Utrecht University

AI Labs

Exploring AI tools to Detect Corruption Risk: Danish Public Procurements

Deliverable 6.2. for the Bridgegap project, submitted to the European Commission, December 2025

Andrea Longobucco

PhD Candidate Applied Economics
School of Economics, Utrecht University

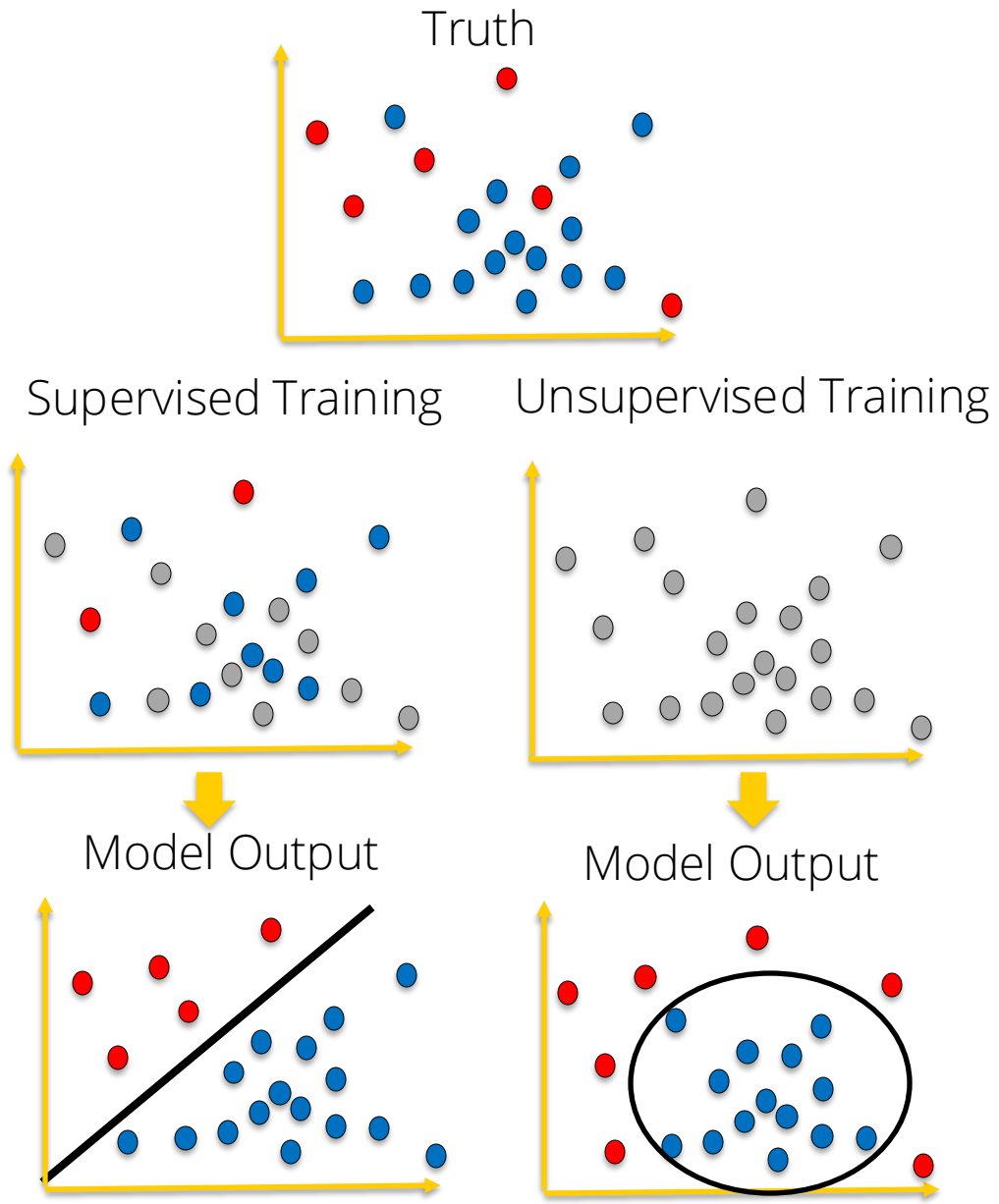
Dr. Joras Ferwerda

Associate Professor Applied Economics
School of Economics, Utrecht University

BRIDGEGAP
BRIDGING THE GAPS IN EVIDENCE, REGULATION AND IMPACT OF ANTICORRUPTION POLICIES

Disclaimer: BridgeGap is funded by the European Commission's Horizon Europe Programme for Research and Innovation, under the Grant Agreement number 101132483. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union, neither can the European Union be held responsible for them.

