Index

Symbols

1NF (first normal form), 59
2NF (second normal form), 59

A

A/B testing, 283-284, 288
accuracy-related metrics, 252
ACID (atomicity, consistency, isolation, durability), 68
active learning, 101-102
ad hoc analytics, 162
adaptability, 31
adversarial attacks, 272
adversarial augmentation, 116
AI (artificial intelligence), ethics, 339, 347-348
  data-driven approach limitations, 349
  irresponsible, case studies, 341-347
  mitigating biases, 353
  model cards, 351-353
  trade-offs, 349
Airflow, 315-316
alert fatigue, 255, 259
alert policies, 259
algorithms
  bandit algorithms, 287-291
  continual learning and, 273-274
  feature importance, 142
analytical processing, 67
Apache Iceberg, 69
architectural search, 174
Argo, 316-318
artifacts, 162
artificial intelligence (see AI)
asynchronous prediction, 198
automated retraining, 275-277
AutoML
  architecture search, 174-178
  hard AutoML, 174-178
  hyperparameter tuning, 173-174
  learned optimizer, 174-178
  soft AutoML, 173-174
autoscaling, 30

B

bagging, ensembles, 158-159
bandit algorithms, 287-291
BASE (basically available, soft state, and eventual consistency), 68
base learners, 156
base model, fine tuning, 100
baselines, offline model evaluation, 179
  existing solutions, 181
  human, 180
  random, 180
  simple heuristic, 180
  zero rule, 180
batch pipeline, 203-205
batch prediction, 197-201
  moving to online prediction, 201-203
batch processing, 78-79
batches, overfitting, 167
binary classification, 37
binary data, 57
binary file size, 57
boosting, ensembles, 159-161
Borg, 314
brand monitoring, 12
browsers, ML (machine learning) and, 222

building versus buying, 327-329
business analysis, 35
business objectives, 26-28

C

calibration, 183-184
canary release, 285
cardinality, classification tasks and, 37
catastrophic forgetting, 264
categorical features, 129-132
champion model, 264
churn prediction, 104
class imbalance, 102
  algorithm-level methods, 110
    class-balanced loss, 112
    cost-sensitive learning, 111
    focal loss, 112
  challenges, 103-105
  evaluation metrics, 106-108
  resampling, 109-110
class-balanced loss, 112
classification
  as regression problem, 107
  binary, 37
  hierarchical, 38
  high cardinality, 37
  multiclass, 37, 38
  multilabel, 38
  sentiment analysis, 120
classification models, 36
cloud computing, 212, 300-302
  elasticity, 300
  multicloud strategy, 302
code versioning, 164
column deletion, 125
column-major formats, 54-56
  pandas, 56
  Parquet, 54
Commuter, 305
compact convolutional filters, 206
computational priorities, 15
compute-intensive problems, 6
concept drift, 238, 241
confidence measurement, 185
containers, 308-310
contextual bandits, 289
continual learning, 35, 264, 268-270
  algorithms and, 273-274
  evaluation and, 272-273
  feature reuse, 277
  fresh data access, 270-272
  stateful training, 265-268
    automated, 277-278
  stateless retraining, 265-268
    manual, 275
  training, automated retraining, 275-277
  versus online learning, 268
convenience sampling, 83
cost-sensitive learning, 111
covariate data distribution shift, 238-240
cron, schedulers, 313-314
cross-functional collaboration, teams, 335
CSV (comma-separated values), row-major format, 54

D

DAG (directed acyclic graph), 312
dashboards, monitoring and, 258
data, 5, 18
  mind versus data, 43-46
  training (see training data)
  unseen data, 6
data augmentation, 113
  adversarial augmentation, 116
  data synthesis, 116-117
  perturbation, 114-116
  simple label-preserving transformations, 114
data distribution shifts
  addressing, 248-250
  detection
    statistical methods, 243-244
    time scale windows, 245-247
  ML system failure, 237
    concept drift, 238, 241
    covariate shift, 238-240
    feature change, 241
    label schema change, 241
    label shift, 238, 240
data duplication, data leakage and, 139
data engineering, 34
data formats, 53
  binary, 57
  column-major, 54-56
  JSON, 54
  multimodal data, 53
  relational model, NoSQL, 63-66
  row-major, 54-56

