Analysis of Online Shopping Behavior

Despoina Iapona & Gerard Palomo

2024-12-14

Contents

1. Introduction

The dataset provides insights into user behavior on an online shopping site. It features various types of data, including continuous, binary, and categorical variables. These variables encompass the count of administrative actions (`Administrative`), interactions related to products (`ProductRelated`), and binary markers indicating whether a visit took place on a weekend (`Weekend`) or led to a purchase (`Revenue`). Furthermore, it includes categorical information such as the user's browser (`Browser`), location (`Region`), and type of visitor (`VisitorType`).

Goals of the Analysis

The primary objective of this analysis is to assess the dataset with different distance metrics and determine the most suitable metric for each type of data. By examining the relationships among observations, this study seeks to improve the analysis of user behaviors and group patterns efficiently.

Importance of Distance Metrics

Distance metrics are essential in data science, particularly for clustering and classification. They assess how similar or different data points are and significantly affect the quality of outcomes. Choosing the right metric is crucial for achieving precise and valuable analyses, especially with mixed-type data.

2. Definition of Distance Metrics

What Are Distance Metrics?

Distance metrics are mathematical tools that measure how similar or different two data points are within a dataset. They play an important role in clustering, classification, and reducing dimensions, as they define the "closeness" or "separation" of two observations in a feature space.

Why Are Distance Metrics Important?

Distance metrics are essential to numerous machine learning algorithms, as they influence how data points are categorized or clustered. Selecting the appropriate metric is crucial for achieving accurate and significant outcomes, especially when dealing with different types of data.

3. Distance Metrics for Different Data Types

Continuous Data

For continuous variables, the following metrics are commonly used:

- **Euclidean Distance**:
  - the simplest measure, determining the direct distance between two points in a multi-dimensional environment.

- Formula:

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

- **Use Case:** Works well for normalized continuous data.
- **Manhattan Distance**:
  - Calculates the total of the absolute differences between matching features.
  - Formula:

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Use Case:** Appropriate for data with lower dimensions or when linear variations are more significant.
- **Canberra Distance**:
- Places greater emphasis on minor differences, particularly for values that are low in sizes.
- Formula:

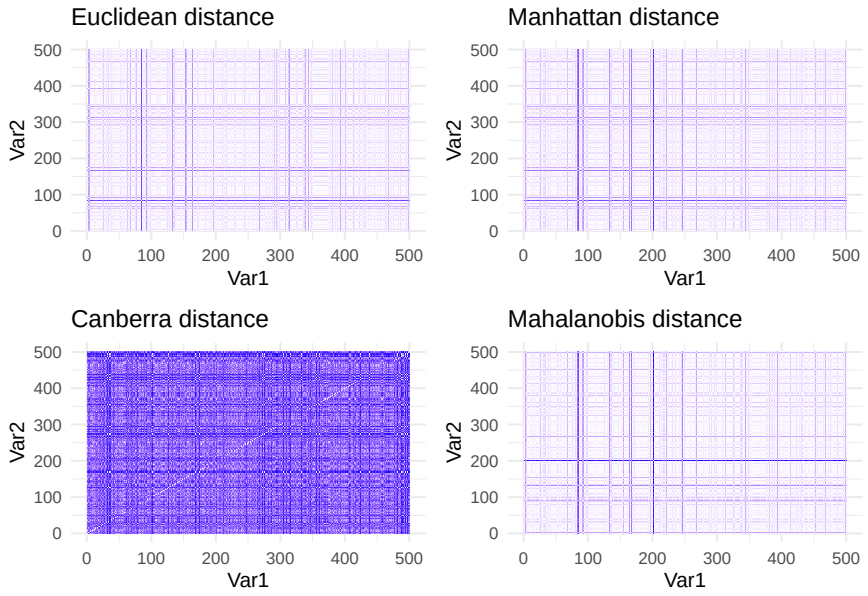
$$d(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

- **Use Case:** Ideal for datasets with variables of small sizes.

- **Mahalanobis Distance**:
  - Considers the relationships between variables and adjusts distances based on variance.
  - Formula:

$$d(x, y) = \sqrt{(x - y)^T S^{-1}(x - y)}$$

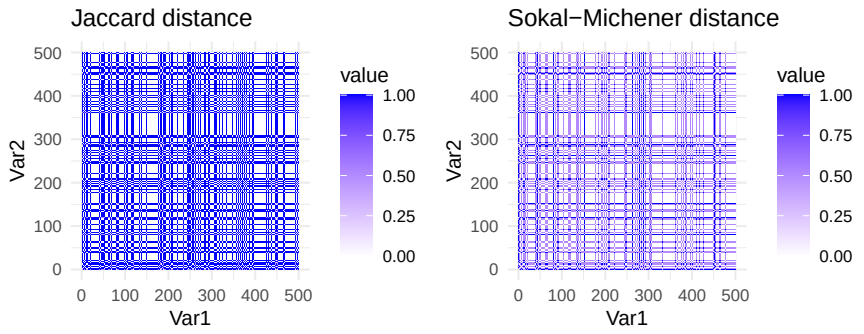
Where S is the covariance matrix.
  - **Use Case:** Ideal for related continuous variables that have different scales.



Binary Data

For binary variables, the following measurements are commonly utilized:

- **Jaccard Distance**:
  - Focuses on mismatches while ignoring shared absences.
  - Formula:
$$d(x, y) = 1 - \frac{a}{a + b + c}$$
  - **Use Case:** Appropriate for sparse binary datasets.
- **Sokal-Michener Distance**:
  - Takes into account both matches and mismatches.
  - Formula:
$$d(x, y) = 1 - \frac{a + d}{p}$$
  - **Use Case:** Ideal for datasets where matches are as important as mismatches.

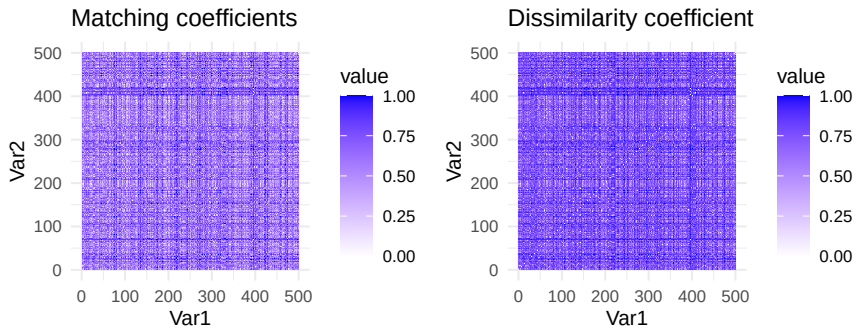


Categorical Data

For categorical variables, the following metrics are effective:

- **Matching Coefficients**:
  - Measure the similarity between categorical variables by counting exact matches.
  - Formula:
$$d(x, y) = \frac{\text{Number of Matches}}{\text{Total Observations}}$$
  - **Use Case:** Suitable for datasets where similarity in categories matters.

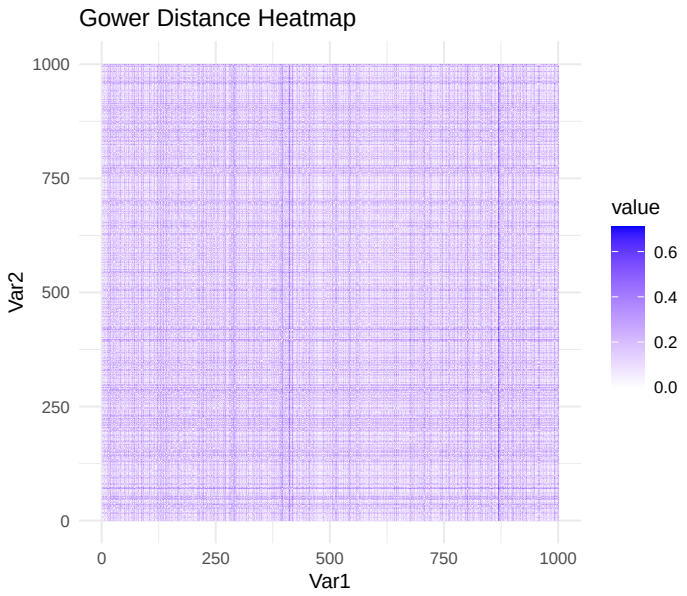
- **Dissimilarity Coefficients**:
  - Focus on differences between categories.
  - Formula:
$$d(x, y) = \frac{\text{Number of Mismatches}}{\text{Total Observations}}$$
  - **Use Case:** Useful when the focus is on categorical diversity.



