

Index

■ A

- Analytics spectrum
 - categories, 4
 - data analysis, 5
 - descriptive analysis, 5
 - diagnostic analysis, 5
 - predictive analysis, 5
 - prescriptive analysis, 6
- Area under curve (AUC), 103
- Average computer
 - memory price, 10
- Azure machine learning, 156
 - algorithms, 31–32
 - automobile price data, 24
 - building and testing, 21
 - components, experiment, 22–23
 - dataset, 25
 - dataset visualization, 26
 - deploying your model, 36–39
 - features, 29–30
 - iterative process, 22
 - palette search, 25
 - prediction, new data, 33–35
 - preprocess, dataset, 26–28
 - staging into production, 39–40
 - visual workspace, 21
 - web service, 40–42
 - web service testing, 39
- Azure Machine Learning, 88–89, 137

■ B

- Bayes point machines (BPMs)
 - average classifier, 78
 - Azure Machine Learning, 79
 - linear classifiers, 78
 - training iterations, 79
- BikeBuyer.csv file, 90

■ C

- Churn models
 - Azure Machine Learning Studio, 107–108
 - consumer business, 107
 - customer (*see* Customer churn model)
 - effective strategy, 107
 - training and testing, 108
- Classification and Regression Tree (CART), 73
- Customer churn model
 - boosted decision tree and forest algorithms, 121
 - classification algorithms, 109, 121
 - confusion matrix, accuracy, precision, recall and F1 scores, 125, 127
 - data preprocessing and selection
 - feature selection modules, 114
 - Metadata Editor, 119–121
 - missing values scrubber, 117–119
 - project columns, 115–116
 - quantize properties, 118
 - select columns, 116–117
 - training data and label, 120
 - decision tree, 121
 - preparing and understanding data
 - descriptive statistics, 113–114
 - KDD Cup web site, 110
 - machine learning model, 113
 - Machine Learning Studio, 110
 - Orange training and labels, 111, 113
 - uploading dataset, 110–111
 - ROC curve, 125–126
 - Score and Evaluate model, 125
 - Split module, properties, 122
 - train model, 124
 - two-class boosted decision tree and forest, 123–124

Customer propensity models
 Bike Buyer dataset, 91
 box-and-whisker icon, 94
 classification algorithm, 99
 confusion matrix, 103
 customer targeting models, 87
 data science process, 88
 demographic variables, 91
 histogram, 92
 IQR, 94
 logistic regression, 99
 log transformation, 93
 prediction error, 99
 predictive models, 105
 project columns, 100
 ROI, 87
 testing and validation, 101
 train module, 100
 true positives, 104
 visualizing data, 91

Customer segmentation models
 companies, 129
 consumer credit score, 130
 data analysts, 129
 K-means clustering
 (*see* K-means clustering)
 learning techniques, 130
 telecommunication industry, 130
 wholesale customers
 (*see* Wholesale customers)

■ **D**

Data analysis
 Azure machine learning, 151
 feature selection, 150
 filter based feature
 selection module, 152
 launch column selector, 150
 missing values scrubber module, 149
 SECOM dataset, 149–150, 152

Data format conversions, 95

Data loading
 labels, 149
 local file system, 146–147
 local machine, 88–89
 non-local sources, 89–90, 147–148

Data mining technologies, 8

Data science
 academic disciplines, 4
 acquiring and data preparation, 11–12

algorithms, 14
 analytics spectrum
 (*see* analytics spectrum)
 classification algorithms, 14
 clustering, 15–16
 competitive asset, 7
 content analysis, 17
 customer demand, 8
 definition, 11
 digital data, 8
 model development and
 deployment, 12
 model's performance, 12
 powerful processes, 3
 practitioners, 4
 processing power, 9
 recommendation engines, 18

Dataset, 145

Data transformation item, 95

Decision tree algorithms
 bagging and boosting, 74
 CART algorithm, 73
 data selection, 74
 Gini impurity, 73
 ID3, 73
 root nodes, 72
 training, 74

