Artificial Intelligence(AI), Automation, and its Impact on Data Science

Richard Boire
Senior Vice President
ENVIRONICS ANALYTICS
Toronto, Canada
Richard.Boire at environicsanalytics dot ca

***Abstract*— AI is transforming the very nature of work and data science is no exception. Will the more high-demand technical skills of today be required 10 years from now? How will the data science discipline evolve to meet the business needs of a marketplace with ever-increasing applications of AI?**

Keywords- artificial intelligence, data science, predictive analytics, data mining

I. Introduction

Artificial intelligence seems to be the latest term which is capturing not only the data science discipline's attention but the general public as well as seen by the many articles in much of the mainstream media. The bestselling business author, Tom Davenport, author of "Competing on Analytics" and "Analytics at Work" [1] shared his perspective on artificial intelligence and its impact in the overall economy at the O'Reilly conference in New York City in the fall of 2016 [2]. During the course of this conference, he emphasized the notion that no one is certain of the exact impact to the overall economy other than that there will be great change. Of particular note, he was looking at these changes and their specific impact on the data science discipline.

No one doubts the fundamental changes caused by artificial intelligence (AI) that will emerge both long-term as well as short-term, particularly as the economy continues to move more towards increased automation. In the last twenty years, we have certainly witnessed this increasing level of automation. Think of the typical office in the 1980's. In creating a document that was sent to another department, 3 people were involved which included the creator of the document, an admin assistant to type it, and a courier to transfer it to the appropriate area. With automation, all these tasks can now be done by one person. No advanced artificial intelligence was utilized here. Instead, the birth of the PC and the internet were the fundamental cornerstones in providing this increased level of office automation.

Another example of increasing automation are call center customer service areas which have developed automated voice messaging systems. These systems purport to achieve better customer service, which is arguably debatable, but from the organizational standpoint, they more importantly achieve significant cost savings. Despite these cost savings, most of us have experienced heightened levels of frustration in trying to talk to an actual live person.

With increasing automation, cost effectiveness has been the mantra of most organizations in the last 20 years. This enhanced focus on cost effectiveness has also resulted in organizations exploring options to outsource tasks to countries with a lower standard of living. Over the years, we have seen how outsourcing of tasks has evolved from the more routine mundane tasks to the more highly advanced knowledge-based tasks. I cannot tell you the number of calls that I have received about outsourcing the technical data science skills in building a predictive model. Under this paradigm, my company becomes the "managers" of data science solutions in applying outsourced solutions towards a given problem. However, our business model has always been to both manage and execute the development of data science solutions and at this point in time we have not bowed to the altar of outsourcing. No need to discuss the merits or failures of this approach as this debate is now becoming a mainstream political issue throughout the world.

Certainly no one doubts the transformative effects of AI within our economy. Much of this discussion presents a positive perspective on AI and its particular effects on the economy. In fact, one such article [3] reports multi-trillion dollar growth in economic output. My concern with this study and other projections is that they fail to project the overall impact on jobs. Growth in economic output cannot occur if over 50% of the population are earning minimum wage.

In this paper, however, we focus on the impact of AI within data science solutions but looking specifically at the data scientist who actually builds the solution. Within data science, the paper explores the great opportunities that will be presented by AI. At the same time, AI can now execute tasks that traditionally required more knowledge-based skills. Data Science is no exception but the discipline will adapt to the new needs of the marketplace which in our opinion will ultimately increase the demand for these type of skills. But what does this ultimately mean? .

