The Data Science Pipeline: Automation and Ethics

Scott Rubey

Computer Science, Portland State University, Portland, Oregon, USA, scrubey@pdx.edu

ABSTRACT

The data science pipeline has historically been labor-intensive and time consuming. It consists of myriad steps, including (but not limited to) data ingestion, cleaning, and modeling. Recent advancements in artificial intelligence and machine learning have eased this burden in many ways, paving a path to faster output and saved dollars. This paper summarizes the steps involved in the data science pipeline, and explores several processes and technologies put forth for the purpose of data science automation.

KEYWORDS

Automation, artificial intelligence, data science, predictive analytics, data mining, data wrangling

1 Introduction

To the outsider, the data science pipeline may appear relatively painless: in order to arrive at a given output, one must simply ingest and process a dataset. The absurdity of this misconception becomes apparent upon examination of the sub-steps required in completing each of these overarching tasks. Once data ingestion is complete, one must clean the data, transform the data (i.e. account for outliers and populate any missing fields), analyze and interpret the data, and finally prepare a model for the client/end user. Two problems make this process all the more arduous for the data scientist. First, the sheer size of the datasets being processed in today's world of big data can be enormous. Tens of millions of records with hundreds of millions of fields can represent a modest domain in some contexts. Secondly, few (if any) datasets are without error in some form. Missing fields, inconsistent formatting and erroneous values are a natural occurrence in most datasets, and such imperfections only grow with the size of the data.

Historically, data scientists were required to perform these tasks manually, a tedious and costly procedure that was prone to human error. Some studies have indicated that 50-80% of a data scientist's time – even today – is tied up in data wrangling, the process of cleaning and processing raw data for the purpose of analysis or modeling later on. Recent technological advancements in artificial intelligence and machine learning have lightened the manual workload substantially in some areas (though the considerable number of tools available to modern users can present obstacles in and of themselves). Automation of this nature provides benefits in several forms. First, less time is required to produce the final model/product, as fewer hours are spent on manual cleaning and other processing of data. Secondly, automation of otherwise complicated tasks makes data science more accessible to those without special training. And third, the resulting model can be produced at a financial discount in the form of fewer human-hours devoted to the aforementioned tasks, benefitting the producer, the client, or both. Indeed, studies have shown that AI automation has the potential to benefit the US economy into the trillions of dollars. [2]

This paper describes the steps involved in the data science pipeline and the technologies being developed for the purpose of automation. Section 2 will provide a description of the steps in the pipeline described above, while Section 3 summarizes several automation technologies put forth in recent years.

2 The Data Science Pipeline

Microsoft Azure's product documentation describes the data science lifecycle in the following terms: 1) Business understanding; 2) Data acquisition and understanding; 3) Modeling; 4) Deployment; and 5)

Customer acceptance.[1] This paper focuses on the subprocesses involved in Step 2, Data acquisition and understanding. These subprocesses are summarized by many in the data science community as "data wrangling." Koehler *et al* [3] describe data wrangling as "the process of preparing potentially large and complex datasets for further analysis or manual examination…" A routine Google search divides this process into the following steps, which are echoed by numerous sources: data extraction, structuring, cleaning, enriching, validating, and publishing. For purposes of this paper, data structuring, cleaning, and enriching will be approached under the umbrella of data "transformation." While data science companies might encourage/deploy proprietary workflows, these steps provide a foundation for understanding pre-modeling phases from a high level. A short description of each step in the data wrangling process follows.

2.1 Data Extraction

Data extraction, simply put, is the process of capturing data from one or more sources. Sources may include (but are not limited to) documents, webpages, images, PDF's, and spreadsheets/CSV files. At this stage, the data is often incomplete, inconsistently formatted, and structured in a way that must be modified prior to analysis. Extracted data may be stored in an intermediate repository for further processing.

2.2 Data Transformation

Data transformation, as mentioned above, is a term that broadly encompasses the structuring, cleaning and enriching of raw data into a state that can be validated and packaged. Those familiar with relational databases are well conditioned toward the notion of structured data. Each record in a relational database exhibits uniformity in terms of the associated data fields and their types. Records can be subdivided into assorted tables that can reference each other through keys. The underlying structure of the database is known as its *schema*.

