

Index

A

Affordable Care Act (ACA), 225, 228
Airline industry
 air carriers, 277–278
 Airport Carbon Accreditation, 279
 Air Travel Consumer Report
 November 2017, 284–285
 big data, 281
 carbon neutrality,
 certification to, 279–280
 cybercriminals attack, 283
 data, 280–281
 defined, 277
 flight delays, 284
 category, 284
 FIM, 286
 multiple linear regression (*see*
 Multiple linear regression models)
 overall causes of, 285
 IATA, 283
 offers, 281–282
 passenger experience, 278, 281–282
 predictive analytics, 281
 safety and security, 278
 Skytrax, 278
 sustainability, 279
 working with NASA, 282
Area Under Curve (AUC), 44, 45, 64
Artificial neural networks (ANNS), 234
Assortment planning process, 104

Augmented Dicky
 Fuller Tests, 137–139, 141
Autocorrelation Factor (ACF), 111–112
AutoRegressive (AR) model, 107–109
Autoregressive integrated moving
 average (ARIMA) model
 ARMA model, 109
 AR model, 107–108
 diagnostic checking stage, 113
 forecasting stage, 114
 identification stage, 111–112
 integrated model, 109, 111
 MA model, 108–109
Auto replenishment system, 102

B

Bank fraud, 32
Banking sector
 cross-selling and up-selling, 29
 customer acquisition, 30
 customer churn, 30
 fraudulent activity, 32–33
 loans, 31
Bank-loan defaults prediction,
 logistic regression
 assumptions, 38
 curve, 37
 equation, 35
 model fit tests, 39
 odds, 36

INDEX

Big data analytics

- agriculture industry, 17
- analytic competitors, 18
- challenges, 10
- energy industry, 17
- finance industry, 16
- health industry, 16
- insurance industry, 15
- Python, 13
- Rapid Miner, 14
- retail industry, 17
- R programming language, 12
- SAS, software suite, 12
- SPSS Modeler and SPSS Statistics, 13
- telecom industry, 16
- tools, 12
- travel and tourism industry, 15
- 5 Vs, 9

Box-Ljung test, 129–131, 133

Brussels attacks, 278

C

Campus area network (CAN), 162

CART algorithm, 173

Chi-Square test, 114

Confusion matrix, 42

Consumer-packaged goods (CPG), *see*

Fast Moving Consumer Good
(FMCG) industry

Convenience stores, 99

Customer churn, 163, 165

D

Data

analytics

application, 18

data analysis, 4

data collection, 3

data preparation, 4

definition, 2

methodology, 2

model building, 5

results, 5

six models, industries, 18, 20

types, 6–7

collection, 3

qualitative data, 7

quantitative data, 8

Decision tree model

advantages, 169

categorical variable, 168

continuous variable, 168

for customer churn, 171–172

definition, 168

dependent variable, 168

internal node, 169

leaf node, 169

limitations, 169

missing values, 170

overfitting problem, 170

postpruning process, 170–171

prepruning process, 170

root node, 168

split algorithms

entropy, 175–176

Gini Index, 173–175

information gain, 177

Variance Reduction

method, 177–179

using R

customer churn, 179

exploratory data analysis, 180

model building and

interpretation on training and

testing data, 184, 186–193

- test data set, 183
- train data set, 183
- using SAS
 - alphabetic list, 194–195
 - contents procedure, 194
 - create library, 193–194
 - customer churn
 - probability table, 216–217
 - frequency procedure, 196
 - goodness-of-fit tests, 200
 - Internet_service, 196–197
 - means procedure, 195
 - model building and interpretation of
 - full data, 200–207
 - model building and interpretation on training and testing data, 208–210, 212, 214, 216
 - normal distribution of
 - Monthly_Charges, 200
 - observations, 198
 - quantiles, 198
 - statistical measures, 197
 - test for location, 198
 - univariate procedure, 197
- Department stores, 99
- Descriptive analytics, 6
- Discount stores, 99
- Durbin-Watson test, 309

E

- Elbow method, 367–368
- Electronics Recording Machine,
 - Accounting (ERMA), 27
- E-tailer, 99

F

- False positive rate (FPR), 44
- Fast Moving Consumer Good (FMCG)
 - industry
 - ASSOCHAM-Tech Si report, 346
 - customer experience
 - and engagement, 347
 - customer segmentation with
 - K-means (*see* K-means clustering)
 - RFM (*see* Recency, Frequency, and Monetary (RFM) model)
 - data, 346
 - definition, 345
 - logistics management, 348
 - markdown optimization, 349–350
 - marketing mix model, 348
- Flight Deck Interval
 - Management (FIM), 285–286
- Forward selection method, 290

