Advanced Data Mining for Retail Intelligence

Online Retail II Dataset

Final Project

Group Members:
Pabitra Bhandari
Haeri Kyoung
Vamsi Krishna Gajulapalli
Prakash Tamang

Instructor: Satish Penmatsa
Semester: Spring 2026

Advanced Big Data and Data Mining
MSCS- 634-M20
University of the Cumberlands
Spring 2026

Project Objective

What Was the Goal?

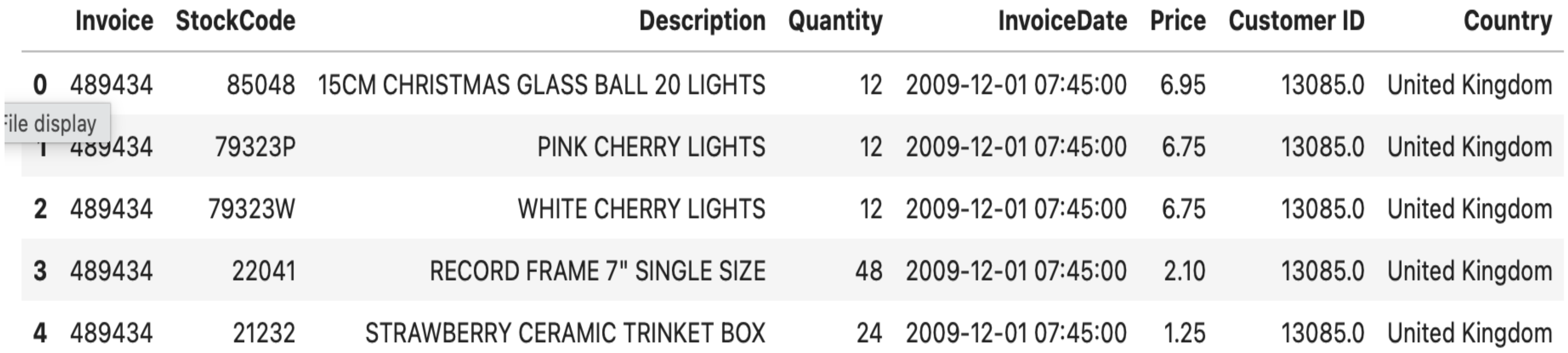
- Apply full data mining lifecycle to real-world data
- Extract actionable business insights
- Compare supervised and unsupervised models
- Evaluate model reliability and generalization

Key Focus:
From raw transactions → structured insights → business recommendations

Dataset Overview

Online Retail II Dataset

- UK-based transactional retail data
- 500+ records and 8+ attributes
- Mix of numerical and categorical features
- Customer purchase behavior data



	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

Why This Dataset?

