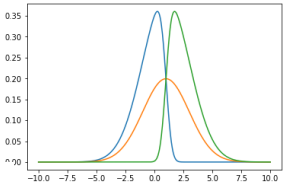
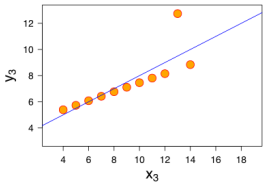
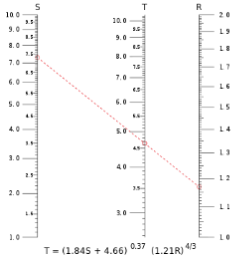
Feature Engineering

Feature Engineering







- Transform Variables
- Extract Features
- Create New Features

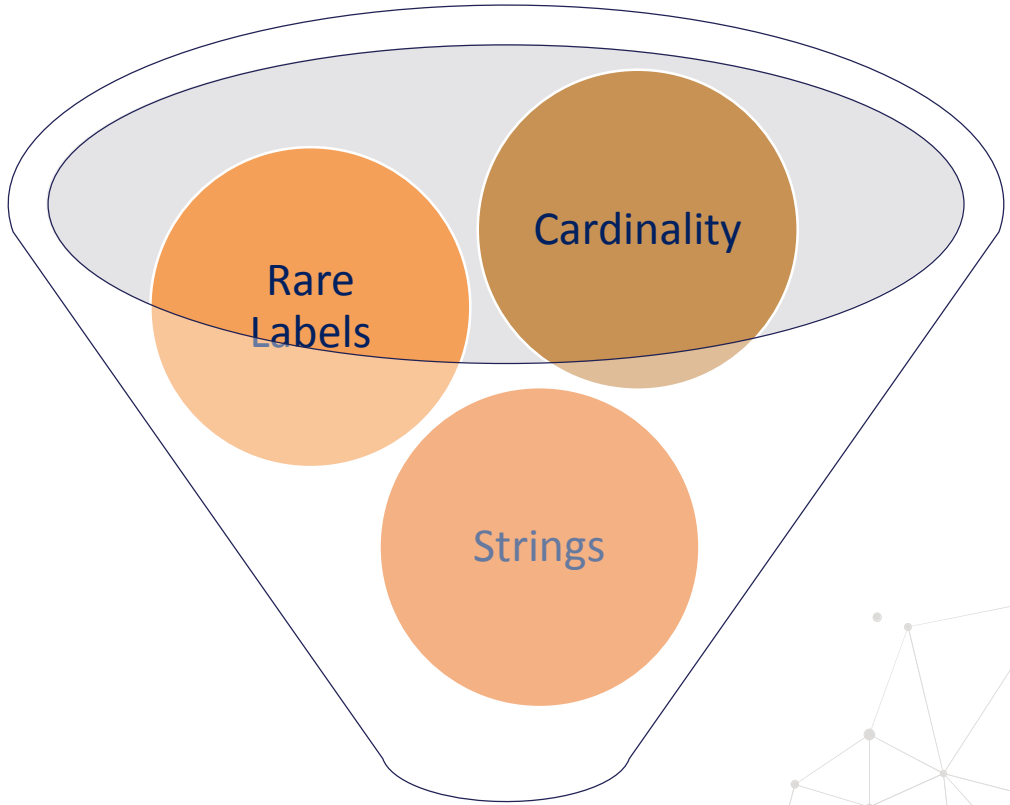
Missing Data

- Scikit-learn and other libraries can’t work with missing data

Missing Data Imputation Techniques

Numerical Variables	Categorical Variables	Both
Mean / Median Imputation	Frequent category imputation	Complete Case Analysis
Arbitrary value imputation	Adding a “missing” category	Adding a “Missing” indicator
End of tail imputation		Random sample imputation

