O'REILLY®

The Evolution of Analytics

Opportunities and Challenges for Machine Learning in Business

Patrick Hall
Wen Phan
Katie Whitson

The Evolution of Analytics

Opportunities and Challenges for Machine Learning in Business

Patrick Hall, Wen Phan, and Katie Whitson

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

The Evolution of Analytics

by Patrick Hall, Wen Phan, and Katie Whitson

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Nicole Tache
Production Editor: Nicole Shelby
Copyeditor: Jasmine Kwityn

Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

May 2016: First Edition

Revision History for the First Edition
2016-04-22: First Release

While the publisher and the authors have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

978-1-491-95471-3

[LSI]

Table of Contents

The Evolution of Analytics: Opportunities and Challenges for Machine Learning in Business

Over the last several decades, organizations have relied heavily on analytics to provide them with competitive advantage and enable them to be more effective. Analytics have become an expected part of the bottom line and no longer provide the advantages that they once did. Organizations are now forced to look deeper into their data to find new and innovative ways to increase efficiency and competitiveness. With recent advances in science and technology, particularly in machine learning, organizations are adopting larger, more comprehensive analytics strategies.

This report provides a guide to some of the opportunities that are available for using machine learning in business, and how to overcome some of the key challenges of incorporating machine learning into an analytics strategy. We will discuss the momentum of machine learning in the current analytics landscape, the growing number of modern applications for machine learning, as well as the organizational and technological challenges businesses face when adopting machine learning. We will also look at how two specific organizations are exploiting the opportunities and overcoming the challenges of machine learning as they've embarked on their own analytic evolution.

Machine Learning in the Analytic Landscape

Machine learning first appeared in computer science research in the 1950s. So why, after all these decades, has it become so popular?

The easy answer is that both the data storage and the data processing capacities have grown tremendously, to the point where it is now profitable for businesses to use machine learning. A smartphone in your pocket now has more storage and compute power than a mainframe in the 80s, and large amounts of complex and unorganized data that is largely dirty, noisy, or unstructured, is now widely available across a nearly infinite network of computing environments. In order to learn, machines need very granular and diverse data. In the past, the data we collected was often too coarse to train machine learning models, but that has changed. A self-driving car, for example, can collect nearly 1 GB of data every second—it needs that granularity to find patterns to make reliable decisions. It also needs the compute power to be able to compute decisions in time.

Machine learning draws from numerous fields of study—artificial intelligence, data mining, statistics, and optimization. It can go by other aliases and consists of overlapping concepts from the analytic disciplines. Data mining, a process typically used to study a particular commercial problem with a particular business goal in mind, uses data storage and data manipulation technologies to prepare the data for analysis. Then, as part of the data mining task, statistical or machine learning algorithms can detect patterns in the data and make predictions about new data.

When comparing machine learning to statistics, we often look to the assumptions about the data required for the analyses to function reliably. Statistical methods typically require the data to have certain characteristics and often use only a few key features to produce results while machine learning models might use millions (or billions) of parameters in a computer-based method to find similarities and patterns among the data. Machine learning models tend to sacrifice interpretability for better predictive accuracy, but usually accept a wider spectrum of data—text, images, and so-called “dirty” (or unstructured) data. A classic example of a machine learning model is one that is used for pattern recognition. We do not really care which pixels drive the prediction, as long as the prediction is accurate on new data. Another significant difference between the methods and algorithms used in machine learning compared to those in related fields is the level of automation; machine learning algorithms often use learned patterns in unassisted ways.

One broad, representative description of machine learning is computational methods that learn from experience to improve performance or to make accurate predictions. Some of the more popular machine learning algorithms include regression, decision trees, random forest, artificial neural networks, and support vector machines. These models are trained on existing data by running large amounts ofdata through the model until it finds enough patterns to be able to make accurate decisions about that data. The trained model is then used to score new data to make predictions. Some applications for these models include churn prediction, sentiment analysis, recommendations, fraud detection, online advertising, pattern and image recognition, prediction of equipment failures, web search results, spam filtering, and network intrusion detection.

There are a number of learning scenarios, or types of learning algorithms, that can be used depending on whether a target variable is available and how much labeled data can be used. These approaches include supervised, unsupervised, and semi-supervised learning; reinforcement learning is an approach often used in robotics but also used in several recent machine learning breakthroughs.

Machine learning gives organizations the potential to make more accurate data-driven decisions and to solve problems that have stumped traditional analytical approaches, such as those involving new sources of unstructured data, including graphics, sound, videos, and other high-dimensional, machine-generated data.

Machine learning is already in use by a variety of industries, including:

Automotive
: For driverless cars, automatic emergency response systems can make maneuvers without driver input.

Banking
: Big data sources provide the opportunity to market new products, balance risk, and detect fraud.

Government
: Pattern recognition in images and videos enhance security and threat detection while the examination of transactions can spot healthcare fraud.

Manufacturing
: Pattern detection in sensor data or images can diagnose otherwise undetectable manufacturing defects.

Retail
: Micro-segmentation and continuous monitoring of consumer behavior can lead to nearly instantaneous customized offers.

Modern Applications of Machine Learning

> *This increased relevance and value stems from a deeper knowledge about customers, leading to stronger customer relationships and higher sales for the business.*

While some applications of machine learning have been used for many years, enhancements in technology are opening up even greater opportunities. Some of the more modern applications of machine learning include recommendation systems, streaming analytics, and deep learning. We will discuss each of these applications in this section.

Recommendation Systems

Recommendation systems are used across a wide range of industries, most notably online shopping sites, and offer significant value to an organization's current or prospective customers by helping them discover new, relevant offers that are based on their behavior rather than a blanket promotion. This increased relevance and value stems from a deeper knowledge about customers, leading to stronger customer relationships and higher sales for the business.

