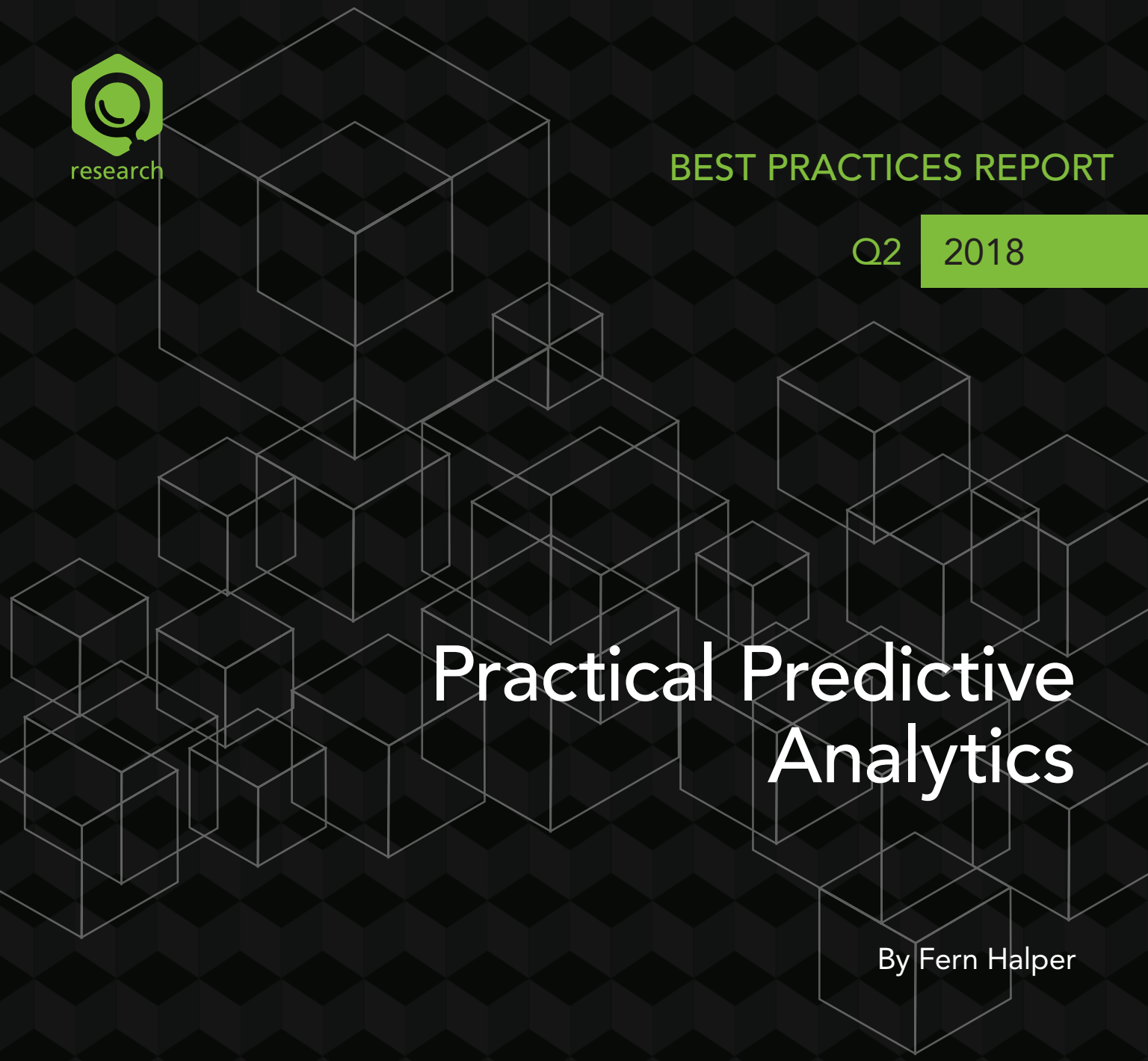


research

BEST PRACTICES REPORT

Q2 2018

Practical Predictive Analytics

By Fern Halper

tdwi

Transforming Data With Intelligence™

Research Sponsors

Hortonworks

SAP

Tellius

Practical Predictive Analytics

By Fern Halper

Table of Contents

About the Author

FERN HALPER, Ph.D., is vice president and senior director of TDWI Research for advanced analytics, focusing on predictive analytics, social media analysis, text analytics, cloud computing, and other "big data" analytics approaches. She has more than 20 years of experience in data and business analysis and has published numerous articles on data mining and information technology. Halper is a coauthor of *Big Data for Dummies* as well as other "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, and service management. She has been a partner at industry analyst firm Hurwitz & Associates and a lead analyst for Bell Labs. Her Ph.D. is from Texas A&M University. You can reach her at fhalper@tdwi.org, @fhalper on Twitter, and on LinkedIn at linkedin.com/in/fbhalper.

About TDWI

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of data management and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

About the TDWI Best Practices Reports Series

This series is designed to educate technical and business professionals about new business intelligence technologies, concepts, or approaches that address a significant problem or issue. Research for the reports is conducted via interviews with industry experts and leading-edge user companies, and is supplemented by surveys of business intelligence professionals.

To support the program, TDWI seeks vendors that collectively wish to evangelize a new approach to solving business intelligence problems or an emerging technology discipline. By banding together, sponsors can validate a new market niche and educate organizations about alternative solutions to critical business intelligence issues. To suggest a topic that meets these requirements, please contact TDWI Senior Research Directors Philip Russom (prussom@tdwi.org), David Stodder (dstodder@tdwi.org), and Fern Halper (fhalper@tdwi.org).

Acknowledgments

TDWI would like to thank many people who contributed to this report. First, we appreciate the many users who responded to our survey, especially those who agreed to our requests for phone interviews. Second, our report sponsors, who diligently reviewed outlines, survey questions, and report drafts. Finally, we would like to recognize TDWI's production team: James Powell, Peter Considine, Lindsay Stares, and Michael Boyda.

Sponsors

Hortonworks, SAP, and Tellius sponsored the research and writing of this report.

Research Methodology and Demographics

Report purpose. While excitement builds around predictive analytics, the reality is that many organizations are struggling to implement it in a meaningful way. The purpose of this report is to explore concrete best practices for building and deploying predictive models in organizations.

Terminology. In this report, predictive analytics refers to the tools and techniques from statistics, computer science, and other quantitative disciplines used to determine the probability of future outcomes using past information. In this context, predictive analytics also includes machine learning.

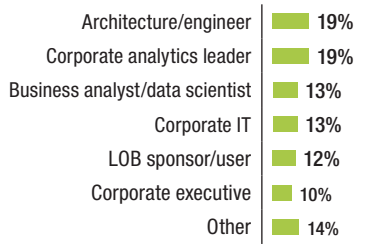
Survey methodology. In January 2018, TDWI sent an invitation via email to the analytics and data professionals in our database, asking them to complete an online survey. The invitation was also posted online and in publications from TDWI and other firms. The survey collected responses from 544 respondents. Half (50%) of respondents already had predictive analytics in use, 40% were planning to implement it, and 10% had no plans. Only 410 respondents completed all questions in the survey, explaining why the number of respondents varies per question.

Research methods. In addition to the survey, TDWI conducted telephone interviews with technical users, business sponsors, and analytics experts. TDWI also received briefings from vendors that offer products and services related to these technologies.

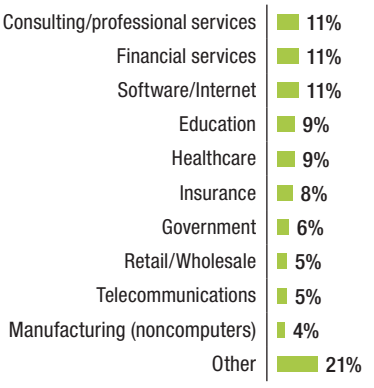
Survey demographics. Respondents act in a variety of roles. The majority of survey respondents are directly involved in analytics (32%), followed by those in architecture and engineering (19%) and corporate IT (13%).

The consulting (11%), financial services (11%), and software/Internet (11%) industries dominate the respondent population, followed by healthcare (9%), education (9%), and insurance (8%). Most survey respondents reside in the U.S. (60%), Europe (11%), or Canada (11%). Respondents come from enterprises of all sizes.

Position

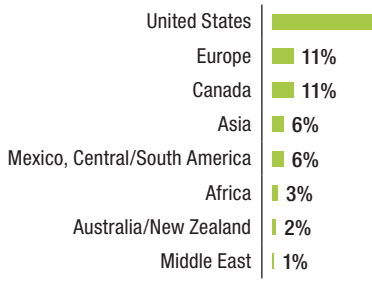


Industry

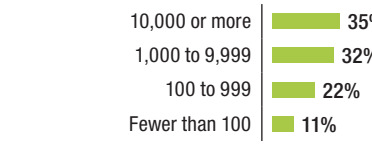


("Other" consists of multiple industries, each represented by less than 4% of respondents.)

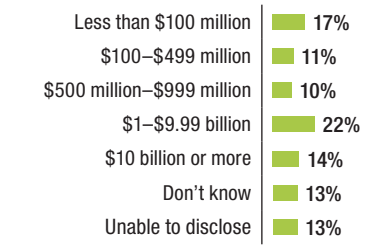
Geography



Number of Employees



Company Size by Revenue



Based on 410 respondents who completed every question in the survey.

Executive Summary

Predictive analytics is now part of the analytics fabric of organizations. TDWI research indicates that it is in the early mainstream stage of adoption. Yet, even as organizations continue to adopt predictive analytics, many are struggling to make it stick. Challenges include skills, executive and organizational support, and data infrastructure issues. Many organizations have not considered how to *practically* put predictive analytics to work, given the organizational, technology, process, and deployment issues they face. Addressing these issues involves a combination of traditional and new technologies and practices. Some practical considerations highlighted in this report include:

Skills development. Organizations are concerned about skills for predictive modeling. These skills include understanding how to train a model, interpret output, and determine what algorithm to use in what circumstance. In fact, skills ranked as *the* biggest barrier to adoption of predictive analytics, with 22% of respondents citing this as *the* top challenge. To address the challenge, organizations are looking to improve the skills of current employees as well as hire externally. They are also looking to use some of the new breed of automated, easy-to-use predictive analytics tools that contain embedded intelligence. Although only 16% use these tools today for predictive analytics and machine learning, an additional 40% are planning to use them in the next few years. Analytics platforms that provide interfaces for multiple personas are also at the top of the list of technologies organizations think can help.

