РАЗКРИВАНЕ НА ЗАКОНОМЕРНОСТИ ПРИ ПАЗАРУВАНЕ В СРЕДА НА ГОЛЕМИ ДАННИ

Доц. д-р Тодор Б. Кръстевич

Резюме

В тази студия представяме възможностите за приложение на избрани алгоритми за бързо откриване на закономерни, често срещани и повтарящи се модели на поведение при пазаруване на продукти или услуги, купувани едновременно и/или в някакво последователност във времето. Подобни анализи често пъти се определят с понятието „анализ на пазарната кошница“. Обект на настоящото изследване е анализът на големи масиви от данни от продажбени трансакции, наблюдавани в множество индивидуални актове на покупка, а предметът – разкриване на възможностите на някои числови алгоритми от областта на машинното обучение за откриване на скрити закономерности в актовете на пазаруване чрез обработка на данни от индивидуалните „пазарни кошници“ на клиенти. Целта е да се убеди читателят във възможностите на извличането на асоциативни правила от големи данни чрез демонстрации с общодостъпни отворени данни. Изложението следва логиката „от концепция към приложение“ и последователното дава отговори на въпросите „защо“ е необходимо да се прави, с „какво“ се прави и „как“ се прави. След кратко въведение в логиката и спецификата на най-популярните алгоритми за откриване на закономерности чрез извличане на асоциативни правила от големи масиви от данни, предлагаме детайлни работни процедури и инструкции за анализ и интерпретация на аналитичните резултати. В края се прави синопсис и се дават насоки и препоръки за използване на аналитичните процедури в маркетингов контекст.

Ключови думи: анализ на пазарната кошница, извличане на асоциативни и секвентни правила, продажбени трансакции, анализ на големи данни, анализ на поведението на купувачите.

JEL: M3, C8, C40.

UNCOVERING SHOPPING PATTERNS IN BIG DATA ENVIRONMENTS

Assoc. Prof. Todor B. Krastevich, PhD

Abstract

In this study, we present the feasibility of applying selected algorithms to rapidly detect regular, common, and recurring patterns of shopping behavior for products or services purchased simultaneously and/or in some temporal sequence. Such analyses are often defined by the term „'market basket analysis“. The object of the present study is the analysis of large data sets of sales transactions observed across multiple individual acts of purchase, and the subject matter is the discovery of the capabilities of some numerical algorithms from the field of machine learning to detect hidden patterns in acts of purchase by processing data from individual customer „market baskets“. The goal is to convince the reader of the possibilities of extracting association rules from big data through demonstrations with publicly available open data. The exposition follows a „concept-to-application“ logic and consistently provides answers to the questions of „why“ it is necessary to do it, „what“ it is done with, and „how“ it is done. After a brief introduction to the logic and specifics of the most popular algorithms for discovering patterns by extracting association rules from large datasets, we provide detailed working procedures and instructions for analyzing and interpreting the analytical results. Finally, we provide a synopsis and provide guidelines and recommendations for using the analytical procedures in a marketing context.

Keywords: market basket analysis, association and sequential rule mining, sales transactions, big data analytics, shopper behavior analysis.

JEL: M3, C8, C40.

Въведение

Определянето на навиците при пазаруване и тяхното изменение във времето е решаващо предизвикателство за ефективното управление на взаимоотношенията с клиентите. В този контекст обещаващи аналитични решения, които компаниите могат да внедрят, са анализирането и прогнозирането на потребителската кошница. Една ефикасна препоръчваща система, състояща се от адекватно извлечени асоциативни правила от анализ на пазарната кошница, може персонализирано да напомня клиента при съставянето на списък за пазаруване онлайн, предлагайки ограничен набор от

артикули, които да привличат вниманието му и да насърчават решението за покупка.

Използването на авангардни аналитични техники за насърчаването и улесняването на пазаруването не е единственият ефикасен начин за повишаване на маркетинговата ефективност. За да накарат повече хора да купуват техните продукти, компаниите трябва също така да разбират възможно най-добре своите клиенти и тяхното поведение. Използването на нови технологии за анализ на данни може да доведе до по-добро разбиране на клиентите. Ако търговците познават профилите на своите клиенти, това ще им позволи да продават по-добре своите продукти, както и да удовлетворяват клиентите (Ribeiro, 2016, p. 2).

Традиционните статистически методи за експлоративен и дескриптивен анализ са силно ограничени във възможностите си за разкриване на скрити закономерности в големи масиви от данни. Вместо тях все по-голямо значение за маркетинговата аналитика добиват методите и алгоритмите от областта на машинното обучение. Машинното обучение се отнася до прилагане на методи или алгоритми, предназначени за изучаване на основни закономерности и модели в данните и за изготвяне на прогнози въз основа на тези модели. Инструментите на машинното обучение първоначално са разработени в областта на компютърните науки, а напоследък навлизат масирано и в бизнес приложенията. Ключова характеристика на техниките на машинно обучение е способността им да произвеждат точни прогнози извън извадката (Dzyabura & Yoganarasimhan, 2018, p. 255), както и да разкриват успешно скрити закономерности и взаимовръзки в големи масиви от данни.

Академичните изследвания в областта на маркетинга традиционно се фокусират върху причинно-следствените закономерности. Фокусът върху причинно-следствената връзка произтича от необходимостта да се правят алтернативни изводи и прогнози. Например към кои потребители да се насочат маркетинговите усилия, коя е конфигурацията на продукта, която купувачът най-вероятно ще избере, коя версия на рекламен банер ще генерира повече кликвания или купуването на кои продукти зависи от търсенето и купуването на други продукти. Това са все прогнозни проблеми. Тези проблеми не изискват разкриване на причинно-следствена връзка, а по-скоро извличането на модели с висока точност на прогнозиране извън извадката. Инструментите на машинното обучение и вчастност генерирането на асоциативни правила могат да се справят с тези видове проблеми.

В настоящата студия представяме възможностите за приложение на алгоритми за бързо откриване на закономерни, често срещани и повтарящи се модели на поведение при пазаруване на набори комбинации от продукти или услуги, купувани едновременно и/или последователно във времето. Обект на изследването е анализът на големи масиви от данни от продажбени трансакции, наблюдавани в множество индивидуални актове на покупка, а

предметът – разкриване на възможностите на някои числови алгоритми от областта на машинното обучение за откриване на скрити закономерности в актовете на пазаруване чрез обработка на данни от индивидуалните „пазарни кошници“ на клиенти. Целта е да се убеди читателят във възможностите на извличането на асоциативни правила от големи данни чрез демонстрации с общодостъпни отворени данни. Постигането на тази цел се конкретизира от две свързани помежду си изследователски задачи: да се разработят процедури за извличане на (1) асоциативни и (2) секвентни правила от продажбени трансакции. Изложението следва логиката „от концепция към приложение“ и последователното дава отговори на въпросите „защо“ е необходимо да се прави, с „какво“ се прави и „как“ се прави. Започваме с кратко въведение в логиката и спецификата на най-популярните алгоритми за откриване на закономерности чрез извличане на асоциативни правила от големи масиви от данни. След това предлагаме детайлни работни процедури и инструкции за анализ и интерпретация на аналитичните резултати с конкретни примери. В края прави се прави синопсис и се дават насоки и препоръки за използване на резултатите в маркетингов контекст.

I. Анализ на пазарната кошница

Използваме понятието „пазарна кошница“ в смисъл на набор от продукти и/или услуги, които се купуват едновременно в отделен акт на покупка. Данните за пазарната кошница описват какво купуват клиентите. Анализът на тези данни е сложен и нито една техника не е достатъчно мощна, за да даде всички отговори на въпроси, които интересуват маркетолозите. Самите данни обикновено описват пазарната кошница на три различни нива: (1) поръчката, т.е. самото събитие на покупката; (2) артикулите, попадащи в отделното събитие на покупката и (3) клиентът, който е свързващото звено между отделните поръчки в хода на времето.

Чрез анализирането на продажбите на продукти във времето е възможно да се отговори на редица важни въпроси за поведението на клиентите, като например кои са най-продаваните продукти или кои артикули, продавали се добре пред предходен период, вече не се продават така добре през текущия период. По-важната за маркетингови решения информация, която се извлича чрез анализ на продажбените трансакции, обаче е за ефекта от маркетинговите интервенции – в смисъл дали продажбите са се увеличили или намалили след определено събитие.

Извличането на асоциативни правила от данни за продажбените трансакции е една от най-мощните и адекватни техники за откриването на устойчиви закономерности в поведението при купуване. С помощта на този клас аналитични техники е възможно да се откриват продукти, при които се наблюдават устойчиви тенденции да се продават заедно, в група. Понякога идентифицирането на подобни често срещани множества е достатъчно за

разбиране и предсказване на поведението на клиента. В други случаи обаче е необходимо да се извлекат друг тип ясни правила, показващи причинно-следствени връзки, т.е. когато присъствието на определени артикули в пазарната кошница в дадена поръчка (акт на покупка, идентифициран в конкретен момент от време), предполага висока вероятност за присъствие на същите и/или други артикули в последващи във времето поръчки.

Основните показатели за оценяване на силата и устойчивостта на асоциативните правила са подкрепа, доверие и интерес. Подкрепата показва колко често дадено правило се среща в данните за трансакциите. Доверието измерва степента на надеждност на извода, изведен от съответното правило, т.е. колко често правилото е вярно. Интересът показва силата на асоциация и се дефинира като отношение между подкрепата на съответното правило и съвместната подкрепа на всяко отделно случайно наблюдавано подмножество, т.е. спрямо ситуацията на липса на правило.

В общия случай извежданите асоциативни правила е възможно да се подразделят в три отделни категории: т.нар. (1) „полезни“ правила, обясняващи закономерност, която може би е била неочаквана; (2) „тривиални“ правила, обясняващи връзки, за които така или иначе се предполага и/или се знае, че съществуват и (3) „необясними“ правила, които нямат логически смисъл. Последните често имат много ниска степен на подкрепа.

Анализът на пазарната кошница чрез извличането на асоциативни правила предоставя възможности за детайлен анализ на закономерностите между отделните продуктови артикули през призмата на поведението при купуване едновременно. Този тип анализи е възможно да бъдат разширени и допълвани чрез идентифициране на устойчиви, закономерни секвенции на актовете на покупка, които освен закономерностите в съдържанието на „кошницата“ отчитат и реда, в който те се извършват.

II. Извличането на асоциативни правила от продажбени трансакции

Когато клиентите си купуват чипс, склонни ли са в същия акт на покупка за вземат шоколад или бира? Ако хората имат модерен смартфон, склонни ли са си купуват безжични слушалки и смарт часовник? Ако домакинствата си сключват автомобилна застраховка, купуват ли си и застраховка на жилище? Отговорите на тези въпроси могат да послужат за основа на позиционирането на марката, рекламата и дори на директния маркетинг. Но как да установим дали съществуват подобни асоциации и как да започнем да ги търсим, когато в клиентските бази данни се съдържат стотици хиляди записи от продажбени трансакции и много полета? Отговори на тези въпроси е възможно да бъдат търсени с помощта на компютърно базирани числови алгоритми.

Концептуално погледнато, алгоритмите за откриване на асоциации предоставят правила[1], описващи кои стойности на полета в клиентските бази данни с продажбени трансакции се срещат заедно и/или в определена последователност. Именно тези правила впоследствие биха могли да се използват за извеждане на приоритети при планиране на промоционални кампании, вземане на маркетингови решения за пакетни предложения, продуктово позициониране, кръстосани продажби и персонализирани препоръки към клиентите в процеса на вземане на решение за покупка.

В областта на машинното обучение са известни два основни числови алгоритъма, приложими за извличането на асоциативни правила от набори от данни за продажбени трансакции – **априорен алгоритъм**, предложен от Агравал и колектив (Agrawal, Imielinski, & Swami, 1993; Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994) и **непрекъснат алгоритъм** за извличане на асоциативни правила (Hidber, 1999), известни под абревиатурата CARMA (от анг. **C**ontinuous **A**ssociation **R**ule **M**ining **A**lgorithm).

Априорният алгоритъм ограничава пространството за търсене на правила, като открива често срещани множества и разглежда само правилата, които са съставени от често срещани множества[2]. С този алгоритъм се обработват елементи и множества от елементи, които съставляват трансакции. Елементите са условия от дихотомен тип (1/0), показващи наличието или отсъствието на определен артикул в конкретна трансакция. Наборът от елементи е група от артикули, които могат да имат или да нямат склонност да се срещат едновременно в рамките на трансакциите. Ако честотата на едновременно наблюдаваните елементи е висока, то това може да се интерпретира като закономерност и респ. да се направи опит, да се изведе условно логическо правило за описание на тази закономерност (т.нар. асоциативно правило). Дясната страна на всяко логическо правило е прието да се нарича антецедент. Антецедентът представлява условната причина. Лявата страна на правилото се нарича консеквент. Консеквентът показва логическото следствие. Концептуално асоциативните правила могат да се опишат във следния примерен формат:

	Консеквент		**Антецедент(-и)**
Правило 1:	Вестник	←	Кафе, Цигари
Правило 2:	Вода	←	Дизел, Антифриз, Сандвич
…	…	…	…
Правило R:	…	←	…

Правило 1 от горния хипотетичен списък би следвало да гласи, че лицата, които купуват бензин, кафе и цигари, вероятно ще купуват и вестник.

[1] С понятието „асоциативно правило“ обозначаваме честотата на трансакциите на един набор от закупени продуктови артикули като условие за закупуване и на друг набор от продуктови артикули.

[2] За подробности вж. Кръстевич (Анализ на пазарната кошница с R, 2021, стр. 40-46)

Процедурата за прилагане на априорния алгоритъм се състои от два етапа. На първия етап се идентифицират често срещани множества в данните, а на втория етап се генерират правила на база списъка с извлечените често срещани множества.

