Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis

Econometrics and Operations Research

in the field of: Business Analytics and Quantitative Marketing

Market Basket Analysis Based On Historical Sales Data

Author: Anne Jasmijn Langerak

Student ID number: 475769

Supervisor: Fu, J

Second assessor: Castelein, A.

Date final version: July 5, 2020

Abstract

Predicting next market baskets has many benefits for companies. Therefore, this thesis investigates various models that predict a customer's next market basket based on historical sales data. My research consists of two parts. First, I analyse the predictions a model that utilizes Wasserstein-based sequence matching. Secondly, I introduce and analyse the predictive power of two additional models that are built from characteristics of the former model and state-of-the-art baseline models (e.g. repurchase last basket). Based on this research it is concluded that models that make use of customer-specific information, (e.g. personal top items), have better predictive power than models that do not use this type of information. Furthermore, it is concluded that the model that has the best predictive power for the analysed data is the personal top items approach.

ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

The views stated in this thesis are those of the author and not necessarily those of Erasmus School of Economics or Erasmus University Rotterdam.

Contents

1 Introduction

Predicting future demand is of great importance for both online and offline stores. It helps to maintain a proper inventory level, which on its own contributes to customer satisfaction and increasing revenue. Gaining a better understanding of customer behaviour is therefore valuable for retailers (Cumby et al. (2004)). A common and extensively researched angle for demand prediction is forecasting future sales volumes using time series analysis over (aggregated) customer sales data (Dekimpe and Hanssens (2000)) . Another angle is market basket analysis (MBA). Although considerably less studied in the current state of the literature, it has numerous potential benefits for retailers. Besides improving their supply chain management, it is also beneficial from a marketing point of view. It allows for personalized advertising, (i.e. recommendations of specific items), which can positively contribute to increasing sales and the customer's shopping experience. The goal of MBA is to predict what the next purchase of a customer will be. In other words, the items that their next market basket will consist of. To this end finding relationships across purchases and identifying connections between products could be beneficial. The latter can identify products that are often purchased together. If the retailer would drop one of these products, it could result in the loss of sales of the co-purchased product(s) (Chen et al. (2005)). This leads to the following research question:

How can we predict the exact items a customer's next market basket will consist of, given historical sales data?

Related studies to market basket analysis make use of pre-coded rules, which are unable to capture cross-customer knowledge in a method. Therefore the underlying information stored in cross-customer knowledge is never used. Research by Kraus and Feuerriegel (2019), claims to overcome this issue by measuring the similarity between purchase histories to find the most similar customers. Based on these nearest neighbours, a next market basket prediction for the customer is formed. They compare their approach to other models that are used in similar research and conclude that their approach outperforms these models in terms of predictive power. In the first part of this thesis the research and thus conclusion are audited. This is done by the use of a dataset from an online supermarket, the same dataset used in the original research. This dataset consists of the shopping behaviour of more than 200,000 customers gathered over a period of 1 year. Based on my obtained results, it can be concluded that Kraus and Feuerriegel (2019) do not outperform all baseline models for the test set that I studied. It is observed that models that incorporate a customer's own purchase history information (i.e. customer-specific information) for their prediction have better predictive power than models that only use the information of all customers to predict a customer's next basket.

The second part of this thesis investigates and leverages this observation in order to improve Kraus and Feuerriegel (2019). I modify one of the latter models that has relatively weak predictive power in such a way that a customer's specific information is used for their prediction. It can be concluded that this modification results in significantly better performance metrics compared to the original model. Building on these results, I modify the Kraus and Feuerriegel (2019) approach and analyse its prediction performance. My approach has a significantly better running time, while performing almost identical to Kraus and Feuerriegel (2019). It is also concluded that the (baseline) model that uses a customer's top items outperforms the Kraus and Feuerriegel (2019) approach, all other baseline models and my own approach. Our contribution to studies into market basket analysis does not only cover the conclusions as to the relevant type of information to include into a prediction approach, but also my extension of Kraus and Feuerriegel (2019) that is more efficient without decaying the predictive power.

In the continuation of this thesis, previously performed research on this topic will be discussed in Section 2. In Section 3, I discuss the data and explain the restrictions for the dataset. Section 4 consists of two parts, where the first part explain the methods that are used for the replication part of this thesis and the second part explains the methods used for two additional models. The results of these methods can be found in Section 5, again split into two parts. Finally, the conclusion that can be drawn from the results are summarized in Section 6, where I also provide a discussion of the limitations of my research and provide ideas and improvements for future research.

2 Literature

Market Basket Analysis requires prediction methods where the outcome can be a dynamic set. Current approaches can be explained in four factors (R. Guidotti et al. (2019)): general, sequential, pattern-based, and hybrid. A general approach in terms of market basket analysis is that of predicting a customer's next market basket by selecting the top sold items across all customers in the given database. I refer to this model as *Global top items* and use it as one the 4 baseline models for comparison. Sequential models use sequential information. For example, studying which items are purchased after each other. Research that studies sequentuality in market baskets is done by Wang et al. (2015). They opt to find an approach that captures the underlying, hierarchical pattern that sequentuality may have with a customer's personal taste and thus influences the content of a customer's next shopping list. Therefore, they introduce the Hierarchical Representation Model (HRM) that uses recurrent neural networks. Which makes this approach hybrid as it uses sequentiality and patterns. Yu et al. (2016) study similar patterns but use a novel model named Dynamic Recurrent Basket. It can capture changes in customers taste over time. Riccardo Guidotti et al. (2017) also provide a hybrid model that uses sequential-information and patterns, they predict the next market baskets using Recurring Sequential Patterns. Their method extracts a Temporal Annotated Recurring Sequence (TARS) pattern that is able to capture decision-influencing factors simultaneously: co-occurrency, sequentuality, periodicity and recurrency of the purchased items. Based on this pattern they develop a special predictor for prediction of the next market basket.

Sequential information is the core behind another baseline models utilized in this thesis, *Association-rules*. Which is a model that tries to find a relation in subsequent market baskets across all customers. The formulation of such rules is often based on the Apriori algorithm (Agrawal, Srikant, et al. (1994)). Apriori Algorithms are used in data mining and are often employed when the goal is to extract patterns of large datasets, such as studying data of a supermarket. A disadvantage of this baseline model is that it is not able to capture the hierarchical structure of products (Kraus and Feuerriegel (2019); Brin, Motwani, and Silverstein (1997)). Again this can be perceived as a hybrid model. A disadvantage of all these models is that they need data on a whole customer base for the prediction a customer. The same goes for Kraus and Feuerriegel (2019) their prediction approach for market basket analysis also uses the entire dataset. Their prediction approach is based on finding customer's that have similar purchasing habits. To find similar customer's subsequential dynamic time warping is utilized. Therefore products are expressed in numerical vectors, which results in the possibility to compute the Wasserstein distance between baskets. This distance metric expresses how similar two baskets are. Which allows for similarity matching between sub-purchase histories of customers.

3 Data

3.1 Dataset

The data used in this research is the *The Instacart Online Grocery Shopping Dataset* (2017). It is an anonymized dataset containing over 3 million orders from more than 200,000 Instacart customers. This dataset is the same as used by Kraus and Feuerriegel (2019). For each customer, there are at least 3 orders and a maximum of 100 orders available, in sequence. The dataset contains 2 files with order information: *order products prior* and *order products train*. The *prior* file contains the purchase histories of all customers except for their last purchase. Information about these baskets is contained in the *train* file. These baskets are used for comparison with the predictions made by the models and thus to analyse the predictive performance of these models. The dataset also provides information about the entire supply of the online supermarket. There are in total 49,688 products, the type of product, product id, aisle id and department id are given.

