Assessing Your Data Readiness for Machine Learning

Robbie Allen
Jul 8

As we write the book Machine Learning in Practice (coming early in 2019), we'll be posting draft excerpts right here.

Let us know what you think, give us a clap down below if you like what you read, and follow @InfiniaML and @RobbieAllen on Twitter for the latest updates!

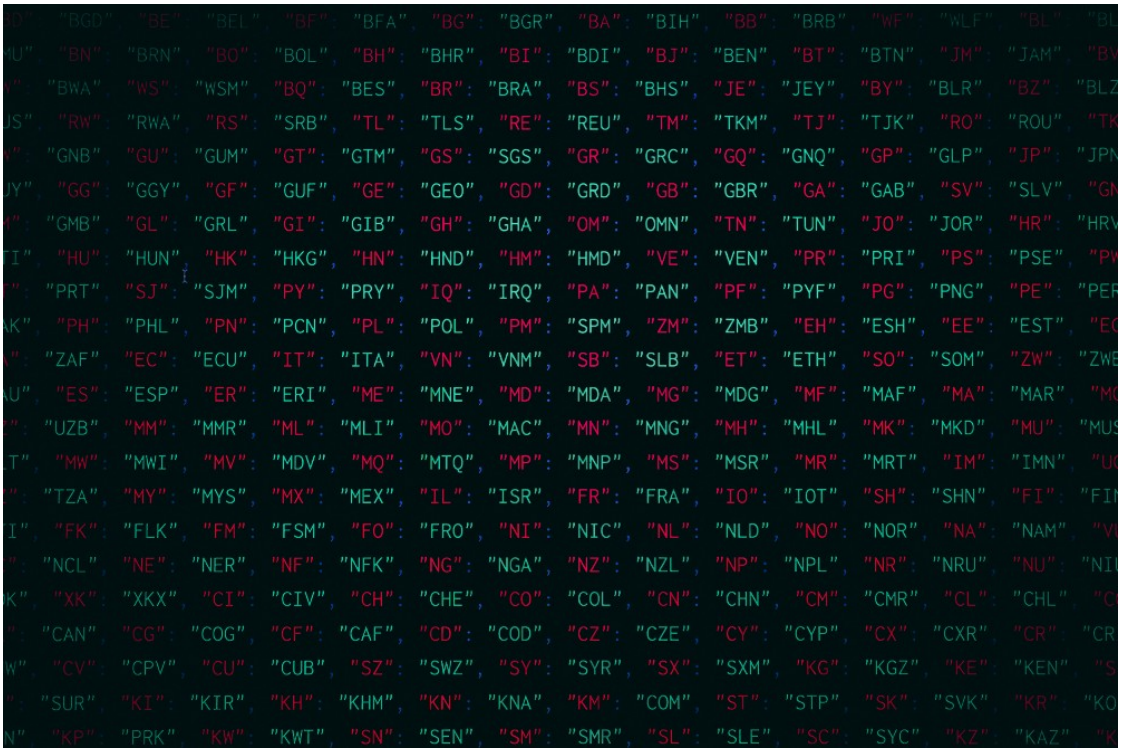

Photo by Pankaj Patel on Unsplash

There are five common reasons why a company's data is not ready for machine learning.

1. The data is not accessible

It may seem obvious that you need to actually *have* the data to do machine learning, but some companies face this initial hurdle right out of the gate. They know the problem they want to solve, but they don't have the data that would actually help them solve it.

Once you've identified a task you want to automate, the first question to ask yourself is whether the data even exists for your model or automation process. A good data strategy

should start with a data inventory. Consider it the raw materials for machine learning. If you don't have the data internally, you'll need to acquire it. Are there external data sets that you can purchase and use or are there internal data sets that you can take advantage of? Does the data exist in any capacity to address the problem you're trying to solve?

Typically, the question of whether data is available isn't a simple binary matter of having the data or not. It's more common that a client has *some* data, but it's not enough of the right kind or it's missing significant pieces.

