Data Wrangling

A White Paper on Data Cleaning

Introduction

The purpose of data wrangling is to ensure that an organization gets the most accurate analytical and predictive results by fixing errors and cleaning the format of the data. This sounds quite basic, and in theory it is. The problem is we cannot handle cleaning our data with a manual process anymore. We live in a world of BIG and CONTINUIOUS data. So, the wrangling must be done in an intelligent way.

Data wrangling allows companies to make informed decisions quickly to serve the company's and client's needs. Skipping this step can lead to many avoidable consequences. Proper data preparation helps organizations deal with future discrepancies in a more efficient way. The old "cleanup" process was time consuming. But, as the digital world becomes more advanced, and data increases in size, we actually have more room for error. Data preparation greatly enhances companies' ability to ace decision making.

Data Preparation Methods

After the data is collected, wrangling involves looking for problems, including any inconsistencies or missing information, as well as skewed data. It is important to format the data to make sure there are no small errors such as abbreviation errors or differences in the data format. For example, times, dates, and names should all be written in the same way. Raw data can then be transformed into a more usable format with categorizing data sets. This ensures that organizations can separate the relevant information, which involves splitting the data for training and evaluation purposes.

There are a number of ways to find issues, or "dirty data". Both quantitative and qualitative methods are implemented. Quantitative data cleaning includes finding statistical errors through visualization, typically using charts and graphs. A qualitative approach uses patterns and logical rules to find the error. For example, when a number in a data set is higher than the mean by three standard deviations, it is a quantitative error. To find a qualitative error, it is important to pay attention to any inconsistent patterns. For example, if salary earnings do not match up with the location or job title, a qualitative data error is detected.

In order to fix these errors, different methods are used in the wrangling process. From a statistical perspective, it is important to understand the methods that repair quantitative errors. Quantitative errors include missing information, which refers to empty entries in the datasets. Outliers are entries that are much higher or lower than the mean of the dataset. This is usually defined as three standard deviations. Duplicate values are detected when two records have an identical value on the key attribute. Mislabeling attributes is also a common error, but this involves feature engineering, which we will discuss in a subsequent paper.

With the detection of various errors comes solutions used to fix them. The solutions include deletion and imputation. Deletion is the process of getting rid of the missing value. Imputation makes use of the mean, median, and mode.

For categorical values, the mode, or most frequent value, is used in imputation as a dummy variable for the missing value. While deletion is an option for fixing errors, dropping information can affect the dataset. The same is true for outliers. Besides altogether removing the errors from the dataset, mean, median, and mode can be used for the outliers to be imputed.

Several methods can also be used for detection. As mentioned, if a value is three standard deviations above or below the mean, it is considered an outlier. The Standard Deviation Method (SD) is just one form of detection. Another common method is the Interquartile Range Method (IQR), which subtracts the 25^{th} and 75^{th} percentile of an attribute. If the value is outside of the IQR range, it is considered an outlier.

Dataset structures

Data wrangling produces different datasets depending on the situation and use case. An Analytic Base Table, or ABT, is the most common structure in data science. This table is used for machine learning, particularly for finding consistent patterns and outcomes. The ABT uses rows representing separate entities, such as a person or location. The columns include inputs, including attributes about a specific entity.

De-normalized transactions are information for business purposes. Prior calls with customers are recorded and saved. Notes on previous calls help to address current concerns. Information about orders, such as date and product information, also fall under transactional information. A transactional structure is summarized to be used by managers. If a specific entity includes attributes over a period, it is considered a time series. Time series need to be divided into increments for analysis. A document library is used for analysis by text mining. The four types of structures play an important role in the process.

Data Wrangling in Various Industries

Proper data is crucial across industries, but healthcare is an industry where clean data is essential. Hospitals keep private and secure medical information which need to be updated, accurate, formatted in a correct and coherent way. Valid raw medical data also mitigates mistakes in the future. For example, it was found that Columbia University Medical Center's Electronic Health Record (HER) data was incomplete. Fifty-two percent (52%) of patients' information was missing the stage of their pancreatic cancer, the size of their tumor and the duration of medical treatments. The organization needed to check additional systems to see if this information was documented elsewhere. The format of EHRs are often different from hospital to hospital, which means they must be converted into a standard format for further analysis. Accurate medical data assists health professionals study disease trends, allowing them to serve patients more effectively both now and in the future. In addition, this discovery of missing information, made the organization internally aware of ensuring the information was collected consistently in the future.

