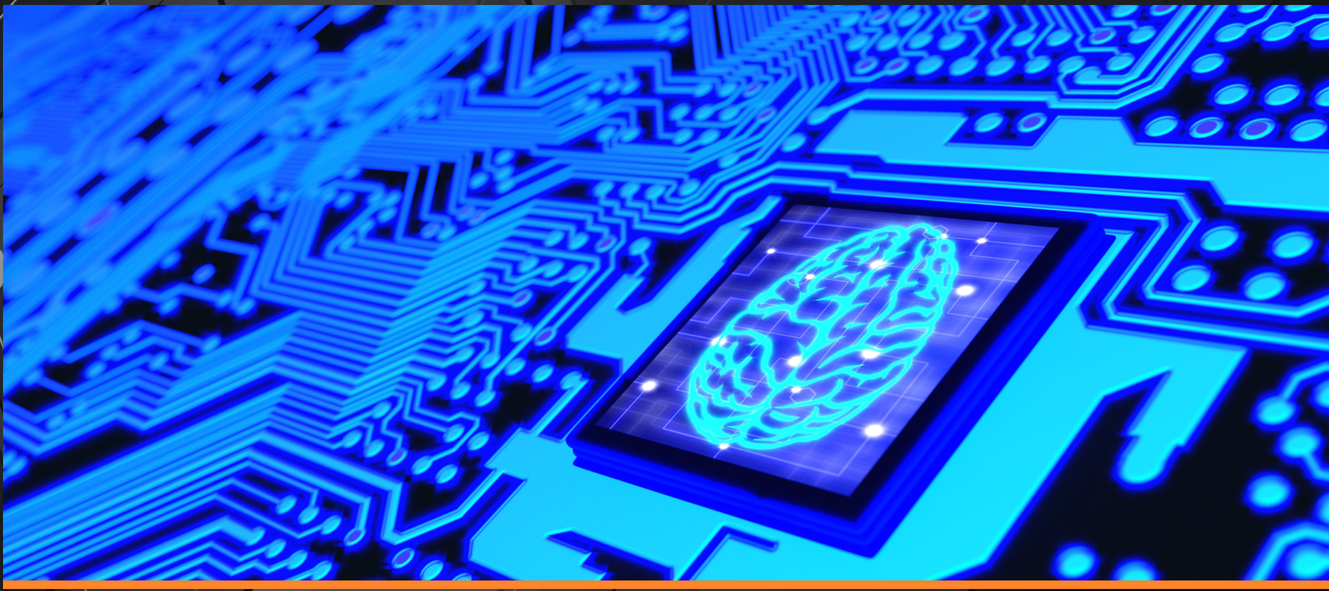
research

CHECKLIST REPORT 2020

# Six Things You Need to Know about Machine Learning to Be Successful



By Fern Halper, Ph.D.

Sponsored by:

snowflake®

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

SEPTEMBER 2020

TDWI CHECKLIST REPORT

# Six Things You Need to Know about Machine Learning to Be Successful

By Fern Halper, Ph.D.

TABLE OF CONTENTS

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

Product and company names mentioned herein may be trademarks and/or registered trademarks of their respective companies. Inclusion of a vendor, product, or service in TDWI research does not constitute an endorsement by TDWI or its management. Sponsorship of a publication should not be construed as an endorsement of the sponsor organization or validation of its claims.

## FOREWORD

Today, more than ever, organizations understand the power of using analytics to adapt to changing business conditions. TDWI research indicates that the vast majority of organizations realize they need to digitally transform to compete; many organizations are looking to accelerate this change. Data science tools such as machine learning are at the heart of this recent wave.

*Machine learning*—the ability of systems to learn from data with minimal human intervention—has been around for decades. Over the past five years, however, organizations have begun to embrace the technology in earnest; it is now in the early stage of mainstream adoption with demand continuing to grow. In a recent TDWI survey, 71 percent of respondents said that demand for machine learning was increasing.[1] That is up from 64 percent the previous year.