Понятието „често срещано множество“ се дефинира като съвкупност от елементи, чиято стойност на показателя „подкрепа“ е по-голяма или равна на зададения от анализиращия минимален праг на „подкрепа“ 3. Подкрепата на даден набор от елементи (продуктови артикули) е броят на записите, в които е намерен наборът от елементи, разделен на общия брой записи в базата данни.

Алгоритъмът стартира със сканиране на данните и идентифициране на множествата от единични елементи (т.е. единични множества), които отговарят на зададения минимален праг на показателя „подкрепа“. Всички единични елементи, които не отговарят на критерия, не се разглеждат по-нататък. След това итеративно се добавят нови и нови елементи, чиято „подкрепа“ надхвърля минималния зададен праг.

След като всички често срещани множества се идентифицират, за всяко от тях се прилага следната процедура:

(1) Изчисляват се всички възможни съдържащи се в него подмножества от елементи;

(2) За всяко подмножество се изчислява показателят „доверие“, показващ колко често елементите, присъстващи в конкретното подмножество Y (консеквент), присъстват в трансакции, които съдържат друго подмножеството X (антецедент). Показателят „доверие“ измерва по същество степента на надеждност на извода, изведен от съответното правило, т.е. колко често правилото е вярно. Колкото е по-висока степента на доверие на дадено правило X→Y, толкова е по-висока вероятността, подмножеството Y да присъства в трансакции, съдържащи подмножеството X;

(3) Ако стойността на показателя „доверителност“ е по-висока от зададен от потребителя минимален праг, правилото се запазва и се приема за закономерно (IBM Corp., 2020, p. 10).

Показателите „подкрепа“ и „доверителност“ измерват силата на всяко идентифицирано асоциативно правило. Освен тях важен показател за значението на правилата има и показателят „интерес“, въведен от Сергей Брин и колектив (Brin, Motwani, Ullman, & Tsur, 1997). Този показател показва силата на асоциация и се дефинира като отношение между доверителността на съответното правило и съвместната подкрепа на всяко отделно подмножество, присъстващо в правилото (Borgelt, 2012, p. 450).

В най-общ смисъл оценяването на извлечените асоциативни правила е възможно да се извърши на базата на едновременното обследване на

[3] Показателят „подкрепа“ е въведен от Агравал и колектив (Agrawal, Imielinski, & Swami, 1993) и показва колко често дадено асоциативно правило е приложимо към конкретен набор от данни.

цитираните показатели, придружени от информация за абсолютната и относителна честота на често срещаните множества от артикули.

Априорният алгоритъм е сравнително ефективен подход за откриване на силни асоциативни правила в пазарната кошница, но има ограничението, че работи само и единствено с категорийни променливи.

Алтернатива на априорния алгоритъм е споменатият CARMA-алгоритъм (Hidber, 1999). Този алгоритъм е относително по-гъвкав, тъй като позволява промяна на минималните прагове на „подкрепа“ в процеса на генериране на асоциативни правила. Освен това позволява извличането на асоциативни правила, съдържащи повече от един консеквенти. CARMA пренася извличането на асоциативни правила в реално време, тъй като дава непрекъсната обратна връзка, контролира се от потребителя и дава детерминирани и точни резултати.

При CARMA-алгоритъма във фазата на оценяване се използва първоначално обработване на данните, чрез което да идентифицират кандидати за често срещани множества от елементи. За съхраняване на информацията за елементите се използват т.нар. „геометрични решетки“, т.е. набор от вектори в евклидовото пространство, образуващи дискретни групи чрез добавяне. Всеки елемент на решетката съхранява елементите и три стойности за съответния набор от елементи: броят на трансакциите, съдържащи набора от елементи, откакто наборът от елементи е добавен в решетката; индексът на записа на трансакцията, за която наборът от елементи е добавен в решетката и горната граница на броя на случаите на присъствие на елементите, преди те да бъдат добавени към решетката. След като кандидатите за често срещани множества са идентифицирани, се извършва повторно обработване на данни, за да се изчислят точните честоти за кандидатите и се определи окончателният списък на често срещаните множества въз основа на тези честоти. След това се използва общ алгоритъм за извличане на закономерности от геометричната решетка, стремящ се да премахне излишните правила (Aggarwal & Yu, 1998).

III. Процедури за извличането на асоциативни правила от продажбени трансакции

Запознавайки се с предходното изложение, прагматичният читател би следвало веднага да се запита, какви са възможностите за бързо и ефективно изчисляване на тези показатели с помощта на големи масиви от данни и как чрез анализирането им да се състави списък от силни асоциативни правила? За да дадем практически насоки, в този раздел ще демонстрираме провеждането на анализ на пазарна кошница с относително малък набор от реални анонимизирани данни, представляващи 786 продажбени трансакции, регистрирани по време на пазаруване в супермаркет. С цел осигуряване на

проследимост и възпроизводимост на резултатите, предоставяме на читателя свободен достъп до анонимизирани първични данни[4]. Файлът съдържа информация за артикулите, закупени от клиентите на супермаркет при еднократно пазаруване в определен времеви диапазон[5]. Наблюдавани са покупки в 10 различни продуктови категории, както и някои демографски данни за клиента (пол, възраст и др.), достъпни от регистрациите на клиентска карта за лоялни клиенти. Всеки запис представлява отделно посещение на магазин, при което е закупен поне един продукт. Първите няколко реда от набора от данни е представен на Таблица 1.

Таблица 1.

Примерна структура на анонимизирани данни за продажбени трансакции (https://bit.ly/2Zkgf3h)

Топла витрина	Замразени храни	Алкохол	Плодове и зеленчуци	Мляко	Тестени изделия	Прясно месо	Тоалетни принадлежности	Шоколадови изделия	Консерви	Пол	Възраст	Семейно положение	Деца до 18 г.	Трудова заетост
1	0	0	0	0	0	0	0	1	0	Жена	18 до 30	Овдовял/-а	Не	Да
1	0	0	0	0	0	0	1	0	0	Жена	18 до 30	Съжителство без брак	Не	Да
1	0	0	0	0	0	0	1	1	0	Мъж	18 до 30	Неженен/Неомъжена	Не	Да
1	0	0	0	1	1	0	0	0	0	Жена	18 до 30	Овдовял/-а	Не	Да
1	0	0	0	0	0	0	0	0	0	Жена	18 до 30	Съжителство без брак	Не	Да
1	0	0	0	0	1	0	0	1	1	Мъж	18 до 30	Неженен/Неомъжена	Не	Не
1	0	0	0	0	1	0	0	0	0	Жена	18 до 30	Неженен/Неомъжена	Не	Не
1	0	0	0	0	0	0	0	0	0	Жена	18 до 30	Овдовял/-а	Не	Не
1	0	0	0	1	0	0	0	1	1	Жена	18 до 30	Неженен/Неомъжена	Не	Не
1	0	0	0	0	0	0	0	0	0	Жена	18 до 30	Неженен/Неомъжена	Не	Не
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
0	1	0	0	0	0	0	0	0	0	Мъж	18 до 30	Неженен/Неомъжена	Не	Да

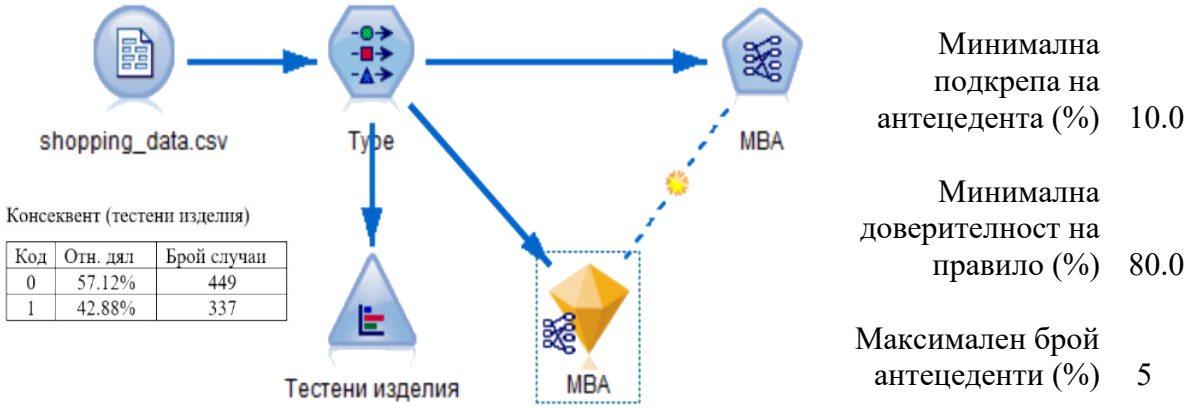
Тъй като изходните данни са дихотомни (вж. защрихованите променливи в Таблица 1), аналитичната процедура следва да се извърши с помощта на априорния алгоритъм. В случая е използвано софтуерното приложение

[4] Данните са достъпни на адрес https://bit.ly/2Zkgf3h

[5] За подробности относно алтернативни форми за структуриране на масиви с данни от продажбени трансакции вж. Кръстевич, Т. (Анализ на пазарната кошница с R, 2021)

IBM SPSS Modeler за извличане на данни и машинно обучение. Алтернативно, същите резултати биха могли да се възпроизведат със SAS Enterprise Miner, както и със софтуер с отворен код (например R/RStudio или Python).

Процедурата за извличане на списък със силни асоциативни правила и тяхното описание чрез ключови показатели е представена на Фигура 1. В случая са извлечени само тези правила, кои отговарят едновременно на зададени ограничителни условия от минимален праг на подкрепа от 10% от извадката, минимален праг на доверителност от 80% и максимален брой 5 на продуктите в антецедента. Въз основа на зададените критерии априорният алгоритъм търси правила и отхвърля тези от тях, които не представляват интерес. На Таблица 2 са изведени идентифицираните правила, които отговарят на зададените критерии.



Фигура 1. Работна процедура за извличане на силни асоциативни правила с априорен алгоритъм

От данните на Таблица 2 е възможно да се разбере, че мляко и замразените храни са били закупени при 85 акта на пазарувания. Припомняме, че файлът има 786 записа на трансакции, така че това е (85/786)*100, или 10,81% от общия брой трансакции („Подкрепа %“). Стойността на показателя „Интерес“ е очакваната възвръщаемост в резултат на използването на съответното правило. Тук тя е изчислена като съотношение между доверителността към базовия процент на консеквента. Например тестените изделия са били закупени при 42,88% от актовете на пазаруване като цяло (илюстрирано на Фигура 1), но са били закупени в 83,5% от случаите, когато са били закупени мляко и замразени храни („Доверителност“). „Интересът“ от използването на това правило се получава, като 83,529/42,88 = 1,948 и означава, че шансовете за закупуване на тестени изделия почти се удвояват, когато се купуват мляко и замразени храни.

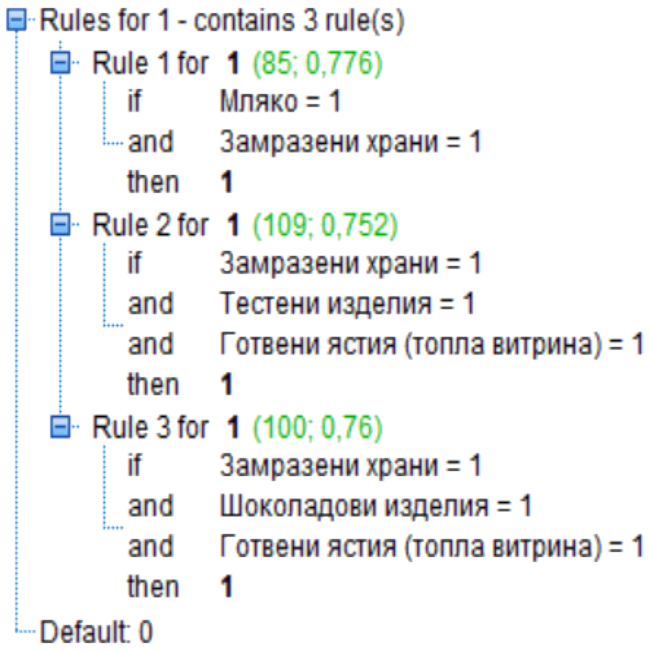
Таблица 2.
Извлечени асоциативни правила от набора с данни shopping_data.csv с априорен алгоритъм

№	Консек-вент	Антецеденти	Чес-тота	Подкре-па антеце-дент %	Дове-ри-тел-ност %	Под-крепа пра-вило %	Ин-терес
R 1	Тестени изделия	Мляко и Замразени храни	85	10.814	83.529	9.033	1.948
R 2	Тестени изделия	Алкохол, Консерви и Готвени ястия (топла витрина)	95	12.087	83.158	10.051	1.940
R 3	Тестени изделия	Замразени храни, Консерви и Шоколадови изделия	90	11.450	82.222	9.415	1.918
R 4	Готвени ястия (топла витрина)	Алкохол, Консерви и Хлебни изделия	97	12.341	81.443	10.051	1.654

Чрез коригиране на минималния праг на доверителност е възможно да се увеличи броят на извлечените асоциативни правила. Например при праг на доверителност 75%, техният брой нараства до 26. Предизвикателството тук е да се разгледат правилата и да се открият онези от тях, които могат да бъдат полезни в маркетингов контекст и/или за постигане на конкретни цели.

