National Research University Higher School of Economics

Faculty of Computer Science

HSE and University of London Double
Degree Programme in Data Science and
Business Analytics

BACHELOR'S THESIS

Research Project

Optimization of client list for communication

Prepared by the student of Group БПАД181, Year 4,

Nechaeva Polina

Thesis Supervisor:

Senior Lecturer, SAS department, Titova Natalia Nikolaevna

Аннотация

Данная работа описывает полный путь построения маркетинговой кампании на основе данных о транзакциях интернет магазина. Работа включает построение сегментации клиентов, формирование наиболее выгодных предложений с использованием маркетинговых механик, оптимальное распределение клиентов по трем каналам коммуникации. Особое внимание уделено предсказанию будущих покупок клиентов. В рамках этой задачи сравнивается качество различных моделей отклика, а также их результативность после применения методов избыточной и недостаточной выборок для работы с несбалансированными выборками. Будущие покупки прогнозируются с помощью анализа рыночной корзины, а затем методом коллаборативной фильтрации, используемым обычно для разработки рекомендательных систем. Эффективность двух методов сравнивается по финансовым показателям итоговых кампаний.

Ключевые слова: электронная коммерция, финансовая оптимизация, сегментация, Анализ рыночной корзины, коллаборативная фильтрация

Annotation

This work describes a full path of developing a market campaign based on data of transactions of an online shop. It includes client segmentation, forming most profitable and appealing offers based on marketing mechanics and distributing communication channels over the target audience. Special attention is paid to predicting the clients’ next purchase. First, different response models are compared with each other and enhanced with methods for dealing with an imbalanced dataset. The future purchases are predicted by Market Basket Analysis and Collaborative Filtering approaches, and their performance is compared by calculating revenue after business optimization.

Key words: e-commerce, financial optimization, clusterization, market basket analysis, collaborative filtering

1. Introduction

In order to exist, all companies need clients to buy their products or services. To attract the customers, companies do different advertisements campaigns. The competition is very high, and the companies have to put a lot of resources into researching and implementing new developments in marketing. In the past years data mining has become a common technique for businesses to use to gain insights based on the data about customers which they can collect mostly from website activity. Companies have large databases with customers' information and can get more data about their shopping behavior and preferences. Good analytics can reveal patterns and turn them into strategies on how to expand business: increase conversion rate, customers' trust and optimize allocation of resources.

In this work we will focus on retaining the current client base. Since there is an enormous number of online stores, most clients have little loyalty for them, because they can always look for more appealing offers in other shops. However each business needs an established customer base, and to get it the companies have to regularly attract the attention of existing customers by providing them with appealing offers and discounts.

Currently there are many well known data mining techniques that can be applied in ecommerce, but there is no common way of how to choose and apply them. Each case should be worked through individually and it requires a lot of experimentation, time and method mixing to get the best result. In this work we will analyze the case of an online toy-store. First, we will preprocess the data. Then we will cluster our customers using 2 different models: K-Means and DBSCAN. Clusterization is used because different recommender models work better with separate segments in which behavioral patterns are more distinct and specific. For each segment we will conduct Market Basket Analysis. It is a technique that generates item combinations for offers based on associations it finds. Market Basket Analysis (MBA) will be compared with a recommendation generation approach called Collaborative Filtering (CF). It is often used in recommendation systems and in this work we want to test how well it will predict products for forming campaign offers. We can also use MBA discoveries to form offers after CF and combine these methods together. Another important step is creating response models. They will be used to predict the likelihood of customers' return. We will compare several models and their performance with oversampling and undersampling techniques because we are dealing with an imbalanced dataset. Finally, both recommended items and probability of return predictions will be used to create client ratings and choose the best clients for communication by formulating and solving Mixed Integer Linear Programming (MILP) problem. Finally, we will calculate the revenue of the campaign. It is necessary because apart from mathematical precision of the models we want to check if the formed campaigns can bring profit to the store.

1.1 Thesis structure

The work is structured as follows: first, we will briefly describe the data and its preprocessing for work and describe several visualizations. Next we will do data clusterization in chapter 4 followed by Market Basket Analysis in chapter 5 and Collaborative Filtering in chapter 6. Next we will discuss response models and their performance on our dataset in chapter 7. Finally, we will describe our approach to solving an optimization problem in chapter 8 and summarize the results.

2. Literature review

Many articles and papers are devoted to data mining in e-commerce. Akter and Wamba [1], mentions that in the e-commerce context, big data enables merchants to track each user's behavior and determine the most effective ways to convert one time customers into repeat buyers. Fan, Lau and Zhao [2], mention that key factors for strategic marketing decisions can be monitored by mining social media data and converted into meaningful insights using big data analytic technologies.

The book [3] describes some of the approaches we will use in this work like building clients' profiles, optimizing marketing expenses, building response models and forming personal recommendations.

Market Basket Analysis was first described by Aguinis et al. [4] It is also known as association rule mining.The main goal is to establish the relationship between items which exist in the market. The method is rather simple but effective and currently many modifications were proposed to enhance the method [5, 6].

Collaborative method is a common method for recommendation systems. Recommendation systems have a wide range of applications [7, 8], including recommending items in online stores, for example books, clothes or movies. It is a memory based method which works on a matrix of user-item interactions. It finds *k* closest users to the test user and then recommends items based on the aggregate of the ratings of these *k* users. Cosine similarity is often used as a measure of distance. In our work we will test the prediction quality of CF on toy shop data and compare its financial results with MBA results.

3. Main part

3.1 Problem statement

Let us have *n* users, each described with *k* features and *c* communication channels among which we want to distribute users. First, for each client we must model ratings $r_1, r_2, ..., r_c$ which will depict the profit level of communicating with client using channels $c_1, c_2, ..., c_c$.

Next we must solve problem of Mixed Integer Linear Programming (MILP):

$$max \sum_{1 \le i \le n;\ 1 \le j \le c} x_{ij} w_{ij}$$

subject to:

$$\sum_{1 \le j \le c} x_{ij} \le 1 \ for \ 1 \le i \le n$$

$$x_{ij} \in 0, 1 \ for \ 1 \le i \le n, 1 \le j \le c$$

$$r_{ij} \in \mathbb{R}$$

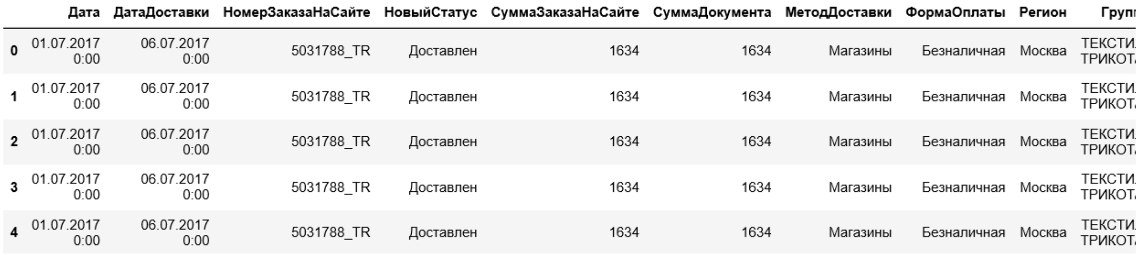
$$n, c \in \mathbb{N}$$

+ additional constraints are possible.

3.2 Data preprocessing

3.2.1 Data cleaning

The data which we will analyze comes from an online store. It includes different information about orders made at the period of 4 months from 07.2017 to 10.2017. There are approximately 1.67 million rows. Each row represents one item bought by one customer. Columns contain prices, marginal profit, date of order, city, payment and delivery types. There is not much information about the customers. Each customer is identified by a phone number; writing an email address is not mandatory so there are some missing values; there are also many incoherent values at the names column and we cannot predict gender accurately. We will mostly rely on purchase history.



	Дата	ДатаДоставки	НомерЗаказаНаСайте	НовыйСтатус	СуммаЗаказаНаСайте	СуммаДокумента	МетодДоставки	ФормаОплаты	Регион	Групп
0	01.07.2017 0:00	06.07.2017 0:00	5031788_TR	Доставлен	1634	1634	Магазины	Безналичная	Москва	ТЕКСТИ ТРИКОТ
1	01.07.2017 0:00	06.07.2017 0:00	5031788_TR	Доставлен	1634	1634	Магазины	Безналичная	Москва	ТЕКСТИ ТРИКОТ
2	01.07.2017 0:00	06.07.2017 0:00	5031788_TR	Доставлен	1634	1634	Магазины	Безналичная	Москва	ТЕКСТИ ТРИКОТ
3	01.07.2017 0:00	06.07.2017 0:00	5031788_TR	Доставлен	1634	1634	Магазины	Безналичная	Москва	ТЕКСТИ ТРИКОТ
4	01.07.2017 0:00	06.07.2017 0:00	5031788_TR	Доставлен	1634	1634	Магазины	Безналичная	Москва	ТЕКСТИ ТРИКОТ

Since the data is raw, we have to carefully preprocess it. The dataset was checked for duplicated or unexplainable rows. For example, there were 2 orders in which the clients paid nothing for the items since the total sum of an order was equal to the shipping price.
Each order has a separate line devoted for shipping. These lines were deleted because they do not affect the analysis.

The dataset was checked for missing values. We found rows without item margin values. It was decided to completely delete orders which include these items. Other missing values are explainable or there is no need to delete them. Not all of the items have a defined 'Группа4' category, but all of them have a non-empty 'Группа3' and 'Номенклатура' so we conclude that some groups do not have leaves in 'Группа4' category. 'ПричинаОтмены' is not compulsory to fulfill as well as 'Электронная_почта_new'. Since some of the clients did not specify an email, when distributing them over communication channels we will have to mark that we cannot contact them with the email channel. Some 'Регион' values are missing but we decided to keep this data and mark the location of such clients as unknown.

Столбец	Кол-во пропущенных значений	Столбец	Кол-во пропущенных значений
Дата	0	МетодДоставки	0
ДатаДоставки	0	ФормаОплаты	0
НомерЗаказаНаСайте	0	Регион	5 432
НовыйСтатус	0	Группа2	0
СуммаЗаказаНаСайте	0	Группа3	0
СуммаДокумента	0	Группа4	24 378

Тип	0	Отменено	0
Номенклатура	0	ПричинаОтмены	557 781
ТипТовара	0	Количество	0
Цена	0	ПВЗ_код	0
СуммаСтроки	0	Статус	0
ЦенаЗакупки	1797	Гео	0
Маржа	1797	СуммаУслуг	0
СуммаДоставки	0	НомерСтроки	0
КоличествоПроданоКлиенту	0	Электронная_почта_new	9 533
Телефон_new	0	Клиент	0
Городмагазина	5 432	МагазинЗаказа	566 999

Table 3.1 The number of missing values by column

For each order a buyout percentage was calculated and added. Based on the buyout percentage we added the revenue, margin and total number of items of the bought out orders.

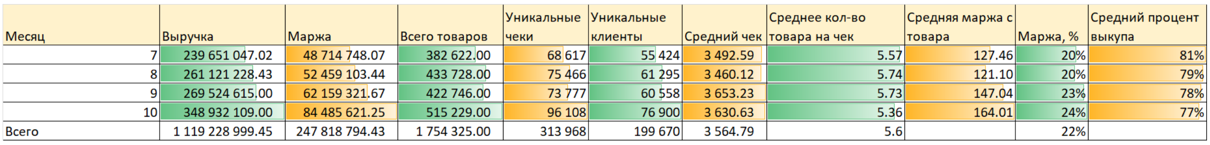
Minor changes were done to the delivery column. Initially there were 6 values and we grouped them into 4 delivery types. ‘PickPoint’ was merged with ‘Самовывоз’ option and ‘DPD’ was merged with ‘Транспортная компания’. Finally, we worked with the region information included in the dataset. It had 2 columns, in one of which there were only 3 values - ‘Москва’, ‘МО’ and ‘Регионы’, which is not informative for the analysis. On the other hand, the second column had too many values, which is also inconvenient to work with. It was decided to create a new column in which each order would belong to 1 of 6 parts of Russia: ‘Центр’, ‘Юг’, ‘Сибирь’, ‘Дальний Восток’, ‘Приволжье’, ‘Урал’ or marked unknown.

We still left some detalization since we created a separate column for average salary in the region where the order was made.[1] 4 salary intervals were formed: ‘<15. тыс’, ‘15-30 тыс.’, ‘30-45 тыс.’, ‘>45 тыс.’.

[1] The data on average salaries was collected by the thesis supervisor and the author in various internet sources.

