White Paper

Good Predictions != Good Decisions

Zachary Beaver
Data Scientist

October 13, 2015

ELDER RESEARCH
DATA SCIENCE · AI · MACHINE LEARNING

Table of Contents

A Fateful Tale

Ted is having a rough week at work. As a call-center employee, his main focus is on customer retention and offering promotions to his company's current subscribers. For some reason, his numbers are horrible this week. Not only are customers rejecting his promotions, they're actually canceling their subscriptions entirely! It's only Wednesday, and he's had as many customers leave in the past three days as in the previous month. He's nervous as he enters the conference room with his coworkers for their mid-week meeting.

Ted's nerves start to subside as he realizes that others are in the same predicament. Customers are churning left and right, and the manager is visibly upset. As it turns out, their abysmal numbers are the result of a change to their weekly call lists based on a mandate from the corporate office; some analysts had built a model that was supposed to provide the "best" customers to target with their phone calls. These analysts were obviously wrong. "How can they know how customers are going to react when they've never even talked to one?" Ted murmurs.

Ted soon returns to work, relieved that his manager has given the order to ignore the new call lists and return to the tried-and-true way of doing things. Ted and his coworkers comply and are relieved as customer churn returns to its normal levels…

The Problem

This anecdote is a classic example of how statistically sound predictions can lead to poor decisions, poor outcomes, and shattered trust between analytics teams and the decision-makers they hope to assist. What happened? Why did the predictions from the data science team at headquarters produce such poor results? In this real-life example, the problem was not an inaccurate statistical model or an incompetent set of employees at the call center. The problem lay in the assumption that customers predicted to be most likely to cancel their subscriptions were the best targets for intervention from a call center employee. In reality, a promotional phone call reminded the most dissatisfied customers of their discontent, causing them to churn at higher rates. Sometimes it's better to let sleeping dogs lie. To generalize, the problem lay in the assumption that "If I can predict 'X', then I will make a better decision about 'Y.'" Good predictions are not automatic precursors to good decisions. There can be a gap between prediction "X" and decision "Y" that needs to be filled.

The Solution

Bridging the gap between good predictions and good decisions can happen in a number of ways, depending on the problem's context. Rather than solely focusing on making good predictions, a data scientist's ultimate goal should be to provide the right information to decision-makers to inform appropriate action. Data scientists assist others. I've always loved the idea of the assist, especially in the game of basketball. Throwing a deft pass is a thing of beauty, but the pass is only useful as a means to scoring. In the same way, predictive models are only useful as a means to effective decision-making. They are not beneficial in their own right, and providing decision-

makers with raw predictions is a bit like throwing a pass to Shaquille O'Neal at the three-point line; we certainly don't want him to try to score from there. As data scientists, our main focus should be to help others "score" (make better decisions), meaning that our predictions should be delivered in a context most conducive to that outcome. But how do we create those contexts?

Run Experiments

An ideal way to understand the effect of model predictions is to conduct randomized control experiments. The "control group" consists of cases where decisions are made as they were prior to the predictive model, and the "treatment group" consists of cases where decisions are made based on model predictions. Experimentation is helpful for two main reasons. First, it helps the organization quantify the effect of model-driven decision making on their baseline. Second, it requires that experimenters engage with decision-makers and think critically about the hypotheses they're testing. In the above churn example, the analytics team could have begun with the hypothesis that calling customers most likely to churn decreases overall churn rates. In testing and ultimately rejecting this hypothesis, the analytics team wouldn't be indicting the statistical model, but rather the practice of targeting customers *within a given score range*. They could then test an alternate hypothesis that churn will decrease by targeting customers with a medium predicted likelihood of churning (i.e. those who are "on the fence"). These experiments would help them to avoid making assumptions and recognize another target variable: customer response to call-center promotions. This sows the seeds for another fruitful analytics effort (possibly using Uplift Modeling to separate "persuadables" from other customer groups). Stating and testing multiple hypotheses in an iterative experimental setting allows decision-makers to effectively operationalize model predictions.

This experimental environment is the bread-and-butter of A/B testers working with websites, where experiments are cheap to run, but iterative experimentation proves more difficult and costly in non-virtual environments. In these situations, experimentation is limited and may have to be supplemented by focusing on hypotheses that are most likely to be valid, or by pulling in observational (not experimentally-derived) data.

Count the Cost

A good way to generate hypotheses about prediction implementation is to inquire about the potential costs and resources that are required for action. Here, we consider (1) the amount of potential gain or loss resulting from true or false positive or negative predictions and (2) the organization's resource constraints.

