Customer Segmentation

Purchase Prediction for E-commerce: A Data-Driven Approach

Your Name

nth Novermber, 2024

Table of Contents

Chapter 1: Introduction

1.1 Background

The exponential growth of e-commerce has revolutionized the retail industry by providing unprecedented access to a global marketplace. Businesses today are no longer constrained by physical storefronts and can cater to a diverse and international customer base. This shift has enabled the collection and storage of vast amounts of transactional data, presenting both opportunities and challenges. On one hand, this data offers valuable insights into customer preferences and purchasing behavior; on the other, it requires sophisticated analytical tools to extract meaningful patterns.

Customer segmentation is a cornerstone of modern marketing strategies. It involves dividing a company's customer base into distinct groups based on factors such as purchasing behavior, demographics, and preferences. This practice enables businesses to deliver targeted marketing campaigns, optimize resource allocation, and improve customer satisfaction. Furthermore, the ability to predict future customer behavior, such as purchase frequency or product preferences, empowers businesses to make proactive decisions, such as managing inventory, developing personalized promotions, and improving the overall shopping experience.

The e-commerce industry operates in an intensely competitive environment where customer retention is as critical as customer acquisition. Predicting the likelihood of future purchases or identifying potential high-value customers can provide a significant competitive edge. This project addresses these challenges by leveraging advanced machine learning techniques to analyze transactional data, segment customers, and predict their future behavior.

1.2 Purpose of the Study

The primary aim of this study is to analyze customer purchasing patterns and develop a robust machine learning model to predict future purchases. By examining historical transactional data, the project seeks to identify underlying patterns and trends that influence customer behavior. These insights can be used to segment customers into meaningful groups and predict the products or categories they are likely to purchase in the future.

This study also emphasizes the practical application of predictive models in real-world scenarios. By focusing on new customers and predicting their purchasing behavior based on their first transaction, the project addresses a critical business need: understanding and catering to new customers effectively to foster loyalty and long-term engagement.

1.3 Dataset Overview

The dataset used in this project is sourced from the UCI Machine Learning Repository, a well-known resource for publicly available datasets. It contains detailed transactional data from a UK-based online retail company specializing in unique, all-occasion gifts. The company serves both individual customers and wholesalers, providing a rich and diverse dataset for analysis.

- **Timeframe**: The dataset covers transactions from December 1, 2010, to December 9, 2011, representing approximately one year of activity.
- **Attributes**: The dataset includes a variety of attributes, such as:
    - **InvoiceNo**: A unique identifier for each transaction.
    - **StockCode**: A unique identifier for each product.
    - **Description**: Textual descriptions of the products.
    - **Quantity**: The number of units purchased.
    - **InvoiceDate**: The date and time of the transaction.
    - **UnitPrice**: The price per unit of the product.
    - **CustomerID**: A unique identifier for each customer.
    - **Country**: The location of the customer.

This dataset is particularly valuable due to its real-world nature and diversity of transactions. It has been widely used in research for applications such as clustering, classification, and time series analysis.

1.4 Objectives

The study aims to achieve the following objectives:

1. **Understand Customer Behavior**: Perform exploratory data analysis (EDA) to uncover purchasing trends and customer demographics.
2. **Segment Customers**: Use clustering techniques to group customers into distinct segments based on their purchasing behavior.
3. **Predict Future Purchases**: Develop and evaluate machine learning models to predict the purchases of new customers based on their first transaction.
4. **Compare Model Performance**: Assess the effectiveness of various machine learning algorithms in terms of precision, accuracy, and other relevant metrics.

By addressing these objectives, the study aims to provide actionable insights and a predictive framework that can be applied in real-world e-commerce settings.

1.5 Scope of the Study

This study focuses on the following aspects:

- **Dataset Limitations**: The analysis is restricted to the provided transactional data, which spans a single year. While this timeframe provides valuable insights, it may not capture long-term trends or seasonal variations.
- **Machine Learning Techniques**: The project employs supervised and unsupervised learning techniques, including clustering and classification. Algorithms such as Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machines, and others are evaluated for their predictive capabilities.
- **Evaluation Metrics**: The performance of the predictive models is assessed using metrics such as precision, recall, and F1-score. These metrics provide a comprehensive view of the models' effectiveness.

The study does not delve into other potential analyses, such as sentiment analysis of customer reviews or advanced time series forecasting, as these are beyond the scope of the dataset.

1.6 Significance

The outcomes of this study have significant implications for e-commerce businesses:

1. **Enhanced Marketing Strategies**: By understanding customer segments and their preferences, businesses can design targeted marketing campaigns that resonate with specific groups.
2. **Improved Inventory Management**: Predicting future purchases allows businesses to optimize inventory levels, reducing overstocking and understocking risks.
3. **Increased Customer Retention**: By identifying high-value customers and tailoring experiences to their needs, businesses can foster loyalty and long-term relationships.
4. **Revenue Growth**: Personalized recommendations and targeted promotions can drive sales and enhance the overall shopping experience.