text, 57
data freshness, model updates and, 279-280
data generation, data leakage and, 140
data iteration, 267
model updates and, 281
data leakage, 135
data duplication prior to splitting, 139
data generation process and, 140
detecting, 140
group leakage, 139
Kaggle competition, 136
scaling before splitting, 138
statistics from test split, missing data and, 138
time-correlated data, 137
data models
relational, 59-62
structured data, 66-67
unstructured data, 66-67
data normalization, 59
data parallelism, distributed training and, 168-170
data scientists, teams, 336-339
data sources, 50
databases, internal, 52
logs, 51
smartphones and, 52
system-generated data, 50
third-party data, 52
user input, 50
data synthesis, 116-117
data-driven approach, AI ethics and, 349
databases and dataflow, 72
dataflow, 72
message queue model, 77
passing through databases, 72
passing through real-time transport, 74-77
passing through services, 73-74
request driven, 75
DataFrame, pandas and, 56
debugging, 165
decision trees, pruning, 208-209
declarative ML systems, 62
deep learning
ML (machine learning) and, 1
ML algorithms and, 150
degenerate feedback loops, ML system failure, 233
correcting, 235-236
dependencies, 312
ML models, model store, 322
dependency failure, 227
deployment, 34, 192
endpoints, exposing, 192
failure, 227
ML models, 320
myths
limited models at once, 194-195
model updating, 196
performance, 195
scale, 196
separation of responsibilities, 193
shadow deployment, 282
development environment, infrastructure, 296, 302
containers, 308-310
setup, 303
IDE, 303-306
standardization, 306-308
directed acyclic graph (DAG), 312
directional expectation tests, 183
discretization, feature engineering and, 128-129
distributed training, 168
data parallelism and, 168
model parallelism and, 170-172
Docker Compose, 310
Docker images, 308-310
Dockerfiles, 308-310
document model, 63
schemas, 64
downtime, 228
driver management service, 73
dynamic sampling, 110

E

edge cases
failure and, 231-231
outliers and, 232
edge computing, 213
model optimization, 214-221
EKS (Elastic Kubernetes Service), 314
embedding
positional embedding, 133-135
word embeddings, 133
endpoint, exposing, 192
ensembles
bagging, 158-159
base learners, 156

boosting, 159-161
spam classifiers, 157
stacking, 161
ethics in AI, 339-347
ETL (extract, transform, load), 70-72
evaluation, offline
confidence measurement, 185
directional expectation tests, 183
invariation tests, 182
model calibration, 183-184
perturbation tests, 181-182
slice-based, 185-188
existing data, 5
experiment artifacts, development and, 323
experiment tracking, 162-163
third-party tools, 163
exporting models, 193

F

F1 metrics, 107
factorization, low-rank, 206-208
fairness, 19
feature change, 241
feature engineering, 120-122
categorical features, 129-132
discretization, 128-129
feature crossing, 132
feature generalization, 144-146
feature importance, 142
missing values and, 123
deletion, 125
imputation, 125-126
MAR (missing at random), 124
MCAR (missing completely at random), 124
MNAR (missing not at random), 124
NLP (natural language processing) and, 122
positional embeddings, 133-135
predictive power of features, 140
scaling, 126-128
useless features, 141
feature scaling, 126-128
feature store, 325-327
features
computation, 326
consistency, 326
extracting, 255
failures and, 166
learned, 120-122
management, 326
monitoring, 253-255
online, 199
reuse, 277
streaming, 199
feedback loops, 288
ML system failure, 234
feedback, users, 93
fixed positional embeddings, 135
fixed-point inference, 210
FLOPS (floating-point operations per second), 298
forecasting customer demand, 11
Fourier features, 135
fraud detection, 11, 104

G

GDPR (General Data Protection Regulation), 164
generalization, features, 144-146
GKE (Google Kubernetes Engine), 314
Google Translate, 1
graph model, 65

H

H20 AutoML, 62
hand labels, 88
lineage, 90
multiplicity, 89-90
hard AutoML, 174-178
hardware failure, 228
hashed functions, 130
heuristics, LFs (labeling functions), 95
heuristics-based slicing, 188
hierarchical classification, 38
human baselines, 180
hyperparameters
failures and, 166
tuning, 173-174
values over time, 163

I

IDE (integrated development environment), 303
cloud dev environment, 307
notebooks and, 304
importance sampling, 87
infrastructure, 293, 295

building versus buying, 327-329
cloud computing and, 300-302
development environment layer, 296, 302
setup, 303-306
fundamental facilities, 295
ML platform layer, 296
requirements, 295
resource management layer, 295
storage and compute layer, 295, 296, 297
compute resources, 297
FLOPS, 298
private data centers, 300-302
public cloud, 300-302
units, 297
input, monitoring, 255
instances on-demand, 300
integrated development environment (see IDE)
interleaving experiments, 285-287
internal databases, 52
interpretability, 20
invariation tests, 182
IR (intermediate representation), 215
iterative processes
model development and, 34
performance check, 149
model updates and, 281
training the model and, 32-33
data engineering, 34
project scoping, 34