Data science technologies such as machine learning are used across many departments in the organization. For instance, machine learning is used in marketing to better understand customer behavior and predict churn. It is being used in operations to predict when a part might fail. It is even used in HR for employee retention.

Why? Because the value of machine learning is real—TDWI research indicates that those organizations that utilize more advanced techniques such as machine learning are more likely to measure a top- or bottom-line impact than those that do not.

At TDWI, we see adoption of advanced technologies such as machine learning as part of a virtuous circle. One part of the organization may start a machine learning initiative. As they gain success with it, other parts of the organization become interested. This success builds on itself. Success, of course, isn't just about the technology; both organizational and technology factors are important for machine learning to succeed.

Organizational factors include supportive leadership, a culture of analytics, the right resources, and organizational models to execute. On the technology front, machine learning requires the capacity to collect, manage, and access large amounts of accurate and diverse data, the ability to create new features and train models, and the capability to deploy, monitor, and update models in production.

This checklist report focuses on six *technology* considerations for successful machine learning and data science. We will see that a cloud data platform can help with many of the items presented in this report.

[1] Unpublished 2020 TDWI survey

## 1 CONSIDER A CLOUD PLATFORM FOR DATA MANAGEMENT

The evolving data landscape is complex, and this can cause issues for an organization that needs to manage many data sources and types. Various data sets often need to be brought together for data science because model training demands diverse data. The traditional on-premises data warehouse was not meant for the kind of large-scale iterative analytics required to develop this training data.

At TDWI, we see many organizations moving towards a multiplatform environment that consists of both on-premises and cloud solutions for data management. In order to support analytics they will need to unify their data. Some organizations unify data with a data fabric approach such as data virtualization. Others put their data into a data lake, a cloud data warehouse, or a data platform. One approach gaining popularity for modern data management and analytics is using a cloud data platform.

A *cloud data platform* is an integrated platform available on public clouds to house diverse data and provide services such as a data warehouse, a data lake, data sharing, analytics, governance, and administrative tooling. The cloud data platform is a single data platform for supporting multiple workloads and data types. It provides one place to store and access data for analytics, data science, and other use cases. As part of the platform, only a single copy of the native data needs to be stored, yet it is made accessible across all nodes.

In addition to providing centralized storage, a cloud data platform can also fulfill compute requirements. Organizations need to provide dedicated compute resources for each workload (ETL, BI, data science, etc.), and a cloud data platform can provide that capability without creating contention for the same resources.

Some of the benefits of a cloud data platform mirror those of the cloud: elasticity, scalability, and flexibility. A cloud data platform can improve productivity, provide agility across various workloads, and improve time to value. In addition to internet-scale capabilities and elasticity, one reason organizations move to a cloud data platform is to simplify the data environment. For instance, the platform automatically deals with software and infrastructure management and updates so the IT team does not need to.

Finally, TDWI research suggests that within the next three years, on-premises systems will no longer be the primary home for analytics data. The mixture of source types for analytics is shifting, which results in less data from traditional enterprise sources and more of a mixture of modern and traditional data. Modern data has a faster growth rate, more diverse data structures and is commonly created in the cloud. If it is generated in the cloud, it often makes sense to also analyze it in the cloud, especially as data volumes grow; a cloud data platform can be advantageous for this.

## 2 COLLECT USEFUL AND DIVERSE DATA

As mentioned, modern data comes in all shapes and sizes. As organizations enhance their analytics capabilities, they are often looking to utilize diverse data. New data types, such as semistructured data, text data, image data, web logs, and machine data are all important for advanced analytics. The data might be internal to the company or come from external sources such as demographic data, competitor data, or data specific to a vertical such as risk-assessment-related data.

The point is that data scientists want to incorporate a variety of data types and sources in their models because that can produce more accurate results. For instance, sentiment data might be a predictor of customer churn. Weather data is important in travel and logistics models. Machine data can be used to develop predictive maintenance models.