AI for Financial Crime Detection



Supervised Data Driven Classification

- Requires labelled examples: e.g., *Positive/Negative/ Unknown*
- AI Models are trained in a **supervised** manner

Unsupervised Anomaly Detection

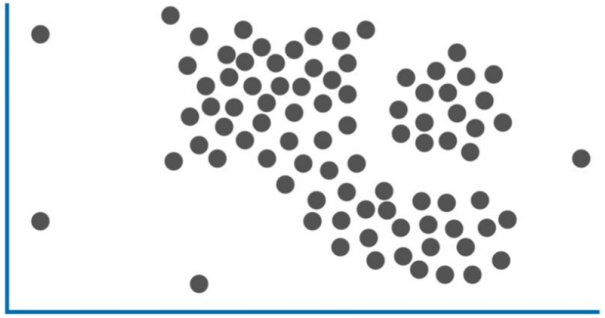
- Works with the assumption that anomalous (outlier) transactions are suspicious enough to warrant an expert opinion
- **Unsupervised** AI Models

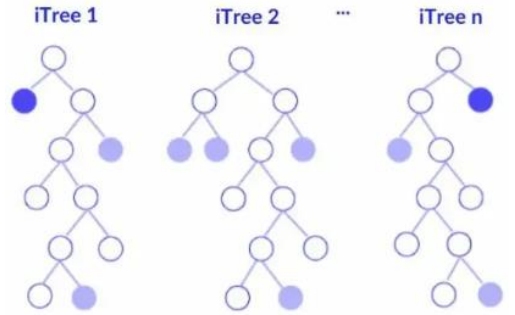
Introduction

- Testing AI Tools for detecting corruption risk with a database of Danish Public Procurements
  - Denmark?
  - Information about procurements supplemented with Beneficial Ownership information
  - 98 variables for 20,877 public procurement contracts, 2016-2022
  - Information about the buyer, the contract, the bidder and their beneficial ownership
  - All procurements are labelled with 5 corruption risk indices: single bidding, short submission period, irregular decision period, non-competitive procurement type, buyer-supplier concentration (Fazekas & Kocsis, 2020)
- Work in progress… These are some first results
- Exploring supervised and unsupervised machine learning models to classify which procurements have a high corruption risk

Method

- Use a bunch of classifier methods
    - 5 unsupervised machine learning methods to determine which procurements are outliers
        - Test: Are the outliers also those classified as high-risk for corruption with CRI?
        - Can help to determine which procurements to check for corruption
    - 6 supervised machine learning methods to determine how ex-ante procurement characteristics can explain high corruption risk (according to CRI)
        - Test: Use a set of observations not used to train the model to test the classification
        - Focus on ex-ante characteristics: can we predict corruption before transferring money?
        - Can help to determine monitoring levels and other mitigating measures





Unsupervised Machine Learning (outlier detection)

1. *Density-based methods (LOF and DBSCAN)* identify observations located in regions of unusually low local density and treat them as potential anomalies relative to their neighbourhood.
2. *Distance-based methods (k-NN distance)* assign higher anomaly scores to observations that lie far away from their k nearest neighbours in the feature space.
3. *Kernel/margin-based methods (one-class SVM)* learn a boundary around the bulk of the data in a high-dimensional feature space and flag points lying outside this boundary as outliers.
4. *Tree-based methods (Isolation Forest)* isolate anomalous observations in randomly grown trees with the idea that outliers are easier to separate from the rest of the data.
5. *Neural Networks, reconstruction-based methods (autoencoder)* train a neural network to compress the input into a low-dimensional latent representation and then reconstruct it. Those with unusually large reconstruction errors are interpreted as anomalous, since their patterns cannot be well explained by the latent structure learned from the majority of the data.

