Index

Note: Page numbers in *italics* refer to Figures; those in **bold** to Tables.

acquisition modeling
- pilot campaign, 95–96, 96
- profiling of high-value customers, 96–97, 97

AIC see Akaike Information Criterion
Akaike Information Criterion, 211

association (affinity) models
- apriori and FP-growth algorithms, 13
- market basket analysis, 13
- rule’s confidence, 12–13
- sequence algorithms, 14

Bagging (Bootstrap aggregation)
- Decision Tree model, 314, 314
- in IBM SPSS Modeler, 61
- in RapidMiner, 61, 62

balancing approach
- balance factor, 49, **50**
- cross-selling, 49, **50**
- in Data Mining for Excel, 53, 53, 54
- disproportionate stratified sampling, 48
- in IBM SPSS Modeler, 49, 51, 52
- oversampling, 49
- in RapidMiner, 50, 52, 53

Bayesian belief networks
- CPT see conditional probability table
- IBM SPSS Modeler, 180, 181, **182**
- Microsoft Naïve Bayes, 182, **183**, 183
- Naïve Bayes models, 176–177
- parent variables, 176
- RapidMiner Naïve Bayes, 184
- Tree Augmented Naïve Bayes models, 177–180

Bayesian Information Criterion, 204, 211

Bayesian networks, 9, 88, 170–171, 191, 300, 302

behavioral segmentation methodology, **114**
- business objective, definition, 115

cluster modeling see cluster modeling, identification of segments
CRISP DM methodology, phases, 113

customer segmentation, 112–113

data exploration and validation, 118–119

data integration and aggregation, 118

data transformations and enrichment, 122–124

deployment of segmentation solution
- customer scoring model, 140–141
- design and deliver of differentiated strategies, 142

distribution of segmentation information, 141
input set reduction, 124

investigation of data sources, 117
modeling process design, 115–117

revealed segments
- marketing research information, 138, 139
- optimal cluster solution and labeling segments, 139
- profiling of, 132–138
- technical evaluation of clustering solution, 128–132

selecting data, 117–118

BIC see Bayesian Information Criterion

BIRCH algorithm, 211

boosting
- Adaptive Boosting (AdaBoost), 62
- in IBM SPSS Modeler, 62–63, 63
- in RapidMiner, 63