J

JSON (JavaScript Object Notation), 54
judgment sampling, 83

K

k-means clustering models, 150
Kaggle, data leakage, 136
knowledge distillation, 208
Kubeflow, 318
Kubernetes (K8s), 310, 314
EKS (Elastic Kubernetes Service), 314
GKE (Google Kubernetes Engine), 314

L

label computation, 271
label schema change, 241
label shift, 238, 240
labeling, 88
class imbalance and, 102
errors, class imbalance and, 105
hand labels, 88
lineage, 90
multiplicity, 89-90
lack of labels, 94
active learning, 101-102
semi-supervision, 98-99
transfer learning, 99-101
weak supervision, 95-98
ML algorithms, 151
natural labels, 91
feedback loop length, 92
recommender systems, 91
perturbation, 114-116
simple label-preserving transformations, 114
language modeling, sampling and, 83
latency, 16
latency versus throughput, 16-18
learning, 3
LFs (labeling functions), 95
heuristics, 95
logs, 51, 51
experiment tracking, 162
monitoring and, 256-257
storage, 51
loop tiling, model optimization, 218
loss curve, 162
loss functions, 40
(see also objective functions)
low-rank factorization, 206-208

M

maintainability, 31
Manning, Christopher, 44
MAR (missing at random) values, 124
MCAR (missing completely at random) values, 124
merge conflicts, 164
message queue, dataflow and, 77
Metaflow, 318
metrics
monitoring and, 250
accuracy-related metrics, 252
features, 253-255
predictions, 252-253
raw input, 255
performance metrics, 162

system performance, 163
mind versus data, 43-46
missing at random (MAR), 124
missing completely at random (MCAR), 124
missing data, test split statistics and, 138
missing not at random (MNAR), 124
ML (machine learning)
browsers and, 222, 223
cloud computing, 212-223
complex patterns, 4
deep learning and, 1
edge computing, 212-223
existing data and, 5
learning, 3
model optimization, 220-221
predictions and, 6
production and, 12-21
repetition, 7
research and, 12-21
scale, 7
smartphones and, 9
unseen data, 6
use cases, 9-12
when to use, 3-12
ML algorithms, 2, 149
deep learning and, 150
labels, 151
versus neural networks, 150
ML model logic, 191
ML models
continual learning, 35
data iteration, 267
debugging, 165
deployment, 320
edge computing, optimization, 214-221
ensembles, 156, 157
bagging, 158-159
base learners, 156
boosting, 159-161
stacking, 161
evaluation, 150
test in production, 281-291
experiment tracking, 162-163
exporting, 193
failures
batches, overfitting, 167
components, 167
data problems, 166
feature choice, 166
hyperparameters and, 166
poor model implementation, 166
random seeds, 167
theoretical constraints, 166
iteration, 267
monitoring, 35
offline evaluation, 178
baselines, 179-181
methods, 181-188
optimization, 220-221
parameters, model store, 322
performance metrics, 162
selection criteria, 151
human biases in, 153
model, 155
performance now and later, 153
simple models, 152
state-of-the-art trap, 152
trade-offs, 154
speed, 163
training, 32-33
data engineering, 34
distributed, 168-172
update frequency, 279
data freshness and, 279-280
data iteration and, 281
model iteration and, 281
updates, 267
versioning, 163-165
ML platform, 319
model deployment, 320
model store, 321-325
ML platform layer, infrastructure, 296
ML system failures
data distribution shifts, 237
addressing, 248-250
concept drift, 238, 241
covariate, 238-240
detection, 242-247
feature change, 241
label schema change, 241
label shifts, 238, 240
ML-system specific
degenerate feedback loops, 233-236
edge cases, 231
production data different from training data, 229-231
operational expectation violations, 227
software

crashes, 228
dependency failure, 227
deployment failure, 227
downtime, 228
hardware failure, 228
ML systems
declarative, 62
failures, 226
iterative processes, 32-35
requirements
adaptability, 31
maintainability, 31
reliability, 29
scalability, 30-31
versus traditional software, 22-23
MLOPs, ML systems design and, 2-3
MNAR (missing not at random) values, 124
model biases, AI ethics, 347-348
model calibration, 183-184
model cards, AI ethics, 351-353
model compression, 206
knowledge distillation, 208
low-rank factorization, 206-208
pruning, 208-209
quantization, 209-211
model development, 34
model implementation, failures and, 166
model parallelism, distributed training and, 170-172
model performance, business analysis, 35
monitoring, 250, 263
(see also test in production)
alerts and, 259
dashboards and, 258
logs and, 256-257
metrics and, 250
accuracy-related metrics, 252
features, 253-255
predictions, 252-253
raw input, 255
multiclass classification, 37, 38
multilabel classification, 38
multimodal data, 53