Results Unsupervised Machine Learning

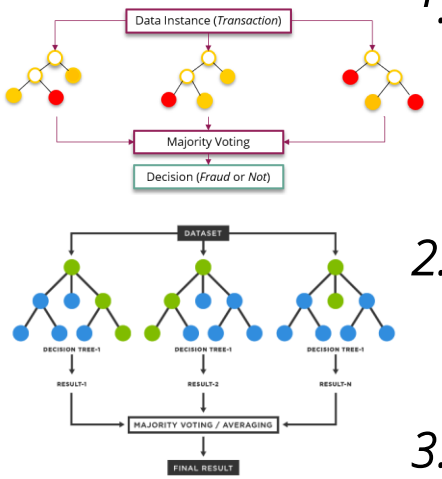
Method	Top percentile	Outlier set size	Single-bid tenders	Outlier single-bid rate	Global outlier single-bid share	Hyperparameters
LOF	1	209	32	0.1531100	0.001532934	minPts=30
DBSCAN (noise)	1	177	29	0.1638418	0.001389222	eps=1.2, minPts=5
One-Class SVM	1	209	57	0.2727273	0.002730539	nu=0.1, gamma=0.1
Autoencoder	1	209	61	0.2918660	0.002922156	hidden=4, epochs=10, l1=0
k-NN dist	1	209	67	0.3205742	0.003209581	k=150
Isolation Forest	1	209	73	0.3492823	0.003497006	ntrees=200, sample_size=256, ndim=1
DBSCAN (noise)	2	414	51	0.1231884	0.002443114	eps=1, minPts=5
LOF	2	418	55	0.1315789	0.002634731	minPts=40
One-Class SVM	2	418	83	0.1985646	0.003976048	nu=0.15, gamma=0.1
Autoencoder	2	418	85	0.2033493	0.004071856	hidden=4, epochs=25, l1=0.0005
k-NN dist	2	418	87	0.2081340	0.004167665	k=200
Isolation Forest	2	418	91	0.2177033	0.004359281	ntrees=200, sample_size=1024, ndim=2
LOF	5	1,044	106	0.1015326	0.005077844	minPts=50
k-NN dist	5	1,044	126	0.1206897	0.006035928	k=10
Isolation Forest	5	1,044	129	0.1235632	0.006179641	ntrees=500, sample_size=512, ndim=1
One-Class SVM	5	1,044	129	0.1235632	0.006179641	nu=0.1, gamma=0.1
DBSCAN (noise)	5	1,125	144	0.1280000	0.006898204	eps=1.2, minPts=30
Autoencoder	5	1,044	150	0.1436782	0.007185629	hidden=4, epochs=25, l1=0

* around 9% of the procurements in the full dataset have single bidding

Results Unsupervised Machine Learning

- Our Isolation Forest model performed roughly 350% better than random auditing when focusing on the top 1% of anomaly scores.
- Focusing on a higher percentage of the anomaly scores (2% and 5%) increases the number of high-risk tenders detected, but lowers the overall detection chance.
- This indicates a cost-benefit trade-off: What to minimize? Monitoring costs or Corruption costs?

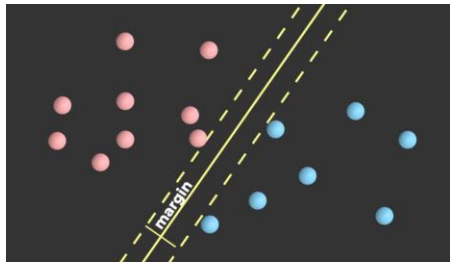
Supervised Machine Learning (train with labels)



1. *Gradient Boosting Machine (GBM)* builds an ensemble of shallow decision trees sequentially, where each new tree is trained to correct the residual errors of the previous ones. This stage-wise learning process enables the model to identify complex nonlinear patterns and interactions between variables, often resulting in high predictive performance.
2. *Distributed Random Forest (DRF)* constructs many decision trees, each of which is trained using a bootstrap sample of the data and a random subset of features. These predictions are then aggregated to reduce variance and overfitting.
3. *Naive Bayes* is a probabilistic classifier that assumes conditional independence between features given the class label. Despite this strong simplifying assumption, it is fast and easy to implement, and often performs surprisingly well when features carry complementary information and the signal is spread across many variables.

