White Paper

The Three-Legged Stool of an Analytics Project

by Kenny Darrell

Lead Data Scientist

February 2015

ELDER RESEARCH

DATA SCIENCE & PREDICTIVE ANALYTICS

Perhaps you have heard rumors going around that analytics is marching into the "trough of disillusionment"? [1] Or pundits who say that analytics may fail to provide results? [2] One reason those claims have a grain of truth is that people have failed to even allow their projects to succeed.

What does an analytics project need to get off on the right foot? It is a short list: a three legged stool. (How have we gone this long without that metaphor? Every field has a 3-legged stool – even Big Data, though it's very different.) [3]
If you have been following along in Predictive Analytics Times, John Elder is doing a great job digging into the Top 10 Data Mining Mistakes. [4, 5, 6, 7, 8, 9, 10, 11, 12] The lessons learned from these are a cornerstone to a great project, but many of them detail things that occur once the project is moving forward. Without addressing a few things up front you may never even get to the point where you have to confront one of these issues. You could fail before you even start.

So what are the three legs? It starts off with -- no surprise -- *data*. "It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts" said Sherlock Holmes. There are many finer points to consider -- do you have the correct data, how was the data sampled, etc. -- but these are moot without some data to start with.

Once you have data you come to the second leg of the stool: a *problem* or a goal to meet. In a typical project the need is explained by a client and a smart data scientist will mold that squishy real-world problem into a crisp approximation that is solvable via data science. If the client can't articulate the problem well, the analyst needs some real freedom to explore the business situation and access to a good bit of subject matter expertise to root out a good goal.

The third leg of the stool, and often the one I find most ignored, is a little hard to define. You need the *ability to take action* on the problem or illuminate the best path at a critical decision point. A human needs to be presented with facts and other data products to steer some action to a desirable outcome (faster, better, cheaper). Or a machine needs to be given a model that can cause the action with the most positive expected outcome given the input.
The key is to do analysis to take action, not for its own sake.

It may seem hard to believe that projects lacking any of these could exist. For the doubtful reader I can assure you they do. I have been involved in projects for every possible permutation of these three binary conditions (legs). And not just ones missing one leg but two or even all three!

Data, Problem, Actionability

In the literature, the closest I could find to the third leg (actionability) is discussion of model feasibility. Is the model practical to use? A strong example is Netflix not using the (million dollar) Netflix Competition Winner because the model was too complicated to actually implement. And, because the business had marched on and the problem had changed by the time the contest was over a year and a half after starting. [13] For the third leg though, I am talking about the ability to actually make a decision.

Let's examine what each of these eight cases look like and, more importantly, what you might do to get them moving forward.

Case NNN

You have nothing; these projects are merely theoretical discussions. They will not fall like most improperly supported stools because they are not even above ground to begin with. They never get anywhere and they should take very little effort. The work they do require though is very frustrating because you cannot ground any of the issues or debates in reality. If you find yourself in this situation there is no hope for success, if your goal is actually solving a problem. Lots of these projects popped off with the rise in hype about big data, and many resulted in the purchase of expensive software (shelfware).

Case NNY (Data)

If you are given data you can start to do analysis, and it often turns into everybody having different conversations at the same time. These usually resolve to "what would be useful to do" or "what algorithms should be tried" - analysis for its own sake. If you have good data scientists they will find the other two legs of the stool (a problem to solve and a set of actions to take), so long as the bureaucracy does not drive them crazy. They start asking questions, finding out where the pain and value are and how they can avoid one and capture the other. You can see who is fully on board when people start to object to these activities.

Case NYN (Problem)

If you are given a problem, the project turns into an exploration, "where could we find something that could help us", looking for existing data to re-purpose. If you have smart data scientists and allow them to think outside of the box they will devise ways to collect the data you need or find some other data source that can be useful. This still does not spell success; you need a clever angle on how to make this practicable, which is often the hardest part because you have to change people and how they think, act and make decisions. This "soft" part of the problem is often much more difficult than the "hard" technical issues. Young techies are often blindsided by "change management" challenges as, without experience, they are completely unexpected.

Case YNN (Actionability)

If you are given a practicable place where decisions are made you need to determine a useful cost function that will define "best". This is trying to find the problem to focus on and how you can come out of it ahead. This is very hard when things are done anecdotally (e.g., "we do things this way because we do them this way"). Look for money to follow, and ask lots of dumb questions. Listen, take notes, repeat back what you're hearing; keep poking.

Case YYN (no Data)

If you are only missing data, you need a scientist mindset. Brainstorm about what data could we collect, or what can we repurpose. I actually like this one; I know I can find data in this

day and age. I can scrape something together from somewhere while I start to concoct experiments to get the rest.

Case YNY (no Problem)

If you are missing a problem, you are working for somebody wishing to use data science to confirm their thoughts. They dislike you telling them anything they do not believe. It is really hard work changing their minds. Best advice here is to run and not look back.

Case NYY (no Actionability)

If you are missing practicable ways to take action on your insights you can write great papers and other data scientist may learn a lot from them. No business value can be gained but you may make some interesting discoveries. A good enough paper may create a place where somebody will be called to action to make decisions. This is also the case where there are lots of regrets. Out there exist tons of data and tons of problems. People conjuring up new types of practicable actions led to Netflix, Uber and Amazon. This case is where you need obscure thinking -- people who try things nobody else thinks of – people who may not know there is a "box". If you want to hear "Why didn't I think of that?" my best advice is to think of 'that'.

Case YYY

All systems are **go**, proceed to read up on the Top 10 Data Mining Mistakes.
I believe the counter-hype for predictive analytics comes from projects that exhibit one of these scenarios. Results were promised, people misunderstood what they needed and failures thereby abounded. If this has happened to you, don't jump ship; just make sure you get started on the correct footing. Heed these warning and you can make your way into the "slope of enlightenment". Most of these the types of projects described above are not doomed, they just need work in the correct place. Don't think about results before you have data. Don't think about algorithms until you know how the solution can be made practicable. Don't think of buying software before you have a problem to solve. If you have all three, proceed to avoiding the ten mining mistakes, and to a smashing success!

About the Author

Kenny Darrell is a Lead Data Scientist at Elder Research, the US's largest and oldest data science consultancy, where he leads projects primarily for federal government clients. He enjoys all aspects of data science; from problem definition and model construction to presenting the results in data products. He tries to keep a balance between hacking code and power points, and is a fan of learning new things and trying to do old things in new ways. Previously, Kenny was a Control Systems Engineer for the Air Force Research Laboratory and CDI Corp working on image recognition, rare event detection and sensor data fusion.
Mr. Darrell earned a BS in Aerospace Engineering and a MS in Quantitative Analysis from the University of Cincinnati, where his research focused on ensemble methods — combining data mining algorithms to increase performance.

www.elderresearch.com

ELDER RESEARCH
DATA SCIENCE & PREDICTIVE ANALYTICS

National Capital Region
2101 Wilson Boulevard
Suite 900
Arlington, VA 22201

855.973.7673

Headquarters
300 W. Main Street
Suite 301
Charlottesville, VA 22903

434.973.7673

Maryland Office
839 Elkridge Landing
Suite 215
Linthicum, MD 21090

855.973.7673