

Financial Data Anomaly Detection Method Based on Residual Explanation for Government Office Operational Expenditure

¹Afrizal Nehemia Toscani

¹Information System, Faculty of Computer Science, Universitas Dinamika Bangsa
Jambi, Indonesia

afrizalnehemia@gmail.com

²Fachruddin

²Master of Information System, Faculty of Computer Science, Universitas Dinamika
Bangsa

Jambi, Indonesia

fachruddin.stikom@gmail.com

Abstract— Ensuring fiscal responsibility and transparency in government offices requires careful control of operating expenditures. However, due to the intricacy and regularity of financial transactions, this work can be difficult, making standard monitoring techniques inefficient and time-consuming. This paper introduces an anomaly detection method for financial data in government office operational expenditures, termed "Residual Explanation". This approach uses sophisticated machine learning techniques to discover anomalous transaction patterns that can point to fraud by analyzing residuals, or the disparities between observed and predicted transaction values. Our method makes use of a Random Forest Regressor model, which is especially well-suited to managing the high dimensionality and non-linear correlations seen in financial datasets. This study focuses on anomaly detection within operational expenses of account 525112 at one of institutions in Indonesia. The results indicate that the optimal model for detecting anomalies operates with a thresholds value of 99,6%. Future improvements to this model could allow its integration of our approach into existing financial systems could enable real-time anomaly detection, which is paramount for preventing fraud and enhancing the financial governance of government expenditures.

Keywords— Financial Data, Anomaly Detection, Residual Explanation, Operational Expenditure.

I. INTRODUCTION

Government offices worldwide manage substantial operational expenditures to ensure the smooth functioning of public services. Detecting anomalies in financial data within these expenditures is crucial for maintaining transparency and accountability. An effective anomaly detection method can help identify irregularities, potential fraud, or inefficiencies in spending patterns. This paper proposes a Financial Data Anomaly Detection Method Based on Residual Explanation specifically tailored for government

office operational expenditure analysis.

To develop this method, insights from various studies on government expenditure, financial management, and economic growth are synthesized. The research by (Kravchuk & Stone, 2010) highlights the predictive value of cash receipts and balances in forecasting fiscal imbalances, which can be instrumental in anomaly detection. Additionally, the work of (Hartanto et al., 2020) emphasizes the factors influencing end-of-year spending behavior in local governments, providing a behavioral perspective that can aid in anomaly identification. Moreover, studies such as that by (Agu et al., 2024) on the effect of government expenditure on human capital index and (Abdullah et al., 2020) on the incentive function of audit opinion shed light on the relationship between expenditure, performance, and oversight mechanisms, which are essential considerations in anomaly detection. Furthermore, the research by (Lang, 2022) on administrative cost data analysis using data mining techniques offers a methodological approach that can be adapted for anomaly detection in operational expenditure.

Anomaly detection is a critical process in various domains, including financial data analysis. It involves identifying data events or observations that significantly deviate from the expected norm (Xie et al., 2023). In the context of financial data, anomalies can indicate irregularities such as fraudulent activities, errors, or unusual patterns that necessitate further investigation (Ahmed & Mahmood, 2015). Detecting anomalies in financial data is crucial for maintaining the integrity and reliability of financial systems, as anomalies can impact decision making processes and financial stability (Davidow & Matteson, 2022). Numerous techniques and methodologies have been developed for anomaly detection in financial data. Machine learning algorithms are widely used in financial market surveillance to detect abnormal price changes and unusual trading behaviors. These algorithms utilize regression tree based ensemble methods to predict future prices and effectively identify anomalies (Tiwari et al., 2021). Additionally, unsupervised anomaly detection algorithms are employed to detect unexpected items or events in multivariate financial datasets (Goldstein & Uchida, 2016).

Anomaly detection in financial data extends beyond fraud detection to encompass risk management and investment decision making. Detecting anomalies in financial statements using machine learning algorithms has been proven effective in identifying unusual financial behaviors and mitigating risks (Lokanan et al., 2019). Moreover, anomaly detection techniques are utilized in forensic accounting to identify unusual corporate financial behaviors and potential financial crimes (Goh et al., 2021).

