Feature Engineering – Variable Transformations
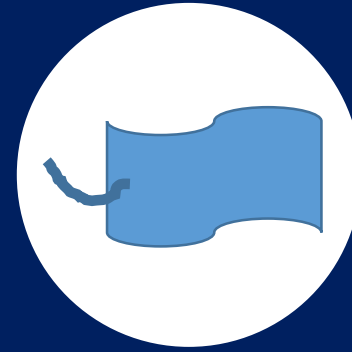
Feature Engineering
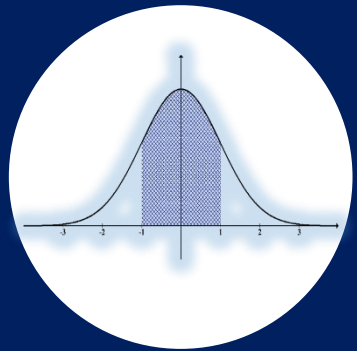

Missing data
Missing values within a variable
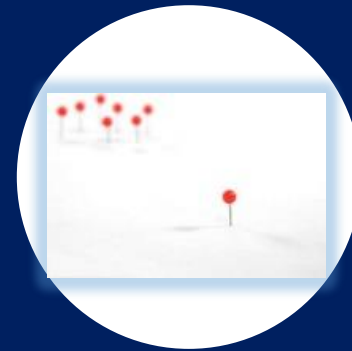

Labels
Strings in categorical variables

Distribution
Normal vs skewed

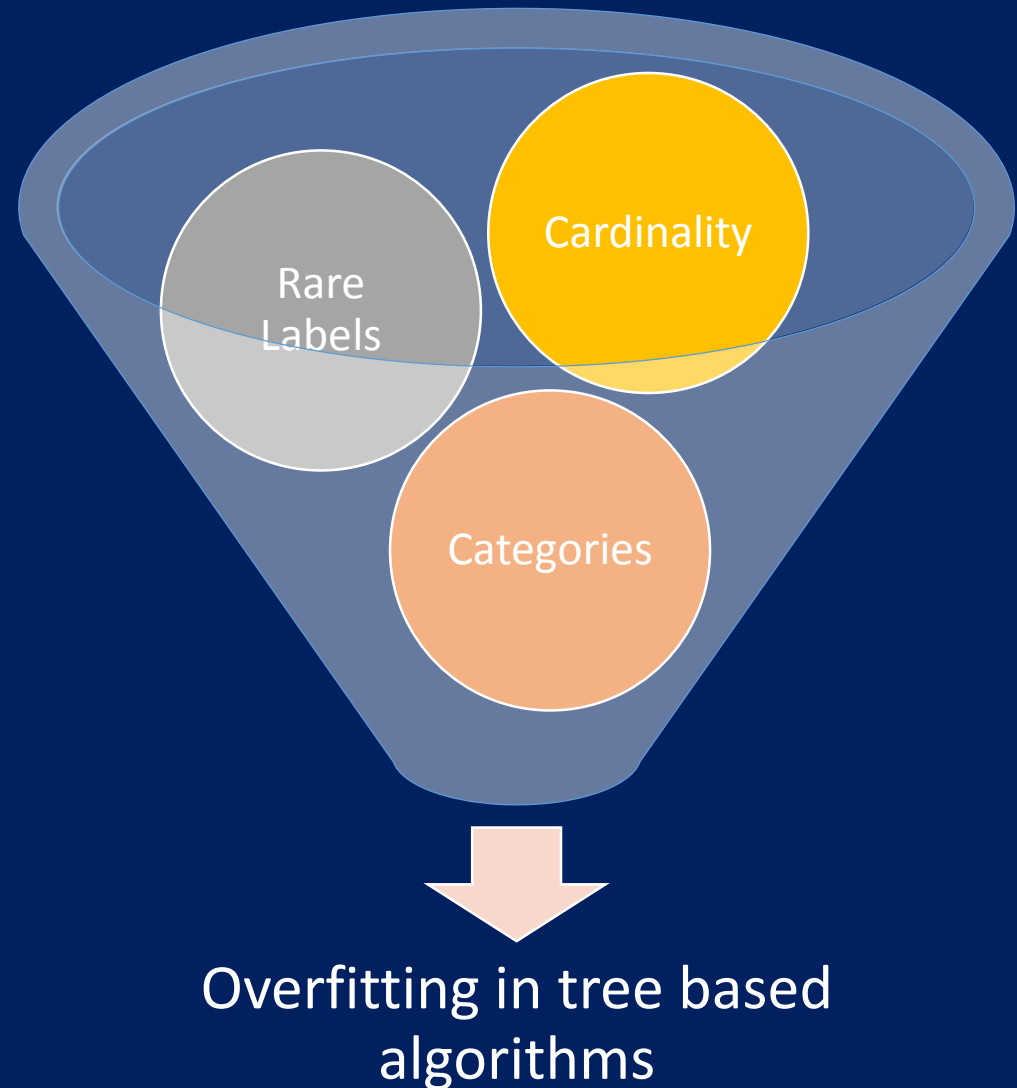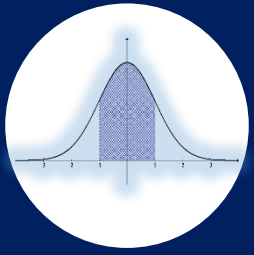
Outliers
Unusual or unexpected values

