



Realizing the Benefits of Automated Machine Learning

Thomas H. Davenport

Realizing the Benefits of Automated Machine Learning

Thomas H. Davenport

Machine learning is a popular idea, but one that's been around for several decades. The idea behind the most common form of it – supervised learning – is to train a model (producing a good fit of the model to the data) on a set of data for which the outcome is known, and then to use the resulting model to predict or classify the outcomes of a different set of data. The first stage of this process is often referred to as “model development”; the second stage as “model deployment” or “scoring.”

What is new about machine learning are three factors:

- The growing popularity of machine learning in organizations;
- The use of multiple, sometimes complex algorithms to create a more effective model;
- The availability of automated machine learning approaches to make the process more efficient and accessible throughout organizations.

The popularity of machine learning is due to multiple factors, including the enormous rise in the amount of data that needs to be analyzed, the need for ever more granular analyses of the data, the availability of a variety of algorithms to analyze the data, and the need to deploy predictive and categorization models into production systems and business processes.

The rise of analytics and big data has led to many new or rediscovered algorithms. Most statistical analyses in the past used linear regression analysis. More recently, logistic regression become much more popular for making predictions of binary outcomes. Now, a wide range of algorithms is available to the machine learning modeler. The data and algorithms



Thomas H. Davenport is the President's Distinguished Professor in Management and Information Technology at Babson College, has taught at Harvard Business School and Dartmouth's Tuck School of Business and directed research centers at Accenture and McKinsey.

An author or coauthor of 15 books and more than 100 articles, he helps organizations to revitalize their management practices in areas such as analytics, information and knowledge management, process management, and enterprise systems.

About This Study

This study of the benefits of automated machine learning was conducted by the author and was sponsored by DataRobot. The company did not influence the content of the report, but did furnish customers for interviews. While the report primarily addresses automated machine learning as a general category, some comments may apply only to DataRobot users, since they were the customers interviewed.

Nine companies that have adopted DataRobot were interviewed (eight by telephone, one by email). The individuals interviewed were either heads of analytics or machine learning groups, IT managers, or heads of functions that make extensive use of analytics and machine learning. Interviews were conducted with companies based in Japan, Canada, and the United States.

are expanding rapidly, but human capabilities — even those of quantitative professionals and data scientists — are not. Therefore, automating many of the activities in machine learning is perhaps the only way for the supply of analytical capabilities to meet the demand. I'll refer to these capabilities as “automated machine learning,” or AutoML.

AutoML promises to transform the machine learning process like nothing since its inception. AutoML typically performs some combination of the following functions:

- Automated data preparation;
- Automated feature engineering;
- Automated competitions among different algorithm types;
- Generation of explanations for why certain variables or features are most influential in models;
- Creation of program code or APIs for model deployment.

The benefits of these functions and of adopting automated machine learning will be discussed below. Details of the study behind this report are in the “About This Study” sidebar.

The remainder of this report addresses the following types of benefits from AutoML, along with a final section about how organizations can overcome challenges to benefit realization:

- Broad Machine Learning Dissemination
- Modeling Productivity
- Reduced Skill Requirements
- Improved Deployment Capabilities
- Model and Feature Explanation
- Overcoming Obstacles to Benefit Realization

Broad Machine Learning Dissemination

Given that machine learning is a powerful resource from which many parts of organizations can benefit, broad dissemination of the approach is a way to improve analytics and organizational performance. AutoML tools can be a key to facilitate greater usage of machine learning throughout an organization. Several companies have undertaken specific initiatives to increase machine learning usage throughout their firms, and AutoML is one of the tools they have employed.

For example, 84.51°, a company wholly owned by Kroger that does analytical and AI projects for the retailer and its suppliers, undertook a structured effort to embed advanced machine learning methods throughout their organization. 84.51° needs to use machine learning because it deals with large volumes of data and models that need to be detailed and updated frequently. For example, it employs a supply chain planning and ordering system that predicts sales for each item in each store for each of the subsequent 14 days. The system generates item level, daily sales forecasts for each of the 2,500+ stores in the Kroger enterprise. The system directly supports orders of thousands of products per store, per day. That scale of analytics couldn't be performed with traditional "artisanal" methods employing substantial human intervention.

