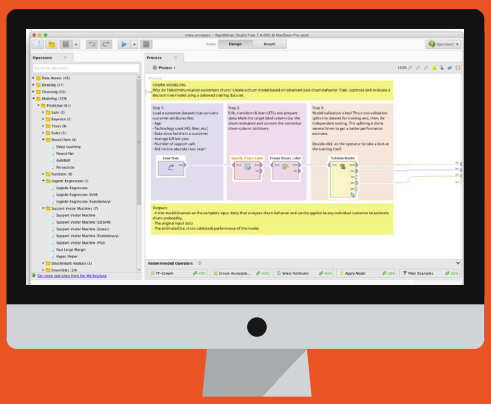
RapidMiner Studio

Lightning Fast Data Science



Key Features

- Visual Programming Environment
- Guided Analytics
- Reusable Building Blocks
- 1500 + Machine Learning & Data Prep Functions
- Easy Integration of R & Python Scripts
- Correct Model Validation Methods
- Access All Types of Data

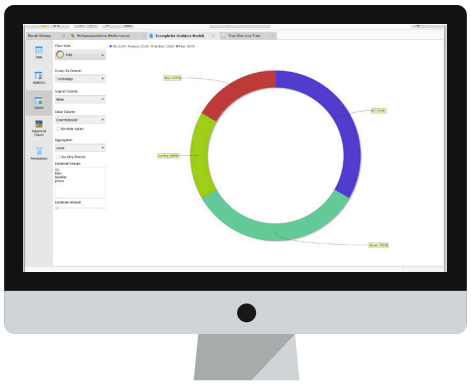
> *We were impressed with the speed with which we were able to use RapidMiner's Data Science Platform to source, cleanse, simplify and cluster our data to solve a complex problem.*

L. Ellenburg,
Senior Manager of Informatics

Maximize Data Science Productivity

RapidMiner Studio is a visual design environment for rapidly building complete predictive analytic workflows. It provides a deep library of machine learning algorithms, data preparation and exploration functions, and model validation tools to support all your data science projects and use cases.

Data science teams can easily re-use existing R and Python code, and add new functionality via a large marketplace of pre-built extensions.



Accelerate Data Prep

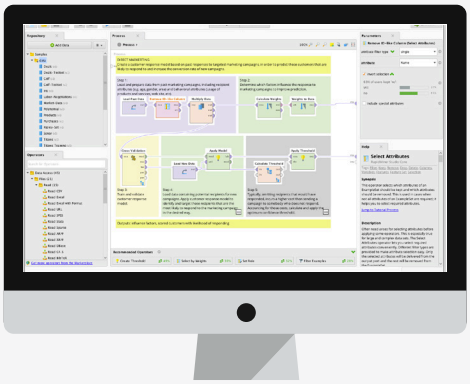
RapidMiner Studio provides a wealth of functionality to speed & optimize data exploration, blending & cleansing tasks – reducing the time spent importing and wrangling your data.

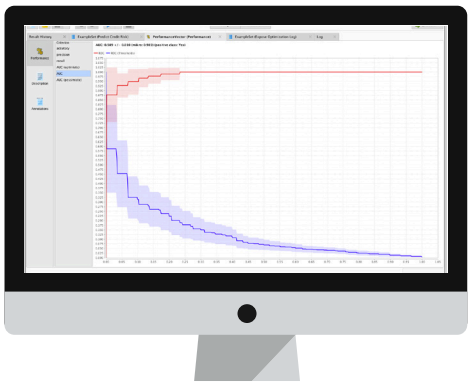
- Statistical overviews, graphs & charts for data exploration
- Powerful functionality for advanced feature weighting, selection & generation
- Anomaly & outlier detection, missing value handling and normalization techniques expertly prepares data

Develop Models Quickly

Hundreds of machine learning, text analytics, predictive modeling algorithims, automation, and process control features help you build better models faster than ever before.

