SEVEN publicações acadêmicas
EM PARCERIA COM: IEMS UNIASSELVI

Predictive modeling for the detection of anomalies in public contracts

Sara Borges Lopes de Sousa[1]

ABSTRACT

Public procurement represents one of the most significant areas of government expenditure and, consequently, one of the most vulnerable to inefficiencies and irregularities. This study proposes and validates a predictive data-science framework designed to detect anomalies in public contracts, specifically focusing on the probability of financial amendments (*aditivos*) as indicators of potential deviations. The research follows an end-to-end methodological structure, encompassing data acquisition from open-government sources, preprocessing, feature engineering, and the implementation of a **Gradient Boosting Machine (GBM)** model optimized for highly imbalanced datasets. Empirical validation revealed strong performance, with the model achieving a recall rate of **0.85**, emphasizing sensitivity over precision to minimize the non-detection of real irregularities. Beyond technical development, the study also discusses the necessity of **Explainable Artificial Intelligence (XAI)** for algorithmic transparency and explores the **Blockchain** technology as a potential foundation for next-generation auditing ecosystems. Ultimately, the paper contributes a reproducible roadmap for algorithmic governance, strengthening proactive oversight mechanisms and supporting data-driven decision-making in the public sector.

Keywords: Data Mining. Public Contracts. Anomaly Detection. Explainable AI (XAI). Blockchain.

1 INTRODUCTION

Public procurement plays a crucial role in national economies, accounting for approximately 12% of global GDP, or nearly USD 13 trillion annually (Williams, 2015). Despite its magnitude, the Brazilian public procurement system remains particularly exposed to management failures and illicit practices, resulting in inefficiency and a loss of institutional credibility. Traditional audit models, often reactive and based on post-factum analyses of a limited subset of contracts, are structurally incapable of coping with the complexity and volume of data generated by the public administration. This reactive logic leads to late interventions, which typically occur only after the damage to public resources has materialized.

The emergence of open government initiatives in Brazil, especially following the Access to Information Law (Law No. 12.527/2011), created an unprecedented data ecosystem. Platforms such as the *Transparency Portal* and the *National Public Procurement Portal (PNCP)* provide a vast repository of data on bids and contracts (Fernandes, Silva;

[1] Master's student in Data Science. Universidade de Brasília (UnB). Brasília, Brazil.
E-mail: sara.blsousa@gmail.com

Oliveira, 2021). However, the mere availability of data does not guarantee transparency. Fragmentation across systems, the absence of unified data standards, and the difficulty of automated access still hinder practical use (Williams, 2015; Nai; Al-Boni; Bifet, 2022).

Comparable challenges related to data fragmentation and underreporting have also been documented in epidemiological monitoring during the COVID-19 pandemic (de Carvalho; de Medeiros; Magalhães, 2024), reinforcing the need for predictive models that enhance public data quality and reliability.

Within this context, data science and machine learning emerge as transformative tools capable of converting massive, heterogeneous public data into actionable intelligence. This research applies such techniques to develop a predictive auditing model that shifts the focus from retrospective assessments to proactive risk evaluation. Rather than replacing the auditor's judgment, the model enhances it by offering a triage system that identifies and prioritizes contracts with a high probability of anomaly — specifically, significant financial amendments (*aditivos*). This paradigm shift redefines public auditing from a *detective role*, centered on investigating past irregularities, to a *sentinel role*, focused on early warning and prevention.

Previous quantitative studies have demonstrated the effectiveness of data-driven approaches in experimental contexts, such as the larvicidal and pupicidal activities of phenolic compounds from *Anacardium occidentale* against arbovirus vectors (de Carvalho et al., 2019).

The article presents four main contributions to both government auditing and applied data science:

(1) a systematic, replicable methodology for predictive modeling of anomalies in public contracts;

(2) the empirical validation of a GBM model optimized for high recall in a realistically imbalanced dataset;

(3) an operational discussion on interpretability, emphasizing XAI as a prerequisite for algorithmic legitimacy in the public sector; and

(4) a critical assessment of Blockchain as a possible future infrastructure for transparency and immutable audit records.