G

- Gini Index, 173–175
- Global Aviation Data
 - Management (GADM), 278
- Goodness-of-fit
 - statistics test, 113

H

- Healthcare industry
 - application of analytics
 - big data technologies, 224
 - data sources, 224
 - healthcare fraud
 - detection, 227–228

INDEX

Healthcare industry (*cont.*)

- improve patient outcomes & lower costs, 228–229
- outbreak of disease and preventative management, predicting, 225
- readmission rate of patients, 225–227

medical devices, 221

Random Forest model, 230

- advantages, 234–235
 - limitations, 235
 - OOB error, 236
 - proximity measures, 237
 - selecting m_{try} , 236
 - selecting N_{trees} , 235
 - using R (*see* Random Forest model, using R)
 - using SAS (*see* Random Forest model, using SAS)
 - variable importance measures, 237
 - working, 230–234
- subsectors, 221
- transitions, financial management, 222–223

Hosmer-Lemeshow test, 39

I, J

IncNodePurity,

see Mean Decrease Gini (MDG)

Integrated model, 109, 111

International Air Transport

Associations (IATA), 277, 283

Internet area network (IAN), 162

Inventory management, 102

K

K-means clustering

- advantages of, 357
- algorithm, 355–357
- limitations of, 357
- overview of, 355
- parameters, 357
- using R

customer loyalty, 366–371, 373–376

customer_seg data set, 358

exploratory data, 359–362

using SAS

- contents procedure, 377
- customer segmentation, 383–384, 386, 391–393
- location tests, 380
- means procedure, 378
- observations, 381
- quantiles, 380–381
- statistical measures, 379
- univariate procedure, 379
- variables and attributes, 377

L

Likelihood ratio test, 39, 41

Local area network (LAN), 162

Logistic regression model/logit model

banking sector using R

- coefficients table, 57, 59
- exploratory data analysis, 47–52
- glm on training data, 56–58
- loan-default data set, 47
- model building and interpretation, 52
- predictive value validation, 61–64
- variance inflation factor, 60

banking sector using SAS
 binary logit model, 87
 Chi-Square test statistics, 88
 coefficient table, 80, 84, 86, 90
 content procedure, 66–68
 default and Emp_status display, 69
 dependent variable, 76
 Fisher's scoring, 77, 87
 goodness-of-fit tests,
 normal distribution, 73
 likelihood ratio, score,
 and Walt test, 78, 88
 means procedure, 65–66
 model fit statistics, 78, 88
 normal distribution for
 Saving_amount, 72–73
 odds ratio, 80, 90
 predicted probabilities and
 observed response, 81, 91
 probability of Default = 1, 78, 87
 procedure frequency, 68, 83
 proc means, 68
 ROC and AUC curve, 91
 survey, select procedure, 82
 testing data set, 74, 82, 83
 test statistics, 71
 training data set, 74
 univariate procedure, 70
 Wald Chi-Square statistic, 84
 bank-loan defaults prediction, 34
 curve, 37
 predictive value validation
 classification table, 42
 confusion matrix, 42–44
 ROC and AUC Curve, 45–46
 sigmoidal curve, 37
 statistical test, individual
 independent variable, 40

M

Marketing mix model, 348
 Mean absolute error (MAE), 117
 Mean absolute percentage
 error (MAPE), 118
 Mean Decrease Accuracy (MDA), 237
 Mean Decrease Gini (MDG), 237
 Metropolitan area network (MAN), 162
 Model fit tests, 39
 Moving average (MA) model, 108–109
 Multiple linear regression models
 assumptions, 288–289
 defined, 286
 equation, 287
 influential observations, 291
 outliers detection, 291
 residual, 290
 R-squared, 291
 selection method, 290
 standard error, 292
 using R
 exploratory data, 293–294, 296–299
 flight_delay data set, 293
 train_data set and test_data set,
 299, 301–303, 305–310
 using SAS
 contents procedure, 312
 corr procedure, 314
 location test, 316
 means procedure, 313
 observations, 316
 partial output of libref, 321–323
 pearson correlation
 coefficients, 314
 quantiles, 316
 sgscatter procedure, 318
 statistical measures, 315

INDEX

Multiple linear regression models (*cont.*)

- survey select procedure, 319
- testing data, 335, 339–340
- training dataset, 324, 326–328, 330–331, 333–334
- univariate procedure, 315
- variables and attributes, 312
- violation of assumptions, 288–289

