CrossBERT: Analysing Market Basket Cross-Effects Using Masked Language Modelling

Bachelor Thesis Business Analytics and Quantitative Marketing
Mark Rademaker, 467978
Erasmus University Rotterdam
Erasmus School of Economics

Supervisor: Luuk van Maasakkers
Second assessor: Dr. Kathrin Gruber

July 2, 2021

Abstract

This paper aims to establish a model that extracts the cross-effects of many product categories in a supermarket. We introduce a model created from a transformer-based machine learning technique called BERT, which stands for Bidirectional Encoder Representations from Transformers, named CrossBERT. We use the attention mechanism of this model to find cross-effects of different product categories. These cross-effects are relevant for making informed marketing decisions that increase sales of a supermarket. To evaluate the performance of the CrossBERT model we compare the model with a multivariate logit model and a restricted Boltzmann machine model, which are models proven to be able to extract cross-effects successfully. The findings in this paper reveal that CrossBERT is better able to extract cross-effect from the data than the other models. From these obtained results we learn that aisle location, diet preferences, and common recipes are drivers of these cross-effects. This paper shows potential applications of CrossBERT in future research.

Contents

1 Introduction

Throughout the years the importance of data has vastly increased. Retail businesses understand that customer behaviour data can be analysed and used in many ways to improve the performance of their business. For supermarkets, Felgate Fearne (2015) have shown that promotions of products do not systematically result in capital growth, indicating that it is highly beneficial for supermarkets to make good informed marketing decisions. As data analysis is essential to retail businesses for making informed marketing decisions, the accuracy and feasibility of the models used become pivotal.

One measure of customer behaviour that is interesting to supermarkets and other retail businesses for making these informed decisions are cross-effects between their products. For customers, purchasing decisions can depend on many factors. These factors include diet preferences, prices, and the location in the supermarket. A customer that wants to make a hamburger will probably buy hamburgers as well as hamburger buns and cheese, these products are complementary goods. Whenever one product negatively impacts the probability of buying the other, the products are substitute goods (e.g., hamburgers and veggie burgers). These correlations between products give valuable information for supermarket holders for their marketing strategy. Whenever a supermarket knows how the purchase of one product affects the probability of buying a different product, an effective marketing strategy can be formed that is. A supermarket can drop the price of one product and may or may not expect to see an increase or decrease in the number of sales of another product, depending on the complementary nature of the products. Whenever a customer buys eggs, the probability of that same customer buying bacon is immediately impacted. These correlations can be used to determine marketing measures that maximize the revenue of a supermarket. In this paper, we try to determine the best method that can be used to extract these cross-effects of many categories from a dataset. Furthermore, we discuss the relevance of the cross-effects for possible marketing measures.

In Russell Petersen (2000), the method multivariate logit model (MVL) is first used to capture the cross-effects of products. This model however requires many calculations and as a consequence, it is limited in the number of products or categories the method can cover. In Hruschka (2014) a restricted Boltzmann machine (RBM) is successfully used as an alternative for deriving the cross-effects with a higher number of product categories. In this paper, we replicate the paper of Hruschka (2014) and compare an RBM with the MVL model. We analyse the performance of the RBM method and extract the cross-effects for many categories on the Instacart dataset, (Instacart, 2017). Furthermore, we try to improve the method by analysing the same dataset withBERT, an abbreviation for Bidirectional Encoder Representations from Transformers, first introduced by Devlin (2019). When training BERT on understanding the compositions of market baskets instead of its usual task of understanding language, we expect to get a model that is able to extract the cross-effects from the data successfully, due to the attention mechanism the model uses (Vaswani ., 2017). To ensure that the BERT model creates an understanding of supermarket baskets we make use of the training task Masked-Language Modelling (MLM).

This paper begins with a literature review. Thereafter we describe the dataset in the data section. In the replication section, we describe the RBM and MVL model and in the extension, section we describe the CrossBERT model. In both parts, we provide the methods used to extract the cross-effects in the methodology with the main findings of the methods described in the results section. The paper ends with the conclusion and a brief discussion. All codes in this paper that are used to obtain the results using CrossBERT can

be found on the Github page: `https://github.com/MarkRademaker/CrossBERT`.

2 Literature Review

The multivariate probit model (MVP) model is in Manchanda . (1999) proven to be outperforming the single-category models in predicting multicategory choice by analyzing 4 categories. In the paper, they are able to separate the variables that drive the multicategory choice that can be controlled, such as cross-price, from those that can not be controlled. From this paper, Russell Petersen (2000) followed by providing an MVL model for extracting cross-effects of 4 categories. This was later done for 6 categories by Boztuğ Hildebrandt (2008). The limited number of categories exposes the drawback of the model. The model is namely inefficient when used to derive the cross-effects of many categories using maximum likelihood (ML).

In Hruschka (2014), an RBM is proven to be a computationally efficient alternative to the MVL model. In case the cross-effects of many products should be derived, the RBM model will be more suitable as the optimisation of the parameters is limited by the introduction of a hidden layer. The hidden layer provides a way to reduce the number of parameters from 3660 for 60 categories when using MVL to 360 when using 5 hidden nodes. A more detailed explanation of the model will be provided in Sec. 4. In Hruschka (2014), they show that 25% of the market baskets can not be adequately modeled without using cross-effects for the RBM. The paper finds that the RBM model outperforms the MVL model for the validation data. The paper is the first that uses ML as an approximation method and therefore differs from the widely used contrastive divergence algorithm, which is an approximation method.

As opposed to the aforementioned papers in Devlin (2019) the main goal is to create a model for natural language processing. Natural language processing (NLP) is the field of computer science that copes with the technology of computers trying to understand the human language. In this paper, they develop the transformer-based machine learning technique BERT. The model is trained to understand language and has proven to be a powerful model in the field of NLP. Its success is partly due to the attention mechanism the model uses. In Vaswani . (2017) the power of the attention mechanism is shown. In this paper the development of a neural network is described that can incorporate the context of a word into the meaning of that word effectively. This mechanism contributes to the BERT model by allowing the language model to be context-aware. The attention mechanism creates possibilities to use the model for a variety of tasks where the context contains critical information.

3 Data

For this research, we make use of the Instacart dataset, (Instacart, 2017). This dataset contains 134 different aisles that consist of many different products in an online supermarket. To keep the MVL computationally feasible we reduce these 134 aisles to 60. The dataset contains 1,048,575 orders from 63,100 customers using Instacart to deliver their groceries. For each order, the dataset contains the products ordered in the time they are added to the market basket by a customer. The supermarket categories that are bought the most based on 30,000 orders of the dataset will be included in the analysis.