Consumers often see the results of recommendation systems on the following types of websites:

- Ecommerce
- Movie, TV, book, music, and other media services
- Supermarkets
- Retail investment firms
- Social networking

While the investments in these systems can be profitable, the biggest challenge is getting started. This so-called "cold start" problem is one of the first obstacles a successful recommendation system must overcome, because many organizations don't have the historical data needed to provide recommendations and must first adapt their business processes to capture this data. In addition, creating a feedback loop that tests users' responses should be built into a recommendation system from the beginning so that recommendations may improve over time.

Streaming Analytics

While hourly or even minute-by-minute batch deployment of models remains a best practice for many applications, high-risk or mission-critical tasks don't have the luxury of time. For some organizations, it's now important to make decisions in real time (or near-real time) based on predictions of machine learning models.

Deploying machine learning models in realtime opens up opportunities to tackle safety issues, security threats, and financial risk immediately. Making these decisions usually involves embedding trained machine learning models into a streaming engine.

There are three types of streaming analytics, based on the location of the data that is being processed:

Edge analytics
: Data collected within device (i.e., gateway, sensor, on machinery, etc.)

In-stream analytics
: Data collected between the device and the server (i.e., network security logs)

At-rest analytics
: Data collected at rest (i.e., database of customer information with real-time streaming events)

Real-time data often arrives from sensors and devices, including network activity, financial transactions, and even online web behavior. The types of sensors, collected data, and forms of real-time analytics can also vary across industries. In healthcare, organizations may measure biometric data, such as blood glucose, blood pressure, and heart rate. In the automotive industry, organizations collect data from onboard car safety systems, including speed, oil pressure, temperature, and g-force. In smart cities, municipalities are using sensors to capture information regarding energy consumption, water consumption, occupancy, and much more. Machine learning on streaming data also offers the opportunity for real-time text analytics, allowing organizations to track sentiment, automatically identifyimportant topics and subjects, and extract entities from unstructured data sources.

Deep Learning and Cognitive Computing

Deep learning is a key component in today's state-of-the-art applications using machine learning to conduct pattern recognition in images, video, and sound. It is proving superior to past pattern recognition approaches due to its improved ability to classify, recognize, detect, and describe data.

The primary algorithm used to perform deep learning is the artificial neural network (ANN). Large volumes of data are often needed to train these networks, so when the number of layers increases a lot of computational power is needed to solve deep learning problems. Because of recent advances in research and computational power, we can now build neural nets with many more layers than just a few years ago.

Deep learning is one of the foundations of cognitive computing, a field in which complex machine learning systems can perform specific, human-like tasks in an intelligent way. By combining natural language processing, large databases, and machine learning approaches such as deep learning, cognitive computing allows computers to translate problems from natural language into machine logic, query a knowledge base for potential solutions, and use machine learning algorithms to make decisions about those potential solutions. While the possible applications of cognitive computing are broad and growing, commercial inroads have been made primarily in the healthcare and customer engagement spaces.

Machine Learning Adoption Challenges Facing Business

Machine learning is not magic. It is simply part of the ever-evolving analytical landscape. Machine learning presents many of the same challenges as other analytic methods; it also presents some unique challenges primarily related to complicated and opaque modeling algorithms. In this section, we will review some of the key organizational, data, infrastructure, modeling, and operational and productionchallenges that organizations must address to successfully incorporate machine learning into their analytic strategy.

Organizational Challenges

> *Long-lasting impact can only be made when a fundamental shift in culture is in place to support the change.*

When incorporating machine learning, perhaps the most difficult challenges an organization must overcome are those around analytic talent and culture. There are simply not enough skilled people to develop and execute analytics projects that involve complex techniques such as machine learning, and the change management required to build a data-driven organization is challenging to execute.

Talent scarcity and engagement

The shortage of deep analytic talent continues to be a glaring challenge that organizations are facing and the need for those who can manage and consume analytical content is even greater. Recruiting and keeping these in-demand technical resources has become a significant focus for many organizations.

Data scientists, the most skilled analytic professionals, need a unique blend of computer science, mathematics, and domain expertise. Experienced data scientists command high price tags and demand engaging projects. They prefer the latitude to explore and be creative with their projects, and this often does not match a siloed departmental structure. One approach to solving this problem that continues to see a lot of interest and success is the development of centers of excellence. These centers allow for the efficient use of analytic talent across the business.

In many instances, acquiring experienced data scientists is just not an option. Therefore, many organizations are building relationships with universities to find recent graduates from a number of new and specialized advanced degree programs in data and analytics. However, to some extent, most organizations must use the talent that they already have. This requires making analytics more approachable and accessible to many different types of business users by providing a holistic machine learning platform that can provide guided or approachable interfaces and also support data scientists, who typically prefer programming. This platform can be homegrown, single-sourced, or a combination of both; the specific choice depends on the organization and industry, as some industries must adhere to regulatory requirements that command certain levels of transparency, auditability, and security in platforms. This approach helps to scale the organization's analytic capabilities from just a handful of experts to a larger group.

Chief analytics officer

Capitalizing on machine learning often requires two competing interests to cooperate: (a) data access, typically owned by information technology (IT) and the chief information officer (CIO), and (b) prioritization of business use cases, which is dictated by lines of business,usually steered by the chief marketing officer (CMO), for example.

