Part A (Question Formation and Exploratory Analysis)

Overview

Understanding how consumers spend money is particularly important for businesses, as it helps them optimize marketing strategies, retain customers, and increase profits. With the development of the technological era, everyone is now accustomed to making decisions based on data. Analyzing a large number of transaction records can give us a clearer view of people's purchasing habits, seasonal changes, and the characteristics of different customer groups. This project intends to use a public dataset to study consumer spending patterns and identify predictive potential based on key demographic and behavioral characteristics.

Problem Statement

Which demographic characteristics (e.g., age, gender, income) and purchase categories are significantly associated with high-spending behavior in retail settings? In addition, can these consumption characteristic patterns effectively identify potential high-value customer groups?

There are broader social issues that can be discussed:

Are there universal high-value customer characteristics in cross-border transaction data?

How do cultural differences in product categories affect value judgment standards?

Can a customer value prediction model applicable to multiple countries be established?

Objective: To identify the key factors that influence customer spending, explore whether future spending can be predicted based on historical data, and develop a repeatable data pipeline for analysis and modeling.

Data Sources & Suitability

UCI Online Retail. With 541,909 cross-border transaction records, the data is large and complex. Its diversity is reflected in the text (product description), time sequence (invoice date), and cross-border (37 countries). The social-related questions answered by the data are reflected in its ability to help small and medium-sized enterprises in cross-border marketing.

Source: UCI

Link: https://archive.ics.uci.edu/dataset/352/online+retail

File: CSV

Dataset Characteristics

Multivariate, Sequential, Time-Series

Instances

541909

Feature Type

Integer, Real

Dataset Information

This is a transactional data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

Has Missing Values?

No

Variable Number

8

Variable Name

InvoiceNo、StockCode、Description、Quantity、、InvoiceDate、UnitPrice、CustomerID、Country

Processing Data

Data cleaning is performed, including the following aspects.

First, missing values are processed: Check the missing value of Description and CustomerID fields. Then outlier processing is performed: delete records with negative Quantity (representing return orders) and exclude records with UnitPrice $\leqslant$ 0. Finally, country standardization is performed: Country fields are uniformly converted to ISO country codes (such as "United Kingdom" $\rightarrow$ "GB")

Determine feature engineering.

Key field mapping: operationalize "high-value customers" as annual purchase frequency > 12 times, and average customer unit price > national median

Pseudocode example

high_value_customer = (annual_spending > national average spending level) & (purchase_frequency > once a month)

Then, culturally related features.

Identify culturally sensitive products through Description keywords (such as "Christmas gift" $\rightarrow$ holiday gifts)

Calculate the cultural consumption index in combination with OECD data:

Cultural consumption index = actual consumption amount / per capita consumption expenditure of the country

Finally, data fusion.

Link Google Trends category popularity data to transaction data on a monthly basis.

Data deficiencies and solutions

Identify possible flaws and pitfalls in the data, mainly the following three points: UK data dominates (87% of transactions), which has limited generalizability to other countries. No demographic information (such as age, gender, etc.): limited customer analysis, limited timeliness (2011 data), which may be somewhat different from now (2025), for example, expensive goods at that time are no longer difficult to obtain.

Solution:

For UK data dominance, oversample non-UK data and clearly mark cultural limitations in the conclusions. .

For demographic information, since there is no user's personal information, we focus on analyzing "behavior". For example, we can analyze: which customers often repurchase, which customers have a high total amount, or which customers buy frequently.

For the limited time span, since the data only includes all transactions that occurred between December 1, 2010 and December 9, 2011, it is still possible to extract feasible strategic recommendations by analyzing quarterly trends and customer life cycles.

Refined Question

"In cross-border e-commerce, can high-value customers be reliably identified based solely on behavioral patterns (purchase frequency, basket composition, spending concentration)? In the absence of demographic data, do these behavioral patterns retain predictive power across cultures?"

Refined Question eliminates reliance on unavailable demographic information, such as age and gender, and emphasizes observable behaviors, transaction frequency, and product affinity across countries. Maintaining research originality while addressing data limitations.

Perform preliminary analysis and visualization of the data. (week6)

Describe the model used to predict the A/B portion of the dataset. (week9)

Give specific suggestions on how to improve the studied system and suggest future work, if any. (week12)

Alternate Question and Dataset

Alternative question

"How to optimize the product mix strategy of cross-border retail through product co-occurrence purchase pattern (Market Basket Analysis)?"

Alternative data source

UCI Online Retail II

https://archive.ics.uci.edu/dataset/502/online+retail+ii

This data source contains 1.06 million transaction records from 2009 to 2011 (an extended version of the main data set) and has the same field structure (InvoiceNo, StockCode, Description, etc.), which can be directly merged with the main data set.

Represents an Original Question

Demographic gap: Customer value assessment using only behavioral data and cultural proxy indicators.

Timeliness paradox: Although using 2011 data, it is innovative through method: weakening outdated patterns and strengthening persistent behavioral characteristics.

REFERENCES

1. **Lau, R.Y.K., Liao, S.Y., Wong, K.F., & Chiu, D.K.W.** (2012) 'Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions', *Expert Systems with Applications*, 39(4), pp. 4691–4700. Available at: https://www.sciencedirect.com/science/article/pii/S0957417412002618 (Accessed: 11 June 2025).
2. **Sun, Y., Zhang, D., Du, S., Cao, Y. and Tian, Q.** (2021) 'Analyzing e-commerce customer purchase behavior using deep learning and association rules', *Data Mining and Knowledge Discovery*, 35(6), pp. 2443–2468. Available at: https://link.springer.com/article/10.1007/s10618-021-00773-5 (Accessed: 11 June 2025).
3. **UCI Machine Learning Repository** (2022) *Online Retail II Data Set*. University of California, Irvine. Available at: https://archive.ics.uci.edu/dataset/502/online+retail+ii (Accessed: 11 June 2025).
4. **UCI Machine Learning Repository** (2019) *Online Retail Data Set*. University of California, Irvine. Available at: https://archive.ics.uci.edu/dataset/352/online+retail (Accessed: 11 June 2025).