II. AI in Image and Text Recognition vs. Consumer Behaviour

With the evolution of artificial intelligence becoming more mainstream, one may ask about what will be the impact in automation and in particular to the more knowledge-intensive tasks such as the discipline of data science. Increasing levels of automation that continue to replace labor may result in outsourcing becoming a moot point as the technology costs become far inferior to even using lower-cost labor from third world countries. But what skills will artificial intelligence (AI) replace within the data scientist's arsenal. In theory with AI, choosing the right mathematical algorithm becomes obsolete as the machine determines the right technique. The machine through its artificial intelligence algorithms outputs the solution which can be immediately applied to a given business problem. In AI, we hear such new concepts as deep learning, which utilize the mathematics of neural nets. Keep in mind, neural nets are not new to data science and have been used by practitioners for the last twenty years. In fact, much of the more recent developments in AI have been about enhancing the developments of neural net algorithms and the literature has provided many examples that have yielded superior results to what was developed fifteen years ago. This is particularly relevant in the areas of image and voice recognition. Yet, recognizing that these algorithms are indeed better than in the past, does that translate to improved performance in predicting consumer behaviour? The reason for this question and its specificity towards consumer behaviour is that much of the historical work done in data science has been in the area of marketing and credit card risk where we are dealing with consumers as the records of interest. In my experience, much of this behaviour is difficult to predict with a high level of accuracy due to the high degree of random error or variation. You can read about this in more detail in a white paper [4] that I published on the topic. Under these kind of scenarios, simple solutions trying to generalize overall trends can work quite well with some of the obvious solutions being techniques such as logistic regression or decision trees. But this does not mean that AI should always be discarded when looking at consumer behaviour. In fact, one could simply broaden the definition of AI to look at all modelling techniques which include neural nets as well as the more traditional techniques. Software already exists and a few of the leading providers in this area are able to provide this kind of capability.

III. DOES AI SOLVE ALL ISSUES RELATED TO DATA MINING

A. *Impact of AI*

So where is the biggest impact of AI on data science and what skills of the data scientist can be done more quickly by the machine? If we look at the four step approach to building data science solutions as mentioned in my book [5] we can better understand where AI will have the biggest impact. Each of these four stages (identifying the business problem, creating the analytical file, applying the right analytical technique, implementation and measurement of the solution) are critical in achieving an end solution. Can AI completely replace all the skills that are required in each of these four stages or are there certain stages and tasks where AI will have the most relevance? Let's look at this more closely..

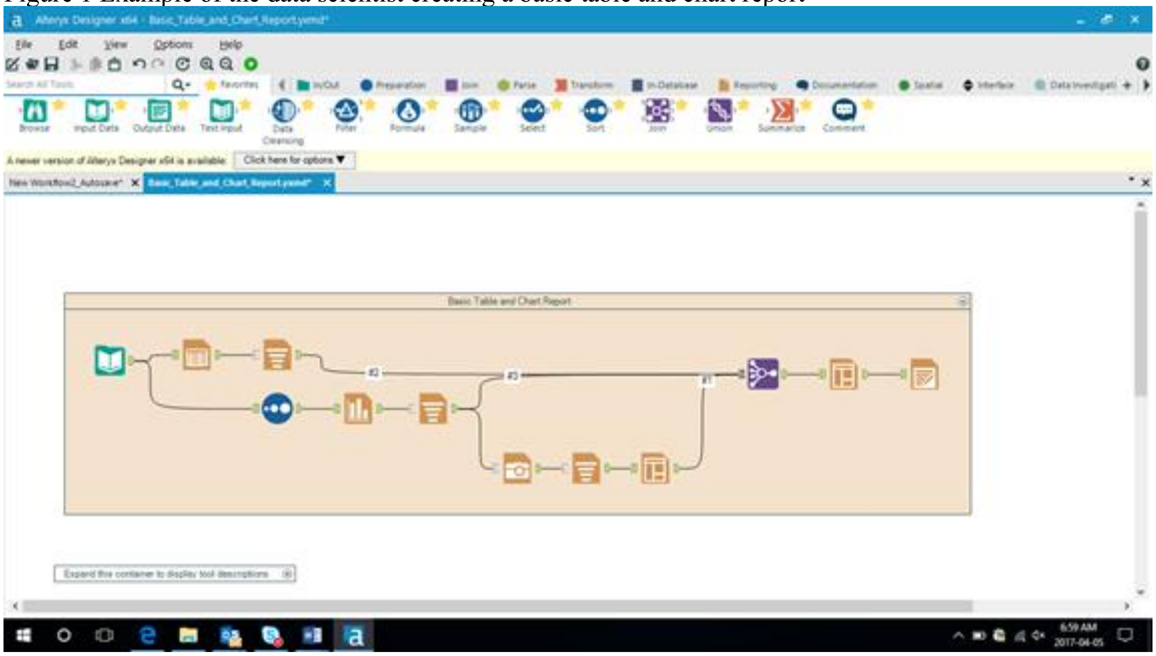
B. *Core Skills of the 2017 Data Scientist today vs. 2027 Data Scientist*

