Fake Job Postings Dataset

Dataset source: Real or Fake? Fake Job Posting Prediction https://www.kaggle.com/datasets/shivamb/real-or-fake-fake-jobposting-prediction (kaggle.com)

Number of Job Postings, Total: 17,880
Number of Fake Postings: 866
Fake Posting (Fraud) Incidence: 4.84%

Model Fitting Results for the Validation/Test Dataset (30% Sample comprised of 5,364 total postings including 260 fakes, a 4.84% fake rate)

Model Iteration	Model Type	TD-IDF Matrix Reduction Scope	Predicted Prob Threshold*	Accuracy	Precision	Recall	F1	ROC-AUC	PR-AUC	Model Notes & Interpretation
1	Logistic	Across all TD-IDF Field Matrices	0.5	0.83	0.2	0.88	0.33	0.914	0.4388	Very high recall, but poor precision (resulting in a lot of false positives, flagging valid job postings as 'fake')
2	Logistic	Across all TD-IDF Field Matrices	0.86	0.95	0.5	0.5	0.5	0.914	0.4388	More balanced precision and recall (but still lower than tree-based methods)
3	Random Forest	Across all TD-IDF Field Matrices	0.5	0.97	0.98	0.4	0.56	0.979	0.824	Extremely high precision and solid accuracy, but recall is only 0.40. Excellent at avoiding false positives, but misses many fraudulent postings
4	Random Forest	Across all TD-IDF Field Matrices	0.276	0.975	0.76	0.71	0.74	0.979	0.824	Very stong candidate. High accuracy, balanced precision and recall and high PR-AUC.
5	XGBoost	Across all TD-IDF Field Matrices	0.5	0.98	0.89	0.62	0.73	0.977	0.834	Strong balance across precision, recall, F1, and high PR-AUC. A reliable performer with good fraud detection and manageable false positives.
6	XGBoost	Across all TD-IDF Field Matrices	0.296	0.977	0.79	0.74	0.76	0.977	0.834	Very stong candidate. High accuracy, balanced precision and recall and high PR-AUC.
7	Logistic	For each TD-IDF Field Matrix	0.5	0.822	0.2	0.85	0.32	0.908	0.354	High recall but precision is low (lots of fals positives). Accuracy is modest. Like iteration 1, it's aggressive in catching fraud but at the expense of flagging valid postings as fake.
8	Logistic	For each TD-IDF Field Matrix	0.877	0.9466	0.45	0.46	0.46	0.908	0.354	More balanced precision and recall (but still lower than tree-based methods)
9	Random Forest	For each TD-IDF Field Matrix	0.5	0.979	0.91	0.63	0.75	0.99	0.886	Strong Solution; Excellent balance of precision and recall, producing a solid F1. High accuracy and PR-AUC. One of the strongest performers, especially for minimizing false positives while still catching frauds.
10	Random Forest	For each TD-IDF Field Matrix	0.317	0.979	0.81	0.77	0.79	0.99	0.886	Strongest Random Forest Candidate; Excellent balance of precision and recall, producing a solid F1. High accuracy and PR-AUC. One of the strongest performers, especially for capturing a high volume of all the fraud cases.
11	XGBoost	For each TD-IDF Field Matrix	0.5	0.984	0.89	0.78	0.83	0.99	0.91	Best overall candidate; Excellent balance of precision and recall, excellent PR-AUC. Among top scoring solution in all evaluation metrics
12	XGBoost	For each TD-IDF Field Matrix	0.342	0.982	0.82	0.82	0.82	0.99	0.91	Strong performer, comparable to the high performing xgboost model in row 11.

* Each technique and TD-IDF candidate feature set were fit the default probability threshold of 0.5 and then an 'optimal' threshold identified by finding the threshold tied to the highest estimated F1 score.