When Translation Problems Arise Between Data Scientists and Business Stakeholders, Revisit Your Metrics

Katie Malone¹

¹Tempus Labs, Chicago, Illinois, United States of America

Published on: Jan 31, 2020

DOI: <https://doi.org/10.1162/99608f92.c2fc310d>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Column Editor's Note: *For the inaugural column in the Industrial Active Learning series, we focus on how communication challenges arise between data scientists and business stakeholders. We argue that the metrics that quantify outcomes are generally very different for data scientists and business stakeholders, making it likely that each side struggles to understand and speak in terms that are familiar to the other side. It's not hopeless, though: great teams prioritize translating results from business terms to data science terms and back, a task that's hard work but well worth the investment.*

According to Nathaniel Hawthorne, easy reading is damn hard writing. Presumably, Hawthorne had literature in mind when he made this observation, but it captures something beyond its original scope: it's also germane to communicating and deploying scientific results (especially data science) to effect change in a business. As frustrated data scientists and business stakeholders alike will admit, communication is hard. Each side speaks a different language, and has different goals, so the translation between sides makes for damn hard writing. On the data science side, it goes deeper than terminology, deeper than how readable a visualization is or how many equations are in the slide deck: it starts in the very foundations of data science training. The quantities that data scientists are trained to optimize, the metrics they use to gauge progress on their data science models, are fundamentally useless to and disconnected from business stakeholders without heavy translation. Business stakeholders bear responsibility for the communication impasse as well, which similarly goes deeper than whether they're willing and able to program a spreadsheet or read an equation in the data scientist's slide deck. Too many business stakeholders expect to be able to hand off a business problem to their data scientists without having to do the hard work of defining and quantifying the outcome they hope to achieve in metrics that a data scientist is able to optimize. Everybody wants easy reading, but few people would say it's their job to do the hard writing.

The heart of data science communication is a handshake between business stakeholders and data scientists: it's a data scientist packaging up a result and handing it off to the stakeholder for use in a business process. The way that this happens, in practice, is via a metric that both sides agree upon. Problems are defined, and solved, in terms of the metrics that measure success, which creates an assumption and reliance on metrics in a way that we might not fully appreciate. Metrics that are easy writing for a data scientist, like area under the curve (AUC) or F1-score (both of which are explained more fully later in this article), make for hard reading for a business stakeholder. Easy reading for a business stakeholder, in terms like customer acquisition or revenue growth, is difficult to tie directly from improvements in predictive modeling. When it comes to what outcomes matter, and how we measure them, data scientists and business stakeholders are speaking different languages.

To understand why, let's start in the realm of the data scientist (building and validating a predictive model in a laboratory environment) and trace the path whereby we end up in the realm of the business stakeholder (measurably improving a business outcome in the real world). Each step along that path requires a change in

thinking and a communication/translation layer, as the outputs of the upstream step get analyzed and transformed to take them progressively out of the lab and into the real world.

The cornerstone scientific work products of data scientists are models and experiments. We'll examine both in turn, but let's start by unpacking modeling. A typical way that a data scientist might build a model is via machine learning algorithms, using data about situations and their outcomes (the training data) to predict the unknown outcomes given a certain situation. Sometimes outcomes are unknown because that data wasn't collected, maybe it's because the outcome hasn't happened yet, but in either case the goal is to build a model that predicts the outcome as accurately as possible. The highest quality models will correctly predict their outcomes, but all kinds of issues from missing data to statistical noise to inappropriate algorithm choices can lead to lower-quality models.

Since we have this notion of some models being more or less 'high-quality,' and since we're all trying to be data-driven here, let's try to quantify what 'model quality' means for the data scientist using that notion to drive their research. A reasonable place to start is with the all-purpose metric taught in most machine learning courses. The AUC is defined as the integral ('area under') the receiver operating characteristic curve ('the curve') traced out by plotting out the true positive rate versus the false positive rate of a model (James, Witten, Hastie, & Tibshirani, 2017). AUC is a metric that ranges from 0.5 to 1.0 with the lower end of this range being interpretable as 'the model does no better than randomly guessing' and with the upper end being the regime of 'the model predicts correctly every time.'

