Maliwat Julius – 864520

Marketing Analytics

Main Objectives

Customer Focus

To increase customer retention and maximize lifetime value through targeted engagement strategies using RFM segmentation and churn prediction.

Product Focus

To enhance revenue through optimized cross-selling opportunities identified by Market Basket Analysis.

Feedback Focus

To improve product and service quality by analyzing customer sentiments and feedback, enabling proactive adjustments and enhancements

Customer Focus

Business Questions

What is the composition of our customer base, and how can we segment it to better understand customer behavior?

What are the key factors driving customer churn, and how can we address them to improve retention?

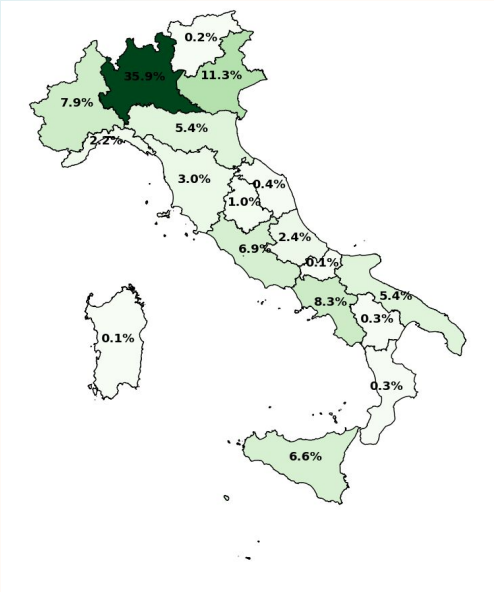
Approaches

RFM Model: This model segments customers based on Recency, Frequency, and Monetary value, enabling precise targeting and customization of retention strategies.

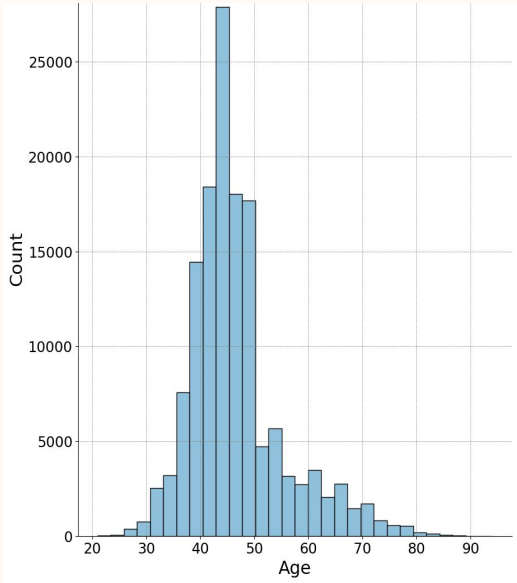
Churn Model: Focused on identifying the likelihood of customer attrition, this model helps in formulating proactive measures to retain high-risk customers.

Customer Demographic Overview

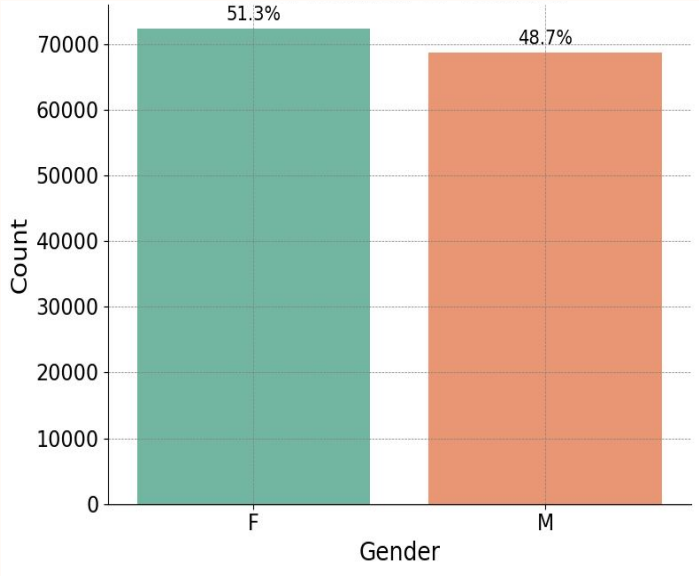
Regional Distribution



Age Distribution



Gender Distribution



Lombardia leads with 35.9% of customers, followed by Veneto (11.3%) and Campania (8.3%), indicating a higher concentration of customers in these regions.

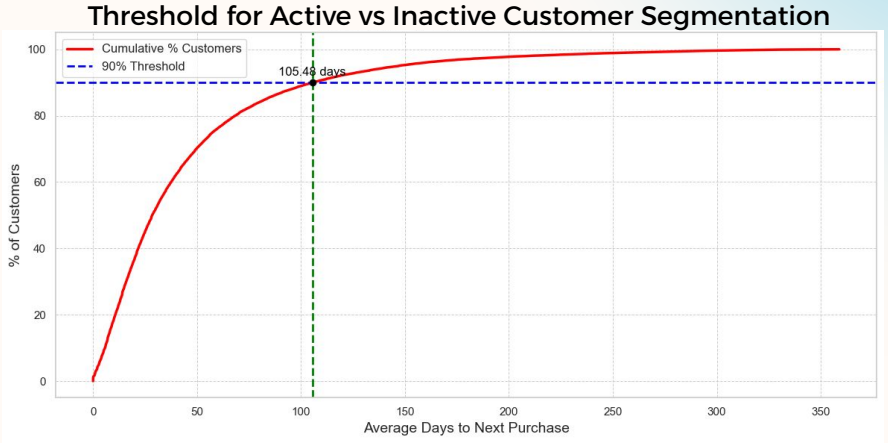
Most customers are between **40–50 years old**, with fewer in the 20–30 and 70+ age groups.

