MACHINE LEARNING FOR DATA MANAGEMENT



Data impacts nearly every part of our lives and is critical for companies to stay relevant. It has transformed almost every industry to drive efficiency, better insights, and business growth.

However, managing this data can take an enormous amount of time and expense. Between security, auditing, organizing, and more, managing data sets can create a significant strain on employee time and energy. In fact, most data scientists and business analysts spend around 80% of their time finding, cleaning, and reorganizing data sets. This leaves just 20% of their time to spend on value-generating activities.

As data scientists become more in-demand and are now difficult to find, their time is even more valuable (and costly) than ever. Streamlining the aspects of their job that are not critical to their expertise can help improve productivity and save money.

Machine learning (ML) can solve this problem. It is a valuable tool for managing and improving efficiency with critical data. The explosion of ML has enabled those with light technical capabilities to handle what was once limited to highly skilled workers.

ML is now one of the biggest trends in data management. The sheer volume and growth of Big Data today have made ML indispensable for most companies. It is ideally suited to play a crucial role in enabling organizations to address their challenges in data management.

Here is what you need to know about Machine Learning, how it improves data management, and tips to successfully implement it.

How machine learning improves data management

Machine learning is a subset of AI that enables computer programming to learn based on experiences. A number of ML and Deep Learning techniques can be employed to help companies perform many crucial tasks, including:

- Address security and compliance challenges
- Schedule SLAs and batch or backup jobs
- Model computations

In the broadest sense, these techniques are divided into three core types:

Supervised learning is where the system is trained with examples of the desired output. Through the use of labeled pairs, the system can map the input to the output based on these examples and decide class labels for the actual inputs. Some of the most common techniques for supervised ML include regression and classification. Also, recommender systems are often based on this type.

Unsupervised learning, where the system learns from unlabeled test data. It recognizes and pinpoints data commonalities and reacts based on either the presence or absence of these similarities in new data. Unsupervised learning can be especially helpful for learning structure in data since users don't have an expected output but instead want to group the data. Some of the most common forms include:

- Neural networks
- Clustering
- Anomaly detection

Reinforcement learning is most often used when action needs to be taken sequentially. One output depends on the one before it, and the next input depends on the current output. In reinforcement learning, the application learns how to achieve a set goal in an uncertain environment. Game development where the game plays against a human player, some recommender systems, and autonomous vehicles use this type of ML.

Each one of these systems helps embed ML-driven intelligence into data management tools.

(Learn more about supervised & unsupervised learning.)

Benefits of ML for managing data

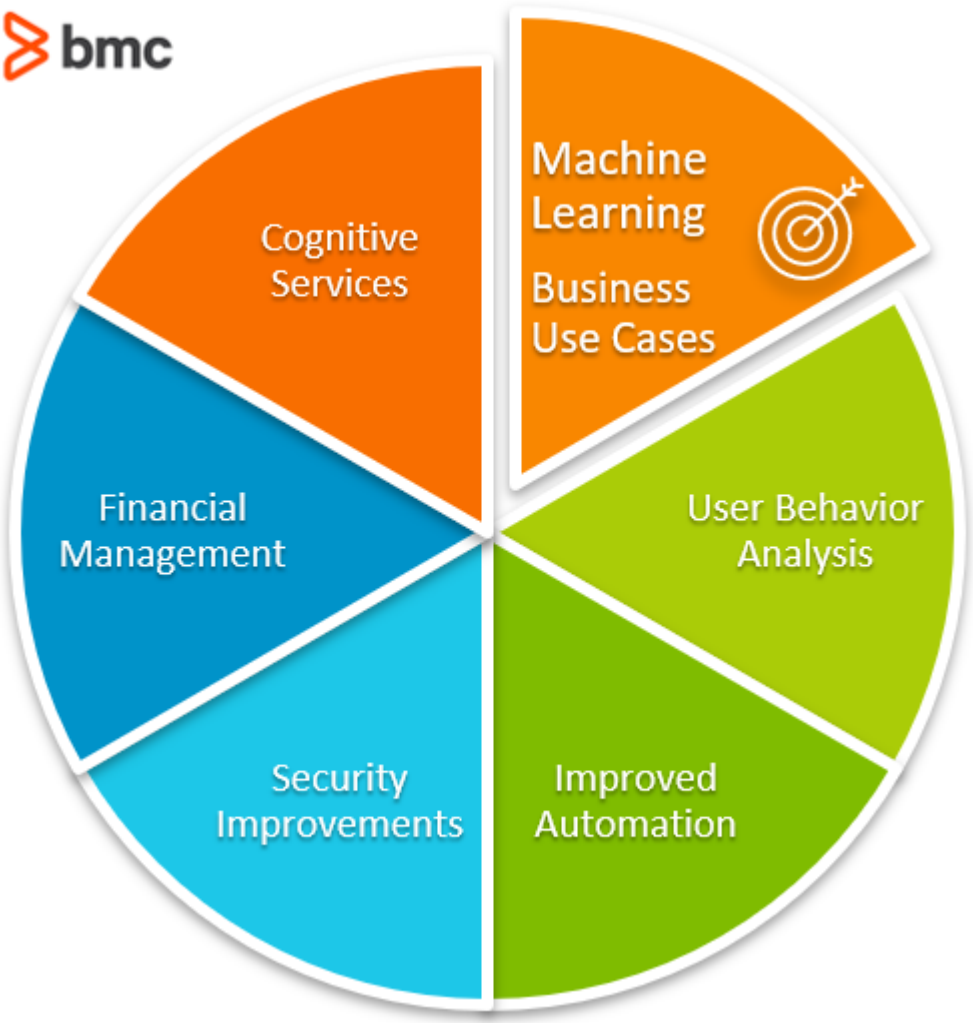
Some of the most significant benefits that ML algorithms can offer in data management include:

