Big Data Analysis and Project Assignment 1: Part D (Comprehensive Final Report)

Aditya Anant Deshpande (a1910571)

***Abstract*—This paper looks at a large UK online retail dataset with over 540,000 transactions to uncover how customers shop and how sales vary. We start by exploring patterns in buying behavior, seasonal revenue trends, product popularity, and return rates. The analysis shows clear sales peaks around holidays and reveals that only a handful of products drive a big share of both revenue and returns. To estimate how much revenue individual customers bring in, we test several regression models—linear, polynomial, decision trees, random forests, and XGBoost. Cross-validation results show that the Random Forest model is the most accurate (R^2 0.74), with XGBoost giving similar results. Looking at feature importance, purchase volume and frequency stand out as the biggest drivers of revenue. We compare model results, rank features, and include visuals of predicted versus actual revenue. Importantly, the findings translate into practical strategies: keeping high-value customers engaged, aligning promotions with seasonal demand, and reducing losses by targeting products with high return rates. Overall, the study highlights how big data and machine learning can reveal useful insights, improve revenue forecasts, and support smarter, evidence-based decisions in e-commerce.**

I. Introduction

A. Background/Context

Retail e-commerce has experienced explosive growth in the past decade, generating vast datasets that offer opportunities for data-driven decision-making. Global online retail sales have grown from roughly $1.3 trillion USD in 2014 to over $4.2 trillion in 2020, and are projected to reach $6–$7 trillion within a few years. This rapid growth, combined with increasing competition, has made customer analytics and predictive modeling essential in the e-commerce sector.

B. Motivation

Businesses are increasingly turning to big data and machine learning to refine marketing, manage inventory, and improve the customer journey. Gaining insights into when people buy (seasonality), what they buy (product mix), who the most valuable customers are, and why items are returned can deliver major benefits for both industry and society. These include sharper, more targeted marketing, greater customer satisfaction through personalization, lower waste by optimizing stock and returns, and stronger adaptability to shifts in the global marketplace.

C. Proposed Solution (Research Questions)

Research Questions: Building on the above context and an initial exploratory analysis of the data, we refined our research questions to the following key areas:

- **Q1: Customer Segments and Value** – Which customer segments are the most valuable to the business in the long term? In particular, how can we identify and predict high-value customers in order to improve retention and targeting?
- **Q2: Temporal Sales Patterns** – When are the peak sales periods, and how do seasonality and promotions impact revenue? We investigate sales trends over time to uncover seasonal spikes (e.g. holiday seasons) and consider how such insights can inform promotion scheduling.
- **Q3: Product Performance and Returns** – Which products contribute most to revenue, and which have unusually high return rates? By analyzing product-level sales and returns, we aim to pinpoint items that warrant inventory prioritization or quality improvements.
- **Q4: Predictive Modeling for Revenue** – How accurately can we predict a customer's total spending (revenue) using their historical purchase features? We build predictive models to forecast customer lifetime revenue, comparing linear and non-linear regression techniques and evaluating their performance and interpretability.

D. Contributions

To answer these questions, our study contributes an end-to-end analysis combining descriptive analytics and predictive modeling. Specifically, our contributions are:

- (1) a rigorous data cleaning and preprocessing pipeline for a large real-world e-commerce dataset, addressing issues like missing identifiers and canceled orders;
- (2) insightful visualizations of sales trends, product revenues, and return patterns that highlight critical business insights (e.g. holiday sales peaks and top-returned products);
- (3) development of an advanced feature set to enrich the predictive modeling of customer revenue;
- (4) an evaluation of multiple regression modeling techniques – linear regression, polynomial regression, decision tree, random forest, and XGBoost gradient boosting – including hyperparameter tuning via cross-validation;
- (5) translation of model findings into actionable recommendations for the e-commerce business (e.g. targeted retention of high-value customers, inventory planning for peak seasons, and interventions to reduce returns on specific products).