Table 1 shows the univariate marginal frequencies of the 60 most frequent categories out of a sample of 30,000 orders, all from different customers in the dataset. These 30,000

orders should give a fair representation of the marginal univariate and bivariate frequencies in the entire dataset. The univariate marginal frequencies are the proportion of the 30,000 orders that contain the specific category. As can be seen in Table 1, vegetables and fruits are bought the most. Table 2 shows the bivariate marginal frequencies of the 60 most frequent pairs, these bivariate marginal frequencies give the proportion of orders that contain the two mentioned categories in Table 2.

From Table **??**, we see that vegetables and fruits are the most common pair in a basket. However, this does not have to imply that the cross-effects between these products are significant. If products are uncorrelated, the bivariate frequency of Table 2 would approximately be the product of univariate marginal frequencies in Table 1. Whenever the bivariate marginal frequency in Table 2 deviates from the product of the univariate marginal frequencies of both categories in Table 1, there could be a correlation between both categories. However, this is not always true as there could be external factors that influence the bivariate marginal frequencies between a pair. For example, two products could have the cross-effects of the same sign with the same products when they do not affect each other and are thus independent. A customer could buy chips and cookies whenever he buys beer. This does not imply that buying cookies affects the probability of buying chips or vice versa. A more precise method is therefore necessary to extract the cross-effects.

Table 1. Relative univariate marginal frequencies of analyzed categories per order

Fresh fruits	0.551	Fresh vegetables	0.455	Packaged vegetables fruits	0.386
Yoghurt	0.252	Packaged cheese	0.238	Milk	0.224
Water seltzer sparkling water	0.205	Chips pretzels	0.174	Soy lactose free	0.174
Bread	0.161	Eggs	0.149	Refrigerated	0.133
Frozen produce	0.129	Ice cream ice	0.122	Crackers	0.119
Lunch meat	0.108	Fresh dips tapenades	0.102	Fresh herbs	0.101
Cereal	0.099	Soft Drinks	0.098	Juice nectars	0.094
Other creams cheeses	0.089	Soup broth bouillon	0.088	Cream	0.088
Hot dogs bacon sausage	0.087	Energy granola bars	0.082	Frozen meals	0.081
Canned jarred vegetables	0.079	Spreads	0.079	Nuts seeds dried fruit	0.079
Paper goods	0.077	Butter	0.076	Baking ingredients	0.076
Packaged produce	0.074	Dry pasta	0.074	Oil vinegars	0.072
Canned meals beans	0.072	Pasta sauce	0.068	Breakfast bakery	0.067
Candy chocolates	0.066	Condiments	0.063	Frozen breakfast	0.062
Cookies cakes	0.062	Instant foods	0.060	Tortillas flat bread	0.059
Frozen appetizers sides	0.059	Spices seasonings	0.056	Coffee	0.054
Tea	0.053	Frozen pizza	0.050	Popcorn jerky	0.048
Fruit vegetables snacks	0.044	Grains rice dried goods	0.044	Asian foods	0.044
Hot cereal pancake mixes	0.043	Baby food formula	0.041	Packaged poultry	0.039
Poulty counter	0.039	Buns rolls	0.037	Preserved dips spreads	0.034

Table 2. Relative bivariate marginal frequencies of analyzed categories per order

Fresh vegetables, Fresh fruits	0.333	Fresh fruits, Packaged vegetables fruits	0.290
Fresh vegetables, Packaged vegetables fruits	0.255	Fresh fruits, Yoghurt	0.183
Packaged cheese, fresh fruits	0.160	Fresh fruits, Milk	0.151
Fresh vegetables, Yoghurt	0.145	Packaged cheese, Fresh vegetables	0.144
Yoghurt, Packaged vegetables fruits	0.134	Packaged cheese, Packaged vegetables fruits	0.125
Fresh fruits, Soy lactosefree	0.123	Fresh vegetables, milk	0.119
Fresh fruits, Water seltzer sparkling water	0.118	Fresh fruits, Bread	0.111
Fresh fruits,Chips Pretzels	0.108	Fresh fruits, Eggs	0.106
Milk, Packaged vegetables fruits	0.104	Fresh vegetables, Soy lactosefree	0.102
Fresh vegetables, Eggs	0.095	Fresh fruits, Frozen produce	0.094
Fresh vegetables, Bread	0.094	Fresh vegetables, Water seltzer sparkling water	0.092
Soy lactosefree, Packaged vegetables fruits	0.090	Fresh vegetables, Chips pretzels	0.088
Fresh vegetables, Fresh herbs	0.086	Fresh fruits, Refrigerated	0.086
Packaged cheese, Yoghurt	0.085	Milk, Yoghurt	0.085
Fresh vegetables, Frozen produce	0.082	Water seltzer sparkling water, Packaged vegetables fruits	0.082
Packaged vegetables fruits, Chips pretzels	0.080	Packaged vegetables fruits, Bread	0.080
Packaged vegetables fruits, Eggs	0.079	Fresh herbs, Fresh fruits	0.078
Fresh fruits, Crackers	0.078	Packaged cheese, Milk	0.076
Fresh fruits, Lunch meat	0.074	Frozen produce, Packaged vegetables fruits	0.074
Fresh fruits, Ice cream ice	0.073	Fresh fruits, Fresh dips tapenades	0.072
Refrigerated, Fresh vegetables	0.068	Fresh vegetables, Lunch meat	0.065
Fresh fruits, Cereal	0.064	Refrigerated, Packaged vegetables fruits	0.063
Fresh vegetables, Fresh dips tapenades	0.062	Fresh vegetables, Canned jarred vegetables	0.061
Soy lactosefree, Yoghurt	0.061	Fresh herbs, Packaged vegetables fruits	0.060
Packaged cheese, Chips pretzels	0.060	Fresh fruits, Juice nectars	0.060
Fresh vegetables, Ice cream ice	0.060	Fresh vegetables, Cracker	0.060
Packaged cheese, Bread	0.059	Fresh dips tapenades, Packaged vegetables fruits	0.059
Fresh vegetables, Soup broth bouillon	0.059	Bread, Yoghurt	0.059
Fresh fruits, Other creams cheeses	0.058	Lunch meat, Packaged vegetables fruits	0.058
Yoghurt, Water seltzer sparkling water	0.058	Fresh fruits, Soup broth bouillon	0.057

4 Replication

In this paper, we first replicate the models of Hruschka (2014) and derive the cross-effects from the RBM and MVL model. Thereafter, a BERT model is implemented, named CrossBERT, that derives the cross-effects by using MLM. This CrossBERT model serves as an extension to the replication.

An RBM model is a two-layer neural network. The network consists of a visible layer and a hidden layer, where the visible layer receives the data as input. The method relies on the conditional probability between the two layers to try to reconstruct the data as good as possible. These conditional probabilities are determined by the weights connecting the nodes of the layers. Compared to a general Boltzmann machine model the restricted Boltzmann machine model has no edges between the hidden nodes or visible nodes. The weights thus represent only the relations between the visible nodes and the hidden nodes. The weights between two different visible nodes and the same hidden node can however serve as a representation of the correlation between visible nodes for one hidden node. Using these weights between the two layers we can find the cross-effects.

