Mariusz Łapczyński

Przygotowanie danych do analizy zachowań e-klientów : droga od danych do wiedzy

Marketing i Zarządzanie (d. Problemy Zarządzania, Finansów i Marketingu) nr 3 (49), 55-65

2017

Marketing i Zarządzanie
nr 3 (49) 2017, s. 55–65
DOI: 10.18276/miz.2017.49-05
ISSN: 1509-0507 | http://wnus.edu.pl/pl/miz/

Mariusz Łapczyński

Uniwersytet Ekonomiczny w Krakowie
Wydział Zarządzania
e-mail: lapczynm@uek.krakow.pl

Przygotowanie danych do analizy zachowań e-klientów – droga od danych do wiedzy[1]

Kody JEL: C80, M31

Słowa kluczowe: proces przygotowania danych, zachowania e-klientów, wykorzystanie internetu, e-commerce

Streszczenie. Celem artykułu jest wskazanie trudności, jakie można napotkać podczas przygotowania danych do analizy. Zaprezentowane przykłady odnoszą się do rzeczywistych danych pozyskanych z e-sklepu oferującego obuwie i dotyczą obszaru *web mining*, nazywanego analizą wzorców zachowań internautów. W artykule przedstawiono wyniki wstępnej eksploracji danych od momentu ich pozyskania, przez sprawdzenie i na przygotowaniu środowiska danych skończywszy.

Wprowadzenie

Przeglądając dowolne procedury analityczne *data mining*, można zauważyć, że w każdej z nich znajduje się co najmniej jeden etap dotyczący przygotowania danych do analizy. W iteracyjnym procesie budowy modeli zaproponowanym przez M.J.A. Berry'ego i G.S. Linoffa (2000) kryje się on pod pojęciem „walidacja,

[1] Wyniki przedstawione w artykule są częścią projektu realizowanego w ramach badań statutowych Katedry Analizy Rynku i Badań Marketingowych Uniwersytetu Ekonomicznego w Krakowie (028/WZ-KAR/02/2017/S/7028).

eksploracja i czyszczenie danych", zaś w procedurze P. Giudiciego (2003) opisany jest terminem „wybór, organizacja i wstępna obróbka danych". Z kolei w schemacie SEMMA (SAS Institute Inc., 2014) odnoszą się do niego dwa etapy: *explore* (eksploracja zbioru obserwacji) oraz *modify* (selekcja, modyfikowanie, usuwanie zmiennych), zaś w procedurze CRISP-DM (Chapman in., 2000) – etap drugi i trzeci, czyli „poznanie struktury danych" i „przygotowanie danych". Autorzy są zgodni, że jest to etap bardzo ważny – niejednokrotnie kluczowy do osiągnięcia satysfakcjonującego wyniku. Jednocześnie potwierdzają, że jest on bardzo czasochłonny i zajmuje co najmniej 50% czasu poświęconego łącznie na przygotowanie danych i budowę modelu.

Celem artykułu jest wskazanie problemów, na jakie można się natknąć podczas przygotowania danych do analizy. Przedstawione w dalszej części przykłady odnoszą się do rzeczywistych danych pozyskanych z e-sklepu, które oferuje obuwie[2]. Egzemplifikacja będzie dotyczyć wybranych etapów obróbki danych.

Etapy przygotowania danych

Tak jak w każdym działaniu zorganizowanym ważne jest odpowiednie planowanie. Procedura przygotowania danych obejmuje osiem etapów (Pyle, 1999, s. 112):

1. Uzyskanie dostępu do danych (*accessing data*) – import danych do programu analitycznego.
2. Sprawdzenie danych (*auditing data*) – sprawdzenie liczby zmiennych, poziomu pomiaru zmiennych, zakresów wartości zmiennych ilościowych itd.; badacz powinien w tym miejscu również zdecydować, czy dostępny zbiór obserwacji pozwoli na rozwiązanie problemów badawczych.
3. Uzupełnienie zbioru zmiennych (*enhancing and enriching data*) – dołączenie zmiennych ze źródeł zewnętrznych lub tworzenie zmiennych pochodnych.
4. Poszukiwanie błędów w doborze próby (*looking for sampling bias*) – sprawdzenie sposobu podziału zbioru danych na zbiór (próbę) uczący i testowy.
5. Ustalenie struktury danych (*determining data structure*) – określenie relacji między zmiennymi w zbiorze danych, charakterystyka super-, makro- i mikrostruktury danych.
6. Przygotowanie środowiska danych do analizy (*building prepared information environment*) – obejmuje m.in. normalizację zmiennych, zastępowanie braków danych czy redukcję liczby zmiennych lub liczby przypadków.

