УДК 685.34
DOI 10.47367/0021-3497_2024_2_80

КЛАСТЕРНЫЙ АНАЛИЗ РЕЗУЛЬТАТОВ АНКЕТИРОВАНИЯ ПОКУПАТЕЛЕЙ ОБУВИ

CLUSTER ANALYSIS OF THE SURVEY RESULTS 0F SHOE BUYERS

А.Н. МАКСИМЕНКО, В.В. КОСТЫЛЕВА, И.Б. РАЗИН

A.N. MAKSIMENKO, V.V. KOSTYLEVA, I.B. RAZIN

(Российский государственный университет им. А.Н. Косыгина (Технологии. Дизайн. Искусство))

(Russian State University named after A.N. Kosygin (Technologies. Design. Art))

E-mail: all-max@mail.ru, kostyleva.vv@mail.ru, igor-razin@yandex.ru

В статье представлены результаты кластерного анализа анкетирования покупателей обуви. Предметом исследования выступают статистические данные, характеризующие потребительские предпочтения опрошенных респондентов. Методологической основой исследования являлись статистические методы, методы извлечения информации и интеллектуального анализа данных. В работе приведены

результаты опроса женской аудитории покупателей обуви в возрасте от 24 до 55 лет, проживающих в Москве и Санкт-Петербурге. Приведено распределение предпочтений покупателей по цене, месту покупки, количеству приобретенных пар обуви за последние полгода и ее назначению. Определены доли покупателей, показавших приверженность бренду при выборе обуви и покупавших обувь отечественного производства. Приведена оценка доверительного интервала, полученная методом Вальда. Проведена иерархическая кластеризация результатов опроса методом полной связи. Для расчета меры сходства между наблюдениями использовано расстояние Хэмминга. В результате анализа выявлено 6 кластеров наблюдений, интерпретация каждого из которых приведена в статье.

The article presents the results of a cluster analysis of a survey of shoe buyers. The subject of the study is statistical data characterizing the consumer preferences of the surveyed respondents. The methodological basis of the study was statistical methods, methods of information extraction and data mining. The paper presents the results of a survey of the female audience of shoe buyers aged 24 to 55 years old living in Moscow and St. Petersburg. The distribution of customer preferences by price, place of purchase, number of purchased pairs of shoes over the past six months and its purpose is given. The shares of buyers who showed commitment to the brand when choosing shoes and who bought shoes of domestic production are given. An estimate of the confidence interval obtained by the Wald method is given. The hierarchical clustering of the survey results by the method of full communication was carried out. The Hamming distance was used to calculate the measure of similarity between observations. As a result of the analysis, 6 clusters of observations were identified, the interpretation of each of which is given in the article.

Ключевые слова: анализ, данные, женская обувь, анкетирование, кластеризация, статистика.

Keywords: analysis, data, women's shoes, survey, clustering, statistics.

Введение

В условиях стремительного развития технологий интернет-продаж [1, 2] и изменений в структуре российского рынка, связанных с уходом западных компаний и появлением новых брендов на рынке [3, 4], особый интерес представляют исследования аудитории покупателей обуви.

Материалы и методы

Нами проведено анкетирование женской аудитории покупателей обуви, проживающих в Москве и Санкт-Петербурге. Опрос проведен в режиме онлайн с помощью сервиса Яндекс.Взгляд [5]. В опросе приняли участие 90 респондентов в возрасте от 24 до 55 лет. Для оценки доверительного интервала для каждой доли ответов из опроса нами использован метод Вальда [6]:

$$\hat{p} \pm z_{\gamma} \sqrt{\frac{\hat{p}_i(1 - \hat{p}_i)}{n}}, \qquad (1)$$

где n – это размер выборки; $\hat{p}$ – доля ответов на вопрос анкеты; z – стандартизированная оценка для доверительной вероятности γ. Доверительная вероятность γ в настоящем исследовании составляет 95%. Результаты анкетирования приведены в табл. 1.

