The Data Quality and Machine Learning Readiness Test

Theresa Kushner and Thomas C. Redman

for www.dataleaders.org

Introduction

It is no secret that machine learning is all the rage. And no wonder. It offers the potential to address areas that, so far anyway, have eluded traditional technologies. Areas like reading MRIs or recognizing faces. At the same time, machine learning algorithms are no better than the data used to train and feed them. Indeed, the quality standards for data used by machine learning algorithms far exceed most other quality requirements. [See "If Your Data is Bad, Your Machine Learning Tools are Useless," (Redman, hbr.org, April 1, 2018). All this means some fairly steep organizational changes will be required as well.

These concerns motivated this "Data Quality and Machine Learning Readiness Test." It is designed to help you understand the most important issues, baseline where you are, sort out which issues you must address in the short term. The test also aims to help you get started on a program that will stand you in good stead for years to come. It may take ten years and hundreds of millions of dollars to become world-class. So, over time, you should ask yourself what is the desired end state for your company. Remember, the goal is to obtain business benefit, not score well on this test.

The test consists of ten "criteria." For each criterion, there are five statements, each reflecting a different rating of where you stand against the criterion. On one end of the scale is "unaware," meaning you don't fully understand the criterion, why it is important, or what you need to do to meet it. Don't be afraid to rate yourself here. Machine learning, and especially data quality requirements, are new and unfamiliar to most. On the other end of the scale is "world-class," reflecting long-time efforts, great results, lots of learning (and almost certainly plenty of failures!) along the way. The statements between reflect steps along the way.

To take the test, you must first select the statement that most closely corresponds to where your organization stands with respect to each criterion. A strong word of caution: It is very easy to rate yourself and/or your company more highly than you deserve. For example, you may reason, "we've had a data quality program in place for a long time. We're pretty good at it" and score yourself a "3." The reality is few companies are good at data quality and, even if they are, that experience may not prepare them for the rigors of machine learning. So be brutally honest. If you are unsure which of two ratings fits best, choose the lower one.

Instructions

Step 1: Form the assessment team. We recommend that you can form an internal assessment team, guided by an independent, skilled professional. Your internal team should consist of five to nine people with a diversity of skills, including data scientists, data quality professionals, senior managers, technologists and those likely to be impacted. Recommendation: Person and/or groups taking this assessment should be at a level that understands the overall data program within their organization and/or company.

Step 2: Take the test. Read each criterion below and spend a moment making sure you understand it. Then read each of the five statements to the right. Select the statement that best corresponds to your current state with respect to the criterion.

Step 3: Revisit your scores. While the Data Quality and Machine Learning Readiness Assessment is a test, no one is keeping score. You will do yourself and your organization far more harm than good by rating yourself higher than you really are. And, as the caution above advises, this is easy to do.

Step 4: Interpret results and take appropriate action. Your success will be driven more by your lowest score than by the average. Thus, take a careful look at the criteria where you rated yourself a 0. In the example above, there are two criteria rated this way. These areas imperil your effort. So first, you must educate about what is required. Do not be embarrassed - many organizations are in similar situations, especially early on. But to be clear, DO NOT PROCEED UNTIL YOU HAVE MAKE YOURSELF FULLY AWARE OF WHAT YOU MUST DO.

Once you have become aware on all elements of the broad scope of effort required, take a look at the criteria where you rated yourself 1. Obviously, this is the best you can score if you are just starting out and in the example above, there are three such criteria. You need to gain some experience. In such situations, many are tempted to start narrow "pilot studies" or "alpha tests." These often focus on gaining some experience with the technologies used to create predictive models. Unfortunately, this will not help you gain the experience you need across the full range. While you can keep your first efforts small, they should be broad, embracing all areas where you need to gain experience.

Note: We are fully aware how difficult it is to make yourself aware, gain experience and figure out how to get to the next level. Indeed, as our cautions make clear, even evaluating where you stand is difficult. In fact, you will be tempted to score yourself often between ratings. Don't do that! If you are in doubt about whether you can confidently put yourself into one of the areas, then rate yourself in the lower. If you find later that you are as good as you thought, it's easy to adjust. Do not hesitate to seek professional help!

Step 5: Use the readiness test on an ongoing basis. It will likely take five to seven years before your data quality and machine learning program advances to a point where you are really happy with it. (Note: You should, of course, expect visible business benefits more quickly, say within two years). So be patient. Re-rate yourself twice a year. And take careful notes. For example, you may find yourself admitting "we thought we were prepared to address quality (or change), but that turned out to be much harder than we expected." This is healthy. Finally, as you score all 2s and 3s, you should ask yourself "how good do we need to be?" The answer will certainly depend on your desired business results and competitive position.

