Comparative Analysis of Clustering-Based Anomaly Detection in Sales Data

Name: Muhammad Ameer Hamza
Student ID: 22034204
Group no: 38
Topic: Anomaly Detection

GitHub Link:
https://github.com/ameerhamza95/Data_Mining_Assignment_2.git

Abstract:

This report explores anomaly detection within the Sales Transactions Weekly Data Set from UCI, employing clustering techniques such as K-means and DBSCAN. The analysis identifies significant sales pattern deviations, aiming to understand their occurrence and potential implications on inventory and sales strategies. Results indicate distinct clustering and anomaly patterns, offering insights for targeted business interventions.

Introduction:

Anomaly detection in sales data is crucial for unearthing non-conforming patterns that could signal significant business insights or operational concerns. Utilizing the Sales Transactions Weekly Data Set from UCI, this report adopts K-means and DBSCAN clustering techniques to surface such anomalies. The choice of DBSCAN—Density-Based Spatial Clustering of Applications with Noise—is motivated by its proficiency in identifying outliers in spatial data. Unlike K-means, DBSCAN does not require pre-specifying the number of clusters, making it adept at handling arbitrary-shaped clusters and noise, thereby offering a robust alternative for anomaly detection [1].

Methodology:

The methodology for anomaly detection in the provided sales data entails several systematic steps:

Data Preprocessing:

The raw sales data is first standardized to ensure uniformity, making it suitable for clustering. StandardScaler from scikit-learn is applied to normalize the dataset, which is critical for the clustering algorithms [2].

Clustering Algorithms:

- K-means: An iterative algorithm that partitions the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. The Elbow Method is employed to determine the optimal number of clusters (K) by identifying the point where the within-cluster sum of squares (WCSS) starts to diminish at a slower rate [3].
- DBSCAN: Stands for Density-Based Spatial Clustering of Applications with Noise. It groups together points that are closely packed together while marking points that lie alone in low-density regions as outliers (anomalies). DBSCAN parameters such as epsilon (eps) and minimum points (minPts) are crucial. The epsilon value is chosen where the greatest curvature (elbow point) is observed in a k-distance plot, which is indicative of a suitable distance to consider for the neighbourhood radius [1].

Performance Evaluation:

- The silhouette score is used to assess the quality of clusters formed by both algorithms. It measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation) [4].

Anomaly Detection:

- For K-means, anomalies (shown in Fig*ure 1*) are detected based on their distance from the nearest cluster centre. points lying beyond a chosen percentile threshold (99th) are classified as anomalies [5].
- In DBSCAN, points labelled as '-1' are considered anomalies (shown in Figure 2), representing points in low-density regions [6].

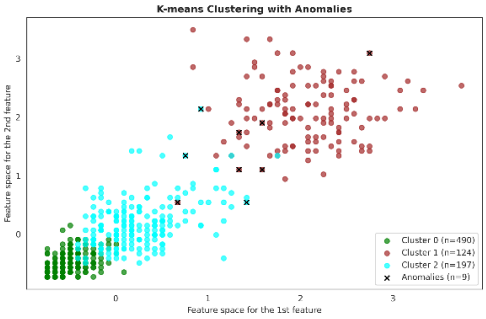

Figure 1: K-means clustering with anomalies

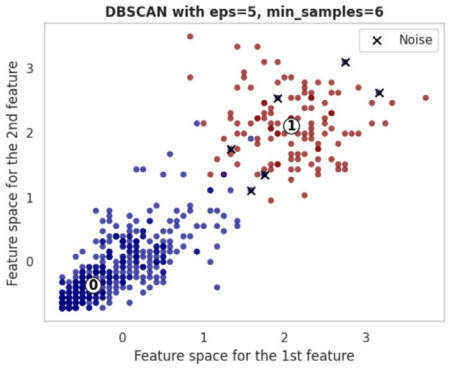

Figure 2: DBSCAN with anomalies

Results:

In the provided analysis of anomaly detection using sales data, DBSCAN and K-means clustering algorithms were applied to discern

irregularities that could influence sales strategies. DBSCAN was favoured for its proficiency in identifying outliers without presuming data distribution, which is crucial in anomaly detection where data may not conform to typical patterns. The methodology entailed data preprocessing to standardize the dataset and the employment of the elbow method and silhouette scores to optimize clustering parameters. The results indicated two optimal clusters for DBSCAN, with a high silhouette score of 0.68 suggesting well-separated clusters. K-means suggested three to five clusters, with the best silhouette score of 0.609 for three clusters. Anomalies were detected in both methods, with K-means identifying nine and DBSCAN six, signifying variations in detection sensitivity. The PCA analysis (shown in *Figure 3*) reinforced these findings, providing a visual representation of clusters and anomalies for comprehensive insight. These methods are instrumental for businesses to detect unusual patterns, aiding in informed decision-making and strategy optimization [7].

