Index

Note: Page numbers followed by "b", "f" and "t" indicate boxes, figures and tables respectively

A

AdaBoost model, 158–159, 159f
ANNs. *See* Artificial neural networks (ANNs)
Anomaly detection
  causes
    click fraud detection, 331b
    credit card transaction fraud monitoring, 331
    data errors, 330
    distributional assumption, 330
    distribution classes, 330
    normal data variance, 330
  characteristics, 332–333
  classification technique, 334
  clustering, 334
  computer network traffic, 329
  density based outlier, 333, 339, 339f–341f
  distance based outlier, 333, 333f
    data preparation, 336–337, 337f
    Detect Outlier (Distances) operator, 337
    distance score, 334–335
    Euclidean distance, 334
    Iris data set, 335, 336f
    k-NN classification technique, 334–335
    Outlier detection output, 338, 338f
    two-dimensional scatterplot, 334, 335f
    unsupervised outlier detection operator, 336
  distribution based outlier, 333–334
  distribution model, 332
  generalized classification model, 344
  high-volume transaction networks, 344
  LOF technique
    minimal points, 342
    output, 342–344, 343f
    RapidMiner process, 342, 343f
    relative density, 341–342
  outlier, 329
  sorting function, 332
  statistical methods
    Mahalanobis distance, 332
    standard normal distribution, 332, 333f
  stratified sampling method, 344–345
  supervised and unsupervised techniques, 414
Anscombe's quartet, 45–46, 46f
Apriori algorithm, 202–206, 202f–203f
  clickstream data set, 204, 204t
  frequent item set support calculation, 204, 205t
  rule generation, 206
  support, 204, 205f
Area under the curve (AUC), 263, 270, 272f
Artificial intelligence, 4–5
Artificial neural networks (ANNs), 13, 18, 126f
  activation function, 126–127
  advantages, 134
  aggregation function, 124–125
  AutoMLP, 130–131
  back propagation, 127
  biological neurons, 125f, 125b
  data preparation, 131
  error calculation, 128–129, 128f
  evaluation, 130f, 132
  hidden layers, 126
  linear mathematical model, 124
  missing values, 133
  modeling operator and parameters, 131–132
  model topology, 124, 124f
  optical character recognition, 127b
  perceptron, 124–125
  performance vector, 132, 133f
  simple aggregation activation function, 128
  transfer function, 124–125
  universal approximator, 126–127
  weight adjustment, 129–130, 129f
Association analysis
  antecedent and consequent, 196
  Apriori algorithm, 202–206, 202f–203f
    clickstream data set, 204, 204t
    frequent item set support calculation, 204, 205t
    rule generation, 206
    support, 204, 205f
  association rules, 197
  confidence, 198–199
  conviction, 198, 200
  cross selling, 196b
  FP-Growth algorithm, 208
    conditional FP-tree, 210, 210f
    data preparation, 211–215, 211f
    frequent paths, 207
    modeling operator and parameters, 212–213, 212f
    rules creation, 213
    transactions 1, 2, and 3, 207, 208f
    transactions list, 207, 207t
    trimmed FP-tree, 209–210, 209f

Association analysis (*Continued*)
lift, 198–200
market basket analysis, 195–196
rule generation process, 200–202, 201f
support, 198–199
Automatic Multilayer Perceptron (AutoMLP), 130–131

B

Bagging model, 156–157, 157f
Bayesian Information Criterion (BIC), 232–234
Bayesian probabilistic theories, 4–5
Big Data, 7
Blog gender classification, 284
key features identification, 296–297, 297f
LibSVM(linear) operators, 297–299
preprocessing text data, 293, 295f
process documents, 293, 295f
Read Excel operator, 293, 294f
test data preparation, 299, 300f
training and testing predictive models, 297–299, 298f
unstructured data, 290t–292t, 288–293
W-Logistic operators, 297–299
X-Validation operator, 297–299, 298f, 299t
Bootstrap aggregating/bagging, 154–155
Boston Housing data set, 170, 171t–172t