- Multitude of classification and regression algorithms facilitate supervised learning
- Broad array of cluster, similarity and segmentation algorithms support unsupervised learning
- Seamless integration of R and Python scripts into workflows provide further extensibility





Confidently Evaluate Performance

Accurately estimate model performance – and confidently validate your results. RapidMiner Studio delivers a correct assessment of the performance of your models before you put them into production.

- Split & cross-validation methods increases the predictive accuracy of machine learning models
- Reuse preprocessing models drastically reduces contamination of your results
- Multiple validation techniques & performance calculations ensure that your results meet expectations

Data Science Workflow Designer
Feature List

Application & Interface

RapidMiner Studio is a visual data science workflow designer accelerating the prototyping & validation of models

- Easy to use visual environment for building analytics processes:
  - Graphical design environment makes it simple and fast to design better models
  - Visual representation with Annotations facilitates collaboration among all stakeholders
- Every analysis is a process, each transformation or analysis step is an operator, making design fast, easy to understand, and fully reusable
- Guided process design leveraging the Wisdom of Crowds, i.e. the knowledge and the best practices of more than 200,000 users in the RapidMiner community
  - Operator recommender suggesting next steps
  - Parameter recommender indicating which parameters to change & to which values
- Convenient set of data exploration tools and intuitive visualizations
- More than 1500 operators for all tasks of data transformation and analysis
- Support for scripting environments like R, or Groovy for ultimate extensibility
- Seamlessly access and use of algorithms from H2O, Weka and other third-party libraries
- Transparent integration with RapidMiner Server to automate processes for data transformation, model building, scoring and integration with other applications
- Extensible through open platform APIs and a Marketplace with additional functionality

Data Access and Management

With RapidMiner Studio, you can access, load and analyze any type of data – both traditional structured data and unstructured data like text, images, and media. It can also extract information from these types of data and transform unstructured data into structured.

- Access to more than 40 file types including SAS, ARFF, Stata, and via URL
- Wizards for Microsoft Excel & Access, CSV, and database connections
- Access to NoSQL databases MongoDB and Cassandra
- Write to Qlik QVX or Tableau TDE files
- Access to Cloud storage like Dropbox and Amazon S3
- Access to text documents and web pages, PDF, HTML, and XML
- Support for all JDBC database connections including Oracle, IBM DB2, Microsoft SQL Server, MySQL, Postgres, Teradata, Ingres, VectorWise, and more
- Access to full-text index & search platform SOLR
- Access to Twitter & Salesforce.com
- Repository-based data management on local systems or central servers via RapidMiner Server
- Connect to Zapier and trigger Zapier tasks
- Access to time series data, audio files images, and many more
- Enhanced data and metadata editor for repository entries

Data Exploration

Immediately understand and create a plan to prepare the data automatically extract statistics and key information.

Descriptive Statistics

- Univariate statistics and plots
  - Numerical attributes: mean, median, minimum, maximum, standard deviation, and number of missing values
  - Nominal / categorical attributes: number of categories, counts, mode, number of missing values
  - Date attributes: minimum, maximum, number of missing values
- Distribution plots
- Bivariate statistics and plots:
  - Covariance matrix
  - Correlation matrix
  - Anova matrix
  - Grouped Anova
- Transition matrix
- Transition graph
- Mutual information matrix
- Rainflow matrix
- Scaled and non-scaled mean-deviation plots
- Plots of attribute weights based on multiple types of connection with targets
- Simple rescaling of axis
- Plots can be easily copied and pasted into other applications or exported as in PNG, SVG, JPEG, EPS or PDF formats
- Choose from a variety of diffrent color schemes

Graphs and Information

- Easy-to-configure charts for fast insight generation from various visualizations
  - Scatter, scatter matrices
  - Line
  - Bubble
  - Parallel
  - Deviation
  - Box
  - 3-D
  - Density
  - Histograms
  - Area
  - Bar charts, stacked bars
  - Pie charts
  - Survey plots
  - Self-organizing maps
  - Andrews curves
  - Quartile
  - Surface / contour plots, time series plots
  - Pareto / lift chart
- Support for zooming and panning
- Additional advanced chart engine for arbitrary definition of multiple charts including: on-the-fly grouping, filtering & aggregation

Data Prep

The richness of the data preparation capabilities in RapidMiner Studio can handle any real-life data transformation challenges, so you can format and create the optimal data set for predictive analytics. RapidMiner Studio can blend structured with unstructured data and then leverage all the data for predictive analysis. Any data preparation process can be saved for reuse.