3.2.2 Visual analytics

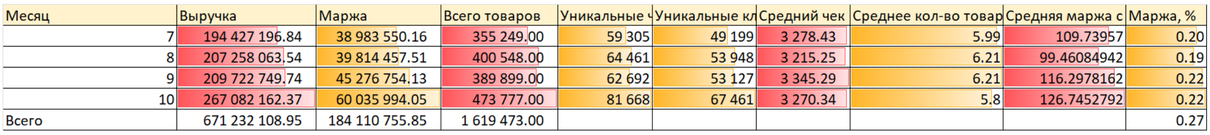
To gain a better understanding of the data different visualizations were made. In total we have transactions of 199 670 customers and 313 968 orders. On average 44 850 people put orders every month and 2 429 put orders every day. The data on the pic. 3.1 depicts all orders, including the ones processed and canceled. Picture 3.2 shows the data only on bought out orders.



Месяц	Выручка	Маржа	Всего товаров	Уникальные чеки	Уникальные клиенты	Средний чек	Среднее кол-во товара на чек	Средняя маржа с товара	Маржа, %	Средний процент выкупа
7	239 651 047.02	48 714 748.07	382 622.00	68 617	55 424	3 492.59	5.57	127.46	20%	81%
8	261 121 228.43	52 459 103.44	433 728.00	75 466	61 295	3 460.12	5.74	121.10	20%	79%
9	269 524 615.00	62 159 321.67	422 746.00	73 777	60 558	3 653.23	5.73	147.04	23%	78%
10	348 932 109.00	84 485 621.25	515 229.00	96 108	76 900	3 630.63	5.36	164.01	24%	77%
Всего	1 119 228 999.45	247 818 794.43	1 754 325.00	313 968	199 670	3 564.79	5.6		22%	

Pic. 3.1 Shop figures by month

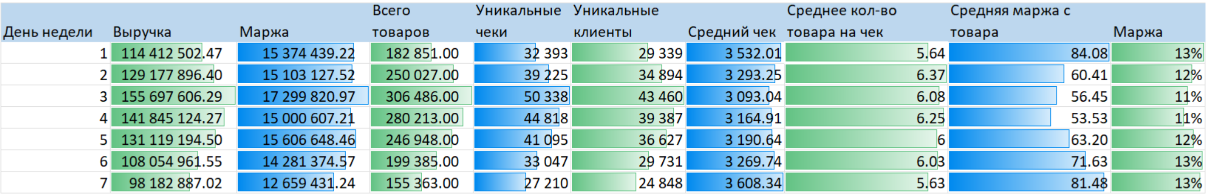
During the 4-month the revenue did not significantly fluctuate except October, when it increased approximately by 27%.



Месяц	Выручка	Маржа	Всего товаров	Уникальные ч	Уникальные кл	Средний чек	Среднее кол-во товар	Средняя маржа с	Маржа, %
7	194 427 196.84	38 983 550.16	355 249.00	59 305	49 199	3 278.43	5.99	109.73957	0.20
8	207 258 063.54	39 814 457.51	400 548.00	64 461	53 948	3 215.25	6.21	99.46084942	0.19
9	209 722 749.74	45 276 754.13	389 899.00	62 692	53 127	3 345.29	6.21	116.2978162	0.22
10	267 082 162.37	60 035 994.05	473 777.00	81 668	67 461	3 270.34	5.8	126.7452792	0.22
Всего	671 232 108.95	184 110 755.85	1 619 473.00						0.27

Pic. 3.2 Shop bought out figures by month

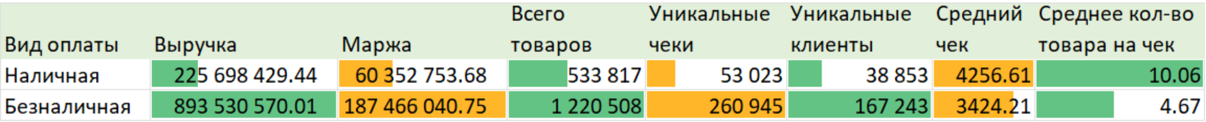
The daily fluctuations are also rather minor. More orders are made during the weekdays rather than weekends.



День недели	Выручка	Маржа	Всего товаров	Уникальные чеки	Уникальные клиенты	Средний чек	Среднее кол-во товара на чек	Средняя маржа с товара	Маржа
1	114 412 502.47	15 374 439.22	182 851.00	32 393	29 339	3 532.01	5.64	84.08	13%
2	129 177 896.40	15 103 127.52	250 027.00	39 225	34 894	3 293.25	6.37	60.41	12%
3	155 697 606.29	17 299 820.97	306 486.00	50 338	43 460	3 093.04	6.08	56.45	11%
4	141 845 124.27	15 000 607.21	280 213.00	44 818	39 387	3 164.91	6.25	53.53	11%
5	131 119 194.50	15 606 648.46	246 948.00	41 095	36 627	3 190.64	6	63.20	12%
6	108 054 961.55	14 281 374.57	199 385.00	33 047	29 731	3 269.74	6.03	71.63	13%
7	98 182 887.02	12 659 431.24	155 363.00	27 210	24 848	3 608.34	5.63	81.48	13%

Pic. 3.3 Shop figures by day of the week

A prevailing number of orders are paid for using cashless methods. The data shows that orders paid in cash have more than 2 times more items on average than cashless, but there is a possibility that such difference happened because of data collection errors.

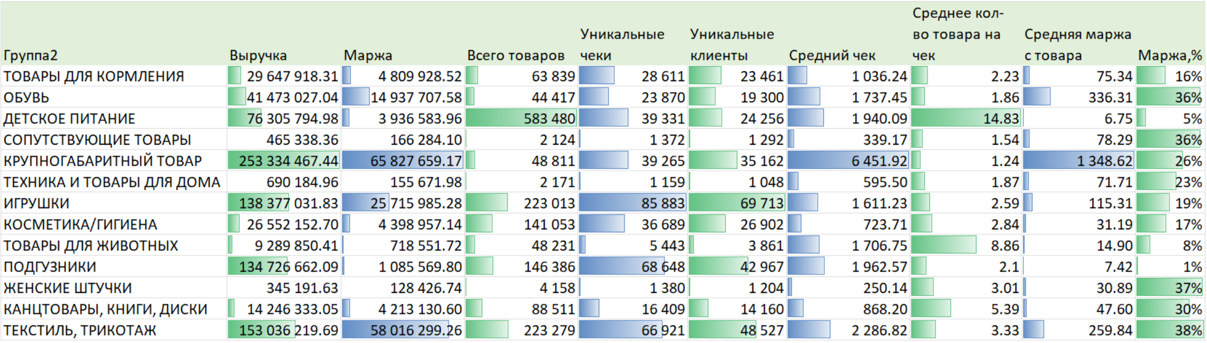


Вид оплаты	Выручка	Маржа	Всего товаров	Уникальные чеки	Уникальные клиенты	Средний чек	Среднее кол-во товара на чек
Наличная	225 698 429.44	60 352 753.68	533 817	53 023	38 853	4256.61	10.06
Безналичная	893 530 570.01	187 466 040.75	1 220 508	260 945	167 243	3424.21	4.67

Pic. 3.4 Shop figures by method of payment

The most remarkable differences can be observed between items of different categories. “КРУПНОГАБАРИТНЫЙ ТОВАР” items bring the most revenue to the shop, followed by ‘ИГРУШКИ’ and ‘ТЕКСТИЛЬ, ТРИКОТАЖ’. Heavy goods also bring the biggest margin which is slightly larger than the margin from textile goods, while toys have a considerable dip. If we want to promote toys in our offer we might choose the toys with higher price or margin to increase the average profit.

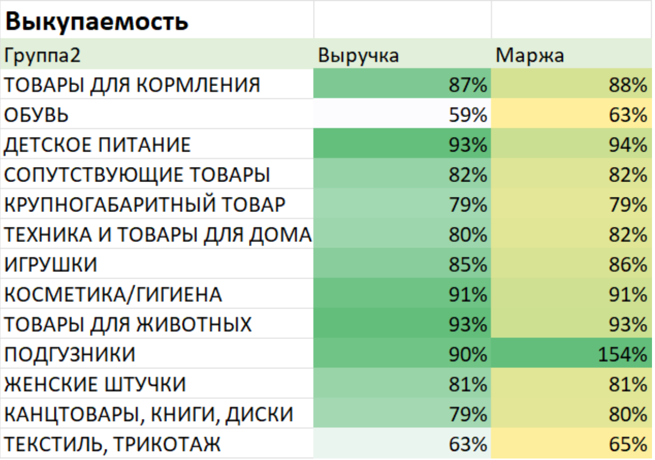
Food items are bought in large quantities, with 14.83 items per check on average, but their margin reaches only 5%. Another important discovery is that the diapers are the KVI or known value items. These are the items that drive the price value perception of customers. In our case diapers have a low price with a very low margin which might be even negative for some of the products. It can be concluded from the pic. 3.6 with the buyout percentages of revenue and margin.



Группа2	Выручка	Маржа	Всего товаров	Уникальные чеки	Уникальные клиенты	Средний чек	Среднее кол-во товара на чек	Средняя маржа с товара	Маржа,%
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	29 647 918.31	4 809 928.52	63 839	28 611	23 461	1 036.24	2.23	75.34	16%
ОБУВЬ	41 473 027.04	14 937 707.58	44 417	23 870	19 300	1 737.45	1.86	336.31	36%
ДЕТСКОЕ ПИТАНИЕ	76 305 794.98	3 936 583.96	583 480	39 331	24 256	1 940.09	14.83	6.75	5%
СОПУТСТВУЮЩИЕ ТОВАРЫ	465 338.36	166 284.10	2 124	1 372	1 292	339.17	1.54	78.29	36%
КРУПНОГАБАРИТНЫЙ ТОВАР	253 334 467.44	65 827 659.17	48 811	39 265	35 162	6 451.92	1.24	1 348.62	26%
ТЕХНИКА И ТОВАРЫ ДЛЯ ДОМА	690 184.96	155 671.98	2 171	1 159	1 048	595.50	1.87	71.71	23%
ИГРУШКИ	138 377 031.83	25 715 985.28	223 013	85 883	69 713	1 611.23	2.59	115.31	19%
КОСМЕТИКА/ГИГИЕНА	26 552 152.70	4 398 957.14	141 053	36 689	26 902	723.71	2.84	31.19	17%
ТОВАРЫ ДЛЯ ЖИВОТНЫХ	9 289 850.41	718 551.72	48 231	5 443	3 861	1 706.75	8.86	14.90	8%
ПОДГУЗНИКИ	134 726 662.09	1 085 569.80	146 386	68 648	42 967	1 962.57	2.1	7.42	1%
ЖЕНСКИЕ ШТУЧКИ	345 191.63	128 426.74	4 158	1 380	1 204	250.14	3.01	30.89	37%
КАНЦТОВАРЫ, КНИГИ, ДИСКИ	14 246 333.05	4 213 130.60	88 511	16 409	14 160	868.20	5.39	47.60	30%
ТЕКСТИЛЬ, ТРИКОТАЖ	153 036 219.69	58 016 299.26	223 279	66 921	48 527	2 286.82	3.33	259.84	38%

Pic. 3.5 Shop figures by item category

There is an interesting pattern: the less diapers were bought, the higher the margin was. That means that diapers are sold at a very cheap price and to attract customers. Customers who made a purchase because of the diapers might also order other items with higher margins and this will result in profit for the store. However, we should avoid using diapers in our offers since they already have a small price.



