12 predictive analytics screw-ups
Make these mistakes and you won't need an algorithm to predict the outcome

Robert L. Mitchell

July 24, 2013 (Computerworld)

Whether you're new to predictive analytics or have a few projects under your belt, it's all too easy to make gaffes. "The vast majority of analytic projects are riddled with mistakes," says John Elder, CEO at data mining firm Elder Research.

Most of those aren't fatal -- almost every model can be improved -- but many projects fail miserably nonetheless, leaving the business with a costly investment in software and time, and nothing to show for it.

And even if you develop a useful model, there are other roadblocks from the business. Elder says that 90% of his firm's projects are "technical successes," but only 65% of that 90% are ever deployed at the client organization.

We asked experts at three consulting firms -- Elder Research, Abbott Analytics and Prediction Impact -- to describe the most egregious business and technical mistakes they're run across based on their experiences in the field. Here is their list of 12 sure-fire ways to fail.

1. Begin without the end in mind.

You're excited about predictive analytics. You see the potential value of it. There's just one problem: You don't have a specific goal in mind.

That was the situation at one large company that engaged Elder Research to start working with its data to predict something -- anything -- that one executive could go out and sell to his business units. While the research consultancy did agree to work with him and developed a model for his use, "No one in those business units was asking for what he was trying to sell," and the project went nowhere, says Jeff Deal, vice president of operations at Elder Research.

The executive "uses the data internally for his own purposes, but to this day he keeps hoping that someone will realize the value of the data," Deal adds.

The lesson: Don't build a hammer and then look for the nail. Have a specific objective in mind before you start.

2. Define the project around a foundation that your data can't support.

A debt-collection business wanted to identify the most successful sequence of actions to take when trying to collect from delinquent debtors. The challenge: The company had a rigid set of rules in place and had followed the same course of action in every single case.

"Data mining is the art of making comparisons," says Dean Abbott,
president of Abbott Analytics, which was retained for the project. Because the company had rules in place that always applied the exact same actions, Abbott had no idea which sequence would work better for collecting debts. "You need historical examples," he says.

And if you don't have those examples, you need to create them through a series of intentionally planned experiments so that you can gather that data. For example, for a given group of 1,000 debtors, 500 might get a threatening letter while the other 500 receive a phone call as the first step. "The predictive models can then be built to predict which characteristics of debtors respond better to the hard letter/call and which characteristics of debtors respond better to getting the call first," he says.

In this case the characteristics might include historical patterns of incurring debt, days to pay past debts, income, ZIP code of residence and so on. "Based on the predictive models, the collections agency would be able to use the best, most cost effective strategy for collecting debts rather than using the same strategy for everyone," he says. But you need to do experiments to get started. "Predictive analytics can't create information from nothing," he says.

**Editor's note:** This story was updated on Thursday, July 25 at around 9:45 AM (eastern time) to correct errors in items #6 and #9. We mistakenly attributed some quotes to the incorrect sources. Computerworld regrets the errors.

3. Don't proceed until your data is the best it can be.

People often operate under the misconception that they must have their data perfectly organized, without any holes, disorder or missing values, before they can start a predictive analytics project.

One global petrochemical company, an Elder Research client, had just begun a predictive analytics project with a great potential return on investment when data scientists discovered that the state of the operations data was much worse than they had initially thought.

In this case, a key target value was missing. Had the business waited to gather new data, the project would have been delayed for at least a year. "A lot of companies would have stopped right there. I see this kill more projects than any other mistake," says Deal.

But data scientists are used to dealing with messy and incomplete data, and they have methodologies that, in many cases, allow them to work around the problem. This time, the
business moved forward, and eventually the data scientists found a way to derive the missing target values from other data, according to John Ainsworth, data scientist at Elder Research.

The project is now on track to deliver major cost savings by accurately predicting failures, avoiding costly shutdowns and identifying exactly where to apply expensive preventive maintenance procedures. Had they waited for perfect data, however, it never would have happened, Deal says, "because priorities change and the data never gets fixed."

4. When reviewing data quality, don’t bother to take out the garbage.

Eric Siegel, president of the consultancy Prediction Impact and author of *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*, once worked with a Fortune 1000 financial services company that wanted to predict which call-center staff hires would stay on the job longest.

At first blush, the historical data appeared to show that employees without a high-school diploma were 2.6 times more likely to stay on the job for at least nine months than were employees with other educational backgrounds. "We were on the verge of recommending that the client begin to prioritize hiring high-school dropouts," Siegel says.

But there were two problems. First, the data, which had been manually keyed in from job applicant resumes, had been labeled inconsistently. One data entry person checked off all educational levels that applied, while another checked only the highest degree completed.

Compounding the problem was the fact that, for some reason, the latter person had labeled data from more of the resumes of people who stayed the longest than did the former. Those issues could have been avoided by making sure labelers were assigned a random group of resumes to key in and that each person used the same labeling methodology.

But the bigger message is this, says Siegel: "Garbage in, garbage out. Be sure to carefully QA your data to ensure its integrity."

