Residency Project Report

Deliverables - 4

Pabitra Bhandari

Haeri Kyoung

Vamsi Krishna Gajulapalli

Prakash Tamang

An assignment submitted in partial fulfillment of the requirements for MSCS-634 as part of the degree Master of Science

In

Computer Science

School of Computing and Information Sciences

University of the Cumberlands

MSCS 634: Advanced Big Data and Data Mining

Dr. Satish Penmatsa

15th February 2026

Spring 2026

Introduction

In today's data-driven environment, organizations generate and store massive amounts of transactional information, yet the real challenge lies in turning that data into meaningful insights. Simply collecting data does not create value unless it is analyzed in a way that supports decision-making, improves operational efficiency, and enhances customer understanding. Through this project, we applied concepts learned in the Advanced Big Data and Data Mining course to explore how analytical techniques can be used to extract practical knowledge from large datasets.

Our project focuses on analyzing a real-world retail dataset to understand purchasing behavior and demonstrate how different data mining methods can be used together to solve business problems. Rather than relying on a single modeling approach, we implemented multiple analytical techniques, including regression, classification, clustering, and association rule mining, to examine the data from different perspectives. This allowed me to see how predictive analytics, pattern discovery, and segmentation complement one another in providing a more complete understanding of customer activity.

Working through this analysis reinforced the importance of the full data analytics lifecycle. We experienced firsthand how raw datasets require extensive preparation before meaningful modeling can begin, and how careful feature engineering can significantly influence analytical outcomes. The project also highlighted the need to interpret model results in a business context rather than treating them solely as technical outputs.

Overall, this work represents an applied learning experience in which theoretical concepts from the course were translated into practical analysis, reflecting how organizations use big data tools to guide strategic and operational decisions.

Dataset Description

For this project, we worked with the **Online Retail II dataset**, which contains transactional data from a United Kingdom–based online retailer. The dataset includes detailed records of individual purchases, such as invoice numbers, product identifiers, descriptions, quantities, pricing, transaction timestamps, and anonymized customer IDs. Each row represents a specific product purchased within a transaction, enabling analysis at both the transaction and aggregated customer levels.

We selected this dataset because it closely reflects the type of real-world data organizations manage in e-commerce and digital sales environments. Unlike simplified academic datasets, this data contains hundreds of thousands of records and includes many of the challenges typically encountered in practice, such as missing values, duplicate entries, and transaction reversals. Working with this dataset provided an opportunity to apply preprocessing and data quality techniques that are essential to any meaningful analysis.

Another reason this dataset was appropriate for the project is that it supports multiple types of analysis. The transactional structure makes it suitable for revenue prediction and purchasing trend analysis, while the presence of customer identifiers enables behavioral modeling, such as segmentation and identification of high-value customers. Additionally, the product-level detail allows for association rule mining to discover relationships among frequently purchased items.

Using this dataset allowed me to simulate a realistic analytics scenario in which raw operational data must be transformed, analyzed, and interpreted to generate insights that could support business strategy, marketing decisions, and customer engagement initiatives.

Data Preparation and Cleaning

Before performing any analysis, it was necessary to prepare the dataset to ensure accurate, meaningful results. Real-world data is rarely ready for immediate use, and this dataset was no exception. During the initial review, we observed missing customer identifiers, incomplete product descriptions, duplicate transactions, and records representing returned items rather than completed purchases. These issues needed to be addressed to prevent misleading conclusions during modeling.

The first step was to remove records that lacked a valid Customer ID or product description. Since the project included customer-level analysis and purchasing behavior, transactions without these fields could not contribute to meaningful insights. We then identified and removed duplicate entries to ensure that revenue and transaction counts were not overstated.

Next, transactions with negative quantities were excluded. These records represented product returns or cancellations rather than actual sales, and including them would distort spending patterns and predictive models. Similarly, transactions with zero or negative pricing values were filtered out to retain only legitimate purchase activity.

After cleaning the dataset, we standardized column names and converted date fields into a proper datetime format. This step enabled time-based analysis and allowed additional features to be extracted later in the project. These preparation steps significantly improved the dataset's reliability and established a consistent foundation for exploratory analysis and modeling.

Through this process, we gained practical experience in recognizing that data preparation is often the most time-consuming yet critical stage of analytics, as the quality of insights depends directly on the quality of the underlying data.