The gender distribution is fairly balanced, with **slightly more female** customers (F) than male customers (M).

Active vs Inactive Customer Segmentation

1. Filter eligible customers:
We began by selecting customers with at least two purchases, ensuring they demonstrate recurring behavior.

2. Segment active vs. inactive customers: For each eligible customer, we calculated their average repurchase interval. Customers with an average interval greater than 105 days (90th percentile of all intervals) were classified as inactive, allowing us to focus on those with more consistent purchasing patterns.

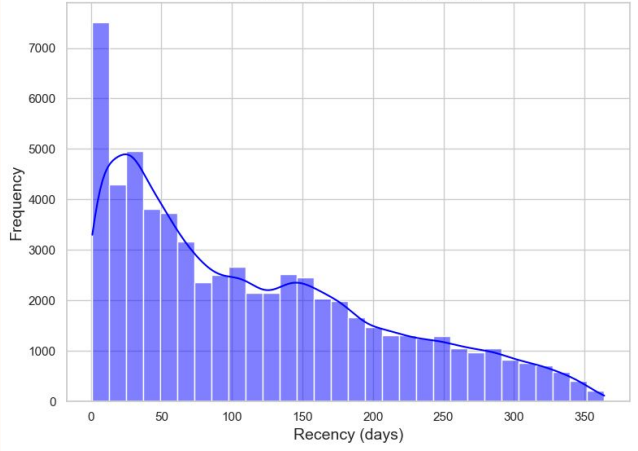


The RFM analysis and churn model were both **applied exclusively to active customers**, focusing retention and marketing efforts on those with the highest engagement

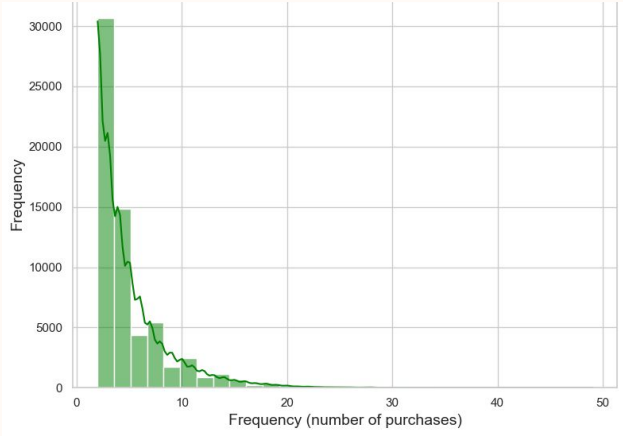
RFM Metrics Overview: Recency, Frequency, and Monetary

Analysis Period: 1st May 2022 – 30th April 2023

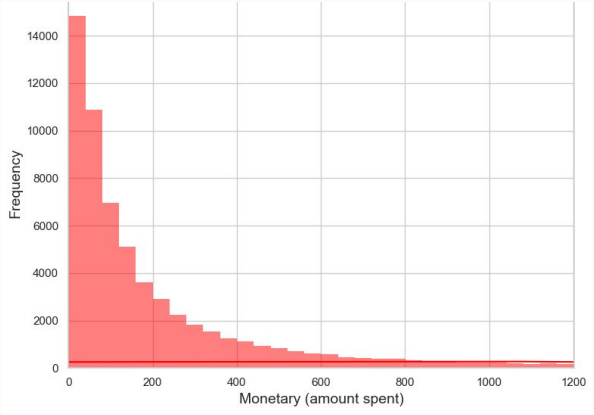
Recency: Days Since Last Purchase



Frequency: Total Purchases per Customer



Monetary: Total Spent per Customer



The recency distribution highlights that a substantial portion of active customers have made their latest purchases very recently, within the first 50 days, reflecting a right-skewed distribution.

The frequency distribution reveals that the majority of active customers make only 2 purchases, indicating a high prevalence of infrequent buyers.

The monetary distribution is heavily skewed with most spending under €200, although there are notable outliers, including several at €200,000 and €300,000, suggesting significant variations in customer spending habits.