The effort, called Embed Machine Learning, or EML, is a formal mission to enable, empower, and engage the organization to better use and embed machine learning. "Enable" meant providing the infrastructure to efficiently use and embed machine learning such as the servers, software, and data connectivity. "Empower" involved identification of the best set of machine learning tools and training analysts and data scientists to use those tools. After evaluating more than fifty tools, 84.51° selected R, Python, and Julia as its preferred programming languages, and settled on DataRobot's AutoML software as its primary machine learning platform. "Engage" meant motivating internal clients to use the tools by demonstrating and socializing the benefits through several proofs of concept, advancing code sharing/examples (via GitHub), and consulting.

At a large U.S. bank, the head of a major business unit became convinced that machine learning could provide substantial benefits for the unit, and that AutoML and DataRobot were important elements in advancing the opportunity. The unit already had a strategic plan, and the leader and his direct reports went back through it and identified three strategic business problems that could be addressed by machine learning. DataRobot's Chief Data Scientist spent several weeks at the unit learning the business. He and some DataRobot colleagues put together a day of training in machine learning and data science for the unit's top 40 executives. It taught them the vocabulary of machine learning and the applications most likely to be supported by it. The group spent an afternoon brainstorming ideas about applications within the bank and the business unit. Seven managers were identified out of the session as champions for machine learning in their functions.

Now there are more than 40 active use cases being pursued in some stage of the lifecycle, all within the same business unit. Two other groups within the bank are now using DataRobot as well. A few projects have stalled, but the majority are already yielding business value, and several are now fully deployed as production capabilities.

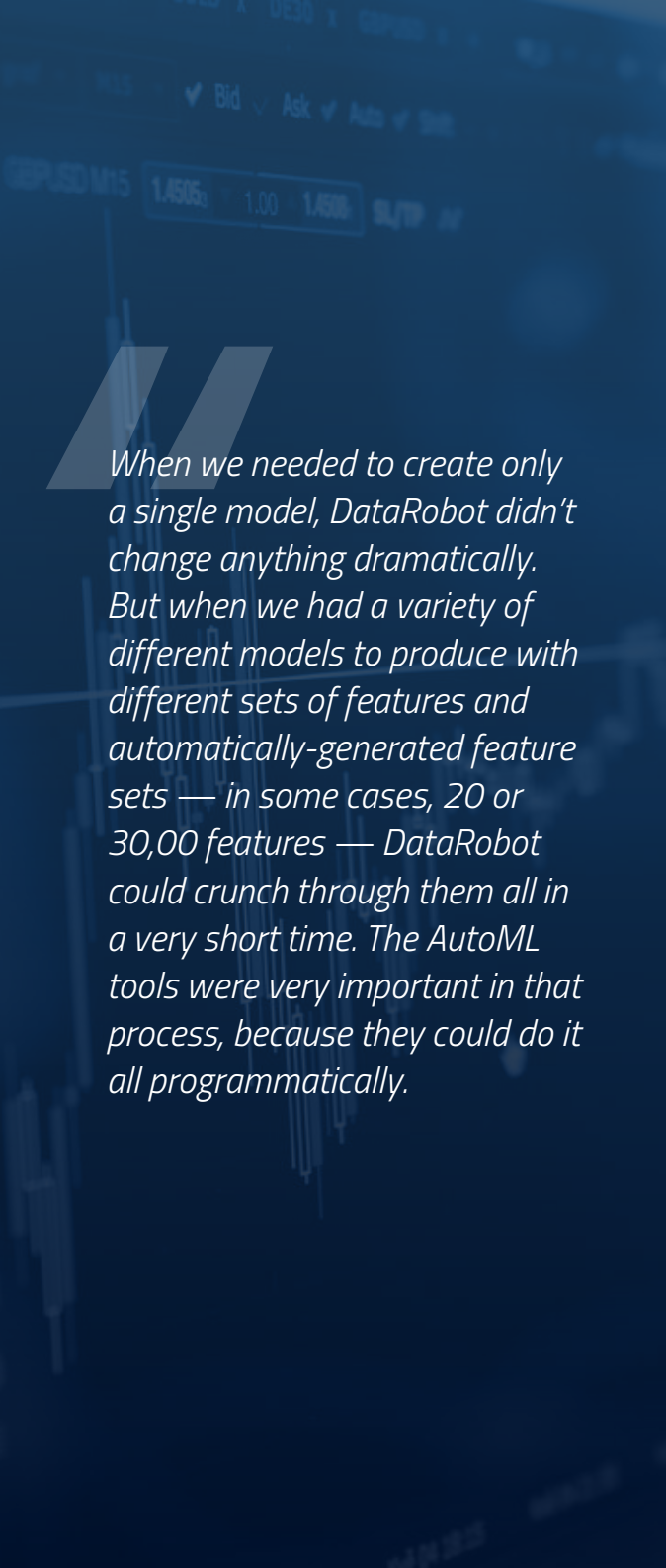
These examples suggest that a concerted intervention to increase the use of machine learning — accompanied by AutoML tools that can make it feasible — can yield a large number of successful projects within a company.