3.2 Restrictions

Kraus and Feuerriegel (2019) apply this dataset on product-level (top 500 most sold items) and category-level (i.e. aisles level). This thesis focuses on the product-level, as this thesis is about predicting the *exact* content of a next market basket. However, the category-level is studied as it can reveal strengths and weaknesses of Kraus and Feuerriegel (2019). Following Kraus and Feuerriegel (2019) I only study the 500 most sold products, therefore all but the top 500 products are neglected from the data set. Kraus and Feuerriegel (2019) propose two additional restrictions, namely only studying purchase history of customers with at least 10 market baskets where each market basket contains at least 5 items. These conditions result in a sub-dataset covering 603,457 orders of 27,139 customers. All products belong to one of the 21 departments, where similar products can be found in the same department. Due to only studying the 500 most sold products, products from 6 of those departments are never studied. The relative selling frequency of the 500 products is shown in the Treemap in appendix A1. The 27,139 customers (i.e. their entire purchase histories except for their last basket) are divided into 3 sets: a training(80%) set, a validation(10%) set and test set(10%). This is called a customer-level splitting strategy in line with the splitting strategy used in Kraus and Feuerriegel (2019). This strategy divides a customer into 1 of the 3 sets (i.e. a customer cannot be in more than 1 set). Although the names of these sets are the same as the file names mentioned in section 3.1, they do not contain the same information. As the sets divide the customers to study the predictive power of the models and the files divide the orders into customers purchase histories and their last order. The training set covers potential neighbours for the customers in the test set, which is the set of customers whose next market basket is predicted by the models. From the test set, 300 random customers are selected to predict the next market basket for. I do not use the entire test set of 2,713 customers as this would increase the running time significantly. For all models and their analysis this sub-test set of 300 customers is used. The training set covering 21,711 customers is not reduced to decrease the running time as it allows for more fair comparison with the results provided in Kraus and Feuerriegel (2019). In a nutshell, for this thesis, a subset of the Instacart dataset is used as I only study the 500 most sold products. This subset is then randomly split into 3 evaluation sets containing customers' entire purchase histories except for their last baskets. The test set, which contains the purchase histories of the customers whose next market basket is predicted by the models, is sized down to a sub-test set containing 300 customers. This is done by drawing a random subset.

Category-level covers the analysis on aisle-level, meaning that each product in a market basket is not expressed by its product id but by its aisle id. All previous statements regarding to the analysis on product-level apply to the analysis on category-level. Meaning that the exact dataset used in the former analysis is used. This is differs from the selection Kraus and Feuerriegel (2019) utilize, which covers 65,710 customers and 1,634,548 orders. Using this sub-dataset would increase the running time significantly, therefore this thesis deviates from their approach.

4 Methodology

As mentioned before, the foundation for this thesis is the Wasserstein-Based Sequence Matching approach of Kraus and Feuerriegel (2019), hereafter abbreviated as K&F approach. Their approach follows four steps:

1. **Model similarity between products**: By building product embeddings
2. **Calculating distance between market baskets**: By utilizing the Wasserstein distance
3. **Calculating the distance between purchase histories**: By means of k-nearest neighbour Subsequence Dynamic Time Warping
4. **Prediction of the next market basket**: Where a threshold is taken into account to possibly revert to a fallback prediction

To measure the predictive power of the approach three performance metrics are computed. To compare this approach to other state-of-the-art approaches for market basket analysis, four baseline models are used following Kraus and Feuerriegel (2019). The first part of this methodology covers the replication of Kraus and Feuerriegel (2019). The second part builds on the models and their conclusions of part 1. I introduce a modified baseline model that only utilizes customer-specific information for the prediction and I built on Kraus and Feuerriegel (2019) by modifying their approach.

4.1 Part 1

4.1.1 Model similarity between products

As products are only represented by their product id, thus prohibiting us from doing any calculations with them, the distance (i.e. similarity) between products cannot be measured. To account for substitute goods, e.g. milk and buttermilk, and simultaneously capture the relation between items that are present in the same market basket, Kraus and Feuerriegel (2019) convert each product into a multi-dimensional vector (a product embedding). The values in the vectors are determined by means of neural embeddings, which are a part of natural language processing. This algorithm starts by randomly initializing a product embedding for each product in the dataset (Tomar (2019)). It requires sentences with words, for the utilized dataset these sentences are market baskets and the words are product id's. The algorithm then goes through each position in the sentences and determines the target word and its context words. This is depended on the window size w adapted in the algorithm. This window indicates how many words next to the target word should be studied (both sides). The product embeddings in a market basket are then optimized by maximizing the following likelihood:

$$\sum_{p \in b_c^i} \sum_{\substack{q \in b_c^i \\ q \neq p}} \log \Pr(p|q), \tag{1}$$

where p and q are product embeddings in market basket i of customer c ($p, q \in b_c^i$ where $p \neq q$ and $c \in C = \{c_1, \ldots, c_h\}$ the set of all customers from a database). The market basket b_c^i is a subset of all available products from the supermarket contained in set I , $b_c^i \subseteq I$. In other words, equation 1 maximizes the probability of predicting the context words given the target word. The probability $\Pr(p|q)$ of a context word given the center word is formulated as:

$$\Pr(p|q) = \frac{\exp\left(u_p^T v_q\right)}{\sum_{r \in I} \exp\left(u_p^T v_r\right)}, \tag{2}$$

where $u_p \in \mathbb{R}^G$ and $v_q \in \mathbb{R}^G$ are latent vectors corresponding to the target and context representation of product p. By averaging u_p and v_q, a G-dimension embedding vector of item p is obtained. To compute the distance between 2 product embeddings, cosine similarity is used. Cosine similarity is computed by the normalized dot product of two product embeddings:

$$\cos(\alpha) = \frac{p \cdot q^T}{||p|| ||q||}, \tag{3}$$

where $\cos(\alpha)$ is the cosine similarity and p and q are product embeddings, where $p \neq q$. A relatively large value indicates that similar goods.

4.1.2 Calculating distance between market baskets

To be able to compare the market baskets on a numerical level, Kraus and Feuerriegel (2019) utilize the *Wasserstein distance* (also referred to as *Earth Mover's Distance*) by expressing market baskets in terms of their product embeddings as calculated in 4.1.1. The Wasserstein distance measures the minimum distance needed to map one basket onto another. This amount, d_{W}, is calculated by the following optimization problem:

$$d_{\mathrm{W}}^{(t)}(X,Y) \overset{\mathrm{def}}{=} \min_{C} \left(\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} d\left(x_i, y_j\right)^t \right)^{\frac{1}{t}}, \tag{4}$$

where t is the order of the distance with $t \geq 1$. X and Y are market baskets containing product embeddings, $x_i \in X$ and $y_j \in Y$, $|X| = m$ and $|Y| = n$ and x_i and y_i are products represented by their embeddings find in the respective baskets. C is a transportation matrix where the elements satisfy:

$$c_{ij} \geq 0 \quad \text{for all } 1 \leq i \leq m \text{ and } 1 \leq j \leq n \tag{5}$$

$$\sum_{j=1}^{n} c_{ij} = \frac{1}{m} \qquad \text{for all } 1 \leq i \leq m \tag{6}$$

$$\sum_{i=1}^{m} c_{ij} = \frac{1}{n} \qquad \text{for all } 1 \leq i \leq n. \tag{7}$$