[2] Ze względu na poufność informacji, niektóre nazwy zmiennych i ich kategorii zmieniono w sposób uniemożliwiający identyfikację podmiotu gospodarczego udostępniającego dane.

7. Przegląd danych (*surveying data*) – ponowne sprawdzenie rozkładów zmiennych, identyfikacja rzadko występujących wartości zmiennych, zakłóceń, naturalnie występujących skupisk itp.
8. Budowa modelu (*modeling data*) (szerzej: Łapczyński, 2012, s. 129–141).

Procedura ta powinna zapewnić efektywne przygotowanie danych, spełniających ustalone wcześniej kryteria.

Problemy na etapie pozyskiwania danych

Zbiór danych zawierał około 30 zmiennych i ponad 3 mln przypadków. Każdy przypadek oznaczał jedno działanie wykonane przez użytkownika (fragment zbioru danych przedstawiono w tab. 1), o czym należy pamiętać podczas budowy modelu statycznego, kiedy jeden wiersz zazwyczaj odnosi się do jednego internauty. Puste pola w tabeli (kolumny: „cena produktu" i „produkt") wskazują na działania, które odnoszą się do produktu wskazanego w pełnym wierszu powyżej. Przykładowo, cztery puste pola dla wizyty nr 1500-b poniżej produktu C oznaczają, że cała sekwencja zdarzeń dotyczy tego właśnie produktu (od godziny 12:37:10 do godziny 12:37:54). Działania te mogły dotyczyć np. dostępnych form płatności, sposobów dostawy, tabeli rozmiarów czy kolejnych fotografii tego samego produktu.

Przy tak dużej liczbie wierszy możliwość przeglądu danych w programie Excel istnieje tylko wtedy, gdy plik zostanie podzielony na kilka mniejszych, o liczbie obserwacji, którą można otworzyć w tym arkuszu kalkulacyjnym (n = 1 048 576). Największym utrudnieniem, na jakie napotkano, było jednak występowanie średników w nazwach kategorii zmiennych. Ze względu na to, że był to plik tekstowy oddzielany średnikami (*.csv), to pojawienie się dodatkowych znaków separujących kolumny przekładało się na większą liczbę kolumn w niektórych wierszach tabeli.

Tabela 1

Fragment zbioru danych

Nr wizyty	Godzina działania	Czas trwania działania (s)	Cena produktu	Produkt	Nazwa przeglądarki	Typ urządzenia	System operacyjny
1500-a	12:33:31	24			Google	desktop	Win 8
1500-a	12:34:01	8			Google	desktop	Win 8
1500-a	12:34:41	40	0,001	A	Google	desktop	Win 8
1500-a	12:34:51	10			Google	desktop	Win 8
1500-a	12:35:45	54	0,178	B	Google	desktop	Win 8

1500-b	12:37:10	46	0,016	C	Google	telefon	Apple 8
1500-b	12:37:20	10			Google	telefon	Apple 8
1500-b	12:37:25	5			Google	telefon	Apple 8
1500-b	12:37:32	7			Google	telefon	Apple 8
1500-b	12:37:54	22			Google	telefon	Apple 8
1500-c	12:38:09	15			nieznana	konsola do gier	Play Station 3
1500-c	12:38:37	28			nieznana	konsola do gier	Play Station 3
1500-c	12:38:38	1			nieznana	konsola do gier	Play Station 3
1500-c	12:38:47	9			nieznana	konsola do gier	Play Station 3

Źródło: dane z e-sklepu.

