Стандартизация данных (Data standardization)

Разделы: Бизнес-задачи

В широком смысле стандартизация данных представляет собой этап их предобработки с целью приведения к определенному формату и представлению, которые обеспечивают их корректное применение в многомерном анализе, совместных исследованиях, сложных технологиях аналитической обработки.

В статистике целью стандартизации является обеспечение возможности корректного сравнения значений наблюдений, собранных одними и теми же методами, но в различных условиях. Например, в магазине, расположенном в курортном городе, число посетителей в сезон отпусков и в «мертвый» сезон может различаться в сотни раз, поэтому выполнить корректный анализ продаж на таких данных проблематично. Кроме этого, наблюдения, собранные в различных условиях, могут происходить из различных вероятностных распределений с различными параметрами, что также усложняет процесс анализа.

В процессе стандартизации происходит формирование стандартизированных шкал. Стандартизация позволяет устранить возможное влияние отклонений по какому-либо признаку. Стандартизация приводит все исходные значения набора данных, независимо от их начальных распределений и единиц измерения, к набору значений из распределения с нулевым средним и стандартным отклонением, равным 1. В результате формируется так называемая стандартизированная шкала, которая определяет место каждого значения в наборе данных, измеряя его отклонение от среднего в единицах стандартного отклонения. Значения стандартизированной шкалы определяются следующим образом:

$$z_i = \frac{x_i - \overline{X}}{\sigma_x},$$

где x_i — исходное значение признака, $\overline{X}$ и σ_x — среднее значение и стандартное отклонение признака, оцененные по набору данных. В стандартизированных шкалах среднее значение величин $\overline{Z} = 0$, стандартное отклонение $\sigma_z = 1$.

Недостатком стандартизированных Z-шкал является возможность присутствия в них отрицательных значений, что в некоторых случая противоречит логике анализа данных. Отрицательные значения могут исключаться путем дополнительных преобразований.