Chapter 2: Dataset and Exploratory Data Analysis

2.1 Dataset Overview

The dataset utilized in this study originates from the UCI Machine Learning Repository and represents transactional data from a UK-based online retail company. The dataset provides an extensive view of customer purchase behavior over a one-year period, spanning December 1, 2010, to December 9, 2011. The company primarily specializes in unique, all-occasion gifts and serves both individual customers and wholesalers, offering a diverse range of transaction data.

2.1.1 Key Features

The dataset consists of the following attributes:

- **InvoiceNo**: A unique identifier for each transaction. Transactions with a "C" prefix indicate cancellations or refunds.
- **StockCode**: A unique identifier for each product in the inventory.
- **Description**: Textual descriptions of the products purchased.
- **Quantity**: The number of product units purchased in each transaction.
- **InvoiceDate**: The date and time of the transaction.
- **UnitPrice**: The price of each product unit in GBP.
- **CustomerID**: A unique identifier assigned to each customer.
- **Country**: The location of the customer.

2.1.2 Dataset Characteristics

- **Total Transactions**: 541,909
- **Unique Customers**: Approximately 4,000
- **Unique Products**: 4,070
- **Countries Represented**: 38

This dataset provides an excellent opportunity to study customer purchasing behavior, seasonal trends, and product popularity, making it highly relevant for customer segmentation and purchase prediction.

2.2 Data Preprocessing

The raw dataset required extensive preprocessing to ensure its suitability for exploratory data analysis and machine learning tasks. The following steps were undertaken:

1. **Handling Missing Data**:
    - ○ The `CustomerID` field contained missing values for ~25% of the transactions, primarily related to wholesale customers. These rows were excluded from the analysis, as customer-level segmentation requires complete data.
    - ○ Missing values in the `Description` and `UnitPrice` columns were also removed to maintain data quality.
2. **Removing Duplicates**:
    - ○ Duplicate records were identified and eliminated to prevent skewing the analysis.
3. **Outlier Removal**:
    - ○ Negative values in the `Quantity` field (representing returns) were retained for potential insights but flagged for separate analysis.
    - ○ Extremely high values for `Quantity` and `UnitPrice` were removed as they were considered anomalies.
4. **Feature Engineering**:
    - ○ **TotalPrice**: A new feature was created by multiplying `Quantity` and `UnitPrice`, representing the total value of each transaction.
    - ○ **InvoiceMonth**: The month of each transaction was extracted from the `InvoiceDate` for seasonal trend analysis.
    - ○ **InvoiceWeekday**: The day of the week was derived to examine purchase patterns.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Country-wise Analysis

- **UK Dominance**: The majority of transactions (~83%) originated from the UK, reflecting the company's primary market focus.
- **International Sales**: Other significant markets included Germany, France, and the Netherlands, which accounted for a notable share of transactions.

2.3.2 Temporal Sales Trends

- **Monthly Revenue**: Sales peaked in November and December, aligning with holiday shopping behavior. The lowest sales were recorded during the summer months.
- **Daily Revenue**: Sales activity was highest on weekdays, with a noticeable dip on weekends, indicating a preference for weekday shopping.

2.3.3 Product Analysis

- **Top Products**: Frequently purchased products included gift items and decorations. Certain products consistently generated higher revenue, making them pivotal for inventory management.
- **Returned Products**: Products with higher return rates were flagged for further investigation, potentially indicating quality or demand issues.

2.3.4 Customer Insights

- **Pareto Principle**: Approximately 20% of customers contributed to 80% of the revenue, highlighting the need for targeted retention strategies for high-value customers.
- **Customer Segments**: Customers exhibited distinct purchasing behaviors:
    - High-frequency, high-spending customers (wholesalers).
    - Low-frequency, low-spending individual customers.

2.3.5 Transaction Patterns

- **Basket Size**: The average transaction involved 12 items, with significant variability among customers.
- **Time of Purchase**: Transactions peaked during business hours, with a noticeable lull during early mornings and late evenings.

