research

CHECKLIST REPORT **2020**

Six Success Factors for Getting Started with Machine Learning Across Your Enterprise

By Fern Halper, Ph.D.

Sponsored by:

H2O.ai

tdwi TRANSFORMING DATA WITH INTELLIGENCE™

MARCH 2020

TDWI CHECKLIST REPORT

Six Success Factors for Getting Started with Machine Learning Across Your Enterprise

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

© 2020 by TDWI, a division of 1105 Media, Inc. All rights reserved. Reproductions in whole or in part are prohibited except by written permission. Email requests or feedback to info@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

Machine learning (ML) is fast becoming part of the fabric of many organizations as they look to embark on an AI transformation. Organizations are embracing machine learning to gain better insights, make better decisions, and improve competitive advantage. In a recent TDWI survey, for instance, over 80 percent of respondents stated that building predictive models using tools such as machine learning was a top AI use case.[1] Annual TDWI surveys reveal that demand for machine learning continues to increase.

Machine learning is one of several technologies that are typically included under the AI umbrella. In machine learning, systems learn from data to identify patterns with minimal human intervention. There are numerous use cases for machine learning across the organization. It is being used to streamline the entire supply chain and improve core business functions such as reducing fraud. It is being used in operations and IT (for predictive maintenance) as well as marketing (for example, for customer behavior analysis or customer churn) and customer service (such as to make next best offers). Machine learning is even used in HR for use cases such as employee churn and employee hiring.

Although machine learning most commonly works on structured data, it is increasingly used against a range of data types. It is a component in natural language processing for text analysis and natural language understanding. In deep learning, machine learning can classify images. It is employed in computer vision. The technology is even embedded in a host of applications across the analytics life cycle to make it easier for organizations to derive insights and build predictive models. At TDWI, we see that those organizations that make use of more advanced technologies, such as machine learning, are also more likely to measure a positive impact from their analytics efforts.

There is a wide range of ML use cases that can help organizations grow. However, the technology is still primarily in the early mainstream adoption stage. At TDWI, we also see that many organizations get stuck when they start to make use of the technology, which can prevent it from being used in other parts of (or even across) the organization. Success with machine learning is part organization-based and part technology-based.

This TDWI Checklist Report discusses key best practices for getting started with AI/ML and making it work across your enterprise.

[1] See the 2019 *TDWI Best Practices Report: Driving Digital Transformation Using AI and Machine Learning*, available at tdwi.org/bpreports.

1 LEADERSHIP AND ORGANIZATIONAL STRATEGIES ARE KEY

Machine learning adoption is people-centered. At TDWI, we see that organizational issues are as significant as technical issues in getting analytics programs off the ground. Machine learning is no exception. It will involve cultural changes that are important to address as part of an organization's strategic plan.

Organizations that succeed in evolving their analytics strategy often have similar characteristics that include being goal-driven, empowering, collaborative, transparent, and skills-based. There are a number of strategies that TDWI survey respondents cite as helpful in advanced analytics efforts:

- **FIND THE RIGHT CHAMPION.** Leadership is critical to any analytics effort. Leaders are responsible for ensuring things get done and they set the vision and tone for the organization. They can help evangelize data-driven concepts, empower users, and put the funding, training, and organizational structures in place to facilitate the move to more advanced analytics such as machine learning. TDWI often sees that it is easier to get an analytics program off the ground and widely accepted if the right leadership is in place. This doesn't have to be a chief analytics officer (although that can help). We see many successful programs led by VPs and directors of analytics or those aligned with strategic lines of business. If machine learning is aligned with the business strategy, it can grow successfully.
- **IDENTIFY THE RIGHT PROJECT.** It sounds obvious, but often one of the biggest obstacles organizations face as they start to adopt machine learning is identifying and articulating the business problem. It is important to understand the use case and then hone that use case. Organizations need to ask questions such as "What is the problem statement?", "What is the outcome?", and "What is the impact?" An enterprise needs to be able to change its mindset from reactive to proactive and articulate the problem in that way. Start any machine learning effort with a real business problem and clear objectives. Ideally, the objectives are measurable so the organization can quantify the impact of the machine learning project. This will help to articulate the impact of a successful project and help to get others on board to help machine learning grow.
- **ORGANIZE TO EXECUTE.** We often see successful organizations utilizing a center of excellence (CoE) model. They typically start with a few data scientists who can build models. There may also be business analysts involved (e.g., using augmented intelligence tooling described below). The CoE can be either centralized or distributed. Some organizations like to distribute talent into business units in an embedded/consultative approach. Some centralize the talent in a hub-and-spoke model. Where there is a CoE, 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.

