Математическое, алгоритмическое и программное обеспечение

DOI 10.24412/2221-2574-2025-3-47-53 УДК 004.85

ПРИМЕНЕНИЕ МАШИННОГО ОБУЧЕНИЯ ДЛЯ СЕГМЕНТАЦИИ ПОЛЬЗОВАТЕЛЕЙ В МАРКЕТИНГОВЫХ ИССЛЕДОВАНИЯХ

Платонова Алла Сергеевна
кандидат технических наук, доцент кафедры физики и прикладной математики Муромского института (филиала) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».
E-mail: allaplatonova@inbox.ru

Рыжкова Мария Николаевна
кандидат технических наук, доцент, декан факультета информационных технологий и радиоэлектроники Муромского института (филиала) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых».
Адрес: 602264, Российская Федерация, Владимирская обл., г. Муром, ул. Орловская, д. 23.

Аннотация: В статье представлены результаты проектирования системы, которая на основе методов искусственного интеллекта автоматизирует сегментацию клиентских данных по уровню покупательской лояльности, что позволяет впоследствии осуществлять персональную работу с различными группами клиентов, а также проводить аналитику статистики покупок. Сегментация клиентов на группы выполнена на основе широко известного метода rfm-анализа. За автоматизацию сегментации клиентов на группы отвечает метод машинного обучения – кластеризация, реализованная в работе с помощью алгоритма k-means. Созданная информационная система реализует персональную работу с покупателями в формате почтовой рассылки персонализированных предложений.

Ключевые слова: искусственный интеллект, машинное обучение, кластеризация, проектирование данных, алгоритмическое обеспечение, RFM-анализ, маркетинг, сегментация покупателей.

Введение

Основная задача машинного обучения заключается в обучении искусственного интеллекта на основе предоставленной ему информации самостоятельно принимать решения, самообучении и постоянном совершенствовании в своём самообучении [1]. Машинное обучение — самая эффективная технология обработки больших массивов данных. В её рамках выполняется задача по поиску в полученных эмпирическим путём данных чётких закономерностей, следуя которым даются прогнозы и выстраиваются обучающие систему алгоритмы [2]. Сферы применения машинного обучения очень обширны и имеют огромные перспективы, поскольку искусственный интеллект стал доступным, а объёмы данных непрерывно увеличиваются.

Одним из классов задач машинного обучения является кластеризация, представляющая собой группировку данных в кластеры на основе общих характеристик. Алгоритмы кластеризации разделяют данные на отдельные группы таким образом, чтобы элементы в одной и той же группе (кластере) были более похожи друг на друга, чем элементы в других группах [3]. Например, в сфере экономики решение задач кластеризации необходимо при поиске новых рынков сбыта, формировании специальных предложений для выделенных групп клиентов с учётом их потребностей, интересов, возраста, выявлении случаев подлога в страховании и т.п.

Согласно результатам международного ис-

следования Microsoft 94% руководителей считают, что технологии искусственного интеллекта важны для решения стратегических задач их организаций. При этом 27% опрошенных уже внедрили соответствующие технологии в ключевые бизнес-процессы, ещё 46% ведут пилотные проекты [4].

К основным тенденциям автоматизации маркетинга в 2023 году относятся инфографика и изображения, чат-боты, воронки продаж, сегментация кликов и персонализированный контент для электронных писем, A/B-тестирование, веб-сайты, ориентированные на мобильные устройства, всплывающее уведомление, многоканальный маркетинг.

В настоящее время одним из самых эффективных трендов автоматизации маркетинга считается сегментирование [5].

Для бизнеса машинное обучение станет «технологией века» благодаря возможности решения сложных аналитических задач. Риски и детали, которые традиционные методы аналитики не выдают, теперь будут заранее известны. В будущем возможно будет просчитать, кто из потенциальных покупателей готов к покупке, кто хочет уйти, кому требуется дополнительная информация о товаре или услуге [6].

Успешность того или иного предприятия основывается не только на производстве качественного продукта и его рекламе, но и зависит от лояльности клиентской базы, которая, в свою очередь, напрямую влияет на объёмы продаж и ожидаемый уровень прибыли. С целью определения лояльности аудитории аналитики прибегают к анализу огромного значения данных о покупателях и их транзакциях. Затем на основе результатов анализа маркетологами выстраивается дальнейшая работа с разными группами клиентов.