Missing Data

- Missing values for certain observations

- Affects all machine learning models

  - Scikit-learn

Random

Systematic

Labels in categorical variables

- Cardinality: high number of labels

- Rare Labels: infrequent categories

- Categories: strings

  - Scikit-learn

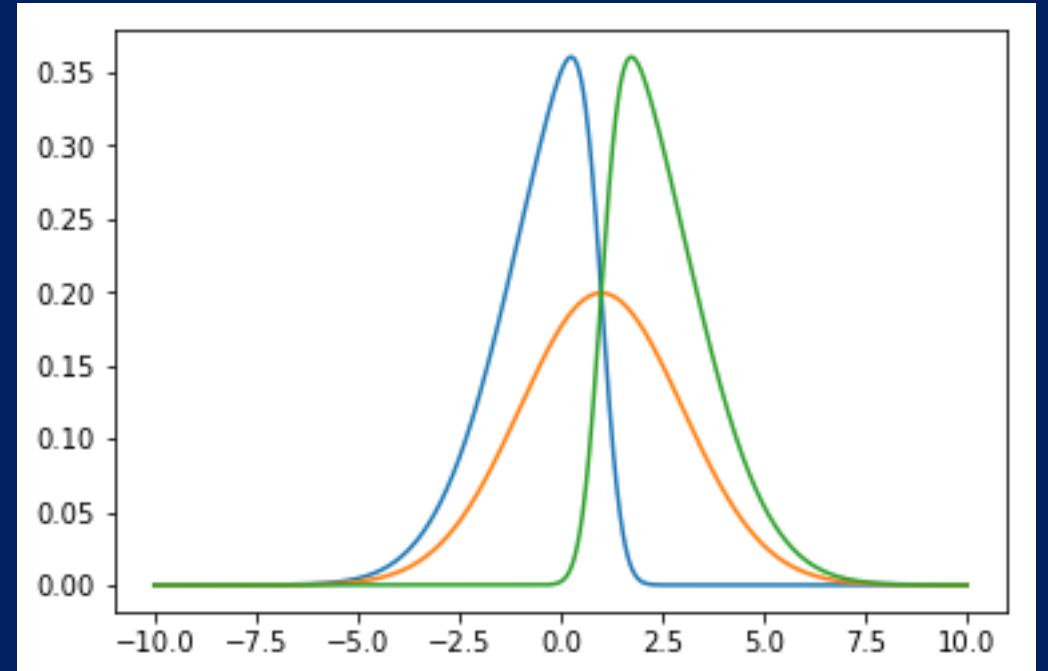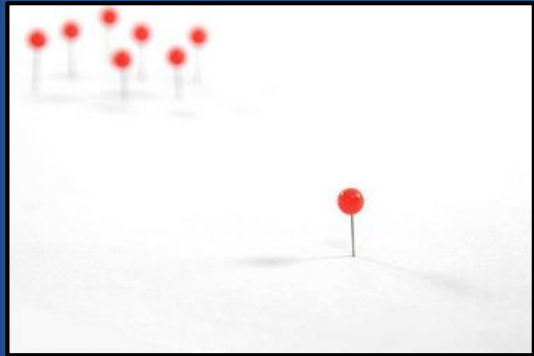
Overfitting in tree based algorithms