Modeling Productivity and Effectiveness

One of the most common benefits cited by companies using AutoML is increased productivity and effectiveness in creating analytical models. Creating models is a core activity for machine learning. It involves such activities as feature (variable) selection and engineering, data preparation, selection of algorithms, and evaluation and comparison of results. DataRobot performs these activities automatically with little need for intervention by a human analyst or data scientist. The result is both substantially greater productivity and more effective models.

At a large U.S. property and casualty (P&C) insurance company, for example, modeling productivity was the primary objective in adopting AutoML. Thus far, notes the head of data science support, “it has been a very helpful throughput tool.” The insurance giant uses DataRobot to get a quick reading on the ROI of alternative machine learning projects. “We get some data, turn DataRobot loose on it, and see what the prediction accuracy is for the model. It’s so quick that we can figure out the value of an analysis without taking a lot of time to assess it,” notes the manager. The company can learn what the key parameters of the model are, what algorithm is best suited to the problem, and what the likely ceiling is on model accuracy. If it seems to be a promising analysis, the company will take it further and perhaps put it into production.

At Sumitomo Mitsui Card Company (SMCC), the largest credit card company in Japan, DataRobot has been applied both to risk modeling and customer insight/marketing applications. In the risk modeling area, some analysts were doing machine learning manually, but it could take up to half a year to build and validate a model. Using DataRobot cut that time to hours or a few days. Hiroki Shiraishi, who leads a group providing machine learning infrastructure to SMCC’s business units, notes that the company wanted to accelerate the process of analyzing credit card data, and there were not enough skilled analysts to meet the need. Therefore, increasing modeling productivity was a key objective.



When we needed to create only a single model, DataRobot didn't change anything dramatically. But when we had a variety of different models to produce with different sets of features and automatically-generated feature sets — in some cases, 20 or 30,00 features — DataRobot could crunch through them all in a very short time. The AutoML tools were very important in that process, because they could do it all programmatically.

A data science group leader in a European insurance firm commented on the most important areas for greater modeling productivity. He noted: “When we needed to create only a single model, DataRobot didn’t change anything dramatically. But when we had a variety of different models to produce with different sets of features and automatically-generated feature sets — in some cases, 20 or 30,000 features — DataRobot could crunch through them all in a very short time. The AutoML tools were very important in that process, because they could do it all programmatically.”

In addition to the automation of key steps in the modeling process, several companies interviewed also cited the DataRobot user interface as a key reason why modeling productivity increased. Some also mentioned the ability to obtain virtually unlimited computing capabilities in the cloud for faster throughput.

At least one user of AutoML has measured the improvement in modeling productivity. In a pilot using DataRobot, the Data Science function at Sompo Holdings, a Japan-based global insurance company, found that the time required to develop models in a particular underwriting domain was reduced by 73% (from 13 person days to 3.5). The predictive ability of the models was also increased by about 5% on average.