Выкупаемость		
Группа2	Выручка	Маржа
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	87%	88%
ОБУВЬ	59%	63%
ДЕТСКОЕ ПИТАНИЕ	93%	94%
СОПУТСТВУЮЩИЕ ТОВАРЫ	82%	82%
КРУПНОГАБАРИТНЫЙ ТОВАР	79%	79%
ТЕХНИКА И ТОВАРЫ ДЛЯ ДОМА	80%	82%
ИГРУШКИ	85%	86%
КОСМЕТИКА/ГИГИЕНА	91%	91%
ТОВАРЫ ДЛЯ ЖИВОТНЫХ	93%	93%
ПОДГУЗНИКИ	90%	154%
ЖЕНСКИЕ ШТУЧКИ	81%	81%
КАНЦТОВАРЫ, КНИГИ, ДИСКИ	79%	80%
ТЕКСТИЛЬ, ТРИКОТАЖ	63%	65%

Pic. 3.6 The bought out percentage of revenue and margin by item category

3.2.3 ABT creation

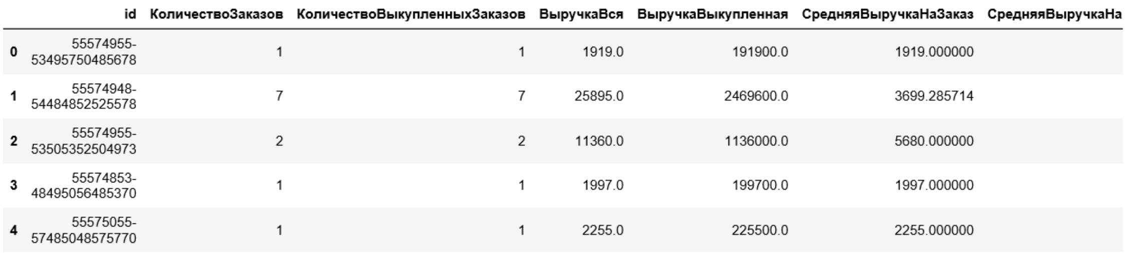
The initial dataset is inconvenient to work with. Each customer and order occur in multiple columns and for the analysis we have to gather the behavior of each client together in one place. Analytical Base Table (ABT) is a flat table used for building analytical models. A single record represents a subject of prediction, the client, and columns store all available information describing the object.

First we have to find an identifier for each client. In our case it will be the telephone number, since it is a necessary attribute to register an account and all clients have it stated. Next we have to aggregate all information into new columns. These columns will include the total number of orders made, total and bought out number of items, total revenue and margin, average revenue and margin per order. Also we calculated the proportion of bought items by categories.

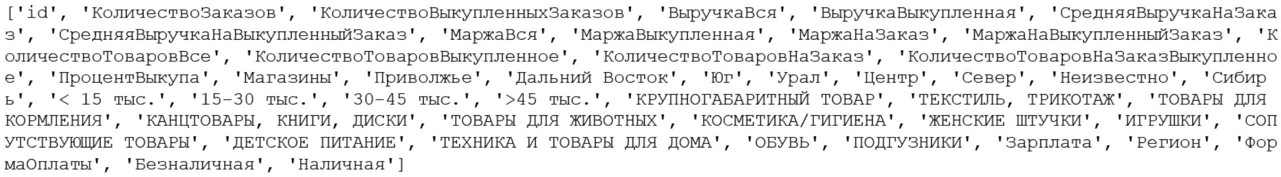
The shop has a product hierarchy: the highest vertex is ‘ Группа2’ which consists of 13 categories. ‘Группа2’ branches into ‘Группа3’ and the leaves of this tree are ‘Группа4’ categories. Although we will conduct several experiments with lower level, we will primary focus on ‘Группа2’. For the ABT we added 13 columns in each of which we calculated the proportion of items bought in this category. The proportion is needed because some items, like heavy goods, are usually bought in small amounts, while baby food is bought in big sets, so we have to scale these items for fair comparison. These columns reflect the item preferences of the clients so we may cluster ones with similar item choices.

Salary and geographical location were one-hot encoded and also added to the table.

Full list of ABT columns can be found on a pic. 3.8.



	id	КоличествоЗаказов	КоличествоВыкупленныхЗаказов	ВыручкаВся	ВыручкаВыкупленная	СредняяВыручкаНаЗаказ	СредняяВыручкаНа
0	55574955-53495750485678	1	1	1919.0	191900.0	1919.000000	
1	55574948-54484852525578	7	7	25895.0	2469600.0	3699.285714	
2	55574955-53505352504973	2	2	11360.0	1136000.0	5680.000000	
3	55574853-48495056485370	1	1	1997.0	199700.0	1997.000000	
4	55575055-57485048575770	1	1	2255.0	225500.0	2255.000000	



```
['id', 'КоличествоЗаказов', 'КоличествоВыкупленныхЗаказов', 'ВыручкаВся', 'ВыручкаВыкупленная', 'СредняяВыручкаНаЗака
з', 'СредняяВыручкаНаВыкупленныйЗаказ', 'МаржаВся', 'МаржаВыкупленная', 'МаржаНаЗаказ', 'МаржаНаВыкупленныйЗаказ', 'К
оличествоТоваровВсе', 'КоличествоТоваровВыкупленное', 'КоличествоТоваровНаЗаказ', 'КоличествоТоваровНаЗаказВыкупленно
е', 'ПроцентВыкупа', 'Магазины', 'Приволжье', 'Дальний Восток', 'Юг', 'Урал', 'Центр', 'Север', 'Неизвестно', 'Сибир
ь', '< 15 тыс.', '15-30 тыс.', '30-45 тыс.', '>45 тыс.', 'КРУПНОГАБАРИТНЫЙ ТОВАР', 'ТЕКСТИЛЬ, ТРИКОТАЖ', 'ТОВАРЫ ДЛЯ
КОРМЛЕНИЯ', 'КАНЦТОВАРЫ, КНИГИ, ДИСКИ', 'ТОВАРЫ ДЛЯ ЖИВОТНЫХ', 'КОСМЕТИКА/ГИГИЕНА', 'ЖЕНСКИЕ ШТУЧКИ', 'ИГРУШКИ', 'СОП
УТСТВУЮЩИЕ ТОВАРЫ', 'ДЕТСКОЕ ПИТАНИЕ', 'ТЕХНИКА И ТОВАРЫ ДЛЯ ДОМА', 'ОБУВЬ', 'ПОДГУЗНИКИ', 'Зарплата', 'Регион', 'Фор
маОплаты', 'Безналичная', 'Наличная']
```

Pic. 3.7 Analytical Base Table; pic 3.8 Columns of ABT

4. Clusterization

Clustering is a task of grouping a set of objects such that the objects in the same cluster are as similar as possible while objects within different clusters are as different as possible.

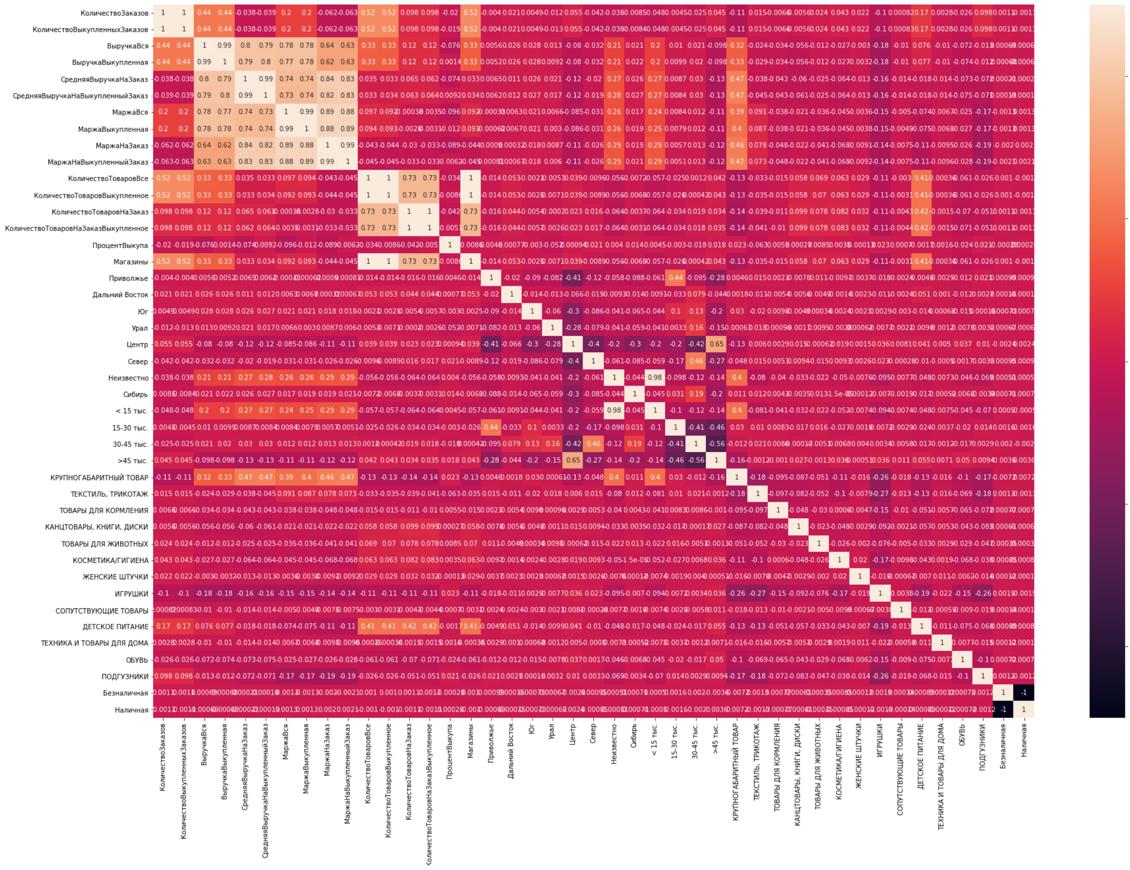
Deep understanding of different customer types that buy from you enables you to more accurately design campaigns to attract buyers. Customers consume items in different ways. Some of them buy a couple of items with high frequency, some buy many items at once but rarely. If we identify main segments of our client base we can form individual communication strategies and make targeted offers which attract customers more efficiently.

To perform segmentation we represent each customer as a vector of some specified features so that we can find the distance between each customer. There exist different distance metrics and algorithms and in this work we will focus on k-means and DBSCAN algorithms with euclidean distance metric.

4.1 Checking for correlations, scaling, PCA analysis

Correlation

Before clusterization we must do some extra transformations with data. For the window we created several features that are similar to each other, for example total margin and margin per orders, total margin and total revenue. We should check the features for correlation. Strong correlation makes it difficult for models to estimate the effect of each dependent variable, because independent variables tend to change in unison. Apart from that we will reduce dimensionality of our data which will decrease the allocated memory and working time and of the algorithms.



Pic. 4.1 Correlation of features

Indeed, there are many correlated features. We will delete the features with correlation higher than 0.45. For One Hot Encoded variables we have to have *k-1* features for *k* categories of variables. In the ABT creation we made columns for all categories, so now we have to delete one geographical location and average salary. The number of remaining features is 40.

Scaling

Next we need to do scaling to equalize the data. Different features have different ranges of value and the distance we do not normalize them the weight of features in the distance will be unequal. We will use a standard scaler which transforms the data by removing the mean and scaling to unit variance.[9]

Principal Component Analysis

Principal Component Analysis is a method to reduce dimensionality of the data losing the least possible amount of information [10]. It projects each data point to a new basis to obtain a lower-dimensional while saving as much variation as possible.

For transformation we can either specify a number of components (dimensions) or proportage of variance to preserve. We set 95% of preserved variance and from 40 features got 24.

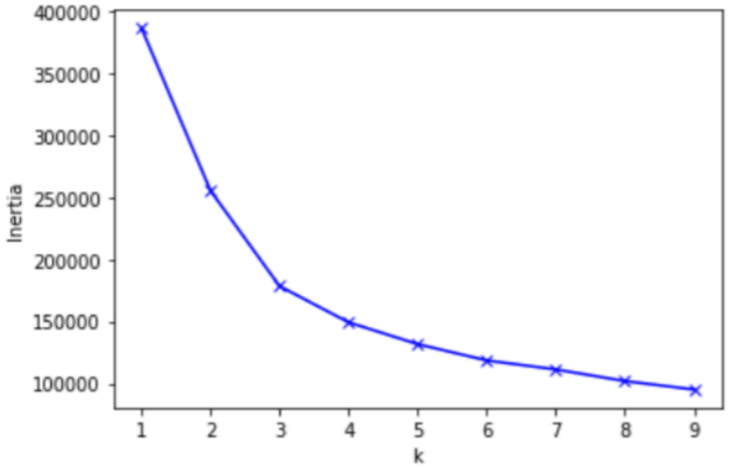
4.2 Determining the optimal number of clusters

To run k-means we have to specify the number of clusters. Too few clusters may mix too different people together and weaken their diversity. We also do not want too many clusters because we can get too specific groups which will be difficult and long to analyze. There are several methods to determine the optimal number of classes although experimentation is also an important part in building the model. We will try the elbow method and the GAP statistic.

Elbow method

Elbow method means plotting the graph of sum of squared distance between each point and its respective cluster centroid for a different number of clusters. An optimal number is called an elbow. It is the point where difference between k and k+1 clusters becomes too insignificant to build additional clusters, or a point after which the diminishing returns are no longer worth an additional cost.

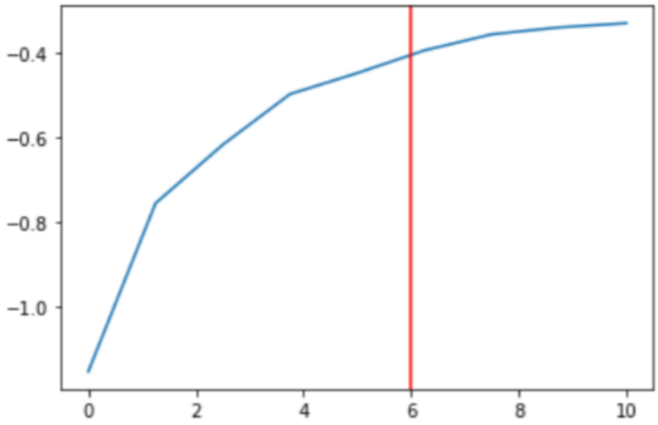
On our data the elbow is at 3 clusters but 4 and 5 clusters are also good options and after that the difference becomes very small.



Pic. 4.2 Graph of Elbow method

The GAP statistic

The GAP statistics compares the intracluster variation of k clusters with their expected variation under a null reference distribution of the data. On our data the GAP statistics did not produce very good results but about 6 clusters is good for the segmentation.



Pic. 4.3 Graph of GAP statistic

4.3 K-means

K-Means is a distance-based algorithm. It minimizes the sum of distances between the points in the cluster and their respective centroid. The distances The algorithm can be shortly summed up:

(1) take a desired number N of clusters as an input;

(2) randomly assign N point from the data to be centroids;

(3) cluster all remaining data by assigning it to the closest centroid using some specified metric (in our case it is euclidean distance);

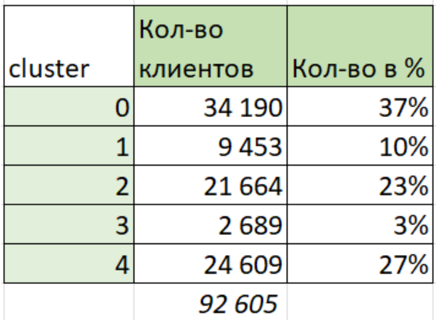
(4) recompute centroids;

(5) repeat steps (3) and (4) until the centroids do not change.

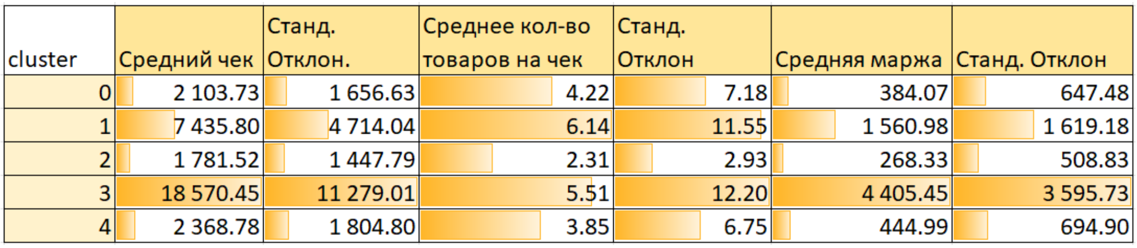
To get the best segmentation which will also have business value we experimented a lot with features, cluster numbers and even deleted some more outliers. We resulted in 5 segments which will be described in the next section.

4.3.2 Kmeans segments

In this section we will describe resulting clusters. In pic. 4.3 the distribution of clients between clusters is depicted. We have 3 big clusters which differ in consumption patterns and 2 small clusters with significantly bigger average checks. To analyze preferred items we made tables which depict the proportion of clients of each cluster who at least once bought an item in the category. (pic. 4.8-4.13)



cluster	Кол-во клиентов	Кол-во в %
0	34 190	37%
1	9 453	10%
2	21 664	23%
3	2 689	3%
4	24 609	27%
	92 605	



cluster	Средний чек	Станд. Отклон.	Среднее кол-во товаров на чек	Станд. Отклон	Средняя маржа	Станд. Отклон
0	2 103.73	1 656.63	4.22	7.18	384.07	647.48
1	7 435.80	4 714.04	6.14	11.55	1 560.98	1 619.18
2	1 781.52	1 447.79	2.31	2.93	268.33	508.83
3	18 570.45	11 279.01	5.51	12.20	4 405.45	3 595.73
4	2 368.78	1 804.80	3.85	6.75	444.99	694.90

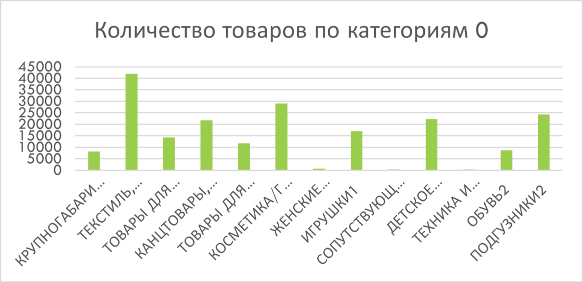
Pic. 4.3 Distribution of customers by segment in k-means;
pic. 4.4 Average figures by segment in k-means

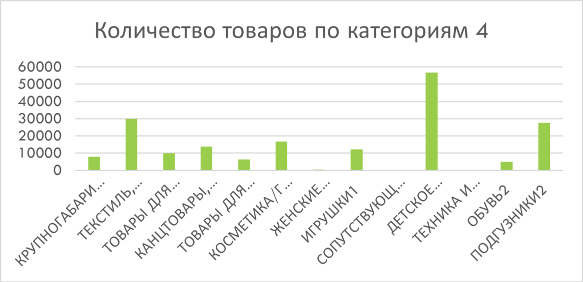
Segment 0

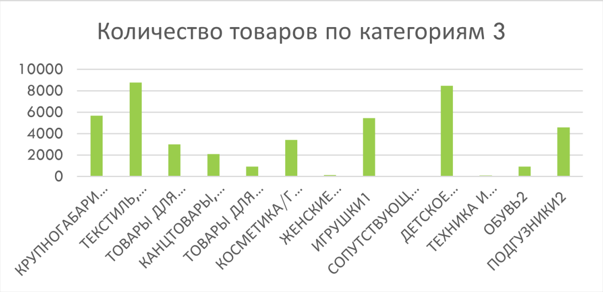
This is the biggest segment with a very low average check. Customers buy on average 4.22 items per check which is 2 100 RUB. In this segment textile is the most popular item. 31.3% of customers bought textile items and on pic. 4.5 we can observe that other items have a lower demand.

Segment 1

This is a cluster with a high average check of 7 435 rubles. There are no outstanding categories, but we can assume that this segment resembles the VIP segment. The average number of orders is big - 2.43 (pic. 4.4) and so is the average number of items per order.







Pic. 4.5-4.7 Number of bought items by categories for segments 0, 4 and 3

Segment 2

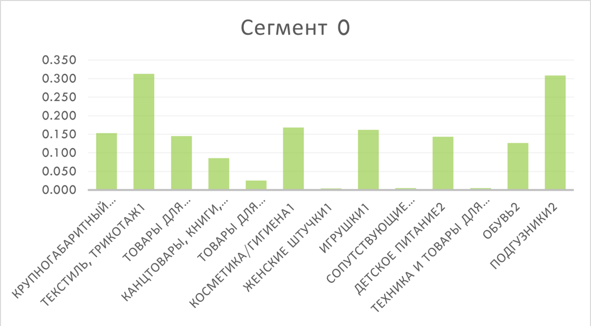
This segment includes 23% of the customers and all of them bought at least one toy. The average check, 1 781 RUB, and the margin, 268 RUB, are the lowest among the clusters so we should try to increase it.

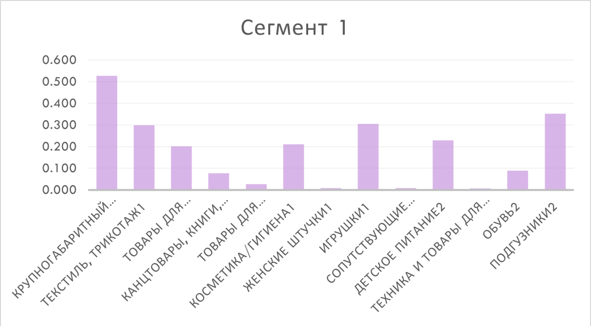
Segment 3

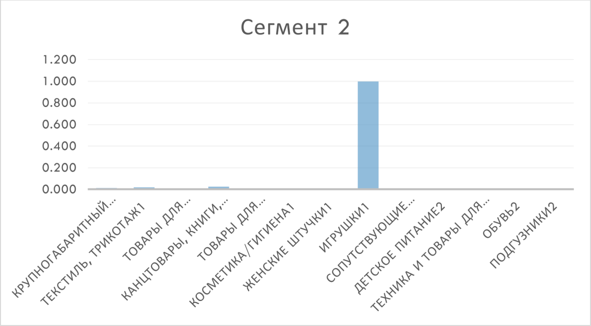
This is the smallest and most expensive segment. There are only 2689 people in it. More than 80% of these customers bought items in category 'КРУПНОГАБАРИТНЫЕ ТОВАРЫ'. The average check is 18 570 rubles. Besides, the average number of items per check and its standard deviation are very high. If we look at the number of bought items by category (pic. 4.7), we will see that these customers also buy many textiles, toys and food items. We can conclude that as well as buying heavy goods these clients make big orders in general.

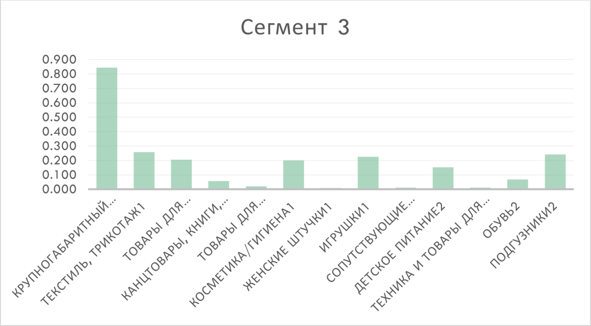
Segment 4

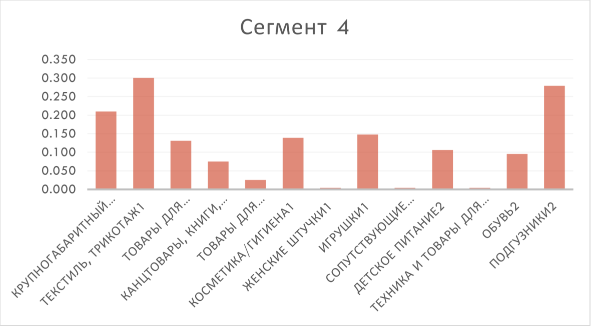
This is a core segment. It consists of 24 609 or 27% clients. The average check and margin are the biggest among big clusters. At first glance the proportion of people who bought specified items looks similar to the 0th segment, but if we look at the amount of items by category (pic. 4.6) we will see that clients buy a lot of baby food and less textile compared to segment 0.

