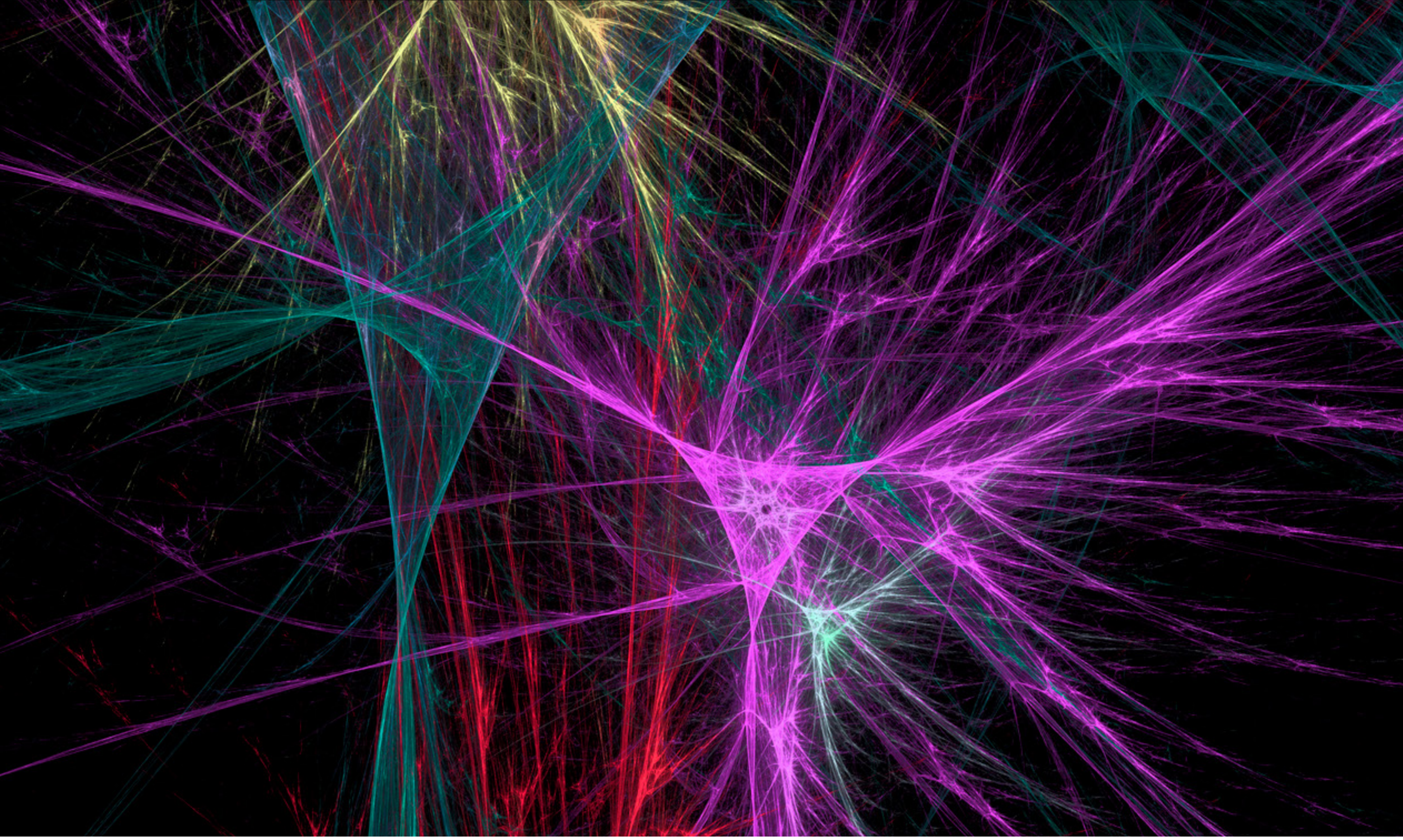
DataInformed

Tap into the Power of Machine Learning

Democratizing Data Science through Automation



Sponsored By: DataRobot

Table of Contents:

Automation Is the New Reality for Big Data Initiatives

by Jelani Harper

The preeminence of data science was inextricably linked to the emergence of big data. Combining business savvy, analytics, and data curation, this discipline was hailed as an enterprise-wide savior for the rapidity of the disparate forms of big data that threatened to overwhelm it.

Numerous developments within the past several months, however, have created a different reality for big data and its future. Its technologies were refined. The self-service movement within the data sphere thrived. The result? Big data came to occupy the central place in the data landscape as critical elements of data science – preparation, analytics, and integration – became automated.

Thanks to the self-service movement's proliferation, even the smallest of organizations can now access big data's advantages. "There's been a lot of discussion about self-service…and having data analysts get at the data directly," MapR Chief Marketing Officer Jack Norris said. "But you also have to recognize, what do you mean by 'the data,' and what has to happen to 'the data' before that self-service takes place?"

The Impact of the Cloud

In many instances, what happens to the data prior to self-service is done by others. Facilitating analytics is one of the chief components of data science, particularly when incorporating big data sources. There are numerous analytics options that end users can access via the cloud that can yield insight into all sorts of data – many of which can do so in nearly real time. Ranging the gamut from conventional historic business intelligence to cutting-edge prescriptive cognitive computing analytics, these services simply require organizations to grant providers access to their data. Cloud analytics decreases physical infrastructure, reduces costs, and effectively outsources potentially difficult and resource-intensive computations. Machine-learning algorithms can provide insight into advisable action based on analytics results (in addition to explanations) and automate the data modeling process, which can prove extremely difficult with time-sensitive big data.

According to Norris, the true benefit of big data analytics is in combining data sources. "If you were looking at just the social media activity of potential prospects out there, you can find some trends. But if you pair that also with your customer information and your customer purchases, you have got a richer view. And then if you add weather data or location information, you can look at different trends there."

Architecture

The data lake concept has been extremely influential in automating pivotal aspects of data science pertaining to integration, data preparation, and analytics. A number of developments within this facet of data management pertaining to big data can expedite the preparation process that can potentially monopolize the time of data scientists. "Data scientists are a rare and precious commodity in an organization," said Cambridge Semantics Vice President of Marketing John Rueter. "You have got these brilliant Ph.D.s who are taking on data science responsibilities. Since information is stored in a data lake in its raw form, they are liable to spend 70 percent of their time just doing the data preparation and data management before you can do any kind of analysis on it."

Data preparation platforms can provide comprehensive views of disparate data sources and their relevance to specific jobs while implementing measures for cleansing and quality. Semantic technologies and machine learning can help identify points of data integration, individual node characteristics, and even facilitate transformative action requisite for specific applications. They also can facilitate much-needed consistency in terms of metadata and schema definitions (on structured and semi-structured data), which Norris said are required "to accomplish self-service so that data analysts can get at the data."

Another means of providing structure and consistency to unstructured big data without data scientists is to leverage JSON-based document stores in data lakes, such as Hadoop. According to Norris, JSON "has the schema built into it. It's basically the data interchange standard of web applications now. It's increasingly the data format produced in the Internet of the Things as the result of sensor data."

Combined with SQL solutions (there are NoSQL ones as well) that interact with JSON to derive schema in real time, "there's no dependency on IT to massage the data before you can do that self-service," Norris explained. Data lakes also provide centralized hubs that are useful for running both operations and analytics simultaneously, with much less need for data-scientist involvement because "having it on a single platform … brings operational agility and results in simplifying your data architecture, simplifying your administration, and simplifying a great deal," Norris said.

> “ ”
> Data science is no longer an arcane discipline that is the domain of only a select few.

