Business Data Mining
Semester 2, 2019

Lecture 6
How to Conduct Modeling Phase?

Kim, Yang Sok

Dept. of MIS, Keimyung University

Table of Content

- **Introduction**
- **Task & Deliverables**
- **Exercises**
  - Exercise 1. Selecting modeling techniques in Rapidminer
- **Conclusion**

Introduction

Introduction

- **In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.**

- **The modeling phase includes four tasks. These are**
  - Selecting modeling techniques
  - Generating test design
  - Build model, and
  - Assess model

- **Rapidminer supports modeling phase through Modeling, Scoring and Validation packages.**

Task & Outputs

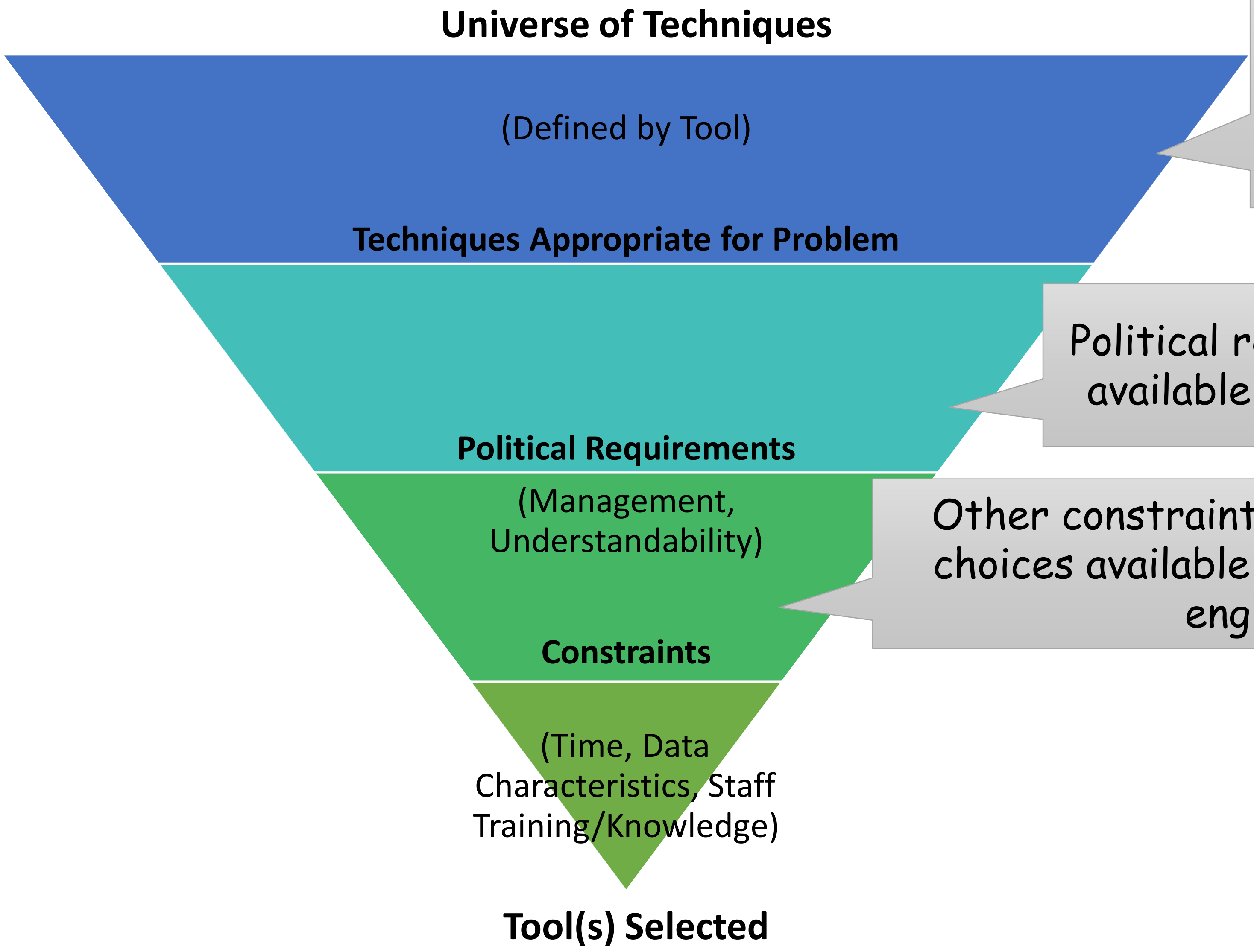
Tasks & Outputs

Select Modeling Techniques

- **Tasks**
  - As the first step in modelling, you'll select the actual modelling technique that you’ll be using.
  - Although you may have already selected a tool during the business understanding phase, at this stage you'll be selecting the specific modelling technique e.g. decision-tree building with C5.0, or neural network generation with back propagation.