Таблица 1

1. Сколько пар обуви Вы купили себе за последние полгода?				
1 пару	2 пары	3 пары	4 пары и более	Ни одной
31,1% (28) ± 9,55%	25,6% (23) ± 8,94%	20% (18) ± 8,26%	12,2% (11) ± 6,71%	11,1% (10) ± 6,46%

2. Где Вы покупали себе обувь в последние полгода?			
Фирменный интернет-магазин	Маркетплейс (Ozon, Wildberries и др.)	Фирменный розничный магазин	Другое
7,5% (6) ± 5,77%	67,5% (54) ± 10,26%	40% (32) ± 10,73%	13,75% (11) ± 7,54%

3. В каком ценовом сегменте Вы покупали обувь в последние полгода?			
до 5000 руб.	5001 – 11000 руб.	11001 – 21000 руб.	21001 руб. и выше
63,75% (53) ± 10,53%	40% (32) ± 10,73%	5% (4) ± 4,77%	1,25% (1) ± 2,43%

4. Как часто Вы делаете возвраты при интернет-заказах обуви?		
Часто (больше 80% возвратов)	Не очень часто (около 50% возвратов)	Относительно редко (менее 20% возвратов)
7% (4) ± 6,35%	16% (10) ± 9,12%	77% (48) ± 10,47%

5. Обувь какого назначения Вы покупали себе в последние полгода?				
Повседневную	Спортивную	Ортопедическую	Домашнюю	Другую
86,25% (69) ± 7,54%	51,25% (41) ± 10,95%	5% (4) ± 4,77%	20% (16) ± 8,76%	11,25% (9) ± 6,95%

6. Вы покупали новые для себя бренды обуви в последние полгода?	
Да	Нет
61,25% (49) ± 10,68%	38,75% (31) ± 10,68%

7. Вы покупали обувь отечественных брендов в последние полгода?	
Да	Нет
57,5% (46) ± 10,84%	42,5% (34) ± 10,84%

8. При покупке обуви Вы руководствовались модными тенденциями?	
Да	Нет
36,25% (29) ± 10,51%	63,75% (51) ± 10,51%

В таблице для каждого вопроса перечислены по нисходящему порядку: варианты ответов, доля каждого ответа (в скобках указано количество респондентов, выбравших данный ответ) и доверительный интервал для каждой доли.

Чтобы получить срез актуальных данных, опрос был ориентирован на респондентов, покупавших обувь в последние полгода. Поэтому в случае, если респондент отвечал, что не покупал обувь в последние полгода, опрос для него на этом заканчивался. В связи с этим выборка наблюдений для дальнейшего анализа составила 80 анкет. Также следует отметить, что вопросы № 2, 3 и 5 анкеты о месте покупки, цене и назначении купленной обуви допускали множественный выбор.

Из полученных данных следует, что за последние полгода 31,1% респондентов купили одну пару обуви, покупки совершались преимущественно на маркетплейсах Ozon, Wildberries и др. (67,5%) при относительно редких возвратах (менее 20%). Приобретена обувь большей частью повседневная (86,25%) стоимостью до 5000 рублей (63,75%) отечественных брендов (57,5%), при этом только 36,25% опрошенных руководствовались модными тенденциями.

Однако такая интерпретация результатов опроса недостаточно объективна и состоятельна, поскольку не дает оценки всему спектру ответов. Поэтому для обработки полученных данных нами использованы алгоритмы иерархической кластеризации – совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) кластеров. «Дерево», представляющее иерархическое слияние кла-

стеров в виде дендрограммы, можно визуализировать. Визуальный осмотр привлекателен для понимания структуры данных, особенно в нашем случае, когда размер выборки небольшой. Вернемся к анализу результатов проведенного нами анкетирования.

На первом шаге для иерархической кластеризации результаты опросов преобразованы нами в двоичные наборы длины n, где значению «1» соответствует утвердительный ответ на вопрос, а «0» – отрицательный. Длина набора «n» определяет число утвердительных ответов. В нашем случае при $n = 23$ получен массив данных размерностью $2 \times 80 \times 23$. На рис. 1 продемонстрирована визуализация массива данных результатов анкетирования покупателей обуви.