By integrating findings from these diverse studies, this paper aims to propose a robust anomaly detection method that leverages residual explanations to uncover irregularities in government office operational expenditure. The method will draw on insights from revenue prediction, spending behavior determinants, human capital impact, audit influence, and data mining analysis to enhance anomaly detection accuracy and effectiveness.

II. LITERATURE REVIEW

Anomaly Detection

Anomaly detection is a crucial area of research spanning various domains such as computer science, artificial intelligence, statistics, hyperspectral imaging, and more. The extensive body of work on anomaly detection has led to the development of numerous algorithms and methodologies aimed at identifying deviations from expected patterns. Researchers have explored innovative approaches to enhance anomaly detection capabilities, such as the integration of advanced models like Auto-Regressive Integrated Moving Average (ARIMA) for online forecasting and anomaly detection (Kozitsin et al., 2021). By leveraging ARIMA models, researchers have devised computationally efficient algorithms to address anomaly detection and forecasting challenges effectively.

A two-dimensional graph that illustrates the concept of various anomalies is shown in Figure 1. The data components in this image constitute two normal zones, N_1 and N_2 , which represent the locations of the majority of the occurrences. Observations that deviate more from the bulk of the other observations, either alone or in small groupings, are called anomalies., such as the region O_3 , points O_1 , and O_2 . Although they are linked, anomaly detection and noise removal—which deal with unwanted noise in data—are not the same.

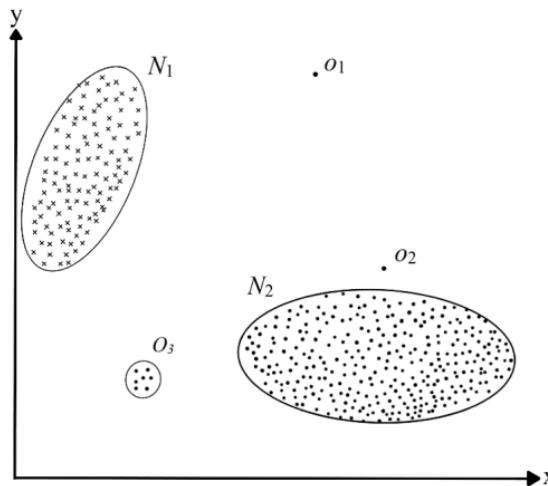


Figure 1. Abnormal data graphically represented in a basic two-dimensional diagram.

Anomalies can arise from a number of causes, including fraud, human mistake, malfunctioning instruments, population variances naturally, behavioral shifts in systems, or flaws inside systems. (Hodge & Austin, 2004). The application area determines how the anomaly detection system handles the anomaly. Let's say, for instance, that the anomaly points to a data entry clerk's typo. In that instance, merely alerting the clerk to the mistake will assist in getting the anomalous data back into a regular entry. Once found, anomalous data from instrument measurements can be easily removed. Furthermore, anomaly detection systems need to be able to identify abnormalities quickly and in real-time in sensitive contexts like intrusion monitoring or fraud detection systems. They also need to be able to provide an appropriate alarm to enable intervention. (Hodge & Austin, 2004)

Financial Fraud

The possible repercussions of undiscovered abnormalities on the sector and on daily life have led to a significant increase in awareness of financial fraud throughout the last 10 years. Due to the billions of dollars in losses that businesses have suffered as a result, there have been significant efforts made to investigate anomaly detection methods for the identification of fraud. Fraud must always be identified as soon as possible since incidents that go unreported cost money over time. Since fraudsters are always changing and inventing new ways to circumvent them, it is necessary to constantly develop and enhance these fraud detection tools (West & Bhattacharya, 2016). The US had 467,361 complaints of cybercrimes, with total damages exceeding \$3.5 billion, according to the cyber Crime Complaint Center (IC3). This represents a roughly 30% increase from the previous year. (Center, 2019). These losses included a variety of crimes, including identity fraud involving fake checks and credit card applications, hacking corporate emails to carry out illegal money transfers, and much more. According to research reported in the ACFE's 2020 Report to the Nations, fraud costs organizations an average of \$8,300 per month and results in revenue losses of 5% worldwide each year. A typical fraud case remains unreported for 14 months. (Association of Certified Fraud Examiners, 2020).

