# **Описание подхода к кластеризации магазинов**

Данные

* Выделение необходимых признаков и последующая кластеризация магазинов основаны только на истории продаж товаров, поскольку в информации о характеристиках магазинов содержится много пропущенных значений практически по каждому пункту.
* Исходные данные о продажах разделены на 2 отдельных блока: за 2146 и 2147 год. Первый блок данных предназначен для непосредственного выделения кластеров, а второй блок – для проверки качества построенной кластеризации.
* Выбросы в данных никак не обрабатывались, - принято решение нивелировать их влияние с помощью подхода к построению признаков.

Признаки

Рассчитываются совокупно за 1 год; по каждой категории товаров; среднее в расчетах объемов продаж заменено на медиану.

Сконструировано несколько групп признаков, характеризующих:

* размер и загруженность магазина;
* насыщенность товарного ассортимента;
* качество ассортимента;
* частоту и периодичность спроса;
* объем спроса.

Перед дальнейшим использованием из данных отсеиваются признаки с околонулевой дисперсией и производится робастное шкалирование переменных.

Факторный анализ

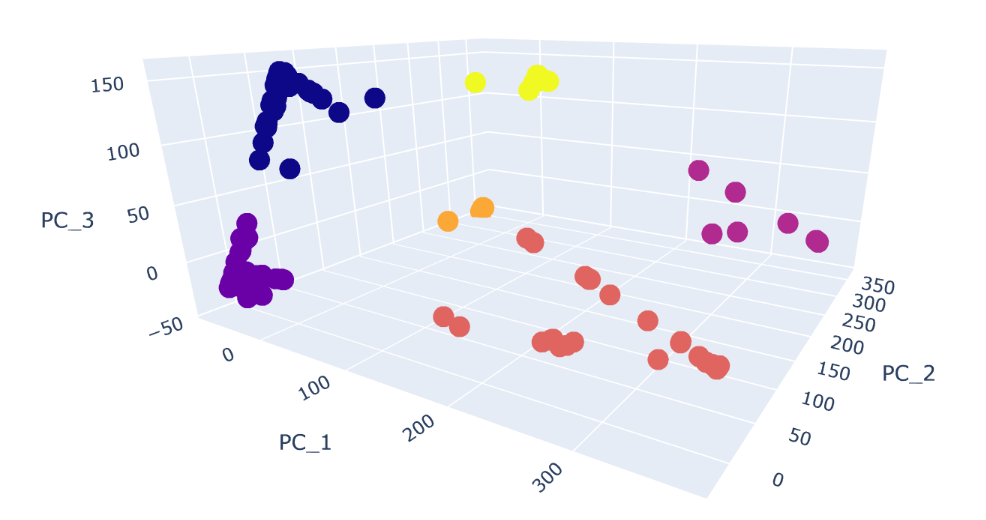
В общей сложности было сконструировано 479 признаков, в связи с чем, для более удобной интерпретации будущих кластеров, было произведено выделение главных компонент с вращением varimax.

Выявлено 3 основных фактора:

* 1-ая компонента: широта востребованного покупателями ассортимента магазина. Большие значения фактора свидетельствуют об узком наборе товаров, продающихся в магазине, в частности: невостребованность хлама, съедобного хлама, брони и одежды, а также патронов, ядер-колы, медикаментов и еды.
* 2-ая компонента: частота и объем продаж солярки. Большие значения фактора свидетельствуют о редких продажах товара.
* 3-я компонента: качество ассортимента магазина. Большие значения фактора свидетельствуют о частых фактах возврата одежды и брони. Также высокие значения компоненты говорят о сниженном уровне продаж патронов, ядер-колы, медикаментов и еды.

Кластеризация

Кластеризация проведена с использованием выделенных факторов, метод - KMeans. Оптимальное количество кластеров было определено по принципу локтя и составило 5 групп магазинов. В процессе визуального анализа было принято решение выделить 6 кластеров для лучшей интерпретации:



1) Наибольший кластер включает 669 магазинов.

* характеризуется широким ассортиментом приобретаемых товаров;
* основной объем продаж приходится на товары: солярка, съедобный хлам, хлам, броня и одежда, а также патроны, медикаменты, ядер-колу;
* низкая частота случаев возврата товаров (броня и одежда).

2) Второй по величине кластер включает 75 магазинов.

* узкий ассортимент приобретаемых товаров;
* основной объем продаж – солярка; низкая частота и объем покупок хлама, съедобного хлама, брони и одежды;
* низкая частота случаев возврата (броня и одежда).

3) Третий кластер - 59 магазинов.

* достаточно широкий ассортимент приобретаемых товаров;
* основной объем продаж - солярка, умеренная частота продаж хлама, съедобного хлама, брони и одежды;
* высокая частота случаев возврата товара (броня и одежда).

4) Четвертый кластер - 18 магазинов.

* широкий ассортимент приобретаемых товаров;
* самые частые покупки хлама, съедобного хлама, брони и одежды, умеренная частота покупок патронов, ядер-колы, медикаментов и еды и самые редкие покупки солярки;
* низкая частота случаев возврата.

5) Пятый кластер - 15 магазинов.

* самый узкий ассортимент покупаемой продукции (в основном, только бензак).

6) Шестой кластер - 9 магазинов.

* достаточно широкий ассортимент приобретаемых товаров;
* частые покупки хлама, съедобного хлама, брони и одежды, редкие покупки солярки, относительно редкие: патронов, медикаментов и ядер-колы;
* частые случаи возврата (броня и одежда).

# **Оценка качества кластеризации**

Устойчивость

Полученных кластеров оценена путем повторной кластеризации магазинов на основе информации за 2147 год. Полученные метки кластеров совпадают примерно на 97%, что говорит об устойчивости выделенных кластеров.

Эффективность

Полученных кластеров по сравнению с делением магазинов просто по городам проверена путем оценки вклада переменной кластера в точность прогнозирования продаж товаров на примере категории "Хлам".

В качестве базовой модели обучения служит Random forest с параметрами max\_depth=30, n\_estimators=150, max\_features='log2' и признаками на основе дат (рабочие/выходные дни, дни недели, месяц) и исходными переменными, доступными из таблицы продаж.

На данных за 2146 год были обучены 2 модели:

* с добавлением признака «город»;
* с добавлением признаков «город» и «кластер».

Анализ точности предсказаний продаж товаров на 2147 год показал, что добавление метки кластера позволило увеличить процент объясненной дисперсии продаж на 4% и увеличить точность прогноза на 14,5% по метрике MAPE.

# **Что можно доработать**

* визуально полученные кластеры могут быть также разделены на несколько новых кластеров, т. е. можно попробовать произвести повторную кластеризацию магазинов уже внутри полученных групп.
* на основе сконструированных переменных можно построить модели для предсказания пропущенных характеристик магазинов. Полученные таким образом характеристики можно использовать в дальнейшем для конструирования новых групп признаков, например, «рейтинг магазина по объемам продаж товара определенной категории в городе», и применить их для кластеризации.