AUC has a catch, though. Machine learning models make probabilistic predictions, like 'there's a 70% chance that this customer will unsubscribe from that service' or 'there's a 20% chance that this medicine will save this patient's life.' Whether to reach out to the dissatisfied customer with a special offer, or put the patient on the medicine, requires someone to decide what the minimum threshold is in order to act on the model's prediction. Data scientists love AUC because it's agnostic to what the threshold is: AUC captures, in one number, the quality of the model across all possible thresholds that could be used. However, when a data scientist goes to tell a business stakeholder how good a model is, speaking just about AUC ignores the reality of using those predictions. Somebody has to take the hard responsibility of deciding what threshold should be used, and that means making a call about what kind of mistakes the business can afford to make. The business stakeholder probably has the relevant context to understand the tradeoffs and decide where to put the threshold, but if she's already exhausted from trying to understand AUC in the first place, that's an awful lot to ask. It's also worth pointing out that we've explicitly changed metrics, from the data scientist's favorite of AUC to something more like precision, recall, accuracy, or F1-score (a metric that averages precision and recall).

But that's just the beginning. In the common case where the business is modeling something about its customers or users, and in many other cases as well, the market is very likely to have segments within the user or customer base. Empirical studies tell us that it's extremely likely that the model doesn't have the same accuracy for all segments, which means that whatever threshold gets selected will manifest as different model

performance for different segments. The problem comes from many possible sources, which a good data scientist can try to investigate and mitigate. For example, sometimes variance in model quality across segments arises because of imbalanced data, sometimes because different segments are easier or harder to learn about (Oaken-Rayner, Dunnmon, Carneiro, & Ré, 2019), and sometimes because of bias in the model or data (O’Neil, 2018). To make the model predictions as useful as possible for the business stakeholder, the data scientist has to work with them to understand what the segments are, separate them in the data, understand how the model performs differently, and tune multiple thresholds. If it sounds like a lot of work, it is. It’s hard writing for the data scientist and hard reading for the business stakeholder. But anything less leaves business value on the table.

Now let’s imagine taking those predictions into the world, where they’re being used to make honest-to-goodness data-driven business decisions. Let’s define a data-driven business decision as one where, based on what we attribute or outcome we predict, we change something, presumably to lead to a better attribute or outcome. Up until this point we are fine, conceptually speaking, with using machine learning methodologies to exploit correlations in the data for making predictions, but now we are in the realm of trying to change outcomes. Machine learning won’t save us here, instead we need to import the statistical infrastructure of causal inference and likely have to think about if experiments are necessary (Taddy, 2019). This is a change in paradigms that will take some heavy lifting, because the data collection processes, methodologies, software libraries, and educational backgrounds involved are likely going to look different for machine learning versus statistics (to say nothing of the metrics the data scientist uses to evaluate the experiments, and their explainability to a business stakeholder: p values are imperfect [Wasserstein & Lazar, 2016] and hardly intuitive). A common shortcut is to interrogate a model for feature importance and use that in lieu of causal inference, thus inviting spurious conclusions and misdirected causal connections that can actually make things worse. For example, finding that the hospital readmission rate is higher for patients who took a certain medicine, ignoring that perhaps those patients needed the medicine because they were sicker to begin with and the medicine is more powerful—refusing the sickest patients a powerful medicine is definitely not the right conclusion, but a machine learning algorithm won’t be able to tell you that.

Yet again, we have to change metrics to make our results relevant. Where before simple predictive accuracy was enough, a business metric like revenue, survival or churn rate, or user satisfaction makes more sense now. We’ve also only now arrived at the *raison d’être* of building the model in the first place: most business users are interested in data science because they want to make better business decisions, leading to better business outcomes, and business outcomes are measured in business metrics. A lot of tension arises between business stakeholders and data scientists here, because data scientists are resistant to being held accountable for business metrics that can be affected by many things other than a data scientist’s model. The anxiety can be particularly pronounced in data scientists who are less experienced with building models meant for real-world decision making—in other words, most academic model-building doesn’t prepare a data scientist well for this part of the job. To editorialize a bit here though, this is something that data scientists need to work to move past and a

place where business stakeholders need to have patience and a long-term view. Better processes do lead to better outcomes, more often than not, in the long term. But that path can be winding and data scientists need to feel safe to explore and iterate without having to constantly defend their work.