An audit report by the Audit Board of Indonesia (BPK) on the 2014 financial statements of a regional government revealed indications of fraud in the procurement of office supplies in 31 out of 33 local government units within the district. In the context of the central government, (Anggraeni, 2018) discovered misstatements in inventory balances at a public service agency (BLU) of a university due to staff inconsistencies in entering the price/value per unit for the same inventory items across different transactions.

Random Forest Algorithm

According to our model, random forest is a sizable and skilled team of analysts that can render more accurate decisions when compared to a decision tree and one experienced professional analyst. This article's decision tree algorithm is comparable to the CART algorithm, and the Gini index is used to determine which characteristics should be separated (Zhang, 2022). An impurity splitting technique is the Gini index. It can be used in binary, continuous value, category, and other fields. The idea behind the particular algorithm is this: if records of k categories are present in the data sample set r at a specific node t , then the Gini indicators are

$$\text{Gini}(t) = 1 - \sum_{j=1}^k [p(j|t)]^2, \quad (1)$$

where the probability from category j to node t is represented by p . When the category field is uniformly distributed, $\text{Gini}(t)$ is the biggest and the usable information at this time is the least; when the minimal Gini is 0, all samples at this node belong to the same category, indicating that the maximum amount of useful information may be retrieved. In the event where the set is divided into l pieces, the division's Gini index is

$$\text{Gini}(T) = \sum_{i=1}^l \left(\frac{n_i}{n} \right) \text{Gini}(i). \quad (2)$$

Following the formation of the forest, each tree determines a corresponding classification conclusion for a fresh sample. The classification result is then determined by a simple majority vote among all the trees. When compared to alternative combination classification methods, the quantity of

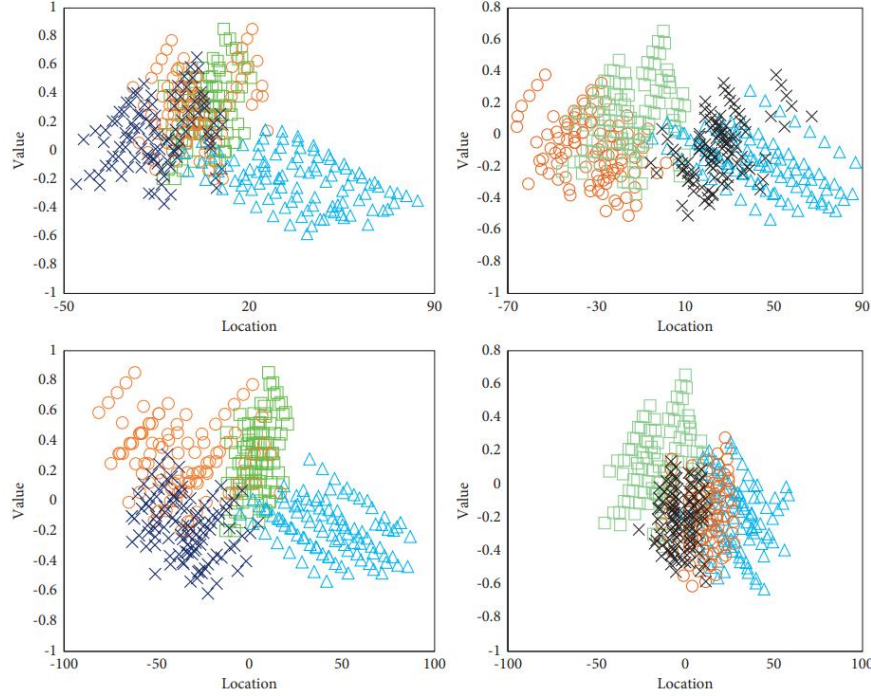


Figure 2. Predicted Value

trees are fairly big; overfitting is not a problem with random forests. It can be demonstrated that the generalization error's upper bound is less than