While there is often well-intentioned collaboration, these parties sometimes don't fully possess the technical or business background to properly communicate with one another. Further, who should own longer-term initiatives that may have wider applicability across business functions? Hence, there is a need for an analytics champion, a chief analytics officer (CAO) who has one foot in IT and one foot in business as well as the technical background to move analytics initiatives like machine learning forward. More and more organizations recognize the need for CAOs, yet this role continues to be defined and organizations are realigning to maximize this leadership.

Data-driven organization

While data scientists are arguably the poster child of this most recent data hype, the pervasiveness and democratization of data in today's world has created the need for savvy data professionals across many levels and functions of an organization. Business processes and technological solutions must enable these less analytically skilled but nonetheless critical members of a pervasive analytic culture.

Ultimately, organizations need to drive toward a data culture, where access to data is commonplace and data-driven decision making is expected. Developing a data culture, implementing the required business processes, and investing in the appropriate technology—all in synergistic fashion—is the primary challenge of creating a data-driven organization that can capitalize on the advances in machine learning. Adopting machine learning can be like any other change management initiative. Long-lasting impact can only be made when a fundamental shift in culture is in place to support the change. Often, this is more about the processes in place than specific capital investments, such as technological infrastructure.

Data Challenges

> *Productive, professional machine learning exercises should start with a specific business need and should yield quantifiable results.*

The sheer volume of data captured by organizations and how it is managed is a serious challenge. Just because you have a lot of data does not mean you have the right data. Flexible, parallel storage and processing platforms such as Hadoop and Spark are common. However, the most state-of-the-art algorithm, the biggest compute cluster, and any magnitude of big data cannot create a trustworthy model without the appropriate data and data preparation.

Data quality

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. If an organization has control over the data-generating source, then quality measures can be built in before the collection process. For example, consider staffing education around business operations with human input. Also, pre-analytics can be applied to raw data during the collection process, cleaning data as it is being collected (e.g., streaming data quality), or in batch after the data has hit persistent storage. Pre-analytics are analytical data quality techniques that are applied to enhance data quality by way of standardization, imputation, and other basic techniques.

Data security and governance

In many regulated industry sectors—such as banking, insurance, healthcare, and government—data security is of paramount importance, and lapses in data security can result in major problems.

For many data science efforts and machine learning software packages, data security is only an afterthought. As data security expertise is typically located in IT, this can cause inefficient and unnecessary conflict between data science and IT groups. To reduce this friction, data security issues should be addressed at the beginning of a machine learning exercise when possible.

Data security, however, is only one part of a large umbrella of data governance, which addresses questions around how an organization's data should be managed. Security aside, how do we ensure data efforts throughout the organization are not unnecessarily replicated, while still maintaining sufficient visibility to capitalize on otherwise unnoticed opportunities? How does an organization create the elusive "single version of the truth"?

Machine learning activities will further exacerbate the situation as the laws of big datakick in—data begets more data. How should the data outputs of machine learning models be used, stored, and reused? There are no easy answers here, especially because they are nuanced for each organization and industry, but bringing data governance concerns to the forefront of any analytics strategy is imperative.

Data integration and preparation

Machine learning algorithms can thrive on more data, as long as the additional data provides more signal and not just more noise. Often, value-added data can exist across disparate data stores (e.g., different technologies, different vendors) and not necessarily synchronized in the same time period (e.g., streaming, always-on, batch ingestion). Ultimately though, all these differences must be reconciled and eventually streamlined to utilize current machine learning algorithms.

After data has been collected and cleaned, it must still be transformed into a format that is logical for machine learning algorithms to consume. While different algorithm implementations may accept different formats of data, the most common representation is to place a single, unique entity of a consistent type in each row of a data set. This entity is often a patient or a customer, but it could also be an insurance quote, a document, or an image. The data scientist and programmer, Hadley Wickham, proposed a general solution to this problem in a paper that he wrote on his tidy data process that is a good resource for those looking for tips.

Clean, comprehensive, and well-integrated data may just contain too many features to be modeled efficiently. This is sometimes referred to as the curse of dimensionality. For an example of a clean but extremely wide data set, consider a term-document matrix used in text mining that may have hundreds of thousands of columns corresponding to each unique term in a corpus of documents. Some processes to address this issue include:

Feature extraction
: Combining the original features (e.g., variables) into a smaller set of more representative features

Feature selection
: Selecting only the most meaningful original features to be used in the modeling algorithm

Feature engineering
: Combining preexisting features in a data set with one another or with features from external data sources to create new features that can make models more accurate

Data exploration

Productive, professional machine learning exercises should start with a specific business need and should yield quantifiable results. Can you be sure the current data can answer the business need or provide quantifiable results? In practice, these needs and desired results often change as a project matures. It is possible that over the course of the project, the original question and goals could become completely invalidated. Where will new questions come from? Can the current data answer these new questions and provide adequate results? The only way to resolve such important, preliminary problems is through multiple cycles of ad-hoc data exploration. Data scientists must have the ability to efficiently query, summarize, and visualize data before and after machine learning models are trained.

Many other best practices exist around data storage and data management. For a more technical overview, Appendix A contains a brief set of suggested best practices for common data quality problems and data cleaning.

Infrastructure Challenges

Managing the various aspects of the infrastructure surrounding machine learning activities becomes 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. Some infrastructure challenges include storage, computation, and elasticity. Let's take a look at each of these challenges.

Storage

When it comes to persistent data storage, no single platform can meet the demands of all data sources or use cases. This is especially true for machine learning projects, which are often conducted on unstructured data, such as text and images, or on large directed or undirected graphs that could define social networks or groups of fraudsters. Storage considerations include data structure, digital footprint, and usage patterns (e.g., heavy write versus heavy read). The usage demand of data is often referred to as data temperature, where "hot" data is high demand and "cold" data is lower demand. Some popular storage platforms include traditional relational databases, Hadoop Distributed File System (HDFS), Cassandra, S3, and Redshift. However, architecting an appropriate, organization-wide storage solution that may consist of various platforms can be an ongoing challenge as data requirements mature and technology continually advances.

