Проект 12 «Разработка рекомендательной системы, основанной на анализе клиентских данных»

Наставник - Наталия Титова

Выполнила - Зыкова Мария

Постановка задачи

Набор данных взят с Kaggle соревнования и содержит около 3х миллионов заказов продуктов от более чем 200 000 пользователей магазина Instacart.

Цели проекта:

- Проанализировать данные о покупках и потребностях покупателей
- Выяснить скрытую связь между продуктами для улучшения перекрестных и дополнительных продаж.
- Выполнить сегментацию клиентов для целевого маркетинга
- Создать модель машинного обучения, чтобы предсказать повторные заказы для покупателей

Общий анализ данных

Общая информация

- 3’421’083 всего заказов
- 49’688 различных продуктов
- 206’209 уникальных клиентов

Три набора:

- prior - заказы всех 206’209 уник. клиентов
- train - 131’209 клиентов
- test - остальные 75’000

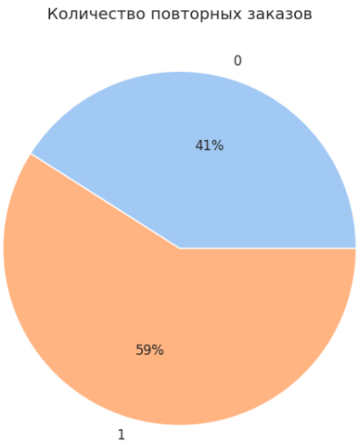
Общая таблица

	order_id	product_id	add_to_cart_order	reordered	product_name	aisle_id	department_id	aisle	department	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
0	2	33120	1	1	Organic Egg Whites	86	16	eggs	dairy eggs	202279	prior	3	5	9	8.0
1	2	28985	2	1	Michigan Organic Kale	83	4	fresh vegetables	produce	202279	prior	3	5	9	8.0
2	2	9327	3	0	Garlic Powder	104	13	spices seasonings	pantry	202279	prior	3	5	9	8.0
3	2	45918	4	1	Coconut Butter	19	13	oils vinegars	pantry	202279	prior	3	5	9	8.0
4	2	30035	5	0	Natural Sweetener	17	13	baking ingredients	pantry	202279	prior	3	5	9	8.0

order_id	*id заказа*	**user_id**	*id клиента*	**department**	*название отдела*
product_id	*id товара*	**eval_set**	*набор данных*	**order_number**	*порядковый номер заказа для этого пользователя*
add_to_cart_order	*порядок добавления товара*	**aisle_id**	*id продуктового прохода*	**order_dow**	*день недели, когда был размещен заказ*
reordered	*был ли перезаказан товар ранее*	**department _id**	*id отдела*	**order_hour_of_day**	*час дня, когда был размещен заказ*
product_name	*название продукта*	**aisle**	*название прохода*	**days_since_prior_order**	*количество дней с момента последнего заказа*

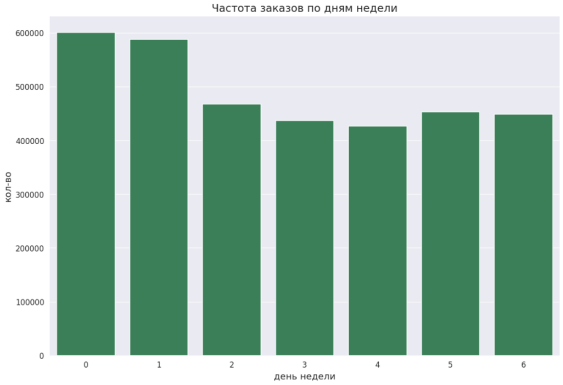
Перейдем к рассмотрению данных более детально

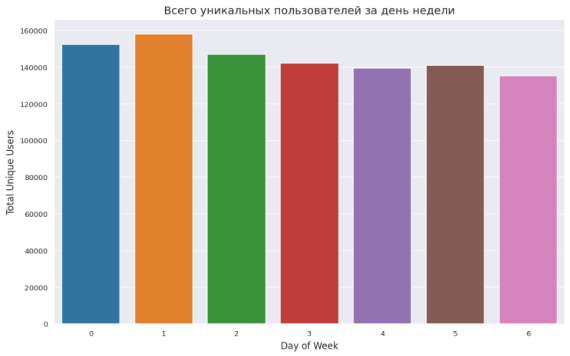
на каждого покупателя приходится минимум 4 заказа, максимум 100. Большая часть клиентов заказывала 4 - 15 раз. Есть 1374 клиентов, сделавшие 100 заказов.



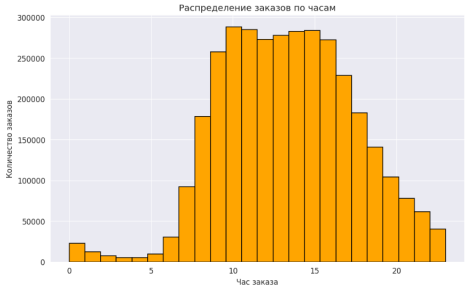
больше часть заказов была перезаказана клиентами. Можно сделать вывод, что пользователи больше склонны покупать те товары, что уже покупали

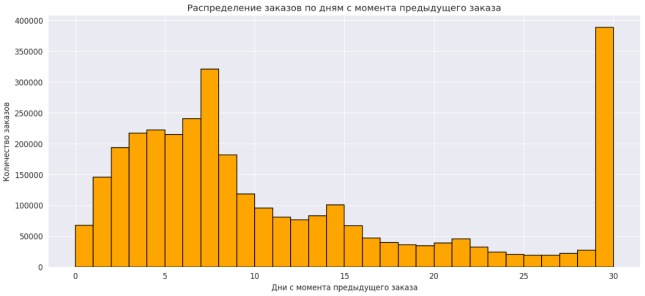
Рассмотрим информацию о времени заказов.