Reduced Skill Requirements

Another key reason why companies adopt AutoML is to enable machine learning work by employees who have less sophisticated data science and analytical skills. It is widely known that there is a shortage of advanced skills, and since AutoML does a lot of the expertise-based tasks in machine learning — feature engineering and algorithm selection, for example — it can help to alleviate the skill shortage. At one large U.S. bank, for example, a junior-level employee whose primary interest in college was to become a ballerina — with an undergraduate degree in sociology — was able to use DataRobot to develop a successful model identifying

investments that would prosper if Donald Trump won the 2016 election. The bank and its clients made money from the model, and the ballerina was promoted rapidly up the investment banking hierarchy as a result.

At the same bank today, the team charged with disseminating the AutoML tool has a choice between presenting DataRobot and AutoML capabilities to highly trained data scientists or to business analysts. They generally move toward business analysts — people who don't code, but have access to data and domain knowledge. They can provide analysts with tools that do all they need for effective machine learning without writing a line of code. The team faces more resistance from data scientists, who “like to do what they like to do — they care about technology and leading-edge data science, but are sometimes poorly integrated with the business.” The team feels that it is easier to provide business domain experts with machine learning horsepower than to get data scientists to engage with the business.

A similar pattern of usage prevails with AutoML at Royal Bank of Canada (RBC), the large Canadian bank. It is investing in artificial intelligence and machine learning, with currently over 200 data scientists working across the bank. Samer Nusier, the bank's Director of Portfolio Management and Credit Strategy, explained that many of the bank's serious data scientists prefer to develop and tune their models on their own. He, however, is an advocate of the “citizen data scientist.” He notes that of the three traditional data science skills — math, computer science, and business domain knowledge — the math and computer science work are increasingly being done by tools like DataRobot. When business analysts who understand the data and customer behavior create the models, they can be as useful as models created by data scientists. “It gives them superpowers,” he notes. Nusier feels that “purple people” — those who understand both some analytics and are business experts — can be equally valuable if supported by AutoML.

Not all users of AutoML found that it should only be targeted at business analysts with relatively few quantitative skills. AirBnB, for example, supplied it to professional data scientists.

Our view is that it is difficult to perform wholesale replacement of a data scientist with an AML [AutoML] framework, because most machine learning problems require domain knowledge and human judgement to set up correctly. In summary, we believe that in certain cases AML can vastly increase a data scientist's productivity, often by an order of magnitude.

Their conclusion is as follows: “Our view is that it is difficult to perform wholesale replacement of a data scientist with an AML [AutoML] framework, because most machine learning problems require domain knowledge and human judgement to set up correctly. In summary, **we believe that in certain cases AML can vastly increase a data scientist’s productivity, often by an order of magnitude.**”¹

Improved Deployment Capabilities

Deployment of machine learning models is, by all accounts, an important aspect of the process of effectively using machine learning. It is the process by which analytical models created in the machine learning process are embedded within other systems and processes for purposes of “scoring” cases where the outcome variables are not known. Production systems with machine learning models have to be available anytime, have low latency, and high throughput. Unfortunately, many machine learning models — even perhaps a majority of them — are never deployed because the requirements are so difficult, and the part of the organization that does deployment is different from the model development group. Fortunately, there are deployment capabilities within DataRobot. While they may be used less frequently than the model development capabilities, some firms have found considerable value from them.

84.51° is perhaps the organization that has made most use of the deployment capabilities within AutoML in this study. Scott Crawford, who heads the Embed Machine Learning initiative at the company, points out that issues around deployment (or “productionalization,” as he refers to it) are often underestimated: “Prior to my current role facilitating the use of machine learning at 84.51°, my work experiences included building and deploying models at one of the nation’s largest insurance companies and one of the world’s largest banks. One commonality across all my experiences is that productionalization is often the most challeng-