Secure Governance

The self-service access to data and its uses that the aforementioned automated aspects of data science facilitate would be useless without adherence to security and governance standards. The metadata and schema consistency of the foregoing methods – which can be augmented by the cataloging capabilities of data-preparation platforms – are useful for restricting and granting access to data based on regulatory concerns, security, and governance policies. They also can provide traceability. Certain governance solutions also are endowed with standards-based semantic capabilities to reinforce policies and procedures while linking data to vital information such as business glossaries.

These methods are also applicable to the varying cloud analytics options, resulting in what Norris referred to as "The four A's: You need to authenticate, and be able to understand who's coming in and tapping into your global directory services. You need to control access. Then you need the ability to audit, so you can understand who did what. Then, lastly, is what we refer to as the architecture. Is that security granular? Is that security at the data level, not by the access method?"

Data science is no longer an arcane discipline that is the domain of only a select few. The self-service movement has succeeded in automating numerous important aspects of data science, which business users can now leverage without understanding the intricacies of algorithm development for analytics or ETL for data preparation and application loading. Significantly, business users can utilize these facets of data science in a way that adheres to governance policies and provides the level of security required for enterprise data. Best of all, automation makes big data initiatives much more affordable and accessible to the enterprise. Automated data science has not obscured the jobs of data scientists, but instead freed them from some of the more time-consuming aspects of their position so they can work on more profound problems. ■

Jelani Harper has written extensively about numerous facets of data management for the past several years. His many areas of specialization include semantics, big data, and data governance.

Intelligent Process Automation: It's About the Data, Not the Robot

by Robert Brown

Robert Brown, Associate Vice President, Cognizant's Center for the Future of Work

Robots, robots everywhere. Automation, whether it's on the factory assembly line, vacuuming floors, unloading ships, or in the latest Hollywood blockbuster, it's as if we are facing an invasion of "all-things-automated" in the zeitgeist of business and the popular press. But haven't we been automating tasks since the dawn of computing? So what's the big deal about today's intelligent process automation tools?

While virtually every process uses technology to make it work, there's still a lot of repetitive, manual data entry, searching, and collating to get things done. And the next wave of process efficiency gain and business outcomes will be driven by automation that helps smart robots complement smart people by using sophisticated software to automate rote tasks.

This has led to our arrival at a profound inflection point in how critical business services will be sourced and, more importantly, delivered. The physical and digital worlds are converging at a speed unfathomable even just a decade ago. It seems as if almost every physical process is getting instrumented with sensors, telematics, and things that drive ever-growing feedback loops of data. This is where the data generated from the automated processes we call intelligent process automation, or IPA, come in.

The insights gathered from IPA are about the interplay of smart robots for knowledge processes that can unlock data, which makes smart people even smarter. We see this happening most predominantly among software tools performing repetitive and rote processes and unleashing a new era of human-machine interface. But because of the analytics potential, the far bigger outcome of these enterprise robots is the potential for analytics: augmenting the creative problem-solving capabilities and productivity of human beings, catalyzing superior process outcomes and business results.

The data generated by these increasingly astute technologies of process automation and digitization is the real prize, for businesses and workers alike. With advances in machine learning, artificial intelligence, and big data, companies can predict rather than react to rapidly changing demands and expectations.

The AI Innovators

Businesses that are already embracing these new technologies are capturing more data, improving processes, and generally empowering workers to be more effective at their jobs.

In logistics, for example, real-time dynamic fleet optimization can work wonders for destination and delivery capacity. Dynamic auto pricing can benefit immeasurably from millions of miles of analysis of "hard brakes." Pharmaceutical trials can be maximized by collating huge volumes of clinical data. Be it claims management in insurance or reconciliation or mortgage processing in banks, process models are becoming less transactional and more interactional, affording multiple opportunities to learn about customers, suppliers, partners, and employees. Given the useful data each one of these interactions generates, process analytics can be a real game changer.

It is this level of service that will increase customer satisfaction. IPA, machine learning, and AI increase the productivity, decision making, and agility of front, middle, and back office teams by quickly supplying them with more accurate and actionable information.

Many early adopters of process automation are now well ahead of the game, embracing innovative ways to create new levels of process efficiency and new possibilities for outdated operational models. From the banking industry to the energy and utilities industry, we are seeing groundbreaking developments that allow companies to make the data they are collecting work harder and smarter and improve the customer experience.

In a recent study by Cognizant's Center for the Future of Work in which more than 500 senior executives across the U.S. and Europe were surveyed, half reported that they saw intelligent process automation as significantly improving their business processes over the next three to five years. Nearly one-fifth reported that they achieved cost savings of greater than 15 percent from intelligent process automation in just the past year. A total of 44 percent saw similar importance for analytics, and nearly 47 percent said that understanding customer requirements is a core strategic goal.

The research also showed that, through these technologies and analytics, people are attaining new levels of process efficiency, such as improved operational cost, speed, accuracy, and throughput volume. A total of 43 percent of respondents cited the use of analytics for better process throughput and quality, and 47 percent said they use analytics to better understand customer requirements. Using the increasingly astute technologies of smart robots is becoming a force multiplier for smart people.

The Future of Business Operations in an AI and Machine Learning Era

Across industries, organizations are using intelligent systems and the rich data they provide to increase speed to market, reduce human error, achieve regulatory compliance and reduce fines, realize faster processing times, and gain immediate ROI through faster implementation times.

Yet, while many businesses – notably banks, insurance companies, and healthcare payers – have gained ground in boosting revenues by bringing analytics to customer-facing processes, most others limit their use of analytics to process optimization alone.

For example, 55 percent of Cognizant's survey respondents reported that the key outcome of analytics efforts today is reducing costs. Well over a third of business leaders said they value better process mechanics more highly than outcomes such as prioritizing business needs, better market penetration, and segmentation or creating new products/services.

It's a new era in business, one in which growth will be driven as much by insight and foresight as by physical products and assets. Process automation is about not only cutting costs, but also optimizing workflows and aligning business processes. And analytics is what determines the direction that businesses must take to effectively translate those improvements into higher customer satisfaction and sales. Cognizant reported the same story in its 2013 study, "The Value of Signal (and the Cost of Noise)," in which the 300 companies surveyed said they achieved a total economic benefit of roughly $766 billion over the previous year based on their use of business analytics.

But sometimes, doing analytics, or merely automating an existing process, falls short. Prompted by innovative competitors, a full digital re-think may be crucial to transforming core processes. By harnessing the power of emerging technologies, such as social, mobile, analytics, sensors and cloud, companies are completely re-imagining customer, supplier, and partner interactions. By igniting the digital information surrounding these entities, organizations can realize business process insights in far greater fidelity than has ever been possible.

So, while a futuristic version of AI continues to fuel the latest sci-fi blockbusters, we already have entered a new era of human-machine cooperation in which software tools have emerged as the robots for administration and knowledge-based tasks. Though we are likely to see the rise of the machines over the coming years, the human touch will never be lost, and automation with sophisticated technologies is here to stay. Like a good science fiction movie, it's coming soon, to a process near you. ■

Robert Hoyle Brown is an Associate Vice-President in Cognizant's Center for the Future of Work, and drives strategy and market outreach for the Business Process Services Practice. He is also a regular contributor to futureofwork.com, "Signals from the Future of Work." Prior to joining Cognizant, he was Managing Vice-President of the Business and Applications Services team at Gartner, and as a research analyst, he was a recognized subject matter expert in BPO, cloud services/ BPaaS and HR services. He also held roles at Hewlett-Packard and G2 Research, a boutique outsourcing research firm in Silicon Valley. He holds a Bachelor of Arts degree from the University of California at Berkeley and, prior to his graduation, attended the London School of Economics as a Hansard Scholar.

