Journal of
*Risk and Financial Management*

MDPI

*Article*

# Comparative Analysis of Ensemble Machine Learning Models for Risk-Oriented Monitoring of Military Procurement

**Tetiana Zatonatska [1], Oleksandr Dluhopolskyi [2,3,*], Oleksandr Artiushenko [1], Isabel Cristina Lopes [4], Anzhela Ignatyuk [1] and Olena Liubkina [1]**

[1] Faculty of Economics, Taras Shevchenko National University of Kyiv, 03-022 Kyiv, Ukraine; tetiana.zatonatska@knu.ua (T.Z.); art.alexander@knu.ua (O.A.); ignatuk@knu.ua (A.I.); liubkina_olena@knu.ua (O.L.)
[2] Faculty of Economics and Management, West Ukrainian National University, 46-027 Ternopil, Ukraine
[3] Institute of Public Administration and Business, WSEI University, 20-209 Lublin, Poland
[4] CEOS.PP, ISCAP, Polytechnic of Porto, 4465-004 São Mamede de Infesta, Portugal; cristinalopes@iscap.ipp.pt
* Correspondence: dlugopolsky77@gmail.com or oleksandr.dluhopolskyi@wsei.pl; Tel.: +380-682-750-430

**Abstract**

This study examines the application of ensemble machine learning methods for identifying and flagging potentially risky transactions in military public procurement in Ukraine, a sector characterized by elevated financial and security sensitivity and limited capacity for comprehensive ex post control. Using an integrated dataset of procurement procedures conducted between 2021 and 2025, enriched with 56 financial, economic, and behavioral indicators of suppliers, the study develops and compares standard logistic and LASSO-penalized regression as econometric benchmarks, Random Forest, XGBoost, XGBoost with SMOTE balancing, and CatBoost classification models. The target variable is defined on the basis of officially detected violations identified through state monitoring. Model performance is evaluated using standard binary classification metrics, with particular emphasis on recall. Model uncertainty and predictive robustness are addressed through partial dependence analysis, temporal stability assessment, and out-of-sample residual diagnostics. The results indicate that the CatBoost model demonstrates the most balanced performance across evaluation measures. Feature importance analysis identifies expected contract value, procurement method, CPV code, and suppliers' financial capacity as significant determinants of procurement-related risk. The findings provide empirical evidence on the usefulness of risk-oriented machine learning tools in supporting earlier detection and monitoring of irregularities in military procurement.

**Keywords:** military procurement; monitoring; AI; machine learning; ensemble models; decision-making; financial control; corruption risks; big data; state audit



Academic Editor: Thanasis Stengos

Received: 11 January 2026
Revised: 13 February 2026
Accepted: 26 February 2026
Published: 28 February 2026



## 1. Introduction

The public procurement sector is one of the most financially demanding and institutionally sensitive components of the public finance management system, as significant amounts of budget funds are redistributed through it every year. In the context of Ukraine's full-scale security challenges, military procurement is becoming increasingly relevant. Unlike civilian procurement, it has a dual purpose (Dluhopolskyi & Chaprak, 2023; Dluhopolskyi & Danyliuk, 2023): on the one hand, to ensure the effective and rational use of public resources, and on the other, to meet critical needs promptly, on which the lives, safety, and health of military personnel and the civilian population directly depend.

The specificity of military procurement lies in its high level of complexity, limited transparency of certain procedures, and significant risks of violations, which complicates the implementation of full control using traditional methods (Zozulya & Matrosov, 2018; Ozsevim, 2023; Chaprak & Dluhopolskyi, 2022; Schwartz, 2004). At the same time, the scale of procurement processes makes it impossible for regulatory authorities to monitor all procedures comprehensively. In this regard, researchers and practitioners are increasingly turning to automated analysis, machine learning, and artificial intelligence tools as an effective mechanism for supporting management decision-making in the field of public procurement. Ensemble models such as Random Forest and gradient boosting (Hancock & Khoshgoftaar, 2020) play a dominant role in contemporary literature on public procurement analysis due to their ability to detect complex nonlinear patterns in large data sets.

A key difference in this paper is the significant expansion of the database, which was made possible by integrating detailed financial, operational, and behavioral indicators of suppliers from the YouControl analytical platform. The use of these indicators, which were not previously considered in procurement analysis models, significantly improves the explanatory power of the models and allows for an assessment of the comprehensive impact of companies' financial condition on the results of procurement procedures. In addition, the paper has formed a logically sound and balanced sample of risky and non-risky military procurements, which ensures more objective training of machine learning models and increases the reliability of the results obtained. The proposed approach creates a basis for building effective control tools and improving the efficiency of military procurement.

## 2. Literature Review

The relevance of using artificial intelligence (AI) and machine learning (ML) in public procurement is growing due to the need to process large amounts of data and improve the efficiency of budget management. This also applies to military procurement procedures, where decisions directly affect the safety, life, and health of citizens. Recent studies show that AI/ML technologies have the potential to transform all stages of the procurement process, from planning to risk assessment and outcome prediction (Balkan & Akyuz, 2025). A review of the literature shows that the use of AI and ML in public procurement is seen as a fundamentally new approach to creating decision support systems (Aboelazm & Dganni, 2025; Sanchez-Graells & O'Connell, 2024). Thus, a comprehensive study (Balkan & Akyuz, 2025) emphasizes that AI/ML can be integrated into virtually all procurement sub-processes, including strategic planning, bid analysis, supplier evaluation, and risk management.

The analysis of the practical application of AI in procurement includes ideas that cost classification, demand forecasting, fraud detection, and procedural transparency can be significantly improved through the implementation of AI algorithms. For example, Ayibam (2025) shows that AI technologies can identify patterns in procurement data, automate the analysis of financial proposals, and assist in the timely detection of anomalies potentially related to violations or fraud (Ayibam, 2025).

Problems relate to data quality and availability, model interpretability, and integration into existing IT infrastructure. Andersson highlights in his study that key barriers to the implementation of AI in public procurement include technological complexities, cultural resistance, lack of technical skills, and organizational factors (Andersson et al., 2025).

Regarding specific AI methods, recent studies show that ensemble models, such as Random Forest and XGBoost, are successfully used to classify and predict risk events in various domains, including supply and logistics. In the context of supply chain security, Wang et al. (2023) demonstrate that XGBoost and other methods can successfully detect anomalies and potential threats when analyzing large data sets (Wang et al., 2023).

Broader research in the field of AI and public procurement also points to hybrid approaches, where AI is combined with multi-criteria decision-making (MCDM) methods and classical statistical models to combine their advantages: interpretability, accuracy, and adaptability (Balkan & Akyuz, 2025). It is important to note that most of these studies focus on global trends in the digitalization of procurement, with only a small proportion devoted specifically to military procurement. However, general conclusions about the effectiveness of ML approaches in identifying risks or anomalies in open data are directly applicable in the context of military contracts, where high-quality forecasts are critical.

In the context of public electronic auctions, increasing attention has been paid to the application of analytical models aimed at detecting irregularities and potentially problematic contracts. In particular, Balaniuk et al. (2013) demonstrated the applicability of a naïve Bayes classifier for evaluating risks arising from cooperation between public authorities and private entities. Similarly, Sales (2013) proposed a risk assessment framework for public contracts, conducting a comparative analysis of logistic regression and decision tree models to evaluate their predictive performance (Balaniuk et al., 2013; Sales, 2013).

Policy-oriented research increasingly emphasizes the transformative potential of artificial intelligence (AI) and machine learning in assessing corruption risks within public procurement systems. Notably, the FALCON (Fight Against Large-scale Corruption and Organized Crime Networks) Horizon initiative demonstrates that AI-driven analytical tools facilitate the processing of large-scale procurement datasets, enabling the detection of complex patterns, correlations, and red flags that traditional rule-based monitoring approaches often fail to identify (FALCON Horizon, 2025).

In the Ukrainian context, public procurement has traditionally been analyzed through the lenses of efficient budget utilization, procedural transparency, and corruption risk mitigation. Since 2019–2020, there has been a marked increase in studies employing big data analytics, economic–mathematical modeling, and machine learning techniques to examine procurement processes.

Ukrainian scholarship has shown increasing engagement with research on the application of artificial intelligence in public procurement, motivated by the need to improve transparency, operational efficiency, and corruption risk identification. In particular, Karpenko et al. (2023) investigate AI as an instrument for mitigating corruption risks in Ukraine's public procurement system. The authors assess regulatory and technological determinants and highlight the advantages of machine learning algorithms in monitoring extensive open datasets and automating routine tasks. The study also underscores the potential of AI-based chatbots to assist procurement participants and foster transparent competition (Karpenko et al., 2023).

Ukrainian research further highlights the rising importance of analytical and data-driven approaches in public decision-making amid wartime conditions. Specifically, studies have explored the evolution of state regulation in Ukraine's energy sector during the Russian–Ukrainian war, employing integrated frameworks that combine regulatory analysis, statistical evidence, and comparative evaluation of EU practices. These analyses stress the necessity for adaptive, evidence-based policy tools capable of functioning effectively under high uncertainty and resource limitations. Although focused on the energy sector, the findings offer transferable insights to other domains of public finance, including defense procurement, where advanced analytical tools and data-driven methodologies can enhance strategic prioritization, efficiency, and transparency in decision processes (Zatonatska & Soboliev, 2024).

Alongside general trends in the application of analytical methods to public procurement and budget management, applied research into defense financing and the optimization of budgetary decisions plays a significant role in the Ukrainian context.

One example of the application of analytical methods in budget management is the study by Sizov et al. (2025), which proposes an analytical approach to assessing the needs of target groups within the financial support system of the Armed Forces of Ukraine. The study employs multi-criteria analysis combined with an expert survey to identify key areas for optimizing budget expenditures and enhancing the efficiency of resource allocation in military structures (Sizov et al., 2025). Another example of analytical decision-making methods applied to military financial management is the work of Artiushenko (2024), which examines the influence of specific factors on defense resource management and financial planning in military contexts. These studies indicate that contemporary Ukrainian research increasingly integrates multi-criteria optimization techniques, expert assessments, and big data analysis to address practical challenges in budget management within the security sector (Artiushenko, 2024).

These contributions provide a valuable applied foundation for the further integration of artificial intelligence in the analysis of military procurement. They delineate the practical context in which such models must operate, including multi-criteria evaluation of alternatives, the requirement for transparency, and the incorporation of expert knowledge in the selection of optimal solutions.

Recent advances in the analysis of public procurement and institutional risk have been driven by the growing availability of administrative data and increased computational capacity. This has led to a broader adoption of data-driven modeling approaches aimed at improving risk identification, monitoring efficiency, and decision support in complex institutional environments. As a result, the choice of methodological framework has become a central issue in the design and interpretation of empirical studies in this field. The methodological distinction between econometric and algorithmic approaches has been widely discussed in the literature. Classical econometric frameworks emphasize model interpretability, parameter identification, and statistical inference under explicit structural assumptions (Wooldridge, 2010). In contrast, statistical learning approaches focus on predictive performance, regularization, and the bias–variance trade-off, particularly in high-dimensional and heterogeneous data environments (Hastie et al., 2009). Recent studies increasingly treat these paradigms as complementary rather than competing, especially in applied institutional risk assessment, where prediction and interpretability serve different analytical purposes.

Overall, the existing scientific literature demonstrates that artificial intelligence and machine learning hold considerable potential for improving transparency, operational efficiency, and risk prediction in public procurement, including mechanisms to counteract corruption and optimize budget expenditures. Nevertheless, there is a clear need for further applied research that tailors these methods to the specific domain of military procurement and rigorously assesses their practical effectiveness using real-world large-scale datasets.

## 3. Materials and Methods

The study adopted a sequential, multi-stage approach to examine public procurement practices in Ukrainian military units from 2016 to 2025. The methodology comprises the stages shown in Figure 1.

