Detection of Anomalous Proposals in Governmental Bidding Processes: A Machine Learning-Based Approach

Higor R. F. Araújo[1], **Paulo F. Leite**[1], **Joaquim J. C. M. Honório**[1,2], **Isabelle M. L. Souza**[1], **Danyllo W. Albuquerque**[1], **Danilo F. S. Santos**[1]

[1]VIRTUS Research, Development and Innovation Centre
Federal University of Campina Grande (UFCG)
Campina Grande – PB – Brazil

[2]Graduate Program in Computer Science – UFCG
Campina Grande – PB – Brazil

{higor.araujo, paulo.leite, isabelle.lima, danyllo.albuquerque, danilo.santos}@virtus.ufcg.edu.br, joaquimhonorio@copin.ufcg.edu.br

Abstract. *Government procurement involves a formal process wherein government bodies select supplier proposals for goods and services to obtain the best possible terms. This study employs three machine learning algorithms to detect irregularities in the Brazilian government's procurement processes, focusing on data from Paraiba state. The efficacy of these algorithms was evaluated using a controlled dataset that contains known anomalies, assessing their ability to identify deviations. The findings demonstrate the effectiveness of these methods, notably the One-Class SVM, which excels at revealing patterns indicative of possible irregularities in government procurement. In conclusion, this research underscores the potential of machine learning algorithms in enhancing the transparency and integrity of public bidding processes.*

1. Introduction

Worldwide, public administrations engage in the procurement of goods, services, and works from companies, with expenditures estimated at $13 trillion annually [1]. The bidding process is crucial for effectively distributing public resources, aiming to foster a fair and transparent environment that supports accountability and integrity in public administration [Grimsey and Lewis 2002, Hamisi et al. 2022]. However, the traditional oversight mechanisms of these processes often face challenges related to opacity and inefficiency [Gallego et al. 2021]. Anomalous proposals within these processes can indicate various irregularities, such as overpricing, bid rigging, inadequate specifications, or conflicts of interest, which may undermine the principles of competitive bidding and equitable treatment. Such irregularities can distort the market, limiting the effectiveness of public spending and compromising the values of fairness and transparency [Gnoffo 2021].

Therefore, detecting irregularities through the analysis of anomalous proposals, such as patterns of co-occurrence, significant geographical distance between the bidding company and the location of the tender, unusually low bids, and the frequency of wins by a single bidder, can serve as flags for auditors. While each of these anomalies, including

[1]World Bank Joins Multilateral Development Banks to Declare a Shared Commitment to Sustainable Procurement. Accessed on February 15, 2024

co-occurrence, may serve as a potential indicator of concern, they can also be legitimate aspects of the bidding landscape, devoid of any fraudulent intent [Wang and Zhu 2020].

The detection of irregularities within governmental bidding processes has historically been fraught with challenges, primarily due to the limitations inherent in manual monitoring and analysis methods [Lyra et al. 2022]. Traditional approaches rely heavily on scrutinizing bidding patterns and pricing behaviors by human auditors, which is time-consuming and prone to oversight due to the complex and subtle nature of collusive arrangements. Moreover, the vast volume of data generated in public procurement processes further complicates the manual analysis, making it difficult for auditors to identify irregularities and patterns indicative of collusion among many bids and participants. Additionally, manual monitoring methods need more scalability and agility to effectively adapt to the increasing strategies employed by colluders to conceal their activities [Althabatah et al. 2023].

Recent advancements in machine learning (ML) have been instrumental in detecting corruption and irregularities within public procurement processes, utilizing diverse datasets and methodologies across various contexts. Studies like [Sun and Sales 2018] illustrate ML's potential to identify favoritism and improper allocations by leveraging specific contract characteristics and quantitative indicators. Similarly, [Ivanov and Nesterov 2019] and [Asaye et al. 2024] introduced innovative approaches for detecting bid leakage and collusive bidding patterns, emphasizing the importance of specialized ML strategies in uncovering procurement irregularities. The study of [Zhao and Wang 2023] further highlights the significant role of Artificial Neural Networks in predicting public tender irregularities, underscoring the impact of new technologies in enhancing financial audits in a big data environment. Despite significant efforts, there remains a need for more comprehensive research in the literature that provides comparative analyses of variables related to company behavior, including co-occurrence, the proximity of the company to the bid site, the discrepancy between the bid proposal and the estimated value, and solo participation in bids. Additionally, a comparative analysis of unsupervised models is also necessary. This approach would deepen our understanding of collusion within geographical contexts, bridging the gap by examining these variables' nuanced interactions and effects on competitive practices.