$$\frac{\bar{\rho}(1 - s^2)}{s^2}, \quad (3)$$

where s is the classification performance of a single tree and ρ is the average correlation coefficient amongst trees (indicating the degree of correlation between the classification trees). Before anything further, create a $N \times N$ zero-element matrix Prox for the training set containing the number of samples IV . Once all the hydrazine trees have been constructed, add I to the corresponding n th row and k th column of the Prox matrix for each pair of samples, n and k , if they show up at the node on the same leaf of the tree. To discriminate between all samples, use the tree that is formed each time. Each sample will reach a different leaf node of the tree. Divide the total number of trees by each element in the matrix. The final Prox matrix is a symmetric matrix with diagonal elements of I , which is obtained by normalizing the processing. The element $\text{Prox}(n, k)$ in the n th row and k th column can be defined as the sample of samples n and k . There are more grounds for us to think that they differ from other samples. Consequently, it always results in a scale of aberrant points that represents the sample's degree of abnormality. Define the initial anomalous point scale of sample n as

$$\text{rawom}(n) = \frac{n_{\text{sample}}}{\bar{P}(n)}. \quad (4)$$

A sample's rawom value will be high in the same class if its p(n) value is low. Compute the mean and variance of the initial outlier scales for every sample in each class to obtain the final outlier scale for every sample following standardization:

$$\text{outliermeasure}(n) = \frac{[\text{rawom}(n) - \overline{\text{rawom}}]}{\sigma} \quad (5)$$

A collection of many decision trees is called a random forest. Counting the frequency of two samples at the same node of each tree can be used to estimate the degree of similarity between samples or the likelihood that they belong to the same class.)The aforementioned adjustment removes the numerical difference caused by the considerable fluctuation in the quantity of different types of samples, making it easier to compare the anomalous point scales of the various sorts of samples. For example, it is known that there are 5% of aberrant samples in each dataset used in the simulation experiment reported in this paper. Figure 2 shows the prediction, which selects the optimal penalty coefficient C using the grid approach. The accuracy of the model has improved to varied degrees after the abnormal samples were removed from the dataset using the three ways, demonstrating the efficacy of the three abnormal sample detection processes. The comparison revealed that the RF-based anomalous sample detection method works better than the other two approaches, improving the model's accuracy and resilience.

III. RESEARCH METHODS

This research uses an unsupervised machine learning anomaly detection technique with the random forest method with residual explanation. The random forest method is a module in the Scikit-Learn machine learning library which runs in the Python programming language.

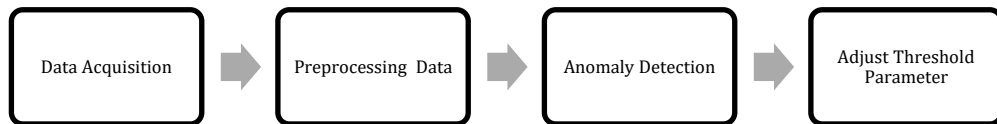


Figure 3. Research Design Anomaly Detection for Financial Data

Data for this study was obtained from a government institution in Indonesia, covering the period from 2022 to 2023. The dataset includes requests for funds related to operational goods spending accounts. The data that will be used is data on requests for funds for operational goods spending accounts.

In the data preprocessing stage, raw data is transformed into a suitable format to enable the extraction of valuable information when fed into the training model for effective clustering of financial data. This stage involves several key steps:

1. Data Cleaning
Removal of missing or inconsistent data to ensure data quality.
2. Data Transformation
Transforming data into the proper formats, such as encoding category variables and normalizing numerical values.

3. Feature Selection

Identifying and selecting relevant features that contribute significantly to the detection of anomalies.

During anomaly detection stage, the model, typically random forest regressor in this context, is used to predict the likelihood of each transaction being an anomaly. The predictions are based on the learned patterns from the historical data. Each data point is evaluated against these learned patterns, and those that do not conform to expected behaviors are flagged.