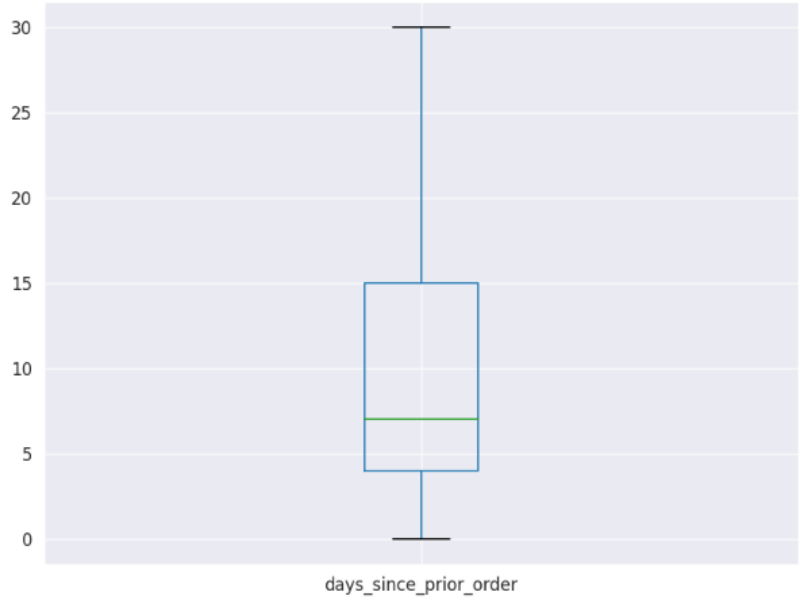
больше всего заказывают в день 0 и 1 - которые скорее всего являются субботой и воскресеньем. Но больше всего покупателей приходится на воскресенье





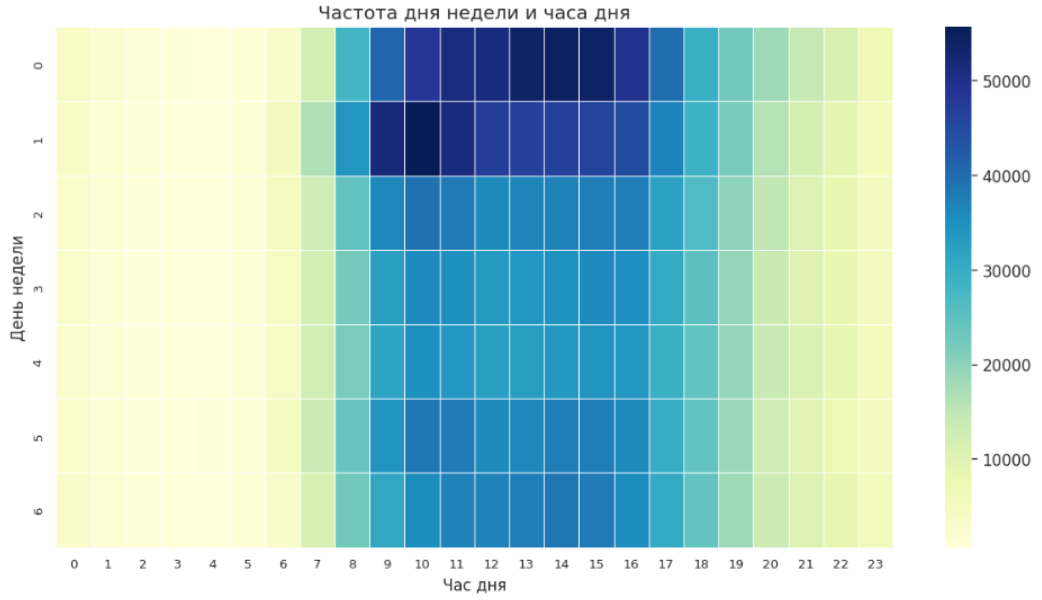
чаще всего люди заказывают с 8 до 18 и склонны заказывать чаще через неделю после предыдущего заказа(подтверждается пиками в 7, 14, 21 и 30 на графике справа)