Basics

- Select attributes operator
- Aggregations for multiple groups and functions like sum, average, median, standard deviation, variance, count, least, mode, minimum, maximum, product, or log product
- Set operators like join, merge, append, union, or intersect
- Operators for handling meta data like rename or attribute role definition
- Filtering rows / examples according to range, missing values, wrong or correct predictions, or specific attribute value
- Filtering outliers according to distances, densities, local outlier factors, class outlier factors, local correlation integrals, or clustering based outlier detections
- Identification and removal of duplicates

Sampling

- Absolute, relative, or probability–based
- Balanced
- Stratified
- Bootstrapping
- Model-based
- Kennard-Stone
- Range

Transformations

- Normalization and standardization
- Z-transformation, range transformation, proportion transformation, or interquartile ranges
- Preprocessing models for applying the same transformations on test / scoring data
- De-normalization making use of preprocessing models
- Scaling by weights
- All kinds of type conversions between numerical attributes, nominal / categorical attributes, and date attributes
- Operator for guessing correct meta data from existing data sets
- Adjustment of calendar dates and times
- Sorting and Pareto sort
- Shuffling
- Rotations of data sets: Pivoting, De-Pivoting, and transposing data sets
- Expression builder for arbitrary transformations on attributes: Statistical functions: round, floor, ceiling, average, minimum, maximum
- Basic functions: addition, subtraction, multiplication, division, less than, greater than, less or equal, greater or equal, equal, not equal, Boolean not, Boolean and, Boolean or
- Log and exponential functions: natural logarithm, logarithm base 10, logarithm dualis, exponential, power
- Trigonometric functions: sine, cosine, tangent, arc sine, arc cosine, arc tangent, hyperbolic sine, hyperbolic cosine, hyperbolic tangent, inverse hyperbolic sine, inverse hyperbolic cosine, inverse hyperbolic tangent
- Text functions: to string, to number, cut, concatenation, replace and replace all, lower, upper, index, length, character at, compare, contains, equals, starts with, ends with, matches, suffix, prefix, trim, escape HTML
- Date functions: parse, parse with locale, arse custom, before, after, to string, to string with locale, to string with custom pattern, create current, difference, add, set, and get
- Miscellaneous functions: if then-else, square root, signum, random, modulus, sum, binomial, missing binomial, missing

Data Partitioning

- Ensure high model quality through hold-out data sets
- Create training, validation, and test data sets
- Default stratification by the class if available
- User-defined partitions possible
- Resulting in example sets usable for modeling or further transformations

Binning

- Interactive binning by user specification
- Simple binning
- Count-based
- Size-based
- Frequency-based
- Entropy-based minimizing the entropy in the induced partitions
- Handling of missing values as its own group

Weighting and Selection

- Attribute weighting
    - 30+ weighting schemes measuring the influence of attributes & forming base or weight-based selections (filter approach)
- Attribute selection
    - Selection of attributes by user specification
    - Removal of "useless" attributes
    - Removal of attributes unrelated to target based on a chi-square or correlation-based selection criterion
    - Removal of attributes unrelated to target based on arbitrary weighting schemes like information gain, Gini index, and others
    - Removal attributes with missing values
    - Selection of random attribute subsets
- Automatic optimization of selections
    - Evolutionary
    - Forward selection
    - Backward elimination
    - Weight-guided
    - Brute-force
- Attribute space transformations
    - Principal Component Analysis (PCA)
    - Singular Value Decomposition
- Support for Fast Map
- Plots for principal components coefficients, Eigenvalues, and cumulative variance of Eigenvalues
- Calculates Eigenvalues and Eigenvectors from correlation and covariance matrices
- Choose the number of components to be retained
- Independent component analysis (ICA)
- Generalized Hebbian Algorithm (GHA)
- Dimensionality reduction with Self- Organizing Maps (SOM)
- Correspondence Analysis

