**Analyzing Consumer Purchase Behavior**

[Original Data Set](https://www.kaggle.com/code/waqi786/electronic-sales-insights-sep-2023-sep-2024/notebook?select=Electronic_sales_Sep2023-Sep2024.csv)

**Background Information**

The electronics retail industry is a highly competitive market with rapid, continual growth marked by its advancements in technology. It encompasses a wide range of products such as electronics, home appliances, computers, etc. The industry was further transformed with the introduction of online shopping. This marked a significant shift for consumers as it provided them with more options and better prices at their fingertips. Traditional brick-and-mortar retailers were forced to revamp their online presence to compete with e-commerce giants like Amazon, Target, Best Buy, etc. According to statistics from ECDB, the top three electronic retailers are Amazon, Apple, and Walmart. Combined, they made over $85 billion in sales in the year 2023, with Amazon leading with $48.8 billion, making up over 50% of last year’s sales (Uzunoglu, 2024). This dominance is likely attributed to Amazon's focus on being an e-commerce company, allowing them to solely focus on leveraging its data analytics into optimized sales strategies.

The unnamed organization from our dataset achieved nearly $65 million in electronic sales last year, retailing products such as headphones, laptops, smartphones, smartwatches, and tablets. Although they appear to operate exclusively online and have competitive sales numbers compared to the major leaders in the industry, they still face industry challenges such as intense competition and adapting to rapid technological change. In this landscape, effective use of analytics not only drives sales but also builds brand loyalty, as customers increasingly seek brands that understand and respond to their unique preferences and needs. Understanding the competitive dynamics and the role of data in shaping strategies is critical for sustained success in the electronics retail sector.

**Business Problem**

In today's fast-paced and information-rich environment, retailers depend heavily on customer insights to boost their sales. With a wealth of data available, businesses need to understand their customers' behaviors, preferences, and shopping patterns. However, using this data effectively comes with challenges. Therefore, it is equally imperative for retailers to invest in data analysis tools to navigate large amounts of data. By prioritizing data-driven decisions, stores can continually adjust to changing consumer needs and market conditions.

Thus, our business challenge lies in understanding consumer buying trends in the electronics retail sector. With technology evolving quickly, electronics retailers face increased competition, especially in e-commerce. The transition to online platforms has given consumers easy access to a wide range of products which only further intensifies competition. However, this change also offers retailers an increased opportunity to gather valuable data from customer transactions and interactions.

Ultimately, electronics retailers can use these insights to boost their competitive advantage and build better relationships with customers. This approach can result in higher sales and ongoing growth, even as the market keeps changing. Adopting strategic data analysis skills will be vital only for addressing current needs but also for tackling future challenges in an ever-evolving landscape.

**Project Objectives**

Our goal for this project is to thoroughly analyze the electronic sales data from a random organization over the last year. With the help of Weka, we will explore different data mining methods and gather valuable insights to solve our business problem. Additionally, we will need to preprocess and adjust the data to meet our research objectives.

Our specific objectives are as follows:

1. Analyze Customer Buying Patterns

* *Identify High-Selling Products*
* Determine which products have the highest sales volumes
* Examine sales data across various categories, such as headphones, laptops, smartphones, smartwatches, and tablets.
* These objectives will help identify market trends and customer preferences.
* *Examine Trends:*
* Implement clustering techniques to group customers based on their buying behaviors and demographics. This can help tailor marketing strategies and enhance personalization.
* Understanding these trends can help the organization optimize marketing strategies and inventory planning.
* *Explore Customer Preferences:*
* Analyze customer demographics, such as age and gender, along with their shopping behaviors (ex: loyalty membership status, shipping preferences, etc).
* This insight will help in constructing a business plan to tailor offerings ultimately with the goal of enhancing customer satisfaction.

1. Enhance Cross-Sales Strategies

* *Identify Products Frequently Purchased Together:*
* Use association rule mining technique to discover combinations of products that are often bought together.
* This insight will highlight potential cross-selling opportunities and allow us to create promotional strategies that bundle products effectively to increase sales.