Once the problem to be solved has been identified and framed, it is important to collect and understand the data that may be used to train the model. This includes:

- **FINDING, COLLECTING, AND ACCESSING THE DATA.** Data science is an iterative practice, which can mean the data scientist needs to go back and collect additional data many times during the course of a single project. Because that data may be scattered among many systems, this is another argument for a unified data approach such as a cloud data platform. Offering data in a single platform may remove hours of latency potentially waiting on an ETL job to bring data in from another environment. Additionally, some cloud data platform vendors provide access to diverse data through a marketplace making unique data from suppliers available to consumers. This third-party data is easily accessible and fresh data is available on demand with no delay.
- **UNDERSTANDING THE DATA.** In addition to collecting and accessing data, good analysis starts with understanding the data in order to create the right features and choose the best models. That means data scientists may want data discovery and visualization tools for slicing and dicing and getting familiar with the data as part of their analytics arsenal.
- **ENSURING DATA INTEGRITY.** Finally, the data for analysis needs to be trustworthy; the old adage "garbage in, garbage out" definitely applies to machine learning. That requires a data quality strategy. Some organizations will look for a unified approach such as a cloud data platform—then they can trust the machine learning data because it is profiled and cleansed before going into the platform. Some platforms provide catalog capabilities so users can more easily find and access their data. The catalog may also provide ratings or rankings of data sources or certifications so users can easily determine which data sources are high quality.

## 3 REMEMBER THAT FEATURE ENGINEERING IS CRITICAL

Remember, machine learning algorithms learn from the data they train on; that means they need curated input data. If the data used to train the model has issues, the model won't be accurate. Data scientists often spend a lot of time on a task called *feature engineering*—constructing new derivative attributes that can better represent the problem being solved. This involves transforming raw data into something more meaningful, i.e., something clearer than what is available in the raw data.

An example of a feature might be a customer score (derived from raw data) for a churn model or a calculated variable called "length of time a customer." These may be based on structured data. Feature engineering must also be done on other kinds of data. For instance, unstructured text data needs to be processed, normalized, and converted into numeric values that a machine learning algorithm can understand. The quality of the features often drives the quality of the model.

Many data scientists argue there is an art to feature engineering and that is true. It can be quite complex and take time to accomplish. Domain experience is often required, as is an understanding of the parameter requirements for each model. The first set of features the data scientist tries might not be predictive. That can mean starting over to create new features as part of the iterative process mentioned previously.

An important trend in vendor products is support for automated feature engineering—i.e., where tooling can automatically develop features. Some tools address tasks such as imputing missing values for certain algorithms, calculating certain functions (mean, etc.), or calculating ratios. Some are more sophisticated. Some only work on relational data. Some are best for image data. Automation can help prevent data leakage where testing data is used as part of the training data set in machine learning due to duplication or preprocessing errors. It can also generate features that data scientists and business users may not have considered.

Although automation does not replace the data scientist, it can assist and cut down on the time spent developing new features. The domain expert can review the features provided by the tool and select the features that may provide predictive value. Given this, some organizations are moving to a hybrid approach that utilizes both manual and automated feature engineering. Where a manual process is in use, some organizations prefer SQL to build features, rather than other tools such as Spark, because they feel it is more efficient—especially if they utilize a cloud data platform.

## 4 SUPPORT MULTIPLE TOOLS FOR DATA SCIENTISTS

Data scientists often utilize many toolsets. For instance, some data scientists are trained in open source tools such as R and Python. They may want to use these tools to develop models or to develop their own applications. Others prefer commercial tools for ease of use, features, functionality, or compliance reasons. When data scientists are viewed as a scarce resource, the organization will often support multiple toolsets to allow the data scientist to use their preferred tools. It is also important to support new tool types such as automated machine learning that can enable data analyst users with data science capabilities.

**AUTOMATED MACHINE LEARNING IS GROWING.** Automated machine learning (AutoML) is becoming more popular, both for data scientists and modern analysts (e.g., business analysts who train to do more data science work). Like automated feature engineering, the idea behind augmented intelligence and AutoML isn't to replace humans but to help them with tasks.