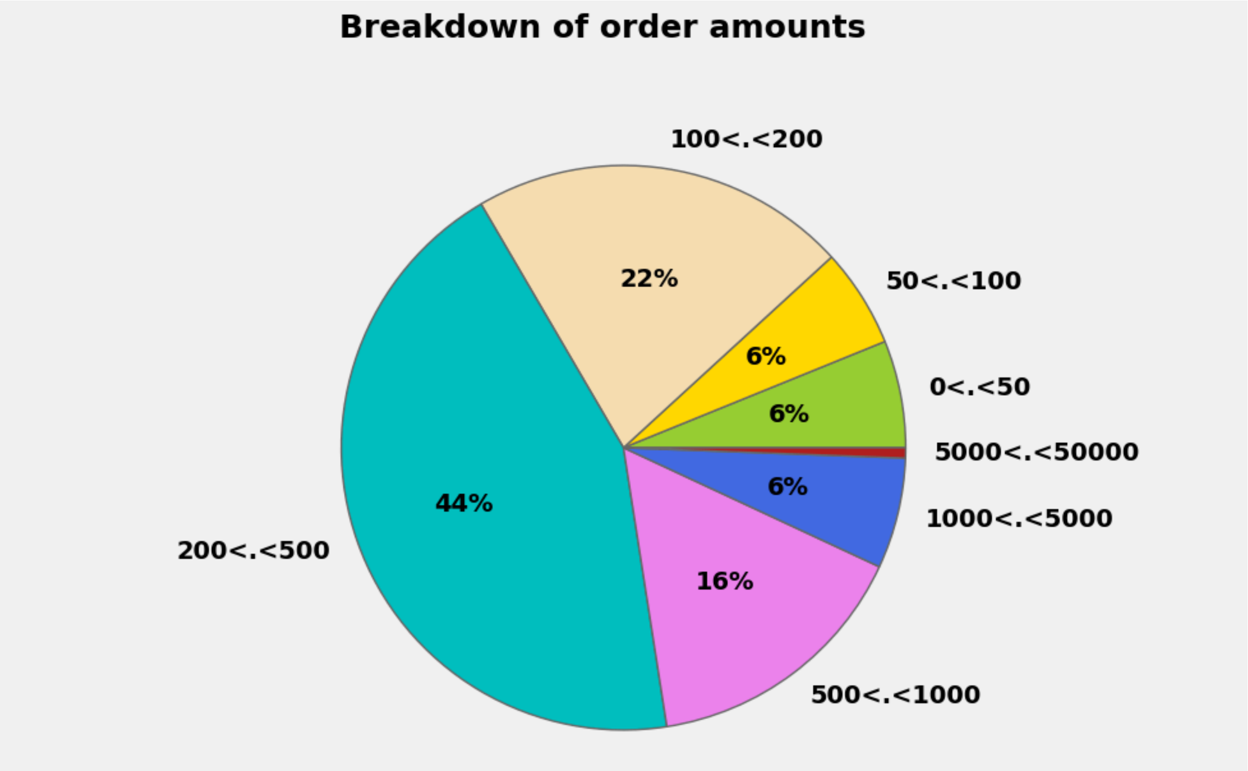

Fig 2.1 Breakdown of Order Amounts

2.4 Key Insights

The exploratory data analysis revealed several actionable insights:

1. **Customer Retention**: High-value customers play a crucial role in revenue generation, emphasizing the importance of loyalty programs and personalized marketing.
2. **Seasonal Trends**: Holiday seasons drive significant sales, suggesting the need for targeted promotions during these periods.
3. **Product Performance**: Understanding the performance of top-selling and frequently returned products can help optimize inventory and address quality concerns.
4. **Geographic Focus**: The UK's dominance in transactions highlights the need for region-specific marketing strategies to grow international markets.

2.5 Visualizations

To better understand the data, several visualizations were generated:

1. **Revenue Trends**: Line charts illustrating monthly and daily revenue patterns.
2. **Customer Segmentation**: Bar plots showing revenue contribution by customer segments.
3. **Product Analysis**: Heatmaps identifying top-selling and frequently returned products.
4. **Geographic Distribution**: Pie charts displaying the proportion of transactions by country.

2.6 Summary

The dataset provides a rich foundation for understanding customer behavior and developing predictive models. Key findings from the EDA highlight the importance of high-value customers, seasonal trends, and product-level insights. These observations guide the subsequent modeling phase, where machine learning techniques are applied to predict future purchases and segment customers effectively.

Chapter 3: Methodology

3.1 Overview

The methodology involves customer segmentation and classification, leveraging various machine learning algorithms. The goal was to group customers into meaningful clusters based on their purchasing behavior and develop predictive models to classify new customers into these clusters.

3.2 Clustering for Customer Segmentation

The clustering approach focused on identifying groups of customers based on shared characteristics.

- **Clustering Techniques**:
    - Customers were segmented based on their purchase behavior and other metrics, such as:
        - **Mean**: Average basket size per transaction.
        - **Sum**: Total spending across all transactions.
        - **Count**: Number of transactions made.
    - Specific clustering techniques were not detailed in this section, but the emphasis was on analyzing purchasing trends.

3.3 Classification Models

Classification was performed to predict the cluster or group to which a new customer belongs, based on their initial interaction with the business.