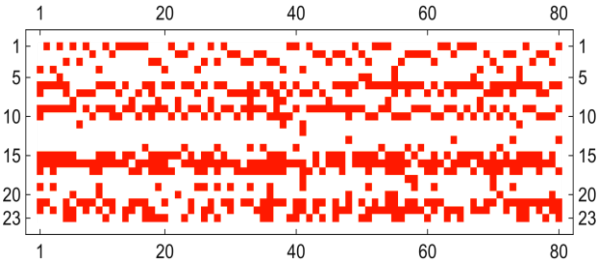

Рис. 1

Для расчета меры сходства наблюдений на двоичных наборах удобно использовать расстояние Хэмминга [7]:

$$d(a,b) = \sum_{i=0}^{n-1} (a_i \oplus b_i), \qquad (2)$$

где n – длина строки; a, b – двоичные наборы.

Рассчитав меры сходства по всем наблюдениям, мы получили матрицу расстояний порядка m=80 – количество исходных кластеров. Далее, найдя наиболее схожие наблюдения, итеративно объединили все наблюдения. Очевидно, что расстояние между идентичными наблюдениями будет равно 0, а между противоположными – 23.

Для оценки расстояния между кластерами наблюдений нами использован агломеративный метод полной связи [8]. Тогда расстояние D(A,B) между кластерами A и B:

$$D(A,B) = \max_{a \in A,\ b \in B} d(a,b), \qquad (3)$$

где $d(a,b)$ – расстояние Хэмминга между $a \in A,\ b \in B$, будет определяться как максимум из множества расстояний между элементом первого кластера и элементом второго.

Результат иерархической кластеризации итогов анкетирования получен в виде дендрограммы (рис. 2), на которой по горизонтальной оси указаны характеристики кластеров в формате $p(k)$, где p – номер кластера, а k – количество наблюдений в кластере.

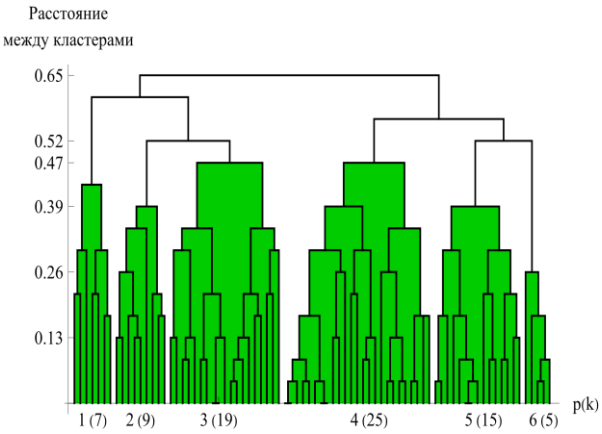

Рис. 2

Порог расстояния, применяемый при формировании новых кластеров, определен на уровне 0,47 на интервале $[0,1]$, поэтому совокупность представляется шестью классами. На вертикальной оси обозначены расстояния между кластерами, где 0 – минимальное расстояние между кластерами, а 0,65 – максимальное расстояние.

Результаты и обсуждения

Оказалась, что совокупность полученных наблюдений можно представить шестью кластерами. Наиболее объемный кластер №4 содержит 25 наблюдений. Высчитав среднее арифметическое по всем ответам на каждый вопрос анкеты, можно определить группирующие переменные каждого кластера и предложить характеристику каждой группе наблюдений.

Для наблюдений из кластера №4 характерны следующие корреляции между признаками «покупки обуви только на маркетплейсах» и «полное отсутствие покупок обуви в рознице». Респонденты этого кластера предпочитают покупать обувь в ценовом диапазоне до 5 тысяч рублей, при этом количество купленной обуви за последние полгода составило 1, 2 пары. Представители этой группы покупают в основном повседневную обувь, в редких случаях – спортивную.