Readiness Test

Business Need/ Objective	Rating against criterion						
Organization	Criterion	0	1	2	3	4	Your Score
1 Program Level	A company must be clear how the machine learning effort contributes to its overall business objectives in terms of revenue growth, cost of risk reduction, etc. It must devote sufficient resources (time, money and people) both short and long-term, to the effort.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time and have learned a lot. We're getting there.	I/we have considerable experience in this regard. Our business and data/analytics/ML efforts are fully in sync.	Our data/analytics/ML efforts have produced considerable benefits, to the degree that we call them out on our annual report.	
2 Project Level	Project teams with a keen sense of urgency must be clear about objectives for their individual projects.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time and have learned a lot. We're getting there.	I/we have considerable experience in this regard. Our business and data/analytics/ML efforts are fully in sync.	Our data/analytics/ML efforts have produced considerable benefits, to the degree that we call them out on our annual report.	
3 Cultural Understanding	Companies must recognize features of their unique cultures that either advance or hinder data quality/machine learning efforts and size their objectives, programs and projects accordingly.	I/we do not fully understand how culture effects our internal efforts	I/we understand what role culture plays in this data initiative, but I/we don't know how to manage it.	I/we know culture is important and I/we take it into consideration with our projects.	I/we are using our culture to help us make the right changes to our organization.	Our culture is data-driven and as such, we value the ML efforts that contribute to our success.	

Readiness Test continued

Business Need/ Objective	Rating against criterion						
Quality Data	Criterion	0	1	2	3	4	Your Score
4 Trusted Sources	Machine learning demands very high accuracy levels. Those using the data must have solid evidence that their sources of data can meet their requirements.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been addressing data accuracy for some time and have learned a lot. We're getting there.	I/we have considerable experience in measuring, improving, and controlling data accuracy. Most of our organization operates against trusted data sources.	Our data is so well trusted that our executives have monetized it and treat it as a valuable company asset.	
5 Understand/ Manage Bias	Organizations should be aware of the biases that could exist in training data and the impact bias can have on predictive models.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've developed our data sources to minimize bias, but not all organizations have clearly adopted the practices.	I/we have considerable experience in this regard. Our data sources are regularly screened for bias and all models are tested for bias regularly. As we discover bias, we refresh our models.	Our machine learning efforts are well managed such that bias is understood and all efforts are made to reduce the impact of bias in our models and reporting. When the possibility of bias is large, we deploy other methods.	

Business Need/ Objective	Rating against criterion						
Quality Data	Criterion	0	1	2	3	4	Your Score
6 Comprehensive, Relevant, Predictive Ability	Available data must form a comprehensive picture, be relevant to the task at hand, and be consistent over time.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time and have learned a lot. We're getting there.	I/we have considerable experience in this regard. Our business and machine learning teams know how to select the data they need from internal sources and how to supplement our data with external sources.	Our machine learning teams are outstanding in continually building and implementing better models as they find more relevant, comprehensive data.	
7 Known quality of data in both training and use	High quality data is required both to train a predictive model and to test it in use. This includes all data: data in motion, streaming data, data stored in data warehouses, marts or lakes, etc.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've have separate data quality initiatives aimed at curating training data and improving data when used in predictive models. We're getting there.	I/we have considerable experience in this regard. All relevant data is of high quality for training, testing and use.	Our data is so well managed and available to our data scientists that we can use data from any source with any velocity throughout the business.	

Readiness Test continued

Business Need/ Objective	Rating against criterion						

Skills	Criterion	0	1	2	3	4	Your Score
8 Leadership	Individually, data quality and machine learning require changes in people's work and together the change is enormous. Organizations pursuing machine learning require a cadre of senior leaders who understand the issues and challenges and can effectively guide programs and projects as well as lead change management efforts.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time and our leaders have learned a lot. We're getting there.	Our senior leaders support the DQ and ML program, understand how difficult change is and are involved in meeting the challenges.	Our senior leaders actively drive change by incorporating DQ and ML efforts. They are public in their support, resolve organizations issues, and walk the talk.	
9 Data Science	Properly developing, training and testing a predictive model requires a diverse team, featuring people who understand the business and the requisite data. The data science team should be composed of data scientists, business analysts, storytellers and data engineers. The team works with a scientific process to determine the right use of predictive analytics for the business.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time and have learned a lot. We've begun a program to teach people basic data science. We're getting there.	I/we have considerable experience in this regard. All relevant data is of high quality for training, testing and use.	Data Science teams have been defined and are actively being deployed across our organization. Data scientists have their own career track and are effectively working with the business.	

Business Need/ Objective	Rating against criterion						
Skills	Criterion	0	1	2	3	4	Your Score
10 Project Management	A business team or project management team, committed to making the predictive models work in practice must be in place to implement, troubleshoot and measure the effectiveness of machine learning programs and projects.	I/we do not fully understand what we need to do to meet this criterion.	I/we understand what I/we need to succeed here, but I/we have never done it before.	I've/we've been at this for some time. Our project teams are diverse, and people are learning to work together. We're getting there.	We have teams that are dedicated to making sure that predictive models are embedded into our operations. We measure success and work to improve continuously.	Our success rate in building predictive models into our operations and decisions is high. Everyone is trained on the basics of data science generally and the use of machine learning in particular.	