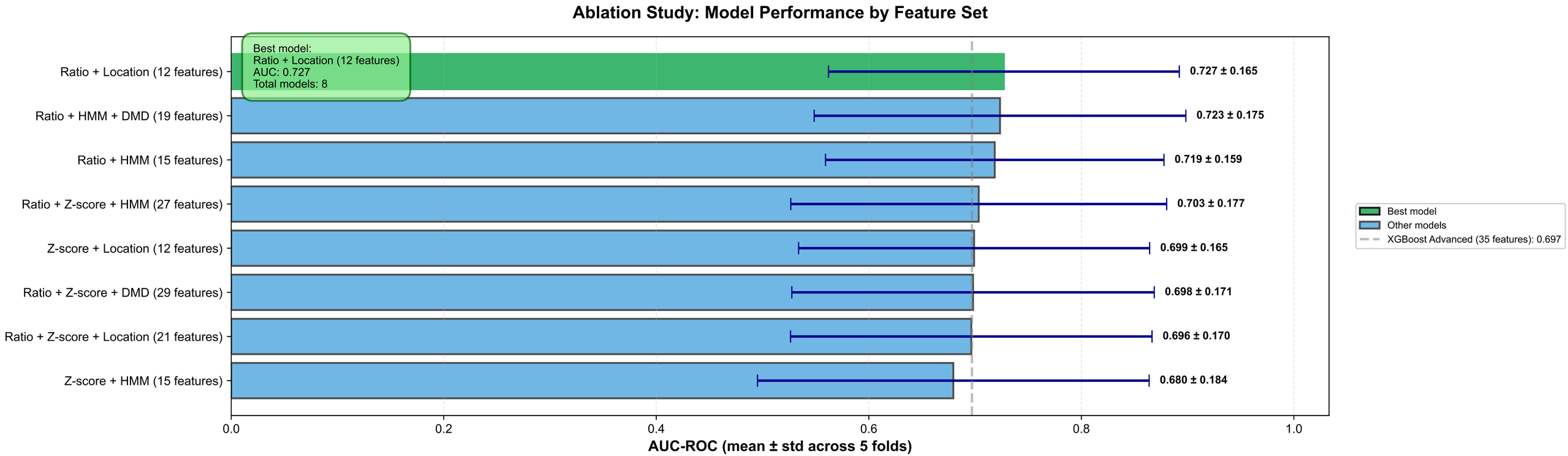
Ablation Study Performance with Z-Score Threshold Sensitivity Analysis



B. Z-Score Threshold Sensitivity Analysis (h=8)

Threshold	Precision	Recall	F1-Score	TP	FP	FN	TN
1.0sigma	0.299	0.366	0.329	5,148	12,061	8,924	27,920
2.0sigma	0.229	0.110	0.148	1,545	5,194	12,527	34,787
3.0sigma	0.150	0.007	0.013	98	554	13,974	39,427

2sigma threshold optimal for h=8 horizon: Balances precision (0.229) and recall (0.110)
Higher thresholds (3sigma) increase precision but sacrifice recall
Lower thresholds (1sigma) increase recall but reduce precision

Panel A: Ablation study comparing 8 Stage 2 XGBoost variants on WITH_AR_FILTER subset (6,553 observations, 534 districts where AR baseline got it wrong—the hardest cases). Ratio + Location (12 features, AUC=0.727) achieves highest discrimination for operational forecasting. Panel B: Z-score threshold sensitivity shows 2σ optimal for h=8 horizon (precision=0.229, recall=0.110, F1=0.148). Advanced features provide complementary scientific value beyond AUC: HMM (15-27 features) applies Bayesian state-space modeling to identify probabilistic regime transitions (HMM transition risk ranks #5 at 3.2% importance), revealing structural shifts in crisis narrative dynamics; DMD (19-29 features) employs spectral decomposition to isolate dominant temporal modes, detecting nonlinear escalation events (largest mixed-effects coefficient +352.38 identifies synchronized multicategory growth). XGBoost Advanced (35 features, AUC=0.697) integrates compositional, stochastic, and modal features for comprehensive crisis driver identification. Discrimination-interpretation tradeoff: parsimonious models optimize classification performance, theoretically-grounded models enable causal inference. 5-fold stratified spatial cross-validation, h=8 months.