Constructing, Exploring, and Preparing a Data Set

Marc S. Galli

Walden University

Professor Dr. Mike Collins

DSCI 2010: Data Science Essentials

August 7, 2020

Constructing, Exploring, and Preparing a Data Set

As previously noted, the data science methodology encompasses all steps from initial business understanding, through the analytic approach of data requirements, collection, understanding, and preparation, to final components of modeling, evaluation, deployment, and feedback. The methodology contains two feedback loops and is an iterative process through to completion with feedback. In the data understanding step exists all of the activities which relate to constructing the data set. This step clarifies whether the data collected is actually demonstrative of the problem for which the data science methodology was originally employed (IBM, 2016). Commonly, to make proper assessment within the data understanding step, descriptive statistics must be construed, to include: mean, median, minimum, maximum, and standard deviation.

Within the Nutri Mondo case study, the data science team began to organize, clean, arrange, and construct the data set for the case study by using IBM's Watson Analytics software. Watson Analytics works out of a web browser and allowed the team to input their data and data analysis to then visualize their findings. The team employed an exploratory approach in taking the data set and analyzing which relationships existed amongst the variables. While examining the data, our team found some information within fragments which was determined to have categorically better performance where a wide range of probability-distribution existed. In other words, statistical analysis proved that these fragments of data were not inappropriately influenced by outliers. Our team termed this data, “robust”, and it was decidedly the most useful data in the data set (Laureate, 2018). Consequently our team focused most of their efforts on that data. Additional steps and measures were taken to clean the remaining data which was extremely important to improve clarity and accuracy of Watson Analytics’ visualizations and data integrity as a whole. This incorporated correction of inaccurate and unclear values where “Nan” and blank values in the data set existed. Our team sought correlations and relationships between values in the data set to construct the data set for the case study.

In the data preparation step, unwanted elements, are removed. This process, despite sounding quick and straightforward, is actually the opposite. It actually takes up 70% to 90% of the project’s overall time. 20% to 40% of the project’s time can be saved through the use of automation of data collection and preparation processes within structured data, and for this reason is the preferred method of preparation. The time saved affords data scientists with more time to focus on actually creating models from the data set. As the scientist transforms the data, the data is brought into a state where it is easier to work with. The correction of missing or otherwise invalid values often occurs right alongside removal of any duplicate records in the data set and properly formatting the rows or columns of data. Lastly, within the four-corners of data preparation, data scientists will use domain knowledge in a process termed 'feature engineering' to create characteristics that will help solve the root problem under consideration and enhance machine-learning algorithms. This greatly aids predictive models but if overdone will influence the data science results (International Business Machines, 2016). When performed in moderation, machine-learning functionality is greatly enhanced.

In a management and team meeting, the data scientists in our Nutri Mondo case study determined that, in light of the findings gleaned from the data understanding step of the data science methodology, the team was only interested in seeing connections between specific values within the data set. They noted a significant amount of outliers that functioned as noise, serving only to drown out the relevant data. The team agreed to control for specific variables related to obesity, access to food, economic inequality, ethnicity, and participation in federal food programs, among others. A total of 28 variables were considered. Implementing these changes required a little time but yielded amazing results with a smaller data subset. Multiple visualizations were used to comprehend datum relationships within the revised subset, to wit: Seaborn and Watson Studio's PixieDust. Our Nutri Mondo team witnessed some initial patterns and drew some initial conclusions, in example: there was a negative correlation between obesity and Asian ethnicity, and a positive correlation between Hispanic ethnicity and soda tax. Additional corrections included the positive correlation between African-American ethnicity and obesity as well as a positive correlation between federal program participation and obesity. One such pattern's corresponding real-world actual data metrics meant that 80% of people using food stamps were Caucasian which was equal with the percentage of Caucasians which made up the total population. By controlling for the specific variables, the team set out to validate appropriate distribution of ethnicities within federal program participation statistics. An additional revelation included Hispanic youths' statistical propensity to be on reduced lunch federal programs. More correlations and patterns emerged and the team jointly reviewed each (Laureate, 2018).

In a final review of our Nutri Mondo team and to conclude this discourse, the data science team explored and prepared the data in the data set through empirical review and analysis utilizing the systematic and exploratory approach encompassed in the data science methodology. Although steps and measures were taken to prune the dataset, which ultimately resulted in a significantly trimmed-down data subset, through the elimination of outliers, unclear, blank, and inconclusive values, our team prevailed in visualizing essential patterns which were the underpinnings of data correlations seen within the data set. Nutri Mondo always kept the predominant business understanding from step one in paramount consideration: its quest to aid local communities in gaining enhanced access to healthy food.

References:

IBM. (2016). *Data understanding* [Video file]. Armonk, NY: Author.

International Business Machines. (2016). *Data preparation* [Video file]. Armonk, NY: Author.

Laureate. (2018). *Nutri Mondo*. Laureate International Universities. https://mym.cdn.laureate-media.com/2dett4d/Walden/DSCI/2010/WorldOfDataScience/index.html