2022

TEACHING DATA QUALITY IN BUSINESS ANALYTICS

Hongwei Zhu
University of Massachusetts Lowell, hongwei_zhu@uml.edu

TEACHING DATA QUALITY IN BUSINESS ANALYTICS

Hongwei Zhu
Department of Operations and Information Systems
University of Massachusetts Lowell
hongwei_zhu@uml.edu

Abstract:

Despite the importance of data quality in business analytics, most degree programs do not offer a course on the topic. It is usually left to the discretion of individual instructors to decide where and how much to cover the topic. In this case study, we describe a course on data quality in a Master of Science in Business Analytics program. Organized in eight modules, the first part of the course covers data preparation and preprocessing. This prepares students with the ability to tackle real datasets in other analytics courses. The second part covers analytics for data quality where algorithms for detecting and resolving data quality issues are covered. The third part addresses large scale and engineering issues of analytics practice where data collection needs to be managed and data quality tasks must be part of the pipeline.

Keywords: Data quality, business analytics, data preparation, curriculum

I. INTRODUCTION

With increasing adoption of data analytics in the digital economy, many colleges offer analytics programs at the undergraduate and graduate levels, typically under the name of data science, data mining, data analytics, or business analytics [Zhu and Wilson, 2017]. It is a common belief that data quality is important when applying data mining and machine learning algorithms. It is also generally known in practice that getting the data ready is a time-consuming step, often taking ~50% of the overall effort of an analytics project. However, few existing programs offer a course to provide a comprehensive and systematic treatment to data quality and data preparation. Students may have a light exposure when a course includes a module on the topic, or they may just have to learn the topic the hard way as they struggle with various data issues in assignments that deal with real datasets. A course has been designed and implemented to address this problem. The course is offered in a master's business analytics program, and it is under continual improvement to incorporate development in technology.

II. DATA QUALITY AND ANALYTICS

Data quality is data's fitness for use [Wang and Strong, 1996; Madnick et al., 2009]. Therefore, in the context of business analytics, data quality is data's fitness for analytics. Various data quality issues can have adverse effects on the outcome and effectiveness of analytics. For example, incomplete datasets can lead to bias in predictive models. The ever-growing amount of data poses challenges that require continued effort to improve algorithm efficiency and to develop new algorithms. At a basic practical level, real world data is not only "dirty" (however it is interpreted), but also usually in a format not readily consumable by analytics algorithms and off-the-shelf tools. As a result, extensive effort must be put on data preparation and cleaning both before and during the analysis and modeling process. Analytics curriculums need to address these challenges to prepare students for their success either as a hands-on analyst or an effective manager.

There are two possible approaches. The first approach is to incorporate data quality topics in different courses. Each instructor decides what and how much to cover depending on their individual course design. For example, a course that requires the use of real-world datasets may cover data preparation and profiling techniques so that students can learn the necessary skills to

turn raw data into a usable format. In contrast, a course that focuses on machine learning algorithms may only use small and precleaned datasets so students would not be distracted by non-algorithm related issues - and there is nothing wrong with this choice. An advantage of this approach is its flexibility. But a disadvantage is the lack of systematic treatment of an important issue in analytics.

The second approach is to create a course dedicated to data quality in the context of business analytics. Students will be exposed to a wide range of issues and methods for addressing the issues. This approach obviously requires a course in the curriculum, which can be considered as a disadvantage given the resources needed. However, in practice, more decisions are data driven and data quality is among the top priorities of Chief Data Officers, a role that can be found in an increasing number of modern organizations [Yang et al., 2014]. Therefore, a dedicated course not only helps students to be effective for various course projects while they are in a degree program, but also prepares them for their future careers.

This case study describes the experience of implementing the second approach in a Master of Science in Business Analytics program at a large national university in the United States.

III. COURSE DESIGN

The course consists of three parts:

- Data quality for analytics. Getting data ready for analytics algorithms is a necessary step; it also helps to identify preprocessing strategies and suitable algorithms to achieve better analytics results.
- Analytics for data quality. This is generally known as data cleaning [1], where algorithms are used to identify and resolve data quality issues.
- Data quality in analytics operations. Methods, systems, and tools for managing data quality in analytics operations.

The course covers these three parts with eight modules:

1. Overview of data quality, types of data, data analytics
2. Crash course on Python, numpy, and pandas
3. Data profiling, exploratory data analysis, shell commands
4. Data transformation, data proximity measures
5. Data merging and integration, deduplication
6. Missing data analysis/imputation; outlier detection
7. Data indigestion, curation, and collection management
8. Systems and tools for data quality and analytics pipeline

Roughly, modules 1 to 4 correspond to Part 1, modules 5 and 6 correspond to Part 2, and modules 7 and 8 correspond to Part 3.

In Module 1, students are introduced to the concept of data quality and are provided with necessary background about data analytics from the perspectives of statistical inference, data mining, and machine learning. The information systems approach to data quality considers a broader range of issues and is more applicable to business analytics. We adapt conventional Total Data Quality Management (TDQM) framework [Wang, 1998] to identify various data quality issues in analytics process. Contemporary issues (e.g., data induced bias in AI) are also covered [European Union, 2019]. Since students enrolled in the class have diverse backgrounds, an overview of various analytics methods is provided in the first module.