Additionally, as organizations begin to scale the number of models they build and put into production, they often need to put staff (such as DataOps, ModelOps, or DevOps) in place

1 LEADERSHIP AND ORGANIZATIONAL STRATEGIES ARE KEY CONTINUED

to deal with machine learning deployment and model monitoring. Data engineers may also need to be involved. These staff members may require new skill sets or may build upon existing skills in IT, but having the right skill set in place will be critical to scale machine learning across the enterprise.

- **MAKE SURE TO COLLABORATE.** It is important to have different parts of the business involved in new analytics efforts. In fact, respondents to TDWI surveys often cite collaboration as a top best practice. For instance, on the technical front in machine learning efforts, IT may be responsible for data engineering and DevOps, while the business may be responsible for building a machine learning model or owning the effort. IT may support organizations across the enterprise in terms of putting models into production. The point is that the effort is cross-functional.

  Additionally, putting models into production will require a change to the organizational mindset if it affects a manual way of doing business or changing a business process—such as a maintenance process or marketing campaign processes. Everyone will need to buy in (e.g., marketing, operations). Such a cultural change will require change management. The more the organization can work together from the get-go, the better the outcome.

2 CONSIDER AUTOMATION WHERE POSSIBLE

As organizations build out their analytics efforts, they will need to deal with ever-increasing amounts of disparate and diverse data. Many organizations are looking to automate portions of their data management and analytics efforts to meet speed and scale requirements. In TDWI surveys, automation is often cited as a top priority for data management.

Two important areas of tooling include automated data pipelines and augmented intelligence. These tools can help you improve productivity, especially with manual, repetitive tasks. Increased productivity will drive adoption across the enterprise.

- **AUTOMATED DATA PIPELINES.** Many organizations still engage in hand coding of extract, transform, and load (ETL) routines, even though modern solutions are available that can automate these steps. Automation tooling is being used in data ingestion/pipelines to establish standard, repeatable processes for collecting, profiling, transforming, and connecting data. These pipelines can be saved, scheduled, and reused.

  Machine learning also often requires data to be prepared and encoded in a particular way for algorithms to work. Machine learning depends on features for model building. Features may be raw attributes or they may be engineered (such as a ratio or an engagement metric). Often, features are categorical variables (e.g., variables with a limited number of values such as blood type). That categorical data will need to be mapped to a numeric value and encoded for use in building a machine learning model. Often more than one encoding scheme is needed depending on the nature of the data (e.g., does it make sense to treat a data element as a continuous variable). Additionally, once that ML model goes into production, fresh data that will be scored must be transformed before it flows through the model. Those same features may need to be encoded on the fly with new data. Look for tooling that can support and automate this.

- **AUGMENTED INTELLIGENCE FOR MACHINE LEARNING.** Given that data scientists and statisticians are often in short supply, some analytics vendors are offering tools that help data scientists and business analysts (and even business users) automate the construction of machine learning models (sometimes referred to as augmented intelligence or autoML). In some of the tools, all the user needs to do is specify the outcome or target variable of interest along with the attributes believed to be predictive. The autoML platform picks the best model. Other tools are even more automated. Some of these tools are in the public cloud; some are licensed on premises. Some provide both options.

  There is increasing interest in augmented intelligence. For instance, in a recent TDWI survey, although less than 20 percent of respondents are using these tools now, an additional 50 percent plan to use them in the next few years.[2] We've recommended that users have the skills to verify the insights produced by these tools. Augmented intelligence tools for ML can be a great productivity booster and can help increase enterprisewide ML adoption, but users need to understand the techniques used to automate model building and should be able to interpret and explain the results themselves.