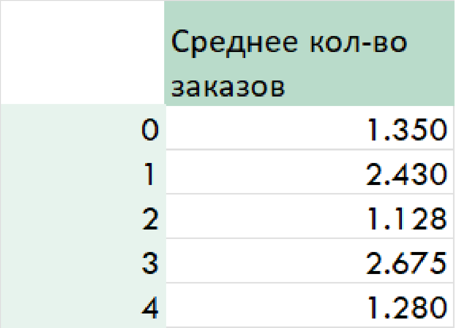












	Среднее кол-во заказов
0	1.350
1	2.430
2	1.128
3	2.675
4	1.280

Pic. 4.8-4.13 The proportion of customers who bought at least 1 item in specific category; pic. 4.14 Average number of orders by segment

4.4 DBSCAN

DBSCAN is a density based spatial clusterization algorithm. It is different from k-means in that it identifies the number of segments on its own. It takes a point and captures the ones that lie in a specified neighborhood until it cannot reach any more points and then moves to the other points to form other clusters. More precisely the action sequence is:

(1) take a distance D and minimum size of cluster N as an input;

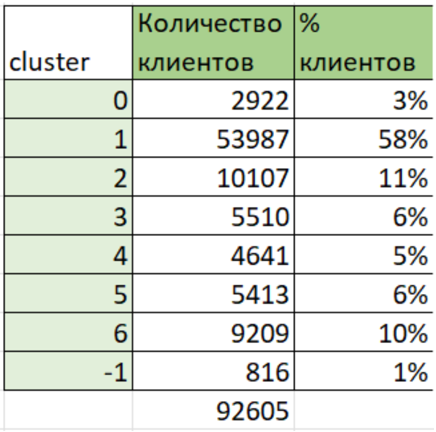
(2) for all points find neighbors in D neighborhood;

(3) choose the core points with N or more neighbors;

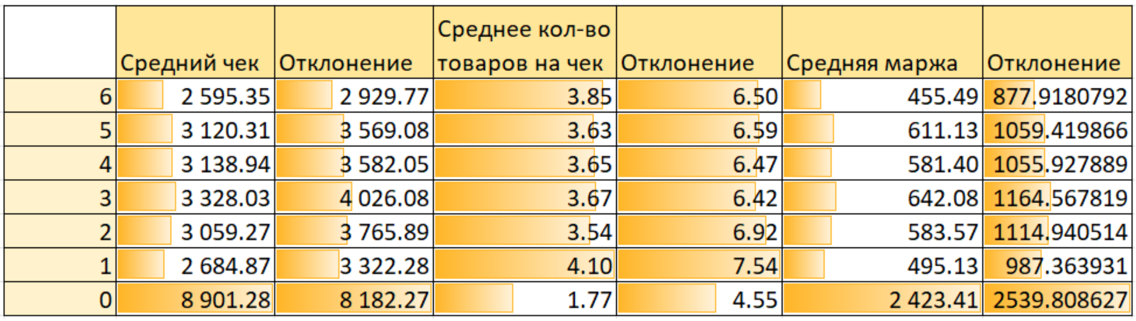
(4) find all connected components of core points which lie in distance smaller or equal to D;

(5) for all other points assign them to a cluster if it lies in the distance smaller than D, otherwise consider it a noize.

On our data DBSCAN performed poorly, probably due to the high density of data points. It either produced an extremely big number of clusters (20+) or one huge cluster and several small ones which were not meaningful.



cluster	Количество клиентов	% клиентов
0	2922	3%
1	53987	58%
2	10107	11%
3	5510	6%
4	4641	5%
5	5413	6%
6	9209	10%
-1	816	1%
	92605	



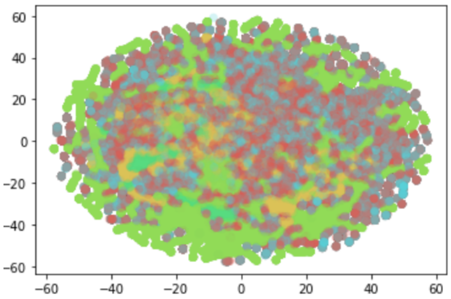
	Средний чек	Отклонение	Среднее кол-во товаров на чек	Отклонение	Средняя маржа	Отклонение
6	2 595.35	2 929.77	3.85	6.50	455.49	877.9180792
5	3 120.31	3 569.08	3.63	6.59	611.13	1059.419866
4	3 138.94	3 582.05	3.65	6.47	581.40	1055.927889
3	3 328.03	4 026.08	3.67	6.42	642.08	1164.567819
2	3 059.27	3 765.89	3.54	6.92	583.57	1114.940514
1	2 684.87	3 322.28	4.10	7.54	495.13	987.363931
0	8 901.28	8 182.27	1.77	4.55	2 423.41	2539.808627

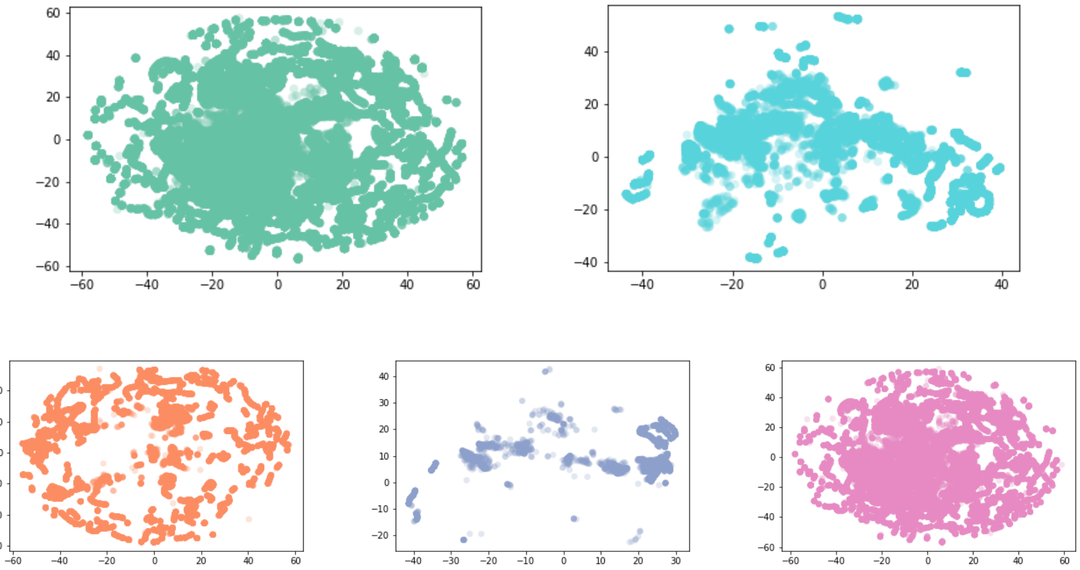
Pic. 4.15 Distribution of customers by segment in DBSCAN;
pic. 4.16 Average figures by segment in DBSCAN

4.5 TSNE

To illustrate why DBSCAN failed we visualized our data using TSNE. TSNE is an algorithm for transforming high dimensional data into low dimensional space. It is very often used for visualization of data to look at its density. TSNE transforms data in a way such that points are located close to similar ones and far from differing ones.

On the pic. 4.17 you can see the visualization of our dataset. Points are located with very high density all over the 2 dimensional space. DBSCAN, being a spatial algorithm, could have failed to identify segments because the distribution of elements is very dense and it continued capturing all points into one segment until only a few small groups are left . Pictures 4.18-4.22 depict how points of resulting k-means clusters are distributed on the TSNE graph.





Pic. 4.17 TSNE visualization of whole dataframe

pic. 4.18-4.22 TSNE visualization of separate k-means segments

5. Market Basket Analysis

Market basket analysis is a data mining technique used in e-commerce to better understand customer purchasing patterns [4]. By analyzing purchase history, it can find products that are likely bought together. This information can help to create personalized offers and campaigns aimed to increase revenue. This technique is easy to understand and implement.

The algorithm takes 2 items and assigns them roles of antecedent, A - left hand of rule, and consequent, B - the right hand of rule. The rule is “If user bought item A then he also bought item B”. For each pair, it calculates the following metrics:

- support = $\frac{Quantity(A+B)}{Quantity(all\ transactions)}$ - shows the proportion of orders with both of the items

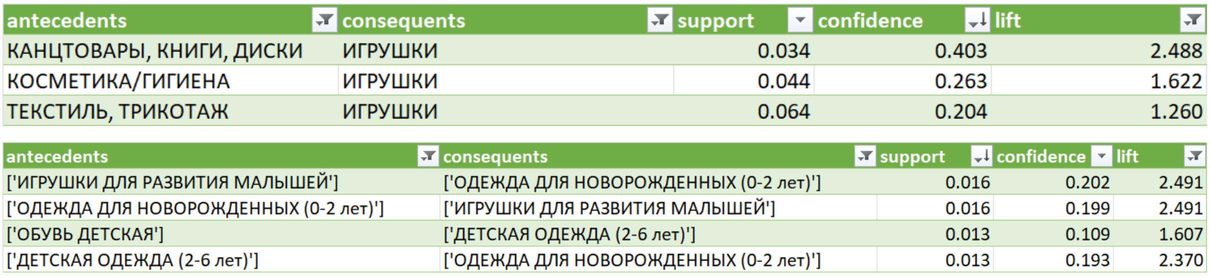
- confidence = $\frac{Quantity(A+B)}{Quantity(all\ transactions)}$ - Shows how many checks with item A also have item B. It shows the legitimacy of the rule.
- expected confidence = $\frac{Quantity(B)}{Quantity(all\ transactions)}$ -
- lift = $\frac{confidence}{expected\ confidence}$ - shows the correlation between items A and B. If *lift* > 1 then item B is bought *lift* times more often if item A is also in the check.

As a result we get a set of common patterns described by 3 main metrics (expected confidence is usually not considered, it is calculated for lift). Then for offers we can choose the items to promote based on some metric criteria or business logic. For implementation we will use a python library MLxtend which has an implementation of Apriori algorithm[12] for extracting frequent itemsets in a library MLxtend[11].

Based on the MBA results we will form offers formulation using common marketing mechanics. For simplicity we will assume that when increasing the average check we take the current average by segment and increase it by 200 rubles and all discounts will be 5% since we do not want the discount to be larger than the margin of the items.
The results and offers for each segment will be described below.

Segment 0

Segment 0 is a segment with many textile buyers. This segment has a very low average check so we will be primarily trying to increase it. The most prominent pattern is that clients who bought textile items also buy toys. “Ваша персональная скидка 5% на категории “ТЕКСТИЛЬ, ТРИКОТАЖ” и “ИГРУШКИ” при покупке на 2300 рублей.”



antecedents	consequents	support	confidence	lift
КАНЦТОВАРЫ, КНИГИ, ДИСКИ	ИГРУШКИ	0.034	0.403	2.488
КОСМЕТИКА/ГИГИЕНА	ИГРУШКИ	0.044	0.263	1.622
ТЕКСТИЛЬ, ТРИКОТАЖ	ИГРУШКИ	0.064	0.204	1.260

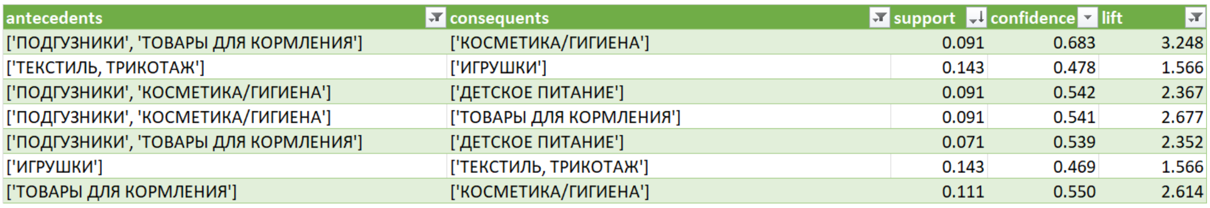
antecedents	consequents	support	confidence	lift
['ИГРУШКИ ДЛЯ РАЗВИТИЯ МАЛЫШЕЙ']	['ОДЕЖДА ДЛЯ НОВОРОЖДЕННЫХ (0-2 лет)']	0.016	0.202	2.491
['ОДЕЖДА ДЛЯ НОВОРОЖДЕННЫХ (0-2 лет)']	['ИГРУШКИ ДЛЯ РАЗВИТИЯ МАЛЫШЕЙ']	0.016	0.199	2.491
['ОБУВЬ ДЕТСКАЯ']	['ДЕТСКАЯ ОДЕЖДА (2-6 лет)']	0.013	0.109	1.607
['ДЕТСКАЯ ОДЕЖДА (2-6 лет)']	['ОДЕЖДА ДЛЯ НОВОРОЖДЕННЫХ (0-2 лет)']	0.013	0.193	2.370

Pic. 5.1-5.2 MBA associations for segment 0, “Группа2” & “Группа3”

Segment 1

This is a rather small cluster of people who have a high average check and buy a large variety of goods. Clients buy many daily necessities, for example, diapers and goods for hygiene. Also people buy baby food and goods for feeding in different combinations.

We want to promote the continuity and regularity of purchases. The mechanics will be “Совершите суммарную покупку на 7500 рублей в течение месяца и Вы получите баллы в размере 5% от накопленной суммы на покупки в следующем месяце!”.



