Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 342 · 2017

Informatyka i Ekonometria 11

Jerzy Korczak

Uniwersytet Ekonomiczny we Wrocławiu
Wydział Zarządzania, Informatyki i Finansów
Katedra Technologii Informacyjnych
jerzy.korczak@ue.wroc.pl

ICT4EDU Wrocław

Maciej Pondel

Uniwersytet Ekonomiczny we Wrocławiu
Wydział Zarządzania, Informatyki i Finansów
Katedra Inteligencji Biznesowej w Zarządzaniu
maciej.pondel@ue.wroc.pl

Unity S.A. Wrocław

METODYCZNE PODEJŚCIE DO ANALIZY I EKSPLORACJI DANYCH MARKETINGOWYCH

Streszczenie: W artykule zaproponowano metodykę realizacji projektu systemu wspomagania decyzji marketingowych z wykorzystaniem metod eksploracji danych i technologii Big Data. Inspiracją podejścia była metodyka eksploracji danych CRISP-DM, która oryginalnie nie była zorientowana na projekty Big Data. Z tego powodu metodykę tę zmodyfikowano pod kątem celu i wymagań funkcjonalnych oraz technologicznych projektowanego przez nas systemu. Główne prace badawcze w projekcie koncentrowały się na analizie i eksploracji dużych, heterogenicznych zbiorów danych o dużej zmienności. W artykule szczegółowo opisano etapy procesu realizacji projektu według rozszerzonej metodyki CRISP-DM, z uwzględnieniem specyfiki procesów analizy i eksploracji dużych baz danych marketingowych przetwarzanych w czasie rzeczywistym. W celu ilustracji podejścia podano też przykłady zadań w trakcie realizacji etapów projektu na konkretnych danych o klientach, transakcjach i produktach sklepu internetowego.

Słowa kluczowe: metodyka realizacji aplikacji informatycznych, eksploracja danych, Big Data, marketing.

JEL Classification: C55.

Wprowadzenie

Eksploracja danych jest procesem automatycznego wykrywania nietrywialnych, nieznanych, potencjalnie użytecznych zależności, reguł, wzorców, schematów, podobieństw lub trendów w dużych zbiorach danych [Witten, 2017]. Najogólniej mówiąc, zadaniem eksploracji jest analiza danych i procesów w celu lepszego ich poznania, zrozumienia i wykorzystania w procesach podejmowania decyzji. Eksploracja danych jest dziedziną multidyscyplinarną, integrującą sze-

reg obszarów badawczych, takich jak: systemy informacyjne, bazy i hurtownie danych, statystykę, sztuczną inteligencję, obliczenia równoległe, badania operacyjne, wizualizację i grafikę komputerową. Systemy eksploracji wykorzystują szeroko technologie informacyjno-komunikacyjne, technologie Web, metody wyszukiwania informacji, techniki geolokalizacji, przetwarzania sygnałów i bioinformatyki.

Głównym celem artykułu jest przedstawienie metodyki analizy i eksploracji danych marketingowych przyjętej w realizacji projektu inteligentnej platformy analizy danych dotyczących wielokanałowej sprzedaży (ang. projekt Real-Time Omnichannel Marketing – RTOM[1]). W projekcie dane są gromadzone głównie w czasie rzeczywistym i przetwarzane w ogromnych ilościach, przy dużej heterogeniczności ich źródeł, formatów, wolumenu i intensywności napływu. Użytkownik platformy (menedżer, analityk marketingu itp.) oczekuje nietrywialnej, nowej i użytecznej wiedzy, którą będzie mógł wykorzystać w procesie podejmowania decyzji. Wiedza wydobyta z zebranych danych została użyta w sposób automatyczny w procesach komunikacji z klientem tak, aby zoptymalizować wybrany parametr biznesowy procesu, np. prawdopodobieństwo zakupu, satysfakcję klienta, ryzyko odejścia klienta, marżę na produkcie i wiele innych. Projekt RTOM nie jest zatem typowym zadaniem dla większości klasycznych systemów Business Intelligence, których realizacja jest relatywnie prosta i znana [Shmueli, Patel, Bruce, 2010].