To replicate the findings in Hruschka (2014) we compare the performance of the MVL model using the with the RBM model in this section using ML estimation on the Instacart dataset. Furthermore, we try to conclude whether the RBM method is a computationally feasible method to get cross-effects for data including many categories.

4.1 Methodology

To compare the MVL and RBM methods we analyze the performance of capturing the cross-effects. The MVL and RBM methods rely on their probability function: equation 4.1 and equation 4.6, respectively. For the data we use the vector of purchasing incidences $y_i = \{0,1\}^J$ where $y_{i,j} = 1$ for category $j = 1, ..., J$ and order $i = 1, ..., I$ is in the basket, $y_{i,j} = 0$ otherwise. First, we show the methods used to derive the probabilities and likelihood function for the MVL method, whereafter the probabilities and likelihood function of the RBM model follow.

The MVL model uses the logit function for the basket probability function:

$$p(y_i) = \frac{exp(\boldsymbol{a'y_i} + \boldsymbol{y_i}'\boldsymbol{V}\boldsymbol{y_i})}{Z_{MVL}} \tag{4.1}$$

where,

$$Z_{MVL} = \sum_{y \in \{0,1\}^J} exp(\boldsymbol{a'y} + \boldsymbol{y'V}\boldsymbol{y}). \tag{4.2}$$

The vector $\boldsymbol{a}$ contains the constant for the categories, which are related to the univariate marginal frequencies of the products. For matrix V the element $V_{i,j}$ contains information on the bivariate frequency of the products. When $V_{ij} > 0$ the products i and j are bought more often compared to the conditional independence, for $V_{ij} < 0$ the opposite applies. The normalization constant Z_{MVL} is obtained by summing over all possible market basket configurations, which is are 2^J market baskets. From equation 4.1 we can derive the conditional probability of a product on the rest of the order for which we use the binary logit functional form:

$$p\left(y_{i,j} \mid y_{i,-j}\right) = 1/\left(1 + \exp\left(-\left(\boldsymbol{a_j} + \sum_{l \neq j} V_{jl} y_{il}\right)\right)\right) \text{ for all } i, j \tag{4.3}$$

where $y_{ij} = 1$ if category j is in basket i, conditional on the other categories ,$-j$, in the baskets. To derive the optimal parameters, $\boldsymbol{V}$ and $\boldsymbol{a}$, for the MVL model we use maximum likelihood estimation (ML). This however is computationally not possible considering equation 4.1 when using the log-likelihood as the number of parameters to optimize are too large. Therefore we estimate the log-pseudo likelihood function. By using the maximum likelihood on the log-pseudo likelihood we are able to obtain $\boldsymbol{V}$ and $\boldsymbol{a}$ that best describe the data. The log pseudo-likelihood value is derived from the conditional probability in equation 4.3 by maximizing the function using the parameters in V and a:

$$LPL = \sum_i \mathrm{LPL}_i \tag{4.4}$$

with,

$$LPL_i \sum_j \left[Y_{ij} \log\left(p\left(y_{ij} \mid \boldsymbol{y}_{i,-j}\right)\right) + (1 - Y_{ij})\left(\log\left(1 - p\left(y_{ij} \mid \boldsymbol{y}_{i,-j}\right)\right)\right)\right] \tag{4.5}$$

where the LPL_i denotes the log-pseudo likelihood calculated for every basket i.

The RBM model captures dependencies between products through hidden nodes. The visible nodes are connected to all hidden nodes and have a weight for every connection. The weights than determine the consequences of the state on the probability of a product being bought. The vector $h = \{0,1\}^K$ are possible configuration of the hidden layer with

$h_{i,k} = 1, ..., K$. Here $h_{i,k} = 1$ if hidden node k is in order i is equal to 1, and $h_{i,k} = 0$ otherwise. The RBM model relies on the joint probability:

$$p(y_i, h_i) = \frac{exp(\boldsymbol{b'y_i} + \boldsymbol{h'_i W y'_i})}{Z_{RBM}} \tag{4.6}$$

with,

$$Z_{RBM} = \sum_{y \in \{0,1\}^J} \sum_{h \in \{0,1\}^K} exp(\boldsymbol{b'y} + \boldsymbol{h'Wy}) \tag{4.7}$$

where Z_{RBM} is the sum over all possible market baskets and vector b the bias vector. The bias vector is a constant for the category, where this vector is determined by the univariate frequency of the product. The weights between the hidden and visible layer contain information about the correlation between categories in a given state. If $W_{j,k}$ and $W_{l,k}$ have the same sign then when state $h_{i,k}$ is active the categories j and l have a positive correlation in market basket i, whenever categories j and l have opposite signs the correlation is negative.

The independence model does not capture any cross-effects. Therefore we set all weights in the independence model to 0. The coefficients for the independence model using ML are:

$$b_j = log(\sum_i y_{i,j}) - log(I - \sum_i y_{i,j}). \tag{4.8}$$

With normalization function:

$$Z_0 = 2^K \prod_j (1 + exp(b_j)) \tag{4.9}$$

where 2^K are all possible market baskets.

To calculate the probabilities of basket i from 4.6 when we do use the hidden variables, we sum over all possible combinations of $h \in \{0,1\}^K$ by using formula:

$$p\left(\boldsymbol{y}_i\right) = \frac{p^*\left(\boldsymbol{y}_i\right)}{Z_{RBM}} \text{ with } p^*\left(\boldsymbol{y}_i\right) = \sum_{\boldsymbol{h} \in \{0,1\}^K} \exp\left(\boldsymbol{b'y}_i + \boldsymbol{h'Wy}_i\right) \tag{4.10}$$

where $p^*(\boldsymbol{y_i})$ is the unnormalized probability.

The conditional probabilities can be derived from Equation 4.6 . For an RBM we make a distinction between a forward pass and a backward pass. The forward pass is where we sample the hidden nodes h_i conditional on the visible nodes y_i in basket i. This results in the conditional probability of the hidden nodes h_i:

$$p\left(h_{ik} \mid \boldsymbol{y}_i\right) = 1/\left(1 + \exp\left(-\sum_j W_{kj} y_{ij}\right)\right) \quad \text{for all } i, k. \tag{4.11}$$

A backward pass is where we sample y_i on h_iThe conditional probability of y_{ij} given the hidden nodes is as follows:

$$p\left(y_{ij} \mid \boldsymbol{h}_i\right) = 1/\left(1 + \exp\left(-\left(b_j + \sum_k W_{kj} h_{ik}\right)\right)\right) \text{ for all } i, j. \tag{4.12}$$

To determine the values for parameters in $\boldsymbol{W}$ and $\boldsymbol{b}$ we use ML of the log-likelihood function (LL). By maximizing the LL in equation 3 we get the weights that give the highest probabilities of encountering the 10,000 baskets in the data. Hruschka (2014) is the first paper to use ML estimation as approximation method of the weights. For RBM's contrastive divergence algorithm is often used. The method depends on the passes through layers given in Equations 4.11 4.12. To obtain the paramaters from ML of the log-likelihood function we use the formula:

$$\mathrm{LL} = \sum_i \log p\left(\boldsymbol{y}_i\right) = \sum_i \log p^*\left(\boldsymbol{y}_i\right) - I \log(Z_{RBM}), \tag{4.13}$$

where we sum over the log probability of all possible market baskets.