Интересен аспект от анализа е създаването на набор от правила, отнасящи се за избран консеквент. Например, ако мениджърът на продуктовата категория „Алкохол“ желае да установи кои комбинации от продукти предсказват покупката на алкохол, възможно е да се идентифицира съответният набор от правила, които да тестват наличието на подобна хипотетична закономерност. За целта е възможно да се използва C5.0 алгоритъм (RuleQuest Research Ltd Pty, 2020), спадащ към класа класификационни алгоритми, основаващи се на правила и работещи с номинални (дихотомни) зависими променливи (Kuhn & Johnson, 2018, p. 392). Моделът C5.0 работи чрез разделяне на извадката въз основа на полето, което осигурява максимален обем информация. След това всяка подизвадка, определена от първото разделяне, се разделя отново, обикновено въз основа на друго поле, и процесът се повтаря, докато подизвадките не могат да бъдат разделени повече. Накрая разделянията на най-ниско ниво се преглеждат отново и тези, които не допринасят значително за стойността на модела, се премахват или орязват.

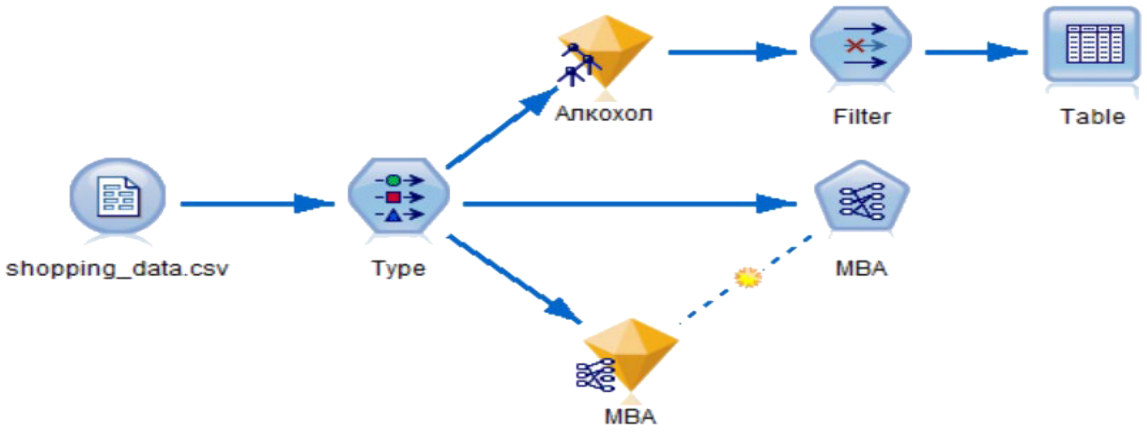
С помощта на C5.0 алгоритъма, приложен върху данните в текущия пример, се идентифицират три такива правила, чиито ключови параметри са интерпретирани на Фиура 2.



Идентифицирани са три правила, чийто консеквент е купуването на алкохол. Първото правило асоциира с купуването на алкохол покупката на мляко и замразени храни. Това правило има честота на антецедента 85 (в 85 акта на пазаруване в пазарната кошница присъства мляко и замразени храни) и 77,6% доверителност (т.е. в 77.6% от случаите, в които са закупувани мляко и замразени храни, е бил закупен и алкохол). Тези два показателя при правило 2 (асоцииращо консеквента с покупката на замразени храни, тестени изделия и готвени ястия)) са 109 и 75,2%, а при правило 3 (асоцииращи алкохола с покупка на замразени храни, шоколадови изделия и готвени ястия) съответно 100 и 76,0%. Това са трите най-силни правила, предсказващи покупката на продукт от категорията „Алкохол“ в набора от наблюдаваните данни от продажбените трансакции.

Фигура 2. Набор от правила, идентифициращи появата на избран консеквент „Алкохол“

С помощта на този набор от правила е възможно да се провери всяка една регистрирана трансакция, дали отговаря или не на условията на някое от трите правила. С помощта на IBM SPSS Modeler работната процедура се проектира по следния начин (Фигура 3):



Фигура 3. Работна процедура за анализ на пълния списък на продажбени трансакции на базата на генерирания набор от три правила с целеви консеквент „Алкохол“

Резултатът от изпълнение на процедурата съдържа списък със 786 наблюдавани трансакции, в които присъства поне един от продуктите, участващи в идентифицирания набор от три правила, представени на Фигура 2. Първите 15 реда от списъка са представени на Таблица 3. Освен информация за продуктите, в списъка присъстват и две нови променливи – $A-Алкохол и $AC-Алкохол. Първата от тях е дихотомна и съдържа кодове 0 и 1. С код 1 са отбелязани трансакциите, за които са валидни поне едно от трите правила от Фигура 2. Втората променлива $AC-Алкохол съдържа стойностите на доверителност за всяко правило. В случаите, когато условията на идентифицираните правила не се прилагат за конкретната трансакция, стойността на доверителност е 0,5.

От Таблица 3 е видно, че при трансакция 12 е приложено третото правило, тъй като доверителната стойност е 0,763. Използването на набор от правила позволява да определите кои клиенти са свързани с дадено правило. Оттук нататък могат да се правят различни видове други анализи (например, по-вероятно ли е клиентите, които отговарят на това правило, да са мъже или жени, по-млади или по-възрастни и т.н.). На Таблица 4 са изведени честотните разпределения на социодемографските променливи на потребителите, в чиято пазарна кошница се съдържа продукт от категорията „Алкохол“ (дефинирани като масов пазар) в сравнение с профила на тези, чието поведение при покупката на алкохол е предсказуемо с помощта на идентифицираните асоциативни правила (дефинирани като сегмент „Алкохол“). Сравнително ясно може да се разграничи профилът на този тип специфични клиенти (163 на брой) – преобладаващо мъже, на възраст между 51 и 60 години, разведени или в съжителство без брак, без деца до 18-годишна възраст в домакинството и активни на пазара на труда. С помощта по-

добно „профилиране“ е възможно да се планират таргетирани промоционални кампании на мястото на продажбите, отчитащи спецификата на купувача.

Таблица 3

Нови полета и идентификатори, генерирани с помощта на процедурата от Фигура 3

ID	Готвени ястия (тпл. витрина)	Замразени храни	Алкохол	Мляко	Тестени изделия	Шоколадови изделия	$A-Алкохол	$AC-Алкохол
1	1	0	0	0	0	1	0	0.500
2	1	0	0	0	0	0	0	0.500
3	1	0	0	0	0	1	0	0.500
4	1	0	0	1	1	0	0	0.500
5	1	0	0	0	0	0	0	0.500
6	1	0	0	0	1	1	0	0.500
7	1	0	0	0	1	0	0	0.500
8	1	0	0	0	0	0	0	0.500
9	1	0	0	1	0	1	0	0.500
10	1	0	0	0	0	0	0	0.500
11	1	0	0	0	0	0	0	0.500
12	1	1	1	1	1	1	1	0.763
13	1	0	0	0	0	0	0	0.500
14	1	0	0	0	0	1	0	0.500
15	1	0	0	0	0	0	0	0.500
…	…	…	…	…	…	…	…	…

Таблица 4

Социодемографски профил на групата клиенти, чието поведение е предсказуемо с помощта на дефинираните асоциативни правила от Фигура 2

		Масов пазар (n = 786)		**Сегмент "Алкохол" (n = 163)**		
		Брой	%	Брой	%	
Пол	Жена	341	54,7%	82	50,3%	
	Мъж	282	45,3%	81	49,7%	▲
Възраст	18 до 30	189	30,3%	47	28,8%	
	31 до 40	153	24,6%	42	25,8%	
	41 до 50	112	18,0%	22	13,5%	
	51 до 60	96	15,4%	34	20,9%	▲
	Над 60	73	11,7%	18	11,0%	
Семейно положение	В брак	152	24,4%	37	22,7%	
	Неженен/Неомъжена	165	26,5%	34	20,9%	
	Овдовял/-а	120	19,3%	30	18,4%	
	Разведен/-а	72	11,6%	26	16,0%	▲
	Съжителство без брак	114	18,3%	36	22,1%	▲
Деца до 18 г.	Да	243	39,0%	30	18,4%	
	Не	380	61,0%	133	81,6%	▲
Трудова заетост	Да	507	81,4%	147	90,2%	▲
	Не	116	18,6%	16	9,8%	

Дотук бяха представени някои основни насоки за анализ на пазарната кошница, основаващи се на най-популярния априорен алгоритъм за извличане на асоциативни правила. Следва да е вече ясно, че алгоритъмът се управлява от три ограничителни критерия – минимална подкрепа, минимална доверителност и максимално допустим брой елементи в антецедента. В зависимост от спецификата на изходните данни с тези три критерия е възможно да се експериментира. Обикновено в началото на анализа подкрепата е резонно да се намали, за да се генерират повече потенциални правила, тъй като е възможно при твърде високи стойности на подкрепа и доверителност да не бъдат изобщо открити силни правила.

Априорният алгоритъм използва подкрепата на антецедента и доверителността на правилата, за да определи кои правила да бъдат запазени. Най-често водещият критерий за определяне на дадено правило като силно е неговата оценка за доверителност. Този показател обаче не е непременно

най-добрият при всички обстоятелства. По същество, ако правилата се подбират само с помощта на доверителността, то алгоритъмът ще генерира списък, съставен само и единствено от точни правила. Точните правила обаче не са непременно най-интересни от маркетингова гледна точка. Нека си представим, че в един супермаркет 40% от всички клиенти купуват минерална вода по време на пазаруване и е намерено просто правила, което гласи, че „ако е купен плодов сок, то е купена и минерална вода“. Да допуснем, че доверителността на това правило е 43%, т.е. 43% от клиентите, които купуват плодов сок, купуват и вода. Дали това правило е съществено от маркетингова гледна точка? Колкото и странно да звучи, отговорът би следвало да е отрицателен, защото купуването на сок не оказва съществено влияние върху това, дали някой купува минерална вода, тъй като процентите са почти еднакви. Ако обаче минималният праг на доверителност бъде фиксиран на 20%, алгоритъмът ще „открие“ това относително тривиално правило на базата на оценката на показателя доверителност.

Друга често срещана ситуация е, когато консеквентът се появява по-рядко, когато се добавят условия. Да предположим, че доверителността в правилото "ако вода, то сок" не е 43%, а 15%. При граница на доверителност 40% това правило не би било избрано, но може да се окаже много важно за търговеца на дребно да знае за тази зависимост! Заедно с минералната вода плодовите сокове се купуват по-рядко, отколкото изобщо се купува. Може би те са някакъв вид заместител на водата? Във всеки случай правило с ниска степен на доверителност би могло да бъде интересно от маркетингова гледна точка, тъй като е еквивалентно на отрицание – тъй като "вода и сок" с 20% степен на доверителност е равнозначно "вода без сок " с 80% степен на доверителност.

В този контекст е възможно при използването на априорния алгоритъм да се задават както верни (1), така и неверни (0) значения на дихотомните променливи на отделните продукти и тогава ще бъдат открити отрицателни връзки[6].

Изводът е, че потенциално интересните правила често са тези, които имат големи разлики в доверителността си в сравнение с базовата стойност на достоверност на самия консеквент. Добавянето на елемент към антецедента е информативно само ако значително променя достоверността на правилото; в противен случай е достатъчно по-простото правило. Показателят „интерес“, който по същество е отношение на доверителността на правилото към базовата стойност на консеквента, също е чувствителен към големи увеличения на доверителността.

[6] Обикновено се откриват толкова много негативни връзки, че тази настройка невинаги е ефективно средство за откриване на асоциации.

За търсенето и идентифицирането на интересни правила е възможно да се ползват четири алтернативни оценъчни показателя: абсолютна доверителна разлика, коефициент на доверителност, информационна разлика и нормализиран хи-квадрат (IBM, 2010, pp. 4-2). Всеки един от тях сравнява текущата доверителност на дадено правило с доверителността на правило с празен антецедент, т.е. само съдържащо консеквент. Доверителността на „празното“ правило представлява просто относителната честота на консеквента и се определя като „априорна“ доверителност. Доверителността на правило с един или повече антецеденти може да се нарече „апостериорна“ доверителност. Показателят „интерес“ по същество представлява частно на апостериорната и априорната доверителност.

Абсолютната доверителната разлика (CD) е най-простият показател и се основава на абсолютната разлика между апостериорната и априорната доверителност (Borgelt, 2017). При използване на априори алгоритъма обикновено се задава първоначална долна граница на тази разлика от 10%, т.е. това да е минималната разлика между двете доверителности. Ограничителното условие може да се представи в следния общ вид:

$$\left|\frac{\text{Апостериорна доверителност}}{\text{Априорна доверителност}}\right| > CD_L,$$

като CD_L представлява долната граница на абсолютната доверителна разлика CD.

Така правило, чието заключение е почти винаги вярно, ще бъде избрано само ако априорната му доверителност е ниска. Освен това, като се има предвид, че се използва абсолютната стойност на разликата, чрез тази мярка могат да се открият и отрицателни правила.

Ако доразвием последния хипотетичен пример с покупките на сок и вода и зададем като долна граница на абсолютната доверителна разлика 50%, при априорна доверителност на сока от 40%, за да бъде избрано правилото „ако вода, то сок“, то доверителността му трябва да надхвърли 90%.

Друг сравнително прост начин за сравняване на двете доверителни стойности е да се изчисли тяхното съотношение и да се извади от 1. В резултат се получава т.нар. коефициент на доверителност. Тъй като този метод използва съотношение, за разлика от абсолютната разлика в доверителните стойности, той е по-чувствителен към съотношенията в по-ниските доверителни области. (Съотношението между 30% (априорна) и 40% (апостериорна) е по-голямо от това между 70% (априорна) и 80% (апостериорна), въпреки че абсолютната разлика е идентична). Коефициентът на доверителност има следния общ вид:

$$1 - \frac{min(\text{Апостериорна доверителност,Априорна доверителност})}{max(\text{Апостериорна доверителност,Априорна доверителност})} > CR_L,$$

в който CR_L представлява долната граница на коефициента на доверителност CR. В случай че апостериорната доверителност е по-висока от априорната, правилото ще бъде избирано, когато:

$$\text{Апостериорна доверителност} > \frac{\text{Априорна доверителност}}{(1-CD_L)}.$$

В противен случай, ако апостериорната доверителност е по-ниска от априорната, правилото ще бъде избрано, когато:

$$\frac{\text{Апостериорна доверителност}}{(1-CD_L)} < \text{Априорна доверителност}.$$

Чрез коефициента на доверителност е възможно да се откриват порядко срещани и негативни ефекти и може да бъде по-прецизно настройван. По същество, ако априорната доверителност е ниска, то тогава е необходимо само малка промяна на апостериорната доверителност, за да бъде избрано съответното правило. Трябва да се отбележи обаче, че този коефициент има склонност към избор на правила с ниска априорна доверителност, както и към негативни правила с висока априорна доверителност.

