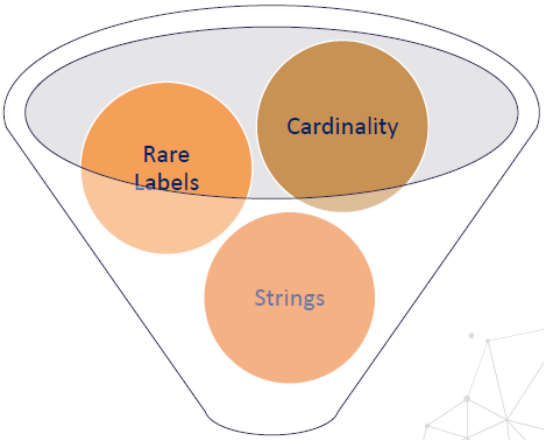
Feature Engineering

- Transform Variables
- Extract Features
- Create New Features

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

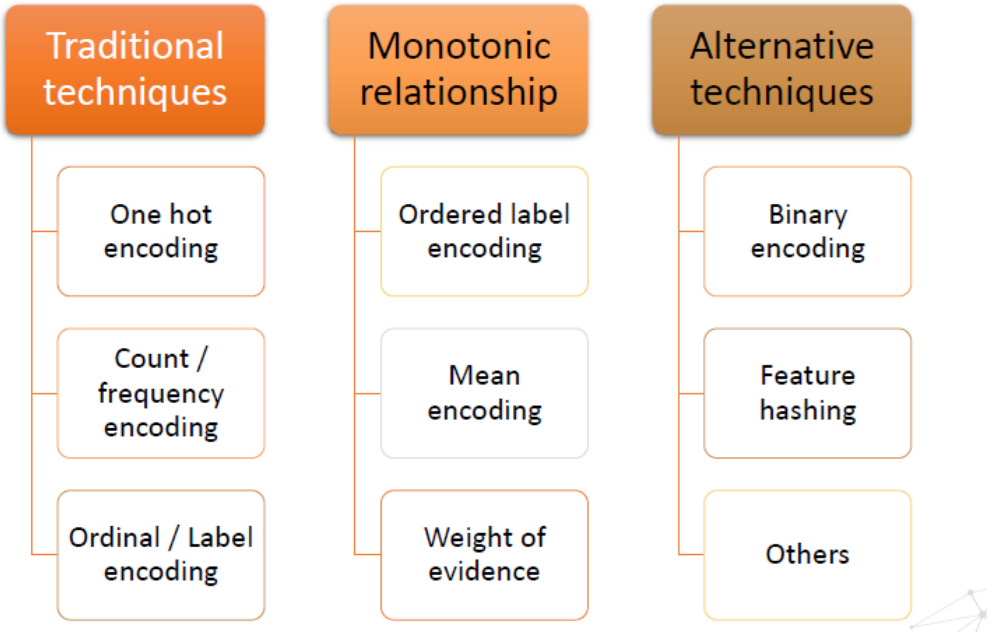
Categorical Variables



Over-fitting

Particularly in decision trees

Categorical Encoding Techniques



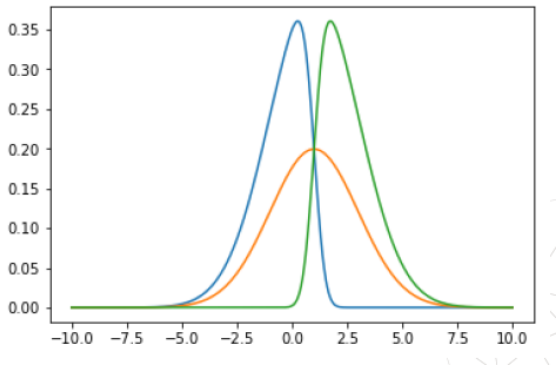
Encoding Techniques: Rare labels

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

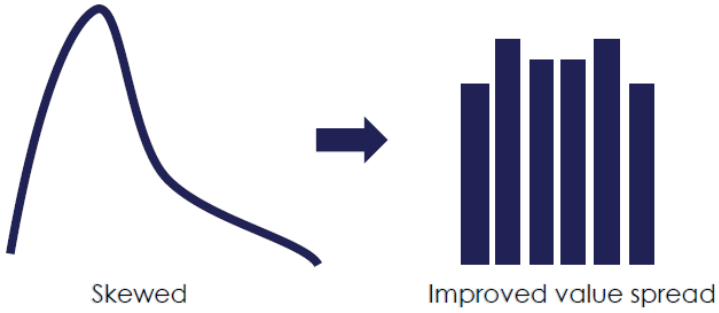
