Gift Bundle Discovery

Kristjan Caius Kasuk

Task 2: Business Understanding

Identifying your business goals

- **Background:** This project analyzes two years of sales data from a UK-based online retailer specializing in unique giftware. The core concept is to use data mining to uncover hidden patterns in customer purchasing behavior.
- **Business Goals:** My goal is to find groups of products that people often buy together. I'll use these findings to suggest gift bundles for my final project poster.
- **Business Success Criteria:** The project is successful if my poster clearly shows a list of product pairs that make sense as gift bundles and have strong numbers to back them up.

Assessing your situation

- **Inventory of resources:**
  - **Data:** The "Online Retail II" dataset (~95 MB).
  - **Tools:** Python (Pandas, NumPy, Scikit-learn), MLxtend, Matplotlib, Seaborn, Plotly, NetworkX.
  - **Personnel:** Kristjan Caius Kasuk
- **Requirements, assumptions, and constraints:**
  - **Requirements:** A final presentation poster and a comprehensive GitHub repository with all code.
  - **Assumptions:** Historical sales data contains meaningful patterns that can be extracted through market basket analysis.
  - **Constraints:** This is a solo project in about 60 hours total.
- **Risks and contingencies:**
  - **Risk:** Data quality is too poor for meaningful analysis.
  - **Contingency:** Spend significant time (20+ hours) to rigorous data cleaning and validation.
  - **Risk:** Market basket analysis results are boring or obvious.

    - **Contingency:** Using a clustering algorithm (like K-Means) to group customers based on their purchasing history and then performing separate market basket analyses on each cluster to find niche, targeted bundles. **This is a work-in-progress idea.**

- **Terminology:** Market Basket Analysis, Association Rules, Support, Confidence, Lift.
- **Costs and benefits:** Costs are my time. Benefit is completing course requirements and learning data analysis skills.

Defining your data-mining goals

- **Data-Mining Goals:**
    1. Clean the data, find product associations using market basket analysis, and identify the strongest patterns.
    2. Implement market basket analysis (Apriori algorithm) to discover product associations.
    3. **(Work-in-Progress):** Explore the use of customer segmentation via clustering as a method to refine bundle recommendations and uncover more specific patterns.
- **Data-Mining Success Criteria:**
    - Successfully process the dataset through a documented cleaning pipeline.
    - Generate a final set of high-quality association rules with strong Lift and Confidence.
    - Produce a set of at least 10-15 bundle recommendations for the poster.

Task 3: Data Understanding

Gathering data

- **Outline data requirements:**
    - The core requirement is transactional data that shows which items were purchased together.
    - The essential fields are InvoiceNo (to group items into a single basket), StockCode or Description (to identify the product), and Quantity.
    - Additional fields like CustomerID and Country are valuable for potential deeper analysis or filtering.
- **Verify data availability:**
    - All required data is confirmed to be present in the chosen "Online Retail II" dataset from Kaggle.
    - The dataset contains over 500,000 rows of transaction data, providing a substantial base for analysis.
- **Define selection criteria:**
    - The initial dataset will be used in its entirety to ensure no patterns are missed.
    - The primary filtering will be during data preparation, where cancelled orders (identified by InvoiceNo starting with 'C') and transactions with negative quantities will be removed.
    - Further use may be focusing on specific markets, like the UK, if the data from other countries is too sparse or introduces noise.

Describing data

The dataset is a structured table with 8 columns:

- InvoiceNo **(Nominal):** A unique 6-digit number assigned to each transaction. This is the key field that defines a "market basket" for the analysis.
- StockCode **(Nominal):** A unique 5-digit code assigned to each distinct product in the inventory.
- Description **(Nominal):** The name and description of the product. This field is crucial for interpreting the results and understanding what the products actually are.
- Quantity **(Numerical):** The number of units of each product purchased in a single transaction.
- InvoiceDate **(DateTime):** The day and time when the transaction was generated. Useful for potential time-based analysis.

- UnitPrice **(Numerical):** The price of a single unit of the product in British Pounds (£).
- CustomerID **(Nominal):** A unique 5-digit number identifying a specific customer. This enables analysis of repeat purchase behavior.
- Country **(Nominal):** The name of the country where the customer resides. The majority of sales are from the United Kingdom.

Exploring data

The initial data exploration will involve several steps to understand the dataset's characteristics:

- I will first check the basic structure, including the number of rows and columns, and the data types of each field.
- I will then look for missing values, with a specific focus on the CustomerID and Description columns, as these are critical for analysis and interpretation.
- Summary statistics for numerical fields like Quantity and UnitPrice will be generated to identify potential outliers, such as negative values or prices that are zero.
- I will identify the top 20 most frequently sold products to get a sense of the retailer's popular inventory.
- The distribution of transaction sizes (number of unique items per invoice) will be analyzed to understand typical customer basket sizes.

Verifying data quality

A systematic check will be performed to assess the quality of the data:

- **Completeness:** I will quantify the amount of missing data, particularly for CustomerID and Description. Rows with a missing Description will be removed, as the product cannot be identified.
- **Accuracy:** The Description field will be checked for inconsistencies like typos, extra spaces, or different capitalizations. These can fragment what should be a single product into multiple entries, skewing the results.
- **Validity:** The data will be scanned for invalid entries. This includes negative Quantity values (which typically indicate cancellations or returns) and a UnitPrice of zero, which may indicate a promotional item or error.
- **Conclusion:** While the dataset has known quality issues, they are well-documented. The issues appear manageable through a careful and thorough data cleaning process, and the volume of valid data should be sufficient for a robust analysis.

Task 4: Planning your Project

Project Plan & Timeline (Total: 62 Hours)

(I made both my actual version and asked AI to make one by sending it my current file and this is a mixture of both)

#	Phase	Task Description	Estimated Hours
1	**Business & Data Understanding**	In-depth background research, defining goals, and EDA as outlined in Task 3.	10
2	**Data Preparation**	**The core of the project.** Advanced cleaning, handling missing data, outlier treatment, additional fields (total price), and data transformation for MBA.	22
3	**Modeling**	Implementing and tuning both Apriori and FP-Growth algorithms. Experimenting with different minimum support and confidence thresholds. Generating a large set of candidate rules.	12
4	**Evaluation**	Analysis of the rules. Applying multi-metric filtering, manual validation for business logic, and segmenting bundles by category/country. Creating a "Gift Bundle Strategy Document."	10
5	**Deployment & Reporting**	Creating high-quality visualizations (network graphs, etc.), finalizing the presentation, and preparing the final GitHub repository with full documentation.	8
		Total	**62**

Methods and Tools

- **Methods:** CRISP-DM, Exploratory Data Analysis (EDA), Data Wrangling, Market Basket Analysis (Apriori & FP-Growth algorithms), Statistical Analysis (Support, Confidence, Lift, Conviction), Data Visualization.
- **Tools:** Python, Pandas, NumPy, MLxtend, Scikit-learn, Matplotlib, Seaborn, Plotly, NetworkX, Jupyter Notebook, Git/GitHub.