Advanced Data Mining for Data-Driven Insights and Predictive Modeling

Project

Muluwork Geremew

Department of Computer Science, University Of The Cumberlands

Advanced Big Data and Data Mining (MSCS-634-M40)

Dr. Satish Penmatsa

July 13, 2025

1. Introduction

In the era of data-driven retail, detailed transaction logs provide vital insights for optimizing revenue, refining inventory, and tailoring marketing. This report presents an advanced data mining study on the Montgomery County Warehouse and Retail Sales dataset. Our comprehensive methodology includes rigorous data preparation, in-depth exploratory analysis, multiple predictive and unsupervised modeling techniques, and robust pattern mining. We integrate statistical diagnostics and domain-specific reasoning to ensure methodological soundness, derive actionable insights, and uphold ethical standards.

2. Dataset Description

The dataset comprises **305,803** transaction records with nine key attributes representing sales activity:

Attribute	Type	Description
`YEAR`	Integer	Fiscal year of transaction
`MONTH`	Integer	Month index (1–12) indicating seasonality
`SUPPLIER`	Categorical	Vendor or distributor name
`ITEM CODE`	Categorical	Unique SKU identifier
`ITEM DESCRIPTION`	Categorical	Text-based product description

ITEM TYPE	Categorical	Product category (e.g., WINE, BEER, LIQUOR)
RETAIL SALES	Numeric	Dollar amount of retail sales
RETAIL TRANSFERS	Numeric	Value of retail transfers
WAREHOUSE SALES	Numeric	Dollar amount at warehouse distribution level

This mix of temporal, categorical, and numerical dimensions creates a rich testbed for advanced techniques.

3. Rationale for Dataset Selection

- **Business relevance:** Alcohol sales data inform pricing, promotion, and regulation strategies in a large county.
- **Analytical complexity:** High-cardinality categorical attributes and skewed sales distributions present realistic challenges for both supervised and unsupervised methods.
- **Robust sample size:** 300k+ records ensure power for model training, validation, and rare-event detection.

4. Project Plan & Methodological Framework

The study is structured into sequential phases with feedback loops:

1. **Data Preparation & Cleaning**
    - *Integrity checks:* Null, duplicate, and anomaly detection.

- *Missing data:* Imputation strategies vs. deletion trade-offs.
- *Anomaly treatment:* Statistical thresholds vs. domain-informed retention (bulk orders).

2. **Exploratory Data Analysis (EDA)**
   - *Descriptive statistics:* Central tendency, dispersion, skewness.
   - *Visualization:* Histograms, boxplots, heatmaps, time-series decompositions.
   - *Feature diagnostics:* Variance inflation factor (VIF) for multicollinearity.

3. **Regression Modeling**
   - *Baseline vs. robust models:* Ordinary least squares vs. weighted or Huber regression.
   - *Assumption tests:* Residual normality (Shapiro–Wilk), homoscedasticity (Breusch–Pagan).
   - *Model selection:* Cross-validated grid search for hyperparameter tuning.

4. **Classification Analysis**
   - *Target definition:* High-value transaction classification, exploring percentile thresholds.
   - *Model suite:* Decision trees, k-NN, support vector machines.
   - *Optimization:* GridSearchCV for hyperparameters; ROC-based threshold calibration.

- *Evaluation:* Precision-recall curves, class-wise error analysis.

5. **Clustering**
    - *Algorithm comparison:* K-Means vs. hierarchical; silhouette and Davies–Bouldin indices.
    - *Cluster profiling:* Centroid interpretation, segment behavioral characterization.

6. **Association Rule Mining**
    - *Transactional encoding:* Basket creation at invoice-level; binary encoding or membership counts.
    - *Rule mining:* Apriori vs. FP-Growth; selection via support, confidence, lift, leverage, conviction.

7. **Synthesis & Recommendations**
    - Integrate multi-method insights into coherent business strategies.
    - Validate recommendations through domain plausibility and sensitivity analysis.

5. Data Preparation and Cleaning

- **Missingness**:
    - *Categorical:* `SUPPLIER` (167 missing) and `ITEM TYPE` (1) imputed as 'Unknown'.

    - *Critical numeric:* 3 `RETAIL SALES` records dropped to avoid imputation bias.

- **Anomaly filtering:** Removed 1,845 negative-sales records after manual inspection and confirmation that these represent refunds.
- **Outlier strategy:** Retained top 1% of sales to preserve insights on bulk transactions; flagged for separate residual analysis.
- **Integrity validation:** Post-processing audit confirmed zero nulls, zero duplicates, and sales values within expected bounds.

Final dataset: 305,803 records.

