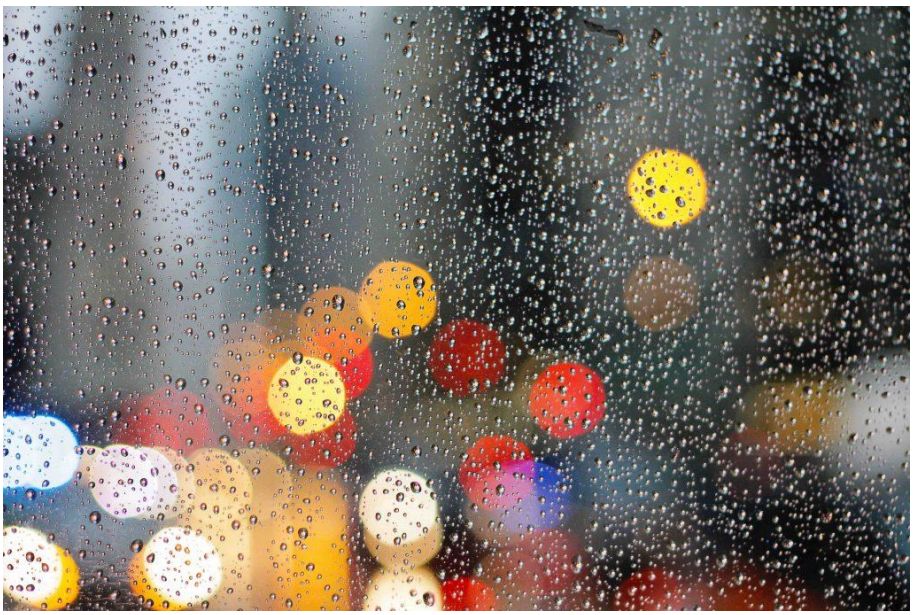
7 ways analytical methods improve data quality

0

By Gerhard Svolba on Hidden Insights 2 APRIL 2019Topics | Analytics Data Management

Data scientists spend a lot of their time using data. Data quality is essential for applying machine learning models to solve business questions and training AI models. However, analytics and data science do not just make demands on data quality. They can also contribute a lot to improving the quality of your data.

Missing value imputation and detection of complex outliers are perhaps the two best-known capabilities of analytics in data quality, but they are by no means the only ones. This post discusses seven ways analytics can improve data quality.



1. Detection of outliers

Analytics plays an important role in detecting outliers based on statistical measures like standard deviation or quantiles. This allows univariate profiling of outliers. Outlier detection can also include methods of cluster analysis and distance metrics. These methods allow you to identify outliers and anomalies in the data from a multivariate viewpoint.

Individual outlier detection with predictive models and time series methods allows you to calculate validation limits and optimal correction values on an individual basis. An overall average might introduce unwanted bias into the analysis, but a within-group average might be a better choice for replacement.

Analytics and data science not only provide methods for profiling and identifying outliers and nonplausible values, but also suggestions for the most probable value to use instead.

2. Imputation of missing values

Analytics can deliver replacement values for missing values in cross-sectional data and time-series data. Imputation methods range from average-based to individual imputation values, which are based on analytic methods like decision trees or spline interpolations for time series. This allows you to use incomplete data in your analysis.

3. Data standardization and deduplication

The identification and elimination of duplicates in a database where no unique key is available for the analysis subjects can be based on statistical methods that describe the similarity between records. These methods provide a measure of the closeness and similarity between records, based on information like addresses, names, phone numbers and account numbers.

4. Handling of different data quantities

Analytics allows you to plan the optimal number of observations for a controlled experiment with sample size and power calculation methods. For small samples or small numbers of events in predictive modelling, analytics provides methods for modelling rare events. For time series forecasting, analytics has so-called intermittent demand models that model time series with only occasional nonzero quantities.

5. Analytic transformation of input variables

Analytical methods can transform variables to a distribution to fit the chosen analysis method. Log and square root transformations are, for example, used to transfer right-skewed data to a normal distribution.

For variables with many categories, analytics provides methods to combine categories. Here, the combination logic for these categories depends on the number of observations in each category and the relationship to the target variables. Examples of these methods include decision trees or weight of evidence calculations.

Text mining allows you to convert freeform text into structured information that analytical methods can then process.

6. Selection of variables for predictive modelling

There are a number of methods for variable selection that allow you to identify a subset of variables with a strong relationship with the target variable in predictive modelling. These methods include simple metrics like R-square and advanced metrics like LARS, LASSO and ELASTICNET.

Many analytical methods allow different options for variable selection in the analysis model itself. Consider, for example, forward, backward and stepwise model selection in regression.

7. Assessment of model quality and what-if analyses

Analytical tools are often designed to assist in model creation and validation. In predictive modelling, for example, it is often important to get a quick initial insight into the predictive power of the available data (this is also referred to as rapid predictive modelling).

These tools also provide measures for rapid assessment of model quality and features for what-if analyses. What-if analyses are especially useful in determining the importance of variables or groups of variables. They estimate the consequences on the predictive power if particular variables are not available