оценим процентное соотношение с помощью графика ящика с усами

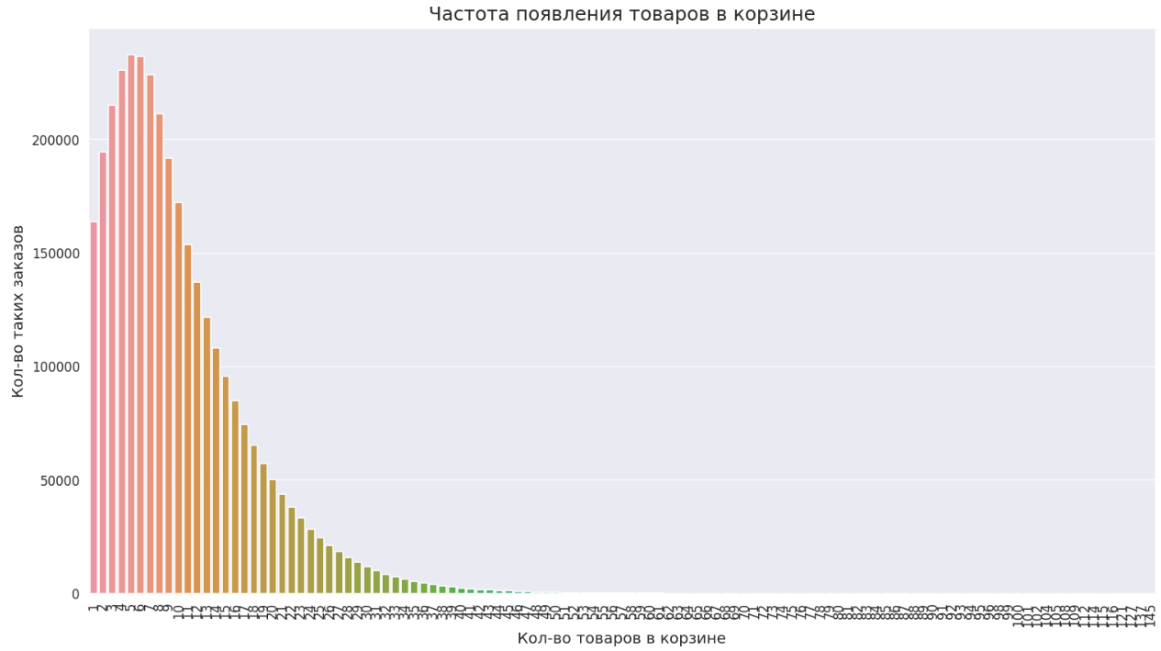


Из приведенного выше графика мы видим, что 25% заказов осуществляют самое большее через 4 дня после их предыдущего заказа. Кроме того, 50% заказов размещаются в период от 4 до 15 дней после их предыдущего заказа.

Рассмотрим связь дня недели и часа дня.

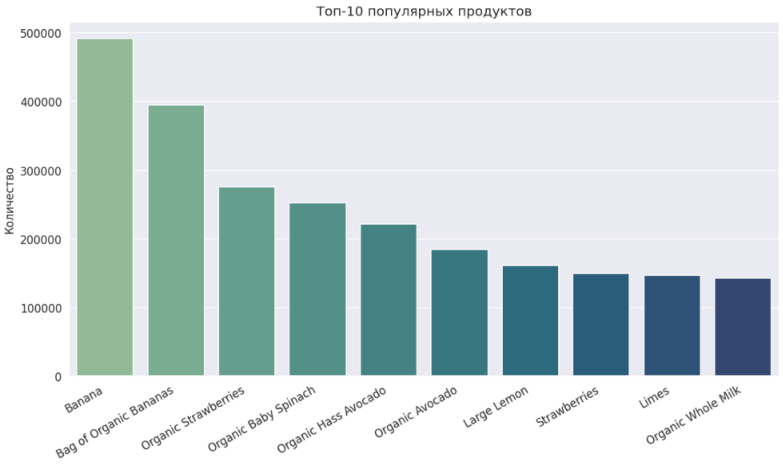


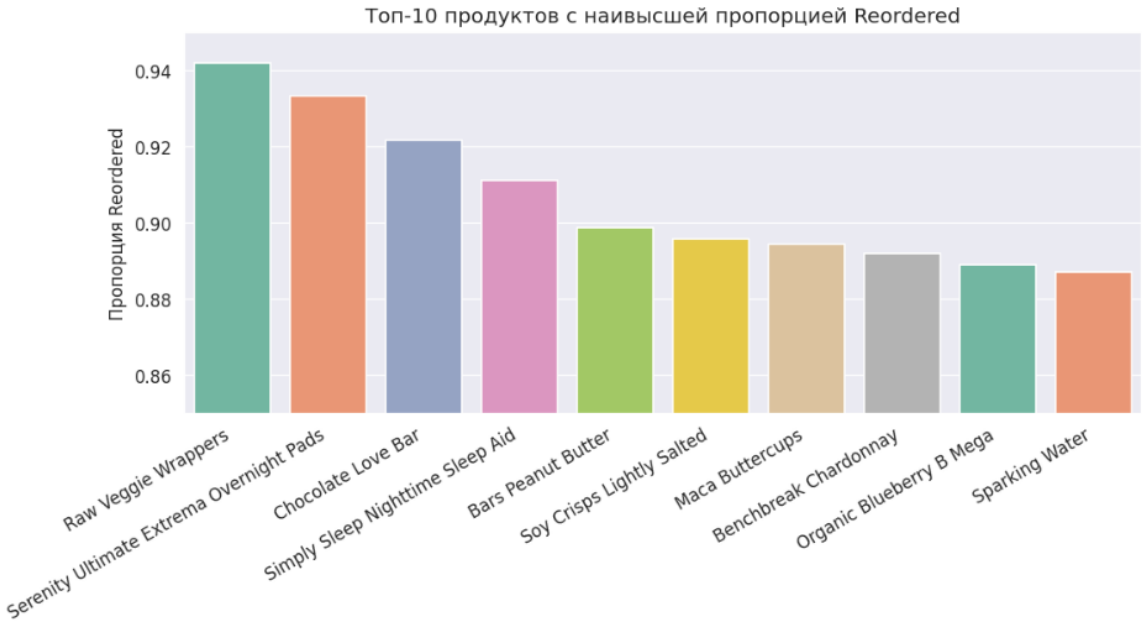
Посмотрим сколько обычно товаров в корзине покупателей.



Чаще всего встречаются заказы в которых от 3 до 8 товаров

Рассмотрим информацию о продуктах



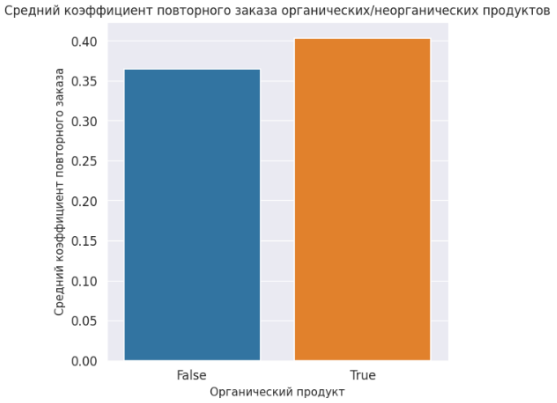


Какой товар люди кладут в корзину в первую очередь?

product_name	pct	count
Экстренная контрацепция	0.792453	42
Rehab Energy Холодный чай с апельсином	0.787879	52
Калифорнийское шампанское	0.777778	14
Ароматизированная водка, Персик	0.708333	51
Каберне Совиньон, Коллекция H3, Horse Heaven Hills	0.7	14
Органический малиновый черный чай	0.692308	27

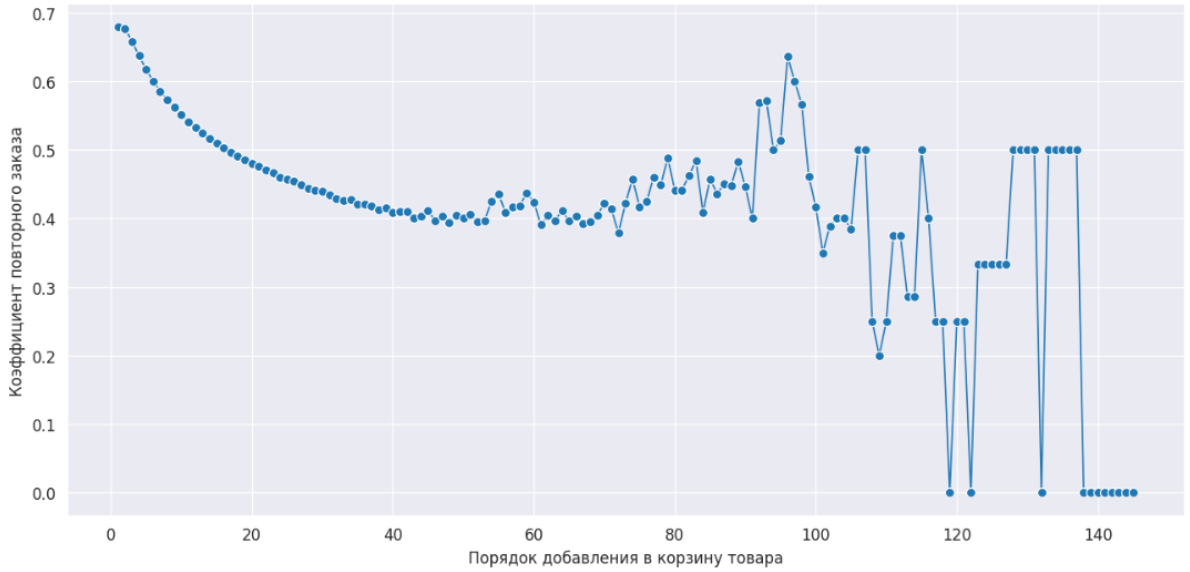
Назальный противозастойный ингалятор лекарственный	0.688889	31
Детская смесь из соевого порошка	0.685714	24
Разливное саке	0.675	27
Детская лекарственная смесь	0.666667	14

А какова доля органических и неорганических товаров?



Доля органических продуктов мала, но их перезаказывают гораздо чаще

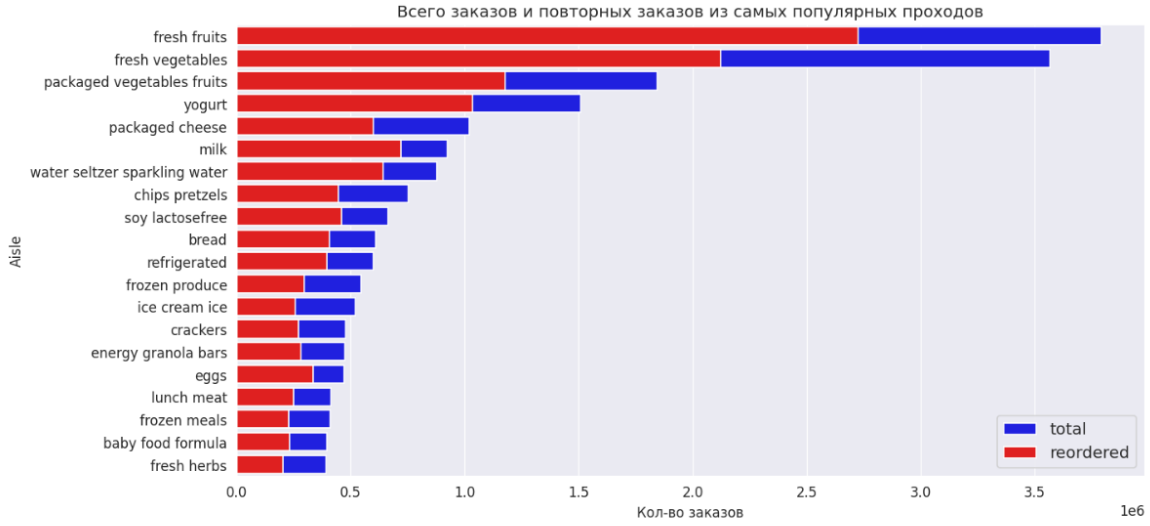
Рассмотрим график заказа на добавление в корзину и средний процент повторного заказа