[2] See the 2019 *TDWI Best Practices Report:Driving Digital Transformation Using AI and Machine Learning*, available at tdwi.org/bpreports.

3 RESPONSIBLE AI WITH EXPLAINABILITY AND INTERPRETABILITY IS CRITICAL FOR ADOPTION AND TRUST

People don't trust what they can't understand and many ML models (such as neural network models) act as black boxes, meaning that their inner workings are not available to the user. This can mean that there is no way for the user to understand how predictions are derived, what they mean, or how to interpret and explain the results.

Yet, interpretability, explainability, fairness, and transparency are important for machine learning models because ML models are being used in decisions that materially impact individuals as well as organizations. They are used to determine whether someone qualifies for a home loan or a credit card. They can pick potential job candidates. The list goes on. Due to the growing impact of machine learning, there is increased academic, governmental, and organizational interest in these issues.[3]

Understanding a prediction helps to build trust in that prediction and, ultimately, in the model. Important concepts include:

- **FAIRNESS.** Models learn from data and often social bias already exists in that data. An important form of bias in machine learning is discriminatory bias based on the data used to train a model. It is important to understand the bias inherent in data to ensure fair predictions across all groups. For instance, if more male candidates have historically been selected for a certain job type, then an ML system trained with the data biased towards males might suggest more males as likely good fits for that job. If the machine learning model is a black box, it is possible that no one would notice this bias until the model is put into production and it becomes clear that no women are being hired for the job. It is critical to guard against discriminatory bias. Because the regulations in this area often fall behind the technology, organizations must be proactive because of the social, economic and ethical implications,[4] including transparency in data used in machine learning as well as companies considering ethics when developing autonomous intelligent systems.[5]

- **INTERPRETABILITY AND EXPLAINABILITY.** Although some people will use the terms interpretability and explainability interchangeably, others argue that they are different terms or that one builds on the other. Interpretability is about understanding an algorithm and how changes to it can change the output. Explainability involves understanding the *why* behind an ML prediction in a way a *human* can understand. For instance, a doctor should be able to understand why a machine learning model provided a specific diagnosis for a patient using logic the doctor would understand. A customer should be able to understand why his loan application was rejected.

  Work is underway in new methods and algorithms to help with explainability using text or graphic illustrations.[6] Some of these are now included in automated machine learning products. Aside from ethical and transparency factors, new regulations also require explainability. For instance, Article 22

[3] For instance, see https://www.whitehouse.gov/ai/ or https://www.nist.gov/topics/artificial-intelligence.

[4] See, for instance, the World Economic Forum's white paper, *How to Prevent Discriminatory Outcomes in Machine Learning*, for guiding principles and questions; available at http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

[5] See the World Commission on the Ethics of Scientific Knowledge and Technology's *Preliminary Study on the Ethics of Artificial Intelligence*, published in 2019; available at https://unesdoc.unesco.org/ark:/48223/pf0000367823

[6] For instance, LIME is an algorithm that can explain the predictions of any classifier or regressor; see Ribeiro, Singh, and Guestrin, "*Why Should I Trust You?" Explaining the Predictions of Any Classifier*, available at https://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf

3 RESPONSIBLE AI WITH EXPLAINABILITY AND INTERPRETABILITY IS CRITICAL FOR ADOPTION AND TRUST CONTINUED

> of the GDPR states that users have the right to review automated decisions. That requires that a model used to derive business decisions be understandable—and that means explainable by those who created it.

As mentioned above, some newer tools on the market provide features to address the interpretability/explainability issue. At a minimum, a tool needs to provide the statistical output from a model (e.g., a confusion matrix), to help the builder understand some details of the model. However, more innovative tools provide explanations for users that include items such as feature importance, derived feature importance, and how variables interact with each other. This information is often presented in a dashboard. Users should determine which features are most important to their organization based on skills, regulatory concerns, and ethical considerations.