Since metrics are so important to quantifying and communicating every step in this process, let's also revisit the business metrics themselves one last time. In the context of measuring business value, it's attractive to pick a metric that's likely already being measured and will allow business stakeholders to see any changes quickly, because it's hard to know if something has improved when that difference takes months or years to materialize. Using a short-term metric as a proxy for a longer-term, more durable quantity seems like a reasonable strategy. For example, click-through rate is a proxy for engagement or interest in an ad or email. Hospital readmission rate is a proxy for suffering a medical complication. Net promoter score is a proxy for customer satisfaction. Among other shortcomings (Thomas, 2019) though, each of these metrics can be artificially pumped up in the short term if you aren't thinking (or don't care) about the long term (Harris & Tayler, 2019). Ideally we'd understand the relationships between metrics that are easy to measure and those that really quantify something of value, or use multiple metrics together to somehow capture a fuller picture than we can get from only measuring outcomes in one way, but these aren't common topics in most data science education programs—we're now into the realm of business strategy. Some efforts (Hohnhold, O'Brien, & Tang, 2015) have been made in propagating changes in short-term metrics to long-term projections, but we long ago left the realm of how most data scientists are trained to validate their models. And just to be clear, the challenge isn't just that there are many steps in that process (although there are), it's that most of the steps are ambiguous, require different people working together to bring the right blend of knowledge, and might require multiple rounds of trial and error to reach the final business goal.

Maya Angelou later amended Hawthorne's quote. She said, "Nathaniel Hawthorne says, easy reading is damn hard writing. . . . It's the other way round, too. If it's slovenly written, then it's hard to read." Every time a data scientist hands off a report with the outcome of an experiment in terms of a quantitative metric with no clear translation into a business outcome, she is the lazy writer asking her reader, the business stakeholder, to do an awful lot of work. Every time a business stakeholder asks for 'business value' from a data scientist without doing the hard work of grappling with what, quantifiably, that's supposed to mean, he is the lazy writer asking his reader, the data scientist, to clean up his disorganized thoughts. But the teams that are able to overcome this, where the business stakeholders make a real effort to understand and speak the language of the data scientists and vice versa, show great returns on their investment in data. Surely part of that success derives from being able to communicate smoothly. But part of it also comes from both sides caring enough, and working hard enough, to express their quantitative goals and results using metrics that are intelligible to their teammates. In the best teams, it's everyone's job to be a good writer.

Disclosure Statement

Katie Malone has no financial or non-financial disclosures to share for this article.

References

- Harris, M., & Tayler, B. (2019). Don't let metrics undermine your business. *Harvard Business Review*. <https://hbr.org/2019/09/dont-let-metrics-undermine-your-business>
- Hohnhold, H., O'Brien, D., & Tang, D. (2015). Focus on the long-term: It's better for users and business. In *Proceedings of the 21st Conference on Knowledge Discovery and Data Mining* (pp. 1849–1858). <https://doi.org/10.1145/2783258.2788583>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in R*. Springer. <https://doi.org/10.1007/978-1-4614-7138-7>
- Oakden-Rayner, L., Dunnmon, J., Carneiro, G., & Ré, C. (2019). Hidden stratification causes clinically meaningful failures in machine learning for imaging. In *Proceedings of the 2020 ACM Conference on Health, Inference, and Learning* (pp. 151–159). <https://doi.org/10.1145/3368555.3384468>
- O'Neil, C. (2018). *Weapons of math destruction: How big data increases inequality and threatens democracy*. London, UK: Penguin Books.
- Taddy, M. (2019). *Business data science: Combining machine learning and economics to optimize, automate, and accelerate business decisions*. McGraw-Hill Education.
- Thomas, R. (2019, September 24). *The problem with metrics is a big problem for AI*. Retrieved December 23, 2019, from <https://www.fast.ai/2019/09/24/metrics/>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
-

©2020 Katie Malone. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.