Perhaps the data has 50 of the attributes we need, but five or ten of those are sparse or not filled in at all. Too many holes make the data unusable—you simply may not have the data you think you do, and you'll need a way to augment it before moving on with machine learning.

These gaps are often due to historical collection methods. Twenty years ago, many businesses didn't care about data. Ten years ago, they started caring about and collecting data without knowing what problems they were going to solve with it. Today, when they know what they want to do with data, the data is missing some attributes that they need. The collectors couldn't predict the future and didn't know what questions they would ultimately be asked.

Often you may find yourself thinking: "if only they had asked for these things, we'd have everything we need." This is more the rule than the exception. It's best to find workarounds or get creative when solving problems because it's rarely the case that you'll have both a well-defined problem and the exact data set you need to solve it.

Programmatic Accessibility

It is not uncommon for internal data sets, especially those coming out of a product or related to customer data, can be difficult to get for non-technical reasons.

Say you were going to use customer data. You know the data exists. In fact, it has everything that you need. But you can't actually use it because your customers never gave you permission. Or perhaps you're inside of a large bureaucratic organization with security restrictions preventing you from accessing the data. Maybe the data has personally identifiable information or other sensitive information and you can't get approvals to get to the data, even though you know it's there.

Again, consider historical collection methods. Originally, when companies established terms of service agreements for their product, they didn't contemplate reusing the resulting data for machine learning. That's because as little as ten or fifteen years ago, few companies were thinking seriously about machine learning at all. Thus, they never got the customer to agree to use it for that purpose.

In the early 2000s, few companies thought about *data* at all. By the late 2000s, "Big Data" became a popular term and companies started to embrace it, so they started to

collect data from customers. Now, in the late 2010s, companies really want to make good use of the data for machine learning. But they may not have the rights to do that because the original terms of service don't allow it.

All of this follows a natural, predictable progression for companies: 1) we don't have data 2) let's start collecting data 3) let's make use of data and 4) we need to go back and ask for permission for the data.

Today, it's common for companies to rewrite their terms of service or otherwise ask customers for permission to use their data.

Another challenge of accessibility is the organizational effort required to actually get what you need. Even if data is ultimately accessible, you may need to account for a bureaucratic lag in doing so. If you have to go through another group that's super busy, and could take months to get to the data. In some cases, this delay may be equivalent to not having the data at all.

One way to speed up the process for sensitive or secure data is to have a firm understanding of security policies and practices, This way, when the inevitable questions come up about why a group should entrust machine learning experts with its data, you already have answers to the questions that their information security departments are going to ask.

Even internally, it's important to know the likely objections others will raise about sharing their data when security is important. Preparing for what you're going to be asked is a way to help address that.

2. The data is not sizeable

Do you have enough data to train on? Are you sure?

There aren't really any consistent rules of thumb for how much data you need to make machine learning work for you. Of course, the conventional wisdom is that machine learning—and especially things like deep learning—are very data hungry. They need lots of data to adequately train on. It makes sense. If the machine is going to understand patterns in data, it needs to see a large number of examples to be able to identify those patterns.

There's been a lot of discourse in the machine learning community about whether that's actually true. Some have shown that you can build relatively interesting models with a small amount of data.

Your data volume needs will vary wildly with the specific use case and implementation. Asking how much data you need is sort of like asking when a model is "good enough."

Machine learning is never going to be a 100% solution where it can make every prediction accurately. There is a gradient for how good is good enough.

It's the same with data. There's not a magic minimum number of rows, be it 1,000 or 100,000. You'll have enough data when you can train a model that is sufficiently accurate for your given use case.

To know if that's true, you need to either be a machine learning expert or consult with one. Of course, if the amount of data is really small, you won't be off to a great start. We had a client give us one record of one person to start out. They told us that 400 more records exist. By definition you can't identify a pattern in a single data point. Even 400 records is pretty small, unless there is little variability across the data.