4 DON'T FORGET DEPLOYMENT AND MODEL MANAGEMENT

Although organizations put considerable emphasis, investment, and resources into the model development stage of the analytics life cycle, putting the model into production is often overlooked. That can lead to project failure or increase the amount of time needed to deploy a model. Yet, putting a model into production is where the real value lies. Experienced organizations understand the steps to take after the model is developed in order to operationalize and manage it. For success, it is important to proactively think about deployment so your organization doesn't build models it can't deploy or manage.

An organization should consider the following technical steps as it starts to build machine learning models:

- **REGISTER:** Registering models provides information (i.e., metadata) such as who built the model, when it was built, who updated it, and important attributes in the model. The registry also stores how many versions of the model have been built. As organizations scale the number of models they put into production, registration becomes more important.

  Additionally, given data scientist turnover, it is important to keep track of potentially valuable models and information about the models. This also promotes sound model governance, which will build trust in the models that are deployed.

- **VALIDATE.** Once a model is built and registered, validate it to make sure it will perform well in production. Aside from testing the model against new data and new conditions, models need to be validated against technical factors such as alignment of data in the target business process—do they work in the system they'll be embedded in as opposed to just with the training environment? Other key factors include the latency in scoring the data and the size of the model. Some deployment environments limit the size of the model in RAM.

- **DEPLOY.** Machine learning models can be deployed in batch or in streaming mode. They can be deployed on premises or in the cloud. Sometimes the biggest challenge is to change a business process to be able to consume the results coming from these models. That often means the model becomes embedded in an application, a web service, an in-memory database, or a legacy system where it will exist in production.

  Some organizations put a wrapper around the model and make it available as an API (e.g., Java API, Python API, REST server, etc.) to the target system. Some DevOps (or ModelOps or DataOps) teams support certain frameworks or deployment standards, such as PMML (Predictive Modeling Markup Language). More often, organizations are using containers.

- **MONITOR AND RETRAIN.** Enterprises must monitor deployed models to see if they degrade over time. Data changes. External conditions change. Models must be updated to reflect this in order to continue performing optimally. 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 a machine learning model is retraining a model once it is in production and

4 DON'T FORGET DEPLOYMENT AND MODEL MANAGEMENT CONTINUED

the organization is monitoring its performance. Look for tooling that helps with this task.

Being able to manage models at scale will be critical for enterprisewide machine learning to succeed. If a model becomes stale and is no longer producing relevant and reliable results, an entire machine learning program can fail. Tooling on the market can help with these steps. Realize, too, that this may call for new roles such as DevOps in your enterprise, especially as the models begin to scale out.

5 MARKET YOUR WINS

Adoption of ML and other advanced technologies can take time. What we often see happen is a virtuous circle—a chain of events that reinforce themselves through a feedback loop. Success begets success. As organizations see success in their analytics program, they start to do more. As they do more and as they have more experience, they tend to see positive results. This success builds on itself to help build out analytics use across the enterprise. However, at TDWI, we also see organizations take concrete steps to market their wins in order to drive excitement and move adoption across the enterprise more quickly.

Communicating results is critical. Some organizations market wins by getting the word out to others in the company via company newsletters where they highlight their successes. Others will post the results on building walls—anywhere that will gain visibility. Some organizations even develop a marketing plan and launch formal rollouts of their analytics, including metrics to measure impact. Part of the plan is communicating how the technology drives value. At other companies, if executives champion the effort, they will often discuss analytics outcomes or cite specific groups that have had successes with analytics at companywide meetings. This often results in other leaders wanting to do something similar, and the success cycle begins.

As interest begins to grow, organizations will try to build skills across the enterprise. Some companies often hold lunch-and-learns, office hours, and other meetings to promote new technologies and to help others build the skills needed to succeed. As companies get more sophisticated, they may hold internal analytics events that include Kaggle-like competitions and demo booths where groups can share what they have done with others. Sometimes vendor partners are involved so they can share their technology with other parts of the organization that might not be using it.

The point is it is important to both make and take opportunities to get analytics efforts noticed.