Model deployment. In this study, respondents are using predictive analytics and machine learning across a range of use cases. Those exploring the technology are also planning for a diverse set of use cases. Yet, many respondents are not considering what it takes to build a valid predictive model and put it into production. Close to 80% have some controls in place (such as any model built is reviewed by an expert), but only about 50% have a DevOps team or another group that puts machine learning models into production, maintains versioning, or monitors the models. Fewer than 25% register their models. Respondents report it can take months to put models into production.

Infrastructure. On the infrastructure side, the vast majority of respondents use the data warehouse, along with a variety of other technologies such as Hadoop, data lakes, or the cloud, for building predictive models. The good news is that these respondents appear to be looking to expand their data platforms to support predictive analytics and machine learning. The move to a modern data architecture to support disparate kinds of data makes sense and is needed to succeed in predictive analytics.

This TDWI Best Practices Report examines how organizations using predictive analytics are making it work. It looks at how those exploring the technology are planning to implement it. Finally, it offers recommendations and best practices for successfully implementing predictive analytics and machine learning in organizations.

The State of Predictive Analytics

The Predictive Analytics Conundrum

Predictive analytics is on the cusp of widespread adoption. Many organizations are excited to make use of the power of predictive analytics (including machine learning) because they understand the value it can provide. However, although numerous organizations do use predictive analytics today, adoption remains elusive to many. In fact, TDWI research indicates that if users stuck to their plans for predictive analytics adoption, 75–80% of organizations would use the technology already although only 35–40% do so currently.

TDWI research indicates that if users stuck to their plans around predictive analytics, adoption would be at 75–80% versus the 35–40% we currently see.

There are numerous factors contributing to this conundrum. Some of the key ones include skill development, issues with data, and the time needed to put models into production.

The top challenge cited by respondents in previous TDWI surveys regarding adoption is most often skills. Some organizations start their predictive analytics journey with one or two people in their organization who can build models. With these skills in demand in the industry, there is often turnover, leaving the organization back at square one. Other organizations have a hard time hiring the right skills in the first place. They may look to improve or expand the skills of business analysts performing less sophisticated analysis, but this can be a slow process if there is no money budgeted for training.

Data infrastructure can also stymie predictive analytics. Many organizations start by building predictive models from data in their data warehouse or data marts. However, the iterative nature of predictive model-building (e.g., refinement and tuning) and the fact that organizations need to manage an ever-increasing amount of data may mean that the data warehouse is not a long-term practical solution. Organizations want to build predictive models using disparate data types. Some of these are "newer" kinds of data including unstructured data or real-time data from sensors, which means assembling and integrating multiple data pipelines.

Other organizations build models using statistical or machine learning approaches but struggle to put them into production due to a myriad of reasons (including rewriting models and politics), making it difficult to act on the insights and realize the benefits of the approach. Funding dries up, as does executive support. On the other hand, organizations are often trying to manage unrealistic expectations by executives. Sometimes issues arise because of a mandate to perform predictive analytics without addressing which business problems to solve first.

These issues swirling around predictive analytics can and should be overcome. Predictive analytics drives significant value. TDWI research consistently finds that organizations using advanced techniques such as predictive analytics are more likely to measure *actual* top- or bottom-line impact for their analytics program than those that do not. As we will see later in this report, we found the same to be true in this survey. Additionally, even those who do not have the budget for predictive analytics *still* see it as a key business opportunity.

Given predictive analytics' value, enterprises must identify ways to address it practically. That may include utilizing some of the newer data and analytics technologies on the market today, developing model-building and deployment processes, and adopting organizational best practices for moving predictive analytics forward and democratizing it in organizations.

The Current State of Predictive Analytics and Machine Learning

To understand the current state of predictive analytics, we asked respondents if they are using predictive analytics or planning to do so. For those who are not using the technologies and have no plans to do so we asked, "Why not?"

Half (50%) of respondents to this survey already use predictive analytics. Another 40% are exploring the technology now.

In this report, the group currently using the technology is referred to as the *active group*. Those who are planning to use the technology are referred to as the *exploring group*. In this survey, half (50%) already use the technology; more than half of this active group has been using it for more than three years. Forty percent are exploring the technology now. Ten percent are not using the technology. This should not be viewed as an adoption rate because respondents tend to gravitate to surveys they can relate to.

The majority of active group enterprises have models in production. Models provide the most value when used in production to make decisions and take action. The majority of active group respondents (73%, not shown) already have models in production. Another 17% plan to have models in production in the next few months. About 10% either don't have models in production or don't know when they might. The fact that the vast majority have put models into production is good news and illustrates that this group of respondents realizes the value of taking action on analytics. In fact, operationalizing models (to make them part of a business process) is one of the top areas of interest for moving programs forward. It is also important for making predictive analytics more pervasive.

Predictive analytics is used across a range of use cases. We asked both the active group and the exploring group what kinds of use cases they are using or plan to use in predictive analytics. As illustrated in Figure 1, there are many popular use cases for predictive analytics for both groups.

- **Marketing applications often lead the way.** Over half (52%) of the active group is using predictive analytics for retention analysis or direct marketing. Cross-sell and up-sell is also popular in marketing. These are popular use cases for the exploring group as well with a third of respondents planning to deploy these use cases in the short term. Previous TDWI

What Are Predictive Analytics and Machine Learning?

Predictive analytics consists of statistical and machine learning algorithms used to determine the probability of future outcomes using historical data. Some people think of machine learning as being completely different from predictive analytics. Machine learning techniques, however, are often used in predictive analytics; they just use a different approach.

Machine learning methods originated in the field of computational science a few decades ago. In machine learning, systems learn from data to identify patterns with minimal human intervention. The computer learns from examples, typically using either supervised or unsupervised approaches. In supervised learning, the system is given a target (also known as an output or label) of interest. The system is trained on these outcomes using various attributes (also called features). In unsupervised learning, there are no outcomes specified.

Popular use cases include churn and fraud analysis, which are also popular for statistical approaches. Popular machine learning techniques include decision trees, neural networks, Naïve Bayes classifiers, and clustering. Many taking our TDWI surveys also cite regression as an important machine learning technique.

In this report, both terms are used.

research indicates that marketing is often one of the first departments to adopt more advanced analytics.[1]

- **Default prediction is also popular.** In addition to marketing use cases, respondents are also utilizing or interested in other kinds of use cases. For example, default prediction ranked high for both groups with 46% of the active group already doing this kind of analysis and 34% of the exploring group planning to do so in the short term. Default analysis is important for a number of use cases including loans, credit cards, premium payments, and tuition payments.
- **Newer use cases such as predictive maintenance are gaining steam.** Thirty-four percent of the active group is already using predictive maintenance and 22% of the exploratory group plans to use it. In predictive maintenance, organizations calculate the probability of an operational asset requiring servicing or even failing. Some organizations make use of sensor data from the Internet of Things (IoT). For instance, a fleet operator might use sensors to collect data from their various trucks. Such data might include the temperature or number of vibrations per second of a particular part or parts. This data can be analyzed using machine learning to determine what precipitates a part failure or when undue wear and tear is occurring. The system "learns" the patterns that constitute the need for repair. That information might be encoded into a set of rules or a model and used to score new data from trucks in order to improve fleet maintenance and operational efficiency. Other use cases include cybersecurity, where 25% of the active group is using predictive analytics (not shown).

Over 30% of the active group claims to be using predictive maintenance.

- **In applications.** An up-and-coming area of interest is embedding predictive analytics and machine learning models in applications that require intelligence. We did not ask specifically about these applications but they are worth mentioning here. These include voice recognition, traffic apps, and chatbots. Although only 12% (not shown) cite building apps as a driver for predictive analytics and machine learning, we expect that percentage to grow.

What is predictive analytics used for/planning to be used for in your organization?

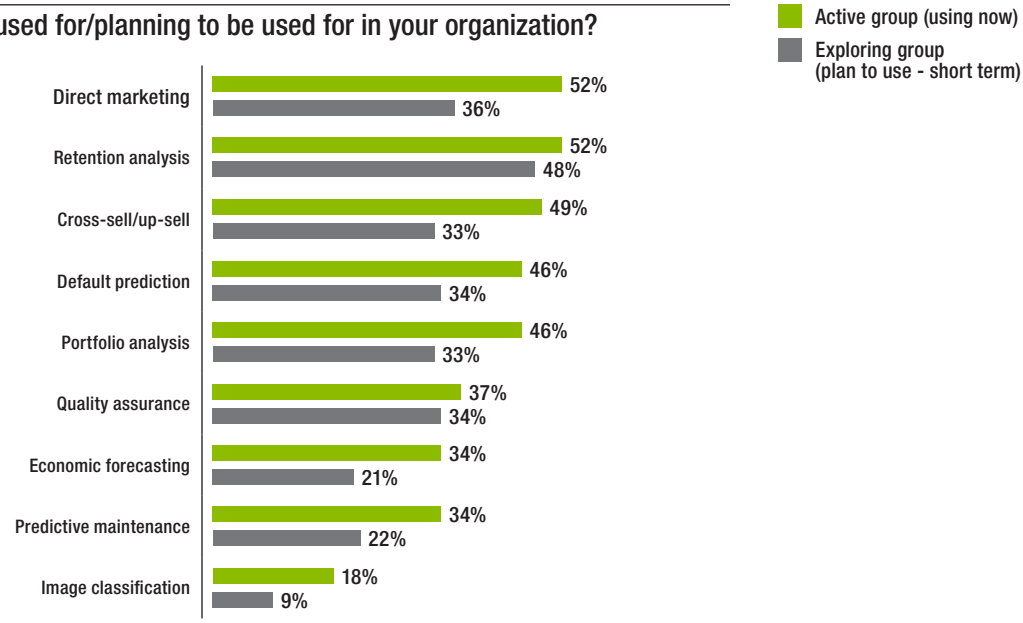


Figure 1. *Based on 244 active group and 180 exploring group respondents. Multiple responses permitted.*

[1] For instance, in a 2014 BPR about predictive analytics, marketing was also a top use case. Percentages have not changed significantly since that 2014 report. See www.tdwi.org/bpreports to access this report.

Enterprises are using disparate data types in predictive analytics, including text data and geospatial data.

Disparate data types for predictive analytics. Organizations often use a range of data types for predictive analytics, a trend TDWI has seen in previous research on advanced analytics. In this survey, almost everyone (96%) in the active group uses structured data, and 71% use demographic data to enrich their customer-related analysis (Figure 2). However, the active group is also making use of other kinds of data such as geospatial data (55%) and internal text data (52%) in their predictive analytics efforts. That makes sense because this kind of data can help to improve a model. For instance, if an insurance company is trying to predict the risk associated with certain policies, it makes sense to understand where the customers are located geographically and if they are near features such as flood zones that increase risk. If a retailer is trying to predict customers' buying patterns, it is helpful to know if a certain product triggered many customer complaints.