Tasks & Outputs

Select Modeling Techniques



Universe of Techniques

(Defined by Tool)

Techniques Appropriate for Problem

Political Requirements

(Management, Understandability)

Constraints

(Time, Data Characteristics, Staff Training/Knowledge)

Tool(s) Selected

Remember that not all tools and techniques are applicable to each and every task. For certain problems, only some techniques are appropriate.

Political requirements limit the choices available to the data mining engineer.

Other constraints further limit the choices available to the data mining engineer.

Tasks & Outputs

Select Modeling Techniques

- **Outputs**
  - **Modelling technique -** Document the actual modelling technique that is to be used.
  - **Modelling assumptions -** Many modelling techniques make specific assumptions about the data, for example that all attributes have uniform distributions, no missing values allowed, class attribute must be symbolic etc. Record any assumptions made.

Tasks & Outputs

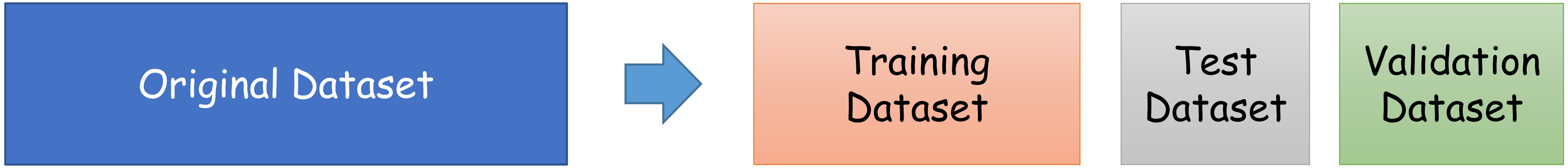
Generating test design

- **Tasks**
  - Before you actually build a model you need to generate a procedure or mechanism to test the model's quality and validity.
  - For example, in supervised data mining tasks such as classification, it is common to use error rates as quality measures for data mining models.
  - Therefore, you typically separate the dataset into train and test sets, build the model on the train set, and estimate its quality on the separate test set.

Tasks & Outputs

Generating test design

- **Outputs**
  - **Test design -** Describe the intended plan for training, testing, and evaluating the models. A primary component of the plan is determining how to divide the available dataset into training, test and validation datasets.



Tasks & Outputs

Build model

- **Tasks**
  - Run the modelling tool on the prepared dataset to create one or more models.

- **Outputs**
  - **Parameter settings** - With any modelling tool there are often a large number of parameters that can be adjusted. List the parameters and their chosen values, along with the rationale for the choice of parameter settings.
  - **Models** - These are the actual models produced by the modelling tool, not a report on the models.
  - **Model descriptions** - Describe the resulting models, report on the interpretation of the models and document any difficulties encountered with their meanings.

Tasks & Outputs

Assess model

- **Tasks**
  - Interpret the models according to your domain knowledge, your data mining success criteria and your desired test design. Judge the success of the application of modelling and discovery techniques technically, then contact business analysts and domain experts later in order to discuss the data mining results in the business context. This task only considers models, whereas the evaluation phase also takes into account all other results that were produced in the course of the project.
  - At this stage you should rank the models and assess them according to the evaluation criteria. You should take the business objectives and business success criteria into account as far as you can here. In most data mining projects a single technique is applied more than once and data mining results are generated with several different techniques.

Tasks & Outputs

Assess model

- **Outputs**
  - **Model assessment** - Summarize the results of this task, list the qualities of your generated models (e.g.in terms of accuracy) and rank their quality in relation to each other.
  - **Revised parameter settings** - According to the model assessment, revise parameter settings and tune them for the next modelling run. Iterate model building and assessment until you strongly believe that you have found the best model(s). Document all such revisions and assessments.