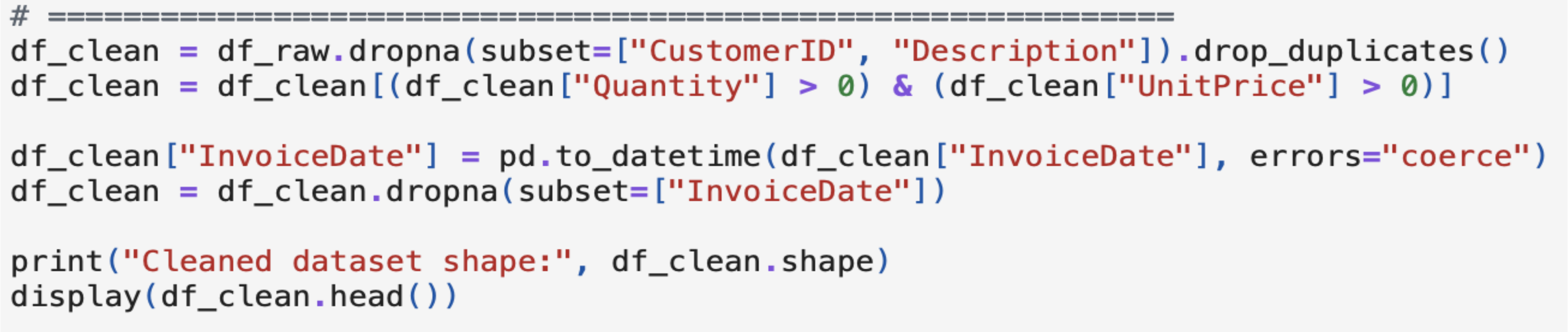
Strategic Reasons for Selection

- Meets academic project requirements
- Supports regression, classification, clustering, and association rules
- Real-world business relevance
- Rich transactional structure
- This dataset allows both predictive modeling and pattern discovery.

Data Preparation

Data Cleaning Steps

- Removed duplicate records
- Filtered canceled transactions
- Handled missing Customer IDs
- Converted InvoiceDate to datetime
- Removed unrealistic negative values
- Engineered feature:
  Total Transaction Value = Quantity × UnitPrice

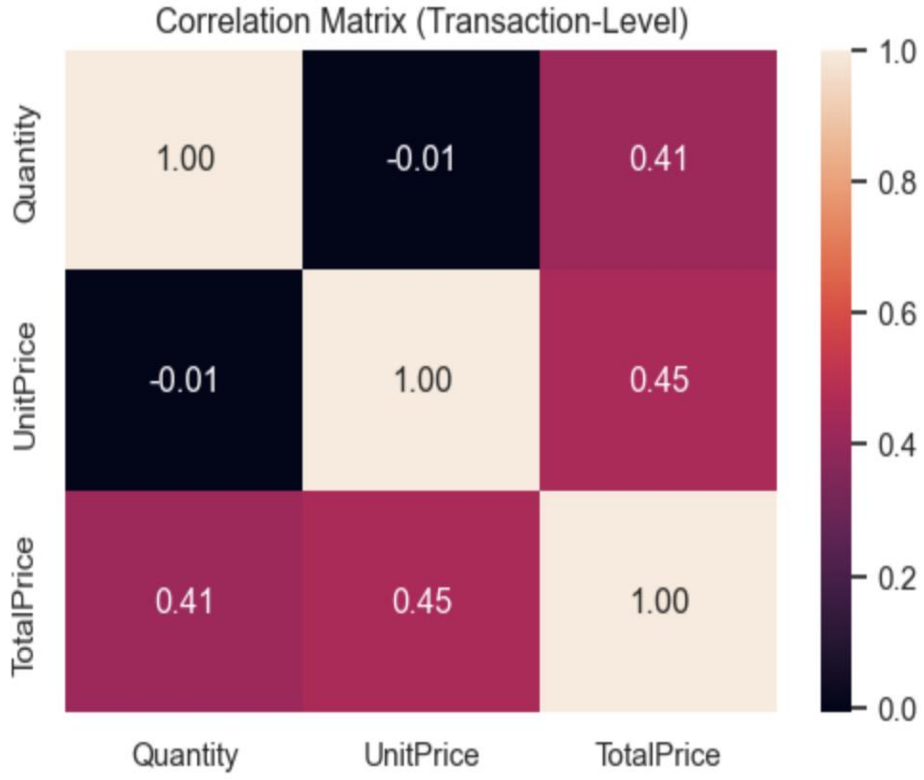


```python
# ============================================================
df_clean = df_raw.dropna(subset=["CustomerID", "Description"]).drop_duplicates()
df_clean = df_clean[(df_clean["Quantity"] > 0) & (df_clean["UnitPrice"] > 0)]

df_clean["InvoiceDate"] = pd.to_datetime(df_clean["InvoiceDate"], errors="coerce")
df_clean = df_clean.dropna(subset=["InvoiceDate"])

print("Cleaned dataset shape:", df_clean.shape)
display(df_clean.head())
```

Exploratory Data Analysis

Key Observations from EDA

- Majority of purchases involve small quantities
- Revenue distribution is right-skewed
- Certain countries dominate transaction volume
- Outliers significantly affect transaction value