Computation

Most machine learning tasks are computationally intensive. Typically, the more data a machine learning algorithm sees during training, the better answers it will provide. While some erroneously believe that the era of big data has led to the end of sampling, the educated data scientist knows that newer, computationally demanding sampling techniques can actually improve results in many cases. Moreover, data preparation and modeling techniques must be refined repeatedly by the human operator in order to obtain the best results. Waiting for days while data is preprocessed or while a model is trained is frustrating and can lead to inferior results. Most work is conducted under time constraints. Machine learning is no different. Preprocessing more data and training more models faster typically allows for more human cycles on the problem at hand and creates more accurate models with higher confidence. 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. Building the correct computation infrastructure can be a serious challenge, but leading organizations are experimenting with and implementing the following approaches:

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). Numerous libraries are available that allow the transparent use of GPUs from high-level languages.

Distributed computing

In distributed computing, data and tasks are split across many connected computers, often reducing execution times. However, not all distributed environments are well-suited for machine learning. And not all machine learning algorithms are appropriate for distributed computing.

Elasticity

Both storage and compute resource consumption can be highly dynamic, requiring high amounts in certain intervals and low amounts in others. For a data scientist, an example of this may be prototyping model development on a laptop with a subset of the data and then quickly requiring more compute resources to train the model on a much larger data set. For a customer-facing web service, an example of this may be adding more resources during times of known peak activity but scaling back during non-peak times. Infrastructure elasticity allows for more optimal use of limited computational resources and/or financial expenditures.

Cloud computing allows organizations to scaleup their computing platform as needed, and to pay for just the hardware and software they need. With the advent of several large and dependable cloud computing services, moving data to one or many remote servers and running machine learning software on that data can be a fairly simple task. Of course, some organizations will choose to keep their data on premise, but they can also benefit from building their own private cloud computing infrastructures due to economies of scale.

Modeling Challenges

For organizations that have already deployed accurate, traditional models or those with regulatory requirements, the complexities and sacrificed interpretability of machine learning models can hinder interest in exploring the benefits that machine learning can provide. In this section, we will discuss several innovative techniques to overcome those challenges.

Model complexity

New analytical algorithms and software continue to appear at a breakneck pace. Many data-driven organizations have spent years developing successful analytics platforms and choosing when to incorporate newer, more complex modeling methods into an overall analytics strategy is a difficult decision when you have already deployed accurate, traditional models. 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.A potential tactic to introduce machine learning techniques into these industries is to position machine learning as an extension to existing analytical processes and other decision-making tools. For example, a bank may be required to use traditional regression in its regulated dealings, but it could use a more accurate machine learning technique to predict when a regression model is growing stale and needs to be refreshed.

For organizations without legacy platforms or regulation obligations, or for those with the ambition and business need to try modern machine learning, several innovative techniques have proven effective:

Anomaly detection
: Anomaly detection is a challenging problem for any analytics program. 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 healthcare, 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. The ability to build models automatically across many segments or even individuals, as in a model factory, allows any gains in accuracy and efficiency to be operationalized.

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 preexisting models can also lead to improved results.

Model interpretability

> *What makes machine learning algorithms difficult to understand is also what makes them excellent predictors: they are complex.*

A major difficulty with machine learning for business applications is that most machine learning algorithms are black boxes. What makes machine learning algorithms difficult to understand is also what makes them excellent predictors: they are complex. In some regulated verticals,such as banking and insurance, models simply have to be explainable due to regulatory requirements. If this is the case, then a hybrid strategyof mixing traditional approaches and machine learning techniques together, such as described in "Predictive Modeling: Striking a Balance Between Accuracy and Interpretability", might be a viable solution to some interpretability problems. Some example hybrid strategies include:

Advanced regression techniques
Penalized regression techniques are particularly wellsuited for wide data. They avoid the multiple comparison problem that can arise with stepwise variable selection. They can be trained on datasets with more columns than rows and they preserve interpretability by selecting a small number of original variables for the final model. Generalized additive models fit linear terms to certain variables and nonlinear splines to other variables, allowing you to hand-tune a trade-off between interpretability and accuracy. With quantile regression, you can fit a traditional, interpretable linear model to different percentiles of your training data, allowing you to find different sets of variables for modeling different behaviors across a customer market or portfolio of accounts.

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 much less accurate than your machine learning model, you've probably missed some important interaction(s) of predictor variables.

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. This model will likely be a more interpretable proxy you can use to explain the more complex logic of the machine learning model.

Because machine learning models provide complex answers in terms of probabilities, interpreting their results can cause friction between data scientists and business decision makers who often prefer more concrete guidance. This challenge requires data scientists, who find the complexity of mathematics appealing, to educate and prepare business leaders, or provide an explanation in visual, monetary, or less technical terms.

For more technical details regarding the suggested usage of specific machine learning algorithms, see Appendix B.

Operational and Production Challenges

Maximum business impact is often achieved by moving machine learning models from an individual or team development environment into an operational system or process.

Model deployment

Moving the logic that defines all the necessary data preparation and mathematical expressions of a sophisticated machine learning model from a development environment, such as a personal laptop, into an operational, secure, production server or database is one of the most difficult, tedious aspects of machine learning. The volumes of data require model logic to be executed where data "lives" (e.g., in-database scoring) because moving massive amounts of data to a development compute engine is impractical. Models may be heavily utilized in operational use cases, required to make millions of decisions a day, each with millisecond response times that are prohibitive in a development environment. Moreover, interpreted languages are probably too slow to guarantee millisecond response times and the model probably needs to be ported into ahigh-performance language, such as C or Java.