Для анализа лояльности клиентов маркетологи применяют сегментацию. Согласно работе R. Shirole сегментация клиентов основана на обнаружении важных дифференциаторов, которые разделяют клиентов на целевые группы [7]. В этих целях широко используется RFM-анализ. В работе Р. Кьяси данному методу даётся следующее описание: «Модель RFM была предложена Хьюзом в 1994 году и используется в маркетинге уже в нескольких десятилетий. Эта модель определяет поведение клиента и представляет характеристики поведения клиента с помощью трёх переменных» [8]. RFM-анализ — способ ранжирования клиентов по трём показателям, начиная от самых активных, приносящих наибольшую прибыль, заканчивая самыми неактивными покупателями. Параметрами RFM-анализа являются Recency (R) — как давно клиенты покупали, Frequency (F) — как часто совершались покупки, Monetary (M) — общая стоимость совершенных покупок [9].

Автоматизация сегментации клиентов возможна с помощью алгоритмов кластеризации. Самый классический и самый популярный алгоритм кластеризации — это метод k-средних (K-means). Метод k-средних — это метод кластерного анализа, целью которого является разделение m наблюдений на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру (центроиду) которого оно ближе всего. Данный метод кластеризации предполагает заведомо известное количество кластеров, на которые будут группироваться данные. В рамках RFM-анализа, кластеризации подлежат рассчитанные для всех клиентов R, F, M-параметры. Результатом кластеризации служит формирование нескольких групп покупателей, в каждой из которых находятся клиенты с близкими (похожими) данными. Оптимальное количество кластеров можно определить «методом локтя», с помощью коэффициента «силуэт» и т.д. Но для проведения маркетинговой работы это число достаточно выбрать равным четырём. Число небольшое, но все же будем делить покупателей на четыре простых и интуитивно понятных группы.

1) Лояльные — это стратегически важные клиенты, которые регулярно совершают покупки.

2) Киты — взаимодействуют с компанией нерегулярно, но при этом тратят большие суммы.

Рис. 1. Контекстная диаграмма

3) Новички — клиенты, недавно впервые сделавшие покупку.

4) Потерянные — давно не совершавшие покупок клиенты.

На сегодняшний день на коммерческом рынке представлены следующие аналитические сервисы: Google Analytics, Adobe Analytics, SAP CRM, Microsoft PowerBI, HubSpot CRM и другие. Например, ключевыми характеристиками решения SAP CRM являются управление маркетинговыми ресурсами, управление кампаниями по всем каналам коммуникации, управление лидами, управление сегментами и списками, управление лояльностью клиентов, маркетинговая аналитика, внутренний маркетинг с управлением предложениями в реальном времени [10]. Анализ существующих систем показывает, что в большинстве своём для небольших организаций они имеют избыточный функционал. Кроме этого, чаще всего правообладателями данных инструментов являются зарубежные разработчики.

Цель работы — проектирование информационной системы, которая автоматизирует сегментацию клиентов.

Создаваемая система должна решать следующие задачи.

1. Импорт в систему данных (загрузка данных клиентской базы в стандартных табличных форматах, таких как: CSV, XML, XLSX).

2. RFM-анализ (анализ клиентов по повторяемости, частоте и стоимости покупок).

3. Кластерный анализ (сегментация клиентов на основе RFM-параметров).

4. Визуализация обработанных данных (предоставление интерактивных диаграммы и графиков).

5. Предоставление функционала для проведения маркетинговой работы с различными группами клиентов.

В данной статье описываются результаты проектирования данных и разработки алгоритмического обеспечения.

Проектирование системы

Этап разработки информационной системы, следующий после проведения предпроектного исследования, — этап проектирования данных. Одной из методологий проектирования является методология структурного анализа и проектирования SADT. Функциональная модель SADT отображает производимые объектом действия и связи между этими действиями [11].

Для создания функциональной модели системы, отражающей структурированное изображение функций производственной системы или среды, а также информации и объектов,

Таблица 1. Входные данные для системы

Название поля в системе	Описание данных	Тип данных
customer_id_col	Идентификатор клиента	Integer
Invoice_date_col	Дата покупки	Date
quantity_col	Количество товара	Integer
unit_price_col	Цена товара	float
transaction_code_col	Номер транзакции	Integer
email_col	Персональные данные	String

связывающих эти функции [12], в данной работе используется нотация IDEF0 методологии SADT. Полученная контекстная диаграмма позволяет определить входные, выходные данные, информацию, управляющую действиями работы, и ресурсы, выполняющие её (рис. 1).

На рисунке представлены взаимодействие основных программных компонентов системы, входные и выходные данные, а также управление и ресурсы. На вход компоненты загрузки и обработки данных должен поступать файл формата .xlsx или .xls, содержащий следующие данные (таблица 1).

Компонент загрузки и обработки данных отвечает за корректность загружаемых данных в систему и работает по следующему алгоритму:

1) получает на вход файл с данными о клиентах,

2) проверяет, соответствует ли расширение файла формату .xsl, .xlsx:

- если нет, то выдаёт сообщение об ошибке,
- иначе переход к шагу 3,

3) считывает данные из файла,

4) проверяет целостность данных:

- если есть пустые строки, то удаляет их,
- если нет, то переходит к шагу 5,

5) проверяет типы данных в строках,

- если не соответствуют требуемым, то выполняет преобразование типов,
- иначе переходит к шагу 6,

6) сохраняет обработанные данные,

7) передаёт данные компоненту RFM-анализа.

Компонент RFM-анализа отвечает за определение параметров для каждого клиента, которые рассчитываются по следующим формулам:

1. *recency*:

$$recency(c) = d - c,$$

где d — самая поздняя дата в таблице; c — дата самой последней покупки клиента.

2. *frequency*:

$$frequency(c) = \sum_{t \in T_c} t,$$

где T_c — набор, содержащий транзакции, совершённые конкретным клиентом; t — идентификатор транзакции в заданном наборе.

3. *monetary*:

$$monetary(c) = \sum_{t \in T_c} amount(t),$$

где $amount(t)$ — функция возвращает сумму, потраченную на транзакцию, идентифицированную по её идентификатору.

Для каждого клиента создаётся отдельная таблица «rfm», содержащая эти три рассчитанных параметра, и эта таблица передаётся на вход компоненты кластеризации данных (таблица 2).

За реализацию кластеризации в компоненте отвечает библиотека Python – Sklearn метод KMeans. Алгоритм кластеризации имеет следующий вид:

Шаг 1. Инициализируется k количество кластеров.

Шаг 2. Каждому кластеру случайным образом присваивается центроид — центр начального кластера.

Шаг 3. Все данные выборки разделяются на k кластеров, центроиды кла-

Таблица 2. Входные данные для компонента кластеризации

Источник данных	Название поля в системе	Описание данных	Тип данных
Загруженные входные данные	customer_id_col	Идентификатор клиента	Integer
Таблица «rfm»	‘Recency’	Recency	Integer
	‘Frequency’	Frequency	Integer
	‘Monetary’	Monetary	float

Таблица 3. Входные данные для компонента работы с клиентами

Источник данных	Название поля в системе	Описание данных	Тип данных
Загруженные входные данные	customer_id_col	Идентификатор клиента	Integer
	email_col	Персональные данные	String
Таблица «rfm»	'Segment'	Название группы клиентов	String

стеров рассчитываются с помощью метрики Евклидова расстояния.

Шаг 4. Центр каждого кластера пересчитывается на основе среднего значения в полученном кластере.

Шаг 5. Третий и четвёртый шаг повторяются, если продолжаются изменения в расстояниях данных в кластерах.

Алгоритм заканчивает работу, когда эти изменения прекращаются.

После завершения работы алгоритма в переменную «rfm» добавляется новое поле данных «Cluster», содержащее метку кластера каждой записи данных. Далее происходит процесс интерпретации полученных данных кластеров, добавляется новое поле «Segment», содержащее название, соответствующее кластеру. Также производится предоставление системой статистики и инфографики по полученным группам клиентов.

Таблица 4. Выходные данные системы

Название поля в системе	Описание данных	Тип данных
customer_id_col	Идентификатор клиента	Integer
Invoice_date_col	Дата покупки	Date
quantity_col	Количество товара	Integer
unit_price_col	Цена товара	float
transaction_code_col	Номер транзакции	Integer
email_col	Персональные данные	String
'Recency'	*recency*	Integer
'Frequency'	*frequency*	Integer
'Monetary'	*monetary*	float
'Cluster'	Метка кластера	integer
'Segment'	Название группы клиентов	String

Для сегментации клиентов на группы, различные по уровню лояльности, система производит следующий порядок шагов для сравнения значений полученных кластеров.

1. Вычисление средних значений параметров *recency*, *frequency*, *monetary* каждого кластера.

2. Определение группы «Потерянные»: выбор кластера с наибольшим средним значением параметра *recency*.

3. Определение группы «Новички»: выбор кластера с наименьшим средним значением параметра *recency* при условии, что значение параметра *frequency* для этого кластера больше, чем у всех остальных кластеров.

