УДК 338.2

5.2.2. Математические, статистические и инструментальные методы экономики (экономические науки)

СЕГМЕНТАЦИЯ КЛИЕНТОВ С ПРИМЕНЕНИЕМ ПРОГРАММНЫХ СРЕДСТВ ДЛЯ МОДЕЛИРОВАНИЯ ИНФОРМАЦИОННЫХ СИСТЕМ И ОБРАБОТКИ ДАННЫХ

Алимова Мария Сергеевна
канд. экон. наук, доцент кафедры прикладной математики
SPIN-код: 4526-5386
ORCID 0000-0002-3643-2409
Scopus Author ID: 57211806316
WoSResearcherID: E-9769-2016
mashasms1@gmail.com

Василак Ростислав Викторович
Аспирант кафедры прикладной математики
rostislavvasilak@gmail.com
Университет «Синергия», Россия, Москва 129090, Мещанская, д. 9/14, стр. 1

Сбалансированная система показателей представляет собой стратегический инструмент управления, включающий четыре ключевых блока: финансовый, клиентский, внутренние бизнес-процессы и обучение, развитие персонала. Комплексная проработка каждого из них способствует росту эффективности деятельности компании и достижению ее стратегических целей. В настоящем исследовании основное внимание уделяется изучению клиентской составляющей в рамках использования сбалансированной системы показателей, в частности – задаче сегментации клиентов. Данный аспект приобретает особую значимость и актуальность в условиях высокой конкурентной среды и необходимости персонализации взаимодействия с целевыми группами. Целью исследования является сегментация клиентов на основе собранных и систематизированных данных анкетирования, полученных из открытых источников, на базе использования программных средств для моделирования информационных систем и обработки данных для решения прикладных задач классификации и кластеризации клиентской базы

Ключевые слова: СЕГМЕНТАЦИЯ КЛИЕНТОВ, СБАЛАНСИРОВАННАЯ СИСТЕМА ПОКАЗАТЕЛЕЙ, МОДЕЛИРОВАНИЕ ИНФОРМАЦИОННЫХ СИСТЕМ, САМООРГАНИЗУЮЩАЯСЯ КАРТА КОХОНЕНА

http://dx.doi.org/10.21515/1990-4665-210-047

UDC338.2

5.2.2. Mathematical, statistical and instrumental methods of economics (economic sciences)

CLIENT SEGMENTATION USING SOFTWARE TOOLS FOR MODELING INFORMATION SYSTEMS AND DATA PROCESSING

Alimova Mariia Sergeevna
Cand.Econ.Sci., associate professor of the Department of Applied Mathematics
RSCI SPIN-code: 4526-5386
ORCID 0000-0002-3643-2409
Scopus Author ID: 57211806316
WoSResearcherID: E-9769-2016
mashasms1@gmail.com

Vasilak Rostislav Viktorovich
Postgraduate student of the Department of Applied Mathematics
Synergy University, Russia, Moscow 129090, Meshchanskaya, 9/14, str. 1

The balanced scorecard is a strategic management tool that includes four key blocks: financial, customer, internal business processes and training, personnel development. Comprehensive development of each of them contributes to the growth of the company's efficiency and the achievement of its strategic goals. This study focuses on the study of the customer component within the framework of the balanced scorecard, in particular - the task of customer segmentation. This aspect is of particular importance and relevance in the context of a highly competitive environment and the need to personalize interaction with target groups. The purpose of the study is to segment customers based on collected and systematized survey data obtained from open sources, using software for modeling information systems and data processing to solve applied problems of classification and clustering of the customer base

Keywords: CUSTOMER SEGMENTATION, BALANCED SCORECARD, INFORMATION SYSTEMS MODELLING, KOHONEN'S SELF-ORGANIZING MAP

Введение

Исследование структурировано и включает в себя четыре ключевых этапа. На первом этапе рассматриваются базовые, фундаментальные теоретические положения, определяющие понятие «сегментация клиентов», а также обоснование актуальности и выбора подходов к ее реализации. Второй этап включает предварительную обработку эмпирических данных, полученных из открытых источников: осуществляется очистка и подготовка выборки к анализу. На третьем этапе проводится моделирование сегментации клиентов, включающее выбор метода, построение модели. Заключительный этап исследования посвящён интерпретации полученных сегментов и их привязке к клиентскому блоку в рамках использования сбалансированной системы показателей.