Biorąc pod uwagę złożoność projektu, jego innowacyjny charakter, a także wielość zespołów i kompetencji oraz zastosowanie nowoczesnych technologii informacyjnych, konieczne było przyjęcie jednolitej metodyki realizacji projektu. W literaturze o ile wiele napisane zostało o algorytmach eksploracji danych generujących wnikliwe analizy biznesowe, o tyle znacznie mniej znaleźć można informacji o metodyce i narzędziach eksploracji [Moutinho, 2015; Witten, 2017; Shmueli i in., 2017]. Metodyka ta wsparta oprogramowaniem powinna umożliwić zespołom bardziej skuteczną i efektywną realizację projektów korzystających z dużych baz danych w czasie rzeczywistym.

Do tej pory opracowano kilka metodyk eksploracji danych i modeli procesów, które spotkały się z różnym stopniem sukcesu w aplikacjach biznesowych [Azevedo, Santos, 2008; Moro, Laureano, Cortez, 2011; Catley i in., 2009; Wheeler, 2016]. Według Gartnera w 2015 roku aż 85% organizacji z Fortune 500 zakończyło się niepowodzeniem zastosowań technologii Big Data. Ci, któ-

[1] Projekt *Real-Time Omnichannel Marketing* (RTOM) jest realizowany przez zespół firmy Unity S.A. w ramach poddziałania RPO WD 2014-2020.

rym się powiodło, charakteryzowali się wysokim stopniem dojrzałości organizacyjnej i dobrym podejściem metodycznym [Piatetsky-Shapiro, 2014].

W ostatnich projektach eksploracji dużych baz danych zdecydowanie dominuje metodyka CRISP-DM [Shearer, 2000] opracowana przez MIT (42% zastosowań), na drugim miejscu są metodyki własne (19%), na trzecim metodyka SEMMA proponowana przez SAS (13%) [Rohanizadeh, Moghadam, 2009; Moutinho, Huarng 2015]. Metodyki pozostałe, takie jak: KDDProcess, My Organizations czy metodyki zorientowane dziedzinowo, posiadają ok. kilku procent rynku [Piatetsky-Shapiro, 2014; Azevedo, Santos, 2008][2].

Przygotowując metodykę dla projektu RTOM, wzięto pod uwagę następujące przesłanki:

1) specyfikę i złożoność projektu, w szczególności procesu eksploracji dużych baz danych w czasie rzeczywistym,
2) konieczność pragmatycznego podejścia do realizacji aplikacji zorientowanej na konkretne problemy zarzadzania sprzedażą i marketingiem,
3) dojrzałość organizacyjną i kompetencje zespołu firmy Unity S.A. w obszarze zastosowań Big Data, nowoczesnych narzędzi analitycznych i technologii informacyjnych.

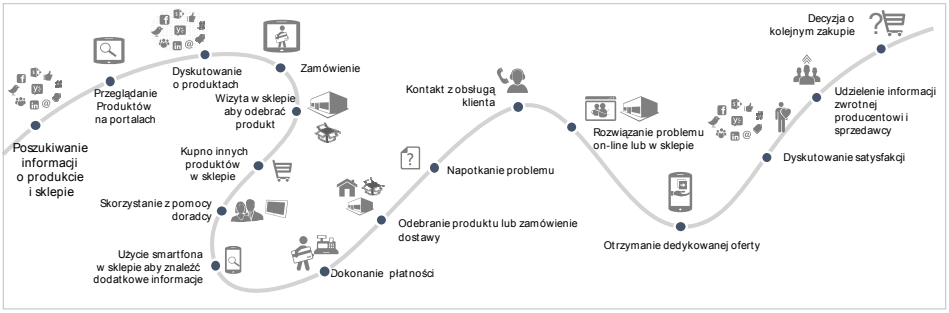
W wyniku przeprowadzonych studiów literaturowych i dyskusji wybrano jako ramę podejścia do realizacji projektu metodykę CRISP-DM. Mimo wielu zastosowań nie jest ona jednakże metodyką zorientowaną na projekty Big Data. Dlatego metodykę tę postanowiono rozszerzyć i dostosować do potrzeb oraz celu i wymagań technologicznych projektu. W następnych punktach tego rozdziału opisano szczegółowo etapy procesu realizacji projektu według rozszerzonej metodyki CRISP-DM, uwzględniającej specyfikę procesów analizy i eksploracji dużych baz danych marketingowych udostępnianych w czasie rzeczywistym.

