index

Symbols

| operator 92

Numerics

68–95–99 rule 44

A

A_ub matrix 234–236
absorbing states 295, 311–317
  obtaining fundamental matrix 313
  predicting 315
  predicting programmatically 316
ACF (autocorrelation function) 197
  plot 199
'add' parameter 218
add_edges() method 290
addition rule 23
agent fees variable 285
agg() method 118, 152
Agile methodology 223–224, 226
AIC (Akaike information criterion) 210
Amount column 336
anytree library 283, 287
append() function 235
Apple 187
apply() method 90, 331, 334
AR (autoregression) 185
  autoregression component 192, 205–207
  moving average component 206
Area variable 125
ARIMA (autoregressive integrated moving average) 185
  class 207–208
  models 192–215
assign() method 321, 326–327, 338
astype() method 332
AUC (area under ROC curve) 137
Augmented Dickey–Fuller (ADF) test 201

B

b_ub vector 234–236
bagging (bootstrap aggregating) 141
basketball games 73
benford library 332
Benford's law 320–324
  distortion factor and z-statistic 343
  mantissa statistics 344
  mean absolute deviation 342
  validating 338
Besni class label 117
beta distribution 355
BIC (Bayesian information criterion) 210
binom module 53
binomial distribution 50
  computing probabilities 54
  properties 51
bln_msk Boolean mask 197
boldface type 285
bounded, defined 37
Breusch–Pagan test 108
budget 350
  constraints 231

C

c-charts 394–396
CDF (cumulative distribution function) 33, 37–39, 245
chance nodes 282
chi-square test 338
  chi-square statistic 339
  degrees of freedom 341
  p-value 341
CL (center line) 381
Class variable 117–118, 121–123, 125, 128
classification metrics 131
  accuracy 136
  area under ROC curve 137
  confusion matrix 132
  false positive/negative rate 135
  misclassification rate 136
  ROC curve 138
  sensitivity 133
  specificity 134
clf object 160–161, 163
close_prices data frame 188, 214, 262, 265
close_prices index 189
coef (coefficients) 95
coeff parameter 211
collectively exhaustive 296
combinations 23–30
  with replacement 24–30
  without replacement 25–27
complement rule for probability 65–66
conditional covariance 198
conditional probabilities 72–77
  examples of 72–73
  formulaic approach to 76–77
    calculating P(A | B) 77
    calculating P(A and B) 76
    calculating P(B) 76
    defining events 76
  independence and 73
  intuitive approach to 74–76
conditional variance 198
confusion matrix 132, 161, 178
conservative approach 277
constraints 231, 233–236
contingency table 74–75
continuous random variables 30–34
  cumulative distribution function 33
  examples 31
  probability density function 32
control charts
  for attributes 388–398
  for variables 398–408
CONVERT class labels 161, 165, 181
CONVERT variable 146, 159
ConvexArea variable 125
copy() method 327
corr() method 123
count variable 326–327
counting rules 22–30
  addition rule 23
  combinations 23–30
  multiplication rule 22
  Pascal's triangle 27
  permutations 23–30
counting, continuous random variables 30–34
  cumulative distribution function 33
  examples 31
  probability density function 32
Counts column 338
covariance (Cov) 198
CPM (critical path method) 349
criterion function 160
cross-tabulation 74
cumulative probability distribution, computing 248–249

D

data
  exploring 115–116, 118–119, 122, 148, 150, 152, 154–155
  importing 186–191
  plotting 190
  understanding 189
data definition 231
data exploration 88
data filtering 144
data transformation 145
data type conversion 146
dataclass decorator 232
decision trees 140–142, 157–173, 271, 282–292
  advantages and disadvantages of 173
  creating schema 283–289
  decision-making without probabilities 273–282
  evaluating model 161
  exploring data 148, 150, 152, 154–155
  fitting 141–142
  fitting model 160
  fitting random forests 174–183
  importing data 142
  interpreting and understanding 164–173
  predicting responses 161
  splitting data 158
  understanding data 143
  wrangling data 144–146

decision variable bounds 231, 236
decision-making without probabilities 273–282
  EV method 279–282
  Maximax method 273–276
  Maximin method 277
  Minimax Regret method 277–279
DecisionTreeClassifier() method 160, 175–176
defects per unit, quality control measures 386
dependencies, defined 351
derived variable 146
DES (double exponential smoothing) 216–217
descriptive statistics 3
DF (distortion factor) 343
diameter variable 399
differencing 195
  purpose 195
  stationarity and 197–205
  types 195
discrete random variables 30, 35–39
  cumulative distribution function 37–39
  examples 35
  probability mass function 36
discrete uniform distribution 56
  computing probabilities 58
  properties of 56
distortion factor 343
distributions, discrete uniform 56
dot product 302
dot() method 305
DOWN variable 144
dropna() method 202, 263
Durbin–Watson test 107

