Different models have different sensitivities to predictor types. How the predictors are entered into the model matter. Data pre-processing is determined by the model. For instance, tree-based models are notably insensitive to the characteristics of predictor data, and opposite of linear regression models which are very restricted.

Definition 1 ((Un-)Supervised Data Processing)**.** *The outcome variable is (or isn't) considered by the pre-processing techniques.*

Definition 2 (Feature Engineering)**.** *The process of how predictors are encoded.*

Examples of Date Encoding

- The number of days from a reference date.
- Isolating the day, day of week, month, year as separate predictors.
- The numeric day of the year (ignoring the calendar year).
- Whether the date was within the school year (opposed to holidays or summer sessions).

0.1 The Chapter's Case Study: Cell Segmentation in High-Content Screening

Contextual Notes From the Reading

- Examination of a sample under a microscope to assess the desired outcome.
- Automated, high-throughput assessment of cell characteristics can produce misleading results.
- Issues related to segmentation of cells resulted in false-positive reading of cell damage.

0.2 Data Transformations for Individual Predictors

Transformations of predictor variables can be used for modeling techniques with strict requirements, or for difficult data used in specific models. The following discusses transformations used in this chapter's case study.

Centering and Scaling

Definition 3 (Centering)**.** *Centering the predictor set is the result of: taking averaging the predictor set, and subtracting this value from all predictor variables. The resulting mean of the set of predictor variables is zero. Let* $\boldsymbol{x}_i = \{x_1, x_2, \ldots, a_n\}$ *be the set of predictor variables, then the centered set of predictor variables is equal to:*

$$S = A - \frac{\sum_{i=1}^{n} \{\boldsymbol{x}_i\}}{n}$$

Definition 4 (Scaling)**.** *Scaling the predictor set is a result of dividing each variable by the standard deviation. This forces all variables to have a common scale, and improve numerical stability of some calculations. The only downside could be a loss of interpretability since the data is no longer in it's original unit.*

Transformations to Resolve Skewness

Definition 5 (Skewness)**.** *The* ***skew*** *of a distribution tells the bias-location of the dataset. For example, right skewed is biased towards larger values, and un-skewed is roughly symmetric. Let* $\boldsymbol{x}_i = \{x_1, x_2, \ldots, x_n\}$*, then the* ***skewness*** *of the set is:*

$$skewness = \frac{\sum (\boldsymbol{x}_i - \bar{x})^3}{(n-1)v^{3/2}},$$

where

$$v = \frac{\sum (\boldsymbol{x}_i - \bar{x})^2}{(n-1)}.$$

You can adjust the skew of a data set by applying a transformation with an equation; ex: log, ln, sqrt, inverse, etc . . . You could also have a system of transformations on different ranges of data.

Box and Cox Transformation Family The use of statistical methods to determine which transform should be used to center a data set.

$$x^* = \begin{cases} \frac{x^\lambda}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases}$$

if $\lambda = 2$, then square transform; $\lambda = 1/2$, then square-root transform; $\lambda = -1$, inverse transform; etc . . .

0.3 Data Transformations for Multiple Predictors

Transformations on groups of predictors on entire sets.

Transformations to Resolve Outliers

Definition 6 (Outliers). *Samples that are exceptionally far from the mainstream of the data.*

When an outlier is suspected:

1. See if the number is feasible / scientifically valid; i.e blood pressure cannot be negative.
2. Verify there haven't been errors in recording the data.
3. Perceived outliers may be due to the skew within small sample sizes.
4. Sampled data could be comparatively clustered compared to the larger set.

Note that tree models are resistant to outliers.

We can minimize outliers by transforming the data's spacial span by dividing the samples by it's squared norm.

$$x_{ij}^* = \frac{x_{ij}}{\sum_{j=1}^{P} x_{ij}^2}.$$

The denominator measures the square distance to the center of the predictor's distribution. This transforms the data as a group, so removing predictors after the transformation can be problematic.

Data Reduction and Feature Extraction

Definition 7 (Data Reduction Techniques). *Reduce data by generating a smaller set of predictors that capture the majority of the information stored in the original variables. New predictors tend to be functions of the original predictors. This is also called* ***Feature Extraction*** *or* ***Signal Extraction****.*

Definition 8 (PC (Principal Components)). *Linear combinations of predictors that capture the largest possible variance from a data set.*

Definition 9 (Loading). *The weight associated with each PC in a Principal Component Analysis, that signifies the importance of each factor on the dataset.*

Definition 10 (PCA (Principal component analysis))**.** *Principal component analysis creates a set of uncorrelated PCs. PCA seeks predictor-set variation without regard to further understanding of model objectives. It can result in irrelevant sets towards the main objective — an unsupervised technique.*

The jth PC can be written as:

$$PC_j = \sum_{k=1}^{n} (a_{jk} * P_k)$$

Where a_k represents the loading, and P_k represents the jth predictor. Loadings close to zero represent predictors with low contributions to the variability in the data set.

0.4 Dealing with Missing Values

0.5 Removing Predictors

0.6 Adding Predictors

0.7 Binning Predictors

0.8 Computing