research

CHECKLIST REPORT

2017

Machine Learning for Business

Eight Best Practices for Getting Started

By Fern Halper

Sponsored by:

SAP®

tdwi

Transforming Data With Intelligence™

APRIL 2017

TDWI CHECKLIST REPORT

Machine Learning for Business

Eight Best Practices for Getting Started

By Fern Halper

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

TABLE OF CONTENTS

FOREWORD

As organizations look to advance with analytics, predictive analytics is frequently on their road map. Businesses are interested in better understanding their customers, predicting behavior, and improving operational processes. They want more accurate insights and the ability to respond faster to change. Machine learning—building systems that can learn from data to identify patterns and predict future outcomes with minimal human intervention—is often on their radar.

Data scientists who engage in analysis are an important piece of the equation. Data scientists can build new models, develop algorithms and applications, and help the organization innovate. However, these data scientists are not always easy to find. TDWI research indicates that organizations are often looking to supplement the data science team by growing the skills of business analysts to use tools such as machine learning. For example, in a recent TDWI survey, 51 percent of respondents said that enhancing business analysts' skills was one of their top two strategies for growing their data science competencies in the organization.[1]

That means that organizations need productivity tools for data scientists as well as a way to equip power users and business analysts to perform advanced analytics. These business analysts can work together with data scientists and other team members to bring machine learning into the organization.

How do businesses get started with machine learning? How do organizations equip business analysts to use machine learning techniques and work in conjunction with data scientists? What do these organizations need to know? This Checklist defines machine learning and discusses best practices for the business as it takes the next step on its analytics journey toward using machine learning.

MACHINE LEARNING FUNDAMENTALS

Though the term *machine learning* has become very visible in the popular press over the past few years—making it appear to be the newest shiny object—the technology has actually been in use for decades. In fact, machine learning algorithms such as decision trees are already in use by many organizations for predictive analytics.

In machine learning, the computer learns from examples, typically in either a *supervised* or an *unsupervised* approach. In a supervised approach, an algorithm is given a set of inputs and makes predictions for a set of corresponding outcomes, or target variables, called *labels*. The target attributes can be classes or numeric values. For instance, in churn prediction the target variable might be a class "leave" or a class "stay." The algorithm uses historical data about customers who churned and those who didn't churn to extract patterns of attributes that relate to outcomes labeled "leave" or "stay." This is the *learning* or *training* phase. The patterns are tested using a test set and then, if the model is valid, it can be used to predict the outcome labels on future data. This is the *application* or *scoring* phase.

One example of a supervised approach is a decision tree, often used where the target variable has a discrete value. Another supervised technique is a regression, which is often best suited for a continuous variable.

In unsupervised learning, an algorithm is given a set of input variables but no labeled outcomes. The algorithm searches automatically for distinct patterns in the input data and groups it into mutually exclusive segments based on similarity. K-means clustering is an example of an unsupervised approach.

With the vast amount of data and compute power available today, machine learning is being used on a range of use cases from marketing to manufacturing. Machine learning is also used in more advanced applications such as in smart cars and in image recognition where the system learns to identify and classify images. The use cases are wide and varied.

[1] For more information, see the 2016 *TDWI Best Practices Report: Data Science and Big Data: Enterprise Paths to Success*, online at www.tdwi.org/bpreports.

NUMBER ONE

LEARN THE THOUGHT PROCESS

Many businesses have done a good job implementing BI in their organization. However, predictive analytics requires a different thought process. BI does a good job slicing and dicing to help answer questions such as "What happened?" or "What is happening?" Data visualization enables the user to explore data and uncover relationships in a multilevel way to gain insight through visuals that include scatterplots, geospatial maps, dashboards, heat maps, and more. Visualization can be very powerful. However, these tools are reactive and examine what *has* happened. Machine learning can help predict what *will* happen. This requires a different thought process in terms of formulating questions, understanding uncertainty and ambiguity, and thinking through results.

Predictive analytics requires a proactive rather than a reactive mindset. In the example for predicting churn described previously, the outcome variables were "leave" or "stay." The goal of the modeler is to build a model that has a high probability of predicting that particular outcome of interest. That means formulating a predictive question, thinking about features and labels that may be predictive, and assembling those attributes.