E

Eccentricity variable 125–126, 128
empirical probabilities 19
ensemble techniques 141
epic constraints 231
epics list 235
equilibrium conditions 295, 307–311
  predicting programmatically 308–311
equilibrium distribution 308
EV (Expected Value) method 272, 279–282
EV data frame 281
evaluating model, confusion matrix 161
EWMA charts 407
examples, number sequences
  payment amounts 336
  street addresses 330–333
  world population figures 334–335
excess kurtosis measures 345
Expected column 338
expected value calculation 255
explanatory variable 81
ExponentialSmoothing class 217–218
Extent variable 125–126, 128

F

F-statistic 103
factorial() method 26, 28–29
false positive/negative rate 135
feature importance 179
features list 233
fetch_stock_data() function 187–188
fillna() method 145
financial data, fetching 187–189
financial investments 73
first_digit variable 331, 334–337
first_probability variable 49
first-digit law 320
first-order differencing 202–205
fit() method 160, 176–177, 208, 217
fitting linear regression
  interpreting and evaluating results 95, 99, 103, 105
  simple linear regression 87
  testing model assumptions 105–108
fitting models 93, 160
  decision trees 141–142
  random forests 141–142
  time series 215–220
  time series models 192–215
    AR and MA components 205–207
    autoregression (AR) component 192
    combining ARIMA components 192
    differencing 195
    evaluating model fit 209–213
    fitting models 207–209
    forecasting 214
    integration (I) component 192
    MA (moving average) component 192
    stationarity and differencing applied 197–205
      ACF plot 199
      Augmented Dickey–Fuller (ADF) test 201
      first-order differencing 202–205
      PACF plot 200
      visual inspection 197
    stationarity, statistical properties 193
floats 189
Forecast variable 214
forecasts, distinguishing from predictions 185
fundamental matrix 313

G

g-charts 396–398
Gini impurity 165
GM (General Motors) 261
groupby() method 118, 165, 326, 400

H

head() method 89, 116–117, 130–131, 147, 189, 195, 326–327, 331, 334, 336, 345–346, 389, 400
heat maps, correlation 122
histograms, logistic regression 119
historical data 259
history() method 188
Holt's linear trend model 217
Holt–Winters exponential smoothing (H-W) 216
Holt–Winters model 218–220
  forecasting 219
  model evaluation and comparison 219
homoscedasticity of residuals 107
HQIC (Hannan–Quinn information criterion) 210

I

I-MR charts 405–407
IDEs (integrated development environments) 6–7
IDLE (Integrated Development and Learning Environment) 6
iloc attribute 75
importing data 142, 186–191
  fetching financial data 187–189
  logistic regression 115
    computing basic statistics 118
    correlation heat maps 122
    histograms 119
    understanding data 115
    viewing data 116
  plotting data 190
  understanding data 189
independence 73
  of residuals 107
inequalities 230
inferential statistics 3
info() method 88–89, 115, 146, 189, 330, 334, 336, 389, 399
integration (I) component 192
intercept 94
internal nodes 164, 169–173
investment_alternatives vector 274
ISO 8601 date formats 188

J

Jarque–Bera test 109
Jupyter Notebook 7

K

Kecimen class label 117
kurtosis() method 346

L

lambda functions 152
law of anomalous numbers 338
LCL (lower control limit) 380
leaf nodes 164, 169–173
len() method 132, 197
line of best fit 85
linear equations 230
linear programming 222
  developing linear optimization framework 229–241
    constraints 233–236
    data definition 231–232
    decision variable bounds 236
    linear equations and inequalities 230
    objective function 233
    result evaluation 239–241
    solving linear programming problem 236–238
  problem formulation 223–229
    approach 225–227
    challenge 224
    feature summaries 227–229
    scenario 224
linear regression 81–87
  conditions for best fit 86–87
  fitting 79
    interpreting and evaluating results 95, 99, 103, 105
    testing model assumptions 105–108
    with simple linear regression 87
  fitting model 93
  goodness of fit 85
  importing and exploring data 88–92
  linear equation 81–83
  vs. logistic regression 113
linearity between variables 106
list comprehensions 233

list() method 321
LLR (log-likelihood ratio) 126
llr p-value 130
log returns 259, 262
log variable 345
log_model regression 125
logistic function 113
logistic regression
  defined 111
  exploring data 115–116, 118–119, 122
  fitting model 125
  fitting, classification metrics 131–139
  importing data 115–116, 118–119, 122
  interpreting and evaluating results 128–130
  multiple logistic regression 115
  vs. linear regression 113
logit() method 125, 128
lower_threshold 92

