information
management

Welcome, Philip | My Account | Log Out

White Papers | Web Seminars | Newsletters | eBooks

- Big Data & Analytics
- Data Management
- MDM & Data Governance
- Infrastructure
- Info Strategy & Leadership
- BI & Data Discovery
- Mobility
- web seminars & white papers
- resource center

Blog

Predictive Analytics or Data Science?

by Steve Miller
FEB 9, 2015 1:53pm ET

Print
Email
Reprints
Comment (1)
Twitter
LinkedIn
Facebook
Google+

I caught up with an old grad school friend a few weeks back. He's a top-notch statistician who's built a successful career working in quants departments of large insurance and health care companies. With little simplification, I'd characterize his role over the last 20 years as a predictive modeling expert. His work is primarily “big iron” -- revolving on Teradata, Oracle and SAS. Besides being a senior statistician, he's also a more-than-capable data integration and statistical programmer.

In the past few years especially, we've had “discussions” on the differences between data science (DS) and statistics/machine learning as disciplines. He's characterized DS as little more than a trumped up moniker marketed by the newest analytics generation to brand themselves with a sexy statistics job title – for work that's indistinguishable from what he's been doing for years.

I've disagreed, not only seeing DS in the service of new data/analytics products, but also arguing the big data, integration, and computation obsessions of data science differentiate it from traditional statistics. More and more, it's data/computation at the forefront, with the likes of Hadoop/Hive/Pig, Spark, R, and Python/Pandas/scikit supplanting SAS in the data warehouse as the go-to tools for DS types. And yes there's a generational divide in usage of the former and the latter. It's millennials vs baby boomers.

This time my friend took a new tack, challenging me to distinguish not statistics from data science, but predictive analytics (PA) from data science. And, I must admit, that gave me pause.

PA & DS both contrast with statistics in their emphasis on prediction over causality and their general use of observational in contrast to experimental methods. In addition, I've always seen predictive analytics as applied statistics/machine learning in the work world, more data-focused and computational than statistics, but less so than data science. When challenged to define the ”point” that separates PA from DS, however, I couldn't, arguing feebly there's a continuum from statistics to data science on a data/computation axis with endpoints “not so much” and “lots” -- and predictive analytics in the middle. That's certainly how my company, Inquidia Consulting, sees it.

It was only after we parted that it occurred to me the differences between PA and DS were clearly manifest in two current Inquidia projects.

The predictive analytics engagement first articulated a classification prediction challenge, after which roughly identifying the source of features from the data warehouse. Early on, there was lots of data extract and munging with SQL and Pentaho, followed by exploratory data analysis and machine learning algorithms in R. In the end, the deliverable was a cross-validated and surprisingly accurate Multivariate Adaptive Regressive Splines model with in-the-future prediction capability. Analytical effort allocation: 25% data/computation, 75% stats/ML.

The data science effort was driven by an association challenge not amenable to collaborative filtering or other off-the-shelf algorithm. Instead, the association was split into a large number of binary classification tasks, which could then be subjected to wide variety of dichotomizers. Indeed, the data/computation work took several mathematical and engineering optimization turns. Ultimately, an appropriate classifier was chosen, a means for training devised, and simplifying assumptions that allowed individual decisions to select the most likely association divined. The prototype was developed in Python, using Apache Spark, Pandas, NumPy, and SciPy, sourcing data from Hadoop/Hive. In the end, a modestly successful model plagued by sparse data was delivered. The production implementation deployment is being written in a combination of Hive and Pig, with Java user-defined functions. Analytical effort allocation: 75% math/data/computation, 25% stats/ML.

Not contented with this soft assessment, I set out to “validate” the observation from a series of analytics LinkedIn job postings. I found eight ads with “data science” in the title, then separated the data/computation verbage from statistics/machine learning. The table below details the cross-tabulation.

Position	Data/Computation	Statistics/Predictive Modeling/ML
1	Experience working with big data platforms like Hadoop, Aster, or equivalent. Experience developing in C, C++, and Java. Experience using SQL, R or SAS analytic tools.	Experience working with real time/near real time analytics. Strong theoretical and practical knowledge of analytical techniques including segmentation creation, time series modeling, change detection.
2	Strong understanding of data engineering, scientific Python, SQL, and data at scale.	Understanding of probability, statistics, machine learning, and data visualization.

3	Outstanding knowledge of and experience in Python, Unix, SQL and NOSQL databases, scikit.learn, virtual environments, javascript, d3, matplotlib, and programming best practices. Well-versed in the data science cycle of identifying problems, retrieving and cleaning data.	Knowledge of Bayesian inference, machine learning algorithms, analyzing data and visually communicating the data to external or internal stakeholders.
4	Experience and proficiency in coding skills relevant for data science, e.g.,R, F#, Python, SQL, Pig, Hive etc.3+ years of experience in Relational database (e.g., SQLServer or PostgreSQL) and NoSQL database (e.g. MongoDB).	Experience in predictive modeling, statistical programming and data visualization. Demonstrated track record and expertise in data mining or machine learning.
5	Familiarity with relational databases and SQL. Fluency in R, Perl, Python, Matlab, SAS, or other tools appropriate for large scale analysis of numerical and textual data.	Significant experience conducting statistical analyses on large datasets. Perform time-series analyses, hypothesis testing, and causal analysis to statistically assess relative impact and extract trends. Experience with data mining, machine learning, statistical modeling tools and underlying algorithms
6	In-depth knowledge on the Hadoop stack (MapReduce, HDFS, Hive, Pig). Experience with scripting languages such as Python.	Understanding of statistics fundamentals. In-depth knowledge on machine learning methods. Understanding of main statistical methods. Understanding of sampling techniques. Understanding of descriptive and inferential statistics concepts.
7	Proficiency in the use of statistical packages like R or Matlab. Proficiency in SQL, Unix/Linux, and a scripting language such as Python, Perl or Ruby. Hands on experience with large datasets and map-reduce architectures like Hadoop.	Experience using machine learning algorithms. Proficiency in statistical analysis, quantitative analytics, forecasting/predictive analytics, multivariate testing, and optimization algorithms.
8	Experience in real-time embedded and networked systems including the proficiency	Hands-on experience in predictive modeling.

	in software languages and tools for embedded and networked distributed systems: Java, Perl/Python, C/C++ Unix / real-time operating system and database management experience is a plus. Experience with large scale distributed programming paradigms like Hadoop.	

The admittedly limited and likely biased sample suggests that data/computation is at least the equal of statistics/ML as a must-have for data science wannabees. Indeed, the data/computation reqs seem more focused than the stats/ML, intimating they have primacy in the early winnowing. It'd be interesting to do a larger examination of characteristics of winning candidates to test if the data/computation “hypothesis” holds up.

Open Thoughts on Analytics

JOIN THE DISCUSSION

(1) Comment

SEE MORE IN

Big Data/Analytics

RELATED TAGS

Analytics,
Predictive Analytics
Comments (1)
In real life, people with all the skills of either column are as rare as hens teeth. Any data science/analytics shop has to piece together skills to get the hoped for results. Someone has to understand what's in the data, someone has to figure out how to move the data from system to system, someone to clean and consolidate the data and someone to run the statistical tools. The skills that aren't listed above are business knowledge and the ability to apply the analytical findings to the problem at hand, or the right brain to apply the findings of your left brainiacs. We usually find at least three people are needed to move from raw data to actionable insights.
Posted by Martha B | Tuesday, February 10 2015 at 8:59AM ET

Add Your Comments: