МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра теории функций и стохастического анализа

СЕГМЕНТАЦИЯ ПОЛЬЗОВАТЕЛЕЙ С ПОМОЩЬЮ МЕТОДОВ КЛАСТЕРИЗАЦИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

Студентки 4 курса 451 группы
направления 38.03.05 — Бизнес-информатика

механико-математического факультета
Быковой Яны Олеговны

Научный руководитель
доцент, к. э. н. _____ А. Р. Файзлиев

Заведующий кафедрой
д. ф.-м. н., доцент _____ С. П. Сидоров

Саратов 2023

ВВЕДЕНИЕ

Актуальность. В современном мире маркетинг становится все более сложным и конкурентным, что требует от компаний использования новых технологий и инструментов для эффективного решения бизнес-задач. Одним из таких инструментов является машинное обучение, которое позволяет компаниям анализировать огромные объемы данных и получать ценные инсайты для разработки маркетинговых стратегий. Сегодня компании используют машинное обучение для решения разнообразных задач, связанных с маркетингом, таких как прогнозирование спроса, определение целевой аудитории, повышение эффективности рекламных кампаний и многое другое.

В данной дипломной работе будет рассмотрен метод кластерного анализа, который является одним из основных методов машинного обучения для сегментации пользователей по различным критериям, таким как интересы, поведение, демографические характеристики и другие. Кластерный анализ позволяет разбить пользователей на группы схожих между собой характеристик, что позволяет компаниям более точно настраивать свои маркетинговые кампании и повышать эффективность взаимодействия с клиентами.

Целью бакалаврской работы является исследование применения методов машинного обучения, в частности кластерного анализа, для решения задач маркетинга. В работе будет проведен анализ существующих методов кластеризации и разработана методика сегментации пользователей. Будет рассмотрен процесс подготовки данных, выбор модели кластеризации и оценка результатов. Все этапы исследования будут проиллюстрированы на практическом примере с использованием реальных данных.

Таким образом, результаты данной работы будут иметь практическое значение для компаний, которые заинтересованы в эффективном использовании методов машинного обучения для решения задач маркетинга и повышения конкурентоспособности на рынке.

Объект исследования - данные компании, которые содержат информацию о поведении и интересах пользователей в отношении товаров или услуг, а также некоторые личные данные. Для проведения кластерного анализа необходимо выбрать подходящий набор, такой как данные покупателей,

данные о посещениях сайта, данные отзывов и т.д.

Предмет исследования - методы машинного обучения, которые могут использоваться для кластерного анализа, такие как метод K-средних, DBSCAN, Hierarchical clustering.

Задачи. Ниже приведены основные задачи, которые необходимо решить в рамках дипломной работы:

— собрать и подготовить необходимые данные для проведения кластерного анализа. Важно учитывать, что качество и точность полученных результатов зависят от качества данных.

— разработать методику сегментации пользователей на основе кластерного анализа. Для этого необходимо выбрать подходящую модель кластеризации и определить критерии, по которым будут группироваться пользователи.

— провести кластерный анализ и получить сегменты пользователей. Для оценки результатов можно использовать метрики, такие как индекс силуэта или дисперсия внутри кластера.

— проиллюстрировать весь процесс исследования и полученные результаты, используя графики и таблицы. Также необходимо представить код и методики, использованные при проведении кластерного анализа.

Практическая значимость. Результаты данной работы будут иметь практическое значение для компаний, которые заинтересованы в эффективном использовании методов машинного обучения для решения проблем маркетинга и повышения конкурентоспособности на рынке посредством качественной сегментации аудитории для целевых продаж.

Структура и содержание работы. Работа состоит из введения, трёх разделов, заключения, списка использованных источников и трёх приложений. Общий объём работы - 54 страницы.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обосновывается актуальность темы работы, формулируется цель работы и решаемые задачи, отмечается практическая значимость полученных результатов.

В **первом разделе** описываются основные теоретические аспекты, которые помогут понять область предстоящего исследования.

Маркетинг - это процесс управления продукцией, ценами, распространением и продвижением товаров и услуг, с целью удовлетворения потребностей клиентов и достижения целей компании.