M

MA (moving average) component 192–207
  autoregression component 205
  moving average component 206
machine learning, overview 4
MAD (mean absolute deviation) 342
MAE (mean absolute error) 141
MajorAxisLength variable 125
majority vote 141
mantissa statistics 344
map() method 117
Markov analysis 294
  absorbing states 311–317
  equilibrium conditions 307–311
  mechanics of 295
  states 296–307
Markov property 295
MAs (moving averages) 185
math.comb() method 28, 30, 70–71
math.factorial() method 55
math.perm() method 26, 71
matplotlib library 83, 179, 322, 390
matrices, of transition probabilities 300–307
  making future state predictions programmatically 305–307
  predicting period 1 given period 0 302
  predicting period 2 given period 0 303–305
  predicting period 2 given period 1 302
max (maximum) 118
max_depth parameter 160
maximax data frame 276
Maximax method 272–276
  identifying possible states of nature 274
  list of payoffs for each alternative and state of nature combination 274
  list of possible alternatives 273
  selecting decision-making method and making decision 275
maximin data frame 277
Maximin method 272, 277
Maximum column 278–279
mds data frame 94
mean 90, 118, 345
  quality control measures 381
mean() method 150, 152, 343, 346
median 90, 118
medical testing 72
metrics_accuracy_score() method 177
min (minimum) 118
Minimax Regret method 272, 277–279
MinorAxisLength variable 125–128
misclassification rate 136
missing values 145
MLE (maximum likelihood estimation) 208
model class 208
model_fit method 208
model_fit.forecast() method 214
model.fit() method 208
models, fitting time series models 192–215
  AR and MA components 205–207
  autoregression (AR) component 192
  combining ARIMA components 192
  differencing 195
  evaluating model fit 209–213
  fitting models 207–209
  forecasting 214
  integration (I) component 192
  MA (moving average) component 192
  stationarity, statistical properties 193
Monte Carlo simulations 242
  applications and benefits of 243
  automating simulations on continuous data 259–270
    analyzing historical data 261
    analyzing results 268–270
    calculating log returns 262
    computing statistical parameters 264
    generating random daily returns 265
    predicting stock prices with Monte Carlo simulations 259
    simulating multiple trials 267
    simulating prices 266
  automating simulations on discrete data 255–259
    plotting and analyzing results 257–259

Monte Carlo simulations *(continued)*
staffing levels example 246–255
analyzing results 254
computing cumulative probability distribution 248–249
establishing interval of random numbers for each variable 250–252
establishing probability distribution 246
expected value calculation 255
generating random numbers 252
simulating series of trials 253
step-by-step process 244
moving range, quality control measures 386
MSE (mean squared error) 141
mu parameter 266
multicollinearity 110
multiple logistic regression 115
multiplication rule 22
multistate systems, vector of state probabilities for 297–300
mutually exclusive 296

N

NaN (Not a Number) 145
naturally occurring number sequences 324
Negative class 75
Newcomb–Benford law 321
nfl_pbp data set 158, 165
nondecreasing 37
nonnegativity 36
norm module 47
normal distribution 42
computing probabilities 48
probability density function 46
properties of 42
normality, testing for 91
of residuals 108
np-charts 392–394
np.arrange() method 51, 327
np.dot() method 306–307
np.log() method 324
np.random_choice() method 327
np.random.normal() method 44
np.random.randint() method 326, 329
np.std() method 344
num_days parameter 266
num_simulations parameter 266
num_trials parameter 256
number of defective items, quality control measures 384–385
number sequences 319
Benford's law 320–324
chi-square test 338–339, 341
distortion factor and z-statistic 343
examples 330–336
mantissa statistics 344
mean absolute deviation 342
naturally occurring 324
plotted distributions 328–330
random distributions 325–330
uniform distributions 325–330
validating Benford's law 338

O

objective function 231, 233
OFFENSIVE_TEAM_VENUE variable 145, 154, 171
oj data frame 389–390, 392
OLS() method 93, 127
opg method 209
optimistic approach 273
optimization 4
optimize module 237
order of magnitude 324
order parameter 207–208
outliers 86
detecting 92