Показателят за информационна разлика (ID) е относително по-сложен от първите два и се основава на информационния принос, изчисляван в битове информация и използван в алгоритъма C5.0. Основната му идея е, че без никаква допълнителна информация за другите елементи в множеството имаме определено разпределение на вероятностите (или по-точно на относителните честоти). След като получим информация, че елементите в антецедента на асоциативното правилото присъстват, имаме различно разпределение на вероятностите. Следователно имаме две апостериорни вероятностни разпределения. Въпросът сега е: колко информация получаваме, като наблюдаваме дали предшестващият елемент на правилото е налице или не. Информацията в случая се измерва като намаляване на ентропията. Следователно ентропиите на двете условни вероятностни разпределения се изчисляват и сумират, претеглени с вероятността за тяхното настъпване. Така се получава (очакваната стойност на) апостериорната или условната ентропия. Разликата на тази стойност с априорната ентропия е придобиването на информация от предшестващото правило или както т.нар. информационна разлика (Borgelt, 2017). Показателят за информационна разлика взема предвид подкрепата на дадено правило, така че при тази мярка, при едни и същи априорни и апостериорни доверителности, дадено правило ще бъде оценено по-високо, ако се отнася за по-голям брой случаи. По този начин алгоритъмът се стреми да елиминира порядко срещаните значими правила, които понякога се откриват от други алгоритми. Въпреки това, тъй като е трудно да се преобразува информационният принос от битове в процентни разлики, използването на тази мярка за оценка обикновено изисква експериментиране, за да се получи полезна настройка.

Като оценъчен показател за извеждането на интересни и значими асоциативни правила е възможно да се използва т.нар. нормиран хи-квадрат. Това на практика е добре познатият от хи-квадрат тест от статистиката за откриване на зависимости между номинални променливи, но нормализиран, за да се отстрани влиянието на броя на трансакциите. След нормализиране хи-квадрат стойността придобива стойности между 0 (липса на връзка) и 1 (перфектна връзка). Така долната граница за този показател се явява силата на асоциацията между анализираните продукти. Както и при информационната разлика спрямо априорната доверителност, нормализираният хи-квадрат критерий не е интуитивно свързан с разликите между априорната и апостериорната доверителност, така че и тук е необходимо да се експериментира с различни настройки. Като цяло този показател при равни други условия е сравнително по-чувствителен към подкрепата.

За да проследим ефекта от използването на четирите гореизложени алтернативни оценъчни показатели, ние репликирахме процедурата от Фигура 1 върху същите данни от продажбени трансакции shopping_data.csv, но с променена минимална доверителност от 75%. Целта ни беше да генерираме сравнително по-голям първоначален списък от асоциативни правила и да използваме този списък като отправна база за сравнение. Като резултат получихме 26 правила, съдържащи до четири антецедента, сортирани от висока до ниска степен на доверителност (>75%). Освен дефинираните основни показатели „честота“, „подкрепа“, доверителност“ и „интерес“, за всяко правило бе изчислен и индекс за „способността за разгръщане“ (DA). Този индекс показва какъв процент от данните отговарят на условията на антецедента, но не и на условията на консеквента и е изчислен по следния начин:

$$DA = \frac{(\text{Подкрепа на антецедента в } n - \text{Подкрепа на правилото в } n)}{n},$$

като с n бележим броя на трансакциите в масива (в случая, $n = 786$). По принцип, колкото са по-ниски стойностите на този индекс, толкова по-добре.

Отправният списък с извлечените 26 асоциативни правила е представен на Таблица 5. Видно е, че не са открити прости правила (т.е. такива, които съдържат само един антецедент. В преобладаващия брой случаи като консеквенти се оказват тестените изделия и готвените ястия.

Таблица 5

Разширен списък с правила, извлечени при ограничение от минимална доверителност от 75%

Консеквент	Антецедент(-и)	F	S (%)	C (%)	RS (%)	L	DA
Тестени изделия	Мляко и Замразени храни	85	10.81	83.53	9.03	1.95	1.78
Тестени изделия	Алкохол и Консерви и Готвени ястия	95	12.09	83.16	10.05	1.94	2.04
Тестени изделия	Замразени храни и Консерви и Шоколадови изделия	90	11.45	82.22	9.41	1.92	2.04
Готвени ястия	Алкохол и Консерви и Тестени изделия	97	12.34	81.44	10.05	1.65	2.29
Готвени ястия	Алкохол и Консерви и Шоколадови изделия	91	11.58	79.12	9.16	1.61	2.42
Тестени изделия	Мляко и Готвени ястия	105	13.36	79.05	10.56	1.84	2.80
Тестени изделия	Мляко и Консерви	100	12.72	79.00	10.05	1.84	2.67
Тестени изделия	Мляко и Алкохол	90	11.45	78.89	9.03	1.84	2.42
Тестени изделия	Замразени храни и Консерви и Готвени ястия	99	12.60	78.79	9.92	1.84	2.67
Готвени ястия	Мляко и Консерви и Тестени изделия	79	10.05	78.48	7.89	1.59	2.16
Тестени изделия	Алкохол и Замразени храни и Консерви	95	12.09	77.89	9.41	1.82	2.67
Готвени ястия	Мляко и Алкохол	90	11.45	77.78	8.91	1.58	2.54
Алкохол	Мляко и Замразени храни	85	10.81	77.65	8.40	1.97	2.42
Тестени изделия	Мляко и Шоколадови изделия	98	12.47	77.55	9.67	1.81	2.80
Замразени храни	Алкохол и Консерви и Тестени изделия и Готвени ястия	79	10.05	77.22	7.76	1.92	2.29
Шоколадови изделия	Алкохол и Консерви и Тестени изделия и Готвени ястия	79	10.05	77.22	7.76	1.63	2.29
Замразени храни	Алкохол и Консерви и Тестени изделия	97	12.34	76.29	9.41	1.90	2.93
Консерви	Алкохол и Тестени изделия и Шоколадови изделия и Готвени ястия	80	10.18	76.25	7.76	1.67	2.42
Алкохол	Замразени храни и Шоколадови изделия и Готвени ястия	100	12.72	76.00	9.67	1.93	3.05
Тестени изделия	Алкохол и Консерви и Шоколадови изделия	91	11.58	75.82	8.78	1.77	2.80
Шоколадови изделия	Алкохол и Консерви и Готвени ястия	95	12.09	75.79	9.16	1.60	2.93
Тестени изделия	Алкохол и Шоколадови изделия и Готвени ястия	106	13.49	75.47	10.18	1.76	3.31
Готвени ястия	Мляко и Тестени изделия	110	13.99	75.45	10.56	1.53	3.44
Консерви	Мляко и Замразени храни	85	10.81	75.29	8.14	1.65	2.67
Готвени ястия	Мляко и Замразени храни	85	10.81	75.29	8.14	1.53	2.67
Алкохол	Замразени храни и Тестени изделия и Готвени ястия	109	13.87	75.23	10.43	1.91	3.44

Легенда: **F** = честота; **S** = подкрепа на антецедента; **C** = доверителност; **RS** = подкрепа на правилото; **L** = интерес; **DA** = способност за разгръщане.

Проведеният експеримент започва с оценяване на ефекта от използване на абсолютната доверителна разлика като критерий за извличане на списък с асоциативни правила. Тъй като този оценъчен показател се основава на априорната доверителност, необходимо е да добием някаква представа за относителната честота на различните продукти в наблюдаваните трансакции. Изчислената априорна доверителност на наблюдаваните продуктови категории е представена на Таблица 6. Разглеждането на тези стойности помага да се изясни защо асоциативните правила, открити в първия модел, не съдържат никакви тоалетни принадлежности, пресни зеленчуци или прясно месо. Техните стойности на подкрепа са твърде ниски, за да бъдат включени. Този пример показва защо е необходимо да се изследват честотите в основните данни, за да се разбере напълно даден набор от асоциативни правила. Например можем да намалим минималната подкрепа на правилата, за да се опитаме да включим и тези продукти в правилата, ако конкретният маркетингов контекст го изисква.

Таблица 6

Априорна доверителност на наблюдаваните продуктови категории

Продуктова категория	Априорна доверителност (%)	Брой трансакции, в които е регистрирана покупка
Готвени ястия (топла витрина)	49.24	387
Шоколадови изделия	47.46	373
Консерви	45.55	358
Тестени изделия	42.88	337
Замразени храни	40.20	316
Алкохол	39.44	310
Мляко	18.83	148
Тоалетни принадлежности	9.92	78
Плодове и зеленчуци	8.27	65
Охладено месо	2.93	23

След като разполагаме с информация за априорната доверителност, вече е възможно да се приложи априорният алгоритъм на базата на оценъчния показател „абсолютна доверителна разлика“. За ограничително условие използваме само долен ограничителен праг за абсолютната разлика $CD_L = 50$ и елиминираме ограниченията за минимална подкрепа на антецедента и за минимална доверителност.

Ключовото решение в този анализ е стойността на долната граница на абсолютната доверителна разлика. Припомняме, че този оценъчен показател се отнася до абсолютната стойност на разликата между априорната и апосте-

риорната доверителност. Позовавайки се на данните от Таблица 6 с априорните стойности на достоверност, имаме като че ли две отделни групи продукти – едната, съдържаща продукти с априорна доверителност между около 40% и 50%, и друга с под 20%. Тъй като най-високата доверителност е малко над 83% (вж. първите редове на Таблица 5), можем да допуснем, че долната граница на показателя „абсолютна доверителна разлика“ от 50 вероятно няма да намери по същество никакви правила за продуктите, които се купуват по-често. Може да се намерят правила за другата група продукти (по-често купуваните), но е по-вероятно да се намерят отрицателни правила.

Абстрахирайки се от тези опасения, стартирането на модела с ограничително условие $CD_L = 50$ води до огромен списък от 774 правила. Първите три от тях, сортирани в низходящ ред на базата на доверителността, са представени на Таблица 7.

Таблица 7

Извлечение на първите три реда от списъка с генерирани асоциативни правила при ограничение $CD_L = 50$

Консеквент	Антецедент(-и)	Rule ID	F	S (%)	C (%)	RS (%)	CD	L	DA
Замразени храни	Охладено месо и Тоалетни принадлежности	2	5	0.64	100.00	0.64	59.80	2.49	0.00
Шоколадови изделия	Охладено месо и Тоалетни принадлежности	3	5	0.64	100.00	0.64	52.54	2.11	0.00
Готвени ястия	Охладено месо и Консерви	6	19	2.42	100.00	2.42	50.76	2.03	0.00
…	…	…	…	…	…	…	…	…	…

Легенда: **F** = честота; **S** = подкрепа на антецедента; **C** = доверителност; **RS** = подкрепа на правилото; **L** = интерес; **DA** = способност за разгръщане.

Първото правило е {Охладено месо, Тоалетни принадлежности} → {Замразени храни}.

Пет клиента са купили двата антецедента и от тези 5, 100% за купили замразени храни (апостериорната доверителност е 100). Априорната доверителност на замразените храни е 40,2 (вж. Таблица 6), така че абсолютната разлика между двете (100 - 40,2 = 59,8) е по-голяма от 50, долната граница на показателя за абсолютната доверителна разлика. Колоната "CD" показва тази разлика и всички стойности в нея ще бъдат по-големи от 50. Разбира се, 5-те купувачи представляват само около 0,6% от файла и това правило може да се определи като рядко срещано. Но от маркетингова

гледна точка и в конкретен контекст то би могло да представлява интерес, тъй като стойността ни показателя „интерес“ (L) е 2,49 (т.е. отношението на апостериорната доверителна вероятност към априорната доверителна вероятност е 100/40,2 или 2,487), което означава, че шансът да се купят замразени храни е почти два пъти и половина по-висок, ако в пазарната кошница на клиента има охладено месо и тоалетни принадлежности.

Ако сравним данните от Таблица 5 и Таблица 7, няма да открием съответствие, тъй като във втората са включени и онези правила, които при първоначалния сценарий са били изключени. Първият списък от правила, основаващ се на доверителността, би следвало като цяло да е по-полезен поради високата подкрепа. Преди да се направят категорични изводи обаче, е препоръчително да се проверяват и други възможни сценарии, основаващи се на други оценъчни показатели.

На Таблица 8 са представени последните три извлечени с априорния алгоритъм правила, но при използване на коефициента на доверителност $CR_L = 50$ като ограничително условие. При този сценарий се идентифицират огромен брой правила (2222), но някои са с различен характер в сравнение с предходния сценарий. Например правилото за охладено месо и замразени храни (Rule ID = 8) има подкрепа от 40,2 % (което е доста висок процент), но доверителността му е само около 6 % (тези, които купуват замразени храни, купуват и охладено месо само при около 6 % от всички актове на покупка). Това „слабо“ правило е включено в списъка, тъй като априорната му достоверност е 2,93% за охладено месо (вж. Таблица 6), но коефициентът на доверителност е около 51,33 (1 - 0.029 / .06)*100) (вж. колоната с етикет CR), което е повече от зададената като ограничителен праг долна граница на показателя. Това е илюстрация на факта как, ползвайки коефициента на доверителност, може да се открият интересни правила с ниски априорна доверителност. Това правило има висока стойност на подкрепа, тъй като много купувачи купуват замразени храни. То изолира асоциация, при която вероятността за купуване на охладено месо се удвоява спрямо базовата стойност и тази закономерност може да бъде полезна за управленските решения в търговския обект.