Кластер №3 содержит 19 наблюдений. В эту группу вошли респонденты, которые, напротив, покупают обувь только в розничных магазинах и не покупают ее на маркетплейсах. Эта же группа отличается в среднем большим количеством пар обуви на человека. К тому же респонденты этой группы покупали обувь во всех предложенных ценовых сегментах. Примечательно, что ее представители очень редко покупают обувь отечественных марок. Это может быть связано с тем, что розничные сети отечественных производителей обуви представлены не так широко. Наряду с этим, большинством респондентов этой группы заявлено о покупках в последние полгода обуви новых для себя брендов.

В кластере №5 оказались покупатели, которые за последние полгода покупали много обуви. В этой группе, объединенной по признаку «покупки обуви в ценовом сегменте до 5000 рублей», не оказалось ни одного человека, кто бы купил только одну пару, но много тех, кто купил 3 и больше. При этом респонденты покупали обувь всех заявленных в анкете назначений. Отметим, что в этой группе респонденты уделяли наименьшее внимание модным тенденциям, покупали обувь как в рознице, так и в интернете.

Респонденты из кластера №2 покупали много обуви, в основном 3 пары и больше, в ценовом сегменте «5001 – 11000 руб.». Почти все респонденты этой группы покупали себе спортивную обувь. Они же больше, чем респонденты других кластеров, при выборе обуви руководствовались модой.

В кластере №6 оказались респонденты, которые на вопрос о месте покупки обуви ответили: «Другое», они же ответили: «Другую» и на вопрос о назначении купленной обуви. Очевидно, что этот случай требует отдельного от данного исследования рассмотрения.

В кластере №1 оказалось мало наблюдений – 7. Эту группу респондентов объединило стремление к покупкам обуви новых марок.

ВЫВОДЫ

Резюмируя вышеизложенное, можно заключить, что общим для всех кластеров является высокая ориентированность респондентов на покупку обуви на маркетплейсах, небольшой процент возвратов обуви при ее покупке через интернет, почти полное отсутствие покупок через фирменные интернет-магазины, относительно высокое стремление к новизне в выборе марки обуви, а также склонность к покупкам обуви отечественного производства.

Заметим, что в нынешних условиях и производители, и поставщики обуви отчасти интуитивно, отчасти по объективным причинам сфокусировались на продажах обуви через маркетплейсы. Выявленные нами позитивный спрос на обувь отечественного производства, тенденции к покупкам продукции новых марок свидетельствуют, что рынок находится в активной стадии формирования и открыт для новых участников. Таким образом, использование алгоритмов иерархической кластеризации как инструмента исследования результатов анкетирования покупателей обуви позволяет не только констатировать текущее состояние рынка, но и прогнозировать направления развития при его мониторинге

В исследовании использовано программное обеспечение: Wolfram Mathematica [9], Jupyter [10], SciPy [11], NumPy [12], Pandas [13].

ЛИТЕРАТУРА

1. *Ларионов В.Г., Шереметьева Е.Н., Балановская А.В.* Векторы цифровой трансформации текстильной промышленности // Изв. вузов. Технология текстильной промышленности. 2022. № 2(388). С. 12…20

2. *Maksimenko A.N., Kostyleva V.V., Volkova G.Y., Razin I.B.* Development of a customized orthopedic shoes application processing system // AIP conference proceedings, Tver, 10 сентября 2021 года. Vol. 2526. – AIP Publishing, 2023. P. 030016. – EDN NVXTLJ.

3. РБК – Бизнес. – https://www.rbc.ru/business/05/10/2023/651d4ba69a7947f63e85ef14 (дата обращения 22.12.2023)

4. *Филатов В.В., Мишаков В.Ю., Ломакина Е.В. и др.* Анализ проекта управления изменениями в рамках стратегии развития легкой промышленности в Российской Федерации на период до 2025 года // Изв. вузов. Технология текстильной промышленности. 2022. № 1 (397). С. 73…85.