It includes being able to think about derived or calculated attributes. For example, you might want to calculate the length of time someone has been a customer or the length of time between purchases because that information could be predictive where individual purchase dates are not. These calculations may involve transforming data, such as a changing a continuous attribute into a discrete attribute—for example, binning revenue values into buckets.

A predictive mindset also requires questioning the results of the model and asking whether they make sense. It is important not to accept results at face value. Correlation does not necessarily mean causation. Predictive analytics also entails dealing with probabilities and uncertainty and adjusting to thinking in this way. For instance, a result might state that the predictive strength is only 70 percent. In some cases, no pattern at all will emerge using machine learning. These "failures" are okay.

All of this is different than traditional BI, where the results of a SQL query are accepted as fact (as long as the data quality is solid). Organizations must get used to this.

NUMBER TWO

FOCUS ON THE USE CASE

Machine learning is a powerful tool that can help businesses gain insight into multiple kinds of behaviors—from people as well as machines. It is used in horizontal and vertical applications to help organizations become more proactive. However, it is important to focus objectives on specific use cases that will have a meaningful impact for the organization. Some of the many use cases for machine learning include:

- **Marketing.** TDWI research indicates that marketing is often one of the first groups in a business to make use of more advanced technology in order to understand customers. In marketing, machine learning is often used for customer segmentation and to provide customers with the "next best" offer. A learning model can be trained on how customers with similar characteristics responded historically to an offer. Other use cases include up-selling, cross-selling, and operationalizing machine learning in recommendation engines.
- **Operations management.** One use case that is becoming popular is preventive maintenance. Here, data from sensors and other devices is used to determine when a part failure might occur. For instance, an oil company might use sensor data from an oil rig (temperature, pressure, etc.) to build a predictive model about failure. As new data is generated and similar conditions are detected, an alert is generated, and maintenance is scheduled. This kind of application is being used in many industries such as manufacturing, healthcare, and transportation. It can be used in IT operations analytics to proactively analyze IT assets and to perform root cause analysis more quickly and automatically take action.
- **Fraud and risk analysis.** Financial institutions and finance departments use historical fraud data to train models to understand patterns associated with fraudulent business transactions. Utilities are implementing machine learning to identify fraudulent patterns of electricity usage.
- **Patient-related analytics in healthcare.** Today, TDWI is often seeing healthcare as an important area for advanced analytics such as machine learning. Use cases include predicting infection, proactive patient heath engagement, and population health analysis.
- **Security.** Machine learning is starting to be used to identify patterns of suspicious activity on a network or in a facility that might indicate a security breach.

NUMBER THREE

LOOK FOR THE RIGHT PREDICTIVE TOOLING

Machine learning algorithms are available through many sources that can provide both data scientists and business analysts with the appropriate tools and interfaces for their needs.

For the business analyst, newer commercial tools are becoming easier to use. Some interfaces are point-and-click and drag-and-drop, where users can string steps together for an analytics workflow. Other tools allow users to call machine learning algorithms through a SQL interface. Many tools provide automated model building through intuitive user interfaces. Here, the user simply specifies the target variable of interest along with the rest of the data attributes and then the software determines the best algorithm and model, given the data. These tools usually explain the output in a way that is easy to understand, such as using graphical representations together with text output. Some allow model export; some do not. Some of these tools are in the public cloud; some are licensed on premises. Some provide either option. Although these tools are quite easy to use, it is still important for the business analyst to understand what the algorithms do—they need to interpret the results and may have to explain or defend the output (see Number Four).

For the data scientist, who may want to build a model from scratch, many of these tools provide scripting interfaces to access models built in open source languages such as Python, which includes a programming language as well as a number of machine learning libraries. These tools also provide interfaces to integrate with popular open source statistical languages and environments such as R, as well as APIs to interface with existing data sources. Many products provide a notebook-style interface for the data scientists where code can be created, executed, and even versioned.

For both business analysts and data scientists, tools often have data preparation facilities. Features can include data profiling and data transformation. Some provide frameworks for facilitating automated data preparation in order to reuse components and help data scientists and business analysts be more productive.

