# Fraud Detection in Sales Transactions: A Comparative Study of DBSCAN/LOF and Isolation Forest

Student: Yunzhen Wu

MSDS411 Unsupervised Learning Methods

Instructor: Dr. Jamie Riggs

Mar. 13, 2025

**Abstract**

Fraudulent sales reporting can result in significant financial losses for companies operating on a commission-based structure. This study evaluates two anomaly detection methods—**DBSCAN/LOF and Isolation Forest**—to detect fraudulent sales transactions. The dataset comprises **133,731 unlabeled sales transactions** and a **test set of 15,732 labeled transactions**, where **1,270** are known fraud cases. Due to the class imbalance, **F1-score, Precision-Recall AUC, and Balanced Accuracy** were used as key performance metrics. The results indicate that **Isolation Forest achieved an F1-score of 0.7555, outperforming DBSCAN/LOF (0.7113)**, making it the preferred method for fraud detection. However, DBSCAN/LOF demonstrated effectiveness in identifying local density anomalies, making it a viable alternative in certain contexts.

## 1. Introduction

Fraudulent sales reporting occurs when sales representatives misrepresent transaction values or quantities to **artificially increase commission earnings**. Traditional fraud detection methods, such as **rule-based systems**, are often ineffective because fraudsters continuously adapt their behavior to bypass detection. As a result, **machine learning-based anomaly detection techniques** offer a more adaptable approach to identifying suspicious transactions.

This study explores two unsupervised anomaly detection methods:**DBSCAN/LOF (Density-Based Clustering & Local Outlier Factor)** and **Isolation Forest (Tree-Based Anomaly Detection).** The primary goal is to evaluate these methods and **recommend the most effective fraud detection technique** based on test-set performance, using precision-recall-balanced evaluation metrics.

## 2. Literature Review

Density-based anomaly detection methods such as **DBSCAN** (Ester et al., 1996) and **Local Outlier Factor (LOF)** (Breunig et al., 2000) have been widely used in fraud detection. These methods identify anomalies based on the density of surrounding data points, making them effective for datasets where fraud cases form distinct, sparse clusters. Prior research has shown that **DBSCAN** performs well when data exhibits **natural clustering,** while **LOF** assigns anomaly scores to individual transactions based on their **relative density compared to nearby points**.

Tree-based anomaly detection techniques, particularly **Isolation Forest** (Liu, Ting, & Zhou, 2008), have demonstrated efficiency in fraud detection due to their **low computational cost and ability to handle high-dimensional data**. The Isolation Forest algorithm works by randomly partitioning data and identifying anomalies based on their susceptibility to **early isolation**.

Comparative research (Campos et al., 2016) suggests that the choice of anomaly detection methods **depends on dataset characteristics and the specific patterns encountered**, because no single approach is universally superior.

## 3. Methods

### 3.1 Exploratory Data Analysis (EDA) and Data Preprocessing

The datasets used in this project contains a systematic sample of sales observations from a case study reported by Torgo (2017, 2021). The train set comprises **133,731 unlabeled sales transactions** and a **test set of 15,732 labeled transactions**, where **1,270** are known fraud cases. The key features used in the models are:

- Prod: Product ID
- Quant: Quantity sold
- Val: Total sales value
- Insp: Inspection result

**Missing values** in both train and test set were minimal (**<3%**). There are only missing data in two columns “Quant”, and “Val”, and the missing values were dropped for consistency.

The class distribution in the test set reveals a significant imbalance, with a **majority (~92%) of transactions labeled as "ok" and only a small fraction marked as "fraud"** (Figure 1). This imbalance necessitates the use of **precision-recall balanced metrics** for evaluation.

Further analysis of **transaction values (Val) and quantities sold (Quant)** distributions indicate that both variables are **highly skewed**, with the majority of transactions occurring at **lower values** (Figure 2). To address this skewness, **log transformations** would be applied to these variables before model training.

Additionally, an analysis of **fraud rate by product** (Figure 3) demonstrates that **some products have disproportionately high fraud rates**, leading to the decision to include **product-level fraud risk** as a feature in the models.

### 3.2. Model Implementation

#### 3.2.1 DBSCAN + LOF

DBSCAN was implemented with **optimized hyperparameters (eps=0.16, min_samples=9)**, while LOF was configured with **a contamination rate of 0.01 and a threshold of 55%**. A transaction was classified as fraud if it was labeled as an anomaly by **both DBSCAN and LOF**, or they were associated with high-risk products.