The analysis draws on a combination of large-scale administrative datasets, regulatory risk indicators, and firm-level information, among other sources.

1. *Creation of the database and supplementation with data on risky purchases identified through regulatory monitoring*. The initial database was constructed using the BI Prozorro analytical module and included more than 500,000 records of procurement procedures conducted by military units. The database contained information on the subject of the

purchase, participants, expected and realized contract values, winning bidders, contract terms, as well as additional monitoring-related variables.

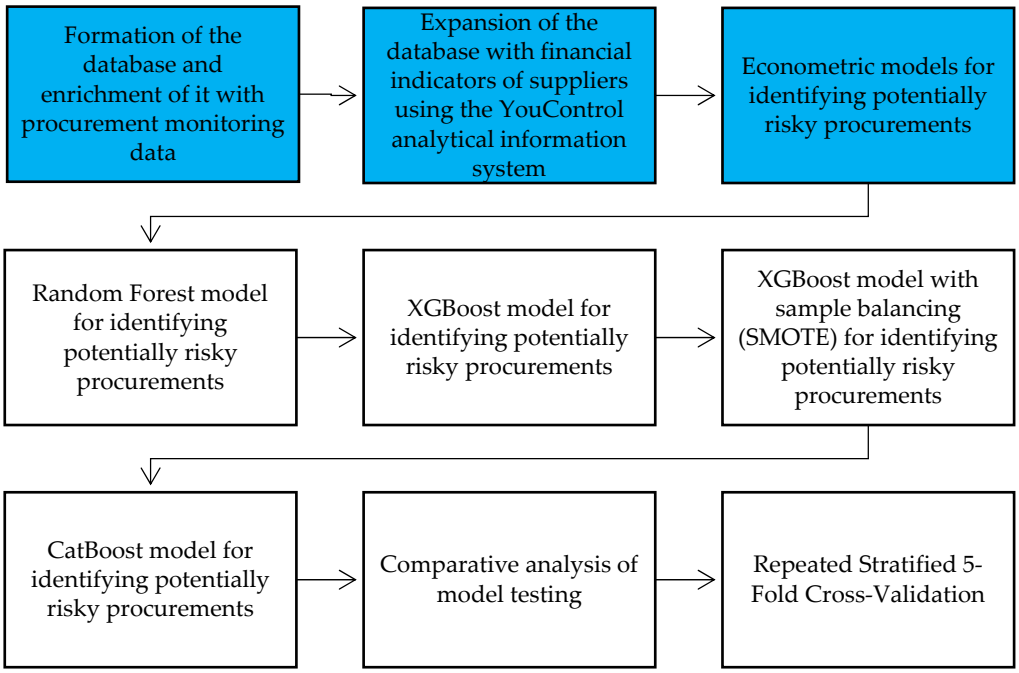


**Figure 1.** Methodological framework for constructing and comparing models for identifying potentially risky procurements.

The next stage involved identifying procurement procedures that had been subject to monitoring by regulatory authorities and in which violations were detected. This process enabled the construction of a dedicated subsample for the development of models aimed at detecting high-risk procurement activities.

2. *Expanding the database through YouControl*. Additional supplier-level characteristics were obtained through the YouControl platform, including financial indicators, ownership structure, litigation history, and tax status, among other variables (YouControl, 2025). This substantially increased the dimensionality of the dataset, which is critical for subsequent machine learning applications. To ensure the structural validity of the predictive framework, a longitudinal data processing approach was adopted. The methodology explicitly addresses potential look-ahead bias by applying an "as-of" feature mapping strategy. For each procurement procedure occurring at time T, all enterprise-level financial predictors sourced from YouControl are lagged to the most recent available reporting period (*T-1*). This ensures that the model operates under realistic informational constraints, such that future financial outcomes cannot influence contemporaneous risk assessments.

The inclusion of a relatively large number of predictors reflects a deliberate methodological choice consistent with contemporary algorithmic modeling practices. The machine learning literature indicates that high dimensionality does not inherently induce overfitting in ensemble-based methods. In particular, Random Forest algorithms are robust to overfitting due to their internal feature subsampling mechanism, whereby stochastic variable selection at each split reduces inter-tree correlation and improves generalization performance (Breiman, 2001).

Furthermore, ensemble algorithms like CatBoost utilize gradient boosting with built-in L2 regularization and shrinkage techniques that effectively perform "automated feature weighting" (Hastie et al., 2009). This allows the model to handle many distinct factors by assigning near-zero importance to redundant or noisy variables. Thus, the model

complexity is controlled not by limiting the number of features, but through algorithmic constraints and out-of-sample validation.

3. *Building econometric models for identifying potentially risky procurements*. Consistent with this dual perspective, the present study integrates econometric and machine learning approaches within a unified evaluation framework. Logistic regression models are treated as econometric benchmarks, valued for their transparency and coefficient interpretability in line with standard econometric practice. At the same time, ensemble machine learning models are employed to capture nonlinear interactions and complex feature dependencies that are difficult to specify parametrically, reflecting the principles of statistical learning outlined in the statistical learning literature. This combined strategy allows the analysis to balance interpretability and predictive accuracy, mitigating the limitations of relying exclusively on either modeling paradigm (Hastie et al., 2009).

To evaluate the incremental value of ensemble learning, we implemented two baseline econometric models:

- Binary Logistic Regression: Serves as the fundamental parametric benchmark to estimate the probability of risk based on the linear combination of predictors. It provides transparency in understanding the signs and significance of individual risk factors.
- LASSO *(Least Absolute Shrinkage and Selection Operator)*: A penalized regression approach used for high-dimensional data.

All econometric models were estimated using the same data preprocessing procedure. Missing values were imputed using median values, and all features were standardized before model estimation. Model performance was evaluated using a time-based validation scheme, identical to that applied to machine learning models. The models were trained on data from 2021–2024 and tested on an independent hold-out sample from 2025.

To assess the temporal robustness of the econometric specifications, an additional coefficient stability analysis was conducted. Specifically, the dataset was divided into two consecutive subsamples corresponding to 2023 and 2024, and both the standard logistic regression and the LASSO-penalized model were re-estimated separately for each period. This procedure allows verification that the identified relationships are not driven by year-specific effects and provides an additional robustness check beyond out-of-sample predictive performance.

4. *Building models for identifying and labelling risky purchases using Random Forest*. In the fourth stage of the study, the Random Forest algorithm, which belongs to the class of ensemble machine learning methods based on decision trees, was used to identify and label risky purchases.

The Random Forest algorithm constructs an ensemble of individual decision trees, each of which is trained on a random subsample of training data with bootstrap sampling and also uses a random subset of features at each tree node. The final decision of the model is determined by aggregating the results of all trees (majority voting method for classification tasks). A description of the algorithm and its probabilistic interpretation can be found in works on machine learning (Breiman, 2001; Hastie et al., 2009; Géron, 2022).

Following Breiman's distinction between the "data modeling" and "algorithmic modeling" cultures, this study does not treat these approaches as mutually exclusive. Econometric models provide a structured and interpretable framework grounded in explicit assumptions about the data-generating process, making them particularly valuable for inference, explanation, and policy interpretation. In this study, logistic and LASSO-penalized regression models are therefore employed as transparent benchmarks that establish a reference level of predictive performance (Breiman, 2001).

At the same time, the algorithmic modeling approach is adopted to complement, rather than replace the econometric analysis. Ensemble machine learning models are used to explore complex nonlinear interactions and heterogeneous patterns that are difficult to capture within standard parametric specifications. The combined use of econometric benchmarks and ensemble models enables us to distinguish genuine nonlinear performance gains from algorithmic complexity and to provide a more balanced assessment of procurement risk. In this sense, the paper contributes to the ongoing econometric–algorithmic debate by demonstrating how both modeling cultures can be jointly applied in a risk-oriented monitoring context.

In this paper, we adopt the algorithmic modeling culture, as defined by Leo Breiman (Breiman, 2001), which stands in contrast to the traditional data modeling (econometric) approach. While the econometric approach assumes a specific stochastic data-generating process (typically linear or logistic regression with a predefined error distribution), the algorithmic approach "treats the data mechanism as unknown". According to this structure, the relationship between the explanatory variables and defence procurement risk is expressed in a general functional form, without assuming a specific parametric or probabilistic data-generating process. This approach is especially appropriate for military procurement data, where relationships are typically non-linear, and the multidimensional distribution of influencing factors is too complex to be adequately captured by standard parametric econometric models. Following the statistical learning framework (Hastie et al., 2009), we represent the relationship (1), where $f(X)$ is an unknown function approximated by the ensemble, and $\epsilon$ reflects the irreducible error without making restrictive parametric assumptions about its distribution:

$$y = f(X) + \epsilon \tag{1}$$

where $y$ is the target variable (procurement risk); $X = (x_1, x_2, \ldots, x_n)$ is the vector of input features (factors); $f(X)$ is the systematic component (structural part) to be estimated using ensemble methods; $\epsilon$ is the random error term (noise), which accounts for unobserved factors and the stochastic nature of the process.

The feature vector $X$ included only numerical indicators characterizing: (1) purchase parameters (expected value, number of bids submitted, number of disqualifications, etc.); (2) financial and economic characteristics of suppliers (revenue, revenue growth rates, number of employees, authorized capital); (3) aggregate indicators of economic activity and participation in public procurement.

To ensure that the Random Forest model is suitable for risk-oriented monitoring of public procurement rather than abstract classification, specific design choices were deliberately made at the data preparation, validation, and hyperparameter levels.

All financial level indicators were constructed using *an* as-of (last available) logic. For each procurement procedure occurring in year $T$, only financial information from year *T-1* or earlier was used. This prevents look-ahead bias and reflects the real informational environment of audit authorities, who must assess procurement risks based solely on data available at the moment of decision-making.

A strict time-aware train–test split was applied. The model was trained on procurements from 2021–2024 and evaluated exclusively on data from 2025. Unlike random cross-sectional splitting, this back-testing strategy simulates real-world deployment, where models are required to predict future risks using historical data only. The hyperparameters of all machine learning models were fixed in advance, based on established empirical practice and results from previous experiments, rather than being tuned within each training window via automated search procedures. This approach was chosen to maintain

consistency with the temporal validation structure and to prevent any potential leakage of information from future data.

The Random Forest hyperparameters were chosen to balance model flexibility and generalization. The number of *trees (*n_estimators = 300*) ensures stable ensemble aggregation and reduces variance, which is critical in procurement data. The maximum tree depth (*max_depth = 10*) constrains model complexity*, preventing excessive memorization of rare procurement configurations while still allowing the capture of nonlinear interactions between contract size, competition indicators, and supplier financial characteristics. Such interactions are central to procurement risk but are typically missed by linear or parametric models.

5. *Formation of models for identifying and labelling risky purchases XGBoost*. Before introducing the XGBoost model, the methodological sequence of models should be briefly clarified. Random Forest is first employed as a robust, low-assumption baseline ensemble that provides stable performance under noisy procurement data. XGBoost is then introduced as a more flexible boosting-based extension, allowing us to test whether sequential error-correction improves the detection of complex risk patterns beyond independent tree aggregation. This approach is a development of ensemble methods and is widely used in tabular data classification tasks due to its high accuracy, efficiency, and ability to work with a large number of features (Hastie et al., 2009; Géron, 2022). The motivation for introducing XGBoost lies in its ability to model complex nonlinear interactions through sequential error correction, which is particularly relevant for procurement risk assessment, where violations often arise from specific combinations of contract size, competition structure, and supplier financial characteristics rather than from isolated factors. At the $t$-th step of boosting, the model looks like this:

$$\hat{y}_t(x) = \Sigma_{k=1}^{t} f_k(x) \quad (2)$$

Unlike Random Forest, where trees are trained independently, XGBoost builds trees sequentially, with each new tree correcting the errors of the previous ensemble by minimizing a predefined loss function. This situation can lead to a bias in machine learning models towards the dominant class and a reduction in the completeness of the detection of risky procurements.