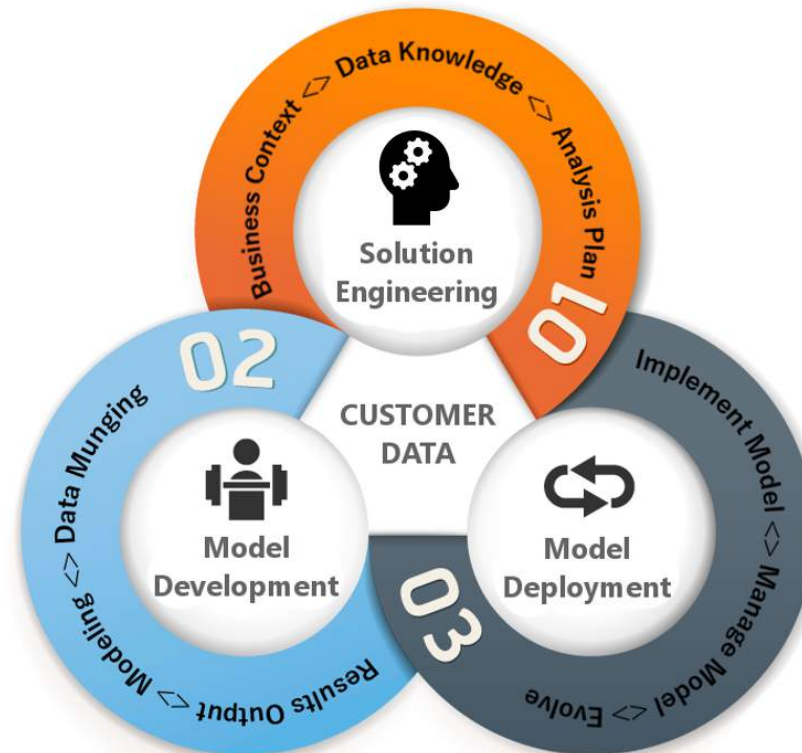
Prior to my current role facilitating the use of machine learning at 84.51°, my work experiences included building and deploying models at one of the nation’s largest insurance companies and one of the world’s largest banks. One commonality across all my experiences is that productionalization is often the most challenging phase of machine learning projects.

¹Hamel Husain and Nick Handel, “Automated Machine Learning — A Paradigm Shift That Accelerates Data Scientist Productivity @ Airbnb,” Medium website, May 10, 2017, <https://medium.com/airbnb-engineering/automated-machine-learning-a-paradigm-shift-that-accelerates-data-scientist-productivity-airbnb-f1f8a10d61f8>

ing phase of machine learning projects. The requirement of a production deployment often severely constrains the viable solutions. For example, productionalization might require code to be delivered in a specific language e.g., C++, SQL, Java, etc.) and/or to meet strict latency thresholds.

84.51° has created a methodology for using machine learning in their business. The first stage is Solution Engineering, in which the model and analytical process are framed and specified. The second stage is Model Development. The third stage is, of course, Model Deployment, in which the chosen model is deployed in production systems and processes. Given the scale of machine learning applications at Kroger, this stage of the process is critical.

8PML Process Model



Automated machine learning tools can help with the deployment process by generating code or APIs that embed the model. 84.51° often makes use, for example, of DataRobot's ability to output Java code for data preprocessing and model scoring. It lifts code out of DataRobot, deploys it into a new system, and then the production system is freestanding and can compute analytically-derived outcomes in a fast and ready fashion.

A European insurance firm's data science leader agreed that outputting Java code was a very useful feature in deployment: "In several client-facing projects we have used the deployment capabilities of DataRobot. You can deploy as an API or export code in Java. We like the Java option better because we are a Java shop and if we have a piece of code we can post somewhere it makes integration with our legacy systems much easier, and there are fewer security challenges compared to external APIs."

Sompo Holdings expects to use AutoML capabilities not only for model development and deployment, but to support the entire "end to end ecosystem" for machine learning. That includes the activities of data preparation, model development, production deployment, monitoring and evaluation of model performance, and updating the model.

Model and Feature Explanation

A final benefit for many organizations using AutoML is model and feature explanation. This feature (originally called “Reason Codes” in DataRobot, and now called “Prediction Explanations”) is important because machine learning models can be difficult to interpret. They may involve many different features or variables, and their relative importance in prediction or classification may be difficult to interpret. In some industries such as financial services, model transparency and explainability (for example, for why a customer is extended or denied credit) are required by regulators. In the European Union, under the General Data Protection Regulations (GDPR), any citizen affected by an analytical model is guaranteed the “right to an explanation.”

