Classifying Consumer Behaviour

Ali Nour (501248744)

Supervisor: Dr. Ceni Baboglu
July 25, 2024

Toronto Metropolitan University

Classification Based on Rating

Study focuses on:

- Identifying key features influencing product ratings
- Examining brand impact on product ratings
- Exploring the relationship between sale price and product rating

The Dataset

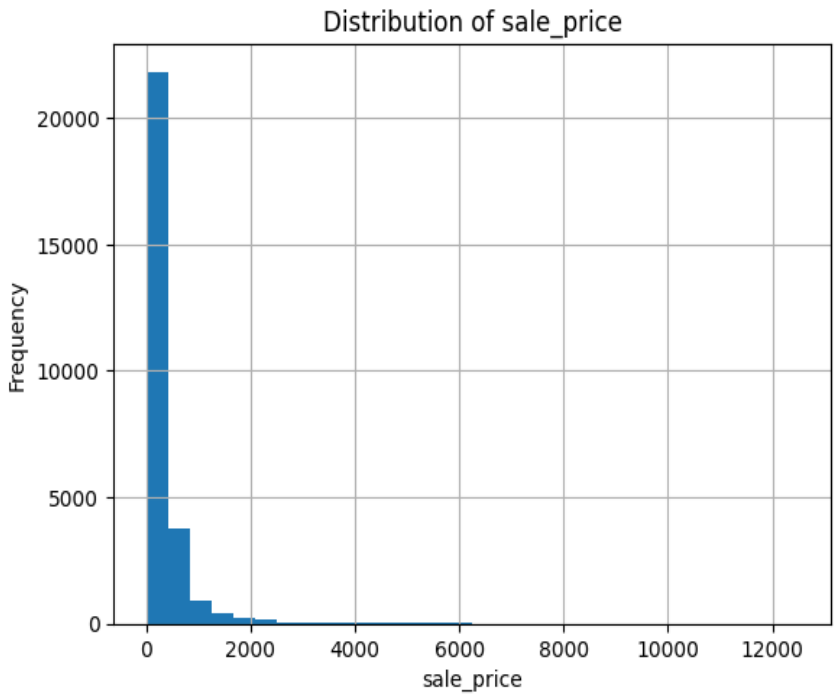
Field	Description	Distinct Values	Missing Values	Mean	Min	Max
Index	Simply the Index!	27,555 (100%)	0 (0.0%)	13,778	1	27,555
Product	Title of the product (as they're listed)	23,540 (85.4%)	1 (< 0.1%)	-	-	-
Category	Category into which product has been classified	11 (< 0.1%)	0 (0.0%)	-	-	-
Sub-category	Subcategory into which product has been kept	90 (0.3%)	0 (0.0%)	-	-	-
Brand	Brand of the product	2,313 (8.4%)	1 (< 0.1%)	-	-	-
Sale price	Price at which product is being sold on the site	3,256 (11.8%)	0 (0.0%)	322.51	2.45	12,500
Market price	Market price of the product	1,348 (4.9%)	0 (0.0%)	382.06	3	12,500
Type	Type into which product falls	426 (1.5%)	0 (0.0%)	-	-	-
Rating	Rating the product has got from its consumers	40 (0.2%)	8,626 (31.3%)	3.94	1	5
Description	Description of the dataset (in detail)	21,944 (80.0%)	115 (0.4%)	-	-	-