1. Improve Inventory Management

* *Optimize Inventory Levels:*
* Analyze sales data to determine optimal inventory levels for each product category.
* By identifying these patterns in product demand, the organization can avoid overstock which ties up capital. Instead, we can refocus those assets into more productive marketing campaigns while ensuring a smoother operational process.

1. Innovate Marketing Campaigns

* *Uncover Insights for Targeted Marketing Initiatives:*
* Use findings from purchasing patterns analysis to create targeted marketing campaigns aimed at specific customer segments.
* These personalized offers based on individual shopping behaviors will increase the likelihood of sales.
* *Enhance Customer Engagement:*
* Similarly, the insights gained from our data analysis can help us develop personalized marketing strategies that don’t go unnoticed.
* This could involve email campaigns, tailored recommendations on the website, or special promotions for loyal customers.
* Enhancing customer engagement through personalized offers can further lead to increased sales and foster long-term customer loyalty.

**Data Mining Techniques**

Method 1: J48 Decision Trees

Objective: To identify which products are most likely to turn regular customers into loyal members and high-selling products. Through this, we can analyze customer buying patterns and identify high-selling products.

Justification: Since decision trees create straightforward and easily understandable visuals, it can illustrate how decisions are made. By applying this method, we can pinpoint important elements that affect customer loyalty and improve our strategies for keeping customers.

Preprocess: The attributes in this data were reduced to 2 attributes: Product Type and their respective Order Quantity. Due to an imbalance in the data, the SMOTE filter in Weka was applied before testing. This helps increase the minority group by generating more instances so that the data is no longer skewed. It is then tested using 10-fold cross-validation with a 0.1% confidence rate.

Results:

* Smartphones: Customers who purchased smartphones were more likely to become loyal members.
* Tablets: Customers who purchased more than 0.5 tablets and bought over 7 units were classified as loyal.
* Smartwatches: Customers who bought more than 1 smartwatch but fewer than 1.5 units in total were considered loyal.
* Laptops: Laptops were omitted from the decision tree as they did not significantly correlate with customer loyalty.

Insights:

Accuracy was relatively low, with only 52.5% of instances correctly classified. A Kappa statistic of 0.0065 indicated minimal agreement, and while the model had a high error rate, this was attributed to the original imbalance in the dataset.

A screenshot of a computer

Description automatically generated

Method 2: K-Means Clustering

Objective: To group customers by specific demographic features and purchasing patterns such as age, gender, and product type to create tailored marketing approaches.

Justification: Since clustering helps us find natural groups within data, we can use this to divide customers into segments. We can then use this information to customize our marketing efforts for different demographics to boost engagement and satisfaction.

Preprocess: Apply Weka's Normalize filter to scale the data. Then, perform clustering using 3 clusters.

Results:

Cluster 0 (25% of data)

Demographics: Primarily Female, non-loyalty members.  
Product Preference: Primarily purchase Tablets (specifically SKU1002).

Purchasing Patterns:

* Most orders are successfully completed, with cash being the preferred payment method.
* Customers prefer Express Shipping.
* Extended warranties and impulse items are frequently purchased as add-ons.
* The Total Price of their orders tends to be lower compared to other clusters.

Cluster 1 (24% of data)

Demographics: Primarily Female, non-loyalty members.  
Product Preference: Focus on purchasing Smartwatches (SKU1003).

Purchasing Patterns:

* Higher order cancellation rate compared to other clusters.
* Primarily use Debit Cards for payments and prefer Overnight Shipping.
* Often buy extended warranties, occasionally buying them more than once, which shows a lack of confidence.
* The Total Price of their orders tends to be higher.

Cluster 2 (51% of data)

Demographics: Primarily Male, non-loyalty members.  
Product Preference: Primarily purchase Smartphones (SWT567).

Purchasing Patterns:

* Higher order completion rate, suggesting they are more decisive buyers.
* Prefer Credit Cards for payment and use Standard Shipping.
* Purchases are generally higher in price, with extended warranties purchased frequently.

**A screenshot of a computer

Description automatically generated**

Method 3: Linear Regression

Objective: To predict the Total Price of electronic product sales based on features such as Gender, Product Type, Quantity, and Purchase Date. This will help forecast future sales trends, optimize marketing strategies, and guide inventory management for electronic products.

Justification: Identifies key factors like product type and time of year that influence total sales, aiding future sales forecasts. Provides clear, actionable insights through coefficients, showing how each factor impacts sales. Quick and accurate for large datasets (20,000 instances), making it practical for real-time use. High correlation (0.936) demonstrates a great match between the model and data, ensuring reliable predictions.

Preprocess: Use Weka’s Filter to Normalize the data. Then, convert numeric values by using the NominalToBinary function under Filters → unsupervised → attribute → NominalToBinary.

Results:

The regression model provided the following equation for predicting the total price of a sale:   
  
Total Price = -0.0018 \* Gender = Male + 0.0123 \* Product Type = Smartphone + 0.0084 \* Product Type = Tablet + 0.0101 \* Product Type = Laptop + 0.0087 \* Product Type = Smartwatch + 0.0067 \* Product Type = Headphones + 0.535 \* Unit Price + 0.4555 \* Quantity + 0.0072 \* Month = December + 0.0056 \* Month = February + 0.2241

Correlation Coefficient: 0.9358  
Mean Absolute Error: 0.0589  
Root Mean Squared Error: 0.0788  
Relative Absolute Error: 32.6236%  
Root Relative Squared Error: 35.242%  
Total Number of Instances: 20,000

Let’s say we want to predict the Total Price for the following situation:

* Female
* Smartphone
* Quantity: 2
* Month: November

We will plug these changes into the linear regression equation:

Total Price = (-0.0018 \* 0) + (0.0123 \* 1) + (0.535 \* 800) + (0.4555 \* 2) - 0.2241

Total Price = 0 + 0.0123 + 428 + 0.911 - 0.2241 = 428.6992

The predicted total price for this customer is approximately $428.70.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Method 4: Random Forest

Objective: Predict whether an order will be Canceled or Completed. This can help identify patterns in customer behavior that lead to cancellations and to reach out with customer support, discounts, or personalized follow-ups to encourage completion.

Justification: Random forests can efficiently manage complicated relationships in data. This will allow us to create more precise conclusions on demand further helping maximize sales.

Preprocess: Resampled our data by going to Filter →supervised →instance →Resample; Clicked into Resample and set the sampleSizePercent to 200.

Results:

* Correctly Classified Instances: 93.31%
* Incorrectly Classified Instances: 6.69%

Class-by-Class Performance:

Canceled (Class A):

* True Positive Rate (0.955): The model correctly identified 95.5% of the Canceled orders
* Precision (0.915): When the model predicted Canceled, it was correct 91.5% of the time.
* F-Measure (0.935): High value, indicates strong overall performance for this class.

Completed (Class B):

* True Positive Rate (0.911): The model correctly classified 91.1% of Completed orders.
* Precision (0.953): When the model predicted Completed, it was correct 95.3% of the time.
* F-Measure (0.932): High value shows excellent performance

Confusion Matrix:

Canceled (A):

19,097 instances of Canceled were correctly classified.

903 instances of Canceled were misclassified as Completed.

Completed (B):

18,227 instances of Completed were correctly classified.

1,773 instances of Completed were misclassified as Canceled.

A screenshot of a computer

Description automatically generated

Method 5: Association Rule

Objective: Explore product pairings that are often purchased together to enhance cross-selling and marketing approaches.

Justification: One of the best techniques for identifying products that are frequently bought together, which is beneficial for cross-selling opportunities. Apriori algorithm is used to identify frequent item sets and generate association rules.  It allows us to determine which items should be promoted together.