NUMBER FOUR

GET SOME TRAINING

Machine learning tools might be easy to use and business analysts are often skilled in data analysis, but it is still important to receive training on machine learning. This does not necessarily mean going back and getting a degree in computer science or statistics, but it does mean taking the time to understand the fundamentals behind the techniques that are used and how to use the tools. Why? At the end of the day, whoever is doing the analysis needs to be able to stand behind the results and explain them. That means understanding what the techniques do and how to interpret the outputs.

There are multiple paths organizations take for training.

- **Internal training.** Some organizations, especially those that are more analytically advanced, have a CoE (center of excellence), which consists of a cross-functional group that provides leadership in analytics. Often, members of these CoEs provide internal training to those businesspeople looking to better understand analytics tools and techniques.
- **Vendor training.** Vendors typically offer training online or at their user conferences, although some hold training sessions outside of their user conferences. These sessions (typically one or more days) can help users understand how to use specific vendor tools and the features and functionalities that the tools provide. Some vendors also provide training on various techniques.
- **External training.** Many organizations send their staff to external training or bring external training to the company via onsite or online classes. Training varies in length. There are boot camps that last several days to a week and introduce business users and analysts to the fundamentals. Online courses can be less than a day in length. Some organizations will send their staff to even longer programs, such as university programs that offer degrees in data science.
- **Self-training.** Although not necessarily the best way to learn, some people teach themselves about machine learning and other data science disciplines. They read books, take online classes, and experiment with software.[2]

The method, of course, will depend on your budget and how serious your organization is about machine learning.

[2] For one example, see Provost and Fawcett, *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking* (O'Reilly Media, 2013).

NUMBER FIVE

REMEMBER—GOOD QUALITY DATA IS STILL IMPORTANT

Some people believe that the sheer volume of data for machine learning can negate data quality concerns—that is, patterns can emerge even from poor quality data. However, data scientists and business analysts want to analyze relatively clean data. Good data quality is critical for models that are put into production; otherwise the models will degrade quickly. The phrase "garbage in, garbage out" applies here.

Poor data quality can affect operational efficiencies, decision making, and reporting—to name just a few areas. Studies have shown that poor data quality can also impact machine learning algorithms.[3] Sound data quality goes beyond dealing with missing data or outliers. It must deal with issues such as data accuracy (is it correct and reliable), completeness (is it provided once and only once), timeliness (is it still relevant), consistency of format, and reasonableness.

Of course, not every analysis requires near-perfect data, and data scientists certainly have tricks up their sleeves to deal with data issues. It can make sense to experiment with various analyses using data in a data lake that may not have been completely vetted. In fact, experimentation is a hallmark of data science. Data scientists and business analysts like an area where they can explore data. However, once it is decided that a data source should be part of the analysis that will be used for decision making, then the organization must determine the necessary level of data quality. This process should definitely involve IT.

Data quality and data governance go hand in hand. That means that a governance process that includes policies and practices needs to be put into place and someone needs to own it. Existing data governance practices may have to be expanded or revamped, but this should be a joint effort between IT and the business.

NUMBER SIX

ESTABLISH MODEL GOVERNANCE PROCESSES

It should be clear now that, given the right tools and training, machine learning models can be built by business analysts as well as data scientists. Successful organizations must often determine which models make sense for the data scientist to build and which make sense for the business analyst. It is a risk/reward calculation. For instance, the organization may not want a business analyst building an image-recognition system for top-level security clearance using deep learning; the risk would be too high. It might be fine, however, for the business analyst to build a marketing campaign model that learns from the results of previous campaigns. In that case, the risk to the company if something goes wrong is lower.

In organizations where business analysts are utilizing machine learning, it is important to institute a series of controls—a kind of model governance process—before a model becomes part of a business process. For example, this might include the use of collaboration features in machine learning platforms that enable analysts to share their work with others. The business analyst can then ask questions of the data scientist as he or she is building the model. Such questions may be about how to interpret the model results, if the data provided to the model makes sense, and so on.

