JÖNKÖPING UNIVERSITY

Forecasting Corruption Risk in EU Procurement Data Using Machine Learning

A Predictive Twist on Bauhr et al. (2020)

Fariba Masoudi
Master's in Applied Economics and Data Analysis
Jönköping University

Course: Predictive Analytics
Instructor: Kristofer Månsson
Assignment: Empirical Replication Assignment (ERA) Project
Submission Date: 25 October 2025

Abstract

This project applies machine learning methods to predict corruption risk in public procurement across EU countries, building on the findings of Bauhr et al. (2020). Using a cleaned dataset from their study, I shift the focus from causal inference to prediction, testing whether features like transparency scores, contract types, and economic indicators can help forecast the share of single-bidder contracts which is a common proxy for corruption risk. I implemented four models: OLS, LASSO, Decision Tree, and Random Forest. While Random Forest slightly outperformed the others, overall predictive performance remained low. These results highlight the complexity of modeling corruption risk and the limitations of available data. Despite weak model fit, the project shows how predictive approaches can complement traditional economic research and help inform policy tools like early warning systems or audit targeting.

1. Introduction

Public procurement is one of the biggest spending areas for governments, and it's often seen as a vulnerable point when it comes to corruption. If contracts are awarded without proper competition or with missing information, the chances of favoritism or misuse of public funds can increase. Because of this, transparency has become a key topic in both research and policymaking.

In their paper *"Lights on the Shadows of Public Procurement"* (2020), Bauhr et al. study how transparency in EU countries affects the risk of corruption. They use the share of contracts with only one bidder as a proxy for corruption and find that higher transparency, especially before awarding the contract, is linked to lower corruption risk. Their work is based on a panel dataset covering 32 EU countries from 2006 to 2015, using standard regression methods to explore causal effects.

In this project, I approach the same topic but with a different method. Instead of focusing on whether transparency causes lower corruption, I look at whether we can **predict corruption risk** using transparency scores and other contract-level features. This changes the focus from explanation to prediction, which fits the goals of the Predictive Analytics course.

I use the same dataset as Bauhr et al. and apply four models: OLS, LASSO, Decision Tree, and Random Forest. These models help test how well transparency indicators and other variables can forecast when a contract is likely to be awarded to just one bidder. This could be useful in practice for example, to help flag risky contracts for closer review even if we can't say exactly what caused the risk.

In many policy areas, prediction can still be valuable even without clear causal relationships. For example, being able to flag contracts that are more likely to involve corruption, can help governments prioritize audits or investigations. This doesn't replace deeper analysis, but it can support decision-making by highlighting potential risks early on. Using machine learning for this kind of task is becoming more common as data becomes more available.

The goal is not to prove a new theory but to see how predictive tools can work alongside traditional analysis. Even if the results aren't perfect, they can show us what's possible and what the limits are when using machine learning for public policy problems like corruption detection.

2. Data

The dataset I used for this project comes from the replication files of Bauhr et al. (2020) available in Harvard Dataverse through this link:

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/IEYEZB, which are based on procurement data collected from the Tenders Electronic Daily (TED) platform (the official source for EU public procurement). The original paper worked with contract-level data across 32 European countries, covering the years 2006 to 2015.

For my analysis, I used a cleaned and aggregated version of the dataset, which summarizes procurement outcomes at the country-year level. The file I worked with was the "collapsed" panel, which was available in Stata format. I converted it to CSV so I could use it in Python. After dropping rows with missing values, the dataset included roughly 200,000 observations. These rows contain the key variables I needed for the prediction task.

Target variable:

- **singleb**: This measures the share of contracts awarded to only one bidder. It's used in the original paper and here as a proxy for corruption risk. The variable is continuous and ranges between 0 and 1.