Чем ниже заказ на добавление в корзину, тем выше процент повторного заказа. Это имеет смысл, поскольку мы обычно сначала покупаем то, что нам необходимо в повседневной жизни.

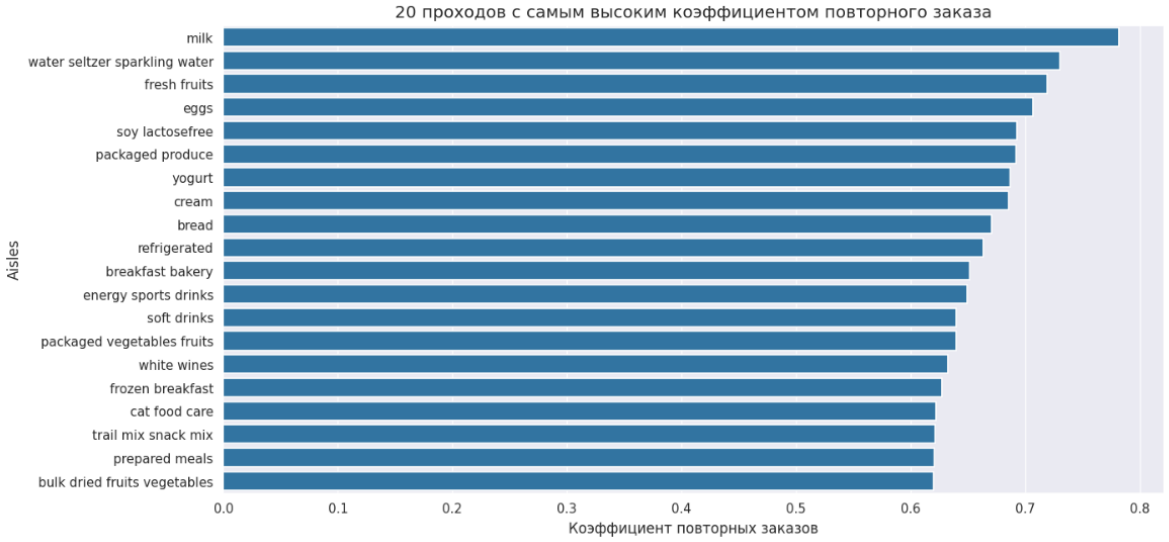
Какие проходы имеют самую большую долю повторных заказов среди всех

график показывает наполнение обычных и повторных заказов среди самых популярных проходов



Это фрукты и овощи, молоко, газированная вода,яйца и так далее

Вот 20 проходов с самым высоким коэффициентом повторного заказа в порядке убывания



а также и проходы с самым низким уровнем повторных заказов

это товары и продукты, в которых нет ежедневной потребности(например, свечи, декор для выпечки, маринад для мяса и тд) и те которые можно купить один раз на длительный срок(витамины, чистящие средства, средства ухода и тд)

Ассоциативные правила для анализа покупательской корзины

Анализ правил ассоциации используется, когда мы хотим найти связь между различными продуктами, чтобы эту информацию можно было использовать для принятия решения, какие продукты следует продавать или продвигать вместе.

Для этой цели я использовала **априорный алгоритм.** Он предполагает, что любое подмножество часто встречающегося набора элементов должно быть частым.

В результате были обнаружены 56 правил ассоциации среди 100 наиболее частых продуктов.

Все правила показали высокий уровень подъема(lift), которая показывает насколько вероятность покупки товара B увеличивается, когда покупают товар A. Мера Lift указывает на сильную связь между товарами.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
35	(Limes)	(Large Lemon)	0.059984	0.065764	0.011860	0.197723	3.006544	0.007915	1.164480	0.709980
34	(Large Lemon)	(Limes)	0.065764	0.059984	0.011860	0.180345	3.006544	0.007915	1.146843	0.714372
52	(Organic Strawberries)	(Organic Raspberries)	0.112711	0.058325	0.014533	0.128940	2.210731	0.007959	1.081069	0.617230
53	(Organic Raspberries)	(Organic Strawberries)	0.058325	0.112711	0.014533	0.249174	2.210731	0.007959	1.181751	0.581582
37	(Organic Avocado)	(Large Lemon)	0.075348	0.065764	0.010538	0.139862	2.126728	0.005583	1.086147	0.572966
36	(Large Lemon)	(Organic Avocado)	0.065764	0.075348	0.010538	0.160244	2.126728	0.005583	1.101097	0.567088
47	(Organic Blueberries)	(Organic Strawberries)	0.042956	0.112711	0.010235	0.238274	2.114024	0.005394	1.164840	0.550621
46	(Organic Strawberries)	(Organic Blueberries)	0.112711	0.042956	0.010235	0.090809	2.114024	0.005394	1.052633	0.593909
49	(Organic Raspberries)	(Organic Hass Avocado)	0.058325	0.090339	0.010966	0.188018	2.081257	0.005697	1.120298	0.551699
48	(Organic Hass Avocado)	(Organic Raspberries)	0.090339	0.058325	0.010966	0.121389	2.081257	0.005697	1.071777	0.571115

Также ниже показаны 10 пар продуктов с наибольшим значением уверенности, которое означает вероятность что будут куплены оба товара.

