Scientific Thinking for Business

Our Story

Data science involves lots of investigation via trial and error. The investigations are based on evidence and this is one of the strongest reasons why data science is considered a "real" science.

You will be using a scientific process with your work at AAVAIL. This will help you to organize your work as well as be able to clearly explain everything you are doing to the AAVAIL leadership.

Let's take a look now at some guidance and best practices for engaging with a scientific mindset.

Science is a process and the route to solving problems is not always direct

A common argument made by statisticians and mathematicians is that data science is not really a science. This is untrue, mainly because data science involves a lot of investigations through sometimes chaotic data sets, in search of meaningful patterns that might help in solving particular problems.

Since data science implies a scientific approach, it is important that all data scientists learn to adopt and use a scientific thought process. A scientific thought process of observation, developing hypotheses, testing hypotheses, and modifying hypotheses is critical to your success as a data scientist.

Pulling in data and jumping right into exploratory data analysis can make your work prone to exactly the types of negative issues that plague data science today. There are a number of well-discussed issues revolving around data science and data science teams not living up to promised potential.

- IBM's Seth Dobrin on how to realize the full ROI of enterprise data.
- Learn to deliver fast ROI with data science

At the heart of this problem is the process of communicating results to leadership. It should begin with a meaningful and well-articulated business opportunity. If that opportunity is stated too simply, as say, *increasing overall revenue* then the central talking point for communication is too vague to be meaningful from the data side.

The business scenario needs to be communicated in a couple of ways:
1. Stated in a testable way in terms of data

The business scenario needs to be communicated in a couple of ways:
2. Stated in a clear way that minimizes the influence of confounding factors

Testable hypotheses

There is no one single best way to articulate a business opportunity as a testable hypothesis. In some cases the statement will be intuitive, but in other cases there will be some back and forth with stakeholders and domain experts.

Guidelines for creating testable hypotheses

Become a scientist of the business

Spend a little bit less time learning new algorithms and Python packages and more time learning the levers that make your specific business go up or down and the variables that impact those levers.

Make an effort to understand how the data are produced

If it comes down to it, sources of variation can be explicitly accounted for in many types of models. If the data come from a database you should ask about the process by which the data are stored. If the data are compiled by another person then dig into the details and find out about the compiling process as well as the details of what happened before the data arrived on their desk.

Make yourself part of the business

Do not under any circumstances become siloed. Proactively get involved with the business unit as a partner, not a support function.

Think about how to measure success

When thinking about what course of action might be most appropriate, keep at the forefront of your mind how you will measure business value when said action is complete.

IMPORTANT: Data Science is NOT Business Intelligence. BI analysts serve to derive business insights out of data. There is without a doubt some overlap, but the job of a data scientist is to investigate the business opportunity and solve it.

There is a balancing act to maintain between directly addressing the business need and ensuring that you have thoughtfully studied the problem enough to ensure that you can account for most of the likely contingencies. The scientific method can be of some guidance here.

Thinking scientifically about the business scenario

A major goal of this process is to make the business objectives clear to leadership. Some of these individuals are technical and some are not, so as a good rule-of-thumb get in the habit of articulating the business problem at a level that everyone can understand. Stakeholders and leadership need to know what you are trying to accomplish before you begin work. They also need to be aware from the start what success would look like. Science is an iterative process and many experiments produce results that some might consider a failure. However, experiments that are properly setup will not fail no matter the result–the result may not useful but you have gained valuable information along the way.

Experiments in this context could refer to an actual scientific experiment (e.g. A/B testing) or it could be more subtle. Let's say you work for a company that collects tolls in an automated way, and you want to identify the make and model of each car in order to modify pricing models based on predicted vehicle weight. After talking with the stakeholders and the folks who implemented the image storage solution you are ready to begin. The experiment here has to do with *how you begin*. You may think that there is enough training data to implement a huge multi-class model and just solve most of the problem. If you approach it that way then you are hypothesizing that the solution will work.