5. Яндекс.Взгляд. – https://surveys.yandex.ru/ (дата обращения 22.12.2023)

6. *Wallis S.* Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods, Journal of Quantitative Linguistics, (2013), 20:3, 178-208. – DOI: 10.1080/09296174.2013.799918

7. *Hamming R.W.* Error detecting and error correcting codes. The Bell System Technical Journal. (April 1950), 29 (2): 147–160. – doi:10.1002/j.1538-7305.1950.tb00463.x. ISSN 0005-8580. S2CID 61141773

8. *Defays D.* An efficient algorithm for a complete link method. The Computer Journal. British Computer Society. (1977), 20 (4): 364–366. – doi:10.1093/comjnl/20.4.364.

9. *Wolfram S.* The Mathematica Book. (1996). – https://api.semanticscholar.org/CorpusID:60959654 (дата обращения 22.12.2023)

10. *Kluyver T., Ragan-Kelley B., Pérez F. et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. Elpub. 2016. P. 87…90.

11. *Virtanen P. et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17, 3 (2020), P. 261…272, arXiv:1907.10121 [cs.MS].

12. *Walt S., Colbert S.C. and Varoquaux G.* The NumPy array: a structure for efficient numerical computation // Computing in Science and Engineering 13, 22 (2011), P. 22…30, arXiv:1102.1523 [cs.MS].

13. *Mckinney W.*, pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer. (2011).

R E F E R E N C E S

1. *Larionov V.G., Sheremetyeva E.N., Balanovskaya A.V.* Vectors of the digital transformation of the textile industry // Izvestiya Vysshikh Uchebnykh Zavedenii, Seriya Teknologiya Tekstil'noi Promyshlennosti. 2022. № 2(388). P. 12…20

2. *Maksimenko A.N., Kostyleva V.V., Volkova G.Y., Razin I.B.* Development of a customized orthopedic shoes application processing system // AIP conference proceedings, Tver, 10 September 2021. Vol. 2526. – AIP Publishing, 2023. P. 030016. – EDN NVXTLJ.

3. RBC – Business. – https://www.rbc.ru/business/05/10/2023/651d4ba69a7947f63e85ef14 (date of application 22.12.2023)

4. *Filatov V.V., Mishakov V.Yu., Lomakina E.V. et al.* Analysis of the change management project under the strategy for the development of light industry in the Russian Federation for the period until 2025 // Izvestiya Vysshikh Uchebnykh Zavedenii, Seriya Teknologiya Tekstil'noi Promyshlennosti. 2022. № 1 (397). P. 73…85.

5. Yandex.Surveys. – https://surveys.yandex.ru/ (date of application 22.12.2023)

6. *Wallis S.* Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods, Journal of Quantitative Linguistics, (2013), 20:3, 178-208. – DOI: 10.1080/09296174.2013.799918

7. *Hamming R.W.* Error detecting and error correcting codes. The Bell System Technical Journal. (April 1950), 29 (2): 147–160. – doi:10.1002/j.1538-7305.1950.tb00463.x. ISSN 0005-8580. S2CID 61141773

8. *Defays D.* An efficient algorithm for a complete link method. The Computer Journal. British Computer Society. (1977), 20 (4): 364–366. – doi:10.1093/comjnl/20.4.364.

9. *Wolfram S.* The Mathematica Book. (1996). – https://api.semanticscholar.org/CorpusID:60959654 (дата обращения 22.12.2023)

10. *Kluyver T., Ragan-Kelley B., Pérez F. et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. Elpub. 2016, P. 87…90.

11. *Virtanen P. et al.*, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods 17, 3 (2020), P. 261…272, arXiv:1907.10121 [cs.MS].

12. *Walt S., Colbert S.C., and Varoquaux G.*, The NumPy array: a structure for efficient numerical computation // Computing in Science and Engineering 13, 22 (2011), P. 22…30, arXiv:1102.1523 [cs.MS].

13. *Mckinney W.*, pandas: a Foundational Python Library for Data Analysis and Statistics. Python High Performance Science Computer. (2011).

Рекомендована кафедрой информационных технологий РГУ им. А.Н. Косыгина. Поступила 31.01.24.