Select Modeling Techniques

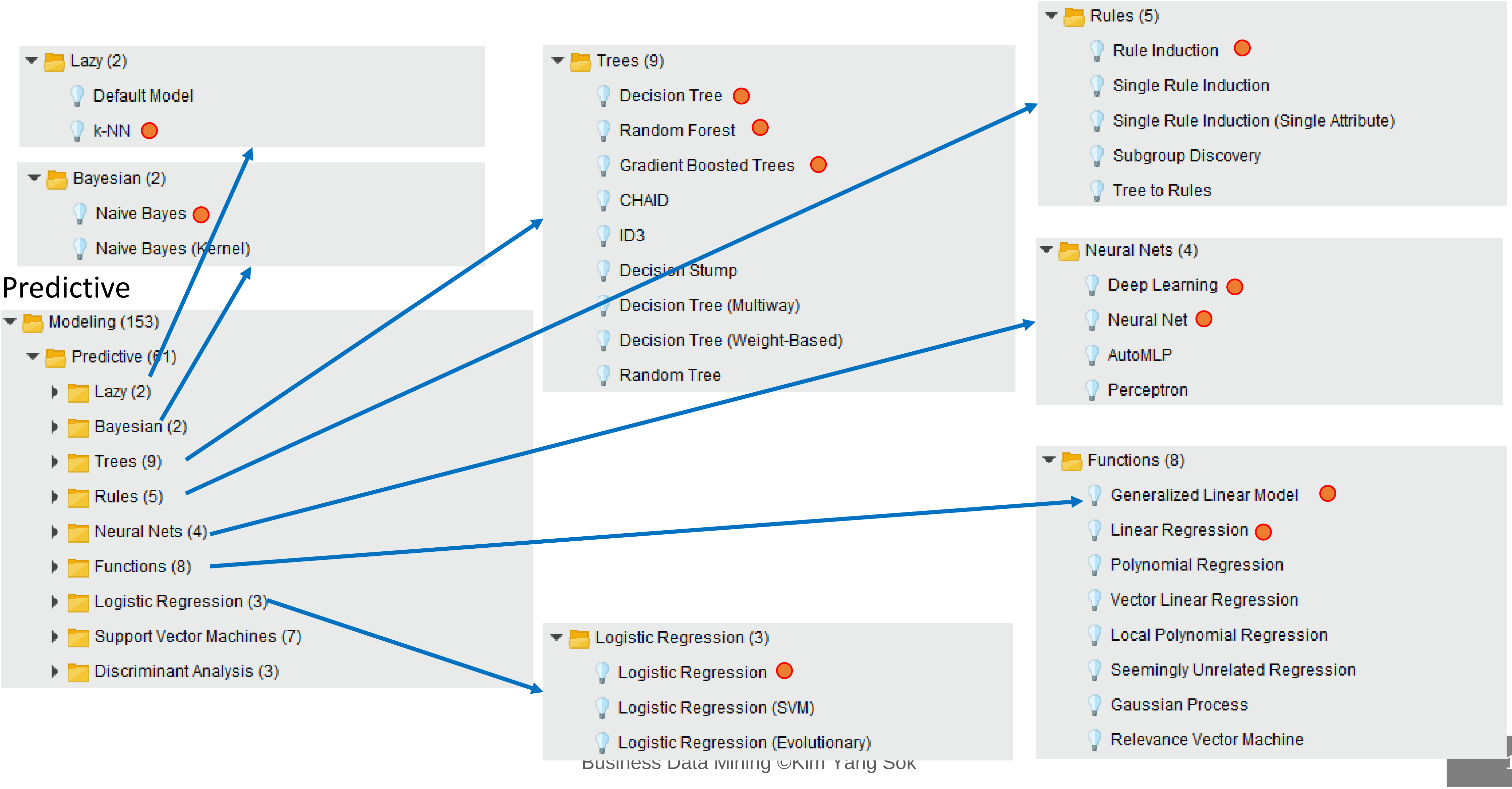
Select Modelling Techniques

- **Data description and summarization**
  - Data description and summarization aims at the concise description of characteristics of the data
- **Segmentation**
  - Segmentation aims at the separation of the data into interesting and meaningful subgroups or classes.
- **Concept descriptions**
  - Concept description aims at an understandable description of concepts or classes.
- **Classification**
  - Classification assumes that there is a set of objects characterized by some attributes or features that belong to different classes.
- **Prediction**
  - Prediction is very similar to classification. The only difference is that in prediction the target attribute (class) is not a discrete qualitative attribute but a continuous one.
- **Dependency analysis**
  - Dependency analysis consists of finding a model that describes significant dependencies (or associations) between data items or events.

