

УДК 004
**ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ
ДЛЯ СЕГМЕНТАЦИИ ДАННЫХ В ПРОГРАММНОМ КОМПЛЕКСЕ
УПРАВЛЕНИЯ СПОРТИВНЫМИ КЛУБАМИ SPORTCRM**

Маслиев Р.О.¹ Григоренко А.Ю.²

1) Канд. геогр. наук, Генеральный директор ООО «Старт Эксперт» г. Краснодар, Россия; rmasliev@ya.ru

2) Инженер технической поддержки ООО «Старт Эксперт» г. Краснодар, Россия; alex.maineng@gmail.com

Аннотация. В статье рассматриваются основные аспекты сегментации данных с использованием алгоритмов машинного обучения, в частности K-Means, в контексте автоматизации RFM-методики (Recency, Frequency, Monetary Value). Подчеркивается важность вовлеченности студентов в реализуемые исследования и практическую работу, с целью получения ими прикладного опыта. В результате был получен визуализированный RFM-анализ, который в дальнейшем позволит оптимизировать маркетинговые стратегии и повысить эффективность коммуникаций с клиентами.

Ключевые термины: машинное обучение, сегментация, кластеризация

**APPLICATION OF MACHINE LEARNING ALGORITHMS FOR
DATA SEGMENTATION IN THE SPORTCRM SPORTS CLUBS
MANAGEMENT SOFTWARE**

Masliev R.O.¹ Grigorenko A.Yu.²

1) Candidate of Geographical Sciences, CEO, Start Expert LLC Krasnodar, Russia; rmasliev@ya.ru

2) Technical Support Engineer, Start Expert LLC Krasnodar, Russia; alex.maineng@gmail.com

Abstract. The article discusses the main aspects of data segmentation using machine learning algorithms, in particular K-Means, in the context of RFM (Recency, Frequency, Monetary Value) automation. The importance of the students' involvement in the realized research and practical work is emphasized, in order to gain applied experience. As a result, a visualized RFM-analysis was obtained, which in the future will allow to optimize marketing strategies and increase the effectiveness of communications with customers.

Keywords: machine learning, segmentation, clustering

Сегментация данных является важным инструментом в области анализа данных и маркетинга, позволяющим разделить обширные массивы информации на более управляемые и однородные группы. Этот процесс помогает выявить скрытые закономерности и различия между группами, что, в свою очередь, способствует более целенаправленному и эффективному взаимодействию с клиентами.

В современном контексте сегментация часто осуществляется с использованием методов машинного обучения, которые позволяют автоматизировать процесс классификации. Такие методы способны обрабатывать большие объемы данных и находить неочевидные и сложные взаимосвязи, незамеченные при традиционном анализе.

Актуальность данной работы обуславливается растущей тенденцией интеграции машинного обучения в бизнес-процессы, с целью повышения эффективности управления большим объемом данных.

В ходе работы трём группам студентов-практикантов было дано задание: выполнить RFM-сегментацию клиентов, основываясь на данных программного комплекса управления спортивными клубами SportCRM, используя разные подходы и алгоритмы, а также методы математической статистики. В своих проектах в ходе производственной практики в ООО «Старт Эксперт», которое является разработчиком SportCRM, студенты экспериментируют с технологиями машинного обучения, такими как TensorFlow и PyTorch. Они изучают возможности работы с API Google для обработки геоданных, аналитики и облачных вычислений [1]. В основу каждого подхода легла кластеризация методом К-Средних (K-Means). Суть метода в формировании из исходных данных K кластеров (групп) на основе их близости к центроидам (центрам кластеров).

Первым этапом стала предварительная подготовка данных. Для начала были выбраны только те столбцы, которые будут использоваться для создания RFM-модели. После чего происходит агрегация данных по уникальному идентификатору клиента (рис.1):

1. Recency: вычисляется как разница между текущей датой и максимальной датой оплаты для каждого клиента. Это показывает, сколько дней прошло с момента последней покупки.