Bootstrap validation method, 47–48

Call Detail Records, 333

candidate model
- CHAID see CHAID model

CHURN model, 246, 246
ensemble model, 249, 250
Gains chart, 244, 245
candidate models, performance measures, 243, 245
CART see classification and regression trees
CDRs see Call Detail Records
C5.0/C4.5
  information gain ratio index
  of child nodes, 155
  Information or Entropy of node N, 155
  2-level C5.0 Decision Tree, 157, 157
  of root node, 155
  split on profession, 155–157, 156
predictors, handling of, 148–149
CHAID model
  and C5.0, 244
  chi-square test see chi-square test, CHAID
  churn scores, 243–244
  and ensemble model, 249
  handling of predictors, 148–149
  parameter, 244
  right branches, 248
  in rules format, 248, 248
  spending frequency, 246
  in tree format, 246, 247
chi-square test, CHAID
  cross-tabulation of target, 158, 159
  Decision Tree algorithm, 158
IBM SPSS Modeler CART
  advanced options, 165, 166
  basics options, 165, 166
  parameters, 165, 167-168
  stopping criteria, 165, 167
IBM SPSS Modeler C5.0 Decision Trees, 164, 164, 164–165
IBM SPSS Modeler CHAID Decision Trees
  advanced options, 160, 161
  2-level, 159, 160
  parameters, 160, 162–163
  stopping rules, 160, 162
independence hypothesis, 158
Microsoft Decision Trees parameters, 170, 170, 171
Pearson chi-square statistic, 158
predictors for split, 159, 160
p-value or observed significance level, 158–159
RapidMiner Decision Trees
  parameters, 168, 169, 169
  recursive partitioning, 165, 168
classification algorithms
  Bayesian belief networks, 176–182, 184
  Bayesian networks, 170–171
  chi-square test, CHAID, 158–160, 164–165, 168, 170
  data mining algorithms, 145
  Decision Tree models, 146–150
  Gini index, CART, 151, 153–154
  information gain ratio index, C5.0/C4.5, 155–157
  Naive Bayesian networks, 172–174
  support vector machines, 184–190
classification and regression trees
  Gini index, 151, 153–154
  handling of predictors, 148
classification modeling methodology, 24, 25
  acquisition modeling, 95–97, 96, 97
  business understanding and process design, 24–33
  combining models, 60–64
  CRISP-DM phases, 24
  cross-selling modeling, 98–100, 99, 100, 101
  in Data Mining for Excel, 57, 59
  data preparation and enrichment, 33–57
  deep-selling modeling, 102–105, 103, 104, 105
  direct marketing campaigns, 94–95
  form of supervised modeling, 23
  in IBM SPSS Modeler, 57, 58
  likelihood of prediction, 23
  meta-modeling, 57
  model deployment, 88–94
  model evaluation, 64–88
  product campaigns, optimization, 100–102, 101, 102
  in RapidMiner, 59, 59–60
  up-selling modeling, 105–108, 106, 107, 108
  voluntary churn modeling, 108–110
classification or propensity models
  Bayesian networks, 9
decision rules, 8
  Decision Trees, 7–8
  logistic regression, 8
  neural networks, 8–9
  support vector machine, 9
class weighting
  class-imbalanced modeling file, 53–54, 55, 57
  in IBM SPSS Modeler, 54, 56
  in RapidMiner, 55, 56
clustering algorithms
  classification modeling, 203
  Expectation Maximization clustering, 203
  with K-means see K-means, clustering algorithms
  with TwoStep, 211–212
cluster modeling, identification of segments
  agreement level, 126
  in Data Mining for Excel, 127, 128
  in IBM SPSS Modeler, 126, 126–127
profiling
  cluster centroid, 133, 133
  cluster separation, 133–134
  in Data Mining for Excel, 135, 136, 137, 138
  effective marketing strategies, development, 132
  in IBM SPSS Modeler, 134, 135, 136
  in RapidMiner, 134, 134
  table of cluster centers, 133, 133
  in RapidMiner, 127, 127
revealed segments
  cohesion of clusters, 129–130
  descriptive statistics and technical measures, 128–129
  in IBM SPSS Modeler, 131, 131
  in RapidMiner, 132, 132
  separation of clusters, 130–131
cluster models, 10
  agglomerative or hierarchical, 11–12
  data preparation
    data miners of organization, 341
    data reduction technique, 340
    explained variance, 341, 342
    interpretation results, 341, 345
    principal components analysis model, 340–341, 341
    rotated component matrix, 341, 343
  expectation maximization clustering, 12
  identifying segments
    cluster models, comparison, 342, 346
    parameter settings, 342, 346
    revealed clusters, distribution, 342, 347