Discussion:

In discussing the results of DBSCAN and K-means clustering methods applied to sales data, we can note several insights. DBSCAN, with its density-based approach, identified fewer anomalies than K-means, which may suggest DBSCAN's higher sensitivity to local data density and potentially greater resistance to noise. The silhouette scores indicate a relatively good separation of clusters, with DBSCAN achieving a higher score, suggesting more cohesive clusters.

When interpreting the PCA plots with anomalies (shown in *Figure 3*) and sales trends (shown in *Figure 4*), it's clear that both clustering methods have flagged products with irregular sales patterns, likely driven by factors like seasonal demand or promotions. The comparative analysis would benefit from further investigation into the reasons behind these patterns for actionable business insights.

However, each method has limitations and potential biases. K-means relies on the assumption of spherical clusters, which may not accurately represent complex data shapes. On the other hand, DBSCAN's performance is highly dependent on the choice of distance measure and density parameters. These inherent biases should be taken into account as they could influence the identification of anomalies and should be considered when generalizing findings [8].

Conclusion:

Analysing sales data with DBSCAN and K-means revealed distinct patterns and anomalies. DBSCAN excels at finding core clusters and outliers, while K-means highlights anomalies but can be biased by assuming specific cluster shapes. This knowledge can power better business decisions in areas like inventory management, marketing, and forecasting, particularly by investigating the identified anomalies to improve sales performance. Further research is recommended to refine these clustering techniques, perhaps by integrating them with predictive analytics to proactively manage and respond to emerging sales trends. Continued exploration of parameter tuning and the incorporation of additional data features could also provide a more nuanced understanding of sales dynamics [9].

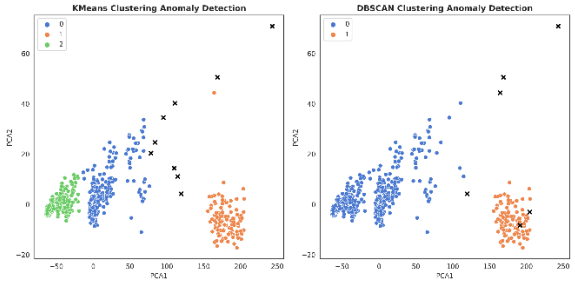

Figure 3: PCA of K-means and DBSCAN

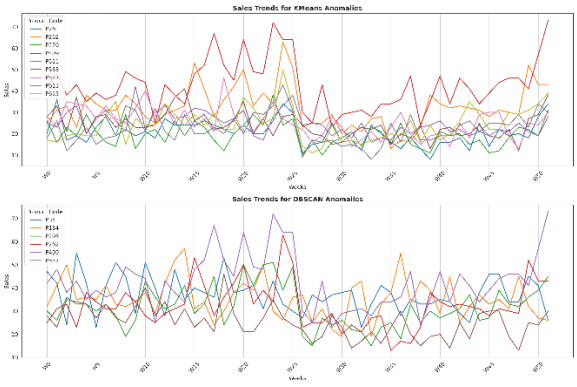

Figure 4: Sales Trend of Anomalous Products

References:

[1] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," in KDD, 1996.

[2] Scikit-learn developers. (n.d.). StandardScaler. scikit-learn. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[3] "K-Means Clustering Algorithm: Applications, Evaluation Methods, and Drawbacks." Towards Data Science. Available at: https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

[4] Rousseeuw, P.J., "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, 1987.

[5] "Anomaly Detection with K-means in Python." Data Tech Notes. Available at: https://www.datatechnotes.com/2020/05/anomaly-detection-with-kmeans-in-python.html

[6] "Anomaly Detection with DBSCAN in Python." Data Tech Notes. Available at: https://www.datatechnotes.com/2020/04/anomaly-detection-with-dbscan-in-python.html

[7] "Comparing DBSCAN, k-means, and Hierarchical Clustering: When and Why To Choose Density-Based Methods." Hex. Available at: https://hex.tech/blog/comparing-density-based-methods/

[8] "DBSCAN vs. K-Means: A Guide in Python." Pierian Training. Available at: https://pieriantraining.com/dbscan-vs-kmeans-a-guide-in-python/

[9] "Predictive Analytics in Sales." Breadcrumbs. Available at: https://breadcrumbs.io/blog/predictive-analytics-in-sales/