Например, примерно в 40% случаев яблоко и банан покупались вместе чаще.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
24	(Organic Fuji Apple)	(Banana)	0.037992	0.200938	0.014378	0.378441	1.883367	0.006744	1.285576	0.487559
15	(Honeycrisp Apple)	(Banana)	0.034078	0.200938	0.012122	0.355725	1.770317	0.005275	1.240249	0.450481
12	(Cucumber Kirby)	(Banana)	0.040789	0.200938	0.013432	0.329296	1.638788	0.005236	1.191377	0.406368
20	(Organic Avocado)	(Banana)	0.075348	0.200938	0.022745	0.301866	1.502282	0.007605	1.144568	0.361591
30	(Seedless Red Grapes)	(Banana)	0.035480	0.200938	0.010534	0.296906	1.477596	0.003405	1.136493	0.335115
5	(Organic Raspberries)	(Bag of Organic Bananas)	0.058325	0.161527	0.017294	0.296508	1.835662	0.007873	1.191874	0.483433
3	(Organic Hass Avocado)	(Bag of Organic Bananas)	0.090339	0.161527	0.026487	0.293199	1.815175	0.011895	1.186294	0.493688
32	(Strawberries)	(Banana)	0.061123	0.200938	0.017661	0.288936	1.437931	0.005379	1.123754	0.324384
16	(Large Lemon)	(Banana)	0.065764	0.200938	0.017603	0.267663	1.332062	0.004388	1.091111	0.266832
53	(Organic Raspberries)	(Organic Strawberries)	0.058325	0.112711	0.014533	0.249174	2.210731	0.007959	1.181751	0.581582

Сегментация клиентов

Сегментация клиентов — это процесс разделения клиентов на группы на основе общих характеристик, чтобы эффективно и адекватно продавать нужные товары каждой группе.

Выполнить сегментацию можно, используя данные о том, какие продукты покупают пользователи.

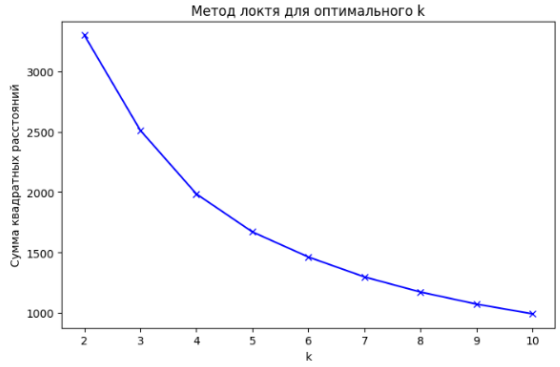
Поскольку существуют тысячи продуктов и тысячи покупателей, я использовала проходы, которые представляют собой категории продуктов, которых всего 134 шт. Это позволит сделать меньше нагрузку на вычисления и повысить качество результата.

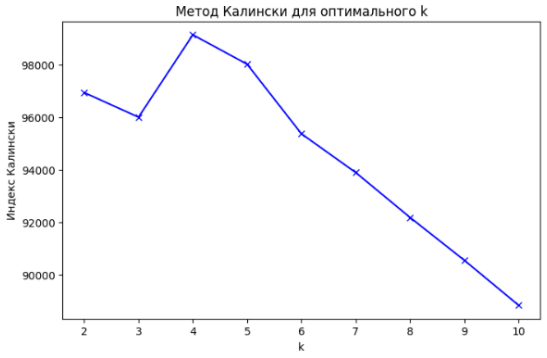
Данные были нормализованы и представлены в нужном формате. Затем я применила анализ главных компонентов, чтобы уменьшить размерности, поскольку алгоритмы кластеризации не дают хороших результатов в более высоких измерениях.

Использован для разделения был метод KMeans, который показал лучшее и чистое разделение

Выбор оптимального K-кластеров при любом количестве основных компонент и кластеров был несколько неоднозначным, но остановилась я на 3х основных компонент.

для поиска кол-ва кластеров использованы были два метода - метод локтя и индекс Калински-Харабаша

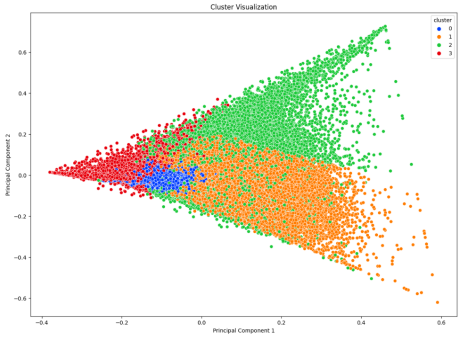




примерно от 4 до 6 кластеров можно посчитать подходящим

4 сегмента оптимально, об этом также говорит пик индекса Калински-Харабаша

Метод KMeans



Метод GaussianMixture

KMeans показал себя лучше

1 кластер:

```
aisle
water seltzer sparkling water    117281
fresh fruits                      19342
soft drinks                       16223
yogurt                            10939
energy granola bars                8730
chips pretzels                     7977
tea                                7473
packaged vegetables fruits         7221
milk                               7144
candy chocolate                    6159
dtype: int64
```