N

n-grams, 120
NAS (neural architecture search), 174
natural labels, 91
feedback loop length, 92
recommender systems, 91
natural language processing (NLP) (see NLP)
neural architecture search (NAS), 174
neural networks, 150
positional embedding, 133
newsfeeds
ranking posts, 41
user engagement and, 41
NLP (natural language processing), 114
data augmentation and, 113
feature engineering, 122
nonprobability sampling, 83
biases, 83
Norvig, Peter, 44
NoSQL, 63
document model, 63
graph model, 65
notebooks, IDE and, 304
NSFW (not safe for work) content filtering, 41
NumPy, 56

O

objective functions, 40-43
observability, 250, 259-261
offline evaluation of models, 178
baselines, 179
existing solutions, 181
human, 180
random, 180
simple heuristic, 180
zero rule, 180
OLAP (online analytical processing), 69
OLTP (online transaction processing) system, 69
on-demand instances, 300
on-demand prediction, 198
One Billion Word Benchmark for Language Modeling, 45
online features, 199
online learning, 268
online prediction, 197-201, 288
moving to from batch prediction, 201-203
streaming pipeline, 203-205
operation expectation violations, 227
operator fusion, model optimization and, 218
orchestrators
HashiCorp Nomad, 314
Kubernetes (K8s), 314
outliers, edge cases and, 232

oversampling
  overfitting, 110
  SMOTE, 110

P

pandas, 56
Papermill, 305
parallelization, model optimization and, 217
parameter values over time, 163
Pareto optimization, 42
Parquet, 54, 57
  binary files, 57
patterns
  changing, 8
  complex, 4
Pearl, Judea, 43
performance metrics, 162
  system performance, 163
perturbation, 114-116
perturbation method of semi-supervision, 99
perturbation tests, 181-182
positional embedding, 133-135
  fixed, 135
precision metrics, 107
prediction, 6, 39
  asynchronous, 198
  batch prediction, 197-201
    moving to online prediction, 201-203
  "mostly correct," user experience, 332-334
  on-demand prediction, 198
  online, 197-201
    streaming pipeline, 203-205
  synchronous, 198
predictions, monitoring, 252-253
predictive power of features, 140
price optimization service, 73
problem framing, 35-43
processing
  analytical, 67
  batch processing, 78-79
  ETL (extract, transform, load), 70-72
  stream processing, 78-79
  transactional, 67
    ACID and, 68
production environment, 192
production, ML and, 12-21
project objectives, 26-28
project scoping, 34
prototyping, batch prediction and, 201
pruning, 208-209
public cloud versus private data center, 300-302

Q

quantization, 209-211
query languages, 60
quota sampling, 83

R

random baselines, 180
real-time transport
  dataflow and, 74-77
  streaming data and, 78
reasonable scale, 294
recall metrics, 107
recommender systems, labels, 91
regression
  class imbalance and, 102
  tasks, 39
regression models, 36
relational databases, 60
relational models, 59-62
  data normalization, 59
  NoSQL, 63
    document model, 63
    graph model, 65
  tables, 59
reliability, 29
repetition, 7
repetitive jobs, scheduling, 311
request-driven data passing, 75
resampling, 109
  dynamic sampling, 110
  oversampling
    overfitting and, 110
    SMOTE, 110
  two-phase learning, 110
  undersampling, 109
reservoir sampling, 86-87
resource management, 311
resource management layer, infrastructure, 295
REST (representational state transfer), 74
ride management service, 73
ROC (receiver operating characteristics) curve, 108
Rogati, Monica, 44
ROI (return on investment), maturity stage of adoption, 28
row deletion, 125

row-major format, 54-56
  CSV (comma-separated values), 54
  NumPy, 56
RPC (remote procedure call), 74