Improve Your New Product Batting Average with Predictive Analytics

by Thomas H. Davenport

Thomas Davenport

Batting averages were an early predictive metric in baseball, and while there are some problems with it, many fans still find it very useful. If a player with a .162 average comes up to the plate, we know we can expect that a hit is unlikely to happen. And if the player maintains that average over the course of the season, the player is unlikely to be with the team next year.

In business, however, the likes of .162 averages are not uncommon. In several industries, when companies introduce new products and services, the majority of those new offerings fail. In pharmaceuticals, for example, just under 10 percent of drug-development projects that undergo clinical trials eventually receive FDA approval. Because of creative accounting, it's difficult to know what percentage of Hollywood films make money, but one economist's analysis suggested that it was only 22 percent. And a low percentage – about 30 percent – of network television programs ever make it to a second season. These industries and many others have grown accustomed to low batting averages for new products and services, and executives in them say that nothing can be done to improve them. But they haven't explored the potential of predictive analytics to raise batting averages.

In the entertainment industry, companies have long believed that predictive analytics on commercial success was impossible. The screenwriter William Goldman once famously noted, "Nobody knows anything. Not one person in the entire motion picture field knows for a certainty what's going to work. Every time out it's a guess and, if you're lucky, an educated one."

Netflix, however, has raised the TV show batting average considerably. You are probably familiar with the company's use of predictive analytics to improve customer recommendation algorithms for movies. But you may not know how the company has used analytics to predict whether TV shows will be home runs, solid base hits, or strikeouts.

The most prominent example of Netflix's bulking up at the plate is the show *House of Cards*, which was the company's first original series. The political drama stars Kevin Spacey and is now entering its fourth season; Netflix has spent at least $200 million producing it thus far, so it's a big decision. Netflix doesn't release viewership figures, but the show is widely regarded as a home run. And it's not by accident. Netflix employed analytics to increase the likelihood of its success.

In applying analytics to decisions concerning *House of Cards*, Netflix used attribute analysis to predict whether customers would like the series. Netflix has identified as many as 70,000 attributes of movies and TV shows, some of which it drew on for the decision about whether to create *House of Cards*:

- Netflix knew that many people had liked a similar program, the UK version of *House of Cards*.
- They knew that Kevin Spacey was a popular leading man.
- They knew that movies produced or directed by David Fincher (*House of Cards*' producer) were well-liked by Netflix customers.

Even knowing these facts, there was, of course, still some uncertainty about investing in the show, but it made for a much better bet. The company also used predictive analytics in marketing the series, creating 10 different trailers for it and predicting for each customer which trailer would be most appealing. And, of course, these bets paid off. Netflix is estimated to have gained more than 3 million customers worldwide because of *House of Cards* alone.

And while we don't know the details of Netflix's analytics about its other shows, it seems to be using similar approaches on them. Virtually all of the original shows produced by Netflix were renewed after their first seasons – putting the company's batting average at well over .900. In addition, Netflix has had many shows nominated for Emmys and has won its fair share as well.

Netflix isn't the only entertainment company to employ predictive analytics. Amazon, which undoubtedly also uses analytics for its Prime Video original shows, has become one of Netflix's primary competitors in original streaming series. Legendary Entertainment uses analytics to predict various aspects of customer behavior relative to its movies, particularly what types of marketing approaches will be effective. And the actor Will Smith is known for his informal use of analytics to pick movies in which he will act, studying the attributes of box office hits.

So if your company's new product/service batting average is low, take a cue from Netflix. Classify some of the key attributes of your past and current products or services. Then model the relationship between those attributes and the commercial success of the offerings. You'll have a predictive model that should give you some sense of how likely a new product or service is to be successful. With these types of predictive analytics, you won't hit a home run every time at bat, but you should be able to become a much better hitter. ■

Tom Davenport, the author of several best-selling management books on analytics and big data, is the President's Distinguished Professor of Information Technology and Management at Babson College, a Fellow of the MIT Initiative on the Digital Economy, co-founder of the International Institute for Analytics, and an independent senior adviser to Deloitte Analytics. He also is a member of the Data Informed Board of Advisers.

How Machine Learning Will Improve Retail and Customer Service

by John O'Rourke

John O'Rourke, Vice President of Marketing, Indix

Technology has transformed how customers and brands interact with each other. Shoppers once relied on face-to-face, in-store interactions to make purchases and receive support. Now, shoppers do their research before entering a store (81 percent of shoppers conduct online research before buying) and seldom rely on salespeople to help them make decisions. Retailers, for their part, have realized that by embracing technology, they can extend their storefronts to their customers' fingertips.

The Internet, buy buttons, mobile payment apps such as Square and Venmo, and couponing and price-matching apps like SnipSnap have changed how we shop. Shoppers can make purchases from within social media apps and compare prices without leaving a store. While these technologies have propelled the retail industry further into the digital age, technology that is still evolving will have the largest impact on the future of the customer service and retail industries.

Embracing Big Data

More retailers are tracking customer shopping habits through data sources such as social media, purchase history, consumer demand, and market trends. By relying on big data technology to gain a deep understanding of shoppers and their buying trends, retailers can maximize customers' spending and encourage customer loyalty.

According to research by Accenture Analytics, 58 percent of retailers described big data as "extremely important" to their organizations, while 36 percent called it "important." Additionally, 70 percent said that big data is necessary to maintain competitiveness, and 82 percent agreed that big data is changing how they interact with and relate to customers. While most retailers recognize that there is power in big data and analytics as it pertains to shoppers and their purchasing habits, few have unlocked the true potential of that data through machine learning.

Matching Products with People

Machine learning technology amplifies and extends the reach of big data analytics and can help create an exceptional shopping experience. Innovative retailers can tap into the power of machine learning algorithms to do things like determine available products from outside vendors or recommend the quantity, price, shelf placement, and marketing channel that would reach the right customer in a particular area.

Applications of this are already being seen. The North Face leveraged IBM's Watson natural-language processing machine learning system to create the Expert Personal Shopper. So when you are in the Jackets and Vests section of the North Face website, you don't have to get overwhelmed by the choices. You can just type, "I'm going on a cabin trip to Iceland in December." Conflating its trove of product information with weather and other data, the app will surface the right product for you. The technology is relatively new, but you can imagine the implications it will have in the future.

Further, the ability to automate everything through advanced analytics and machine learning soon will mean that basic customer service will be performed by bots that can predict our needs and provide service in the fastest, most immediate way possible: by offering us items we didn't know we needed. As retailers gain more insight into their customers and products, machine learning will be able to match buyers and sellers based on buyers' needs and product availability.

Shopping is becoming increasingly programmatic. In the future, services like digital assistants (Siri, Cortana, Facebook's M) will learn more about us and offer us relevant and personalized product offers. Say, for example, you use a particular brand of razor. Your digital assistant will learn your shopping and usage habits and offer you the best deal on the product at the right time. It might even place the order for you.

Improving the Backend

Machine learning and advanced analytics will not only change how we shop and provide customer service, but also simplify how retailers perform basic operations. Data science and machine learning give us the ability to automate so much of the heavy lifting required to find insight within heaps of data. With these tools, retailers can find useable and useful data to change the shopping experience for consumers.

Technology enables us to create an index of every product in the world, enabling retailers to offer customers the best prices, keep products adequately stocked, and track competitors' minimum-advertised-price violations. A central database of the world's product information enables retailers to offer the best shopping experience for buyers.

An innovative-technology approach to customer service and commerce will combine data about our behaviors and choices with data about products and product attributes to create the optimal shopping experience. This approach takes the guesswork out of purchasing and makes the shopping experience more enjoyable for everyone. ■

John O'Rourke is Vice President of Marketing at Indix. He earned his undergraduate degree and MBA from the University of Washington.