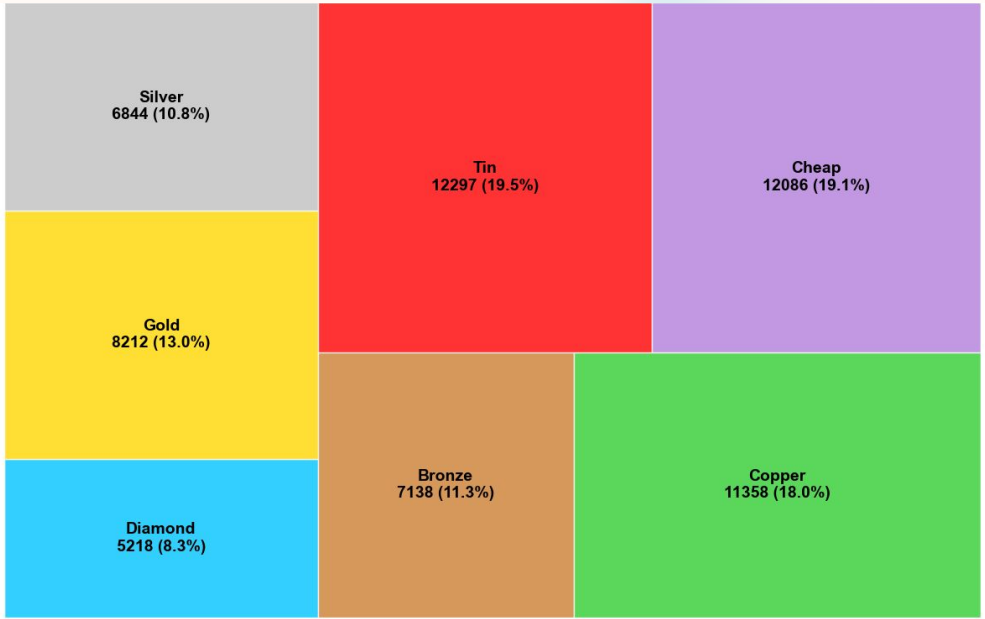
RFM Segmentation: Customer Value Distribution

Click here to access the Dashboard

1. Tertile Division: Split active customers into tertiles for Recency, Frequency, and Monetary values, creating 27 distinct RFM groups.

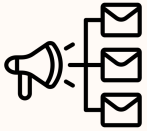
2. RF Matrix Creation: Combine Recency and Frequency scores to categorize customers into distinct loyalty statuses

3. RFM Segmentation: Cross these RF classes with Monetary values to generate the final 7 RFM segments.



Low-value customers ('Tin,' 'Cheap') make up 38.6%, mid-value customers ('Copper,' 'Bronze,' 'Silver') are 40.1%, and high-value customers ('Gold,' 'Diamond') account for 21.3%, representing the most engaged and profitable segments.

Action Steps (RFM Analysis)



Launch Automated Re-Engagement Campaign

Target the 38.6% of low-value customers ('Tin' and 'Cheap') with an automated email re-engagement campaign. Focus on delivering personalized offers tailored to their past purchase behaviors. This strategy is designed to increase their purchase frequency and engagement at a low cost, with the potential to boost their overall value to the business.

Implement Tiered Loyalty Program

Introduce a tiered loyalty program to the 40.1% of mid-value customers ('Copper,' 'Bronze,' 'Silver'). These customers show moderate engagement but can be incentivized to increase their purchase frequency through loyalty rewards based on their spending. The objective is to encourage their transition into higher-value segments, thus driving greater long-term revenue.

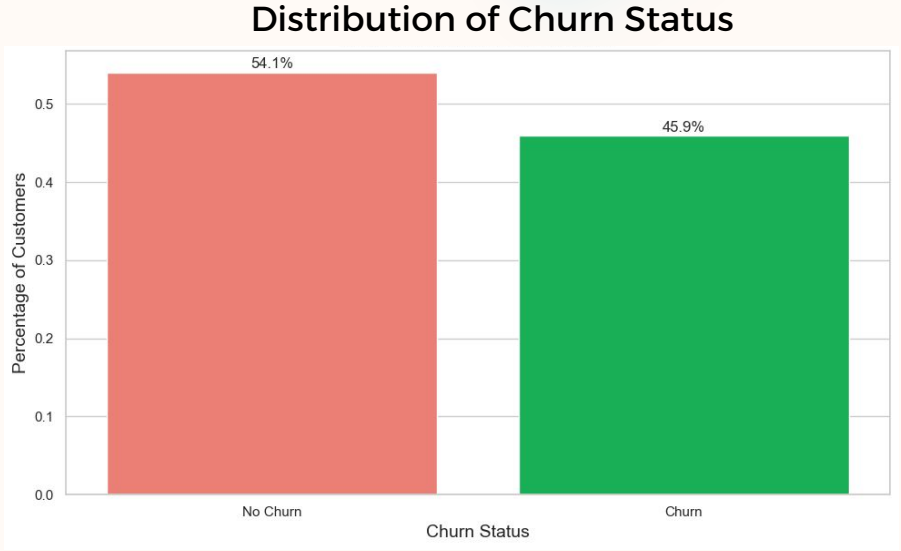
Develop VIP Program

Focus on the 21.3% of high-value customers ('Gold' and 'Diamond') by offering a VIP experience that includes exclusive offers, early access to new products, and premium customer service. This high-touch strategy is essential for strengthening their loyalty and emotional connection to the brand, reducing churn risk, and ensuring long-term profitability

Churn Status

1. Churn Definition: We used the average repurchase interval (105 days) from the RFM analysis as the threshold to identify churners, based on the typical time between purchases

2. Threshold Application: Customers who have not made a purchase within 105 days, starting from May 1st, 2023, were categorized as 'Churn.'



The bar plot shows a nearly balanced distribution, with 45.9% classified as 'Churn' and 54.1% as 'No Churn.'