1. Zarys projektu RTOM

Platforma Real-Time Omnichannel Marketing umożliwia zautomatyzowaną, personalizowaną analizę i eksplorację w czasie rzeczywistym danych marketingowych o kliencie i jego zachowaniu w modelu wielokanałowej sprzedaży i marketingu, z wykorzystaniem algorytmów sztucznej inteligencji i geotargetowania.

[2] Badania przeprowadzono w 2014 r. na 200 firmach, głównie z Ameryki Północnej (45%), Europy (28%) i Azji (14%) [www 1].

Podstawowe założenie strategii sprzedaży wielokanałowej opiera się na fakcie, że pojedyncza transakcja klienta może zostać przeprowadzona przy wykorzystaniu większej niż 1 liczby kanałów kontaktu klienta z dostawcą. Nie można jej mylić z podejściem wielokanałowym, które oznacza, że sprzedawca dysponuje wieloma odseparowanymi od siebie kanałami kontaktu z klientem (np. sklepy naziemne, witryna internetowa, sklep online, aplikacja mobilna, *Contact Center* i wiele innych). Podejście *omnichannel* ma na celu poprawę współpracy z klientem (ang. *customer experience*). Implementacja podejścia *omnichannel* wymaga pełnej integracji kanałów offline z tymi online na poziomie biznesowym, a także informatycznym. Aktualnie ścieżka klienta (ang. *customer journey*) angażuje różne aktywności oraz prowadzona jest w wielu kanałach komunikacji, co prezentuje rys 1. *Omnichannel* to zatem duże wyzwanie biznesowo-informatyczne, ale przede wszystkim szansa na pełne poznanie potrzeb i zachowań klientów [por. Frazer, Stiehler, 2014; Masterson, Tribby, 2009; Rigby, 2011]. W pełnej realizacji strategii pomóc muszą zatem zadania eksploracji danych oraz technologia Big Data [Marz, Warren, 2015].



Rys. 1. Ścieżka klienta

Źródło: [www 2].

Podstawowe wymagania dotyczące systemu RTOM to:

– **opracowanie zunifikowanego profilu klienta** w oparciu o koncepcję *Master Data Management* [Chorianopoulos, 2016], z implementacją różnego rodzaju referencji między danymi, dotyczącymi np.:
  - preferencji produktowych: jaki rozmiar kupuje klient, jakie kolory wybiera, jego ulubione marki produktów itp.,
  - kanałów, w których klient zamawia lub odbiera produkty,
  - czasu, kiedy zamawia (np. urodziny, okazje, początek roku szkolnego, wakacje itp.),
  - finalnego odbiorcy (czy kupuje dla siebie, partnera, małżonka lub małżonki, dziecka czy innej osoby);

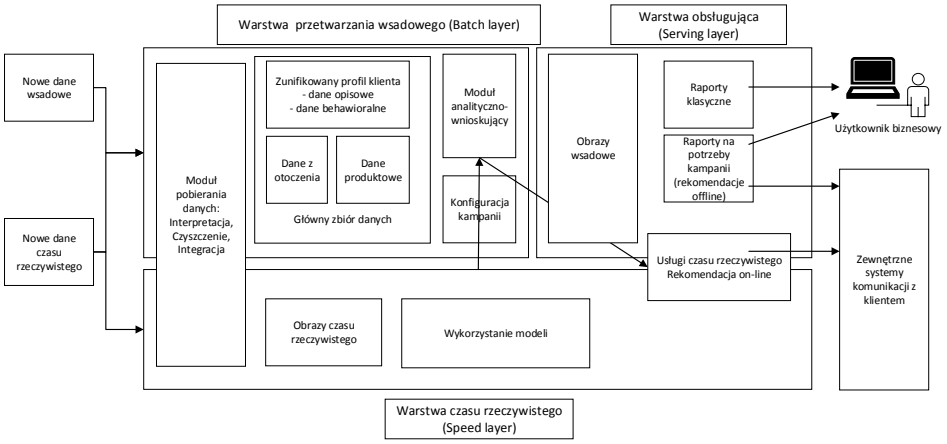
– **otwartość na nowe modele** Sztucznej Inteligencji. Dostępne na rynku mechanizmy rekomendacji bazują w większości na analizie statystycznej lub prostych regułach asocjacyjnych. W RTOM zaproponowano dodatkowo metody uczenia nienadzorowanego: w wielopoziomowych regułach asocjacyjnych, mechanizmach grupowania (ang. *clustering*), a także na zadaniach klasyfikacji (metody uczenia z nadzorem). System będzie pozwalał na implementację własnych modeli predykcyjnych;
– **możliwość analizy danych behawioralnych** pochodzących nie tylko z transakcji sklepowych, ale również z danych opisujących sposób poruszania się klienta po witrynie internetowej, wyszukiwania, filtrowania itp. oraz kanałów offline (rejestracji wizyt w sklepach, analizy danych z kanałów mobilnych itp., reklamacji, *Contact Center*). W analizie wykorzystana jest wiedza dziedzinowa dotycząca branży i charakterystyk produktów sprzedawanych przez wybraną sieć handlową, np.:
  - identyfikacja produktu poprzez jego cechy charakterystyczne a nie identyfikator (np. opis *męskie białe buty do biegania marki X rozmiaru Y* jest dla nas ważniejszy niż *produkt o id = 34…02047*),
  - dopasowanie asortymentu do czynników zewnętrznych, m.in. pory roku – ważne w przypadku np. *Kurtek*, ale w przypadku portfeli już nie, w przypadku *Koszul* może częściowo ważne,
  - uwzględnienie faktu, dla kogo przeznaczony jest produkt (*Damski*, *Męski*, *Dziecięcy*), czego nie ma np. w przypadku *RTV*.