Raw data, by nature, has no explicit schema. In order to assemble data from varied sources in a fashion that allows for analysis and modeling, the data scientist (historically speaking) has been required to possess extensive knowledge of the problem domain such that they can manually map raw data features to a set of target data features. For example, suppose the problem requires the determination projected property tax revenue from a given jurisdiction; from a collection of public records obtained through local title/escrow companies along with assessment and taxation records from county and state web documents, the data scientist might create a target schema for a structured database of property addresses, assessed values, and millage rates.

Once the raw data has been compiled and appropriately structured, the data must be "cleaned." Data cleaning refers to the process of accounting for missing data fields, identifying and rectifying incorrect records, and ensuring uniform formatting in the target dataset. Extending our example from the previous paragraph, let's suppose the end user requires that all zip codes be represented in their abbreviated five-digit form, but the raw data includes nine-digit zip codes. The data scientist must develop a method of locating the offending fields and transforming them to meet the user's specification. Furthermore, missing fields may be inferred from other data of similar nature or proximity; it the data scientist's challenge to determine when and how to apply such techniques, as misapplication can be costly.

Data enrichment, an optional third step in our data transformation procedure, involves augmenting the previously assembled dataset with further information; for some applications (consumer behavior analytics, for instance), enriching a target dataset in this fashion can provide a broader, more complete picture of a given subject prior to packaging for analysis and modeling.

---

[1] https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle

2.3 Validation / publication

Data validation and publication require little explanation, which is why they appear bundled in the same section. Validation refers to the process of ensuring the efforts made in our prior steps were fruitful, i.e. that the target dataset is complete, accurate, and consistently formatted – in short, that it is ready for further analysis. A simple validation might include type checking of fields (making sure street numbers are integers, for instance), along with checking for inclusion of certain character groupings in text fields (such as Blvd., St., Rd.). Upon validation, the data is "published," making it available for use in subsequent steps in the data science pipeline.

3 Automating Workflow

Section 2 provided a description of the steps involved in the data science pipeline. By no means is this an all-inclusive list, but it provides a foundation for understanding the complex workload one must undertake in order to arrive at a useable model from raw data. Reflecting upon the fact that each of these tasks was undertaken manually for decades, it should come as no surprise that a great deal of research has been performed with regard to task automation. Such studies have targeted very specific steps in the pipeline, such that there now exists a plethora of tools providing assistance in each of the above referenced tasks. This can be a double-edged sword, as data scientists must now possess knowledge of the many tools at their disposal, how to use them, and how one tool may provide better results than a similar one given the dataset in question.

Automation algorithms have only grown more powerful in the modern age of artificial intelligence and machine learning. Koehler *et al* [3] provide a description of one such methodology as it applies, specifically, to data transformation and validation (i.e. "wrangling"). Their technique initiates via a process known as "schema matching," which seeks to address the problem presented by schema-less raw data, as discussed in Section 2. In the schema matching phase of their automation workflow, the algorithm is provided with metadata compiled and managed by domain experts. This metadata may include keywords, data types, or other elements that may be useful in training the algorithm to recognize congruences between raw and target data items. Such recognition takes place through analysis of parsed tokens, which are then scored based on a confidence factor and output accordingly.

Schema matching, in this workflow, is followed by schema *mapping*, which uses the information accumulated in the previous step to rearrange the raw data such that it attains a structured, cohesive state. Prior to executing this step, the AI is provided with a training dataset; the algorithm evaluates this training data, which is effectively a smaller subset of the target dataset. With the information extracted from this training data, the algorithm can then initiate the process of distributing the raw data over a target schema. This process is handled iteratively. Candidate mappings are evaluated and scored according to a confidence metric; upon completion, the algorithm returns the most viable candidate.

The next step in Koehler's data transformation process is referred to as *value format transformation*. This is the phase in which nine-digit zip codes might be reformatted as five-digit zip codes, "St." may be expanded to "Street," and so on. The algorithm that provides this functionality does so by automatically detecting data that requires transformation, thereby triggering the appropriate modifications. As in prior steps, this process is informed through the evaluation of training data, as prepared and supplied by the data scientist.