Lastly, d is a metric that represents the distance between two products embeddings of dimension $Gx1$. The metric employed by Kraus and Feuerriegel is the Euclidean distance, which is defined as:

$$d(x_i, y_j) = \sqrt{\sum_{g=1}^{G} (x_{i,g} - y_{i,g})^2}. \tag{8}$$

For efficiency reasons, a lower bound for the Wasserstein distance is taken into acccount. If the value of the lower bound exceeds the distance of previously calculated nearest neighbours, the computation of the exact Wasserstein distance is skipped. This lowerbound is the maximum of the outcome of:

$$LB_1 = \sum_{j=1}^{n} \min_{k=1,\ldots,m} d\left(x_k, y_j\right)^p \frac{1}{n} \tag{9}$$

$$LB_2 = \sum_{i=1}^{m} \min_{k=1,\ldots,n} d\left(x_i, y_k\right)^p \frac{1}{m}. \tag{10}$$

Or mathematically, $LB^* = \max\{LB_1, LB_2\}$. The lower-bound is thus the maximum of the mean minimum distance of all products in a basket compared to the items in the other basket.

4.1.3 Calculating the distance between purchase histories

To determine the distance between two purchase histories (i.e. customers) from the complete set of purchase histories $\mathcal{B}$, Dynamic Time Warping (DTW) is used. The complete set $\mathcal{B}$ contains the purchase histories of all h customers in a dataset, mathematically $\mathcal{B} = \{B_1, B_2, \ldots, B_h\}$. The purchase history of a customer consists of all their baskets b_c, mathematically $\left[b_c^1, b_c^2, \ldots, b_c^m\right] = B_c \in \mathcal{B}$. DTW finds the shortest path between two purchase histories. It is possible that the two histories differ in size and only form a proper match for a sub-sequence of the history (e.g. B_c and $B_d\left[i_s : i_e\right]$ for $1 \leq i_s \leq i_e \leq n$). To account for this possibility DTW makes use of Star-Padding at the beginning of B_c, which means that B_c has zero distance to all market baskets of B_d. The distance between two purchase histories $d_{\mathrm{DTW}}\left(B_c, B_d\right)$ is equal to minimum value of the matrix D, where the dimension of the D depends on the cardinality of the purchase histories. The elements of D are recursively determined by:

$$D_{ij} = d_{\mathrm{W}}^{(t)}\left(b_c^i, b_d^j\right) + \min\left\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\right\} \text{for all } i \text{ =1,...,m and } j \text{ =1,...,n}, \tag{11}$$

where $D_{i,0} = D_{0,0} = 0$ and $D_{0,j} = \infty$. The $d_{\mathrm{W}}^{(t)}\left(b_c^i, b_d^j\right)$ is the Wasserstein distance as described in section 4.2. The distance between the two customers (i.e. their purchase histories) d_{SDTW} is then computed by $\min_i D_{in}$ where n is the number of baskets in the purchase history of customer d, B_d.

4.1.4 Prediction of the next market basket

To actually predict the content of the next market basket of a customer c, the k-nearest neighbours of the customer are needed. The k-nearest neighbours are the customers that have the closest distance between their (sub-)purchase history and the purchase history of customer c (i.e. their purchase histories are most similar). The k-nearest neighbours are found by computing the d_{SDTW} of customer c with all customers and take the smallest k-values:

$$\left(d^*, j_s^*, j_e^*\right) = \underset{\substack{1 \leq j_s \leq j_e \\ d=1,\ldots,h}}{\arg\min} d_{\mathrm{SDTW}}\left(B_c, B_d\left[j_s : j_e\right]\right), \tag{12}$$

where d^* is the nearest neighbour, j_s^* is the index of the first basket of the sub-purchase history of d^* and j_e^* is the index of the last basket included in the sub-purchase history of d^*. Similar, j_s is the index of the first basket en j_e is the index of the last basket included in the sub-purchase history of a customer d.

After locating the k-nearest neighbours, the average distance $\bar{d}_{SDTW}$ is then computed by taking the mean of the d_{SDTW} of the k-neighbours:

$$\bar{d}_{SDTW} = \sum_{d=1}^{k} \frac{d_{d,SDTW}}{k}. \tag{13}$$

If $\bar{d}_{SDTW}$ exceeds a certain threshold τ, the next market basket of customer c consists of the top n_c items from their purchase history B_c. Where n_c is equal to the average basket size across customer c's purchase history B_c. This fallback prediction is used when the customer differs significantly from all other available customers in C. However, when enough similarity between customer c and the k-nearest-neighbours is observed, the prediction is equal to the top $|B_{d^*}\left[j_e^* + 1\right]|$ items in the baskets of the k-neighbours ($|B_{d^*}\left[j_e^* + 1\right]|$ is the number of items in the basket of the nearest neighbour). The fallback prediction is thus used when $\bar{d}_{SDTW}\left(B_c, B_{d^*}\right) > \tau$.

Kraus and Feuerriegel (2019) do not describe the way they chose a value for the threshold τ. I therefore study the course of the performance metrics against a range of value's for the threshold to find the optimal value for τ. This range is determined by combining the proposed threshold value's as described in Kraus and Feuerriegel (2019) and by analyzing the distribution of the average distances $\bar{d}_{SDTW}$ of the customers in the test set.

4.1.5 Performance metrics

To measure the performance of a model, I use 3 performance metrics as described by Kraus and Feuerriegel (2019).

1. *Wasserstein distance*

   This metric is equal to the average Wasserstein distance of the predicted versus realized market baskets across all customers in the test set. The Wasserstein distance is calculated as described in section 4.1.2.

2. *F1-score*

   This metric is equal to the average F1-score of the predicted market baskets versus realized market baskets across all customers in the test set. The F1-score is defined as the harmonic mean of precision and recall (Rendle, Freudenthaler, and Schmidt-Thieme (2010)):

   $$F_1 = 2 \cdot \frac{\text{precision } \cdot \text{recall}}{\text{precision } + \text{recall}}, \tag{14}$$

   where precision and recall are calculated by:

   $$\text{Precision} = \frac{|b_c^{m+1} \cap \hat{b}_c^{m+1}|}{|b_c^{m+1}|} \tag{15}$$

$$\text{Recall} = \frac{|b_c^{m+1} \cap \hat{b}_c^{m+1}|}{|\hat{b}_c^{m+1}|}. \tag{16}$$

The b_c^{m+1} represent the predicted basket and $\hat{b}_c^{m+1}$ is the actual basket.

3. *Jaccard coefficient*

   This metric defines the average ratio of co-occurrences to non-co-occurrences between the baskets. The Jaccard coefficient for each prediction-realization pair is calculated as (Musalem, Aburto, and Bosch (2018)):

   $$J = \frac{p}{p+q+r}, \tag{17}$$

   where p is the number of items in both baskets, q is the number of items present in the predicted basket but not in the realized basket and r is the number of items present in the realized basket but not in the predicted basket. Mathematically this can be seen as: $p = |b_c^{m+1} \cap \hat{b}_c^{m+1}|$, $q = |b_c^{m+1} \setminus \hat{b}_c^{m+1}|$ and $r = |\hat{b}_c^{m+1} \setminus b_c^{m+1}|$.