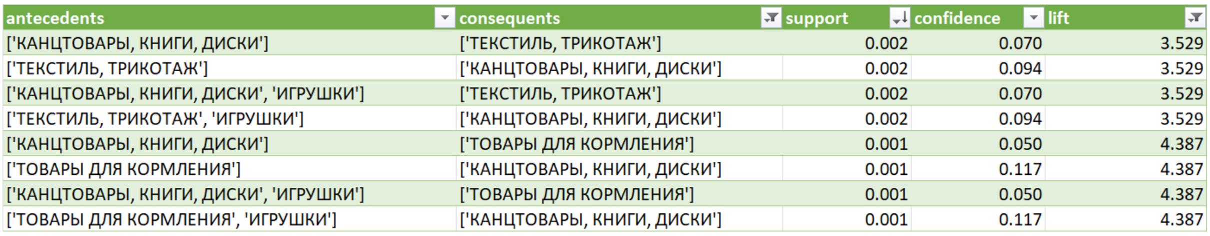
antecedents	consequents	support	confidence	lift
['ПОДГУЗНИКИ', 'ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['КОСМЕТИКА/ГИГИЕНА']	0.091	0.683	3.248
['ТЕКСТИЛЬ, ТРИКОТАЖ']	['ИГРУШКИ']	0.143	0.478	1.566
['ПОДГУЗНИКИ', 'КОСМЕТИКА/ГИГИЕНА']	['ДЕТСКОЕ ПИТАНИЕ']	0.091	0.542	2.367
['ПОДГУЗНИКИ', 'КОСМЕТИКА/ГИГИЕНА']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.091	0.541	2.677
['ПОДГУЗНИКИ', 'ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['ДЕТСКОЕ ПИТАНИЕ']	0.071	0.539	2.352
['ИГРУШКИ']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.143	0.469	1.566
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['КОСМЕТИКА/ГИГИЕНА']	0.111	0.550	2.614

Pic. 5.3 MBA associations for segment 1, “Группа2”

Segment 2

All clients in segment 2 bought toys. They have the lowest average check and number of items per check and the toys category has a very low margin on average. The MBA does not have any outstanding patterns but there are many triplets in which there are toys. We will not focus on the MBA association but on making toys bring more revenue. We will try to increase the average check by making customers buy more expensive items. The average price of toys is 675 rubles. We want each client to buy 2 items per check, and, increasing the average check by 200 rubles, we get an offer:

“Купите 2 товара в категории “ИГРУШКИ” на сумму от 2000 рублей и получите скидку 5%”.



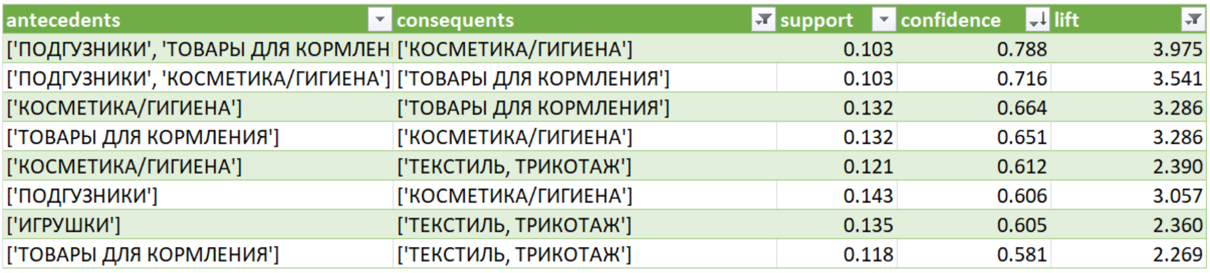
antecedents	consequents	support	confidence	lift
['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.002	0.070	3.529
['ТЕКСТИЛЬ, ТРИКОТАЖ']	['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	0.002	0.094	3.529
['КАНЦТОВАРЫ, КНИГИ, ДИСКИ', 'ИГРУШКИ']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.002	0.070	3.529
['ТЕКСТИЛЬ, ТРИКОТАЖ', 'ИГРУШКИ']	['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	0.002	0.094	3.529
['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.001	0.050	4.387
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	0.001	0.117	4.387
['КАНЦТОВАРЫ, КНИГИ, ДИСКИ', 'ИГРУШКИ']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.001	0.050	4.387
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ', 'ИГРУШКИ']	['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	0.001	0.117	4.387

Pic. 5.4 MBA associations for segment 2, “Группа2”

Segment 3

This is a segment in which customers bought many big and expensive items. We also noticed that these clients buy many daily goods like textile, toys and hygiene items.

Since expensive items are usually bought in small quantities, there are not many patterns with them. The rules show that hygiene and makeup products were bought with different other categories regularly with baby food, textile and feeding accessories, so we chose “КОСМЕТИКА/ГИГИЕНА” as a category for promotion. The goal of the offer is in increasing the variety of carts and promoting regularity of purchases. The mechanics will be: “Совершите покупку в разделе ‘КОСМЕТИКА/ГИГИЕНА’ на сумму от 2200 рублей и получите бонусы в размере 5% от покупки.” Since most of the clients bought expensive items, their average check is very high and it should not be used as the minimum price. Instead I will set 2200 rubles as the minimum price since it balances the clients who are able to make expensive orders and the low average price of the category.

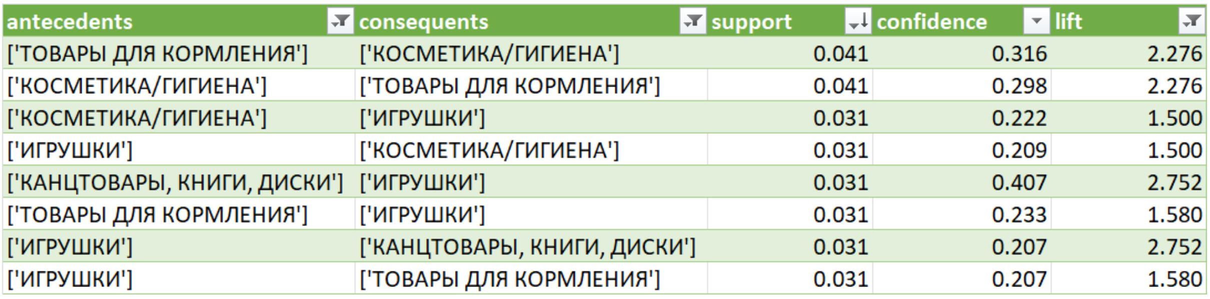


antecedents	consequents	support	confidence	lift
['ПОДГУЗНИКИ', 'ТОВАРЫ ДЛЯ КОРМЛЕН	['КОСМЕТИКА/ГИГИЕНА']	0.103	0.788	3.975
['ПОДГУЗНИКИ', 'КОСМЕТИКА/ГИГИЕНА']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.103	0.716	3.541
['КОСМЕТИКА/ГИГИЕНА']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.132	0.664	3.286
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['КОСМЕТИКА/ГИГИЕНА']	0.132	0.651	3.286
['КОСМЕТИКА/ГИГИЕНА']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.121	0.612	2.390
['ПОДГУЗНИКИ']	['КОСМЕТИКА/ГИГИЕНА']	0.143	0.606	3.057
['ИГРУШКИ']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.135	0.605	2.360
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['ТЕКСТИЛЬ, ТРИКОТАЖ']	0.118	0.581	2.269

Pic. 5.5 MBA associations for segment 3, “Группа2”

Segment 4

This is a core segment in which customers buy a wide range of items. The average check is bigger compared to the toys and textile segment and we will focus on increasing the frequency of orders with the help of MBA. The combination of “КОСМЕТИКА/ГИГИЕНА” and “ТОВАРЫ ДЛЯ КОРМЛЕНИЯ” has a high lift and both these categories have associations with other items, so we will make an offer with them: “Соверши покупку на сумму от 2500 рублей и получи скидку 5% на категории КОСМЕТИКА/ГИГИЕНА” и “ТОВАРЫ ДЛЯ КОРМЛЕНИЯ” при следующей покупке”.



antecedents	consequents	support	confidence	lift
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['КОСМЕТИКА/ГИГИЕНА']	0.041	0.316	2.276
['КОСМЕТИКА/ГИГИЕНА']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.041	0.298	2.276
['КОСМЕТИКА/ГИГИЕНА']	['ИГРУШКИ']	0.031	0.222	1.500
['ИГРУШКИ']	['КОСМЕТИКА/ГИГИЕНА']	0.031	0.209	1.500
['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	['ИГРУШКИ']	0.031	0.407	2.752
['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	['ИГРУШКИ']	0.031	0.233	1.580
['ИГРУШКИ']	['КАНЦТОВАРЫ, КНИГИ, ДИСКИ']	0.031	0.207	2.752
['ИГРУШКИ']	['ТОВАРЫ ДЛЯ КОРМЛЕНИЯ']	0.031	0.207	1.580

Pic. 5.6 MBA associations for segment 4, “Группа2”

6. Collaborative Filtering Recommendations

Collaborative filtering is a method of making automatic predictions about the interests of a user by collecting preferences or taste information from many users (collaborating)[13]. It is a common approach for recommendation systems. Recommendation systems predict sets of items which a person may like based on their past preferences. Two main approaches to model them are Content-Based Filtering and Collaborative Filtering-based recommender systems. The content-based algorithm uses user preferences for item characteristics to come up with recommendations. For example, if you liked a fantasy film from 2000 the model will search other movies from this genre and year. Our dataset does not contain much information about items apart from their type and nomenclature, so it was decided to use Collaborative Filtering which does not require features of the items to be given. In this approach the model looks for users who rated items similarly to you. If you highly rated Harry Potter and Lord of the Rings, the model will recommend a movie that was liked by users who also highly rated these films.

There are 2 main techniques of CF (Collaborative Filtering): memory-based approach and latent factor models. First approach relies on calculating the distance between feature vectors of users and finding the nearest neighbors. Second approach tries to characterize both users and items by transforming them into the same latent factor space. The latent factor space represents hidden features of films and users and multiplication of these features will give the predicted ratings.

In the context of an online shop, we will use the nearest neighbors approach. Our goal is to predict the future purchases of clients, and we will test whether customers with similar behavior will buy the same items in the future.

To model this problem mathematically, we create a binary table of all items bought by users during the first 3 months, where 1 indicates “bought at least 1 item” and 0 - “did not buy items”. Next we create a table for the 4th month: each column will represent a category of an item and ‘1’ will mean that the client bought that item in the 4th month and ‘0’ if not. The model will be trained on the first 3 months and prediction will be made for the 4th one. For each test user we will find *k* users with closest feature vectors, and get their purchases in the 4th month. For each category, if more than some fixed proportion of the neighbors bought an item, we will conclude that a target user will also buy this item. As a result, we will get a binary vector with predictions for 13 categories.

We will train the model only on users who have bought items both in the first 3 months and in the 4th. We will focus on prediction accuracy and the task of calculating the churn rate will be left to the response models. The tests will be conducted on ‘Группа2’, ‘Группа3’ and ‘Группа4’ levels of items. For final predictions we will use ‘Группа2’. We want to construct personalized offers yet with an opportunity to choose items of liking by clients. Other levels can produce more customized predictions, but it is also harder to make them accurate because of the sparsity of matrices. In recommendation systems we have to make good predictions to maintain the users’ trust level, but there is some room for inaccuracies because usually there are many items which a user may possibly like. In this work, however, we want to test the model's ability to predict for as many clients as it is possible, partly due to inability to conduct field tests and reevaluate the model after them. Besides, we will also try reducing the dimensionality of the data using PCA to test whether it increases quality of predictions.

6.1 Results

We constructed ABTS similar to the ones used for Market Basket Analysis.