Согласно классическому определению, предложенному Ф. Котлером, сегментирование рынка представляет собой процесс разделения потенциальных покупателей на группы, различающиеся по своим потребностям, характеристикам или поведенческим признакам, и, следовательно, требующие различных продуктов или маркетинговых стратегий. В научной и прикладной литературе выделяется ряд подходов к сегментированию, среди которых дескриптивный, психографический, поведенческий, кластерный и другие. В рамках настоящего исследования акцент будет сделан на кластерном подходе, поскольку анализируемые данные получены методом анкетирования, и априорное распределение респондентов по группам неизвестно.

Методы и материалы.

В области клиентской сегментации отсутствует универсальный алгоритм, обеспечивающий однозначное и воспроизводимое решение. В зависимости от характеристик выборки, методов обработки и применяемых критериев, возможны различные стратегии сегментирования

и неоднозначные результаты, даже при работе с одними и теми же исходными данными.

Для реализации задачи сегментации в исследовании используются данные анкетирования, полученные из открытых источников, а также программные средства для моделирования информационных систем и обработки данных (Python, Microsoft Excel), что позволит обеспечить как аналитическую, так и прикладную обоснованность результатов.

Для кластерного сегментирования характерно применение факторного анализа и метода главных компонент. Оба метода позволяют представить исходные переменные в виде ограниченного набора новых латентных переменных (факторов или компонент), обладающих интерпретируемой структурой.

Метод главных компонент представляет собой линейное преобразование, при котором множество взаимосвязанных переменных преобразуется в меньший набор некоррелированных компонент, упорядоченных по убыванию объясняемой дисперсии. Первая главная компонента захватывает максимально возможную дисперсию в данных, вторая – оставшуюся дисперсию, ортогональную первой, и т.д.

Факторный анализ, в отличие от PCA, направлен на выявление скрытых (латентных) факторов, стоящих за наблюдаемыми переменными. Предполагается, что каждая наблюдаемая переменная является линейной функцией от нескольких факторов и уникальной (специфической) составляющей.

Особое внимание следует уделить самоорганизующейся карте Кохонена (SOM, Self-Organizing Map) – разновидности обучающейся без учителя нейронной сети, предназначенной для отображения многомерных данных в пространство меньшей размерности, с сохранением топологической структуры исходных данных.

Универсального метода решения поставленной задачи не существует, что усложняет исследование и одновременно подчеркивает его актуальность. В связи с этим, будет представлен один из возможных подходов, продемонстрировавший результаты, пригодные для интерпретации в прикладном контексте. Эмпирическая база включает ответы 540 респондентов на 66 вопросов, представленных в шкале от 1 до 5. Первичный разведочный анализ данных (EDA) осуществлялся в среде Jupyter Notebook.

Результаты исследования.

На начальном этапе было проведено общее ознакомление с данными: определены типы переменных, выявлено наличие категориальных признаков, а также проанализированы пропущенные значения в каждом из столбцов. Поскольку все переменные представлены в числовом формате, работа с категориальными признаками не потребовалась. Пропуски были обнаружены в двух столбцах; их доля составила менее 2% от всей выборки, что позволяет применить стратегию удаления. Однако перед этим был осуществлён ручной просмотр соответствующих строк. Установлено, что среди них встречаются как сбалансированные, так и несбалансированные респонденты, но ввиду отсутствия уникальной информации они были исключены из анализа. Следует подчеркнуть, что исключение респондентов из выборки при работе с реальными данными является методологически неоднозначной процедурой. В условиях прикладных исследований, где сбор каждого ответа сопряжён с финансовыми затратами для компании, принятие решения об удалении наблюдений требует наличия обоснованных и строго аргументированных причин. Исключение данных респондентов допустимо лишь в тех случаях, когда сохранение их в выборке может существенно исказить результаты анализа или привести к некорректным выводам.

```
Признак: 53, кол-во уникальных значений: 4
Уникальные значения:
53
3.0    49
2.0    12
1.0     2
4.0     1
Name: count, dtype: int64

Признак: 80, кол-во уникальных значений: 4
Уникальные значения:
80
3.0    21
2.0    20
4.0    17
1.0     6
Name: count, dtype: int64
```