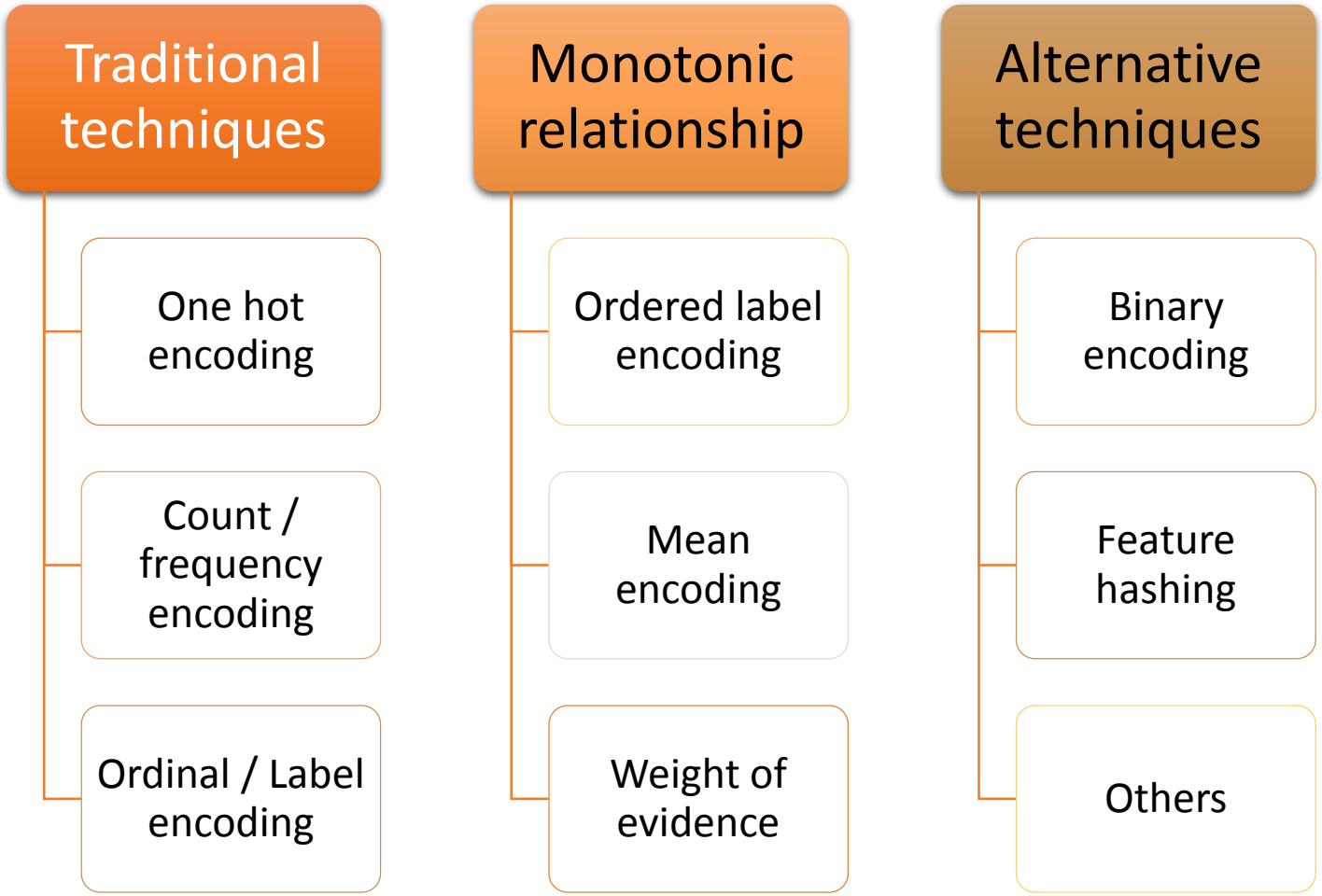
Categorical Variables



Over-fitting

Particularly in decision trees

Categorical Encoding Techniques



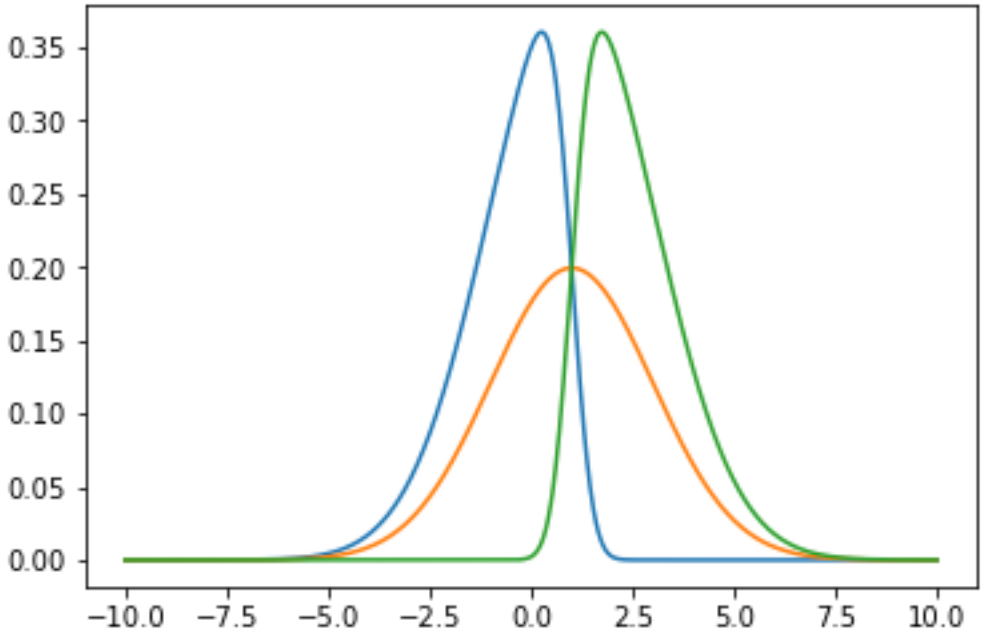
