КЛАСТЕРНЫЙ АНАЛИЗ ПОЛЬЗОВАТЕЛЕЙ ДЛЯ ПЕРСОНАЛИЗАЦИИ ПРОДУКТОВ

Воронкин Р.А., Горшков В.И.

Постановка задачи: задача данной работы заключается в применении кластерного анализа пользователей для персонализации продуктов, чтобы повысить эффективность продаж и улучшить опыт пользователей. Задача состоит в том, чтобы определить, какие пользователи имеют схожие характеристики и какие продукты и услуги могут быть наиболее эффективными для каждой группы пользователей.

Цель работы: Целью данной работы является проведение кластерного анализа пользователей для определения групп пользователей с похожими характеристиками для дальнейшей персонализации продуктов.

Используемые методы: Для проведения кластерного анализа пользователей были использованы следующие методы:

1. Подготовка данных - сбор и очистка данных с использованием SQL и Python.

2. Выбор количественных и категориальных признаков для анализа и стандартизация данных.

3. Определение числа кластеров с использованием метода локтя и силуэтного коэффициента.

4. Проведение кластерного анализа с помощью метода K-Means.

5. Оценка качества кластеризации с помощью F-силуэта и анализ с помощью классификации на основе дерева решений.

Результат: Путем проведения кластерного анализа пользователей было обнаружено, что клиенты разделяются на четыре различных кластера на основе сходных характеристик. В первый кластер входят пользователи, которые активно используют продукты и услуги в течение последних 6 месяцев. Во втором кластере находятся пользователи, которые после первоначального использования продуктов и услуг стали использовать их реже. В третьем кластере - пользователи, которые прекратили использование продуктов и услуг, а в четвертом кластере - новые пользователи, которые только начали использовать продукты.

Дальнейший анализ показал, что наиболее перспективной категорией пользователей являются пользователи первого кластера. На основе этого анализа была выработана маркетинговая стратегия, которая была нацелена на поощрение пользователей резидентов к повторному использованию продуктов и услуг**.**

Практическая значимость: Результаты данной работы могут быть использованы

для персонализации продуктов и услуг, что увеличит эффективность продаж и улучшит пользовательский опыт. Пользова- тельская сегментация может помочь компаниям определить, какие продукты и услуги наиболее стратегически важны для каждой группы пользователей, и адаптировать свою маркетинговую стратегию соответствующим образом. Использование кластерного анализа также может помочь в более глубоком понимании нужд и предпочтений пользователей, что в свою очередь может помочь компаниям разработать более эффективные продукты и услуги.

Ключевые слова: Кластерный анализ, кластеризация, анализ данных, методы кластеризации, алгоритм k-средних, график, изображения.