- **Optimization.** ML can automatically table join approaches, select distribution methods for data, resource management schemes, and choose query optimization strategies. This can create faster and more responsive system performance.
- **Capacity management.** As data increases, scaling becomes an issue for most organizations. ML can perform workload-aware autoscaling and spot instance purchasing.
- **Automation.** ML can cut down on some of the more time-intensive development tasks associated with data management. A few of the functions it can achieve include mapping sources to targets, onboarding new sources, and cataloging data.

Most of all, ML allows companies the chance to back away from the more traditional rule-based management. Rule-based management heavily relies on human oversight and the ability to predict every potential scenario. Instead, ML works out the right way to help achieve company goals, eliminating much of the heavy workload.

Because of these benefits, ML can offer organizations a distinct advantage for many different users:

- Users who are not highly technical, for example, can do advanced functions that were once limited to data scientists.
- Developers can delegate many of their tasks to others so that they can improve their productivity and concentrate on high-value tasks.
- ML can also help improve system performance that requires less administrator involvement.
- IT will have a significantly reduced burden since they no longer need to manage massive amounts of data.



Where to use machine learning

As more companies realize the benefits Machine Learning offers data management, its use is exploding across all industries. It can be utilized in nearly every business vertical to help improve productivity and accuracy.

With its advantages, ML has a number of use cases to automate and optimize data management:

Anomaly detection

Data collection is only as good as its accuracy. However, identifying outliers and points that don't belong can be a time-consuming process. It is also an area that rarely scales well as volumes of data increase rapidly. ML can work accurately and sift through large datasets. Plus, it constantly adapts to be more precise as it learns with time.

(Try this anomaly detection introduction.)

Data cataloguing

Data collection continues to surge as volume increases each year. ML can help ease the time and energy usually spent organizing the search and discovery, governance, and curation. ML can identify patterns and utilize ML to make the data more user-friendly as it learns user behavior.

It can also help improve GDPR compliance and better ensure privacy functionality.

Data mapping

With ML, businesses can utilize their data more easily because it is organized in manageable and easy-to-understand systems. Organizations can better personalize their marketing efforts and segment their data, as the ML algorithms can identify data and categorize it for future purposes. Plus, it can cleanse data effectively with unification and data cleaning.

Security

Data security is one of the most prominent concerns organizations have today: the average cost of a data breach for a U.S. company is $4.24 million. Machine Learning can help prevent a breach by detecting malicious activity, analyzing mobile endpoints, and automating repetitive security tasks.

Data domains

With ML algorithms, businesses can automatically recognize and catalog data structures and sources into specific domains. It enables people to browse and search the domains that concern them, such as customer or product domains. In some cases, advanced ML can detect domain relationships across various datasets to make browsing and searching even more effortless.

As ML and data continue to grow, the use cases also expand. ML has implications for capacity planning, governance, system performance, and more.

Tips for using ML for Data Management

To get the most out of Machine Learning for data management, consider taking these three steps to start:

1. **Begin with your current domain-specific knowledge.** Consider which processes and rules your employees handle manually to figure out where to start. For example, you might currently have contracts that have been open for too long that need to be addressed. With that understanding, you could build a model to find unmatched contracts.
2. **Find new patterns by automating.** With unsupervised learning and automation, you can utilize ML to pick up on incorrect sequences, typos, or other potential mistakes.
3. **Identify patterns that add value.** Not all patterns and trends in your data are valuable from a business perspective. For example, you might not need to know where your customers are located at this point in your online business. Determine which patterns are useful for your company and validate them against common-sense checks.

These are not one-time-only steps. Instead, continue evaluating where you can implement and incorporate Machine Learning models to improve the learning process continuously. As organizations grow and change, their need for ML will as well. Recognize new areas where ML can improve performance and productivity and assess whether your current use is helpful.

It is also vital that IT teams do not feed all of the data into unsupervised learning models with ML. Teams still need to be involved and ensure that they are not over-fitting models to derive too many insights.

Improving data performance with machine learning

Machine learning has the potential to radically transform how organizations collect, organize, and utilize their data. Companies can better use their data to provide deeper insights and make finding the information they need quickly. With the use of ML, companies can become more agile, adaptable, and efficient.

As businesses collect more data to remain relevant, the productivity of their IT teams often suffers. ML can provide a valuable tool for organizing their data and scaling their operations without compromising accuracy or security. ML can provide a critical role in data management by continuously evaluating ML needs and keeping IT involved.

Related reading

- BMC Machine Learning & Big Data Blog
- Data Management vs Data Governance: Main differences
- Data Analytics vs Data Analysis: What's The Difference?
- Using Python for Big Data & Analytics (Python is Perfect for Big Data)
- Best Books on Big Data & Data Science