XGBoost was implemented without any class rebalancing techniques to evaluate its intrinsic ability to handle moderately imbalanced procurement data. This design choice serves two purposes. It allows for a direct comparison with the Random Forest baseline under identical class distributions. It provides a benchmark for assessing whether boosting alone is sufficient to improve sensitivity to violations, or whether additional imbalance-handling mechanisms are required.

The model hyperparameters were selected with the following parameters. A relatively large number of trees (n_estimators = 300) ensures stable convergence of the boosting process. Tree depth was constrained (max_depth = 10) to prevent excessive specialization on rare procurement configurations, while a reduced learning rate (learning_rate = 0.05) was used to allow gradual refinement of the ensemble. Subsampling of observations (subsample = 0.8) and features (colsample_bytree = 0.8) introduces stochasticity into the learning process, mitigating overfitting and improving robustness in the presence of correlated procurement indicators.

The reported hyperparameter values represent an initial baseline configuration. During model development, these parameters were systematically varied within empirically reasonable ranges, and alternative configurations were evaluated using time-aware out-of-sample testing. A reduced learning *rate (*learning_rate = 0.05*) served as a starting point for gradual model refinement. Subsampling of observations (*subsample = 0.8*) and features (*colsample_bytree = 0.8*) was applied to reduce overfitting.*

6. *Formation of models for identifying and marking risky procurements XGBoost (SMOTE)*. The SMOTE method consists of artificially increasing the number of objects in the minority class by generating synthetic observations based on existing data. The theoretical foundations of SMOTE and its impact on classification quality are described in (He & Garcia, 2009), as well as in textbooks on applied machine learning (Hastie et al., 2009; Géron, 2022).

XGBoost model, with the SMOTE, is particularly relevant for military procurement analysis, where detected violations constitute a minority class and the institutional cost of missing a risky procurement is substantially higher than that of generating additional false alerts. While XGBoost alone is capable of handling moderate class imbalance, this stage was designed to explicitly test whether targeted rebalancing improves sensitivity to violations in a defense procurement context. Parameter values were selected based on established empirical practice in applied machine learning and prior studies on imbalanced classification problems in the public sector and financial risk analysis (Hastie et al., 2009; Géron, 2022). During model development, these parameters were systematically varied within reasonable ranges, and alternative configurations were evaluated using time-aware out-of-sample testing. *The SMOTE procedure was applied only to the training data using a fixed random seed (*random_state = 42*) to ensure reproducibility. Given the presence of synthetic observations, the depth of individual trees (*max_depth = 6*) is* compared to the baseline XGBoost model. This constraint limits the model's capacity to over-specialize on artificial patterns introduced during oversampling. The remaining hyperparameters were treated as initial reference values and were subsequently varied within reasonable ranges during experimentation.

7. *Formation of models for identifying and marking risky purchases, CatBoost*. To build models for identifying risky procurements, the study used the CatBoost algorithm, which belongs to the Gradient Boosted Decision Trees (GBDT) family and is designed to work with heterogeneous tabular data containing numerical and categorical features. The feasibility of using CatBoost in the classification of large and heterogeneous data sets is justified in recent review studies on machine learning The use of CatBoost in this study is due to its ability to work effectively with heterogeneous data, automatically process categorical features, and demonstrate stable results in classification tasks on large datasets, as confirmed by empirical results in various application areas, including finance, cybersecurity, and anomaly detection (Hancock & Khoshgoftaar, 2020; Hastie et al., 2009).

Using gradient boosting algorithms, CatBoost builds a prediction in the form of an additive model, which is the sum of individual base algorithms (decision trees):

$$\hat{y}_t(x) = f_1(x) + f_2(x) + \ \ldots \ + f_t(x) \tag{3}$$

where $f_k(x)$—$k$-th decision tree forecast, $t$—number of boosting iterations, $\hat{y}_t(x)$—aggregated ensemble forecast for an object $x$.

Unlike classical implementations of gradient boosting, modern approaches to working with categorical variables consider the risk of target leakage and the associated prediction shift that arise when using statistics calculated using the target variable for the same object. To reduce this effect, ordered or leave-one-out schemes for calculating statistics are used, in which only data available without considering this observation are used for each observation (Hancock & Khoshgoftaar, 2020; Micci-Barreca, 2001).

The CatBoost model configuration was defined as an initial setup and subsequently refined through experimental evaluation. The number of boosting iterations (iterations = 400) and tree depth (depth = 6) were selected to balance model expressiveness and generalization in the presence of heterogeneous procurement data. A key configuration element was the use of class weighting to explicitly reflect asymmetric risk preferences in procurement monitoring.

To further examine the sensitivity of the model to different risk preferences, an additional methodological step was introduced by varying the class weights assigned to the target variable. In particular, the weight of the violation class (y = 1) was systematically increased relative to the non-violation class to analyze how this adjustment affects the balance between false positive and false negative errors. This approach allows the model to explicitly reflect different levels of risk tolerance by penalizing missed violations more heavily than incorrectly flagged non-violating procurements. Such an analysis is especially relevant in the context of public procurement monitoring, where the relative costs of these two types of classification errors are inherently asymmetric.

The impact of alternative class-weight configurations was assessed by comparing the resulting number of flagged procurements, true positive detections, false positives, and missed violations across model runs. Importantly, this methodological design does not treat the model output as a final audit decision but rather as an analytical support tool. The ultimate interpretation of risk signals and the selection of acceptable operating points remain the responsibility of control authorities, who may adjust class weights in accordance with available resources, capacity, and priorities, thereby ensuring flexibility in practical implementation.

To enhance model interpretability, partial dependence analysis was applied to all key predictors in the final model specification. For numerical variables, standard partial dependence plots were constructed using the model-based partial dependence framework. This approach estimates the average marginal effect of a given feature on the predicted probability of a violation while holding other variables constant.

For categorical variables, partial dependence was approximated at the category level by computing the average predicted risk within each category. To ensure statistical reliability and reduce noise, the analysis focused on the most frequent categories within each variable.

8. *Analysis of model testing results*. At the final stage of the study, the results of testing the constructed machine learning models were analyzed to comparatively assess their ability to identify potentially risky purchases. The analysis was performed on a test sample that was not used during the model training stage. The quality of classification was assessed using standard metrics for binary classification tasks, in particular, precision, recall, F1-score, and ROC-AUC. Particular attention was paid to the recall indicator for the "violation" class ($y = 1$), as it was assumed that the false non-detection of a risky purchase has more significant consequences than the false classification of a non-risky purchase as potentially risky.

For each model, the following were analysed: confusion matrices; precision, recall, and F1-score for both classes.

In addition, for the ensemble models, feature importance analysis was conducted to identify the factors exerting the strongest influence on the predicted probability of detecting procurement violations. This enhances the interpretability of the findings and supports the practical deployment of the models within risk-oriented monitoring systems.

The performance results of the individual models were compared to identify the approach offering the most balanced performance with respect to sensitivity, forecast stability, and feasibility of real-world implementation.

To assess the reliability of the ensemble models and account for possible stochastic dependencies, a diagnostic framework was applied. In contrast to conventional parametric models that rest on prior assumptions about error distributions, algorithmic modelling evaluates stability via empirical inspection of classification residuals, defined here as the difference between the observed outcome and the predicted probability of a breach.

The diagnostic process addresses three principal dimensions.

Bias Assessment: The mean residual is computed on the out-of-sample test set (2025 data) to verify the absence of systematic over- or underestimation of risk.

Temporal Stability: The structure of errors is examined over the annual cycle. Annual resolution is adopted for temporal analysis, since the principal financial indicators and enterprise-level data from the YouControl system follow a yearly reporting structure. The analysis, therefore, retains annual granularity while confirming robustness across fiscal years.

Quantile Analysis: The residual distribution is inspected to assess error concentration and to confirm that the stochastic component is contained within acceptable limits.

By prioritizing out-of-sample residual stability over conventional goodness-of-fit measures, the methodology emphasizes predictive consistency on previously unseen future data rather than dependence on theoretical assumptions about the noise process (Hastie et al., 2009).

9. *To ensure the reliability and predictive robustness of the results, the study employed a dual validation strategy*. The primary evaluation follows a time-aware back-testing (out-of-time) approach, in which the dataset is split chronologically: models are trained on historical data from the 2021–2024 period and tested on an independent hold-out set comprising procurements from 2025. This design replicates a realistic operational setting in which the algorithm must forecast future risks using only information available up to the present, thereby eliminating look-ahead bias.

To further ensure model stability, year-stratified cross-validation was applied to the training corpus (2021–2024). Rather than random shuffling, folds were constructed by preserving entire annual periods. This procedure validates the model across distinct yearly cycles and procurement conditions within the training window. By evaluating performance on each year while training on the remaining years, the approach confirms that learned patterns are not artifacts of year-specific noise (Hastie et al., 2009).

To assess the discriminatory power of the models, receiver operating characteristic (ROC) curves were constructed. The area under the ROC curve (ROC-AUC) serves as a summary measure of model quality, representing the probability that the model correctly ranks a randomly chosen positive instance (“1”, i.e., violation) higher than a randomly chosen negative instance (“0”). ROC-AUC values near 0.5 indicate performance no better than random guessing, whereas values approaching 1 reflect strong classification ability (Fawcett, 2006).

## 4. Results

The initial database was created using the BI Prozorro analytical module. The data was downloaded step by step from separate sections of the module: first from the ‘Procurement’ sheet, then from “Suppliers” and “Contracts”. Each of these files was a large Excel spreadsheet with detailed information about procurement procedures, participants, contracts, costs, deadlines, etc. After the initial collection, all tables were integrated into a single aggregated database using Excel functions. This made it possible to consolidate disparate tables into a single structure with unified variables, convenient for further analysis in Python 3.12 and machine learning. The stepwise process of filtering and constructing the final analytical sample of military procurements is illustrated in Figure 2.

Step 1. The initial database was constructed using the BI Prozorro analytical module and included 2 million+ procurement records conducted by budget spending units during the 2016–2025 period.

Step 2. Filter by main spending authority (Ministry of Defense). Given that the study focuses exclusively on military procurement, the first thematic filter was applied to exclude procurements conducted by non-defense entities, such as local governments,

educational institutions, municipal councils, and other civilian organizations. Only procurements associated with the Ministry of Defence as the main budget spending authority were retained.

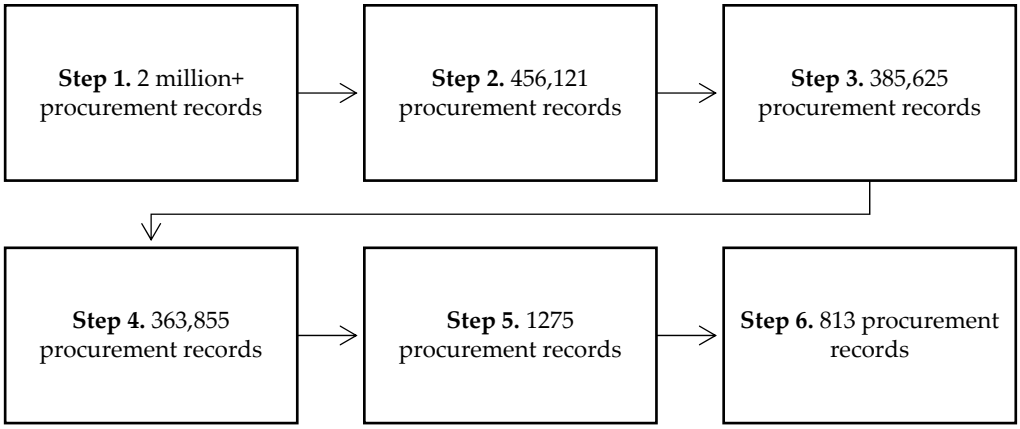


**Figure 2.** Stepwise formation of the analytical sample of military procurements.

Step 3. At the next stage, procurements were further restricted to cases where the contracting authority was a military unit, excluding transactions conducted by other subordinate institutions or affiliated organizations that are not directly involved in military procurement activities.