    Silhouette measure, 342, 346
K-means, 12
K-medoids, 12
Kohonen network/self-organizing map, 12
profiling
  behavioral profile, 352–353
  distribution of factors for Cluster 5, 344, 348
  profiling chart, 349, 350, 351, 352
  profiling of clusters, 344, 348, 349
structures, 349
  table of centroids, 344, 347
RapidMiner process, 354, 354
TwoStep cluster, 12
conditional probability table, 176, 177, 306
  of gender input attribute, 178, 179
  probabilities of output, 178, 178
  of profession input attribute, 178, 178
  of SMS calls input attribute, 178, 179
  of voice calls input attribute, 178, 178
confusion matrix and accuracy measures
  in Data Mining for Excel, 70, 71
  error rate, 67
  F-measure, 67
  in IBM SPSS Modeler, 68, 68, 69
  misclassification or coincidence matrix, 66, 66
  Performance operator, 69, 69, 70
  Precision measure, 67
  in RapidMiner, 68, 69
  Recall measure, 67
  ROC curve, 67
  sensitivity and specificity, 67
CPT see conditional probability table
CRISP-DM see Cross Industry Standard Process for Data Mining process model
CRM see customer relationship management
Cross Industry Standard Process for Data Mining process model
  business understanding, 4
  data preparation, 4
  data understanding, 4
  deployment, 5
  evaluation, 4–5
  modeling, 4
  phases, 4–6, 5, 24
Cross or n-fold validation method
  in Data Mining for Excel, 45, 47
  modeling dataset, 45, 46
  n iterations, 45
  in RapidMiner, 45, 46, 47
cross-selling modeling
browsing the model
  C5.0 model, 304, 305, 305
  CHAID model, 301–303, 306
  CPTs, 306
  FLAG_GROCERY attribute, 307, 308
  Gains chart, 302–303, 304
  gains chart of ensemble model, 308, 308
  performance metrics for individual models, 301, 304
response rate, 304–305
ROI chart, 307, 307
TAN, 305–306, 306
campaign list
  Modeler deployment stream, 308, 309
  scored customers, estimated fields, 308–309, 309
Data Mining for Excel, 329, 331
  Accuracy Chart wizard, 327, 328
  campaign response, 322–323
  classification algorithm and parameters, 320, 321
  Classification Matrix wizard, 325, 326
  Classify Wizard, 319–320, 320, 321
  confusion matrix, 327, 327
cumulative percentage of responders, 329, 329, 330
Decision Tree model, 322–323, 323, 324, 325, 325, 327, 327
dependency network of BDE tree, 322, 323
Gains charts for two Decision Tree models, 327, 328
model deployment, 329, 330, 331
Split (Holdout) validation, 322, 322, 323
validation dataset, 325, 326
validation of model performance, 325, 327, 329
  development of, 295
  mining approach, 296–297
modeling procedure
  IBM SPSS Modeler procedure, 296, 297
  setting roles of attributes, 299–300
  Split (Holdout) validation, 298–299
  test and loading campaign responses, 298
model deployment, 329, 330, 331
parameters, 300, 303
  pilot campaign, 98, 99
  product uptake, 98–99, 100
  profiling of owners, 99–100, 101
modeling procedure
  IBM SPSS Modeler procedure, 296, 297
  setting roles of attributes, 299–300
  Split (Holdout) validation, 298–299
closure date and reason, 221
card level, voluntary churn model, 221, 222
card usage attributes, 221
carding procedure
  assessment of data quality, 36–38
  continuous (range) attributes, 35–36
tool of IBM SPSS Modeler, 36
data exploration
  assessment of data quality, 36–38
categorical attributes, 36
  continuous (range) attributes, 35–36
tool of IBM SPSS Modeler, 36
data management procedure, churn model
  from cards to customers
    closing date, 224
  cards’ closing dates, 223
  filtering out cards, 224, 225
  flagging cards, 224, 224
  IBM SPSS Modeler node, 226, 227
  initial card usage data, 222, 223
  open at end, observation period, 224, 225
customer relationship management
  customer development, 1
customer satisfaction, 1
data mining, 20
customer scoring model
  in Data Mining for Excel, 140–141
  Decision Tree, 141
customer scoring model
  in IBM SPSS Modeler, 140, 140
  in RapidMiner, 140, 141
customer segmentation
  behavioral, 112–113
customer scoring model
  definition, 112
  loyalty based, 113
  needs/attitudinal, 113
  propensity based, 113
  sociodemographical, 113
  value based, 112
customers grouping, value segmentation
  binning node, 287, 288
  binning procedure, 287, 287
data dictionary
  card level, voluntary churn model, 221, 222