How Machine Learning and Predictive Analytics are Shaping the New Marketing Landscape

By Scott Etkin

Once the stuff of science fiction, machine learning is experiencing rapid growth and generating widespread interest due to its ability to deliver predictive insights and potential to transform virtually every industry. Marketing, for example, is seeing traditional approaches and processes upended by machine learning and unprecedented insight into customer preferences, needs, and behaviors.

Saatchi & Saatchi Wellness is full-service advertising and marketing services agency that provides wellness and pharmaceutical brands with everything from television advertising to CRM services. Within the company sits an analytics consultancy that provides, among other services, dashboards, reporting, data integration, and predictive and prescriptive analytics solutions.

Kevin Troyanos, SVP, marketing analytics at Saatchi & Saatchi Wellness, reviewed the machine-learning platform from DataRobot and said the product enables companies to develop robust models that perform well in a short period of time.

Troyanos spoke with *Data Informed* about how clients are using predictive analytics, demand trends, and evolving applications for predictive analytics.

Data Informed: How are your clients using your predictive analytics services? What kinds of challenges are they trying to address?

Kevin Troyanos: The key challenge that we are helping our clients address is around the question of commercial spend optimization. Our predictive-analytics services range across a multitude of distinct analysis offerings, with this specific goal in mind: Whether in predicting customers who are most likely to adopt a new product (or remain loyal to a mature one), determining the optimal mix of channels in order to maximize profitability, predicting macro-level responsiveness to broad national campaigns, or predicting micro-level responsiveness to personalized, targeted direct marketing tactics, we want to help our clients utilize their dollars in the most effective and efficient way possible by eliminating waste and uncovering growth opportunities. Predictive analytics enables us to, with a degree of probabilistic certainty, make smart, prescriptive recommendations on where

Kevin Troyanos, SVP, Saatchi & Saatchi Wellness Analytics

investment should be made and, perhaps more importantly, where wasted dollars can be safely divested or re-allocated.

What tasks do you see being automated as part of the predictive analysis process and how has this impacted how predictive analytics is used in organizations?

Troyanos: We see companies automating significant portions of their predictive analysis and model development. In years past, it may have taken weeks to develop and validate a predictive model for a specific use case. As a result, a high percentage of time was utilized for "in-the-weeds" model development rather than higher-level data strategy. Automation like the kind DataRobot affords increases capacity. With this tool, organizations can streamline our model-building process and exponentially speed up time to delivery while maintaining a high level of accuracy. This level of efficiency will help companies apply predictive analytics to a much wider variety of business problems, and enable organizations to leverage the power of big data.

How is demand for your analytics group's services trending?

Troyanos: Over the past year, the demand for our group's services has grown exponentially. More and more, we are being tapped for additional opportunities to leverage our data science capabilities for client needs – both within our agency and through partnerships with our sister agencies throughout the Publicis Health network. Our clients are continually looking for ways to do more with less and fewer resources. Our predictive analytics services help us to enable our clients to do just that: find ways to optimally invest marketing dollars in order to drive bottom-line results.

Looking ahead, have you identified additional areas and applications for predictive analytics?

Troyanos: One of the key areas of future focus that we have identified is in optimizing the efforts of large national field forces. Field force teams typically represent a significant marketing investment. These teams are effective in driving top-line growth, but it is critical to optimize their efforts. Machine-learning based predictive analytics can help to drive targeting models to optimize bottom-line efficiency of this investment – particularly in prescribing reach and/or frequency targets.

We also have identified the digital space as another key area of focus – in particular, around developing tools to scenario plan and dynamically benchmark website conversion funnel activity. Conversion rates are often a moving target for brands that drive various forms and levels of media investment. Predictive analytics applied to conversion activity can allow brands to predict this moving target in the face of differential media spend – and inversely, to plan accordingly. ■

Scott Etkin is the editor of Data Informed. Email him at Scott.Etkin@wispubs.com. Follow him on Twitter: @Scott_WIS.

Predictive Analytics Captures Value in Every Business Sector

by Giles Nelson

Giles Nelson, Senior Vice President, Product Strategy and Marketing, Software AG

For a long time, the use of data was essentially a matter of recording the past.

We had 40 years of the database management system and 20 years of data warehousing and business intelligence.

In the last 10 years, the development of event processing has enabled businesses to focus on the present, reacting to information immediately. In the financial markets, for example, it is possible to spot inappropriate trading as it happens.

From dealing with the present, we moved on to facing the future, analyzing historical data to reveal significant patterns. Now, we have reached the point at which we can advance to the next level of capability in analytics: that is, to combine the power of prediction with our immediate understanding of what is happening this instant.

Take the example of a telecommunications company with 30 million subscribers. It has built a predictive model that uses information on how users are accessing the network right now, whether it is making a phone call, accessing a website, or using particular apps.

The company combines the power of prediction and event processing to send offers at the most appropriate time. The model it has built will notice, for example, that a customer regularly makes international calls and is close to upgrading to a more suitable plan. Instead of sending that customer an offer when he or she is likely to be busy with other matters, the company sends the offer just after the customer has ended an overseas call, making the offer far more relevant and timely.

Sending the offer at the right time can be very powerful. Take, for example, a business traveler who is regularly at Heathrow Terminal 5. In this example, the traveler may have been shopping online with a retailer or has an account with the retailer that includes handing over a mobile phone number. The traveler also may have given the retailer permission to share her location with trusted partners.

Then as she walks past the store in the airport mall, she receives an offer for a product she was looking at before she left the office, informing her that it is in-store and that if she buys it within the next 20 minutes, she will receive a 10 percent discount. A slick operation of this nature is about the ability to interact immediately with customers.

The same kind of data processing and analysis that is behind this interaction already is deployed in manufacturing. Where companies rely on their own generators, for instance, maintenance is highly disruptive, taking as much as three months for the overhaul of each engine.

However, by using data from sensors monitoring cylinder pressure, temperature, or fuel consumption, it is possible to indicate when individual parts are at the point of failure. Using streaming analytics, the model will pick up a drop in pressure and, by correlating with records, predict that the current rate of wear will lead to malfunction within, say, four weeks.

Equally, in capital markets, the capacity to analyze vast streams of data from trading operations and other unstructured, external sources such as news feeds, chat rooms, and employment records will expose inappropriate activity when matched against norms established from historical records. The same techniques also will detect when automated trading algorithms deviate without cause, permitting interventions to prevent events such as the infamous Knight Capital loss of $440 million in 30 minutes, when its algorithms bought at ask prices and sold at bid prices.

In consumer finance, predictive analytics is allowing banks to detect credit card fraud as it happens and to take immediate action by blocking transactions rather than becoming aware of the crime after the event.

Train companies can use predictive and streaming analytics to optimize use of their track infrastructure. They can ensure that passengers are directed to the most suitable station platforms so that delays are minimized.

Further applications are set to occur in utility companies, where predictive analytics will allow them to operate their grids more efficiently in the age of renewable energy. As more residences and businesses generate electricity from solar installations, it remains difficult to match production and demand. Using predictive analytics in combination with live data, it is possible to smooth out the peaks and troughs, reducing waste.

> “ ”
> Predictive analytics is allowing banks to detect credit card fraud as it happens.

The organization that knows most about us is our bank. If banks can persuade us to let them use our data, they could predict what we might want to buy, the services they could provide, and much more.

But perhaps some of the most significant impacts will be in healthcare and medicine, where the advanced use of predictive analytics is set to change many practices. For example, in cases of patients with chronic conditions who wish to maintain independence by living in their homes, their movements can be tracked via a smart badge or wristband so that if they depart from their normal pattern, remote caretakers can be notified to check on them.