Step 4. Successful procurement procedures. Only successfully completed procurement procedures were included in the analytical sample. Cancelled, unsuccessful, or terminated tenders were excluded to ensure consistency in outcome-based analysis.

Step 5. To identify potentially risky procurements, the study used the "Monitoring" tab of the BI Prozorro module, which contains information on tenders audited by the State Audit Service of Ukraine and other authorized control bodies. Both proactive monitoring (initiated automatically or manually) and complaint-based audits were considered. The selection criterion was the official detection of violations based on monitoring results. This step resulted in the identification of 1275 procurements with recorded legal violations, which formed the core subsample for building risk detection models.

Step 6. At the final stage, the sample was restricted to the 2021–2025 period, as supplier-level financial and operational indicators from the YouControl platform—used to enrich the dataset—are available only for this timeframe. After applying this temporal filter and ensuring time-consistent data alignment, the final modeling sample consisted of 813 procurement procedures.

It should be noted that the total number of procurements conducted through the electronic procurement system demonstrates a general upward trend over the 2016–2025 period. The only exception is 2022, corresponding to the onset of the full-scale military invasion of Ukraine, when procurement activity declined sharply. An in-depth analysis of this disruption lies beyond the scope of the present study.

As a result, the database was supplemented with some important variables, including: monitoring identifier; date of commencement of monitoring by the control authority; presence of violations identified during monitoring; date of tender announcement; CPV (common procurement vocabulary); number of unique participants in the lot; number of disqualifications per lot; number of questions asked about the lot; name of the winner; tax ID code of the winner.

These parameters enabled a qualitative characterization of risky purchases and served as the foundation for constructing the target variable used in subsequent machine learning model training.

For a more thorough examination of procurements and the development of comprehensive supplier profiles, extended data from the YouControl analytical platform was incorporated into the consolidated database. This integration not only provided greater detail on companies' participation in public tenders but also facilitated the inclusion of financial and operational indicators in machine learning models aimed at identifying potentially risky procurements or those offering opportunities for budget savings.

More than 50 variables were added to the database, encompassing the following categories:

1. General company characteristics: KVED code (Ukrainian Classification of Economic Activities, aligned with EU NACE Rev. 2), authorized capital, number of employees, reporting form type (annual), presence of sanctions, and owned transport assets.
2. Financial ratings: MarketScore, FinScore, and express analysis indicators.
3. Tender activity: number of bids submitted and number of tender wins.
4. Interactions with the public sector: number of transactions and total volume of public funds received over 2021–2025.
5. Financial statements: revenue, total income, assets, cash and cash equivalents, and net profit for the periods 2021–2025 (reported annually and quarterly where available).
6. Growth dynamics: year-on-year relative revenue growth (in percentage terms).
7. Foreign economic activity: total exports and imports, with goods classified according to UKTZED (Ukrainian Classification of Goods for Foreign Economic Activity).

The inclusion of these data elements permits a multifaceted assessment of suppliers, encompassing both their current financial position and historical growth trajectories—factors essential for detecting patterns associated with corruption risks or inefficiencies in procurement processes.

To accommodate the temporal structure of the data and mitigate the risk of spurious correlations, a strict "as-of" feature engineering protocol was implemented, combined with time-aware validation. Key financial indicators (e.g., revenue, assets, net profit) were converted to lagged variables (*T-1*), ensuring that, for any procurement in period T, only information available before or at T was utilized. This procedure eliminates look-ahead bias, in which future data could otherwise leak into historical predictions (Hastie et al., 2009).

Following the creation of the _last_available data block, the original unlagged variables (_2021, *_2022, *_2023, *_2024, *_2025) were removed to maintain feature consistency across training and test sets.

It should be noted that financial indicators for counterparties, sourced from the YouControl system, were available only for the period 2021–2025. Consequently, the application of the time-aware approach and the derivation of features based on the last available reporting period necessitated the exclusion of procurements for which matching financial data were either unavailable or could not be temporally aligned with the tender year. After reconciling the temporal structure and eliminating potential information leakage, the final analytical sample comprised 813 tenders. This reduction preserved methodological integrity and ensured the validity of out-of-time testing.

Importantly, the final sample of 813 procurements is not intended to represent the full population of military procurement procedures statistically. Rather, it comprises a purposively selected set of information-rich cases for which reliable, temporally consistent financial, procedural, and monitoring data are available. Consequently, these procurements are likely to differ systematically from the broader population in terms of contract value, procedural complexity, and extent of regulatory scrutiny, and they disproportionately include procedures that have undergone monitoring by the State Audit Service of Ukraine. The proposed model is therefore intended for risk-based prioritization and early identification of anomalies within high-risk segments of military procurement, rather than for

comprehensive screening of all procurement activities. Results and inferences should be interpreted accordingly, within the defined analytical scope of this study.

This sampling strategy reflects the methodological constraints and objectives of the present research focus. By training the model on cases characterized by high audit visibility, the learning process draws on transactions whose risk profiles have already been substantiated through regulatory oversight. Such an approach enables the model to identify structurally meaningful configurations of risk indicators. The resulting system is thus positioned as a specialized, risk-oriented screening instrument targeted at high-risk segments of military procurement, rather than as a general-purpose filter applicable across all categories of public procurement.

At the first stage of the empirical analysis, econometric binary classification models—namely, standard logistic regression and LASSO-penalized logistic regression—were estimated to establish a baseline for subsequent comparison. The econometric models were implemented in Python using the scikit-learn library. The models were trained and evaluated on a sample of 813 observations, each described by 21 quantitative features capturing procurement parameters and the financial and economic characteristics of suppliers.

The target variable was binary: it took the value 1 if regulatory monitoring identified violations and 0 otherwise.

To ensure practical relevance and predictive robustness, the conventional random 80/20 stratified split was replaced with a time-aware back-testing validation strategy. Random splitting in longitudinal data can introduce data leakage, whereby future information inadvertently influences predictions on past observations. Accordingly, the training set comprised procurement procedures conducted during 2021–2024, while the test set consisted of all procedures from 2025.

This chronological partitioning enables evaluation of the model's ability to predict risks in previously unseen future transactions using only historically available information. The class distribution within the training sample remained reasonably balanced (approximately 59.3% risky and 40.7% non-risky procurements), and the 2025 test set served as an independent hold-out sample to assess the performance stability of the ensemble approach under realistic monitoring conditions.

A standard logistic regression model was estimated as a baseline econometric approach for identifying potentially risky procurements. When evaluated on the independent 2025 test sample, the model demonstrated acceptable average discriminatory performance, with a ROC-AUC of approximately 0.69 and an overall classification accuracy of around 0.65. For procurements with detected violations, the model achieved a precision of about 0.74, a recall of approximately 0.58, and an F1-score close to 0.65. In contrast, recall for the non-violation class was higher (around 0.74), indicating an uneven classification performance across classes.

On the next step, the LASSO-penalized logistic regression model was additionally estimated. The results of the regularized specification were largely consistent with those of the standard logistic regression. The ROC-AUC remained at approximately 0.69, while overall accuracy was again close to 0.65. For violation cases, precision reached about 0.73, recall was approximately 0.57, and the corresponding F1-score was around 0.64. These results suggest that while LASSO regularization improves model parsimony and stability, it does not substantially enhance the ability to detect risky procurements.

Overall, the econometric models provide a transparent and interpretable baseline with acceptable average discrimination. However, the relatively moderate recall values for violation cases (around 0.57–0.58) highlight the limitations of linear parametric models in risk-oriented procurement monitoring. To address these limitations, the next stage of

the analysis applies ensemble machine learning models, with a comparative evaluation presented in Sections 5 and 6.

Coefficient stability analysis for econometric benchmarks was implemented using Python and the scikit-learn library. Standard logistic regression and LASSO-penalized logistic regression were estimated separately for yearly subsamples (2023 and 2024). Data preprocessing was performed within a unified Pipeline framework.

The results of the coefficient stability analysis for the standard logistic regression and the LASSO-logistic regression indicate that a subset of variables exhibits a high degree of intertemporal consistency in the estimated effects. In particular, in the LASSO specification, for instance, the number of employees retains a positive sign in both 2023 and 2024 (0.20 and 0.26, respectively), with a minimal absolute difference in coefficients ($\approx$0.06). Similarly, charter capital preserves a positive direction of influence (0.13 and 0.07; difference $\approx$ 0.06), while the cost of goods sold shows a stable negative sign ($-$0.10 and $-$0.12; difference $\approx$ 0.02). In the standard logistic regression, the smallest interannual deviations are observed for the number of clarification requests submitted during the tender process ($-$0.038 and $-$0.002; difference $\approx$ 0.04) and the number of disqualifications ($-$0.156 and $-$0.420; difference $\approx$ 0.26), indicating relative stability in the direction of these effects over time.

At the same time, a substantial share of financial and behavioral variables displays pronounced interannual variation in coefficient magnitudes and, in some cases, sign changes. The largest absolute deviations in the standard logistic regression are recorded for the number of winning tenders (0.60 in 2023 and $-$0.05 in 2024; difference $\approx$ 0.65), the supplier's gross profit ($-$0.46 and 0.19; difference $\approx$ 0.65), and the number of vehicles owned ($-$0.34 and $-$0.94; difference $\approx$ 0.60). In the LASSO specification, the most pronounced temporal instability is observed for the number of vehicles owned (0.00 and $-$0.77; difference $\approx$ 0.77), total supplier revenue ($-$0.42 and 0.00; difference $\approx$ 0.42), and the number of bidders ($-$0.09 and 0.24; difference $\approx$ 0.32). Such coefficient behavior is characteristic of wartime environments and highly volatile defense procurement systems, where linear effects cannot be assumed to be time-invariant.

The coexistence of directionally stable effects with pronounced context dependence and temporal nonlinearity provides a strong rationale for moving beyond purely parametric specifications. In this setting, ensemble machine-learning methods are better suited to capture the complex and time-varying structure of procurement risk that cannot be adequately represented by static linear models. It is precisely the presence of directionally stable, yet context-dependent and nonlinear effects that justifies the further use of ensemble machine learning methods capable of more adequately reflecting the complex dynamics of risks over time.

At the next stage of verification, the Random Forest model was employed to identify potentially risky procurements. The Random Forest model was trained using 300 decision trees. A maximum tree depth of 10 was imposed to limit overfitting and enhance generalization performance. A fixed random state was specified to ensure full reproducibility of the experimental results.

To avoid confusion between results obtained under different evaluation settings, the presentation of empirical findings is structured according to the purpose of each performance assessment. Below, we briefly clarify the role of each table, after which the results are analysed and discussed in detail.

Tables 1–4 report detailed performance metrics for each model, including class-specific classification measures. Table 5 presents aggregated performance indicators for all four models used in the study and serves as the primary basis for their comparative evalua-

tion. Tables 1–6 report results obtained from the out-of-time test using 2025 data, which constitutes the main empirical evidence of model performance.

Tables 7 and 8 present the results of stratified 5-fold cross-validation for all four models and are used exclusively to assess robustness and stability across different data splits.

Table 9 reports comparative diagnostics of residuals and temporal stability for the models based on the 2025 test set and is intended for diagnostic purposes rather than as a measure of core predictive performance.

