11 Data Science Skills for Machine Learning and AI

As companies realize the power of data, they're tasked with finding data science practitioners with AI and ML skill sets to help them use the data to make better business decisions. Advanced analytics techniques, big data infrastructure and powerful algorithms are providing organizations with the ability to utilize their data for significant business value. The challenge, however, is finding the right skill sets and talent to make better use of the data.

The growth of data science requires a deeper set of skills and capabilities from data science practitioners. While the fundamentals of data science have been around for decades, only recently have the tools and techniques matured to provide the capabilities necessary to accomplish more advanced data analytics, AI and machine learning (ML) goals.

As such, there are many skills for machine learning and AI that data scientists, IT professionals and ML engineers should hone before embarking on AI and ML projects. Some of these skills are general math and statistics skills, while others are more specific to the technology and platforms being used.

Statistics and math

Rather than programming their way to an explanation, data scientists need to derive insights from raw data by using statistics, probability and various functions. Therefore, it should come as little surprise that the fundamental competencies data scientists need are a core understanding of statistics, probability and methods for data derivation.

1. Statistics and probability skills

The cornerstones of deriving insights from data are the mathematical areas of statistics and probability. Advanced levels of statistics are the mainstay of data science and are applied throughout the profession with data visualization, data modeling, identification of correlations, regression, feature transformation, data imputation and dimensionality reduction, among others.

Data scientists require a firm grasp on concepts such as the following:

• mean, median and mode;

• standard deviation and variance;

• correlation coefficients and the covariance matrix

• probability distributions -- Binomial, Poisson, Normal;

• p-value;

• Bayes' Theorem; and

• aspects of the confusion matrix including precision, recall, positive predictive value, negative predictive value, receiver operating characteristic (ROC) curves, Central Limit Theorem, R2 score, Mean Square Error, A/B testing and Monte Carlo Simulation.

2. Multivariable calculus and linear algebra

Linear algebra and multivariable calculus are widely applied by organizations in data science to manipulate and transform data and derive insights. Linear algebra is applied in areas such as data processing and transformation, dimensionality reduction and model evaluation. Core linear algebra topics data scientists need to be familiar with include vectors, norms, matrices, matrix transpositions and manipulations, dot products, eigenvalues and eigenvectors. ML models, particularly deep learning approaches, rely on matrix math and multivariable calculus. It's key that data scientists are familiar with multivariable calculus concepts such as derivatives and gradients, step functions, sigmoid functions, logit functions, cost functions, min/max values, Rectified Linear Unit functions and function plotting.

3. Optimization methods

In addition to core statistics and probability knowledge, data scientists need to understand how to optimize functions, data and algorithms to achieve end objectives. Many ML algorithms that focus on predictive applications achieve their goals by minimizing an objective function and learning the weights applied to the testing data to obtain final predictions.

Other optimizations include the needs for cost and error functions, methods for rapidly determining values from big data and iterating for better performance and accuracy. Key areas data scientists should have knowledge in include cost function and objective functions, likelihood and error functions, gradient descent algorithms and their variants.

Data analysis and wrangling

While being able to derive insights from data is the core of data science, the ability to present that information in ways that can provide value to an organization is equally important. Likewise, analysis of data requires access to sufficient volumes of structured data on which to base those insights. As a result, data scientists also need to have key skills around data visualization, data manipulation, data preparation and data wrangling.

4. Data visualization

Taking the numerical or classification insights from data and presenting them in a way that can be understood by decision-makers is a vital skill for data scientists. Data visualization, which embodies the concept of creating charts, graphs, diagrams and other illustrations of data, is helpful for people who are better with visual information than numerical or quantified data. In many ways, data visualization is a creative aspect of data science, and appeals to those with design-thinking or UX priorities. The most important outcome of data visualization is successfully building a story from the data using visualizations that people can easily understand.

Data scientists should have experience with a variety of data plotting and charting approaches including the following:

• histograms;

• bar and area charts, pie and line charts, waterfall charts, thermometer and candlestick charts;

• segmentation and clustering diagrams;

• scatter plots and bubble charts;

• visualizations of classification space;

• methods for visualization during exploratory data analysis;

• frame and tree diagrams;

• funnel charts, word clouds, heatmaps, video and image annotations;

• map and geospatial visualizations; and

• the use of a wide range of gauges, metrics and measures.

5. Data manipulation, preparation and wrangling

Working with big data requires a lot of plumbing to get access to data in the right quantity and of the right quality. While data engineering is its own field, data scientists

need at least some core knowledge and experience in how to access pools of big data and manipulate it into the shape needed for analysis and processing. According to AI market intelligence firm Cognilytica (where this writer is an analyst), over 80% of AI and data analytics project time is spent on data wrangling and manipulation tasks.

One of the requirements for data wrangling is data access and collection. Data scientists should have some experience in big data access through state-of-the-art and widely accepted data platforms including Hadoop and Spark, as well as more traditional data access approaches including SQL and NoSQL methods. They should also have experience with common databases such as MongoDB and Postgres. In addition, data scientists should be familiar with manipulating the data with data selection, data extraction and methods by which huge data sets are filtered to the relevant portions.

In fact, 56% of data scientist positions list SQL as a requirement, according to a report from Villanova University on the talent gap in data analytics.