After obtaining the estimations for coefficients, choosing a good evaluation method of the models remains. As we have used two different methods to obtain coefficient estimations we can not use the comparison of the likelihood values to make any conclusion. To compare the accuracy of the models we analyse the prediction accuracy of the model based on the absolute deviation(AD). To get the AD statistic we measure the ability of the models to reconstruct the baskets. The hit rate determines the quality of the reconstruction. The statistic is calculated using the following formula:

$$\mathrm{AD} = \frac{1}{500} \sum_{s=1}^{500} \sum_i \sum_j \left| y_{ij} - \hat{y}_{ijs} \right| \tag{4.14}$$

where 500 is the number of artificial datasets and y_{ijs} is the sth sampled purchase incidence value of category j in basket i. These artificial datasets are created through Gibbs sampling Hruschka (2014). This is done by drawing samples on the conditional probability from equations 4.11 and 4.12 and then convert the probabilities to binary values through Bernoulli sampling. The sampling procedure consists of a forward and a backward pass. In the forward pass, the hidden nodes are sampled based on the conditional probability in equation 4.11, here a market basket in the data is given as input for the visible nodes. In the backward pass, the market basket given as input is recreated by using the conditional probability in equation 4.12 from the sampled hidden nodes. The recreated market baskets are called artificial orders in this paper.

The AD statistic will give a good indication on whether the cross-effects are accurate. To derive the cross-effects we use these artificial datasets. To calculate these marginal cross-effects we multiply the weights of the visible nodes to one of the hidden node, we do this for all hidden nodes. The formula will look as follows:

$$\begin{aligned} \frac{\partial \langle y_j \rangle}{\partial \langle y_l \rangle} &= \langle y_j \rangle \left(1 - \langle y_j \rangle\right) \sum_k W_{kj} \frac{\partial \langle h_k \rangle}{\partial \langle y_l \rangle} \\ &= \langle y_j \rangle \left(1 - \langle y_j \rangle\right) \sum_k W_{kj} W_{kl} \langle h_k \rangle \left(1 - \langle h_k \rangle\right) \end{aligned} \tag{4.15}$$

where $\langle y_j \rangle$ and $\langle h_k \rangle$ are the proportions for which $y_j = 1$ and $h_k = 1$. For $\langle y_j \rangle$ we use the average marginal univariate frequency of a visible node j of the 500 artificial datasets. The same is done for $\langle h_k \rangle$ where we derive the average marginal univariate frequency of the hidden nodes on the 500 artificial datasets. The marginal cross effects are positive if $W_{kj} W_{kl} > 0$ for each k, and are negative vice versa.

4.2 Results

The performance of the estimated model is measured on a dataset that is split into two sets of equal size of 10,000 orders, the estimation set and the validation set. The weights are determined by the ML on the estimation set. The LL statistics are shown in Table 3

Number of hidden variables	Number of parameters	Log-likelihood (LL)		Bayesian Information Criterion (BIC)
		Estimation Data	Validation Data	
Independence	60	-185,089	-185,473	371,499
1	120	-183,257	-183,686	368,477
2	180	-181,725	-182,166	365,990
3	240	-180,798	-181,257	364,724
4	300	-179,723	-180,273	363,309
5	360	-178,940	-179,606	362,528

Table 3. Log-likelihoods rounded to the nearest integer number

As can be seen in Table 3 the more hidden variables are included the higher the log-likelihood. This would mean that indeed cross-effects play a role for customers buying products in a supermarket. For the MVL the number of parameters that are estimated is 3,600, significantly more than the RBM models. To compare the performance of the RBM model compared to the MVL model we look at the AD statistics calculated using equation 4.14. The MVL has an AD of 94,977 compared to a statistic of 96,372 for the RBM model with 5 hidden variables. For this AD statistic, the weights of the highest log-likelihood out of 50 optimisations are used. For the initialisation of the weights, we start with an initialisation where the weights are drawn from a normal distribution around 0 with a 0.5 standard deviation. For the initialisation of the bias vector, we use the obtained vector from the independence model. The MVL performs better than the RBM model, however this relatively small difference proves that the RBM model is a suitable alternative method for analysing cross-effects in high dimensions.

When optimising the RBM model we see that the optimisation is not robust for different initialisations. Therefore we use 50 random initialisation and use the 10 best optimisations based on the log-likelihood. We calculate the cross-effects of these 10 best optimisations using the obtained parameters. As the optimisations are dependent on the initialisation we try to remove any large outlying cross-effect of the data. By removing the outliers the obtained average cross-effects will not be influenced by the initialisation. Identifying outliers based on the Z-score is in this case insufficient, as the average and standard deviation are influenced too much by the presence of the outliers. Therefore we use the interquartile range for the cross-effects to remove the outliers. All the cross-effects between the first and third quartile of the obtained results will be included in calculating the average. The average cross-effects of these optimisations are shown in Table 10. This table shows the largest cross-effects according to the RBM with 5 hidden variables. We chose 5 hidden variables as this is the best model based on the BIC statistic in 3, this statistic measures the quality of the model.

Fresh Herbs, Fresh Vegetables	3.112	Poultry, Fresh Vegetables	2.783
Poultry Counter, Fresh Herbs	2.262	Canned Meals Beans, Fresh Vegetables	2,174
Fresh Herbs, Packaged Fruits Vegetables	2.114	Canned Meals Beans, Fresh Herbs	1.894
Fresh Herbs, Packaged Fruits, Vegetables	1.072	Pasta Sauce, Fresh Vegetables	1.538
Soft Drinks, Fresh Vegetables	-1.445	Canned Jarred Vegetables, Fresh Vegetables	1.409
Canned Jarred Vegetables, Fresh Herbs	1.394	Packaged Poultry, Fresh Vegetables	1.389
Fresh Herbs, Canned Meals Beans	1.332	Poultry Counter, Canned Meals Beans	1.301
Pasta Sauce, Fresh Herbs	1,213	Packaged Poultry, Fresh Herbs	1.208
Canned Meals Beans, Dry Pasta	1.204	Dry Pasta, Canned Meals Beans	1.171
Poultry Counter, Canned Jarred Vegetables	1.131	Popcorn Jerky, Fresh Vegetables	-1.105
Dry Pasta, Fresh Herbs	1.103	Fresh Vegetables, Fresh Herbs	1.082
Fresh Herbs, Canned Jarred Vegetables	1.079	Preserved Dips Spreads, Dry Pasta	1.078
Soft Drinks, Yoghurt	-1.064	Canned Jarred Vegetables, Fresh Dips Tapenades	1.062
Cookies Cakes, Fresh Vegetables	-1.024	Poultry Counter, Pasta Sauce	1.013
Poultry Counter, Packaged Fruits Vegetables	1.000	Water Seltzer Sparkling Water, Fresh Vegetables	-0.996

Table 4. 30 largest cross-effects of RBM using 5 hidden nodes

In Table 4 we see several unhealthy products having negative cross-effects on healthy products. To look further into this we show in table 5 the 12 largest negative cross-effects.