In an acute-care hospital ward, sensors that measure a patient’s vital signs, such as heartbeat and blood pressure, can feed data into a predictive model that will alert staff to the imminent danger of a bad event such as a heart attack. The model achieves its insights by matching current information against the corpus of data about patient reactions in the past.

So while the deployment of predictive analytics is already having a major impact, its effect is only just beginning.

For the consumer, handing over data in this way may require a constant process of negotiation with brands and third parties. It will be a question of exchanging data in return for better service. As this data market develops, it may be the kind of blockchain technology behind Bitcoin that emerges as a convincing protection mechanism.

However, provided that organizations put effort into building trust and security, they should be able to accumulate ever greater volumes of consumer, instrument, or trading data, with which they can build the predictive models that bring major gains in revenue and efficiency. ■

Giles Nelson is Senior Vice President, Product Strategy and Marketing, at Software AG.

For Greater Data Value, Spread Analytics to the Business

by Stephen Rohrer

Stephen Rohrer, Enterprise Analytics Data Scientist, TIAA-CREF

Ready to uncover massive insights from data, but don't have a data scientist on board? You are not alone. With deep talent gaps and enormous demand for these data superheroes, many organizations struggle to achieve success with analytics. In fact, McKinsey and Company predicts that the United States alone may face a 50 percent to 60 percent gap between supply and requisite demand of deep analytic talent.

However, all hope is not lost! By breaking analytics out of its IT silo and spreading it to the broader business, organizations can still garner enormous value.

We all have heard the excuses before: "That's nice, but it won't help us," "Maybe next year, when we have budget," or, "We'll take it into consideration." These comments come from well-meaning leaders who may not understand the possibilities for business growth that analytics can offer. Many feel overwhelmed at the mere glance of a complex algorithm, don't see the actionable insights from data, or simply have their focus elsewhere.

Even successful analytics programs have burdens and roadblocks. Our exposure to data science is cluttered with complex statistics and black-box algorithms, immediately building a 10-foot wall for business users wanting to make sense of the data. While developing statistics and algorithms is not easy, delivering them in a user-friendly and understandable way is even harder. Everyone has different questions based on their unique perspectives, is inundated with status meetings, drawn-out projects, and boring reports, which rarely change processes. As a result, many big data first-timers are left feeling overwhelmed with material and reluctant to embrace the analytics.

Shift Gears

It's time for a paradigm shift. What if we adapted the analytics to the business' needs and consumption, instead of trying to fit business needs into traditional analytics? By considering what the business users' experience will be with the analytics, IT generates more interest in the data and, as a result, provides more value to the business.

It is possible to have success with analytics with a few key ingredients:

- **The right people.** Build a team whose members complement each other and bring unique skills and abilities to the table.
- **The right tools.** By selecting technologies that make analytics easy for both analytics professionals and business users to understand, IT teams can "teach to fish" and eliminate back-and-forth communication to explain tools and do away with the headaches for both parties.
- **The right approach**. Do not be afraid to educate and advocate throughout the organization. This is a critical factor in determining the overall success of analytics.

Identify the Ingredients for Success

One of the most important factors to consider is the analytics team – emphasis on "team." In bringing together a group of professionals (not always the elusive unicorn!) with the right mix of analytical and creative skills, organizations can pave the way for new insights.

For instance, having a "dreamer" on board lets you imagine different scenarios based on the data or predict possible future outcomes. He can bring the team out of the technical weeds to help meet broader business objectives. The "liaison" serves as the central communicator between IT and business users, interpreting business needs and maintaining a strong relationship with the data science team. Without this bridge, business users remain stuck on the status quo and can fail to find use cases for analytics. "Artisans" and "storytellers" both play a critical role in pulling interesting nuggets from the data, which are then visualized and conveyed through actionable dashboards, intuitive data products, and behind the scenes machine learning.

Success also requires both imagination and creativity. As Bill Franks, chief analytics officer at Teradata, has said, "There's no need to be creative if you have the exact data for the exact problem … but the reality is that you have to make assumptions, deal with inconsistencies, and then choose a model that may not be perfect." Organizations must bring together a team that embraces this philosophy of creativity and brings necessary soft skills like commitment and curiosity to the table.

It is also important to build an ecosystem that the business side will use. Think about how others will consume the analytic insights, and deliver an intuitive experience via data products that invites them to come back for more. Many analytical tools still bring a learning curve, so be prepared to educate users on how to work with them. Take the time to work with business professionals on how to interpret the data and results. Stay committed to and engage with the entire organization to build lasting relationships, and don't forget the keyword: advocate!

Enjoy the Ride

By turning analytics into an experience for the enterprise, you can reduce project time from months and years to mere weeks. Insights that your team discovers become available to others, and analytics becomes as easy to consume as scrolling through your newsfeed. Your team can shift focus from reactive analysis to proactive solutions, while the business becomes more engaged and stakeholders participate and learn throughout the process. With analytics that are easily digested, distributed to everyone, and visually communicated, decisions are made instinctively and naturally. Businesses gain not only a solution, but also an understanding that can hold tremendous benefits for years to come. ■

Stephen Rohrer is an Enterprise Analytics Data Scientist at TIAA-CREF, a leading financial services provider to those who serve others. He holds a Master of Science in Analytics from NC State's Institute of Advanced Analytics, is a SAS certified professional, and is an active analytics advocate. Stephen has worked in multiple industries and fields including logistics, government, healthcare, marketing, IT, and financial services. When he's not concocting data products, Stephen enjoys time with his wife and two children, and is an avid gamer who scrapes streaming and server statistics for fun.

Better Predictions. Faster.

A Practical Guide to Automated Machine Learning

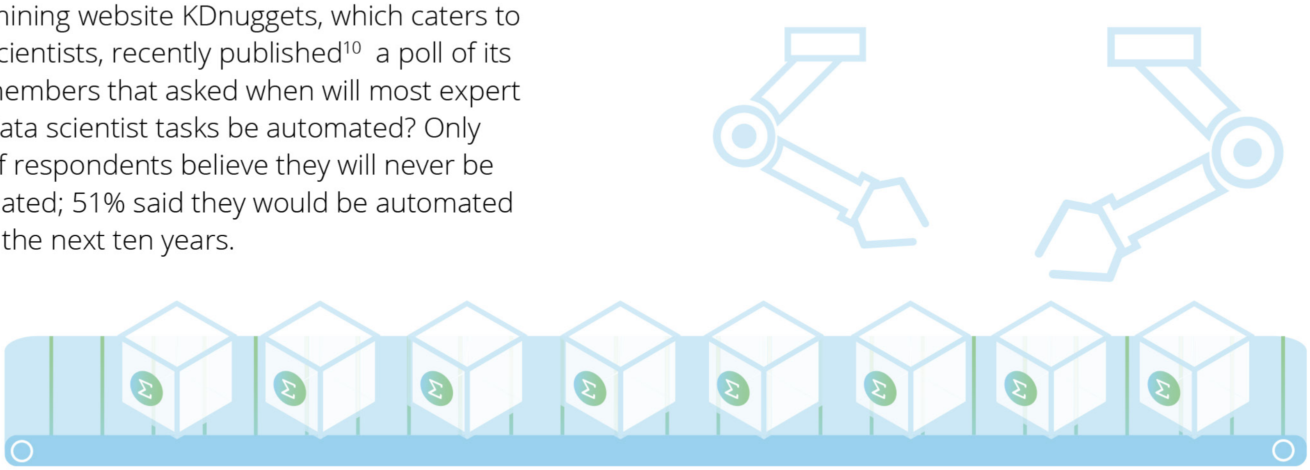
Data Science: A Scarce Resource.