The process begins with the model making predictions on the dataset. For each transaction, the model calculates an anomaly score, which quantifies how much a given transaction deviates from what the model considers normal. Transactions with high anomaly scores are flagged as potential outliers.

The threshold parameter is crucial in anomaly detection models as it determines the cut-off point at which data points are classified as anomalies. Adjusting this parameter is a critical step in fine-tuning the model's performance to ensure it effectively identifies outliers without generating too many false positives.

IV. RESULTS AND DISCUSSION

Data for this study was obtained from a government institution in Indonesia, covering the period from 2022 to 2023. The data used is expenditure transaction data for purchasing goods or in Indonesia it is usually called SP2D data. In this data there is some information regarding the money spent, tax deductions, units submitted, and sub-output codes for budget items. In particular, we convert the tax deduction data into percentage form.

Table 1. List of Features

Name	Description
val_sp2d	Expenditure value
tax_deduction	Tax deduction
soutput_code	Budget item code
Trans_type	Type of transaction (GUP,LS, LS-KONTRAK, LS-KONTRAK-UNIT)

The author only filtered all expenditure data, only budget line account 525112 (goods expenditure) was used and the results of this filter produced data of 4,457 transactions.

Table 2. Data Type of Features

Feature	Data Type	Description
val_sp2d	Number	Value
tax_deduction	Decimal	Value
soutput_code	Text	Categorical Variable
Trans_type	Text	Categorical Variable

these features collectively provide a comprehensive view of each transaction, allowing the anomaly detection model to analyze and identify irregularities effectively. The

numerical features (`val_sp2d` and `tax_deduction`) capture the quantitative aspects of transactions, while the categorical features (`soutput_code` and `unit_code`) offer context and classification, enhancing the model's ability to detect anomalies that deviate from established patterns. This detailed analysis ensures that the model can effectively flag suspicious transactions, contributing to more robust financial oversight and fraud prevention.

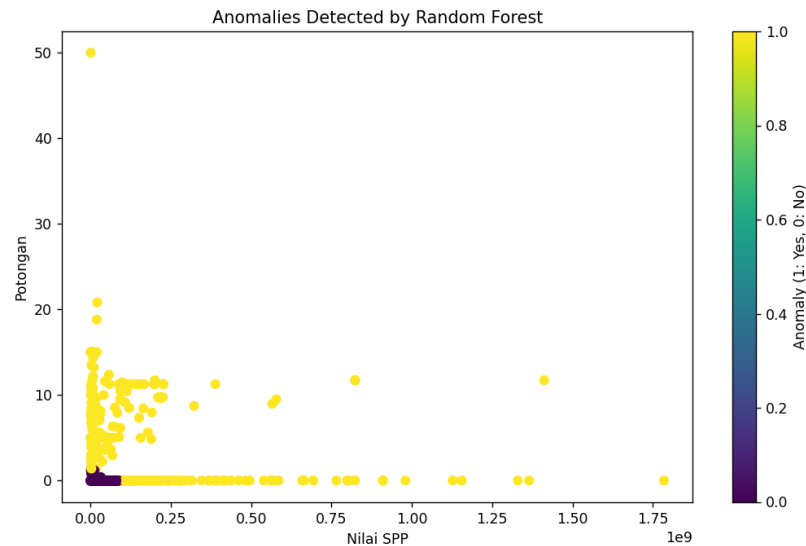


Figure 4. Anomaly detected by random forest for threshold 0.95

As seen in Figure 4, using a threshold of 0.95 resulted in too many transactions being classified as anomalies. Upon reviewing several transaction records, it was found that many valid transactions were incorrectly identified as anomalies.