Attribute Generation

- Operators for generating IDs, copies, concatnations, aggregations, products, Gaussian distributions, and more
- Automatically optimized generations and detection of latent variables: Evolutionary weighting
- Forward weighting
- Backward weighting
- Multiple algorithms for the automatic creation of new attributes based on arbitrary functions of existing attributes
- Genetic programming

Modeling

RapidMiner Studio comes equipped with an un-paralleled set of modeling capabilities and machine learning algorithms for supervised and unsupervised learning. They are flexible, robust and allow you to focus on building the best possible models for any use case.

Similarity Calculation

- Calculation of similarities between data points
- Cross Distances operator computes similarities between data points of two data sets
- Numerical distance measures
  - Euclidean
  - Camberra
  - Chebychev
  - Correlation
  - Cosine
  - Dice
  - Dynamic Time Warping
  - Inner product
  - Jaccard
  - Kernel-Euclidean
  - Manhattan
  - Max-Product
  - Overlap
- Nominal / categorical distance measures
  - Nominal
  - Dice
  - Jaccard
  - Kulczynski
  - Rogers-Tanimoto
  - Russel-Rao
  - Simple Matching
- Mixed Euclidean distance for cases with numerical & nominal attributes
  - Bregman divergences
  - Itakura-Saito
  - Kullback-Leibler
  - Logarithmic loss
  - Logistic loss
  - Mahalonobis
  - Squared Euclidean
  - Squared Loss

Clustering

- User defined clustering or automatically chooses the best clusters
- Support Vector Clustering
- Several strategies for encoding class into the clustering
- k-Means (for all available distance and similarity measures)
- k-Medoids (for all available distance and similarity measures)
- Kernel k-Means
- X-Means
- Cobweb
- Clope
- DBScan
- Expectation Maximization Clustering
- Self-organizing maps
- Agglomerative Clustering
- Top Down Clustering
- Operators for flattening hierarchical cluster models
- Extraction of prototypes for centroid-based cluster models

Market Basket Analysis

- Associations and sequence discovery
- Measuring quality of rules by support, confidence, La Place, gain, ps-value, lift or conviction
- Interactive filter for frequent item sets
- Interactive visualization of association rules as a network graph
- Rules description table
- User defined rule filtering depending on minimum value for the above criteria or matching criteria for specific items
- FP-Growth (similar to Apriori but far more efficient)
- Generalized sequential patterns

Market Basket Analysis Cont'd

- Modular operators for the creation of frequent item sets or association rules only
- Post-processing to unify of item sets
- Application of association rules to deploy as a recommendation engine

Decision Trees

- Easy-to-understand models
- Supported methods: classification and regression trees (CART), CHAID, decision stumps, ID3, C4.5, Random Forest, bagging and boosting
- Support for multi-way trees
- Gradient Boosted Trees (GBT)
- Pre-pruning and pruning
- Split criteria include information gain, gain ratio, accuracy, and Gini index
- Error-based and confidence-based pruning
- Distribution shown at tree leaves
- Height of distribution bars correlate to number of examples in each leaf
- Majority class shown at tree leaves
- Class counts shown as tool tip at tree leaves
- The darkness of connections correlates with the number of examples on this path
- Graphical and textual representation of trees
- Interactive visualization of trees including selecting and moving of nodes

Rule Induction

- Recursive technique with easy-to-read results
- Especially useful for modeling rare events like for subgroup discovery
- Supported methods: rule induction, single rule induction, single attribute, subgroup discovery, tree to rules
- Supported splitting criteria include information gain and accuracy
- Definition of pureness of rules
- Error-based pruning
- Easy to read and parse representation of rule sets as textual descriptions or tables

Bayesian Modeling

- Naïve Bayes
- Kernel Naïve Bayes
- Bayes models can be updated and are therefore especially suitable for large data sets or online stream mining