customers’ closing dates, 223
data enrichment
  customer signature, 39
data reduction algorithm, 39
  feature selection, 39
  informative KPIs, 38–39
  naming of attributes, 40
data exploration
  assessment of data quality, 36–38
categorical attributes, 36
  continuous (range) attributes, 35–36
tool of IBM SPSS Modeler, 36
data management procedure, churn model
  from cards to customers
    cards’ closing dates, 223
    filtering out cards, 224, 225
    flagging cards, 224, 224
    IBM SPSS Modeler node, 226, 227
    initial card usage data, 222, 223
    open at end, observation period, 224, 225
total number of transactions, calculation, 226, 226
data management procedure, churn model (cont’d)
enrichment, customer data
  balances, 229
deltas, spending, 228, 231
limit ratios, 229
monthly average number, transactions, 229
spending amount, monthly average, 227
spending frequency, 228
spending recency, 227–228, 230
tenure, customer, 227
trends, card ownership, 228, 232
modeling population and target field
  defined, 231, 233
latency period, 231
REFERENCE DATE, 230
in scoring phase, 230–231
selection, 231, 232
short-term churners, 231, 233
reference, customer data
balances, 229
deltas, spending, 228, 231
limit ratios, 229
monthly average number, transactions, 229
spending amount, monthly average, 227
spending frequency, 228
spending recency, 227–228, 230
tenure, customer, 227
trends, card ownership, 228, 232
modeling population and target field
  defined, 231, 233
latency period, 231
REFERENCE DATE, 230
in scoring phase, 230–231
selection, 231, 232
short-term churners, 231, 233
data mining
  algorithms, 145
CRISP-DM, 4–6, 5
CRM strategy, 1
customer life cycle management, 1–2
customer segmentation, 2
datamart, 16, 17, 18, 19
direct marketing campaigns, 2–3, 3
market basket and sequence analysis, 4
marketing reference table, 19–20
personalized customer handling, 15
required data per industry, 16
supervised models, 6–10
unsupervised models, 10–15
Data Mining for Excel
  balancing approach, 53, 53, 54
churn model
  accuracy and error rate, 263, 265
  approaches, 256
  churners and nonchurners, cumulative
distribution, 263, 267
classification algorithm, 259
Classify Wizard, 258, 258–259, 259
confusion matrix, 263, 265
Decision Tree model, 261, 262, 263, 265
mining structure, storing, 261
Query wizard, 265, 267
scored customers and model derived
  estimates, 265, 268
Split (Holdout) validation, 259, 260
validation, performance, 261, 264
classification modeling methodology, 57, 59
cluster modeling, 127, 128, 135, 136,
  137, 138
confusion matrix and accuracy measures,
  70, 71
Cross or n-fold validation method, 45, 47
in cross-selling model see cross-selling
modeling
customer scoring model, 140–141
Gains/Response/Lift charts, 78, 78, 79
K-means, 208, 209, 210
Naive Bayesian networks
  Classify Wizard, 174, 174
  Dependency network, 175, 175
receiver operating characteristic curves, 82,
  84, 85, 86
scoring customers, 91, 92, 93
Split (Holdout) validation method, 43, 44
data preparation procedure
  aggregating at customer level, 269, 271,
  275, 275
  aggregating at transaction level
    adding demographics using merge node,
    276, 276
    at customer level, 275, 275
    invoice level, 269, 271
  categorizing transactions into time zones,
    273, 274
classification modeling, tasks, 33–34
customer level, usage aspects, 268
data exploration and validation, 35–38
data integration and aggregation, 35
data transformations and enrichment, 38–40
enrich customer information
  average basket size, 277–278, 279
  basket diversity, 278
customer tenure, 276
deriving new fields, 272
flags of product groups, 278
frequency, 279, 280
monetary value, 279, 280
monthly average purchase amount, 277
ratio of transactions, 277
recency, 279, 279
relative spending, 276–277, 277
IBM SPSS Modeler Derive nodes, 272
imbalanced outcomes, 48–57
initial transactional data, 268, 269
investigation of data sources, 34
Modeler datetime_weekday() function, 273
pivoting transactional data, 269, 270
payment type, 273, 274
series of Restructure nodes, 274, 275
selecting data sources, 34
validation techniques, 40–48
data transformations
  event outcome period, 38
  label or target attribute, 38
  optimal discretization or binning, 38
  data validation process, 38