Particularly important for model deployment

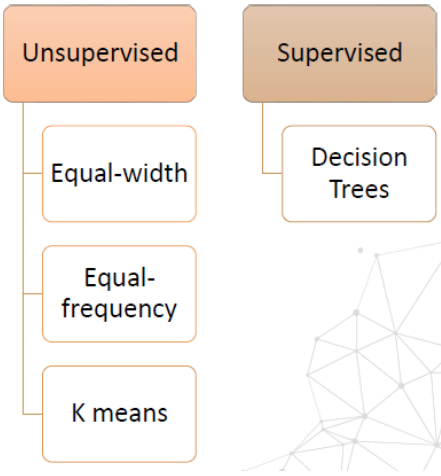
Distributions

- Some models make assumptions on the variable distributions

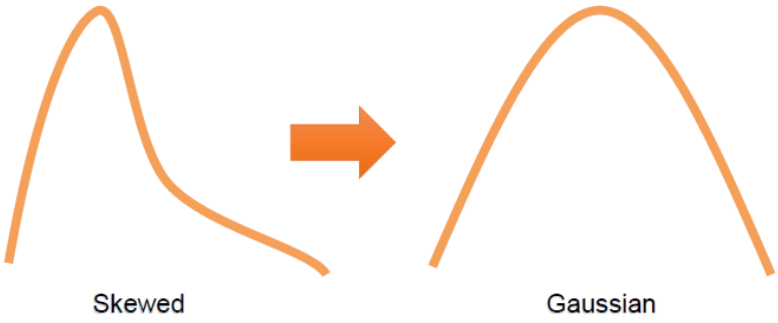


Discretisation





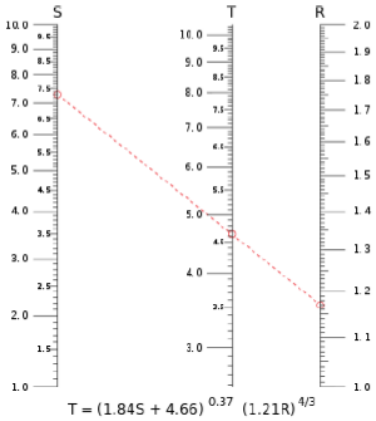
Mathematical transformations



Variable transformation

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

Variable Magnitude



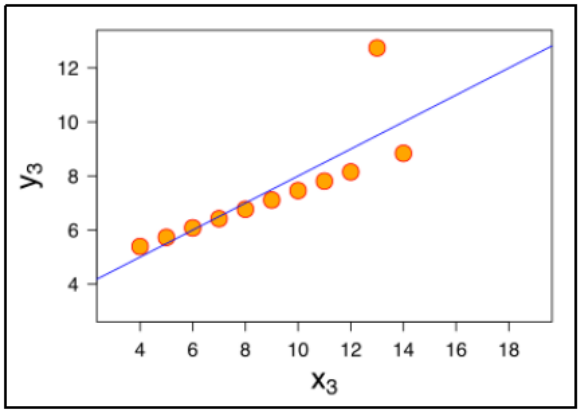
The machine learning models affected by the magnitude of the feature:

- 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

Outliers

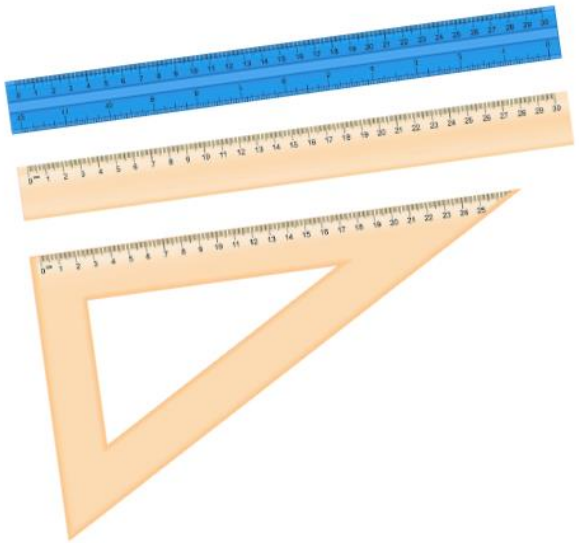


- Discretisation
- Capping / Censoring
- Truncation

Datetime Variables

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

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

Transactions and Time Series

Aggregate data

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

Text

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

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

Open-source for Feature engineering

Category Encoders

Feature-engine

Featuretools

Scikit-Learn Transformers

- Missing Data Imputation
  - SimpleImputer
  - IterativeImputer
- Categorical Variable Encoding
  - OneHotEncoder
  - OrdinalEncoder
- Scalers
  - Standard Scaler
  - MinMaxScaler
  - Robust Scaler
  - A few others
- Discretisation
  - KBinsDiscretizer
- Variable Transformation
  - PowerTransformer
  - FunctionTransformer
- Variable Combination
  - Polynomial Features
- Text
  - Word Count
  - TFiDF

Feature Engine Transformers

Discretisation methods

- EqualFrequencyDiscretiser
- EqualWidthDiscretiser
- DecisionTreeDiscretiser
- ArbitraryDiscreriser

Variable Transformation methods

- LogTransformer
- ReciprocalTransformer
- PowerTransformer
- BoxCoxTransformer
- YeoJohnsonTransformer

Scikit-learn Wrapper:

- SklearnTransformerWrapper

Variable Combinations:

- MathematicalCombination
- CombineWithReferenceFeature

Imputing Methods

- MeanMedianImputer
- RandomSampleImputer
- EndTailImputer
- AddMissingIndicator
- CategoricalImputer
- ArbitraryNumberImputer
- DropMissingData

Encoding Methods

- OneHotEncoder
- OrdinalEncoder
- CountFrequencyEncoder
- MeanEncoder
- WoEEncoder
- PRatioEncoder
- RareLabelEncoder
- DecisionTreeEncoder

Outlier Handling methods

- Winsorizer
- ArbitraryOutlierCapper
- OutlierTrimmer

Category Encoders

```
import category_encoders as ce

encoder = ce.BackwardDifferenceEncoder(cols=[...])
encoder = ce.BaseNEncoder(cols=[...])
encoder = ce.BinaryEncoder(cols=[...])
encoder = ce.CatBoostEncoder(cols=[...])
encoder = ce.HashingEncoder(cols=[...])
encoder = ce.HelmertEncoder(cols=[...])
encoder = ce.JamesSteinEncoder(cols=[...])
encoder = ce.LeaveOneOutEncoder(cols=[...])
encoder = ce.MEstimateEncoder(cols=[...])
encoder = ce.OneHotEncoder(cols=[...])
encoder = ce.OrdinalEncoder(cols=[...])
encoder = ce.SumEncoder(cols=[...])
encoder = ce.PolynomialEncoder(cols=[...])
encoder = ce.TargetEncoder(cols=[...])
encoder = ce.WOEEncoder(cols=[...])
```