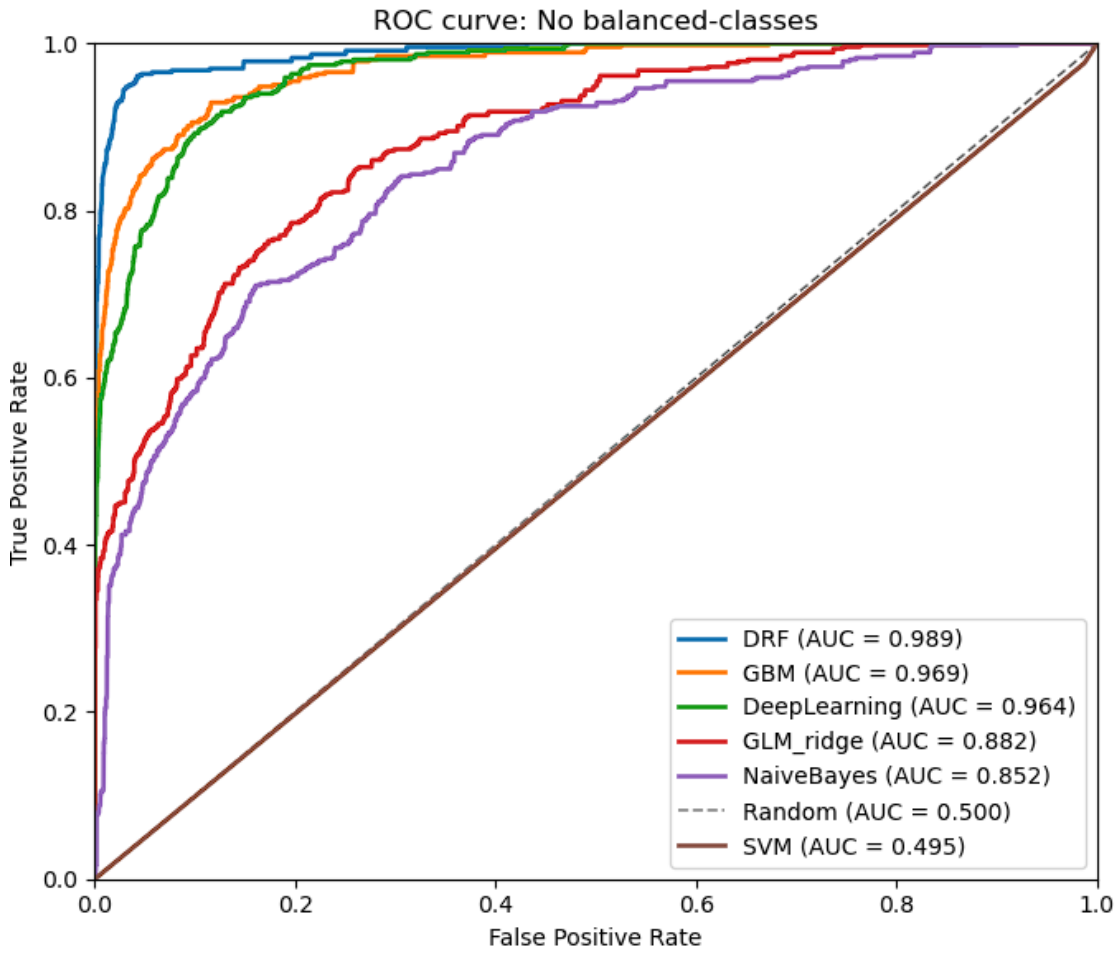
Supervised Machine Learning (classify with labels)

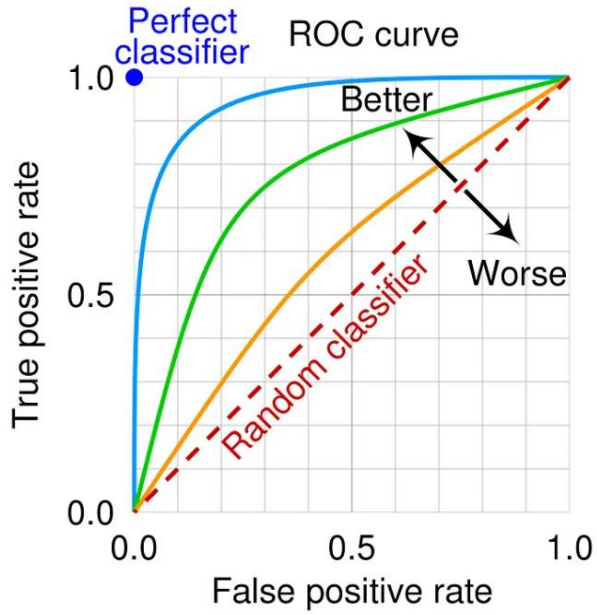
4. *GLM ridge (Regularised Logistic Regression)* uses a logistic regression (binomial Generalized Linear Model) with L2 regularisation (also known as the ridge penalty). This shrinks the coefficients towards zero, stabilising multicollinear estimates and reducing overfitting. (preferred over normal logistic regression because of the many features, which could lead to data sparsity)
5. *Support Vector Machine (SVM, Gaussian Kernel)* searches for a decision boundary that maximises the margin between classes in a transformed feature space. SVM can model nonlinear relationships by implicitly mapping the original predictors into a higher-dimensional space where a linear separator becomes possible, while relying on a relatively compact set of support vectors.