The active group is also making use of real-time event data. Nearly half (48%) are using it now, with another 41% planning to use it in the next few years. Some of this may be IoT sensor data (21% use it today) for use cases such as predictive maintenance, as described above. However, there are other use cases for real-time data including those in risk analysis (e.g., fraud detection) and marketing (e.g., recommendation engines).

Those exploring the technology plan to use structured data (83%) and demographic data (50%) when they start building predictive models. They are looking at geospatial and internal text data as well (both at 32%). If users stick to their plans, the vast majority will be using these kinds of data as well as real-time data, external text data, and log data several years after they start deploying predictive analytics (not shown). About half would be using IoT data (also not shown).

The trend is clear: organizations want and need to make use of disparate data types for predictive analytics because it enriches the data set and can add value to the analysis.

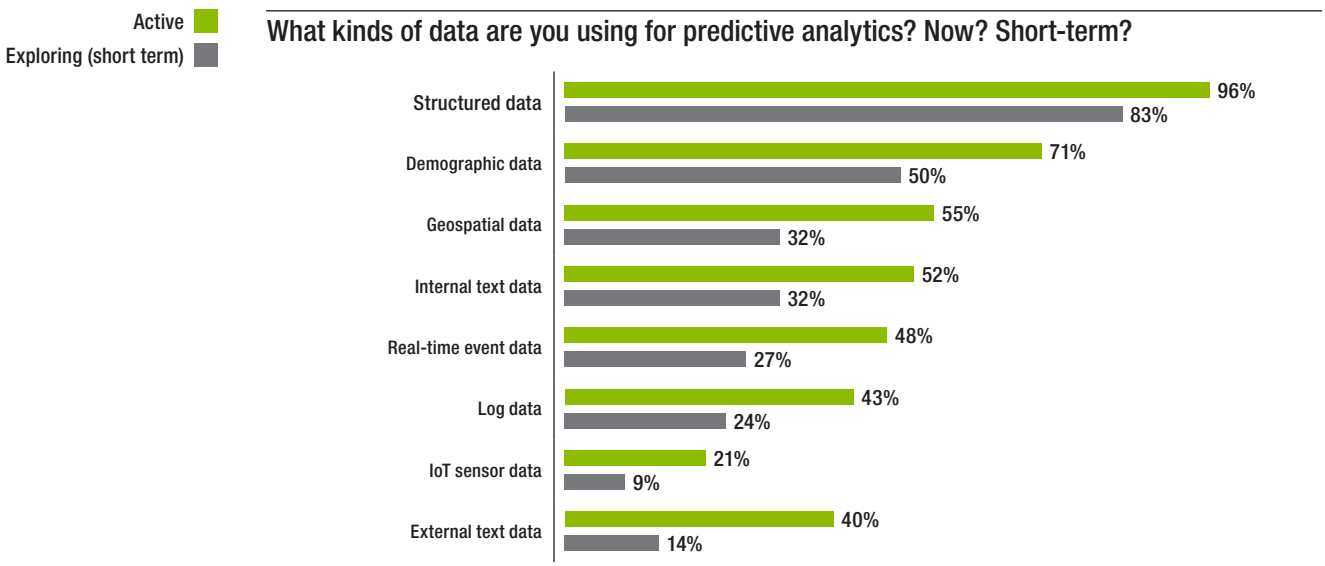


Figure 2. *Based on 244 active and 180 exploring respondents. The term "now" was used in the question for the active group; the term "short term" was used with the exploring group.*

Why not use predictive analytics? A very small group of respondents (10%) is not making use of the technologies. As noted, this is not a reflection of the market adoption but more likely due to the nature of the survey itself.

We wanted to know why those not deploying or planning to deploy predictive analytics are staying on the sidelines. The top reason given: they are still focused on BI activities (e.g., reports and dashboards) according to over half of respondents in this group. Another top answer was that the organization didn't have the skills for predictive analytics and machine learning (33%, not shown).

Although these respondents weren't performing predictive analytics, they do find it valuable. In fact, the vast majority (75%, Figure 3) said they believe that predictive analytics is valuable and they wish they could use it in their organization. Fewer than 2% said it wasn't valuable for their organization.

Three-quarters (75%) of those organizations with no plans to use predictive analytics still believe it is valuable.

Do you believe that predictive analytics is valuable?

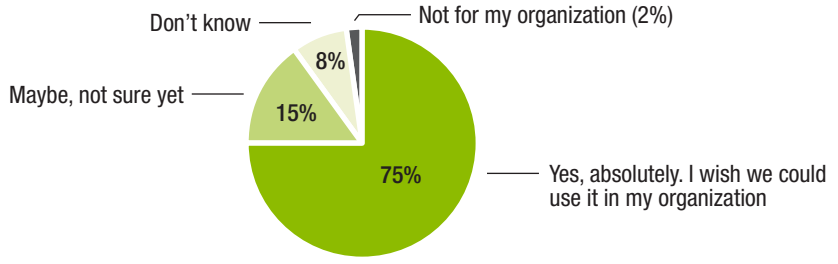


Figure 3. *Based on 52 respondents not using predictive analytics.*

New Practices for Predictive Analytics and Machine Learning

Several technology trends and practices affect predictive/machine learning model building and are important for making predictive analytics more pervasive. Vendors are offering a range of options to address skills issues, helping "democratize" analytics (e.g., making predictive modeling more accessible to non-statisticians and data scientists). These include technologies supporting ease of use, automation, and open source.

Vendors offer a range of options to address skills issues, helping "democratize" analytics. These include ease-of-use and automation features.

New Technologies Promise Ease of Use

Because predictive analytics and machine learning skills are in such high demand, vendors are providing tooling to help make predictive modeling easier, especially for new users. Important to ease of use are these features:

- **GUIs.** Many users do not like to program or even write scripts; this spurred the movement toward GUIs (graphical user interfaces) decades ago in analytics products. Today's GUIs often provide a drag-and-drop and point-and-click interface that makes it easy to build analytics workflows. Nodes can be selected, specified, dragged onto a canvas, and connected to form a predictive analytics workflow. Some vendor GUIs allow users to plug in open source code as a node to the workflow. This supports models built in R or Python, for instance.

- **Workflows and versioning.** Many products provide workflows that can be saved and reused, including data pipeline workflows for preparing the data as well as analytics workflows. If a data scientist or another model builder develops a model, others can reuse the model. This often includes a point-and-click interface for model versioning—critical for keeping track of the latest models and model history—and for analytics governance.
- **Collaboration features.** Anyone—from a business analyst to a data scientist—building a model often wants to collaborate with others. A business analyst may want to get input from a data scientist to validate a model or help build a more sophisticated one. Vendors provide collaboration features in their software that enable users to share or comment on models. We will see later in this report that collaboration among analysts is an important best practice to help democratize predictive analytics.
- **Persona-driven features.** Different users want different interfaces. A data scientist might want a notebook-based interface, such as Juypter notebooks (e.g., "live" Web coding and collaboration interfaces) or simply a programming interface. A business analyst might prefer a GUI interface. A business analyst might want a natural language-based interface to easily ask questions and find insights (even predictive ones) in the data. New analytics platforms have tailored environments to meet the needs of many personas while maintaining strong data integrity beneath the platform. This makes building the models more efficient. As we will see later in this report, analytics platforms are important to respondents.

Automation

Automation is another way of making predictive analytics and machine learning easier to use and opening up model building and deployment to more people across the organization. In automation, software is infused with "smarts"—rules or advanced analytics such as machine learning that perform tasks for the analyst.

In this survey, 16% of respondents are using some of these automated tools and 40% percent expect to use them in the next few years (see Figure 4).

Is your organization using automated predictive analytics tools?

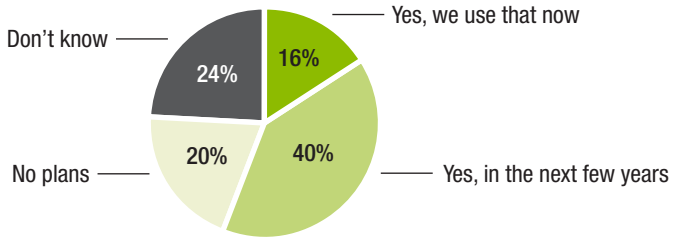


Figure 4. *Based on 387 respondents from both the active and exploring group.*

Automation is occurring across the analytics life cycle. Some examples include:

- **Data preparation.** Data preparation has a number of elements and objectives. The process typically includes data sourcing and ingestion, moves through making data suitable for use through transformation and enrichment, and then integrates with governance and stewardship. The steps are interdependent and overlap. Machine learning algorithms are being embedded into software for data preparation. For example, machine learning is being used to assess data quality and accuracy (e.g., address mapping). Automation can transform data for analysis, such as automatically creating common ratios from attributes in the data file for use in analysis. Such automation can help reduce the time needed to prepare data for analysis—which can be 60-80% of an overall analysis.[2]

[2] See TDWI's Best Practices Report on data preparation for more information. tdwi.org/bpreports

- **Model building.** As noted, because data scientists and statisticians are often in short supply, many vendors offer tools that help business analysts and even business users automate the discovery of insights and construct predictive models. In some tools, all users need to do is supply the target or outcome variables of interest along with the attributes they believe are predictive. The software picks the best model. Many of these tools provide details about the statistics and math used; some do not. In this survey, the 16% of respondents (Figure 4) using automated model-building tools jumps to over 20% (not shown) for the active group.

  Respondents cited a number of benefits for these tools (Figure 5). For instance, 31% stated that it enables more people to build models and that means more insight and action. Twenty-four percent said it takes less time to create a model, improving time to value. Fourteen percent noted that if business analysts can build models, data scientists have the time to do more sophisticated work.

- **Monitoring models.** As important as building a model is, monitoring it once it goes into production is critical to predictive analytics success. Models get stale and degrade over time, so it is important to keep track of how well they are performing. Some vendors are providing ways to automate the monitoring process. For instance, some tools schedule updates for model building and introduce a champion/challenger approach to make sure the model is still current. Other tools provide automated alerts if a model is degrading. Model monitoring is another best practice that we will discuss in more detail later in this report.

In some tools, all users need to do is supply the target or outcome variables of interest, along with the attributes they believe are predictive.

What do you believe/have you seen to be the biggest benefit of these kinds of tools?