Our study addresses the aforementioned limitations, focusing on detecting irregularities by analyzing anomalous proposal submissions. To achieve this, we utilized real bidding data spanning the last ten years (2014 to 2023) from the Brazilian state of Paraiba, applying the Isolation Forest, Local Outlier Factor (LOF), and One-Class SVM unsupervised algorithms due to their capabilities in identifying outliers in large and complex datasets. This approach enables the effective pinpointing of proposals significantly deviating from typical patterns, potentially indicating procurement fraud, including but not limited to price fixing, bid rigging, and other corrupt practices.

This study makes the following contributions:

(i) We highlight the application of machine learning algorithms in detecting anomalies within bidding processes, demonstrating their ability to uncover patterns indicative of potential irregularities, with One-Class SVM specifically tailored for this purpose;
(ii) By applying these algorithms to real bidding data from the Brazilian state of Paraiba over the last ten years, we offer practical insights and methodologies that can be

adapted and scaled for use in various jurisdictions;

(iii) In addition, our study contributes to public management by detecting anomalous proposals, which can serve as indicators for uncovering potential irregularities and fraudulent activities in governmental bidding processes.

To facilitate replication of the contents of this article, all used instruments and procedures are made available according to practices of the Open Science Framework in the external repository [2].

2. Background and Related Work

This section outlines the key concepts necessary for understanding this study (Sections 2.1 and 2.2) and reviews related work to highlight our research's significant contributions and advancements (Section 2.3).

2.1. Problem Definition and Background

In the context of Brazilian public bidding, ensuring the integrity and competitiveness of the process is crucial. However, unethical practices can compromise these principles, necessitating the detection of anomalous tenders that might indicate fraudulent activities.

Anomalous proposals are identified through features that may indicate potential collusive practices or other anti-competitive alliances. The dataset $\mathcal{D}$ is a subset ($\mathcal{D} \subseteq \mathcal{T}$) of a larger database $\mathcal{T}$, chosen randomly to reduce processing costs. It comprises proposals P_i, each belonging to a bidding process $\mathcal{B}_j$, and is represented as a collection of selected proposals from all bidding processes within $\mathcal{D}$:

$$\mathcal{D} = \bigcup_{j=1}^{|\mathcal{B}|} \{P_{j1}, P_{j2}, ..., P_{jn_j}\}$$

where $|\mathcal{B}|$ denotes the total number of bidding processes in the dataset, and n_j is the number of proposals within the bidding process $\mathcal{B}_j$.

Each proposal P_i in $\mathcal{D}$ is described by a feature vector $\mathbf{x}_i$, encapsulating aspects such as distance from company to bidding location d_{Ci}, co-occurrence among companies ($Cooc$), the difference between the bid value and the bidder's proposal t_{Ci}, time of existence of the company (e_{Ci}), and the single company participates in a bidding process ($Solo_{Ci}$).

The distance from a company to a bidding location, d_{Ci}, is mathematically defined as: $d_{Ci} = \sqrt{(x_{Bi} - x_{Pi})^2 + (y_{Bi} - y_{Pi})^2}$ with (x_{Bi}, y_{Bi}) and (x_{Pi}, y_{Pi}) representing the geographical coordinates of the company and the bidding location, respectively.

The co-occurrence is defined based on the repeated participation of a subset of companies $\mathcal{C}_{sub}$, where $|\mathcal{C}_{sub}| \geq 2$, in multiple proposals. This cumulative co-occurrence is calculated by summing the number of proposals in which these subsets of companies participate together. Formally, the adjusted definition is expressed as:

$$Cooc_{Ci} = \sum_{\{P_{i1}, P_{i2}, ..., P_{im}\} \subseteq \mathcal{D}} (\exists\, \mathcal{C}_{sub} \subseteq \mathcal{C}(P_{ik}), |\mathcal{C}_{sub}| \geq 2, \forall k \in \{1, 2, ..., m\}, m > 1)$$

[2]https://encurtador.com.br/tzBIN

The difference between the bid value and the bidder's proposal, t_{Ci}, is calculated as the absolute difference between the bid value submitted by a company and the estimated value of the bidding: $t_{Ci} = |v_{Bi} - v_{Pi}|$ where v_{Bi} is the bid value submitted by the company for proposal i and v_{Pi} is the estimated value of the proposal.

The time of existence of the company, e_{Ci}, is determined by the difference between the year of the bidding and the year the company was founded: $e_{Ci} = y_{Bidding} - y_{Foundation}$ where $y_{Bidding}$ is the year of the bidding and $y_{Foundation}$ is the year the company was established.