2 кластер

```
aisle
fresh fruits                     1080777
fresh vegetables                  786799
yogurt                            776662
packaged vegetables fruits        644826
packaged cheese                   514999
chips pretzels                    460429
milk                              452604
water seltzer sparkling water     420946
refrigerated                      337412
bread                             317634
dtype: int64
```

3 кластер

```
aisle
fresh vegetables                 2261697
fresh fruits                     1445058
packaged vegetables fruits        775896
yogurt                            408050
packaged cheese                   325278
milk                              265126
fresh herbs                       251391
soy lactosefree                   237669
frozen produce                    210488
water seltzer sparkling water     196434
dtype: int64
```

4 кластер

```
aisle
fresh fruits                     1097011
fresh vegetables                  363646
packaged vegetables fruits        337370
yogurt                            256692
milk                              166141
packaged produce                  142393
packaged cheese                   134588
water seltzer sparkling water     106872
soy lactosefree                   101960
chips pretzels                     87998
dtype: int64
```

Модель машинного обучения для прогнозирования повторных заказов продуктов

Для обучения я решила рассмотреть три распространенные модели машинного обучения - Logistic Regression, Random Forest и XGBoost.

Решается задача классификации, т.е. будет ли перезаказан товар ранее заказаный клиентом.

Чтобы построить модель, данные были подготовлены и извлечены дополнительные признаки такие как:

```
статус повторных заказов последней покупки каждого пользователя
количество каждого продукта
количество повторных заказов со стороны клиента
```

Набор для обучения был составлен из двух множеств данных представленных в датасете - prior и train.

prior содержит информацию о всех клиентах, train крайние покупки 131’209 клиентов. Объединив эти наборы я составила датасет для обучения, содержащий около 8 миллионов данных, который был затем разделен на обучающую и валидационную выборки. На валидационной выборке был подсчитан accuracy, а уже после предсказание сделано на тестовом наборе.

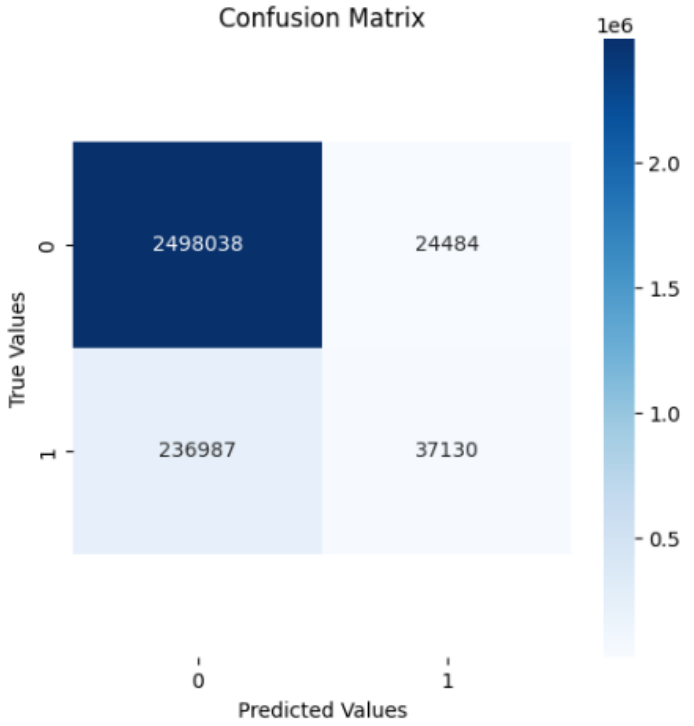
	Logistic Regression	**Random Forest**	**XGBoost**
accuracy	0.9019	0.90547	0.9065
f1-score	0	0.17	0.22
AUC	0.5	0.548	0.563

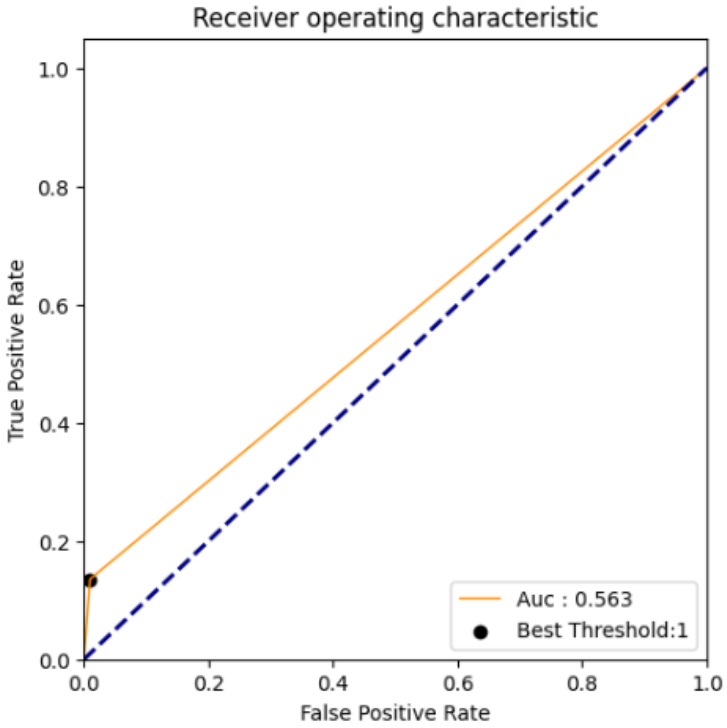
Лучше всего показала себя модель **XGBoost**