W projekcie przewidziano również możliwość generowania wiedzy z zebranych danych w postaci:

- interaktywnych raportów, umożliwiających potwierdzenie lub odrzucenie hipotez,
- rekomendacji komunikatów marketingowych dla poszczególnych segmentów klientów wynikających z modelu predykcyjnego,
- zaleceń działań marketingowych dokonywanych w czasie rzeczywistym w stosunku do konkretnego klienta.

Biorąc pod uwagę wspomnianą wcześniej heterogeniczność źródeł danych, ogromną ilość danych oraz konieczność generowania odpowiedzi na zapytania w czasie rzeczywistym, postanowiono oprzeć system RTOM na architekturze Lambda, stanowiącej architekturę referencyjną dla skalowalnych systemów przetwarzania danych w czasie rzeczywistym [Karau i in., 2015; Marz, Warren, 2015]. Jak pokazano na rys. 2, platforma składa się z 3 warstw charakterystycznych dla architektury Lambda, mianowicie:

- warstw przetwarzania wsadowego (ang. *batch layer*) – przechowywanie danych historycznych opisujących działania klientów (zunifikowany profil klienta). Zbiór ten stanowi główną kopię zbioru danych (ang. *master dataset*), na podstawie którego przeliczane są obrazy wsadowe. Repozytorium jest zbudowane w oparciu o technologię Apache Hadoop i HDFS oraz dostępne mechanizmy odczytu danych (Hive/Impala, HBase, Cassandra, inne);
- warstwy obsługującej (ang. *serving layer*) – obrazy wsadowe umożliwiające generowanie raportów oraz wyniki wnioskowania wykonywanego przez modele predykcyjne;
- warstwy przetwarzania czasu rzeczywistego (ang. *speed layer*) – obrazy czasu rzeczywistego uzupełniające obrazy wsadowe danymi czasu rzeczywistego.



Rys. 2. Architektura systemu RTOM

Źródło: Opracowanie własne na podstawie: Marz, Warren [2015].

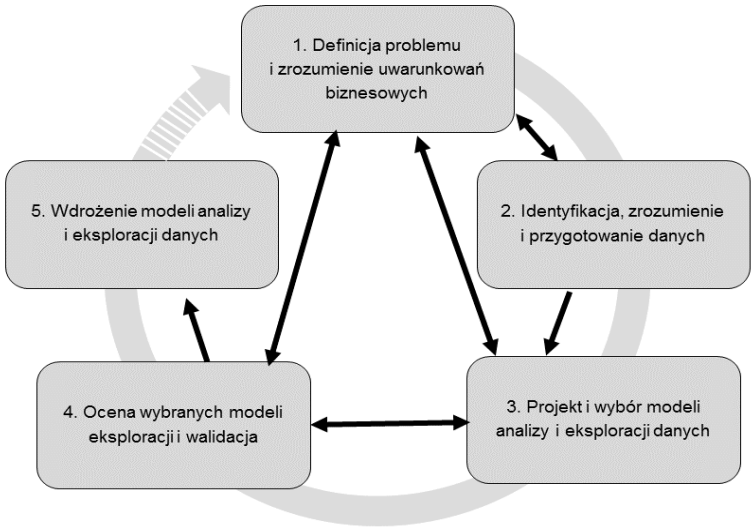
W dalszej części artykułu omówiono metodykę realizacji projektu RTOM.