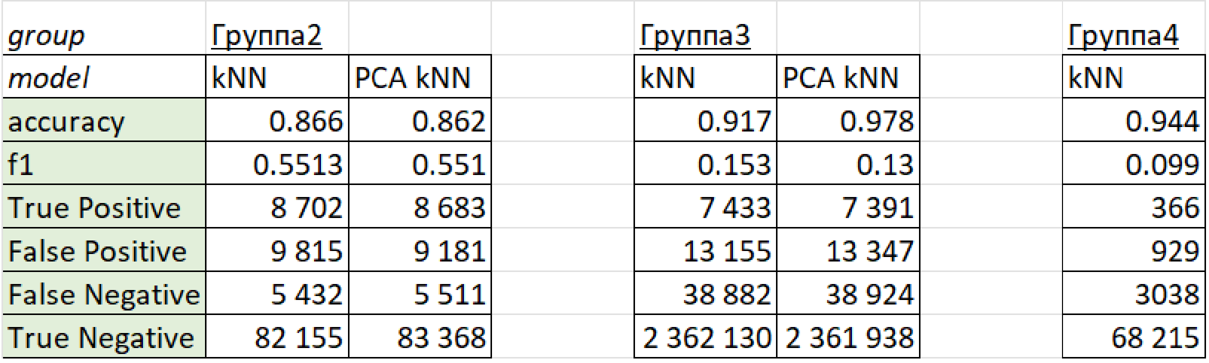
The metric we chose for testing is an F1-score with confusion matrices. We are primarily interested in the True Positives (correctly predicted positive values), False Positives (incorrectly predicted positive values) and False Negatives (incorrectly predicted negative values. We need to maximize the TP and minimize the FP and FN. F1 score is a harmonic mean of the recall and accuracy, metrics, defined as following:

Precision = $\frac{TP}{TP+FP}$

Recall = $\frac{TP}{TP+FN}$

F1 = $\frac{2*Precision*Recall}{Precision + Recall}$

F1 is a good metric to calculate on imbalanced datasets. It lies in range [0; 1] and we want to maximize this metric.



group	Группа2			Группа3			Группа4
model	kNN	PCA kNN		kNN	PCA kNN		kNN
accuracy	0.866	0.862		0.917	0.978		0.944
f1	0.5513	0.551		0.153	0.13		0.099
True Positive	8 702	8 683		7 433	7 391		366
False Positive	9 815	9 181		13 155	13 347		929
False Negative	5 432	5 511		38 882	38 924		3038
True Negative	82 155	83 368		2 362 130	2 361 938		68 215

Pic. 6.1 Different metric for CF model

The tests show that on the 'Группа2' F1 score reaches 0.55. It is a positive result considering that the model is unsupervised. About 60% of the positive values were correctly guessed although quite a big number of negative values were predicted as positive. The PCA transformation have not made any significant changes in quality.

To understand better whether this model is good for deployment we should conduct test fields and analyze how clients respond to offers. Some customers might like False Positive recommendations, some may not respond because we did not identify correct positives. For this work we consider the results satisfactory.

The tests on 'Группа3' and 'Группа4' show worse results. They have missed many positive values and marked as positive many negative values. Yet we will consider these results satisfactory. The mathematical performance is average but the practical value can be determined only with field tests. Besides, we will analyze the campaign from the financial standpoint.

It should be mentioned that the method suffers from a cold-start problem, but in our context we generate offers for our current database so our model is not affected by this problem.

To generate offers, we run the model on all of the clients who made purchases during the first 3 months, even if they did not return in the 4th. The probability of return will be predicted by response models and for now we want only to get item recommendations.

6.2 Forming offers

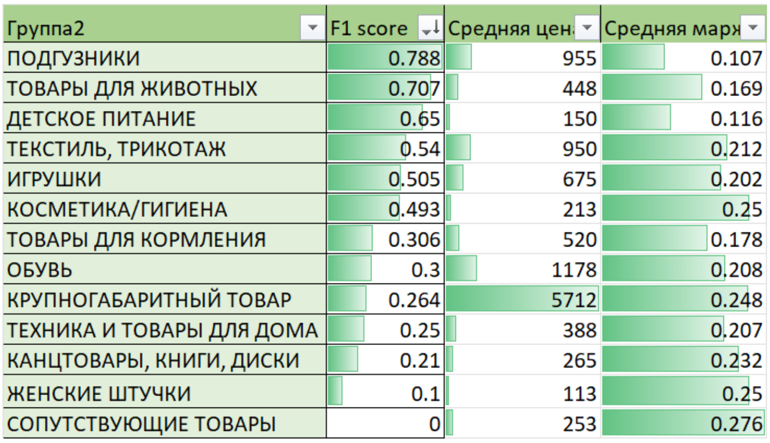
To form offers for CF we will rely on the quality of predictions by category and average checks of clients and among clusters. Some clients can have more than one category predicted and we do not want to use predictions which have a very low F1 score. Besides, for clients from segment 1 we will send an offer which we chose in MBA offers. Clients there have a very high average check so when comparing revenue MBA method may have an advantage. This offer does not include any MBA patterns so we could have come up with it even without using MBA technique. (The offer is “Совершите суммарную покупку на 7500 рублей в течение месяца и Вы получите баллы в размере 5% от накопленной суммы на покупки в следующем месяце!”)

For other clients we will use all categories with F1 score ≥ 0.3 except diapers, since we do not want to promote them. For other categories we will give a 5% discount. We also have to set a price under which clients will get a discount. We will take an average check of the clients and check whether it is not too small or big by restricting it with a range [1800 RUB; 2400 RUB] which is a minimum and maximum check among general segments (segments 0, 2, 4). We will add 200 RUB to each value with the goal of increasing an average check.

The examples of offers:

- ТОВАРЫ ДЛЯ ЖИВОТНЫХ - “‘Купите товары в категории ТОВАРЫ ДЛЯ ЖИВОТНЫХ на сумму от X RUB и получите скидку 5%’
- ДЕТСКОЕ ПИТАНИЕ - ‘Купите товары в категории ДЕТСКОЕ ПИТАНИЕ на сумму от X RUB и получите скидку 5%.’
- ТЕКСТИЛЬ, ТРИКОТАЖ - ‘Купите товары в категории ТЕКСТИЛЬ, ТРИКОТАЖ на сумму от X RUB и получите скидку 5%’

and so on.



Группа2	F1 score	Средняя цен	Средняя марж
ПОДГУЗНИКИ	0.788	955	0.107
ТОВАРЫ ДЛЯ ЖИВОТНЫХ	0.707	448	0.169
ДЕТСКОЕ ПИТАНИЕ	0.65	150	0.116
ТЕКСТИЛЬ, ТРИКОТАЖ	0.54	950	0.212
ИГРУШКИ	0.505	675	0.202
КОСМЕТИКА/ГИГИЕНА	0.493	213	0.25
ТОВАРЫ ДЛЯ КОРМЛЕНИЯ	0.306	520	0.178
ОБУВЬ	0.3	1178	0.208
КРУПНОГАБАРИТНЫЙ ТОВАР	0.264	5712	0.248
ТЕХНИКА И ТОВАРЫ ДЛЯ ДОМА	0.25	388	0.207
КАНЦТОВАРЫ, КНИГИ, ДИСКИ	0.21	265	0.232
ЖЕНСКИЕ ШТУЧКИ	0.1	113	0.25
СОПУТСТВУЮЩИЕ ТОВАРЫ	0	253	0.276

Pic. 6.1 Quality of chosen CF model and average prices by category

7. Response models

The response model is a classification model. It is used to predict the likelihood of some specified customer’s action using information collected about clients. Ehese models can predict a wide range of to predict the churn rate, the likelihood of buying special items and **

In this work we will predict the likelihood of a customer returning. Each customer will be assigned a binary class: 1 if a client has returned and 0 otherwise. For the training data we will construct a new ABT, similar to the one used in segmentation, but this time we will include data only from the first 3 months and the behavior from the 4th month will be used as a target variable or indicator of return.

It is important to mention that since the timeframe of training data is 3 months, we should predict the behavior for the next month using the data from exactly the past 3 months. We must keep the same timeframe because the mode using different time periods will negatively affect the quality of predictions.

The data was also cleaned from the customers who made purchases only at the targeted month. They are newbies and we cannot analyze their behavior patterns; they will only add noise to the model.

7.1 Model Choice

We will compare 3 models for classification: Decision Tree, Random Forest and XGBoost. These are popular models for solving classification problems. Decision Tree is an easy to understand algorithm which does not require any specific domain knowledge. It also results in good accuracy which is enough for solving many real world tasks.It is a non parametric supervised learning method which classifies the data into branch-like segments that result in an inverted tree.

Random Forest is an estimator which fits a number of decision trees on different subsamples and averages or combines in any other way their results. It takes Decision Tree’s generally good results and improves them by controlling overfitting and increasing accuracy even more.

XGBoost is known as one of the models with best performances on structural and tabular data. It enhances decision trees with a gradient boosting framework and provides a parallel tree boosting that solves many problems in a fast and accurate way. [14]

7.2 Metric Choice

The most simple metric for classification is accuracy, which denotes the percentage of correctly predicted answers. But accuracy poorly evaluates models trained and tested on an imbalanced dataset. Imbalanced datasets have an uneven distribution of observations in each class. Often we are interested in correctly identifying the observations from the minority class, and accuracy does not reflect this ability of a model. In such cases a good metric to use is an F1-score, which will be the main focus of this work. The metric was described in section 6.1.

7.4 Dealing with imbalanced data

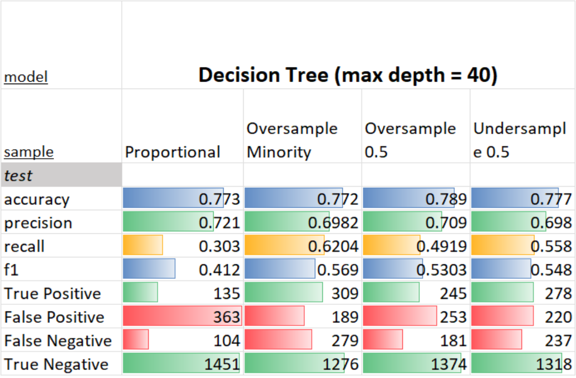
In our dataset there are 83390 clients in total and only 10623, or 12.7%, returned to the shop in the last month. This can be considered a moderate imbalanced distribution: clients from class '1', or returned, belong to the minority class, while clients from '0' class are from the majority class. There exist different techniques for improving model training on such distributions and the best approach is to empirically find the ones which give the best result. In this work we focus on oversampling and undersampling techniques.

- Oversampling - technique in which we randomly select examples from the minority class (with replacement) and add them to the training set. The proportion to which we want to oversample the minority class can be varied.
- Undersampling - technique in which we randomly select examples from the majority class and delete them from the training dataset. The proportion of deleted objects can also be varied.

When we have not too much data, oversampling is a preferable technique. By undersampling we can exclude from the dataset point from the majority class with very distinguishable features. When datasets are large the probability of excluding important samples becomes smaller. We will compare the performance of both techniques as well as performance on simple random samples and stratified samples. In stratified sampling we first define a number of stratas and then choose objects proportionally to the difference in size of stratas. In our case stratas are returned and not returned customers.

7.5 Comparing performance

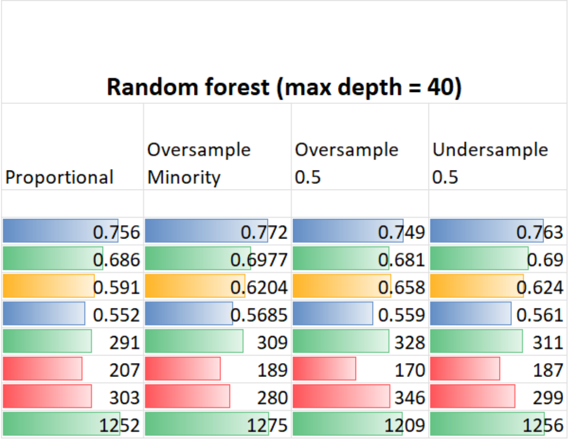
The dataset was divided into train set, validation set and test set according to 4 techniques described in the previous section. The parameters of models were tuned using the validation set and the performance on the test set will be described below.



model	Decision Tree (max depth = 40)			
sample	Proportional	Oversample Minority	Oversample 0.5	Undersampl e 0.5
test				
accuracy	0.773	0.772	0.789	0.777
precision	0.721	0.6982	0.709	0.698
recall	0.303	0.6204	0.4919	0.558
f1	0.412	0.569	0.5303	0.548
True Positive	135	309	245	278
False Positive	363	189	253	220
False Negative	104	279	181	237
True Negative	1451	1276	1374	1318

Pic. 7.1 Quality of Decision Tree

The best depth for Decision tree is 40. It provides a poor F1-score which varies from 0.412 to 0.57. Oversampling improves the accuracy but by a very small amount.

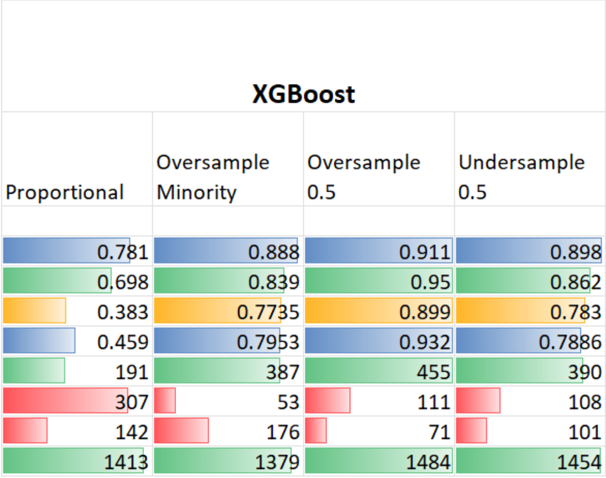


Random forest (max depth = 40)			
Proportional	Oversample Minority	Oversample 0.5	Undersample 0.5
0.756	0.772	0.749	0.763
0.686	0.6977	0.681	0.69
0.591	0.6204	0.658	0.624
0.552	0.5685	0.559	0.561
291	309	328	311
207	189	170	187
303	280	346	299
1252	1275	1209	1256

Pic. 7.1 Quality of Random Forest

Random Forest produces more stable results which are a little better compared to Decision Tree but the F1-score is still average.
XGBoost gives the best scores. On a simple proportional sample its performance is similar to previous methods but oversampling greatly boosts its performance up to reaching F1-score equal to 0.932. The best option for all models is an oversampling in which a minority class is sampled until it is a half off the size of a majority class. Another option is sampling the minority class until its size is equal to the majority class, but on our data it did not increase the quality of predictions.