Of course, if you're talking about millions of rows of data, that *tends* to be fine. If you have orders of magnitude more than that, then you're *likely* okay.

But even then, it depends. If you have a million rows of data, but 90% of them are similar or exactly the same, that's probably not enough. If the million rows only represent 20% of the possible output, that's almost certainly not enough.

It's helpful to remember that you need not be limited to your company's internal data. Other types include broader industry or demographic data—census data, for example, or large repositories of natural image data. All of these can fit together: for example, general household information augments industry-specific customer data, which itself augments individual customer data.

3. The data is not useable

Is your data clean enough to be useful? Remember, "garbage in, garbage out" applies to machine learning. The accuracy of your machine learning model is directly related to the cleanliness of the data it trained on.

Sometimes data isn't missing—it's just messed up. For example, there may be 20 different spellings for a given label. Perhaps a given field is supposed to be a string that maps to something specific, but that data has several typos throughtout.

A lack of cleanliness often makes machine learning much harder. First, you have to recognize that there is in fact a problem. Second, you have to find a solution. All of this must happen before any machine learning even begins. The more time you spend cleaning up data, the less time you have to do useful work that could result in something of value.

The best way to solve this problem is to never have it, and to follow good data practices from the beginning, even if it's more painful upfront. Again, bad data yields bad machine learning outcomes. Garbage in, garbage out.

It's impossible to catch every problem in data, especially if it's been collected through bad practices. You won't always know all the corner cases for how the data could be screwed up, and so you could miss stuff. By far, the best approach is to ensure that you have the highest possible quality from the beginning through good practices of storing, measuring, and validating your data.

4. The data is not understandable

Do you know what each data field means, and can you communicate it effectively?

There's a misunderstanding sometimes that you can give a data scientist a bunch of data and a problem, walk away, and them come back in a few weeks once she's solved it. We wish it were that easy, but almost never is. The first step once you have a data set is explaining the data set. And we're not even talking about the intricacies of domain-specific knowledge. You first need a basic understanding of what each column in the data set means. If the name of column 5 is "Column 5" and it has some numbers in it, that's not very helpful.

Machine learning experts typically rely on the client to explain their data. You'd be surprised how difficult of a task this often becomes. It's not uncommon for us to get a big spreadsheet, database dump, or JSON file of data with 200 columns of information, and the columns will be named things like, "NM." What does that mean? "Name?" "New Mexico"? Something else? Sometimes the only definitions of the data are in the name of the column, and we're supposed to infer what that field contains.

Such confusion is very common, especially if you're getting the data from another group. If you don't have direct hands-on contact with the data, it's unlikely that you know what all the data fields are. We typically ask for data definitions and schemas up front to better understand what the columns contain and the relationships between them. Sometimes, there's a very important relationship that need to be defined.

Not understanding each data element is a huge impediment to actually doing interesting things with the data. The lack of data clarity is one reason that machine learning projects need to be a real, ongoing partnership between client and data scientists working on the project. There must be an ongoing dialog to resolve any misunderstandings.

5. The data is not maintainable

Can you reliably produce the data set in an ongoing manner?

A lot of work goes into data preparation for a machine learning project. The data needs to be accessed, cleaned, and described as we've covered so far. Often, you need to provide the data to a third party—external or internal—to begin work on creating machine learning models and that requires an initial data dump.

A data scientist doesn't always need the full data set to get started. A sample is enough. Generally the sample data can be pulled together in an ad-hoc fashion. That's ok to get started, but it's critical to keep in mind that if this machine learning model will be part of a long-term solution, you need to be able to provide the same access to data repeatedly.

Complex data solutions often require the creation of a data pipeline, which consist of a set of tasks that pull together and transform data into a particular format useful for machine learning. You don't necessarily need to build a production data pipeline to get a data scientist started, but it is something you should be aware of being needed before a machine learning solution can make it into a production environment.