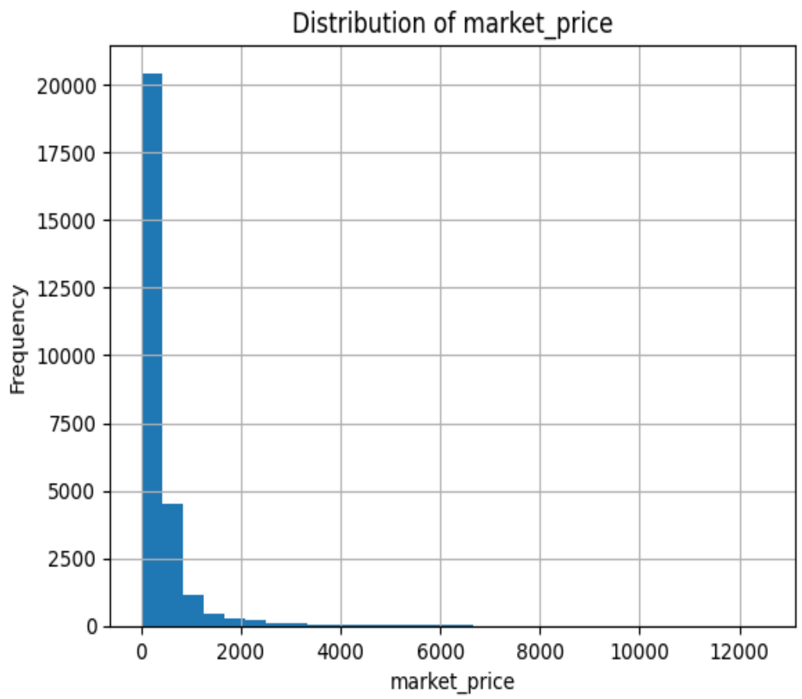
Toronto Metropolitan University

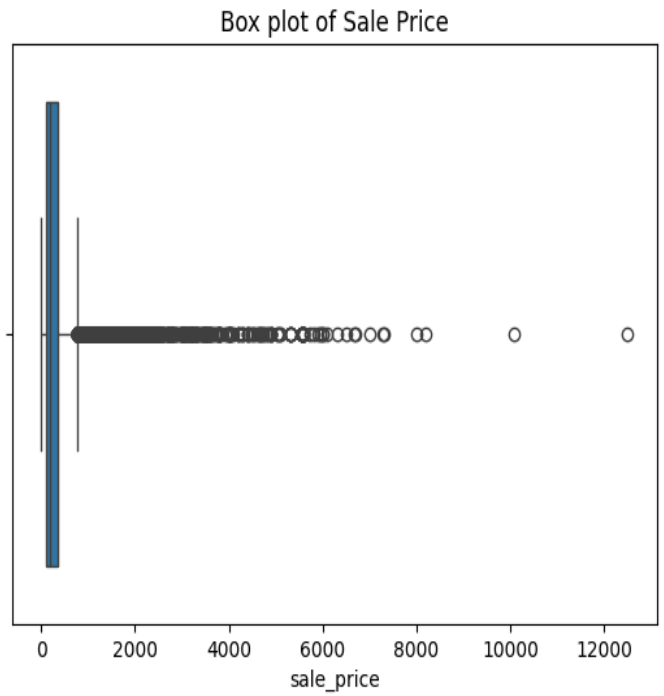
https://www.kaggle.com/datasets/surajjha101/bigbasket-entire-product-list-28k-datapoints/data