2 METHODOLOGY

The methodological design followed best practices in data science to ensure rigor, reproducibility, and operational relevance. The process encompassed the full analytical cycle, from data acquisition and cleaning to model development, optimization, and validation. In this study, an **anomaly** is defined as any contractual modification that produces a direct financial impact on public resources, particularly through **significant additive terms** (*aditivos contratuais*).

2.1 DATA ACQUISITION AND PREPROCESSING

The dataset was built using from two main sources: the *Transparency Portal* and the *National Public Procurement* **open government data** *Portal (PNCP)*. These sources provided structured information (e.g., contract value, dates, bidding modality, supplier ID, and contracting entity) and unstructured data (e.g., textual descriptions of contract objects and amendment justifications).

Due to the lack of standardized formats and robust APIs, **web scraping** was employed to automate data collection (Fernandes et al., 2021). The resulting raw data underwent several preprocessing steps:

- **Missing Value Treatment:** Median imputation for numerical variables and mode imputation for categorical ones.
- **Normalization:** Standardization of numeric variables (zero mean, unit variance) to prevent magnitude bias.
- **Categorical Encoding:** One-Hot Encoding for nominal variables (e.g., “*Pregão*,” “Concorrência,” “Dispensa”).

These transformations ensured consistent and analyzable inputs for model training.

2.2 EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

Exploratory Data Analysis (EDA) was conducted to understand variable distributions, correlations, and class imbalance — where anomalous contracts represented less than 3% of the total sample.

Feature Engineering proved decisive in improving model performance, as anomalies tend to emerge from **contextual patterns** rather than isolated features (IBM, 2022). The most relevant feature categories included:

This approach resembles biological research where the interaction among multiple compounds determines the overall outcome, as observed in the insecticidal activity of *Persea americana* extracts against *Aedes aegypti* larvae (de Carvalho et al., 2011).

- **Temporal Features:** Time between bidding and contract signing, interval to the first amendment, and seasonality indicators (month, weekday).
- **Relational Features:** Historical frequency of supplier contracts, total contracts per auditor, and prior amendment counts (Rodríguez et al., 2022).
- **Deviation and Ratio Features:** Ratios between amendment and original values, accumulated amendment totals, and deviation from historical averages for similar contracts.
- **Textual Features:** Text mining via TF-IDF (Term Frequency–Inverse Document Frequency) on descriptive and justification texts, revealing linguistic patterns linked to risk (Grace et al., 2016).

2.3 PREDICTIVE MODELING AND CLASS IMBALANCE TREATMENT

The prediction task was structured as a **binary classification problem**, where the dependent variable *is_anomaly* equals 1 if the contract exceeded a predefined amendment threshold (≥ 25% of the original value) and 0 otherwise.

The chosen algorithm was the **Gradient Boosting Machine (GBM)** (Friedman, 2001), compared against variants such as **XGBoost, LightGBM,** and **CatBoost** (Abidi et al., 2021). To address severe class imbalance, **class weighting** was implemented, assigning higher penalties to false negatives, a critical adjustment in audit contexts where missing an anomaly entails higher social costs than a false alarm.

To provide a comparative overview of the main gradient boosting frameworks, (**Table 01**) summarizes their core technical characteristics and typical use cases. This comparison guided the model selection, confirming GBM as the most appropriate algorithm for the dataset's structure and audit-oriented objectives.