Рисунок 1 – Уникальные значения определенных респондентов с пропусками

Далее были рассчитаны базовые статистические характеристики набора данных. Средние значения и стандартные отклонения переменных оказались сопоставимыми по шкале, при этом сами переменные представлены четырьмя значениями в единой шкале, что делает дополнительную стандартизацию необязательной. Дополнительно была проведена проверка наличия выбросов по показателям суммы, среднего и стандартного отклонения ответов респондентов, по результатам которой наблюдения с аномальными значениями стандартного отклонения были удалены. Одной из особенностей рассматриваемого массива данных является монотонность ответов отдельных респондентов, что может негативно сказаться на качестве кластеризации. Для устранения подобных

случаев была использована мера энтропии Шеннона, позволившая исключить нерепрезентативные наблюдения.

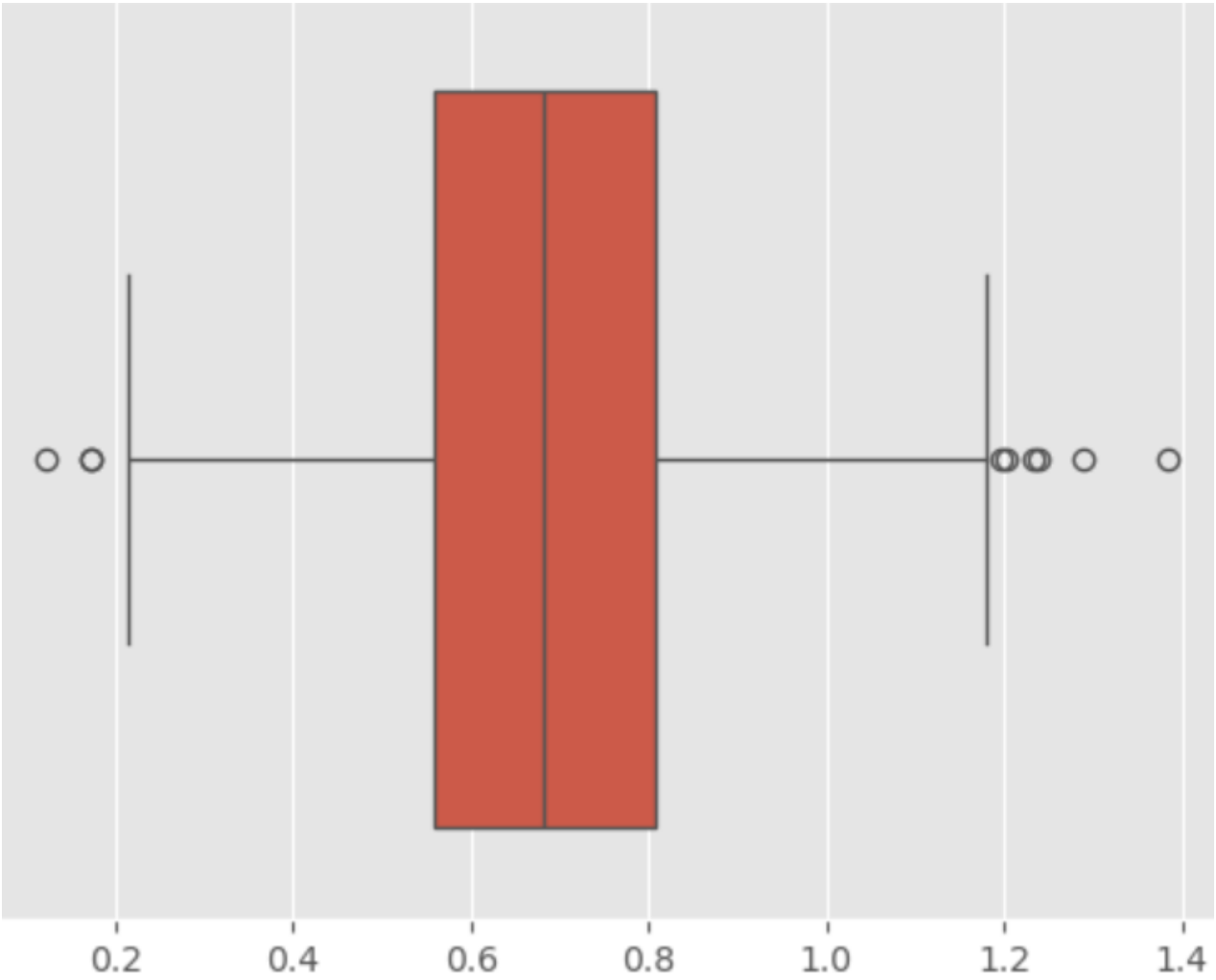