4. Определение группы «Киты»: выбор кластера с наибольшим средним значением параметра *monetary*.

5. Определение группы «Лояльные»: выбор кластера, у которой среднее значение параметра *monetary* больше, чем у кластера «Киты», при условии, что значение параметра *frequency* для этого кластера больше, чем у кластера «Киты».

На вход компоненты работы с группами клиентов поступают следующие данные (таблица 3).

Данный компонент отвечает за предоставление статистики и инфографики по группам клиентов и функционал маркетинговой работы. В качестве одной из форм работы маркетолога с сегментами в создаваемой системе будет реализована почтовая рассылка персональных сообщений (акций, скидок, персональных предложений и т.п.).

Выходные данные системы представляют набор следующих данных (таблица 4).

Заключение

Каждый четвёртый email-маркетолог испытывает трудности с контентом: они просто не знают, что ценного сообщить своим клиентам. Что же нужно маркетологам для отправки действительно релевантных писем? Сегментация. Для отправки эффективных, высоко персонализированных рассылок систему можно

настроить таким образом, чтобы проводить автоматические замеры по ключевым параметрам через равные промежутки времени и сегментировать покупателей на необходимое число групп. Хотя сегментация открывает гораздо больше возможностей, связанных с бизнес-аналитикой и стратегическим планированием.

Система, результаты проектирования которой представлены в данной статье, призвана помочь в вопросе автоматизации анализа клиентской базы, сегментации покупателей и отправке электронных писем в рамках готовой почтовой маркетинговой стратегии. Использование современных информационных технологий и методов машинного обучения для разработки такой системы позволит повысить качество маркетинговой работы:

1) эффективность — автоматизация процесса сегментации клиентской базы позволяет сэкономить трудовые и временные ресурсы;

2) точность — сегментация методом кластеризации обеспечивает более точное и объективное разделение клиентов на сегменты;

3) индивидуализация — сегментация на основе модели rfm-анализа позволит бизнесу предлагать более персонализированные продукты и услуги.

Литература

1. *Haron H., Golovachyova V., Tomilova N.* Principles of machine learning operation // Труды университета, 2022. №4 (89). С. 457–462.

2. *Ахмедова С.З.* Большие данные и машинное обучение. Научный аспект. 2023. Т. 18. № 5. С. 2304–2310.

3. *Кононова Н.В., Шиян Н.В., Плешешников П.И., Козина Н.И.* Машинное обучение. Использование неконтролируемого обучения. В сборнике: Вызовы современности и стратегии развития общества в условиях новой реальности. Сборник материалов XIV Международной научно-практической конференции. Москва, 2023. С. 150–153.

4. *Иванов М.Ю., Сыготина М.В., Надршин В.В., Дербенёва А.В.* Технологии интеллектуального анализа данных в решении экономических задач // Baikal Research Journal. 2022. Т. 13. № 2. С. 1–14.

5. *Wentley S.* Marketing Automation Trends to Know in 2023. 2024. [Электронный ресурс]. URL: https://www.constantcontact.com/blog/marketing-automation-trends/ (дата обращения 07.11.2024).

6. 4 модели сегментации клиентов для персонализации контента в 2024 году. Российская CDP платформа «Altcraft Platform». 2024. [Электронный ресурс]. URL: https://altcraft.com/ru/blog/modeli-segmentacii-klientov-dlya-personalizacii (дата обращения 07.11.2024).

7. *Shirole R., Salokhe L., Jadhav S.* Customer Segmentation using RFM Model and K-Means Clustering // International Journal of Scientific Research in Science and Technology. 2021. No. 8(3). Pp. 591–597.

8. *Qiasi. R., Baqeri-dehnavi. M., Minaei-bidgoli B.* Developing A Model For Measuring Customers Loyalty And Value With Rfm Technique And Clustering Algorithms // Journal of Mathematics and Computer Science. - 2020. No. 4(2). Pp. 172–181.

9. *Гончарук С.И., Воробьев С.П.* Описание RFM-анализа при сегментации клиентов интернет-магазина // Инновационная наука. 2020. №2. С. 59–62.

10. AI innovations at SAP. [Электронный ресурс]. URL: https://www.sap.com/products/crm/what-is-crm.html (дата обращения 07.11.2024).

11. Методология SADT. [Электронный ресурс]. URL: https://ami.nstu.ru/~vms/SADT_Ross/html/index.html (дата обращения 07.11.2024).