4.1.6 Baseline models

The K&F model is compared to four state-of-the-art baseline models using the performance metrics:

1. *Global top items*

   This model predicts the next market basket of a customer c based on the average basket size n_c of that customer. The items in the basket are determined by the top n_c items across all customers (Wang et al. (2015)).

2. *Personal top items*

   This model predicts the next market basket of customer c based on the specific taste of the customers. The basket consists of the top n_c items purchased in the customer's history (Cumby et al. (2004)). Again n_c is determined by the average basket size of that customer.

3. *Repurchase last basket*

   This model predicts the next market basket of customer c by equalizing it to the past basket of that customer (Cumby et al. (2004)). Given the purchase history of customer c, $B_c = \left[b_c^1, b_c^2, \ldots, b_c^m\right]$, the next market basket is given by:

   $$b_c^{m+1} \overset{\text{def}}{=} b_c^m \tag{18}$$

4. *Association rules*

   The goal of this model is to find a pattern in subsequent market baskets, within the history of the customer database. A matrix S is constructed, where the elements are the summation of the Cartesian products as determined by:

$$C = \cup_c C_c = \big\{(a, b) \text{ for } a, b \text{ in } b_c^i \cdot b_c^{i+1} \text{ for } 1 \leq i \leq m - 1\big\} \tag{19}$$

In other words, $S_{a,b}$ denotes the number of times product a is in market basket b_c^i and product b is in the subsequent basket b_c^{i+1}, for all customers c across all baskets in a given dataset. The shape of S is given by the amount of unique products n across the entire customer base (i.e. $S^{n \times n}$). The prediction of the next market basket b_c^{m+1} is then determined by studying the items in the last basket b_c^m and searching for the items most often purchased after the items in that basket (Kraus and Feuerriegel (2019)).

4.1.7 Category-level analysis

The before mentioned methodology is also applied on category-level, following Kraus and Feuerriegel (2019). When comprising baskets into products' categories, baskets become more similar. As the core of the K&F approach is finding similarity between purchase histories, and thus baskets, this could lead to different conclusions when comparing the conclusions drawn from the analysis on product-level.

4.2 Part 2

For the extension part of this thesis, I first modify the *association rules* model by incorporating customer-specific information. The motivation behind for this analysis is explained in section 5.3. Besides that, I also extent the K&F approach. Again, the motivation behind this extension are described in section 5.3.

4.2.1 Individual-specific association rules

The *individual-specific association rules* approach is a modification of the *Association rules* model. Again, the goal of the model is to find a pattern in subsequent market baskets. The difference, however, is that I exclusively study the customer history of the customer that I predict the next market basket in this model instead of analysis of the entire customer database. The elements of the matrix S_c are constructed by the summation of the Cartesian products as determined by:

$$C_c = \big\{(a, b) \text{ for } a, b \text{ in } b_c^i \cdot b_c^{i+1} \text{ for } 1 \leq i \leq m - 1\big\} \tag{20}$$

To emphasize the difference with the *Association rules* model, $S_{a,b}$ in the modified version denotes the number of times customer c purchased item b in the next market basket when item

a was purchased in the previous market basket. The shape of S_c is $d \times n$. Given the purchase history B_c of customer c, $B_c = \left[b_c^1, b_c^2, \ldots, b_c^m\right]$, d is the number of unique products in baskets b_c^1 up to and including b_c^{m-1} and e is the number of unique products in baskets b_c^2 up to and including b_c^m. The prediction of the next market basket b_c^{m+1} is then determined by studying the items in the last basket b_c^m and searching for the items most often purchased after the items in the last basket.

4.2.2 My approach

To predict the next market basket of a customer c, the last basket known of that customer is compared to all baskets in the dataset except for the last basket of each customer. These baskets should not be included as the market baskets after them are not known yet. The k-nearest neighbours are baskets that have the smallest distance to that of the customer's last basket b_c^m, in term of the Wasserstein distance $d_{\mathrm{W}}^{(t)}$. Mathematically, the most similar basket b^* is computed as follows:

$$b^* = \underset{\substack{1 \leq j \leq |B_d|-1 \\ d=1,\ldots,h}}{\arg\min} \; d_{\mathrm{W}} \left(b_c^m, b_d^j\right) \tag{21}$$

The index d denotes the customer whose purchase history is being analysed and the index j denotes for which basket of the purchase history the distance is computed. As the last basket of each customer d is not compared, the j ranges between the first and the second to last basket present in the purchase history of customer B_d. The next basket b_c^{m+1} is then predicted as follows:

1. **Determine the basket length**: The length of the next market basket of n_c is equal to the average basket length of the purchase history of that customer.

2. **Products in basket**: The items b_c^{m+1} will consist of is determined by the top n_c most frequent items in the baskets of the k-nearest neighbours of c, if the average Wasserstein distance $\bar{d}_W$ of the neighbours is below a threshold ϕ. Where $\bar{d}_W$ is computed as:

$$\bar{d}_W = \sum_{d=1}^{k} \frac{d_{d,W}}{k} \tag{22}$$

   If the average distance is above the threshold, $\bar{d}_W > \phi$, this approach reverts back to the same fallback prediction as used by Kraus and Feuerriegel (2019) (i.e. *the personal top items*). The value of ϕ is found by studying the course of the performance metrics against a range of values for this threshold. The range is determined by analysing the distribution of the average distances of the k-nearest neighbours.

Again, to measure the predictive performance the performance metrics described in section 4.1.5 are used.

5 Results

For all computations the first Wasserstein distance is used (i.e. $t = 1$), following Kraus and Feuerriegel (2019).

5.1 Product embeddings

The product embeddings are valued using the entire dataset, which is the dataset before applying the imposed restrictions described in section 3. Each product embedding has a dimension of 50×1 and the window used is 5 ($w = 5$), following Kraus and Feuerriegel (2019). A product is translated to an embedding if it is present at least 50 times in the entire dataset. In appendix A2 an example of the performance of the product embeddings on product-level is displayed for 3 randomly chosen products.

5.2 Prediction performance product-level

Table 1 show the values of the performance metrics for the K&F approach and the four baseline models for my sub-test set. The results of the K&F approach are obtained with the following values for the parameter:

- The number of nearest neighbours k is equal to 5, which is the middle value of the tuning range proposed by Kraus and Feuerriegel (2019). The reason that I only chose one value is due to time limits.
- The threshold τ for the fallback prediction is set to 35, which is the largest value in tuning range for this parameter proposed by Kraus&Feuerriegel.

It can be concluded from Table 1 that the *Association rules* model performs the worst compared to the other in terms of all metrics. The Jaccard coefficient is only 0.068 meaning that the ratio of co-occurrences to non-co-occurrences is very low. The same conclusion holds for the *Global top items* model, although it performs slightly better than the former. The *Repurchase last basket* already performs twice as good, but the best model seems to be that of *Personal top items*. It can be seen that this model has the exact same values for the metrics as that for the K&F approach. This can be explained by the fallback prediction. As described in section 4 the next market basket of some customer c will be determined by a fallback prediction if their average neighbour distance is larger than the threshold. Meaning it will be predicted by the *Personal top items* model. This indicates that a threshold of 35 would exclude all predictions made by the K&F and thus explain why the performance metrics are identical.