6 GOVERNANCE WILL BECOME ESSENTIAL

Governance rules and policies set out how an organization protects and manages its data. These rules and processes include understanding regulatory issues, dealing with data quality, maintaining standards, ensuring accountability, and keeping data secure and compliant in what is rapidly becoming a large data ecosystem. A good governance process addresses both business compliance (such as the GDPR cited previously and newer initiatives such as CCPA) as well as technical standards for data and data management.

In TDWI surveys, data governance is always cited as a top challenge in data management. For machine learning, model governance will also become part of the governance process. Although machine learning adoption is still relatively early, organizations undertaking ML should start to think through governance for both data management and analytics to support it. This will help to build trust with the data used for analytics as well as the analytics itself.

On the data front, organizations will be capturing more diverse data for use in machine learning. This will include structured as well as unstructured data coming from multiple sources, both on premises and in the cloud. It will be important that sound data governance principles are in place. This includes technology and processes to deal with data quality because data quality will be key for building trust. As the old saying goes, “garbage in, garbage out.” It also includes capturing metadata about the data that is used in machine learning models as well as data lineage to understand how data has been changed and transformed. All of this will ultimately be important for trusted machine learning across the enterprise.

On the modeling front, models will need to be tracked and managed, as described previously. Registries will become important for collecting and tracking information about the models. So, too, will the need for models to be interpretable and explainable. Both will become part of the model governance process. Additionally, organizations will need to govern how the building process occurs in light of the move to augmented intelligence.

In organizations where business analysts are utilizing machine learning, institute a series of controls—another facet of the model governance process—before a model becomes part of a business process in production. For instance, some organizations use data scientists and statisticians working with business analysts to act as the control point. Other organizations have a comprehensive checklist of items that must be completed if a model is to go into production.

FINAL THOUGHTS

Interest in machine learning continues to grow. Currently organizations may only have a few models in production in one or two business units. To succeed and grow machine learning across the enterprise, organizations will need to adopt new technologies and processes. This includes tooling to automate manually intensive tasks to help increase productivity and scale. It will involve new models to organize and execute. It will also include new processes and tooling to build trust in the predictions, such as dashboards for explainability and interpretability that incorporate new techniques and start to lead to responsible AI.

Success may necessitate getting out of your comfort zone in terms of marketing wins and evangelizing the technology. The effort will be worth it. TDWI finds that those organizations that put advanced technologies such as machine learning to work are more likely to measure an actual top- or bottom-line impact from their analytics efforts.

ABOUT THE SPONSOR

H2O.ai is an open source leader in AI and machine learning with a mission to democratize AI for everyone. H2O.ai is transforming the use of AI to empower every company to be an AI company in financial services, insurance, healthcare, telco, retail, pharmaceutical, and marketing. H2O.ai is driving an open AI movement with H2O, which is used by more than 20,000 companies and hundreds of thousands of data scientists. H2O Driverless AI, an award-winning and industry-leading automatic machine learning platform for the enterprise, is helping data scientists across the world in every industry be more productive and deploy models in a faster, easier, and cheaper way.

H2O.ai partners with leading technology companies such as NVIDIA, IBM, AWS, Intel, Microsoft Azure, and Google Cloud Platform and is proud of its growing customer base which includes Capital One, Nationwide Insurance, Walgreens, Wells Fargo, and MarketAxess. H2O.ai believes in AI4Good with support for wildlife conservation and AI for academics. Learn more at http://www.H2O.ai.

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 vice president and senior director of TDWI Research for advanced analytics. She is well known in the analytics community, having been published hundreds of times on data mining and information technology over the past 20 years. Halper is also coauthor of several Dummies books on cloud computing and big data. She focuses on advanced analytics, including predictive analytics, text and social media analysis, machine learning, AI, cognitive computing, and big data analytics approaches. She has been a partner at industry analyst firm Hurwitz & Associates and a lead data 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 research and advice for BI professionals worldwide. TDWI Research focuses exclusively on analytics and data management issues and teams up with industry practitioners to deliver both broad and deep understanding of the business and technical issues surrounding the deployment of business intelligence and data management solutions. TDWI Research offers reports, commentary, and inquiry services via a worldwide membership program and provides custom research, benchmarking, and strategic planning services to user and vendor organizations.