Рисунок 2 – Выбросы по показателю стандартного отклонения ответов респондентов

После предварительной обработки респондентов следующим этапом является работа с переменными. На первом шаге была построена матрица корреляций между переменными. Результаты корреляционного анализа показали отсутствие высоко коррелированных признаков (коэффициент корреляции не превышал порогового значения 0,7), что исключило возможность исключения избыточных признаков на данном этапе.

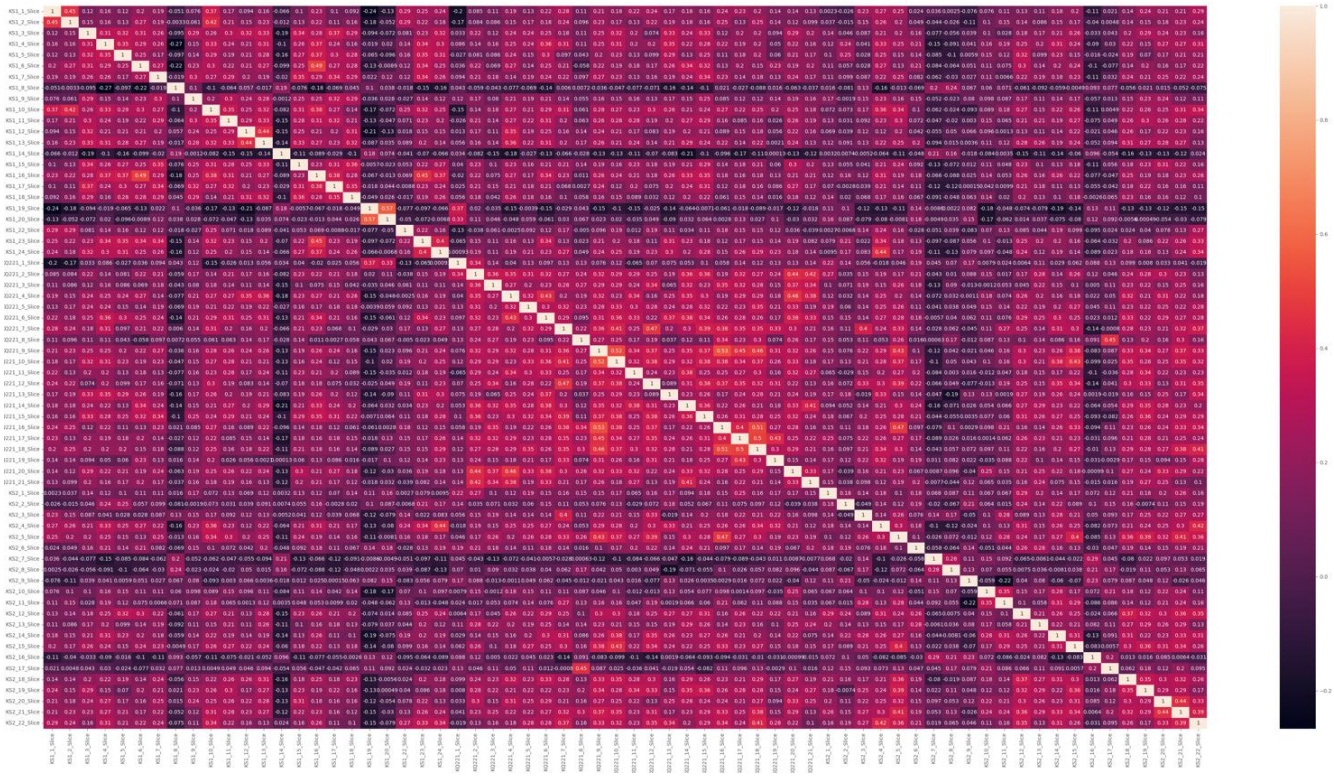

Рисунок 3 – Корреляционная матрица признаков

В условиях слабой линейной зависимости между переменными было принято решение о применении алгоритма кластеризации KMeans ко всему множеству признаков. Несмотря на варьирование параметров модели, результаты кластеризации оказались неэффективными. Альтернативные алгоритмы кластеризации, в том числе DBScan и иерархическая кластеризация, также не продемонстрировали удовлетворительной устойчивости и интерпретируемости кластеров.