In exploring anomalous behaviors within competitive bidding processes, a noteworthy feature emerges to capture instances wherein a single company participates in a bidding process without competition from other entities. This phenomenon is encapsulated through the feature $Solo_{Ci}$, which delineates the solitary participation of a company in a given bidding process. Formally, the feature $Solo_{Ci}$ is defined for each proposal P_i as follows:

$$Solo_{Ci} = \begin{cases} 1 & \text{if } |\mathcal{C}(P_j)| = 1 \\ 0 & \text{otherwise} \end{cases}$$

where $|\mathcal{C}(P_j)|$ denotes the number of companies participating in the bidding process $\mathcal{B}_j$ associated with proposal P_i. A value of 1 indicates that the company has participated alone in the bidding process, highlighting a lack of competitive bids, whereas a value of 0 denotes the presence of one or more competing proposals from different companies within the same process.

The models are not intended to replace manual validation efforts but to be used in conjunction to enhance the detection of irregularities in public tenders. Model utilization, denoted by $\sum_{i=1}^{|\mathcal{D}|} f(\mathbf{x}_i)$, where f is the algorithmic function and $\mathbf{x}_i$ is the feature vector for proposal i, serves to complement manual verification, enhancing the detection of irregularities in public tender processes. In this expression, $|\mathcal{D}|$ is the number of proposals in the dataset, and the summation iteratively applies the function f across all feature vectors, symbolizing the aggregate analysis conducted by the models in conjunction with manual efforts.

2.2. Bidding Process and Traditional Detection of Irregularities

The bidding process in Brazil is a structured framework established to ensure transparency, competitiveness, and fairness in public entities' procurement of goods and services. This process is governed by Federal, also known as the Bidding Law, which outlines the procedures that federal, state, and municipal governments must follow to conduct their purchases and contracts [3]. The law mandates using public tenders to select the most advantageous proposal, which is not always the lowest price but offers the best cost-benefit ratio considering quality, technical performance, and other relevant factors.

The Brazilian bidding process begins with the public entity's need to acquire goods or services. This need leads to preparing a bidding notice specifying the procurement details, including the requirements that bidders must meet and the selection criteria. The

[3] In Brazil, law of Biddings and Administrative Contracts no. 14,133/2021

process then moves to the publication of this notice, allowing potential bidders to participate. The bidding procedures can vary, including competitive bidding, price taking, invitation, contest, and auction, each suitable for different circumstances based on values, objects, and complexity.

To illustrate the bidding process steps and machine learning application for detecting anomalous tenders, which may indicate potential irregularities such as overpricing, fraud, or non-compliance with procurement laws, the following diagram is presented in Figure 1.

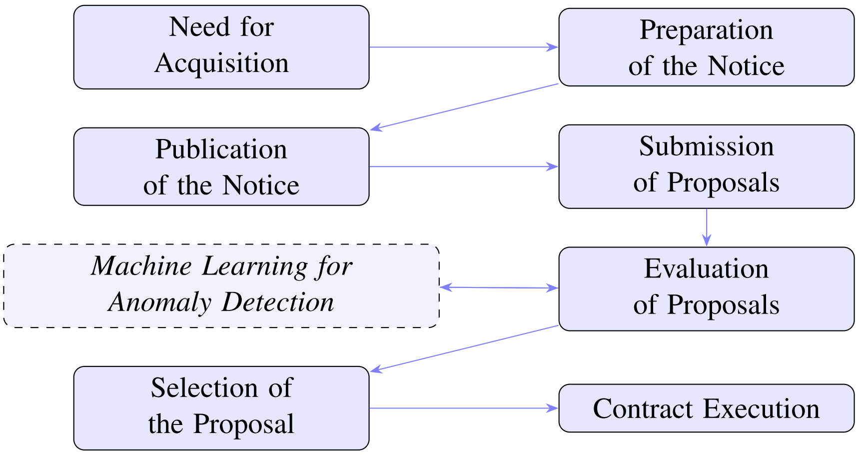

Figure 1. Integrative diagram of the bidding process highlighting traditional stages and the insertion point of ML algorithms for anomaly detection.

Traditional detection of irregularities in this process has heavily depended on manual audits and supervision by internal and external control bodies, such as the Union Court of Auditors (TCU) at the federal level and similar courts at the state level, like the State Court of Auditors of Paraiba. These bodies are responsible for examining the legality, legitimacy, and efficiency of the bidding processes and contracts. The audit process involves reviewing documentation, verifying compliance with legislation, and assessing the economic feasibility of the acquisition. Inspections and investigations are carried out to detect overpricing, fraud, and other irregularities that may occur during the bidding process or the execution of contracts.