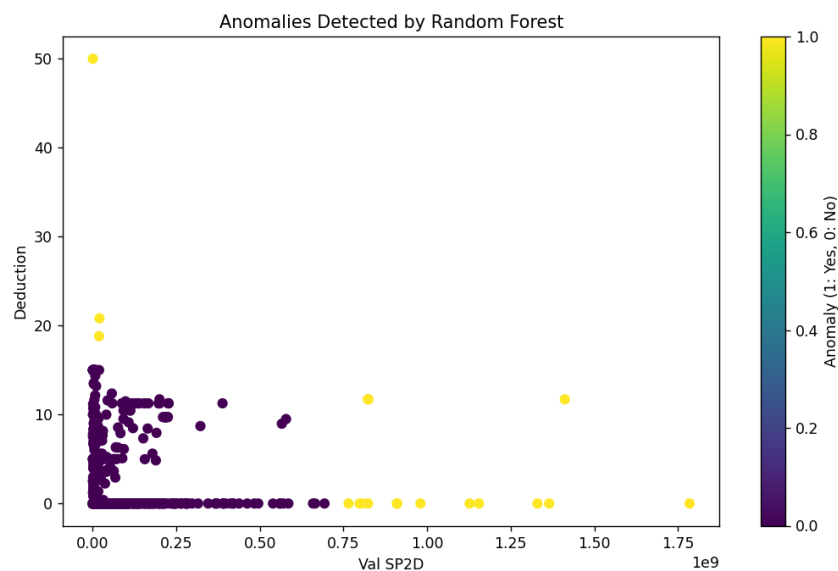


Figure 5. Anomaly Detected by random forest for thresholds 0.994

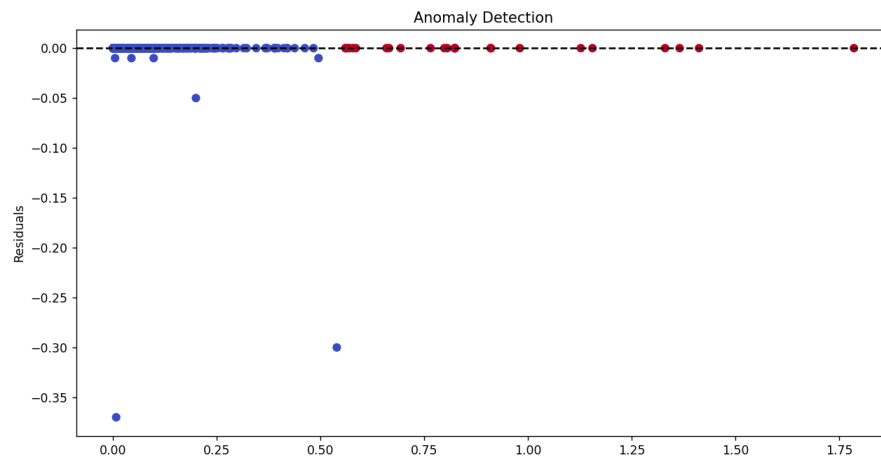


Figure 6. Anomaly detection with residual for threshold 0.994

After several adjustments to the threshold, the results with the smallest residuals and the most accurate anomaly detection, as observed by experts, were achieved. The results using a threshold of 0.996 can be seen in Figure 5. Additionally, the comparison between anomalies and residuals is shown in Figure 6.

	val_sp2d	deduction	anomaly	residual
51	300000	15.000000	1	0.0
186	805000000	0.000000	1	0.0
188	1155000000	0.000000	1	0.0
189	980000000	0.000000	1	0.0
191	1127000000	0.000000	1	0.0
192	1365000000	0.000000	1	0.0
196	1785000000	0.000000	1	0.0
198	560000000	0.000000	1	0.0
207	910000000	0.000000	1	0.0
212	560000000	0.000000	1	0.0
215	658700000	0.000000	1	0.0

Figure 7. Transactions identified as anomalies with a threshold of 0.994

	val_sp2d	deduction	anomaly	residual
0	212600000	9.713193	1	0.0
7	6750000	9.741333	1	0.0
8	209900000	9.712831	1	0.0
43	218000000	9.713890	1	0.0
46	111986600	10.454545	1	0.0
...	...	...	...	...
4404	18600000	3.629032	1	0.0
4405	51150000	4.728739	1	0.0
4409	371784250	0.000000	1	0.0
4418	199589100	11.711712	1	0.0
4419	245000000	0.000000	1	0.0

Figure 8. Transactions identified as anomalies with a threshold of 0.95

From the anomaly detection process, using a threshold of 0.95 identified 377 data points as anomalies, while using a threshold of 0.994 identified only 54 data points. The results of the anomaly detection can be seen in Figures 7 and 8. The residual values mostly show 0, it indicates that the model has high accuracy.