Predictors used:

- exante_transp100: A transparency score based on whether key information was available **before** contract award (like estimated value, duration, etc.).
- expost_transp100: Similar but checks whether this information was provided **after** contract award.
- nocft: Number of contract notices published.
- ca_contract_type31, 32, 33: Dummy variables indicating contract types.
- econ_2gdp_mio_eur: Total contract value as a share of GDP.
- demo_d3dens: Population density.

Before modelling, I scaled the numerical variables and kept the dummy variables for contract type as 0is.

Data Cleaning and Scaling

Before modeling, I dropped any rows that had missing values in either the target variable or the predictors. Then, I standardized the numerical variables using a StandardScaler so they would all be on the same scale (mean = 0, standard deviation = 1). This helps models like LASSO work better.

Overall, this dataset gave me a mix of transparency, economic, and structural features that seemed suitable for trying out the predictive models from class. The fact that it was used in a published paper made it especially useful, since I knew it had already been cleaned and validated.

3. Methodology

In this project, I treated the task as a predictive modelling problem. The main goal was to see how well corruption risk, measured by the share of single-bidder contracts (singleb), could be predicted using the transparency and contract-level features in the dataset. I used a supervised machine learning approach, which means that the model learns patterns from past data and then tries to predict outcomes for new data.

To make sure the models could be properly tested, I split the data into two parts — 80% for training and 20% for testing. The training set was used to fit the models and learn relationships, and the test set was used to evaluate how well they perform on unseen data. This setup helps check how generalizable the models are and avoids overestimating their performance.

Models Used

I tested four different models that represent both linear and non-linear methods. This was important because corruption patterns might not follow simple linear trends, and using different model types helps capture different kinds of relationships.

- **Ordinary Least Squares (OLS):** A simple linear regression model that I used as a baseline to compare with more advanced methods.
- **LASSO Regression:** Similar to OLS, but it adds a penalty that pushes weak variables toward zero. This helps prevent overfitting and highlights which features are most important.

- **Decision Tree Regressor:** A model that splits the data into smaller groups based on variable values. It can detect non-linear patterns and interactions that linear models might miss.
- **Random Forest Regressor:** This model combines many decision trees and averages their predictions. It's generally more stable and accurate because it reduces the overfitting that a single tree might have.

These models were chosen because they cover both traditional econometric methods (OLS, LASSO) and more flexible machine learning methods (Decision Tree, Random Forest). Comparing them helps show whether more complex models can actually improve prediction for this type of data.

Evaluation

I evaluated the models using two common metrics:

- **Root Mean Squared Error (RMSE):** Measures how far the predictions are, on average, from the real values. Smaller RMSE means better predictions.
- **R-squared (R^2):** Shows how much of the variation in the target variable can be explained by the model.

Comparing RMSE and R^2 across the four models helped me see both the size of prediction errors and the amount of variation explained. I expected the Random Forest to perform best since it can capture more complex relationships, while OLS served as a baseline for comparison.

The models were trained using default parameters, except for LASSO, which includes built-in cross-validation to automatically find the best penalty value. I didn't fine-tune hyperparameters because the focus of this project was to understand how these models perform relative to one another rather than achieving the absolute best prediction score.

Overall, this setup made it possible to test both simple and advanced models in a fair and consistent way, showing how different approaches perform when predicting corruption risk using real procurement data.

4. Results

After running all four models, I compared their performance on the test data using RMSE and R-squared (R^2). These results show how well each model did at predicting the share of single-bidder contracts.

The table below summarizes the outcomes:

Model	RMSE	R^2
OLS Regression	~0.280	~0.052
LASSO Regression	~0.280	~0.050
Decision Tree	~0.278	~0.080
Random Forest	~0.275	~0.092

The **Random Forest model** gave the best results overall. It had the lowest RMSE and the highest R^2 but even then, it only explained about **9% of the variation** in the target variable. This means that while the model picked up on some patterns, it still couldn't capture most of what drives corruption risk as measured here.

LASSO and OLS had very similar performance, suggesting that the linear relationships between the predictors and the outcome were weak. LASSO didn't really shrink many coefficients to zero either, which means it didn't find strong signals to filter out.