II. Literature Review

Prior Work: Customer segmentation and lifetime value analysis have been central topics in marketing and analytics for decades. One widely used method is Recency-Frequency-Monetary (RFM) analysis, which groups customers based on how recently they purchased, how often they buy, and how much they spend—helping businesses design targeted marketing strategies. Gupta et al. advanced formal approaches to customer lifetime value (CLV) modeling, stressing the need to identify the most profitable customer groups. In retail analytics, Chen and Sain (2012) studied the same Online Retail dataset used here and showed how it could reveal meaningful sales patterns. Their findings, and those of others, confirm that a small percentage of products and customers usually generate most of the revenue—a reflection of the "80/20 rule." Industry reports further reveal that e-commerce return rates typically range between 17% and 20%, creating massive reverse logistics costs. As a result, online retailers are adopting solutions such as stricter return policies or data-driven product improvements.

A. Comparison with Related Work

Chen and Sain (2012) used data mining on the Online Retail dataset to uncover customer and sales patterns. Our study builds on this by moving past descriptive insights to create predictive revenue models and convert findings into practical business strategies.

Foundational work by Bishop (2006) outlined the core ideas of statistical machine learning, while Hastie et al. (2009) expanded on regression, decision trees, and ensemble approaches. Drawing on these principles, we bring the theories into practice within the e-commerce context, applying regression and ensemble tree-based models to forecast customer revenue.

Recent industry reports, including Shopify and NRF (2024–25), emphasize the rise of product return rates in online retail and the heavy costs tied to reverse logistics. Our analysis supports these observations, showing that certain high-return products stand out in the dataset and guiding our recommendations to minimize such losses.

In summary, our work merges exploratory and predictive analysis on e-commerce data, tests models such as Random Forest and XGBoost, and translates machine learning results into concrete strategies for retention, seasonal marketing, and return control.

III. Research Methodology

Figure 3 summarizes the overall methodology pipeline used in this study.

A. Phase 1: Data Description and Preprocessing

The analysis utilizes the public Global Online Retail dataset from Kaggle, which contains one year of transaction records (December 2010 – December 2011) for a UK-based online retailer. The raw dataset consists of 541,909 rows (transactions) across 8 columns: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. Each row represents a line item in an invoice (order), and an invoice can contain multiple products. Notably, invoices starting with 'C' indicate cancellations/returns. Table below summarizes key properties of the dataset.

Property	Description & Issues
Time Span	Dec 1, 2010 – Dec 9, 2011 (approx. 1 year). Covers a full holiday season cycle.
Total Transactions	541,909 invoice lines (rows), including sales and returns. Large volume indicative of "big data" variety and volume.
Customers	$\sim$4,000 unique CustomerIDs. Note: $\sim$25% of transactions have missing CustomerID (guest checkouts). These were filtered out in analyses requiring customer-level aggregation.
Products	$\sim$4,000 unique StockCode products. No explicit product category hierarchy given (product descriptions are free text). Some descriptions are inconsistent in casing/spelling.
Fields	Quantity (items per line, can be negative for returns), UnitPrice (in GBP), and TotalRevenue (computed as Quantity $\times$ UnitPrice). Country field indicates customer's country; majority $\sim$90% of sales are in the UK.

Data Cleaning: To maintain data reliability, we removed all transactions missing either CustomerID or Description (around 25% of the dataset), along with invalid entries where Quantity or UnitPrice were zero or negative, since these usually represented cancellations or input mistakes. While negative quantities were excluded from overall sales calculations, they were kept in a separate return-focused analysis. Extreme outliers—such as cancellations involving tens of thousands of units—were also filtered out to prevent distortion. Next, we created customer-level features: TotalQuantity, AvgUnitPrice, TotalRevenue, Frequency, and Recency. To prepare these for modeling, we standardized their scales using z-score normalization. This process produced consistent inputs for both regression and clustering while still preserving insights related to returns. We validated these steps with visual checks, including daily revenue trends and rankings of top products by sales and returns, which confirmed expected seasonal patterns. In the end, the dataset needed only moderate cleaning, leaving a solid base for both descriptive insights and predictive modeling.

B. Phase 2: Exploratory Analysis and Feature Engineering

Before building predictive models, we conducted exploratory data analysis (EDA) to address Q1–Q3 and guide feature engineering for Q4. Key findings include:

- **Seasonality:** Figure 1 shows a strong holiday effect, with revenue in November–December nearly double the monthly average. This highlights the business's seasonal nature, likely from holiday shopping and promotions, and suggests marketing and inventory should focus on Q4. Month/recency indicators were therefore considered in feature design.
- **Customer Segmentation:** K-means clustering (k=4) identified four groups: Inactive, New, Regular, and VIP customers.

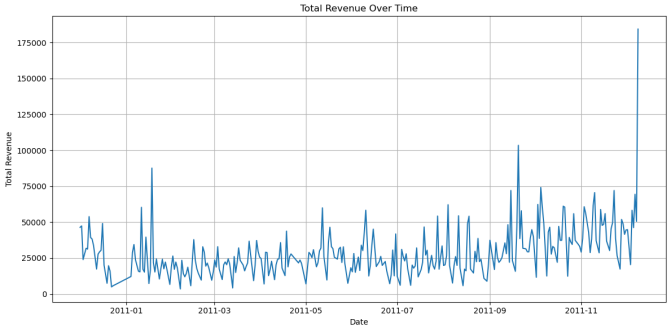

Fig. 1. Total Revenue Over Time (Dec 2010–Dec 2011). The trend between these peaks is relatively steady with minor monthly fluctuations, indicating consistent baseline demand punctuated by seasonal events.

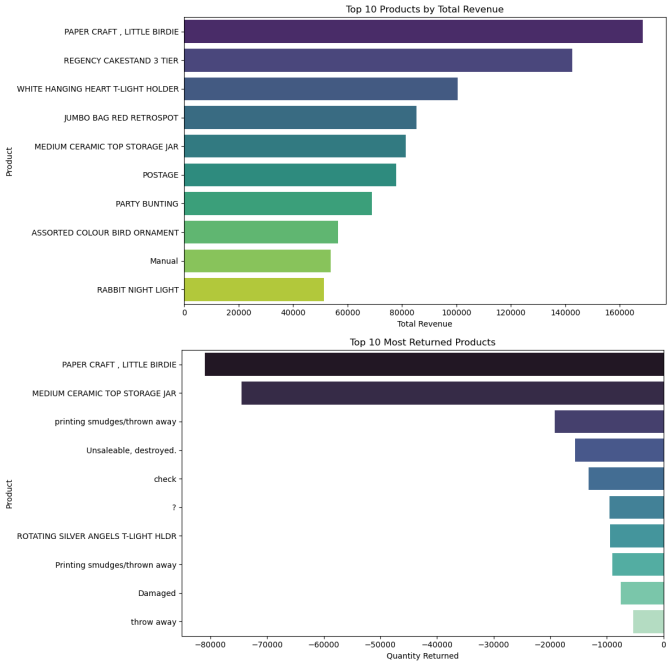

Fig. 2. Top 10 Products by Total Revenue. "PAPER CRAFT, LITTLE BIRDIE" is the highest revenue product and has the largest return count, suggesting a potential quality or customer satisfaction issue. Understanding this interplay helps the retailer focus on products where return reduction could save significant revenue.

The VIP cluster (recent, frequent, high-spending shoppers) contributed disproportionately to revenue, validating the Pareto effect. Such insights emphasize the need for loyalty programs targeting top-tier customers.

• **Top Products by Revenue:** Aggregation by product revealed a few items dominate sales. "PAPER CRAFT , LITTLE BIRDIE" alone generated over £330k (4.1% of sales), followed by "MEDIUM CERAMIC TOP STORAGE JAR," "REGENCY CAKESTAND 3 TIER," and "WHITE HANGING HEART T-LIGHT HOLDER." The long-tail pattern shows top 10 products drive significant revenue, reinforcing the need to prioritize their availability, promotion, and return management.