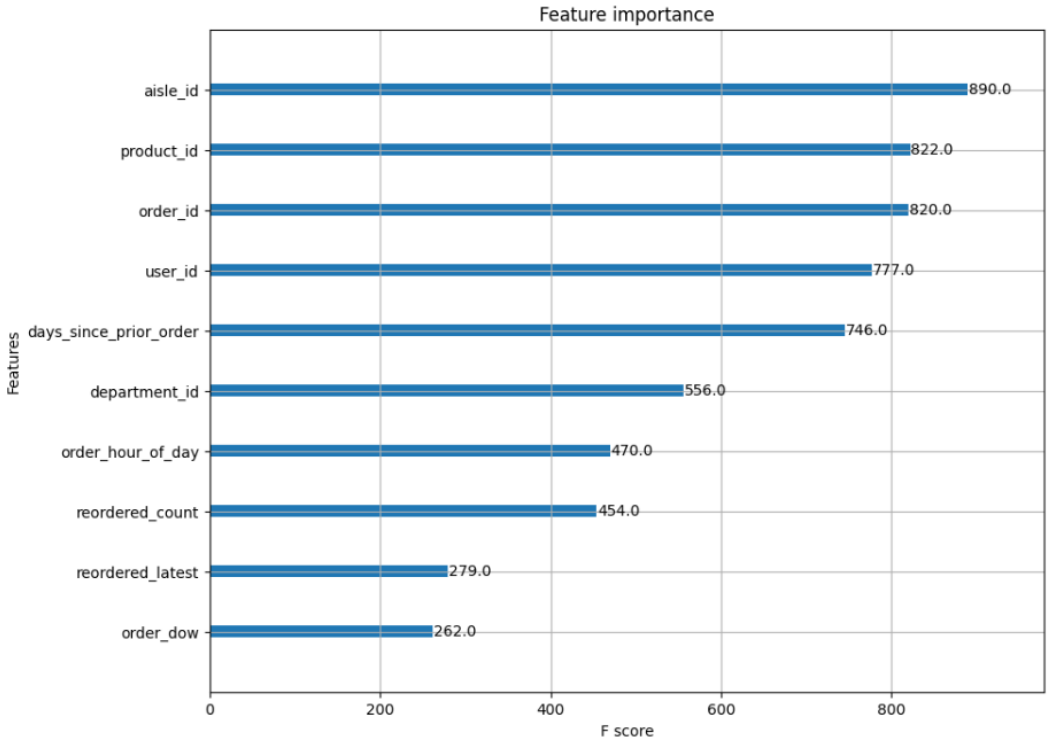
```
Classification report :
              precision    recall  f1-score   support

         0.0       0.91      0.99      0.95   2522522
         1.0       0.60      0.14      0.22    274117

    accuracy                           0.91   2796639
   macro avg       0.76      0.56      0.59   2796639
weighted avg       0.88      0.91      0.88   2796639
```


Confusion Matrix
1e6
2.0
1.5
1.0
0.5
2498038
24484
236987
37130
True Values
0
1
0
1
Predicted Values


Receiver operating characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Auc : 0.563
Best Threshold:1

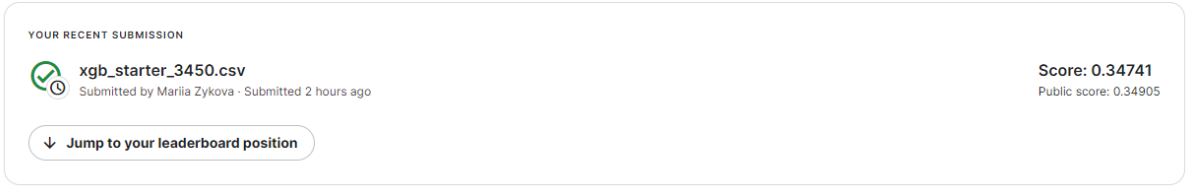


Результат записан в такую таблицу:

	order_id	user_id	order_dow	order_hour_of_day	days_since_prior_order	product_id	reordered_count	reordered_sum	reordered_latest	aisle_id	department_id
3052321	2634088	11335	3	18	8.0	6873	1	0	0.0	83	4
5273294	3379345	104957	2	9	30.0	15923	1	0	0.0	67	20
6564246	1344587	174152	0	7	30.0	45692	1	0	0.0	107	19
7705447	2206969	196469	1	9	20.0	31205	1	0	0.0	26	7
6947171	1103780	57331	1	20	15.0	28599	1	0	0.0	130	14
...	...	...	...	...	...	...	...	...	...	...	...
1965286	1099888	79357	5	12	4.0	8277	9	8	0.0	24	4
4473121	256984	62140	0	20	8.0	42959	1	0	0.0	74	17
8158334	355337	70111	3	6	30.0	24341	1	0	0.0	31	7
6278263	1455864	170500	6	14	29.0	11097	1	0	0.0	30	6
4360443	692804	192171	0	11	7.0	47788	5	4	0.0	24	4

1000 rows × 11 columns

Результат в Kaggle



максимальный скор лидера в данном соревновании был 0.40