Usually, the data mining project involves a combination of different problem types, which together solve the business problem.

Select Modeling Techniques

Operators in RapidMiner



Predictive

- Modeling (153)
  - Predictive (61)
    - Lazy (2)
    - Bayesian (2)
    - Trees (9)
    - Rules (5)
    - Neural Nets (4)
    - Functions (8)
    - Logistic Regression (3)
    - Support Vector Machines (7)
    - Discriminant Analysis (3)

- Lazy (2)
  - Default Model
  - k-NN ●

- Bayesian (2)
  - Naive Bayes ●
  - Naive Bayes (Kernel)

- Trees (9)
  - Decision Tree ●
  - Random Forest ●
  - Gradient Boosted Trees ●
  - CHAID
  - ID3
  - Decision Stump
  - Decision Tree (Multiway)
  - Decision Tree (Weight-Based)
  - Random Tree

- Rules (5)
  - Rule Induction ●
  - Single Rule Induction
  - Single Rule Induction (Single Attribute)
  - Subgroup Discovery
  - Tree to Rules

- Neural Nets (4)
  - Deep Learning ●
  - Neural Net ●
  - AutoMLP
  - Perceptron

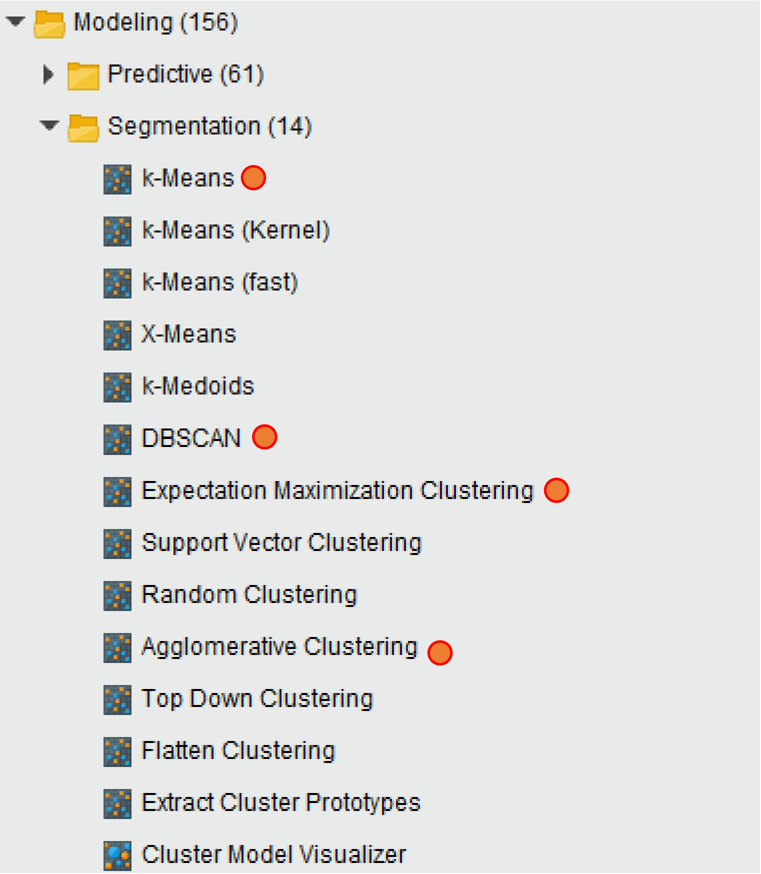
- Functions (8)
  - Generalized Linear Model ●
  - Linear Regression ●
  - Polynomial Regression
  - Vector Linear Regression
  - Local Polynomial Regression
  - Seemingly Unrelated Regression
  - Gaussian Process
  - Relevance Vector Machine

- Logistic Regression (3)
  - Logistic Regression ●
  - Logistic Regression (SVM)
  - Logistic Regression (Evolutionary)