Table 1: Performance metrics of the four baseline models and the Kraus&Feuerrieggel approach for the sub-test set from the Instacart dataset on product-level

Model	**Wasserstein distance**	**F1-Score**	**Jaccard coefficient**
Global top items	9.467	0.159	0.091
Personal top items	7.274	0.351	0.232
Repurchase last basket	8.219	0.258	0.163
Association rules	10.945	0.121	0.068
K&F approach	7.274	0.351	0.232

The average neighbour distances of all customers in my sub-test set range between [39.808, 87.241], thus never reaching below the threshold. This can also be seen in the histograms provided in appendix A3. Taking for instance the median value of these distances as the threshold (τ=73.211) results in a Wasserstein distance of 8.249, an F1-Score of 0.242 and a Jaccard coefficient of 0.148. To determine a fair threshold for the K&F approach, Figure 1 is studied. A steep deterioration is observed for a threshold value around 60. Therefore, I choose τ equal to 60 results in a Wasserstein distance of 7.313, an F1-Score of 0.348 and a Jaccard coefficient of 0.229. Performing slightly less than the *Personal top items* approach. A conclusion that can be drawn from the above results is that a customer's personal taste is a relatively important factor for next market baskets prediction, as the models *Personal top items* and *Repurchase last basket* predict relatively well. Both these models rely solely on an individual's purchase history for prediction. This could explain why the *Association rules approach* does not predict well as the main source of prediction information is looking at all customers.

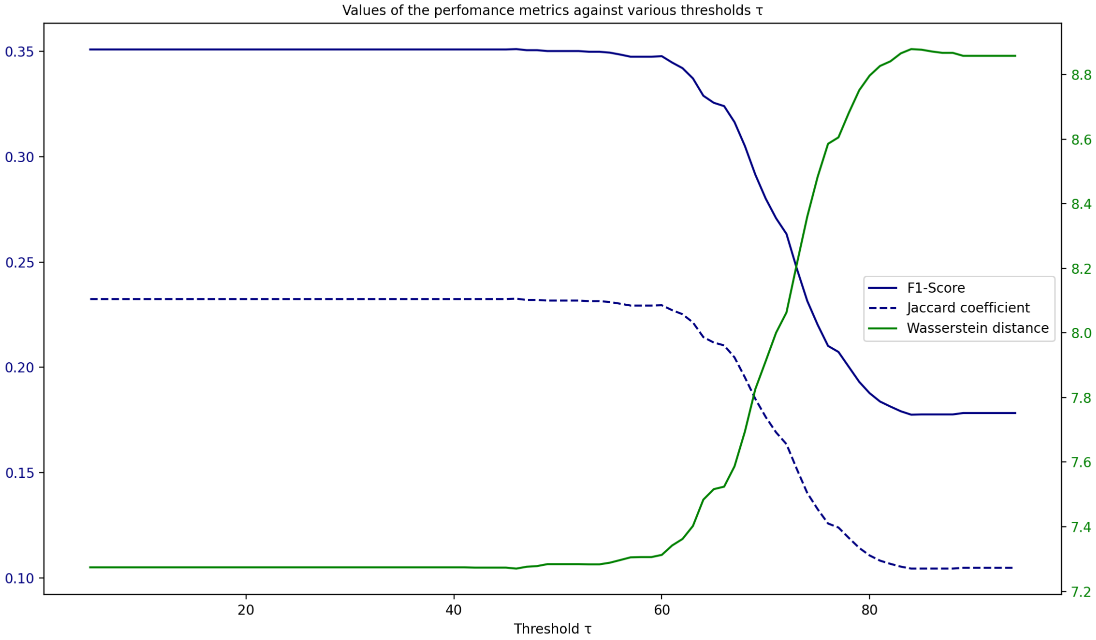


Figure 1: Course of the performance metrics as a function of the threshold τ for the K&F approach

5.3 Prediction performance category-level

Following from product-level analysis on the employed dataset, the K&F approach thus performs relatively well. Table 2 displays the values of the performance metrics for the four baseline models and the K&F approach on category-level. The F1-Score and Jaccard coefficient for all models are significantly better compared to their values in Table 1. This can be explained by the fact that when studying category-level, the products are comprised to their categories. Meaning that all baskets become more similar making prediction easier, as there are less options to predict. Again, the *personal top items* perform best.

Table 2: Performance metrics of the four baseline models and the Kraus&Feuerrieggel approach for the sub-test set from the Instacart dataset on category-level

Model	**Wasserstein distance**	**F1-Score**	**Jaccard coefficient**
Global top items	7.464	0.395	0.266
Personal top items	5.366	0.584	0.444
Repurchase last basket	6.091	0.513	0.373
Association rules	8.483	0.281	0.176
KF approach ($\tau = 60$)	5.415	0.581	0.441

5.4 Part 2

Following from part 1, it can be observed that including a customer's specific information for that customer's prediction results in overall better predictive power. To test this observation further, I modify the *Association rules approach* by utilizing a customer's history for their prediction instead of that of all customers. This model, called *individual-specific association rules*, has the following predictive performance on the same sub-test as mentioned before: a Wasserstein distance of 9.833, an F1-Score of 0.218 and a Jaccard coefficient of 0.133. This modified version outperforms its original approach by 80.2% in terms of the F1-score and 95.6% in terms of the Jaccard coefficient. Based on these results and the conclusions drawn in part 1, it can be concluded that the includement of a customer's own information for that customer's prediction results in significantly better predictive performance.

To potentially improve the K&F approach, this thesis introduces a modification of this approach, *my approach*, where I combine individual-specific information (the computation of the length of the next market basket prediction) with parts of the K&F approach (e.g. Wasserstein distance, fallback prediction and cross-customer knowledge). A noticeable difference between my approach and the K&F approach is that I do not compare (sub-)purchase histories of potential neighbours, but compare individual baskets. The reasoning behind this is that the running time of the K&F approach is long (see section 5.5 for detailed information about the computation performance) as all possible sub-purchase histories of a potential neighbour are compared. The performance metrics of my approach, again for the same sub-test set, are: a Wasserstein distance of 7.929, an F1-Score of 0.284 and a Jaccard coefficient of 0.183. The value of ϕ is the median value of the average neighbour distances, $\phi = 4.931$. To allow for fair comparison between all models in this thesis, I investigate what the optimal threshold ϕ should be for my approach. The average distances of the 5 nearest neighbours range between [1.451, 6.726]. This can also be seen in appendix A4. To clarify the big difference between the range of my approach and the range of the K&F, my approach' range represent the distance between baskets, whereas K&F's range represent the distance between (partial) purchase histories.

To determine an actual value for the threshold ϕ, Figure 2 is studied. The performance metrics deteriorate fast after a threshold of about 3.5. This value for ϕ results in the performance metrics being equal to 7.338 for the Wasserstein distance, an F1-Score of 0.345 and a Jaccard coefficient of 0.227. Performing slightly less than the *Personal top items* and almost identical to the K&F approach with the estimated threshold, following from Table 3.