Machine learning drives business advantage everywhere. With better predictions, banks reduce losses from credit defaults and fraud; insurers develop more competitive pricing; retailers personalize offers to consumers; hospitals improve patient outcomes; and telecommunications providers optimize bandwidth.

Businesses collect massive amounts of data, build data science departments, buy data science technology, and hire data scientists all to realize the power of machine learning. Hadoop and other technologies make the collection of data commonplace, but data scientists who make sense of it are still a scarce resource. VentureBeat[1], The Wall Street Journal[2], The Chicago Tribune[3] and many others all note the scarcity; a McKinsey report[4] projects a shortage of people with analytical skills through 2018. The scarcity is so pressing that Harvard Business Review suggests[5] that you stop looking, or lower your standards.

One way to alleviate the pressure, may be to automate some of the work data scientists currently perform. In IT Business Edge, Loraine Lawson wonders[6] if artificial intelligence will replace the data scientist. In a Sloan Management Review article headlined Data Scientist in a Can, Michael Fitzgerald argues[7] that companies are already trying to automate the function; however, he fails to distinguish between outsourcing analytics – which companies have done for years – and automating analytics, which is quite different. In Forbes, technology thought leader Gil Press confidently asserts[8] that the data scientist will be replaced by tools; Scott Hendrickson, Chief Data Scientist at social media integrator Gnip, agrees[9].

Data mining website KDnuggets, which caters to data scientists, recently published[10] a poll of its own members that asked when will most expert level data scientist tasks be automated? Only 19% of respondents believe they will never be automated; 51% said they would be automated within the next ten years.



[1] http://venturebeat.com/2015/03/17/why-data-scientists-and-marketing-technologists-are-the-hottest-jobs-of-2015/

[2] http://blogs.wsj.com/cio/2014/11/10/for-cios-universities-cant-train-data-scientists-fast-enough/

[3] http://www.chicagotribune.com/business/ct-indeed-survey-0514-biz-20150514-story.html

[4] http://www.mckinsey.com/features/big_data

[5] https://hbr.org/2014/09/stop-searching-for-that-elusive-data-scientist/

[6] http://www.itbusinessedge.com/blogs/integration/will-artificial-intelligence-replace-the-data-scientist.html

[7] http://sloanreview.mit.edu/article/data-scientist-in-a-can/

[8] http://www.forbes.com/sites/gilpress/2012/08/31/the-data-scientist-will-be-replaced-by-tools/

[9] https://blog.gnip.com/data-scientist-vs-data-tools/#

[10] http://www.kdnuggets.com/polls/2015/analytics-data-science-automation-future.html

The March of Automation:

Evidence from other fields is encouraging.

- The Washington Post summarizes[11] successful efforts to automate anesthesia during surgery.
- The MIT Technology Review reports[12] on a machine learning algorithm that classifies paintings more accurately than trained art historians.
- A report from consulting firm A.T.Kearney projects[13] that automated "Robo Advisors" will run $2 trillion in investment portfolios by 2020.
- An article in The Atlantic notes[14] that nearly half of American jobs could be automated.

Automation is neither an all-or-nothing phenomenon, nor does it happen overnight. Consider the automobile, a highly evolved technology. In the era of the Model T Ford, driving was hard; only a fraction of the population could operate the vehicle, change a tire or troubleshoot frequent breakdowns. Over time, auto manufacturers simplified the driving process and made cars more reliable, expanding the pool of potential drivers.

Today, most adults can drive; mass-market cars are loaded with automated features: adaptive cruise control, lane-keeping systems, driver alert systems and blind spot indicators. Full automation is just over the horizon; earlier this year, the state of Nevada cleared[15] Freightliner's self-driving[16] truck for use on public highways.

Reflecting the incremental and progressive nature of automation, the National Highway Transportation Safety Administration's policy[17] on automated vehicles defines five levels of autonomy:

- Level 0: the driver is in complete command at all times.
- Level 1: one or more specific control functions automated.
- Level 2: at least two primary control functions automated to work together.
- Level 3: the driver can relinquish full control under certain conditions.
- Level 4: the vehicle performs all functions for the entire trip.

Suppose we apply a similar model to machine learning; our "levels of autonomy" look something like this:

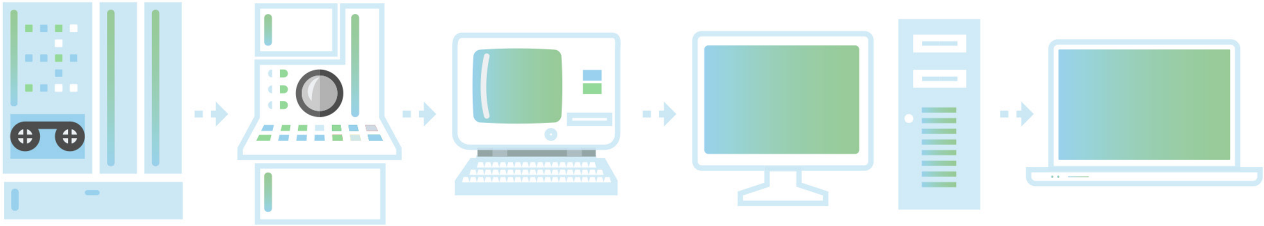
- Level 0: the analyst manually codes all steps in the modeling process.
- Level 1: the software automates one or more individual steps, such as sample splitting.
- Level 2: the software automates multiple steps, such as training and pruning a tree.
- Level 3: the analyst defines a test plan and the software executes the plan.
- Level 4: the analyst specifies a target and the software builds a production-ready model.

This perspective reveals two key insights. First, before we can automate machine learning, we have to define the steps in the process and the desired outcomes – otherwise, we don't know what it is that we're automating. Second, it's clear that machine learning is already automated to a considerable degree. Even the most hardcore

[11] https://www.washingtonpost.com/news/the-switch/wp/2015/05/15/one-anesthesiology-robot-dips-its-toes-into-whats-possible-this-one-jumps-all-in/
[12] http://www.technologyreview.com/view/537366/the-machine-vision-algorithm-beating-art-historians-at-their-own-game/
[13] http://www.bloomberg.com/news/articles/2015-06-18/robo-advisers-to-run-2-trillion-by-2020-if-this-model-is-right
[14] http://www.theatlantic.com/business/archive/2014/01/what-jobs-will-the-robots-take/283239/
[15] https://www.newscientist.com/article/dn27485-autonomous-truck-cleared-to-drive-on-us-roads-for-the-first-time/
[16] http://www.ccjdigital.com/behind-the-wheel-of-freightliners-inspiration-autonomous-truck/
[17] http://www.nhtsa.gov/About+NHTSA/Press+Releases/U.S.+Department+of+Transportation+Releases+Policy+on+Automated+Vehicle+Development

data scientists who work in programming languages use Integrated Development Environments (IDEs) that automate routine programming tasks. Tools like this are at least at the second level of automation.

Automated Machine Learning: A Brief History



Automated machine learning is not new. Before Unica launched its successful suite of marketing automation software, the company's primary business was machine learning, with a particular focus on neural networks. In 1995, Unica introduced Pattern Recognition Workbench (PRW), a software package that used automated trial and error to optimize model tuning for neural networks. Three years later, Unica partnered with Group 1 Software (now owned by Pitney Bowes) to market Model 1, a tool that automated model selection over four different types of predictive models. Rebranded several times, the original PRW product remains as IBM PredictiveInsight, a set of wizards sold as part of IBM's Enterprise Marketing Management suite.