Exploratory Data Analysis (EDA)

Once the dataset was cleaned, we conducted exploratory data analysis to better understand purchasing patterns, data distribution, and potential relationships among variables. This step was important for identifying trends that could guide feature engineering and model selection.

The analysis showed that most transactions involved relatively small purchase quantities, while a small number involved bulk purchases. This created a right-skewed distribution, common in retail environments where a few high-volume orders account for a significant share of sales. A similar pattern emerged in transaction revenue, with a limited number of purchases accounting for disproportionately high values. Because of this skewness, a logarithmic transformation was later applied to stabilize the data for regression modeling.

Geographic analysis revealed that the majority of transactions originated from the United Kingdom, indicating that the retailer's customer base was highly concentrated in one region. This insight suggested that purchasing trends and promotional strategies would likely be influenced by regional demand patterns.

Time-based analysis also revealed noticeable variations in purchasing activity across months, suggesting seasonal effects in customer behavior. Such trends are valuable for businesses because they can inform inventory planning and targeted marketing campaigns during peak demand periods.

Correlation analysis further confirmed a strong relationship between purchase quantity and total transaction value, consistent with expectations, since revenue is directly influenced by how many items customers buy.

During this exploratory phase, we moved beyond simply viewing the dataset and began to understand the behavioral patterns embedded within it. These insights directly informed the feature engineering process and helped shape the modeling strategies used later in the project.

Feature Engineering

After understanding the data's structure and behavior, we developed additional features to better represent customer activity and purchasing patterns. Feature engineering was necessary because raw transactional data often lacks sufficient context for predictive modeling. By transforming existing variables and creating new ones, we were able to capture more meaningful behavioral information.

One of the first features created was the total transaction value, calculated by multiplying quantity and unit price. This variable represented the monetary impact of each purchase and served as the primary outcome variable for regression analysis. Because revenue values were highly skewed, we applied a logarithmic transformation to reduce the influence of extreme values and improve model stability.

We also extracted several time-based features from the transaction timestamp, including the month, day of the week, and hour of purchase. These variables allowed the models to account for temporal patterns such as seasonal demand and purchasing behavior at different times.

To better capture shopping behavior within individual orders, we generated basket-level features, including the number of unique items per invoice and the total quantity purchased per transaction. These features helped describe whether a customer was making a small targeted purchase or a larger bulk order.

At the customer level, we constructed behavioral metrics based on Recency, Frequency, and Monetary value (RFM). These measures are summarized as the customer's last purchase

date, how often they bought products, and how much they spent overall. RFM-style features are widely used in industry because they provide an interpretable way to measure engagement and customer value.

This feature engineering process enabled the models to analyze behavior rather than just transactions, making subsequent regression, classification, and clustering analyses more informative and better aligned with real business applications.

Modeling Approaches

After understanding the data's structure and behavior, we developed additional features to better represent customer activity and purchasing patterns. Feature engineering was necessary because raw transactional data often lacks sufficient context for predictive modeling. By transforming existing variables and creating new ones, we were able to capture more meaningful behavioral information.

One of the first features created was the total transaction value, calculated by multiplying quantity and unit price. This variable represented the monetary impact of each purchase and served as the primary outcome variable for regression analysis. Because revenue values were highly skewed, we applied a logarithmic transformation to reduce the influence of extreme values and improve model stability.

We also extracted several time-based features from the transaction timestamp, including the month, day of the week, and hour of purchase. These variables allowed the models to account for temporal patterns such as seasonal demand and purchasing behavior at different times.

To better capture shopping behavior within individual orders, we generated basket-level features, including the number of unique items per invoice and the total quantity purchased per

transaction. These features helped describe whether a customer was making a small targeted purchase or a larger bulk order.

At the customer level, we constructed behavioral metrics based on Recency, Frequency, and Monetary value (RFM). These measures are summarized as the customer's last purchase date, how often they bought products, and how much they spent overall. RFM-style features are widely used in industry because they provide an interpretable way to measure engagement and customer value.

This feature engineering process enabled the models to analyze behavior rather than just transactions, making subsequent regression, classification, and clustering analyses more informative and better aligned with real business applications.

Classification Modeling

In addition to predicting transaction value, the project aimed to identify customers who contribute the most to overall revenue. This type of analysis is commonly used in industry to support targeted marketing, retention strategies, and personalized services. To accomplish this, I formulated a classification problem in which customers were labeled as “high value” if their total spending fell within the top quartile of all customers.