The results in Table 1 indicate that the model exhibits a satisfactory capacity to detect procurements exhibiting signs of violations. The recall for the violation class (y = 1) reaches 0.77, meaning that the model correctly identifies more than three-quarters of risky cases in the test sample. This level of sensitivity is particularly relevant in risk-based monitoring systems, where the priority is to minimize undetected violations.

**Table 1.** Results of testing the Random Forest model for identifying potentially risky purchases.

| Type | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (No violations) | 0.59 | 0.49 | 0.54 | 172 |
| 1 (Violations detected) | 0.69 | 0.77 | 0.73 | 215 |
| Accuracy | | | 0.66 | 387 |
| Macro average | 0.64 | 0.63 | 0.63 | 387 |
| Weighted average | 0.65 | 0.66 | 0.65 | 387 |
| ROC-AUC | | | 0.6935 | |

The precision for class (y = 1) equals 0.65, indicating a moderate rate of false positives, whereby some non-risky procurements are incorrectly classified as potentially risky. The overall classification accuracy reaches 0.61, while the ROC-AUC value of 0.67 demonstrates that the model discriminates between risky and non-risky procurements noticeably better than random guessing. At the same time, the lower recall observed for the non-violation class (y = 0) suggests a systematic tendency of the model to overestimate risk. Such behavior is typical and acceptable in risk-screening applications, where reducing the likelihood of missing potentially problematic procurements is prioritized over minimizing false alarms. These limitations justify the further use of more complex ensemble models and sample balancing methods, in particular XGBoost and SMOTE, as well as algorithms capable of working effectively with categorical features.

At the next stage of the study, the XGBoost model was applied to identify potentially risky procurements. The model was trained without employing class-balancing techniques, thereby allowing assessment of its baseline performance under the observed class imbalance. Only numerical features were included in model construction.

Classification performance was evaluated on the test sample using standard binary classification metrics. The results reported in Table 2 indicate that the XGBoost model exhibits greater discriminatory power than the baseline Random Forest model.

The recall for the violation class (y = 1) is 0.66, indicating that the model correctly identifies approximately two-thirds of procurements in which violations were detected. Precision for this class is also 0.66, reflecting a balanced trade-off between the detection of true risky cases and the rate of false positives.

Overall classification accuracy is 0.62, while the ROC-AUC of 0.68 confirms that the XGBoost model exhibits stable discriminatory power under the time-aware validation framework. Despite the absence of explicit class balancing during training, the model demonstrates consistent ability to separate risky from non-risky procurements, rendering it suitable for preliminary risk screening.

These results support the use of XGBoost as a comparatively effective tool for identifying potentially risky procurements. At the same time, the moderate class imbalance observed justifies exploration of sample-balancing techniques, which are examined in the subsequent stage of the analysis.

**Table 2.** Classification performance of the XGBoost model for identifying potentially risky procurements.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (No violations) | 0.57 | 0.57 | 0.57 | 172 |
| 1 (Violations detected) | 0.66 | 0.66 | 0.66 | 215 |
| Accuracy | | | 0.62 | 387 |
| Macro-average | 0.61 | 0.61 | 0.61 | 387 |
| Weighted-average | 0.62 | 0.62 | 0.62 | 387 |
| ROC-AUC | | 0.675 | | |

At the next stage of the study, to enhance the detection of potentially risky procurements, the XGBoost model was retrained in combination with the SMOTE oversampling method.

The resulting model achieved an overall accuracy of 0.62 (Table 3) and a ROC-AUC of 0.681, the highest value recorded among the configurations evaluated at this stage. These figures indicate an improvement in discriminatory performance and greater capacity to capture nonlinear relationships between procurement features and the presence of violations.

**Table 3.** Classification performance of the XGBoost model with SMOTE for identifying potentially risky procurements.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (No violations) | 0.63 | 0.37 | 0.47 | 172 |
| 1 (Violations detected) | 0.62 | 0.82 | 0.71 | 215 |
| Accuracy | | | 0.62 | 387 |
| Macro-average | 0.62 | 0.60 | 0.59 | 387 |
| Weighted-average | 0.62 | 0.62 | 0.60 | 387 |
| ROC-AUC | | 0.681 | | |

At this stage of the experimental analysis, the CatBoost model was applied to identify potentially risky procurements. CatBoost is particularly well-suited to tabular data and can process categorical features natively without requiring prior encoding. This capability represents a substantial advantage in the context of public procurement analysis, where a large proportion of variables are discrete or categorical in nature.

The categorical features incorporated into the model included CPV codes, KVED codes, procurement procedure type, type of contracting authority, and UKTZED codes—variables commonly encountered in public procurement datasets and summarized in Table 6. For the violation class (y = 1), the recall reached 0.82, indicating that more than four-fifths of risky procurements in the test sample were correctly identified. Precision for this class was 0.62, reflecting a moderate rate of false positives. The F1-score for the violation class was 0.71, one of the highest values observed across the models evaluated. Overall classification accuracy stood at 0.62, while the ROC-AUC of 0.681 confirmed the model's adequate discriminatory power in distinguishing between risky and non-risky procurements.

Table 4 presents the detailed performance results for the CatBoost model, including precision, recall, and F1-score for each class, together with the overall ROC-AUC. In addition, the table reports the ranking of feature importance derived from the model's internal importance scores, thereby highlighting the primary risk factors associated with public procurement violations.

**Table 4.** Classification performance of the CatBoost model for identifying potentially risky procurements.

| Class | Precision | Recall | F1-Score | Support |
|---|---|---|---|---|
| 0 (No violations) | 0.58 | 0.61 | 0.59 | 172 |
| 1 (Violations detected) | 0.67 | 0.64 | 0.76 | 215 |
| Accuracy | | | 0.63 | 387 |
| Macro average | 0.63 | 0.63 | 0.63 | 387 |
| Weighted average | 0.63 | 0.63 | 0.63 | 387 |
| ROC-AUC | | 0.727 | | |

To facilitate clear interpretation and effective presentation of the analysis results, data visualization techniques were employed. The graphical representations were generated using the matplotlib and seaborn libraries (Figure 3).

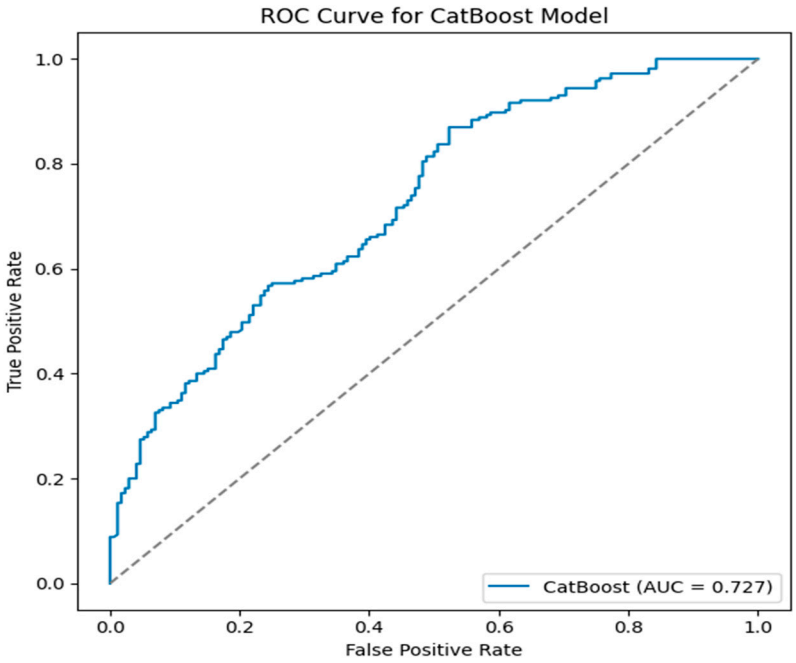


**Figure 3.** ROC curve of the CatBoost model for identifying potentially risky procurements.

Figure 3 displays the receiver operating characteristic (ROC) curve for the CatBoost model, depicting its capacity to discriminate between procurements with detected violations and those without across varying classification thresholds. The area under the ROC curve (AUC = 0.727) reflects satisfactory discriminatory performance and confirms that the model substantially outperforms random classification. The curve exhibits a steady increase in the true positive rate as the false positive rate rises, a pattern characteristic of models designed for risk-oriented monitoring applications.

As part of the empirical analysis, a series of experiments was conducted by retraining the CatBoost model under varying class-weight configurations for the violation class, enabling a quantitative evaluation of the trade-off between recall and false positive rates.

In the baseline configuration (no additional weighting, ratio 1:1), the model exhibited moderate sensitivity: it flagged 205 of 387 procurements as risky, correctly identifying 138 actual violations while leaving 77 undetected. Raising the violation-class weight to the range 1:1.5–1:2 reduced the number of missed risky cases to 62–53, while the total number of flagged procurements increased to 232–243.

Further increases in violation weights (1:3–1:4) produced a marked improvement in recall, with undetected violations falling to 45–31 cases—a reduction of more than half relative to the baseline. This gain in sensitivity, however, came at the expense of a higher false positive rate: the number of non-violating procurements incorrectly flagged rose from 67 in the baseline to 93–105 under the higher-weight configurations. Taken together, the results reveal a systematic pattern: progressively higher penalties for violation misclassification reduce the number of missed risks consistently, albeit at the cost of greater operational burden from additional false alerts—an anticipated characteristic of risk-oriented screening systems.

At the same time, the selection of specific class-weight values remains context-dependent, hinging on the acceptable level of residual risk and the operational resources available to audit authorities. This consideration is examined further in Section 5.

Table 5 presents the performance results of the tested models, including class-specific precision, recall, and F1-score values together with the overall ROC-AUC. These metrics, combined with the accompanying visualizations, provide a clear basis for assessing each model's capacity to detect procurements exhibiting signs of violations and enable direct comparison across the Random Forest, XGBoost, XGBoost with SMOTE balancing, and CatBoost configurations.

**Table 5.** Comparative performance of the models for identifying potentially risky procurements.

| Model | Precision (y = 1) | Recall (y = 1) | F1-Score (y = 1) | Accuracy | ROC-AUC |
|---|---|---|---|---|---|
| Random Forest | 0.69 | 0.77 | 0.73 | 0.66 | 0.694 |
| XGBoost | 0.66 | 0.66 | 0.66 | 0.62 | 0.675 |
| XGBoost + SMOTE | 0.62 | 0.82 | 0.71 | 0.62 | 0.681 |
| CatBoost | 0.67 | 0.64 | 0.76 | 0.63 | 0.727 |

The comparative analysis of the machine learning models reveals distinct trade-offs between sensitivity to violations and overall classification stability. The Random Forest model attains a relatively high recall for the violation class (0.77), enabling it to detect the majority of risky procurements. This performance, however, is offset by a higher rate of false positives, as evidenced by moderate precision and ROC-AUC values. Such behavior indicates a propensity to overestimate risk, which may be tolerable in preliminary screening contexts but constrains its overall discriminatory efficiency.

Comparison of the results highlights the strengths of non-parametric ensemble methods in the context of procurement risk monitoring. The XGBoost model without sample balancing exhibits more conservative classification, yielding balanced yet comparatively lower recall, F1-score, and ROC-AUC values. Although the application of SMOTE increases sensitivity to violations (recall = 0.82), this gain is not matched by a commensurate improvement in overall discriminatory power or accuracy, suggesting that synthetic oversampling provides limited enhancement of model robustness in the present setting.

In this study, the application of SMOTE should therefore be interpreted primarily as an exploratory robustness check addressing class imbalance, rather than as a method that substantively improves the practical applicability or institutional usability of the model.