Decision Tree model, 7–8, 88, 147, 191, 300, 302
  algorithms, 158
    attribute selection method, 146
    classes of target attribute, 146
    decision rules, 147–148
    handle predictors, 146
    root node and, 148
    user-specified terminating criteria, 146
  with bagging, 314, 314, 315, 316
    Bagging (Bootstrap aggregation), 314, 314
    churn model, 261, 262, 263, 265
    classification algorithms, 146–150
    classification or propensity models, 7–8
    cross-selling model, 322–323, 323, 324, 325, 325, 327, 328
    customer scoring model, 141
    “divide-and-conquer” procedure, 146
  handling of predictors
    binary splits, 149
    C5.0/C4.5 and CHAID, 148–149
    Classification and Regression Trees, 148
  IBM SPSS Modeler C5.0, 164, 164, 164–165
  IBM SPSS Modeler CHAID, 159, 160, 160, 161, 162, 162–163
  Microsoft Decision Trees parameters, 170, 171
  modeling dataset, 151, 152–153
  Random Forests, 63, 64
  RapidMiner, 165, 168, 169, 169
  supervised segmentation, 146
  tree pruning, 150
  using terminating criteria
    prepruning or forward pruning, 149
    split-and-grow procedure, 149
  deep-selling modeling
    pilot campaign, 102–103, 103
    profiling of customers, 104–105, 105
    usage increase, 103–104, 104
  dimensionality reduction models, 11
  direct marketing campaigns, 94, 94–95

