Corruption Detection in Public Procurement

PU Learning on ANAC data for Corruption Risk Scoring

Davide Bortoletto, Giorgia Caicchiolo, Gabriele Paganelli & Gianmarco Rosa • University of Padova

∼9.5M	3	781	2%	nnPU
ANAC Contracts	Datasets (M1/M2/M3)	Labeled Positives	Estimated Prior $\hat{\pi}$	Best Model

Overview

We develop a **corruption risk scoring system** for Italian public procurement, covering ∼9.5M contracts from ANAC (2008 to 2025). The problem is framed as **Positive-Unlabeled (PU) Learning**: 781 confirmed corrupt contracts, 1 817 confirmed clean ones, and over 9.4M unlabeled. The output is not a binary label but a **relative risk ranking**, designed to help control authorities concentrate limited inspection resources on the contracts most likely to be irregular.

Data

Sources: ANAC (National Public Contracts Database), ISTAT (provincial socioeconomic indicators), OpenGA (TAR and Council of State rulings).

3 temporal datasets reflecting procurement lifecycle phases:

- **M1 (ex ante):** 9.47M contracts, 24 variables, publication-time info only.
- **M2 (during):** 3.47M contracts, 40 variables, adds award data (bids, discount).
- **M3 (ex post):** 1.13M contracts, 60 variables, adds execution data (variants, cost overruns).

Datasets are nested (M3 $\subset$ M2 $\subset$ M1). Missing values mostly treated as informative categories.

Class Prior Estimation

The true fraction of corrupt contracts π is unobservable. Four estimators compared; **PULSNAR** adopted as the reference, the only one designed for the SAR (*Selected At Random*) regime, necessary because labeled positives are systematically biased by contract type and geography.

Dataset	Median	95% CI
M1	1.35%	[0.80%, 4.24%]
M2	1.53%	[0.78%, 6.35%]
M3	0.56%	[0.24%, 12.59%]

PULSNAR estimates converge with empirical literature ($\pi \approx 2\%$; Decarolis & Giorgiantonio 2022). Operational value: $\hat{\pi}_{\text{op}} = 0.02$, with sensitivity grid over $[0.005, 0.05]$.

Models

Four PU learning families, each in **PU** and **PNPU** (with confirmed negatives) variants:

- **Biased Learning:** unlabeled treated as down-weighted negatives. Applied with LR, SVM, RF, LightGBM.
- **EM-like (soft/hard):** iterative weight updates on unlabeled, collapse to iteration zero with judicial negatives; reduce to supervised LightGBM on P vs N.
- **nnPU:** Kiryo et al. (2017) risk estimator with non-negative correction. Applied to LR, LightGBM, Neural Net.
- **Bagging PU / PUET:** ensemble over unlabeled subsamples. PUET uses Extra Trees with nnPU-based split criterion.

4-fold stratified cross-validation, shared across all models.

Results

Metrics: **Lift@1%** (operational relevance) and **PR AUC** (ranking quality). Random baseline: Lift = 1, PR AUC = 0.02. PNPU consistently matches or outperforms PU across all families.

Model	M1		M3	
	Lift	PR AUC	Lift	PR AUC
Bagging LGB	17.80	0.402	21.43	0.591
LGB P vs N	20.95	0.602	17.86	0.885
nnPU LGB	**17.41**	**0.355**	**21.90**	**0.556**
PUET	15.75	0.408	14.76	0.591

Selected model: LightGBM nnPU. Combines high and stable lift, best-in-class calibration (ECE < 0.06 after Platt scaling), conformal coverage $> 90\%$, and SHAP importances consistent with known corruption mechanisms. Decisive advantage: the only tree-based model that explicitly optimizes a PU learning objective.

Top SHAP variables: contract amount (all datasets); bid window, award delay, admitted bids in M1 and M2; cost overrun in M3, coherent with competition-suppression mechanisms in the literature.

Interactive Application

APPalti makes the model accessible to non-technical users. The interface is structured in three sections mirroring the M1/M2/M3 phases, with dynamic fields that appear only when relevant data is available. Output is expressed as a **global percentile rank** relative to the historical ANAC distribution, translated into four qualitative risk classes (low / moderate / high / very high). Additional features include per-contract SHAP explanations, bootstrap-based score stability analysis, and a batch CSV mode for large-scale screening.

Technologies

Languages & tools: R, Python, LightGBM, scikit-learn, PULSNAR, SHAP, tidyverse.
Data: ANAC BDNCP (CIG 2008 to 2025), ISTAT, OpenGA.

Future work: live data ingestion (ANAC, TAR rulings), municipality-level legality indicators, finer prior estimation with larger compute.