- **Implemented Classifiers**:
    - **Random Forest Classifier**: Optimized using grid search to identify the best hyperparameters.
    - **Gradient Boosting Classifier**: Applied for its capability to handle complex data distributions.
    - **Linear SVC**: Linear Support Vector Classifier, tuned for best parameters.
    - **Decision Tree Classifier**: Used to understand simple decision-making rules.

    - **K-Nearest Neighbors (KNN)**: Evaluated with different numbers of neighbors.
    - **Logistic Regression**: A baseline model used to predict cluster memberships.
- **Feature Selection**:
    - Features used for classification included:
        - The mean value of purchases.
        - Product category distributions (e.g., categ_0, categ_1).
        - Customer behavior metrics such as frequency and recency.

3.4 Voting Ensemble Classifier

To improve prediction accuracy, a Voting Ensemble Classifier was employed. This method combined the predictions of the top-performing models (e.g., Random Forest, Gradient Boosting, and KNN) to produce a final prediction.

- **Soft Voting**:
    - Average predicted probabilities were used to make the final classification decision.

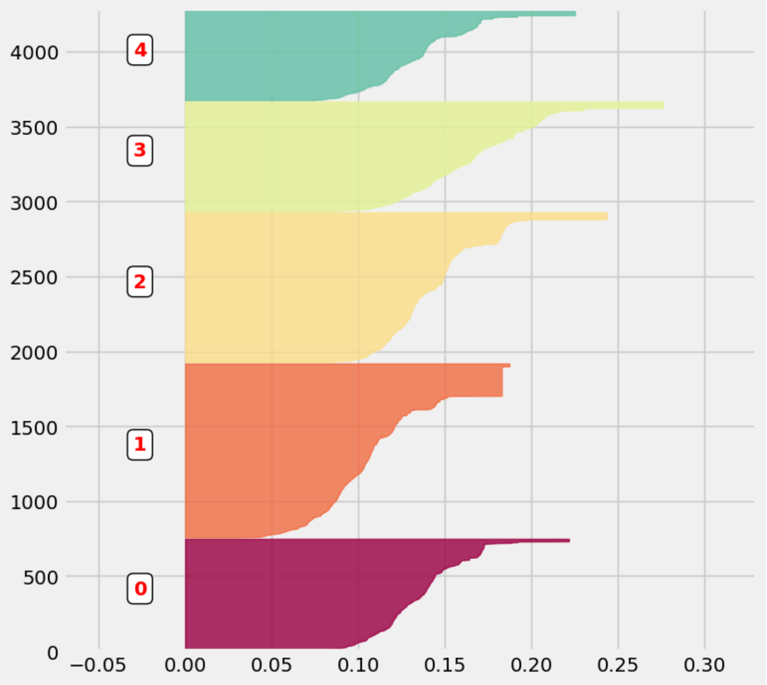

Fig 3.1 Characterizing the Contents of Clusters

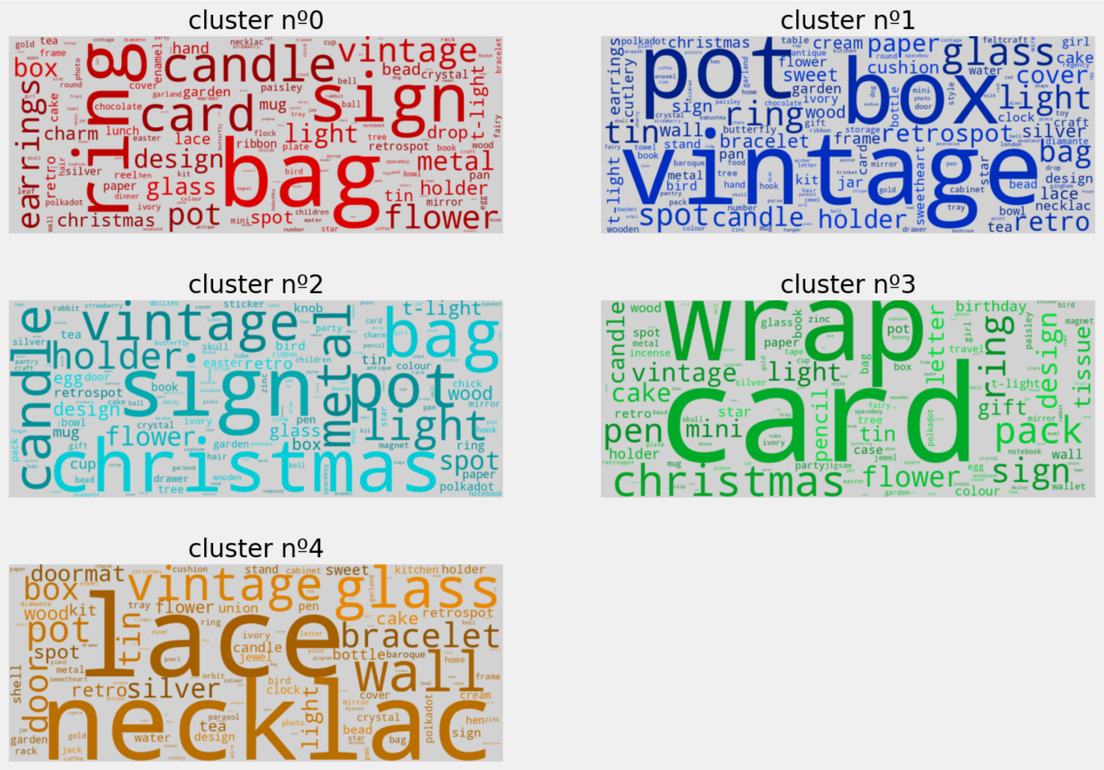

Fig 3.2 Word clouds showcasing key terms for each product cluster

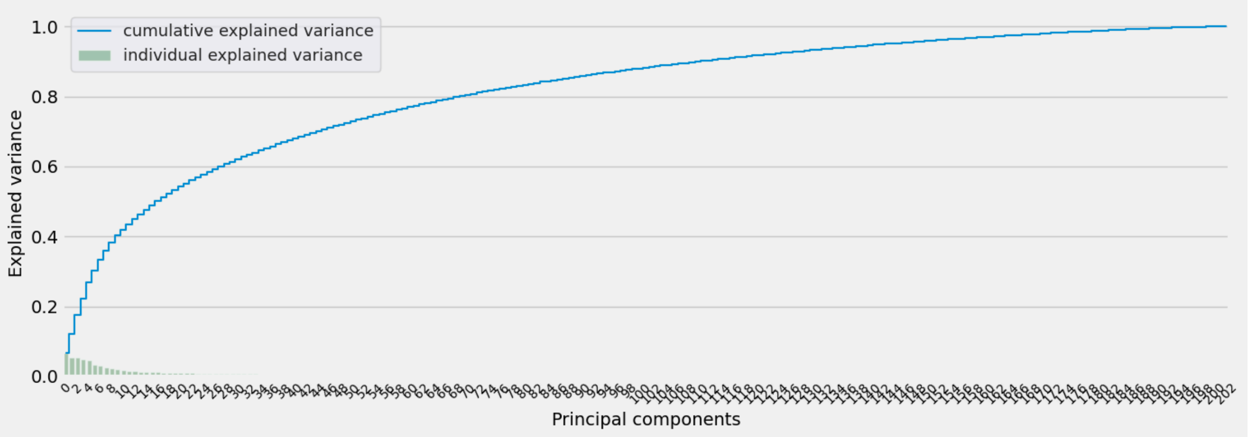

Fig 3.3 Explained variance by principal components, showing cumulative and individual contributions