Preprocess:  The data set was converted to Market Basket.  Data was filtered by NumericToNominal type using default settings and we found 10 rules (top 5 are below)

Results:

Rule 1

Condition: Laptop = 0 AND Smartphone = 1 AND Headphones = 0  
Confidence: 93%  
Lift: 1.1  
Insights: Customers who buy a smartphone but not a laptop are 93% likely to not purchase headphones.

Rule 2:

Condition: Laptop = 0 AND Smartphone = 1 AND Smartwatch = 0 AND Headphones = 0  
Confidence: 93%  
Lift: 1.1  
Insights: Customers who buy a smartphone, but no laptop or smartwatch are 93% likely to not purchase headphones.

Rule 3:

Condition: Smartphone = 1 AND Headphones = 0  
Confidence: 93%  
Lift: 1.09  
Insights: Customers who buy a smartphone are 93% likely to not purchase headphones.

Rule 4:

Condition: Laptop = 0 AND Tablet = 1 AND Headphones = 0  
Confidence: 90%  
Lift: 1.06  
Insights: Customers who buy a tablet but not a laptop are 90% likely to not purchase headphones.

Rule 5:

Condition: Smartwatch = 1 AND Tablet = 0 AND Headphones = 0  
Confidence: 90%  
Lift: 1.06  
Insights: Customers who buy a smartwatch but not a tablet are 90% likely to not purchase headphones.

A screenshot of a computer

Description automatically generated

**Data Mining Process**

Data Cleansing & Data Transformation:

Data cleansing is the first essential step in the data mining process because it ensures the dataset is accurate, consistent, and complete therefore further providing accurate results. Prior to running the dataset in our data processing tool, we went through several steps to ensure the data was optimized for processing. First, we confirmed any missing values were not present by checking the “Missing: 0 (0%)” line under “Preprocess.” Apart from the “Add-Ons Purchased” attribute, the rest all were complete with 0% missing values. The “Add-Ons Purchased” category had 4868 instances or 24% missing values. To ensure this didn’t negatively disrupt our outcomes, we replaced the blank spots with a “0.” Secondly, we checked for duplicate entries to ensure that each sale is counted only once by loading the dataset into Excel and using the “Remove Duplicates” function. We also standardized the “Total Price,” “Unit Price,” “Add-On Total,” and “Purchase Date” in Excel format to be the same.

We also scaled numerical attributes like “Total Price” and “Quantity” so that they fall within a consistent range (e.g., 0-1) to ensure fair weighting in models like K-Means or regression. Lastly, since some algorithms require uniform data types, we had to preprocess two to either numeric or nominal. Since K-Means clustering and linear regression method both require all attributes to be numeric, therefore we transformed the dataset to NominaltoNumeric. For Apriori, we used the NumerictoNominal function.

Attribute Selection and Transformation:

Once data cleansing was complete, we proceeded with evaluating the most relevant attributes to use for each data mining technique. In the J48 method, the attributes product type and quantity were chosen. The product type column was then expanded and replaced by their respective devices which resulted in a transformation where we have: ProductLaptop, ProductTablet, ProductSmartphone, and ProductSmartwatch. We used these specific measures to find out which products at how many units sold are associated with customer loyalty. Similarly for Random Forest, 5 attributes were selected: CustomerID, Age, Gender, Unit Price, Add-Ons Purchased.

**Data Description**

This data set is from a company that sells electronics, the record spans over an entire year from Sep 2023 - Sep 2024. It seems that sales made are online due to every transaction having different types of shipping. Each customer has unique identifiers with purchases from different dates implying that this data is based on customer accounts on their website. There are 20,000 records. The attributes in the data are the following:

| **Attribute Name** | **Description** | **Data Type** |
| --- | --- | --- |
| Customer ID | Unique Identifier | Numeric/Integer |
| Age | Age of Customer | Numeric/Integer |
| Gender | Gender of Customer | Categorical |
| Loyalty Member | Yes/No | Categorical |
| Product Type | Type of Electronic Product Sold | Categorical |
| SKU | Unique Code for Product | Categorical |
| Rating | Customer Rating (1 - 5 Stars) | Numeric/Integer |
| Order Status | Status of Order | Categorical |
| Payment Method | Method Used for Payment | Categorical |
| Total Price | Total Price of Transaction | Numeric/Float |
| Unit Price | Price/Unit of Product | Numeric/Float |
| Quantity | Number of Units Purchased | Numeric/Integer |
| Purchase Date | Date of Purchase | Categorical |
| Shipping Type | Type of Shipping | Categorical |
| Add - Ons Purchased | Additional Items Purchased | Categorical |
| Add - On Total | Total Price of Add - Ons | Numeric/Float |