2. Frequency: определяется как количество уникальных транзакций (или покупок) для каждого клиента. Это позволяет понять, как часто клиенты совершают покупки.

3. Monetary Value: суммируется общая сумма всех покупок клиента, что позволяет оценить его ценность для бизнеса.

	id	recency	frequency	monetary_value
0	100685	77	1	8500.0
1	100686	77	1	8500.0
2	100704	91	1	12500.0
3	100705	91	1	7650.0
4	100706	91	1	12500.0

Рисунок 1 – Преобразованная таблица данных

Следующий этап – определение количества кластеров. Это необходимо для корректного объединения данных в естественные группы, а также во избежание переобучения или недообучения модели.

Для этого был использован метод локтя (Elbow method). Его суть заключается в следующем: мы выполняем кластеризацию для различных значений k и строим график зависимости суммарной внутрикластерной дисперсии (WCSS) от количества кластеров. Обычно на графике видно, что WCSS резко снижается при небольшом увеличении k , а затем начинает снижаться более плавно. Точка, где эта "резкая" часть графика переходит в "плавную", называется "локтем", или точкой насыщения (saturation point). Это и есть оптимальное количество кластеров.

Для реализации метода на языке программирования python была использована библиотека YelloBrick. Сначала указывается диапазон значений количества кластеров (k). Далее для каждого значения применяется алгоритм кластеризации, вычисляется WCSS и строится график. (рис. 2)

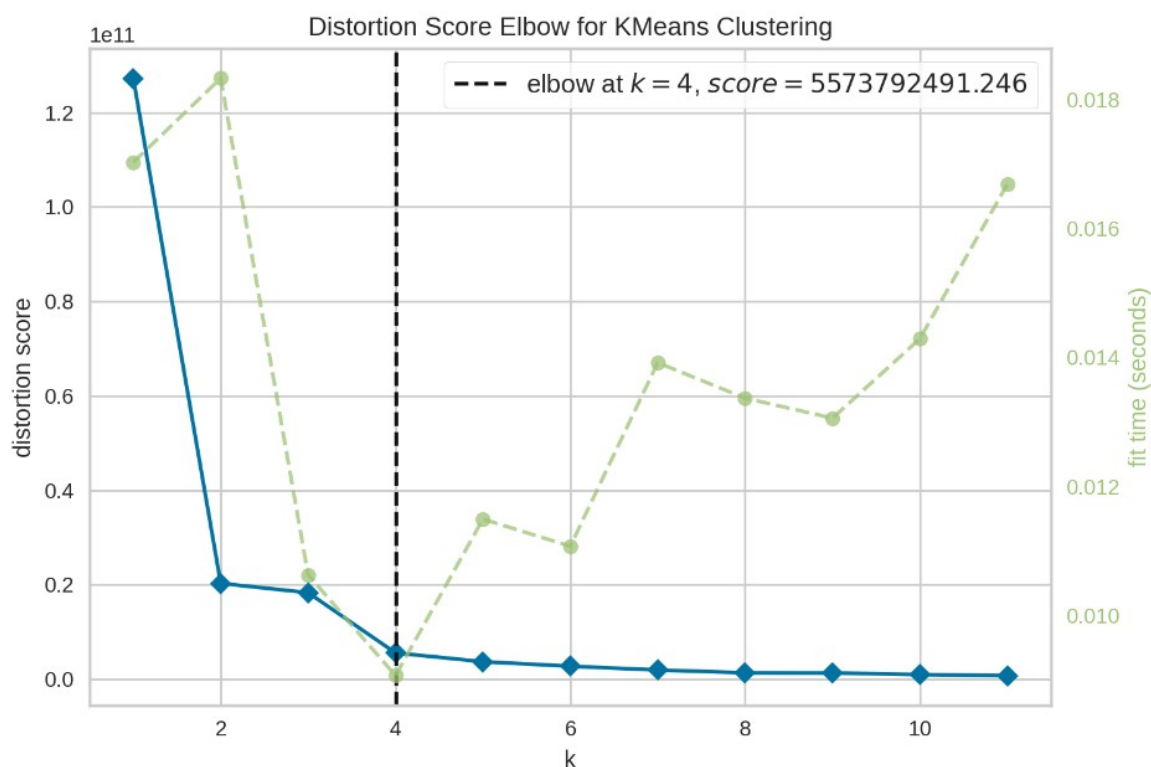


Рисунок 2 – График WCSS

Как можно увидеть по графику – оптимальное количество кластеров равно четырем.

