СЕГМЕНТАЦИЯ АБОНЕНТСКОЙ БАЗЫ ТЕЛЕКОМУННИКАЦИОННОЙ КОМПАНИИ

УДК 519.816

ЛЕПА Евгений Владимирович
к.т.н., доцент кафедры информационных технологий Херсонского национального технического университета.
Научные интересы: системы принятия и поддержки решений, технологии интеллектуального анализа данных.
e-mail: e.lepa@mail.ru

ВВЕДЕНИЕ

В настоящее время накоплено большое количество данных, извлечение знаний из которых, могут быть использованы для повышения эффективности деятельности компаний. Для решения этой задачи широко используются методы интеллектуального анализа данных [1-2].

В работе использована аналитическая платформа Deductor, на которой была проведена сегментация абонентов телекоммуникационной компании по возрастным группам для решения задач менеджмента и маркетинга.

ПОСТАНОВКА ЗАДАЧИ

Для телекоммуникационной компании по предоставлению услуг мобильной связи, необходимо выполнить анализ ее деятельности за определенный период времени. Целями такого анализа является построение профилей абонентов, продолжительности и времени звонков, ежемесячных расходов, а также выявление наиболее доходных сегментов услуг.

Результаты анализа могут быть использованы для разработки маркетинговых акций и новых тарифных планов, оптимизации затрат и предотвращения оттока клиентов компании.

ОСНОВНОЙ МАТЕРИАЛ

Для решения поставленной задачи использованы методы кластеризации. В аналитической платформе Deductor реализованы три основных метода кластеризации - k-средних, g-средних и карты Кохонена. Первые два метода близки по алгоритму для решения задачи кластеризации. Основное отличие их заключается в том, что в методе k-средних задается аналитиком количество кластеров, а в методе g-средних оно определяется автоматически. Карты Кохонена являются разновидностью искусственных нейронных сетей [3].

В результате использования любого из методов могут быть получены профили кластеров, таблица их параметров и статистические данные, Кроме того, аналитическая платформа позволяет выполнить прогнозирование. Применение карт Кохонена позволяет повысить качество визуализации результатов решения задачи.

Для анализа использованы данные биллинговой системы. Эти данные включают возраст абонентов, средние затраты и длительность разговоров, количество звонков в разное время суток, количество звонков в другие города и страны, на стационарные телефоны, а также количество SMS.

Предварительно все абоненты коммуникационной сети были разделены по возрасту на группы, которые характеризуются определенным статусом и активностью деятельности.

1 группа. Абоненты до 20 лет - школьники, студенты начальных курсов обучения.

2 группа. Абоненты возраста в пределах 21-35 лет – молодые специалисты.

3 группа. Абоненты возраста в пределах 36-55 лет – наиболее активный период деятельности.

4 группа. Абоненты возраста в пределах 56-60 лет - убыль активности деятельности.

5 группа. Абоненты старше 60 лет – пенсионеры.

Таким образом, предварительно определено пять сегментов (кластеров).

В результате анализа эти предположения могут подтвердиться полностью, частично или отвергнуты, как не соответствующие реальности.

При использовании метода k-средних задано пять кластеров и получены их профили, определяющие номер кластера, количество примеров, попавших в кластер и их процент, а также статистику по каждому входному параметру (полю) исходных данных. При использовании метода g-средних автоматически было определены четыре кластера, для которых получены профили и статистика.

По статистике кластеров можно определить, насколько первоначальные предположения о разбиении абонентов на возрастные группы оказались оправданными. Эти данные представлены в табл. 1, в которой каждый метод имеет свою нумерацию кластеров.

Таблица 1 –

Статистические данные кластеров

Номер кластера	Минимальный возраст	Максимальный возраст	Средний возраст
Пять кластеров			
0	12	15	13,5
1	51	70	61,3
2	16	24	19,6
3	23	34	28,9
4	35	54	43,7
Четыре кластера			
0	12	22	17
1	23	34	28,5
2	35	54	44,5
3	55	70	62,5

Из таблицы видно, что для пяти кластеров вторая группа (абоненты возраста 21-35 лет) практически полностью попадают в третий кластер. То же можно сказать и о третьей группе (абоненты возраста 36-55 лет) которые попадают в четвертый кластер. Четвертая группа (абоненты возраста 56-60 лет) попадает в первый кластер, который имеет диапазон 51-75 лет. Это только частично соответствует принятым предположениям относительно возрастных групп.