Mixed Data

For datasets with mixed variable types (continuous, binary, and categorical), a flexible metric is required:

- **Gower Distance**:
  - Combines different metrics depending on the data type of each feature.
  - Formula:
$$d(x, y) = 1 - \frac{\sum_{h=1}^{p} s_h(x_h, y_h)}{p}$$
    Where s_h(x_h, y_h) is a similarity score for each variable h.
  - **Use Case:** Ideal for datasets with mixed data types.



4. Comparison of Distance Metrics

Final Remarks

Analyzing different distance metrics gave important information about how the observations in the dataset are related. Each metric has its own strengths suited for certain types of data.

- **Continuous Data**: Looking at different distance metrics like Euclidean, Manhattan, Canberra, and Mahalanobis provided various views, showing how scale and correlations matter.
- **Binary Data**: Jaccard and Sokal-Michener metrics highlighted the importance of presence/absence and balanced matches, respectively.
- **Categorical Data**: Matching and dissimilarity coefficients showed different but helpful views on how categories are similar and different.
- **Mixed Data**: The Gower Distance proved to be the best metric, integrating input from continuous, binary, and categorical variables into one clear measure.

Although individual metrics gave more insight into specific data types, the Gower Distance was chosen for the final Multidimensional Scaling (MDS) analysis. This decision guarantees that all types of variables are accurately represented, reflecting the mixed nature of the dataset and providing a complete view of the underlying patterns.

5. Multidimensional Scaling (MDS)

Use of Distance Matrices in MDS

Multidimensional Scaling (MDS) is a technique that reduces dimensions by using distance matrices to show high-dimensional data in a simpler form. Its aim is to keep the distances between pairs of observations as close as possible in the new space.

The steps are:

- 1. Create a distance matrix (D) for all observation pairs using a selected distance method (Gower).
- 2. Adjust the points in the lower-dimensional space so that the distances between them reflect the original distance matrix.

MDS is particularly useful for visualizing patterns, clusters, and relationships in high-dimensional data.

Multidimensional Scaling (MDS) Using Gower's Distance

Objective

Reduce the dimensions of a mixed-type dataset with MDS and Gower's distance.

Results and Insights

1. Cumulative Variance Explained

- The cumulative variance plot (Figure 1) indicates that the first two dimensions account for 18.54% of the overall variability.
- This suggests that more dimensions are necessary to fully understand the data's structure.

2. MDS Configuration

- The scatterplot (Figure 2) shows the first two dimensions (PC1 and PC2).
- Clusters and groups suggest that some observations share similar traits.
- The MDS setup reveals patterns in the data that match its mixed characteristics.

Figures

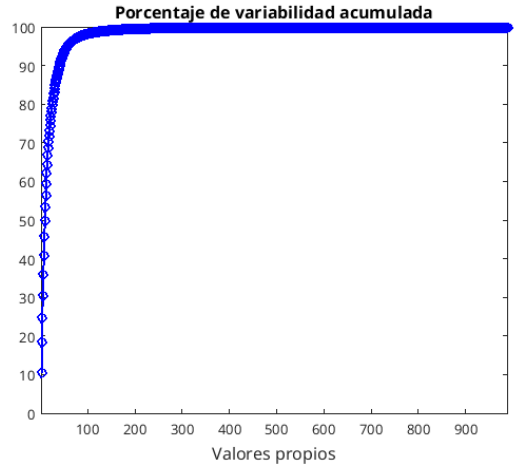


Figure 1: Cumulative Variance Explained

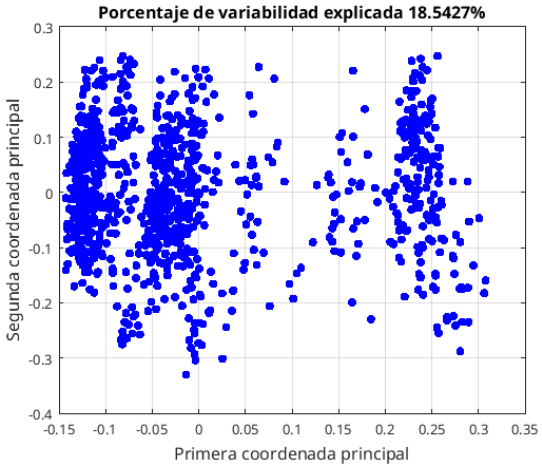


Figure 2: MDS Configuration