Week 7 Summary: Imbalanced Evaluation, Cross-Validation, and Ensembles

Z3003 Data Mining: Scaling to Large Data · 4–5 November 2026 · Dr. Innocent Nyalala

Lecture 19: Evaluating Classifiers Beyond Accuracy

On 56,988 test farm-seasons with roughly 10% failed seasons, a constant majority-class predictor scores 89.8% accuracy while catching 0 failures. Evaluate from the confusion matrix (TP, FN, FP, TN):

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \qquad F_1 = \frac{2PR}{P+R},$$
$$\text{MCC} = \frac{\text{TP}\cdot\text{TN} - \text{FP}\cdot\text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}}.$$

Lowering the classification threshold on the depth-8 tree from 0.5 to 0.2 lifts recall from 12% to 53% at 33% precision. Use ROC-AUC for threshold-free ranking and Precision-Recall curves when positives are rare.

Lecture 20: Cross-Validation, Bias-Variance, and Leakage Prevention

Single splits fluctuate (66.8% to 72.2% across 20 splits); k-fold cross-validation evaluates every record out-of-fold to estimate mean and spread. Low training and validation scores signal high bias; high training score with low validation score signals high variance. Prevent leakage via `Pipeline` (fit scalers per fold), `GroupKFold` (keep each farm in one fold), and `TimeSeriesSplit`.

Lecture 21: Bagging, Random Forests, and Boosting

Averaging n independent predictors reduces variance by σ^2/n. Bagging trains deep trees on bootstrap samples (leaving $\approx$ 36.8% out-of-bag); Random Forests also subsample $\approx \sqrt{d}$ features per split. Sequential boosting (AdaBoost reweighting errors, Gradient Boosting fitting residuals) reduces bias. On the farm-seasons data, pruned tree (72.1%), Random Forest (72.2%), and Gradient Boosting (72.3%) plateau at the information ceiling set by missing soil and pest attributes.