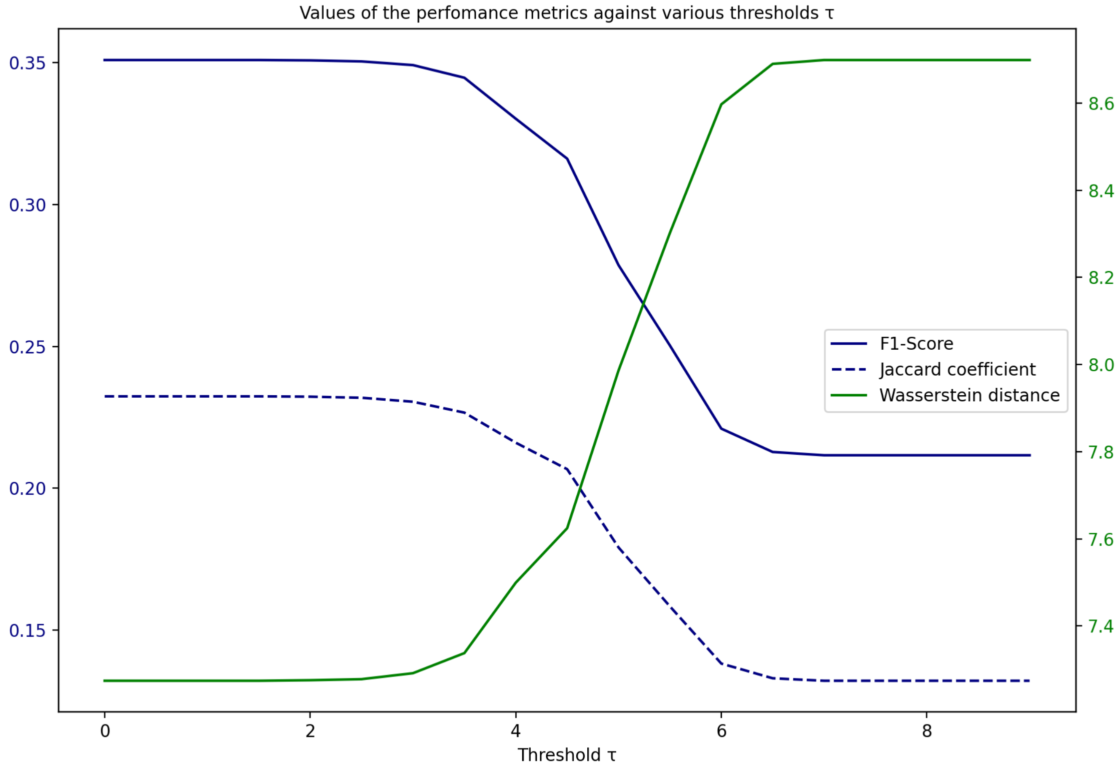


Figure 2: Course of the performance metrics as a function of the threshold τ for my approach

Table 3: Performance metrics of the all models for the sub-test set from the Instacart dataset

Model	**Wasserstein distance**	**F1-Score**	**Jaccard coefficient**
Global top items	9.467	0.159	0.091
Personal top items	7.274	0.351	0.232
Repurchase last basket	8.219	0.258	0.163
Association rules	10.945	0.121	0.068
KF approach ($\tau = 60$)	7.313	0.348	0.229
Individual association rules	9.833	0.218	0.133
my approach ($\phi = 3.5$)	7.338	0.345	0.227

5.5 Computational performance

The running time of the K&F approach on my 300 customer covering sub-test set was about 39 hours for the product-level analysis. The running time of my approach was about 13.5 hours. All other models have a running time of a few seconds. The running time of the K&F approach on my 300 customer covering sub-test set was about 50 hours for the category-level analysis. The difference between running time for the K&F approach on both levels could be explained by basket similarity. As the baskets on category-level are more similar than they are on product-level, less computations are skipped by the lower-bound threshold (section 4.1.2). I use two methods to speed up my code. First of all, I use numba's jit package. Which allows functions that strictly use *numpy* to run significantly faster. My second method is running the code using multi-processing. The results are based on an Intel Core i5 2.9 GHz processor with 4 cores and 16 GB RAM. All used Python packages and their version are specified in appendix A5.

6 Conclusion

Predicting future demand has many benefits for both online and offline stores. It helps to optimize the inventory management and marketing strategies. It, therefore, increases the revenue which is one of the, if not most, important goals of a commercial company. But how can we predict what the next order of a customer will consist, given historical sales data? In the first part of this thesis I replicated a part of Kraus and Feuerriegel (2019), where I not only recreated their model but also investigated and compared four baseline models. It can be concluded that K&F approach does not outperform all of these state-of-the-art models. The performance metrics for predictions based on a customer's top items out-performed the prediction made by the K&F approach. This is because the latter approach uses the former approach as a fallback prediction when the neighbours average distance exceeds a certain threshold. The smaller the threshold value the more predictions will be made by the fallback prediction, thus making the two approaches more similar. However, if I increase this threshold the model's predictive power decreases, meaning that the more predictions made by the actual approach the worse the performance metrics will be. Kraus and Feuerriegel (2019) concluded that their approach outperforms all baseline model although the difference with *Personal top item* was small. A possible explanation for the difference in conclusions is the size of the used test-set. I utilize the same dataset as them, but my results only cover the predictions for 300 customers (sub-test set). Which is only about ten percent of the number of customers Kraus and Feuerriegel (2019) used. The fallback prediction the K&F approach uses is based on the *Personal top item* model. Meaning that it is possible that only a few customers whose predictions are better when estimated by Wasserstein-sequence matching instead of the fallback prediction, are included in the sub-test set. Explaining that they conclude that their approach out-performed the *Personal top item* by a little. It could be that I only used a small number, or none at all, of customers where the prediction based on Wasserstein-sequence matching was better than the prediction made by the approach' fallback prediction. Resulting in the conclusion that the *Personal top item* model is better.

The second conclusion that is drawn is the importance of customer-specific information in a prediction model. The models that performed relatively well used customer-specific information for their predictions. To further investigate this observation I modified a baseline model that does not use customer-specific information by including this type of information. It is concluded that the modified version increases the prediction performance by 80.17% in terms of F1-score compared to the original model. This reinforces the presumption of the importance of including customer-specific information. Building on this conclusion and the fact that the Kraus and Feuerriegel (2019) performs relatively slow, I introduce my own approach that is a modification of Kraus and Feuerriegel (2019). It can be concluded that this approach performs almost identical compared to the K&F approach, while having a running time that is 65% faster than the K&F approach. Although, my approach incorporates customer-specific information it does not perform better than the K&F approach. A possible explanation is that including customer-specific information is only useful when it contributes to the prediction of the exact items in the next market basket. Whereas I only used this type of information to determine the length of the next market basket.

Based on the performance metrics evaluated over my sub-test set, the best method to predict a customer's next market basket for the employed dataset is the *Personal top items* model. It outperforms all other models analysed in this thesis, although sometimes the differences are small. Generally, the models that performed best used customer-specific information for the predictions. It can be concluded that customer-specific information thus contributes positively to the prediction performance of a model.

Discussion & future research

The size of my utilized test set is small (sub-test set). Which could explain some of the differences I experienced compared to the results of Kraus and Feuerriegel (2019). I, therefore, recommend using a larger test set for future research. I compared the K&F approach and my own approach by using threshold values that are determined by analysing the predictive performance of my sub-test set. In other words, the thresholds give the best results for the sub-test set. This does not guarantee that these thresholds are optimal for the rest of the data. Which could alter the conclusions. For future research, this should be taken into account either by employing a sub-test set or using k-fold cross validation. I only investigated the K&F approach with five as the number of nearest neighbours that should be used in the prediction. Again, it is possible that a different number could lead to different conclusions. Another improvement for future research is the use of different datasets. Can the same conclusions be drawn on those other datasets? If not, that would mean that the model only works proper for some datasets. And naturally vice versa, why would there be a significant difference in model performance on different datasets? My final remark for future research is adjusting the K&F approach by incorporating more customer-specific information, for instance determine the next market basket length by averaging the lengths of the customer's purchase history instead of using the length of the nearest neighbour. The final limitation in my research is the way duplicate items in market baskets are handled. As of now, it is not possible to predict the frequency of a product in a market basket. Meaning that if a product should be included, I predict that it is only purchased one time. This is not realistic as it possible that customers purchase multiples of a specific product during one shopping trip.

