Sign in

Developer Network

Developing Analytics Solutions with the Data & AI Global Practice

What is the role of a data scientist?

★★★★★

February 23, 2017 by Michele Usuelli // 1 Comments

Share 0 0 0

By Michele Usuelli, Lead Data Scientist

Data Science has been around for decades, but it recently increased in popularity among companies. Although the tools and techniques existed already, there are some changes. Digital technologies generate more data that can drive new advanced analytics use-cases. Also, there are more success stories show-casing the value in data, making companies more keen to invest resources into new solutions. Because of the hype, phrases like "Data Science" and "Big Data" became buzzwords. However, their meaning is loosely defined and it's not entirely clear to many businesses. From a practical perspective, what is data science? Is the role of the data scientist the same as the statistician, or are there new challenges?

The scope of this article is to understand what the role of a data scientist consists in. Depending on the company and on the specific role, there are lots of differences, so it's challenging to come out with a universal profile. However, it's still possible to understand the role from a high-level perspective. Breaking down the core responsibilities of a data scientist, we can have a better understanding of the related skillset.

The starting point of a data science project is "**why**". Why does the company need an advanced analytics solution? Why is it willing to allocate a budget? Why will the project have an impact? To address these queries, the data scientist is capable of understanding the business context, brainstorm solutions, and identify what's valuable. The related skills are business acumen, experience in applying a solution in a specific industry, and soft skills like stakeholder management and listening. Being knowledgeable about the field is definitely useful although not mandatory, since data scientists can interact with subject matter experts to get the information they need. Therefore, the core skill is being capable of collecting business information and defining advanced analytics use-cases accordingly.Also, the data scientist should be able to present the solution and its value, and good presentation skills help in that.

After having defined the target, the next question is "**what**". What do we need to do to solve the challenge? What techniques can we use? What are the main steps from the current situation to the final solution? To address these queries, the data scientist should be able to define the logical steps to build an end-to-end solution. The required knowledge is about statistics, data processing, machine learning techniques, model validation. However, being knowledgeable about the separate steps is not enough as the data scientist needs to be capable of designing an end-to-end solution that every time is different depending on the data and on the target. The main challenges are to

- prepare the data: blend the original data sources, structure the data in the required format
- apply a machine learning model: define the machine learning model addressing the business challenge
- validate the model: according to the context, define a meaningful way to measure the success of the model

Each step requires thinking outside the box and using common sense, in addition to some knowledge about the techniques.Knowing what are the logical step of an advanced analytics solution doesn't imply being able to build it. The final question is "**how**". How can we implement the solution? How can we put the data together? How can we prototype and deploy the solution?

This part is more technical and the skillset is diverse. The main areas are

- dealing with data challenges: the data can be incomplete, have a large volume, have a challenging structure (e.g., text, images). The data scientist should be able to identify and use the tools required by the data. For example, some tools included into the Hadoop ecosystem became popular recently.
- coding: although there are pre-built high-level tools, every data science solution is unique and it will involve some coding. Being able to use specific languages like R and Python is useful, but not always necessary. The core skill is knowing how to code, no matter the language. In presence of large amount of data, it's also useful to be able to design parallel algorithms to scale across large data volumes.
- organising the data into databases: knowing how to store and organise the data, and extract the relevant information. This part is performed using SQL/NoSQL databases although managing them can be out of the scope of a data scientist.
- prototyping: quickly build a good-enough solution that works.
- deploying: put the solution into production.

The technical skills depend a lot on the context, so there is more diversity in the "how" area.

The data science process requires a broad expertise and the data scientist can’t go very deeply into each component of the solution. That’s especially true for data scientist consultants, given that they join new projects where the customer has already a deep knowledge about the context and the tools. To design and build the solution, the data scientist needs to interact with

- Customer stakeholders: to define the scope of the solution and to measure its success, the data scientist should have a conversation with the stakeholders.
- Customer subject-matter experts (SMEs): data science often consists in improving an existing solution using advanced analytics methodologies, so the starting point is to understand the current solution and use it as a starting point. Also, the data scientist needs to know how to interpret the data. To get some help on that, the data scientist needs to work close to the SMEs.
- Academic researchers: although data scientists are experts of machine learning techniques, their knowledge is not as deep as academic researchers. Also, data scientists are focused on bringing value to projects, so they often don’t have enough time to develop complex algorithms. The most common way to work together with academic researches is to use tools developed by open-source community. A good example is CRAN providing with its R 10000+ packages, providing the most cutting edge

statistical tools and machine learning techniques. Also, in larger projects there might be some researchers working together on the machine learning part of the solution.

- Solution architects: the data scientist defines and builds advanced analytics models that are utilised by the solution. Usually an architect designs the overall solution, taking into consideration the business context and the technological infrastructure.
- Data engineers: the "how" part of the engagement is usually the most time consuming and it required a broad range of skills. Data engineers help building and deploying the solution, designing the pipeline translating the data into actions and insights. In some engagement, the data scientists build the prototype and the data engineers put it into production.
- Software SMEs: if the solution integrates other technologies, there might be other people involved, especially solution architects.

This article shows what's common across most of the data scientists and aims to provide more clarity about the role. Depending on the specific case, the skillset can be more detailed and it varies a lot depending on the industry, seniority, level, team. Also, in larger teams there will be people specialised in different aspects of the solution, so it's less important to have a person having the full skillset.

Popular Tags

#ArtificialIntelligence #DataScience AI AML Analytics Apache Spark APS Artificial Intelligence Azure Data Factory Azure Key Vault AzureML Azure SQL Data Warehouse Basket analysis Center of Excellence Centre of Excellence CoE Data Data Science Data Scientist Role Decision forests Event Hubs k-fold Machine Learning Measurability Model Goodness Modelling Power BI PowerShell ML Scoring R Random Projection R Services Scikit SQL Server 2016 SQL Server R Services Stream Analytics Visualizations Whitepaper

Archives

November 2018 (1)
All of 2018 (4)
All of 2017 (3)
All of 2016 (11)
All of 2015 (11)

Tags Data Scientist Role

Join the conversation

Add Comment

Rajwant Kaur

2 years ago

Thank you for sharing this piece of information Michele. I have been going around many blogs to understand this area however couldn't make it clear. With the details mentioned above I will be able to put the things in practice in a better way. Keep guiding and sharing knowledge with us. Thank you once again.

Dev centers

- Windows
- Office
- Visual Studio
- Nokia

Microsoft Azure

More...

Learning resources

- Microsoft Virtual Academy
- Channel 9
- Interoperability Bridges
- MSDN Magazine

Community

- Forums
- Blogs
- Codeplex

Support

- Self support

Programs

- BizSpark (for startups)
- DreamSpark
- Imagine Cup

Newsletter Privacy Terms of use Trademarks