In order to better understand the impact of AI on data science, it is important to understand current roles and responsibilities of the data scientist. In 2017, the specific programming demand would have been for individuals with knowledge in R, Python, SAS, or a number of other more traditionally intensive computer-based languages such as Java or C++. High levels of mathematical and statistically-based knowledge would also be expected to be one of the core skills requirements for a junior data scientist. No question that the key skills requirements would have more of a technical bent rather than the softer skills which might translate to business knowledge and how these technical solutions would be applied in a business setting. The thinking in 2017 would be that the tech skills are the immediate need. Meanwhile the organization's training and internal development programs would build those softer skills of domain knowledge and how to practically apply these solutions within the given business. The development of this hybrid would be ever-evolving with tech skills as the initial foundation complemented by increasing domain knowledge. The more successful hybrids would comprise those data scientists who would ultimately end up in executive-level positions.

Now let's forward to 2027 and what might be the requirements of the junior data scientist. In an age of artificial intelligence and increased automation, the need for coding and programming will be minimized. But what does that mean to the junior data scientist of 2027? Information and data will still be analyzed and the need for an analytical file will remain as one of the core steps within the data science/data mining process and certainly in the development of any data science solutions. Yet, it is the tools which will improve in order to better enable the data scientist to create the analytical file. We are already observing evidence of this through a number of vendors that offer GUI interfaces where the user clicks on icons that represent a certain data function. At the end of this process, the user ends up with a map of all the different processes and tasks which were required to create the analytical file. The analytical file can then be used to develop models or to produce reports and tables. In the example below, the data scientist is simply trying to create a basic table and chart report, Fig. 1.

Note the need for no programming as the tasks and functions in creating both the analytical file and the required reports/tables are represented by drop-down icons.

Figure 1 Example of the data scientist creating a basic table and chart report



C. Creating the Analytical File

This ability to facilitate analytical file creation is and will continue to be a key deliverable amongst software vendors moving forward. As a result, the data scientist needs to spend less time in the creation of programming code and more time focusing on aligning the right data to solve the business problem. Gone are the days when the data scientist would simply extract all the data. In a world of Big Data, access to data is no longer the challenge. Instead, the challenge for the data scientist is to be focused on the data that will be relevant and meaningful in solving a given business problem. For example, if I am building a claim risk model, how relevant is the social media commentary related to insurance policies. The relevance in building a predictive model would only be significant if we can match a high portion of these policies back to their social media commentary. Yet, in another example in looking at insurance fraud, the use of social media might be used to detect patterns of communication where evidence of fraud seems to be most relevant. In both these cases, a stronger level of domain knowledge would help to guide the analyst in what data would be most meaningful.

The human technical skills of the past in creating the analytical file can now be augmented by software with the data scientist now focusing on the business problem, the approach to solving the problem and of course using the right data. The emphasis for data scientists will be on thinking rather than coding. As a result, more business challenges can now be tackled which in the past may not have been addressed due to data limitations or programming resource constraints. But data science in 2027 will still require that the individual have deep knowledge on data and how it can be used for data science projects. For example, the wide array of procedures and tasks that are used when manipulating data are core areas of knowledge to the data scientist. Keep in mind, data manipulation is typically well over 80% of the data scientist's time within a given data science project. This will not change in 2027 but tools will allow the individual to do this quicker thereby allowing more data science projects to be undertaken. Although programming will be less of a need, familiarity with the use of analytics software as well as data will be a core requirement. Better tools allow the data scientist to focus more on thinking through the problem rather than on programming. In other words, what kind of analytical file needs to be created to solve the given business problem. More importantly, no AI algorithm is going to automatically generate the "right analytical file". Instead the reliance is still on the human being or data scientist.