Encoding Techniques: Rare labels

- ✓ One hot encoding of frequent categories
- ✓ Grouping of rare categories

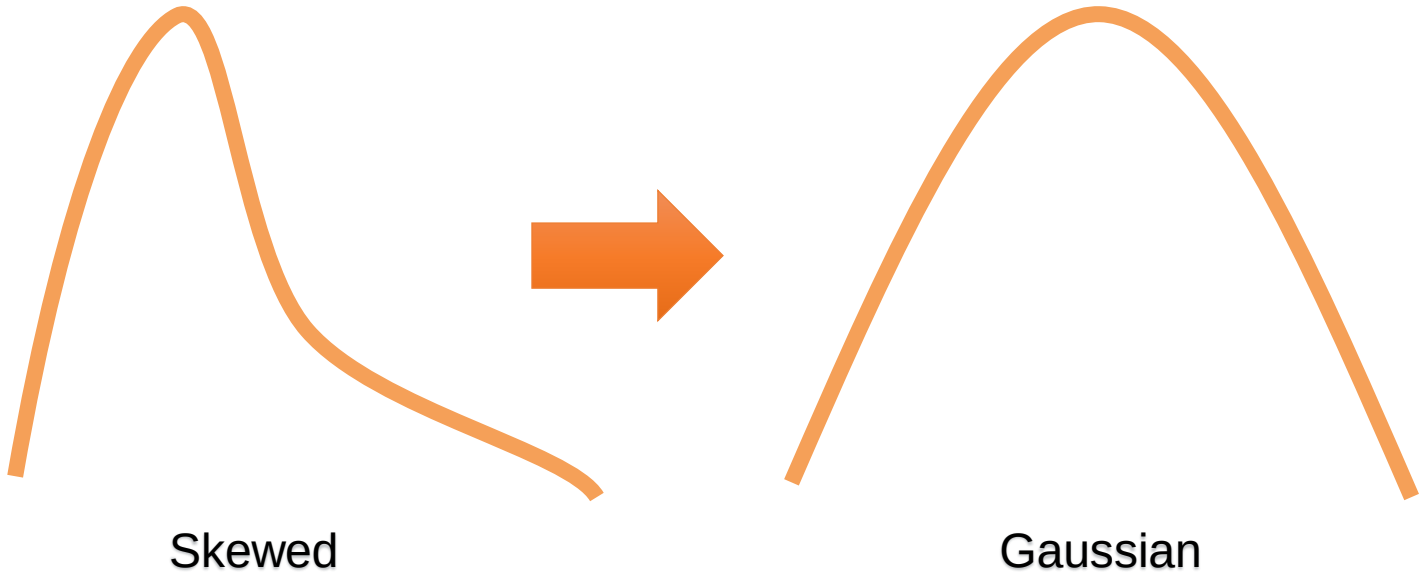
Particularly important for model deployment