Churn Model Features

For the features used in our churn model, we selected several metrics to predict customer behavior effectively:

1. **Frequency Score**: Tertiles (low, medium, high) measuring purchase frequency.
2. **Monetary Score**: Tertiles (low, medium, high) reflecting total customer spend.
3. **Gender**: Evaluates churn behavior differences by gender.
4. **Is Reviewer**: Binary indicator of whether a customer has left a review.
5. **Has Refunded**: Binary indicator tracking if a customer has requested refunds.
6. **Number of Unique Products Purchased**: Tertiles (low, medium, high) analyzing product variety.
7. **Percentage of Discounted Products**: Tertiles (low, medium, high) measuring the proportion of discounted purchases.
8. **Average Monetary Value Per Order**: Tertiles (low, medium, high) evaluating spend per order.
9. **Average Number of Products Per Order**: Tertiles (low, medium, high) measuring product variety per transaction.
10. **Total Products Purchased**: Tertiles (low, medium, high) representing total purchase volume.

Tertiles were used to manage the effect of outliers, particularly rare bulk orders, ensuring that extreme cases do not disproportionately skew the analysis.

Recency Score was excluded as our churn definition is based on customer activity within 105 days, preventing recency from dominating the model.

Training Setup

Validation Strategy: Data split into 70% training, 15% validation, and 15% testing using the holdout method.
Models Evaluated: Performance of Logistic Regression, Naive Bayes, Decision Tree, and KNN analyzed to determine the most effective model for churn prediction.
Metrics: Accuracy, Precision (Positive Class), Recall (Positive Class), and F1 Score (Positive Class) utilized to comprehensively assess each model's capability in accurately identifying churn.

Validation Results

	Accuracy	Precision (Positive Class)	Recall (Positive Class)	F1 Score (Positive Class)
Logistic Regression	**0.620**	**0.589**	**0.572**	**0.581**
Naive Bayes	0.608	0.567	0.625	0.594
Decision Tree	0.613	0.577	0.591	0.584
KNN	0.572	0.536	0.501	0.518

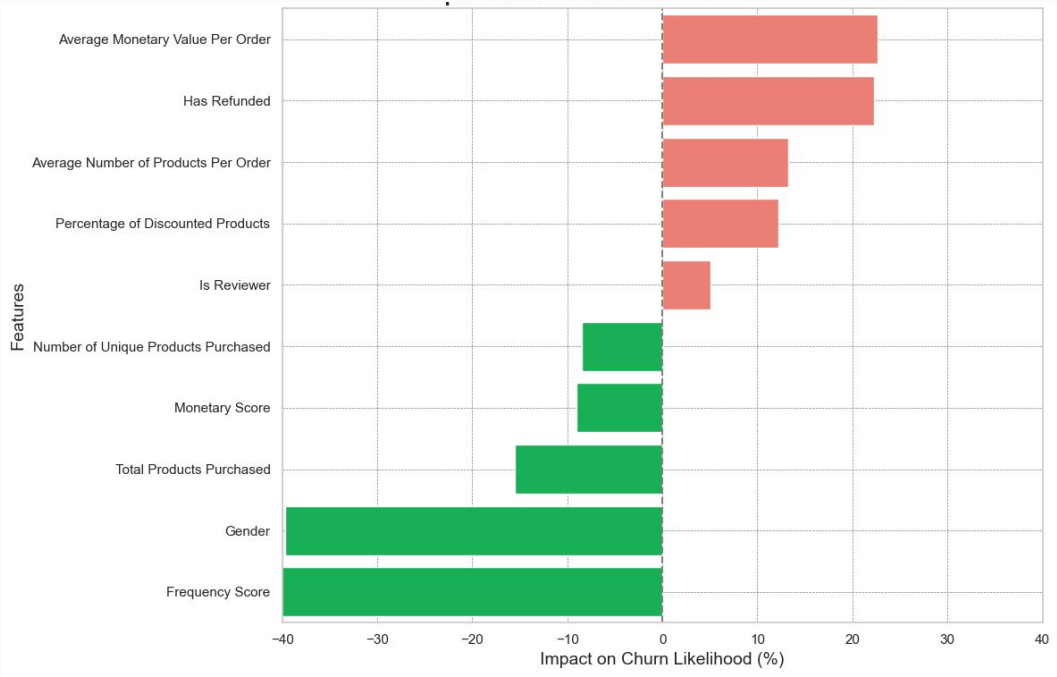
Selected Logistic Regression despite its slightly lower F1 score in the positive class due to its superior accuracy and interpretability. This balance allows for clearer insights and more effective churn reduction strategies.

Test Results

	Accuracy	Precision (Positive Class)	Recall (Positive Class)	F1 Score (Positive Class)
Logistic Regression	0.64	0.61	0.61	0.61

The Logistic Regression model demonstrated solid performance on the test set, achieving an accuracy of 0.64, with precision, recall, and F1 score each at 0.61.

Key Drivers of Customer Churn



Factors Increasing Churn Likelihood: Higher churn is linked to **Average Monetary Value Per Order**, **Has Refunded**, suggesting that high spending and refunds may push customers away.

Factors Reducing Churn Likelihood: **Frequency Score**, **Total Products Purchased**, and **Gender** lower churn risk, indicating that frequent purchases and product diversity strengthen customer loyalty.