N

- National Aviation System (NAS), 284
- Non-constant variance score test, 310

O

- Oscars of the aviation industry, 278
- Out-of-bag (OOB) error, 236, 243

P, Q

- Partial Autocorrelation
 - Factor (PACF), 111, 113
- Pearson moment correlation method, 296
- Permutation measure, 237
- Plain Old Telephony Services (POTS), 161
- Postpruning process, 170–171
- Predictive analytics, 6
- Prepruning process, 170
- Prescriptive analytics, 7
- Proximity, 237

R

- Random Forest model
 - advantages, 234–235
 - limitations, 235
 - OOB error, 236

- proximity measures, 237
- selecting m_{try} , 236
- selecting N_{trees} , 235
- using R
 - data, 238
 - model building & interpretation, 244–248
 - performing data
 - exploration, 239–243
 - splitting data set, 243–244
- using SAS
 - Bare_Nuclei outcome, 252
 - freq procedure, 251
 - model building & interpretation on full data, 253–257, 259–261
 - model building & interpretation on training and testing data, 261–271
 - procedure content, 249–250
 - variables and attributes, 250
 - variable importance measures, 237
 - working, 230–234
- Receiver Operating Characteristics (ROC), 44
- Recency, Frequency, and Monetary (RFM) model
 - attributes, 351–352
 - scores calculation, 353–354
 - using R
 - customer_seg data set, 358
 - exploratory data, 359–361
 - RFM table, 362–366
 - using SAS
 - contents procedure, 377
 - customer segmentation, 383–384, 386, 391–393
 - location tests, 380
 - means procedure, 378
 - observations, 381

- quantiles, 380–381
 - statistical measures, 379
 - univariate procedure, 379
 - variables and attributes, 377
- Retailing industry
- assortment planning process, 104
 - convenience stores, 99
 - department stores, 99
 - discount stores, 99
 - e-tailer, 99
 - inventory management, 102
 - market, 97
 - predictive analytics
 - customer engagement, 100–101
 - definition, 100
 - supply chain management, 102
 - price optimization, 103
 - SARIMA models (*see* Seasonal ARIMA (SARIMA) models)
 - sector, 97
 - specialty stores, 99
 - supermarket stores, 99
 - supply chain, 98
 - Walmart, 97–98, 100
- Reverse logistics management, 349
- Root mean squared error (RMSE), 117
- S**
- Seasonal ARIMA (SARIMA) models
- ACF and PACF correlogram, 116
 - ARIMA model (*see* Autoregressive integrated moving average (ARIMA) model)
 - create new library, 134
 - non-stationary to stationary series conversion, 116
 - representation, 115
- sales data of food and beverages
 - Box-Ljung test, 129–131, 133
 - error measures, 128
 - forecasting retail sales, 130
 - import dataset, 119
 - non-stationary series, 122
 - seasonally differenced retail sales, 123–126
 - stationary series, 121
 - time series data, 119–120
 - SAS for sales forecasting
 - alphabetic list, 135
 - Augmented Dicky Fuller test, 137–139, 141
 - create new library, 133
 - with different non seasonal terms, 150–154, 156, 158
 - maximum likelihood method, 141–145, 147, 149
 - proc content, 134
 - timeseries procedure, 135–137
 - variable information, 135
 - time series data, 105, 107, 115
 - Segment-of-one
 - marketing approach, 347
 - Skytrax, 278
 - Specialty store, 99
 - Statistical Analysis System (SAS), 1
 - Statistical test, individual
 - independent variable logistic regression model likelihood ratio test, 41
 - predictive value validation, 41–46
 - Wald statistic test, 40
 - Supermarket stores, 99
 - Supply chain management, 98, 102
 - Support vector machines (SVMS), 234

T, U

- Telecommunications (telecom)
 - companies, 161
 - customer churn, 163, 165
 - decision tree
 - (*see* Decision tree model)
 - definition, 161
 - fraud detection and prevention, 166
 - network analysis, 165
 - predictive analytics, 165
 - price optimization, 166, 168
 - types of data networks, 162–163
- Time series data, 105, 107

V

- Variance inflation
 - factor (VIF) test, 60, 289, 310
- Variance Reduction method, 177–179
- Virtual private network (VPN), 162

W, X, Y, Z

- Wald statistic test, 40–41
- Walmart Stores Inc., 97–98, 100
- Warehouse management system, 102
- Wide area network (WAN), 162
- Wireless industry, 161