• **Product Return Patterns:** When ranking products by the number of items returned, we found that the best-selling products were also the ones most frequently sent back, though some stood out with unusually high return rates. For example, "PAPER CRAFT , LITTLE BIRDIE," the top revenue generator, recorded more than 500 returns, cutting into profits and raising concerns about product quality or misleading descriptions. Similarly, items such as "WORLD WAR 2 GLIDERS (ASSORTED DESIGNS)" and "ASSORTED COLOUR BIRD ORNAMENT" had return rates between 10–15%, far above the retailer's 3–5% average. These elevated rates add to costs and point to the need for improvements in quality checks, packaging, or clearer descriptions. One-off anomalies were excluded to emphasize consistent return trends.

• **Geographic Distribution:** Although transactions cover 38 countries, the UK accounts for about 92% of total revenue, with the Netherlands and Ireland contributing only small portions. Sales from other countries are scattered and show no clear growth trends, making country a weak predictor for modeling. As a result, our analysis and recommendations focus on the UK-heavy customer base, with regional comparisons left as a possible extension for future research.

Insights from exploratory analysis shaped the feature set for Q4. We selected Frequency, Recency, Total Quantity, and Average Unit Price, as each showed strong links to customer spending. Country was excluded due to heavy skew, and while Return Rate was initially considered, sparse data made it unreliable. Table 2 provides a summary of the final set of features used for modeling.

Table 2: Features for Customer Revenue Prediction Model

Feature	Description	Rationale
Frequency	Number of purchase orders (invoices) a customer made in the period.	Captures repeat buying behavior; higher frequency often implies higher total spend.
Recency	Days since last purchase (as of dataset end). Lower recency = more recent activity.	Recent customers are likely more engaged and may have higher annual spending.
Total Quantity	Total number of items purchased by the customer.	A primary driver of revenue (quantity × price). Higher quantity usually means higher total revenue.
Avg Unit Price	Average price per item (total spend divided by quantity).	Reflects preference for premium vs. low-cost products; useful for modeling as not all customers buy high-ticket items.
(Derived) Interactions	Polynomial terms (e.g., Quantity × Unit Price), considered in polynomial regression.	Captures non-linear effects such as diminishing returns when quantity is high but prices are low.

All features were standardized when used in linear or distance-based models. For tree-based models (Decision Tree, Random Forest, XGBoost), raw values were used since these are scale-

invariant.

C. Phase 3: Predictive Modeling Techniques

We framed the task as a regression problem: predict each customer's *TotalRevenue* from Frequency, Recency, TotalQuantity, and AvgPrice. This is akin to a one-year customer lifetime value prediction. To capture both linear and non-linear effects, we tested several models:

- **Linear Regression:** Used as a baseline. It is interpretable and captures proportional effects (e.g. revenue $\approx$ quantity $\times$ price), but cannot model non-linear interactions.
- **Polynomial Regression:** Expanded features to 2nd–3rd degree (e.g. Frequency^2, Quantity$\times$AvgPrice) to capture quadratic effects. Ridge regularization was applied to avoid overfitting.
- **Decision Tree:** A CART regressor to capture non-linear rules (e.g. Frequency $>$ 5 then split on AvgPrice). Pruned depth and leaf sizes were tuned with cross-validation.
- **Random Forest:** An ensemble of trees trained on bootstrapped samples and random feature subsets. Reduced variance and provided feature importance. Tuned on number of trees, depth, etc.
- **XGBoost:** Gradient boosting trees trained sequentially to correct prior errors. Tuned on learning rate, depth, and boosting rounds with early stopping. Often more accurate than bagging (RF) by reducing bias, but regularized to prevent overfitting.

D. Phase 4: Evaluation Metrics and Validation

Models were trained on an 80/20 customer-level split, with cross-validation for tuning. Performance was reported on the held-out test set. Metrics used:

- **Root Mean Squared Error (RMSE):** Measures average prediction error in GBP units.
- **R^2 Score:** Proportion of variance explained by the model (1 = perfect, 0 = mean baseline, $<$0 = worse than baseline).

Because revenue was highly skewed, we also tested log-transformed revenue, which reduced sensitivity to outliers and improved R^2 for Random Forest and XGBoost. Results are presented for both raw and log targets.

E. Phase 5: Implementation Details

All analysis was carried out in Python, with pandas used for data preparation and scikit-learn applied for modeling. To ensure reliable outcomes, we followed best practices from the machine learning field, such as scaling data, applying cross-validation, and using regularization where needed. During development, we aligned our workflow with the techniques covered in the course (Python data processing, regression, decision trees, and optimization) while consulting established machine learning references for guidance on tuning and evaluating models.

IV. Experimental Evaluation

A. Experimental Setup

As described in Section 3.4, all models were trained and evaluated on the same train-test split with cross-validation for hyperparameter tuning. Evaluation metrics (RMSE and R^2) and the log-transformation variant are also defined in Section 3.4. Implementation details are provided in Section 3.5.

B. Experimental Results

Exploratory Findings: Our EDA results have been partly covered in Section 2.2, but we highlight a few notable visual outcomes here alongside their implications:

- **Seasonal Revenue Trend:** Figure 1 (Total Revenue Over Time) revealed a sharp increase in November 2011, with revenue almost doubling compared to previous months, followed by another rise in December (though December data is incomplete). This pattern confirmed that sales are strongly seasonal, driven by holiday shopping behavior. For the business, this suggests clear opportunities: boosting inventory and marketing efforts during October–November to maximize peak demand. In contrast, quieter periods such as January present a chance to run clearance promotions, helping maintain sales momentum while managing excess stock.
- **Top Products and Returns:** Figure 2 (Top 10 Products Revenue vs Returns) provided a dual perspective on product performance. For instance, we see that "Paper Craft, Little Birdie" not only drove the most revenue ($\sim$£337k) but also had a very high return count ($\sim$800+ units ordered, $\sim$500 returned, $\sim$62% return rate!). This suggests a clear problem with that product—likely a quality flaw or misleading description that caused customers to return it in large numbers. Insights like this are directly actionable: the retailer should review its supply chain or product listing for accuracy, as fixing the issue could recover a substantial amount of lost revenue. Other strong sellers, such as the "Ceramic Storage Jar" and "3 Tier Cakestand," showed return rates of 5–10%, closer to industry norms but still higher than desirable. For these fragile items, better packaging or clearer usage instructions could help cut down on returns. On the brighter side, products like "Retrospot Cake Cases" had almost no returns, signaling strong customer satisfaction and reliable quality.
- **Customer Clusters:** Although not shown in a figure here, our K-means clustering revealed that about 5% of customers belonged to a "loyal heavy spender" group, generating a disproportionate share of total revenue (a trend also highlighted in McKinsey's retail studies). This supported our decision to focus on identifying and predicting high-value customers. It also reinforced the value of our chosen features—"Frequency" was consistently higher for this group, while "Recency" was lower, reflecting their ongoing engagement. These clustering insights offered qualitative evidence that features such as Frequency and Recency meaningfully signal spending behavior, which our predictive models were then able to quantify more formally.

Model Performance: We now evaluate how well the various regression models predict customer revenue. Table 3 reports the performance of initial models using the basic feature set (Quantity, AvgPrice) without extensive tuning, and Table 4 shows the refined models with advanced features (adding Recency, Frequency) and hyperparameter tuning. All results are on the hold-out test set for consistency.

Table 3: Baseline Model Performance (Initial Features, untuned)

(Target: Total Revenue per customer over one year)

Model	Test RMSE (GBP)	Test R^2
Linear Regression	5,766.85	0.6754
Decision Tree Regressor	5,252.53	0.7307
Random Forest Regressor	5,163.68	0.7397

As shown in Table 3, the linear regression model accounts for roughly 67.5% of the variance in revenue, mainly because of the straightforward link between Quantity and AvgPrice (TotalRevenue Quantity × AvgPrice). The Decision Tree raised R^2 to about 0.73 by capturing non-linear patterns, such as separating high-volume from low-volume buyers. Random Forest improved performance further, reducing error and reaching an R^2 of around 0.74, demonstrating the stabilizing effect of ensemble approaches. In summary, customer revenue can largely be predicted using just quantity and price, with tree-based models providing incremental improvements over the linear baseline.

