Advanced Analytics Is More Than Machine Learning

January 2022

The hype around machine learning seems extreme. Between job descriptions, missions for corporate data science teams and training materials you would think that machine learning is the only tool in the data science toolbox. Experienced practitioners probably won't find anything new here. For clarity, I am going to use the term advanced analytics to mean any methods beyond just reporting but usually involves modeling.

The reality is much different. For those readers who have never heard of The Data Science Handbook, I would like to draw your attention to this text. This is not a book review. But I wanted to use this piece of work to remind us of skills that are required in dealing with the "real world", as the author Field Cady puts it. This is for anyone who is asking "how do I prioritize the topics for my professional development?"

Cady considers a wide array of technical skills and theory useful for building comprehensive advanced analytic solutions. He also delves into considerable real world insight and considerable practical knowledge not provided in textbooks.

Take for example so many assumptions about various values being normally distributed, usually for the convenience of the introductory method applied. This text takes time to review exploration methods that are constantly on guard for this overly simplistic assumption and embrace the distributions that are the useful representation of reality. An intuition about signal vs noise and outlier management is well developed.

The acceptance of a custom data science solution requires the customer to evolve their understanding of methods to be applied to their business problem. One of the most important chapters in this text is on communication. The business customer must be thoughtfully educated and maybe persuaded to adapt their thinking. From the business model to the tech stack, the changes may be as dramatic as any other changes adopted in the business. All the presentation techniques here are frequently used by the best in the business but have rarely been documented. As a result this chapter may be one of the most important. "Trust with the business that usually involves communication".

An accepted solution on paper is of no value. A solution that only works on a laptop or "in the lab" is of little value. Twice I have seen situations where data science teams say "I have it working on my laptop, I now need seven more laptops to run our model on all the enterprise data". Frequently these teams had a good theory and implemented a prototype. However, they never had experience beyond the tools for exploring and prototyping.

This text covers all the foundations of computer science that are critical for delivering a business solution. Some of these questions are asked in the planning stages of every successful project. How much time do we have to train our data or determine the parameters of our model? How fast do we have to provide answers when the business queries the model or system? What are the volumes and velocities of the training data or the inference questions?

This text covers the details around the most efficient use of cpu, memory and IO. From hardware to software, a standout practitioner will know how to double the performance without doubling the cost. Then there is the big jump to consider. Can my solution or solution pieces run in containers? Or do parts of my solution have to scale up to distributed processing or a distributed database?

One of his main areas is software and hardware. He considers languages and how to choose. Various ways data is structured and stored. What are the implications of various choices for both clarity and performance. He also covers various database types and appropriate use.

Basically we are moving data between CPU, memory and disk. Some processing is so large it needs to run on more than one computer. This provides a nice foundation to the most important concepts that need to be addressed to handle performance issues.

This foundation in performance is required before moving on to problems that are not possible on a single machine. What if you are asked to develop a clustering model that may take several hours on a single machine but the business needs to produce results run weekly in less than an hour. A large array of our standard techniques can now be distributed across multiple machines or have processing offloaded to a GPU.

Solutions to business problems frequently require a collection of models and transformations to happen in a coordinated fashion. This text covers a variety of methods that have become common requirements on projects and frequently benefit from the fall cost of processing power. Bayesian statistics is now possible because of the falling cost of number crunching. Natural language processing is a common way to create features for many of our models. So many business processes with noise and other uncertainty benefit from approaches in stochastic modeling. Finally, so many business questions involve some necessary time series analysis. Cady's introductions to these topics help early practitioners create a larger toolbox.

Cady does a nice introduction to how ML uses optimization. Optimization has always had important uses beyond statistical learning. Once we can forecast things we frequently have to make various business decisions.

Significant components of this book cover all the basics of machine learning. The greatest value is how the machine learning material is integrated with the rest of the content. This provides considerable context. For example, your project may be to "determine the optimal number and type of staff for each shift". The first part of this problem usually involves the

question of “how much demand will we have for each shift?”. The second part of the problem then usually goes “now that we know demand let's determine optimal staffing levels”. This is an optimization problem. Cady starts to provide a foundation for both the theory and practical considerations involved in advanced analytics projects.

There is much to learn beyond his topics required to get solutions to work in the real world. However, there is a mix of foundational theory and cut your teeth, getting started with details. All of this seems to be delivered in a balanced way. But the tip of the iceberg. Each of these topics is a great starting point for learning and practicing at a greater depth. All covered with appropriately brief passages. Brief because he is touching on so many tools and important perspectives.

There are probably a couple of major topic areas that should at least be noted that were not covered in Cacy’s work. Cady does note that he didn’t cover GPUs, their related power and the opportunities to use that for deep learning. We are also adding linear algebra for GPU, graph theory and graph databases, stream processing and online learning to be at least noted for use in a practitioners toolbox. Discovering at least the benefits of these are left to the reader.

In summary, solutions to business problems require a well formulated solution in the context of the business, a good foundation in the math and a solid technical implementation. This is a considerable amount of change for a business. Leadership in this area requires a considerable range of skills. Cady does a marvelous job of considering a foundation for all the skills required to deliver these types of business capabilities.