6. *Deep Learning (feed-forward neural network)* with multiple hidden layers learns a hierarchy of nonlinear transformations of the input variables. It can approximate highly complex decision boundaries and capture subtle interactions between features by combining many simple units, although this comes at the expense of reduced interpretability compared to simpler models.

Results Supervised Machine Learning





Results Supervised Machine Learning

- The Distributed Random Forest model (DRF) clearly stands out as the best performer
  - It correctly identifies almost nine out of ten high-risk tenders while keeping the number of false alarms relatively low, achieving high precision and balanced accuracy

- Support Vector Machine performs particularly bad
  - It achieves a balanced accuracy close to random (0.495), an extremely low overall accuracy (0.085) and precision (0.074); however, recall is almost 1 (0.976)
  - In practice, this means that the SVM classifies almost all tenders as high risk

Concluding remarks

- AI tools can help to identify corruption risk, but research is needed
  - Effectiveness, Explainability, Fairness
- Goal should not be to replace human investigation, but to assist human investigation with AI

- We tested the use of beneficial ownership (BO) data by repeating the best classifier (DRF) 1000 times on data without BO data: dataset with BO outperforms without BO 1000 times
  - Preliminary: first tests on data from Denmark and Ukraine (pre-war) show that the importance of BO data is higher in environments where more corruption is expected
  - (paper under review for Regulation & Governance)

- These results are based on labels from other research, not ground truth
  - We are now matching our data with STR-data from the Danish FIU to have alternative labels

Thank you!

Utrecht University

Sharing science,
shaping tomorrow

Dr. Joras Ferwerda
Utrecht University School of Economics
AI & Finance Lab
j.ferwerda@uu.nl

Results Supervised Machine Learning

Model	Balanced Accuracy	Accuracy	Precision	Recall	F1-score
DRF	0.931	0.978	0.838	0.876	0.856
GBM	0.867	0.965	0.776	0.751	0.763
Deep Learning	0.822	0.955	0.713	0.665	0.688
GLM Ridge	0.708	0.94	0.64	0.436	0.519
Naïve Bayes	0.696	0.924	0.49	0.427	0.456
SVM	0.495	0.085	0.074	0.976	0.138

	Actual corruption	Not corruption
Classified as high corruption risk	True positive	False positive (type I error)
Classified as low corruption risk	False negative (type II error)	True negative

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

(Goal = minimize false positives)

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

(Goal = minimize false negatives)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

(How often is the model correct?)

When dealing with heavily imbalanced data—such as 99% negative (e.g., non-fraud) and 1% positive (e.g., fraud)—a model that always predicts "negative" will have 99% accuracy but 0% utility.
Balanced accuracy will correctly report 50% in this scenario, as it treats both classes equally.

F1-score is the harmonic mean of precision and recall (robust measure of performance with imbalanced data)

Context: Unboxing ChatGPT

Artificial Intelligence

Machines that **mimic human behaviour** such as:
- Reasoning
- Learning
- Natural language understanding
- Multi agent systems

Machine Learning

Machines that **learn patterns** from data to make predictions:
- Forecasting
- Fraud Detection
- Risk management
- Data analysis

Deep Learning

Machines that learn **highly complex patterns** based on deep neural networks

Generative AI

Machines that learn how to **create new**, realistic content such as text, images, videos, music, etc

Large Language Models

Machines that learn how to **generate text** based on web scale data and billion parameter models

ChatGPT

A conversational system based on a proprietary (OpenAI) LLM (GPT)