Two other commercial attempts at automated machine learning date from the late 1990s. The first, MarketSwitch[18], consisted of a solution for marketing offer optimization, which included an embedded "automated" machine learning capability. In sales presentations, MarketSwitch bragged that it had hired former Soviet rocket scientists to develop its software, and promised customers they would be able to "fire their SAS programmers". Experian acquired MarketSwitch in 2004, repositioned the product as a decision engine and replaced the "automated machine learning" capability with its own outsourced analytic services.

KXEN, a company founded in France in 1998, built its machine learning engine around an automated model selection technique called structural risk minimization[19]. The original product had a rudimentary user interface, depending instead on API calls from partner applications; more recently, KXEN repositioned itself as an easy-to-use solution for Marketing analytics, which it attempted to sell directly to C-level executives. This effort was modestly successful, leading to sale of the company in 2013 to SAP for an estimated[20] $40 million.

Early efforts at automation from Unica, MarketSwitch and KXEN failed to make an impact for two reasons. First, they "solved" the problem by defining it narrowly; limiting the scope of the solution search to a few algorithms, they minimized the engineering effort at the expense of model quality and robustness. Second, by positioning their tools as a means to eliminate the need for expert analysts, they alienated the few people in customer organizations who understood the product well enough to serve as champions.

In the last several years, the leading analytic software vendors (SAS and IBM SPSS) have added automated modeling features to their high-end products. In 2010, SAS introduced SAS Rapid Modeler[21], an add-in to SAS

[18] http://www.experian.com/decision-analytics/marketswitch-optimization.html
[19] http://www.svms.org/srm/
[20] https://451research.com/report-short?entityId=79713
[21] https://support.sas.com/resources/papers/proceedings10/113-2010.pdf

Enterprise Miner. Rapid Modeler is a set of macros implementing heuristics that handle tasks such as outlier identification, missing value treatment, variable selection and model selection.

The user specifies a data set and response measure; Rapid Modeler determines whether the response is continuous or categorical, and uses this information together with other diagnostics to test a range of modeling techniques. The user can control the scope of techniques to test by selecting basic, intermediate or advanced methods. In 2015, SAS introduced a new generation of this product branded as SAS Factory Miner.

IBM SPSS Modeler includes a set of automated data preparation features as well as Auto Classifier, Auto Cluster and Auto Numeric nodes. The automated data preparation features perform such tasks as missing value imputation, outlier handling, date and time preparation, basic value screening, binning and variable recasting. The three modeling nodes enable the user to specify techniques to be included in the test plan, specify model selection rules and set limits on model training.

All of the software discussed so far is commercially licensed. The caret[22] package in open source R includes a suite of productivity tools designed to accelerate model specification and tuning for a wide range of techniques. The package includes pre-processing tools to support tasks such as dummy coding, detecting zero variance predictors, identifying correlated predictors as well as tools to support model training and tuning.

The training function in caret currently supports 192 different modeling techniques; it supports parameter optimization within a selected technique, but does not optimize across techniques. To implement a test plan with multiple modeling techniques, the user must write an R script to run the required training tasks and capture the results.

Auto-WEKA[23] is another open source project for automated machine learning. First released in 2013, Auto-WEKA is a collaborative project driven by four researchers at the University of British Columbia and Freiburg University. In its current release, Auto-WEKA supports classification problems only. The software selects a learning algorithm from 39 available algorithms, including 2 ensemble methods, 10 meta-methods and 27 base classifiers. Since each classifier has many possible parameter settings, the search space is very large; the developers use Bayesian optimization to solve this problem[24].

Challenges in Machine Learning (CHALEARN)[25] is a tax-exempt organization supported by the National Science Foundation and commercial sponsors. CHALEARN organizes the annual AutoML[26] challenge, which seeks to build software that automates machine learning for regression and classification. The most recent conference[27], held in Lille, France in July, 2015, included presentations[28] featuring recent developments in automated machine learning, plus a hackathon.

[22] http://topepo.github.io/caret/index.html
[23] http://www.cs.ubc.ca/labs/beta/Projects/autoweka/#papers
[24] http://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/autoweka.pdf
[25] http://www.chalearn.org
[26] http://automl.chalearn.org
[27] https://sites.google.com/site/automlwsicml15/
[28] https://indico.lal.in2p3.fr/event/2914/

Designing an Automated Machine Learning Platform.

Requirements for a modern automated machine learning platform fall into two categories: support for the machine learning process, and support for enterprise computing.

Automated machine learning software should support the machine learning process from beginning to end:

- Rapid Data Ingestion. Data scientists need to leverage data from across the organization and from external sources. For speedy data ingestion, automated machine learning software should support open-standards based interfaces with relational databases and Hadoop, as well as text files and common desktop formats.

- Automated Data Preparation. Data scientists use expert judgment to examine raw data and make decisions about how to work with it in a predictive model. Automated machine learning software should perform this diagnosis and present the results in clear and concise visuals.

  Data scientists routinely split raw data into learning, validation and holdout samples so they can validate models and protect against overfitting. Automated machine learning software should perform this step for the user and protect the holdout sample.

- Automated Test Planning. There are hundreds of potential algorithms; a recent benchmark study tested[29] 179 for classification alone. The best way to determine the right algorithm for a given problem and data set is a test and learn approach, where the data scientist tests a large number of techniques and chooses the one that works best on fresh data. (The No Free Lunch No Free Lunch[30] theorem formalizes this concept).

  Using information about the target and predictor variables, automated machine learning software should select the most appropriate techniques from the available universe, generate a test plan and execute the plan.

- Automated Feature Engineering. Each machine learning technique has a distinct way that it handles data, which may require pre-processing before model training can begin. Also, data scientists use best practices in feature engineering to prepare data for better results. Automated machine learning software should incorporate and use this expertise

- High-Performance Model Training. Model training is computationally intensive. Moreover, even with heuristics and bootstrapping, a comprehensive experimental design may require thousands of

[29] http://jmlr.org/papers/v15/delgado14a.html
[30] http://www.no-free-lunch.org

model train-and-test cycles. Automated machine learning software should leverage state-of-the art computing for high performance and rapid learning.

- Transparent Model Evaluation. For "hard money" predictive models that have strategic implications for a company, no executive will approve deployment without first understanding the model's behavior and validity.

  Automated machine learning software should provide tools so that expert and business users can evaluate the results of a modeling experiment, check for bias, compare models and, if appropriate, override the automatic model selection.

- Real-Time and Batch Deployment. The greatest predictive model in the world is worthless if it is not used. That is what happened with the Netflix Prize winner[31], a sophisticated application of a technique called Pragmatic Chaos that now resides in the dustbin of history; Netflix paid out the prize and buried[32] the solution because it was too expensive to deploy. Automated machine learning software should include scalable engines for both batch scoring and for real-time predictions.

To serve the needs of the modern enterprise, there are three additional requirements:

- Open Source Analytics. Automated machine learning software should build on a foundation of open source software. The cadence of innovation in open source analytic languages, such as Python and R, is much faster than in commercial software. Moreover, an open source foundation simplifies integration with Big Data stacks and reduces cost of ownership.

- Business and Expert User Interfaces. Automated machine learning software should support diverse user personas, including:

  - Expert users who want to write custom code.
  - Analysts with deep statistical training but limited programming skills.
  - Sophisticated business users who want to engage with the model development process.
  - Business users who want to visualize the characteristics of a model and how it behaves.