Table 3. Feature Importance

Name	Feature Importance
val_sp2d	0.5040
tax_deduction	0.4942
trans_type_GUP	0.0
trans_type_LS	0.0
trans_type_LS_KONTRAK	0.000949
trans_type_LS_KONTRAK_UNIT	0.0
soutput_code_4471.DBA.001	0.000369
soutput_code_4471.DBA.003	0.000539
soutput_code_4471.DBA.004	0.0

The features "val_sp2d" (50.40%) and "tax_deduction" (49.42%) remain the primary contributors to the model. The nearly 50-50 proportion indicates that these two features are almost equally important in the model's predictions. This likely reflects a strong relationship between the transaction value ("val_sp2d ") and the discounts given ("tax_deduction ") in the model's decision-making process for detecting anomalies.

For the feature " trans_type," the contribution of "trans_type_LS-KONTRAK" is now slightly higher (0.0949%). Although still very small, this indicates a slight increase in its influence on the model. This may suggest that in certain conditions or data subsets, the LS contract type has characteristics that might be associated with anomalies. Meanwhile, " trans_type_GUP," " trans_type_LS," and " trans_type_LS-KONTRAK-UNIT" still have zero or very low contributions, indicating that they are less informative in the context of this anomaly detection.

Regarding the "soutput_code" feature, codes like " soutput_code_4471.DBA.001" and " soutput_code_4471.DBA.003" show small increases in importance (0.0369% and 0.0539%, respectively). Although these increases are relatively minor, they suggest some variation in how these codes may relate to anomalous behavior. " soutput_code_4471.DBA.004" remains non-contributory, consistent with previous observations that this code might not be relevant to what the model is trying to predict or detect.

V. CONCLUSION AND SUGGESTION

Conclusion

The testing results indicate that the random forest with residual explanation model has the potential to detect anomalies in transaction data for inventory expenditure accounts (525112). Based on the trial of various threshold parameter values, a threshold of 96.64% was found to be the most optimal for detecting anomalies in the chosen model. Regardless of whether the detected anomalies actually indicate fraud, the findings from this study can be incorporated as an additional feature to serve as an early warning system in the institution's expenditure authorization application.

The features "val_sp2d" and "tax_deduction" are the most important for the model, indicating that transaction values and tax deductions play a key role in detecting anomalies. The "trans_type_LS-KONTRAK" feature has a minor increase in importance, suggesting it may be somewhat relevant in certain cases. Other transaction types have very little impact. Some specific "soutput_code" features show slight importance, hinting at potential links to anomalies, though one remains irrelevant.

Overall, these findings highlight the main factors influencing the model's anomaly detection.

Suggestion

The insights gained from this analysis emphasize the importance of focusing on the most influential features to improve the accuracy of anomaly detection. These findings can guide further refinement of the model and the development of an early warning system for expenditure authorization in institutional applications. Future work could explore the potential for integrating these results into real-time monitoring systems, enhancing the ability to detect and respond to anomalies more effectively.

Acknowledgments

We would like to express our sincere gratitude to the institution that provided the data necessary for this research. Their support and collaboration have been invaluable in enabling us to conduct this study. We also extend our thanks to the team members who contributed their expertise and time to ensure the success of this project.