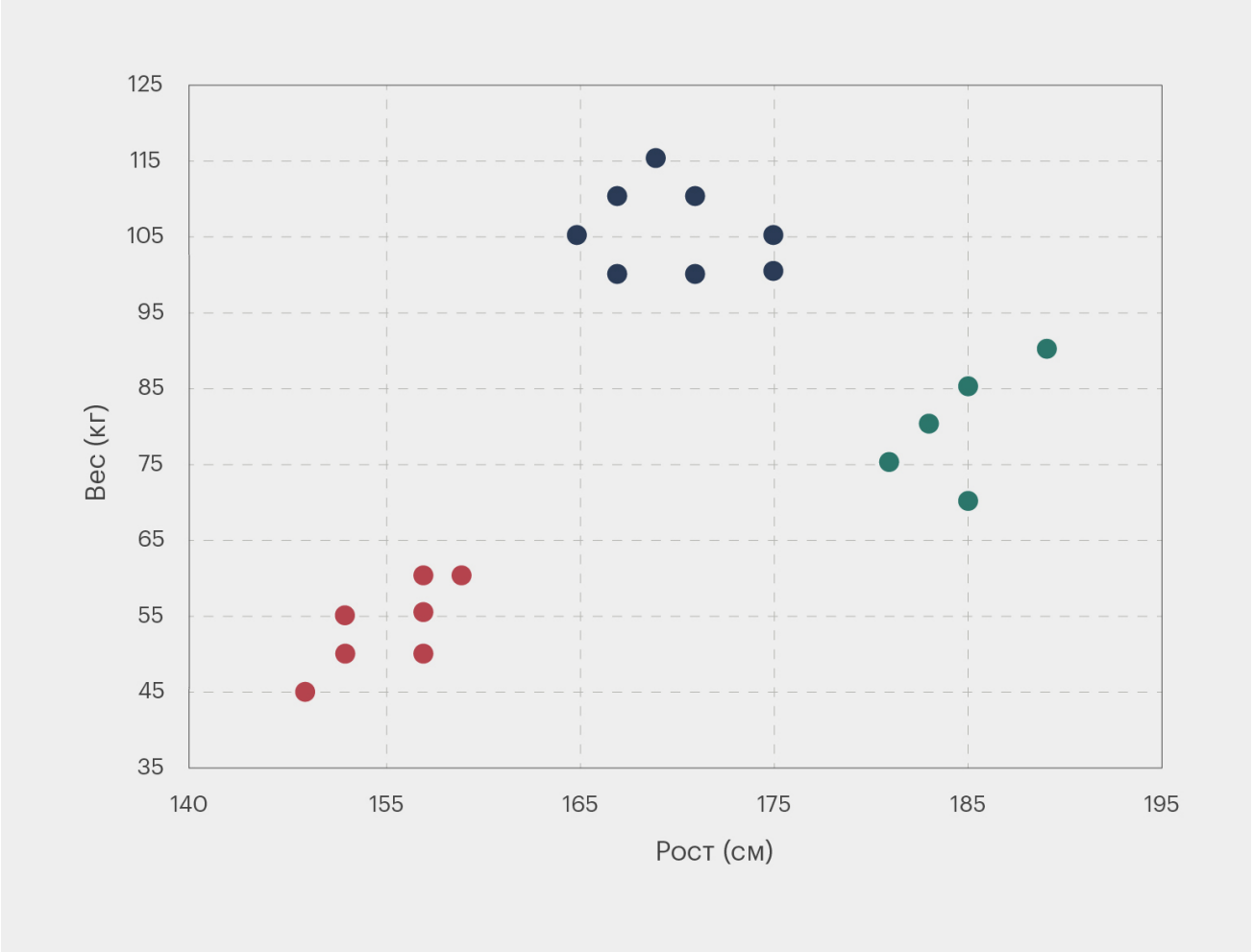
Свою статью я позиционирую, как обобщённую обзорную статью на не- сколько статей взятых с habr.com и других сайтов, вся полученная информация была пропущена через себя и преобразована для дальнейшей работы с ней. Ссылки на статьи приведу в дополнительной литературе.

Основная часть

Как работает кластеризация?

Кластеризация является методом машинного обучения, не требующим контроля и заданных целей. Он позволяет данным самим проявить естественные структуры. Простым примером может служить управление компанией по производству футболок, где данные клиентов – их рост и вес – могут быть использованы для построения графика с ростом по одной оси и весом по другой. Клиенты на графике будут представлены в виде точек, что позволяет группировать их по подобию. Этот метод помогает лучше понять структуру данных и глубже изучить характеристики и потребности групп

пользователей.



Пример кластеризации на основе роста и веса

Зная, что данные клиентов группируются в несколько кластеров, вы можете определить, какие размеры футболок нужно производить в каждом из кластеров, чтобы удовлетворить предъявляемые разными группами потребности. Кластеризация помогает выявить группы из данных, которые могут быть рассмотрены по отдельности для лучшего понимания их особенностей и улучшения процессов управления бизнесом.

Однако, необходимо учитывать, что для получения точных результатов при кластеризации важны как выбранные характеристики признаков, так и методы обучения и анализа данных. Кроме того, результаты кластеризации могут отличаться в зависимости от выбранного количества кластеров и критериев их определения. Поэтому важно использовать правильный подход для определения числа кластеров и критериев, чтобы получить наиболее точные результаты при кластеризации данных.