Table 1

Comparison of Gradient Boosting Implementations

Characteristic	**GBM (Traditional)**	**XGBoost**	**LightGBM**	**CatBoost**
Training Speed	Moderate	High (parallelized)	Very High (histogram-based)	High

Characteristic	**GBM (Traditional)**	**XGBoost**	**LightGBM**	**CatBoost**
Memory Consumption	High	Moderate	Low	Moderate
Categorical Feature Handling	Requires preprocessing	Requires preprocessing	Native support	Native support
Regularization	Basic	Advanced (L1/L2)	Advanced (L1/L2)	Advanced (ordered boosting)
Recommended Use Case	Balanced datasets	General-purpose ML	Large tabular data	Mixed categorical–numeric data

Source: Adapted from Friedman (2001), Abidi et al. (2021), and Rodríguez et al. (2022).

Hyperparameter optimization was performed through **Grid Search** and **Stratified K-Fold Cross-Validation,** ensuring balanced class proportions across folds and preventing overfitting.

3 RESULTS AND DISCUSSIONS

3.1 DATASET CHARACTERIZATION

The final dataset included **13,550 contract records** from multiple Brazilian federal agencies (2018–2024), with **45 engineered features** and an anomaly rate of **2.36%.**

3.2 MODEL PERFORMANCE

The optimized GBM achieved a **Recall of 0.85, Precision of 0.72,** and an **F1-score of 0.78**, with an **AUPRC of 0.81.** These metrics indicate strong sensitivity to anomalies while maintaining acceptable precision — suitable for a risk-screening system (Carvalho, 2021).

The confusion matrix revealed that of *320 reais* anomalies, **272 were correctly identified,** while **48 remained undetected.** Though the model produced 87 false positives, this trade-off favored proactive detection over omission.

3.3 FEATURE IMPORTANCE

Feature importance analysis confirmed that **contextual and deviation-based variables** dominated predictive power, particularly the **Amendment-to-Original Ratio, Supplier Amendment History**, and **Contract Frequency per Supplier** (Dantas et al., 2024). These insights validate the hypothesis that relational and proportional patterns provide more reliable anomaly indicators than static contract attributes.

Similarly, predictive insights have been observed in toxicological models where multiple biological indicators determine ovicidal and deleterious effects across species (Ferreira de Carvalho et al., 2019).

3.4 OPERATIONAL TRADE-OFFS AND STRATEGIC RISK GOVERNANCE

The integration of predictive modeling into public auditing represents a significant shift from *reactive control* to *preventive governance*. However, this transition is not merely technical — it redefines how institutions conceptualize **risk, accountability, and resource allocation**. As demonstrated, prioritizing **Recall (0.85)** over **Precision (0.72)** maximizes the system's ability to capture real anomalies, even at the cost of investigating some false alarms.

From a governance perspective, this strategy aligns with the principle of **administrative prudence**: it is preferable to audit an additional compliant contract than to overlook a fraudulent one (Grace et al., 2016). The trade-off, therefore, is not purely statistical but **ethical and economic**, balancing efficiency with justice. According to Rodríguez et al. (2022), similar trade-offs have been adopted in anti-fraud systems across the European Union, where the cost of false negatives in public spending often outweighs operational overheads.

The proposed model can thus be interpreted as an **early-warning mechanism**, guiding auditors' focus toward high-risk entities and transactions. This optimization reflects an **intelligent governance approach** (Agostino et al., 2022), where data-driven insights inform human judgment rather than replace it. Future institutional adoption should emphasize **human-in-the-loop frameworks**, ensuring that machine learning outputs support, and never supplant, auditor discretion.

Contrasts traditional auditing methods with the new paradigm of predictive, AI-driven auditing **(Table 02)**. The comparison highlights not only the operational differences but also the ethical and governance implications of adopting intelligent oversight mechanisms in the public sector.

Table 2

Comparative Perspective: Traditional vs. Predictive Auditing Models

Dimension	Traditional Auditing (Reactive)	Predictive Auditing (Proactive, AI-driven)
Timing	Post-event investigation	Real-time anomaly detection
Scope	Limited sample of contracts	Entire public procurement databases

Dimension	**Traditional Auditing (Reactive)**	**Predictive Auditing (Proactive, AI-driven)**
Decision Basis	Human judgment and reports	Data-driven risk prediction (ML models)
Main Limitation	Late identification of irregularities	Possible opacity and false positives
Governance Focus	Accountability after damage	Prevention and continuous monitoring
Ethical Challenge	Bureaucratic delay	Algorithmic transparency and fairness
Supporting Technologies	Manual audits and spreadsheets	Machine Learning, XAI, Blockchain

Source: Adapted from Carvalho (2021); Grace et al. (2016); Moura et al. (2020); Agostino et al. (2022).

