

---

# DESIGN DOCUMENT

Prepared by : Dilay Sevgili

Submitted to : Hepsiburada  
Recommendation Team

June 2, 2022

# Contents

|          |                                    |          |
|----------|------------------------------------|----------|
| <b>1</b> | <b>Introduction</b>                | <b>3</b> |
| 1.1      | Purpose . . . . .                  | 3        |
| 1.2      | Project Scope . . . . .            | 3        |
| <b>2</b> | <b>Data Exploration</b>            | <b>4</b> |
| 2.1      | Data Types . . . . .               | 4        |
| 2.2      | Missing Values . . . . .           | 4        |
| <b>3</b> | <b>Recommendation Applications</b> | <b>5</b> |
| 3.1      | Recommendation Model . . . . .     | 5        |
| 3.2      | API Application . . . . .          | 5        |
| <b>4</b> | <b>Further Works</b>               | <b>6</b> |

# **1 Introduction**

## **1.1 Purpose**

A product recommendation is a system that aims to predict and show the products that a user would like to purchase. In this study, "Related Products" recommendation for the cart page of an online grocery store is studied. When a customer is landed on the cart page, top 10 recommended products with highest scores are shown. This document covers details of recommendation system.

## **1.2 Project Scope**

Recommendation model with an API which retrieves products in the cart as request parameters and returns recommended products (top ten) with scores and its application are shared through this document.

## 2 Data Exploration

### 2.1 Data Types

There are two types of data for this assignment. One of them is 'event' data which includes cart events of users and the other one is 'meta' data that provides information about products. All features in two data set have text features that are suitable for NLP techniques. Since 'productid' is the common variable in data sets it will be used to get information about the product that each user purchase.

### 2.2 Missing Values

Before starting any analyses, we need to be sure about missingness in data sets. Events data set has 6 rows which includes missing price and productid information. On the other hand, meta data has 1 rows that contains all missing features and 'brand' category has 459 missing attributes. Since 'brand' is also important feature missing brands are deleted before application. However, alternative ways for handling with missing values are discussed in 'Alternative Approaches' section.

## 3 Recommendation Applications

### 3.1 Recommendation Model

Main recommendation application for this assignment is Content Based Recommendation. This approach creates the properties and the metadata of a particular item to suggest other items with similar characteristics with cosine similarity. To be able to calculate similarities one vector was created by 'brand', 'category', 'subcategory' and 'name' and these were combined in 'phrase' field. Other vector that we calculate similarities is 'name' of product. After calculating similar contents for each products in session id, recommendation models that I used return best 10 products that related with sessions. To be able to make more clear, step by step algorithms is following:

- **Step 1:** First, we create phrase column with `PhraseData(data)` function by using 'brand', 'category', 'subcategory' and 'name'.
- **Step 2:** `TransformedData(data_phrase, data)` function creates vector transformed text columns by using 'phrase' and 'name'.
- **Step 3:** `FindProductsWithRelatedSessionId(sessionId, dataframe)` returns all the products in the same sessionid.
- **Step 4:** Finally, `RecommendedProducts(cardlist, data, combine, transform)` returns top 10 products for sessionid. Although each product have top 10 similar product list, this function gives a list for whole session with top 10. For example, if a session has 4 different products with top 10 products, this function creates a list for top 10 product with highest scores within whole 40 products. This algorithm creates top 10 products according to each session items in the cart.

### 3.2 API Application

Since it is asked to serve the model with an API which retrieves products in the cart as request parameters and returns recommended products (top ten) with scores. To be able to run and test API Application 'Postman' can be used as it stated in Readme.md file.

## 4 Further Works

Since only Content Based Recommendation with cosine similarity was used in this assignment, alternatives approaches may be used for better results. Although this model works well for some sessionids, we had also some irrelevant results for other sessions. Since there is time limitation for this assignment, alternative approaches is not used but they are listed in following:

- Not removing missingness in brand, alternatively not using brand in 'phrase' since brands mostly are stated in name columns.
- I used Content Based Recommendation that finds similar products but it could be use User-User Recommendation which finds similar users and recommend based on those similar users products.