The **feature importance** scores from the Random Forest gave some insight into which variables mattered most. The top predictors were:

- econ_2gdp_mio_eur (economic size)
- expost_transp100 (post-award transparency)
- demo_d3dens (population density)

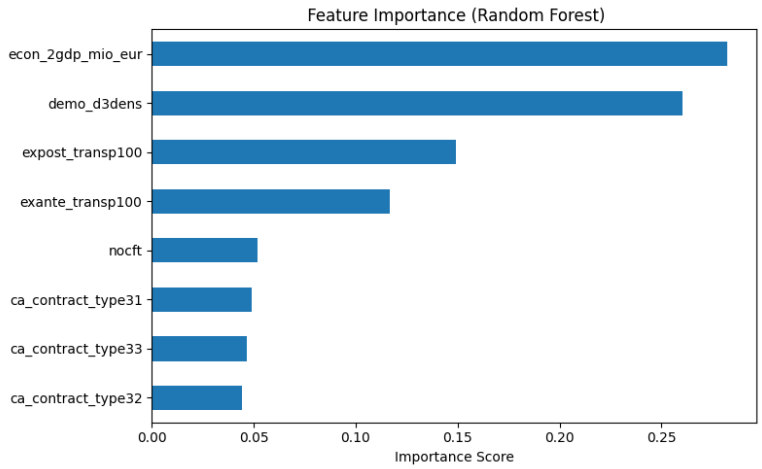

Figure 1. Feature Importance from Random Forest Model.

Economic output (GDP) and population density emerged as the most important predictors of single bidder share.

Although these variables were relatively more important than others, the differences weren't dramatic and none of them had overwhelming influence.

The plot of actual vs. predicted values also showed that predictions were clustered close to the mean and didn't capture much variation. The decision tree model found some small non-linear splits, but overall, the models struggled to predict with high accuracy. In short, all the models worked as expected in terms of code and logic, but their ability to predict the outcome was limited. Even though the predictive results were weak, this comparison is still informative. It shows that even complex models like Random Forest can only do so much when the underlying data doesn't contain strong predictive signals. In this case, corruption risk likely depends on many hidden institutional and behavioral factors that are not directly captured by transparency or contract-level features. This highlights a key challenge in applying machine learning to social and economic problems meaning models can only be as good as the data they are trained on.

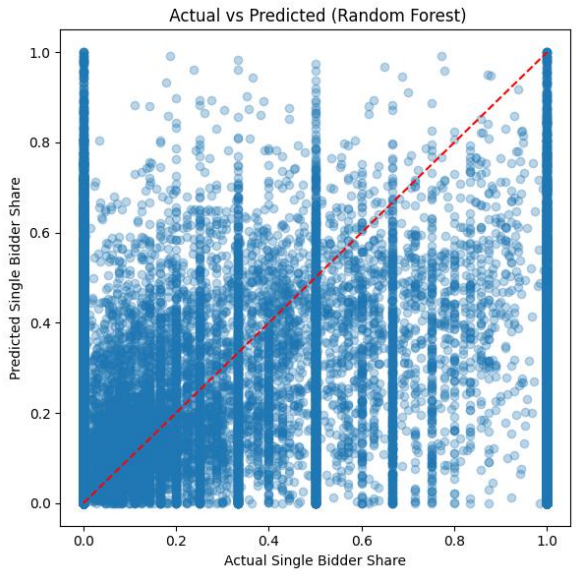

Figure 2. Actual vs. Predicted Single Bidder Share (Random Forest).

The scatter shows poor alignment with the diagonal, indicating that the model struggles to accurately predict single bidder shares.

5. Discussion

The results from this project bring up a few important points about both the data and the challenge of predicting corruption risk.

