you do not have an analytics team, engage a consultant or analytic services provider who can help you build the capability and provide interim support. If your analytics team expresses no interest in driving business value, examine the team's leadership and incentives.

ENGAGE YOUR IT ORGANIZATION. Your IT organization plays a critical role getting your application into production, so it's important to engage them early. IT organizations are sometimes reluctant to introduce advanced analytics into production systems, fearing that "rocket scientists" will tie up the system or bring it down. To address these concerns, make sure that IT helps define the technical requirements for your machine learning software. Your IT team will be very concerned about such things as Hadoop support, the ability to run in the cloud and other things that can make or break your application.

CHOOSE YOUR SOFTWARE OPTIONS. You may be told that your organization already has the software it needs to deliver your application. That's likely not true; machine learning is a rapidly developing field, with significant advances in the past year. Think of it this way: *if your organization already has the software it needs to deliver your application, why isn't your application built already?*

In the section below headed Software Considerations, we outline the most important things to look for in machine learning software. Your analysts and your IT organization will fill in details about such things as Hadoop distributions and cloud platforms.

DEFINE EVALUATION CRITERIA. Your business problem defines your measures of success. Work with your analysts and IT representatives to develop specific and measureable criteria, which should include:

- Measures of prediction success
- Runtime performance for model training and model scoring
- Scaling requirements, measured in data volume (rows and columns)
- Output requirements

If there is an existing predictive model in production, your evaluation criteria should specify the performance thresholds any new software should meet.

PLAN A TRIAL. Work with your analysts and IT team to plan a trial, or Proof of Concept (POC). If you limit the scope to open source software, as we recommend, your out-of-pocket costs will be minimal. (Commercial software vendors ordinarily do not charge for evaluation software; however licensing costs for commercial machine learning software that scales to Big Data starts at seven figures.)

MACHINE LEARNING SOFTWARE REQUIREMENTS

Software for machine learning is widely available, and organizations seeking to develop a capability in this area have many options. The following requirements should be considered when evaluating machine learning:

- Speed
- Time to Value
- Model Accuracy
- Easy Integration
- Flexible Deployment
- Usability
- Visualization

Let's review each of these in turn.

SPEED: Time is money, and fast software makes your highly paid data scientists more productive. Practical data science is often iterative and experimental; a project may require hundreds of tests, so small differences in speed translate to dramatic improvements in efficiency. Given today's data volumes, high-performance machine learning software must run on a distributed platform, so you can spread the workload over many servers.

TIME TO VALUE: Runtime performance is just one part of total time to value. The key metric for your business is the amount of time needed to complete a project from data ingestion to deployment. In practical terms, this means that your machine learning software should integrate with popular Hadoop and cloud formats, and should export predictive models as code you can deploy anywhere in your organization.

MODEL ACCURACY: Accuracy matters, especially so when the stakes are high; for applications like fraud detection, small improvements in accuracy can produce millions of dollars in annual savings. Your machine learning software should empower your data scientists to use all of your data, rather than forcing them to work with samples.

EASY INTEGRATION: Your machine learning software must co-exist with a complex stack of Big Data software in production. Open source software is easier to deploy, modify and integrate into your production workflows. Additionally, look for machine learning software that runs on commodity hardware, and does not require specialized HPC machines or exotic hardware like GPU chips.

FLEXIBLE DEPLOYMENT: Your machine learning software should support a range of deployment options, including co-located in Hadoop or in a freestanding cluster. If cloud is part of your architecture, look for software that runs in a variety of cloud platforms, such as Amazon Web Services, Microsoft Azure and Google Cloud.

USABILITY: Your data scientists use many different software tools to perform their work, including analytic languages like R, Python and Scala; your machine learning platform should integrate easily with the tools your data scientists already use. Well-designed machine learning algorithms include time-saving features.

- Ability to treat missing data
- Ability to transform categorical data
- Regularization techniques to manage complexity
- Grid search capability for automated test and learn
- Automatic cross-validation (to avoid overlearning)

VISUALIZATION: Successful predictive modeling requires collaboration between the data scientist and business users. Your machine learning software should provide business users with tools to visually evaluate the quality and characteristics of the predictive model.

ABOUT H2O.AI

H2O.ai is focused on bringing AI to businesses through software. Its flagship product is H2O, the leading open source platform that makes it easy for financial services, insurance and healthcare companies to deploy AI and deep learning to solve complex problems. More than 9,000+ organizations and 90,000+ data scientists depend on H2O for critical applications like predictive maintenance and operational intelligence. The company -- which was recently named to the CB Insights AI 100 -- is used by over a third of Fortune 500 enterprises, including 8 of the world's 10 largest banks, 7 of the 10 largest insurance companies and 4 of the top 10 healthcare companies. Notable customers include Capital One, Progressive Insurance, Transamerica, Comcast, Nielsen Catalina Solutions, Macy's, Walgreens, Kaiser Permanente, and Aetna.