Select Modeling Techniques

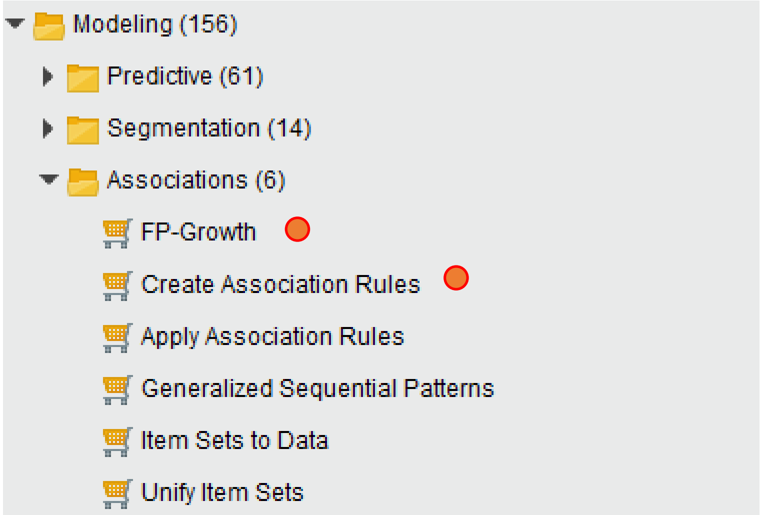
Operators in RapidMiner

Segmentation



- Modeling (156)
  - Predictive (61)
  - Segmentation (14)
    - k-Means ●
    - k-Means (Kernel)
    - k-Means (fast)
    - X-Means
    - k-Medoids
    - DBSCAN ●
    - Expectation Maximization Clustering ●
    - Support Vector Clustering
    - Random Clustering
    - Agglomerative Clustering ●
    - Top Down Clustering
    - Flatten Clustering
    - Extract Cluster Prototypes
    - Cluster Model Visualizer

Association

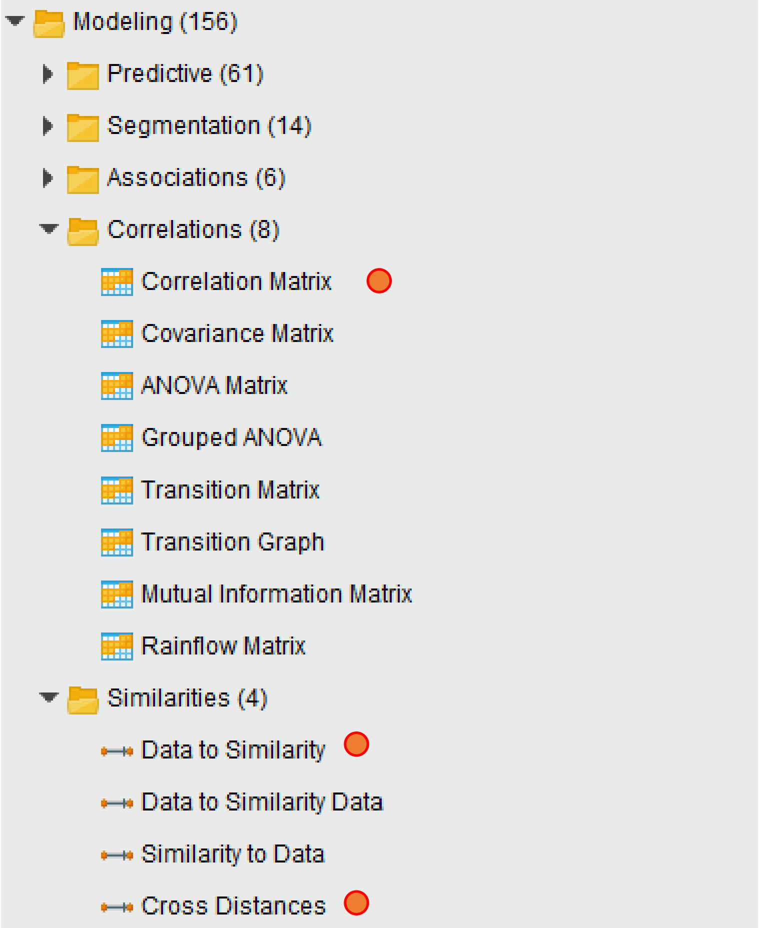


- Modeling (156)
  - Predictive (61)
  - Segmentation (14)
  - Associations (6)
    - FP-Growth ●
    - Create Association Rules ●
    - Apply Association Rules
    - Generalized Sequential Patterns
    - Item Sets to Data
    - Unify Item Sets