S

sampling, 82
  importance sampling, 87
  nonprobability, 83
    biases, 83
  reservoir sampling, 86-87
  simple random sampling, 84
  stratified sampling, 84
  weighted sampling, 85
scalability, 30-31
  autoscaling, 30
scale, 7
  deployment myths, 196
schedulers, 313-314
  Borg, 314
  Slurm, 314
schemas, document model, 64
scoping a project, 34
self-training, 98
semi-supervision, 98-99
sentiment analysis classifier, 120
serialization, 193
services
  dataflow and, 73-74
  driver management, 73
  price optimization, 73
  ride management, 73
SGD (stochastic gradient descent), 169
shadow deployment, 282
SHAP (SHapley Additive exPlanations, 142
simple heuristic, offline evaluation, 180
simple label-preserving transformations, 114
simple random sampling, 84
Simpson's paradox, 186
skewed distribution, feature scaling and, 127
slice-based evaluation, 185-188
slicing
  error analysis, 188
  heuristics based, 188
  slice finders, 188
Slurm, 314
smartphones
  data sources and, 52
  ML (machine learning) and, 9
smooth failing, user experience, 334
SMOTE (synthetic minority oversampling technique), 110
Snorkel, 95
snowball sampling, 83
soft AutoML, 173-174
software system failure
  crashes, 228
  dependency, 227
  deployment, 227
  hardware, 228
spam filtering, 41
splitting
  data duplication, 139
  data leakage and, 138
SQL, 60
SQL databases, 61
SSD (solid state disk), 297
stacking, ensembles, 161
stakeholders, research projects, 13-15
state-of-the-art models, 152
stateful training, 265-268
  automated, 277-278
stateless retraining, 265-268
  manual, 275
stochastic gradient descent (SGD), 169
storage and compute layer, infrastructure, 295, 296, 297
  compute resources, 297
  FLOPS (floating-point operations per second), 298
  private data centers, 300-302
  public cloud, 300-302
  units, 297
storage engines, 67
stratified sampling, 84
stream processing, 78-79
streaming data, real-time transport, 78
streaming features, 199
streaming pipeline, 203-205
structured data, 66-67
Sutton, Richard, 44
synchronous prediction, 198
synthetic minority oversampling technique (SMOTE), 110
system performance metrics, 163
system-generated data, 50

T

tags, model store, 323
tasks
  classification, 36
    binary, 37
    high cardinality, 37
    multiclass, 37, 38
    multilabel, 38
  labels, 91
  regression, 36, 39
teams
  cross-functional collaboration, 335
  data scientists, 336-339
  production management, 336
telemetry, 260
test in production, 263, 281
  A/B testing, 283-284
  bandits, 287-291
  canary release, 285
  interleaving experiments, 285-287
  shadow deployment and, 282
text data, 57
text file size, 57
theoretical constraints, failures and, 166
third-party data, 52
time-correlated data, data leakage and, 137
training
  automated retraining, 275-277
  distributed, 168
    data parallelism and, 168-170
    model parallelism and, 170-172
  stateful, 265-268
    automated, 277-278
  stateless retraining, 265-268
    manual, 275
training data, 81
  class imbalance, 102
    algorithm-level methods, 110-113
    challenges, 103-105
    evaluation metrics, 106-108
    resampling, 109-110
  data augmentation, 113
    perturbation, 114-116
    simple label-preserving transformations, 114
  data distributions, 229
  data leakage, 135
  labeling, 88
    hand labels, 88-90
    lack of labels, 94-102
    natural labels, 91-94
    user feedback, 93
  n-grams, 121
  noisy samples, 116
  sampling, 82
    importance sampling, 87
    nonprobability, 83-84
    reservoir sampling, 86-87
    simple random sampling, 84
    stratified sampling, 84
    weighted sampling, 85
training the model, iteration and, 32-33
  data engineering, 34
  project scoping, 34
transactional processing, 67
  ACID and, 68
transfer learning, 99-101
two-phase learning, 110

U

undersampling, 109
unseen data, 6
unstructured data, 66-67
updates, deployment myths, 196
use cases, 9-12
user experience, 331
  consistency, 332
  predictions, mostly correct, 332-334
  smooth failing, 334
user feedback, 93
user input data, 50

V

vCPU (virtual CPU), 299
vectorization, model optimization, 217
versioning, 163-165
  code versioning, 164

W

WASM (WebAssembly), 223
weak supervision, 95-98
  Snorkel, 95
weighted sampling, 85
word embeddings, 133
workflow management, 314
  Airflow, 315-316
  Argo, 316-318

DAG (directed acyclic graph), 312
Kubeflow, 318
Metaflow, 318

X

XGBoost, 142

Z

zero rule baselines, 180
zero-shot learning, 100