После идентификации точки насыщения (saturation point) производилась процедура кластеризации с оптимальным числом групп. Кластеризация и последующая визуализация результатов группами студентов выполнялась разными методами: с помощью языка программирования python и с помощью программы Microsoft Excel.

Далее была сформирована сводная таблица, преобразующая данные в формат удобный для визуализации (рис. 3).

	A	B	C	D	E	F	G	H	I	J	K
1	sportsmen	lastpay	MIN из days	SUM из count	SUM из sum	R	F	M	RFM	Обозначение	Кластер
2	6	2025-02-28	25	5	38100	2	2	1	221	Дрейфующие	3
3	18	2025-03-01	24	3	37500	2	2	1	221	Дрейфующие	3
4	19	2025-02-28	25	2	11900	2	3	3	233	В зоне риска	5
5	25	2025-03-05	20	6	41160	1	1	1	111	Лояльные	1
6	49	2025-02-27	26	6	41160	3	1	1	311	В зоне риска	5
7	68	2025-03-03	22	6	41160	1	1	1	111	Лояльные	1
8	69	2025-03-03	22	6	41160	1	1	1	111	Лояльные	1
9	77	2025-02-28	25	5	28980	2	2	2	222	Дрейфующие	3
10	80	2025-03-13	12	14	56720	1	1	1	111	Лояльные	1
11	97	2025-03-13	12	9	48720	1	1	1	111	Лояльные	1
12	135	2025-03-04	21	13	53250	1	1	1	111	Лояльные	1
13	141	2025-02-28	25	4	49220	2	2	1	221	Дрейфующие	3
14	142	2025-02-28	25	4	49220	2	2	1	221	Дрейфующие	3
15	211	2025-02-28	25	6	41160	2	1	1	211	Перспективные	2
16	212	2025-03-01	24	6	36570	2	1	1	211	Перспективные	2
17	221	2025-03-03	22	7	53660	1	1	1	111	Лояльные	1
18	235	2025-03-13	12	7	53460	1	1	1	111	Лояльные	1
19	240	2025-03-11	14	3	7500	1	2	3	123	Дрейфующие	3

Рисунок 3 – Сводная таблица в программе Microsoft Excel

Визуализация кластерного RFM анализа была выполнена с использованием тепловой карты (heatmap) (рис. 4) и пузырьковой диаграммы (bubble chart) (рис. 5).



Рисунок 4 – Визуализация RFM анализа с помощью тепловой карты

Тепловая карта использована для визуализации нормализованных средних значений RFM-метрик, где цветовой градиент отражает количество пользователей в каждом кластере.