Wybrane problemy podczas sprawdzania danych

Podczas analizy rozkładów zmiennych zaobserwowano, że niektóre z nich mają kategorie tekstowe i wartości liczbowe. Było to wynikiem – wspomnianej wcześniej – niejednakowej liczby zmiennych w poszczególnych przypadkach. Okazało się, że etykiety zmiennych jakościowych zostały przesunięte do sąsiednich kolumn, w których znalazły się zmienne ilościowe (i na odwrót). Określenie skali pomiarowej było możliwe dopiero po uporządkowaniu całego zbioru danych.

W tabeli 2 przedstawiono przykładowe zmienne i kategorie/wartości, które w nich występowały. W celu wyczyszczenia danych konieczne było przesunięcie kolumn, wskazanie symbolu określającego brak danych (tutaj: \N), zmiana formatu daty albo wskazanie separatora miejsc dziesiętnych wartości liczbowych. W tym ostatnim wypadku pozostawienie kropek było tożsame z traktowaniem tej zmiennej przez program analityczny jako zmiennej jakościowej z etykietami tekstowymi. Gdyby badacz chciał zinterpretować statystyki opisowe (np. średnią arytmetyczną, odchylenie standardowe), to byłyby one liczone z kodów przypisanych do tych etykiet, czyli np. z liczb 101 i 102 zamiast 49.99 i 59.99.

Prosta analiza zmienności za pomocą rozstępu i wykresów ramka-wąsy pozwoliła wskazać w niektórych zmiennych ilościowych obserwacje podejrzane o nietypowość. Jedną z nich była zmienna informująca o czasie wykonywania działania w obrębie witryny. Okazuje się, że 0,05% działań trwało od 1700 do 1800 sekund, czyli od około 28 do 30 minut. Tak duże wartości mogą oznaczać błąd, ale mogą również wskazywać na to, że internauta po wyborze produktu na stronie

e-sklepu odszedł od komputera i powrócił do przeglądania oferty po pół godzinie. W innej zmiennej – „liczba oglądanych podczas wizyty produktów" – 0,008% wartości było wyższych od 114, co oznacza, że taki procent odwiedzających przeglądał ponad 114 produktów podczas jednorazowej wizyty na stronie sklepu. Wydaje się to mało prawdopodobne, jednak nie można wykluczyć, że stronę e-sklepu przeglądali przedstawiciele konkurencji, którzy chcieli zapoznać się z ofertą rynkową firmy. Być może były to roboty indeksujące (*web crawlers*) albo pracownik e-sklepu odpowiedzialny za projektowanie i aktualizację strony www.

Tabela 2

Przykłady zmiennych o nieokreślonym poziomie pomiaru

Kategorie/ wartości zmiennej 1	Liczba obserwacji	Kategorie/ wartości zmiennej 2	Liczba obserwacji	Kategorie/ wartości zmiennej 3	Liczba obserwacji
8	112	;1171"	3	\N	6501
9	97	;282890"	1389	199.99	6419
10	3477	8/31/1900	176	209.99	275
WIN	2043	9/2/1900	1	89.99	1559
AND	673	9/4/1900	34	229.99	1172
IOS	1169	9/6/1900	6276	269.99	2579

Źródło: opracowanie własne.

Warto dodać, że obserwacje nietypowe mogą znajdować się również w zmiennych jakościowych. W zbiorze danych znajduje się informacja o tym, jaki język był ustawiony w przeglądarce internetowej internauty oraz z jakiego kraju łączył się on z siecią. Przykładowo, skrót en-us oznacza, że użytkownik używał oprogramowania w języku angielskim (en-) i łączył się z internetem z terenu Stanów Zjednoczonych (-us). Za obserwacje nietypowe należy uznać języki: kataloński z Hiszpanii (ca-es), prowansalski z Francji (oc-fr), ormiański (hy) czy gruziński (ka).