Module 2 is to help students who have not prior experience with Python to get up to speed on using Python to work with data.

Module 3 focuses on data profiling and inspection. Data profiling can reveal many characteristics of data, including data quality issues [Abedjan et al., 2015]. Before brining data into analytical tools, it is important to take a "peek" at the data even for simple things such as does the data file contain a header row and what is used for field separator. For this task, Shell commands can be quite effective.

Module 4 covers various data transformations and their impact on proximity measures. Various data preparation methods and their impact on model performance are also discussed [Salvador et al., 2016; Kristof et al., 2017].

Modules 5 and 6 cover data integration [Dong, 2018] and various data cleaning techniques [Ilyas and Chu, 2019]. Additionally, the modules also discuss the impact of various data quality issues on analytics [Blake and Mangiameli, 2011] as well as the impact of various types of cleaning on machine learning performance [Li et al., 2021].

Modules 7 and 8 cover a range of issues of data quality management for analytics operations, such as managing data collections [Roh et al., 2021], continuous data quality monitoring [Polyzotis et al., 2017], data processing tools [Khalajzadeh, 2022], and processing data in various formats.

IV. COURSE IMPLEMENTATION

A module can be delivered in 1-2 weeks. Currently, it is offered in two formats: a 13-week in-person instruction and an 8-week accelerated online delivery.

There is no text for the course. Reading assignments for each module are primarily from journal and conference articles. Through assigned readings, students learn relevant concepts and useful methods.

Problem sets that involve datasets and programming are designed to reinforce concepts and allow students to gain essential skills.

V. DISCUSSION AND CONCLUSION

A challenge we face is that students without analytics background may find the second part difficult. This is partially addressed by placing the course in the second term of the program. During the first term, students typically take courses that cover statistical learning, data mining, and decision analytics. These courses introduce analytics methods such as regression, optimization, clustering, and classification. For sophisticated data cleaning techniques, we give a high-level introduction so that students understand the intuition when they apply these methods.

Another challenge is that some students need more time to catch up on Python programming and shell commands. This is partially addressed by assigning self-paced training courses on DataCamp, which offers time-limited free access to academic institutions.

Overall, the course offers a systematic treatment on data quality in the context of business analytics. It fills a gap in analytics curriculum. In addition to improving the course, we plan to evaluate the course's impact on student learning.

REFERENCES

Abedjan, Z., Golab, L, Naumann, F. 2015. Profiling relational data: a survey. The VLDB Journal, 24, 4, p.557-581.

Blake, R., Mangiameli, P. 2011. "The effects and interactions of data quality and problem complexity on classification", ACM Journal of Data and Information Quality, 2(2), Article 8.

Dong, X.L. and Rekatsinas, T. 2018. "Data Integration and Machine Learning: A Natural Synergy". VLDB 2018, p.2094-2097.

European Union Agency for Fundamental Rights (2019) “Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights,” https://fra.europa.eu/en/publication/2019/data-quality-and-artificial-intelligence-mitigating-bias-and-error-protect

Ihab F. Ilyas and Xu Chu. 2019. *Data Cleaning*. Association for Computing Machinery, New York, NY, USA.

Khalajzadeh, H., Abdelrazek, M., Grundy, J., Hosking, J., He, Q. 2022. "Survey and Analysis of Current End-User Data Analytics Tool Support" in IEEE Transactions on Big Data, vol. 8, no. 01, pp. 152-165

Kristof Coussement, Stefan Lessmann, Geert Verstraeten, "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry". *Decision Support Systems*, Volume 95, 2017, p.27-36

Lee, Y. W., Madnick, S. E., Wang, R. Y., Wang, F., Zhang, H. 2014. "A Cubic Framework for the Chief Data Officer: Succeeding in a World of Big Data." MIS Quarterly Executive 13(1), Article 6.

Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., Zhang, C. 2021. "CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks," in *IEEE 37th International Conference on Data Engineering (ICDE)*, Chania, Greece, 2021 pp. 13-24

Madnick, S. E., Wang, R. Y., Lee, Y. W., Zhu, H. 2009. "Overview and Framework for Data and Information Quality Research", ACM Journal of Data and Information Quality, 1(1), Article 2.

Polyzotis, N., Roy, S., Whang, S. E., Zinkevich, M.. 2017. Data Management Challenges in Production Machine Learning. In Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD ’17). Association for Computing Machinery, New York, NY, USA, 1723–1726

Roh, U., Heo, G., Whang, S. E. 2021. "A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective," IEEE Transactions on Knowledge and Data Engineering, 33(4), 328-1347

Salvador Garcia, Julian Luengo, Francisco Herrera, “Tutorial on practical tips of the most influential data preprocessing algorithms in data mining”. *Knowledge-Based Systems*, 98, 2016, p.1-29

Wang, R. Y. (1998) “A Product Perspective on Total Data Quality Management,” *Communications of the ACM*, 41(2), pp. 58-63

Wang, R. Y. AND Strong, D. M. 1996. "Beyond accuracy: What data quality means to data consumers", *Journal of Management Information Systems,* 12(4), 5–34.

Zhu, H., Wilson, E. V. 2018. "AMCIS 2017 Panels Summary Report", Communications of Association of Information Systems, vol. 43, Article 16.