Another industry that leans heavily on data is the transportation industry. Public transportation, such as buses and trains, require proper data for the number of people using the service, times and duration of service, payments and more. Data wrangling organizes each type of data into specific sets which helps ensure riders have an easy experience booking reservations, paying for tickets and using the transportation services. Overall productivity of the services increase with information gathered, for example number of parking spots needed at particular stations. The transportation industry has improved over the years and one of the main reasons why, is that they have made great use of analytics.

The media, as another focus industry, constantly relies on data to put out information to the public and evaluate their performance in a week, month, or year. Statistics on the number of people that view an article, which stories are trending, and the amount of people who return to a website are all considered in order to improve the quality of the platform or content. Disorganized information can lead to a misunderstanding of the statistics. In order to keep up with the continuous flow of information, data wrangling helps catch mistakes in the data, which increases the performance of the organization. For the information that is relayed to the public, after the data is cleaned, it is often shared in a visualization. More and more mainstream media news outlets are opting to use infographics and charts to display any statistics to the public.

In addition to performance statistics within the media industry, organization data, such as deadlines, publishing dates and archived articles, are beneficial to have both cleaned and accessible.

Challenges

One of the common data wrangling challenges is making sense of the results. It is difficult to figure out the important information of the analysis. To make these observations, many times industry knowledge is incorporated. Although this can be a process, it is overall beneficial to the project at hand.

Data Wrangling Jobs

It is difficult for business analysts to prepare their own data. There is a the lack of visibility into the raw data. Organizations are often not able to hire data scientists because of the expense, due to high demand and an inadequate number of data science professionals. By 2020, it is predicted that the demand for data scientists will rise to 700,000, according to an IBM report. Other organizations lack the knowledge of the vast benefits data wrangling and, subsequently, implementing machine learning brings to organizations.

Data Visualization

Data visualization presents datasets through charts or graphs to help identify new patterns. Visualization techniques can be used to draw conclusions from patterns without using tedious spreadsheets. Faster analysis means businesses can address problems more quickly and find correlations or trends that were previously missed. This gives businesses an edge on improving both customer satisfaction and product quality. Visualizations also help communicate findings to a less technical audience. Visual representation is proven to get messages across faster, and with less confusion. But, the main point to remember, is even if visualizations look aesthetically pleasing, the only effective visualizations are the ones that start with clean and accurate data.

Conclusion

With all of the fancy analytics, machine learning and predictive analytics, we must not forget what all of it is built on. Your raw data. Make sure it is clean. Or your analytics will be telling you a false tale.

References

https://www.springboard.com/blog/data-wrangling/
https://www.forbes.com/sites/traceywelsonrossman/2018/06/25/wrangling-data-in-service-to-digital-marketing/#127d8f94c9ac
https://www.medicaldesignandoutsourcing.com/ai-is-data-hungry-the-challenges-of-data-prep-for-medtech/
http://vis.stanford.edu/files/2011-DataWrangling-IVJ.pdf
https://www.trifacta.com/data-preparation/
https://www.kdnuggets.com/2018/12/six-steps-master-machine-learning-data-preparation.html
https://analyticsconsultores.com.mx/wp-content/uploads/2019/03/Data-Preparation-Challenges-Facing-Every-Enterprise-TDWI-SAS-2017.pdf
https://www.bbntimes.com/en/technology/4-challenges-faced-by-organizations-before-venturing-into-machine-learning
https://www.betterevaluation.org/sites/default/files/data_cleaning.pdf
https://theappsolutions.com/blog/development/data-wrangling-guide-to-data-preparation/
https://www.import.io/post/what-is-data-cleansing-and-transformation-wrangling/
https://www.cs.sfu.ca/~jnwang/papers/sigmod2016-datacleaning-tutorial.pdf
https://arxiv.org/pdf/1904.09483.pdf
https://www.sas.com/en_us/insights/big-data/data-visualization.html
https://infogram.com/page/data-visualization
https://www.elderresearch.com/blog/what-is-data-wrangling
https://www.talend.com/resources/data-wrangling/
https://www.elderresearch.com/blog/what-is-data-wrangling

DATA-CORE SYSTEMS