P

p_bar variable 390
p-charts (proportion charts) 379, 388–392
p-value, chi-square test 341
PACF (partial autocorrelation function) 197
plot 200
pandas library 88, 115–116, 143
Pascal's triangle 27
payments data frame 336
payoff table 275–278, 280–281
pd.concat() method 165
pd.DataFrame() method 75, 321
pd.read_csv() method 88, 115, 143, 330, 334, 336, 389, 399
pd.set_option() method 147
PDF (probability density function) 32, 42
Perimeter variable 125, 128
permutations 23–30
with replacement 24
without replacement 25
PERT (program evaluation review technique) 349
pessimistic approach 277
pistons data frame 399–400

planning, overview 350
PLAY_TYPE variable 144, 155, 165, 171
plot_acf() method 199
plot_pacf() method 200
plot_tree() method 163
plot() method 149
plotted distributions 328–330
plotting
- data 190
- decision tree 163

PMF (probability mass function) 35–36, 51
pmf_values variable 58
Poisson distribution 60, 247
- computing probabilities 62
- properties of 60

poisson module 62
poisson_table data frame 247
poisson.cdf() method 249
poisson.pmf() function 62
populations data frame 334, 338, 342, 345, 347
Positive class 75
pow() method 25, 66
predicting responses 161
prediction variable 131
predictions, distinguishing forecasts from 185
predictor variable 81
print() function 338
print() method 89, 161, 177, 233–234, 237
priority scores 225
Prob (f-statistic) 105
probability 3, 16
- complement rule for 65–66
- computing 48
- converting and measuring 19
- counting rules 22–30
- examples and solutions 68–71
- of on-time completion 349
- reference guide for 67
- types of 19

probability density function 46
probability distributions 42
- and conditional probabilities 41
- binomial distribution 50–51, 54
- establishing 246
- normal distribution 42
- Poisson distribution 60, 62

probability_not_4 variable 65
probability_not_4_both variable 66
process capability indices, quality control measures 387
programming, making future state predictions programmatically 305–307
project crashing 349
project management 350
- estimating activity times with PERT 354–357
- WBS (work breakdown structure) 350–354

proportion defective, quality control measures 383
proportion variable 390
pseudo R-squared 129
PyCharm 7
PyDev 7
Python 4–6
- ease of learning 5
- ecosystem 4
- IDEs 6
- industry adoption 6
- online support and community 5
- using for precision and efficiency 13
- versatility 6

Q

Q–Q plot 108
quality control
- control charts for attributes 388–398
  - c-charts 394–396
  - g-charts 396–398
  - np-charts 392–394
  - p-charts 388–392
- control charts for variables 398–408
  - EWMA charts 407
  - I-MR charts 405–407
  - r-charts 402
  - s-charts 403
  - x-bar charts 400
- visualizing 378

quality control measures 380–388
- defects per unit 386
- mean and center line 381
- moving range 386
- number of defective items 384
- number of defects 385
- process capability indices 387
- proportion defective 383
- range 382
- sample size 383
- standard deviation 381
- UCL and LCL 380
- upper control limit and lower control limit 380
- z-score 386

quantitative methods 2–4
QUARTER variable 148

R

r-charts 402
R-squared 8, 82, 85, 99
random daily returns 260
random distributions 325–330
random forest classifier object 175–177
random forests 140–142
  fitting 141–142, 174–183
  importing data 142
  understanding data 143
  wrangling data 144
random numbers
  establishing interval of 250–252
  generating 252
random variables 30, 59
  continuous 30–34
  discrete 35–39
random_distribution variable 327
random_state parameter 160
RandomForestClassifier() method 175
range, quality control measures 382
rank variable 346
recall_score() method 134
rectangular distribution 56
regression
  analysis 4
  line 81, 85
  table 94
regret data frame 278
reset_index() method 326
residual error 83
residuals 211
  homoscedasticity of 107
  independence of 107
  normality of 108
response variable 81
result evaluation 231
right-continuous 38
ROC (receiver operating characteristic) curve 138
roc_auc_score() method 137
root node 164–167
round() method 321
row_number variable 325, 327