Exercise: Selecting Modelling Techniques

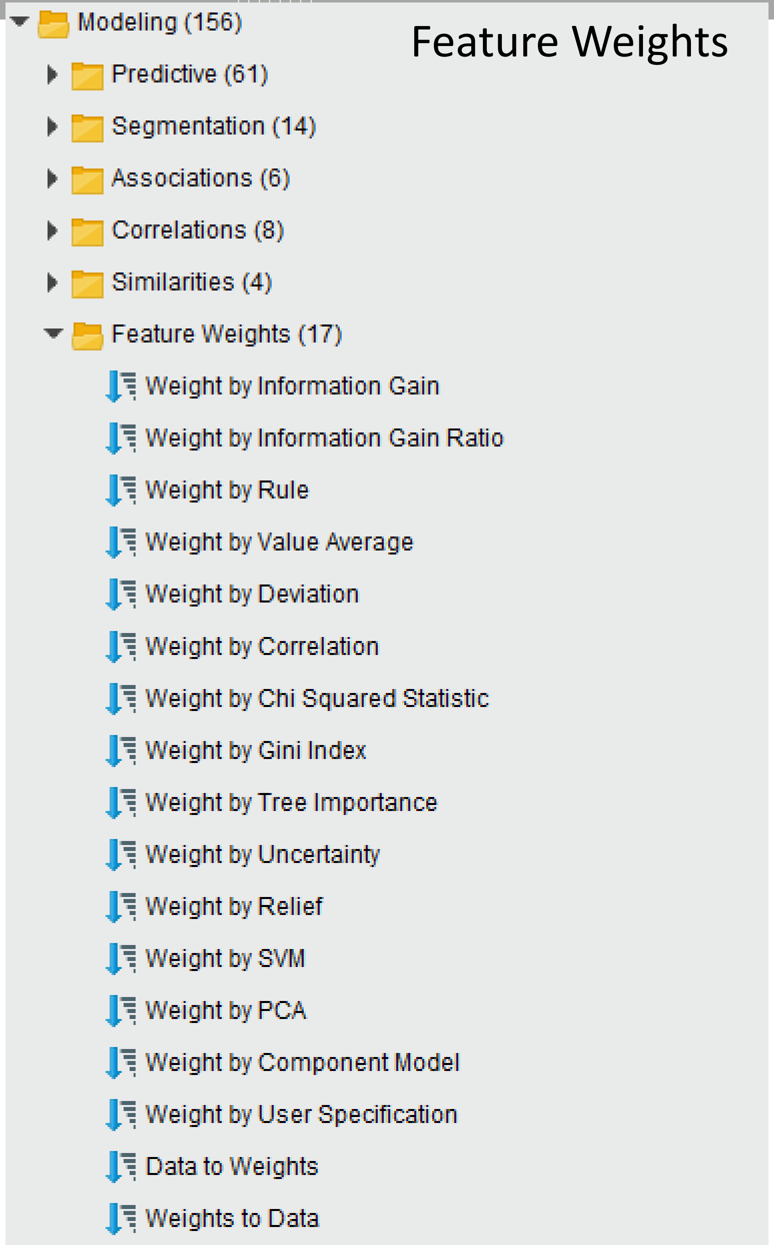
Operators in RapidMiner

Correlation & Similarity



- Modeling (156)
  - Predictive (61)
  - Segmentation (14)
  - Associations (6)
  - Correlations (8)
    - Correlation Matrix ●
    - Covariance Matrix
    - ANOVA Matrix
    - Grouped ANOVA
    - Transition Matrix
    - Transition Graph
    - Mutual Information Matrix
    - Rainflow Matrix
  - Similarities (4)
    - Data to Similarity ●
    - Data to Similarity Data
    - Similarity to Data
    - Cross Distances ●

