A MEASURED APPROACH TO DATA SCIENCE

Strike a Balance Between Self-Service and Governance

Author: JD Anderson
Email: janderson@fusionalliance.com

Introduction

Data science is an important field of study as a means of analyzing big data. The success stories of how data science and machine learning provide organizations with new insights that stimulate the growth of customer service, productivity and profitability by leaps and bounds are true.

The initial steps for integrating data science into your organization need not be costly. The focus is often on finding a "data scientist" who will find ways to provide immediate insight to your data. But a more thoughtful, measured approach to incorporating data science into your organization may be more efficient and effective.

What Is Data Science?

It's not easy to pin down the definition of data science. Depending on whom you talk to, the meanings can be radically different.

A strong definition was offered by Jeff Leek in the Marcy 17, 2015 *Simply Statistics* blog (www.simplystatistics.org). Leek said, "Data science is the process of formulating a quantitative question that can be answered with data, collecting and cleaning the data, analyzing the data and communicating the answer to the question to the relevant audience."

Leek's definition is pertinent because it avoids relying on specific concepts,

The Nuts and Bolts of Data Science

Why data science?

It's a big-data world, and data science helps businesses answer questions by examining this vast wealth of information. It's an important asset that empowers business leaders to make informed decisions based on truth.

Why should my business care?

You should care is because data science is critical to remaining competitive. By applying the insights generated about internal and external criteria, organizations can strengthen customer service, improve productivity, increase revenue and more. The data will drive results.

On the flipside, those who don't leverage this process compete at a disadvantage, making decisions based on guesswork.

What does the process involve?

Data science is broken down into five steps. You'll begin with a question or hypothesis and eventually use data to develop answers that inform your business decisions. More details are in this paper.

Who is qualified to perform these functions?

Several skillsets are required, but you be able to assemble a qualified data sciences team from your own, internal resources.

such as big data and machine learning, or specific tools, such as Hadoop, R and Python.

Data science can actually be performed using any number of tools and on many types of data, regardless of the size. The classic data sets used to develop and test statistical processes are actually very small. While big data is often a wonderful resource, in reality one of the first steps will always be to aggregate and/or reduce the data to a smaller size in order to be useful.

The specific types of statistical modeling tools and algorithms are many and varied, and new ones continue to be developed all the time. The most important consideration is not which tool or algorithm is used, but that the correct solution is applied to the problem. Many business questions can be answered through simple summaries, counts or percentages.

The trick is understanding the data enough to decide the best approach and having the skill sets and tools available to implement that approach. This is heavily dependent on process, which brings us to a second reason Leek's definition is so apt -- it emphasizes that data science is a process with multiple parts that all need to work together.

Data Science as a Process

The process of data science can be broken down into five parts.

1. Know Your Use Case

Knowing your use case delivers actionable information about the core needs of the organization, and it's absolutely key to driving the entire process. Knowing your use case defines what data is required, how it will be gathered, how it will be looked at and how the results need to be reported. Data science works best when there is a question or a hypothesis that needs to be answered or proven.

2. Acquire and Clean the Data

The data you acquire can come from inside the company and/or outside the company (public domain data sets, social media feeds, etc.), but it must be driven by the needs of the use case question. Acquiring and cleaning the data is often time consuming and resource intensive, but it is the most important part of the process.

Surveys of data and statistical analysts often state that this step consumes 80-90% of their time, leaving only 10-20% for the actual statistical analysis, but it is absolutely critical that this part of the process is done with great care. Accuracy of analysis is tightly related to the quality of the initial data sources.

3. Understand the Data

Once you have the data, you need to understand what you have. This includes:

- What it is and what it is not
- What it contains that is useful
- What it contains that might be problematic or misleading

Exploratory analysis of the data, i.e., learning the properties within the data that relate and can be applied to the use case question at hand, is important. And information about the source of the data and how it was processed is critical in assessing its usefulness.

Spending time sampling and profiling your data pays great dividends in two key areas, using the data in analysis and being able to assess the validity of the results.

4. Use the Data to Answer the Question

This is the step where the high-end skill sets of a statistical analyst are applied and is often the quickest and seemingly

IN THIS DAY AND AGE, DATA SCIENCE IS A NECESSITY.

SAVVY ORGANIZATIONS USE IT TO SHARPEN BUSINESS STRATEGY AND STRENGTHEN THEIR ABILITY TO COMPETE.

easiest part of the data science process.

When the data has been ingested into an environment by a load process, deep analysis begins. This includes using statistical modeling, machine-learning algorithms, clustering techniques and other appropriate tools to see if the question can be answered.

If the previous steps have all been done well (a clear question exists, the data was properly cleansed and is fully understood), then selecting and implementing the analysis can be fairly straightforward to the skilled statistician.

5. Communicate the Results

It is vital to make the results of this seemingly arcane and mathematically dense process understood at the business level. Interesting and actionable results are of no use if no one knows about them or can understand them.

Resourcing Data Science as an Organization

Looking at the process outlined above, it's clear that finding a single technologist or engineer or mathematician who can accomplish all steps is not likely.

Rather, a data science team of several people who cover all of the necessary skill sets would be the most viable solution. Building such a team is not difficult. Most organizations already have employees with many of the required abilities.

1. Know Your Use Case: Business Analysts and Subject Matter Experts

The business analysts (BAs) and subject matter experts (SMEs) will hopefully already have a firm grasp of the organization's internal data and know the current use case questions being asked by the business. The key here will be for them to expand their horizons to other data sources and wider questions.

They will need to start looking beyond internal systems to other externally available data sources and consider how these might be used to gain new insights into how the organization is relating to the outside world.

Thinking creatively about what other information may be available and how it might be used can lead to even more intriguing use case questions.

2. Getting and Cleaning Data: Database/Data Warehouse Architects and ETL Programmers

Like BAs and SMEs, architects and programmers will need to expand their activities to include both external and highly unstructured data. They will also need to understand the more specific requirements of how a statistical analyst needs the data formatted and delivered.

Fortunately, getting and cleaning data is generally part of these architects' and programmers' everyday lives, and leveraging their knowledge and skills will be critical to providing the analysts the information they need.

3. and 5. Understand the Data and Communicate the Results: Data Analysts, Data Stewards, Report Developers

Data analysts, data stewards and report developers should already have a good handle on the organization's internal data. Like BAs and SMEs, the analysts, stewards and report developers will need to expand their horizons to other data sources. They will already have a history of bridging the communications gap between IT and the business, and that will help the statistical analyst understand the data and the business understand the results.

4. Use the Data to Answer the Question: Statistical Analyst/Data Scientist

Unless the organization already employs statisticians, the skill set of a statistical analyst or data scientist will most likely need to be added. This can either be done by bringing in an outside resource or developing the skill sets internally.

Do not discount your existing data analysts when looking to fill this role. Their current knowledge of the data is a huge running start, and an intermediate level of statistical training will provide them with a variety of new tools to utilize. It will not make them Ph.D. statistician rock stars and they might not fully understand the underlying theories, but not all use case questions require deep statistics to answer, and the practical application of regression modeling and machine learning tools can go a long way.

Conclusion

Data science can provide an organization with new and surprising insights into both internal processes and interactions with the outside world. Take time to build the correct structure and resources to implement data science so it can become an integral and productive asset to the organization.

janderson@fusionalliance.com