#### 3.2.2 Isolation Forest

Isolation Forest was trained with **400 trees (n_estimators=400), a contamination rate of 0.07, and a sampling rate of 0.9**. Transactions were labeled as fraud if the Isolation Forest model labeled them as anomalies, or they were associated with high-risk products.

### 3.3 Model Evaluation Metrics

Due to the dataset's **class imbalance**, the models were assessed using:

- **F1-Score (Macro & Weighted Average)**: To balance precision and recall.
- **Precision-Recall AUC**: A metric suited for imbalanced classification.
- **Balanced Accuracy**: Adjusted accuracy that accounts for the dataset's skewed distribution.

## 4. Results

| Model | Macro F1 | Weighted F1 | PR AUC | Balanced Accuracy |
|---|---|---|---|---|
| DBSCAN + LOF | 0.7113 | 0.9097 | 0.5084 | 0.7542 |
| Isolation Forest | 0.7555 | 0.9214 | 0.5985 | 0.8229 |

The results indicate that **Isolation Forest outperformed DBSCAN/LOF across all evaluation metrics** (Figure 4 & Table above). Isolation Forest achieved **higher balanced accuracy (82.29%), higher recall (0.9293) and F1-score (Macro: 0.7555, Weighted: 0.9214)**, making it **the preferred model for fraud detection**. DBSCAN/LOF, while slightly less effective, might capture local anomalies that Isolation Forest overlooked.

## 5. Conclusion

This study compared **DBSCAN/LOF and Isolation Forest** for fraud detection in sales transactions. The findings suggest that **Isolation Forest is the superior method**, given its

**higher recall, precision, and overall performance**. However, **DBSCAN/LOF remains useful in cases where fraud follows localized density variations**.

For deployment, the following recommendations apply:

- **If minimizing false positives is critical**, Isolation Forest is the better choice.
- **If detecting nuanced fraud patterns is the priority**, DBSCAN/LOF may be beneficial.

Future improvements could include a **hybrid model combining both techniques**, as well as **feature engineering using external metadata** such as sales regions.

**References**

Breunig, Markus M., Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. “LOF: Identifying Density-Based Local Outliers.” *ACM SIGMOD Record* 29 (2): 93–104.

Campos, Guilherme O., Arthur Zimek, Jörg Sander, Rodrigo J. G. B. Campello, Barbora Micenková, Erich Schubert, Ira Assent, and Michael E. Houle. 2016. *“On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study.”* Data Mining and Knowledge Discovery 30 (4): 891–927.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD’96)*, 226–31.

Liu, Fei Tony, Kai Ming Ting, and Zhi-Hua Zhou. 2008. “Isolation Forest.” *In ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, December 2008*, 413–422.

Torgo, Luis. 2017. *Data Mining with R: Learning with Case Studies* (2nd ed.). Boca Raton, Fla.: Chapman & Hall/CRC Press. [ISBN-13: 978-1482234893] Data and software available from the Comprehensive R Archive Network.

Torgo, Luis. 2021. "Package DMwR2." Comprehensive R Archive Network. Source of the full data set from which the systematic sample has been drawn.