■ **E**

Ensemble models
 algorithms, 18
 applications, 18–19
 building, 19

Evaluate model module, 155–157

■ **F, G, H**

Feature selection, 96, 98
 Filter based feature selection, 154
 Fisher score, 151

■ **I, J**

Iterative Dichotomizer 3 (ID3), 73

■ **K**

K-means clustering
 categories, 131
 experiment samples, 132

- feature hashing, 133
- properties, 135, 137–138
- right features, 134–135
- segmentation of companies, 130

L

- Learning algorithms
 - agglomerative and division, 79
 - BPMs, 78–79
 - centroid initialization methods, 82
 - density-based algorithms, 80
 - K-means clustering, 80–81
 - mapping, 75
 - partitioning-based clustering, 80
 - regression algorithms
 - (*see Regression algorithms*)
 - supervised learning, 74
 - SVMs, 76–78
 - telecommunication, 75
- Linear correlation module, 96

M

- maml.mapInputPort(1) method, 48
- Metadata Editor module, 138
- Missing values scrubber
 - properties, 28–29

N, O

- Neural networks
 - ART, 70
 - hidden nodes, 71
 - input and output nodes, 70
 - propagations, 70
 - rate of convergence, 71
 - self-organizing maps, 70
 - sigmoidal activate function, 70

P, Q

- PCA. *See* Principal Component Analysis (PCA)
- Predictive maintenance models
 - business problem, 145
 - components, 143
 - deployment, 158
 - manufacturing industry, 143
 - model-based condition, 144

- repairs, 143
- staging, 158–160
- testing and validation, 154
- training, 152–153
- transmission, 144
- vibration analysis, 144

- Principal Components
 - Analysis (PCA), 50, 52–53, 134
- Project Columns properties, 27

R

- Receiver Operating
 - Curve (ROC curve), 125, 155–126
- Regression algorithms
 - decision tree algorithms, 72–74
 - linear regression, 68–69
 - neural networks, 70–71
 - numerical outcomes, 67
- Regression model experiment, 35
- Regression techniques, 16–17
- ROC curve. *See* Receiver Operating Curve (ROC curve)
- R, statistical programming language
 - actuarial sciences, 43
 - Azure Machine Learning, 44–45, 64
 - bioinformatics, 43
 - building and deploying
 - Execute R Script module, 46–48
 - language modules, 45
 - maml.mapInputPort(1)
 - method, 48
 - ML Studio, 47–48
 - visualization, 48, 50
 - data preprocessing
 - components, 53–54
 - Execute R Script module, 51–52
 - machine learning algorithm, 50
 - Metadata Editor module, 51
 - missing values scrubber
 - module, 51
 - PCA, 50, 52–53
 - sample and CRM dataset, 50
 - decision tree
 - Adult Census Income Binary Classification dataset, 59–61
 - library(), 61
 - ML Studio, 62
 - rpart, 59, 64
 - view output log, 62, 64

■ INDEX

R, statistical programming language (*cont.*)
finance and banking, 43
script bundle (zip)
 Execute R Script module, 56, 58
 folder containing, 54–55
 package, 55
 uploading, dataset, 55–56
telecommunication, 43

■ **S**

Score model module, 33
Simulation, 17
Split module, 153
Staging
 Azure machine learning studio, 158
 into production, 160–161
 publishing, 160
 training modules, 159
Support vector machines (SVMs)
 hyperplane, 76
 kernel-based learning, 76

random number seed, 78
telecommunication, 76

■ **T**

Two-class boosted decision
tree module, 104

■ **U, V**

UCI Machine Repository, 138

■ **W, X, Y, Z**

Wholesale customers
 cluster assignment, 141–142
 clustering model, 138
 Euclidean distance, 140
 K-means clustering, 139
 Metadata Editor, 140
 train clustering model, 139
 UCI Machine Repository, 138