Distributions

- Better spread of values may benefit performance

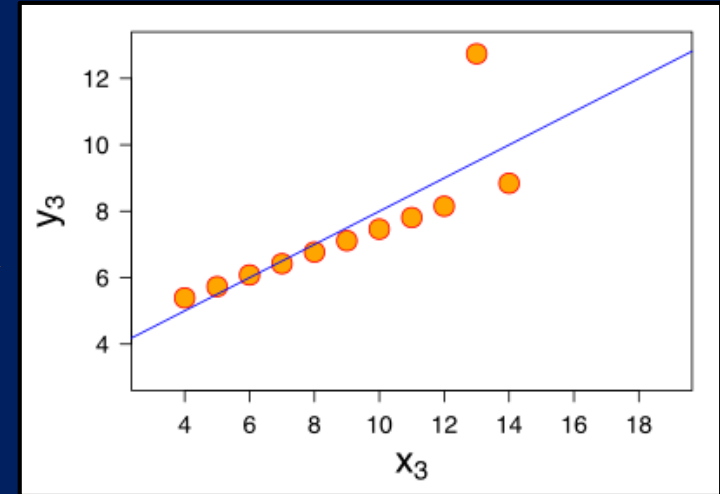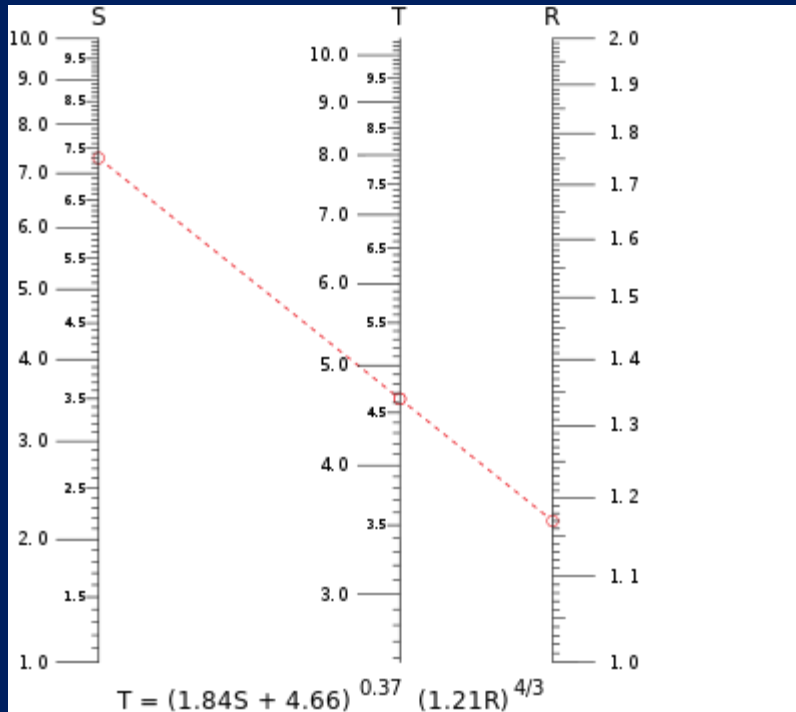Gaussian vs Skewed

Outliers

Linear models

Adaboost

Tremendous weights

Bad generalisation

Feature Magnitude - Scale

$$T = (1.84S + 4.66)^{0.37} (1.21R)^{4/3}$$

Machine learning models sensitive to feature scale:

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

Tree based ML models insensitive to feature scale:

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