AI technologies are being applied across the analytics life cycle, including in building predictive models. In some AutoML products, all the user needs to do is specify the outcome or target variable of interest and supply the raw data. The software determines multiple potential models, conducts automatic feature engineering, then trains and suggests the best fit model. Other tools may be less automated and only provide a subset of these capabilities.

At TDWI, we see AutoML as a big growth area if users stick to their plans. Data scientists are interested in these tools because they help improve their productivity. Modern analysts are interested because they can help build data science models without requiring technical expertise. TDWI recommends that organizations using AutoML tools still learn the machine learning techniques in order to understand and defend their models.

**SUPPORT FOR MULTIPLE PERSONAS IS CRITICAL.** Support for multiple tools will most likely result in support for multiple personas involved in machine learning. This can include the data scientist as well as others who support the data science process, such as data engineers, DevOps, and advanced analysts who are building models—perhaps using some of the automated tooling described above. The interface should support these multiple personas, and the data underneath should be common and trusted.

Regardless of the kinds of tools used, the key is that the organization has data management tools that can support these machine learning tools. The machine learning space is evolving rapidly; it's imperative that data scientists can access the evolving tools, frameworks, and libraries they require (e.g., a deep learning tool or a natural language processing tool) without needing to re-platform their data.

## 5 DON'T FORGET ABOUT THE BACK END

Organizations often think about hiring data scientists to build models; they are concerned about accomplishing the front-end steps in model building—collecting data, preparing data, transforming data, developing features, and model training. Although these steps are obviously important, organizations often don't consider what is necessary for putting the model into production. This oversight can lead to project failure or increase the amount of time needed to deploy a model. Production steps include managing the model, deploying the model, monitoring the model, and retraining it.

**MODEL MANAGEMENT.** Model management includes registering and versioning models that are put into production. Just as a piece of software is versioned, a machine learning model should be too. It is important to version the model in the model-building stages as well as in the case where you deploy new versions immediately when you've made a change. That means registering and versioning the model and capturing metadata about the model—including when it was built, who built it, and what data was used to train the model. In this way, the organization will be able to track models that are put into production and know what version of the model they are running. Versioning models is also important for audit and compliance purposes. There are both commercial and open source tools available for model management.

**MODEL DEPLOYMENT.** Once a model is built and validated, it can be deployed into production. This involves exporting the model as well as developing the pipeline to score fresh data in batches or streaming. There are different approaches to model deployment. For instance, some organizations will rewrite the model so it fits into a production system or application. This is generally not advisable because it can introduce many errors into the process. Others export models using APIs. Some export them in containers. Still others use certain frameworks or deployment standards such as PMML (predictive model markup language) or the newer ONNX (open neural network exchange), which provides an open source format for AI models (deep learning and machine learning).

Regardless of the method used, it is important to be able to feed data to the model in production. This will require gathering data, doing any preprocessing, and recalculating the features that need to be input into the model. In other words, having a consistent data pipeline is important for building the model as well as operating it in production.

**MODEL MONITORING.** Once a model is created it is good for a certain period of time—be it months, weeks, or days (or shorter, depending on the use case). Environmental and external factors are always changing. Products change. Competitors change. Models can get stale. The degradation of model accuracy is a serious problem, so organizations need to monitor models in production to see if they are drifting. This may be done manually until the organization starts to scale the number of models in production. At that point, the organization will need to deploy some sort of automated technology, preferably one that runs multiple models simultaneously.

**RETRAINING THE MODEL.** What happened in the past isn't necessarily what will happen in the future. Data for models tends to drift from the original training set, often because the assumptions used to build the model have changed. The last step of operationalizing analytics is retraining the models once they've been in production and the organization is monitoring their performance.

## 6 PUT THE MODEL BACK INTO THE BUSINESS

For a data science initiative to be successful, it is important that the new models and their outputs drive action. This is the operationalizing aspect of the data science process described in Number Five. This can involve embedding analytics to bring the results of an analysis to the decision maker and/or other applications. Embedding analytics also opens analytics up to more users—it makes analytics more pervasive, more actionable, and more valuable.