Undersampling, on the other hand, does not improve the model. In some cases during training the performance was even worse compared to proportional samples. This is probably due to the small size of dataset.



XGBoost			
Proportional	Oversample Minority	Oversample 0.5	Undersample 0.5
0.781	0.888	0.911	0.898
0.698	0.839	0.95	0.862
0.383	0.7735	0.899	0.783
0.459	0.7953	0.932	0.7886
191	387	455	390
307	53	111	108
142	176	71	101
1413	1379	1484	1454

Pic. 7.3 XGBoost

For final predictions we will use the XGBoost model trained on an oversampled with parameter 0.5 training set. We will run these models on the ABT with behavior from months 2-4 and save results into the final ABT.

8. Business optimization & Result

Now we have a final ABT in which for each client we have 2 different offers, recommended by MBA and CF methods and probability of return. We have to find the best customers to send offers to. For it we will give a rating to each client. The formula for ratings will be constructed as follows:

$r\ =\ probability\ of\ return\ \ *\ \ minimum\ check\ of\ the\ offer$

We have 3 communication channels: phone call, sms and email. Each channel has price per client and efficiency, both parameters can be varied. For simplicity, we will assume that:

Phone call: $p_{call}\ =\ 100\ RUB,\ e_{call}\ =\ 0.6$

SMS: $p_{call}\ =\ 15\ RUB,\ e_{call}\ =\ 0.4$

Email: $p_{call}\ =\ 4\ RUB,\ e_{call}\ =\ 0.15$

Besides, some of the clients have not filled out their email, so we must set their ratings for email equal to 0.

As a result, we will have a matrix of size $n * 3$, where $r_{final_{ij}} = e_j * r_{ij}$ for $1 \le i \le n,\ 1 \le j \le c$.

We will have 2 different matrices with different ratings for MBA and CF. Now we have to optimize this matrix for communication. We can formulate this task as a Mixed Integer Linear Programming problem. The mathematical formulation was described in 3.1.

We may also want to put constraints on our campaign. For example, we cannot make more than *x* calls or send more than *y* emails or we want exactly *l* % of customers to be called by the call center. We may also set constraints on the budget of the campaign. The examples of constraints:

- Number of clients for channel 1 equals 1000:

$$\sum_{1 \le i \le n} x_{ij} = 1000 \text{ for j} = 1$$

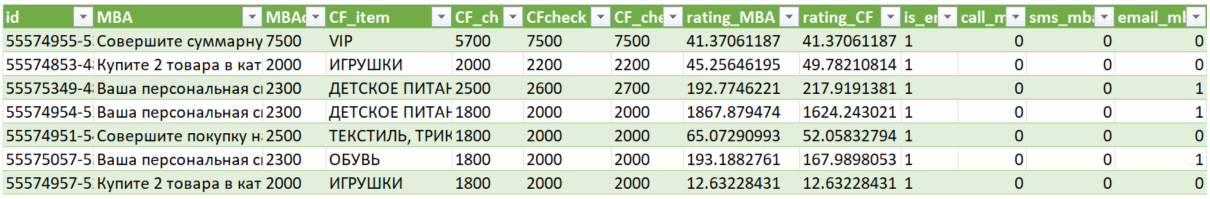
- The marketing budget for campaign is 250000 RUB:

$$\sum_{1 \le i \le n, 1 \le j \le c} x_{ij} * p_j \le 250000$$

To solve the problem we will use a python library called OR Tools [15]. It is a library for solving different linear programming problems. To illustrate the constraints we set that maximum number of phone calls is 3000 and maximum number of sms is 6000.

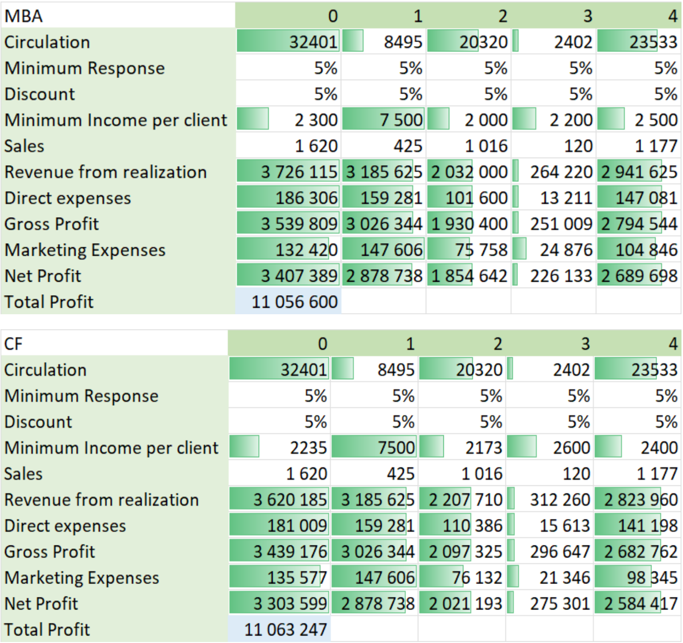
We gathered offer formulations, response probabilities, distributed channels in one final ABT shown at pic 8.1.

As a last step we need to assess the work from the financial point of view. We made a simple financial table which shows the minimum total profit from the campaign under the assumption that 5% of the clients used the offer. Both methods resulted in very similar items. The main difference is in minimum average income per client. In some segments it is bigger for MBA, in some for CF. For CF we can further increase revenue by choosing more marginal categories of items. We can also set a constraint for total marketing expenses so we will form a smaller group of customers for field tests. Overall, the estimated profit exceeds 11 million rubles which is definitely an indicator to attempt deploying the campaign.



id	MBA	MBA	CF_item	CF_ch	CFcheck	CF_che	rating_MBA	rating_CF	is_er	call_m	sms_mba	email_ml
55574955-5	Совершите суммарну	7500	VIP	5700	7500	7500	41.37061187	41.37061187	1	0	0	0
55574853-4	Купите 2 товара в кат	2000	ИГРУШКИ	2000	2200	2200	45.25646195	49.78210814	1	0	0	0
55575349-4	Ваша персональная с	2300	ДЕТСКОЕ ПИТАН	2500	2600	2700	192.7746221	217.9191381	1	0	0	1
55574954-5	Ваша персональная с	2300	ДЕТСКОЕ ПИТАН	1800	2000	2000	1867.879474	1624.243021	1	0	0	1
55574951-5	Совершите покупку н	2500	ТЕКСТИЛЬ, ТРИК	1800	2000	2000	65.07290993	52.05832794	1	0	0	0
55575057-5	Ваша персональная с	2300	ОБУВЬ	1800	2000	2000	193.1882761	167.9898053	1	0	0	1
55574957-5	Купите 2 товара в кат	2000	ИГРУШКИ	1800	2000	2000	12.63228431	12.63228431	1	0	0	0

Pic. 8.1 Final Analytical Base Table



MBA	0	1	2	3	4
Circulation	32401	8495	20320	2402	23533
Minimum Response	5%	5%	5%	5%	5%
Discount	5%	5%	5%	5%	5%
Minimum Income per client	2 300	7 500	2 000	2 200	2 500
Sales	1 620	425	1 016	120	1 177
Revenue from realization	3 726 115	3 185 625	2 032 000	264 220	2 941 625
Direct expenses	186 306	159 281	101 600	13 211	147 081
Gross Profit	3 539 809	3 026 344	1 930 400	251 009	2 794 544
Marketing Expenses	132 420	147 606	75 758	24 876	104 846
Net Profit	3 407 389	2 878 738	1 854 642	226 133	2 689 698
Total Profit	11 056 600				

CF	0	1	2	3	4
Circulation	32401	8495	20320	2402	23533
Minimum Response	5%	5%	5%	5%	5%
Discount	5%	5%	5%	5%	5%
Minimum Income per client	2235	7500	2173	2600	2400
Sales	1 620	425	1 016	120	1 177
Revenue from realization	3 620 185	3 185 625	2 207 710	312 260	2 823 960
Direct expenses	181 009	159 281	110 386	15 613	141 198
Gross Profit	3 439 176	3 026 344	2 097 325	296 647	2 682 762
Marketing Expenses	135 577	147 606	76 132	21 346	98 345
Net Profit	3 303 599	2 878 738	2 021 193	275 301	2 584 417
Total Profit	11 063 247				

Pic. 8.2-8.3 Financial calculations for MBA&CF approaches

9. Conclusion

In conclusion, we have successfully solved both tasks formulated in section 3.1. We constructed ratings by multiplying the probability of return, predicted by response models, and minimum offer check, calculated after offer formulation. For offer formulation we used MBA and CF.

Both methods gave similar results meaning both of them can be successfully deployed in production. Both methods also have potential to bring more profit. With MBA we can try using different item associations and with CF we can try choosing more expensive items. For both methods we can also try increasing offer checks. We can further conduct field testing after which we can collect real response data and analyze it to adjust offers. Overall, many parameters can be further tuned. E-Commerce is a fast-changing sphere so there is plenty room for improvement. However, in terms of these work, we have completed all the tasks and showed that Collaborative Filtering can be used outside recommendation systems.

All of the code can be found at the open GitHub repository[16].

Bibliography

[1] Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How ‘big data’ can make BIG IMPACT: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, *165*, 234–246. https://doi.org/10.1016/j.ijpe.2014.12.031

[2] Fan, S., Lau, R. Y. K., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of Marketing Mix. *Big Data Research*, *2*(1), 28–32. https://doi.org/10.1016/j.bdr.2015.02.006

[3] Artun, O., & Levin, D. (2015). *Predictive marketing: Easy ways every marketer can use customer analytics and Big Data*. John Wiley & Sons.

[4] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington, D.C., 1993, pp. 207 – 216.

[5] Chen, Y.-L., Tang, K., Shen, R.-J., & Hu, Y.-H. (2005). Market basket analysis in a multiple store environment. *Decision Support Systems*, *40*(2), 339–354. https://doi.org/10.1016/j.dss.2004.04.009

[6] Raeder, T., Chawla, N.V. Market basket analysis with networks. *Soc. Netw. Anal. Min.* 1, 97–113 (2011). https://doi.org/10.1007/s13278-010-0003-7

[7] R. Baraglia, F. Silvestri, An online recommender system for large web sites, in: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, 2004, pp. 199–205, doi:10.1109/WI.2004.10158.

[8] D.R. Fesenmaier, U. Gretzel, C. Knoblock, C. Paris, F. Ricci, S. Stabb, H. Werther, A. Zipf, Intelligent systems for tourism, Intelligent Systems 17 (6) (2002) 53– 66.

[9] *Sklearn.preprocessing.StandardScaler*. scikit. (n.d.). Retrieved May 19, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[10] Abdi, H. and Williams, L., 2010. Principal component analysis. Wiley Interdisciplinary

[11] Raschka, S. (n.d.). Home - mlxtend. Retrieved May 19, 2022, from http://rasbt.github.io/mlxtend/

[12] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, pp. 478 – 499

[13] Francesco Ricci and Lior Rokach and Bracha Shapira, Introduction to Recommender Systems Handbook, Recommender Systems Handbook, Springer, 2011, pp. 1-35

[14] *XGBoost documentation*. XGBoost Documentation - xgboost 1.6.1 documentation. (n.d.). Retrieved May 19, 2022, from https://xgboost.readthedocs.io/en/stable/

[15] Google. (n.d.). *OR-tools | google developers*. Google. Retrieved May 19, 2022, from https://developers.google.com/optimization/

[16] *VeggieRiceBalls/optimization-in-e-commerce: DSBA Diploma*. GitHub. Retrieved May 19, 2022, from https://github.com/veggieRiceBalls/Optimization-in-e-commerce