С другите два оценъчни показателя „информационна разлика“ и „нормализиран хи-квадрат“ също биха могли да се съставят подобни аналитични сценарии. Подчертаваме обаче, че самата интерпретация на правилата и решението, дали те са важни и интересни за маркетингови прозрения, следва да се реши само след съобразяване с тяхната априорна достоверност. Възможностите за експлоративен анализ тук са на практика неограничени и включват дори и сценарии, при които целта да бъде установяването на негативни правила (IBM, 2010, pp. 4-13).

Таблица 8

Извлечение на последните три реда от списъка с генерирани асоциативни правила при ограничение $CR_L = 50$

Консек-вент	Антецедент(-и)	Rule ID	F	S (%)	C (%)	RS (%)	L	CR	DA
…	…	…	…	…	…	…	…	…	…
Охладено месо	Замразени храни	8	316	40.20	6.01	2.42	2.05	51.33	37.79
Охладено месо	Плодове и зеленчуци и Мляко и Алкохол и Консерви и Тестени изделия	1522	17	2.16	5.88	0.13	2.01	50.25	2.04
Охладено месо	Тоалетни принадлежности и Алкохол и Консерви и Тестени изделия и Шоколадови изделия	1767	17	2.16	5.88	0.13	2.01	50.25	2.04

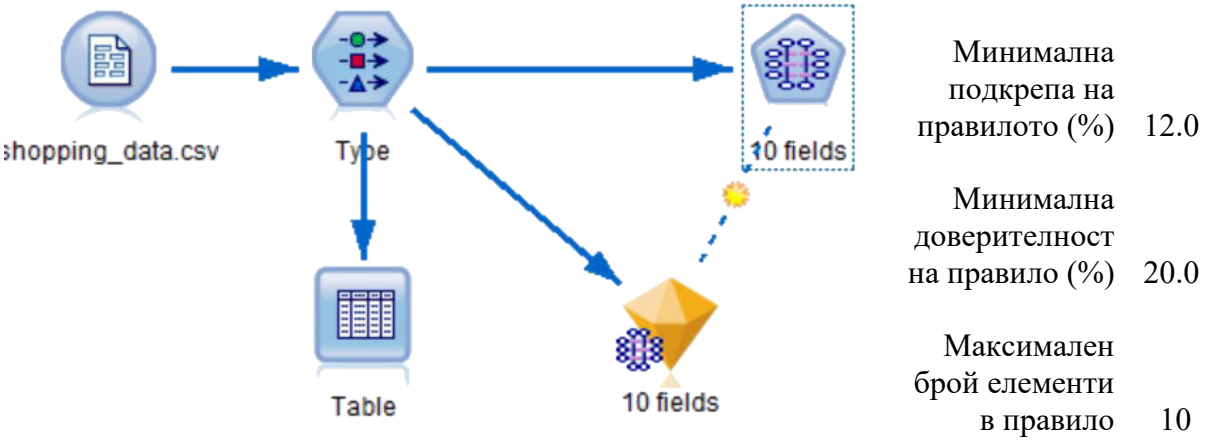
Легенда: **F** = честота; **S** = подкрепа на антецедента; **C** = доверителност; **RS** = подкрепа на правилото; **L** = интерес; **DA** = способност за разгръщане.

Дотук за извличането на асоциативни правила с цел анализ на пазарната кошница бе използван основният и най-популярен априорен алгоритъм, който работи само с номинално скалирани променливи, кодирани дихотомно. За да се намали допълнително времето за изчисление, алгоритъмът извършва интелигентно индексиране и минимализира преходите през целия набор от данни, за да генерира пестелив списък от асоциативни правила[7]. Алтернатива на априорния алгоритъм се явява непрекъснатият алгоритъм за извличане на асоциативни правила, известен с абревиатурата CARMA. Както бе вече споменато, CARMA осигурява непрекъсната обратна връзка, докато списъкът с покупки се сканира. За всяко извлечено правило се поддържат намаляващи, детерминистични интервали за неговата подкрепа и достоверност. Алгоритъмът сканира масива с продажбени трансакции два пъти. По време на първото сканиране потребителят може да променя праговете на подкрепа и доверителност "в ход". Така в комбинация с непрекъснатата обратна връзка потребителят може да определи "правилните" прагове интерактивно. CARMA гарантира, че създава всички правила за асоцииране след най-много 2 сканирания и че за всяко правило има точна стойност на подкрепата и достоверност (Hidber, 1999). По отношение на

[7] За повече информация и подробно описание на изчислителните процедури вж. Агравал и колектив (Agrawal, Mannila, Srikant, Toivonen, & Verkamo, 1996), както и Боргелт (Apriori: Find Frequent Item Sets and Association Rules with the Apriori Algorithm, 2017)

производителността CARMA превъзхожда традиционния априорен алгоритъм при задаване на ниски прагове на подкрепа и е по-ефективен по отношение на изчислителните процедури при наличието на трудно идентифицируеми правила. Всичко това дава по-добри възможности за контрол от страна на потребителя.

В настоящото изследване бяха тествани и възможностите на CARMA върху същия набор от трансакционни данни за продажбите shopping_data.csv. Работната процедура, разработена с IBM SPSS Modeler, следва следния логически модел (вж. Фигура 4).



Фигура 4. Работна процедура за извличане на силни асоциативни правила с CARMA

По подобие на априорния алгоритъм, CARMA също така използва трите параметъра „подкрепа“, „доверителност“ и максимален размер на елементите (общ брой на антецедентите и консеквентите), за да контролира създаването на правила. Различното обаче е, че минималната подкрепа се отнася както за антецедентите, така и за консеквентите. В този илюстративен анализ използваме ограничителни стойности, съответно 12%, 20% и 10, и идентифицирахме 146 правила (вж. Таблица 9). Обикновено при експлоративния анализ се започва с отправни стойности от 20%, 20% и 10 и се експериментира най-напред с намаляване на поддръжката, което неминуемо води до експоненциално нарастване на броя на извлечените асоциативни правила.

В сравнение с априорния алгоритъм (вж. Таблица 2) CARMA конструира доста по-различни асоциативни правила. Това се дължи донякъде и на обстоятелството, че като критерий се задава минимален праг на подкрепа на цялото правило, а не само на антецедента. Повечето правила при CARMA са опростени и съдържат само по един антецедент. Причината за това е доста строгото изискване, минималната подкрепа на правилата да бъде поне 12%, което означава, че дадено правило трябва да се прилага за поне една дванадесета от файла. В резултат на това най-високата доверителна стойност е 74,32, а подкрепата на правилото е само 13,99. Ако погледнем обаче долната

част на списъка от Таблица 9, ще открием правила, съдържащи повече от един консеквент (напр. правило с ID 146 има вид {Готвени ястия} → {Алкохол, Консерви}), което е невъзможно за идентифициране с класическия априорен алгоритъм.

Таблица 9

Извлечение на първите и последните шест правила от списъка със 146 с CARMA

Консеквент(-и)	Антецедент(-и)	Rule ID	F	S (%)	C (%)	RS (%)	L	DA
Тестени изделия	Мляко	1	148	18.83	74.32	13.99	1.73	4.83
Тестени изделия	Замразени храни и Консерви	2	163	20.74	71.78	14.89	1.67	5.85
Тестени изделия	Алкохол и Консерви	3	136	17.30	71.32	12.34	1.66	4.96
Готвени ястия	Мляко	4	148	18.83	70.95	13.36	1.44	5.47
Готвени ястия	Алкохол и Консерви	5	136	17.30	69.85	12.09	1.42	5.22
Замразени храни	Алкохол и Консерви	6	136	17.30	69.85	12.09	1.74	5.22
…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…
Готвени ястия и Алкохол	Консерви	141	358	45.55	26.54	12.09	1.25	33.46
Замразени храни и Алкохол	Консерви	142	358	45.55	26.54	12.09	1.15	33.46
Мляко	Шоколадови изделия	143	373	47.46	26.27	12.47	1.40	34.99
Замразени храни и Шоколадови изделия	Готвени ястия	144	387	49.24	25.84	12.72	1.21	36.51
Замразени храни и Консерви	Готвени ястия	145	387	49.24	25.58	12.60	1.23	36.64
Алкохол и Консерви	Готвени ястия	146	387	49.24	24.55	12.09	1.42	37.15

Легенда: **F** = честота; **S** = подкрепа на антецедента; **C** = доверителност; **RS** = подкрепа на правилото; **L** = интерес; **DA** = способност за разгръщане.

Логично е обаче да се стигне до естествения въпрос, кой от двата модела за извличане на асоциативни правила да бъде използван с приоритет и за кои ситуации единият е по-подходящ от другия. Като се има предвид колко много правила са възможни в един файл с данни с много променливи и хиляди записи и като се има предвид, че всеки модел прилага различен подход за пресяване на възможните кандидати за правила, тези два метода не гарантират, че ще намерят абсолютно еднакви правила в даден набор от продажбени трансакции. Донякъде априорният алгоритъм предлага по-голям избор от оценъчни критерии, чрез които да се контролира процесът на селекция на силни и интересни правила. Ако желаете да имате избор на метод за избор на правила, очевидно априорният модел ще бъде предпочетен, защото дава на потребителя по-голям контрол върху метода за

генериране на правила. CARMA алгоритъмът обаче е по-гъвкав и понякога по-бърз при работа с огромни масиви от данни. Въпреки че CARMA може да бъде по-ефективна от априорния подход, ако наборите от данни не са много големи, скоростта обикновено не е основен проблем при генерирането на набори от асоциативни правила и вероятно не е определящ фактор. Ако предпочитанията на анализатора е да се съсредоточи върху това, колко универсално е дадено правило, включващо както антецедентите, така и консеквенти, CARMA позволява директен контрол на тази настройка. Ако вместо това желанието му е да се съсредоточи върху честотата на антецедентите, препоръката е да се използва априори алгоритъмът. И накрая, ако стремежът е да се фокусира само върху определени консеквенти или антецеденти, априорният модел може да осигури по-голям контрол, тъй като се съобразява с ролята на променливите.

Освен тези общи насоки, за да се реши как да се определи подкрепата и доверителността, трябва да вземете предвид следните по-конкретни изследователски въпроси на анализа (IBM, 2010, pp. 4-20):

- Дали целта е да се идентифицират рядко срещани комбинации;
- Дали интересът е към извличане само на правила с висока степен на доверителност и/или подкрепа;
- Дали интерес представляват правила, съдържащи само един или повече броя антецеденти;
- Дали стремежът е да се отрият правила с голяма абсолютна или относителна разлика между априорната и апостериорната доверителност.

Въз основа на отговорите на тези въпроси може да се избира метод и някои първоначални настройки за различните параметри. Но неизменно ще трябва да се експериментира и да се променят настройките, за да се намери оптимален набор от правила. И вероятно правилната изследователска стратегия е да се използват и двата модела и да се сравняват констатираните резултати.

IV. Извличането на секвентни правила от продажбени трансакции

Приведените подходи и примери за извличане на асоциативни закономерности предполагат, че едни и същи елементи (асортиментни позиции) могат да се появят както в лявата, така и в дясната страна на правилото, т.е. това са един вид „ненасочени“ техники за аналитично извличане на знания от данни. Асоциативните правила обаче могат да се използват и по директен начин, като лявата и дясната страна съдържат различни типове елементи. Например в контекст на една имейл маркетингова кампания, провеждана по поръчка от търговска банка, получателят на непоисканото търговско съобщение с предложение за потребителски кредит има три възможни

алтернативи за реакция: да не предприема нищо (което е най-срещаното действие), да кликне върху имейла (с което показва интерес към търговското предложение) и да подаде жалба срещу полученото съобщение, декларирайки го като нежелан спам (Berry & Linoff, 2011, p. 569). Първото от тези действия не носи никаква полза и по същество не струва нищо (защото изпращането на електронна поща е пренебрежително евтино). Второто генерира приходи за компанията, така че е доста важно. Третото действие е разход и то голям. Ако твърде много клиенти се оплакват от даден доставчик на интернет услуги, тогава той може да отхвърли всички предложения за електронна поща от компанията. В този пример банката натрупва исторически данни, съдържащи многобройни примери за клиенти, които щракват върху електронната поща, и много обикновено по-малко случаи, при които клиентите се оплакват. Обикновено още при първото оплакване на клиент неговият имейл се премахва от всички списъци с имейли. Кликовете и оплакванията по същество са различни събития, въпреки че и двете действия са провокирани от едно и също съдържание (в случая оферта). Традиционните асоциативни правила не могат да бъдат адекватен инструмент в тази ситуация. Вместо тях целта на анализа следва да бъде фокусирана не върху едновременното възникване на събитията, а върху причинно-следствената им връзка и по-специално да се намери отговор на въпроса „кои видове оферти водят до оплаквания, когато клиентите вече са кликнали върху други оферти?“. Асоциативното правило би трябвало да има следния общ вид:

$$\{[A\ ]\text{:клик } + [B\ ]\text{: клик}\} \rightarrow \{[\boldsymbol{C}]\text{: оплакване от спам}\}$$

В случая очевидно става дума за предсказване на верижно събитие, което е краен елемент на логическа секвенция. В следващите редове ще бъде представена концептуалната рамка за идентифициране на секвентни модели на поведение при покупка в среда на големи масиви с данни от продажбени трансакции.