На следующем этапе был осуществлён переход к методам снижения размерности. В частности, был проведен факторный анализ, который, согласно теоретическим предпосылкам, является оптимальным инструментом для обработки анкетных данных, а также метод главных компонент (PCA), направленный на сохранение переменных, обладающих наибольшей долей объяснённой дисперсии. Предварительная проверка пригодности данных для применения этих методов осуществлялась с помощью коэффициента Кайзера – Мейера – Олкина (KMO) и критерия

сферичности Бартлетта, результаты которых подтвердили допустимость проведения факторизации.

Тем не менее, полученные результаты вновь оказались неудовлетворительными. Проведённый факторный анализ позволил выделить 16 факторов, которые в совокупности объясняли лишь 45% общей дисперсии. При этом вклад каждой переменной в объяснение дисперсии был приблизительно равным, что указывает на отсутствие чётко выраженных латентных структур. На основании этих наблюдений была сформулирована гипотеза о возможной нелинейной зависимости между переменными. Данная гипотеза получила подтверждение в ходе последующего анализа с использованием теста на мультиколлинеарность, а также построения полиномиальных регрессионных моделей для отдельных признаков.

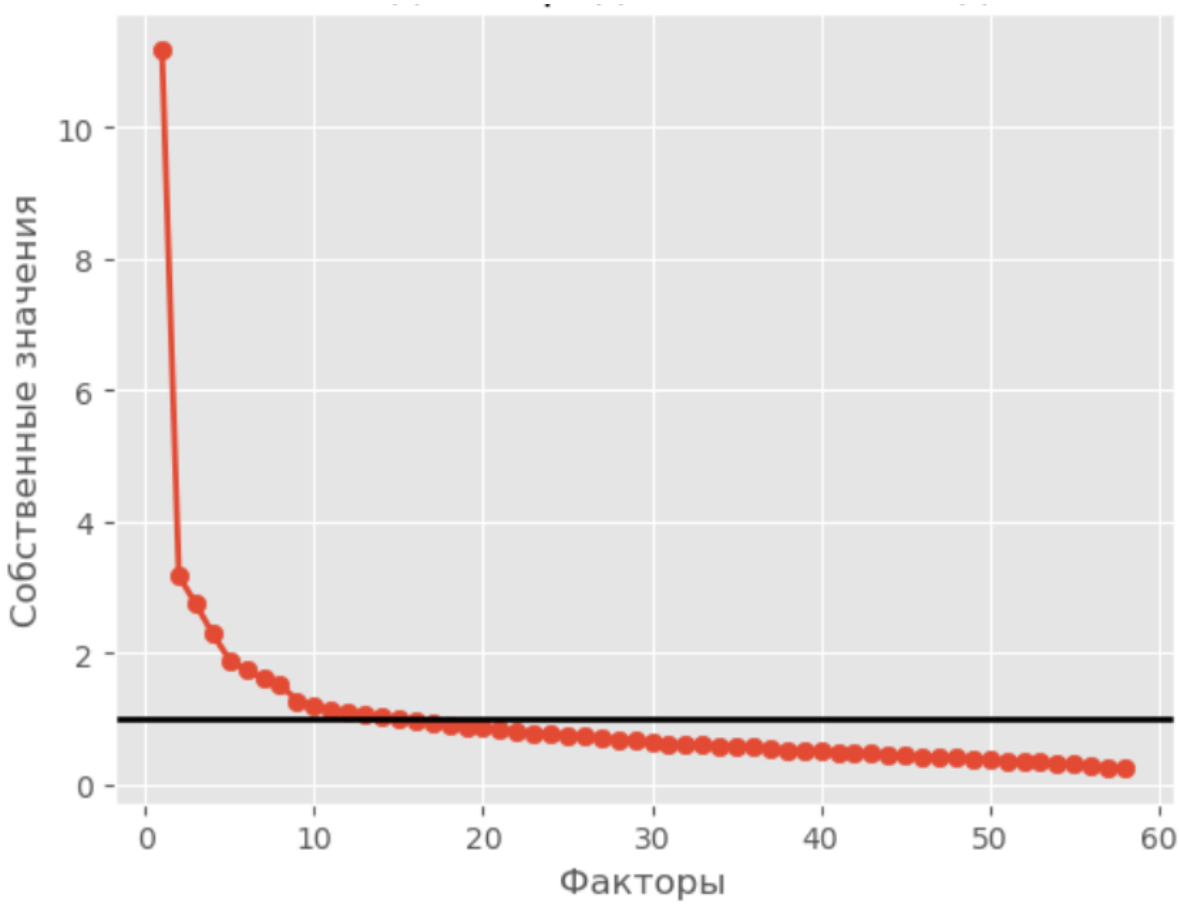

Рисунок 4 – Применение факторного анализа для первичных данных

Для более эффективного снижения размерности был реализован оригинальный алгоритм отбора информативных переменных. В частности, одна переменная выбиралась в качестве целевой, а в качестве обучающей выборки рассматривались все возможные комбинации по пять переменных. Из всех таких комбинаций отбиралась та, которая обеспечивала наилучшее предсказание целевой переменной. После этого в отобранную пятёрку последовательно добавлялись комбинации по две, три и четыре переменных, и переменные с наибольшим вкладом в предсказательную способность включались в итоговую выборку.