The lack of parity in available model execution engines between development and production environments is another challenge. For example, models may be developed in R or Python, but the target production environment requires a high-performance implementation. In addition, the IT process also needs to ensure that the data required as input for the model predictions is available at time of application and that the results (predictions) are fed appropriately into downstream decision engines. For a model to be truly useful, the model application process needs to be implemented into the organization's IT infrastructure to provide answers to the business when needed and in an understandable format.

Model and decision management

As machine learning and predictive analytics become more prevalent throughout the organization, the number of models produced will rapidly increase and once those models are deployed they must be managed and monitored, including providing traceability and version control. These models have often been trained on static snapshots of data so their predictions will typically become less accurate over time as the environment shifts away from the conditions captured in the training data due to competitive offers, changing customer base, and changing customer behavior. After a certain amount of time, the error rate on new data surpasses a predefined threshold and models have to be retrained or replaced.

Champion/challenger testing is a common model deployment practice where a new challenger model is compared against a currently deployed model at regular time intervals. When a challenger model outperforms a currently deployed model, the deployed model is replaced by the challenger model and the champion/challenger process is repeated. Another approach to refreshing a trained model is through online updates that continuously change the value of model parameters or rules based on the values of new, real-time streaming data. The smart move is to assess the trustworthiness of real-time data streams before implementing an online modeling system.

Model monitoring and management is just one part of the larger field of analytical decision management. In analytical decision management, model predictions are often combined with practical business rules to determine the appropriate business decision. Decision management combines the automatically generated predictions of a statistical or machine learning model with common sense business rules usually constructed by humans. Decision management also seeks to organize metadata, such as model and business rule versioning, lineage, authorship, documentation, and assessment into an approachable software application. For highly regulated or analytically maturing organizations, decision management can help ensure that analytical operations continue to provide maximum value to the organization.

Agility

Interacting with operational and production systems requires care due to the mission-critical nature of these systems; failure or disruption of service in these systems can lead to substantial negative business impact. Hence, there is a need for tighter controls and governance. However, we must balance the weight of operational responsibility with the agility to innovate and leverage the dynamic machine learning landscape. New algorithms, libraries, and tools surface every day. Successful organizations will quickly evaluate new innovations, consuming those that bear fruit into standard practice. Likewise, paths for managed models to reenter development should be frictionless, whether it be for simple retraining or a more involved model rebuilding.

Real Impacts of Machine Learning in Business

The following two case studies represent examples of how organizations are overcoming some of the challenges that machine learning presents to businesses, and how organizations are using machine learning to improve their analytical results.

Geneia: Transforming Healthcare for the Better

> It's never been a question about whether or not we should invest in these technologies.
>
> —Heather Lavoie, Geneia

Geneia, a healthcare technology and consulting company, helps healthcare organizations deliver better care at a lower cost through a user-friendly data and analytics platform called Theon. With Theon, clinical, claims, physiological, biometric, and consumer data are all woven together and the analytics are customized to the role of the care team member and personalized for patients using constant streams of data to look for any early, subtle signs of trouble that must be investigated. For practitioners, Theon provides advanced clinical, analytical, and technical solutions to hospital systems, physician organizations, and nursing coordinators. At the Clevel, it provides information about a population's health goals. Theon also allows insurers to assess quality and cost measures, and to compare their contracted hospital and physician networks against benchmarks.

Moving toward more advanced analytic methods like machine learning was a natural evolution for Geneia. "It's never been a question about whether or not we should invest in these technologies," said Heather Lavoie, President and Chief Operating Officer of Geneia. "It is foundational to our business model and fundamental to where the United States needs to go as it relates to getting better quality healthcare and better quality of life."

Addressing organizational challenges

With everyone competing for the limited supply of experienced talent, Geneia focuses human resources on developing partnerships with universities to recruit and cultivate recent graduates. Geneia recognizes that retaining existing talent is incredibly important. "Data scientists are celebrated and highly valued in our organization, and we're incredibly excited about them. What has been extremely important for us is to continue to give them time for exploration, building and testing the limits of their knowledge," says Lavoie.

While giving data scientists the time and access to the data required to maintain their skills, data security is another area of significant focus for Geneia. Security involves not only who has access to certain data but also how their teams are structured. According to Lavoie, there are legitimate constraints on how the company operates across states. "We have to be diligent about how we segregate the data across all of our environments, including development, quality assurance, training, as well as production. Data security impacts the structure of the teams in terms of who can have access to what information, like on-shore or off-shore resources, interns, and identifiable or de-identified information."

Impacts of machine learning

Geneia uses its understanding of the market and its unique regulatory privacy requirements to drive its product roadmap. It currently uses statistical and machine learning techniques—including principal component analysis—to assimilate claims, clinical, consumer, actuarial, demographic, and physiologic data, and to see patterns or clusters in the data and try to make inferences about causality and combinations. Binary logistic regression, a straightforward classification technique, helps identify predictors and develop risk-scoring algorithms to identify where sick patients fall on a wellness spectrum and the most appropriate remediation. Additional classification methods, such as support vector machines, help to make inferences about incorrect coding that may cause erroneous reimbursement issues. These methods also validate that a patient's prescription is warranted based on their diagnosis and help to determine the next best action based on their propensity to engage in certain programs. It helps in the identification of disease cohorts, such as those for diabetes or congestive heart failure, to determine who belongs in a particular group. Artificial neural networks also allow Geneia to learn more about the data when the relationship between variables is unknown.