For those of you who have done much image analysis work, you could guess that approach would likely result in a significant loss of time. If we take a step back and think scientifically, we could approach the solution from an evidence driven perspective. Before investing a significant amount of time you may try to see if you can distinguish one make and model from the rest before adding more classes. You may want to first pipe the images through an image segmentation algorithm to identify the make of the car. There are many possible ways to build towards a comprehensive solution, but it is important to determine if either of these piecemeal approaches would have any immediate business value.

This might be a good time for a reminder about the steps in the scientific method.

The Scientific Method

It is the process by which science is carried out. The general idea is to build on previous knowledge to in order to improve an understanding of a given topic.

1. Formulate the **question**
2. Generate a **hypothesis** to address the question
3. Make a **prediction**
4. Conduct an **experiment**
5. **Analyze** the data and draw a conclusion

We will continue with an interactive example, but first it is important to note that **Scientific experiments must be repeatable in order to become reliable evidence.**

Question

The question can be open-ended and generally it summarizes your business opportunity. Let's say you work for a small business that manufactures sleds and other winter gear and you are not sure which cities to build your next retail locations. You have heard that Utah, Colorado and Vermont are all states that have high rates of snowfall, but it is unclear which one has the highest rate of snowfall.

Hypothesis

Because the Rocky mountains are higher in elevation and they are well-known for fresh powder on the ski slopes you hypothesize that both Utah and Colorado have more snow than Vermont.

Prediction

If you were to run a hypothesis test Vermont would have significantly less snow fall than Colorado or Utah

Experiment

You hit the NOAA weather API to get average annual snowfall by city. We have compiled these data for you in snowfall.csv.

snowfall.csv

You could use a 1-way ANOVA to test the validity of your prediction, but let's start by looking at the data.

First we read in the data

```
import pandas as pd
df = pd.read_csv("../data/snowfall.csv")
```

Next, subset the data to focus only on the states of interest

```
mask = [True if s in ['CO','UT','VT'] else False for s in df['state'].values]
df1 = df[mask]
```

Finally, create a pivot of the data that focuses only on the relevant summary data

```
pivot = df1.groupby(['state'])['snowfall'].describe()
df1_pivot = pd.DataFrame({'count': pivot['count'],
                          'avg_snowfall': pivot['mean'],
                          'max_snowfall': pivot['max']})
print(df1_pivot)
```

```
         count  avg_snowfall  max_snowfall
state
CO         5.0         37.76          59.6
UT         2.0         51.65          58.2
VT         1.0         80.90          80.9
```

Analyze

There is not enough data to do a 1-way ANOVA. The experiment is not a failure; it has a few pieces of information.

1. There is not enough data
2. There is a small possibility that VT gets more snow on average than either CO or UT
3. Our degree of belief in the conclusion drawn from (2) is very small because of (1)

The notion of *degree of belief* is central to scientific thinking. It is somehow a part of our human nature to believe statements that have little to no supporting evidence. In science the word belief, with respect to a hypothesis is proportional to the evidence. With more evidence available, ideally, from repeated experiments, one's degree of belief should change. Evidence is derived from the process described above and if we have none then we are stuck at the *question* stage and a proper scientific *hypothesis*cannot be made.

The other important side to *degree of belief* is that it never caps out at 100 percent certainty. Some hypotheses have become *laws* like Newton's Law of Gravitation, but most natural phenomena in the world outside of physics cannot be explained as a *law*.

A hypothesis is the simplest explanation of a phenomenon. A scientific theory is an in-depth explanation of the observed phenomenon. Do not be mistaken with the word *theory*, there can be sufficient evidence that your degree of belief all but touches 100%, and is plenty for decision making purposes. A built-in safeguard for scientific thought is that our degree of belief does not reach 100%, which leaves some room to find new evidence that could move the dial in the other direction.

There are additional factors like external peer review that help ensure the integrity of the scientific method and in the case of implementing a model for a specific business task this could mean assigning reviewers for a pull request or simply asking other qualified individuals to check over your work.

Download the full working file here:

comparing-snowfall.py

More resources

- Science for social good
- IBM research home