References

[1] Rakesh Agrawal, Ramakrishnan Srikant, et al. *Fast algorithms for mining association rules.* 1994.

[2] Sergey Brin, Rajeev Motwani, and Craig Silverstein. *Beyond market baskets: Generalizing association rules to correlations.* 1997.

[3] Yen-Liang Chen, Kwei Tang, Ren-Jie Shen, and Ya-Han Hu. *Market basket analysis in a multiple store environment.* 2005.

[4] Chad Cumby, Andrew Fano, Rayid Ghani, and Marko Krema. *Predicting customer shopping lists from point-of-sale purchase data.* 2004.

[5] Marnik G Dekimpe and Dominique M Hanssens. *Time-series models in marketing:: Past, present and future.* 2000.

[6] R. Guidotti, G. Rossetti, L. Pappalardo, F. Giannotti, and D. Pedreschi. *Personalized Market Basket Prediction with Temporal Annotated Recurring Sequences.* 2019.

[7] Riccardo Guidotti, Giulio Rossetti, Luca Pappalardo, Fosca Giannotti, and Dino Pedreschi. *Market basket prediction using user-centric temporal annotated recurring sequences.* IEEE, 2017.

[8] Mathias Kraus and Stefan Feuerriegel. *Personalized Purchase Prediction of Market Baskets with Wasserstein-Based Sequence Matching.* 2019.

[9] Andres Musalem, Luis Aburto, and Maximo Bosch. *Market basket analysis insights to support category management.* 2018.

[10] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. *Factorizing personalized markov chains for next-basket recommendation.* 2010.

[11] *The Instacart Online Grocery Shopping Dataset.* Published on Kaggle. 2017.

[12] A. Tomar. *A math-first explanation of Word2Vec.* `https://medium.com/analytics-vidhya/maths-behind-word2vec-explained-38d74f32726b`. Accessed: 2010-07. 2019.

[13] Pengfei Wang, Jiafeng Guo, Yanyan Lan, Jun Xu, Shengxian Wan, and Xueqi Cheng. *Learning hierarchical representation model for nextbasket recommendation.* 2015.

[14] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. *A dynamic recurrent model for next basket recommendation.* 2016.

Appendix

A1: Treemap displaying the relative selling frequency of the departments of the 500 most sold products (product-level)

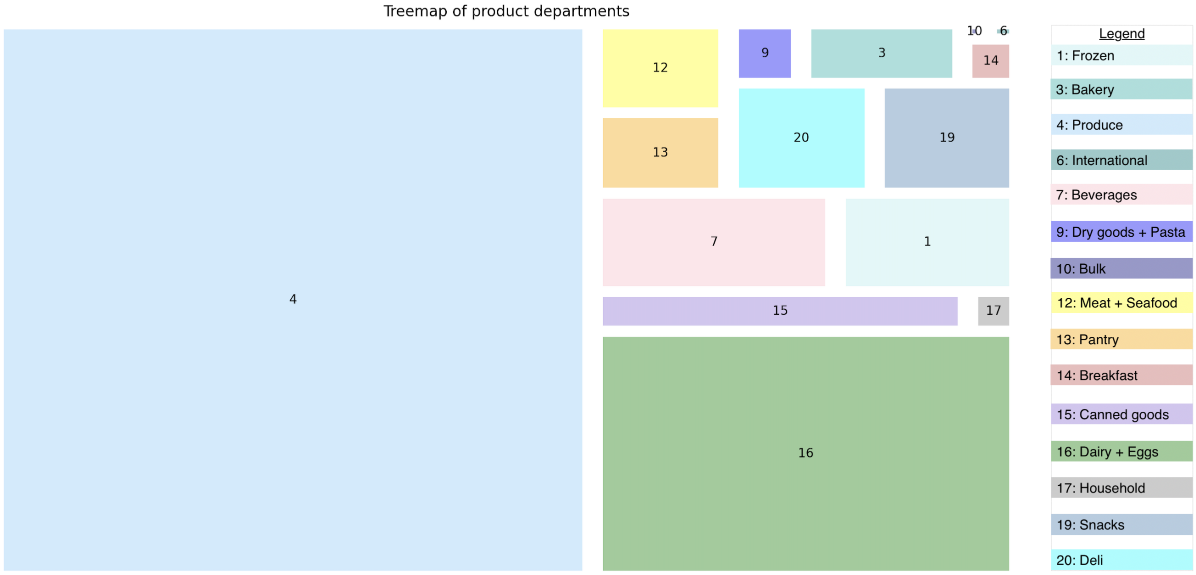
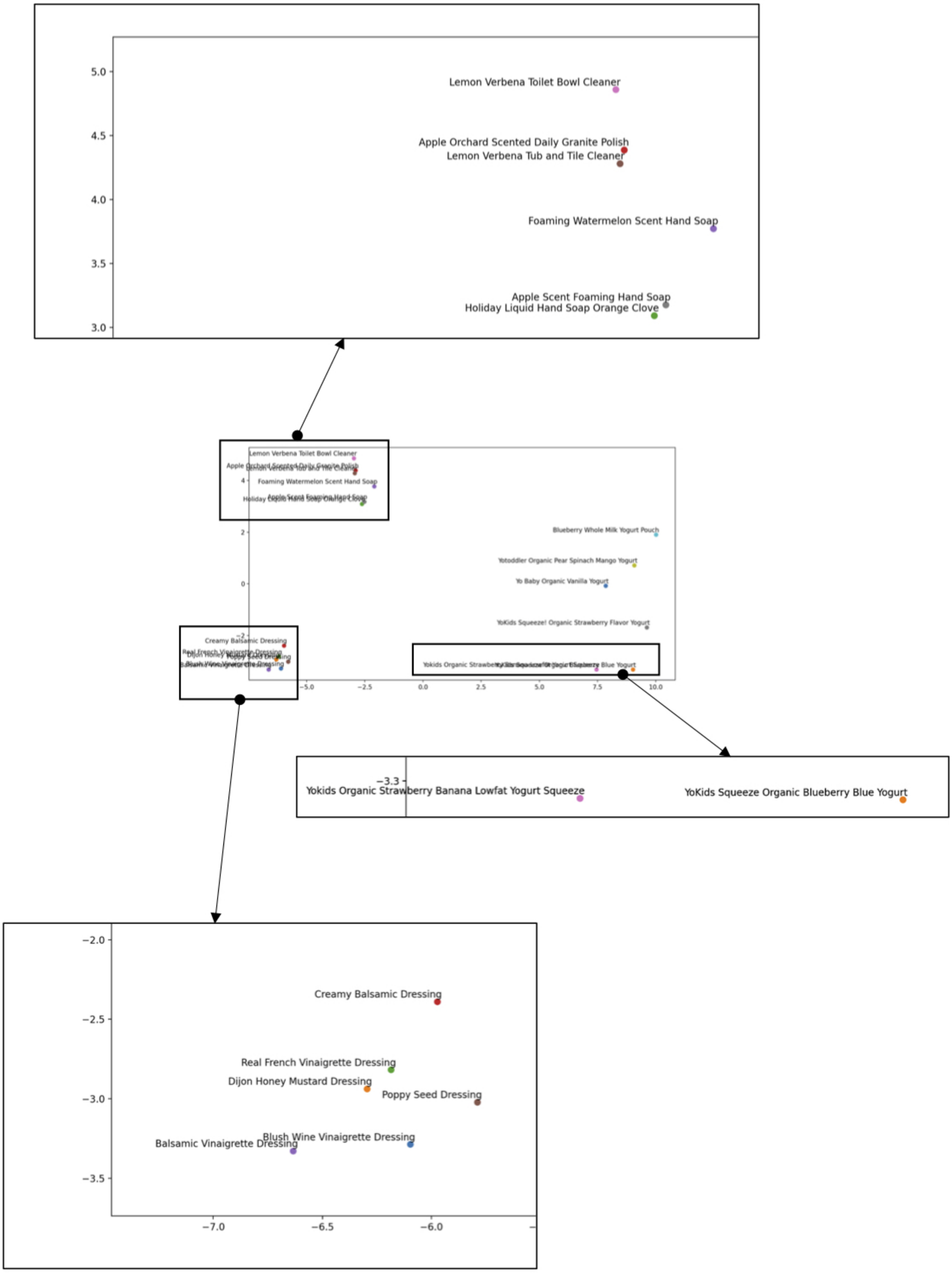