3.5 EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI) AND ALGORITHMIC LEGITIMACY

High-performing machine learning models like GBM, XGBoost, and LightGBM are powerful but inherently **opaque**. Their internal logic is difficult for non-specialists to interpret, creating the so-called “black box problem” (Lundberg; Lee, 2017). In the public sector, this opacity challenges fundamental principles of **due process, transparency**, and **citizen accountability** (Carvalho; Cunha, 2020).

Explainable Artificial Intelligence (XAI) seeks to reconcile predictive accuracy with interpretability. By using frameworks such as **SHAP (Shapley Additive Explanations)** and **LIME (Local Interpretable Model-Agnostic Explanations)**, decision-makers can visualize which features most influenced a model’s prediction in specific cases. This level of transparency allows **auditors to trace the reasoning behind algorithmic alerts**, evaluate their plausibility, and question potential biases (Abidi et al., 2021).

The adoption of XAI is not merely a technical enhancement but a **democratic necessity**. As Bryson and Winfield (2017) argue, any system that impacts public resource allocation must offer “meaningful human control. ” In the context of auditing, this means that automated risk assessments must be **comprehensible, contestable, and correctable.** Furthermore, the incorporation of interactive dashboards and XAI visualizations can serve as **training tools**, enhancing auditors’ analytical literacy and promoting organizational learning.

Another essential dimension is **bias detection and fairness**. Models trained on historical procurement data may inadvertently replicate systemic biases — for instance, overflagging certain regions or suppliers. Hence, the use of fairness metrics and **algorithmic auditing protocols** (Raji et al., 2020) should accompany any deployment to avoid reinforcing inequities in the public procurement system.

3.6 BLOCKCHAIN AND THE FUTURE OF IMMUTABLE AUDITING

Blockchain technology offers an innovative framework for **transparent, tamper-proof audit trails.** Each transaction or decision within the auditing process can be recorded as a cryptographically sealed block, ensuring **traceability and immutability** (Moura et al., 2020; Araújo et al., 2021). When applied to public auditing, this could mean that the entire life cycle of an audit, from anomaly detection to corrective action — becomes publicly verifiable, promoting **citizen trust** and **institutional integrity.**

However, the **implementation challenges** are considerable. Public blockchain networks, such as Ethereum or Bitcoin, face scalability constraints and high energy costs. **Private or consortium blockchains**, on the other hand, may reduce transparency by reintroducing centralized control (Alghamdi; Khan, 2023). Policymakers must therefore design governance models that balance **decentralization with accountability**, possibly through **hybrid architectures** combining permissioned blockchains with public verification layers.

A particularly complex dilemma is the **conflict between blockchain immutability and data protection rights.** The Brazilian *Lei Geral de Proteção de Dados (LGPD)*, much like the EU's GDPR, guarantees the *right to be forgotten*. This legal right contradicts blockchain's foundational immutability (Abreu, Pereira; Gomes-Jr, 2024). A potential solution lies in **off-chain storage mechanisms** that separate personal data from immutable records while maintaining audit integrity.

3.7 ETHICAL AND INSTITUTIONAL IMPLICATIONS

Beyond technical and legal challenges, the adoption of AI and blockchain in auditing raises deep **ethical questions**. Public oversight functions are inherently tied to notions of **justice, legitimacy, and equality before the law** (Carvalho, 2021). Delegating parts of this responsibility to algorithms requires **strong institutional safeguards, i**ncluding ethical review boards, algorithmic accountability frameworks, and clear delineations of human responsibility (Bryson; Winfield, 2017).