However, despite these traditional mechanisms for detecting irregularities, challenges persist due to the complexity of the processes, the volume of transactions, and sometimes the methods used to commit fraud [Adobor and Yawson 2023].

2.3. Related Work

Various studies have explored approaches, algorithms, and datasets to identify irregularities in public procurement and contracts. A systematic review highlighted the evolution of the number of publications in this area, indicating a growing interest and a diversity of scientific approaches to combat fraud, corruption, and collusion in public tendering processes. This research identified 48 relevant articles, of which 36 were published in scientific journals and 13 in conference proceedings, covering a wide range of domains, from computer science to public administration and law [Lyra et al. 2022].

Recent studies highlight the application of Machine Learning (ML) in detecting corruption and irregularities in public procurement processes across different contexts, employing diverse datasets and techniques. [Sun and Sales 2018] discussed the potential of ML, especially Random Forest algorithms, in identifying improperly allocated educational tenders due to corruption or political influences, demonstrating the efficacy of combining quantitative indicators with ML. [Rodríguez et al. 2022] examined various ML algorithms to identify collusive auctions in public procurement, showing that ML can effectively detect illegal agreements among firms in datasets from five countries.

Studies demonstrate advancements in machine learning and statistical methods to detect irregular procurement processes and bid-rigging cartels across different jurisdictions and sectors. [Huber et al. 2022] explore the transnational application of statistical screens combined with machine learning to identify bid-rigging cartels, achieving high classification accuracy within the same country but facing challenges when applying models across different countries due to institutional differences. [Ivanov and Nesterov 2019] introduce a machine-learning approach to detect bid leakage in Russian procurement auctions by employing a Positive-Unlabeled Classification strategy, estimating a significant presence of bid leakage. [Asaye et al. 2024] study proposes a framework using pattern mining techniques to detect collusive bidding patterns in highway project procurements, addressing the lack of effective models for identifying red-flag bidding behaviors. Lastly, research by [Zhao and Wang 2023] highlights the use of Artificial Neural Networks to predict irregularities in public tenders, emphasizing the role of new technologies in enhancing financial audits within a big data context.

However, despite significant advances, there is a gap in the literature regarding the use of data from specific localities and their specificities. Many studies focus on data from specific countries or sectors, allowing exploration in other regions or contexts. Moreover, while some ML techniques have been widely explored, others still need more investigation to assess their effectiveness in public procurement and contracts.

3. Methodology

This study adopts an experimental approach to explore potential irregularities in government tender processes by detecting anomalous tenders. Recognizing the importance of research ethics, we ensured data anonymization, concealing names and identifiable information of the entities involved to protect their privacy and confidentiality.

3.1. Data Collection and Sources

Our research strategy involves analyzing tender data from the state of Paraiba, Brazil, collected over a span of ten years (2014-2023). These data were sourced from the Court of Accounts of Paraiba (TCE-PB) Portal [4], which provides access to records of public tender processes, including tender notices, bidding outcomes, contract values, and supplier details. To enhance our analysis and improve the accuracy of our findings, we also integrated data from Brazil's Federal Revenue (Receita Federal)[5] and the Brazilian Institute of Geography and Statistics (IBGE) [6].

[4]TCE-PB (Accessed February 18, 2024): https://tce.pb.gov.br/

[5]Receita Federal (Accessed February 18, 2024): https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/dados-abertos

[6]IBGE (Accessed February 18, 2024): https://www.ibge.gov.br/acesso-informacao/dados-abertos.html

The database from the Federal Revenue offers additional insights into the companies participating in these tenders, including their location and date of establishment. With this information, it is possible to assess the lifespan of a company involved in the tender process and its municipality of origin.

We utilized data from the IBGE, which provides latitude and longitude coordinates for each municipality in Brazil. These geographical details allow, beyond the spatial analysis of companies, the execution of calculations about the Euclidean distance between the tender location and the contracted company.

Finally, in this study, a subset of 626 proposals (where 36% of data was manually classified as anomalies) was selectively utilized due to the requirements of labeling the predictive class and to mitigate processing costs. This refined dataset ensured a focus on the most relevant instances by excluding empty fields, outlier or erroneous values, and entries about state purchases and individual persons. A detailed proposal data analysis can be found in an external repository[7].