Utilizing a model in production can take a number of forms, all of which ensure that the model is actionable. Approaches for putting the model back into the business include:

- **SURFACING THE RESULTS IN A DASHBOARD.** Many organizations are comfortable working with dashboards. The results of a machine learning model can be displayed in an interactive visual dashboard to help users across the organization act on the results. For instance, a model built using customer-related data can be surfaced in a dashboard to display the customers at risk of churn. That same dashboard can be used to highlight the top predictive factors (e.g., high monthly charges or a short-term customer). Marketers can then use that information to devise a retention plan targeted to those customers.

- **EMBEDDING THE MODEL INTO A SYSTEM.** Likewise, a machine learning model can be embedded into a system, such as a call center system. For instance, a call center agent at a wireless company might use the output of a machine learning model to make an offer to a customer at risk of dropping a service. The call center agents do not have to understand the model operating behind the scenes, although they do have to know what to do with the information they may be seeing on their screens.

- **AUTOMATING THE MODEL.** In many cases, decisions can be determined and acted upon in an automated fashion. For instance, a model that predicts the probability that a credit card transaction is fraudulent based on historical data of fraudulent and non-fraudulent transactions can be embedded into the system, and as new transactions flow through the system, they are scored for probability of fraud. If the probability meets a certain threshold then an action is automatically taken—such as the transaction being declined. Another example of model automation is a recommendation engine. In this case, historical buyer behavior data from online activity might be used to create a machine learning model that helps predict what customers might buy based on what other products they (and those like them) bought in the past. So, when users interact with the site they are given personalized recommendations based on customers with similar characteristics.

- **EMBED THE MODEL INTO A DEVICE.** Of course, analytics can also be embedded in devices. One example of this is in the area of predictive maintenance. Here, a model is developed using data from previous malfunctions. This data might come from sensors, such as those that measure temperature, pressure, load, or vibration of parts. The model is then deployed into a device to either send an alert or take an action—such as shutting down a piece of equipment, if the model predicts there will be a critical failure—before a problem occurs.

Many organizations are building their applications to run in the cloud. A cloud data platform can provide the foundation for these applications. Models can be scored in the cloud platform and then passed to the application for use.

## CONCLUDING THOUGHTS

Machine learning is here to stay; demand is only increasing. TDWI sees certain technology best practices as important for machine learning. These include support for the machine learning workflow including data collection, data visualization, feature engineering, model building, management, and deployment.

Many organizations are looking to support this workflow using a cloud data platform that provides one place to store and access data for machine learning and other advanced use cases.

## ABOUT OUR SPONSOR

snowflake®

Snowflake's cloud data platform shatters barriers that have prevented organizations of all sizes from unleashing the true value from their data. Thousands of customers deploy Snowflake to advance their businesses beyond what was once possible by deriving insights from their data by all their business users. Snowflake equips organizations with a single, integrated platform that offers the data warehouse built for the cloud, instant, secure, and governed access to their network of data, and a core architecture to enable many types of data workloads, including a single platform for developing modern data applications. Snowflake: Data without limits. Find out more at Snowflake.com.

For more machine learning resources and to learn how Snowflake can help, please visit:

- A Cloud Data Platform for Data Science
- Five Things A Data Scientist Can Do To Stay Current

## ABOUT TDWI CHECKLIST REPORTS

TDWI Checklist Reports provide an overview of success factors for a specific project in business intelligence, data warehousing, analytics, or a related data management discipline. Companies may use this overview to get organized before beginning a project or to identify goals and areas of improvement for current projects.

## ABOUT THE AUTHOR

**Fern Halper, Ph.D.,** is VP 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 coauthor of "Dummies" books on cloud computing, hybrid cloud, service-oriented architecture, and service management, and Big Data for Dummies. 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 by email (fhalper@tdwi.org), on Twitter (twitter.com/fhalper), and on LinkedIn (linkedin.com/in/fbhalper).

## ABOUT TDWI RESEARCH

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