5. Use data from the future to predict the future.

The problem with data warehouses is that they’re not static: Information is constantly changed and updated. But predictive analytics is an inductive learning process that relies on analysis of historical data, or "training data," to create models. So you need to recreate the state the data was in at the earlier time in the customer lifecycle. If data is not date-stamped and time-stamped, it’s easy to include data from the future that generates misleading results.

That’s what happened to a regional auto club when it set about the task of building a model it could use to predict which of its members would be most likely to buy its insurance product.

For modeling purposes, the club needed to recreate what the data set was like early on, prior to when members had bought or declined to buy insurance, and exclude subsequent data. The organization had created a decision tree that included a text variable containing phone, fax or email data. When the variable contained any text, there was 100% certainty that those members would later buy the insurance.
"We were assured that the indicator was known at the time" -- before the members had purchased the insurance -- but auto-club staffers "couldn't tell us what it meant," says Elder, who worked on the project. Knowing this was too good to be true, he continued to ask questions until he found someone in the organization who knew the truth: The variable represented how members had cancelled their insurance -- by phone, fax or email. "You don't cancel insurance before you buy it," Elder says. So when you do modeling you have to lock up some of your data.

6. Don't just proceed, but rush the process because you know your data is perfect.

Between 60% and 80% of the time spent on a new predictive analytics project is consumed by preparing the data, according to Elder Research. Analysts have to pull data from various sources, combine tables, roll things up and aggregate, and that process can take as much as a year to get everything right. Some organizations are absolutely confident that their data is pristine, but Abbott says he's never seen an organization with perfect data. Unexpected issues always crop up.

Consider the case of the pharmaceutical business that hired Elder Research for a project, but balked at the time allocated for data work and insisted on speeding up the schedule. The consultancy relented, and the project moved forward with a shortened schedule and smaller budget. But soon after the project started, the firm discovered a problem: The ship dates for some orders preceded the dates when the orders had been called in. "Those weren't problems we couldn't overcome, but they took time to fix," Deal says -- time that was no longer in the budget.

Once he pointed out the issue, the executive realized there was a problem and had to go back to the management team to explain why the project was going to take longer. "It became a credibility
issue for him at that point," Deal says. Lesson learned: No matter how good you think your data is, expect problems: It's better to set expectations conservatively and then exceed them.

7. Start big, with a high-profile project that will rock their world.

A large pharmaceutical company had grandiose plans that it thought were too big to fail. As it began to build an internal predictive analytics service, the team decided to do something that would "revolutionize the health care industry," Deal recalls them proclaiming in an initial meeting.

But the project's goals were just too big and required too large of an investment to pull off -- especially for a new team. "If you don't see results quickly you don't have anything to encourage you to maintain that level of investment," he says.

Eventually the project collapsed under the weight of its own ambitions. So don't swing for the fences, especially your first time at bat. "Set small, realistic goals, succeed with those and begin to build from there," Deal advises.

8. Ignore the subject matter experts when building your model.

It's a common misconception that to create a great predictive model you simply insert your data into a black box and turn the crank -- and accurate predictive models just pop out. But data mining experts who take the data, go away and come back with a model usually end up with flawed results.

That's what happened at a computer repair business that worked with Abbott Analytics. The business wanted to predict which parts a technician should bring for each service call based on the text description of the problem from the customer call record.

"It's hard to pull out key concepts from text in a way that's useful for predictive modeling because language is so ambiguous," Abbott says. The business needed a 90% accuracy rate in predicting a parts requirement, and the first models attempted to make predictions based on certain keywords that appeared in the text. "We created a variable for each keyword and populated it with a "1" or "0" indicating the existence of that keyword in the particular problem ticket," which included the text of the customer call.

"We failed miserably," Abbott says.

So Abbott went looking for more data -- from the technicians themselves. "The secret sauce is taking the data you have and augmenting it so that the attributes have more information in them," he says. After speaking with the domain experts, his team came up with an approach that was successful.

"Instead of having hundreds of sparsely populated variables, we condensed this into dozens more information-rich variables, each tied to the historic relationships to parts being needed," Abbott explains. Essentially, they matched up the occurrence of certain keywords in repair histories to discover what percent of the time a part had been needed.

"What we were doing was reworking the data to be more aligned with what an expert would be
thinking, instead of relying just on the algorithms to pull things together. This is a trick we use a lot because the algorithms are only so good at pulling together those patterns," he says.

9. Just assume that the keepers of the data will be fully on board and cooperative.

Many big predictive analytics projects fail because the initiators didn't cover all of the political bases before proceeding. One of the biggest obstacles can be the people who own the data, who control the data or who control how business stakeholders can use the data. One Elder Research client -- a payday lending firm, which offers short term loans to tide people over until their next paycheck -- never got past the project kickoff meeting due to internal dissent.

"All along the way we were challenged by the IT person, who was insulted that he had not been asked to do the work," Deal says. All of the key people who were integral to the project should have been on board before the first meeting started, he says.

Then there was the case of a debt collection firm that had big plans for figuring out how to improve its success rate. Abbott attended the initial launch meeting. "The IT people had control of the data and they were loath to relinquish any control to the business intelligence and data mining groups," he says.