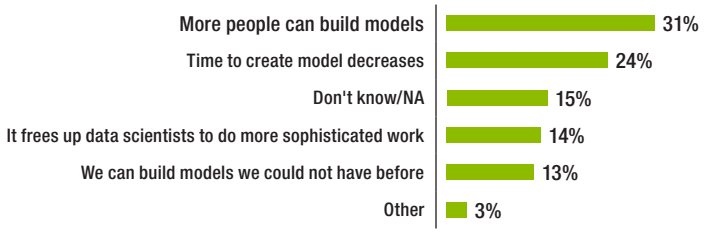


Figure 5. *Based on 385 respondents in both the active and exploring group*

Open Source Is Used Frequently

Open source provides a community of innovation that many data scientists like. The open source model is a collaborative development model where code is made available for free and the copyright holder has the rights to study, change, or distribute the code. Open source has become popular, especially for big data and data science, because it is a low-cost source community for innovation, which appeals to many data scientists and analytics application developers.

In this survey, open source was popular among all respondents (Figure 6). The vast majority of those currently building models as well as those exploring predictive analytics use or plan to use open source. R leads the way, followed by Python. Survey results suggest that organizations typically use open source in conjunction with commercial products. Open source is often used to experiment with and build models, especially by data scientists who have experience with and are comfortable with it. Use of open source in production depends on the skill level of the organization and how the DevOps team wants to standardize models. Some organizations rewrite models; others put them into commercial products (as discussed above). Some organizations do put open source models into production.

The vast majority of those currently building models *and* those exploring predictive analytics use or plan to use open source.

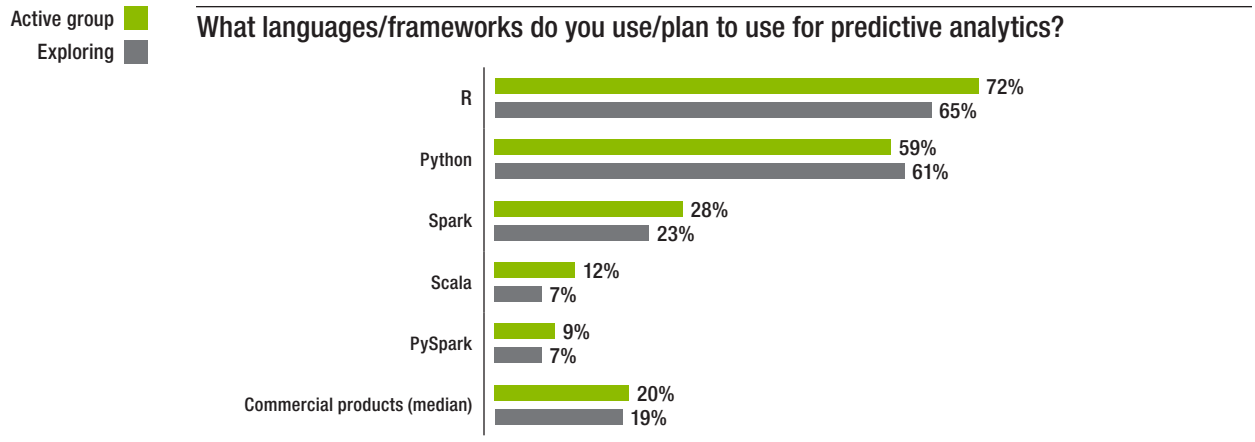


Figure 6. *Based on 244 active group respondents and 180 exploring group respondents. Multiple responses allowed.*

Open source can provide the flexibility many data scientists like to help them "fail fast" when building models. Open source is often also used to build apps that utilize analytics.

There is a trade-off with open source, however. The organization needs to have the skills to use open source or else pay for support. Some organizations have the skill set to utilize open source—many students are trained in R and Python. Others use commercial open source platforms that provide the flexibility of open source along with support. Some enterprises use a combination of open source and commercial products (including easy-to-use products).

Best Practices for Building and Deploying Models

A number of best practices for building and deploying predictive and machine learning models are important for getting models out to an organization. These practices span the entire analytics life cycle.

Make Sure You Understand the Business Problem

This sounds obvious, but it is often one of the biggest obstacles organizations face as they start to adopt predictive analytics and machine learning. Understand the use case and then hone it. Ask questions: What is the problem statement? What is the outcome? What is the impact?

Measuring impact will help an organization launch its first predictive model. Once the enterprise addresses a few use cases, it can extend them, which is how predictive modeling grows.

USER STORY THE IMPORTANCE OF SELECTING THE RIGHT BUSINESS PROBLEM

"A signature project that addresses real business problems can help drive predictive analytics," said a data warehouse executive for a large Midwest insurance company. In this case, it was predicting (and then contacting) households that had a high probability of payment lapse. Because insurance companies don't make money until customers have paid a few years of premiums, this is a big area of loss—often costing insurance companies millions of dollars a year. This seemed like a natural first project for predictive analytics and one that could sell executives on the technology and the purchase of a commercial product.

That was over 10 years ago. With that success, the company was able to work on many other predictive analytics initiatives with large ROIs. For instance, agent turnover is a big problem in insurance. The team built risk models to predict which insurance representatives were likely to leave the company. They created analytics reports using the models and distributed them to the responsible divisional VPs to try to retain more personnel. "It was uncanny about how well models were doing," said the data warehouse executive. The company was also able to use predictive analytics to move from paper to Web-based insurance approval and cut down on the number of questions they asked prospective customers by understanding which questions were most important to predict risk. Other projects followed.

Of course, a good data infrastructure was also necessary to make these models work. In this case, the data warehouse includes data about customers, demographics about households, and customer value information. "Once we understood customers and their value, we could move to predictive analytics," said the executive.

He recommends a process for being successful with predictive analytics. "First, come up with a list of business problems not being solved. Analytics in any form is not a crystal ball. You need to say, 'Look, we have business problems that need predictive analytics.' Second, try to identify diverse areas in the company that could use predictive analytics. This will build momentum. Third, try to get people engaged in the process and get people using the results. Finally, let the results be known. Don't keep them to yourself."

Data Quality Is a Must

Previous TDWI research has pointed to data quality being a big issue for analytics, and the same holds true for predictive analytics and machine learning. We asked respondents what technical challenges have been or do they think will be the most difficult to overcome. The top answer was data quality (33%) followed by data integration (23%, both not shown).

The old saying "Garbage in, garbage out" applies to predictive modeling as well as other kinds of analytics. In fact, aside from lost revenue due to incorrect models based on bad data, poor data quality can reduce trust in the results of the analysis and even damage a company's reputation.

As we've pointed out, machine learning is being used in some newer products to help find issues in the data. This might include deduplicating data or looking for gaps, outliers, and anomalies in the data. For instance, a machine learning algorithm can be trained to recognize when two names that appear to be different are actually the same. These same algorithms can determine if observed data values differ from predicted values.

These kinds of tools can help organizations meet data quality goals of reasonableness, consistency, timeliness, and relevancy.

Put Controls in Place

As more people in the organization build models, it is important to put controls or a quality assurance (QA) process in place to make sure that a sound model is built—either for active use to drive insights or to put into production. This is true even for the easy-to-use tools that can democratize predictive analytics, as mentioned earlier.

Almost 4 in 5 (79%) respondents stated that models are reviewed by an inside expert once built.

We asked the active group what controls it puts in place when building a predictive model. The results (Figure 7) indicate that the active group does utilize some sort of mechanism to make sure models are properly built. Seventy-nine percent of the active group respondents stated that models are reviewed by an inside expert once built. Some organizations hire a few data scientists or statisticians who will help the business analysts. Ensuring that someone with expertise is reviewing the models is important. It is one thing to experiment with a model, but it is another to use it in production.

Other controls include not letting certain users build certain models at all. Almost half (48%) of the active group respondents stated that "certain models can only be built by experts." For example, if a pricing model might be material to a company, someone with expertise would build that model. A model for a regional marketing campaign might be considered "safe" for a business analyst to build as skills improve.

Which of the following controls does your organization use when building a predictive model?

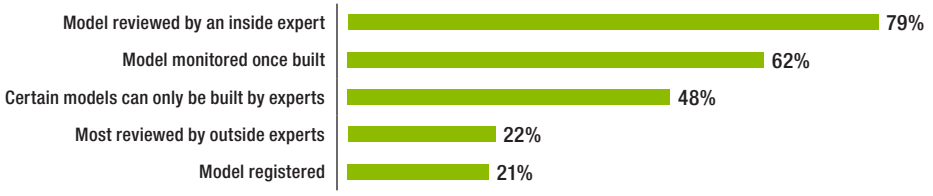


Figure 7. *Based on 219 active group respondents. Multiple responses allowed.*

Plan for Production

To gain value from predictive analytics and machine learning, plan to operationalize the models—i.e., make them part of a business process. Operationalizing models is the top strategy for those who have already built predictive models (see Figure 8). There are numerous best practices for putting models into production. These include.

Close to 50% of the active group have a DevOps team in place.

Get DevOps involved. If a model is going to be deployed in a system, such as in a call center system, you must understand that target system. Get the DevOps group involved; you need to work with the people responsible for managing and monitoring models in production. Close to half of the active group (46%) have a DevOps group in place. Another 28% plan to have one in place in the next 6 months (both not shown).

Model versioning. Putting models into production also requires keeping track of the models, so a model registry or versioning system is also important. It is easy to keep track of a few models, but once many models go into production, you need to register and track them, including metadata about the models (date created and by whom, data sources used, etc.). Only 21% of the active group has registered their models (Figure 7), suggesting that many users still do not have many models in production. This is an area companies need to watch as they build more models.

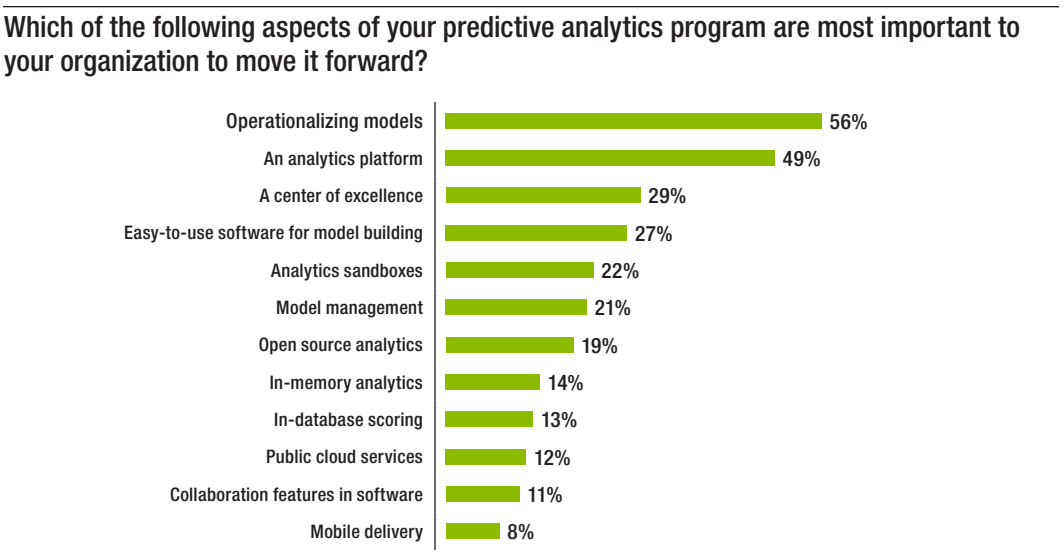


Figure 8. *Based on 219 respondents in the active group. Three choices allowed per respondent.*

Exporting models. Deploying models into production can take time. For the majority of respondents in this survey, it took at least a few months to put models into production. For 40%, it took more than six months (not shown). As we will see later, organizations are not necessarily satisfied with the ease of moving models into production.