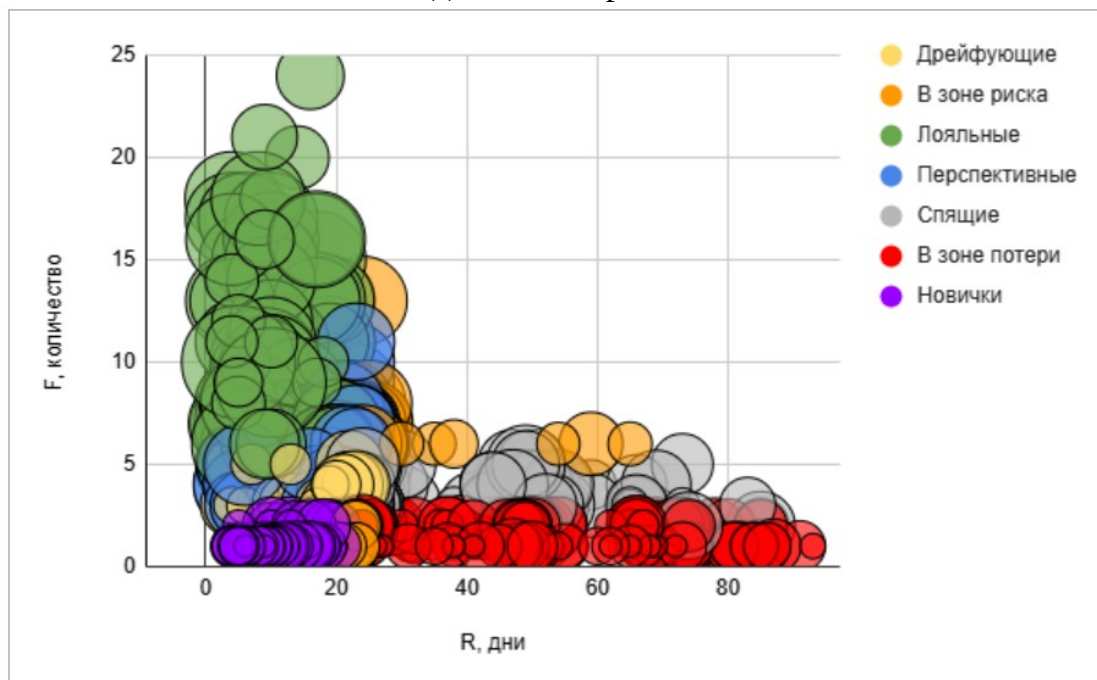


Рисунок 5 – Визуализация RFM анализа с помощью пузырьковой диаграммы

Пузырьковая диаграмма использована для многомерной визуализации и отображает Recency (Ось X), Frequency (Ось Y), Monetary (размер пузыря) и принадлежность к кластеру (цвет).

Данные методы визуализации обеспечили многокритериальный анализ, позволяя количественно оценить взаимосвязь трёх ключевых параметров (Recency, Frequency, Monetary Value) в многомерном признаковом пространстве, а также выявить компактность кластеров и их распределение в пространстве признаков.

Таким образом, итогами практической работы стал наглядный RFM анализ, который в дальнейшем позволит оптимизировать маркетинговые стратегии и повысить эффективность коммуникаций с клиентами при использовании программного комплекса SportCRM. Он формирует основу для персонализации коммуникаций: от точечных предложений до прогнозирования LTV (пожизненной ценности), а интеграция RFM с ML-алгоритмами, как показала практика, углубляет эти возможности, автоматизируя прогнозирование оттока, выявление скрытых паттернов (например, сезонных предпочтений) и динамическую адаптацию стратегий под меняющиеся данные.

Этот результат также наглядно демонстрирует, что совместная работа студентов-практикантов и компаний взаимовыгодна: бизнес получает свежие идеи и перспективные кадры, а студенты — реальный опыт и возможность применять свои знания на практике.

Список использованных источников:

1) Маслиев Р.О., Калинина М.Е. ОСОБЕННОСТИ ОРГАНИЗАЦИИ ПРОИЗВОДСТВЕННОЙ ПРАКТИКИ ДЛЯ СТУДЕНТОВ IT - СПЕЦИАЛЬНОСТЕЙ В МАЛЫХ ТЕХНОЛОГИЧЕСКИХ КОМПАНИЯХ НА ПРИМЕРЕ ООО СТАРТ ЭКСПЕРТ В сборнике: Образование - Наука - Практика. сборник статей I Всероссийской научно-практической конференции обучающихся, преподавателей, практических работников. Краснодар, 2024. С. 10-13.

2) Вахромеева Е.Н., Зензинова Ю.Б. АВТОМАТИЗАЦИЯ КЛАСТЕРИЗАЦИИ КОМПАНИЙ ПО ФИНАНСОВЫМ ПОКАЗАТЕЛЯМ С ИСПОЛЬЗОВАНИЕМ K-MEANS АЛГОРИТМА НА БОЛЬШИХ ДАННЫХ Дискуссия. 2024. № 5 (126). С. 46-50.