Distributions

- Some models make assumptions on the variable distributions



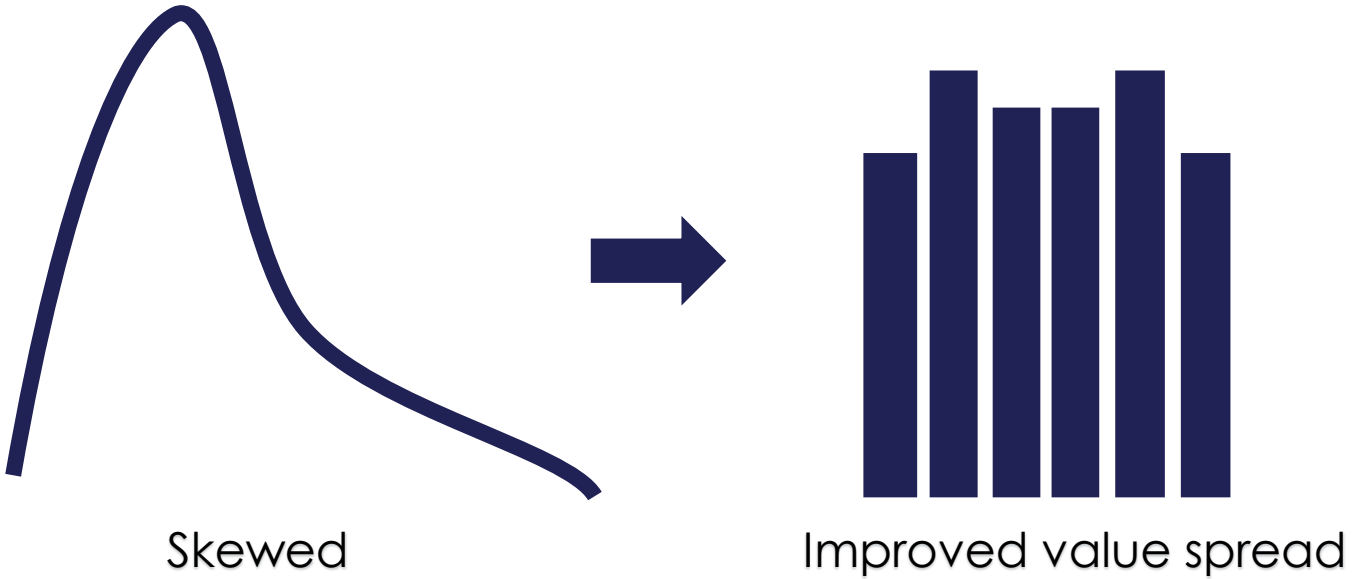
Mathematical transformations



Variable transformation

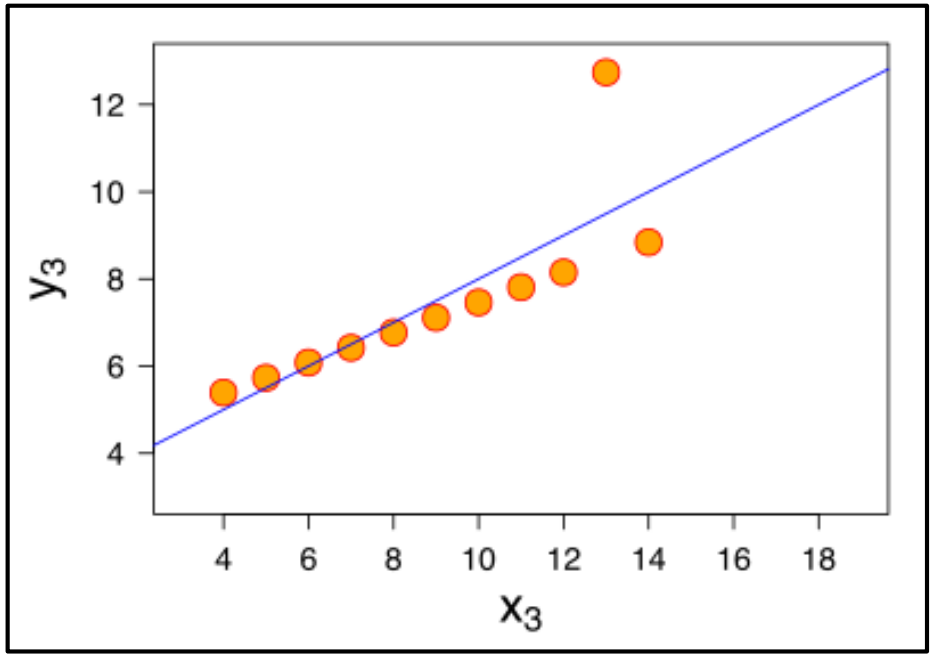
