Get started   Open in app

towards
data science

Follow   557K Followers

Problem identification in data science projects

Three questions to ask yourself to achieve SMART success.
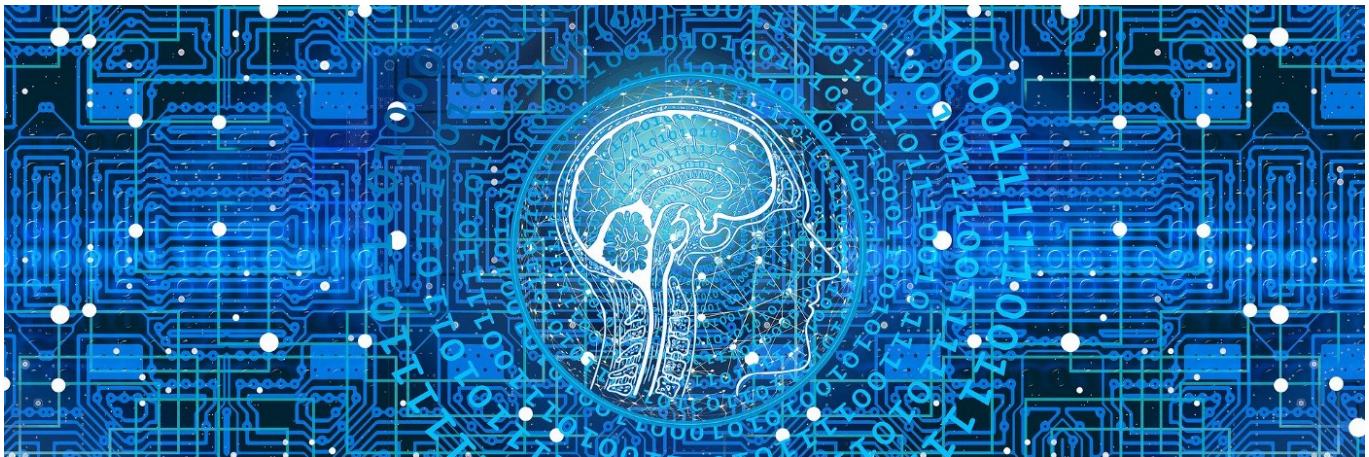
Guy Maskall · Jul 15, 2020 · 3 min read ★

Image by Gerd Altmann from Pixabay

Solve the right problem!

Many of you will be familiar with Kaggle, the home of machine learning competitions. It's the place you can can go to hone your data science skills, right? Wrong. It's the place you can go to hone your machine learning skills. At least that's how it used to be. There'd be a training data set complete with a column of labels and a test data set lacking said labels. You train and improve your best machine learning models, predict the labels on the test set, submit, and wait and watch your name/team rise up the leader board!

Well sorry, that's not data science. Not the full gamut of data science anyway. In fact, really just about the smallest part. A fun part, sure. A key part, arguably. But definitely not the be all and end all. It counts for nothing if your best machine learning model predicts the wrong thing, or simply something not directly useful to the business. Indeed, it's possible to identify a problem (or, perhaps rather, a metric) that's positively bad for your business, as Microsoft's Bing team found out (eventually). Your business problem is to build out a great search engine. A great search engine is one that users want to use more often, right? So you correctly identify the business problem to be one of driving up the number of searches that users make, right? Nope. Your correct problem identification for a search engine is to <u>optimize the finding, not the searching</u>!

With that example above, and I do encourage you to read the linked blog, I'm almost tempted to end this article here, because it's such a great lesson that succinctly encapsulates the key message. Identifying the wrong problem really will set you off in the wrong direction, and any subsequent attempts to optimize will simply get you more wrong faster. But, salutary war stories aside, what constructive techniques are there to guide you towards success?

Be SMART!

A good checklist to run through, when drafting your problem statement, is <u>SMART</u>:

- Specific, not general

- Measurable

- Action-oriented

- Relevant (to the key problem)

- Time-bound

If you find yourself writing a problem statement with phrases such as "improve sales" or "reduce churn", you should catch yourself and review the SMART checklist. Improve by how much? By when? The end of the second quarter? Third? Next year? What is "sales" — units sold or revenue? What is "churn"? When the team managing "pay as you go" customers look at customers who left their system, should it really be called churn if a customer upgrades to a 12 month contract with your company (and so just moves onto

another team's books)? Perhaps your company would actually like more of that sort of thing!

Is SMART enough? Arguably, the Bing team, as highlighted above, thought they were being SMART. Unfortunately, they'd picked the wrong "specific". You can start to deconstruct a SMART problem statement. You can start to talk about business context and stakeholders and constraints and risks and… If you're new to this and your head is starting to spin, I can't blame you. When you feel there are too many factors to think about, I offer you three simple questions that will help orient you:

1. who is it for?

2. what do they want?

3. how will you know when you've got what they want?

For all the fancy talk of SMART measures and context and stakeholders, this is what it's about. That and "keeping it real". There's little point delivering a great solution a week after it was needed — would a good-enough solution delivered in time be better? Did you really think things through when you created a model to predict customer churn a month in advance using features that include information only known a week in advance? Understanding what is wanted, and how you're going to measure your solution in the context of delivering that want, and in a way that is realistic, really is the key to setting up a data science project for success.

About this article

This is the second article of a linked series written to provide a straightforward introduction to getting started with the data science process. You can find the introduction here, and the next article in the series here.

Get this newsletter   newsletter.

Data Science

●|| Medium

About   Write   Help   Legal

Get the Medium app