Among the approaches evaluated, the CatBoost model delivers the most balanced and methodologically sound performance. Without recourse to synthetic data augmentation, it achieves a recall of 0.64 for the violation class, the highest F1-score (0.76), and the strongest ROC-AUC value (0.727). These outcomes underscore CatBoost's effectiveness in managing heterogeneous tabular datasets that contain a substantial proportion of categorical features and exhibit complex nonlinear relationships.

Taken together, the comparative analysis indicates that no single model dominates across all performance dimensions simultaneously. The selection of an appropriate model, therefore, depends on the specific priorities of the monitoring system—whether to maximize the detection of violations or to maintain an optimal balance between accuracy and sensitivity. The findings provide a scientifically grounded foundation for the practical adoption of risk-oriented machine learning approaches in public procurement.

The feature importance analysis obtained from the CatBoost model enables identification of the principal factors that most strongly influence the predicted probability of a procurement being classified as potentially risky. Notably, several of the highest-ranked features reported in Table 6 are categorical in nature. Their substantial influence was effectively captured due to CatBoost's native support for categorical variables, which eliminates the need for prior encoding or transformation. In contrast, conventional models that depend on numerical feature representations may systematically underweight or entirely fail to account for the contribution of such categorical predictors.

**Table 6.** Feature importance ranking derived from the CatBoost model.

| Rank | Feature | Feature Type | Importance |
|---|---|---|---|
| 1 | expected_value | numerical | 13.11 |
| 2 | procurement_method | categorical | 9.34 |
| 3 | revenue_growth_last_available | numerical | 8.03 |
| 4 | cpv_code | categorical | 6.20 |
| 5 | employees | numerical | 4.71 |
| 6 | company_size_last_available | categorical | 4.17 |

The feature importance analysis derived from the CatBoost model enables identification of the principal factors exerting the strongest influence on the predicted probability of classifying a procurement as potentially risky (Table 6). The most influential feature was the expected procurement value (expected_value), providing empirical support for the hypothesis that the risk of violations rises with the financial scale of the contract.

High importance was also assigned to the procurement method (procurement_method) and the Common Procurement Vocabulary code of the procurement item (cpv_code), highlighting the structural character of risks linked to procedural type and the specificity of the procured goods or services.

To strengthen the economic interpretation of the modeling results, a partial dependence analysis was performed for all six key predictors identified in the feature importance assessment (Table 6). Given the mixed data structure, different interpretation tools were applied depending on the nature of the variables. For numerical features, partial dependence plots (PDPs) were constructed using a standard partial dependence framework, reflecting the average marginal effect of each variable on the predicted risk, with other features held constant. For categorical variables, partial dependence at the category level was approximated by comparing the average predicted risk between groups.

The analysis was implemented in Python using standard scientific and machine-learning libraries. Data processing relied on NumPy 2.0.2 and Matplotlib 3.10.0, while

model estimation, validation, and interpretability analysis were conducted using scikit-learn and CatBoost.

An analysis of partial dependence at the category level shows a clear order of procurement methods according to the risk assessments predicted by the model. The highest predicted risk is associated with the negotiated procurement procedure, which is on the left side of the distribution and shows the highest average predicted probability of violation. Slightly lower—but still elevated—risk levels of predicted risk are observed for open tenders, including open tenders with publication in English and open tenders with specific procedural features. These procurement formats consistently receive higher risk scores from the model, indicating that they contain a richer set of risk-related signals that can be detected by the algorithm.

As we move rightward along the distribution, predicted risk scores gradually decrease. Simplified procedures are associated with significantly lower average predicted risk. The lowest predicted risk is observed in procurements conducted without the use of an electronic system, where the model yields only minimal average risk scores.

This pattern likely stems from both data characteristics and institutional features of certain procurement types. Procurements conducted outside electronic systems typically involve significantly less data disclosure and a much more limited set of observable attributes. As a result, they provide fewer risk-relevant signals for algorithmic learning and classification. Consequently, the model assigns lower predicted risk scores to these types of procurements, reflecting reduced detectability of potential violations rather than their actual absence. This finding underscores the need for enhanced oversight and complementary monitoring mechanisms for non-electronic procurement, as algorithmic risk assessment tools are inherently limited by the information environment in which they operate.

The partial dependence analysis further supports an industry-oriented interpretation of procurement risks. At the category level, based on CPV codes, certain procurement segments exhibit systematically higher predicted risk scores, even after controlling for procedural and financial characteristics. In particular, construction-related works, major repair services, energy supply, and technically complex goods show the highest average predicted risk.

At the same time, empirical data on the observed frequency of detected violations provide a complementary perspective. In the study sample, several CPV categories, such as footwear (1881), medical devices (3312), pharmaceutical products (3369), and repair and maintenance of motor vehicles (5021), account for the largest number of recorded violations, with up to 20 cases observed in some categories. This concentration reflects not only sector-specific risk intensity but also procurement volume and repeated exposure over time. Taken together, these findings suggest that procurement risk should be interpreted as a combination of sector vulnerability (as captured by the model's predicted risk levels) and empirical exposure (as reflected in the observed frequency of violations), rather than solely from a procedural lens.

The identification of risk indicators in public procurement is not a novel undertaking, given the existence of institutionalized automated risk detection mechanisms in Ukraine. In particular, the Ministry of Finance has endorsed a formal methodology for the application of automatic risk indicators within the electronic procurement system, which functions as the principal trigger for initiating monitoring by the State Audit Service. Within this framework, the majority of monitoring cases are initiated not based on direct evidence of violations but through the activation of predefined automated markers. Public complaints and manually identified irregularities assume a secondary role, while automated risk scoring constitutes the primary mechanism for prioritizing procurement procedures for further examination (Ministry of Finance of Ukraine, 2024).

The existing Ukrainian methodology establishes a hierarchical system of risk indicators assigned priority levels (high, medium, low) and ranks procurements according to the number and severity of triggered indicators. These indicators are predominantly based on procedural and transactional attributes of individual procurements, including parameters such as procurement method, number of bidders, rejection patterns, contract amendments, payment terms, procurement value, and timing characteristics. In some instances, limited historical data are incorporated, for example, the number of procurements conducted by a supplier. However, the financial, operational, and economic characteristics of suppliers themselves are largely excluded from automated risk assessment (Ministry of Finance of Ukraine, 2024).

This design reflects a deliberate regulatory orientation toward ensuring formal compliance with procurement legislation and detecting procedural deviations that may indicate breaches of public procurement law. As a consequence, risk indicators are assessed largely in isolation, with each triggering independently according to predefined thresholds. Interactions among multiple risk factors, as well as the broader financial profile and behavioral patterns of suppliers, are not explicitly modeled within the current institutional framework.

The findings of the present study indicate that this approach captures only a subset of the overall risk landscape. The feature importance analysis reveals that procurement risk is driven not by isolated indicators but by the joint effect and interaction of multiple characteristics. Expected contract value, procurement method, and CPV classification emerge as key predictors, yet their contribution to risk is particularly pronounced when considered in combination rather than individually. For instance, large contract values are not inherently risky in isolation; however, when paired with less competitive procurement procedures or specific CPV categories, they substantially elevate the probability of detected violations—an interaction effect that rule-based indicator systems are unable to accommodate.

Moreover, the analysis underscores the importance of supplier-level financial dynamics. Revenue growth proves more predictive than absolute revenue levels, as rapid expansion may reflect capacity constraints, aggressive market entry, or opportunistic behavior amid heightened demand in defense-related procurement. Such dynamics can heighten execution risks or create incentives for non-compliance. These patterns are consistent with concerns routinely identified in audit practice and financial oversight, yet they remain absent from prevailing automated monitoring frameworks.

The results demonstrate that risk classification is driven not by the activation of a single indicator but by the synergistic effect of multiple factors acting concurrently. Comparable insights appear in international risk frameworks. The FALCON project and the R2G4P training manual highlight contract value, procedural discretion, limited competition, and sector-specific characteristics as central drivers of corruption and inefficiency risks (FALCON Horizon, 2025; OECD & European Commission, 2021). While these frameworks acknowledge that risk often emerges through interactions among multiple indicators rather than isolated signals, most practical implementations of indicator-based systems continue to employ additive or rule-based logic that assesses each marker independently.

The incorporation of supplier-level financial dynamics introduces a dimension largely overlooked in existing automated monitoring frameworks. The greater informativeness of revenue growth relative to absolute revenue levels suggests that rapid expansion may signal elevated execution risk or opportunistic behavior during periods of increased public demand. From a risk management standpoint, these patterns align with execution and sustainability concerns documented in audit practice and prior machine learning applications to procurement red flags (Lima et al., 2023).

Empirical evidence on the association between CPV codes and the frequency of detected violations supports interpreting procurement risks from a sectoral rather than

exclusively procedural perspective. This uneven distribution can be attributed to the structural features of these procurement segments. Many of the identified categories are characterized by technical complexity, regulated pricing mechanisms, or a restricted pool of qualified suppliers. These conditions amplify information asymmetry between contracting authorities and bidders, hinder objective price benchmarking, and weaken competitive discipline. Consequently, even formally compliant procedures in these sectors may entail elevated underlying risks. The prominent position of CPV codes among the most influential predictors in the CatBoost model (Table 6) should not be construed as evidence that specific product categories are inherently problematic. Rather, it reflects the reality that procurement risk materializes at the intersection of the procurement object, procedural discretion, and supplier characteristics.

To assess the robustness of ensemble learning results beyond headline performance metrics, feature-importance stability was evaluated using a time-aware cross-year comparison. Feature importance scores were obtained from CatBoost models trained on separate annual subsamples using the CatBoost library, while rank-based stability was assessed via Spearman rank correlation (SciPy implementation). For each year, feature importance rankings were extracted and compared pairwise across years: the overlap of the top-10 most important features was also computed.

From a methodological perspective, Spearman's rank correlation values above 0.4–0.5 are commonly interpreted as indicating moderate stability in applied empirical research, particularly in settings characterized by structural heterogeneity and temporal volatility. In this study, pairwise Spearman correlations of feature importance range from approximately 0.39 to 0.59, with an average of about 0.51 across year pairs—placing the results within or above the range typically considered acceptable for robust inference. In parallel, the overlap among the top-10 most important features remains substantial, averaging 6 and 7 shared predictors across different years. Taken together, these results suggest that while the relative ordering of individual predictors varies over time, the model consistently relies on a stable core set of features.

To evaluate the stability and generalization capability of the machine learning models, stratified 5-fold cross-validation was performed using the StratifiedKFold implementation. The procedure was applied to three models: Random Forest, XGBoost, and XGBoost with SMOTE-based class balancing. The aggregated cross-validation results are reported in Table 7.

**Table 7.** Aggregated results of Stratified 5-Fold Cross-Validation for machine learning models.

| Model | Mean ROC-AUC | Std. ROC-AUC | Mean Recall (Class 1) | Std. Recall (Class 1) |
|---|---|---|---|---|
| Random Forest | 0.706 | 0.029 | 0.711 | 0.038 |
| XGBoost (no SMOTE) | 0.696 | 0.037 | 0.706 | 0.045 |
| XGBoost + SMOTE | 0.674 | 0.033 | 0.748 | 0.061 |

The aggregated results from stratified 5-fold cross-validation and subsequent out-of-sample testing on the 2025 hold-out set indicate that none of the evaluated models exhibits unequivocal superiority across all performance metrics. Nevertheless, the analysis identifies distinct strengths associated with each approach in the context of risk-oriented procurement monitoring.

The Random Forest model recorded the highest mean ROC-AUC (0.706) and, at the same time, the lowest variability across folds. These results confirm its superior stability and robustness under repeated resampling and heterogeneous data partitions. In addition, Random Forest consistently achieved high recall for the violation class, underscoring its

particular suitability for risk-screening applications in which the minimization of false negatives (undetected violations) is essential to safeguarding national security and public financial resources.