Анализът на секвентните модели в процеса на пазаруване добавя елемента време към анализа на пазарната кошница. Анализът разглежда не само асоциациите между елементите, но и поредиците от елементи, при които подреждането във времето е важно. При него основен обект на наблюдение са действия, които се извършват в определен ред във времето. Идентифицирането на секвентни закономерности в поведението на клиентите предполага идентифициране на субекта, извършващ всеки конкретен акт на покупка. По същество секвентният анализ на процесите на пазаруване няма смисъл при анонимни трансакции, тъй като те не могат да бъдат свързани помежду си във времето (Berry & Linoff, 2011, p. 574).

Моделирането на секвенции добива широка популярност при извличането на информация от уебстраници и анализа на потока от кликвания

върху уебстраници; то започва като начин за анализ на данни от уебрегистри, за да се разберат моделите на навигация в уеб и да се идентифицират маршрутите на сърфиране, които водят до определени страници, например страницата, в която се извършва плащането на продукта. Употребата на тези алгоритми е разширена и в днешно време те могат да се прилагат за всички бизнес проблеми, при които е интересна „последователността“. Могат да се използват като средство за прогнозиране на следващите очакваното „движение“ на клиентите или следващата фаза в жизнения цикъл на клиента. В банковото дело е възможно да се прилагат за идентифициране на поредица от събития или взаимодействия с клиенти, които могат да бъдат свързани с преустановяване използването на даден продукт; в телекомуникациите – за идентифициране на типичните пътища на покупка, които са силно свързани с покупката на определена допълнителна услуга; в производството и контрол на качеството, за да се открият признаци в производствения процес, които водят до дефектни продукти.

За маркетинговите специалисти е важно да разберат съществуват ли закономерности в последователността на действията на обслужваните клиенти. Ако клиентите посещават често една уебстраница, посещават ли след това друга? Ако купуват един продукт, дали по-късно винаги или сравнително често купуват друг? Ако имат някакъв конкретен опит с продукта, как това променя последващото им продуктово или пазарно поведение? Съществува голям набор от аналитични и статистически методи за решаване на такива въпроси, вариращи от времеви редове до модели на Марков, от каузално моделиране до динамично клъстеризиране (Chapman & Feit, 2019, p. 399). Известни са и множество научни изследвания, предлагащи алгоритмични решения на подобни проблеми (Eihinger, Nauck, & Klawonn, 2006; Cabanes, Bennani, & Dufau-Joël, 2009; Gupta & Han, 2012; Mukherji, Srinivasan, & Welbourne, 2014; Bermingham & Lee, 2014; Tsai & Lai, 2015; Goel & Mallick, 2015). Сред тях особен интерес предизвиква предлаганият от Гуидоти и колектив (Guidotti, Rossetti, Pappalardo, Giannotti, & Pedreschi, 2019) подход за персонализирано прогнозиране на пазарната кошница с времево анотирани повтарящи се секвенции. Авторите разработват алгоритъм за проектиране и внедряване на ефективна препоръчваща система, състояща се от адекватно извлечени асоциативни правила от големи масиви с данни от продажбени трансакции. Идеята е, тези правила да се използват автоматизирано за напомняне и съставяне на персонализирани списъци за пазаруване, предлагайки артикули, от които клиентите вероятно имат нужда.

Принципите на секвентните алгоритми за извличане на асоциативни правила се базират на CARMA-алгоритъма (Agrawal & Srikant, 1995). При него е важно да се дефинира понятието „секвенция“. Под секвенция в контекста на анализа на пазарната кошница се обозначава регистрирана във

времето поредица от актове (събития) на покупка на продуктови артикули и/или услуги от страна на отделните клиенти.

Алгоритъмът за секвениране анализира последователностите от събития, за да се открият общи и често срещани секвенции. Той се използва за идентифициране на асоциации на събития/покупки/атрибути във времето. Отчита се редът на събитията и се откриват последователни асоциации, които водят до конкретни изводи.

Този алгоритъм се изпълнява в две стъпки:

(1) Идентифицират се често срещани секвенции на актове на покупка, съдържащи еднотипни елементи (артикули) и се съхраняват във възлите на геометрична решетка;

(2) След това на базата на специфични критерии (като напр. минимална „подкрепа“, минимална „доверителност“ или поставяне на ограничение върху секвенцията на антецедента) се извличат силни и значими секвентни правила.

Секвентният алгоритъм генерира правила, аналогични на априорния алгоритъм, но с една разлика – при него се откриват силни асоциации между секвенции от предшестващи събития и появата на последващо събитие. С други думи, когато някакви събития се случват в определен ред, то е възможно да е налице повишена вероятност за възникване на определено последващо събитие (Chorianopoulos, 2016, p. 14). Това би могло да се изрази като логическа операция:

АКО {АНТЕЦЕДЕНТИ, подредени в хронологичен ред), ТО {КОНСЕКВЕНТ}

Ако разгледаме подобно правило в контекст на клиенти на една хипотетична банка, то би могло да гласи, че тези клиенти на банката, които започват взаимоотношенията си с нея с откриване на спестовна сметка и впоследствие придобиват кредитна карта и краткосрочен депозит, имат повишена вероятност да инвестират в акции.

Основната структура на данните за анализ на пазарната кошница, представена по-рано на Таблица 1, е подходяща за извличане и на секвентни правила, стига записът на поръчката (или редът) да съдържа времеви индекс или някакви маркери, индикиращи последователност. На Фигура 5 е представен илюстративен логически модел на данни от продажбени трансакции, позволяващ извличане на често срещани секвенции в актовете на покупка. С етикета SID е отбелязан идентификатор на отделния клиент. Използвани са буквени идентификатори за разграничаване на отделните асортиментни позиции (купувани артикули). Под „времеви маркер“ се разбира ординална променлива, отразяваща последователните моменти във времето, в които са регистрирани отделните продажбени трансакции. Например клиент със SID 1 има регистрирани четири продажбени трансакции. Първата съдържа артикули, C и D, следващата A, B и F и т.н. Целта на

анализа е да се идентифицират тези секвенции, които се срещат най-често и следователно отразяват някаква закономерност в последователността на купуване на конкретни артикули. Въведените вече показатели „подкрепа“, „доверителност“ и „интерес“ са приложими и при извличането на устойчиви и силни правила за описание на секвенциите.

Данни от продажбени трансакции

SID	**Времеви маркер на трансакцията (брой периоди от последната покупка)**	**Артикул**
1	**10**	**C D**
1	**15**	**A B C**
1	**20**	**A B F**
1	**25**	**A C D F**
2	**15**	**A B F**
2	**20**	**E**
3	**10**	**A B F**
4	**10**	**D G H**
4	**20**	**B F**
4	**25**	**A G H**

Често срещани секвенции…

Чести секвенции от 1 елемент

A	4
B	4
D	2
F	4

Чести секвенции от 2 елемента

AB	3
AF	3
B→A	2
BF	4
D→A	2
D→B	2
D→F	2
F→A	2

Чести секвенции от 3 елемента

ABF	3
BF→A	2
D→BF	2
D→B→A	2
D→F→A	2

Чести секвенции от 4 елемента

D→BF→A	2

Фигура 5. Логически модел на данни от продажбени трансакции, позволяващи извличане на често срещани секвенции в актовете на покупка
Източник: (Zaki, 2001, p. 33)

В някои случаи дори и поведението при купуване само на един елемент (конкретен артикул, продукт или услуга) може да бъде интересно за наблюдение в хода на времето. Например много фармацевтични продукти

са предназначени за хронични състояния и трябва да се приемат последователно във времето. Съществува обаче риск за предложителя, конкретният продукт за бъде заменен с конкурентен продукт, респ. бранд. Основателният маркетинговият въпрос, който може да бъде зададен в този случай, е: какъв модел на последователно във времето купуване на съответния артикул демонстрира отделният клиент. По-интересни за изучаване и предсказване обаче са сценариите, при които се наблюдава поведение на „превключване“ между различни елементи. Разбира се, наличието на повече алтернативи за купуване означава и огромен брой теоретично възможни секвенции. Оттук следва и целта на анализа – идентифициране и извличане на най-често срещаните от тях и (евентуалното последващо) типологизиране на клиентите на базата на тази информация. По-конкретно, с помощта на идентифицирани закономерности в последователността на купуване на разнородни артикули е възможно да се предскажат както обектът на следващата покупка, делът от клиенти, които е вероятно да преминат към друг продукт или бранд при следващото си решение за покупка, както и евентуалното прекъсване на секвенцията от последователни актове на покупка.

С увеличаването на броя на елементите (т.е. асортиментните позиции) броят на потенциалните секвенции се увеличава драстично и за тяхното идентифициране и анализиране са необходими сравнително сложни числови алгоритми и софтуерни програми.

Дотук бе представена накратко логиката на алгоритмите за секвениране, предназначени за анализиране на последователности на събитията с цел откриване на общи и често срещани секвенции. Тези алгоритми могат да се използват за идентифициране на асоциации на събития/покупки/атрибути във времето. Те отчитат реда на събитията и откриват последователни асоциации, които водят до конкретни резултати. А как това би могло да се реализира в реална среда, е обект на изложението в следващия раздел.

V. Процедури за извличането на секвентни правила от продажбени трансакции

Основните идеи, заложени в алгоритмите за извличане на асоциативни правила и разгледани в предходните раздели, е възможно да бъдат използвани и разширени за целите на анализа на секвентни поведенчески модели. За да се обработват секвентни данни, данните за продажбените трансакции трябва да имат две допълнителни характеристики:

- Времеви маркер (или информация за поредността на наблюдението), за да се определи кога са се случили трансакциите една спрямо друга (вж. Фигура 5);

- Идентифицираща информация като например номер на банкова сметка, идентификационен номер на домакинство или идентификационен номер на клиент, която идентифицира различните трансакции като принадлежащи на един и същ клиент или домакинство (вж. Фигура 5, SID).

Извличането на секвенционни правила е подобно на процеса на извличане на прости ненасочени асоциативни правила, а именно:

(1) Всички артикули, закупени от клиент, се разглеждат като една поръчка и всеки артикул запазва времевия маркер, указващ позицията му в съответната секвенция;

(2) Процесът е същият и за набор от артикули, които се купуват едновременно. Всеки набор от едновременно купувани артикули би трябвало да съдържа времеви маркер (или друг тип информация за поредност), както и отделните артикули;

(3) Извеждат се само тези правила, при които елементите от лявата страна (т.нар. антецеденти) са се появили преди елементите от дясната страна (т.нар. консеквенти).

В резултат на следването на тази „проста“ логика се съставят списъци с асоциативни правила, които могат да разкрият секвенционни модели на поведение при купуване.

Секвентният анализ на данни от продажбени трансакции се родее с някои алтернативни техники за аналитично извличане на знания от данни, като например анализът на „оцеляването“ (от англ. “survival analysis”). Двете техники имат много сходства. Анализът на секвентните модели също както и анализът на оцеляването изучават последователности в поведението на клиентите. Основната разлика е в определението за време. При анализа на оцеляването времето се измерва в общи единици, като дни, седмици и месеци. При секвенционния анализ подредбата е по-важна от конкретната продължителност. Въпреки че анализът на оцеляването е по-силно фокусиран върху фактора време, конкуриращите се рискове в комбинация с повтарящи се събития са вид анализ на секвентни модели. От гледна точка на статистиката анализът на секвентни модели е пример за верижен анализ (от англ. „path analysis“). Верижният анализ обаче се използва типично в по-специфичния контекст на анализа на връзките за анализ на уебсайтове (Berry & Linoff, 2011, p. 579).

Най-популярните алгоритми[8] за аналитично извличане на секвентни закономерности от големи масиви от данни е разработеният от Агравал и Срикант (Agrawal & Srikant, Mining Sequential Patterns, 1995) „двуетапен процес за извличане на секвентни модели“, реализуем с IBM SPSS Modeler (IBM Corp., 2020, pp. 337-346). Впоследствие този алгоритъм е разширен от Заки (Zaki, 2001) и става популярен с абревиатурата SPADE (от анг.

[8] Разширена таксономия на алгоритмите за извличане на секвентни правила от големи масиви от данни предлагат Моброке и Езейф (Mabroukeh & Ezeife, 2010, p. 3:17).

Sequential **PA**ttern **D**iscovery using **E**quivalent **C**lass). Алгоритъмът на SPADE, е реализуем със софтуер с отворен код в R пакета arulesSequences (Buchta & Hahsler, 2020). Представяната по-долу процедура следва първоначалния алгоритъм на Агравал. С цел осигуряване на прозрачност, проследимост и възпроизводимост използваме публично достъпни реални анонимизирани данни за поведението на клиентите на голяма търговска банка[9].