Action Steps (Churn Model)

Refund Management Initiative

Establish a dedicated support team to proactively manage customers who have requested refunds, as this group is at a higher risk of churn. This team will promptly address and resolve refund issues, ensuring that customer concerns are resolved efficiently. Follow-up communications will be implemented to confirm satisfaction, aiming to reduce churn and improve overall retention.

Targeted Discount Strategies

Develop a program that offers personalized discounts to customers identified with high average spending per order, as this group showed a higher likelihood of churn. The discounts should be tailored based on the customer's purchasing history to enhance perceived value and foster loyalty. For instance, offering a 10% discount on the next purchase after a high-value order could mitigate the churn risk and encourage repeat business.

Product Focus

Business Questions

What are the most common product combinations purchased together?

How can we utilize these patterns to enhance marketing strategies and increase sales?

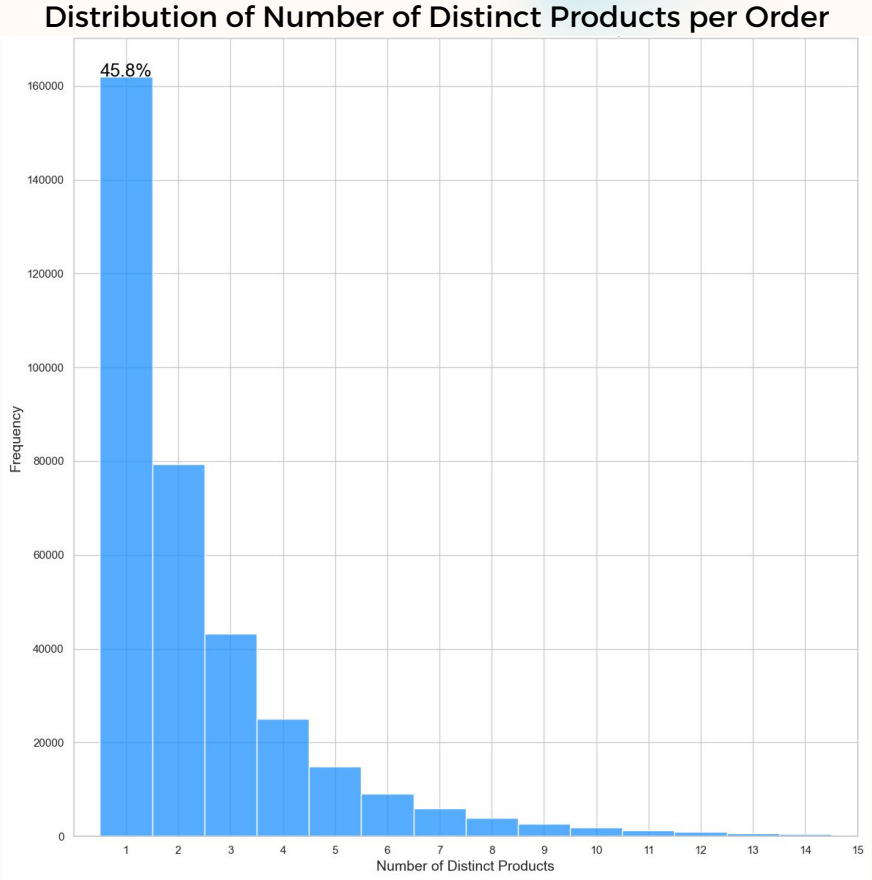
Approach

Market Basket Analysis: This technique analyzes the co-occurrence of products in transactions to discover patterns of items often purchased together, facilitating effective cross-selling and upselling strategies.

Product Diversity in Orders

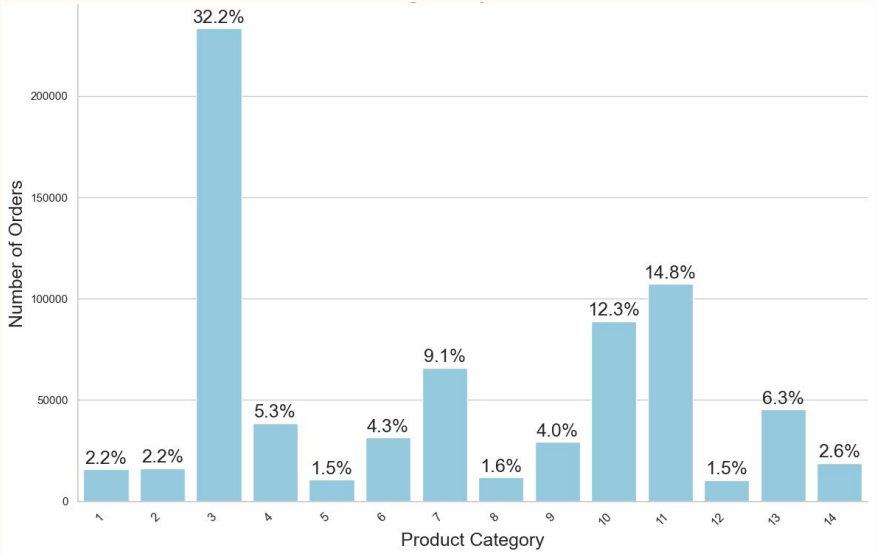
Single Product Orders
45.8% of all orders include only a single product, indicating a preference for limited-item purchases.

