MWAIS 2022 Proceedings

Midwest (MWAIS)

5-5-2022

The Truth About Numbers: Subjectivity in the CRISP-DM Process

Natalie Gerhart
Creighton University, nataliegerhart@creighton.edu

Russell Torres
University of North Texas, russell.torres@unt.edu

The Truth About Numbers: Subjectivity in the CRISP-DM Process

Natalie Gerhart
Creighton University
NatalieGerhart@creighton.edu

Russell Torres
University of North Texas
Russell.Torres@unt.edu

ABSTRACT

Analytics is undoubtedly changing the efficiency and decision making of businesses. Because analytics is inherently numeric, there is a false sense of security in their certainty. While numbers seem concrete, those that work closely with numbers are quick to explain numbers can be manipulated (Agarwal 2020).

Within data science, there are several subjective biases that an analyst will include in the analysis, inadvertently or purposefully (Agarwal 2020). While some propose machine learning approaches to mitigate uncertainty in analytics (i.e. Hariri, Fredericks, and Bowers 2019), this research suggests there are human biases that are outside of the scope of the machines. The human subjectivity is rarely studied, but has received some attention in niche literature (i.e. Cooray 2021).

To build high quality analytical models, analysts go through the CRoss-Industry Standard Process for Data Mining (CRISP-DM) process. This is the dominant analytics process and it involves five steps: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman et al. 2000). In each of these phases, data analysts make decisions. In the business understanding phase, the analyst (in collaboration with other stakeholders) will be offered a question, which they must then apply background knowledge to determine if the question is answerable. Determining an appropriate question is a challenge itself. Following this, data understanding requires evaluation of available data. The availability of data is subjective depending on the amount of effort an analyst is willing to apply to gather data. Data preparation involves many subjective decisions as an analyst determines how to handle common dirty data issues such as missingness and outliers. This step might also involve feature reduction, which might be a subjective decision. In the modeling phase, the analyst determines the appropriate models to use and also makes subjective decisions about when the iteration phases have achieved their best result. In the evaluation phase, analysts must determine if the model is good enough, which is a vague term to determine if it will meet the acceptability threshold of the business stakeholders. In the deployment phase, the results of the modeling must be made useful to the business.

In this research, we propose a qualitative approach of interviews with data analysts to help determine subjectivity in each phase of CRISP-DM. While it is understood within analyst networks that some decisions are still made by human agents, this is more ambiguous to decision makers. Business leaders should have pristine understanding of the CRISP-DM process to better understand why some models might be misleading in different settings. Our contributions are primarily practical in nature to inform business decision makers. Secondary to this, we expect our contributions will be meaningful to concretely explain subjectivity in the modeling process, which allows analysts to better pinpoint their weaknesses and provide better insights. Finally, these insights can help inform Information Systems educators about key areas to focus on when teaching future analysts about their own subjectivity.

Keywords

Data analysts, subjectivity, modeling.

REFERENCES

1. Agarwal, R. (2020) Five Cognitve Biases in Data Science (and How to Avoid Them). Retrieved November 1, 2021 (https://www.kdnuggets.com/2020/06/five-cognitive-biases-data-science.html).
2. Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. (2000) *CRISP-DM 1.0*.
3. Cooray, S. (2021) The Subjectivity of the Social Network Analyst. *AMCIS 2021 Proceedings*, August 9-13, Montreal.
4. Hariri, Reihaneh H., Fredericks, E., and Bowers, K. (2019) Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges, *Journal of Big Data*, 6, 1.