Feature Weights

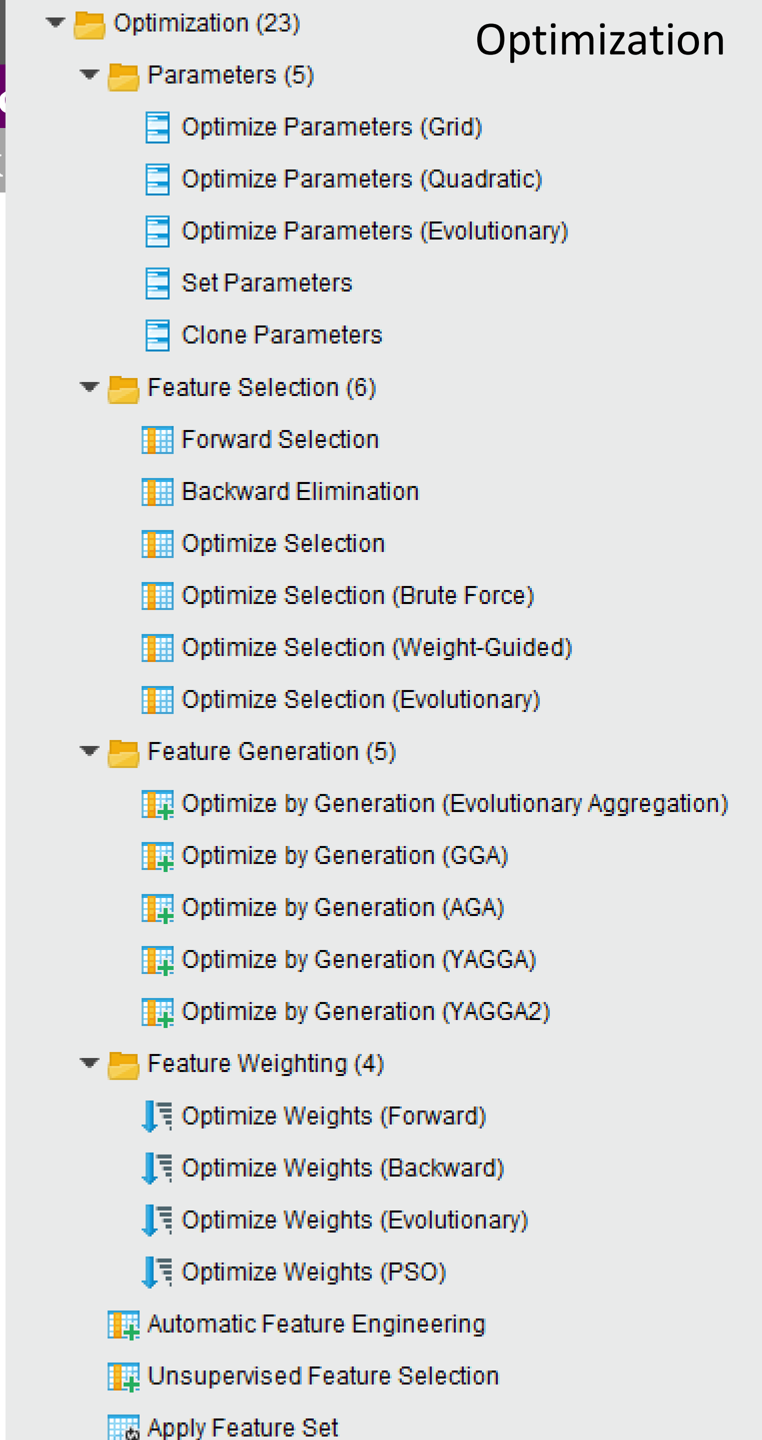


- Modeling (156)
  - Predictive (61)
  - Segmentation (14)
  - Associations (6)
  - Correlations (8)
  - Similarities (4)
  - Feature Weights (17)
    - Weight by Information Gain
    - Weight by Information Gain Ratio
    - Weight by Rule
    - Weight by Value Average
    - Weight by Deviation
    - Weight by Correlation
    - Weight by Chi Squared Statistic
    - Weight by Gini Index
    - Weight by Tree Importance
    - Weight by Uncertainty
    - Weight by Relief
    - Weight by SVM
    - Weight by PCA
    - Weight by Component Model
    - Weight by User Specification
    - Data to Weights
    - Weights to Data