Market Basket Analysis
For a deeper understanding of purchasing behaviors, we will analyze orders containing **at least two products** to identify opportunities for cross-selling and bundling.



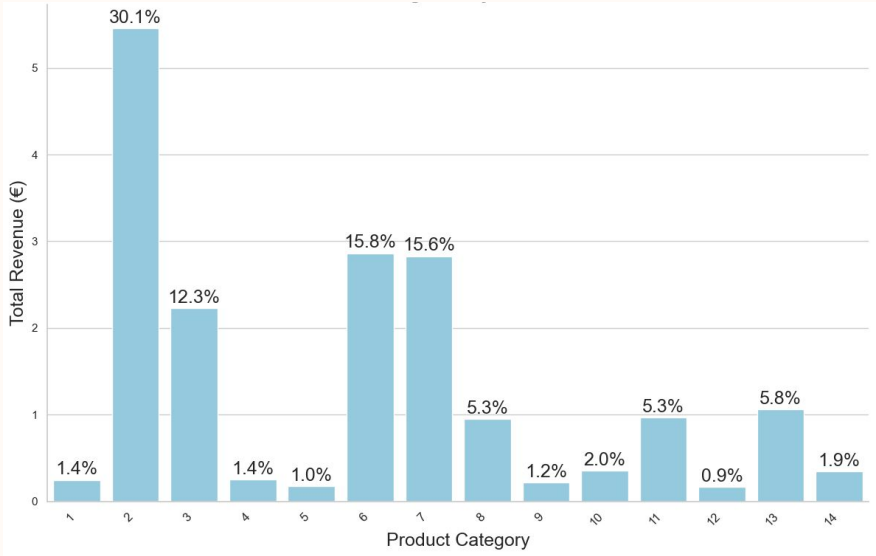
Category Analysis in Multi-Item Orders

Category Popularity in Multi-Item Orders



Category 3 emerges as the most popular, capturing 32% of orders, significantly impacting our sales strategy.

Total Revenue by Category in Multi-Item Orders

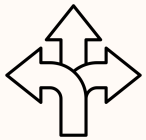


Category 2 dominates the revenue, contributing to 30.1% of total sales. The next significant contributors are **Categories 6 and 7**, each generating around 16% of the revenue.

Market Basket Analysis Setup

Data Filtering: Focused on orders with at least two distinct products, excluding returns, to ensure meaningful product associations.
Basket Format: Converted the data so each row represents a transaction, with columns indicating product purchases.
Apriori Algorithm: Applied the algorithm to uncover frequent itemsets and association rules based on minimum support and confidence thresholds.

Market Basket Analysis Approaches



All Categories: Included all product categories in the analysis.
Excluding Most Popular: Excluded the most popular category (Category 3) to focus on less dominant categories, reducing computational load and uncovering unique associations.
High-Value Categories: Concentrated on baskets from the top three highest-value categories (Categories 2, 6, and 7) to identify high-impact opportunities and apply more stringent thresholds.

Market Basket Analysis: Parameter Settings

Approach	Min Support	Min Confidence	Min Lift
All Categories	0.003	0.75	1
Excluding Most Popular	0.001	0.85	1
High-Value Categories	0.001	0.85	1

By removing the most popular category, we optimized our computational efficiency, which enabled us to lower the support thresholds for both the Excluding Most Popular and High-Value Categories approaches, uncovering valuable but less frequent associations more effectively.

Market Basket Analysis Results

All Categories Approach

Category 3, as the most popular, naturally generated the most rules, signifying its dominant influence in market interactions.

Antecedent Category	Consequent Category	Number Of rules
3	3	47

Excluding Most Popular Approach

Removing Category 3 highlighted less dominant categories, revealing 41 new rules. **Category 7** emerged as the most productive with 23 new rules, suggesting its strong potential for targeted marketing strategies

Antecedent Category	Consequent Category	Number Of rules
7	7	23
8	8	16
5	5	2

High-Value Categories Approach

Focused on Categories 2, 6 and 7, this approach showed significant rule generation in **Category 7** but none for Category 2, likely due to bulk ordering patterns that obscure meaningful associations

Antecedent Category	Consequent Category	Number Of rules
7	7	38
6	6	1

All generated rules were within the same category, indicating limited interactions between different categories

Action Steps (Market Basket Analysis)

Targeted Marketing for Category 7

Launch an email campaign targeting frequent buyers from Category 7, offering a 10% discount on next purchases that include complementary items from the same category. Highlight specific combinations that have shown frequent co-purchases in the 23 identified rules to encourage higher transaction values.

Promotional Campaign for Category 8

Implement a social media advertising campaign for Category 8, focusing on the unique product combinations found in the 16 association rules. Promote a 15% discount on these bundles to increase visibility and attract more buyers, enhancing both the popularity and profitability of this category.

Profitability Enhancement for Category 3

Use the 47 association rules identified to promote premium products within Category 3, which despite its high popularity, ranks lower in profitability. Offer a 5% discount on high-margin items within this category when bundled with more frequently purchased items, aiming to enhance overall profitability through upselling

Feedback Focus

Business Questions

What is the overall sentiment of customer reviews, and how does it impact our brand perception?