To build this model, I aggregated the transactional data to the customer level and used behavioral features such as Recency, Frequency, average purchase quantity, average unit price, and the number of unique products purchased. These variables provided a summary of each customer’s engagement and purchasing habits.

Two classification algorithms were implemented: Logistic Regression and Random Forest. Logistic Regression served as a baseline model that provided interpretability and showed how individual features influenced the probability that a customer was high value. Random

Forest, an ensemble learning method, was used to capture more complex, nonlinear relationships within the data.

Model performance was evaluated using accuracy, F1-score, and ROC curve analysis. The Random Forest model achieved stronger performance, indicating that customer behavior is influenced by interactions among multiple variables rather than simple linear relationships. This result reinforced the value of using ensemble models when analyzing behavioral data.

This phase demonstrated how classification techniques can help organizations move beyond descriptive analytics to actionable decision-making, such as identifying which customers should receive loyalty incentives or targeted promotions.

Clustering Analysis

While classification required predefined labels, clustering was used to discover natural groupings within the customer base without prior assumptions. This approach is valuable because it allows patterns to emerge directly from the data rather than being imposed through predefined categories.

Using the same customer-level behavioral features (Recency, Frequency, Monetary value, average quantity, average unit price, and number of unique products), I applied the K-Means clustering algorithm to segment customers into distinct groups. Before clustering, the data was standardized to ensure that variables with larger numerical ranges did not dominate the segmentation process.

The clustering results revealed several distinct behavioral patterns. One group consisted of highly engaged customers who purchased frequently and generated substantial revenue. Another cluster included low-frequency customers with higher recency values, suggesting

limited or declining engagement. Additional clusters represented moderate spenders and occasional buyers with lower transaction activity.

To better interpret these segments, I visualized the clusters using Principal Component Analysis (PCA), which reduced the feature space to two dimensions while preserving variance. This visualization helped confirm that the clusters were meaningfully separated and not randomly assigned.

From a business perspective, these segments provide practical insights. High-value, high-frequency customers could be prioritized for loyalty programs, while less engaged customers might benefit from re-engagement campaigns. Clustering, therefore, complements classification by offering a broader view of customer behavior patterns across the entire population.

Association Rule Mining

The final stage of the analysis focused on identifying relationships among products that were frequently purchased together. Unlike regression or classification, which attempt to predict outcomes, association rule mining is used to discover hidden patterns within transactional data. This technique is commonly applied in retail environments to support recommendation systems, product placement strategies, and cross-selling opportunities.

To perform this analysis, I applied the Apriori algorithm to transaction data, focusing on purchases made within the United Kingdom to maintain a manageable dataset size while still capturing meaningful behavior. The transactional data was transformed into a basket format, where each invoice represented a set of purchased items. This structure enabled the algorithm to assess how often products co-occurred across transactions.

Using support, confidence, and lift metrics, the model identified product combinations with strong purchasing relationships. High-support itemsets indicated commonly purchased

product groups, while high-lift rules revealed associations that occurred more frequently than expected by chance. These findings suggest opportunities to bundle related products, improve product recommendations, and design marketing strategies that encourage complementary purchases.

During this phase, I observed that pattern mining provides a different type of insight than predictive modeling. Rather than forecasting behavior, it helps explain how customers naturally combine products, offering actionable intelligence for merchandising and sales optimization.

Key Insights and Recommendations

The objective of this project was to analyze transactional retail data to understand purchasing behavior, predict transaction value, identify high-value customers, segment customers into meaningful groups, and uncover product relationships. By integrating regression, classification, clustering, and association rule mining, the analysis provided a comprehensive view of both customer spending and behavior.

Key Insights from Data Preparation and Exploration

Data preprocessing revealed several important characteristics of the dataset. A significant portion of the raw records contained missing customer identifiers, duplicate transactions, or negative quantities representing product returns. Removing these records was essential to ensure that the analysis reflected true purchasing behavior rather than operational noise.

Exploratory analysis showed that:

- Transaction values were **highly right-skewed**, meaning a small number of purchases contributed disproportionately to total revenue.
- Sales activity displayed **seasonal variation**, indicating that customer demand fluctuated over time.

- Revenue was **geographically concentrated**, with the United Kingdom accounting for the majority of transactions.
- Log-transforming transaction values significantly improved model stability by reducing the influence of extreme outliers.