Soft Drinks, Fresh Vegetables	-1.445	Popcorn Jerky, Fresh Vegetables	-1.105
Soft Drinks, Yoghurt	-1.064	Cookies Cakes, Fresh Vegetables	-1.024
Water Seltzer Sparkling Water, Fresh Vegetables	-0.996	Dry Pasta, Fresh Fruits	-0.924
Candy Chocolate, Fresh Vegetables	-0.886	Packaged Produce, Fresh Vegetables	-0.876
Spices Seasonings, Fresh Fruits	-0.846	Paper Goods, Fresh Vegetables	-0.814
Soft Drinks, Packaged Fruits, Vegetables	-0.778	Energy Granola Bars, Fresh Vegetables	-0.731

Table 5. 12 largest negative cross-effects of RBM using 5 hidden nodes

We see in Table 5 that soft drinks, cookies caked, candy chocolate, and energy granola bars all have negative effects on the probability of buying fresh vegetables. This shows that products containing a lot of added sugar tend to have negative cross-effects with vegetables.

5 Extension

To extend the analysis of cross-effects we make use of BERT, a machine-learning technique used for natural language processing. To our knowledge, we are the first to use BERT to find cross-effects.

BERT understands language as the model is pre-trained on a large number of texts from books and Wikipedia. By making use of Masked Language Modelling (MLM) and Next Sentence Predictions (NSP) the model trains itself in understanding word correlations and sentence structures. For this extension, we focus on MLM as this is a method that can be implemented in the same way for understanding cross-effects between market categories. We call this model CrossBERT. By building CrossBERT we can lay the foundation of the BERT model for cross-effects. Through fine-tuning, which means post-training the CrossBERT model on a smaller dataset, the model can be applied for different situations as we assume that cross-effects for different supermarkets are never the same but are in general fairly

similar. Through fine-tuning, the model can learn aspects of the market basket structure that are more specific to the setting of the supermarket.

BERT contains 12 hidden layers in its neural network compared to only 1 hidden layer for an RBM. As every hidden layer contains nodes that capture features of the data the complexity of the model increases notably. Both a BERT MLM model and an RBM model try to reconstruct the input as good as possible through their layers. In essence, both models use the input and the weights between the layers to reconstruct the input, and if the reconstruction is poor the weights will be changed according to the loss function. This loss function determines the quality of the reconstruction. In this paper, the weights of the RBM model are not determined by passes through the network but by the maximum likelihood. Where the BERT model could outperform the RBM model is through its complexity, as mentioned before the BERT model has 12 layers, all these layers contain weights that can capture features of the data.

To extract cross-effects the model should understand a part of customer behaviour in supermarkets, this is done using the attention mechanism, introduced in Vaswani . (2017). The attention mechanism enables the model to incorporate the context in the analysis of the specified task. By incorporating the context, which is the products in the basket, the model can learn the correlations between products. CrossBERT can be trained to understand that when a customer buys burgers and tomatoes, the probability that hamburger buns will be bought as well is higher. This can go into great detail because of the many features of customer behaviour the 12 layers of BERT can capture. The complexity of the BERT model allows the model to understand multiple combinations of products, this means that the model could find the effects of a combination of two products on a dependent product. In this section, we limit the model to analysing cross-effects to evaluate the performance of this single task compared to the other models.

5.1 Methodology

As mentioned in this paper we focus on training CrossBERT on MLM. The training process works by covering an arbitrary word in a sentence by '[MASK]' and predicting the masked word. Based on the quality of this prediction the model can learn from it. The training process is unsupervised. As BERT is context-dependent it will use all words in the context of the '[MASK]' to make this prediction. A similar training strategy can be used to let BERT learn basket compositions. Arbitrarily one of the items of a basket will be covered with '[MASK]' and BERT will predict the masked category using a score per category.

To create a BERT model, the weights in the network need to be updated according to a loss function of a prediction. For CrossBERT we update the weights based on the prediction of the masked category by using the cross-entropy loss function. The weights in the neural network of BERT will be adjusted such that the cross-entropy loss function is minimized. The cross-entropy loss function is calculated as follows:

$$CE = -\sum_{j=1}^{L} \boldsymbol{y_j} log(f(z_j)) \tag{5.1}$$

where C is the set of the possible categories, and $y_j \in \{0, 1\}$ is an element of the vector for the next category in the basket, this vector contains only one element $y_j = 1$. $f(z_j)$ denotes the score for category j and is calculated using the function in equation 5.2. The last layer in the neural network is a layer with a softmax function that calculates the predicted relative

probability of every possible category being the masked category given the other products that are in the basket. The softmax function will convert vectors to probabilities and is calculated as follows:

$$f(z_j) = \frac{e^{z_j}}{\sum_{j=1}^{L} e^{z_j}} \tag{5.2}$$

with z_j being the vector for category j divided by the sum of the vectors for all categories L.

For measuring the performance of the BERT model and compare the performance with the MVL and RBM model on the market baskets we make use of the same evaluation criterion as mentioned in equation 4.14. To create the artificial baskets we use the probability function of the order on a masked token given in Equation 5.2. Here the z_i will be the vector created from the network and is thus conditional on the rest of the market basket. Note that the model uses the masked token predictions, this implies that it measures the conditional probability only on one specific token. Therefore we can only derive the conditional probability specific to position m.

We then apply Bernoulli sampling to transform the probabilities to binary numbers of either in the basket, $y_{ij} = 1$, or not in the basket, $y_{ij} = 0$. Similar to Wang Cho (2019) where they use Gibbs sampling to recreate a dataset, however in our case an artificial sentence starts by masking out an entire sentence. When recreating an order we mask out only the predicted product. This way we sample based on the conditional probability on the rest of the market basket, through the attention mechanism. The product of which we calculate the probability of being the '[MASK]' is left out of the market basket. This is because CrossBERT is trained on only observing either 0 or 1 for the presence of a category in a market basket. For example, whenever an order contains fresh vegetables the model will never predict fresh vegetables in the artificial dataset as the conditional probability of encountering another fresh vegetables product is 0. To remove this bias we sample on the conditional probability of all other categories in the market basket.