C

Categorical data types, 40
Chi-square-based filtering
attribute weighting, 362, 363f
contingency table, 361–362, 362t
expected frequency table, 361–362, 362t
observed *vs.* expected frequencies, 362
Classification
ANNs. *See* Artificial neural networks (ANNs)
categorical target variable prediction, 408–410
classes, 63
decision trees. *See* Decision trees
ensemble learners. *See* Ensemble learners
K-nearest neighbors. *See* K-nearest neighbors (k-NN)
naïve bayesian. *See* Naïve bayesian
rule induction. *See* Rule induction
SVMs. *See* Support vector machines (SVMs)
Clustering, 286, 288f–289f
DBSCAN clustering, 240f. *See* Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering
dimensionality reduction, 218
document clustering, 218
k-means clustering, 230f. *See also* k-means clustering.
object reduction, 219
SOM, 242–243, 242f. *See also* Self-organizing map (SOM)
types
customer records segmentation, 222t, 222b
DBSCAN, 221
density clustering, 221
Euclidean distance measurement, 219
exclusive/strict partitioning clusters, 219
fuzzy/probabilistic clusters, 220
hierarchical clustering, 219, 221
model-based clustering, 221
overlapping clusters, 219
prototype-based clustering, 220–221
SOM, 221
Confusion matrix, 189–190, 189f
Cross Industry Standard Process for Data Mining (CRISP-DM), 17–18, 18f

D

Data discovery techniques, 5
Data exploration
data, definition, 37
data preparation, 38
data sets
categorical/nominal, 40–41
numeric/continuous, 40
types, 40–41
visual exploration, 39–40
data understanding, 38
data visualization, 37
Andrews curves, 57–59, 59f
box whisker plot, 48–49
bubble chart, 55, 55f
class-stratified histogram, 48, 49f
class-stratified quartile plot, 49, 50f
cognitive thinking, 47
density charts, 55, 56f
deviation chart, 57, 58f
distribution chart, 47–49, 48f
histogram, 48, 48f
parallel chart, 56–57, 57f
quartile plot, 49, 50f
scatter matrix plot, 53–55, 54f
scatter multiple plot, 53, 53f
scatterplot, 52–53, 52f
descriptive statistics, 37
multivariate exploration. *See* Multivariate exploration
univariate exploration. *See* Univariate exploration
Data mining process, 19f
anomaly detection algorithm, 15
application
assimilation, 34
production readiness, 32–33
remodeling, 34
response time, 33
technical integration, 33
artificial neural network, 18
automated clustering, 18
classification model, 4–5, 10, 13
CRISP-DM, 17–18, 18f
data preparation
EDA, 23
feature selection, 26
missing values, 24
outliers, 25–26
pivot/transpose functions, 22–23
quality, 24
sampling, 26–27
transformation, 25
types and conversion, 25
descriptive/explanatory modeling, 18–19
DMAIC, 17–18
modeling, 28f
abstract data representation, 27
classification, 27
decision tree techniques, 28
ensemble modeling, 31–32
model evaluation, 31, 31t
regression model, 30, 30f
simple linear regression technique, 28