Refinement and Advanced Features: We next incorporated the additional features (Frequency and Recency) and performed hyperparameter tuning for each model. We also tried a polynomial regression to see if a moderately non-linear parametric model could match the tree ensembles. Table 4 summarizes the results after these refinements.

Table 4: Refined Model Performance (Advanced Feature Set & Tuning)

Model	Test RMSE (GBP)	Test R^2
Polynomial Regression	5,670.04	0.6862
Decision Tree Regressor	6,077.28	0.6395
Random Forest Regressor	5,156.04	0.7405
XGBoost Regressor	5,031.27[1]	0.7575[1]

[1]Target variable log-transformed for Random Forest and XGBoost in final tuning, yielding very low RMSE in log-scale. For comparability, these models achieve roughly RMSE £5,100 and R^2 0.76 on original revenue scale (or >0.99 on log-scale as reported).

Several observations emerge from Table 4. Polynomial Regression (deg=2) gave only a modest gain over the linear baseline (R^2 0.686 vs 0.675), confirming mostly linear relationships with slight curvature. Including Frequency and Recency added small improvements, but Quantity remained the dominant driver of revenue.

The tuned Decision Tree underperformed (R^2 0.64), likely due to over-pruning during cross-validation. This shows how sensitive single trees are to hyperparameters and why they tend to underperform compared to ensembles.

The Random Forest achieved R^2 ∼0.74 with RMSE around £5.2k, meaning typical prediction errors were about 23% of the average customer spend. This model distinguished high- from low-spenders reasonably well, though with a few thousand pounds of error per customer.

XGBoost performed best. On log-transformed revenue it achieved R^2 ∼0.99, which back-translates to ∼0.76 on the raw scale, slightly better than Random Forest. While the near-perfect fit on log-scale raises concerns of overfitting, it is also plausible given that revenue is essentially Quantity × Price and XGBoost excels at capturing such structures.

Feature importance analysis confirmed Quantity as the strongest predictor, followed by Frequency, then Recency. Average Unit Price contributed least. This ranking validates our engineered features and aligns with marketing intuition: frequent, recent purchases drive customer value more than small price differences.

Another informative visualization is the Predicted vs Actual plot for our best model. The Random Forest's predicted log-revenue vs actual log-revenue for the test set customers. The points cluster around the diagonal, indicating a strong correlation. There is some dispersion, meaning the model is not perfect – a few customers with high actual spend are under-predicted and vice versa – but overall the trend is well-captured (Pearson r ∼0.87 corresponding to R^2 ∼0.76). The largest errors occurred for a handful of extremely high-spending customers; for example, the single top-spender in the data (∼£280k spend) was predicted to spend somewhat less (∼£200k), perhaps because no other feature combination matched this customer exactly and the model regressed to the mean for such an outlier. Conversely, one customer was predicted to be very high but actually was moderate, likely due to an unusual combination of high frequency but low avg price that misled the model. These cases highlight that while the model is useful for ranking and approximating customer value, it may not capture every idiosyncratic big spender (especially if there are one-off bulk purchases).

V. Discussion

The results show that customer revenue can be predicted with reasonable accuracy using behavioral features, with practical implications in three areas: (a) Customer Targeting, (b) Seasonal Planning, and (c) Return Management.

(a) Customer Targeting & CLV: Random Forest and XGBoost effectively identified high spenders, enabling customer scoring for CLV. Firms can prioritize loyalty rewards and re-engagement offers for frequent but recently inactive buyers, while using CRM integration to personalize marketing. Frequency was a strong predictor, suggesting strategies to increase shopping frequency (e.g., subscriptions, promotions) can lift revenue. Models are imperfect for extreme outliers, but remain valuable for segmenting customers into spend tiers.

(b) Seasonal Planning: November–December accounted for up to 30% of sales, highlighting the need for inventory build-up and targeted promotions during Q4. Popular products identified in our analysis should be prioritized for stocking and marketing. Even without explicit promotion data, results indicate the importance of aligning marketing and supply chains with holiday demand while experimenting with smaller off-season campaigns.