There are numerous reasons why this task takes so long. Some causes are organizational; others are technical.

- **Validating models.** Those who took a few months or more to put models into production noted that it could take time to collect data and validate and tune a model. Some models are complex. If model building is new for an organization, it can take time (a few months or much longer) to build consensus, prove that the model works, then secure approval. For instance, those organizations with more than three years of experience with predictive analytics were more likely to be satisfied with their organizational support than those with less than one year (not shown).
- **Making sure someone is responsible.** Additionally, it can take time to get models into production because no one is assigned to be responsible for this task. That is why it is important to start planning for production early. Interestingly, there appeared to be no significant difference in time to put models into production for those using the easy-to-use tools versus those using methods that are more traditional. This may have been because the companies were using multiple methods for building models. However, although some newer tools can make the models easier to build, there is still the issue of actually deploying the model into production. Therefore, respondents using these tools still needed to get buy-in, organize a team, and resolve technical issues.

It can take time to put models into production when no one is responsible for it.

- **Export language.** On the technical front, sometimes an organization rewrites the model so it is in the same language as the target system. Thirty-five percent of the active group has done that (Figure 9), and this clearly adds time to any project, even if DevOps is involved. Some respondents use Predictive Modeling Mark-up Language (PMML), a development standard introduced in the 1990s. PMML is an XML-based interchange format that provides a way to represent models in order to share them with systems and applications. The idea behind PMML is to provide a vendor-neutral standard. Although PMML is arguably still the most popular standard in the market, some vendors support a newer standard called

Portable Format for Analysis (PFA), but as illustrated in Figure 9, it hasn't caught on yet. Other vendors are enabling deployment using a language that is readable. For instance, some vendors deploy models in C++ or C#. Some deliver output in Java.

Forty-five percent of the active group uses APIs to put models into production.

Many organizations use application programming interfaces (APIs), software intermediaries that allow two applications to talk to each other. For instance, some vendors support deploying models into applications as microservices via Restful APIs.

Of course, organizations are looking to deploy models everywhere—in applications, in devices, as well as in systems and processes. For example, in IoT scenarios, an organization might want to deploy a model to an edge device. That requires some thought up front because there will be security concerns.

The reality is that it may take time to initially put models into production. This is because organizations need to buy in and processes and teams need to be put in place. Although some of the tools on the market can help, it is also important to plan for production, including monitoring models once in production to ensure they stay current and reliable.

How do you export models you put into production?

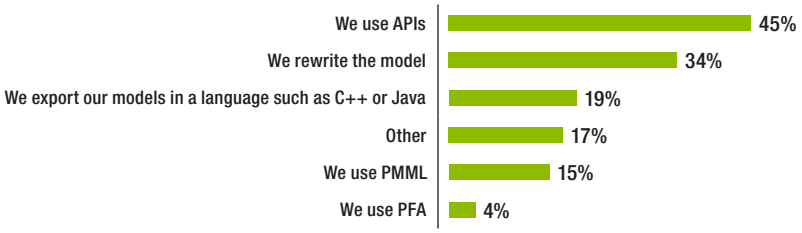


Figure 9. Based on 173 respondents in the active group. Multiple responses allowed.

New Trends and Best Practices in Data and Analytics Infrastructure

Several data infrastructure trends are worth noting because they can help your organization make more data available to a greater number of users. Trends include data lakes, public cloud, and analytics platforms. Data lakes and the public cloud are especially helpful when it comes to managing big data. These are all part of a modern data architecture that supports analytics.

Excitement About Data Lakes

Over half (51%) of the exploring group rank data lakes in the top three infrastructure technologies for predictive analytics and machine learning.

Organizations are adopting data lakes because lakes provision the kind of raw data that users need for exploration and discovery-oriented forms of advanced analytics. A data lake can also be a consolidation point for both new and traditional data, thereby enabling predictive analytics and machine learning. In this survey, the data lake ranked just behind the data warehouse in terms of a top infrastructure component for predictive analytics with 40% and 51% of the active group and exploring group ranking it in the top three, respectively (Figure 10), although it is not yet widely deployed (Figure 11).

There are many reasons why organizations look to data lakes for their predictive analytics and machine learning needs. These include:

- **Scalability.** Data lakes provide repositories for large amounts of disparate data, often from multiple data sources. Enterprises can use tools such as machine learning to explore data in a data lake because these tools are designed to uncover patterns over massive amounts of data even in the discovery/exploration phase before any real model building occurs. At TDWI, we often see data scientists and others are using data in a data lake to develop models.
- **Performance considerations.** Many organizations are building models against their data warehouse. However, the data warehouse was not built to support the iterative kind of analysis needed for modeling where models are refined and tuned. That is one reason why organizations look to a data lake to support analytics such as machine learning.
- **Data lakes ingest data quickly and store data in raw form.** Some analysts building models like to see the data in its raw form. Some data in the data lake is captured and stored with little or no improvement, so the ingestion process is very quick. Of course, other users like more refined data, which is why modern data lakes often have multiple zones. Data is often cleansed and structured so the lake doesn't become a swamp.

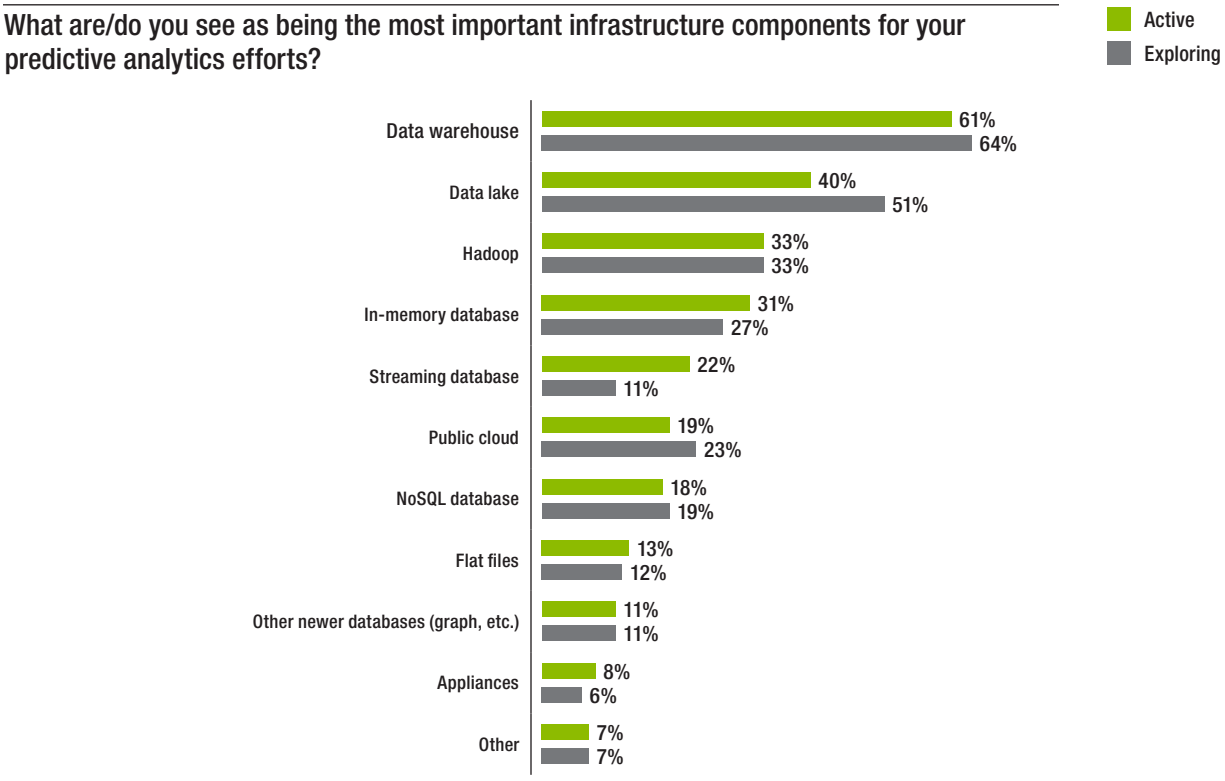


Figure 10. *Based on 206 active group respondents and 171 exploring group respondents. Up to three answers allowed.*

Public Cloud Grows in Importance for Data and Analytics

Organizations are putting plans in place for data warehouses, data lakes, and analytics platforms on the public cloud.

Public cloud infrastructure is increasingly used for predictive analytics and machine learning, and it is set to grow (see Figure 11). The public cloud provides elasticity, flexibility, and scalability for both data infrastructure and analytics. Data is often generated in the cloud and it makes sense to land and analyze it there. There are analytics services in the cloud that enable developers to buy marketplace analytics (sometimes called analytics services) such as machine learning models. These services can be incorporated into an application.

Although public cloud services ranked behind other aspects of a predictive analytics program in Figure 8, it appears that organizations are planning to use the public cloud for data warehouses, data lakes, and even analytics platforms. For example, although 28% of organizations currently use a data warehouse on a public cloud, another 37% expect to do so in the near future. Our survey found that 13% use Hadoop in the public cloud now, and this number will likely more than double if users stick to their plans. The same is true for the data lake, where 13% use it now in the public cloud and another 28% plan to use it soon. The public cloud appears to be a growth area for data infrastructure.