test data set, 28, 29f, 29t
training data set, 28, 29f, 29t
predictive modeling, 18–19
prior knowledge, 19
attribute, 22
causation *vs.* correlation, 22, 22t
data point/record/data object, 21
data set, 21, 21t
identifiers, 22
label, 22
objective, 20
subject area, 20–21
quantitative analysis, 35
RapidMiner software, 12. *See also* RapidMiner
SEMMA, 17–18
SOM, 14–15
types
anomaly/outlier detection, 9. *See also* Anomaly detection.
classification, 9. *See also* Classification
clustering, 9. *See also* Clustering
market basket analysis, 9. *See also* Association analysis
regression techniques, 9. *See also* Regression methods
supervised/unsupervised learning models, 8, 10
time series forecasting, 9. *See also* Time series forecasting
wrapper-type methods, 15–16
Data storage, 6
Data transformation tools
Append operator, 391–392
data type conversion operators, 387
De-pivot operator, 388
dichotomization, 386
discretization output, 387–388, 390f
discretize operator, 387–388, 389f
Join operator, 391–392
pivot tables, 388, 390f–391f
Davies-Bouldin index, 229
Decision trees, 13, 28
accuracy, 83
advantages, 87–88
aggregate measures, 86
baseline model performance measures, 85, 85f
credit default identification process, 85–86, 86f
credit scoring, 72
data preparation
attribute value replacement, 76–77, 78f
data transformation, 77–79, 80f
German credit data, 74–75, 75t
disadvantages, 88
entropy, 65, 65f
uncertainity reduction, 65b
gain ratio, 82–83
Gini index, 83
definition, 66
Golf data set, 66, 67t, 69–70, 70f
information gain, 68, 69t
split data, 66
subsets/branches, 69, 69f
information gain, 82, 88
Meta Cost, 85
minimal gain value, 83
minimal leaf size, 85
overfitting, 70
post-pruning, 70–71
pre-pruning, 70–71
prospect filtering, 72–73
prospect scoring data, 83–84, 84f
pruning, 70–71
regression trees, 64
scale normalization, 87
Shannon entropy, 71
splitting data, 73
supervised learning algorithm, 74
target variable, 64
Dendrogam, 221
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering
border points, 236, 236f
center-based density, 234, 235f
centroid methods, 242
core points, 236, 236f
data preparation, 238
density-clustering algorithm, 234
epsilon and MinPoints, 236
evaluation (optimal), 239
high-density and low-density space, 235
k-distribution chart, 237, 238f
noise points, 236, 236f
operator and parameters, 239
prototype-based clustering, 220–221
varying densities, 237, 238f
visual output, 239–241, 241f
Density-clustering algorithm, 221, 234
Descriptive statistics, 5, 37
characteristics, 41
multivariate exploration. *See* Multivariate exploration
univariate exploration. *See* Univariate exploration
Dimensional slicing, 5
Dimensions, 7

E

Ensemble learners, 14
AdaBoost model, 158–159, 159f
aggregate hypothesis/model, 148
Bagging meta model, 156–157, 157f
Bagging operator, 155–156, 156f
boosting, 157–158
bootstrap aggregating/bagging, 154–155
conditions, 151–152
drought prediction, 150b
error rate, 151
generalization error, 162
meta learning, 148–149
probability mass function, 150
Random Forest operator, 160–161, 161f
voting, 153–154, 153f–155f
Ensemble modeling, 31–32
Euclidean distance, 102–105
Exclusive/strict partitioning clusters, 219
Exploratory data analysis (EDA), 23. *See also* Data exploration

F

Feature selection, 26
attributes selection, 415–416
chi-square-based filtering
attribute weighting, 362, 363f
contingency table, 361–362, 362t
expected frequency table, 361–362, 362t
observed *vs.* expected frequencies, 362
rank attributes, 362, 363f
dimension reduction, 347, 370
filter type, 347
information theory

Feature selection (*Continued*)
information exchange, 358
numeric Golf data set, 358–359, 360f
motivation, 348b
multicollinearity, 348
multiple regression, 348
naïve Bayesian classifiers, 348
PCA, 349, 357f. *See also* Principal component analysis (PCA)
remove independent variables, 348
types, 347
wrapper type feature selection, 347
aggressive feature selection, 369, 369f
Backward Elimination operator, 364–365, 367f
computational resource consumption, 364
forward selection, 364
maximal relative decrease, 368
permissive feature selection, 369, 369f
preset stopping criterion, 369–370
regression model, 364, 364t
Split Validation operator, 368
squared correlation, 368
Frequent Pattern (FP)-Growth algorithm, 208
conditional FP-tree, 210, 210f
frequent paths, 207
modeling operator and parameters, 212–213, 212f
results interpretation, 213–215, 213f–214f
rules creation, 213
transactions list, 207, 207t
trimmed FP-tree, 209–210, 209f
Fuzzy/probabilistic clusters, 220