S

s-charts 403
sample size, quality control measures 383
sample variable 389, 399
samples array 45
schedule 350
scipy library 47, 59, 339
scipy.stats 59, 62
scope 350
SCORE_DIFFERENTIAL variable 145, 152, 171–172
seaborn library 123
second_probability variable 49
sensitivity 133
SES (simple exponential smoothing) 216–217
set_option() method 116
set() method 331
shape attribute 159
shapiro() method 92
shift(1) method 263
sigma parameter 266
sigma2 parameter 211
sigmoid function 113
SimpleExpSmoothing class 217
simulating multiple trials 260
simulating prices 260
simulation 4
simulation_df data frame 266
size variable 389
size() method 165, 326
skew() method 346
skewness measures 345
sklearn (scikit-learn) library 134
sklearn.tree module 163
sort_values() method 346
specificity, defined 134
splitting data 158
sprints, defined 352
Spyder 7
SSE (sum of squares error) 103
SSR (sum of squares regression) 99–100, 103
SST (sum of squares total) 99–100
stage1 predictor variable 94
standard deviation 90
  quality control measures 381
states 296–307
  matrix of transition probabilities 300–307
  vector of state probabilities for multistate systems 297–300
stationarity 193
  differencing applied 197–205
    ACF plot 199
    Augmented Dickey–Fuller (ADF) test 201
    first-order differencing 202–205
    PACF plot 200
    visual inspection 197
  statistical properties 193
statistical parameters 260

statistics
  benefits of learning 8–11
  book structure 12–14
  overview 1
  probability distributions and conditional probabilities 41
  quantitative methods 2–4
  what this book does not cover 14
stats module 92, 339
stats.chisquare() method 339–341
stats.randint() method 59
statsmodels library 93, 125, 127, 199–200, 207
statsmodels.stats package 107
statsmodels.stats.diagnostic module 108
statsmodels.tsa.arima.model module 207
statstools module 107
Stays Healthy node 292
std (standard deviation) 118
std err parameter 211
steady-state probabilities 308
stock_data data frame 188
street addresses 330–333
Strong column 278
strong_econ estimate 281
strong_economy vector 275
subjective probability 19
subtasks 351
subtrees 288
sum() method 97, 322
summary statistics 90
summation, defined 36
symptoms variable 75

T

tail() method 89, 116–117, 130–131, 189, 326–327, 346
target variable 81
tasks 224, 351
terminal nodes 164
test data frame 197, 214
test_results variable 75
theoretical probabilities 19
three-sigma rule 44
Ticker object 188
ticker symbols 187
time series models 184
  distinguishing forecasts from predictions 185
  importing and plotting data 186–191
time series, fitting models 192, 207–220
  applicability 216
  AR and MA components 205–207
  autoregression (AR) component 192
  choosing between ARIMA and exponential smoothing 217
  combining ARIMA components 192
  differencing 195
  evaluating model fit 209–213
  forecasting 214
  Holt–Winters model 218–220
  integration (I) component 192
  MA (moving average) component 192
  mathematical properties 216
  model structure 216
  SES and DES models 217
  stationarity 193
  stationarity and differencing applied 197–205
  types of 216
total_cost variable 234
total_outcomes variable 58
traffic 72
train data frame 197, 202, 208
train_diff data frame 204
train_test_split() method 159
transition probabilities, matrix of 300–307
  making future state predictions programmatically 305–307
  predicting period 1 given period 0 302
  predicting period 2 given period 0 303–305
  predicting period 2 given period 1 302
trees, decision trees 157–173
  advantages and disadvantages of 173
  evaluating model, confusion matrix 161
  fitting model 160
  interpreting and understanding 164–173
  plotting 163
  predicting responses 161
  splitting data 158
trials, simulating series of 253
two-way table 74

U

UCL (upper control limit) 380
uniform distribution of mantissas 324
uniform distributions 325–330
  discrete 56, 58
uniform module 59
uniform_distribution vector 326
unique_values variable 331
upper_threshold 92
usecols parameter 143, 330
user stories 224

V

value_counts() method 148
var (variance) 118
variance measures 345
vector of state probabilities 296
VIF (variance inflation factor) 126
visual inspection 197
VS Code (Visual Studio Code) 7
vstack() function 235

W

WBS (work breakdown structure) 349–354
weak_econ estimate 281
weak_economy vector 275
weather forecasting 72
world population figures 334–335
wrangling data 142, 144
  data filtering 144
  data transformation 145
  data type conversion 146
  derived variable 146
  missing values 145

X

X_test test data 177
x-bar charts 400

Y

y_pred object 161, 177
y_train test data 177
YARDS_GAINED variable 145–146, 159
yf alias 187
yfinance library 187

Z

z parameter 211
z-score, quality control measures 386
z-statistic 343