Виды маркетинга включают:

1. Исследовательский маркетинг.
2. Сегментационный маркетинг.
3. Продуктовый маркетинг.
4. Ценовой маркетинг.
5. Рекламный маркетинг.
6. Продвижение (promotion) товара.
7. Маркетинговые коммуникации.
8. Дистрибуционный маркетинг.
9. Маркетинговые исследования и аналитика.
10. Маркетинг услуг.

Функции маркетинга включают изучение и анализ рынка, определение целевой аудитории, разработку продукта и стратегии ценообразования, управление продажами и распространением, а также создание и продвижение бренда. В ходе анализа в данной работе будет определяться целевая аудитория.

Сегментация клиентов является важным инструментом маркетинга, который позволяет выделить группы потребителей с похожими потребностями, предпочтениями и характеристиками. Это позволяет компании более точно настраивать свои маркетинговые стратегии и акции, чтобы привлечь и удержать клиентов.

Во **втором разделе** описывается набор данных, выбранных для проведения исследования. Данные были взяты с платформы kaggle. Набор данных состоит из 2000 записей и 8 столбцов:

CustomerID: Уникальный идентификатор покупателя;

Gender: Пол покупателя (мужской или женский);

Age: Возраст покупателя;

Annual Income: Годовой доход покупателя в тысячах долларов;

Spending Score: Оценка трат покупателя, основанная на различных таких факторах как поведение покупателя, расходы и предпочтения (1-100);

Profession: Род деятельности пользователя;

Work Experience: Опыт работы (в годах);

Family Size: Количество членов семьи.

Был проведён разведочный анализ данных, заполнены пропуски, визуализированы некоторые зависимые переменные.

Третий раздел посвящён описанию рассматриваемых методов кластеризации и методам оценок моделей.

В **первом подразделе** третьего раздела реализуется первый рассматриваемый метод - **метод k-средних** (англ. k-means) — это алгоритм машинного обучения для кластеризации данных таким образом, чтобы объекты внутри каждой группы были максимально похожи друг на друга, а объекты из разных групп были как можно более различны.

Перед началом работы установлены следующие библиотеки: pandas, numpy, matplotlib, seaborn, scikit-learn, Kneed. Также, прежде чем приступить к построению модели, необходимо стандартизировать все переменные.

Определять качество данной модели будет коэффициент силуэта или показатель силуэта - это показатель, используемый для оценки качества кластеров, созданных алгоритмом. Баллы за силуэт варьируются от -1 до +1. Чем выше оценка силуэта, тем лучше модель. Оценка силуэта измеряет расстояние между всеми точками данных в одном кластере. Чем меньше это расстояние, тем лучше оценка силуэта. Он также измеряет расстояние между объектом и точками данных в ближайшем кластере. Чем выше это расстояние, тем лучше. Оценка силуэта ближе к +1 указывает на хорошую производительность кластеризации, а оценка силуэта ближе к -1 указывает на плохую модель кластеризации.

К данным был применен метод PCA. PCA - это метод, который помогает нам уменьшить размер набора данных. Когда мы запускаем PCA для фрейма данных, создаются новые компоненты. Эти компоненты объясняют максимальную дисперсию модели. После применения данного метода оценка повысилась до 0,38. С данным результатом можно заявить о целесообразности использования данной модели в составлении стратегий.

При построении модели выделяется оптимальное количество кластеров - 5. На основе всех полученных выводов можно объеденить результаты, чтобы выстроить портрет целевой аудитории по выделенным группам и дать некоторые рекомендации.

- Кластер 0: «Бюджетный покупатель». Низкий доход, низкий рейтинг трат. Этот кластер включает покупателей с низким уровнем дохода и низким рейтингом трат. Возможно, эти покупатели являются более экономичными и ориентированы на покупку необходимых товаров, а не на роскошь. Чтобы привлечь к себе внимание данной группы, магазин может предлагать недорогие товары, регулярные скидки и акции на необходимые товары. Также могут быть полезны рекламные кампании, направленные на экономических покупателей.

- Кластер 1: «Умеренный покупатель». Средний доход, средний рейтинг трат. Этот кластер включает покупателей со средним уровнем дохода и средним рейтингом трат. Эти покупатели могут быть более умеренными в своих покупках и ориентированы на покупку товаров необходимых для их жизни. Магазин может привлечь данную группу покупателей, предлагая разнообразие товаров, которые соответствуют их потребностям и предпочтениям, а также выгодные условия покупки товаров, например, скидки на большие покупки.