Пятая группа (абоненты старше 60 лет) вообще не является отдельным кластером. Кроме того, есть первый кластер для возрастного диапазона 20-25 лет, наличие которого не предполагалось.

Выполненная кластеризация в полной мере не подтвердила предположения о разбивке абонентов на возрастные группы. Прежде чем внести определенные изменения в первоначальные предположения или сделать окончательные выводы, необходимо воспользоваться другими методами кластеризации.

Для четырех кластеров первоначальные предположения о возрастных категориях практически полностью подтвердились. Исключение представляет только третий кластер, куда вошли абоненты четвертой и пятой групп.

Выполненный анализ показывает, что наиболее реальным является предположение о четырех возрастных группах абонентов. Для окончательных выводов о сегментации абонентов коммуникационной сети был использован еще одним методом кластеризации – карты Кохонена.

При использовании этого метода было задано четыре кластера и получены их статистические данные, которые представлены в табл. 2.

Таблица 2 –

Статистические данные кластеров

Номер кластера	Минимальный возраст	Максимальный возраст	Средний возраст
0	12	22	17
1	23	34	28
2	51	70	63
3	35	54	43

Статистические данные для четырех кластеров, полученных двумя различными методами, практически полностью совпадает. Кроме статистических данных, при использовании различных методов кластеризации получены профили кластеров. В табл. 3 приведены статистические данные и профили кластеров, полученные методами g-средних и с помощью карт Кохонена. Размер кластера определяется по количеству вошедших в него примеров из исходных данных.

Таблица 3

	Метод g-средних	Карта Кохонена
Размер кластера	11	11
Возрастной диапазон, лет	12-22	12-22
Размер кластера	12	18
Возрастной диапазон, лет	23-34	23-34
Размер кластера	20	19
Возрастной диапазон, лет	35-54	35-54
Размер кластера	16	11
Возрастной диапазон, лет	55-70	51-70

Несмотря на практически полное совпадение возрастных диапазонов абонентов, размеры кластеров немного отличаются в двух возрастных диапазонах.

Дальнейший анализ абонентов коммуникационной сети проводится по карте Кохонена, которая построена в виде отдельных пластов «пирога». Каждый пласт соответствует определенному параметру (полю) входных данных. На основании этой карты и статистики полученных профилей можно выполнить необходимый анализ абонентов сети по любому параметру исходных данных. В качестве такого параметра выбран возраст.

В табл. 3 приведенный средний возраст абонентов для каждого кластера. Для каждого кластера выбирается соответствующая точка на карте **Возраст** и появляются оценки для всех остальных параметров на всех других картах в этом же кластере. Это дает возможность оценить зависимости одного параметра (поля) исходных данных от других. Результаты выполненного анализа представлены в табл. 4.

Для среднего возраста абонента в каждом кластере определены средние значения остальных параметров входных данных, что является основанием для дальнейшего анализа, используя также другие инструменты анализа и наборы исходных данных.

Таблица 4 –

Анализ карт Кохонена

	Номер кластера			
	0	1	2	3
	Средний возраст (лет)			
	17	28	63	43
Расходы, грн.	24	84	35	60
Продолжительность, мин	17	1	16	13
Звонков днем	120	150	25	160
Звонков вечером	180	25	81	120
Звонков ночью	20	10	6	9
Звонков в другие города	15	30	61	80
Звонков в другие страны	4	10	2	18
Звонков на стационарные телефоны	15	50	8	45
SMS	170	95	9	55

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

Для решения задачи сегментации абонентов телекоммуникационной сети использованы три различных метода кластеризации, реализованных в аналитической платформе Deductor. Анализ выполнен только для одного из параметров исходных данных – распределение абонентов на возрастные группы.

Первоначальное предположение о наличии пяти возрастных групп абонентов не подтвердилось. Наиболее реальным является распределение абонентов по четырем возрастным группам, что подтверждается результатами, полученными двумя из трех используемых методов.

Использование подобного анализа позволяет решить ряд задач менеджменту и маркетингу, направленных на повышение эффективности работы компании.

ЛИТЕРАТУРА:

1. Palkin N.B., Oreshkov V.I. Business of analyst: from information to knowledges. – SPb: Piter, 2010. – 352 s.
2. Dyuk V., Samoylenko P. Data Mining. Educational course. – SPb.: Piter, 2002. – 368 s.
3. Guidance of analyst. Version of 5.2. – M.: Basegroup Labs, 2010. – 122 s.

Рецензент: *д.т.н., проф. Соколова Н.А., Херсонский национальный технический университет.*