Theon is updated dynamically as new data becomes available. In instances where real-time data is available, such as remote patient monitoring and wearable devices, it is updated in real time. Geneia's data scientists do model improvement on a regular basis. One of the things that Lavoie is most excited about is their active learning feedback loop. "We have built feedback screens into the workflow so that those who are interacting with the patients can provide direct feedback that we can use to improve the models," she said. "If we have an alert that comes up for a patient and a nurse realizes that it is not applicable, the healthcare professional can indicate that. Now, we can understand that the alert was a false positive, and the model can use that moving forward to make better predictions. This ongoing human element of improvement through our clients is improving the model as part of their regular workflow."

Healthcare is a highly regulated environment, so it is critical that Geneia is able to defend improvements made to their models both internally and to their customers when new models are created. To overcome this challenge, Geneia has developed approaches such as using historical data to compare the predictions of a model to what actually happened.

The future is bright

The future holds exciting opportunities for Geneia in leveraging the Internet of Things (IoT) to understand subtle changes in health. According to Lavoie, "we will have much better information about people in terms of their movement in the home, their toileting, and adherence to prescriptions by capturing data rather than asking them to recount the information. We will be able to better help people interpret their own health status more effectively."

Geneia expects genomics and precision medicine to be game changers. "Today there is so much waste, so much unnecessary treatment that is not helpful to patients because we use treatment as a blunt instrument," said Lavoie. "We try things now based on evidence, but not individualized evidence. When we can get to the point where we have a specific intervention and a better understanding about whether or not it is actually going to work for that individual, the promise is so great." These innovations in healthcare could allow providers to understand the effects of nutrition and medication at an individual level, thus avoiding adverse drug reactions, eliminating unnecessary costs, and improving patient outcomes.

Equifax: Turning Big Ideas Into Products and Insights

Once a consumer credit company, Equifax now organizes, assimilates, and analyzes data on more than 800 million consumers to help its customers across different industries make informed decisions. "I'm a data and analytics geek who works for a data and analytics company," said Peter Maynard, Senior Vice President of Enterprise Analytics at Equifax. According to Maynard, the big picture of his job almost always involves analyzing data and making it useful for both internal teams and external customers. For example, one upcoming customer engagement will involve analyzing billions of social media posts and relating posts' content to consumer buying behavior. Of course, developing ever more detailed and nuanced models of consumer behavior is not the only way Equifax makes data useful. Teams that Maynard works with also guide clients in their analytics ventures, build analytics solutions, and research analytical innovations.

Machine learning driving innovation

For Maynard and team, machine learning is an important part of their innovation strategy. In general, Maynard feels that machine learning allows for purer data-driven solutions than conventional approaches. In the past, common-sense business rules were used in combination with more straightforward mathematical models to make decisions based on data. Heuristic rules derived from human experience can be a powerful addition to modeling logic, but they also come with the baggage of human bias and judgment errors. Machine learning models are able to shoulder more of the intellectual work that humans did in the past, allowing decisions to be made more directly from the data, and leading to potentially more accurate and objective conclusions.

At Equifax, these big ideas are being turned into products and insights, and teams working with Maynard have cracked an extremely difficult problem in the credit scoring industry. It has long been understood that machine learning algorithms could make more accurate credit lending decisions than currently established approaches; however, machine learning algorithms are typically seen as uninterpretable black boxes and can be very difficult to explain to both consumers and regulators. Using Equifax's patent-pending Neural Decision Technology (NDT), artificial neural networks with simple constraints can be used for automated credit lending decisions, which are measurably more accurate than logistic regression models and also produce the mandatory reason codes that explain the logic behind a credit lending decision. The NDT's increased accuracy could lead to credit lending in a broader portion of the market, such as new-to-credit consumers, than previously possible. The NDT is able to generate its decisions with less human input than contemporary approaches as logistic regression models often require segmentation schemes to achieve the best results. The NDT is able to implicitly derive this grouping with no manual human input and while also generating better predictions.

Internal and external challenges

Inside Equifax, Maynard sees two primary challenges in the continued adoption of machine learning into their overall analytics strategy. First, Equifax has highly optimized processes built around traditional analytical approaches. Trying new things just takes more time, and Maynard believes that it's management's responsibility to create space for this research and development to occur. The second challenge is increasing awareness. Maynard maintains that data scientists must keep management informed of research projects and that management must in turn encourage data scientists to move forward on the best and most actionable ideas. Organic buzz around success is another important way to raise awareness, and some new machine learning projects at Equifax have produced the kind of great results that have spread through the data and analytics groups on their own.

Maynard sees his customers facing a different kind of challenge. Many have the same target destination: closing the feedback loop between model predictions and consumer behavior to make accurate real-time decisions. This is the "test and learn cycle on steroids," says Maynard. Organizations want to observe consumer behaviors, predict whether these behaviors will lead to an outcome that will affect revenue, such as defaulting on debt, increasing spending, or churning to a competing product, and take actions to address the anticipated behavior. Realistically, automated real-time decision-making capabilities require significant hardware, software, and policy infrastructure, which can take a great deal of time to build out. Maynard observes that customers are able to move toward this target destination at different speeds. Perhaps not surprisingly, companies with less legacy technologies and processes to maintain are often able to update their infrastructure the fastest.

Finding and retaining analytical talent

When asked about the perceived shortage of qualified data scientists, Maynard responded that he feels the talent gap is beginning to be filled. There are now many analytics graduate programs producing new talent, and highly educated professionals from other quantitative fields are being lured into data science. However, Maynard was quick to point out that Equifax still has a lot more data than data scientists, and that data and the work it enables are probably some of their greatest assets in attracting and retaining talent. Given Equifax's track record of implementing analytical solutions, the models trained by its data scientists will likely have a real impact on consumers. He also highlighted the importance of preserving the experimental graduate school mindset in the workplace. "If you're a curious person with a hypothesis, let's hear it."