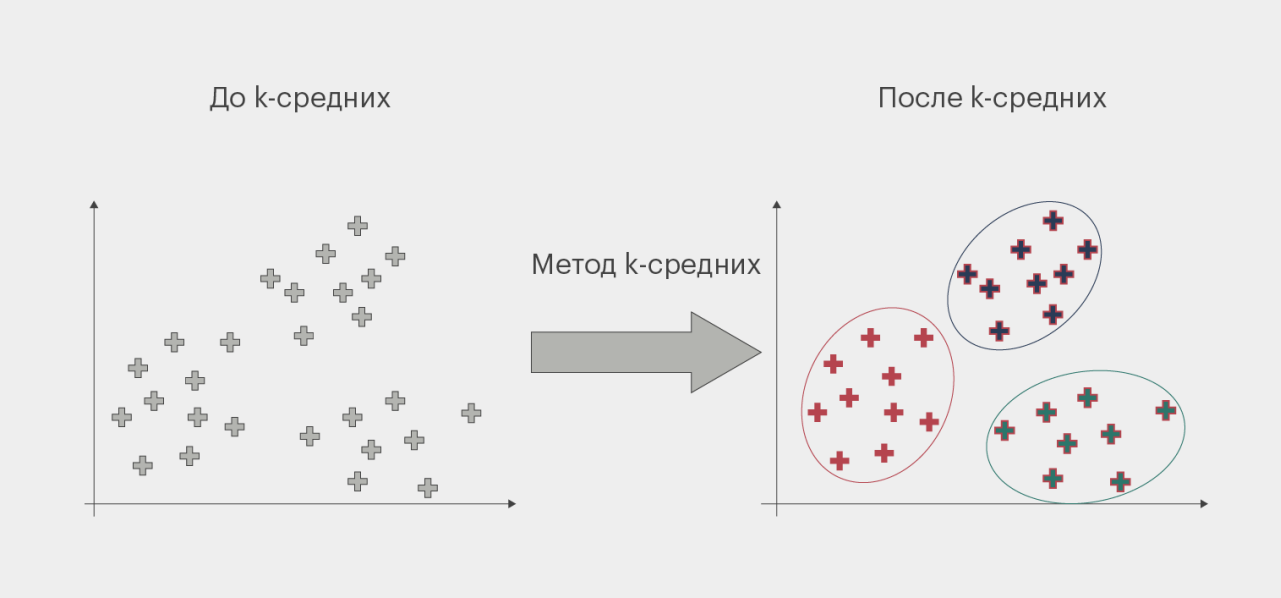
Алгоритм кластеризации тоже «строит» график и помечает на нём точкой каждого клиента, о котором есть информация. Потом рассчитывает расстояние между каждой парой точек. В основе этого расчёта лежит теорема Пифагора: если у вас есть значения x и y двух точек, вы можете рассчитать расстояние между ними.

На основе этих расчётов алгоритм выявляет схожесть точек. Чем меньше расстояние между парой точек, тем больше они похожи. Чем больше расстояние, тем сильнее они различаются. В результате получаются группы, точки в которых находятся близко к друг другу. Группа — это кластер. В кластере находятся клиенты с похожим ростом и весом. Алгоритм окрашивает кластеры в разные цвета, чтобы было понятно, к какой группе относится покупатель.

Если вы используете всего две переменные — такие как рост и вес, то кластерный анализ кажется простым и интуитивно понятным. Если начать добавлять переменные, всё станет сложнее. Тогда можно использовать алгоритм k-средних.

Алгоритм k-средних — метод кластеризации, который позволяет разбивать данные на группы, похожие по пяти, десяти и более признакам. Его идея в том, что кластеризация выполняется не за один раз.

Если простыми словами, то алгоритм работает так. Ему задают, сколько нужно выделить кластеров, и он делает множество подходов (итераций), чтобы найти их. Во время первой итерации он находит две удалённые друг от друга точки и формирует кластеры вокруг них. Во время следующих берёт другие точки и строит новые кластеры. Так он ищет группы точек с наиболее близкими средними значениями. Алгоритм завершается, когда при очередной итерации кластеры не изменяются.



Визуально работу метода k-средних можно представить так

Кластерный анализ впервые ввёл математик Роберт Трион[5] в 1939 году, его монограмма 1939 года *«Кластерный анализ»[6]* была одной из первых работ, в которой был описан метод кластерного анализа . Вообще, кластерный анализ охватывает множество дисциплин, таких как математика, архитектура, маркетинг, археология, медицина, философия, психология

По своей сути Кластерный анализ[1] - это метод многомерного анализа данных в статистике, который позволяет разбивать множество объектов на группы (кластеры) на основе сходства между ними. При кластеризации мы стремимся рассортировать объекты на кластеры таким образом, чтобы объекты внутри кластера были максимально похожи друг на друга, а объекты, относящиеся к разным кластерам, - максимально различались.

Какая роль у маркетолога в этом процессе?

Маркетолог задаёт переменные — показатели, по которым формируют кластеры. Например, это могут быть не «рост» и «вес», а «доход клиента», «возраст», «стоимость покупки» и другие. Также маркетолог описывает кластеры, созданные алгоритмом, и определяет, можно ли использовать полученные результаты.

Маркетолог может экспериментировать: добавлять или удалять переменные и повторно запускать алгоритм. Это позволяет проверить, создаёт ли алгоритм более осмысленные кластеры.

Как использовать метод кластеризации? рассказываем пошагово

Вот что нужно сделать, чтобы провести кластерный анализ.