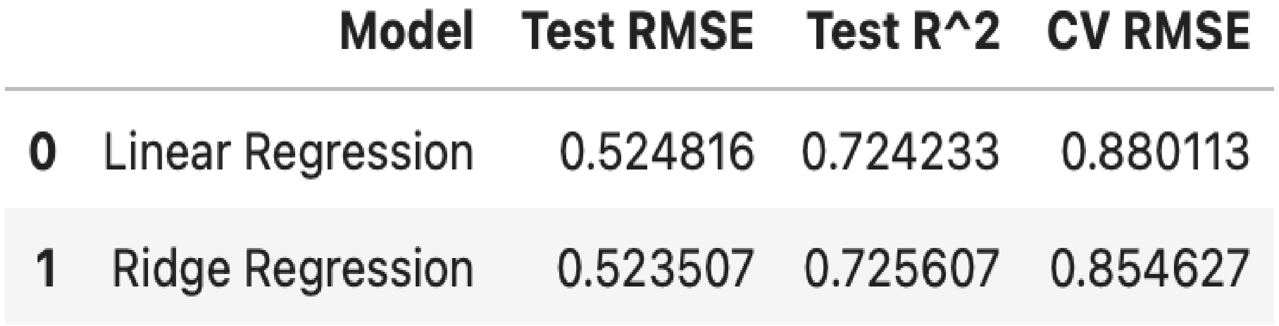
Regression Modeling

Models Developed

- Linear Regression
- Ridge / Lasso Regression

Evaluation Metrics

- R²
- MSE
- RMSE
- Cross-validation



	Model	Test RMSE	Test R^2	CV RMSE
0	Linear Regression	0.524816	0.724233	0.880113
1	Ridge Regression	0.523507	0.725607	0.854627

Model comparison table

Regression Results & Insights

Key Findings

- Regularized regression generalized better
- Linear regression was sensitive to extreme values
- Feature engineering significantly improved performance
- Cross-validation confirmed model stability

Insight: Controlling model complexity improves predictive reliability.

Classification Modeling

Objective

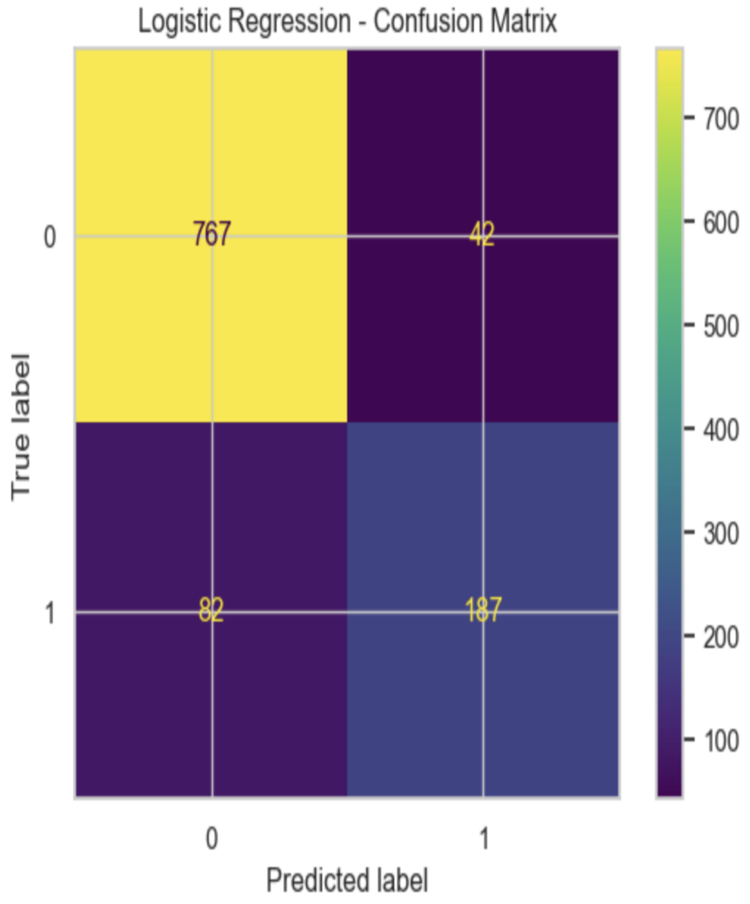
Predict categorical customer or transaction outcomes.