В процессе итераций была произведена дополнительная группировка переменных: изначально были сформированы две группы, включающие переменные второй степени, а затем в одной из этих групп было осуществлено дополнительное деление с использованием признаков пятой степени. В результате многочисленных итераций удалось выделить несколько устойчивых блоков переменных, характеризующихся высокой внутренней согласованностью и почти полной взаимной объяснённостью. Это позволило обоснованно исключить ряд переменных из дальнейшего анализа, сократив размерность признакового пространства.

Заключительный этап исследования был направлен на формирование кластеров респондентов. Попытки осуществить кластеризацию с применением перебора возможных комбинаций наблюдений в ручном режиме оказались неэффективными вследствие высокой вычислительной сложности задачи. Действительно, число возможных комбинаций, например, из 40 элементов при общей выборке в 400 респондентов, оказывается чрезмерно велико, что делает такой подход практически неприменимым.

В связи с этим, опираясь на результаты предварительного анализа, был разработан альтернативный алгоритм кластеризации респондентов. Предложенный механизм основан на последовательном использовании

самоорганизующейся карты Кохонена (Self-Organizing Map, SOM) для выделения кластеров внутри отдельных сегментов выборки. Суть подхода заключалась в следующем: на первом этапе один из сегментов данных подвергался кластеризации с применением SOM, при этом идентифицировался кластер с наибольшим числом респондентов, обладающих максимальными средними значениями признаков. Индексы строк, соответствующие этому кластеру, сохранялись.

На следующем шаге из оставшихся данных (второго сегмента) исключались строки, относящиеся к ранее выделенному кластеру, после чего процедура кластеризации с использованием SOM повторялась. Данный процесс осуществлялся итеративно до получения четырёх кластеров респондентов.

После завершения кластеризации был проведён этап ручной маркировки кластеров: каждому респонденту в общей выборке был присвоен индекс соответствующего кластера на основании ранее сохранённых индексов строк. Итогом данного процесса стало построение профилирующей таблицы, содержащей распределение респондентов по выделенным кластерам.

