Customer Segmentation & Anomaly Detection in Retail Transactions

Team 24 - Matthew Nguyen, Kevin Su; University of Houston - College of Natural Science and Mathematics - Houston, Texas

ABSTRACT

We present a data science pipeline for customer segmentation and anomaly detection using the **Online Retail II** dataset (UK online gift retailer, 2009–2011). From invoice-level data we engineer intuitive RFM-style features—recency, frequency, monetary value, and basket size—to summarize each customer's purchasing behavior. A **compact autoencoder** (AE) learns non-linear embeddings of these features, which are clustered with **K-Means** to form interpretable customer segments. We then flag anomalous customers using both **Isolation Forest** (IF) and **AE reconstruction error**. The analysis confirms a strong **Pareto effect** in revenue, identifies a small group of extremely high-value customers, and finds a consistent set of unusual buyers where IF and AE agree. Together, these results show how **classical RFM** ideas and modern representation learning can be combined to prioritize retention, marketing, and risk-management actions from raw transaction data.

BACKGROUND

In online retail, a small fraction of customers contributes a disproportionate share of revenue, often described by the **Pareto principle** ("~20% of customers drive ~80% of sales"). Knowing who those customers are—and how they differ from the rest—is essential for targeted marketing, inventory planning, and fraud prevention. Classical approaches rely on **RFM** features (Recency, Frequency, Monetary value) with simple clustering, but high-dimensional, noisy data can make clusters unstable and hide subtle structure. Unusual purchasing patterns may also signal fraud, data quality issues, or rare but valuable customer behavior. This project extends the RFM framework with **autoencoders** and **Isolation Forest** to learn richer representations and detect anomalies in a label-free setting on the **Online Retail II** dataset. It also serves as a focused case study of unsupervised learning, representation learning, and **anomaly detection** techniques covered.

OBJECTIVE

Build a pipeline that segments customers and detects anomalous purchasing behavior in the Online Retail II dataset.

- Create interpretable customer segments using engineered RFM / basket features and K-Means clustering on autoencoder embeddings.
- Produce a ranked anomaly list using Isolation Forest and AE reconstruction error, highlighting cases where both methods agree and are therefore most trustworthy.
- Summarize results in business-friendly visuals (Pareto curve, segment profiles, segment lift, anomaly agreement) that can directly inform marketing, retention, and risk-management decisions.

METHODS

Data & Feature Engineering

- Start from the Online Retail II CSV. Drop cancellations, non-positive quantity/price lines, and rows with missing customer IDs. Aggregate invoices to the customer level, computing: transaction count, total and mean spend, total item quantity, mean basket size, recency_days (days since last purchase), and RFM scores based on recency, frequency, and monetary quantiles. Persist the cleaned customer-level table for reruns and downstream modeling.

Representation Learning & Segmentation

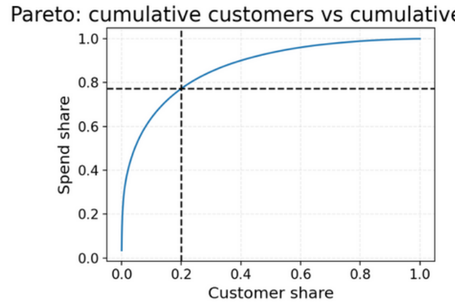
- Standardize numeric features and train a shallow autoencoder (TensorFlow/Keras) to compress them into a low-dimensional latent vector. Run K-Means on AE embeddings for k ∈ [2, 10]; select k by silhouette score and compare against clustering on raw features and sequence-based embeddings. Profile each segment using z-scored feature means to create interpretable segment "fingerprints" (Figure 2).

Anomaly Detection

- Train Isolation Forest on the same feature space to obtain an anomaly score per customer. Use AE reconstruction MSE as a second anomaly score; work on a log scale to stabilize the heavy-tailed error distribution. Define "high-risk" customers as those above high quantiles of both IF score and AE log-MSE and visualize agreement with a hexbin plot (Figure 4). Export all scores, cluster labels, and metrics to CSV for transparency and reuse in further analysis.

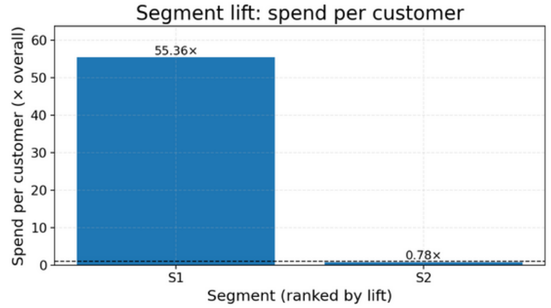
RESULTS

Figure 1: Pareto Curve of Customer Spending



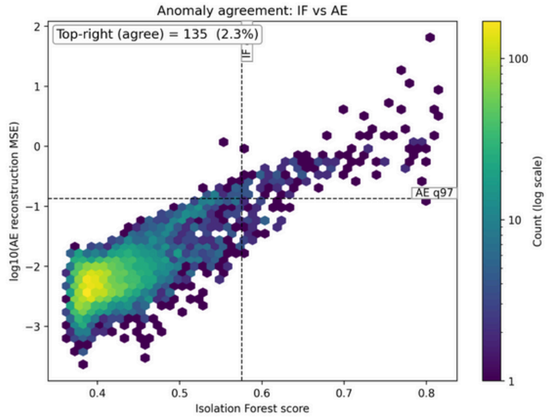
Roughly the top ~20% of customers account for about 75–80% of total revenue, confirming a strong Pareto effect and motivating focused retention strategies for this group.

Figure 3: Segment Lift Chart



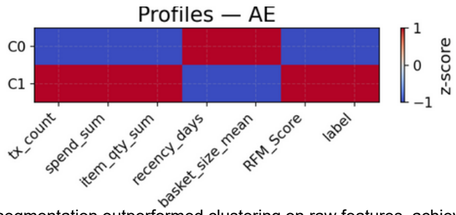
When we rank segments by spend per customer, the top segment spends dozens of times more than the overall average, while lower segments spend at or below the mean. This makes the high-value segment an obvious target for tailored promotions, VIP-style benefits, and retention programs

Figure 4: Anomaly Agreement Hexbin



Using high quantiles of IF score and AE log-MSE, only ~2–3% of customers fall in the top-right "agree" region. These anomalies often correspond to unusually large baskets or very rare spending patterns compared with the learned "normal" behavior, and therefore deserve manual review

Figure 2: AE Segment Profile Heatmap



AE-based segmentation outperformed clustering on raw features, achieving a silhouette score of ~0.92. Cluster profiles show clear separation in spend, frequency, recency, and basket size, with one segment behaving as a "VIP core" and others aligning with more occasional or discount-driven buyers. Across multiple random seeds and K-Means initializations, silhouette scores varied only slightly, indicating that the discovered segments are stable rather than artifacts of initialization.

CONCLUSIONS

Autoencoder-based segmentation provided sharper and more stable clusters than traditional feature-only K-Means while remaining easy to interpret through RFM-style summaries. The resulting segments map naturally onto business actions, especially prioritizing a small, extremely high-value customer group for targeted retention and experience upgrades. Combining Isolation Forest with AE reconstruction error allowed us to flag a compact set of "high-agreement" anomalies that are both rare and behaviorally extreme, making them strong candidates for fraud checks, data quality review, or bespoke customer outreach. Overall, the project demonstrates how DSII techniques in clustering, representation learning, and anomaly detection can be applied to a real retail dataset to support both strategic and operational decision-making.