The combination of SMOTE with XGBoost produced a substantially higher mean recall for the violation class (0.748), reflecting enhanced sensitivity to risky procurements. This improvement in recall, however, was offset by increased variability across folds and a modest decline in precision, leading to a higher incidence of false positives.

To evaluate the stability and generalization capability of the CatBoost model, repeated stratified 5-fold cross-validation was performed using the StratifiedKFold implementation (Table 8). The reported metrics include the area under the ROC curve (ROC-AUC) and recall for the positive class (class 1), which together capture the model's effectiveness in identifying procurements involving violations.

**Table 8.** Results of stratified 5-fold cross-validation for the CatBoost model.

| Fold | ROC-AUC | Recall (Class 1) |
|---|---|---|
| 1 | 0.781 | 0.714 |
| 2 | 0.785 | 0.733 |
| 3 | 0.789 | 0.789 |
| 4 | 0.720 | 0.722 |
| 5 | 0.821 | 0.846 |
| Mean | 0.779 | 0.761 |
| Std. Dev. | 0.033 | 0.051 |

The results of cross-validation indicate a stable performance of the CatBoost model across different data subsamples. The mean ROC-AUC value of 0.779 suggests a more than satisfactory (good) ability of the model to distinguish procurements with signs of violations from non-risky cases. The relatively low standard deviation of the ROC-AUC (0.033) confirms the absence of significant performance fluctuations across individual folds. The average recall for the violation class equals 0.779, indicating that the model is capable of identifying approximately 77% of risky procurements.

The applied machine learning literature emphasizes that a low standard deviation of performance metrics within cross-validation is an indicator of a model's generalization capability and the absence of excessive dependence on a specific data split (Hastie et al., 2009). Standard deviation values not exceeding 0.05–0.10 for key classification metrics are generally considered acceptable for real-world applied tasks involving noisy data (Géron, 2022). According to the results obtained, the standard deviation value of 0.051 indicates an acceptable level of variability and the absence of critical performance degradation across individual data splits.

To check the reliability of the forecasts and verify the models for compliance with econometric stability requirements, an in-depth analysis of the residuals on an out-of-sample test set (out-of-time test set 2025) was performed (Table 9). Since in machine learning the validity of non-parametric methods is proven not by a priori assumptions about noise distribution, but by the stability of errors on new data, we applied a method for diagnosing classification residuals. The purpose of this stage was to determine the level of systematic bias of each model and to test its stability against heavy tails and data volatility. The assessment was carried out using statistical analysis of mean values, standard deviation, and percentile distribution of errors.

**Table 9.** Comparative diagnostics of residuals and temporal stability of models.

| Model | ROC-AUC | Mean Residual | Std. Residual | $Q_{0.05}$ | $Q_{0.95}$ |
|---|---|---|---|---|---|
| Random Forest | 0.666 | 0.024 | 0.477 | −0.663 | 0.666 |
| XGBoost | 0.675 | 0.017 | 0.507 | −0.846 | 0.868 |
| XGBoost + SMOTE | 0.683 | 0.053 | 0.501 | −0.821 | 0.876 |
| CatBoost | 0.915 | 0.031 | 0.349 | −0.613 | 0.590 |

The comparative residual diagnostics presented in Table 9 provide an explicit assessment of model uncertainty and predictive stability on an out-of-sample test set (2025). Beyond standard discrimination metrics, the analysis evaluates the distributional properties of prediction errors through mean residuals, residual standard deviations, and empirical quantiles. Across all models, mean residuals remain close to zero, indicating the absence of systematic bias in predictions. At the same time, substantial differences emerge in the dispersion of residuals: while Random Forest and XGBoost-based specifications exhibit relatively high residual standard deviations (approximately 0.48–0.51), the CatBoost model demonstrates markedly lower dispersion (0.349), suggesting tighter concentration of prediction errors and greater reliability of probabilistic outputs.

Importantly, the use of residual quantiles ($Q_{0.05}$ and $Q_{0.95}$) provides an empirical approximation of uncertainty bounds around model predictions, directly addressing concerns related to variability and robustness. The narrower inter-quantile range observed for CatBoost (approximately $-0.61$ to 0.59) compared to other architectures indicates substantially lower tail risk and reduced volatility in extreme prediction errors. This evidence shows that the superior performance of CatBoost is not limited to higher ROC-AUC values but is also accompanied by more stable and predictable error behavior. As such, the results already incorporate a systematic evaluation of model uncertainty and out-of-sample stability, aligning with best practices for policy-relevant risk assessment—where the reliability of predictions is as important as their average accuracy.

The comparative analysis demonstrates the superior performance of the CatBoost model, which not only achieved the highest discriminative ability but also exhibited the greatest stability in its error structure. CatBoost recorded the lowest residual standard deviation, reflecting greater prediction density and substantially lower error variance relative to the other architectures considered.

The application of SMOTE in conjunction with XGBoost yielded a modest improvement in ROC-AUC; however, it produced the highest systematic bias and retained considerable error variability, rendering it less suitable for reliable, long-term institutional monitoring. Random Forest displayed moderate overall performance but was outperformed by CatBoost in terms of both predictive stability and consistency.

The results confirm that the models exhibit no critical quality deficiencies across individual cross-validation folds. The observed variation in performance metrics among folds appears to reflect the inherent structural heterogeneity characteristic of real-world procurement datasets, rather than model instability. Furthermore, no fold displayed anomalously poor performance.

## 5. Discussion

The results of this study highlight the value of combining econometric and algorithmic approaches in risk-oriented monitoring of military procurement. Econometric models, including standard logistic regression and its LASSO-penalized specification, provide a transparent and interpretable baseline for risk assessment and exhibit stable average discriminatory performance. Such models remain valuable for establishing initial

risk benchmarks and for interpreting the effects of individual predictors within a linear modeling framework. At the same time, the comparative analysis shows that ensemble machine learning models extend the analytical capacity of procurement monitoring by capturing nonlinear interactions and complex data structures. In particular, Random Forest and XGBoost demonstrate higher sensitivity to procurements with detected violations, while the CatBoost model offers the most balanced trade-off between detection sensitivity, overall classification quality, and predictive stability. These findings do not challenge the relevance of econometric methods; rather, they indicate that ensemble models can effectively complement traditional approaches in data-rich and structurally complex procurement environments.

The comparative analysis of Random Forest, XGBoost, XGBoost with SMOTE, and CatBoost (see Table 5) shows that advanced gradient boosting algorithms, particularly CatBoost, deliver the most balanced trade-off between sensitivity to violations and overall discriminatory power in a highly heterogeneous and institutionally sensitive setting. CatBoost stands out as the most effective model overall, achieving the highest F1-score (0.76) and ROC-AUC (0.727) while retaining a recall of 0.64 for detected violations. This performance is especially notable because CatBoost attains these results without relying on synthetic oversampling, underscoring its strength in natively handling categorical variables and complex feature interactions. These findings align with existing literature highlighting the strengths of gradient boosting techniques in high-dimensional tabular data involving mixed data types (Hancock & Khoshgoftaar, 2020; Hastie et al., 2009).

Random Forest exhibits relatively strong recall (0.77), reflecting high sensitivity to potential violations. However, this advantage is offset by lower overall accuracy (0.66) and F1-score (0.73), suggesting a propensity to over-identify procurements as risky. From a policy and operational standpoint, such behavior may be tolerable—or even desirable—in early-warning systems where false negatives carry particularly high costs. Nevertheless, in resource-constrained audit environments, an excessive rate of false positives can impair the efficiency of control processes, highlighting the value of more balanced classifiers.

The results for XGBoost and XGBoost with SMOTE further illustrate the trade-offs involved in addressing class imbalance. Although SMOTE raises recall for the violation class to 0.82, this gain is accompanied by only marginal improvement in F1-score and no meaningful increase in overall accuracy. This pattern implies that synthetic oversampling yields limited benefits in settings where risk patterns exhibit structural complexity rather than mere underrepresentation. Such observations are consistent with prior studies cautioning against the routine application of oversampling methods in institutional and corruption-related datasets (He & Garcia, 2009).

Taken together, the comparative findings indicate that CatBoost achieves an optimal balance among sensitivity, precision, and discriminatory ability. This supports the conclusion that AI-supported procurement monitoring systems should favor algorithms adept at capturing nonlinear relationships, categorical complexity, and institutional nuances, thereby improving both analytical reliability and practical utility in defense procurement oversight.

From a methodological perspective, the results of this study confirm the complementary interpretation of econometric and algorithmic modelling paradigms, aligning with Braiman's concept of the 'two cultures.' Econometric models remain indispensable for analysis, particularly when clear assumptions about the data-generating process are required. In contrast, algorithmic models offer practical advantages in capturing nonlinear interactions and structural heterogeneity in complex, data-rich environments. Rather than favouring one paradigm over the other, this study demonstrates how econometric models and machine learning ensemble models can be used in combination to enhance both the analytical rigour and operational efficiency of procurement risk monitoring.

The feature importance analysis (Table 6) provides substantive validation of the model results. The dominant role of expected contract value confirms well-established theoretical and empirical findings that higher-value contracts are associated with elevated risks of corruption and manipulation (Sales, 2013; Schwartz, 2004).

The significance of procurement method and CPV code indicates that risks are structurally embedded in procedural design and sectoral specificity. Certain procurement methods and CPV categories are systematically more exposed to discretion, limited competition, and information asymmetry, which is consistent with prior evidence on military logistics and defense procurement (Zozulya & Matrosov, 2018; Ozsevim, 2023).

Competition-related indicators, including the number of bids and disqualifications, further support the interpretation that distorted competitive environments signal heightened manipulation risks. This aligns with studies emphasizing the role of AI-based tools in detecting abnormal competitive patterns as early indicators of corruption (Ayibam, 2025; Karpenko et al., 2023). Finally, the importance of supplier-level financial characteristics—such as revenue growth, employment size, and financial scores—demonstrates the added value of integrating firm-level data via YouControl. This extends existing Ukrainian research (Artiushenko, 2024; Sizov et al., 2025) by showing that supplier financial dynamics are active determinants of procurement risk, particularly in fragile wartime economic conditions.

The conclusions drawn from Table 6 about the importance of indicators in the model have a direct impact on the development of risk-based monitoring systems. They indicate that CPV codes should not be considered as isolated warning signals, but should be included in sectoral risk profiles that interact with other dimensions, such as expected contract value, procurement method, and the financial characteristics of the supplier.

Integrating historical patterns of violations by CPV category into automated risk assessment systems will enable oversight bodies to better identify "sectoral blind spots" where standard procedural controls are less effective. Combined with machine learning techniques, this approach enables a shift from static selection toward more context-dependent identification of high-risk procurements. The significance of procurement method and CPV code further indicates that risks are structurally embedded in procedural design and sectoral specificity. Certain procurement methods and CPV categories are systematically more exposed to discretion, limited competition, and information asymmetry, which is consistent with prior evidence on military logistics and defense procurement (Zozulya & Matrosov, 2018; Ozsevim, 2023).

As demonstrated by the experiments (CatBoost) with alternative class-weight configurations, increasing the weight of the violation class from 1:1 to 1:3–1:4 raises the number of procurements flagged as risky from approximately 205 to 263–289 cases, while the number of false positives increases from 67 to 93–105 cases. This directly translates into a higher operational burden for audit authorities, as each falsely flagged procurement requires additional review without leading to the detection of an actual violation. Therefore, the use of aggressive class-weighting schemes is justifiable only when sufficient institutional capacity exists to process the increased volume of risk signals.

At the same time, the feature importance analysis (Table 6) indicates that procurement risks are structurally embedded, as expected contract value, procurement method, and CPV code emerge among the most influential determinants. This suggests that class weighting should not be applied in isolation but rather integrated into a broader risk monitoring framework that combines probabilistic risk scores with procedural and market-specific characteristics. Such an integrated approach enables the prioritization of audit efforts across procurement types and market segments, thereby mitigating excessive operational burden while preserving high sensitivity to genuinely high-risk procurements.