Итоговые результаты кластеризации и характеристика полученных групп респондентов представлены на рисунке 5.

segment	Qu1	Qu2	Qu3	Qu4	Qu5	Qu6	Qu7	Qu8	Qu9	Qu10	Qu11	Qu12	Qu13	Qu14
0	2,61	2,9	3,79	3,37	3,37	3,54	3,34	2,31	3,43	3,12	3,48	3,72	3,66	2,2
1	2,524193548	2,733870968	3,169354839	2,612903226	2,685483871	2,943548387	2,870967742	2,58870968	2,95967742	2,637096774	2,919354839	3,233870968	3,120967742	2,532258065
2	3,235294118	3,431372549	3,754901961	3,421568627	3,460784314	3,568627451	3,509803922	2,68627451	3,5	3,333333333	3,490196078	3,735294118	3,676470588	2,343137255
3	3,032967033	3,241758242	3,549450549	3,186813187	3,285714286	3,417582418	3,450549451	2,45054945	3,43956044	3,43956044	3,428571429	3,648351648	3,637362637	2,241758242
segment	Qu15	Qu16	Qu17	Qu18	Qu19	Qu20	Qu22	Qu23	Qu24	Qu25	Qu26	Qu27	Qu28	Qu29
0	3,64	3,42	3,75	3,9	2,59	2,79	2,28	3,09	3,15	2,93	3,66	3,33	3,77	3,42
1	3,088709677	2,661290323	3,322580645	3,282258065	2,766129032	2,814516129	2,258064516	2,62903226	2,35483871	2,927419355	3,322580645	3,177419355	3,282258065	2,879032258
2	3,578431373	3,568627451	3,715686275	3,754901961	2,568627451	2,843137255	2,588235294	3,55882353	3,37254902	3,421568627	3,823529412	3,617647059	3,81372549	3,588235294
3	3,648351648	3,351648352	3,78021978	3,67032967	2,450549451	2,703296703	2,406593407	3,13186813	3,03296703	2,945054945	3,604395604	3,472527473	3,846153846	3,054945055
segment	Qu30	Qu31	Qu33	Qu35	Qu36	Qu42	Qu45	Qu48	Q50	Qu52	Qu53	Qu54	Qu55	Qu56
0	3,54	2,72	3,11	3,66	2,62	2,85	3,64	2,82	2,86	2,67	2,7	3,27	3,16	2,38
1	3,10483871	2,5	2,661290323	3,39516129	2,548387097	2,596774194	3,314516129	2,66129032	2,60483871	2,790322581	2,774193548	3,153225806	2,637096774	2,266129032
2	3,803921569	3,401960784	3,558823529	3,803921569	3,37254902	3,62745098	3,725490196	3,20588235	3,30392157	2,774509804	2,931372549	3,333333333	3,225490196	2,745098039
3	3,516483516	3,076923077	3,340659341	3,868131868	2,824175824	3,21978022	3,736263736	3,02197802	3,2967033	2,78021978	2,725274725	3,252747253	2,989010989	2,758241758
segment	Qu57	Qu58	Qu59	Qu60	Qu61	Qu62	Qu63	Qu64	Qu65	Qu66				
0	3,42	3,45	2,71	3,06	3,16	3,01	3,34	3,52	3,41	2,67				
1	2,822580645	3,161290323	2,274193548	2,709677419	3,129032258	2,85483871	3	3,16129032	2,91935484	2,306451613				
2	3,401960784	3,529411765	3,12745098	3,529411765	3,039215686	3,019607843	3,578431373	3,61764706	3,5	3,235294118				
3	3,285714286	3,43956044	2,516483516	3,230769231	2,89010989	2,923076923	3,494505495	3,68131868	3,3956044	2,769230769				

Рисунок 5 – Профилирующая таблица с сегментами данных

В результате проведённого анализа были выделены четыре кластера респондентов, для каждого из которых удалось сформулировать обобщённые характеристики:

1. Женщины, часто пользующиеся транспортными услугами, располагающие достаточными финансовыми средствами.
2. Мужчины, редко прибегающие к использованию транспортных услуг, с ограниченными финансовыми возможностями.
3. Мужчины, часто использующие транспортные услуги, обладающие высоким уровнем дохода.
4. Женщины, редко пользующиеся транспортными услугами, характеризующиеся низким уровнем дохода.

Заключение

Таким образом, в ходе проведённого исследования была осуществлена сегментация клиентов, а полученные результаты демонстрируют соответствие типовым кейсам, наблюдаемым в практической деятельности. Это подтверждает релевантность предложенного подхода для решения прикладных задач классификации и группировки респондентов.

