Data Profiling

Data Profiling *is the process of examining the data available from an existing information source (e.g. a database or a file) and collecting statistics or informative summaries about that data.*

The goals of data profiling are:

- Get **insights** about the data;
- Assess **data quality**;
- Identify **data problems**;
- Recognize opportunities.

Analysis can be classified into two categories:

- **Univariate** - analysis of a single variable;
- **Multivariate** - analysis of multiple variables.

We usually consider four **perspectives** of analysis:

- **Granularity** - the level of detail of the data and precision;
  - e.g. if the data was collected daily, or hourly, per city or per country;
- **Distribution** - the distribution of the data;
  - e.g. normal, uniform, skewed, etc;
- **Sparsity** - analysis of the coverage of the data;
  - e.g. how many missing values;
- **Dimensionality** - the number of variables.

False predictor is a variable that is **highly correlated** with the target variable, but it is not available at the time of prediction.

Removing false predictors from the model is important to avoid **overfitting**, and improve hte model performance.

Granularity

Granularity *is the level of detail of the data.*

- The **finer the granularity**, the **more detailed** the data;
- Data at a finer granularity can be **aggregated** to a coarser granularity; **Aggregation** are made through:
  - **Discretization and composition** for numeric data;
  - **Concept hierarchies** for symbolic data - **taxonomies**.

Distribution

Distribution *is the distribution of the data.*

- Understand data **centrality** and **dispersion**;
- Identify **missing values** and **outliers**;
- **Central Tendency** - mean, median, mode;
- **Histogram** - a graphical representation of the distribution of the data;
- **Discrete Distributions**:
  - **Uniform** - all values are equally likely;
  - **Bernoulli** - binary variable;
  - **Binomial** - number of successes in a sequence of `n` independent experiments;
  - **Poisson** - number of events occurring in a fixed interval of time or space;
  - **Hypergeometric** - number of successes in a sequence of `n` draws without replacement from a finite population of size `N` that contains exactly `K` objects with that feature;
- **Continuous Distributions**:
  - **Normal** - the most common distribution;
  - **Exponential** - the time between events in a Poisson process;
  - **Log-Normal** - the logarithm of the variable is normally distributed;
  - **Chi-Square** - the sum of squares of `k` independent standard normal random variables.
- **Outliers** - values that are far from the rest of the data;

- `X` is outlier if `X < ` μ ` - n`σ or `X > ` μ ` + n`σ, where:
    - μ - mean;
    - σ - standard deviation (square root of the variance);
    - `n` - number of standard deviations.

- **Measuring Dispersion**:
  - **Variance** - the average of the squared differences from the mean;
    - $var(D) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$;
  - **Standard Deviation** - the square root of the variance;
    - $std(D) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$;
  - **Interquartile Range** - the difference between the third and first quartiles;
    - $IQR = Q3 - Q1$;
  - **5-Number Summary** - the minimum, first quartile, median, third quartile and maximum;
  - **Boxplot** - a graphical representation of the 5-number summary.

- **Skewed distribution** - when the mean is not equal to the median.

Sparsity

> ***Sparsity** is the percentage of missing values.*

- Only **present** values are considered in the analysis;
- **Scatter Plot** - a graphical representation of the data - allow the identification of subspaces of the data domain;
  - Allow the identification of dispersion and outliers;
  - Allow the identification of **correlation** between variables;
- **Heat Maps** - graphical representation of matrices, which each cell is colored according to its value;
  - Always **symmetric**, since the correlation between `X` and `Y` is the same as the correlation between `Y` and `X`;
  - The **diagonal** is always **1**, since the correlation between `X` and `X` is always `1`;

Dimensionality

Dimensionality *is the number of variables.*

- **Extrinsic Dimensionality** - the number of variables in the data - `dim(D) = d`;
- **Intrinsic Dimensionality** - the number of variables that are relevant to the analysis - `k` (`k < d`);
- `n « d` - the **number of records** is much smaller than the **dimensionality** - data tends to be **highly sparse**;
- **Curse of Dimensionality** - the **number of records** required to **cover** the **data domain increases exponentially** with the **dimensionality**;
- **Hughes Phenomenon** - the **accuracy** of the **model decreases** with the **dimensionality**.

Similarity Measures

Similarity measures concern with **quantifying how alike two records are**. They show **higher values for more similar records**.

- **Euclidean Distance** - the most common distance measure. It is the **square root of the sum of the squared differences** between the values of the attributes;
  - $d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$
- **Manhattan Distance** - the sum of the absolute differences between the values of the attributes. Also known as **block distance** or **Minkowski distance**;
  - The **generalization** of the Euclidean distance - uses the axes of the space to define the distance;
  - Interesting for **categorical non-ordinal** variables;
  - $d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$;
- **Chebyshev Distance** - the maximum absolute difference between the values of the attributes;
  - **Do not weight** the attributes differently;
  - $d(x, y) = \max_{i=1}^{n} |x_i - y_i|$;

- **Cosine Distance/Similarity** - the cosine of the angle between the two vectors;
  - Adequate when the **magnitude of the vectors is not important**;
  - Bounded measure - always between `-1` and `1`;
  - $sim(x,y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$;
- **Contingency Table** - a table that shows the **joint frequency distribution** of two binary variables;

	y	$\neg y$
x	α	ϵ_1
$\neg x$	ϵ_0	β

- $sim(x,y) = \frac{\alpha}{\alpha+\epsilon_0+\epsilon_1}$
- $d(x,y) = \frac{\epsilon_0+\epsilon_1}{\alpha+\epsilon_0+\epsilon_1}$
- **Jaccard Similarity** - the ratio between the size of the intersection and the size of the union of two sets;
  - More useful in presence of **asymmetric (unbalanced) variables**.
  - $sim(x,y) = \frac{|x \cap y|}{|x \cup y|}$;

  Dummy variables are used to **represent** the **presence** or **absence** of a **categorical variable**.