Inna zmienna jakościowa – „typ urządzenia" – informowała o rodzaju urządzenia, za pomocą którego internauci łączą się z siecią. Oprócz popularnych komputerów stacjonarnych i laptopów (77% wizyt), były tam smartfony (20%) i tablety (2,5%), ale także konsole do gier (Xbox, Nintendo, Play Station) i telewizory z dostępem do internetu. Należy dodać, że 0,13% wizyt realizowanych za pomocą konsol i telewizorów to – w ujęciu absolutnym – prawie 1000 odwiedzin sklepu.

Na tym etapie czyszczenia danych należy również zidentyfikować etykiety kategorii zmiennych jakościowych. Wydaje się, że brak polskiej czcionki nie stanowi dużej przeszkody, gdyż kategorie mogą być opisane z użyciem przypadkowych znaków (np. Tenis\\u00f3wki i trampki, TenisA3wkiitrampki, TenisĂłwki

i trampki, Tenisﾃｳwki i trampki), bez spacji (np. buty Reebok i butyReebok) czy za pomocą nieokreślonych symboli (np. S失, S粽,T\,ﾁmヤL, ﾁ豊, ツ2彝, %/, ﾌ, %/@ﾌ).

Podczas sprawdzania danych należy również poznać definicje zmiennych i wstępnie ustalić ich przydatność do rozwiązania problemu badawczego. W zbiorze danych występuje zmienna „nazwa wyszukiwarki", która powinna zawierać etykiety odnoszące się do nazw wyszukiwarek internetowych, np. Google, Yahoo czy Bing. Okazało się jednak, że znajdowało się w niej 658 kategorii odwołujących się do nazw marek produktów, konkretnych adresów stron internetowych niebędących wyszukiwarkami (np. lapajciucha.pl, wbutach.pl) czy do nazw kampanii promocyjnych. Etykiety tekstowe są niespójne, co może wskazywać na nieprecyzyjną definicję tej zmiennej.

Niektóre zmienne, jak np. „numer wizyty", określa sposób wczytywania danych przez programy analityczne. Pozwala ona na identyfikację pojedynczych wizyt na stronie, a nie pojedynczych użytkowników. Przykładowo, wizyty jednego internauty, który odwiedzał stronę sklepu 10 razy, traktowane są jako 10 odrębnych wizyt. W badanym okresie zarejestrowano około 3 mln działań w obrębie witryny, które wykonano podczas około 700 tys. wizyt. Okazało się, że prawie połowa wizyt cechowała się tylko jednym działaniem, co może oznaczać, że odwiedzający stronę opuścił ją natychmiast po jej otwarciu. Mogli to być internauci odesłani do e-sklepu przez łącze lub baner reklamowy z innej strony, internauci, którzy znaleźli się tu przypadkowo albo ci, którzy zdecydowali, że będą kontynuować przeglądanie oferty w innym czasie.

Przykładowe problemy związane z przygotowaniem środowiska danych

Najczęściej pojawiającym się problemem w tym zbiorze danych była nadmierna liczba kategorii zmiennych jakościowych. Zmienna „model urządzenia" zawiera 1218 kategorii, które odnoszą się do modelu urządzenia, za pomocą którego użytkownik łączy się z internetem. Nazwy urządzeń przyporządkowano jedynie do około 20% wszystkich wizyt. Pozostałe 80% to braki danych, które mogą odnosić się do komputerów stacjonarnych. Duża liczba kategorii bardzo utrudnia analizę i w niektórych przypadkach uniemożliwia wprowadzenie takich zmiennych do modelu. Dodatkowo, wiele symboli odnosi się do tej samej marki i modelu telefonu, np. GALAXY S5, SM-J500FN, SM-G531F i SM-G903F to odmiany telefonu Samsung Galaxy. Zdecydowano się zatem na utworzenie zmiennej pochodnej – „marka", której liczba kategorii została zredukowana do 76.