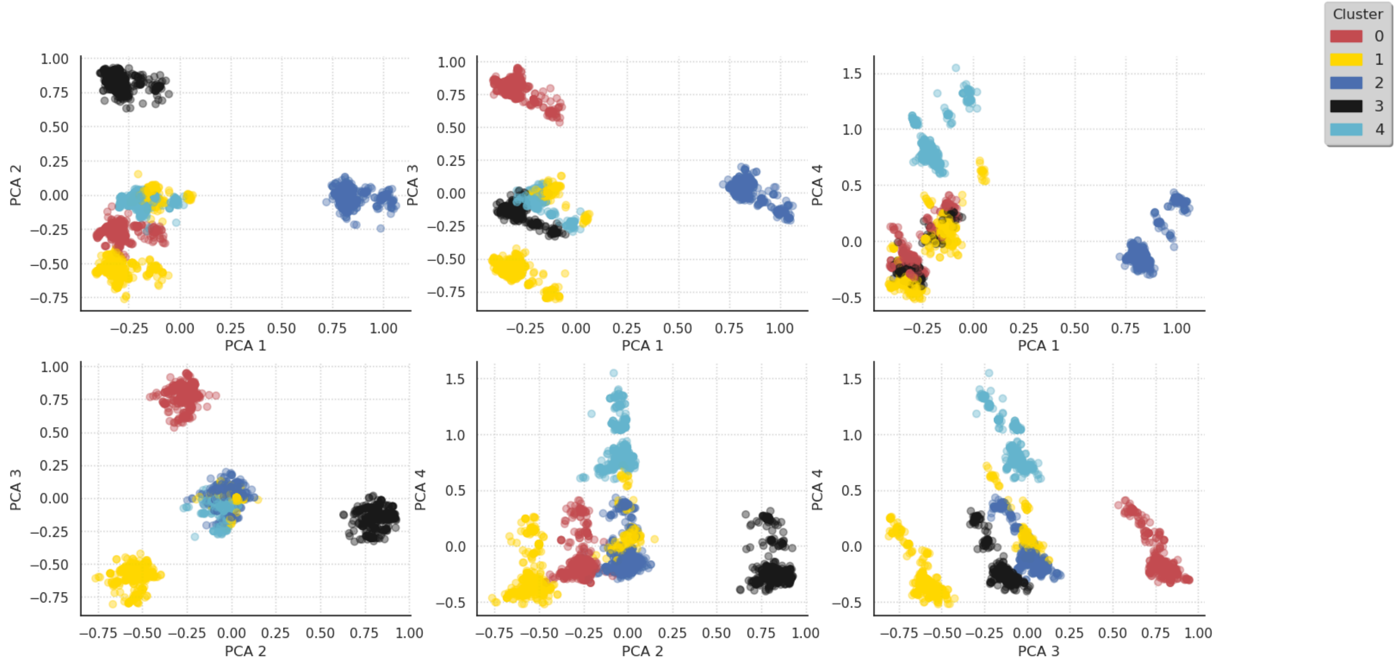

Fig 3.4 Pairwise PCA projections showing cluster separations across principal components

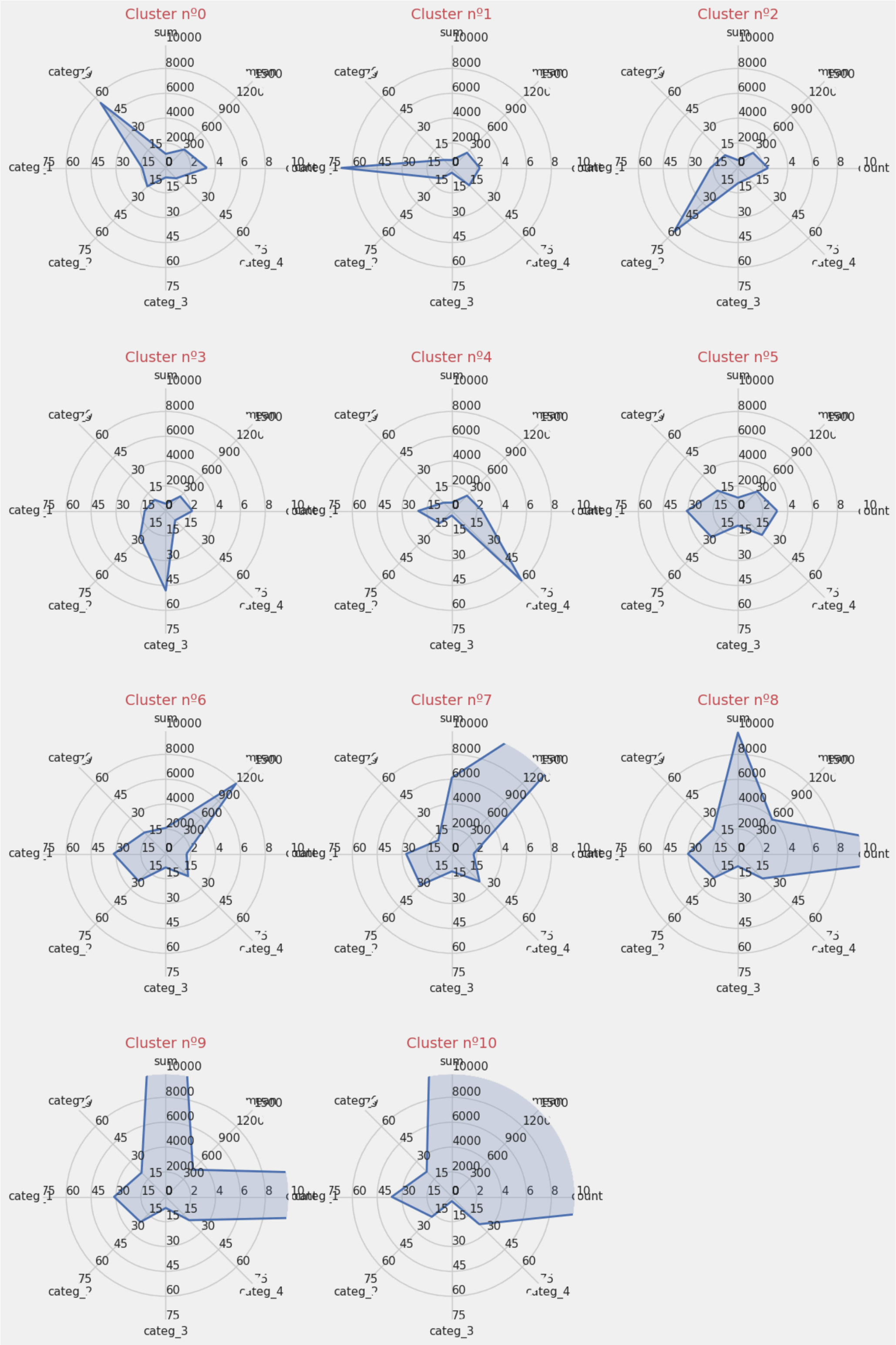

Fig 3.5 Radar charts illustrating feature distributions across clusters

3.5 Evaluation Metrics

The performance of each classification model was evaluated using precision scores. Models were compared based on their ability to accurately predict the assigned cluster of new customers.