## Appendix

**Figure 1:** Distribution of Fraud vs. Normal Transactions.


**Figure 2:** Distribution of Quant and Val in Train & Test Sets.

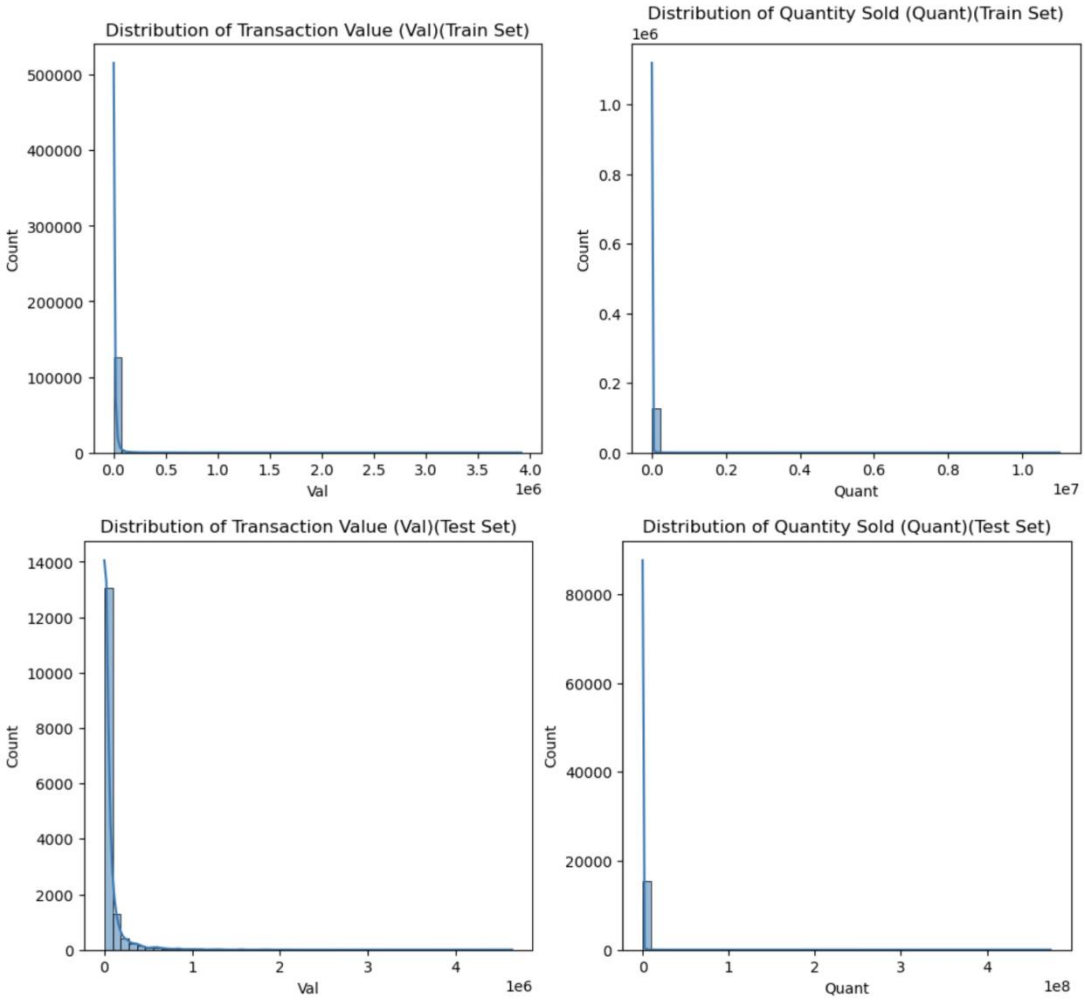

**Figure 3:** Fraud Rate Distribution by Product.

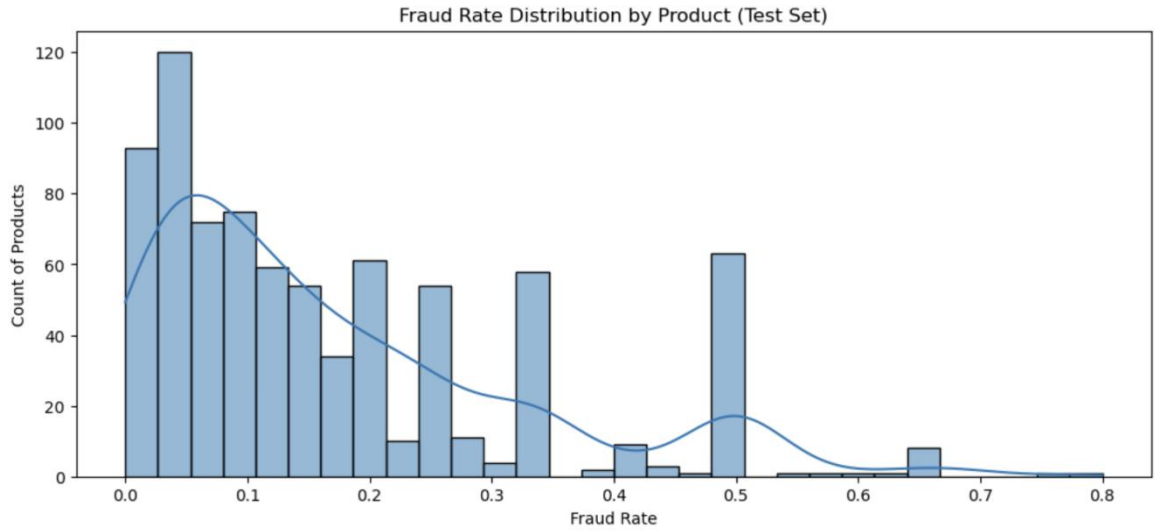


**Figure 4:** Model Performance Comparison.