Regression

- Linear
- Logistic
- Generalized Linear Model (H2O)
- Kernel Logistic Regression
- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)
- Regularized Discriminant Analysis (RDA)
- Stepwise forward and backward selection
- Selection with M5', t-test or iterative t-test
- Seemingly unrelated regression
- Vector linear regression
- Polynomial regression
- Local polynomial regression
- Gaussian Processes

Neural networks

- Flexible network architectures with different activation functions
- Multiple layers with different numbers of nodes
- Different training techniques
- Perceptron
- Multilayer Perceptron
- Deep Learning (H2O)
- Automatic optimization of both learning rate and size adjustment of neural networks during training

Support Vector Machines

- Powerful and robust modeling techniques for large numbers of dimensions
- Offers overfitting control by regularization
- Especially suitable for modeling unstructured information like text data
- More than 10 different methods for support vector classification, regression, and clustering
- Support Vector Machine
- Relevance vector machine
- Linear, Evolutionary, PSO, Fast Large Margin, Hyper Hyper
- Kernel functions include dot, radial basis function, polynomial, neural, Anova, Epachnenikov, Gaussian combination, or multiquadric
- Simple support vector machines for boosting support
- Linear-time support vector machine for fast training also for large numbers of dimensions and examples

Memory-Based Reasoning

- k-Nearest Neighbors for classification and regression
- Locally weighted learning
- Optimized scoring through ball trees data search structure

Model Ensembles

- Hierarchical models
- Combination of multiple models to form a potentially stronger model
- Vote
- Additive regression
- Ada boost
- Bayesian boosting
- Bagging
- Stacking
- Classification by regression
- Meta cost for defining costs for different error types and detecting optimal models avoiding expensive errors

Validation

RapidMiner Studio provides the means to accurately and appropriately estimate model performance. Where other tools tend to too closely tie modeling and model validation, RapidMiner Studio follows a stringent modular approach which prevents information used in pre-processing steps from leaking from model training into the application of the model. This unique approach is the only guarantee that no overfitting is introduced and no overestimation of prediction performances can occur.

Performance Criteria

- Many performance criteria for numerical and nominal / categorical targets, including:
  - Accuracy
  - Classification error
  - Kappa
  - Area under curve (AUC)
  - Precision
  - Recall
  - Lift
  - Fallout
  - F-measure
  - False positives
  - False negatives
  - True positives
  - True negatives
  - Sensitivity
  - Specificity
  - Youden index
  - Positive predictive value
  - Negative predictive value
  - PSEP

Performance Criteria Cont'd

- Correlation
- Spearman rho
- Kendall tau
- Squared correlation
- Absolute error
- Relative error
- Normalized absolute error
- Root mean squared error (RMSE)
- Root relative squared error (RRSE)
- Squared error
- Cross entropy
- Margin
- Soft margin loss
- Logistic loss

- Calculating significance tests to determine if and which models performed better
  - T-test
  - Anova
- Find threshold operator to determine optimal cutoff point for binominal classes
- Performance estimation for cluster models based on distance calculations, density calculations, or item distributions

Validation Techniques

- Embed pre-processing steps into the validation
- Display multiple results in history to help better evaluate model performance
- Various techniques for the estimation of model performance: Cross validation (with parallel execution of the folds)
- Split validation
- Bootstrapping
- Batch cross validation
- Wrapper cross validation
- Wrapper split validation
- Visual evaluation techniques
- Lift chart
- ROC curves
- Confusion matrix

Scoring

RapidMiner's Studio makes the application of models easy, whether you are scoring them in the RapidMiner platform or using the resulting models in other applications.

- Operator for applying models to datasets (Scoring)
- Support of predictive models, cluster models, preprocessing models, transformation models, and models for missing value imputations
- Storing of models in central repositories for reuse in other processes and projects
- Applying a model creates optimal scores by ignoring unused attributes and handling previously unseen values
- Import and export of RapidMiner models, R models, and Weka models from repository or files
- Support of PMML 3.2 and 4.0

Automation and Process Control

Unlike many other predictive analytics tools, RapidMiner Studio covers even the trickiest data science use cases without the need to program. Beyond all the great functionality for preparing data and building models, RapidMiner Studio has a set of utility-like process control operations that lets you build processes that behave like a program to repeat and loop over tasks, branch flows and call on system resources. RapidMiner Studio also supports a variety of scripting languages.