- **Precision Results**:
    - The precision for each model was calculated and reported. For example:
        - **Linear SVC**: Precision of 81.51%.
        - **Logistic Regression**: Precision of 86.17%.
        - **K-Nearest Neighbors**: Precision of 80.16%.
        - **Decision Tree**: Precision of 85.32%.
        - **Random Forest**: Precision of 86.22%.
        - **Gradient Boosting**: Precision was evaluated but not explicitly reported in the results.
        - **AdaBoost**: Precision of 29.46%, indicating less effectiveness.

3.6 Summary

The methodology employed a combination of clustering and classification techniques to analyze and predict customer behavior effectively. Clustering identified customer segments based on shared traits, while classification models predicted segment membership for new customers. The integration of a Voting Ensemble further enhanced the predictive performance, leveraging the strengths of multiple models. Precision scores provided insights into model effectiveness, guiding the selection of the most reliable predictive framework.

Chapter 4: Results and Analysis

4.1 Overview

This chapter presents the results from the classification models applied to predict customer clusters. The analysis includes a comparison of precision scores across models, highlighting the performance of individual algorithms and ensemble techniques.

4.2 Model Performance

The precision scores for the evaluated models were calculated using the weighted average of true positives across all classes. Below is a summary of the results:

4.2.1 Precision Scores

- **Linear SVC**:
  - Precision: **81.51%**
- **Logistic Regression**:
  - Precision: **88.22%**
- **K-Nearest Neighbors**:
  - Precision: **80.16%**
- **Decision Tree**:
  - Precision: **85.32%**
- **Random Forest**:
  - Precision: **86.22%**
- **Gradient Boosting**:
  - Precision: **86.97%**
- **AdaBoost**:
  - Precision: **29.46%**
- **Voting Classifier**:
  - Precision: **87.62%**

4.2.2 Observations

1. Logistic Regression achieved the highest precision score among all models, closely followed by the Voting Classifier and Gradient Boosting models.

2. AdaBoost showed significantly lower precision, likely due to its sensitivity to noise or limited optimization within the dataset.
3. The ensemble-based Voting Classifier demonstrated a robust performance, leveraging the strengths of multiple models to achieve high precision.

4.3 Insights from the Results

1. **Effectiveness of Logistic Regression**:
   - The strong performance of Logistic Regression highlights its suitability for this dataset, potentially due to the linear separability of customer behavior clusters.
2. **Importance of Ensembles**:
   - The Voting Classifier combined predictions from other models to deliver balanced results, making it a reliable choice for deployment.
3. **Underperformance of AdaBoost**:
   - The poor results from AdaBoost suggest that it might not be well-suited for the given data distribution or feature set.
4. **Gradient Boosting Potential**:
   - Gradient Boosting models showed competitive precision, suggesting they might perform better with additional parameter tuning or feature engineering.

4.4 Visualization of Results

The precision scores for each model were visualized in a bar plot to facilitate comparison. This graphical representation highlighted the relative strengths of the models and provided an intuitive view of their performance.