**Interpretation of Results**

J48: Although the accuracy was low, the results still provide insight into the types of products that could potentially drive customer loyalty. Our findings reveal that specific product categories, including smartphones, tablets, and smartwatches, are closely linked to higher customer loyalty, indicating that focused marketing strategies may improve customer retention. Nonetheless, we need to bear in mind the model's low accuracy of 52.5% and a Kappa statistic of 0.0065 which highlights challenges in its effectiveness, largely due to initial data imbalance. The company should focus on optimizing marketing strategies for smartphones and tablets to increase member loyalty. Overall, the analysis showed that smartphone purchases are strongly associated with customer loyalty. Targeted promotions and loyalty programs for smartphone buyers could further enhance customer retention.

Clustering (K-Means): The results show distinct gender segmentation and varying purchasing behaviors among the three clusters. For Cluster 0, this group probably consists of practical Female shoppers. They prioritize Express Shipping and tend to make swift purchasing decisions, due to their inclination towards impulse buys. They are mindful of prices, which is shown by their lower overall spending, and likely appreciate the added security of extended warranties for their Tablet purchases. Marketing efforts should focus on promoting fast shipping and extended warranty offers. Cluster 1 group also consists of female shoppers who are interested in smartwatches but often struggle with making decisions, resulting in increased cancellation rates. They prefer quick shipping options, such as overnight delivery, and are open to spending more, likely because smartwatches tend to be expensive. Marketing strategies should focus on providing reassurance about their purchase choices, such as offering a "no-questions-asked" return policy, while highlighting fast, reliable shipping and extended warranty options. Lastly, Cluster 2 consists of a male customer segment that prioritizes smartphone buying. They prefer standard shipping, showing less urgency than Cluster 1, yet they still appreciate extended warranties. Marketing efforts should showcase the latest smartphone models and promote extended warranty options. Implementing loyalty programs could further drive repeat purchases, as this segment makes up a large part of the customer base.

Linear Regression: The model demonstrates strong accuracy in predicting total sales, with key factors like product type, quantity, and time of year having the most significant impact. Its high correlation and low error rates make it effective for forecasting future sales trends. With low error rates (mean absolute error of 0.0589 and root mean squared error of 0.0788), the model provides accurate sales predictions. Furthermore, it indicates that Unit Price (0.535) and Quantity (0.4555) have the largest positive impact on the Total Price. In other words, higher unit prices or larger quantities purchased significantly drive up the overall sales. When Gender is Male, the coefficient is -0.0018. A negative coefficient decreases the total price; therefore, men are less likely to spend compared to women. Since gender is a categorical variable, in this case, Male is coded as 1 and Female as 0. Additionally, certain months, particularly December and February, also show a negative contribution to sales totals. Lastly, when it comes to Product Type, Smartphones (0.0123) yields the highest, positive coefficient indicating the highest impact on total. This makes it useful for forecasting future trends, guiding decisions on inventory, marketing, and sales strategies. Its simplicity and ability to be easily updated with new data also make it practical for ongoing business analysis.

The Random Forest: The model can help with customer retention, the creation of loyalty programs, assist with product marketing, cross-selling/upselling, the amount of inventory that you keep, and even marketing budgets. For customer retention, we can predict whether an order will be completed or canceled and can adopt a strategy to mitigate by providing customer support, discounts as well as running a target email campaign. Loyalty programs can use this model to segment customers based on their likelihood to complete and their loyalty status. We can identify high-value customers and increase the benefits of loyalty programs which can give discounts, access to new products early, or even increase loyalty rewards. In terms of product marketing, this model shows that certain products are more likely to be canceled so you can focus on highlighting the product. Random Forest can identify when customers are likely to purchase add-ons and can target cross-selling and upselling at the point of purchase. Finally, this model helps with operational needs like keeping the correct amount of inventory (meet demand and not overstocked) and can help efficiently use marketing budgets.