В качестве направления для дальнейших исследований представляется целесообразным рассмотреть возможность увеличения числа выделяемых сегментов, например, до шести, с целью более детальной дифференциации групп потребителей. Кроме того, перспективным является применение альтернативных программных средств для реализации процедуры кластеризации (таких как Statistica и других специализированных пакетов) с последующим сравнением качества и устойчивости полученных результатов между выбранными инструментами.

Литература

1. Big Data [Электронный ресурс]. – Режим доступа: https://www. uplab.ru/blog/big-data-technologies.
2. Fabrigar, L.R., Wegener, D.T. Exploratory Factor Analysis / L.R. Fabrigar, D.T. Wegener. – Oxford University Press, 2011. – 176 p.
3. Jolliffe, I.T. Principal Component Analysis / I.T. Jolliffe. – Springer Science & Business Media, 2002. – 487 p.
4. Kohonen, T. Self-Organizing Maps (3-rd ed.) / T. Kohonen. – Springer Series in Information Sciences, Vol. 30. – Springer, 2001. – 502 p.
5. Kotler, P. Marketing Management / P. Kotler. – Toronto: Prentice Hall Canada, 2001. – 775 p.
6. Вдовин, В.М. Предметно-ориентированные экономические информационные системы / Вдовин В.М., Суркова Л.Е., Шурупов А.А. – изд. 3-е. – М.: Дашков и К, 2016. – 388 с.
7. Герасименко, В.В. Маркетинг: учебник / В.В. Герасименко, М.С. Очковская и др. – 3-е изд., перераб. и доп. – Москва: Проспект, 2016. – 512 с
8. Жданов, М.В. Моделирование информационного обеспечения системы управления предприятием / М.В. Жданов // Международный журнал гуманитарных и естественных наук. – 2024. – № 3-2(90). – С. 59-61.
9. Саенко, И.И. Применение динамического метода оценки конкурентоспособности при разработке стратегии устойчивого развития мясной отрасли Краснодарского края / И.И. Саенко, С.А. Дьяков // Вестник Академии знаний. – 2020. – № 1 (36). – С. 206-216.
10. Управление бизнес-процессами предприятия: учеб. пособие / сост. Е.В. Пирогова. – Ульяновск: УлГТУ, 2017. – 107 с.

References

1. Big Data [Elektronnyj resurs]. – Rezhim dostupa: https://www. uplab.ru/blog/big-data-technologies.
2. Fabrigar, L.R., Wegener, D.T. Exploratory Factor Analysis / L.R. Fabrigar, D.T. Wegener. – Oxford University Press, 2011. – 176 p.
3. Jolliffe, I.T. Principal Component Analysis / I.T. Jolliffe. – Springer Science & Business Media, 2002. – 487 p.
4. Kohonen, T. Self-Organizing Maps (3-rd ed.) / T. Kohonen. – Springer Series in Information Sciences, Vol. 30. – Springer, 2001. – 502 p.
5. Kotler, P. Marketing Management / P. Kotler. – Toronto: Prentice Hall Canada, 2001. – 775 p.
6. Vdovin, V.M. Predmetno-orientirovannye ekonomicheskie informacionnye sistemy / Vdovin V.M., Surkova L.E., Shurupov A.A. – izd. 3-e. –M.: Dashkov i K, 2016. – 388 s.
7. Gerasimenko, V.V. Marketing: uchebnik / V.V. Gerasimenko, M.S. Ochkovskaya i dr. – 3-e izd., pererab. i dop. – Moskva: Prospekt, 2016. – 512 s
8. Zhdanov, M.V. Modelirovanie informacionnogo obespecheniya sistemy upravleniya predpriyatiem / M.V. Zhdanov // Mezhdunarodnyj zhurnal gumanitarnyh i estestvennyh nauk. – 2024. – № 3-2(90). – S. 59-61.
9. Saenko, I.I. Primenenie dinamicheskogo metoda ocenki konkurentosposobnosti pri razrabotke strategii ustojchivogo razvitiya myasnoj otrasli Krasnodarskogo kraya / I.I. Saenko, S.A. D'yakov // Vestnik Akademii znanij. – 2020. – № 1 (36). – S. 206-216.
10. Upravlenie biznes-processami predpriyatiya: ucheb. posobie / sost. E.V. Pirogova. – Ul'yanovsk: UlGTU, 2017. – 107 s.