

Homework 1-2 + Lab 1 Reading Checklist

Mapped to Week 1 / Week 2 / Week 3 milestones for HW1, HW2, and Lab 1

Instructions: Directly search for each reading online and summarize 2-3 actionable takeaways.

Week 1 - Dataset Selection, Advanced EDA, Leakage Control, and Baseline Evaluation

- scikit-learn: Common pitfalls (data leakage, improper validation, preprocessing outside cross-validation)
- scikit-learn: Train/test splitting - train_test_split (random_state, stratify, and when to use time-based splits)
- scikit-learn: Pipelines - Pipeline (preprocessing + model fit without leakage)
- scikit-learn: ColumnTransformer (mixed numeric/categorical features without leakage)
- Variance Inflation Factor (VIF): how to compute it and interpret multicollinearity; practical thresholds and remedies
- Advanced EDA visuals: pair plots with KDE + clustered correlation heatmaps (and what conclusions are valid vs. overreach)
- Paper: "Datasheets for Datasets" (Gebru et al., 2021) - documentation, recommended uses, and risk analysis
- Optional: Kaggle tutorial: "Data Leakage" (concrete leakage examples to avoid) - <https://www.kaggle.com/code/residentmario/leakage-especially-knowledge-leakage>

Week 2 - Logistic Regression Objective + Optimization; Trees/Ensembles; Proper CV and Tuning

- Logistic regression objective: MLE derivation, log-loss, and assumptions (independence, linear decision boundary)
- MAP for logistic regression: priors -> regularization (L2/L1); how MAP differs from MLE and when it helps
- Gradient descent variants: batch vs. SGD vs. mini-batch; learning-rate selection and convergence intuition
- Paper: "Adam: A Method for Stochastic Optimization" (Kingma & Ba, 2015) - plus brief notes on Momentum and RMSProp
- scikit-learn: Cross-validation overview (what CV estimates and common misuses)
- scikit-learn: StratifiedKFold (recommended for imbalanced classification)
- scikit-learn: GridSearchCV / RandomizedSearchCV (tuning protocol + fair comparisons)
- scikit-learn: Metrics reference (precision, recall, F1, ROC-AUC; regression: MAE, RMSE)

- Paper: "Model Cards for Model Reporting" (Mitchell et al., 2019) - intended use, evaluation, limitations, ethics
- Optional: Paper: "Underspecification" (D'Amour et al., 2020) - why similar test scores can hide different behavior
- scikit-learn: DecisionTreeClassifier - depth/complexity controls, visualization, and cost-complexity pruning
- Paper: "Bagging Predictors" (Breiman, 1996) - why bagging reduces variance
- Paper: "Greedy Function Approximation: A Gradient Boosting Machine" (Friedman, 2001) (boosting fundamentals)
- Sensitivity analysis: what it is, how to do it responsibly (feature perturbations, partial dependence cautions)

Week 3 - Statistical Comparison, Interpretation, Calibration/Thresholding, and Shift/Monitoring

- Model comparison basics: confusion matrix, classification report, and metric selection under class imbalance
- Statistical testing for model differences: paired t-test basics, when it is (and is not) appropriate; alternatives (e.g., McNemar)
- Bias-variance trade-off: diagnosing under/overfitting across trees, bagging, and boosting
- scikit-learn: Permutation importance (model-agnostic importance + caveats)
- scikit-learn: Partial dependence / ICE plots (communicating feature effects responsibly)
- scikit-learn: Probability calibration + operating threshold selection using a cost model (Platt vs. isotonic)
- Paper: "Probabilistic Outputs for Support Vector Machines" (Platt, 1999) - classic calibration approach
- Paper/book: "Dataset Shift in Machine Learning" (Quionero-Candela et al., 2009) - drift, shift types, and monitoring
- Drift metrics: PSI or KL divergence after binning; how to interpret drift vs. model performance degradation
- Practice: Write 15-25 error-case notes (classification) OR analyze largest residuals (regression) using a failure-mode taxonomy
- Optional: imbalanced-learn Pipeline guidance (resampling must occur inside CV folds)
- Optional: Paper: "XGBoost: A Scalable Tree Boosting System" (Chen & Guestrin, 2016) - for the graduate extension

Report tip: List the readings you used (by title) and include 2-3 takeaways that directly shaped your choices (split strategy, pipeline, metrics, tuning, ablations, and error analysis).