EDA: Visual Analyses

Distribution of Sale Price

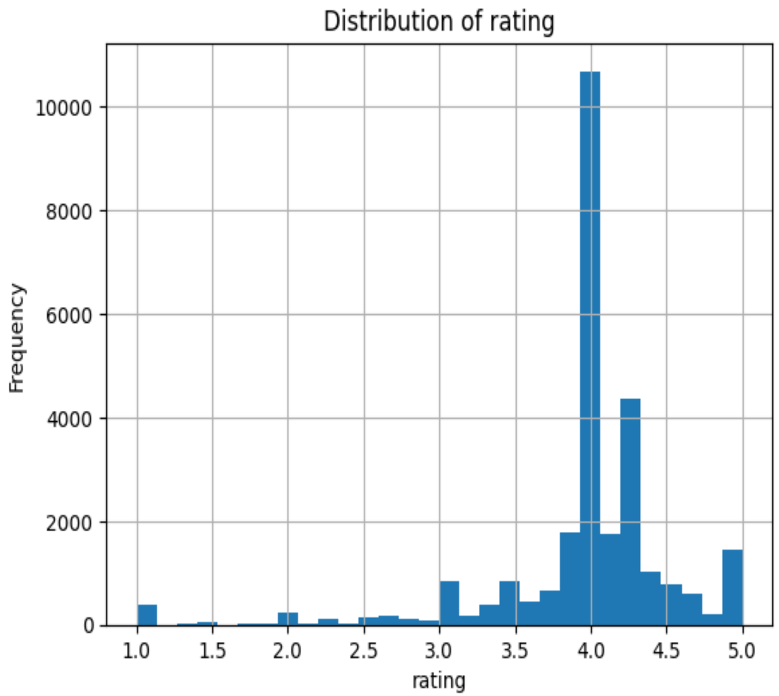


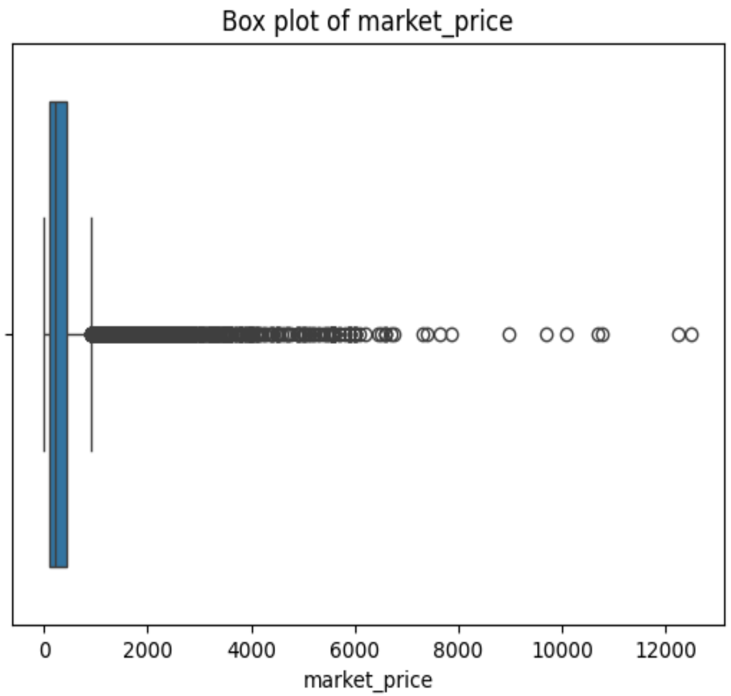


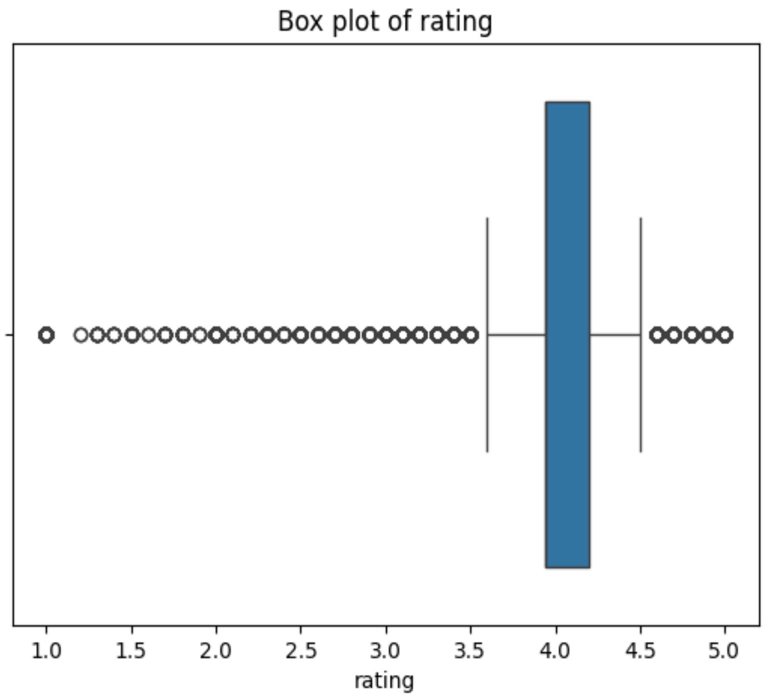


EDA: Visual Analyses

Distribution of Rating