- Кластер 2: «Экономичный». Высокий доход, низкий рейтинг трат. Этот кластер включает покупателей с высоким уровнем дохода, но с низким рейтингом трат. Эти покупатели могут быть более экономичными и предпочитать покупать только необходимые товары. Для привлечения данной группы покупателей магазин может предлагать товары по доступным ценам, а также скидки и акции на необходимые товары.

- Кластер 3: «Рискованный покупатель». Низкий доход, высокий рейтинг трат. Этот кластер включает покупателей с низким уровнем дохода, но с высоким рейтингом трат. Эти покупатели могут быть предпочитать более дорогие товары в ущерб своей финансовой стабильности. Для привлечения данной группы покупателей магазин может предлагать рассрочку или кредит на покупку дорогих товаров, а также скидки на такие товары.

- Кластер 4: «Роскошный покупатель» .Высокий доход, высокий рей-

тинг трат. Этот кластер включает покупателей с высоким уровнем дохода и высоким рейтингом трат. Эти покупатели могут быть более склонны к роскоши и покупке более дорогих товаров. Для привлечения данной группы покупателей магазин может предлагать эксклюзивные товары и услуги, а также роскошные товары по выгодным ценам, например, на распродажах.

Во втором подразделе реализуется следующий метод - DBSCAN (Density-Based Spatial Clustering of Applications with Noise) — это алгоритм кластеризации данных, который позволяет идентифицировать кластеры на основе плотности точек в пространстве данных. В отличие от метода k-средних, который требует указания числа кластеров заранее, DBSCAN может автоматически определять количество кластеров и обнаруживать выбросы.

Основная идея метода DBSCAN заключается в том, чтобы найти области в пространстве данных, где плотность точек выше определенного порога. В DBSCAN есть два основных параметра: радиус эпсилон и минимальное количество точек в радиусе (MinPts). Если точка имеет MinPts соседей в радиусе эпсилон, то она считается основной точкой, и все точки в радиусе эпсилон от нее считаются принадлежащими к ее кластеру. Точки, которые находятся вне радиуса эпсилон от всех основных точек, считаются выбросами.

Обучение модели DBSCAN. Можно обучить модель DBSCAN. Для этого импортируется класс DBSCAN из библиотеки scikit-learn и задаем параметры алгоритма. Далее, устанавливается радиус эпсилон (eps) равным 5 и минимальное количество точек в радиусе (min samples) равным 10. Эти значения можно настроить в соответствии с конкретными требованиями проекта. После того, как мы получили метки кластеров, мы можем визуализировать результаты. Для этого используется библиотека matplotlib. После, визуализируются кластеры по осям Annual Income (k$) и Spending Score (1-100). Каждый кластер обозначен уникальным цветом. Также добавлены метки осей x и y. После того, как были визуализированы кластеры, можно проанализировать результаты.

Параметры алгоритма были заданы следующим образом: eps = 0.5, min samples = 5. Были получены 4 кластера. В данном примере мы можем сделать следующие выводы на основе кластеризации методом DBSCAN:

- Кластер -1: шумовые точки, которые не принадлежат ни одному кла-

стеру.

- Кластер 0: покупатели с низким доходом и низким возрастом.

- Кластер 1: покупатели с высоким доходом и высоким возрастом.

- Кластер 2: покупатели со средним доходом и средним возрастом.

Магазин может привлечь покупателей из кластера 0, предлагая товары по более низким ценам, а покупателей из кластера 2 - предлагая разнообразие товаров, которые соответствуют их потребностям и предпочтениям, а также выгодные условия покупки товаров, например, скидки на большие покупки. Для привлечения покупателей из кластера 1 магазин может предлагать роскошные товары по выгодным ценам, а покупателям из кластера 3 - рассрочку или кредит на покупку дорогих товаров, а также скидки на такие товары.

Хотя алгоритм DBSCAN показал средние результаты в этом примере, в каждом конкретном случае необходимо тщательно выбирать гиперпараметры и оценивать качество полученной кластеризации. Например, слишком большое значение параметра eps может привести к объединению несвязанных точек в один кластер, тогда как слишком маленькое значение eps может привести к получению большого количества мелких кластеров. Оценка модели - 0,223.