3.2. Feature Engineering

The Feature Engineering process was guided by the need to capture the characteristics of proposals that may indicate anomalous behavior, as elaborated in the Problem Definition and Background section (2.1). Based on the dataset $\mathcal{D}$, each proposal P_i is represented by a feature vector $\mathbf{x}_i$, comprising various aspects for identifying potentially anomalous proposals. These features aim to encapsulate both the spatial aspects of the bids, the relational dynamics among bidders, financial discrepancies, and historical performance metrics that might collectively signal irregularities.

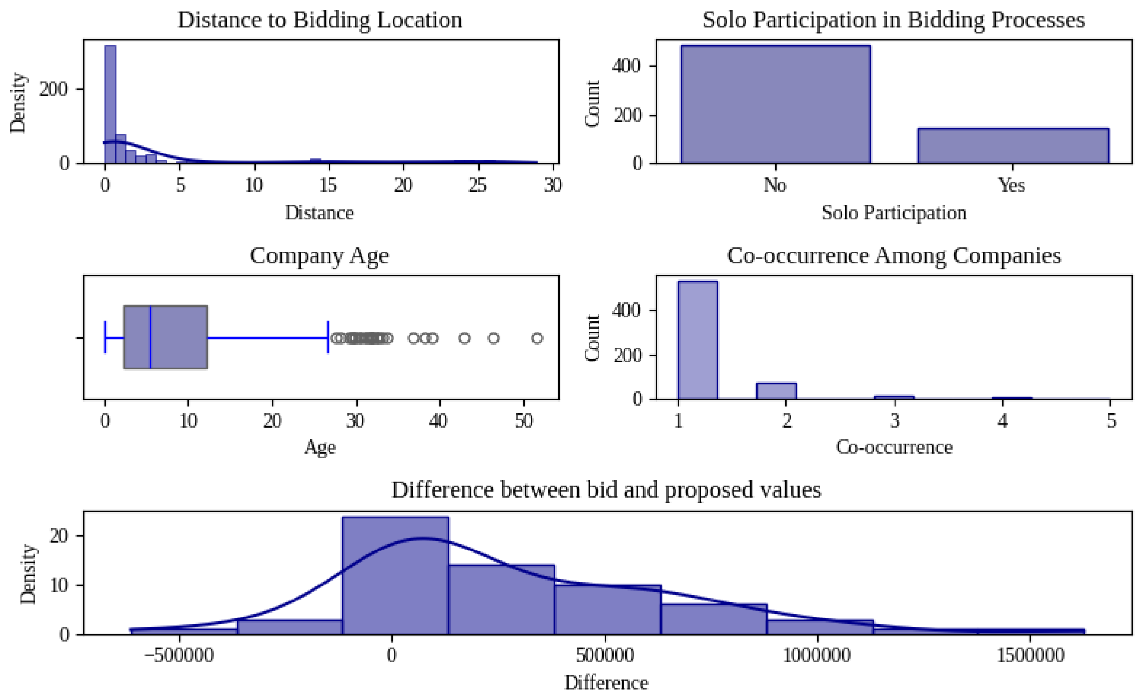

Figure 2. Overview of bidding characteristics in procurement processes.

[7]Project in the Open Science Framework: https://encurtador.com.br/tzBIN

Figure 2 presents a compilation of five graphs, each depicting a characteristic related to a dataset on bidding processes. The histogram detailing the distance to the bidding location (d_{Ci}) suggests a skewed distribution, with most companies situated a short distance from the bidding site. The bar graph concerning solo participation in bidding processes ($Solo_{Ci}$) exhibits a significant disparity between the number of companies that have participated alone versus those that have not, with solo participation being less common. The boxplot for company age (e_{Ci}) indicates a broad range of ages among the companies, with some outliers significantly older than the median. The bar graph displaying the co-occurrence among companies ($Cooc_{Ci}$) illustrates the frequency at which companies bid together, with the majority having a low rate of co-occurrence. Finally, the density plot for the difference between the bid and proposed values (t_{Ci}) shows a distribution centered around zero, suggesting that bid and proposed values are often close, with notable exceptions where the difference is considerable.

3.3. Model Training and Validation Metrics

In anomaly detection within a specific dataset, we employed three machine learning algorithms: Isolation Forest, Local Outlier Factor, and One-Class Support Vector Machine. Each model was selected for its unique approach to identifying outliers or anomalies in the data. Anomaly labels were generated through a manual evaluation process by assessing the extreme values of each feature vector. The models were trained on a subset of the data (70% training split, 30% test following the temporal order to avoid information leakage), ensuring that each model was evaluated on unseen data to validate its generalization capability.