By providing training data at each step in the aforementioned process, Koehler's team managed to achieve 95-98% accuracy when comparing their AI output with a reference ground-truth dataset.[2] Given these numbers, it is no mystery as to why the data science industry is seeing an explosion of AI-based data cleaning,

[2] There are additional processes included in this method that go beyond the scope of this paper. As of this writing, one may view their paper through IEEE's archive for a complete description.

transformation, analysis and modeling tools. With such an abundant variety of mechanisms to choose from, how can the data scientist begin to determine the appropriate tool for a given project?

Fear not: there are even tools that help us select tools. Biem *et al* [1] detail one such mechanism in their paper "Towards Cognitive Automation of Data Science." The solution depicted therein is an AI for extreme front-end preprocessing of raw data, along with a dynamic repository of algorithms used for automation of subsequent steps such as data transformation. The process is initiated by prompting the user for certain data and preferences relative to their project. This information is provided as input to an analytics repository, which is effectively a search engine for algorithms. (Indeed, their mechanism is capable of combing web-based resources, such as research papers and articles, in order to gain insights regarding the best procedures and algorithms for the given project or dataset.) Selected algorithms are then analyzed by a learning controller, which outputs the top candidates to the user. At each step in the process, the user is allowed interaction with the mechanism for the purpose of providing real-time feedback. For instance, if the user likes certain aspects of a recommended tool but not others, she may provide this as input to the learning controller, which can subsequently attempt to find a superior alternative.

These are simply a few of the AI tools available to the modern data scientist. Specialized tools are becoming widely available for all aspects of the data science pipeline. A comprehensive analysis is very likely impossible and goes beyond the scope of this paper.

4 Conclusion

Worthy of note is that each of the automated processes outlined above relies heavily on input from the data scientist for optimal results to be obtained. Modern automation tools have substantially alleviated the manual nature of the data scientist's workflow, however they have not replaced the data scientist wholesale. In his paper "Artificial Intelligence(AI), Automation, and its Impact on Data Science," Boire [2] hypothesizes that the industry of the following decade will diverge, with one faction focusing on technical advancements, and the other focusing on business knowledge. Indeed, data science was born a hybrid industry, melding algorithms and statistical studies with business profitability and community development. Predictive modeling can now be used to win the Super Bowl or to enact social change, and artificial intelligence has made the field accessible to a broader body of users. If the data science community maintains sight of the fact that the industry was created by people for the betterment of people – and can balance technological advancements with a wealth of human ingenuity – its future will remain bright.

ACKNOWLEDGMENTS

I would like to thank Professor Kristin Tufte, the instructor for this course, for providing advice, ideas, and resources for this project.

REFERENCES

[1] Alain Biem, Maria A. Butrico, Mark D. Feblowitz, Tim Klinger, Yuri Malitsky, Kenney Ng, Adam Perer, Chandra Reddy, Anton V. Riabov, Horst Samulowitz, Daby Sow, Gerald Tesauro, Deepak Turaga. 2015. Towards Cognitive Automation of Data Science. AAAI Conference on Artificial Intelligence. [Online]

[2] Richard Boire. 2017. "Artificial Intelligence(AI), Automation, and its Impact on Data Science." *2017 IEEE International Conference on Big Data (Big Data),* Boston, MA, pp. 3571-3574 DOI: 10.1109/BigData.2017.8258349

[3] Martin Koehler, Alex Bogatu, Cristina Civili, Nikolaos Konstantinou, Edward Abel, Alvaro A.A. Fernandes, John Keane, Leonid Libkin, Norman W. Paton. 2017. "Data Context Informed Data Wrangling," *2017 IEEE International Conference on Big Data (Big Data),* Boston, MA, pp. 956-963, DOI: 10.1109/BigData.2017.8258015

[4] Thomas C.H. Lux, Stefan Nagy, Mohammed Almanaa, Sirui Yao, Reid Bixler. 2019. "A Case Study on a Sustainable Framework for Ethically Aware Predictive Modeling," *2019 IEEE International Symposium on Technology and Society (ISTAS),* Medford, MA, pp. 1-7, DOI: 10.1109/ISTAS48415.2019.8937885