Equally as important as data selection is dealing with questionable quality data. Collecting data from multiple data sources can lead to many issues including missing, incorrect, conflicting or potentially biased data. With data programming and collection approaches, data wrangling and preparation requires addressing all of these forms of data imperfections and applying transformations, manipulations, formatting changes and augmentation to improve overall data quality.

Data scientists will need to know how to identify missing or erroneous data, methods of data imputation, approaches for augmenting or enhancing data sets, methods for data transformation and multiplication as needed, identification and treatment for outliers, correction of data types, data scaling and normalization, detection of potential bias in data, data deduplication and data anonymization.

Data analysis and modeling

With the right amount of data at the right level of quality and with the prerequisite understanding of statistics and mathematical approaches for dealing with data, data scientists must apply those skills to build models that an organization can use effectively for analysis and prediction. Data scientists need to know how to create models, build analytics offerings and craft implementations that are put into action by the organization.

6. Data analysis

Using open source as well as commercial offerings, data scientists need to know how to build analytics products that can generate predictive, descriptive and projective results from data. These models are built using existing data to generate results from future data. Data analysis helps the organization apply its knowledge of data to new information to generate better insights and provide stronger decision-making.

These models are generally relevant to the line of business and to organizational needs ranging from recommending products to customers to projecting sales and inventory, from understanding trends in customer or patient data to classifying a wide range of data into categories.

Data analysis is done using a wide range of tools including Excel, big data analysis tools like Hadoop and Spark, commercial analytics offerings such as SAS and MATLAB and open source offerings using R, Python, Java, Julia and other languages. Knowledge of these tools and using them to achieve data analysis objectives is incredibly important to a data scientist's success.

7. ML algorithms, modeling and feature engineering

ML has become the most visible aspect of the modern data scientist's job as it requires them to build models from data using their skills for machine learning methods and algorithms. Data scientists need to understand the vast range of ML algorithms including:

• decision trees, random forests, bagged and boosted tree approaches;

• Bayesian methods;

• k-nearest neighbors;

• support vector machines;

• ensemble methods;

• clustering approaches including k-means, gaussian mixture and principal component analysis;

• Markov models; and

• recurrent neural networks, convolutional neural networks and Boltzmann machines.

Data scientists must ensure they are up to speed with the

innovation taking place in ML algorithms.

Additionally, data scientists need to understand how to perform model evaluation and hyperparameter optimization. This means performing cross-validation and model optimization steps, as well as understanding ROC and learning curves.

Data scientists aiming to have skills for machine learning will also need to know how to use third-party models for their own needs to shorten overall model development time. This means having an understanding of transfer learning and how to enhance models.

Platform and technology proficiencies

By putting all these pieces together, data scientists will also need to be proficient in programming and technology to effectively do their job in their work environment.

8. Programming skills

Much of the technology around data science has evolved over the past few decades. Open source as well as commercial offerings provide a plethora of tools, libraries, frameworks and support functionality across the full lifecycle of the data scientist's job responsibilities.

Data scientists need proficiency in a range of languages including Python, R, Julia and Java-based languages. Python in particular has been the star of the data science world. In 2018, 66% of data scientists reported using Python every day, overtaking R as the most popular language for data science. Julia and other languages are helpful for high-speed and big data processing, and even the use of commercial offerings from SAS and MATLAB are helpful to accomplish a range of data science and analytics tasks, particularly in enterprise settings where the ability to scale up projects is important.

The more technical knowledge and skills data scientists have, the better.

9. Analytical and big data processing tools

Big data and data access technologies and tools are needed to extract meaningful insights from big data. Data scientists should have some knowledge of platforms and frameworks for big data processing including SQL, Spark, Hadoop, Hive and Pig. According to Villanova University's report, 49% of data scientists ranked Apache Hadoop as the second most important skill for a data scientist.

10. Cloud-based platforms and machine learning as a service (MLaaS)

Increasingly, much of the work of data science and ML engineering is done in the cloud. Data scientists should have some experience with cloud-based MLaaS environments from Amazon, Microsoft, Google and IBM, among others, with specific expertise in whichever of those environments are utilized by their organization. Many of these platforms have a wide range of tools, pretrained models and additional support for the full lifecycle of model development and data science activities.

11. Data engineering and manipulation tools

To accomplish many of those data preparation and manipulation tasks, data scientists should have experience using tools for big data manipulation, including open source offerings such as Pandas, as well as those provided by commercial or cloud-based providers. Data scientists should also have experience working with unstructured data such as images, videos, emails and documents that come from different channels and sources.

Growing your skills

At this point you might be asking yourself, "How can it be possible for a single data scientist to know all these things, let alone be proficient?"

No doubt, just reading this list of skills can be overwhelming. The experienced data scientist who possesses all these skills is in extremely high demand, as it's hard to find people who not only have all these skills, but also the experience and knowledge in how to apply them effectively. So, keep all these skills in mind when you're building up your experience but approach data science as a team sport.

The organization needs to make sure that each of these areas of knowledge and expertise are covered. If you can be that one "unicorn" with all these skills and capabilities, you'll find yourself with tremendous job prospects. If you are a member of the hiring organization, be aware that competition will be intense and requested salaries will be high. But if you approach the need for data science skills as a group effort, you'll find greater opportunities to not only meet your team's immediate needs, but also grow your team's capabilities over time.