The analysis of the partial dependence plot for both numerical and categorical predictors included in the model reveals the following patterns in the data. In particular, procurements conducted outside electronic procurement systems are associated with significantly lower predicted risk scores. This should be interpreted not as indicating lower intrinsic risk, but rather as reflecting reduced information content and weaker risk signals available for algorithmic detection. This finding highlights an important limitation of data-driven monitoring tools and underscores the need for supplementary oversight mechanisms in procurement segments characterised by limited transparency and data availability.

At the same time, the study provides a basis for further scientific research. A promising direction is to deepen the financial analysis of suppliers by including indicators of liquidity, asset structure, cash flows, and dependence on budget funds. Special attention should be paid to the analysis of textual procurement monitoring reports, which contain detailed descriptions of identified violations. The use of natural language processing (NLP) methods will make it possible to identify typical patterns of violations and to link the qualitative characteristics of violations to quantitative procurement indicators. In addition, combining machine learning results with industry and sector data is promising, as it will allow market-specific characteristics to be considered.

## 6. Conclusions

The comparative analysis presented in this study indicates that no single model achieves unequivocal superiority across all performance metrics. Rather, the findings highlight distinct trade-offs among sensitivity in detecting violations, overall classification accuracy, and prediction stability. Within this framework, the CatBoost model stands out as the most balanced and robust approach for identifying potentially high-risk military procurement contracts.

Among the ensemble methods evaluated, Random Forest demonstrates high sensitivity to violations (recall = 0.77), rendering it particularly suitable for early-warning systems in which the minimization of false negatives represents a primary objective. This strength, however, comes at the cost of a greater propensity to overestimate risk, resulting in elevated false-positive rates and diminished discriminatory power. By contrast, the XGBoost model without sample rebalancing exhibits more conservative and stable performance, delivering moderate yet balanced values for recall, F1-score, and ROC-AUC. The incorporation of SMOTE alongside XGBoost elevates recall to 0.82; nevertheless, this improvement is not matched by commensurate gains in overall accuracy or ROC-AUC and coincides with diminished prediction stability. These patterns suggest that synthetic oversampling via SMOTE offers limited incremental benefit in the present institutional context.

The CatBoost model exhibits the most advantageous trade-off between sensitivity, precision, and overall discriminatory performance. Without the use of synthetic data augmentation, it attains a recall of 0.64, the highest among the models evaluated, along with the highest F1-score (0.76) and the highest ROC-AUC (0.727). Cross-validation results presented in Table 8 further substantiate the stability and generalization capability of CatBoost, underscoring its robustness in the presence of temporal variability and data heterogeneity. Such attributes are especially pertinent to operational risk-oriented monitoring systems that function amid uncertainty and diverse data regimes. The stability analysis confirms that the ensemble model identifies procurement risks in a sufficiently consistent manner over time, with feature-importance rankings exhibiting moderate, methodologically acceptable Spearman correlations and substantial overlap of key predictors across years. This indicates that the model's conclusions are not driven by year-specific noise while remaining flexible enough to adapt to changing institutional and market conditions.

Table 9 reports comparative diagnostics of residual behavior and temporal stability across the evaluated models, using the 2025 test set. The findings reveal that, notwithstanding variations in primary predictive performance documented earlier, the models exhibit marked differences in error distributions and consistency over time. Notably, CatBoost records the lowest standard deviation of residuals, signifying greater stability and reduced volatility in prediction errors relative to competing methods. The correspondingly high ROC-AUC value for CatBoost in Table 9 attests to its robust capacity to distinguish between correctly and incorrectly classified instances at the residual level.

Examination of prevailing procurement monitoring practices in Ukraine indicates that activation of an "automatic marker" within the Prozorro system constitutes the principal trigger for initiating monitoring. The prevailing set of indicators, however, predominantly encompasses procedural dimensions, such as participant disqualifications, post-award price modifications, and compliance with document publication deadlines. While the system incorporates certain metrics of supplier activity, it largely omits analysis of core financial indicators, thereby neglecting the economic capacity of suppliers to meet contractual obligations. In contrast to conventional indicator-based frameworks, which assess procurement risks via discrete procedural alerts, the approach proposed herein illustrates that risk arises from the interplay among multiple determinants, encompassing contract value, procurement procedure, sectoral attributes (as captured by CPV codes), and supplier-specific financial trajectories. Crucially, these factors are intended not to supplant existing indicators but to augment them.

**Author Contributions:** Conceptualization, T.Z. and O.D.; methodology, T.Z. and O.A.; software, A.I. and O.L.; validation, I.C.L., T.Z. and O.D.; formal analysis, O.A. and T.Z.; investigation, O.D.; resources, A.I. and O.L.; data curation, I.C.L.; writing—original draft preparation, T.Z., O.D. and O.A.; writing—review and editing, I.C.L., A.I. and O.L.; visualization, I.C.L.; supervision, T.Z.; project administration, T.Z. and O.D.; funding acquisition, T.Z. and O.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** YouControl platform. https://youcontrol.com.ua/ (accessed on 1 December 2025).

**Acknowledgments:** During the preparation of this manuscript, the authors used Grammarly Desktop 1.144.1 (Superhuman) and ChatGPT (OpenAI, GPT-5.1 Thinking model) for language polishing and minor text editing. The authors reviewed and edited all AI-assisted outputs and took full responsibility for the final content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

Aboelazm, K. S., & Dganni, K. M. (2025). Public procurement contracts futurity: Using of artificial intelligence in a tender process. *Corporate Law & Governance Review*, *7*(1), 60–72. [CrossRef]

Andersson, P. E., Arbin, K., & Rosenqvist, C. (2025). Assessing the value of artificial intelligence (AI) in governmental public procurement. *Journal of Public Procurement*, *25*(1), 120–139. [CrossRef]

Artiushenko, O. (2024). Application of multi-criteria optimisation method for decision-making in the sphere of financial support of military troops. *Visnyk Taras Shevchenko National University of Kyiv. Military-Special Sciences*, *2*(60), 53–57. [CrossRef]

Ayibam, J. N. (2025). Artificial intelligence in public procurement: Legal frameworks, ethical challenges, and policy solutions for transparent and efficient governance. *Alkebulan: A Journal of West and East African Studies*, *5*(2), 54–69. [CrossRef]

Balaniuk, R., Bessiere, P., Mazer, E., & Cobb, P. R. (2013). Corruption risk analysis using semi-supervised naïve Bayes classifiers. *International Journal of Reasoning-Based Intelligent Systems*, *5*(4), 237–246. [CrossRef]

Balkan, D., & Akyuz, G. A. (2025). Artificial intelligence (AI) and machine learning (ML) in procurement and purchasing decision-support: A taxonomic literature review and research opportunities. *Artificial Intelligence Review*, *58*(11), 341. [CrossRef]

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. [CrossRef]

Chaprak, Y., & Dluhopolskyi, O. (2022, December 26–27). *Features of public procurement during wartime: The case of Ukraine*. The 1st International Scientific Conference (pp. 159–165), Prague, Czech Republic. Available online: https://ojs.publisher.agency/index.php/RR/issue/view/12 (accessed on 1 December 2025).

Dluhopolskyi, O., & Chaprak, Y. (2023). Key research directions and dilemmas in public procurement: A theoretical framework. *Innovation and Sustainability*, *1*, 51–63. [CrossRef]

Dluhopolskyi, O., & Danyliuk, I. (2023). Assessment of savings in the electronic public procurement system: The PROZORRO case, 2018–2022. *Socio-Economic Relations in the Digital Society*, *4*(50), 95–111. [CrossRef]

FALCON Horizon. (2025). *Combatting corruption in public procurement: Policy brief No. 03*. FALCON—Fighting against Large-Scale Corruption and Organized Crime Networks. Available online: https://www.falcon-horizon.eu/wp-content/uploads/sites/6/2025/11/2025_11_04-FALCON-Policy-brief-03_-Combatting-Corruption-in-Public-Procurement.pdf (accessed on 1 December 2025).

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. [CrossRef]

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (3rd ed.). O'Reilly Media.

Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: An interdisciplinary review. *Journal of Big Data*, *7*, 94. [CrossRef]

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. [CrossRef]

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. [CrossRef]

Karpenko, O., Karpenko, Y., & Herman, D. (2023). Artificial intelligence as a tool for reducing corruption risks in public procurement. *Public Administration Aspects*, *11*(2), 129–135. [CrossRef]

Lima, W., Lira, R., Paiva, A., Silva, J., & Silva, V. (2023, June 18–23). *Methodology for automatic extraction of red flags in public procurement*. International Joint Conference on Neural Networks (IJCNN) (pp. 1–8), Gold Coast, Australia. [CrossRef]

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, *3*(1), 27–32. [CrossRef]

Ministry of Finance of Ukraine. (2024). *On approval of the methodology for determining automatic risk indicators in public procurement* (Order No. 476, 27 September 2024). Official Bulletin of Ukraine. Available online: https://zakon.rada.gov.ua/laws/show/z1474-24 (accessed on 1 December 2025).

OECD & European Commission. (2021). *Training manual on analysing public procurement risks: Reducing the Risk of Corruption in Public Procurement (R2G4P)*. OECD Publishing. Available online: https://seldi.net/wp-content/uploads/2022/01/R2G4P_Training-manual-on-analyzing-public-procurement-risks.pdf (accessed on 1 December 2025).

Ozsevim, I. (2023). Public military procurement lacks agility and is "broken". *Procurement Magazine*. Available online: https://procurementmag.com/articles/public-military-procurement-lacks-agility-and-is-broken (accessed on 1 December 2025).

Sales, L. (2013). Risk prevention of public procurement in the Brazilian government using credit scoring. *OBEGEF Working Papers*, *19*, 1–27. Available online: https://ideas.repec.org/p/por/obegef/019.html (accessed on 1 December 2025).

Sanchez-Graells, A., & O'Connell, E. (2024). *Demystifying AI in public procurement: Following a beginner's curiosity*. European Institute of Public Administration (EIPA). Available online: https://www.eipa.eu/blog/demystifying-ai-in-public-procurement-following-a-beginners-curiosity/ (accessed on 1 December 2025).

Schwartz, J. I. (2004). The centrality of military procurement: Explaining the exceptionalist character of United States federal public procurement law. *GW Law Faculty Publications & Other Works, Faculty Scholarship*. Available online: https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2024&context=faculty_publications (accessed on 1 December 2025).

Sizov, A., Artiushenko, O., & Dyachenko, S. (2025). Analysis of the needs of the target group in the financial support system of the Armed Forces of Ukraine and estimation of the economic efficiency of project solutions. *Herald of Khmelnytskyi National University. Economic Sciences*, *1*, 238–250. [CrossRef]

Wang, H., Sua, L. S., & Alidaee, B. (2023). Enhancing supply chain security with automated machine learning. *Preprint*. [CrossRef]

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). MIT Press.

YouControl. (2025). *YouControl platform*. Available online: https://youcontrol.com.ua/ (accessed on 1 December 2025).

Zatonatska, T., & Soboliev, O. (2024). Transformation of state regulation in the energy sector under the influence of the Russian-Ukrainian war. *Visnyk of Taras Shevchenko National University of Kyiv. Economics*, *2*(225), 15–26. Available online: https://econom.bulletin.knu.ua/uk/article/view/2713/3191 (accessed on 1 December 2025).

Zozulya, A., & Matrosov, M. (2018). Analysis of the specificity of the interaction of the military unit with suppliers in logistic processes of military logistics. *ScienceRise*, *10*(10), 6–9. [CrossRef]