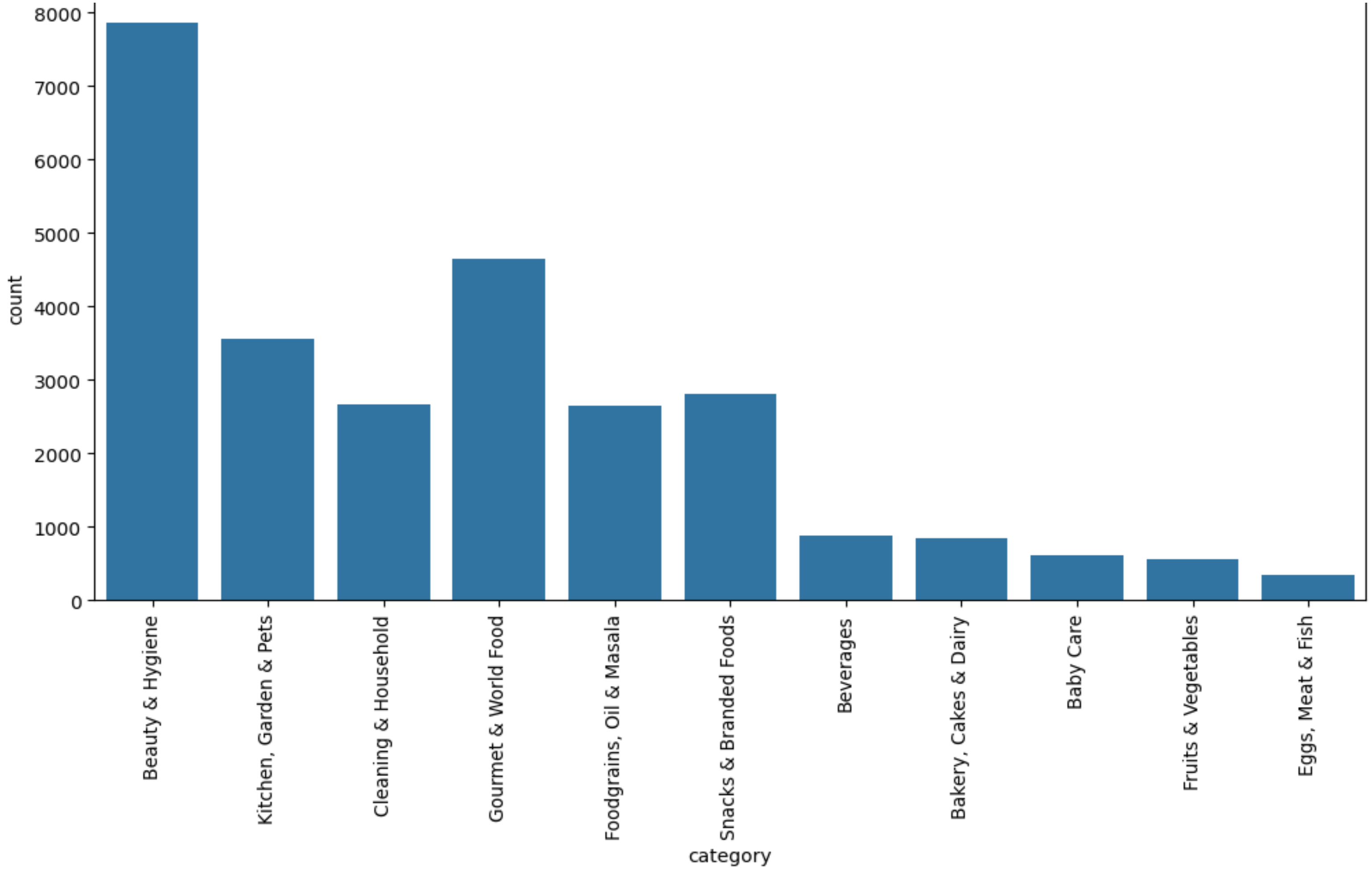




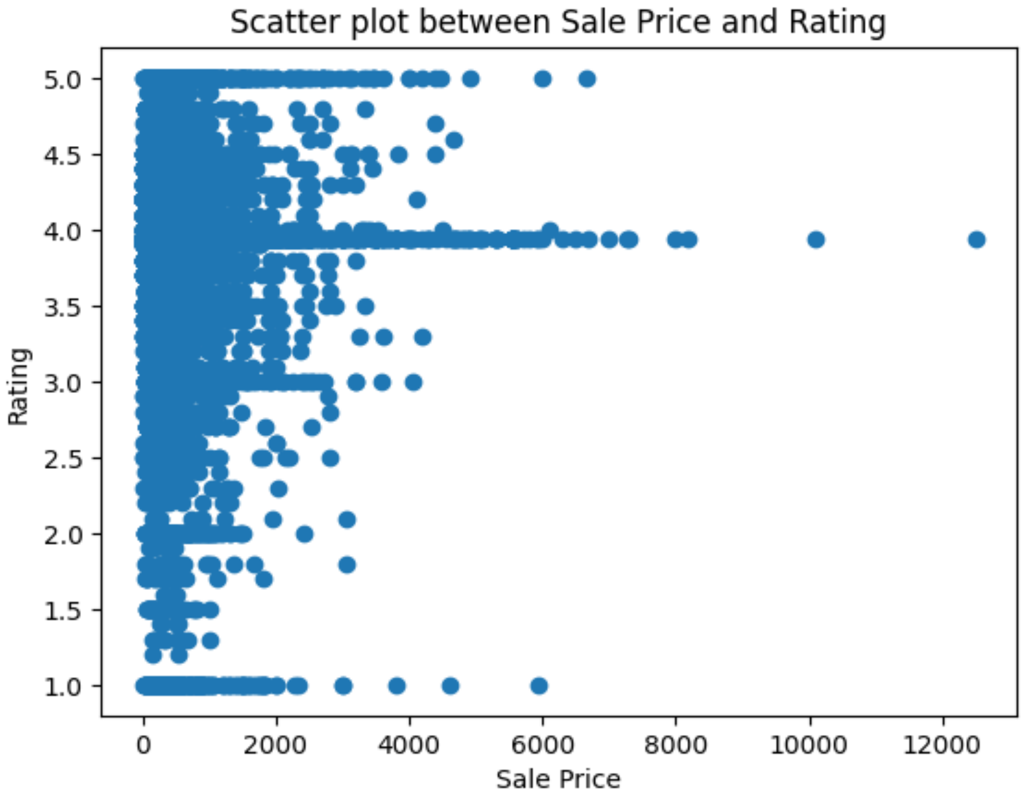


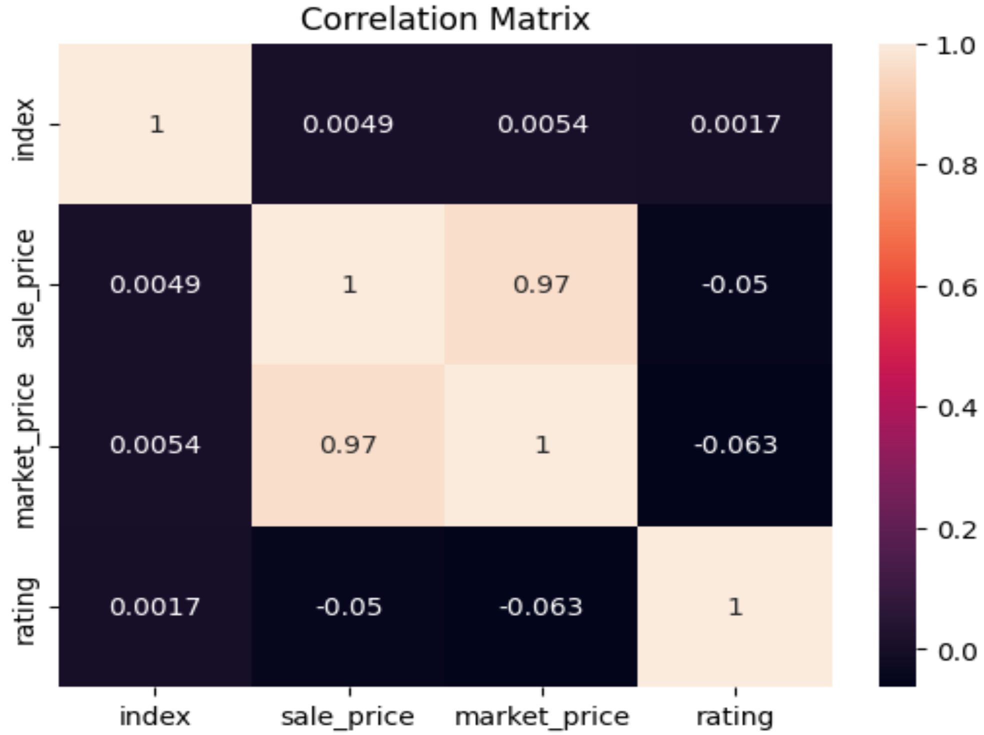
EDA: Visual Analyses

Count Plot of Category



EDA: Visual Analyses


Scatter plot between Sale Price and Rating


Correlation Matrix

	index	sale_price	market_price	rating
index	1	0.0049	0.0054	0.0017
sale_price	0.0049	1	0.97	-0.05
market_price	0.0054	0.97	1	-0.063
rating	0.0017	-0.05	-0.063	1