As in toxicological research, where strict ethical and methodological protocols govern experimental procedures (de Carvalho, 2010), algorithmic auditing also demands robust institutional safeguards to ensure scientific integrity and accountability.

Moreover, institutions must develop **capacity-building strategies**. Many public agencies lack personnel trained in data science or AI governance. Partnerships with universities and technology agencies can bridge this gap, promoting **co-production of knowledge** and **digital transformation** grounded in public values.

Finally, there is an urgent need for **interdisciplinary dialogue** — uniting experts in computer science, public administration, law, and ethics — to construct a coherent framework for algorithmic auditing in Brazil. Such dialogue can ensure that predictive governance not only improves efficiency but also **strengthens democracy and trust.**

4 CONCLUSION

This research proposed and validated a **predictive modeling framework** for the detection of anomalies in public contracts, demonstrating the feasibility of transforming government auditing into a **proactive, data-driven process**. The implementation of a **Gradient Boosting Machine (GBM)** optimized for highly imbalanced data produced robust results; particularly a **recall rate of 0.85**, indicating strong capability to identify potential irregularities.

However, the study also highlights that **technical performance alone is insufficient** for legitimacy in public administration. True innovation in government auditing depends equally on **transparency, interpretability, and ethical governance of algorithms**. The integration of **Explainable AI (XAI)** provides a critical bridge between computational intelligence and human accountability, ensuring that model-driven insights can be understood, justified, and improved upon by auditors and managers.

The exploration of **Blockchain technology** opens additional prospects for immutable, publicly verifiable audit systems. Nevertheless, successful implementation will require overcoming **technical, legal, and organizational barriers**, including challenges of scalability, legal compliance with data protection standards, and institutional adaptation to decentralized paradigms.

This study's limitations include the **use of data restricted to a single administrative level** (federal government) and **offline model validation**, which constrain generalization. To address these, future research should pursue three key directions:

1. **Pilot implementation with integrated XAI dashboards**, enabling real-time interaction between auditors and model-generated explanations;

2. **Unsupervised anomaly detection**, using algorithms such as *Isolation Forest* and *Autoencoders* to uncover unknown patterns of collusion, inefficiency, or fraud;
3. **Techno-regulatory assessment of Blockchain adoption** in Brazil, with interdisciplinary collaboration to design feasible, transparent, and legally compliant audit systems.

Ultimately, this research reinforces that **data science can empower public governance** when accompanied by ethical design, transparency, and a commitment to social accountability, transforming auditing from a reactive instrument into a preventive guardian of public integrity.

REFERENCES

Abidi, W. U. H., et al. (2021). Real-time shill bidding fraud detection empowered with fused machine learning. IEEE Access, 9, 113612–113621.

Abreu, B. M. de, Pereira, T. H. S., & Gomes-Jr, L. (2024). Detecção de fraudes em licitações públicas: Uma comparação de modelos de detecção de anomalias. In Anais da XIX Escola Regional de Banco de Dados (ERBD) (pp. 81–90). Farroupilha, RS, Brazil. https://doi.org/10.5753/erbd.2024.238821

Agostino, D., et al. (2022). Data science in promoting the participation of SMEs in public biddings. ResearchGate Preprint. https://www.researchgate.net/publication/385421403

Aldana, A., Falcón-Cortés, A., & Larralde, H. (2022). A machine learning model to identify corruption in Mexico’s public procurement contracts. arXiv preprint, arXiv:2211.01478.

Alghamdi, R., & Khan, F. (2023). Blockchain as a driver for transformations in the public sector. Information Polity, 28(1), 1–20.