2. Metodyka CRISP-DM – propozycje rozszerzeń

W projekcie i implementacji platformy RTOM wykorzystano najnowsze rozwiązania technologiczne i programowe. W odróżnieniu od większości istniejących rozwiązań w systemach CRM skoncentrowano się na analizie danych heterogenicznych, semistrukturalnych, dostępnych w czasie rzeczywistym. Wymagało to nie tylko szerokiego uwzględnienia technologii Big Data, sztucznej inteligencji, technologii mobilnych, ale i też przyjęcia oraz konsekwentnego respektowania właściwej metodyki projektowania i implementacji platformy. Jak zaznaczono, przyjęta metodyka jest w znacznym stopniu oparta na metodyce CRISP-DM.

Metodyka *Cross-Industry Standard Process for Data Mining* (CRISP-DM) zakłada, że każdy projekt eksploracji danych rozwija się w określonym cyklu życia. Rys. 3 przedstawia schemat procesu realizacji platformy RTOM według rozszerzonej metodyki CRISP-DM. Strzałki na przedstawionym diagramie pokazują zależności pomiędzy poszczególnymi etapami. Otoczenie kołem wszystkich etapów symbolizuje ciągłe dopasowywanie rozwiązań do nowych warunków otoczenia. W stosunku do oryginalnej wersji CRISP-DM, w projekcie wprowadzono kilka istotnych modyfikacji i rozszerzeń. Pierwszą z nich jest połączenie w jednym etapie, w miejsce dwóch, wszystkich prac związanych z identyfikacją problemu, zrozumieniem uwarunkowań i przygotowaniem danych. Druga istotna modyfikacja polegała na wprowadzeniu już na pierwszych etapach narzędzi wstępnego projektowania eksperymentów oraz oceny realności implementacji modeli analitycznych. Szczegółowy opis poszczególnych etapów przedstawiono w dalszej części artykułu.

Zaletą metodyki jest to, że po pięciu etapach następuje kolejna iteracja tego procesu, która pozwala na ciągłe ulepszanie i doskonalenie modeli eksploracji oraz podnosi jakość rezultatów.



Rys. 3. Etapy metodyki CRISP-DM

Źródło: Opracowanie własne.

Pierwszy etap polega na zebraniu i zrozumieniu założeń projektu z perspektywy biznesowej i wstępnym zaplanowaniu działań zmierzających do osiągnięcia celu projektu. Zrozumienie uwarunkowań biznesowych obejmuje:

- jasne sformułowanie celów i wymagań projektu w terminologii biznesowej,
- wykorzystanie sformowanych celów i ograniczeń do opracowania szczegółowej definicji problemu,
- sformułowanie wstępnych hipotez i metod ich walidacji,
- zebranie opinii i ocen zaproponowanych metod osiągniecia celów przez kadrę kierowniczą firmy, akcjonariuszy i ekspertów dziedzinowych,
- identyfikację źródeł pozyskania i zakresu niezbędnych danych,
- określenie koniecznych narzędzi i technologii informacyjnych,
- stworzenie wstępnego planu działań potrzebnych do osiągnięcia tych celów.