eigenvalue (or latent root) criterion of components, 197
  percentage of variance/information, 196, 196
  z-score method, 197

EM see Expectation Maximization clustering

estimation (regression) models, 7, 9–10
  linear or nonlinear functions, 9
  ordinary least squares regression, 10
  Expectation Maximization clustering, 12, 203
factor analysis, 14
  feature selection (field screening), 7, 10

Gains/Response/Lift charts
  binary classification problem, 70
  churn model, 72
  creation of charts, 72
  cumulative Lift or Index chart, 75, 76
  in Data Mining for Excel, 78, 78, 79
  Evaluation Modeler node, 75, 76
  Gains Chart, 73, 74
  in IBM SPSS Modeler, 75, 76, 77
  Kolmogorov–Smirnov statistic, 74
  performance measures, 72–73
  in RapidMiner, 77, 77
  Response chart, 75, 75

Gini index, CART
  child nodes, 151
  distribution of target classes, 151
  IBM SPSS Modeler CART, 153, 154
  purity improvement, 151
  of root node, 153
  splits and predictors, 153, 154
  voice and SMS usage, 153

IBM SPSS Modeler
  Bagging (Bootstrap aggregation), 61
  balancing approach, 49, 51, 52
  Bayesian belief networks, 180, 181, 182
  boosting, 62–63, 63
  CART, 153, 154, 165, 166, 167, 167–168
  Classification and Regression Trees, 164, 164, 164–165
  classification modeling methodology, 57, 58
  class weighting, 54, 56
  cluster modeling, 126, 126–127
  confusion matrix and accuracy measures, 68, 68, 69
  cross-selling modeling, 296, 297
  customer scoring model, 140, 140
  data exploration, 36
  derive nodes, 272
  Gains/Response/Lift charts, 75, 76, 77
  K-means, 206, 207, 207
  mobile telephony, 336, 340
  principal components analysis, 201, 202, 202
IBM SPSS Modeler (cont’d)
  profiling, 134, 135, 136
  receiver operating characteristic curves, 80, 82, 83, 84
  revealed segments, 131
  scoring customers, 89, 89, 90
  Split (Holdout) validation method, 42, 42, 43
  stream (procedure), churn modeling
    Auto-Classifier node, 241, 242
    Balance node, 239, 240
derived fields/candidate predictors, 234, 235–238
description, 234, 238
initial and balanced distribution, 239, 240
modeling steps, 234
parameters, 242, 244
Split (Holdout) validation and Partition node, 234, 239
Three Decision Tree and SVM model, 242, 243
Type node, setting, 241, 241
undersampling, 239
support vector machines, 188, 188, 189
Tree Augmented Naïve Bayesian network, 177, 177
TwoStep, 211, 212, 212
ICA see independent component analysis
imbalanced outcome distribution
  balancing, 48–53
class weighting, 53–57
independent component analysis, 14
K-means, clustering algorithms, 12, 205
  Bayesian Information Criterion, 204
centroid-based partitioning technique, 204
  centroids of identified clusters, 204, 206
  in Data Mining for Excel, 208, 209, 210
  Euclidean distance, 204
IBM SPSS Modeler, 206, 207, 207
K-medoids, 204
  Modeler’s, 204, 206
  RapidMiner K-means and K-medoids cluster, 208, 208, 208–209
K-medoids, 12
  Kohonen network/self-organizing map, 12
Kolmogorov–Smirnov statistic, 74
KS see Kolmogorov–Smirnov statistic
logistic regression, 8, 88
market basket analysis, 13
marketing reference table
  aggregations/group by, 18
deltas, 20
derive, 19
filtering of records, 18
flag fields, 20
joins, 18
ratios (proportions), 20
restructure/pivoting, 19
sums/averages, 19–20
maximum marginal hyperplane, 185–186, 186
meta-modeling or ensemble modeling, 57, 60
mining approach
cross-selling model
  data and predictors, 296
  modeling population and level of data, 296
  target population and attribute, 296
  time periods and historical information, 296, 297
and data model
cross-selling campaign, 26
  pilot campaign approach, 26
  product possession approach, 26
  product uptake approach, 26
voluntary churn propensity model
  data sources and predictors, selection, 220
  modeling population and data level, 220
  target population and churn definition, 220–221
  time periods and historical information, 221
mining datamart, 16
  marketing reference table, 17–20
  of mobile telephony operator, 16, 18, 333
  of retail banking, 16, 17
  of retailers, 16, 19
MMH see maximum marginal hyperplane
mobile telephony
  behavioral segmentation, 332
  Call Detail Records, 333
  clustering, IBM SPSS Modeler procedure, 336, 340
  core segments, 333–335
  high-level quality services, 332
  modeling steps, 340
  organization’s mining datamart, 333
  segmentation fields, 336, 337–339
  SMS and MMS messages, 332
model deployment
churn propensities
  defined, churn, 250
  ensemble model, 250
  propensity-based segmentation, 252, 252
  scored customers, sample, 251, 251
  voluntary churn model scoring procedure, 250, 251
$XF-CHURN field, 251–252

direct marketing campaigns
procedure and results, 94
scoring customers, marketing campaign, 88–94

model evaluation procedure
accuracy measures and confusion matrices, 66–70
gains, response, and lift charts, 70–78
precampaign model validation, 64–66
profit/ROI charts, 81–85

RapidMiner modeling process
confidence(T) field, 256
model deployment, 255
ROC curve, 255–256
Split Validation operator, 255
ROC curve, 78–81
test-control groups, 85–88

modeling process design
behavioral segmentation methodology
determining segmentation level, 116
selecting observation window, 116–117
selecting segmentation population, 115–116
selection of appropriate segmentation criteria, 116
classification modeling methodology
defining modeling population, 27–28
determining modeling (analysis) level, 28
target event and population, 28–29
time frames, 29–33

Naïve Bayesian networks
Apply Model operator, 174, 174
Attribute Characteristics, 175, 176
Attribute Profiles output, 175, 176
Data Mining for Excel, 174, 174, 175, 175
conditional probability, 172
normal distribution assumption, 173
RapidMiner process, 173, 173

Naïve Bayes model
Create Lift Chart operator, 256, 257
with Laplace correction, 254
ROC curve, 255, 256