Of course, some organizations use a much more formal process—especially in the case of a model that has material implications. This makes good sense. Here, it might be necessary for the business analyst to obtain sign-off from the data scientist before the model is put into production. This can help avoid issues down the road if there is a problem with the model or if the analyst did not think through data issues clearly.

[3] See research by Sessions and Valtorta, "The Effects of Data Quality on Machine Learning Algorithms," (International Conference on Information Quality, 2006).

NUMBER SEVEN

PUT MACHINE LEARNING INTO ACTION

Organizations often find that the two most challenging areas for advanced predictive analytics efforts are defining objectives and deploying a model into production. TDWI research has found that it can take upwards of six to nine months to put models into production. Yet, what good is a model if there is no action taken on it?

Of course, action can take many forms. For example, a machine learning model can be used to generate rules that are then manually enforced. However, the goal of many organizations is to automate machine learning models in production, such as to identify fraud or determine recommendations. That means that the model must be operationalized as part of a business process, and that means there are a number of factors to consider.

- **Consider new tooling.** Modern decision-management software allows organizations to register, deploy, monitor, and reuse models that might be incorporated into a business process. When you only have a few models you want to operationalize, it might be acceptable to store them in a directory and ask IT or your development team to recode them. However, this is not scalable or practical in the long term. Think about how hard it would be to manage hundreds or even thousands of models manually (see Number Eight).
- **Think about the design.** Whatever kind of operationalized analytics your organization puts in place, the predictive models need to fit into that workflow. That often means customizing the front end so that it fits—especially for embedded analytics. If people are diverted away from their way of doing things, adoption will likely be more difficult.
- **Plan for deployment.** Analytics are worthwhile only if they are used to take action. That means that your organization needs to organize to execute around these analytics. If the analytics include models deployed into a system or application, someone needs to do that work. Successful organizations plan for deployment and assign deployment teams to this work.

NUMBER EIGHT

MANAGE, MONITOR, AND OPTIMIZE CONTINUOUSLY

The reality in the world of machine learning is that there may be many models built. It is important to keep track because models can get old and stale, and accuracy will degrade. Depending on the business problem, a model may need to be updated frequently. It is also important to keep track of the models to maintain institutional knowledge.

Some organizations will manage the models in directories. However, this is not a good long-term solution because as they build more models it might be hard to monitor them manually. As stated previously, it is hard to keep track of hundreds or thousands of models. Some tools provide model management facilities that have some sort of model registry built in to them. Other tools are able to manage models automatically and detect when a model goes stale because the accuracy of the model degrades. An alert is then sent to the appropriate person. If your organization is going to be building many machine learning models, these kinds of tools are worth looking into.

The frequency of model update depends on how frequently data and market conditions are changing. Seasonality may affect data and model accuracy. It is important to think about how often data becomes irrelevant. For instance, say you have a model for predicting the probability that a customer will buy your product. It may depend on new competitors and new products and will almost certainly depend on what products you are selling. If your product line changes and a product is discontinued, it is time to change the model. As a result, some models change daily, though some can obviously be in production much longer than that.

CONCLUSION

TDWI has consistently seen that machine learning and predictive analytics can provide both top- and bottom-line value to organizations. With the right tools, training, and processes, organizations can take the first step to utilizing machine learning for a range of objectives and use cases.

ABOUT OUR SPONSOR

www.sap.com

As the market leader in enterprise application software, SAP is at the center of today's business and technology revolution. SAP helps you streamline your processes, giving you the ability to use live data to predict customer trends—live and in the moment. Across your entire business. When you run live, you run simple with SAP. SAP innovations help 345,000 customers worldwide work together more efficiently and use business insight more effectively.

The SAP BusinessObjects Analytics portfolio provides a comprehensive set of modern analytics capabilities, on premises and in the cloud, that work together to analyze data wherever it resides for enterprises of all sizes and across every industry. These solutions help users better understand their business, plan and predict the future, and simplify and transform the enterprise in the digital age. By adding new mobile, predictive, data visualization, and big data analytics capabilities, as well as streamlined packaging and promotions, SAP continues its leadership and track record of offering customers the analytics solutions that deliver business value and innovation throughout an organization.

To learn more visit sap.com/predictive

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.

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.