BIBLIOGRAPHY

- Abdullah, M. I., Furqan, A. C., Yamin, N. Y., & Oktora, F. E. (2020). Incentive Function of Audit Opinion for the Increase of Regional Operational Expenditure and Own-Source Revenues Through Sensitivity Analysis in Indonesia. *Research in World Economy*, 11(1), 20. <https://doi.org/10.5430/rwe.v11n1p20>
- Agu, P. C., Inyama, O. I., & Ubesie, C. M. (2024). Effect of Government Expenditure on Human Capital Index in Nigeria. *European Journal of Accounting, Auditing and Finance Research*, 12(1), 18–33. <https://doi.org/10.37745/ejafr.2013/vol12n21833>
- Ahmed, M., & Mahmood, A. N. (2015). Novel Approach for Network Traffic Pattern Analysis using Clustering-based Collective Anomaly Detection. *Annals of Data Science*, 2(1), 111–130. <https://doi.org/10.1007/s40745-015-0035-y>
- Anggraeni, D. P. (2018). Pendeteksian fraud pada tata kelola penatausahaan persediaan badan layanan umum. *Jurnal Ekonomi Modernisasi*, 14(3), 168–176.
- Association of Certified Fraud Examiners. (2020). *Report to the Nations*. <https://acfepublic.s3-us-west-2.amazonaws.com/2020-Report-to-the-Nations.pdf>
- Center, I. C. C. (2019). *Internet Crime Report*. https://pdf.ic3.gov/2019_IC3Report.pdf
- Davidow, M., & Matteson, D. S. (2022). Factor analysis of mixed data for anomaly detection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(4), 480–493. <https://doi.org/10.1002/sam.11585>
- Goh, C., Lee, B., Pan, G., & Seow, P. S. (2021). Forensic analytics using cluster analysis: Detecting anomalies in data. *Journal of Corporate Accounting & Finance*, 32(2), 154–161. <https://doi.org/10.1002/jcaf.22486>
- Goldstein, M., & Uchida, S. (2016). A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. *PLOS ONE*, 11(4), e0152173. <https://doi.org/10.1371/journal.pone.0152173>
- Hartanto, S., Sunarto, H., & Supramono, S. (2020). The Determinants of the End-of-Year Spending Behavior of Local Governments' Financial Managers: A Lesson Learn

- from Indonesia. *Montenegrin Journal of Economics*, 16(1), 193–206. <https://doi.org/10.14254/1800-5845/2020.16-1.13>
- Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22, 85–126.
- Kozitsin, V., Katser, I., & Lakontsev, D. (2021). Online Forecasting and Anomaly Detection Based on the ARIMA Model. *Applied Sciences*, 11(7), 3194. <https://doi.org/10.3390/app11073194>
- Kravchuk, R. S., & Stone, S. B. (2010). How and when do structural deficits reveal themselves? The case of Indiana. *Journal of Public Budgeting, Accounting & Financial Management*, 22(4), 487–510. <https://doi.org/10.1108/JPBAFM-22-04-2010-B002>
- Lang, L. (2022). Analysis of Administrative Cost Data Based on Data Mining. *Academic Journal of Science and Technology*, 3(3), 202–205. <https://doi.org/10.54097/ajst.v3i3.2982>
- Lokanan, M., Tran, V., & Vuong, N. H. (2019). Detecting anomalies in financial statements using machine learning algorithm: The case of Vietnamese listed firms. *Asian Journal of Accounting Research*, 4(2), 181–201. <https://doi.org/10.1108/AJAR-09-2018-0032>
- Tiwari, S., Ramampiaro, H., & Langseth, H. (2021). Machine Learning in Financial Market Surveillance: A Survey. *IEEE Access*, 9, 159734–159754. <https://doi.org/10.1109/ACCESS.2021.3130843>
- West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. *Computers & Security*, 57, 47–66. <https://doi.org/10.1016/j.cose.2015.09.005>
- Xie, Y., Wang, M., Zhong, Y., Deng, L., & Zhang, J. (2023). Label-Free Anomaly Detection Using Distributed Optical Fiber Acoustic Sensing. *Sensors*, 23(8), 4094. <https://doi.org/10.3390/s23084094>
- Zhang, Q. (2022). Financial Data Anomaly Detection Method Based on Decision Tree and Random Forest Algorithm. *Journal of Mathematics*, 2022, 1–10. <https://doi.org/10.1155/2022/9135117>