First, the low R-squared values across all models show that the predictors in the dataset — even though they seem relevant, just don't explain much of the variation in single-bidder share. This isn't necessarily a failure of the models themselves, but it does suggest that there are probably other unobserved factors driving corruption risk that aren't included in the data.

For example, things like political connections, enforcement strength, informal networks, or even regional practices could influence whether a contract goes to just one bidder. These kinds of factors are hard to measure and weren't part of the dataset I used, which only included transparency indicators and a few economic and demographic controls.

Also, it's worth noting that **singleb** is just a proxy for corruption. While having only one bidder might signal corruption in some cases, it could also just reflect market conditions, lack of competition, or technical requirements in certain sectors. So, the outcome itself might be noisy and hard to predict even with good data.

Another interesting finding is that transparency scores didn't stand out as strong predictors in this modelling approach. That doesn't mean transparency doesn't matter, it just means that, in terms of prediction, these scores alone weren't enough to produce accurate forecasts. This is different from the original paper, which used regression to show significant causal relationships. It's a reminder that what works in a causal model doesn't always translate into strong predictive performance. Still, even though the models didn't perform well in terms of prediction, they were helpful for testing which features had more influence and exploring how well machine learning methods apply to real-world governance data.

At the same time, the results highlight how machine learning can still play a role even when predictions are weak. The process of testing different models helps reveal which parts of the data carry the most information and where the gaps are. For policy-related research, this can be valuable, of course not for making precise forecasts, but for improving data collection and identifying which variables might deserve closer attention in future studies. In that sense, the exercise is less about perfect prediction and more about learning what the data can and cannot tell us.

Even though these models followed a standard predictive process, their weak performance can partly be explained by data limitations. The dataset didn't include important political or institutional factors that likely influence corruption, and the singleb variable itself is only a rough proxy. These issues show how difficult it is to capture complex governance dynamics using open procurement data alone.

6. Conclusion

This project examined whether corruption risk in public procurement which is measured by the share of single-bidder contracts, can be predicted using transparency indicators and other contract-level features. I used the same dataset as Bauhr et al. (2020), but instead of applying causal inference methods like in their paper, I focused on prediction by testing four models: OLS, LASSO, Decision Tree, and Random Forest.

The results showed that even the best-performing model, the Random Forest, had low predictive power. None of the models explained more than about 10% of the variation in the outcome. This suggests that while transparency and procurement features are relevant, they are not sufficient on their own to accurately predict corruption risk. There are likely many

unobserved factors, such as political and institutional characteristics, that play an important role but are not captured in the data.

The project also illustrates the difference between **causal inference and prediction**. A variable can be statistically significant in a regression without necessarily being useful for forecasting. Prediction requires consistent and measurable patterns in the data, and in this case, those patterns were weak. Recognizing this gap helps clarify both the potential and the limits of using machine learning in applied policy contexts.

Even with these limitations, the project was still valuable. It provided hands-on experience with applying different machine learning models to real-world policy data and showed how predictive tools can complement traditional methods. In future work, expanding the dataset to include institutional or political indicators might improve predictive performance. But even if the models stay weak, they can still serve as useful early warning tools, not to replace audits or investigations, but to help identify which contracts may deserve closer attention.

Overall, this project reinforced the importance of understanding both the possibilities and the boundaries of data-driven research in economics. While predictive models can't fully capture complex behaviors like corruption, they can still help researchers and policymakers think differently about how to use data for transparency and accountability.

8. References

- Bauhr, M., Czibik, Á., Licht, F., & Fazekas, M. (2020). Lights on the shadows of public procurement: Transparency as an antidote to corruption. *Governance*, *33*(3), 495-523. https://doi.org/10.1111/gove.12432
- Czibik, Agnes, Mihály Fazekas, Monika Bauhr, and Jenny de Fine Licht. 2019. "Replication Data for: Lights on the Shadows of Public Procurement - Transparency as an Antidote to Corruption." Harvard Dataverse. https://doi.org/doi:10.7910/DVN/IEYEZB.