These findings justified the feature engineering strategy used in later modeling stages.

Insights from Regression Modeling

Linear Regression and Ridge Regression were used to predict the log-transformed transaction value.

- Both models achieved **similar predictive performance**, with Ridge Regression slightly outperforming Linear Regression.
- The comparable results suggest that the engineered features (quantity, unit price, time features, and invoice-level aggregates) already captured most of the explainable variance.
- Regularization in Ridge Regression helped reduce minor overfitting, as reflected in improved cross-validation RMSE.

Interpretation:

Transaction value in this dataset is largely driven by straightforward behavioral signals such as basket size and purchase timing, rather than complex nonlinear relationships.

Insights from Classification Modeling

Classification models were developed to identify **high-value customers**, defined as those in the top 25% of total spending.

- Logistic Regression provided strong baseline performance with high accuracy and F1-score.

- Random Forest achieved near-perfect classification performance, indicating clear separation between high-value and low-value customer groups based on RFM-style features (Recency, Frequency, Monetary).
- Features such as **purchase frequency and cumulative spending** were the strongest indicators of customer value.

Interpretation:

High-value customers exhibit consistent purchasing patterns that can be reliably detected using behavioral metrics, making them suitable targets for loyalty programs and retention strategies.

Insights from Customer Segmentation (Clustering)

K-Means clustering identified distinct customer segments:

- A large segment consisted of **moderate-frequency customers with steady spending**.
- A small but critical segment represented **extremely high-value customers with very large purchase volumes**.
- Another segment showed **infrequent or dormant customers**, suggesting churn risk.

Interpretation:

Customer behavior is not uniform. Segment-based strategies are more appropriate than one-size-fits-all marketing approaches.

Insights from Association Rule Mining

A priori analysis uncovered recurring product combinations purchased together.

- Certain items consistently appeared in the same baskets, indicating natural product affinities.

- These associations provide a data-driven foundation for **product bundling and recommendation systems**.

Interpretation:

Customers often purchase complementary goods together, enabling retailers to design cross-selling strategies that increase average order value.

Practical Recommendations

Based on the analytical results, the following recommendations can be made:

1. **Implement Customer Segmentation Strategies**

   Tailor marketing campaigns to specific clusters, particularly focusing on retaining high-value customers and re-engaging dormant ones.

2. **Develop Loyalty Programs for High-Value Customers**

   Classification models can proactively identify valuable customers and support targeted incentives, improving long-term retention.

3. **Use Basket Insights for Cross-Selling**

   Association rules can inform product placement, bundling, and recommendation engines to increase transaction size.

4. **Leverage Predictive Models for Revenue Forecasting**

   Regression models can help estimate expected transaction values, supporting financial planning and demand forecasting.

5. **Focus on Behavioral Features in Future Analytics**

   Features derived from purchase behavior (frequency, diversity, basket size) proved more informative than raw transactional attributes.

Ethical Considerations

While this project focused on technical modeling and data-driven insights, it is equally important to consider the ethical implications of analyzing customer transaction data. Responsible data use ensures that analytical outcomes benefit organizations while protecting individuals and avoiding unintended harm.

Data Privacy and Confidentiality

The dataset used in this project contains transactional information linked to anonymized customer identifiers rather than personally identifiable information (PII). However, even anonymized behavioral data can reveal sensitive patterns about customer habits.

To address privacy concerns:

- The analysis relied solely on **aggregated behavioral features** (e.g., spending totals, purchase frequency) rather than on individual-level data.
- No effort was made to re-identify customers or merge this dataset with external data sources.
- Customer identifiers were treated strictly as analytical keys for grouping transactions.

These steps align with responsible analytics practices and principles outlined in modern data protection frameworks such as GDPR-style data minimization.

Bias and Representativeness

The dataset is heavily dominated by transactions from the United Kingdom. As a result:

- Models trained on this data may not generalize well to other geographic regions or retail environments.

- Customer segmentation and purchasing patterns reflect a **specific market context**, not universal behavior.

Recognizing this limitation prevents misapplication of the results to populations that were not represented in the data.

Fairness in Predictive Modeling

Classification models were used to identify “high value” customers. While such models are valuable for business decision-making, they can introduce fairness concerns if used improperly.

For example:

- Over-targeting high-value customers could unintentionally neglect new or lower-spending customers.
- Automated decision systems could reinforce existing spending inequalities.