When recreating the market baskets, we can not use a fixed size for the artificial basket. This would cause biased results compared to the MVL and RBM models. Therefore we must consider longer and shorter orders. To determine the size of an artificial basket we draw from the distribution of the order lengths. This process is done to create 500 artificial baskets to measure the AD in equation 4.14. As can be seen in Figure 1 the order sizes distribution is right-skewed.

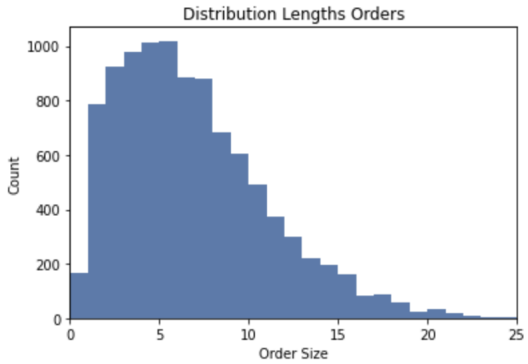

Figure 1. Histogram containing the order sizes of the training set

To obtain the cross-effects 25 random datasets are created of a length drawn from the distribution in Figure 1 by picking random categories. Half of these orders contain independent category k and the other half do not. All orders will have a [MASK] token in an arbitrary position. From the softmax function in equation 5.2 we then derive the probability of the '[MASK]' token being the target category $y_{i,m,j}$ given the rest of the random order. We measure the difference of $y_{i,m,j}$ between an order with the independent category, $y_{i,-m,k} = 1$, and an order without the independent category, $y_{i,-m,k} = 0$. Here $-m$ refers to any of the other positions than the masked token. We repeat this 25 times and average the effects to remove any possible influence of the positions of the target and independent categories on the probability. The obtained cross-effects can be interpreted as the average cross-effects of independent category k on dependent category j for a position m in the order. This is calculated from 25 random datasets by using:

$$\frac{\partial \langle y_{m,j} \rangle}{\partial \langle y_{-m,k} \rangle} = \frac{1}{25} \sum_{i=1}^{25} p(y_{i,m,j}|y_{i,-m,k} = 1) - p(y_{i,m,j}|y_{i,-m,k} = 0) \tag{5.3}$$

here we denote $y_{m,j}$ as the average cross-effect. We look at the difference in probability of $y_{i,m,j}$ when the independent category is not in the order, $y_{i,-m,k} = 0$, and the independent category is in the order, $y_{i,-m,k} = 1$.

We also derive the percentage increase of the probability. The average percentage change of the probability of dependent category j is in basket i when category k is basket i is calculated as follows:

$$\%\text{ change}_{k,j} = \frac{\sum_{i=1}^{25} p(y_{i,m,j}=1|y_{i,-m,-j},y_{i,k}=1) - p(y_{i,m,j}=1|y_{i,-m,-j},y_{i,-m,k}=0)}{\sum_{i=1}^{25} p(y_{i,m,j}=1|y_{i,-m,-j},y_{i,-m,k}=0)} \times 100\% \tag{5.4}$$

where $y_{i,-m,-j}$ are all other categories besides the independent and dependent categories on all remaining positions $-m$. In equation 5.3 we derive the percentage point difference whereas in equation 5.4 we derive the percentage change. We provide the cross-effects in two ways because in Equation 5.4 the cross-effects are not dominated by categories with a high marginal frequency. In equation 5.3 small correlations between products with both high marginal univariate frequencies could have cross-effects that exceed the cross-effects of large correlations between products with only small univariate marginal frequencies. This way no cross-effects get overlooked containing relevant marketing information.

5.2 Preprocessing Data

Since pre-trained BERT is trained in analysing texts, the new model should be trained from scratch to understand the new domain of supermarket baskets. The BERT base model developed by Google is trained on next-sentence predictions (NSP) besides MLM. Because market baskets are considered independent for different customers this training task is dropped from the training. The training dataset contains only orders from different customers to avoid any bias that could occur when the training set contains multiple orders of the same customer. We can not assume that these orders are independent. This training set consists of 30,000 orders of the Instacart dataset. The training dataset will be critical for the performance of the model as overfitting and underfitting for the training data are common in machine learning. Therefore, models with different epochs and masking probability are considered.

To preprocess the data for MLM, first, the dataset needs to be tokenized. A BERT model reads every line from the data as tokens. A token is an element that the model has in

its vocabulary. The created vocabulary of CrossBERT contains all of the 60 most frequent categories. Special tokens are added to the training data and are used to illustrate the beginning of a market basket, [CLS], the end of a market basket, [SEP], and the masked product, [MASK].

Second, the number of epochs needs to be determined as well as the masking probability. An epoch is one iteration of training through the data. The masking probability determines the percentage of tokens in the dataset that are being masked. These parameters of the model are determined through the loss function on the validation set. The validation set consists of 10,000 orders from all different customers.

In contrast to an RBM, BERT takes into account the order in which the products are bought, as mentioned before. The model can thus calculate the cross-effects specific to the position of the product in the order. This analysis could be relevant as the order in which products are added to the basket is also dependent on the aisle in the supermarket. Since the MVL and RBM models do not cover the analysis of this feature of the model, we try to remove this feature from the training data. The data is copied and shuffled 6 times which avoids that the CrossBERT model trains on the order of products in the basket. The model is therefore trained on a dataset of 30,000 orders which are shuffled 6 times. The training set will thus contain 180,000 orders. As we use binary variables we can not calculate the effects of repeating categories. CrossBERT could be able to find the probability of buying multiple products when buying at least one, $P(y_{ij} \geq 2 | y_{ij} \geq 1)$, when trained accordingly. This could be interesting for a supermarket for marketing measures such as 'buy one get one free'. To compare the model with MVL and RBM to evaluate its accuracy, we limit the training to binary variables. We thus remove all multiples of categories in an order.

5.3 Results

To develop CrossBERT, first, the model needs to be trained such that it performs sufficiently well in predicting masked tokens by using cross-effects. Different training samples and the number of epochs are considered to provide a model that can accurately describe the full dataset. Table 6 includes the training loss of using a different number of epochs and the validation loss.

	4 epochs	6 epochs	8 epochs	10 epochs
Training Loss	3.177	3.151	3.110	**3.163**
Validation Loss	3.146	3.153	**3.145**	3.163

Table 6. Training loss using batch-size 64 and 0.15 masking probability, with a training set of 30.000 orders and a validation set of 10.000 orders

From Table 6 we see that the loss function is decreasing when training with more epochs. This suggests that the model is learning the cross-effects of products as it better predicts the masked tokens. After training the model the validation accuracy is the highest when performing 8 epochs. As can be seen in Table 6, 10 epochs have a higher training accuracy but a lower validation accuracy. This proves that overfitting on the training data occurs. To create the best model, the masking probability needs to be taken into account. A masking probability that is too large could lead to missing relevant bivariate frequencies in the training data and a low masking probability could lead to a model that is underfitting

the training data. Therefore we consider the masking probability for a model with 8 epochs with different masking probabilities.