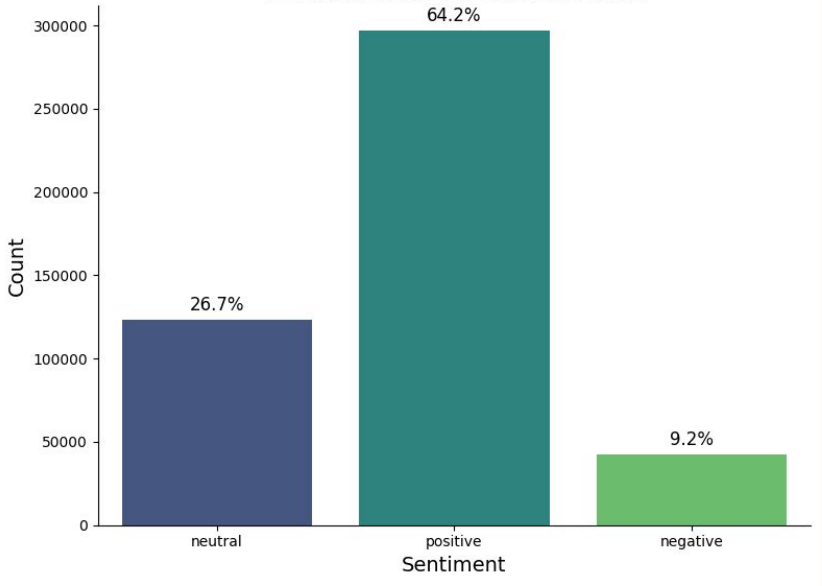
What specific actions can we take to address neutral and negative feedback while capitalizing on positive reviews?

Approach

Sentiment Analysis Model:
Analyzes the emotional tone of customer reviews to gauge sentiment, aiding in enhancing customer satisfaction and managing brand reputation effectively

Sentiment Distribution

64.2% of pre-labeled reviews display positive sentiment, showcasing customer satisfaction. Neutral and negative sentiments represent 26.7% and 9.2% respectively.



Model Training Objective

Aim to train a **Sentiment Analysis model** using the pre-labeled review data to effectively categorize and understand diverse customer sentiments.

Model Application

Utilize the trained model to **analyze unlabeled customer review** data, enabling targeted marketing and support interventions based on detected sentiments, and to refine customer engagement strategies.

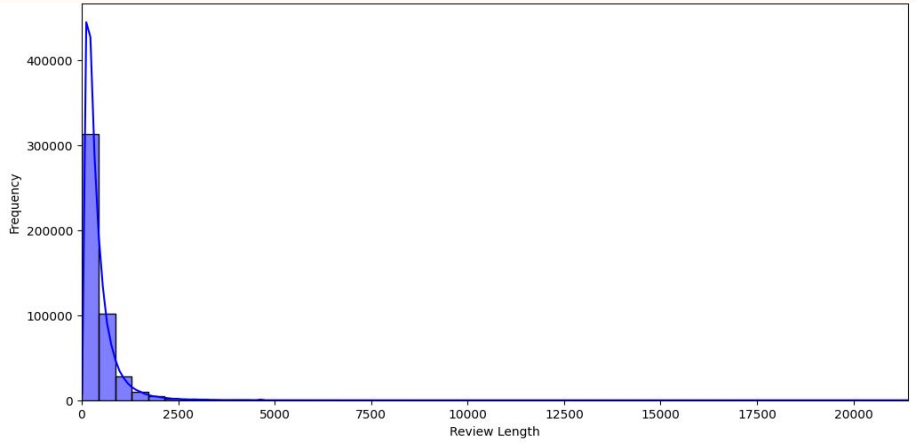
Preprocessing

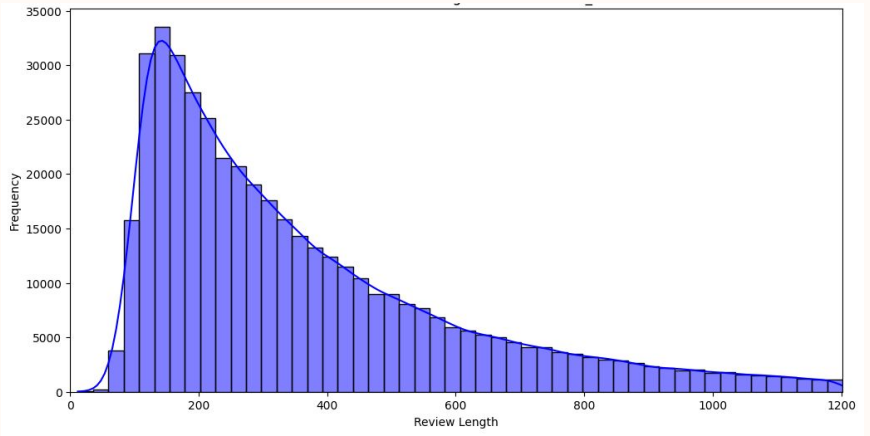
Outlier Removal: Removed reviews with text lengths above the 95th percentile to filter out overly long reviews.

Text Cleaning: Eliminated punctuation and special symbols to clean the dataset.

Lemmatization: Reduced words to their base forms to ensure consistency in text analysis.