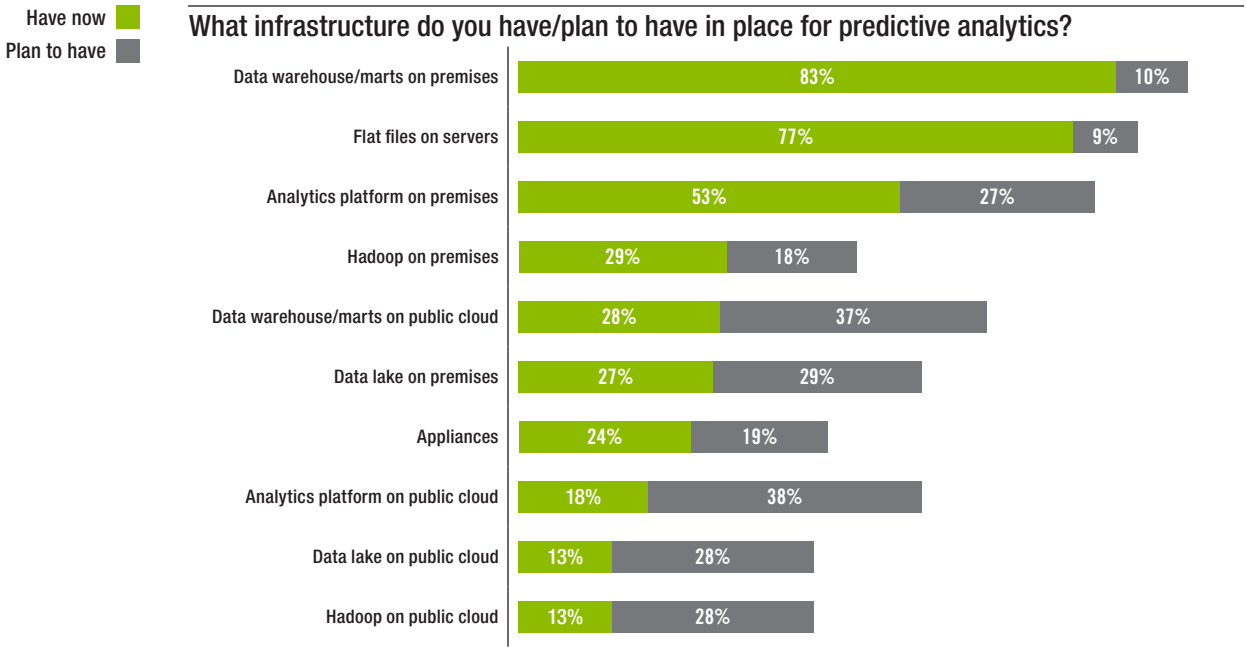


Figure 11. *Based on 377 respondents.*

This is also the case for analytics platforms on the public cloud. Although fewer than one in five (18%) use analytics platforms on the public cloud today, 38% more plan to utilize analytics platforms in the public cloud.

We will see in the satisfaction and value section of this report that infrastructure has a big role to play in affecting satisfaction and measuring value.

Analytics Platforms Are at the Top of the List

As seen in Figure 11, analytics platforms, both on premises and the public cloud, are used today and organizations are planning to use them in the future. There are several benefits to be enjoyed from an analytics platform, which is probably why it ranked second on the list of important aspects of a predictive analytics program in Figure 8.

Analytics platforms integrate and bring data together in a shared resource. That means that everyone is working off the same data when they analyze it. A platform also unifies the analytics toolset in one place. That is important for the various personas mentioned earlier in this report. For instance, a data scientist might want a notebook interface and make use of sophisticated models, but a business analyst might want a GUI interface to access prebuilt models. Data engineers want access to consistent data for feature development. The platform may also contain some of the model management capabilities described earlier to support DevOps. Vendors are building analytics platforms with multiple personas in mind on both open source and proprietary software, so there are many options.

Organizational Best Practices for Predictive Analytics and Machine Learning

Organizations face challenges when they embark on predictive analytics and machine learning. Previous TDWI research points to the fact that organizational issues can often prevent an analytics project from moving forward. At the top of the list for organizational challenges in this research is lack of skilled personnel. In this survey, 22% of the respondents said that was the biggest barrier to adoption, followed by responses such as "the business case isn't strong enough" (16%) and "not enough budget" (15%, all not shown). Organizational issues are often key stumbling blocks to getting predictive analytics and machine learning working.

Twenty-two percent of respondents said a lack of skilled personnel is the biggest barrier to predictive analytics adoption.

Talent-Building Strategies

A big part of the issue, of course, is that data scientists are an expensive resource that can be hard to find. When asked who in their organization is currently building models using predictive analytics/machine learning, 72% of the active group cited data scientists, 49% cited business analysts, and only 15% cited business units (all not shown). The numbers were not significantly different in the exploring group.

Many organizations are trying to grow talent from within, especially within the business analyst community. Others are hiring externally, and most use a combination of approaches. In an open-ended question, we asked, "How are you building talent for predictive analytics in your organization?" The responses fell into several categories:

- **Training.** This includes both internal and external training. For instance, some organizations send their business analysts to conferences and other external training programs. They also train from within in workshops and clinics. Some hold courses each week on tools. Some have informal mentoring programs. Some hire external consultants (who may be building models) to train internal staff.
- **Hiring.** Some organizations hire young graduates from universities—often at the master's degree or Ph.D. level. Some hire established data scientists and some hire a mix. For instance, one respondent said, "We have mix of alliances with universities as well as hiring some data scientists."

- **A combination of approaches.** Most organizations utilize a combination of approaches; they train from within and hire externally. One respondent explained, "We are hiring predictive analysts with [a master's degree] in data science/statistics in our advanced analytics center of excellence and are training savvy analysts in all areas of the business."

In the next section, we will see that lack of skills often affects satisfaction with and value of predictive analytics and machine learning.

USER STORY TALENT AND COLLABORATION ARE KEY FOR PREDICTIVE ANALYTICS

"One of the challenges we've had with predictive modeling is that people have left the organization and that makes model maintenance a problem," said a BI manager at a telco. "For instance, we had built a model using R that predicted call volumes for use with call center staffing. The person who built it was smart—a Ph.D. in math—but he left the company and there have been some struggles with the model since then as it is complex."

"Right now, we have some other models under construction," he added. This includes a model that is being used to help address where the company has network capacity to attract new customers. "There is sort of a gray area between a business analyst and a data scientist. The person building this new model is a quantitative business analyst who is using a GUI-based commercial software tool to build the model." The company is looking to attract more talent by providing a career progression in the group that is more than a move to a managerial role. "Managerial roles might not be attractive to very quantitative types," said the BI manager.

The company is also overcoming the challenges it has had with collaboration. "A few years ago, the IT group told us we could use business tools, but if we wanted IT's help then IT needed to do it." That made it hard for the business teams to move forward with higher-level analytics. Part of the problem was that IT didn't realize the extent of the analysis that was happening with low-level tools. "Our analyst groups are spread out over the company—we have a team in marketing, call center, etc., so representatives from all of the groups got together with IT to work with them." Since then, the company has been moving forward with a more collaborative model. "We are now looking for quick wins to show ROI and build from there."

Consider a Center of Excellence

Previous TDWI research indicates that a CoE can be an effective organizational strategy.

Previous TDWI research indicates that a center of excellence (CoE) can be an effective organizational strategy. The CoE ranked among the top best practices in this survey, as well (see Figure 8). The CoE typically consists of a group of people with expertise in a specific area—in this case, analytics including predictive analytics. With regard to a best practice for predictive analytics, one respondent stated, "We created an analytics center of excellence that supports enterprise analytics tools and provides predictive modeling expertise for business functions without this skill. We also created an analytics community to educate interested employees regarding tools, modeling techniques, resources, and networking opportunities." Members of the CoE often help to build models as well as educate the organization.

The CoE can be either centralized or distributed. There are pros and cons to each approach. For instance, some organizations like to distribute the talent into business units in an embedded or consultative approach. However, that can mean that the quantitative analyst or data scientist feels isolated and unable to collaborate with others. More often, organizations are following a hub and spoke model, where there is a CoE, and teams are sent from the CoE to work with the business. Sometimes there is a dotted-line reporting structure in the shared model. The jury is still out as to the best model for a CoE.

One point to note is that the CoE does not need to be called by this name. Interestingly, some organizations feel that using the term *center of excellence* makes the group seem unapproachable. They often give the CoE a name that resonates more with the organization.

Collaboration

Collaboration is critical when it comes to advanced analytics such as predictive analytics and machine learning. This collaboration happens on several levels. First, different groups, such as business and IT, need to collaborate. In this way, IT will understand the business's issues and the business will talk to IT about the data. Second, analyst groups need to collaborate. As one respondent said, "It helps to have data scientists mentor and work collaboratively with other analysts." This can happen in a number of ways. An analyst can give a model to a data scientist to make sure it is working well. The data scientists can take that model to the next level. Alternately, a data scientist can formulate use cases and write an algorithm. The data scientist can then hand the model to the business analyst to explore with a similar data set.

In this survey, 54% of respondents felt that collaboration among different groups in the organization was an important best practice to help democratize analytics. Thirty-two percent felt that collaboration among analysts was important (Figure 12).

Fifty-four percent felt that collaboration among different groups was important to help democratize analytics.

Of course, collaboration is not something that happens overnight. It requires building a collaborative culture based on trust.

Based on your previous experience in analytics, what are the top three best practices for helping to democratize analytics in an organization?

Figure 12. *Based on 414 respondents. Up to three responses per respondent allowed.*

Build Trust with Evidence

Building trust happens on a number of levels. There is building trust with data as well as trusting analytics output. Some of the trust has to do with the technology. People are often hesitant to adopt what they don't understand, making it important to socialize predictive analytics so people understand the benefits. For instance, executives and other leaders can help to build trust by evangelizing the concepts and practicing what they preach. As one respondent noted, "Recognition from executive management that predictive analytics can inform decisions and a commitment to be open-minded and not rely on gut feeling" is a best practice for moving analytics forward.

In this survey, nearly half of respondents cited publicizing results as a strategy to move analytics forward.

Successes with the technology can also help to build trust. In fact, publicizing results throughout the organization is often cited as a top best practice for helping to democratize analytics. In this survey, close to half of respondents cited publicizing results as a strategy to move analytics forward (Figure 12). As one respondent put it, "You need great use cases that continue to show value." Other comments included:

- "A team of dedicated experts without hidden agendas."
- "Need the ability to show ROI, especially over the first hurdle."
- "Building proofs of value (not only proofs of concept), with minimal viable products to prove value early and often."
- "Find use cases with clear value and that can be fulfilled quickly for a group(s). Identify individuals that have real influence in the organization [who] will help sell the ideas and evangelize the successes."

As these comments illustrate, it is important to provide evidence that the predictive models are making an impact. That can mean quantifying the impact and comparing that impact against past results.