IV. The Increasing Importance of the Data Science Hybrid in an AI World

As the world continues to evolve towards a more Knowledge-Based economy with increasing automation and advancing artificial intelligence algorithms, there will be an ever-increasing need for data science practitioners who can act as the bridge or hybrid between the data/technology/mathematics versus the needs of the business. Certainly, there will always be the need for the hard-core data scientist technician who can write code to improve existing data science solutions which of course would include AI algorithms. But with open source becoming the norm in accessing new techniques and approaches, the data scientist simply has access to a much wider variety of solutions. The traditional analytics vendors such as SAS, IBM, and Microsoft are now competing against open source platforms such as R and Python. Some of these commercial software vendors are establishing partnerships to take advantage of the open source options. With more options available, the business expectations will evolve towards data science hybrids who can use the right data, and the right level of mathematics to solve specific business problems. This will require the ever-increasing use of AI algorithms where the data science hybrid needs to understand the output and to determine what it means in improving overall business results.

At the academic level, colleges and universities will need to emphasize more of those softer skills in training students how to think through a given business exercise. Emphasis on courses with case studies will comprise a large portion of course outlines in any data science discipline. Yet, there will still be the need for the more technical courses for those more hard-core technical programmers who can write code and the algorithms to solve problems which might not be solvable when using a more GUI-based analytics software. Some academic institutions may in effect design two tracts where one tract is geared more towards the creation of these data science hybrids as discussed above and another tract which emphasizes the more technical programming languages. In both tracts(technical vs. hybrid), the data scientist needs to understand data and how to "work" it to create an effective analytical file. On the mathematics side, both tracts need to understand the output. For the data science hybrid, this means the ability to interpret mathematical results such that these results can deliver incremental business results.

.

.

V. CONCLUSION

With academic institutions continuing to evolve their programs alongside the internal training and mentorship programs provided by many organizations, data science as a profession in 2027 will indeed be bright. Many new developments, techniques and approaches will emerge which is a natural outcome of our discipline. The use of the hard core technical data scientist in discovering these new developments complemented by the hybrid data scientist who can apply the relevant learning ensures that we are always looking at new solutions but with a view on how they provide incremental value over the status quo. This scenario will just continue to grow as organizations seek more problem solvers with AI technology representing another option in the data scientist's toolkit.

REFERENCES

[1] Thomas H. Davenport, Jeanne G. Harris, Robert Morison, "Analytics at work: smarter decisions, better results," Harvard Business Review Press, 2010

[2] Thomas H. Davenport, "Only humans need apply: Adding value to the work of very smart machines," session at O'Reilly Conference on AI in New York City, https://conferences.oreilly.com/artificial-intelligence/ai-ny-2016/public/schedule/speaker/251535.

[3] Nicholas Chen, Lau Christenson, Kevin Gallagher, Rosamond Matte, and Greg Rafert, "Global Economic Impacts Associated with Artificial Intelligence," http://www.analysisgroup.com/uploadedfiles/content/insights/publishing/ag_full_report_economic_impact_of_ai.pdf, 2016.

[4] Richard J. Boire, "Is Predictive Analytics for marketers really that accurate,"https://link.springer.com/article/10.1057/jma.2013.8, Journal of Marketing Analytics, vol 1, issue 2, pp.118-123, May 2013.

[5] Richard J. Boire, "Data mining for managers-how to use data (big and small) to solve business challengers," https://www.amazon.ca/Data-Mining-Managers-Business-Challenges/dp/1137406178. Palgrave Macmillan., 2014.