

MACHINE LEARNING QUICK REFERENCE: BEST PRACTICES

Topic	Common Challenges	Suggested Best Practice
Data Preparation		
Data collection	<ul style="list-style-type: none"><li>Biased data</li><li>Incomplete data</li><li>The curse of dimensionality</li><li>Sparsity</li></ul>	<ul style="list-style-type: none"><li>Take time to understand the business problem and its context</li><li>Enrich the data</li><li>Dimension-reduction techniques</li><li>Change representation of data (e.g. COO)</li></ul>
“Untidy” data	<ul style="list-style-type: none"><li>Value ranges as columns</li><li>Multiple variables in the same column</li><li>Variables in both rows and columns</li></ul>	Restructure the data to be “tidy” by using the melt and cast process
Outliers	<ul style="list-style-type: none"><li>Out-of-range numeric values and unknown categorical values in score data</li><li>Undue influence on squared loss functions (e.g. regression, GBM, and <i>k</i>-means)</li></ul>	<ul style="list-style-type: none"><li>Robust methods (e.g. Huber loss function)</li><li>Discretization (binning)</li><li>Winsorizing</li></ul>
Sparse target variables	<ul style="list-style-type: none"><li>Low primary event occurrence rate</li><li>Overwhelming preponderance of zero or missing values in target</li></ul>	<ul style="list-style-type: none"><li>Proportional oversampling</li><li>Inverse prior probabilities</li><li>Mixture models</li></ul>
Variables of disparate magnitudes	<ul style="list-style-type: none"><li>Misleading variable importance</li><li>Distance measure imbalance</li><li>Gradient dominance</li></ul>	Standardization
High-cardinality variables	<ul style="list-style-type: none"><li>Overfitting</li><li>Unknown categorical values in holdout data</li></ul>	<ul style="list-style-type: none"><li>Discretization (binning)</li><li>Weight of evidence</li><li>Leave-one-out event rate</li></ul>
Missing data	<ul style="list-style-type: none"><li>Information loss</li><li>Bias</li></ul>	<ul style="list-style-type: none"><li>Discretization (binning)</li><li>Imputation</li><li>Tree-based modeling techniques</li></ul>
Strong multicollinearity	Unstable parameter estimates	<ul style="list-style-type: none"><li>Regularization</li><li>Dimension reduction</li></ul>
Training		
Overfitting	High-variance and low-bias models that fail to generalize well	<ul style="list-style-type: none"><li>Regularization</li><li>Noise injection</li><li>Partitioning or cross validation</li></ul>
Hyperparameter tuning	Combinatorial explosion of hyper-parameters in conventional algorithms (e.g. deep neural networks, Super Learners)	<ul style="list-style-type: none"><li>Local search optimization, including genetic algorithms</li><li>Grid search, random search</li></ul>
Ensemble models	<ul style="list-style-type: none"><li>Single models that fail to provide adequate accuracy</li><li>High-variance and low-bias models that fail to generalize well</li></ul>	<ul style="list-style-type: none"><li>Established ensemble methods (e.g. bagging, boosting, stacking)</li><li>Custom or manual combinations of predictions</li></ul>
Model Interpretation	Large number of parameters, rules, or other complexity obscures model interpretation	<ul style="list-style-type: none"><li>Variable selection by regularization (e.g. L1)</li><li>Surrogate models</li><li>Partial dependency plots, variable importance measures</li></ul>
Computational resource exploitation	<ul style="list-style-type: none"><li>Single-threaded algorithm implementations</li><li>Heavy reliance on interpreted languages</li></ul>	<ul style="list-style-type: none"><li>Train many single-threaded models in parallel</li><li>Hardware acceleration (e.g. SSD, GPU)</li><li>Low-level, native libraries</li><li>Distributed computing, when appropriate</li></ul>
Deployment		
Model deployment	Trained model logic must be transferred from a development environment to an operational computing system to assist in organizational decision making processes	<ul style="list-style-type: none"><li>Portable scoring code or scoring executables</li><li>In-database scoring</li><li>Web service scoring</li></ul>
Model decay	<ul style="list-style-type: none"><li>Business problem or market conditions have changed since the model was created</li><li>New observations fall outside domain of training data</li></ul>	<ul style="list-style-type: none"><li>Monitor models for decreasing accuracy</li><li>Update/retrain models regularly</li><li>Champion-challenger tests</li><li>Online updates</li></ul>