Exer es

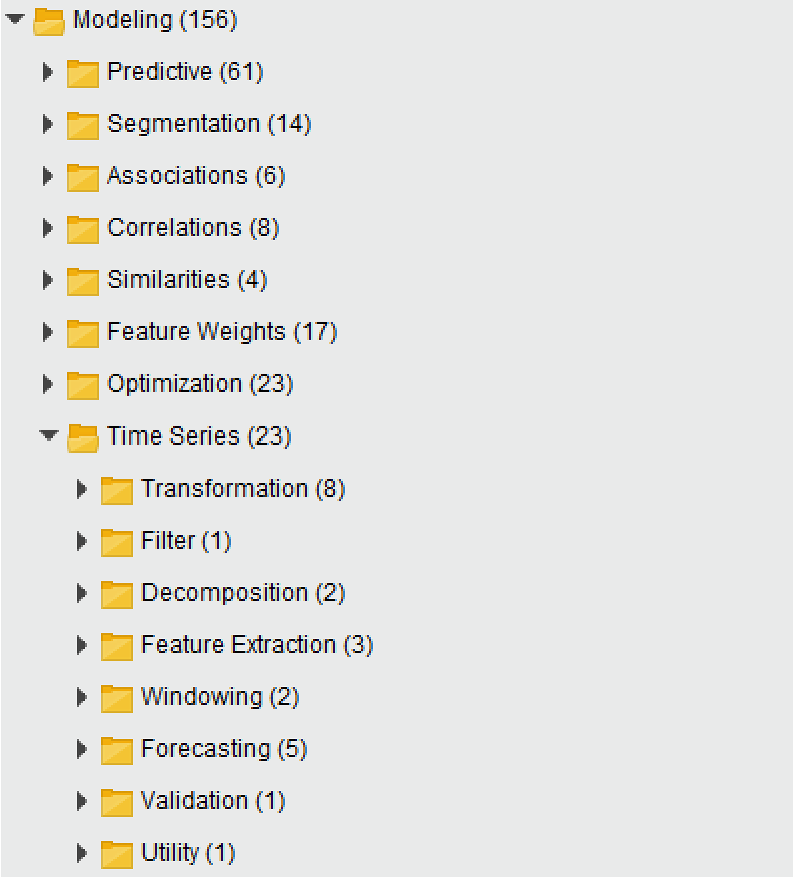
Task

Optimization



- Optimization (23)
  - Parameters (5)
    - Optimize Parameters (Grid)
    - Optimize Parameters (Quadratic)
    - Optimize Parameters (Evolutionary)
    - Set Parameters
    - Clone Parameters
  - Feature Selection (6)
    - Forward Selection
    - Backward Elimination
    - Optimize Selection
    - Optimize Selection (Brute Force)
    - Optimize Selection (Weight-Guided)
    - Optimize Selection (Evolutionary)
  - Feature Generation (5)
    - Optimize by Generation (Evolutionary Aggregation)
    - Optimize by Generation (GGA)
    - Optimize by Generation (AGA)
    - Optimize by Generation (YAGGA)
    - Optimize by Generation (YAGGA2)
  - Feature Weighting (4)
    - Optimize Weights (Forward)
    - Optimize Weights (Backward)
    - Optimize Weights (Evolutionary)
    - Optimize Weights (PSO)
  - Automatic Feature Engineering
  - Unsupervised Feature Selection
  - Apply Feature Set

Time Series



- Modeling (156)
  - Predictive (61)
  - Segmentation (14)
  - Associations (6)
  - Correlations (8)
  - Similarities (4)
  - Feature Weights (17)
  - Optimization (23)
  - Time Series (23)
    - Transformation (8)
    - Filter (1)
    - Decomposition (2)
    - Feature Extraction (3)
    - Windowing (2)
    - Forecasting (5)
    - Validation (1)
    - Utility (1)

Conclusion

Conclusion

- **There are four tasks in the modeling phase**
  - Selecting modeling techniques, Generating test design, Build model, and Assess model.
- **This lecture focuses on "Selecting modeling techniques", because other task can be explained with real modeling process.**
- **There are six problem types in data mining**
  - Data description and summarization, Segmentation, Concept descriptions, Classification, Prediction and Dependency analysis
- **In the next lectures, you will learn how to generate test design with k-NN algorithm.**

QUESTIONS?