To mitigate this risk, predictive outputs should **support**, not replace, human decision-making.

Responsible Use of Insights

The insights from clustering and association rules could be used to influence purchasing behavior through personalized marketing. Ethical use requires that:

- Recommendations remain transparent and non-manipulative.
- Customers retain autonomy in purchasing decisions.
- Data is not used to exploit vulnerable groups.

Analytics should aim to enhance customer experience rather than pressure or mislead consumers.

Project-Level Ethical Safeguards

During this project, the following safeguards were maintained:

- Only publicly available academic data was used.
- No sensitive attributes (e.g., demographics, income, or identity markers) were modeled.
- Results were interpreted at a **behavioral trend level**, not at an individual level.
- Models were evaluated strictly for analytical understanding, not deployment.

Conclusion of Ethical Review

Ethical data science is not only about achieving accurate models but also ensuring that insights are applied responsibly, transparently, and within the context of the dataset's limitations. By acknowledging privacy, bias, and fairness considerations, this project demonstrates an approach that balances analytical rigor with ethical awareness.

Conclusion

This project explored how data mining and machine learning techniques can be applied to real-world retail transaction data to extract meaningful business insights. Using the **Online Retail II dataset**, the analysis followed a structured pipeline: data preprocessing and exploratory analysis, predictive modeling, customer segmentation, and pattern discovery.

The preprocessing phase played a critical role in ensuring analytical reliability. Removing missing customer identifiers, duplicate transactions, return records, and invalid pricing entries yielded a clean dataset that reflects true purchasing behavior. Feature engineering—such as transaction-level revenue (TotalPrice), temporal variables (Month, DayOfWeek, Hour), and invoice-level basket characteristics—enabled models to better capture behavioral patterns embedded in the data.

Regression modeling demonstrated that transaction value can be predicted with reasonable accuracy using engineered behavioral features. Both **Linear Regression** and **Ridge Regression** achieved similar performance, with Ridge slightly outperforming due to its regularization capability, which helped stabilize coefficients and improve generalization, as shown by cross-validation results.

Classification analysis shifted the focus from transactions to customers. By constructing RFM-style behavioral metrics (Recency, Frequency, Monetary), the project successfully identified high-value customers. The **Random Forest** classifier achieved strong predictive performance, highlighting that customer value can be inferred from purchasing patterns without requiring demographic data.

Unsupervised learning further expanded the understanding of customer behavior. **K-Means clustering** revealed distinct customer segments ranging from loyal high-frequency buyers to low-engagement customers. These segments provide actionable insights for targeted marketing, retention strategies, and resource allocation.

Finally, **association rule mining** uncovered frequently co-purchased products, offering insights into basket composition and opportunities for product bundling or recommendation systems.

Together, these approaches demonstrate that combining supervised, unsupervised, and pattern-mining techniques creates a comprehensive analytical framework capable of supporting real-world retail decision-making.

Future Work

While the project achieved its analytical objectives, several extensions could enhance both accuracy and applicability:

1. **Incorporating Additional Time-Based Modeling**

   Seasonal decomposition and time-series forecasting (e.g., ARIMA or Prophet) could better capture demand fluctuations and improve sales prediction.

2. **Advanced Machine Learning Models**

   Gradient boosting models such as XGBoost or LightGBM may improve predictive performance compared to linear methods while handling non-linear relationships more effectively.

3. **Customer Lifetime Value (CLV) Modeling**

   Extending the classification framework to estimate long-term customer value would allow organizations to prioritize retention strategies more strategically.

4. **Scalable Recommendation Systems**

   Association rules could be integrated into a collaborative filtering or hybrid recommendation engine to deliver personalized product suggestions.

5. **Geographic and Demographic Enrichment**

   Adding external contextual data could improve segmentation accuracy and reduce the regional bias observed in this dataset.

6. **Model Deployment Considerations**

   Future work could include building a production pipeline using APIs or dashboards to operationalize insights for business stakeholders.

References

Agrawal, R., & Srikant, R. (1994). *Fast algorithms for mining association rules*. Proceedings of the 20th International Conference on Very Large Databases.

Chen, D., Sain, S. L., & Guo, K. (2012). *Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining*. Journal of Database Marketing & Customer Strategy Management.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

McKinney, W. (2010). *Data structures for statistical computing in Python*. Proceedings of the 9th Python in Science Conference.

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