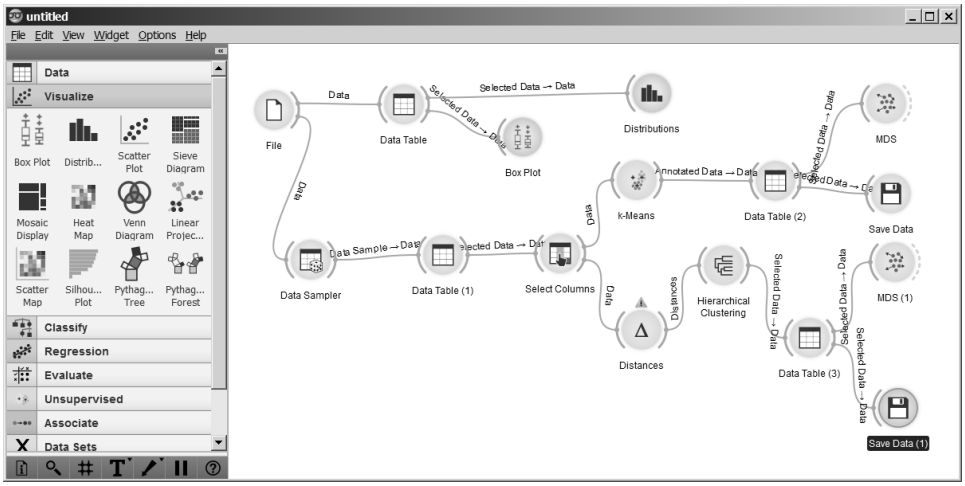
Jak wcześniej zaznaczono, istotną modyfikacją oryginalnej metodyki było wprowadzenie wstępnej walidacji sformułowanych hipotez na próbce danych źródłowych. Zadanie to było wykonane przez analityka danych przy wykorzystaniu platformy eksploracji danych Orange[3]. Prace te powinny być udokumentowane i przedstawione jako pierwsza wersja modelu wraz z opisem WE/WY (w tym określeniem zmiennych, niezbędnych normalizacji i transformacji). Kamień milowy tego etapu to opracowanie dokumentacji zawierających odpowiedzi na powyżej wymienione punkty oraz dokumentacji prototypu modelu (modeli) zrealizowanego na platformie Orange.

W celu ilustracji podejścia skorzystano z przykładu jednego z zadań rozwiązywanych na platformie RTOM – problemu klasteryzacji lub inaczej grupowania. Klasteryzacja może dotyczyć zarówno klientów, produktów, transakcji, jak i kontaktów klientów ze stronami WWW. Dla przykładu, w bazie danych systemu istnieje kilka tysięcy klientów, każdy opisany przez kilkadziesiąt atrybutów o różnym stopniu znaczenia. Celem klasteryzacji jest wyszukanie skupień – inaczej klastrów podobnych klientów, do których możemy się zwrócić z ofertą lub promocją określonych produktów. Wymaga się, aby otrzymane klastry charakteryzowały się określonymi właściwościami statystycznymi (jak np. minimalną wariancją) oraz użytecznością w procesie podejmowania decyzji marketingowych (np. przy określeniu grupy lojalnych klientów). Oczekuje się, że

[3] Platforma Orange jest łatwym do opanowania narzędziem eksploracji danych z bogatym interfejsem graficznym i licznymi funkcjami (ang. *widgets*) analizy danych, klasyfikacji, klasteryzacji i predykcji. Zaproponowana idea wizualnego projektowania procesu eksploracji wraz z możliwością rozbudowy funkcji w języku Python sprawia, że Orange jest narzędziem bardzo często stosowanym przez analityków. Więcej informacji o Orange można znaleźć na stronach Uniwersytetu w Lubljanie [www 3].

dzięki klasteryzacji osiągnie się lepiej adresowaną i bardziej efektywną promocję produktów sklepu, wyrażoną konkretnie we wskaźnikach rentowności sprzedaży. Na tym etapie zdefiniowano też źródła danych; w naszym przypadku są to systemy transakcyjne, CRM, dane geolokalizacyjne, dane sieci społecznościowych i logi serwisów internetowych sklepu.

W pracach tego etapu niezwykle ważnym zadaniem jest sformułowanie wstępnych hipotez oraz zebranie opinii i ocen zaproponowanych metod osiągniecia celów przez kadrę kierowniczą firmy, akcjonariuszy i ekspertów dziedzinowych. Innowacyjne metodycznie jest tu opracowanie prototypu modelu i przeprowadzenie wstępnej walidacji na uproszczonym przykładzie, przy wykorzystaniu łatwego narzędzia eksploracji danych. Jednym z takich narzędzi jest ogólnodostępna platforma Orange. Schemat procesu klasteryzacji pokazany jest na rys. 4.



Rys. 4. Schemat procesu klasteryzacji

Źródło: Opracowanie własne z wykorzystaniem platformy Orange.

Otrzymane wyniki wraz z ilustracją klastrów umożliwiły nie tylko lepsze zrozumienie problemu oraz uściślenie celów i ograniczeń biznesowych, ale też pozwoliły na dokonanie wstępnej walidacji modeli analizy i eksploracji danych.