- Logarithmic
- Exponential
- Reciprocal
- Box-Cox
- Yeo-Johnson

Discretisation



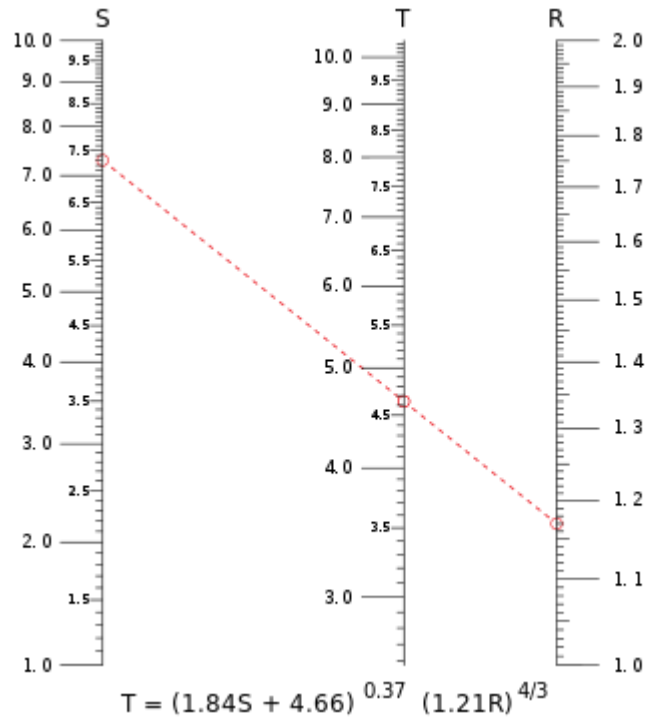
- Unsupervised
  - Equal-width
  - Equal-frequency
  - K means
- Supervised
  - Decision Trees