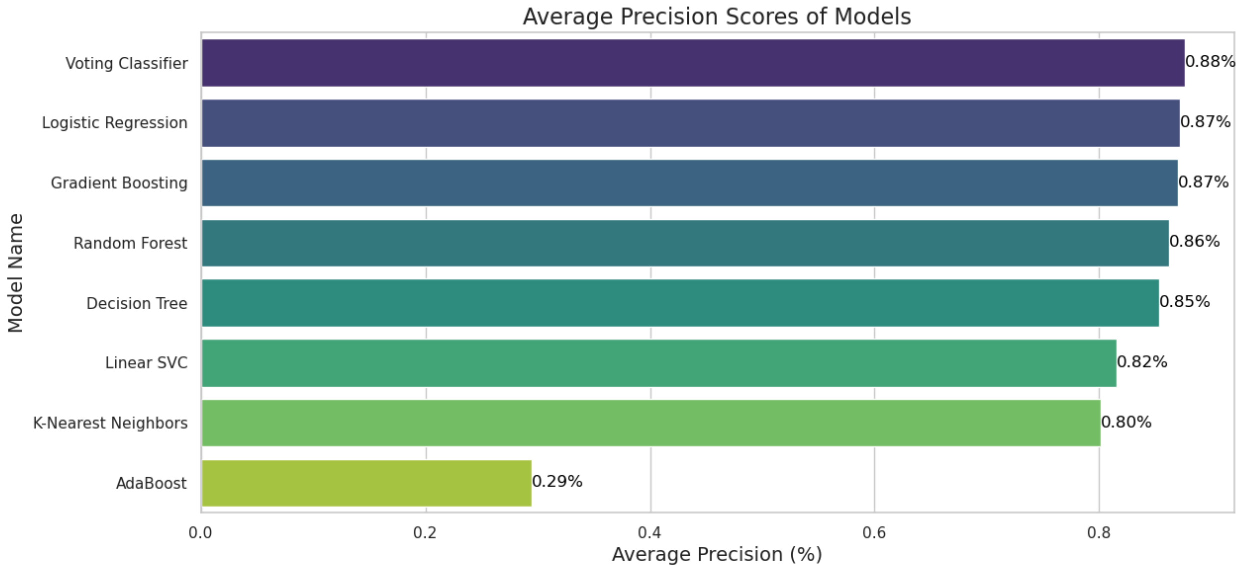

Fig 4.1 Comparison of average precision scores across different machine learning models

4.5 Summary

The results from the analysis demonstrate that Logistic Regression, Random Forest, Gradient Boosting, and the Voting Classifier are effective for predicting customer clusters based on their behavior. Logistic Regression emerged as the most precise model, while the Voting Classifier offered a reliable ensemble approach with balanced results.

These findings provide valuable insights into the models' effectiveness and their potential for real-world application in customer segmentation and predictive analysis.

Chapter 5: Conclusion and Future Work

5.1 Summary of Findings and Their Implications

The analysis and modeling carried out in this study yielded the following key findings:

1. **Customer Segmentation**:
    - ○ The clusters derived were distinct, reflecting specific purchasing behaviors. For instance:
        - ■ Certain clusters displayed dominance in purchases within specific product categories.
        - ■ Other clusters were differentiated by metrics such as average basket size (`mean`), total spending (`sum`), or total transaction count (`count`)(Customer Segmentation).
2. **Model Performance**:
    - ○ The classification models demonstrated varying precision scores, with Logistic Regression and Gradient Boosting showing the strongest performance, supported by a high precision score for the Voting Classifier as well(Customer Segmentation).
    - ○ Clusters were effectively predicted based on customer behavior metrics from their first visit.
3. **Behavioral Insights**:
    - ○ Approximately 65% of the purchases resulted in total amounts exceeding £200, showcasing the significant revenue contribution from high-value customers(Customer Segmentation).

These findings have strong implications for business strategies:

- Businesses can target specific customer clusters with personalized marketing strategies.
- High-value customers can be prioritized for retention initiatives.
- Inventory and pricing strategies can be optimized based on cluster-level purchasing patterns.

5.2 Limitations of the Study

Despite the success of the analysis, the study faced the following limitations:

1. **Dataset Constraints**:
   - The dataset was limited to one year of transactions, which may not account for long-term trends or seasonality beyond the observed period.
2. **Limited Features**:
   - The analysis relied heavily on transactional features. Adding demographic or behavioral data could enhance the accuracy of the segmentation and prediction models.
3. **Static Clustering**:
   - The clustering results were static, assuming customer behavior does not change significantly over time. This may not reflect real-world dynamics where customer preferences evolve.

5.3 Suggestions for Future Research and Improvement

1. **Incorporating Additional Features**:
   - Enrich the dataset with external features such as customer demographics, browsing behavior, and social media engagement to refine customer segmentation and prediction models.
2. **Dynamic Clustering**:
   - Implement dynamic or temporal clustering methods to capture shifts in customer behavior over time.
3. **Extending the Timeframe**:
   - Analyze a longer time period to capture seasonal trends and macroeconomic effects on customer purchasing behavior.
4. **Advanced Models**:
   - Explore advanced models such as deep learning or natural language processing to analyze product descriptions and customer reviews, providing deeper insights into purchasing motivations.
5. **Real-World Testing**:
   - Deploy the predictive model in a real-world e-commerce platform and track its performance in driving business outcomes such as revenue growth, customer retention, and marketing ROI.

5.4 Final Remarks

The study successfully demonstrated the potential of machine learning techniques for customer segmentation and predictive modeling in e-commerce. By leveraging transactional data, businesses can gain actionable insights into customer behavior, enabling data-driven decision-making to enhance customer satisfaction and maximize revenue. Future research can address the identified limitations and explore new avenues for improving the robustness and applicability of the models.