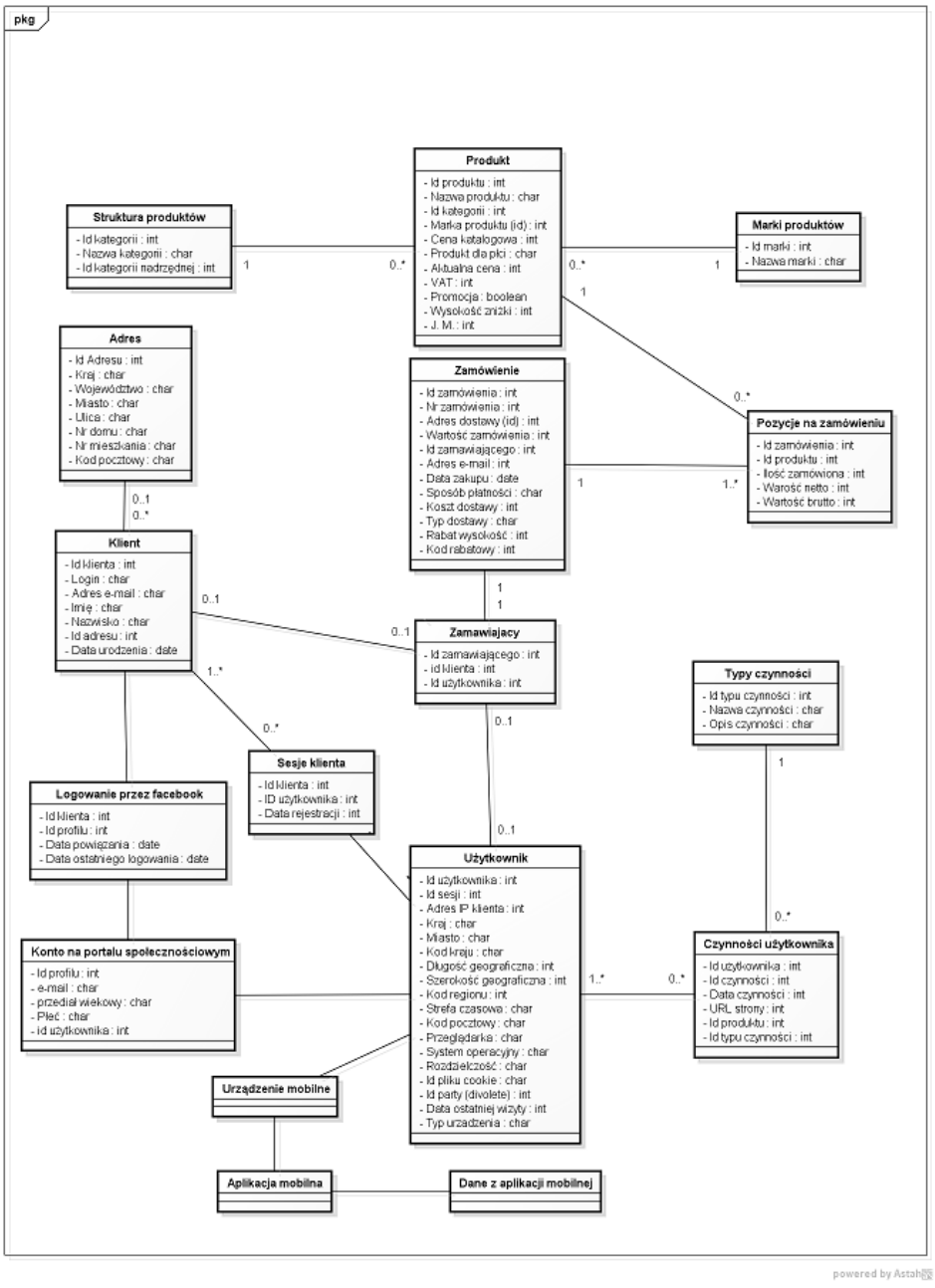
Drugi etap obejmuje prace związane ze identyfikacją, zrozumieniem i przygotowaniem danych. Jak wcześniej zaznaczono, w projekcie, w porównaniu z oryginalna wersją CRISP-DM, połączono dwa etapy: zrozumienia i przygotowania danych. Ze wszystkich etapów jest on najbardziej iteracyjny i kosztowny czasowo. Głównym zadaniem jest zebranie danych oraz ich wstępne przetworzenie pod kątem narzędzi i algorytmów eksploracji. W kontekście technologii Big Data dane są zbierane w tzw. *data sandboxes*. Technologicznie *data*