The left plot shows the original distribution of review lengths, which included significant outliers, some extending beyond 17,500 characters. Following the removal of these extreme cases, the right plot displays a refined distribution, capped at 1,202 characters





Training Setup

Validation Strategy: Data split into 60% training, 20% validation, and 20% testing segments using the holdout method.
Feature Extraction: Utilized TF-IDF vectorization considering both single words and bigrams to capture nuanced language patterns in reviews.
Models Evaluated: Logistic Regression, Linear SVM, and Naive Bayes used to assess effectiveness in sentiment classification.
Metrics: Selected Accuracy, Macro F1, and Weighted F1 scores to provide a comprehensive evaluation of model performance across multiple classes

Validation Results

Model	Accuracy	Macro F1	Weighted F1
Logistic Regression	**0.84**	**0.77**	**0.83**
Linear SVM	0.84	0.77	0.83
Naive Bayes	0.67	0.32	0.55

Although Logistic Regression and SVM performed equally well across all metrics, **we selected Logistic Regression** due to its interpretability. This allows us to better understand feature importance, making it easier to extract actionable insights for targeted strategies

Test Results

	Accuracy	Macro F1	Weighted F1
Logistic Regression	0.84	0.77	0.83

The model achieved an accuracy of 0.84, with a macro F1 score of 0.77 and a weighted F1 score of 0.83, indicating a strong overall ability to classify sentiments correctly.

Test Confusion Matrix

Negative: 5235 correct, misclassified as neutral (1778) and positive (995), indicating difficulty in distinction.

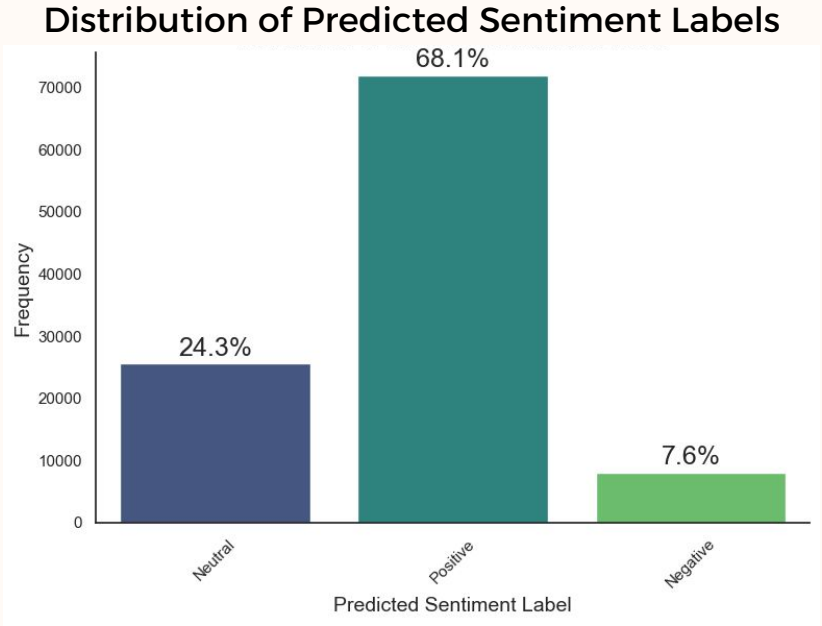
Neutral: 14640 correct, with 7300 misclassified as positive, showing overlap with positive sentiments.

Positive: Highest accuracy (53672 correct), though 3021 confused with neutral, suggesting minor confusion.

Unlabeled Reviews: Sentiment Predictions

Application on Unlabeled Data:
The retrained model was applied to a separate dataset of unlabeled customer reviews to predict their sentiments.

Predicted Sentiment Distribution:
The distribution of sentiment labels shows 68.1% positive reviews, 24.3% neutral, and 7.6% negative.



Visualizing Sentiments: Word Clouds from Model Predictions

Word Cloud of Model-Labeled Positive Reviews

Most Influential Words in Label Prediction in order
best, delicious, great, love, perfect

Word Cloud of Model-Labeled Negative Reviews

Most Influential Words in Label Prediction in order
Worst, terrible, awful, disgusting, horrible

The green word cloud for positive sentiment highlights frequent terms like "good," "delicious," and "enjoy," mainly focused on food-related topics, with influential words such as "best" and "perfect" driving the model's positive classification. In contrast, the red word cloud for negative sentiment features terms like "flavor" and "coffee," often paired with negative descriptors like "disappointed" and "terrible," with influential terms such as "worst" and "horrible" strongly contributing to negative classifications.

Action Steps (Sentiment Analysis)

Reward and Amplify Loyal Customers

Since the majority of customer feedback (68.1%) is positive, launch a campaign focused on rewarding and amplifying these loyal customers. Offer personalized discounts or exclusive offers to encourage them to continue their positive relationship with the brand. This could further strengthen brand loyalty and promote positive word-of-mouth

Convert Neutral Customers into Promoters

Addressing the 24.3% of feedback that is neutral, we suggest implementing follow-up surveys or targeted communications to convert neutral customers into promoters. This approach aims to shift their neutral stance to a more positive one, enhancing overall customer satisfaction.

Resolve Negative Feedback to Improve Retention

Given that negative reviews constitute 7.6% of the feedback, it is crucial to engage these customers directly. By designing specialized customer service campaigns that offer resolutions or incentives, we aim to improve customer satisfaction and retention significantly.