12. *Щаников С.А.* и др. Актуальные задачи теории и практики системной инженерии // Радиотехнические и телекоммуникационные системы. 2020, №4, С. 42–55.

13. *Рыжкова М.Н., Кутарова Е.И.* Когнитивное моделирование результатов образовательной деятельности студентов радиотехнического направления подготовки // Радиотехнические и телекоммуникационные системы. 2016. №2. С. 79–86.

Поступила 7 ноября 2024 г.

English

APPLICATION OF MACHINE LEARNING FOR USER SEGMENTATION USING THE EXAMPLE OF ANALYZING DATA ON PURCHASES MADE

Alla Sergeevna Platonova — PhD, Associate Professor, Department of Physics and Applied Mathematics, Murom Institute (branch) "Vladimir State University named after A.G. and N.G. Stoletovs".
E-mail: allaplatonova@inbox.ru

Maria Nikolayevna Ryzhkova — PhD, Associate Professor, the Dean of the Faculty of Information Technologies and Radioelectronics, Murom Institute (branch) "Vladimir State University named after A.G. and N.G. Stoletovs".

Address: 602264, Russian Federation, Vladimir region, Murom, Orlovskaya str., 23.

Abstract: The article presents the results of creating a system that, based on artificial intelligence methods, automates the segmentation of customer data by the level of customer loyalty, which subsequently allows for personal work with various groups of customers, as well as for analyzing purchase statistics. Customer segmentation into groups is performed based on the well-known RFM analysis method. The machine learning method - clustering, implemented in the work using the k-means algorithm, is responsible for automating the segmentation of customers into groups. The created information system implements personal work with customers in the format of mailing personalized offers.

Keywords: artificial intelligence, machine learning, clustering, data design, algorithmic support, RFM analysis, marketing, customer segmentation.

References

1. *Haron H., Golovachyova V., Tomilova N.* Principles of machine learning operation. Trudy universiteta, 2022. No. 4 (89). Pp. 457–462.

2. *Akhmedova S.Z.* Big data and machine learning. Scientific aspect. 2023. Vol. 18. No. 5. Pp. 2304–2310.

3. *Kononova N.V., Shiyan N.V., Plesheshnikov P.I., Kozina N.I.* Machine learning. Using unsupervised learning. In the collection: Challenges of our time and strategies for the development of society in the context of the new reality. Collection of materials of the XIV International scientific and practical conference. Moscow, 2023. Pp. 150–153.

4. *Ivanov M.Yu., Sygotina M.V., Nadrshin V.V., Derbeneva A.V.* Data Mining Technologies in Solving Economic Problems. Baikal Research Journal. 2022. Vol. 13. No. 2.

5. *Wentley S.* Marketing Automation Trends to Know in 2023. 2024. [Electronic source]. URL: https://www.constantcontact.com/blog/marketing-automation-trends/ (access date 07.11.2024).

6. 4 4 Customer Segmentation Models for Content Personalization in 2024. 2024. [Electronic source]. URL: https://altcraft.com/ru/blog/modeli-segmentacii-klientov-dlya-personalizacii (access date 07.11.2024).

7. *Shirole R., Salokhe L., Jadhav S.* Customer Segmentation using RFM Model and K-Means Clustering. International Journal of Scientific Research in Science and Technology. 2021. No. 8(3). Pp. 591–597.

8. *Qiasi. R., Baqeri-dehnavi. M., Minaei-bidgoli B.* Developing A Model For Measuring Customers Loyalty And Value With Rfm Technique And Clustering Algorithms. Journal of Mathematics and Computer Science. 2020. No. 4(2). Pp. 172–181.

9. *Goncharuk S.I., Vorobyov S.P.* Description of rfm-analysis for segmentation of clients of online store. Innovatsonnaya nauka. 2020. No. 2. Pp. 59–62.

10. AI innovations at SAP. [Electronic Source]. URL: https://www.sap.com/products/crm/what-is-crm.html (access date 07.11.2024).

11. SADT Methodology. [Electronic Source]. URL: https://ami.nstu.ru/~vms/SADT_Ross/html/index.html (access date 07.11.2024).

12. *Shchanikov S.A.* Current problems of the theory and practice of system engineering. Radio engineering and telecommunication systems. 2020. No. 4. Pp. 44–55.

13. *Ryzhkova M.N., Kutarova E.I.* Cognitive modeling of the results of educational activities of students of the radio engineering direction of training. Radio engineering and telecommunication systems. 2016. No. 2. Pp. 79–86.