Conclusion

Machine learning is moving into the mainstream. As documented in this report, 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. Consider a holistic view of machine learning inside your organization. To extract continuous value from machine learning projects and technologies, data must move from ingestion all the way to impactful business decisions in a streamlined process. Coming years will certainly bring improvements to today's already impressive state-of-the-art capabilities, and perhaps even discontinuous jumps in progress. Don't get left behind!

Further Reading

For further reading about the field of machine learning, its place among other analytical disciplines, and its historical evolution, the following papers are recommended:

- “50 Years of Data Science” by David Donoho (*http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf*)
- “Statistical Modeling: The Two Cultures” by Leo Breiman (*http://projecteuclid.org/euclid.ss/1009213726*)

Appendix A. Machine Learning Quick Reference: Best Practices

Topic	Common Challenges	Suggested Best Practice
Data Preparation		
Data collection	• Biased data • Incomplete data • The curse of dimensionality • Sparsity	• Take time to understand the business problem and its context • Enrich the data • Dimension-reduction techniques • Change representation of data (e.g., COO)
"Untidy" data	• Value ranges as columns • Multiple variables in the same column • Variables in both rows and columns	Restructure the data to be "tidy" by using the melt and cast process
Outliers	• Out-of-range numeric values and unknown categorical values in score data • Undue influence on squared loss functions (e.g. regression, GBM, and *k*-means)	• Robust methods (e.g. Huber loss function) • Discretization (binning) • Winsorizing
Sparse target variables	• Low primary event occurrence rate • Overwhelming preponderance of zero or missing values in target	• Proportional oversampling • Inverse prior probabilities • Mixture models
Variables of disparate magnitudes	• Misleading variable importance • Distance measure imbalance • Gradient dominance	Standardization
High-cardinality variables	• Overfitting • Unknown categorical values in holdout data	• Discretization (binning) • Weight of evidence • Leave-one-out event rate
Missing data	• Information loss • Bias	• Discretization (binning) • Imputation • Tree-based modeling techniques
Strong multicollinearity	Unstable parameter estimates	• Regularization • Dimension reduction

Topic	Common Challenges	Suggested Best Practice
Training		
Overfitting	High-variance and low-bias models that fail to generalize well	• Regularization • Noise injection • Partitioning or cross validation
Hyperparameter tuning	Combinatorial explosion of hyper-parameters in conventional algorithms (e.g., deep neural networks, Super Learners)	• Local search optimization, including genetic algorithms • Grid search, random search
Ensemble models	• Single models that fail to provide adequate accuracy • High-variance and low-bias models that fail to generalize well	• Established ensemble methods (e.g., bagging, boosting, stacking) • Custom or manual combinations of predictions
Model Interpretation	Large number of parameters, rules, or other complexity obscures model interpretation	• Variable selection by regularization (e.g., L1) • Surrogate models • Partial dependency plots, variable importance measures
Computational resource exploitation	• Single-threaded algorithm implementations • Heavy reliance on interpreted languages	• Train many single-threaded models in parallel • Hardware acceleration (e.g., SSD, GPU) • Low-level, native libraries • Distributed computing, when appropriate
Deployment		
Model deployment	Trained model logic must be transferred from a development environment to an operational computing system to assist in organizational decision making processes	• Portable scoring code or scoring executables • In-database scoring • Web service scoring
Model decay	• Business problem or market conditions have changed since the model was created • New observations fall outside domain of training data	• Monitor models for decreasing accuracy • Update/retrain models regularly • Champion-challenger tests • Online updates

Appendix B. Machine Learning Quick Reference: Algorithms

Penalized Regression	
Common Usage • Supervised regression • Supervised classification	Common Concerns • Missing Values • Outliers • Standardization • Parameter tuning
Suggested Scale • Small to large data	Interpretability • High
Suggested Usage • Modeling linear or linearly separable phenomena • Manually specifying nonlinear and explicit interaction terms • Well suited for $N << p$	

Naïve Bayes	
Common Usage • Supervised classification	Common Concerns • Strong linear independence assumption • Infrequent categorical levels
Suggested Scale • Small to extremely large data sets	Interpretability • Moderate
Suggested Usage • Modeling linearly separable phenomena in large data sets • Well-suited for extremely large data sets where complex methods are intractable	

Decision Trees

Common Usage

- Supervised regression
- Supervised classification

Common Concerns

- Instability with small training data sets
- Gradient boosting can be unstable with noise or outliers
- Overfitting
- Parameter tuning

Suggested Scale

- Medium to large data sets

Interpretability

- Moderate

Suggested Usage

- Modeling nonlinear and nonlinearly separable phenomena in large, dirty data
- Interactions considered automatically, but implicitly
- Missing values and outliers in input variables handled automatically in many implementations
- Decision tree ensembles, e.g., random forests and gradient boosting, can increase prediction accuracy and decrease overfitting, but also decrease scalability and interpretability

k-Nearest Neighbors (kNN)

Common Usage

- Supervised regression
- Supervised classification

Common Concerns

- Missing values
- Overfitting
- Outliers
- Standardization
- Curse of dimensionality

Suggested Scale

- Small to medium data sets

Interpretability

- Low

Suggested Usage

- Modeling nonlinearly separable phenomena
- Can be used to match the accuracy of more sophisticated techniques, but with fewer tuning parameters

Support Vector Machines (SVM)