USER STORY BUILDING TRUST IS KEY TO ADOPTING PREDICTIVE ANALYTICS

"Have a close relationship with the business. Find a business sponsor you can partner with to do a proof of concept (POC). You'll have to show them the value of the product; don't take anything for granted," said a BI manager at a marketing organization. The company had been outsourcing the predictive models associated with marketing campaigns. However, the BI team thought they could save the company money and get better results if they built the model in-house. "We had to send the data to the vendor. That could add weeks to the project and impact results," said the BI manager.

The group built a campaign model using commercial tools with a GUI-based interface. "We felt that open source would be a steep learning curve," said the BI manager. The team did a test and compared the partner model with the in-house model during a six-week campaign. "We could see on initial responses that we could beat the outsourced model. In fact, we ultimately showed that our lowest-performing segment did better than the best-performing segment from the outsourced model." The team let the process play out and then presented the results to the business. Over a period of time, the business transitioned from being solely dependent on the outsourced vendor to doing the modeling in-house. Now, the in-house team is moving to other areas.

"One key was being able to tell a compelling story with the model," said the BI manager. "The partner would be able to tell the story, but it's not really one they would want to tell. We would be able to show *all* the scores and how our selections performed against our hypothesis for best responders in addition to how we performed against the partner." Another key was having the authority to determine the methodology and build the model. That means having a team that is comfortable with predictive analytics. In his case, this is a small team: one with model-building experience and a few others building competency in predictive analytics.

Satisfaction and Value with Predictive Analytics

We have been citing best practices for predictive analytics and machine learning throughout this report. However, how satisfied are organizations that use the technology? Are they measuring value?

Satisfaction

Fifty-six percent of the active group was satisfied with their use of predictive analytics and machine learning.

Overall, the active group appears to be satisfied with the use of predictive analytics in their organization. Fifty-six percent of this group stated they were satisfied with the use of predictive analytics and machine learning. Only 11% were dissatisfied. The rest (32%) were neutral (no figures shown). However, there are areas for improvement. The active group was typically satisfied with executive and organizational support for predictive analytics (64% and 59% respectively), but they were less satisfied with their infrastructure, organizational skills, and their organization's ability to easily put models into production.

Infrastructure. Twenty-one percent of respondents were dissatisfied with their infrastructure and only 44% were satisfied (Figure 13). That is understandable. Many organizations are still making use of their data warehouse for predictive analytics. Some are moving to Hadoop. However, TDWI sees the need for organizations to adopt modern data architectures in order to be successful analyzing ever-increasing amounts of disparate data. Why? Integration and data quality are top challenges cited by respondents. Organizations are using data warehouses but know that they need to incorporate other platforms to manage their data and perform iterative analysis. Furthermore, they need to be able to integrate disparate data across these platforms. That means that organizations need to develop a well-architected (and well-governed) multiplatform data environment that consists of the data warehouse and other platforms such as the data lake, streaming platforms, and cloud. This appears to be where organizations are headed, based on Figure 11. The key will be to bring the data together in a coherent fashion and understand how the organization wants to treat newer forms of data.

Close to half of the active group is using or planning to use business analysts to help build models.

Organizational skills. Lack of skilled personnel is often cited as the biggest barrier to the adoption of predictive analytics and machine learning. Although 72% of the active group has data scientists on board as well as statisticians (43%, not shown), lack of skills is still an issue. Close to half (not shown) of the active group is using or planning to use business analysts to help build models using predictive analytics and machine learning. This number is about the same in the exploring group. That means that those people will need to be trained as well as perhaps work with easier-to-use tools. As seen in Figure 13, fewer than half (46%) of the active group is satisfied with the ease of use of their software.

Ability to easily put models into production. Fully one-fourth (25%) of the active group was dissatisfied with the ability of their organization to put models into production; less than half expressed satisfaction (41%). This goes hand-in-hand with the fact that it can take time to do so, as mentioned previously. However, many respondents in the active group report that models are rewritten and the majority don't make use of APIs. Sometimes models need to be rewritten, but vendors are trying to provide ways to make deployment simpler. Likewise, as we saw, about half of the respondents don't have a designated group for model deployment. This is an area for improvement.

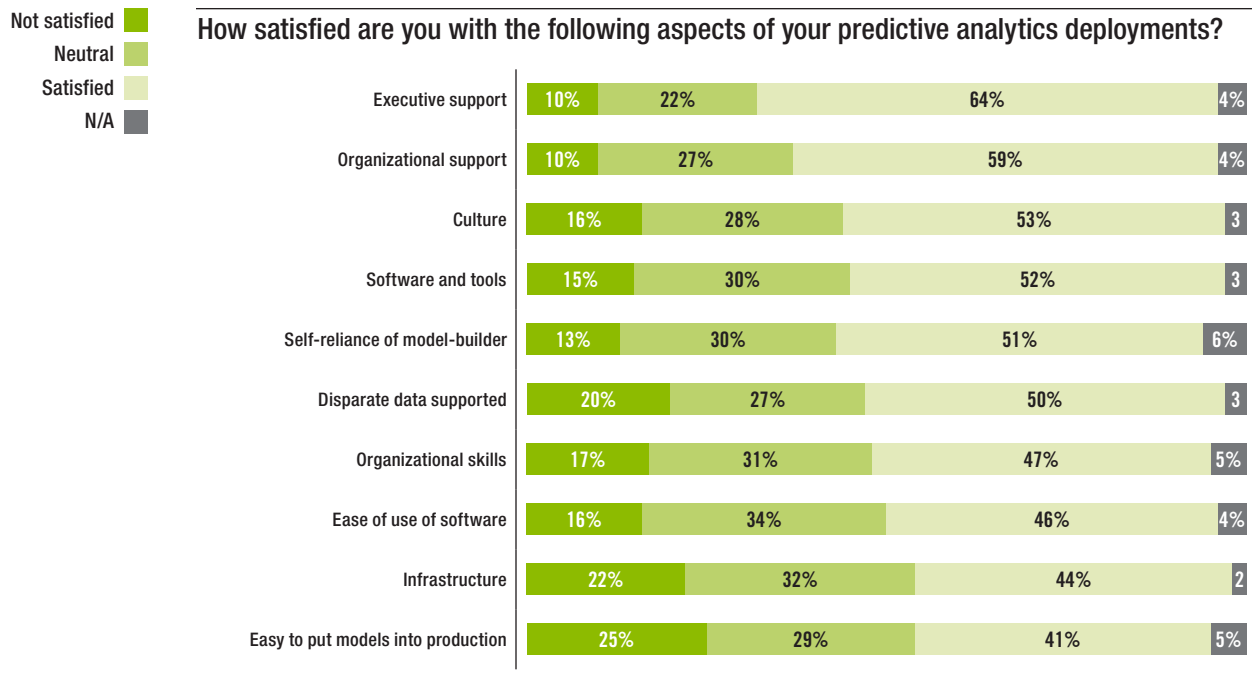


Figure 13. *Based on 219 respondents from the active group.*

Value

It is one thing to be satisfied with a technology, it is another to measure impact. To explore best practices further, we wanted to understand the characteristics of those companies that obtain measureable value from predictive analytics and if they differed from companies that were not reporting such value.

We separated the active group further into two groups—those who measured either a top- or bottom-line impact (or both) or attained ROI objectives (80 respondents) and those who did not measure an impact (57 respondents). Both groups contained members who were using the technology from less than one year to greater than three years. The rest of the active group thought they had positive outcomes but had not measured them. Other respondents believe they have gained value, but have not measured value specifically.

We compared the two groups across a number of dimensions, including demographics, organizational support, and technology. The sample size is small, but some interesting findings emerged that support the satisfaction points made earlier in this report.

Executive support. Previous TDWI research on advanced analytics points to the need to have executives help move the effort forward. The same is true in this research. Those respondents with executive support are more likely to measure value than those who do not. This makes sense; as we have discussed, executives can provide funding and evangelize the importance of predictive analytics and machine learning.

Not surprisingly, those who have not measured value are also more likely to have issues with skills.

Skills play a vital role. Not surprisingly, those who have not measured value in their predictive analytics efforts are also more likely to have issues with skills. When we asked respondents to rate how satisfied they were with particular aspects of their analytics deployments, 60% of the group that measured value were satisfied with team skills but only 35% of the group that didn't measure value were satisfied (Figure 14). This is a much higher percentage than the overall

satisfaction of the active group with skills (Figure 13). Additionally, this is a statistically significant difference in the value analysis and underscores the importance of putting the right skills in place if predictive analytics is going to work practically in an organization. This group was also more likely to be satisfied with the self-reliance of their model builders. Interestingly, the group that had not measured value was more likely to be looking to easy-to-use tools for model building as an important aspect of their predictive analytics effort (54% versus 43%, not shown). These tools might help, but only if there are ultimately some skilled people in the organization. These do not need to be Ph.D. data scientists, but they do need to understand the methodologies. Organizations that succeed do so because they have the talent to succeed.

Putting models into production is key. Those who measure value are generally more likely to be satisfied with how they put models into production than the group that didn't measure value (Figure 14). This supports the satisfaction analysis above as well as previous TDWI research about the importance of operationalizing models. Those who do so can act on the results and measure the impact. Those who simply take the results of predictive models and use it for insight have a harder time linking the outcome to the use of the model. As noted earlier, it is important to put a process in place to get models into production and then manage them in production. The group that didn't measure value may not yet have put those processes in place.

Those who measure value are more likely to be satisfied with how they put their models into production than those who didn't.