- The **Isolation Forest** technique identifies outliers by isolating observations [Liu et al. 2008, Li et al. 2021]. It uses a forest of random trees, where each tree is grown by randomly selecting a feature and a split value. This method inherently assumes that anomalies are easier to isolate than normal points. We used 100 trees for our application, setting the contamination level based on the dataset's expected outlier proportion, thereby effectively guiding the algorithm to distinguish anomalies.
- The **Local Outlier Factor** quantifies how much an observation deviates in density from its neighbors, considering the idea that normal data points occur in dense neighborhoods, whereas anomalies are located in sparser regions [Alghushairy et al. 2020]. In this case, the algorithm was configured to examine the 20 nearest neighbors to determine the density discrepancy, using the dataset's anomaly ratio to adjust its sensitivity towards identifying outliers, ensuring a balanced detection approach across varied data densities.
- The **One-Class SVM** [Qiao et al. 2021] approach is tailored for anomaly detection by establishing a boundary around data points considered "normal". It employs a radial basis function kernel to shape this boundary, with the 'auto' setting for gamma and the nu parameter mirroring the dataset's anomaly fraction. This setup helps mitigate overfitting, particularly in imbalanced datasets, by focusing the model on capturing the majority class's pattern while allowing for some flexibility in defining what constitutes an outlier. The data were normalized using the StandardScaler to ensure better performance and mitigate the risk of overfitting.

To assess the performance of each model, we used the ROC-AUC score and the confusion matrix as our primary metrics. The ROC-AUC score, derived from the Receiver

Operating Characteristic (ROC) Curve, quantifies a model's ability to distinguish between the anomaly and normal classes as the decision threshold varies. Mathematically, the area under the ROC curve (AUC) represents the probability that a model will correctly classify a random sample from the positive class (anomaly) as more likely to be positive than a random sample from the negative class (normal) across all possible decision thresholds. This metric ranges from 0 to 1, where 0.5 indicates no better performance than chance, and 1 indicates a perfect discrimination capability between classes.

4. Results and Discussion

This study investigated the effectiveness of three anomaly detection algorithms: Isolation Forest, Local Outlier Factor, and One-Class SVM. The evaluation was based on the ROC-AUC metric, focusing on their capability to distinguish between normal and anomalous instances.