Common Usage

- Supervised regression
- Supervised classification
- Anomaly detection

Common Concerns

- Missing values
- Overfitting
- Outliers
- Standardization
- Parameter tuning
- Accuracy versus deep neural networks depends on choice of nonlinear kernel; Gaussian and polynomial often less accurate

Suggested Scale • Small to large data sets for linear kernels • Small to medium data sets for nonlinear kernels	Interpretability • Low
Suggested Usage • Modeling linear or linearly separable phenomena by using linear kernels • Modeling nonlinear or nonlinearly separable phenomena by using nonlinear kernels • Anomaly detection with one-class SVM (OSVM)	
Artificial Neural Networks (ANN)	
Common Usage • Supervised regression • Supervised classification • Unsupervised clustering • Unsupervised feature extraction • Anomaly detection	Common Concerns • Missing values • Overfitting • Outliers • Standardization • Parameter tuning • Accuracy versus deep neural networks depends on choice of nonlinear kernel; Gaussian and polynomial often less accurate
Suggested Scale • Small to large data sets for linear kernels • Small to medium data sets for nonlinear kernels	Interpretability • Low
Suggested Usage • Modeling linear or linearly separable phenomena by using linear kernels • Modeling nonlinear or nonlinearly separable phenomena by using nonlinear kernels • Anomaly detection with one-class SVM (OSVM)	
Association Rules	
Common Usage • Supervised rule building • Unsupervised rule building	Common Concerns • Instability with small training data • Overfitting • Parameter tuning
Suggested Scale • Medium to large transactional data sets	Interpretability • Moderate
Suggested Usage • Building sets of complex rules by using the co-occurrence of items or events in transactional data sets	

k-Means

Common Usage

- Unsupervised clustering

Common Concerns

- Missing values
- Outliers
- Standardization
- Correct number of clusters is often unknown
- Highly sensitive to initialization
- Curse of dimensionality

Suggested Scale

- Small to large data sets

Interpretability

- Moderate

Suggested Usage

- Creating a known a priori number of spherical, disjoint, equally sized clusters
- *k*-modes method can be used for categorical data
- *k*-prototypes method can be used for mixed data

Hierarchical Clustering

Common Usage

- Unsupervised clustering

Common Concerns

- Missing values
- Standardization
- Correct number of clusters is often unknown
- Curse of dimensionality

Suggested Scale

- Small data sets

Interpretability

- Moderate

Suggested Usage

- Creating a known a priori number of nonspherical, disjoint, or overlapping clusters of different sizes

Spectral Clustering

Common Usage

- Unsupervised clustering

Common Concerns

- Missing values
- Standardization
- Parametertuning
- Curse of dimensionality

Suggested Scale

- Small data sets

Interpretability

- Moderate

Suggested Usage

- Creating a data-dependent number of arbitrarily shaped, disjoint, or overlapping clusters of different sizes

Principal Components Analysis (PCA)

Common Usage

- Unsupervised feature extraction

Common Concerns

- Missing values
- Outliers

Suggested Scale

- Small to large data sets for traditional PCA and SVD
- Small to medium data sets for sparse PCA and kernel PCA

Interpretability

- Generally low, but higher sparse PCA or rotated solutions

Suggested Usage

- Extracting a data-dependent number of linear, orthogonal features, where $N >> p$
- Extracted features can be rotated to increase interpretability, but orthogonality is usually lost
- Singular value decomposition (SVD) is often used instead of PCA on wide or sparse data
- Sparse PCA can be used to create more interpretable features, but orthogonality is lost
- Kernel PCA can be used to extract nonlinear features

Nonnegative Matrix Factorization (NMF)

Common Usage

- Unsupervised feature extraction

Common Concerns

- Missing values
- Outliers
- Standardization
- Correct number of features is often unknown
- Presence of negative values

Suggested Scale

- Small to large data sets

Interpretability

- High

Suggested Usage

- Extracting a known a priori number of interpretable, linear, oblique, nonnegative features

Random Projections

Common Usage

- Unsupervised feature extraction

Common Concerns

- Missing values

Suggested Scale

- Medium to extremely large data sets

Interpretability

- Low

Suggested Usage

- Extracting a data-dependent number of linear, uninterpretable, randomly oriented features of equal importance

Factorization Machines	
Common Usage • Supervised regression and classification • Unsupervised feature extraction	Common Concerns • Missing values • Outliers • Standardization • Correct number of features is often unknown • Less suited for dense data
Suggested Scale • Medium to extremely large sparse or transactional data sets	Interpretability • Moderate
Suggested Usage • Extracting a known a priori number of uninterpretable, oblique features from sparse or transactional data sets • Can automatically account for variable interactions • Creating models from a large number of sparse features; can outperform SVM for sparse data	

About the Authors

Patrick Hall is a senior staff scientist at SAS and an adjunct professor in the Department of Decision Sciences at George Washington University. He designs new data mining and machine learning technologies. He is the 11th person worldwide to become a Cloudera certified data scientist. Patrick studied computational chemistry at the University of Illinois before graduating from the Institute for Advanced Analytics at North Carolina State University.

Wen Phan is a senior solutions architect with SAS. He leads cross-functional teams to architect practical data-driven solutions and deliver high value outcomes for some of SAS's largest customers. Most recently, he has led solutions around analytics centers of excellence, decision management, cloud-first architectures, and machine learning. Wen holds a Master of Science in business analytics and a Bachelor of Science in electrical engineering.

Katie Whitson is a senior marketing specialist at SAS where she leads the go-to-market strategy for SAS's Analytics Portfolio, focusing primarily on SAS's Advanced Analytics capabilities including machine learning. She holds a Bachelor of Science in business management from North Carolina State University.