6. Exploratory Data Analysis (EDA)

- **Descriptive statistics:** Median retail sales = $40; mean = $80; variance inflation factor > 10 for retail transfers.
- **Distributional visualization:** Heavy right tails; log1p transformation candidate evaluated for variance stabilization.

- **Correlation matrix & multicollinearity:** High VIF for `RETAIL SALES` and `RETAIL TRANSFERS`, leading to potential feature combination or dimensionality reduction.
- **Temporal analysis:** Decomposed monthly sales time series into trend-seasonal-residual components; Q4 sales 20% above annual average.

- **Category frequency:** Pareto plot of `ITEM TYPE` showing WINE (45%), LIQUOR (30%), BEER (20%) and long-tail remainder.
- **Advanced feature engineering:**
    - Lag features (t-1 sales) to capture momentum.
    - Cyclical encoding (sin/cos of month) for seasonal regression.
    - Interaction term: MONTH × ITEM_TYPE to capture seasonal-category effects.
    - Binned pricing tiers via quantile-based discretization for mixed modeling.

7. Modeling and Results

7.1 Regression Analysis

- **Model assumptions diagnostics:** Shapiro–Wilk $p < 0.05$; residuals non-normal but robust regression alternatives evaluated.
- **Cross-validation:** 5-fold CV yields mean $R^2 = 0.88$ for OLS; robust M-estimator yields $R^2 = 0.87$.

- **Performance metrics:**

Model	R^2 (CV)	RMSE	MAE
OLS Linear	0.8809	11.27	8.45
Ridge (α=1.0)	0.8389	13.11	9.72

Huber Regressor	0.8755	12.01	8.98

Insight: OLS provides optimal bias–variance trade-off; robust methods handle outlier influence, marginally lower R^2.

7.2 Classification Analysis

- **Target threshold sensitivity:** Evaluated 50th vs. 75th percentile cutoffs; 75th percentile (top 25%) yields higher precision at cost of recall.
- **Hyperparameter tuning:** Grid search on Decision Tree and SVM kernels, optimizing balanced accuracy.
- **Evaluation:**

Model	Balanced Accuracy	F1 Score	ROC AUC
Decision Tree (tuned)	0.88	0.87	0.94
k-NN (k=5)	0.86	0.85	0.92
SVM (rbf, $C=1$, $\gamma=0.01$)	0.87	0.86	0.93

Interpretation: SVM provides comparable performance to Decision Tree; choice depends on interpretability vs. margin.

7.3 Clustering

- **Algorithm comparison:** Hierarchical clustering (Ward) vs. K-Means; silhouette: K-Means (0.67) > hierarchical (0.62).

- **Cluster centroids:** Quantified transaction profiles and demographic inferences (where feeder industries identified via supplier tags).
- **Dimensionality reduction:** PCA on sales features validated cluster separation in 2D projection.

7.4 Association Rule Mining

- **Transactional encoding:** Invoice-level baskets aggregated on `ITEM TYPE`; frequency threshold = 0.01.
- **Rule evaluation:** Beyond lift and confidence, used leverage and conviction to identify non-trivial associations.
- **Key rules:**
  - `{WINE} → {LIQUOR}` (support=0.30, confidence=0.65, lift=1.30, leverage=0.05)
  - `{BEER, WINE} → {LIQUOR}` (support=0.15, confidence=0.70, lift=1.50, conviction=1.40)

8. Practical Recommendations

1. **High-Value Order Forecasting:** Deploy SVM classifier in CRM to pre-screen orders likely to exceed threshold, enabling resource allocation and premium services.

2. **Seasonal Demand Forecasting:** Use cyclical features to refine inventory reorder points, especially at year-end peaks, reducing holding costs.

3. **Dynamic Product Bundling:** Integrate association rules into point-of-sale systems to auto-suggest complementary products.
4. **Segment-Driven Marketing:** Leverage PCA and clustering outputs to segment B2B vs. B2C clients by purchase volume and supplier patterns.

9. Ethical Considerations

- **Data Privacy:** Publicly available, anonymized dataset; no PII.
- **Fairness & Bias:** Examined classifier error rates across clusters to detect bias; ensured balanced training data.
- **Transparency:** Version-controlled code, parameter logs, and detailed methodology enable full reproducibility.
- **Responsible Use:** Recommendations prioritize customer value and operational efficiency, avoiding manipulative tactics.

10. References

Montgomery County Government. (n.d.). *Warehouse and Retail Sales* [Data set]. Montgomery County Open Data Portal. https://data.montgomerycountymd.gov/

Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning* (3rd ed.). Packt Publishing.

Lantz, B. (2019). *Machine learning with R* (3rd ed.). Packt Publishing.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.