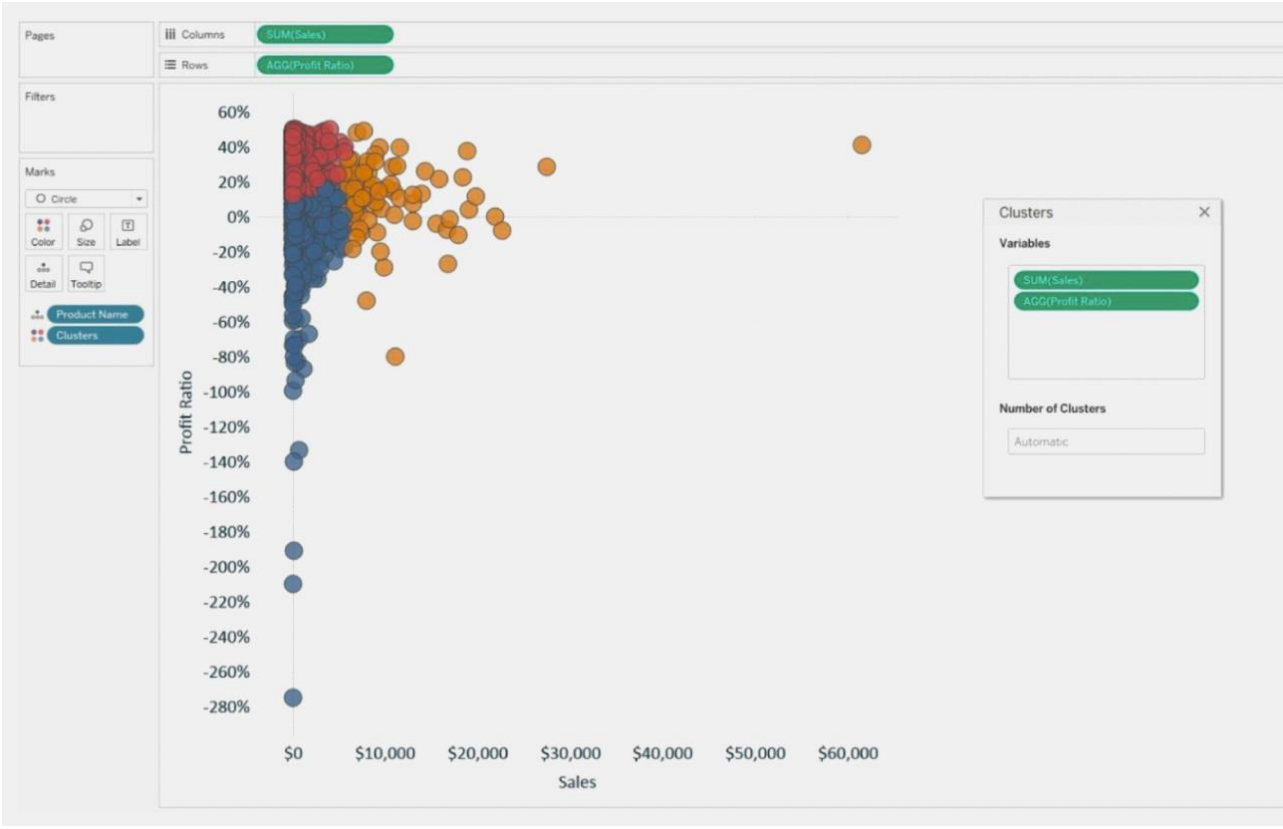
- Подготовить данные. Вы должны убедиться, что у вас есть все необходимые сведения. Это должны быть детализированные данные по каждому клиенту или продукту. Агрегированные данные не подойдут.
- Перевести данные в цифры. Это нужно, чтобы можно было посчитать расстояние между «точками» — объектами, которые нужно кластеризовать. Например, если один из параметров — город, можно присвоить Москве код 402, Санкт-Петербургу — 403 и так далее.
- Объединить данные в хранилище. Это нужно скорее для удобства. Например, можно собрать данные в BigQuery от Google.

Возможно, вам также понадобится преобразовать данные, если они выражены в разных единицах измерения. Например, можно стандартизировать все значения так, чтобы они находились в диапазоне от 0 до 1.

Когда данные обработаны, можно использовать алгоритм. Есть несколько способов кластеризации:

- использовать программный метод — например, если в команде есть специалисты по анализу данных, они могут применять для кластерного анализа языки R или Python;
- обратиться к аналитическим сервисам вроде Tableau — в них есть встроенные инструменты для кластеризации;
- работать с хранилищами данных — например, в BigQuery можно визуализировать результаты, если вы знаете синтаксис языка SQL;
- использовать Excel и считать всё вручную, но это подходит только для небольшого числа объектов — например, если нужно разнести семь объектов с двумя параметрами на две группы.