OLSR see ordinary least squares regression
ordinary least squares regression, 10

PCA see principal components analysis
principal components analysis, 14, 39, 193–194
clustering, 193

components to extract
behavioral fields, 194, 194, 196

eigenvalue (or latent root) criterion, 196, 196–197
interpretability and business meaning, 194, 198
pairwise correlation coefficients, 194, 195
percentage of variance criterion, 197–198, 198
scree test criterion, 198
data reduction, 192
linear correlation between continuous measures, 193
meaning of component interpretation process, 201
Modeler, 199, 200
in RapidMiner, 199
rotation techniques, 199, 200
model, 340–341, 341
reduction of dimensionality, 193

Random Forests
Decision Tree models, 63, 64
in RapidMiner, 64, 65

RapidMiner modeling process
Attribute operator, 253
Bagging (Bootstrap aggregation), 61, 62
balancing approach, 50, 52, 53
boosting, 63
chi-square test, CHAID, 165, 168, 169, 169
classification modeling methodology, 59, 59–60
class weighting, 55, 56
cluster modeling, 127, 127, 354, 354
confusion matrix, 68, 69, 317, 317
Cross or n-fold validation method, 45, 46, 47
cross-selling model, 312, 312–313
customer scoring model, 140, 141
Decision Tree model with bagging, 314, 314, 315, 315, 316, 316
Gains/Response/Lift charts, 77, 77
K-means and K-medoids cluster, 208, 208, 208–209
model evaluation procedure, 255–256
Naïve Bayes model, 173, 173, 184, 254, 255, 256, 257
predictors, 235–238, 253
principal components analysis, 199, 201, 203, 203
profiling, 134, 134
Random Forests, 64, 65
receiver operating characteristic curves, 80, 80–81, 81, 317, 318
RapidMiner modeling process (cont’d)
retail case study see retail case study, RapidMiner
revealed segments, 132, 132
scoring customers, 90, 90, 91, 317, 318, 319, 319
Set Role settings, 253, 253
Split (Holdout) validation method, 42, 43, 44, 44, 254, 254, 255, 313, 313–314, 314
SVM models, 189, 190, 190
value segmentation and RFM cells analysis, 310, 310–312
receiver operating characteristic curves, 59, 67
area under the curve measure, 79
confusion matrix and accuracy measures, 67
Gains chart, 79
Gini index, 79
in IBM SPSS Modeler, 80
model evaluation, 78–81
Naïve Bayes model, 255, 256
performance of model, 317, 318
Profit/ROI charts
customers, 82
in Data Mining for Excel, 82, 84, 85, 86
in IBM SPSS Modeler, 82, 83, 84
marketers, 81
in RapidMiner, 80, 80–81, 81
sensitivity, 78
recency, frequency, and monetary analysis
cell segmentation procedure, 268
data preparation phase, 293
distribution, 293, 294
clustering model, 293
components, 291
cross-selling models, 268, 291
grouping (binning) of customers, 291
indicators, construction of, 291, 292
monitoring consuming behaviors, 290
quintiles, grouping customers, 291, 292
in retail industry, 291
scatter plot, 293, 295
regression models see estimation (regression) models
retail case study, RapidMiner
cross-selling model, 312, 312–313
Decision Tree model with bagging
bagging operator, 314, 314
parameter settings, 314, 315, 316
in tree format, 315, 316
performance of model
confusion matrix, 317, 317
ROC curve, 317, 318
scoring customers
model deployment process, 317, 318
prediction fields, 319, 319
Split (Holdout) validation, 313, 313–314, 314
value segmentation and RFM cells analysis, 310–312
RFM see recency, frequency, and monetary analysis
ROC see receiver operating characteristic curves
scoring customers, marketing campaign
binary classification problems, 88
Create Threshold operator, 91, 91
in Data Mining for Excel, 91, 92, 93
Gains/Profit/ROC charts and tables, 89
in IBM SPSS Modeler, 89, 89, 90
probabilistic classifiers, 88
propensity segmentation, 93–94
in RapidMiner, 90, 90, 91
scree test criterion, 198
segmentation algorithms
clustering algorithms
with K-means, 204, 206, 208
with TwoStep, 211
with data mining algorithms, 192
PCA see principal components analysis
sequence algorithms, 14
Split (Holdout) validation method
churn modeling, 234, 239, 259, 260
cross-selling modeling, 40–41, 41, 298–299, 322, 322, 323
in Data Mining for Excel, 43, 44
distribution of target attribute, 41
in IBM SPSS Modeler, 42, 42, 43
model training, 41
performance metrics, 40
random sampling, 40
in RapidMiner, 42, 43, 44, 44
retail case study, 313, 313–314, 314
supervised modeling, 23
classification or propensity models, 7–9
estimation (regression) models, 7, 9–10
feature selection (field screening), 7, 10
support vector machines, 9, 191, 242, 244, 244
linearly inseparable data
IBM SPSS Modeler, 188, 188, 189
Kernel functions, 187–188
Polynomial transformation, 187
RapidMiner SVM models, 189, 190, 190
linearly separable data
linear discriminant function, 184
maximum marginal hyperplane, 185–186, 186
separating hyperplane, 184–185, 185, 186 nonlinear mappings for classification, 145
SVM see support vector machines

