# 5 machine learning mistakes – and how to avoid them

Machine learning gives organizations the potential to make more accurate data-driven decisions and to solve problems that have stumped traditional analytical approaches. However, machine learning is not magic. It presents many of the same challenges as other analytics methods. In this article, we introduce some of the common machine learning mistakes that organizations must avoid to successfully incorporate this technique into their analytics strategy.

**Machine learning mistake 1: Planning a machine learning program without data scientists**

The shortage of deep analytics talent continues to be a glaring challenge, and the need for employees who can manage and consume analytical content is even greater. Recruiting and keeping these in-demand technical experts has become a significant focus for many organizations.

Data scientists, the most skilled analytics professionals, need a unique blend of computer science, mathematics and domain expertise. Experienced data scientists command high price tags and demand engaging projects.

How to solve it?

Develop an analytics center of excellence. These centers function as an analytics consultancy inside the organization. The center can consolidate analytical talent in one place and allow for the efficient use of analytical skills across the business.

Build relationships with universities. Create an internship program or a university recruiting program to find new talent. You can also tap into university programs that pair students with businesses to help solve problems.

Develop talent from within. Look for employees who have a natural aptitude for mathematics and problem solving, and invest in data science training.

Make analytics more approachable. If your data visualization tools are user friendly and data is easy to explore, others in the business can solve problems with data, too, not just data scientists.

**Machine learning mistake 2: Starting without good data**

While improving algorithms is often seen as the glamorous side of machine learning, the ugly truth is that a majority of time is spent preparing data and dealing with quality issues. Data quality is essential to getting accurate results from your models. Some data quality issues include:

Noisy data. Data that contains a large amount of conflicting or misleading information.

Dirty data. Data that contains missing values, categorical and character features with many levels, and inconsistent and erroneous values.

Sparse data. Data that contains very few actual values, and is instead composed of mostly zeros or missing values.

Inadequate data. Data that is either incomplete or biased.

Unfortunately, many things can go wrong with data in collection and storage processes, but steps can be taken to mitigate the problems.

How to solve it?

Data security and governance. Address data security issues at the beginning of a machine learning exercise, especially if support from other departments is required. Likewise, early plans for data governance should consider how algorithms will be used, stored and reused.

Data integration and preparation. After data has been collected and cleaned, it must still be transformed into a format that is logical for machine learning algorithms to consume.

Data exploration. Productive, professional machine learning exercises should start with a specific business need and yield quantifiable results. Data scientists must have the ability to efficiently query, summarize and visualize data before and after machine learning models are trained, and build algorithms as new data is added.

**Machine learning mistake 3: An insufficient infrastructure for machine learning**

For most organizations, managing the various aspects of the infrastructure surrounding machine learning activities can become a challenge in and of itself. Trusted and reliable relational database management systems can fail completely under the load and variety of data that organizations seek to collect and analyze today.

How to fix it?

Planning for the following areas can ensure your infrastructure is built to handle machine learning.

Flexible storage. Design an appropriate, organizationwide storage solution that meets data requirements and has room to mature with technology advances. Storage considerations should include data structure, digital footprint and usage.

Powerful computation. A powerful, scalable and secure computing infrastructure enables data scientists to cycle through multiple data preparation techniques and different models to find the best possible solution in a reasonable amount of time. The following approaches have shown success for machine learning:

Hardware acceleration. For I/O-intensive tasks such as data preparation or disk-enabled analytics software, use solid-state hard drives (SSDs). For computationally intensive tasks that can be run in parallel, such as matrix algebra, use graphical processing units (GPUs).

Distributed computing. In distributed computing, data and tasks are split across many connected computers, often reducing execution times. Make sure you are using a distributed environment that’s well suited for machine learning.

Elasticity. Storage and compute resource consumption can be highly dynamic with machine learning, requiring high amounts in certain intervals and low amounts in others. Infrastructure elasticity allows for more optimal use of limited computational resources and/or financial expenditures.

Many data-driven organizations have spent years developing successful analytics platforms. Choosing when to incorporate newer, more complex modeling methods into an overall analytics strategy is a difficult task.

**Machine learning mistake 4: Implementing machine learning too soon or without a strategy**

Many data-driven organizations have spent years developing successful analytics platforms. Choosing when to incorporate newer, more complex modeling methods into an overall analytics strategy is a difficult task. The transition to machine learning techniques may not even be necessary until IT and business needs evolve. In regulated industries, interpretation, documentation and justification of complex machine learning models adds an additional burden.

How to fix it?

Position machine learning as an extension to existing analytical processes and other decision-making tools. For example, a bank may use traditional regression in its regulated dealings but use a more accurate machine learning technique to predict when a regression model is growing stale and needs to be refreshed.

For organizations with the ambition and business need to try modern machine learning, several innovative techniques have proven effective:

Anomaly detection. While no single approach is likely to solve a real business problem, several machine algorithms are known to boost the detection of anomalies, outliers and fraud.

Segmented model factories. Sometimes markets have vastly different segments. Or, in health care, every patient in a treatment group can require special attention. In these cases, applying a different predictive model to each segment or to each patient may result in more targeted and efficient actions. Using a model factory approach to build models automatically across many segments or individuals allows the implementation of any gains in accuracy and efficiency.

Ensemble models. Combining the results of several models or many models can yield better predictions than using a single model alone. While ensemble modeling algorithms – such as random forests, gradient boosting machines and super learners – have shown great promise, custom combinations of pre-existing models can also lead to improved results.

**Machine learning mistake 5: Difficulties interpreting or sharing model methodologies**

What makes machine learning algorithms difficult to understand is also what makes them excellent predictors: They are complex. A major difficulty with machine learning is that most machine learning algorithms are seen as black boxes. In some industries, such as banking and insurance, models simply have to be explainable due to regulatory requirements.

How to fix it?

A hybrid strategy of traditional approaches and machine learning techniques can be a viable solution to some interpretability problems. Some example hybrid strategies include:

Advanced regression techniques. Knowing when to use advanced techniques is essential. For example, penalized regression techniques are well suited for wide data. Generalized additive models allow you to fine-tune a trade-off between interpretability and accuracy. With quantile regression, you can fit a traditional, interpretable linear model to different percentiles of training data, allowing you to find different sets of variables for modeling different behaviors.

Using machine learning models as benchmarks. A major difference between machine learning models and traditional linear models is that machine learning models usually take a large number of implicit variable interactions into consideration. If your regression model is less accurate than your machine learning model, you’ve probably missed some important interactions.

Surrogate models. Surrogate models are interpretable models used as a proxy to explain complex models. For example, fit a machine learning model to your training data. Then train a traditional, interpretable model on the original training data, but instead of using the actual target in the training data, use the predictions of the more complex algorithm as the target for this interpretable model.

Effective use of machine learning in business entails developing an understanding of machine learning within the broader analytics environment, becoming familiar with proven applications of machine learning, anticipating the challenges you may face using machine learning in your organizations, and learning from leaders in the field.