Outliers



- Discretisation
- Capping / Censoring
- Truncation

Variable Magnitude



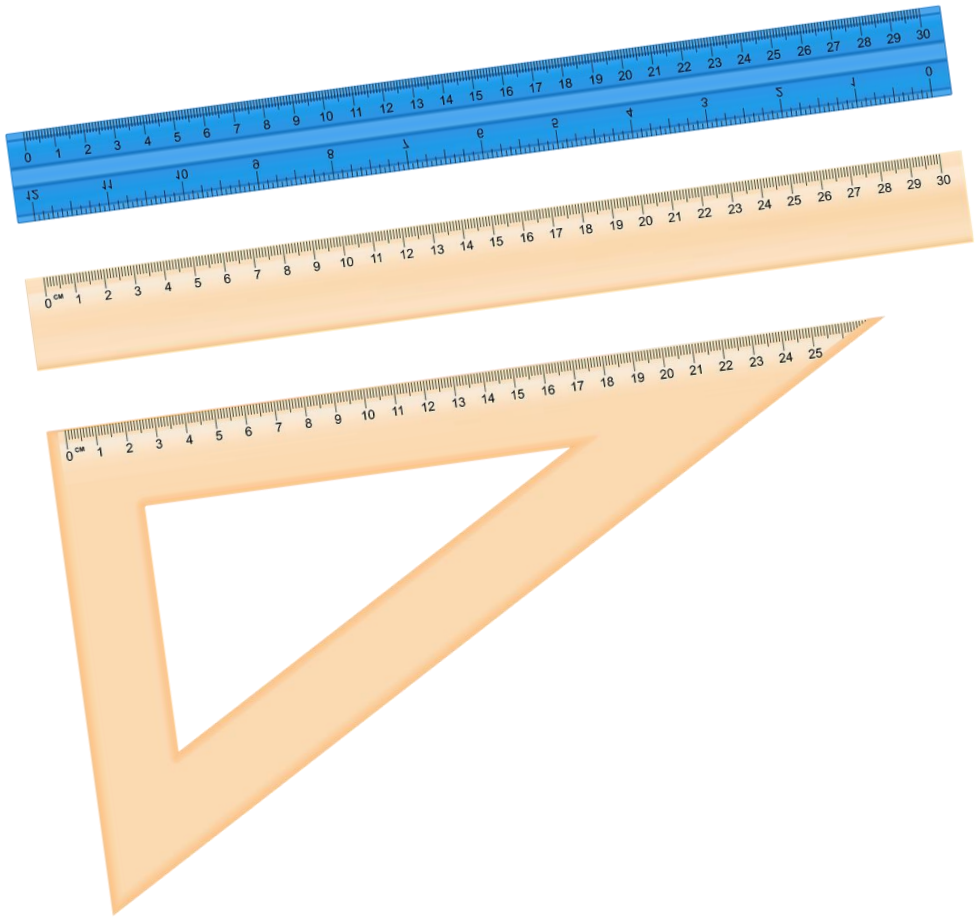
The machine learning models affected by the magnitude of the feature:

- Linear and Logistic Regression
- Neural Networks
- Support Vector Machines
- KNN
- K-means clustering
- Linear Discriminant Analysis (LDA)
- Principal Component Analysis (PCA)

Machine learning models insensitive to feature magnitude are the ones based on Trees:

- Classification and Regression Trees
- Random Forests
- Gradient Boosted Trees

Feature scaling methods



Scaling methods

- **Standardisation**
- Mean normalisation
- **Scaling to maximum and minimum**
- Scaling to absolute maximum
- Scaling to median and quantiles
- Scaling to unit norm

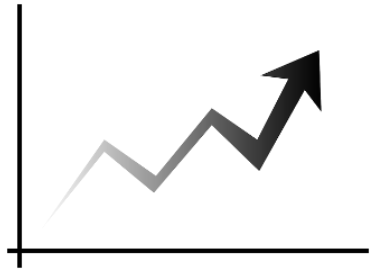
Datetime Variables

- Day, Month, semester, year
- Hour, min, sec
- Elapsed Time
  - Time between transactions
  - Age

Text

- Characters, words, unique words
- Lexical diversity
- Sentences, paragraphs
- Bag of Words
- TFiDF

Transactions and Time Series



Aggregate data

- Number of payments in last 3, 6, 12 months
- Time since last transaction
- Total spending in last month

Geo Data

- Distances

Feature Combination

- **Ratio**: Total debt with income ➔ Debt to income ratio
- **Sum**: Debt in different credit cards ➔ total debt
- **Subtraction**: Income without expenses ➔ disposable income

Train In Data

Thank you

www.traininindata.com