Models Used

- Random Forest
- Logistic Regression

Metrics

- Accuracy, Precision, Recall, F1-score, and Confusion Matrix



Confusion Matrix

Classification Insights

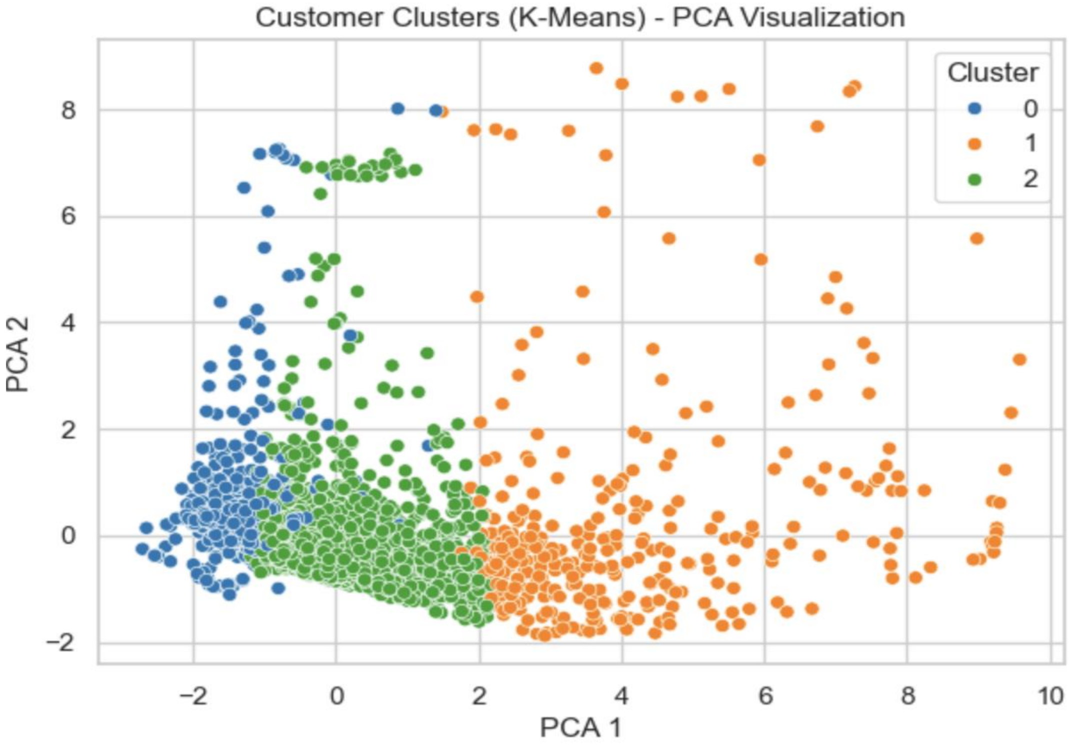
- Tree-based models captured nonlinear patterns effectively
- Class imbalance required careful handling
- Feature selection influenced classification accuracy

Insight: Behavioral patterns can be predicted with structured modeling.

Clustering Analysis

K-Means Customer Segmentation

- Identified distinct customer groups
- High-value customers form separate cluster
- Low-frequency buyers cluster separately