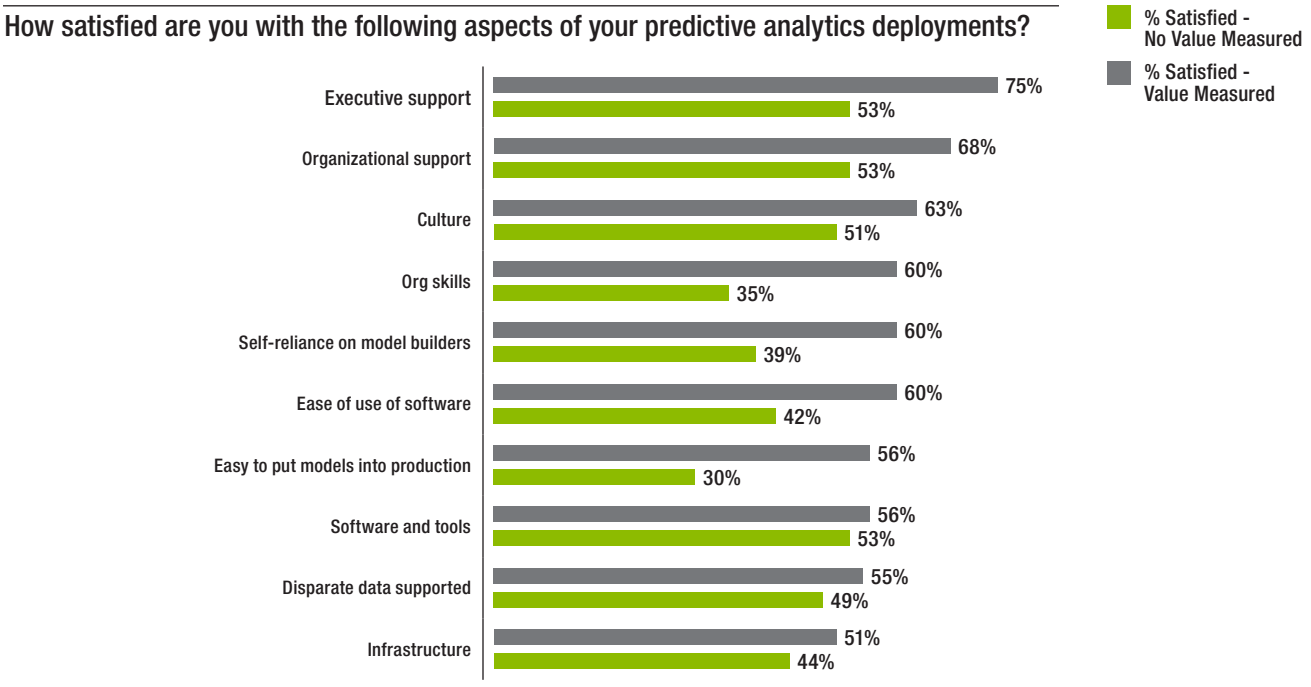


Figure 14. *Based on 80 respondents who measured value and 57 who did not measure value. Note that this is based on the same question as Figure 13.*

Vendor Solutions

The firms that sponsored this report are among the leaders in BI, analytics, and data management. To get a sense of the direction of the industry as a whole, this section looks at the portfolio of these vendors. (Note: The vendors and products mentioned here are representative; the list is not intended to be comprehensive.)

Hortonworks

Hortonworks offers 100 percent open source data management platforms, services, and solutions that support modern big data architectures. The Hortonworks Data Platform (HDP) is powered by the company's Apache Hadoop distribution and is designed to scale out for distributed clusters and managed data lakes. The platform is completely open source with YARN at the core and supports connecting data over multicloud and on-premises environments.

The company also offers an Apache NiFi-based stream processing/analytics platform called Hortonworks DataFlow. This platform provides data collection, transformation, and content routing from the edge of the network to the enterprise. HDF also supports event-stream processing using Kafka and Storm to perform real-time analysis. This supports moving models to the edge as well. The Hortonworks DataPlane Service (DPS) extends the company's data management, security, and governance offerings across federated data fabrics. DPS provides security controls as well as a method of secure communication across hybrid environments. Currently, Hortonworks provides a data life cycle manager that controls and manages data between and across multiple tiers. The company will offer more services in the future.

SAP

Software giant SAP offers a number of predictive analytics products and services. SAP Predictive Analytics is an on-premises solution that supports the entire analytics life cycle—from data preparation to model development and model management. The company also offers automation features for data preparation and model building and reuse. Targeted at the citizen data scientist and the data scientist, SAP Predictive Analytics provides connectivity to a range of data sources as well as to the SAS ecosystem. It can run in-memory, on HANA. SAP also offers a cloud version with a subset of the functionality.

SAP applications such as its CRM and ERP are enabled with predictive analytics to help end users who aren't that familiar with tools such as machine learning. SAP has picked key use cases in these S/4 HANA enterprise applications. For instance, SAP S/4 HANA has several machine learning models for stock and transit that will automatically analyze the data in the application to forecast supplier performance based on past performance with a particular supplier.

The goal of software giant SAP is to enable the intelligent enterprise. The company offers multiple approaches to machine learning and AI in order to address end users' needs. It directly embeds machine learning capabilities into its ERP solutions and other line-of-business applications to help different kinds of users take advantage of machine learning technology with no additional training required. For more advanced business analysts, SAP delivers machine learning capabilities for building customized models without requiring advanced knowledge of data science techniques. This solution is delivered in a cloud environment, called the SAP Analytics Cloud, which brings BI, machine learning, and planning together in a single place. SAP Analytics Cloud will also provide capabilities for embedding customized models into multiple SAP applications.

Tellius

Founded in 2015, Tellius is a next-generation AI-powered software solution targeted at business users and business analysts. The company provides an easy-to-use business analytics platform that embeds machine learning into its software to enable business users and business analysts to get to insight faster. This includes natural language-based search analytics that provides suggestions for analytics questions to users as they type.

Recently, the company launched the Tellius Genius Insights engine that provides automated discovery of insights, patterns, and trends using machine learning. It also offers a machine learning capability that evaluates and tunes multiple algorithms to determine the best-fit predictive model. APIs can be used to embed the models into applications.

Apache Spark powers the platform.

Recommendations

This report has detailed many best practices for predictive analytics and machine learning. In closing, we summarize the report by listing the top 10 best practices for making modeling practical, along with comments about why each is important. Think of the best practices as recommendations that can guide your organization into successful predictive model implementations.

Know your business problem. It is important to start with a *real* business problem and clear objectives when embarking on a more advanced analytics project. You should still experiment with data—exploration is part of analysis. However, understanding the business problem at hand is more likely to lead to success than looking for patterns in a sea of data.

Get executives on board. We have seen in this report that organizations that measure value are more likely to be satisfied with their executive support than those who do not. It is easier to get predictive analytics and machine learning funding if you can get an executive to sponsor the effort. The executive can also help to evangelize the results, use the results, and help build trust in model outputs. This can go a long way in helping predictive analytics to become more widespread in your company.

Make sure you have the skills. There are excellent tools on the market that can help everyone become more productive. However, the reality is that if your organization really wants to perform sophisticated analytics, you're probably going to have to hire at least a few data scientists. As this report notes, organizations that have measured an impact with predictive analytics are more likely to be satisfied with the skills in their organization than those who have not. In addition to helping execute the work and potentially build apps, these people can also provide guidance to those business analysts trying to come up to speed. So, if you're planning to train existing talent, make sure to set aside funding for training whether it is internal or external.

If your organization really wants to do sophisticated analytics, you're probably going to have to hire at least a few data scientists.

Build a center of excellence. A CoE can be a great way to make sure your infrastructure and the analytics you implement are coherent. CoEs can help your organization disseminate information, provide training, or maintain analytics governance (such as reviewing models that will be put into production).

Think about the architecture, including the cloud. The data warehouse is not going anywhere anytime soon. However, predictive analytics and machine learning necessitate moving beyond the data warehouse to platforms that can support multistructured data and iterative analytics.

As we have seen in this report, there is room for improvement in terms of data infrastructure. These multiplatform data architectures might include the data warehouse, Hadoop, and other platforms, both on premises and in a public cloud.

Think about analytics platforms that support multiple personas. Organizations are interested in analytics platforms, and they should be. These platforms help provide unified data and support multiple personas using interfaces tailored to individual needs. Data scientists get their access to open source programming environments and business users can access user-friendly tools. Many of these platforms support collaboration, which is important for successful model building. These platforms also provide some tooling to help support model management, which is key to predictive analytics success in the long-term.

Consider newer tools. To get your analytics program off the ground and gain visibility, it is important to consider some of the newer analytics tools available. This includes open source as well as tools that embed advanced analytics functionality (such as machine learning) into the software. Traditional vendors as well as newer entrants are offering these tools. Some vendors offer the tools as part of an analytics platform.

Model governance is critical. Technologies that support model versioning and send automated alerts if a model goes stale can be very helpful.

Govern the models. This includes establishing controls for model development and deployment as well as monitoring models once in production. Model governance is critical. Technologies that support model versioning and send automated alerts if a model goes stale can be helpful here.

Operationalize predictive models. Analytics provides the greatest amount of value when someone can take action on the results. Think about embedding predictive analytics into a system, process, or application. Consider what that will take and how you will do it. Will you need to rewrite the model? Alternatively, can you use APIs to do the job? Additionally, think through who is going to put the models into production. Do you have a DevOps group or another group that is responsible for model deployment? As you put more models into production, someone will need to be responsible.

Don't declare victory too soon. As many of the respondents stated, it is important to show value with predictive analytics before it becomes widely accepted in an organizations. The first insight from a POC is not when you should be saying the job is done.

Hortonworks
hortonworks.com

Hortonworks is a leading innovator in the industry, creating, distributing and supporting enterprise-ready open data platforms and modern data applications. Our mission is to manage the world's data. We have a single-minded focus on driving innovation in open source communities such as Apache Hadoop, NiFi, and Spark. We along with our 1600+ partners provide the expertise, training, and services that allow our customers to unlock transformational value for their organizations across any line of business. Our connected data platforms powers modern data applications that deliver actionable intelligence from all data: data-in-motion and data-at-rest. We are Powering the Future of Data.

SAP
sap.com

As the market leader in enterprise application software, SAP is at the center of today's business and technology revolution. Our innovations enable approximately 300,000 customers worldwide to work together more efficiently and use business insight more effectively. SAP helps organizations of all sizes and industries overcome the complexities that plague our businesses, our jobs, and our lives. With Run Simple as our operating principle, SAP's nearly 77,000 employees focus on a singular purpose that inspires us every day: to help the world run better and improve people's lives.

Tellius
tellius.com

Tellius is a business analytics platform powered by machine learning so anyone can ask questions of their data and discover hidden insights with a single click. The platform is based on the high performance computation engine of Apache Spark and scales up to big data use cases. Tellius connects to a wide variety of data sources—databases, files, applications, and big data repositories. The platform can be deployed on-premises or hosted in the cloud.

The Tellius Genius AI Engine utilizes embedded machine learning algorithms for business users, data analysts, and data scientists to automatically perform advanced analysis such as key drivers, segmentation, trends, and anomalies from billions of data points and complex data. Automated machine learning allows data professionals to build, test, and evaluate machine learning models, and these models can be operationalized via the Predict API. Search-based analytics also gives users a simple interface to explore their data and perform ad hoc query.

TDWI Research provides research and advice for data professionals worldwide. TDWI Research focuses exclusively on data management and analytics issues and teams up with industry thought leaders and practitioners to deliver both broad and deep understanding of the business and technical challenges surrounding the deployment and use of data management and analytics solutions. TDWI Research offers in-depth research reports, commentary, inquiry services, and topical conferences as well as strategic planning services to user and vendor organizations.

tdwi

Transforming Data With Intelligence™

555 S. Renton Village Place, Ste. 700
Renton, WA 98057-3295

T 425.277.9126
F 425.687.2842
E info@tdwi.org

tdwi.org