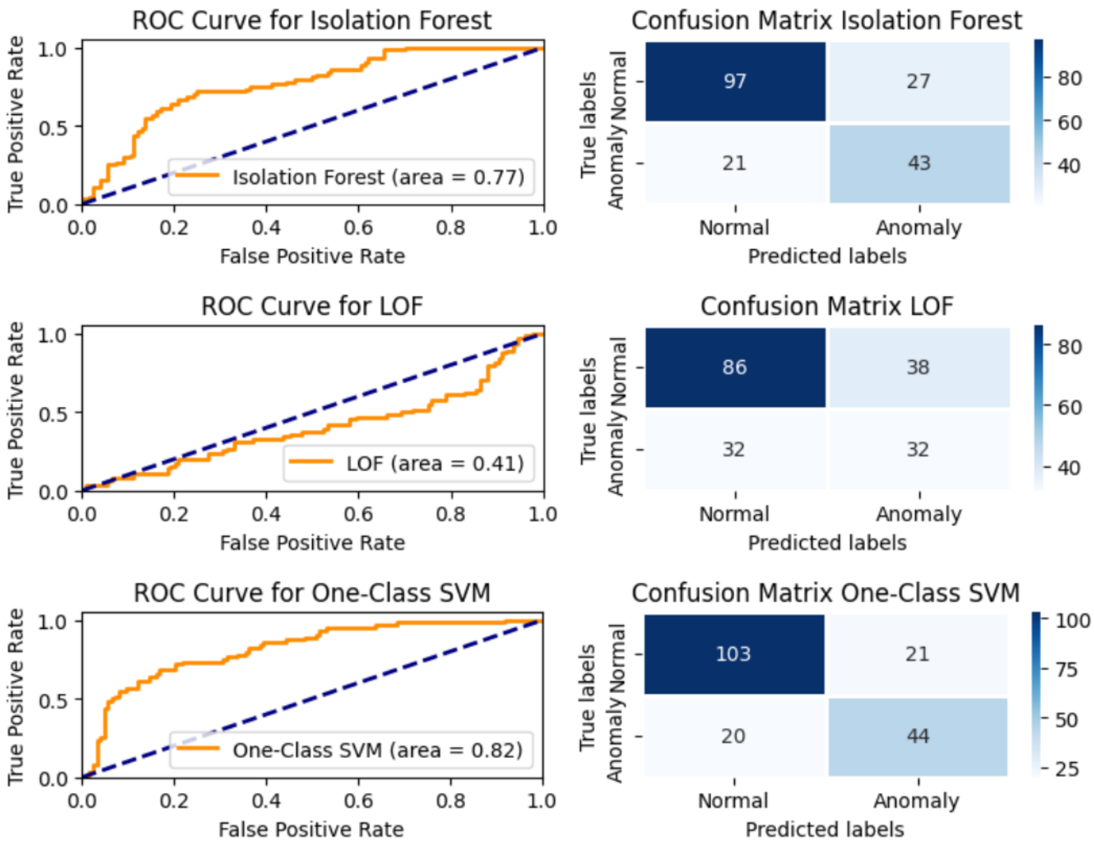

Figure 3. Performance comparison of anomaly detection models.

Figure 3 summarizes the performance comparison of the anomaly detection models. The Isolation Forest algorithm demonstrated a ROC-AUC score of 0.769, indicating a strong ability to differentiate between the two classes. The confusion matrix for this model showed that it correctly identified 97 normal instances and 43 anomalous instances, while it misclassified 27 normal instances as anomalous and 21 anomalous instances as normal. This performance suggests a solid foundation but highlights areas for improvement in minimizing false positives and false negatives. The LOF (Local Outlier Factor) algorithm achieved an ROC-AUC score of 0.406, reflecting a limited capability to distinguish between normal and anomalous instances. According to the confusion matrix, it accurately recognized 86 normal instances and 32 anomalous instances. However, it also incorrectly labeled 38 normal instances as anomalous and 32 anomalous instances as normal, indicating significant challenges in achieving accurate classification. The One-Class SVM model exhibited the highest performance among the three algorithms, with a ROC-AUC score of 0.821. This suggests it has good efficiency in detecting anomalies. The confusion matrix for One-Class SVM indicates that it successfully identified 103 normal instances and 44 anomalous instances, with fewer instances of false positives (21) and false negatives (20) compared to the Isolation Forest and LOF algorithms. These results underscore the effectiveness of the One-Class SVM in managing imbalanced datasets and its capability to identify anomalies.

Finally, the permutation-based approach observed that the variables' difference in bid value and distance from the company to the bidding location play the most significant roles in influencing the model's outcome. This analysis highlights the impact of financial competitiveness and geographical proximity on the predictive model's performance.

5. Threats to Validity

The limitations of this study are categorized into four areas: construct validity, internal validity, external validity, and conclusion validity. Each category presents specific challenges and actions to mitigate these threats.

Construct Validity: The primary threat to construct validity lies in accurately operationalizing anomalies in government procurement. Misclassification of normal activities as anomalies or overlooking subtle irregularities can skew the analysis. To address this, the study employed a comprehensive feature engineering process incorporating various data aspects to represent the procurement activities accurately. Expert validation was also used to ensure the features adequately captured the nuances of procurement irregularities.

Internal Validity: The threat to internal validity includes potential biases in the machine learning algorithms that could affect their performance detecting irregularities. The study mitigated these threats by utilizing diverse algorithms to compare their effectiveness and employing a systematic approach to parameter tuning and model validation to reduce overfitting and bias.

External Validity: External validity threats arise from the generalizability of the findings. The study's focus on Paraiba's procurement data may limit the applicability of the results to other regions or countries with different procurement practices and legal frameworks. To enhance generalizability, the study suggests that future research should apply the methodology to a broader range of geographical locations and procurement systems and compare the findings to identify common patterns and anomalies.

Conclusion Validity: The risk to conclusion validity involves the reliability of the findings, particularly the risk of drawing incorrect conclusions about the effectiveness of the machine learning algorithms in detecting procurement irregularities. The study ensured rigorous statistical analysis and validation of the machine learning models to strengthen conclusion validity. Additionally, it adopted a multi-model approach to cross-validate the findings and highlight consistent patterns across different algorithms.

6. Final Remarks

This study scrutinizes the identification of irregularities in government procurement processes, analyzing anomalous tenders within the procurement data of Paraiba, Brazil. It integrates data from the Brazilian Federal Revenue and the IBGE to conduct a thorough analysis, enabling a comprehensive approach to anomaly detection. This approach considers both the proposals' financial details and the companies' geographical distribution, aiding in identifying potential fraud and collusion.