Figure 3: Treemap of the frequency of products categorized by their departments. The legend indicates the department name belonging to each color + number combination

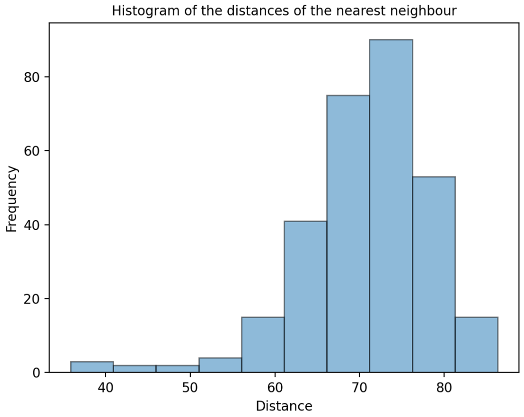
A2: Product embedding plot

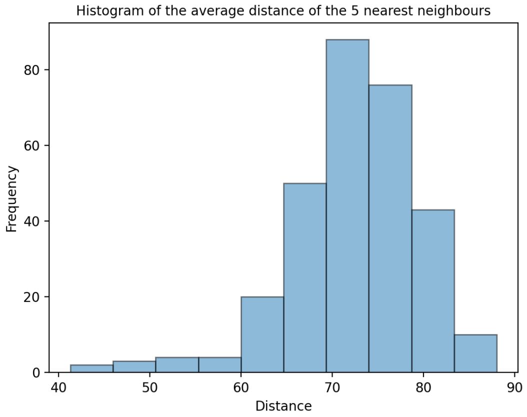
The below figure plots 3 randomly chosen products and their 5 most similar one's out of the dataset, in terms of cosine distance, in a two-dimensional space. It can be seen that there are 3 groups, all with similar products. The group at the right contains different types of yoghurt, the group at the bottom-left consists of types of dressings and the last group, top-left, contains household products. Meaning that similar products are indeed closer together.


5.0
4.5
4.0
3.5
3.0
Lemon Verbena Toilet Bowl Cleaner
Apple Orchard Scented Daily Granite Polish
Lemon Verbena Tub and Tile Cleaner
Foaming Watermelon Scent Hand Soap
Apple Scent Foaming Hand Soap
Holiday Liquid Hand Soap Orange Clove
4
2
0
−2
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0
Blueberry Whole Milk Yogurt Pouch
Yotoddler Organic Pear Spinach Mango Yogurt
Yo Baby Organic Vanilla Yogurt
YoKids Squeeze! Organic Strawberry Flavor Yogurt
−3.3
Yokids Organic Strawberry Banana Lowfat Yogurt Squeeze
YoKids Squeeze Organic Blueberry Blue Yogurt
−2.0
−2.5
−3.0
−3.5
−7.0
−6.5
−6.0
Creamy Balsamic Dressing
Real French Vinaigrette Dressing
Dijon Honey Mustard Dressing
Poppy Seed Dressing
Blush Wine Vinaigrette Dressing
Balsamic Vinaigrette Dressing

A3: Distributions of the distances of the 5 nearest neighbours for the product-level analysis

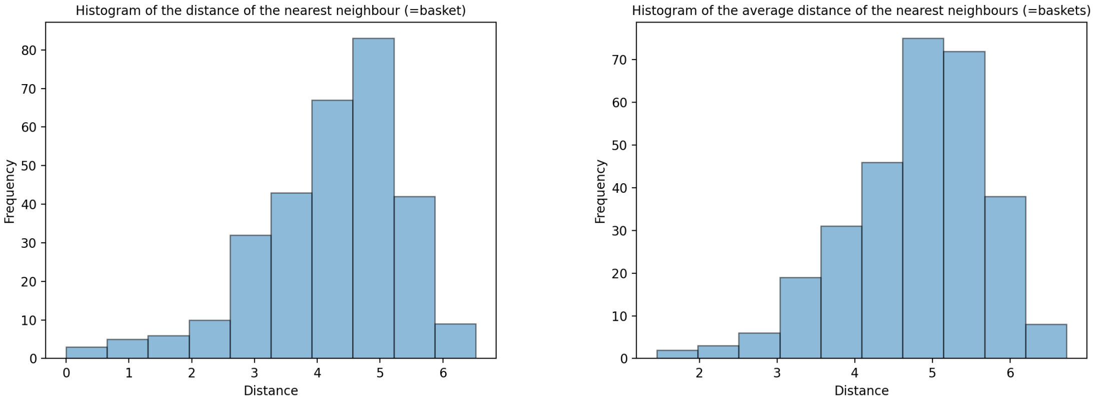
The below histograms o display the distances of the nearest neighbour and of the average distance of the 5 nearest neighbours for the K&F approach. The left histogram in this figure shows the distribution of the distances of the nearest neighbour, showing that the distance of the nearest neighbour is above 35. The histogram on the right shows the distribution of the average neighbour distances. The majority of values are above 50.





A4: Distributions of the distances of the 5 nearest neighbours for the category-level analysis

The below figure displays histograms of the distances of the nearest neighbour and of the average distance of the 5 nearest neighbours for my approach. The left histogram in this figure shows the distribution of the distances of the nearest neighbour, showing there are a few baskets that are exactly the same as the basket of the customer in the sub-test set (distance = 0). The histogram on the right shows the distribution of the average neighbour distances. The majority of values are above 4.



A5: Packages version

Python 3.8	
Package	**Version**
boto	2.49.0
boto3	1.13.16
botocore	1.16.16
certifi	2020.4.5.1
chardet	3.0.4
cycler	0.10.0
Cython	0.29.19
docutils	0.15.2
gensim	3.8.3
idna	2.9
jmespath	0.10.0
joblil	0.15.1
kiwisolver	1.2.0
llvmlite	0.32.1
matplotlib	3.2.1
numba	0.49.1
pandas	1.0.3
pip	19.0.3
POT	0.7.0
pyparsing	2.4.7
python-dateutil	2.8.1
pytz	2020.1
PyYAML	5.3.1
requests	2.23.0
s3transfer	0.3.3
scikit-learn	0.23.1
scipy	1.4.1
setuptools	40.8.0
six	1.15.0
sklearn	0.0
smart-open	2.0.0
threadpoolctl	2.0.0
tqdm	4.46.0
urllib3	1.25.9