TAN see Tree Augmented Naïve Bayesian network
telecommunications, segmentation application
data dictionary and segmentation fields, 336
data preparation procedure, 335–336
mobile telephony, 332–333
modeling procedure, 336, 340
identifying segments with cluster model, 342
preparing data for clustering, 340–341
profiling and understanding clusters, 344, 348–349
segmentation deployment, 354
segmentation procedure
deciding level, 335
dimensions, 335
population, mobile telephony core segments, 333–335
time frames and historical information analyzed, 335
using RapidMiner and K-means cluster
Cluster Distance Performance operator, 355
clustering with K-means algorithm, 354–356, 359
Euclidean distance, 355
K-means parameter settings, 355, 357, 358
mobile telephony segments, 355, 358
PCA algorithm, 354–355, 355
profile of clusters, 356, 359
variance/information, by components, 354–355, 356

test-control groups
direct marketing campaign, 85, 86
Model Holdout group, 87
Random Holdout group, 87, 88
recorded response rate, cross-selling campaign, 87, 87
time frames
in churn model, 31, 31
customer profiles, 29
event outcome period, 30, 31, 32–33
latency period, 30–31
multiple time frames, 33
observation (historical) period, 30, 30, 32

potential voluntary churners, identification of, 32
validation phase, 31
Tree Augmented Naïve Bayesian network, 305–306, 306
IBM SPSS Modeler, 177, 177
structure, 177, 177
training dataset, 179–180, 180
TwoStep cluster, 12
Akaike Information Criterion, 211
Bayesian Information Criterion, 211
IBM SPSS Modeler, 211, 212, 212
preclusters, 211

unsupervised models
association (affinity) and sequence models, 12–14
cluster models, 10–12
dimensionality reduction models, 11
record screening models, 14–15
up-selling modeling
pilot campaign, 105–106, 106
product upgrade, 107, 107
profiling of premium product owners, 107–108, 108

validation techniques
Bootstrap validation, 47–48
Cross or n-fold validation, 45–47
Split (Holdout) validation method, 40–44

value segmentation
and cross-selling in retail
data dictionary, 272, 279–285
data preparation procedure, 268–279
evolution and marketing usage, 287–290

and cross-selling in retail
data dictionary, 272, 279–285
data preparation procedure, 268–279
evolution and marketing usage, 287–290
grouping customers see customers grouping, value segmentation mining approach, 295–296
modeling procedure, 296–301
predictive accuracy of classifiers, 301–308
recency, frequency, and monetary analysis, 290–295
retail case study using RapidMiner, 309–319
transactional data, 267–268
and RFM cells analysis, 310, 310–312
discretization of numeric fields, 312
Map operator, value segments, 310
MONETARY attribute, 310
relevant binned attributes, 312, 312
total purchase amount, 310, 311