Таблица 10

Примерна структура на анонимизирани данни за ползваните банкови услуги в трансакционен формат (извлечение на първите редове от пълния набор от данни за 32367 клиенти, достъпен на https://bit.ly/3jTOB4t)

(ACCT)	(SERVICE)	(VISIT)	(DATE)
500026	CKING	1	1.2.2018
500026	SVG	2	15.2.2020
500026	ATM	3	6.4.2021
500026	ATM	4	31.5.2021
500075	CKING	1	15.12.2015
500075	MMDA	2	21.1.2016
500075	SVG	3	28.1.2016
500075	ATM	4	13.3.2020
500075	TRUST	5	30.4.2020
500075	TRUST	6	15.6.2020
500129	CKING	1	26.9.2020
500129	SVG	2	30.11.2020
500129	IRA	3	4.1.2021
500129	ATM	4	7.2.2021
500129	ATM	5	4.5.2021
500256	CKING	1	20.8.2020
500256	SVG	2	15.9.2020
500256	CKCRD	3	26.9.2020
500256	CKCRD	4	12.12.2020
500341	CKING	1	19.2.2019
…	…	…	…

Легенда	**Етикет**
Номер на клиента	(ACCT)
Поредност на ползване на услугата във времето	(VISIT)
Дата на ползване на услугата	(DATE)
Наименование на продукта/ услугата	(SERVICE)
Дебитна карта	ATM
Кредит за покупка на автомобил	AUTO
Кредитна карта	CCRD
Срочен депозит	CD
Пътнически чекове	CKCRD
Разплащателна сметка	CKING
Стоков кредит	HMEQLC
Индивидуална пенсионна сметка	IRA
Електронно банкиране	MMDA
Ипотечен кредит	MTG
Потребителски кредит	PLOAN
Спестовна сметка	SVG
Доверителна сметка със специално предназначение	TRUST

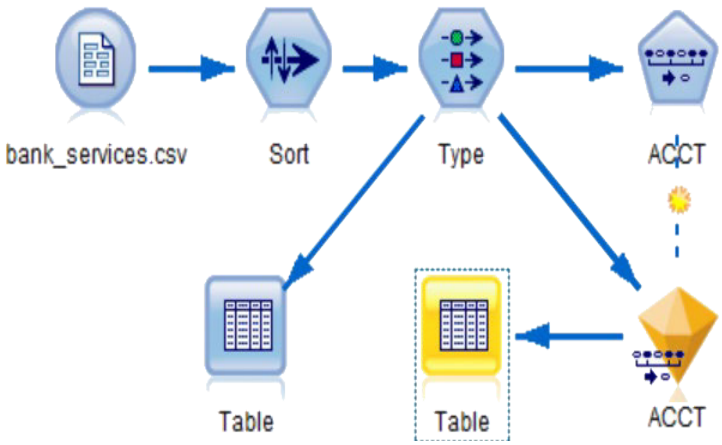
На Таблица 10 е представено извлечение от първите 20 реда от анонимизирани данни за поведението на потребители на банкови услуги в

[9] Данните са достъпви в csv-формат на https://bit.ly/3jTOB4t.

хипотетична търговска банка. Пълният набор от данни съдържа 32376 записа на регистрирано ползване на конкретната услуга по вид и по дата, от страна на 7991 клиенти на банката. Данните са подложени на предварителна обработка и почистване и включват четири променливи: номерът на клиента (ACCT), описание на ползваната услуга (SERVICE), датата на която е ползвана съответната услуга (DATE), и последователността, в която са ползвани услугите от страна на всеки клиент (VISIT)10. Данните са предварително сортирани по номера на клиента и след това по последователността, в която е ползвал във времето съответна услуга. Например първите четири реда съдържат информация за поведението на клиент с номер 500026, който на четири последователни дати е ползвал първо разплащателна сметка, след това спестовна сметка, след това е ползвал два последователни пъти дебитни карти. За този клиент в базата данни има четири поредни регистрирани операции, за следващия клиент с номер 500075 са регистрирани шест последователни операции, за последващия 500129 съответно 5 операции и т.н. Въпреки че значенията на променливата ACCT са в числов вид, на практика тази променлива се третира като номинална. Номинална е и променливата SERVICE, тъй като съдържа тринадесет категории, съответстващи на услугите, предлагани от банката (вж. легендата на Таблица 10).

Работната процедура за извличане на често повтарящи се секвенции от описания масив е илюстрирана на Фигура 6. Освен познатите минимални прагове за подкрепа, доверителност и минимален брой елементи в правило, процедурата позволява и контрол на броя на най-добрите правила за предсказване, които ще бъдат използвани за генериране на прогнозни полета. В примера прогнозите са съставени за трите най-добри секвентни правила. Освен тези ограничителни условия, при по-задълбочен анализ е възможно да се контролира и максимално допустимата продължителност между две поредни събития в една секвенция, ако вместо променливата VISIT се използва информацията за действителното време между две операции от променливата DATE.

[10] За ориентация в програмния R-код, който би могъл да бъде използван за извличане на поредността на ползваните услуги от всеки клиент на базата на променливата DATE вж. Koenecke (2019).



Минимална подкрепа на правилото (%)	15.0
Минимална доверителност на правило (%)	20.0
Максимален брой елементи в секвенция	10
Брой прогнози, които да се добавят в потока	3

Фигура 6. Работна процедура за извличане на секвентни правила

Резултат от изпълнението на процедурата при описания по-горе сценарий е представен на Таблица 11. Видно е, че при така зададените ограничения са извлечени само осем секвентни правила. Правилата са подредени в низходящ ред на базата на оценката на доверителността. Нека разгледаме първото правило, което, описано вербално, би звучало така: Ако в предходен момент е ползвана разплащателна сметка {CHING}, то много е вероятно в следващ момент да бъде ползвана и спестовна сметка {SVG}. Това правило е вярно за 4329 клиента (вж. „Честота“). Тези 4329 клиенти представляват 54,17% от общия брой клиенти на банката (вж. „Подкрепа на правилото“). 85,78% от клиентите (вж. показателя „Подкрепа“) имат секвенция с антецедент CKING (ползвана разплащателна сметка) и консеквент SVG (ползвана спестовна сметка). Подкрепата показва дела на клиентите, за които предшестващите елементи (антецедентите) са верни. За разлика от асоциативните модели тук подкрепата не се основава на броя на случаите, а на броя на идентификаторите в променливата ACCT. С други думи 85,7% от всичките 7991 клиенти са ползвали разплащателни сметки, т.е. 6848. От тези 6848 клиенти 63,15% ползват в непосредствената последваща банкова операция спестовна сметка (SVG), което закръглено е 4324 клиента (вж. показателя „доверителност“). Това по същество представлява най-често срещаната секвенция, описваща типичния „път“ на клиента.

Има обаче и по-сложни правила, при които секвенцията може да съдържа повече от една предходни операции. Ако обследваме например четвъртото правило, то гласи, че ако е регистрирана операция по използване на разплащателна сметка, последвана от операция със спестовна сметка, то с вероятност от 54,17% може да се очаква следваща операция с дебитна карта.

Обръщаме внимание на обстоятелството, че повечето секвенции завършват с използване на дебитна карта. Човек с познания в конкретната област и контекст би могъл да разгледа тези последователности, за да определи дали в тях има нещо интересно или неочаквано (като напр. поредица от едни и същи операции в рамките на една секвенция, както е при второто

правило). Ако например се открие секвенция, която има повече от 3, 4 или 5 еднотипни операции (напр. ползване на дебитна карта), това за банката би могло да бъде сигнал за злонамерено действие и картата да бъде автоматично блокирана. Въпреки че в конкретната база данни умишлено не са включени социодемографските признаци на клиентите, при наличието им е възможно да се идентифицират, клъстеризират и профилират групи потребители, със сходни модели на поведение.

Таблица 11.
Извлечени секвентни правила с алгоритъма на Агравал и Срикат (Agrawal & Srikant, 1995)

Антецедент(-и)	Консек-вент	Честота	Подкрепа на антецедента (%)	Довери-телност (%)	Подкрепа на правилото (%)
CKING	SVG	4329	85.78	63.15	54.17
ATM	ATM	1709	38.46	55.61	21.39
CKING > ATM	ATM	1546	36.19	53.46	19.35
CKING > SVG	ATM	1986	54.17	45.88	24.85
CKING	ATM	2892	85.78	42.19	36.19
SVG	ATM	2053	61.87	41.53	25.69
SVG	CD	1256	61.87	25.40	15.72
CKING	CD	1677	85.78	24.46	20.99

Създаденият набор от често срещани секвенционни правила е възможно да бъде използван за предсказване на поведението на индивидуално равнище. Обикновено се използват първите три правила, отличаващи се с най-висока доверителност (IBM, 2010, pp. 5-12). На Таблица 12 в колоните с префикс "$S-" са поместени трите най-надеждни прогнози за очакваните банкови операции, които да се появят впоследствие в секвенцията и предсказани на базата на наблюдаваните до този момент събития. Стойностите на показателя „доверие" за всяка прогнозна оценка са представени в колоните с префикс "$SC-". Таблицата илюстрира само първите 15 записа от общо 32376. От първия запис (клиент с номер 500026 и първи случай на ползване на услуга) личи, че е била ползвана разплащателна сметка (CKING). Най-вероятната услуга, която се очаква да ползва този клиент при следващия контакт с банката, при условие че в предходния е ползвал разплащателна сметка, е някаква операция със спестовна сметка (вж. етикета в колона $S-ACCT-1). Вероятността за това е 63,2% Това правило може да се види и на първия ред от Таблица 11. Тъй като повечето секвенции, предсказани с това правило, завършват със SVG, втората и третата по сигурност предсказани операции са в известен смисъл по-интересни в конкретния случай. Така например следващата най-вероятна

услуга, която може да се очаква да бъде използвана от този клиент, при условие че е извършвал някаква операция със спестовна сметка, е ползването на дебитна карта (вж. етикета в колона $A-ACCT-2). Доверителността за това събитие е 42,2%. Третата най-вероятна операция е със срочен депозит. Вероятността за това е 24,5%. По този начин се генерират трите най-сигурни бъдещи прогнози, базирани на наблюдаваната секвенция. Разглеждайки прогнозите за клиента с номер 500026, забележете, че най-вероятната банкова операция, която ще се появи по-късно, може да се промени с промяната на наблюдаваната секвенция. Това е логично, тъй като с получаването на повече информация за дадена последователност могат да се прилагат допълнителни правила.

Таблица 12

Трите най-добри прогнози за секвенциите от банкови операции

ACCT	SERVICE	VISIT	$S-ACCT-1	$SC-ACCT-1	$S-ACCT-2	$SC-ACCT-2	$S-ACCT-3	$SC-ACCT-3
500026	CKING	1	SVG	0.632	ATM	0.422	CD	0.245
500026	SVG	2	SVG	0.632	ATM	0.459	CD	0.254
500026	ATM	3	SVG	0.632	ATM	0.556	CD	0.254
500026	ATM	4	SVG	0.632	ATM	0.556	CD	0.254
500075	CKING	1	SVG	0.632	ATM	0.422	CD	0.245
500075	MMDA	2	SVG	0.632	ATM	0.422	CD	0.245
500075	SVG	3	SVG	0.632	ATM	0.459	CD	0.254
500075	ATM	4	SVG	0.632	ATM	0.556	CD	0.254
500075	TRUST	5	SVG	0.632	ATM	0.556	CD	0.254
500075	TRUST	6	SVG	0.632	ATM	0.556	CD	0.254
500129	CKING	1	SVG	0.632	ATM	0.422	CD	0.245
500129	SVG	2	SVG	0.632	ATM	0.459	CD	0.254
500129	IRA	3	SVG	0.632	ATM	0.459	CD	0.254
500129	ATM	4	SVG	0.632	ATM	0.556	CD	0.254
500129	ATM	5	SVG	0.632	ATM	0.556	CD	0.254
…	…	…	…	…	…	…	…	…

Получените предвиждания и стойностите на доверителност са ценни, тъй като идентифицират най-надеждните прогнози въз основа на целия набор от правила. От резултативния набор от данни, част от който е илюстриран на Таблица 12, с минимални усилия и програмен код е възможно да се извлекат конкретни правила и/или да се намери конкретна секвенция, интересуваща изследователя.

Синопсис и препоръки

Разгледаните подходи и алгоритми за анализ на пазарната кошница и поведението при купуване на потребителите могат да се класифицират едновременно и като експлоративни, и като дескриптивни изследователски методи. Целта им най-общо е, на базата на асоциации да идентифицират модели (или правила), съставени от малък брой елементи (продуктови артикули). Анализът на асоциациите дава лесни за разбиране резултати, формулирани като логически правила. С други думи това са техники за откриване на връзките на един или повече елементи (набор от асортиментни позиции) с друг елемент. Всяка от тези закономерности се нарича правило и обикновено се генерират голям брой правила за всяка съвкупност от данни със сравнително разнообразен вид и брой трансакции. За да се получат надеждни резултати, са необходими входни данни с достатъчен брой случаи във всяка категория. С други думи изкривените или недостатъчни по обем извадки могат да доведат до лоши решения. Освен това откритите взаимовръзки при анализа на асоциациите са само такива и не предполагат причинно-следствена връзка. Изключение правят моделите за извличане на секвенции. При всички случаи обаче при интерпретиране на резултатите трябва да се изхожда от бизнес познания, здрав разум и друга съпътстваща информация, за да оценените напълно какви модели и закономерности биват откривани.

Анализът на пазарните кошници чрез извличането на асоциативни правила е подходящ, когато е налице сравнително голям набор от данни от дискретни продуктови артикули, които е необходимо да бъдат групирани в малки по размер групи (често срещани множества). Подобни сценарии се наблюдават най-често в търговията на дребно, но няма причина, приложенията да се ограничават само до тази област. Застрахователни искове, банкови трансакции или медицински процедури са други потенциални области на приложения.

Анализът на пазарните кошници обикновено се извършва върху много големи масиви от данни с десетки хиляди трансакции и стотици позиции. Но с нарастването на размера на файловете, особено на броя на елементите (купуваните артикули), броят на потенциалните връзки нараства бързо, както и необходимото време за изчисления. Например само за 50 елемента има 1225 различни комбинации от два елемента, 19 600 от три елемента и т.н.[11]. Така че на практика броят на случаите (клиентите) може да бъде голям, но броят на отделните елементи обикновено е много по-малък, както и броят на елементите, които трябва да бъдат свързани заедно. Следова-

[11] За конкретните формули на намиране размера на допустимите комбинации вж. Кръстевич, Т. (Анализ на пазарната кошница с R, 2021, стр. 30 и 41).