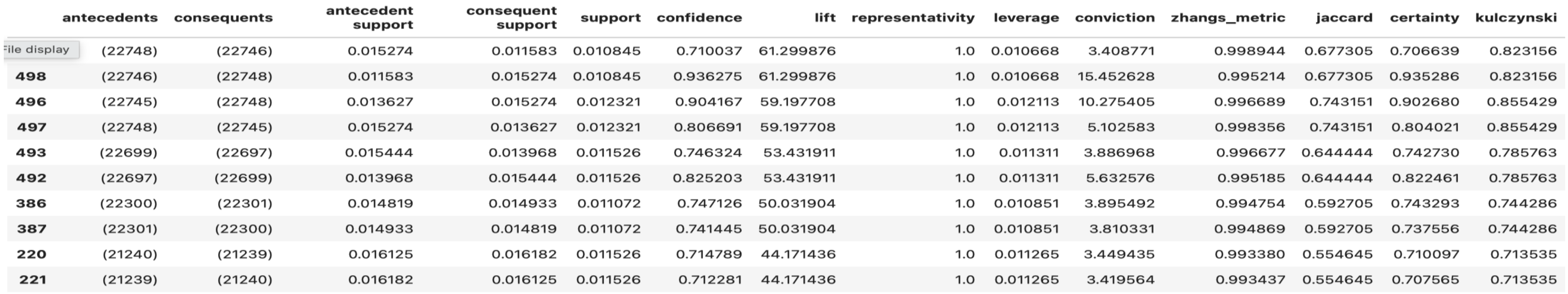


Cluster visualization plot

Association Rule Mining

Product Co-Purchase Patterns

- Applied Apriori / FP-Growth
- Evaluated support, confidence, lift
- Identified strong bundling opportunities



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
File display	(22748)	(22746)	0.015274	0.011583	0.010845	0.710037	61.299876	1.0	0.010668	3.408771	0.998944	0.677305	0.706639	0.823156
498	(22746)	(22748)	0.011583	0.015274	0.010845	0.936275	61.299876	1.0	0.010668	15.452628	0.995214	0.677305	0.935286	0.823156
496	(22745)	(22748)	0.013627	0.015274	0.012321	0.904167	59.197708	1.0	0.012113	10.275405	0.996689	0.743151	0.902680	0.855429
497	(22748)	(22745)	0.015274	0.013627	0.012321	0.806691	59.197708	1.0	0.012113	5.102583	0.998356	0.743151	0.804021	0.855429
493	(22699)	(22697)	0.015444	0.013968	0.011526	0.746324	53.431911	1.0	0.011311	3.886968	0.996677	0.644444	0.742730	0.785763
492	(22697)	(22699)	0.013968	0.015444	0.011526	0.825203	53.431911	1.0	0.011311	5.632576	0.995185	0.644444	0.822461	0.785763
386	(22300)	(22301)	0.014819	0.014933	0.011072	0.747126	50.031904	1.0	0.010851	3.895492	0.994754	0.592705	0.743293	0.744286
387	(22301)	(22300)	0.014933	0.014819	0.011072	0.741445	50.031904	1.0	0.010851	3.810331	0.994869	0.592705	0.737556	0.744286
220	(21240)	(21239)	0.016125	0.016182	0.011526	0.714789	44.171436	1.0	0.011265	3.449435	0.993380	0.554645	0.710097	0.713535
221	(21239)	(21240)	0.016182	0.016125	0.011526	0.712281	44.171436	1.0	0.011265	3.419564	0.993437	0.554645	0.707565	0.713535

Insight: Data reveals cross-selling potential.

Key Business Takeaways

- High-value customers require targeted marketing
- Product bundling can increase sales
- Feature engineering significantly improves models
- Regularization enhances generalization
- Combining multiple techniques gives deeper insight

This project moved beyond prediction to actionable strategy.

Challenges & Ethical Considerations

Challenges

- Noisy transactional data
- Outliers affecting regression
- Overfitting risk
- Feature consistency across models

Ethical Awareness

- Customer data privacy
- Avoiding demographic bias
- Transparent model evaluation

Future Improvements & Conclusion

Future Improvements

- Time-series sales forecasting
- Customer lifetime value modeling
- Deep learning approaches
- Real-time recommendation systems

Conclusion

This project demonstrates how structured preprocessing, feature engineering, supervised learning, and unsupervised learning together produce reliable and actionable retail intelligence.

Thank you…!!