Methodologically, the research employs three machine learning algorithms: Isolation Forest, Local Outlier Factor, and One-Class SVM, each known for its effectiveness in detecting data outliers within complex datasets. These algorithms were rigorously tested and optimized to identify significant deviations from established bidding patterns, indicating potential procurement anomalies. The methodology encompassed a data preprocessing phase for cleaning, normalizing, and integrating the data into a cohesive framework, followed by the algorithm application and performance evaluation against a controlled dataset.

The study's results particularly highlight the One-Class SVM's efficiency in identifying potential irregularities in procurement processes, showcasing its adaptability to the specificities of the Brazilian context. Academically, the research contributes to the literature on machine learning applications in public procurement and anomaly detection, underlining the importance of integrating diverse data sources for a holistic analysis. Practically, it offers valuable insights for public managers on how advanced data analysis can bolster the integrity and efficiency of procurement processes.

For future work, the study aims to expand its analytical scope by incorporating tender data from other Brazilian states and assessing new model features to detect geographical and operational patterns of irregularities. This expanded analysis intends to unveil regional disparity in procurement irregularities, enhancing national-level understanding. Additionally, the research plans include a comparative study on different machine learning algorithms' effectiveness in anomaly detection and engaging with stakeholders through interviews and pilot projects to ascertain these technologies' practical applicability and benefits in the tendering processes. Efforts will also be directed towards educating stakeholders with limited AI knowledge, employing Large Language Models to elucidate anomaly detection's intricacies in public procurement.

References

Adobor, H. and Yawson, R. (2023). The promise of artificial intelligence in combating public corruption in the emerging economies: A conceptual framework. *Science and Public Policy*, 50(3):355–370.

Alghushairy, O., Alsini, R., Soule, T., and Ma, X. (2020). A review of local outlier factor algorithms for outlier detection in big data streams. *Big Data and Cognitive Computing*, 5(1):1.

Althabatah, A., Yaqot, M., Menezes, B., and Kerbache, L. (2023). Transformative procurement trends: Integrating industry 4.0 technologies for enhanced procurement processes. *Logistics*, 7(3):63.

Asaye, L., Moriyani, M. A., Le, C., and Le, T. (2024). Detecting red-flag bidding patterns in low-bid procurement for highway projects with pattern mining. *Journal of Management in Engineering*, 40(1):04023060.

Gallego, J., Rivero, G., and Martínez, J. (2021). Preventing rather than punishing: An early warning model of malfeasance in public procurement. *International Journal of Forecasting*, 37(1):360–377.

Gnoffo, S. (2021). A dynamic performance management approach to frame corruption in public procurement: a case study. *Journal of Public Procurement*, 21(1):75–96.

Grimsey, D. and Lewis, M. K. (2002). Evaluating the risks of public private partnerships for infrastructure projects. *International Journal of Project Management*, 20(2):107–118.

Hamisi, S., Philibert, N., and Paul, N. (2022). Transaction costs and competitive tendering in public procurement: Moderating role of integrity. *African Journal of Business Management*, 16(6):130–146.

Huber, M., Imhof, D., and Ishii, R. (2022). Transnational machine learning with screens for flagging bid-rigging cartels. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(3):1074–1114.

Ivanov, D. and Nesterov, A. (2019). Identifying bid leakage in procurement auctions: Machine learning approach. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 69–70.

Li, C., Guo, L., Gao, H., and Li, Y. (2021). Similarity-measured isolation forest: Anomaly detection method for machine monitoring data. *IEEE Transactions on Instrumentation and Measurement*, 70:1–12.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.

Lyra, M. S., Damásio, B., Pinheiro, F. L., and Bacao, F. (2022). Fraud, corruption, and collusion in public procurement activities, a systematic literature review on data-driven methods. *Applied Network Science*, 7(1):83.

Qiao, Y., Wu, K., and Jin, P. (2021). Efficient anomaly detection for high-dimensional sensing data with one-class support vector machine. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):404–417.

Rodríguez, M. J. G., Rodríguez-Montequín, V., Ballesteros-Pérez, P., Love, P. E., and Signor, R. (2022). Collusion detection in public procurement auctions with machine learning algorithms. *Automation in Construction*, 133:104047.

Sun, T. and Sales, L. J. (2018). Predicting public procurement irregularity: An application of neural networks. *Journal of Emerging Technologies in Accounting*, 15(1):141–154.

Wang, C. and Zhu, H. (2020). Representing fine-grained co-occurrences for behavior-based fraud detection in online payment services. *IEEE Transactions on Dependable and Secure Computing*, 19(1):301–315.

Zhao, H. and Wang, Y. (2023). A big data-driven financial auditing method using convolution neural network. *IEEE Access*, 11(11):41492–41502.