	Masked Probability			
	0.10	0.15	0.20	0.25
Validation Loss	3.165	**3.145**	3.234	3.338

Table 7. Validation Loss with different masking probabilities for a model trained with 8 epochs

From Table 6 and Table 7 we conclude that a CrossBERT model trained on 8 epochs of a dataset containing 30,000 orders with a masking probability of 0.15% performs best on the validation set of 10,000 orders. Therefore we continue with this model to build artificial datasets from which we derive the AD statistic.

Model	AD statistic
Multivariate Logit	94,977
Restricted Boltzmann Machine	96,372
CrossBERT	**94,814**

Table 8. AD statistic

From Table 8 we conclude that the CrossBERT model is able to outperform the MVL and RBM model based on the absolute deviation of the artificial dataset. The CrossBERT model is therefore better able to calculate the marginal cross effects. For these cross-effects between the categories, we make a distinction between two measurements. One is the percentage change of the probability of the 30 pairs that change the most percentage-wise, shown in Table 10. The second is the marginal change in probability, where Table 9 contains the 30 categories with the 30 highest cross-effects.

Specialty Cheeses, Fresh fruits	-0.095	Packaged Fruits Vegetables, Fresh Fruits	0.085
Fresh Vegetables, Fresh Fruits	0.078	Energy Granola Bars, Fresh Vegetables	-0.074
Fresh Fruits, Fresh Vegetables	0.068	Canned Jarred Vegetables, Fresh Vegetables	0.064
Poultry Counter, Fresh Vegetables	0.064	Packaged Produce, Fresh Fruits	0.058
Baby Food Formula, Fresh Fruits	0.051	Packaged Fruits Vegetables, Fresh Vegetables	0.049
Fresh Vegetables, Packaged Fruits Vegetables	0.044	Yoghurt, Fresh Fruits	0.044
Preserved Dips Spreads, Chips Pretzels	0.040	Fresh Fruits, Packaged Fruits Vegetables	0.039
Fresh Dips Tapenades, Chips Pretzels	0.039	Soft Drinks, Fresh Vegetables	-0.038
Milk, Fresh Vegetables	0.035	Popcorn jerky, Packaged Vegetables Fruits	-0.033
Fresh Vegetables, Soft Drinks	-0.033	Grains Rice Dried Goods, Fresh Vegetables	0.032
Baby Food Formula, Yoghurt	0.031	Packaged Cheese, Cereal	-0.031
Ice Cream Ice, Fresh Vegetables	-0.030	Refrigerated, Fresh Fruits	0.030
Paper goods, Fresh Fruits	-0.028	Canned Jarred Vegetables	-0.027
Pasta Sauce, Fresh Fruits	-0.027	Frozen Produce, Fresh Vegetables	0.027
Fruit Vegetable Snacks, Chips Pretzels	0.027	Condiments, Fresh Fruits	-0.026

Table 9. Marginal probability increase of independent variable (1st) on dependent variable (2nd) from CrossBERT

Dry Pasta, Pasta Sauce	347.6%	Fruit Vegetable Snacks, Energy Granola Bars	208.2%
Hot dogs Bacon Sausage, Buns Rolls	207.8%	Canned Jarred Beans, Canned Jarred Vegetables	170.9%
Canned Jarred Vegetables, Canned Meals Beans	163.9%	Pasta Sauce, Dry Pasta	163.6%
Chips Pretzels, Fresh Dips Tapenades	162.5%	Spices Seasonings, Oils Vinegars	155.3%
Frozen Meals, Frozen Pizza	154.9%	Fresh Vegetables, Fresh Herbs	149.7%
Fresh Herbs, Fresh Vegetables	149.2%	Frozen Appetizers Sides, Frozen Pizza	139.7%
Frozen Meals, Frozen Pizza	132.1%	Soup Broth Bouillon, Canned Jarred Vegetables	129.1%
Frozen Appetizers Sides, Frozen Meals	127.8%	Other Creams Cheeses, Breakfast Bakery	121.9%
Butter, Baking Ingredients	115.4%	Oils Vinegars, Spices Seasonings	110.4%
Breakfast Bakery, Other Creams Cheeses	110.0%	Spices Seasonings, Baking Ingredient	108.9%
Chips Pretzels, Preserved Dips Spreads	106.3%	Preserved Dips Spreads, Chips Pretzels	104.9%
Baking Ingredients, Butter	104.4%	Condiments, Asian Foods	99.2%
Packaged Fruits Vegetables, Packaged Produce	96.0%	Canned Meals Beans, Preserved Dips Spreads	95.3%
Buns Rolls, Hot Dogs Bacon Sausage	93.1%	Fresh Vegetables, Packaged Vegetables Fruits	93.0%
Spices Seasonings, Pasta Sauce	92.5%	Frozen Meals, Frozen Breakfast	91.6%

Table 10. Probability increase (%) of independent variable (1st) on the dependent variable (2nd) from CrossBERT

From Table 9 one thing that stands out is the negative cross-effects between products considered healthy and products considered unhealthy, similarly to the obtained cross-effects using the RBM model. Soft drinks decrease the probability of buying fresh vegetables by 3.3 percentage point. Variables such as age, time of the day, day of the week, price, and gender are perhaps causal factors for the choice between healthy and unhealthy products. Also the occasion wherefore the products are bought can highly influence the choice for healthy or unhealthy foods and drinks. For example, a customer may buy fresh vegetables for dinner on Tuesday and buy soft drinks on Friday for a party. The cross-effects show that both products are rarely bought together.

Table 10 shows the percentage increase of the probability that the independent category has on the dependent category. For example whenever dry pasta is in the basket the probability of pasta sauce also being in that basket increases by 348%. Many of the high cross-effects in Table 10 are because the products are part of a common recipe. Hot dogs for instance are normally eaten with buns, explaining the high increase in the probability of purchasing buns whenever a customer buys hot dogs. Another explanation could be the location in the store. From Table 10 we see that canned jarred vegetables and canned meals beans are highly correlated as well as frozen products. Many stores keep the canned products in the same vicinity as well, similar to frozen products. Whenever a customer buys one of the products the customer could be intrigued by other products which catch their eyes. This is where shelve placement in the supermarket can contribute to effective marketing measures. An online supermarket such as Instacart could use the same layout but online instead. Moreover, we see that in Table 10 there are no negative percentage changes in the highest 30 cross-effects based on absolute value, meaning that the complementary goods dominate the cross-effects of categories. This could be attributed to the fact that many substitution goods fall in the same category because these products is often partly identical.