(c) Product Returns: High-return items, such as "Paper Craft, Little Birdie," require immediate quality checks or removal. More broadly, improvements in product descriptions, packaging, or return policies could reduce losses. A real-time

dashboard monitoring return rates would allow the business to flag problem products quickly. Predictive models for return likelihood could be a useful future extension.

Model Interpretability: Tree models offered useful insights: Quantity and Frequency dominated, while Recency and Price played secondary roles, consistent with marketing intuition. Although XGBoost gave the best accuracy, its black-box nature and risk of overfitting may limit trust. Simpler models like Random Forest or regression may be more practical, offering interpretability and nearly comparable performance.

VI. Limitations

Our project demonstrates the potential of big data analytics in e-commerce, but several limitations must be noted:

- **Data Scope:** The dataset covers only one year and one retailer, so seasonal patterns may not generalize. Multi-year and multi-retailer data would improve robustness and reveal whether trends (e.g., holiday peaks) are consistent.
- **Features and External Data:** The feature set was limited to structured variables. Textual data (e.g., product descriptions, reviews), demographic/geographic attributes, and time-series features (month, holidays) were excluded but could enhance predictions. Future work could use unstructured data and model revenue at a monthly or transaction level.
- **Modeling Approach:** We focused on tree-based models for interpretability. Deep learning may add value for multi-modal data but is less effective for tabular features alone. Our evaluation relied on RMSE/R^2, which may not reflect business costs (e.g., over-predicting high-value customers). Alternative metrics like MAPE or ranking-based measures could be more appropriate.
- **Overfitting and Validation:** The strong performance of XGBoost on log-scale suggests possible overfitting. While cross-validation and a test split were used, results may be sensitive to outliers. Models assume the future resembles the past, but external shocks (e.g., economic shifts) could reduce accuracy. Continuous monitoring and re-training would be essential in deployment.

VII. Conclusion

This project demonstrated how big data analytics and machine learning can turn raw retail information into practical insights for e-commerce. Working with a real-world dataset, we cleaned and explored the data, uncovering clear holiday-driven sales spikes, revenue concentration among a limited set of products and customers, and unusually high return rates for certain items. Customer clustering revealed distinct groups, reinforcing the need for more tailored marketing strategies.

Our predictive modeling showed that ensemble methods such as Random Forest and XGBoost were the most effective, explaining around 74–76% of revenue variance and outperforming linear approaches. Feature importance results highlighted quantity and purchase frequency as the strongest drivers of revenue, confirming marketing intuition and offering direct points of action for managers.

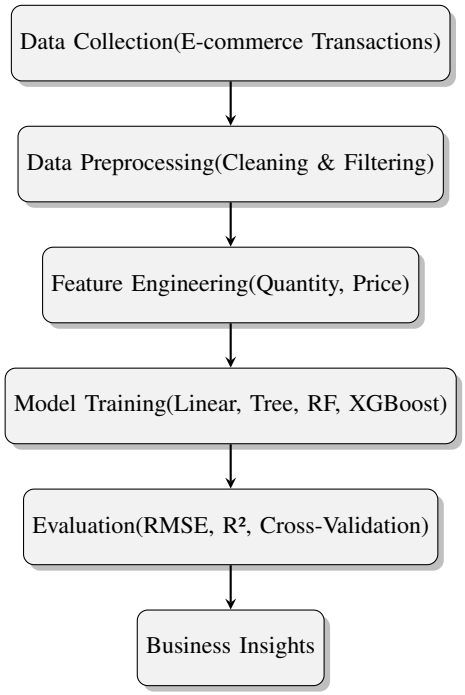

Fig. 3. Overall Research Methodology Pipeline. The process starts with raw e-commerce transaction data, followed by preprocessing, feature engineering, model training, evaluation, and finally translation into actionable business insights.