телно може да е трудно да се определи априори правилният брой и вид елементи, така че обикновено, както при всички автоматизирани методи, могат да се изпробват няколко различни решения с различни набори от правила. Препоръчваме също така стандартна валидация, т.е. оценяване на генерираните правила, които представляват интерес, върху извадка от данни за валидиране (SPSS Inc., 2003, pp. 7-8).

Що се касае до моделите за разкриване на секвентни закономерности и зависимости, те представляват по същество разширение и допълнение на класическия анализ на асоциациите чрез отчитането на фактора време. Тук аналогия с методите за анализ на времеви редове обаче е недопустима, тъй като времевите секвенции не отчитат метрично дистанцията между моментите на извършване на продажбените трансакции. Асоциативните правила, извличани с помощта на алгоритмите за анализ на секвенциите, имат същия общ формат като правилата при традиционния анализ на пазарната кошница. С добавянето на времевия компонент обаче откритите правила дават допълнителна информация за вземане на адекватни маркетингови решения.

Данните, подходящи за анализ на асоциациите, трябва да се събират във времето от едни и същи лица или единици. Променливата ID служи за групиране на случаите от една и съща единица. Например верига супермаркети, с помощта на карти за редовни клиенти, може да проследява покупките, направени при многократни пазарувания от един и същ клиент. Последователният анализ на тези данни може да даде отговор на въпроси като: „Купуват ли се пресни хлебни изделия при последователни пътувания? “ или „След като е купил замразени зеленчуци, купува ли потребителят отново замразени зеленчуци в рамките на три седмици?“.

Анализът на секвенциите може да се прилага към всякакви подходящи данни, така че да се проследява работата на оборудването с течение на времето, като се търсят модели, които водят до повреда, или да се изследват транзакциите с кредитни карти, за да се търсят последователности, които могат да подскажат дали картата е подновена или не.

Анализът на секвенциите е подходящ винаги, когато е налице набор от дискретни елементи, които трябва да се групират или да изследват за закономерности във времето. Това се случва най-често в търговията на дребно, но няма причина да се ограничава само до тази област. Както и при анализа на асоциациите, други потенциални приложения са застрахователни искове, банкови трансакции, медицински процедури и последователни активности в посещенията на уебсайтове по време на пазаруване онлайн.

Едно от най-големите предимства на анализа на пазарната кошница чрез извеждане на асоциативни правила е, че резултатите от него са лесно разбираеми за всеки. Правилата, които той открива, могат да бъдат изразени на обикновен разговорен език и не изискват никакви статистически тестове. Обикновено графичните резултати, с изключение на малки набори от

артикули, не са толкова полезни, колкото цифровите обобщения, показващи действителната сила на асоциацията.

Създадените правила за асоцииране и секвениране могат да бъдат кодирани като оператори в SQL. Това означава, че при необходимост те могат да се прилагат директно към бази данни. Въпреки това внедряването на модела при тези техники обикновено не включва прилагане на правилата към нов файл с данни. Вместо това резултатите са директно приложими, тъй като могат да се вземат решения за кръстосани продажби и промоции чрез просто разглеждане на асоциациите и определяне на това, кое ниво на асоциация е достатъчно високо, за да изисква действие.

Използвани източници

Aggarwal, C. C., & Yu, P. S. (1998). Online Generation of Association Rules. *Proceedings of the 14th International Conference on Data Engineering* (pp. 402-411). Los Alamitos, CA: IEEE ComputerSociety Press.

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, (pp. 487-499).

Agrawal, R., & Srikant, R. (1995). Mining Sequential Patterns. *Proceedings of the Eleventh International Conference on Data Engineering* (pp. 3-14). Los Alamitos, CA: IEEE Computer.

Agrawal, R., Imielinski, T., & Swami, A. (1993, December). Database Mining: A Performance Perspective. *IEEE Transactions on Knoledge and Data Engineering, 5*(6), 914-925.

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proc. of ACM SIGMOD Intl. Conf. Management of Data*, (pp. 207-216). Washington, DC.

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast Discovery of Association Rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, & R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining* (pp. 307-328). Menlo Park, CA: AAAI Press.

Bermingham, L., & Lee, I. (2014). Spatio-temporal sequential pattern mining for tourism scineces. *Procedia Computer Science*, *29*, pp. 379-389. doi:10.1016/j.procs.2014.05.034

Berry, M. J., & Linoff, G. S. (2011). *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (3 ed.). Indianapolis: John Wiley and Sons.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 437-456.

Borgelt, C. (2017). *Apriori: Find Frequent Item Sets and Association Rules with the Apriori Algorithm.* Retrieved 10 28, 2021, from borgelt.net: https://borgelt.net/doc/apriori/apriori.html#diff

Brin, S., Motwani, R., Ullman, J. D., & Tsur, S. (1997). Dynamic Itemset Counting and Implication Rules for Market Basket Data. *ACM SIGMOD Record, 26*(2), 255-264. doi:10.1145/253262.253325

Buchta, C., & Hahsler, M. (2020). „*arulesSequences: Mining frequent sequences*“. *R package version 0.2-25'.* CRAN. Retrieved 10 9, 2021, from https://bit.ly/3AoTgAX

Cabanes, G., Bennani, Y., & Dufau-Joël, F. (2009). Mining Customers' Spatio-temporal Behavior Data using Topographic Unsupervised Learning. *8th International Conference on Machine Learning and Applications, ICMLA 2009*, (pp. 372-377). doi:10.1109/icmla.2009.23

Chand, C., Thakkar, A., & Ganatra, A. (2012). Sequential Pattern Mining: Survey and Current Research Challenges. *International Journal of Soft Computing and Engineering (IJSCE), 2*(1), 185-193.

Chapman, C., & Feit, E. M. (2019). *R for Marketing Research and Analytics* (2 ed.). Cham: Springer.

Chorianopoulos, A. (2016). *Effective CRM Using Predictive Analytics.* Chichester: John Wiley & Sons.

contributors, W. (11 8 2021 r.). *Data Mining Algorithms In R.* (T. F. Wikibooks, Ред.) Изтеглено на 9 10 2021 г. от wikibooks.org: https://bit.ly/3DEOj9p

Dzyabura, D., & Yoganarasimhan, H. (2018). Machine learning and marketing. In N. Mizik, & D. M. Hanssens, *Handbook of Marketing Analytics* (pp. 255-279). Cheltenham,: Edward Elgar.

Eihinger, F., Nauck, D., & Klawonn, F. (2006). Sequence Mining for Customer Behaviour Predictions in Telecommunications. *Proceedings of theWorkshop on Practical Data Mining at ECML/PKDD*, (pp. 3-10).

Goel, A., & Mallick, B. (2015). Customer Purchasing Behavior using Sequential Pattern Mining Technique. *International Journal of Computer Applications, 19*(1), 24-31.

Guidotti, R., Rossetti, G., Pappalardo, L., Giannotti, F., & Pedreschi, D. (2019). Personalized Market Basket Prediction with Temporal Annotated Recurring Sequences. *IEEE Transactions on Knowledge and Data Engineering, 31*(11), 2151-2163. doi:10.1109/TKDE.2018.2872587

Gupta, M., & Han, J. (2012). Applications of Pattern Discovery Using Sequential Data Mining. In P. Kumar, P. R. Krishna, & S. B. Raju, *Pattern Discovery Using Sequence Data Mining. Application an Studies* (pp. 1-23). IGI Global.

Hidber, C. (1999). Online Association Rule Mining. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data - SIGMOD '99* (pp. 145-156). New York: ACM Press.

IBM. (2010). Clustering and Association Models with IBM SPSS Modeler. *Course Code: 0A042*. IBM Corp.

IBM Corp. (2020). *SPSS Modeler Algorithms Guide.* IBM Corporation.

Koenecke, A. (2019, January 22). *Tutorial: Sequential Pattern Mining in R for Business Recommendations*. Retrieved October 7, 2021, from

https://blog.revolutionanalytics.com/: https://blog.revolutionanalytics.com/2019/02/01/

Kuhn, M., & Johnson, K. (2018). *Applied predictive modeling.* New York: Springer.

Mabroukeh, N. R., & Ezeife, C. I. (2010). A Taxonomy of Sequential Pattern Mining Algorithms. *ACM Computing Surveys, 43*(1), 1-41.

Mazarbhuiya, F. A. (2015). Mining Sequential Patterns from Super Market Datasets. *International Journal of Computer Trends and Technology (IJCTT), 30*(4), 206-212.

Mukherji, A., Srinivasan, V., & Welbourne, E. (2014). Adding Intelligence to Your Mobile Device via On-Device Sequential Pattern Mining. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct Publication - UbiComp '14 Adjunct*, (pp. 1005-1014). doi:10.1145/2638728.2641285

Peng, W.-C., & Liao, Z.-X. (2009). Mining sequential patterns across multiple sequence databases. *Data & Knowledge Engineering, 68*(10), 1014-1033. doi:10.1016/j.datak.2009.04.009

Reps, J. M., Garibaldi, J. M., Aickelin, U., Soria, D., Gibson, J. E., & Hubbard, R. B. (2012). Discovering sequential patterns in aUKgeneral practice database. *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*, (стр. 960-963).

Ribeiro, J. M. (2016, January). *Sequence Mining analysis on Shopping Data.* Master's Dissertation, Porto.

RuleQuest Research Ltd Pty. (2020). *Data Mining Tools See5 and C5.0.* Retrieved 10 28, 2021, from www.rulequest.com: https://www.rulequest. com/see5-info.html

SPSS Inc. (2003). *Data Mining: Overview.* Chicago, IL: SPSS Inc.

Tsai, C. -Y., & Lai, B. -H. (2015). A Location-Item-Time sequential pattern mining algorithm for route recommendation. *Knowledge-Based Systems, 73*, 97-110. doi:10.1016/j.knosys.2014.09.012

Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2014). The use of sequential pattern mining to predict next prescribed medications. *Journal of Biomedical Informatics, 53*, 73-80. doi:10.1016/j.jbi.2014.09.003

Wright, A. P., Wright, A. T., McCoy, A. B., & Sittig, D. F. (2015, February). The use of sequential pattern mining to predict next prescribed. *Journal of Biomedical Informatics, 53*, 73-80.

Zaki, M. J. (2001). SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning Journal, 42*, 31-60.

Кръстевич, Т. Б. (2021). *Анализ на пазарната кошница с R* (Том 143). Свищов: Библиотека „Стопански свят".

том 30, 2022 г.

ИНСТИТУТ ЗА НАУЧНИ ИЗСЛЕДВАНИЯ
ПРИ СТОПАНСКА АКАДЕМИЯ
„Д. А. ЦЕНОВ" - СВИЩОВ

АЛМАНАХ
НАУЧНИ ИЗСЛЕДВАНИЯ

ИКОНОМИЧЕСКИ И УПРАВЛЕНСКИ ИЗМЕРЕНИЯ НА ОБЩЕСТВЕНАТА ТРАНСФОРМАЦИЯ

том 30, 2022 г.

АЛМАНАХ
НАУЧНИ ИЗСЛЕДВАНИЯ

Академично издателство „ЦЕНОВ"
Свищов - 2022 г.

ИНСТИТУТ ЗА НАУЧНИ ИЗСЛЕДВАНИЯ
СТОПАНСКА АКАДЕМИЯ „Д. А. ЦЕНОВ” – СВИЩОВ

АЛМАНАХ НАУЧНИ ИЗСЛЕДВАНИЯ

ИКОНОМИЧЕСКИ И УПРАВЛЕНСКИ ИЗМЕРЕНИЯ НА ОБЩЕСТВЕНАТА ТРАНСФОРМАЦИЯ

ТОМ 30
2022

АКАДЕМИЧНО ИЗДАТЕЛСТВО „ЦЕНОВ” – СВИЩОВ

Издава се със средства от целевата субсидия за научна дейност на СА „Д. А. Ценов”, съгласно Наредбата за условията и реда за оценката и планирането, разпределението и разходването на средствата от държавния бюджет за финансиране на присъщата на държавните висши училища научна или художествено творческа дейност.

ISSN 1312-3815

INSTITUTE FOR SCIENTIFIC RESEARCH
D. A. TSENOV ACADEMY OF ECONOMICS – SVISHTOV

SCIENTIFIC RESEARCH ALMANAC

ECONOMIC AND MANAGERIAL DIMENSIONS OF SOCIAL TRANSFORMATION

VOLUME 30
2022

TSENOV ACADEMIC PUBLISHING HOUSE

This issue is funded by the state „Ordinance on the terms and procedure for the evaluation and planning, allocation and spending of the state budget funds in financing scientific or artistic activities, intrinsic to state higher schools" for the inherent to the „D. A. Tsenov" Academy of Economics scientific activity.

ISSN 1312-3815

СЪДЪРЖАНИЕ

Раздел I

Пазари, управление и иновации в икономиката на знанието

Раздел II

Финансова стабилност, икономически политики, регулации и устойчиво развитие

Раздел III

Глобализация, конкурентоспособност и сътрудничество за интелигентен растеж

CONTENTS

Section I

Markets, Management and Innovations in Knowledge Economy

Section II

Financial Stability, Economic Policies, Regulations and Sustainable Development

Section III

Globalisation, Competitiveness and Cooperation for Intelligent Growth

СТОПАНСКА АКАДЕМИЯ „Д. А. ЦЕНОВ”

АЛМАНАХ НАУЧНИ ИЗСЛЕДВАНИЯ

ТОМ 30

ИКОНОМИЧЕСКИ И УПРАВЛЕНСКИ ИЗМЕРЕНИЯ НА ОБЩЕСТВЕНАТА ТРАНСФОРМАЦИЯ

Даден за печат на 09.02.2022 г., излязъл от печат на 24.03.2022 г.
Поръчка № 18798, тираж: 100 бр.

Издателство и печат: Академично издателство „Ценов”
Свищов, ул. „Цанко Церковски“ 11А

ISSN 1312-3815