Retail Sales Data Analysis Report

Project Overview

This report provides a detailed analysis of customer spending behavior based on the Retail Sales Dataset. The goal was to determine if customer demographics and purchase attributes can predict whether a customer is a high spender. The analysis used various statistical methods and a K-Nearest Neighbors (KNN) model for classification.

Business Case

In a competitive retail market, understanding customer behavior is vital for effective marketing and resource allocation. Classifying customers as high or low spenders can help businesses tailor their strategies, optimize investments, and increase revenue.

Phases of Analysis

Phase 1: Data Cleaning and Preprocessing

1. The dataset was cleaned by:
    - Standardizing column names.
    - Converting dates to a uniform format.
    - Removing rows with missing or irrelevant values (e.g., incomplete dates and 2024 data).
    - Dropping identifiers like transaction and customer IDs.
    - Normalizing numerical columns such as age, quantity, price per unit, and total amount.
2. **Key Outcome**: A processed and normalized dataset ready for exploratory and predictive analyses.

Phase 2: Exploratory Data Analysis (EDA)

1. Spending Patterns

- Customers spend the most on **Beauty** and **Electronics** products.
- **Clothing** sees the lowest spending.

2. Class Imbalance

- **80% of customers** were low spenders, and only **20%** were high spenders.
- This imbalance necessitated adjustments for modeling.

3. Relationship Analysis

- **Gender and Spending**: Minimal variation in spending between males and females.
- **Age and Spending**: Spending decreases consistently with age. Younger customers spend more, especially on Beauty and Electronics.
- **Product Categories**: Electronics and Beauty are the highest revenue generators across all age groups.

4. Statistical Insights

- A Chi-Square test showed no significant association between gender and product categories.
- Correlation analysis revealed:
    - A strong positive correlation between price per unit and total amount.
    - A moderate correlation between quantity and total amount.

5. Visual Highlights

- **Class Distribution:** Bar chart illustrating the imbalance between high and low spenders.
- **Box Plots:** Spending by gender and product category
- **Bar Charts**: Average spending by age groups and product categories.
- **Correlation Heatmap**: Showcasing relationships among numerical features.

Phase 3: Predictive Modeling

K-Nearest Neighbors (KNN) Model

- **Objective**: Classify customers as high spenders based on demographics and purchase attributes.
- **Results**:
    - **Overall Accuracy**: 95%
    - **Recall for High Spenders**: 98%
    - **Precision for High Spenders**: 85%

- **Confusion Matrix**:
  - True Positives: 52
  - False Positives: 9
  - True Negatives: 138
  - False Negatives: 1

Model Interpretation

- The model effectively identifies high spenders.
- High recall ensures most high spenders are captured, although precision could be improved to reduce false positives.

Visual Highlights

- **Confusion Matrix Heatmap**: Demonstrates model performance.
- **Bar Chart**: Comparison of precision, recall, and F1-scores for high and low spenders.

Key Insights

1. **Demographics and Spending**:
   - Age is a significant predictor of spending behavior.
   - Gender has minimal influence on spending patterns.
2. **Purchase Attributes**:
   - Product categories and quantities are strong indicators of spending levels.
3. **Model Performance**:
   - The KNN model provides robust predictions but can benefit from fine-tuning to enhance precision.

Recommendations

1. **Marketing Strategies**:
   - Focus efforts on younger demographics for Beauty and Electronics products.

- Develop targeted campaigns to convert low spenders in other categories.

2. **Future Enhancements**:
    - Address class imbalance through oversampling or synthetic data generation (e.g., SMOTE).
    - Explore hyperparameter tuning for improved model precision.
    - Incorporate additional features such as seasonality or promotional effects.
3. **Visual Design**:
    - Continue using inclusive and accessible design principles for visualizations.
    - Simplify data representations to cater to diverse audiences.

Conclusion

The analysis demonstrates that customer demographics and purchase attributes are strong predictors of spending behavior. The insights and predictive model can help retail managers make informed decisions to enhance customer engagement and maximize revenue.

For more details, the full dataset and visualizations can be accessed via the GitHub repository.