Background process execution

- Execute multiple processes in parallel
- Long-running processes can be run in the background, while continuing to work on other process in the foreground for faster and more effective development iterations.
- Processes running in the background can be monitored. Results and logs can be reviewed once they are available.
- The maximum number of allowed processes running simultaneously can be configured to adapt to the hardware resources and the demand of the processes being executed. The default is the number of cores minus one.

Scripting

- Write scripts for easy-to-complex data preparation and transformation tasks where existing operators might not be sufficient
- Incorporate procedures from other processes or projects
- Develop custom models
- Augment scoring logic by custom post-processing or model application procedures
- Easy-to-use program development interface: Predefined imports for common data structures
- Syntactic sugar for simplified data access and alteration
- Interactive code editor and syntax high-lighting
- Execute command line programs and integrate results and result codes in processes
- Execution of SQL statements directly in database
- Seamless integration of the various programming languages into the RapidMiner Studio user interface: Execution of Groovy scripts within RapidMiner Studio processes
- Execution of OS scripts within RapidMiner Studio processes
- Execution of R scripts within RapidMiner Studio processes
- Execution of Python scripts within RapidMiner Studio processes
- Predefined scripted models & transformtions available as operators
- Custom scripts can be stored and executed as own operators within a process

Process Control

- Organize segments in sub-processes and reuse them in different projects
- Repeat execution over a segment of a process
- Support for loops
    - Loop (basic loop, with parallel execution of the iterations)
    - Attributes (parallel execution of the iterations)
    - Labels
    - Subsets
    - Values (parallel execution of the iterations)
    - Examples
    - Clusters
    - Batches
    - Data Sets
    - Data Fractions
    - Parameters
    - Files (parallel execution of the iterations)
    - Repository entries

Process Control Cont'd

- Branches (if-then-else) based on:
  - Data values
  - Attribute existence
  - Numbers of examples
  - Performance values
  - Existence of files and process inputs
  - Definition of macros
  - Arbitrary expressions
- Creation of collections of the same type
- Collection handling: selection, flattening, or looping
- Remembering and recalling (intermediate) process results for complex process designs
- Handling expected and unexpected errors and exceptions

Automatic Optimization

- Automatic selection of best performing sub processes
- Measuring the influence of preprocessing steps by nested cross validations / other validations
- Automatic selection of best model type and parameters
- Automatic selection of best attribute subsets
- Automatic optimization of process para-meters, including modeling parameters
  - Grid
  - Quadratic
  - Evolutionary

Macros

- Centralized definition of macros / variables containing arbitrary textual or numerical content
- Usage of macros everywhere in the process design, especially as value for parameters
- Macros can be defined during the process or in the process context
- Definition of macros in the context allows for parameterization of complete processes, e.g. for transforming processes into customizable web services
- Extraction of macro values from data values, meta data or statistics supported
- Expression engine for calculating arbitrary macro values from existing macros

Logging

- Logging can be introduced at arbitrary places within a process
- Logging can collect parameter values, performance values, or specific values for each operator, e.g. the current generation for evolutionary algorithms
- Data values can be logged
- Macro values can be logged
- Logged values can be transformed into several formats including: data sets and weights which can be stored, transformed, analyzed, or visualized like any other data set.

Process-Based Reporting

- In cases where logging alone is not sufficient, a complete process-based reporting engine allows for the collection of arbitrary results in static reports
- Different formats like PDF, Excel, HTML, or RTF supported
- Different reporting styles including a sequential report or portals
- Support of sections with up to 5 levels
- Arbitrary process results as well as intermediate results can be transformed into different types of visualizations like tables, charts etc.
- Support for page breaks and other style information
- Combination with loops or other process control structures allows for highly-detailed result overviews even for complex process designs