Araújo, V. S., Freitas, M. G., & Martin, M. V. A. (2021). Blockchain e o futuro dos contratos administrativos. Revista Quaestio Iuris, 14(1), 481–503. https://doi.org/10.12957/rqi.2021.48956

Bryson, J. J., & Winfield, A. F. T. (2017). Standardizing ethical design for artificial intelligence and autonomous systems. Computer, 50(5), 116–119. https://doi.org/10.1109/MC.2017.154

Carvalho, G. G. de A., & Cunha, M. A. R. de A. (2020). Potenciais aplicações e consequências do uso da blockchain para a administração pública. Revista de Administração Contemporânea, 24(2), 164–178.

Carvalho, S. S. T. (2021). Impacto da inteligência artificial na atividade de auditoria. Cadernos de Finanças Públicas, 21, 1–25.

Dantas, F. F. C. de A., Cerqueira, A. L. O. de, & Aguiar, R. A. de. (2024). Governança algorítmica e inteligência artificial na auditoria governamental: Desafios e oportunidades do sistema Alice. Repositório Institucional da Enap. https://repositorio.enap.gov.br/handle/1/8764

de Carvalho, G. H. F., et al. (2010). Toxicological effects of ethanolic extract of seed and bark of Persea americana (Lauraceae), on larvae and pupae of Aedes albopictus (Skuse, 1894) (Diptera, Culicidae). Vita et Sanitas, 4(1), 21–33.

de Carvalho, G. H. F., et al. (2011). Atividade inseticida do extrato bruto etanólico de Persea americana (Lauraceae) sobre larvas e pupas de Aedes aegypti (Diptera, Culicidae). Revista de Patologia Tropical/Journal of Tropical Pathology, 40(4), 348–361.

de Carvalho, G. H. F., et al. (2019a). Larvicidal and pupicidal activities of eco-friendly phenolic lipid products from Anacardium occidentale nutshell against arbovirus vectors. Environmental Science and Pollution Research, 26(6), 5514–5523.

de Carvalho, G. H. F., et al. (2019b). Ovicidal and deleterious effects of cashew (Anacardium occidentale) nut shell oil and its fractions on Musca domestica, Chrysomya megacephala, Anticarsia gemmatalis, and Spodoptera frugiperda. Chemistry & Biodiversity, 16(5), e1800468.

de Carvalho, G. H. F., de Medeiros, G. G., & Magalhães, R. de L. B. (2024). Subnotificação de doença de Chagas no Estado do Amapá no período da pandemia de COVID-19. Caderno Pedagógico, 21(9), e7609.

Fernandes, L. S., Silva, J. P. R. da, & Oliveira, L. F. de. (2021). Mineração de dados no Portal da Transparência para análise de licitações. Revista de Sistemas e Computação, 11(2).

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

Grace, E., et al. (2016). Detecting fraud, corruption, and collusion in international development contracts: The design of a proof-of-concept automated system. In IEEE International Conference on Big Data (pp. 1444–1453). Washington, DC: IEEE.

IBM. (2022). What is feature engineering? IBM Knowledge Center. https://www.ibm.com/think/topics/feature-engineering

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4768–4777). Long Beach, CA: NIPS.

Moura, L. M. F., et al. (2020). Blockchain e a perspectiva tecnológica para a administração pública: Uma revisão sistemática. Revista de Administração Contemporânea, 24(3), 259–274.

Nai, R., Al-Boni, M. S., & Bifet, A. (2022). Public procurement fraud detection and artificial intelligence techniques: A literature review. In Workshop on Legal Data Analysis and Mining (LeDAM).

Raji, I. D., et al. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT) (pp. 33–44). Barcelona, Spain: ACM.

Rodríguez, M. J. G., et al. (2022). Collusion detection in public procurement auctions with machine learning algorithms. Automation in Construction, 133, 104055.

Vieira, A. R. M. (2025). Protótipo de um sistema de geração automatizada de pareceres de auditoria baseado em aprendizado de máquina. Repositório Institucional da UFC. https://repositorio.ufc.br/handle/riufc/82469

Williams, P. (2015). Government data does not mean data governance: Lessons from a public sector audit. Government Information Quarterly, 32(3), 324–331.