|  |
| --- |
| HEPSİBURADA |
| Related Products Recommendation – Dataset Preparation Report |
| Data Scientist Job Assignment |

|  |
| --- |
| Ali GÜVEN  9-21-2022 |

Contents

[1. Abstract 2](#_Toc114654042)

[2. Introduction 2](#_Toc114654043)

[3. Dataset Distribution 3](#_Toc114654044)

[4. Methods 3](#_Toc114654045)

[4.1. Drop NA Values 4](#_Toc114654046)

[4.2. Finding Unique Carts, Product Categories and Subcategories 4](#_Toc114654047)

[4.3. Finding Total Number of Products 4](#_Toc114654048)

[4.4. Finding Total Price of the Products in Cart 4](#_Toc114654049)

[4.5. Time Zone Classification 4](#_Toc114654050)

[4.6. Category and Sub-Category Classification 4](#_Toc114654051)

[5. Improvements 5](#_Toc114654052)

Table of Figures

[Figure 1: Flow cart of the process of the dataset 2](#_Toc114652611)

[Figure 2: Dataset distribution 3](#_Toc114652612)

# Abstract

Dataset was created with events and meta files. Events file includes events of the users that added products to their carts. Each cart separated by “sessionid” unique. After separated the carts, feature extraction process was started. There are 156 features that extracted from the files according to total products, sum of total price, time zone of the buying (morning, afternoon, night, and evening), categories of the product bought by user (extracted like one hot encoding), sub categories of the product bought by user (extracted like one hot encoding) in a cart.

# Introduction

The custom features extracted are total products and sum of total price in a cart. They are the most observed features for product recommendation models. Therefore, the most important features are the rest of the feature set. Time zone is important because, needs of a person can be vary during the day. He or she would be thirsty, hungry, or just need a battery. For example, we mostly realize that we need bulb just before night.

Classifying categories and subcategories according to their unique is helping the model to understand what the user order mostly and recommendation model decides what products the user can be add to the cart. Process of obtaining dataset can be demonstrated by Figure 1. Total number of samples after extracted features and prepared the dataset is 54442.

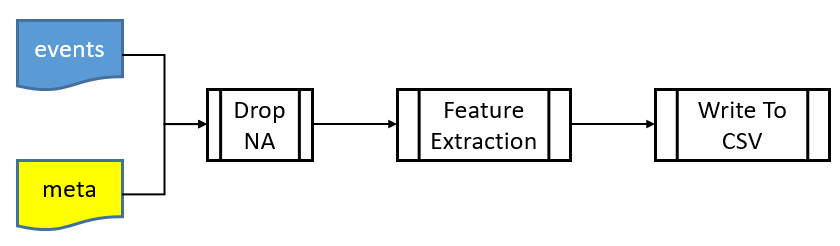


Figure 1: Flow cart of the process of the dataset

# Dataset Distribution

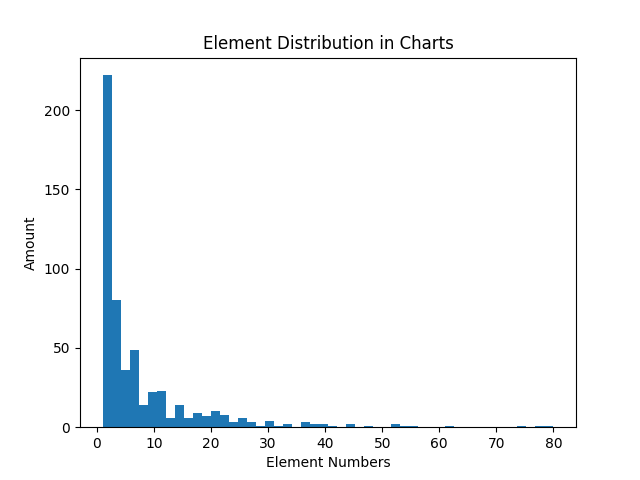


Figure 2: Dataset distribution

As we can see in Figure 2 that users mostly chose to buy products in small number. Hence, the distribution seems like exponential distribution rather than Gaussian. It shows that the probability of the next users would buy products in small numbers is higher.

# Methods

For preparing dataset and the feature extraction process, 7 methods were developed and used. These are (i) Drop NA values in the events and meta files, (ii) finding unique carts, product categories and its subcategories, (iii) finding total number of the products in a cart, (iv) finding sum of prices of the products in a cart, (v) time zone classification, (vi) category classification, and (vii) sub category classification.

# Drop NA Values

There are some NA values in events and meta files. The amount of these values was small then instead of changing or filling the values, dropping was the best choice.

# Finding Unique Carts, Product Categories and Subcategories

I decided to develop my model by taking each carts as a sample. So that if a new cart would come and related products would be wanted to find, just cluster it according to previous data and decide the related products with the nearby clusters. Advantage of this style is no matter how many products would be bought by users; the model is independent from the number of products.

Product categories and sub-categories have unique values. These values were used for extracting features.

# Finding Total Number of Products

It is a conventional feature that is used in most of the related works. Finding is the number of the total products in a cart.

# Finding Total Price of the Products in Cart

It is a conventional feature that is used in most of the related works. Finding is the total price of the products in a cart.

# Time Zone Classification

It is an important feature because of the users’ decisions can be changed by the time zone of a day. A day is divided into four parts which are morning ( 6 – 12), afternoon (12-18), evening (18 – 24), and night (24 – 6). Features extracted as four features each represents the four parts. For example, if a user buy in morning the sample includes the features as [1, 0, 0, 0].

# Category and Sub-Category Classification

It is an important feature because of the related products should be choose according to the categories and sub-categories in the cart. The features were extracted as unique of the categories and sub-categories, and represented in the dataset as integer. For example, there are 2 categories and 3 sub categories in total, and the user buy first category and its sub category, then the features would be [1,0,1,0,0]. Some products do not have category or sub-category, if so I put them in “Other” class.

# Improvements

Dataset includes mostly the carts with small number of products. This can be problem for the users who buy large number of products. Dataset should be balanced.

Dataset includes NaN values; it should be filled if necessary.

Some products do not have category or subcategory, for healthy clustering they should be put in a category or sub-category.

Dataset was created in summer month, June. For gathering more information, whole days in a year must be considered. So that, we can use seasons, months as features.