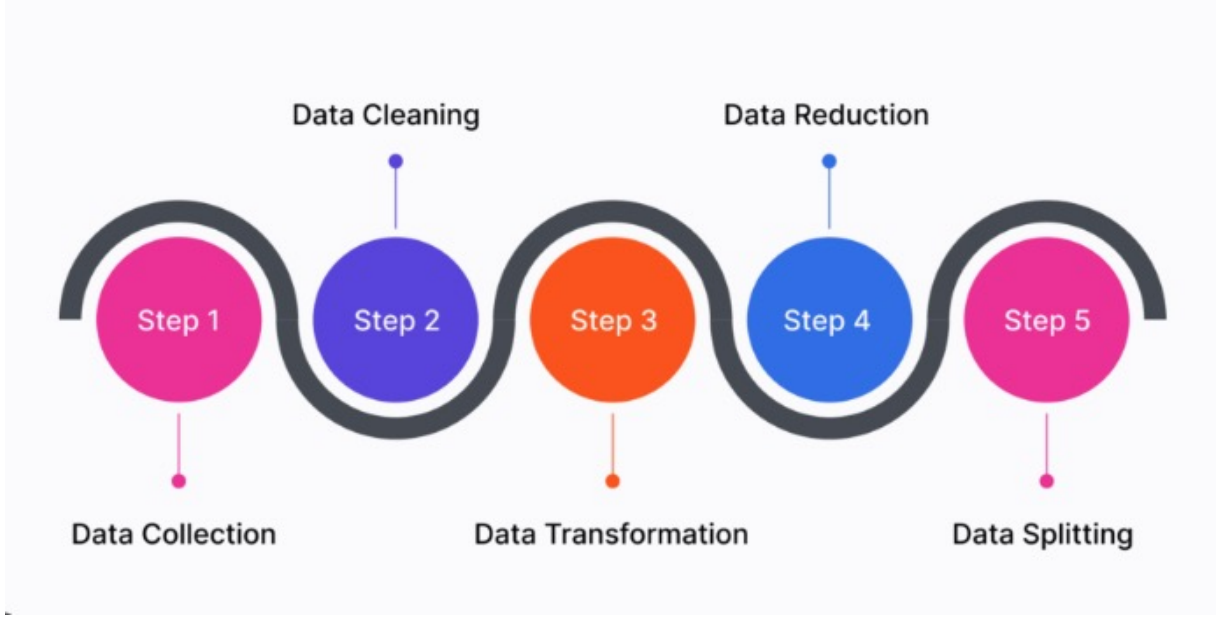
Methods

Data Collection: Open-source dataset from Kaggle.

Data Preparation cleaning: Addressed missing values, duplicates, normalized text, standardized values.

Feature Engineering: Extracted and created relevant features.

Data Preprocessing: One-hot encoding, creating a data-frame with selected features.



Methods

Machine Learning Models

- Decision Tree Classifier
- Random Forest Classifier
- Logistic Regression

Model Training and Evaluation:

- Training: 80% training set, 20% testing set.
- Evaluation: Accuracy, precision, recall, F1-score.

Findings

Model Performance

- Decision Tree: Cross-validation accuracy 0.616, test accuracy 0.625.
- Random Forest: Cross-validation accuracy 0.638, test accuracy 0.648.
- Logistic Regression: Cross-validation score 0.570, test accuracy 0.556.

Findings

Model	Cross-Validation Accuracy	Test Accuracy	Precision	Recall	F1-Score	Misclassification Rate	Performance in Lower Ratings
Decision Tree	0.616	0.625	0.62	0.62	0.62	High	Moderate
Random Forest Classifier	0.638	0.648	0.64	0.65	0.64	Reduced	Improved
Logistic Regression	0.570	0.556	.54	0.56	0.40	Significant	Poor

Conclusion

- Random Forest model outperforms Decision Tree and Logistic Regression
- Decision Tree shows reasonable performance
- Logistic Regression struggles with dataset complexity
- Product category, brand, and sale price significantly influence ratings.
- Insights valuable for businesses to optimize product ratings and understand consumer behavior
- Identifying patterns in consumer behavior helps improve product development and customer service strategies.