G

Gain curves, 264, 268f

H

Hamming distance, 102–105
Hierarchical clustering, 219, 221
Hypothesis-driven techniques, 7–8
Hypothesis testing, 5–6

I

Integer data type, 40
Iterative algorithms, 4–5

K

Keyword clustering
Crawl Web operator, 285
document-clustering problem, 284
Get Pages operator, 285
k-medoids operator, 286, 288f
medoid clustering, 284–285
process, 286, 289f
unstructured data, 285–286, 285f
k-means clustering
BIC, 232–234
centroid prototype approach, 234
centroids output, 231, 232f
cluster centroid/mean data object, 223
Cluster Distance Performance operator, 231
cluster label, 229–230
Davies-Bouldin index, 229
empty clusters, 228–229
Euclidean distance, 224
evaluation parameter, 229
initiation, 228
labeled example set, 232
limitations, 232–234
local optimum, 227–228
new centroids location, 226, 226f–227f
operator and parameters, 231
outliers, 229
performance criterion, 227–228
performance vector, 232, 233f
postprocessing, 229
prototype-based clustering and boundaries, 224, 225f
prototype data point, 223
sum of squared errors, 226
termination, 227
visual output, 232, 233f
Voronoi partitions, 223, 223f
K-nearest neighbors (k-NN), 334–335
eager learners, 99
forest type prediction, 100b
lazy learners, 99, 108
eager learners, 111
execution and interpretation, 109f–110f, 110
modeling operator and parameters, 108–109
nonparametric method, 99
proximity measure
correlation similarity, 106
cosine similarity, 107
distance, 102–105, 104f
Jaccard similarity, 106–107
SMC, 106
unseen test record, 102

L

Linear regression
average error, 169
data separation, 172, 173f
dependent and independent variable, 167–168
feature selection option, 173, 174f
gradient descent, 169
"greedy" feature selection, 174–175, 176f
linear regression operator, 173, 174f
median price, 169
model validity, 180
null hypothesis, 177–178
p-values, ranking variables, 177–178, 177f
RapidMiner, 170, 171t–172t
simple regression model, 167–168, 168f
split validation operator, 172–173, 173f
squared correlation, 177, 177f
unseen test data, 178–179, 179f
"wrapper" functions, 176
Linear regression model, 7
Local outlier factor (LOF) technique
Binominal operator, 342–344
data preparation, 342
Detect Outlier operator, 342
minimal points, 342
output, 342–344, 343f
RapidMiner process, 342, 343f
relative density, 341–342
Logistic regression, 1–2
binomial response variable, 191
confusion matrix, 189–190, 189f
credit scoring exercise, 188
data preparation, 188
likelihood function, 185
linear model, 182, 183f
logit function, 180–181, 181f, 184
MetaCost operator, 190–191, 190f
modeling operator and parameters, 188, 189f
nonlinear curve, 182, 183f
nonlinear optimization techniques, 185

odds ratio analysis, 186b
parameters, 184
"sigmoid" curve, 182
S-shaped curve, 182
SVM, 191, 191f
Titanic wreck, 186b, 187f

M

Machine learning, 4–5
Mahalanobis distance, 332
Manhattan distance, 102–105
Meaningful patterns extraction, 3
MetaCost operator, 190–191, 190f
Minkowski distance, 102–105
Mixture of Gaussians, 221
Model-based clustering, 221
Model evaluation
AUC, 263, 270, 272f, 273
classification performance metrics, 264–267, 269f
confusion matrix/truth table, 257, 259t
accuracy, 259
binary/binomial classification, 258
definition, 258
evaluation measures, 260, 260t
precision, 259
recall, 259
relevance, 259
sensitivity, 258–259
specificity, 259
data partitioning, 267
Direct marketing (DM), 257b–258b
evaluation, 267
Generate Direct Mailing Data operator, 264–267
Lift Chart operator creation, 267
lift charts, 257, 270, 271f
by RapidMiner, 270, 272f
lift curves, 263–264, 266f, 268f
modeling operator and parameters, 267
performance operator, 267
ROC curves, 257, 260, 262–263, 262f, 266f, 270, 272f, 273
Split Data operator, 267
Split Validation operator, 267
Moore's Law, 1–2
Multiple linear regression (MLR), 170
Multivariate exploration
central data point, 44
correlation
Anscombe's quartet, 45, 46f
Cartesian coordinate, 45
Pearson correlation coefficient, 44–45, 45f
quadratic functions, 45