The first dimension of cost is familiar to data scientists. In a binary classification problem, most algorithms allow the analyst to set a threshold, between 0 and 1, where predictions above the threshold are labeled as one class (e.g. "customer predicted to churn") and those below are labeled as another ("customer not predicted to churn"). The default is to set this threshold such that the overall accuracy of the model is maximized,

but this assumes that all modeling errors are created equal. In our customer churn example, it's likely much more costly to misclassify a departing customer as "not going to churn" (and take no action) than it is to misclassify a staying customer as "going to churn" (and take unnecessary action). The cost of losing a subscribing customer greatly outweighs the cost of marketing needlessly to one who is staying. One should lower the cutoff threshold until the total expected cost is minimized and thereby capture more "churners" at relatively little expense. A model's success metrics should be informed by the costs attributable to false positives and false negatives.

Considering the cost of modeling errors is a good and necessary step, but it still doesn't encompass much of the complexity and nuance of most decision-making contexts. Take the example of searching for contract fraud, and pretend that I've built an accurate model while following the previous suggestions. Now it's time to give my predictions to a fraud investigator:

Me: "Here are all the contracts we believe have a high likelihood of being fraudulent." [Hands list to investigator]

Investigator: "Great! …Wait, there are over a thousand contracts here! I've never made it through more than three investigations in a year. Which ones am I supposed to choose?"

Me: "Oh, well…look, I can order these based on their prediction probabilities. Just give me a second.

[Returns with ranked list]

Me: "Here you go. Just start with the contracts at the top, and work your way down."

Investigator: "OK…Now hold on. All of these contracts at the top are kind of small. Sure, they might be committing fraud, but it's chump change. I can only investigate THREE of these contracts each year. I don't want to waste my time with small fish."

Me: "Oh, alright. Well, what about this one? It looks pretty big."

Investigator: "Are you kidding?!? That contract is held by E. Corp, the largest contracting company in the agency. Investigating them would cause chaos and might cripple our agency's operations. If they're doing enough wrong, OK…, but I'd need a lot more than a model score to start that trouble. There are a lot of factors you're not considering here."

This example, based on a real scenario, shows the potential complexity of decision-making. More generally, it demonstrates an important principle: **predictions are beholden to an organization's resource constraints**. Considering these constraints is necessary for generating predictions capable of driving effective action. At the most basic level, this can involve following simple rules, such as "investigators will not inspect

issues below ‘X’ dollars.” However, a more data-driven approach is to formulate decision-making as an optimization problem (taking a page out of the Operations Research playbook). For example, use information about expected investigative return, limits on investigative hours, expected investigative hours per contract, criticality of contracts, etc., to maximize return subject to these constraints.
Notice that the optimization depends on well-calibrated prediction probabilities (i.e. if a model output is 0.55, it roughly corresponds to a 55% chance in reality). So, the data scientist should employ model calibration techniques — post-processing predicted probabilities to more accurately reflect real probabilities. In general, **extending data science from the prediction to decision-making realm is a two-way street**; not only should prediction results affect the decision-making process, but knowledge of that process should also affect the modeling and delivery of results.

Establish Feedback Loops

So far, these suggestions have focused on *initializing and establishing* predictions as verified drivers of successful decision-making, but that’s just the beginning. Situations, data, and what constitutes a good decision might change over time, and the goal should be to facilitate good and sustainable decision-making through the establishment of relational and technical feedback loops. Decision-makers should be able to report back on what’s working and what’s not (which provides great insight for future model features), and data scientists should be able to communicate warnings and caution based on what they’re seeing in new data. These cautions from data scientists are facilitated by technical feedback loops, including monitoring:

- Model inputs for shifts from their prior distributions (potentially indicating that “the game is changing” and that the model’s predictions may be less valid)
- Model prediction distributions (again, potentially yielding insight about population changes)
- Model performance on new data (allowing the data scientist to determine when the model needs to be re-trained, re-factored, or retired)
- The effect of reinforcement bias. In other words, when a model’s predictions are used to guide decision-making, decision outcomes are model-driven. When the model is then re-trained on these outcomes, its view of the world is “reinforced,” or biased, by its previous predictions. Over multiple re-trainings, this leads to less effective models.

Establishing feedback from decision-makers and data scientists creates a proactive approach to sustainable good decisions rather than a reactionary stance toward aging data products. Throughout an analytics project, the focus should be to drive effective action. In predictive analytics, it’s easy to oversimplify this goal to just “building an accurate model” because the lion’s share of time is spent on the model itself. We tend to assume that others will know how to use the results properly, which may not be the case. As data scientists, we need to resist the urge to be content with accurate predictions and extend analytics to embrace the business problem to ensure that beneficial outcomes are realized and sustained by stakeholders. An accurate model without an effective action is an effort in futility and can degrade organizational trust in

analytics. So don’t stop short of good decisions; extend the reach of analytics, and drive effective action!

About the Author

Zach Beaver is a Data Scientist at Elder Research in the Washington DC office, where he assists government agencies in leveraging their data to support better decision-making. His recent work includes fraud detection, entity resolution, and predicting upcoming pitches in MLB games. He is a graduate of UC Berkeley’s Master of Information and Data Science program and holds degrees in Biology and Computer Science from Wofford College.

www.elderresearch.com

ELDER RESEARCH

DATA SCIENCE · AI · MACHINE LEARNING