The firm spent hundreds of thousands of dollars developing the models, only to have management put the project into a holding pattern "for evaluation" -- for three years. Since by then the information would have been useless, "holding pattern" was effectively a euphemism for killing the project. "They ran the model and collected statistics on its predictions, but it never was used to change decisions in the organization, so was a complete waste of time."

"The models were developed but never used because the political hoops weren't connected," Abbott says. So if you want to succeed, build a consensus -- and have C-suite support.

10. If you build it they will come: Don't worry about how to serve it up.

OK, you've finally got a predictive model that actually works. Now what?

Organizations often talk extensively about the types of models they want built and the return on investment they expect, but then fail to deploy it successfully to the business.

When consultants at Elder Research ask how the business will deploy the models in the work environment, the response often is "What do you mean by deployment? Don't I just have models that are suddenly working for me?" The answer is no, says Deal.
Deployment strategies, or how the models will be used in the business environment once they are built, can range from very simple -- a spreadsheet or results list given to one person -- to very complex systems where data from multiple sources must be fed into the model.

Most organizations fall into the latter category, Deal says: They have complex processes and huge data sets that require more than just a spreadsheet or results list to make use of the output. Not only do companies have to invest in appropriate analytics software, which could cost $50,000 to $300,000 or more, but they may need software engineering work performed to connect the data source to the software that runs the models.

Finally, they may need to integrate the outputs into a visualization or business intelligence tool that people can use to read and interpret the results. "The deployment of a successful model is sometimes more work than building the model itself," he says.

Even then, the deployment strategy may need to be tweaked to meet the needs of users. For example, the Office of Inspector General for the U.S. Postal Service worked with Elder Research to develop a model for scoring suspicious activities for contract-fraud investigators.

At first the investigators ignored the predictive models. But the tool also gave them access to data they needed for their investigations.

Then the team decided to present the information in a more compelling way, creating heat maps to show which contracts on a map had the highest probability of fraud. Gradually, investigators started to appreciate the head start the scoring gave to their investigations.

Today, some 1,000 investigators are using it. It was a learning moment even for the experts at Elder Research. "We learned a lot about how people use the results, and how they develop an appreciation for the predictive models," Deal says.

11. If the results look obvious, throw out the model.

An entertainment-based hospitality business wanted to know the best way to recover high-value, repeat customers who had stopped coming. Abbott Analytics developed a model that showed that 95% of the time most of those customers would come back.

"The patterns the model found were rather obvious for the most part. For example, customers who had been coming to the property monthly for several years but then stopped for a few months usually returned again" without any intervention, Abbott says.

The business quickly realized that it didn't need the model to predict what offers would get those customers back -- they expected to recover them anyway -- while the other 5% weren't likely to come back at all. "But models can be tremendously valuable if they identify who deviates from the obvious," Abbott says.

Rather than stop there, he suggested that they focus on the substantial number of high-value former customers...
who the model had predicted would return, but didn't. "Those were the anomalies, the ones to treat with a new program," Abbott says.

"Since we could predict with such high accuracy who would come back, someone who didn't come back was really an anomaly. These were the individuals for whom intervention was necessary."

But the business faced another problem: It didn't have any customer feedback on why they might have stopped coming and the models could not predict why the business had not recovered those customers. "They're going to have to come up with more data to identify the core cause of why they're not returning," Abbott says. Only then can the business start experimenting with emails and offers that address that reason.

12. Don't define clearly and precisely within the business context what the models are supposed to be doing.

Abbott once worked on a predictive model for a postal application that needed to predict the accuracy of bar codes it was reading. The catch: The calculation had to be made within 1/500 of a second so that an action could be taken as each document passed through the reader.

Abbott could have come up with an excellent algorithm, but it would have been useless if it couldn't produce the desired result in the timeline given. The model not only needed to make the prediction, but had to do so within a specific time frame - and that needed to be included in defining the model. So he had to make trade-offs in terms of the algorithms he could use. "The models had to be very simple so that they met the time budget, and that's typical in business," he says.

The model has to fit the business constraints, and those constraints need to be clearly spelled out in the design specification. Unfortunately, he adds, this kind of thinking often doesn't get taught in universities. "Too many people are just trying to build good models but have no idea how the model actually will be used," he says.

**Bottom line: Failure is an option**

If, after all of this, you think predictive analytics is too difficult, don't be afraid, consultants advise. Abbott explains the consultants' mindset: "You make mistakes along the way, you learn and you adjust," he says. It's worth the effort, he adds. "These algorithms look at data in ways humans can't and help to focus decision making in ways the business wouldn't be able to do otherwise."

"We get called a lot of times after people have tried and failed," says Elder. "It's really hard to do this right. But there's a lot more that people can get out of their data. And if you follow a few simple principles you can do well."

*Robert L. Mitchell* is a national correspondent for Computerworld. Follow him on Twitter at [twitter.com/rmitch](https://twitter.com/rmitch), or email him at rmitchell@computerworld.com.