Q

Queries, 6

R

Random Forest operator, 160–161, 161f
RapidMiner
attributes, 375–376, 376f
data importing and exporting tools
CSV file, 377–379, 380f
data import wizard, 379, 381f–382f
Import Configuration Wizard, 377–379
data scaling and transformation tools, 371
data set, 375–376
data transformation tools
Append operator, 391–392
data type conversion operators, 387
De-pivot operator, 388
dichotomization, 386
discretize operator, 387–388, 389f
Join operator, 391–392
label variable, 386–387
logistic regression, 386–387
machine learning algorithms, 386
pivot tables, 388, 390f–391f
data types, 375–376
data visualization tools, 383
bivariate plots, 383–386
multivariate plots, 386
results, 382–383, 384f
Statistics, 383, 385f
univariate plots, 383
decision tree, 376, 377f
example set, 375–376
graphical user interface
RapidMiner 6.0, 372, 373f
views, 372, 374f
operator, 376
optimization tools
attributes upper bound and attributes lower bound parameters, 399
configuration, 397, 399f
disadvantage, 402
Generate Data operator, 397
genetic search optimization, 402, 404f
grid search optimizer, 399–400, 400f–401f
inner process, 397, 398f
mutation and cross-over, 402
"nested" operator, 396
Optimize operator, 396
Optimize parameters, 397
polynomial function, 396–397, 397f
quadratic greedy search optimization, 402, 403f
process, 377, 378f
RapidMiner Studio GUI, 371
repository, 375, 375f
sampling and missing value tools
balanced accuracy, 394
balancing data sets, 392
bootstrapping, 395
imbalanced data set, 392, 393f
rebalance subprocess, 394, 395f
replace missing values, 395–396
unbalanced data, 392–394, 394f
YALE, 371
Ratio data type, 40
Read Excel operator, 75, 76f
Receiver operator characteristic (ROC) curves, 257, 260, 262–263, 262f, 266f, 270, 272f, 273
Regression methods, 4–5, 30, 30f
feature selection methods, 165–166
function fitting, 165–166
linear regression, 165, 174, 175f. *See* Linear regression
logistic regression, 165. *See also* Logistic regression
RapidMiner, 165–166
Rename operator, 77–79
Repeated Incremental Pruning to Produce Error Reduction (RIPPER), 91
Replace (Dictionary), 76–77, 78f

Representative models, 3–4, 4f
Rule induction, 13
  antecedent/condition, 89
  class selection, 91
  conjunct, 89
  data preparation, 94–95
  Decision Tree operator, 97, 98f
  disjunct/classification rule, 89
  exhaustive rule set, 90
  Golf data set, 88, 90f
  learn-one-rule, 93
  modeling operator and parameters, 95, 96f
  mutually exclusive set, 89–90, 98–99
  results interpretation, 95–97, 96f
  RIPPER, 91
  rule generation, 89f, 90–93, 92f
  rule set, 89
  sequential covering approach, 91
  split conditions, 88
  Tree to Rules operator, 97, 97f