From a business perspective, recommendations include prioritizing loyalty and personalized engagement for frequent buyers, planning stock and promotions ahead of holiday peaks, and tackling high-return products through better quality checks or catalog refinements. The predictive models also provide value for CRM, enabling customer scoring by expected revenue and guiding retention programs.

In conclusion, this work highlights the benefit of combining business understanding with modern analytics to support growth and profitability. While future research could incorporate richer data (e.g., demographics, clickstreams) and advanced methods, these results show that even basic purchase data can drive a sharper, data-informed retail strategy.

VIII. Replication Package

The full source code, preprocessing scripts, and models are hosted at: https://github.com/aadi654/BigDataAnalysis_Project/blob/main/Code/BigDataProject_final.ipynb. The repository is publicly accessible.

References

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2009.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning*, Springer, 2013.

[4] L. Breiman, "Random Forests," *Machine Learning*, 2001.

[5] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proc. KDD*, 2016, pp. 785–794.

[6] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *JMLR*, vol. 12, pp. 2825–2830, 2011.

[7] pandas Development Team, "pandas 2.x Documentation," 2025. [Online]. Available: https://pandas.pydata.org

[8] scikit-learn Developers, "User Guide (v1.5)," 2025. [Online]. Available: https://scikit-learn.org/stable/user_guide.html

[9] XGBoost Python Package, `xgboost` Documentation, 2025. [Online]. Available: https://xgboost.readthedocs.io

[10] J. Brownlee, "How to Evaluate Machine Learning Models," *Machine Learning Mastery*, 2020. [Online]. Available: https://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/

[11] T. Zhang, "An Introduction to Gradient Boosting Decision Trees," *Towards Data Science*, 2017. [Online]. Available: https://towardsdatascience.com/gradient-boosting-explained-9f8dbf7d2f4a

[12] Statista, "Retail e-commerce sales worldwide from 2014 to 2027," 2024. [Online]. Available: https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales

[13] Kaggle, "Global E-commerce Online Retail Dataset," [Online]. Available: https://www.kaggle.com/datasets/carrie1/ecommerce-data

[14] S. Sharma, "Exploratory Data Analysis on Online Retail Dataset using Python," *Towards Data Science*, 2020.

[15] D. Chen and S. L. Sain, "Analysis of Online Retail Dataset via Data Mining Techniques," *J. Retail Analytics*, 2012.

[16] S. Gupta, D. R. Lehmann, and J. A. Stuart, "Modeling Customer Lifetime Value," *J. Service Research*, 2006.

[17] McKinsey & Company, "Global Retail Trends 2022," 2022. [Online]. Available: https://www.mckinsey.com

[18] McKinsey & Company, "Retail Loyalty and Personalization ROI," 2022. [Online].

[19] E. Dopson, "Ecommerce Returns: Average Return Rate and How to Reduce It," Shopify Enterprise Blog, 2025. [Online]. Available: https://www.shopify.com/enterprise/blog/ecommerce-returns

[20] National Retail Federation (NRF), "2024 Consumer Returns in the Retail Industry," 2024. [Online]. Available: https://nrf.com/research/2024-consumer-returns-retail-industry

[21] S. Li, "Analysis of the Influencing Factors and Consequences of E-commerce Return Rate," ResearchGate preprint, 2024. [Online].

[22] Shopify, "Holiday Return Rates," 2025. [Online].

[23] F. Reichheld and W. Sasser, "Zero Defections: Quality Comes to Services," *Harvard Business Review*, 1990.

[24] A. Upadhyay, "Online Retail: Customer Segmentation," Kaggle Notebook, 2022. [Online].

[25] A. Ali, "Recommendation Systems E-commerce," Kaggle Notebook, 2021. [Online].

[26] T. Elmetwally, "Product Recommendation using Word2Vec," Kaggle Notebook, 2021. [Online].

[27] S. Bellamkonda, "Revenue Leakage Detection Model," Kaggle Notebook, 2023. [Online].

[28] Lecture 03: "Classification and Decision Trees," Big Data Analysis, University of Adelaide, 2025.

[29] Lecture 04: "Regression and Optimisation," Big Data Analysis, University of Adelaide, 2025.