What is Data Preparation

Edris Safari

Bellevue University, Nebraska U.S.A.

The adage “garbage in garbage out” is commonly known by most not all data scientists as something to be cognizant of., However, with good data preparation, we may just be able to extract valuable information even with garbage data. Data preparation comes after Data understanding which is preceded by business understanding. According to the CRIPS-DM model, data preparation and modeling are very much interactive. We therefore can surmise the importance of this stage in the process. Indeed, the outcome of the data preparation can determine success or failure of the project in its entirety. We will examine the process of data preparation and identify key factors that play a part in this crucial practice.

Data come in different formats and contents, but at the end of the day, they will be tabulated in rows and columns; where columns are features of the dataset and each row contains data for each feature as an instance of that feature or a set of features. For example in a table with name, age, salary as columns(features) would have a row with ‘Joe’, ‘28’,’$35,000’, and another with ‘nancy’,’32’,’$45,000’, and another thousand rows each with the name, age, and salary of the thousand people in the organization. Most, not all the time, data comes to the data scientist(or data engineer in this case) not so organized, or the data does not reflect a good sample of values, missing and incorrect values(age less than one when surveying adults), outliers((age greater than

80 when surveying young adults). This make cleaning the data imperative to the ultimate success of the project. Once data is cleaned or during this process, the data scientist must be cognizant of the list of features she/he will propose to put in the predictive model. As depicted in the CRISP-DM model, what makes the data preparation and modeling a repetitive process is indeed this process of cleaning the data, selecting the feature, running appropriate mode until an optimal dataset and a good model is selected.

Data cleaning and feature selection are not trivial by any stretch of imagination. In the process of data cleaning, many aspects must be considered. The type of data (numeric, categorical) is important. The very context of data is important in cleaning data. Feature selection is like cleaning in that the features that are selected might make sense in some aspect, and not in other. Sometimes new feature must be created either from existing features or derived entirely based on a combination of feature or deciphered by some other phenomena that the data(cleaned) may expose.

All these uncertainties and use-case specific criteria that make data preparation a scary proposition can be alleviated by the availability of tool, techniques and best practices that are evolving rapidly in the field of data science. For example, a common variable cleaning technique is to handle missing values. In this case, the missing values are categorized as MCAR(Missing Completely At Random), MAR(Missing at Random) and MNAR(Missing Not At Random) [Ref2]. The BKMs for these types of missing data is[Ref2]:

- Listwise and columns deletion
- Imputation with a constant
- Mean and median imputation for continuous variables
- Imputing with distributions

- Random imputation from own distribution
- Imputing missing values from a model
- Dummy variables indicating missing values

In conclusion, data preparation is an arduous, and time-consuming process and can make or break the project. However, just as in regular projects, a sound requirement gathering and design process make development and deployment simple and reliable, so do data preparation and modeling for data science projects.

References

1. Abbott, Dean. **Applied Predictive Analytics**. Wiley. Kindle Edition.
2. How to Prepare Data for Machine Learning and A.I.
3. James, Gareth. **An Introduction to Statistical Learning** (Springer Texts in Statistics). Springer New York. Kindle Edition.