S

Sample, Explore, Modify, Model, and Assess (SEMMA), 17–18
Self-organizing map (SOM), 14–15
  centroid update, 244–246, 245f–246f
  country data set, 248f, 249
  data preparation, 249
  data transformation, 247, 247f
  execution and interpretation, 251
  grid space, 243
  initialization, 244
  location coordinates, 252, 254f
  modeling operator and parameters, 250–251, 250f
  neural network, 243–244
  termination, 246
  topology specification, 244
  visual model, 251–252, 252f–253f
Shannon entropy, 71
Simple matching coefficient (SMC), 106
Simple regression model, 167–168, 168f
Stochastic model, 7–8
Subject matter expertise, 4
Support vector machines (SVMs), 14, 191, 191f
  advantages, 148
  boundary, 136, 136f
  complex nonlinear dataset, 134
  disadvantage, 147
  hyperplane, 135, 135f
  Kernel functions, 138
  linearly separable, 136, 137f
  margin, 136, 136f
  penalty, 136, 136f
  prediction accuracy, 144, 145f
  quadratic polynomial, 138
  Scatter 3D Color plot, 145, 146f
  support vectors, 135
  two-ring nonlinear problem, 145–147, 147f

T

Term frequency-inverse document frequency (TF-IDF), 277–279
Text mining
  clusters, 284
  customer relationship management software, 276
  data warehousing and business intelligence, 276
  IBM's Watson program, 276b
  key features identification, 296–297, 297f
  keyword clustering
    Crawl Web operator, 285
    data preparation, 286, 287f
    document-clustering problem, 284
    Get Pages operator, 285
    k-medoids operator, 286, 288f
    medoid clustering, 284–285
    unstructured data, 285–286, 285f
    website keyword clustering process, 286, 289f
  Lexical substitution, 280–282
  meaningful n-grams, 283, 283f
  preprocessing operator, 293, 295f
  preprocessing steps, 283, 283t
  preprocessing text data, 293, 295f
  Read Excel operator, 293, 294f
  similarity mapping, 279
  stemming, 282–283
  stopword filtering, 280, 282f
  stopwords, 280
  term filtering, 280–282
  term frequencies, 281t
  test data preparation, 299, 300f
  TF–IDF, 277–279
  token, 279
  tokenization, 279–280
  trained models, 299–302, 300f–301f
  unstructured data, 288–293, 290t–292t
  X-Validation operator, 297–299, 298f, 299t
Time series forecasting, 308f
  autocorrelation, 306
  cross-sectional data, 305, 306f
  data-driven forecasting methods, 305–306
  decomposition, 306–307
  descriptive modeling, 306–307
  forecasting demand, 307f, 307b
  Holt's two-parameter exponential smoothing, 311–312
  Holt-winters' three-parameter exponential smoothing, 312–313
  model-driven forecasting methods, 306, 314f
    autoregression models and ARIMA, 316t, 317–318
    independent variables, 321
    Inner level process, 323–324, 325f
    label variable, 320, 322f
    limit time box, 323–324
    linear regression, 313–317, 315f, 316t
    polynomial regression, 313–314, 315f
    prediction horizon controls, 318–319
    regression equation, 321–322
    set iteration macro, 323–324
    Set Role operator, 319–320
    step size, 320
    Windowing operators, 318–320, 320f
    windowing transformation, 317f–319f, 318, 320, 321f
    Window size, 320
  multiple linear regression model, 305
  naïve forecast. *See* Naïve forecast
  neural network model, 305
  notation system, 308
  predictive modeling, 306–307
  predictor variables, 305
  supervised model, 305
Trend,-seasonality, 311, 312f

U

Univariate exploration, 41, 42t
  descriptive statistics, 43, 43f
  deviation, 43
  mean, 42
  median, 42
  mode, 42
  range, 43
  standard deviation, 43
  variance, 43
Unsupervised process, 413

V

Vote meta modeling operator, 153–154, 154f

W

Wrapper type feature selection, 15–16, 347
  aggressive feature selection, 369, 369f
  Backward Elimination operator, 364–365, 367f
  forward selection, 364
  maximal relative decrease, 368
  multiple regression model, 365
  permissive feature selection, 369, 369f
  preset stopping criterion, 369–370
  Split Validation operator, 368

Y

Yet Another Learning Environment (YALE), 371