Association Rule Mining: There is a consistent negative association between purchasing headphones and other tech products such as smartphones and tablets. The company could utilize this insight to implement product bundling strategies. Offering discounts or promotions on headphones for customers purchasing smartphones or tablets could encourage additional sales and increase overall revenue.

**Managerial and Practical Implication**

Analyzing Customer Buying Patterns: The J48 Decision Tree model revealed that customers who bought smartphones and tablets were significantly more likely to become loyal members, achieving an accuracy rate of 52.5%. This finding enables managers to concentrate their marketing strategies on these valuable products, including developing targeted promotions and loyalty programs specifically for smartphone and tablet purchasers. Furthermore, the K-Means Clustering analysis divided customers into specific segments based on their buying habits, identifying three main groups: practical female shoppers, indecisive female shoppers interested in smartwatches, and male customers who prioritize smartphones. This segmentation allows managers to customize marketing messages and offers to align with the unique characteristics and needs of each segment, thereby enhancing the effectiveness of marketing initiatives. We have also gathered that customers in Cluster 0 prefer express shipping and tend to make quick purchasing decisions. Management can adopt strategies that align with these preferences, such as improving shipping options and highlighting features like hassle-free returns, ultimately enhancing the overall customer experience.

Enhance Cross-Sales Strategies: By applying the Apriori algorithm to find frequent itemset, the business can identify products that are commonly purchased together, such as smartphones and accessories. This information can be used to design targeted marketing campaigns that bundle these products, encouraging cross-sales and increasing average transaction values. Additionally, insights from both J48 Decision Trees and K-Means Clustering methods indicate that smartphone buyers are a crucial customer segment. Managers can develop loyalty programs that reward repeat purchases among this group, potentially increasing retention and driving sales. For example, implementing tiered rewards based on spending thresholds could encourage higher-value purchases.

Improve Inventory Management: The Linear Regression Model revealed that unit price (0.535) and quantity (0.4555) are strong predictors of total sales. By understanding these relationships, managers can make informed decisions about inventory levels, ensuring that they stock enough of high-demand items while reducing overstock on less popular products. The regression model's ability to predict sales with low error rates (mean absolute error of 0.0589) enables ongoing assessment and refinement of inventory strategies. By continually updating the model with new data, management can adapt to changing customer preferences and seasonal trends more effectively. Additionally, the Decision Tree technique highlighted the importance of smartphones in driving customer loyalty. This insight can guide product development and inventory decisions to ensure that popular items are adequately stocked, aligning offerings with customer demand.

Innovate Marketing Campaigns: K-Means Clustering identified that different customer segments have varying preferences for shipping options. By enhancing shipping methods based on these insights, such as offering express shipping for practical female shoppers, management can significantly boost customer satisfaction and retention. Furthermore, insights from J48 Decision Trees and K-Means Clustering methods can inform the establishment of feedback channels tailored to specific customer segments. For instance, the indecisive smartwatch buyers in Cluster 1 may benefit from targeted surveys to gather input on their purchasing hesitations, allowing the business to address concerns and refine marketing strategies accordingly.

**References**

Ali, W. (2024, September 23). 📊✨ Electronic Sales Insights: Sep 2023 - Sep 2024. Kaggle.com; Kaggle. <https://www.kaggle.com/code/waqi786/electronic-sales-insights-sep-2023-sep-2024/notebook>

Uzunoglu, C. (2024, July 19). Electronics Market in the U.S.: Revenue, Growth, Market Size | ECDB.com. Ecommercedb.com. https://ecommercedb.com/insights/u-s-electronics-media-market-10-biggest-online-stores-in-the-united-states/4632