На изображении ниже видно, как результат кластерного анализа может выглядеть на практике. Это пример из сервиса Tableau, в котором есть функция кластеризации. Большое количество продуктов сгруппированы в три отдельных кластера по цене и рентабельности.



Пример кластеризации товаров на основе их цены и прибыли в Tableau

В алгоритм кластеризации можно включить гораздо больше переменных. Но даже при использовании только двух переменных результат анализа может быть действительно информативным. Например, если вы отвечаете за маркетинг и стратегию, вы можете на его основе определить, какие продукты стоит продвигать в первую очередь, а от каких лучше отказаться.

Таким образом одним из основных преимуществ кластерного анализа[3] является возможность разбивки значительного количества данных на более мелкие группы, которые более подробно и удобно изучать. Кластерный анализ может служить начальным этапом для более глубокого анализа данных, также он может быть использован для поиска подопытных групп в экспериментах или тестировании.

Плюсы и минусы кластерного анализа

Кластерный анализ — не идеальное решение для всех. Вот плюсы и минусы, о которых стоит помнить.

Плюсы:

- данные просто визуализировать и интерпретировать;
- анализ легко масштабировать на миллионы записей;
- система динамичная — если изменить данные, то кластеры тоже изменятся.

Минусы:

- разные выполнения алгоритмов могут давать разные результаты;
- при использовании алгоритма k-средних маркетолог должен заранее определить, сколько кластеров должно быть;
- перед применением кластерного анализа нужно подготовить данные.

Главное о кластерном анализе

- Кластеризация подходит для деятельности, при которой важно разделять данные на группы.
- Если данных мало, кластерный анализ не нужен — можно использовать «Яндекс Метрику» или другие простые аналитические инструменты. Кластерный анализ подходит, когда данных много.
- Кластеризация данных происходит с помощью алгоритмов. Они разделяют все объекты на группы на основе расстояния между «точками».
- Для кластеризации важно подготовить детализированные данные и собрать их в одном месте.
- Кластерный анализ хорош тем, что с его помощью можно легко анализировать большой объём данных и визуализировать их.

Вывод: кластерный анализ пользователей представляет собой мощный инструмент для персонализации продуктов и услуг, улучшения

пользовательского опыта и повышения эффективности продаж. В данной работе был проведен кластерный анализ пользователей с применением метода K-Means, который позволил определить группы пользователей с похожими характеристиками. Дальнейший анализ показал, что наиболее перспективной категорией пользователей являются пользователи первого кластера. Разработанная маркетинговая стратегия была нацелена на поощрение пользователей резидентов к повторному использованию продуктов и услуг. Результаты данной работы могут использоваться компаниями для персонализации продуктов, увеличения эффективности продаж и усовершенствования пользовательского опыта. Важно отметить, что данные, используемые для кластерного анализа, должны быть правильно структурированы и очищены перед проведением анализа для получения точных результатов. Кластерный анализ является динамичным инструментом, который требует постоянного обновления и адаптации на основе изменяющихся потребностей и характеристик пользователей.

Таким образом, компании, которые стремятся улучшить свои продукты и услуги и повысить свою конкурентоспособность на рынке, должны использовать кластерный анализ пользователей как важный инструмент для персонализации продуктов и услуг, улучшения пользовательского опыта и повышения эффективности продаж.

Список используемой литературы:

1. Статья на тему: Кластерный анализ и сегментация. Url: https://tidydata.ru/segmentation (Дата обращения: 26.05.2023)

2. Статья на тему: Как кластерный анализ работает в маркетинге — разбираем методы и алгоритмы на примере. Url: https://skillbox.ru/media/marketing/kak-klasternyy-analiz-rabotaet-v-marketinge-razbiraem-metody-i-algoritmy-na-primere/ (Дата обращения: 26.05.2023)

3. Статья на тему: примеры кластерного анализа в реальной жизни. Url:

https://www.codecamp.ru/blog/cluster-analysis-real-life-examples/ (Дата обращения: 26.05.2023)

4. Статья на тему: кластеризация клиентов. Url: https://newtechaudit.ru/klasterizacziya-klientov-analiz-lichnosti-klienta/ (Дата обращения: 26.05.2023)
5. Статья на тему: Открытый курс машинного обучения. Тема 9. Анализ временных рядов с помощью Python. Url:https://habr.com/ru/companies/ods/articles/327242/ (Дата обращения: 26.05.2023)
6. Статья на тему: Кластерный анализ. Url:https://xn----dtbjkdrhdlujmd8i.xn--p1ai/azbuka/klasternyj-analiz/