6 Conclusion

In this paper, we have developed CrossBERT, a model capable of extracting cross-effects from market basket data. The results in Table 8 show that CrossBERT can outperform the RBM and MVL models and is therefore a feasible and good alternative for analysing

the cross-effects for multiple categories. From the results in Table 9 we see that multiple healthy categories have negative cross-effects on unhealthy categories and vice versa. This implies that marketing measures in one of the two categories could not or negatively impact the sales of the other. Moreover, in Table 10, we focus on the cross-effect of products with a lower univariate marginal frequency, by using the percentage change. Here we see that common recipes are one of the drivers of the cross-effects. Chips pretzels has a positive cross-effect on fresh dips tapenades. We also see that dry pasta has a large positive cross-effect on pasta sauce. These common recipes can be a good target for marketing measures. In Russell Kamakura (1997), it is shown that promotions will lead to a higher number of sales in the positively related categories of the promoted category. The promotion of pasta sauce will have positive effects on the sales of pasta and vice versa. The same applies to all other positive cross-effects in Table 10.

Moreover, it stands out that the frozen products have high positive cross-effects in Table 10. The positive cross-effects of frozen appetizers and frozen pizza will probably not be caused by the complementary nature of these categories, as appetizers are not commonly eaten before a pizza. Instead, the driver of the positive cross-effects seems to be the in-store location of these categories or for online supermarkets the location on the webpage. Frozen products are mostly located at the same aisles, as it is more sustainable to place the freezers at the same place rather than spread around the store. By concentrating the freezers less power is needed to keep the products frozen. This affects the behaviour of customers as in Bezawada . (2009) it is proven that locating the aisles closer to each other, increases the cross-effects between the products. Similarly, the relation between aisle location and cross-effects can describe the high positive cross-effects of the canned products as well. To limit the search costs of a customer, online supermarkets may choose a similar location layout on the web as in a physical supermarket. This can explain that these effects are shown also in the Instacart dataset.

To conclude, in this paper we created CrossBERT to analyse the cross-effects of categories in a supermarket. We have shown that CrossBERT better recreates the dataset than the MVL and RBM models by using its understanding of basket compositions. This proves that CrossBERT is therefore a computationally feasible alternative to analyse cross-effects of many categories. The CrossBERT model provides supermarket owners the necessary tool to apply a detailed analysis of the cross-effects that enables them to make informed marketing decisions.

7 Discussion

CrossBERT consists of a 12-layer neural network, which is useful because it is able to extract complex patterns from data. These complex patterns contain many features which can be stored in the weights of these 12 layers. However because the model is built to understand human language, which is more complex than the marginal cross-effects, the CrossBERT model could be improved by reconsidering the number of layers that are used. Future studies could point out that fewer layers decrease the number of computations while still obtaining accurate cross-effects of the products. Furthermore, the positional value of the CrossBERT model is removed from the training by shuffling the datasets several times. However, for a new model removing the positional encoding that is passed to the network could provide a better method for dropping this task from the training.

The power of the CrossBERT model lies in the potential of the model to be used in far

more detail. CrossBERT could be trained to find patterns in product purchase data and would possibly be a suitable model to analyse cross-effects for products instead of categories. Analysing the cross-effects on a product level would arguably show more negative cross-effects because many categories consist of substitution goods. The category soft drinks for example consists of many competitive brands of which the drinks are rarely bought together because of customer preferences.

Moreover, the CrossBERT model could have a lot of potential because of the possibility of fine-tuning. Fine-tuning is a term used for training the model after pre-training. When this is done on the data of the supermarket, the cross-effects of this specific supermarket can be obtained precisely. The model uses the pre-training phase in order to give the model an understanding of general correlation by training on a large dataset. The fine-tuning phase then adjusts the weights specifically to fit the customer behaviour data of the supermarket. The fine-tuning can be done on a relatively small dataset, which is useful for retail businesses having only a small dataset at hand. Future studies could also pre-train CrossBERT on multiple datasets to avoid any bias of the dataset of Instacart.

Another part of customer behaviour analysis where CrossBERT could be a good model is predicting the next basket of a customer. As mentioned before the market baskets of one customer are not necessarily independent. CrossBERT could potentially be trained to find patterns in customer purchases at the supermarket. To do this the next sentence prediction task BERT uses should not be dropped from the training tasks of the model.

References

bezawada2009crossBezawada, R., Balachander, S., Kannan, PK. Shankar, V. 2009. Cross-category effects of aisle and display placements: a spatial modeling approach and insights Cross-category effects of aisle and display placements: a spatial modeling approach and insights. Journal of Marketing73399–117.

boztuug2008modelingBoztuğ, Y. Hildebrandt, L. 2008. Modeling joint purchases with a multivariate MNL approach Modeling joint purchases with a multivariate mnl approach. Schmalenbach Business Review604400–422.

devlin2018bertDevlin, J. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding Bert: Pre-training of deep bidirectional transformers for language understanding.

felgate2015analyzingFelgate, M. Fearne, A. 2015. Analyzing the impact of supermarket promotions: a case study using Tesco Clubcard data in the UK Analyzing the impact of supermarket promotions: a case study using tesco clubcard data in the uk. The Sustainable Global Marketplace The sustainable global marketplace (471–475). Springer.

hruschka2014analyzingHruschka, H. 2014. Analyzing market baskets by restricted Boltzmann machines Analyzing market baskets by restricted boltzmann machines. OR spectrum

instacartInstacart. 2017. `https://www.kaggle.com/c/instacart-market-basket-analysis/overview/description`.

manchanda1999shoppingManchanda, P., Ansari, A. Gupta, S. 1999. The “shopping basket”: A model for multicategory purchase incidence decisions The “shopping basket”: A model for multicategory purchase incidence decisions.

russell1997modelingRussell, GJ. Kamakura, WA. 1997. Modeling multiple category brand preference with household basket data Modeling multiple category brand preference with household basket data. Journal of Retailing734439–461.

russell2000analysisRussell, GJ. Petersen, A. 2000. Analysis of cross category dependence in market basket selection Analysis of cross category dependence in market basket selection. Journal of Retailing763367–392.

vaswani2017attentionVaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN.Polosukhin, I. 2017. Attention is all you need Attention is all you need.

wang2019bertWang, A. Cho, K. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model Bert has a mouth, and it must speak: Bert as a markov random field language model. arXiv preprint arXiv:1902.04094.

Abbreviation	Definition
AD	absolute deviation
BERT	Bidirectional Encoder Representations from Transformers
LL	log-likelihood
LPL	log-pseudo likelihood
ML	maximum likelihood
MLM	masked language modelling
MVL	multivariate logit model
MVP	multivariate probit model
NLP	natural language processing
RBM	restricted Boltzmann machine