The large U.S. P&C insurance firm’s machine learning group uses this capability extensively. The head of data science support notes: “Reason Codes are my favorite feature. To see which features are contributing to the model at what level is extremely valuable. It’s very helpful to explain why a particular customer, for example, is popping up as a likely sale for commercial insurance.”

At the Canada-based bank RBC, model explanations are very helpful to both internal audiences and regulators. Samer Nusier, who heads Portfolio Management and Credit Strategy for the bank, says that the capability to explain models is “incredible for so many reasons. We sometimes have massive gradient-boosted tree models. Variables can be very predictive for small groups. Reason Codes let us understand why we make credit decisions, which helps with our regulators. It also helps to dictate actions from some of our marketing models. If we find out that customers who respond to our offers shop at another retailer, we might influence the offer by including a gift card from that retailer. Models can be impenetrable, but the ability to explain them helps a lot.”

*Models can be impenetrable,
but the ability to explain them
helps a lot.*

At SMCC in Japan, Hiroki Shiraishi, head of a machine learning infrastructure group, notes that the use of machine learning in financial institutions is not very far along. At the moment, there are no explicit regulations about its use. In the future, he expects that if regulators want to know how particular analytical decisions are made, model explanation capabilities could be very helpful in discussing models with regulators. Today, they are important within the credit card company to understand why particular types of customers are identified as targets for marketing.

Overcoming Obstacles to Benefit Realization

These benefits do not, of course, come automatically to organizations. There are a variety of obstacles to their realization, and organizations need to address them if they're going to be successful with AutoML. None of the obstacles seem insurmountable, and several of the companies interviewed for this study have already overcome them.

Some barriers are related to general issues in machine learning that AutoML simply can't address. For example, machine learning (at least the supervised variety, which is most common in business) requires labeled data in substantial volumes to train the model, and AutoML doesn't really address that issue. It can improve the efficiency of an analysis, however, by minimizing the amount of data used for training and validation.

There are two organizational or cultural barriers to the widespread adoption of AutoML in organizations. One is simply lack of awareness about AutoML and machine learning in general. Business managers in particular may not know what kinds of problems to which machine learning is suited and how AutoML can make the process substantially easier. Educational interventions (as used, for example, at the large U.S. bank described above) of various types can provide insights about the approach and stimulate demand for machine learning throughout the organization.

At SMCC in Japan, part of Hiroki Shiraishi's job is to make people aware of what machine learning can do for the credit card firm. To that end he offers training, seminars, external speakers, and individual sessions with senior executives to explain the concepts. He notes that although AutoML automates a lot of machine learning processes, it still requires basic understanding of what machine learning does.

The other organizational/cultural issue is resistance to AutoML from data scientists. Some perceived it as threatening to their profession; others said they were not convinced that AutoML-based modeling could be as successful as their own efforts. This issue was mentioned at about half the companies interviewed; some encountered no objections from even highly-trained data scientists. Among those who did encounter some resistance, the most effective strategy seemed to be to create some informal competitions between AutoML and data scientist-driven analyses. Some found that when the AutoML models turned out as well or better and were generated in substantially less time, most data scientists were more accepting of the technology. Companies might also try to encourage a general posture of openness to new technologies by data scientists, and to reward those who embrace them.

To get the greatest amount of value from AutoML and tools like DataRobot, it's helpful for organizations to adopt a disciplined, end-to-end process approach to machine learning — similar to that used by 84.51° or by Sompo Holdings in Japan. This allows for more exploitation of the less common benefits of AutoML, such as in the deployment and model explanation subprocesses of machine learning. It also allows for better measurement of benefits and improvements from AutoML.

It seems inevitable that machine learning (and other analytical approaches as well) will be increasingly automated over time. The organizations that adopt this technology are likely to be much more successful and productive with this powerful tool for prediction and classification. They will also be able to accomplish positive results without needing (as many) data scientists to do the necessary analysis. As in other areas of business and life, approaches and processes that employ greater levels of intelligence and automation will eventually prevail.

Contact Us

DataRobot
One International Place, 5th Floor
Boston, MA 02110
www.datarobot.com
info@datarobot.com