Zmienna „kategoria produktu" miała pierwotnie około 300 wariantów, które przyporządkowywały produkt do bardziej ogólnej klasy. Część nazw odnosiła się

do konkretnych marek obuwia, a ich niewielka częstotliwość występowania spowodowała, że zostały zastąpione jedną kategorią – „buty inne markowe". Niektóre warianty zostały zbyt ogólnie sformułowane, np. „buty i obuwie", „buty męskie" czy „buty damskie", co utrudnia ich porównanie z konkretnymi typami obuwia, np. sandały, baleriny, koturny, kalosze. Wszystkie warianty, w których nazwie wskazano na kanał dystrybucji (nazwę sklepu internetowego) zamieniono na stosowne kategorie (np. buty damskie od firmy ABC, buty męskie od firmy ABC) nawet wówczas, gdy pokrywały się z podobnymi kategoriami produktów dystrybuowanymi w inny sposób (np. buty damskie, buty męskie). Marki obuwia, które charakteryzowały się stosunkowo dużą częstotliwością wyboru i jednocześnie uznano za charakterystyczne i unikatowe, utworzyły odrębne kategorie tej zmiennej: Dr. Martens, Hunter czy Crocs. Wydaje się, że sposób wprowadzania danych do bazy uniemożliwia rozłączną klasyfikację produktów. Przykładowo, buty do trekkingu mogły zostać wprowadzone do bazy danych jako: trapery i trekkingi, trapery i tekkingi marki ABC, buty ABC, buty Merrell, buty Timberland, trapery i trekingi damskie na zimę od firmy ABC. Pojawia się tu jednak problem z klasyfikacją, ponieważ:

- firmy Merrell i Timberland produkują również sandały,
- firma ABC ma w ofercie buty podobne do obuwia Dr. Martens (glany, półbuty) i do obuwia Vans (tenisówki); te typy obuwia mają swoje odrębne klasy („glany", „półbuty", „tenisówki i trampki"),
- wskazanie na kanał dystrybucji (firma ABC) włącza buty trekkingowe do ogólnej kategorii „buty damskie od firmy ABC".

W części zmiennych jakościowych dostrzeżono trudne do zidentyfikowania kategorie. Przykładowo, zmienna „nazwa przeglądarki" występuje w 62 różnych wariantach odnoszących się do nazwy przeglądarki internetowej, z której korzysta użytkownik. Najpopularniejsze były oznaczone symbolami: CH (najprawdopodobniej Chrome), FF (najprawdopodobniej Firefox), OP (najprawdopodobniej Opera) i MF (najprawdopodobniej Mozilla Firefox). Trudność w ustaleniu właściwej nazwy wynika z tego, że niektóre skróty mogą oznaczać kilka różnych programów, np. BB to Bitty Browser, Baidu Browser albo Brisk Bard. Skrót CO może z kolei oznaczać Comet Bird, Comodo Dragon, Comodo Ice Dragon, Conkeror albo Cốc Cốc Browser. Jeżeli z kolei FF i MF oznaczają ten sam program, to powinny być połączone w jedną klasę.

Kłopotliwa może być również nadmierna ziarnistość (rozdrobnienie, granularność) danych. Wśród adresów stron internetowych, których wizyta poprzedzała wizytę w e-sklepie, znajdują się m.in.:

- pobrane z aplikacji mobilnych, np. aplikacja Google dla systemu Android (com.google.android.googlequicksearchbox),

- porównywarki cenowe, np. okazje.info,
- wyszukiwarki produktów, np.allani.pl, salesandshopping.pl,
- wyszukiwarki sklepów, np. mapahandlu.pl,
- wyszukiwarki internetowe, np. bing.com, r.search.yahoo.com,
- Facebook, z którego wizyty mogą być identyfikowane na kilka sposobów: facebook.com (wizyta z FB bez zabezpieczeń typu LinkShim), l.facebook.com (wizyta z FB z zabezpieczeniami typu LinkShim), m.facebook.com (wizyta z telefonu komórkowego bez zabezpieczeń typu LinkShim) oraz lm.facebook.com (wizyta z telefonu komórkowego z zabezpieczeniami typu LinkShim).

Wątpliwości dotyczą wejść klientów na stronę sklepu bezpośrednio ze strony Facebooka: czy informacja o posiadaniu zabezpieczeń bądź ich braku pozwoli lepiej profilować nabywców?