Третий подраздел посвящен реализации метода иерархической кластеризации - это один из методов кластеризации данных, который позволяет группировать наблюдения (объекты, элементы) в кластеры на основе их сходства друг с другом. Метод иерархического анализа строит дерево кластеров, называемое дендрограммой, которое показывает, как кластеры объединяются на каждом уровне иерархии. Существуют два типа иерархической кластеризации: агломеративный и дивизионный.

Преимуществами метода является то, что он требует предварительного задания количества кластеров, так как иерархия формируется автоматически, позволяет обнаруживать кластеры различных размеров, форм и плотностей и дает наглядную визуализацию иерархии кластеров с помощью дендрограммы.

Получившиеся кластеры:

-Кластер 1 с низким доходом и низким уровнем расходов: В этом кластере находятся клиенты с низким годовым доходом и низким уровнем расхо-

дов, как правило большая семья, средний возраст. Они могут быть отнесены к экономически менее обеспеченным группам, которые ограничены в своих возможностях для трат.

-Кластер 2 с высоким доходом и низким уровнем расходов: В этом кластере находятся клиенты с высоким годовым доходом, но низким уровнем расходов. Это может указывать на то, что эти клиенты тщательно контролируют свои расходы или имеют другие финансовые обязательства, из-за которых они ограничены в трате, как правило более взрослый сегмент.

-Кластер 3 с средним доходом и средним уровнем расходов: Этот кластер содержит клиентов со средними значениями годового дохода и уровня расходов. Они могут быть среднестатистическими клиентами, которые имеют некоторую свободу в трате, не находясь в крайности низкого или высокого уровня дохода и расходов.

-Кластер 4 с высоким доходом и высоким уровнем расходов: В этом кластере находятся клиенты с высоким годовым доходом и высоким уровнем расходов. Они могут быть клиентами, которые имеют возможность тратить больше денег, их финансовые возможности позволяют им позволить себе роскошь и дорогие покупки.

По подсчётам, оценка качества модели кластеризации метрика силуэта равна 0.559.

Третий раздел посвящен сравнению трёх используемых методов и подведению итогов.

По итогам сравнения, иерархическая кластеризация демонстрирует наивысшую оценку силуэтного коэффициента (0.558), это говорит о том, что кластеры имеют некоторую степень различия между собой и могут быть полезны для разработки целевых маркетинговых стратегий.

Были даны рекомендации к продвижению товаров среди данных групп:

Кластер 1 :

Предлагать товары и услуги с низкой ценой, акции, скидки и специальные предложения, чтобы привлечь их внимание. Рекламировать продукты, которые предоставляют экономическую выгоду, помогают сэкономить деньги или предлагают доступные варианты. Обратить внимание на товары, которые соответствуют потребностям семей с детьми или предлагают практичные

решения для повседневной жизни.

Кластер 2 :

Предлагать уникальные и эксклюзивные товары, которые подчеркивают статус и индивидуальность клиентов этой группы. Акцентировать внимание на высоком качестве, долговечности и престиже товаров. Предлагать решения для оптимизации финансовых обязательств и инвестиционные возможности.

Кластер 3 :

Разнообразить ассортимент товаров, чтобы удовлетворить различные потребности и предпочтения этой группы клиентов. Предлагать товары и услуги с разумной ценой и хорошим соотношением цена-качество. Использовать персонализированный маркетинг и рекомендации, чтобы предложить клиентам товары, основываясь на их предыдущих покупках.

Кластер 4 :

Рекламировать роскошные и дорогие товары, подчеркивая их эксклюзивность и превосходство. Организовывать специальные мероприятия, VIP-сервисы и предложения, чтобы привлечь внимание и создать чувство привилегии у клиентов этой группы. Сотрудничать с престижными брендами и предлагайте коллекции и ограниченные выпуски товаров.

В **заключении** подведены итоги бакалаврской работы, расписаны задачи, которые были выполнены.

Основные результаты

1. Определены основные понятия, необходимые для проведения исследования и постороения моделей. Проведена предварительная работа с данными.

2. Разработан программный код для реализации кластерного анализа данных различными методами машинного обучения (k-means, DBSCAN, иерархический кластерный анализ).

3. Интерпритированы результаты исследования. Проведен сравнительный анализ и дана индивидуальная оценка для всех моделей.

4. Выделены основные атрибуты каждого из выявленных сегментов. Выстроены образы вокруг атрибутов каждого кластера и даны рекомендации для маркетинговых стратегий.