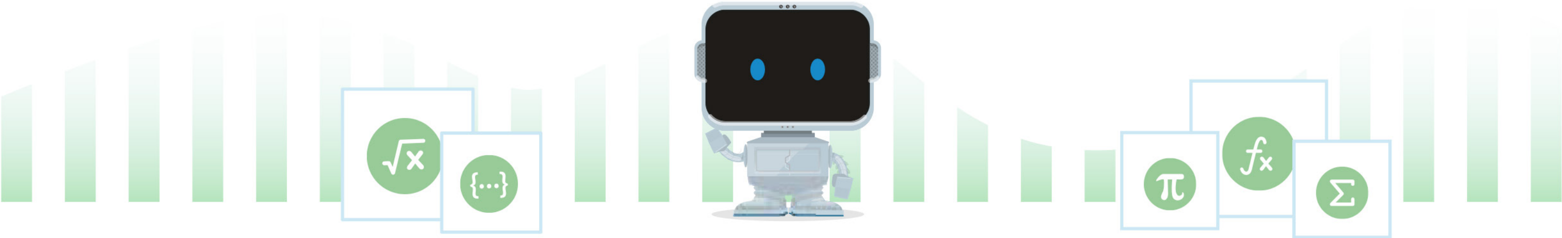
- Enterprise Scalability. Automated machine learning software must scale to the enterprise level, on many dimensions, measured by users, projects, models and data volume. In practical terms, this means that the software should support deployment in Hadoop, standards-based integration with databases and support for low-maintenance provisioning: on premises or in the cloud.

  Any automation is better than no automation. But for a truly agile and automated workflow, machine learning software should meet all of these requirements.

[31] http://www.wired.com/2009/09/how-the-netflix-prize-was-won/

[32] http://thenextweb.com/media/2012/04/13/remember-netflixs-1m-algorithm-contest-well-heres-why-it-didnt-use-the-winning-entry/

Introducing DataRobot:



DataRobot, a data science and machine learning company located in Boston, Massachusetts, offers a platform for users of all skill levels to build and deploy accurate predictive models in a fraction of the time needed using conventional tools and methods.

DataRobot uses massively parallel processing to train and evaluate thousands of models in R, Python, Spark MLlib, H2O and other open source libraries. It searches through millions of possible combinations of algorithms, pre-processing steps, features, transformations and tuning parameters to deliver the best models for your dataset and prediction target.

How does DataRobot stack up against the requirements for automated machine learning software outlined earlier in this chapter? Let's review:

- Rapid Data Ingestion. DataRobot loads data from text files, from URLs, through ODBC connections and from HDFS. You can deploy DataRobot in Hadoop, so your data never leaves HDFS.
- Automated Data Preparation. When DataRobot reads data, it automatically detects duplicate variables and variables with no information value – such as blanks and constants. DataRobot profiles target and predictor variables, and uses this information later in the process to build optimal test plans. DataRobot automatically partitions your input dataset into learning, validation and holdout dataset. It locks the holdout dataset, so that users do not accidentally use it prior to final model selection.
- Automated Test Planning. With built-in expertise, DataRobot uses information about your target variable and predictors to define a list of models to test.
- Automated Feature Engineering. DataRobot builds feature engineering into the modeling test plan, so it automatically preprocesses data for best results with the technique to be tested. At the most basic level, DataRobot actually tests pipelines, or "blueprints" consisting of multiple processing steps defined by a team of world-class data scientists.
- High-Performance Model Training. DataRobot's modern software architecture builds on the most current tools and methods to enable rapid parallel execution of large scale modeling experiments.
- Transparent Model Evaluation. Once DataRobot has trained and tested a battery of models, users can work with a number of tools to assess accuracy, understand how the model behaves and identify possible

bias. Users can evaluate model accuracy with many different metrics including AUC, Log-Loss and RMSE. DataRobot uniquely provides partial dependency plots, a technique that helps users visualize how individual variables affect the prediction, a vital tool in areas such as credit risk, where bias can seriously affect the value of a predictive model.

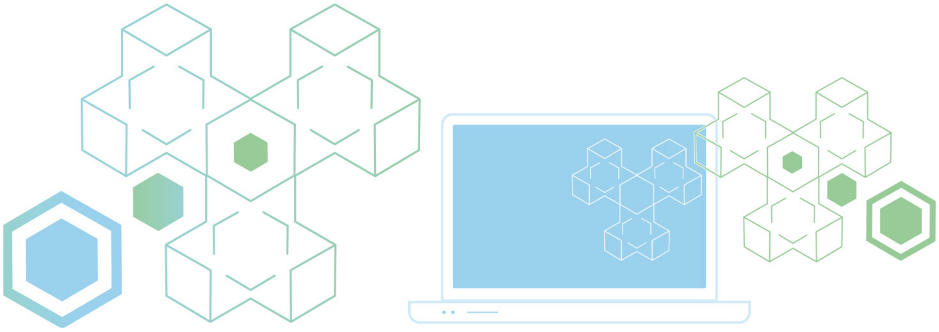
- Real-Time and Batch Deployment. DataRobot offers a real-time prediction engine and a batch prediction engine. The real-time prediction engine is designed so that organizations can implement multiple instances with load balancing to achieve a desired level of speed and throughput. The batch prediction engine runs in Spark, and it can run either in a freestanding Spark instance or it can be co-located with Hadoop.

- Open Source Analytics. DataRobot builds on R, Python, H2O, Spark and XGBoost for machine learning techniques, and open source technologies such as Docker, Hadoop, Spark, GlusterFS, Redis and MongoDB for implementation.

- Business and Expert User Interfaces. For the business user, DataRobot offers an easy-to-use web interface complete with visualization tools to evaluate predictive models. For the expert user, DataRobot offers R and Python APIs for modeling and separate APIs for prediction. Developers can build production-ready processes that call the prediction API to incorporate the power of predictive modeling into applications.

- Enterprise Scalability. DataRobot scales out, not up. Organizations deploy DataRobot on premises in free-standing clusters or in Hadoop, or in the cloud. DataRobot scales easily as usage grows; all components support distributed deployment across multiple virtual or physical machines. DataRobot distributes and manages workloads across the datacenter resources allocated to it.

Implications for Data Science.



The power and impacts of automation on business, and the profession of data science is undeniable. Organizations that embrace this new technology will have a sustainable competitive advantage, and now that the automated machine learning ‘genie is out of the bottle’ time is of the essence. Those that move fastest, will gain the best advantage.

Automating many of the processes associated with machine learning and predictive analytics empowers the data scientist to shift focus from routine coding and data wrangling to tasks that add real value: understanding the business problem, the deployment context and explaining results. Automation also changes the mix of people engaged in the predictive analytics process, with more focus on business skills and domain knowledge than pure coding skills.

Applying automation to data science is a transformative process, and like any form of digital transformation, should be undertaken in stages. Start by identifying manageable yet impactful initial projects, develop focused small teams, move and learn quickly, celebrate success, and scale as organizational skill and business needs dictate.

We can't eliminate the need for human expertise from predictive analytics, for the same reason that robotic surgery does not eliminate the need for surgeons. Someone has to understand the business problem and explain the results of analysis to executives; models don't explain themselves. We can build tools that improve the productivity of data scientists and help them build better models.

Check out additional content on Data Informed

Find other articles like these and more at Data Informed: data-informed.com

Data Informed gives decision makers perspective on how they can apply big data concepts and technologies to their business needs. With original insight, ideas, and advice, we also explain the potential risks and benefits of introducing new data technology into existing data systems. Follow us on Twitter, @data_informed

DataInformed

Data Informed is the leading resource for business and IT professionals looking for expert insight and best practices to plan and implement their data analytics and management strategies. Data Informed is an online publication produced by Wellesley Information Services, a publishing and training organization that supports business and IT professionals worldwide. © 2016 Wellesley Information Services.