Inny przykład dotyczący nadmiernej ziarnistości danych odnosi się do zmiennej „system operacyjny". Najpopularniejsze kategorie zdają się potwierdzać intuicyjną znajomość rynku – najwięcej użytkowników używa programów firmy Microsoft, potem Android (dotyczy urządzeń mobilnych), a na końcu Apple. Pod rozwagę należy wziąć możliwość grupowania programów opartych np. na tym samym jądrze Linuxa: Ubuntu, Kubuntu, Edubuntu czy Xubuntu. Wydaje się, że niewielka popularność tych systemów pozwala potraktować to jako jedną kategorię – „inne systemy operacyjne".

Kolejny przykład związany z nadmiernym rozdrobnieniem danych dotyczy, wspomnianej wcześniej, wersji językowej przeglądarki użytkownika. Łącznie zmienna posiadała 143 kategorie, jednak nie oznacza to, że użytkownicy mieli do dyspozycji aż tyle wersji językowych w przeglądarkach, ponieważ:

- symbol pl-pl oznacza język polski (pierwszy człon skrótu „pl-") z kraju Polska (drugi człon skrótu „-pl",
- symbol „pl" informuje jedynie o języku polskim,
- symbol en-us i en-gb informuje o języku angielskim z odpowiednio: Stanów Zjednoczonych i Wielkiej Brytanii itd.

Z punktu widzenia właściciela witryny, tworzenie i obsługa nowych wersji językowych e-sklepu powinny uwzględniać wyłącznie statystyki dotyczące języka, którym posługują się potencjalni klienci, a nie kraj, z którego łączą się z internetem. Wydaje się, że wejście z serwerów Jamajki (en-ja), Irlandii (en-ia), Australii (en-au), Kanady (en-ca) czy Belize (en-bz) jest kwestią mniej ważną niż posługiwanie się przez tych użytkowników językiem angielskim. Zdarzają się również sytuacje, gdy użytkownik łączy się z Polski, a jego przeglądarka ma ustawiony język angielski. Może to oznaczać, że klient anglojęzyczny (turysta, emigrant, pracownik konsulatu

itp.) łączy się z internetem z naszego kraju albo Polak używa komputera, który ma oprogramowanie bez polskiej wersji językowej. Trudno tu obiektywnie zdecydować, jaka to ma być kategoria zmiennej (język polski czy język angielski). Na szczęście niewielka liczba takich wizyt na stronie nie obciąża znacząco wyników analizy.

Podczas obróbki danych, ze zbioru obserwacji usuwa się zmienne o nieprecyzyjnych definicjach bądź zmienne, które nie przyczyniają się do rozwiązania problemu badawczego. Przykładem może być zmienna „liczba dni od pierwszej wizyty", która zawiera informację o tym, ile dni upłynęło od pierwszej wizyty internauty w sklepie. Niespełna 99% wizyt na stronie sklepu odbywało się po raz pierwszy, co może oznaczać, że rzeczywiście tak było albo że zdecydowana większość użytkowników usuwa z pamięci przeglądarki ciasteczka, czyli pliki pozwalające na analizę ich aktywności sieciowej w przeszłości. Możliwe jest również, że dane o pierwszej wizycie na stronie sklepu nie były gromadzone. Jeżeli klienci nie muszą się logować podczas odwiedzin sklepu, to śledzenie ich przeszłych zachowań w sieci jest właściwie niemożliwe ze względu m.in. na używanie różnych komputerów w domu i w pracy, rotacyjny adres IP, czyszczenie ciasteczek, zmianę systemu operacyjnego, zmianę dostawcy internetu, wymianę komputera czy łączenie się z siecią na przemian z urządzeń stacjonarnych i mobilnych. Zmienna została wykluczona z analizy. W zbiorze danych pozostawiono natomiast zmienną „liczba dni od ostatniego zamówienia", która informuje o liczbie dni, jaka upłynęła od ostatniego zakupu. Identyfikacja kolejnych odwiedzin jest bowiem bardziej prawdopodobna, jeżeli odwiedzający dokonał już kiedyś zakupu w sklepie (może mieć już swoje konto; jest możliwe, że loguje się na stronie).

