CONSENSUS SHAP: ACTIONABLE RECOMMENDATIONS

FINDING 1: EXPLANATION AGREEMENT IS MODERATE

- Overall agreement: ρ = 0.57 ± 0.30
- Only 79% of instances have reliable explanations
- Single-model SHAP should NOT be trusted blindly

FINDING 2: MODEL FAMILY MATTERS

- Tree models agree with each other: ρ = 0.68
- Tree models disagree with Logistic Regression: ρ = 0.41
- XGBoost-LightGBM have highest agreement: ρ ≈ 0.79
- Difference is statistically significant (p < 0.001)

FINDING 3: DATASET CHARACTERISTICS PREDICT DISAGREEMENT

- Number of features affects agreement
- Prediction agreement correlates with explanation agreement
- Can predict ~-130% of variance in agreement using dataset features

RECOMMENDATION 1: USE TREE-ONLY CONSENSUS

- Average SHAP values from: XGBoost + LightGBM + CatBoost + Random Forest
- Exclude Logistic Regression (different explanation space)
- Improves reliability significantly

RECOMMENDATION 2: REPORT EXPLANATION UNCERTAINTY

- Compute std of SHAP rankings across models
- Flag instances with std > 0.3 as "unreliable"
- Report confidence intervals, not just point estimates

RECOMMENDATION 3: VALIDATE ON YOUR DOMAIN

- Agreement varies by dataset (0.35 to 0.80)
- Test consensus approach on your specific data
- Consider domain-specific reliability thresholds

RECOMMENDATION 4: REGULATORY COMPLIANCE

- EU AI Act requires explanations for high-risk AI
- Single-model explanations may be legally insufficient
- Consensus SHAP provides more defensible explanations