Ze zbioru danych usunięto również zmienną „wersja przeglądarki internetowej", której przykładowe kategorie to: 999.1, 11.64, 48.0 czy 50.14. Łącznie zmienna ta ma 409 kategorii, których wartość informacyjną oceniono jako niską. Podobnie postąpiono ze zmienną „klient powracający", określającą, czy internauta jest użytkownikiem nowym (kod 0) czy powracającym (kod 1). Analiza rozkładu tej zmiennej wskazuje jednak, że jest ilościowa, a jej wartości oznaczają liczbę wizyt na stronie sklepu. O jej wyłączeniu z dalszych analiz zadecydowała, wspomniana wcześniej, trudność w identyfikacji adresu IP.

Podsumowanie

Etap przygotowania danych do analizy poprzedza budowę modeli prognostycznych i opisowych – zarówno tych konfirmacyjnych, jak i eksploracyjnych. Liczba czynności, które należy uwzględnić, jest dosyć duża i obejmuje m.in. zastępowanie braków danych, zastępowanie obserwacji nietypowych, redukcję liczby kategorii

zmiennych jakościowych, dyskretyzację zmiennych ilościowych, uzupełnianie zbioru danych o zmienne ze źródeł zewnętrznych, tworzenie zmiennych pochodnych, transformację zmiennych, przekształcenia zmiennych odnoszących się do czasu, redukcję liczby zmiennych czy redukcję liczby przypadków. Wszystko to sprawia, że etap ten jest bardzo czasochłonny i może zająć nawet do 80% czasu przeznaczonego na analizę danych. Jest to jednocześnie etap bardzo ważny, którego nie można pominąć, i który niestety nie zapewnia, że otrzymany model będzie wartościowy.

W artykule przedstawiono przykładowe problemy, na które napotkano podczas przygotowania danych do analizy wzorców zachowań klientów e-sklepu. Ze względu na poufność danych, autor nie mógł ujawnić nazwy sklepu, a w wielu przypadkach konieczna także była zmiana nazw opisywanych zmiennych i ich kategorii. Nie zmniejsza to jednak wartości opracowania, ponieważ opisywane przykłady są prawdziwe i wskazują na skalę zjawiska, tj. czasochłonność i szerokie spektrum procedury przygotowawczej. Wątpliwości, które pojawiły się na tym etapie analizy danych, potwierdzają konieczność ścisłej współpracy między analitykami, decydentami i osobami odpowiedzialnymi za obsługę baz danych.

Bibliografia

Berry, M.J.A., Linoff, G.S. (2000). *Mastering data mining. The art and science of customer relationship management*. New York: John Wiley & Sons.

Chapman, P. i in. (2000). *CRISP-DM 1.0. Step-by-step data mining guide*. Materiały szkoleniowe firmy SPSS.

Giudici, P. (2003). *Applied data mining. Statistical methods for business and industry.* West Sussex: John Wiley & Sons.

Łapczyński, M. (2012). Charakterystyka wybranych etapów procedury przygotowania danych do budowy modeli data mining. *Zeszyty Naukowe Uniwersytetu Ekonomicznego w Krakowie*, *884*, 129–141.

Pyle, D. (1999). *Data preparation for data mining.* San Francisco: Morgan Kaufmann.

SAS Institute Inc. (2014). *Introduction to SAS Enterprise Miner.* Pobrano z: http://support.sas.com/documentation (10.02.2017).

The Data Preparation for Analysis of E-customers – The Way from Data to Knowledge

Keywords: data preparation process, e-customers behaviour, internet usage, e-commerce

Summary. The purpose of this paper is to indicate the difficulties that one may encounter when preparing data for analysis. The presented examples refer to real data obtained from the e-shop offering footwear. They are related to web mining area that is called web usage

mining. The author presents the results of data preparation process from the moment of obtaining the data, through auditing data to building prepared information environment.

Translated by Mariusz Łapczyński

Cytowanie

Łapczyński, M. (2017). Przygotowanie danych do analizy zachowań e-klientów – droga od danych do wiedzy. *Marketing i Zarządzanie*, *3* (49), 55–65.