sandbox składa się z masowo równoległych procesorów, obszernej pamięci i mechanizmów WE/WY zapewniających skalowalność procesu zbierania danych i niezależność od operacyjnych systemów baz danych [White, 2015]. Dzięki temu *sandbox* zapewnia możliwość przeprowadzania złożonych analiz danych bez interrupcji działania systemów informacyjnych firmy. Zbierane dane mogą być różnego rodzaju: pochodzić z systemów transakcyjnych, urządzeń mobilnych, kostek systemów OLAP, logów telefonicznych, logów Web i Internetu. Szacuje się, że rozmiar *data sandbox* może przekraczać dziesięciokrotnie wielkość hurtowni danych firmy. Należy zaznaczyć, że *data sandbox* jest współdzielony przez analityków projektu i moduły eksploracji, przy czym wymaga się równocześnie, aby platforma *sandbox* zapewniła bezpieczeństwo i poufność danych.

Drugim ważnym zadaniem tego etapu jest przygotowanie i transformacja danych według schematu ELT (ang. *Extract–Load–Transform*). Korzyścią ELT jest zachowanie danych w ich oryginalnej postaci w bazie danych. Wówczas analityk może je dowolnie przekształcać bądź pozostawić w niezmienionej formie. W tym zadaniu należy także zbadać jakość zbieranych danych i przedstawić statystycznie użyteczne miary. Ostatnim zadaniem jest organizacja i projekt procesu transformacji surowych danych. Wśród typowych operacji transformacji można wymienić procesy analizy zmiennych, filtrację, normalizację danych, uzupełnienia informacji brakujących itp.

W projekcie platforma *data sandbox* działa pod systemem Linux z wykorzystaniem technologii bazodanowych NOSQL dostępnych na platformie Hadoop oraz przetwarzaniem zgodnym z paradygmatem MapReduce w silniku przetwarzania Spark [Ryza i in., 2015]. Kamienie milowe etapu to opracowanie dokumentacji technicznej i utworzenie *sandbox* dla RTOM. Dla przykładu, w projekcie RTOM głównym źródłem danych jest system transakcyjny sklepu i logi kontaktów klientów z aplikacją internetową. Schemat bazy danych ilustruje rys. 5.

Oprócz danych transakcyjnych do *data sandboxa* ściągane są informacje ze wszystkich kanałów kontaktu z klientem. Są to m.in. dane geolokalizacyjne klientów czy dane opisujące aktywności klientów w mediach społecznościowych.



Rys. 5. Schemat konceptualny bazy danych

Źródło: Opracowanie własne.

Trzeci etap procesu koncentruje się na projekcie i wyborze modelu eksploracji danych. O ile w poprzednim etapie przygotowania danych położono większy nacisk na jakość danych, to w tym etapie głównie modeluje się zależności między zmiennymi w obszarze określonych problemów biznesowych. Skorzystano tu z dokumentacji wstępnych wersji modelu (modeli) wcześniej przygotowanych na platformie Orange. Bezcenny jest tutaj udział ekspertów dziedzinowych, którzy są w stanie podpowiedzieć zmienne mogące mieć wpływ na rozwiązanie problemu oraz na przyjęcie lub odrzucenie zdefiniowanych w pierwszym etapie hipotez. W szczególności może to dotyczyć rozróżnienia w interpretacji związków korelacyjnych i przyczynowo-skutkowych.

Wybór zmiennych ma istotne znaczenie dla jakości eksploracji. Analityk musi być otwarty na wykorzystanie różnych algorytmów eksploracji, ich parametryzacje i konstrukcje wektora wejściowego. Wybór wektora wejściowego i modelu eksploracji jest procesem iteracyjnym. Testowanie modelu na wszystkich możliwych zmiennych jest z reguły niepraktyczne. W celu redukcji wymiarowości przestrzeni analityk może tu posłużyć się wiedzą ekspertów, którzy zasugerują istotne zmienne lub skorzystać z algorytmów rangujących zmienne, według kryteriów takich jak: indeks Gini, zysk informacyjny, ANOVA, wskaźnik redukcji entropii.

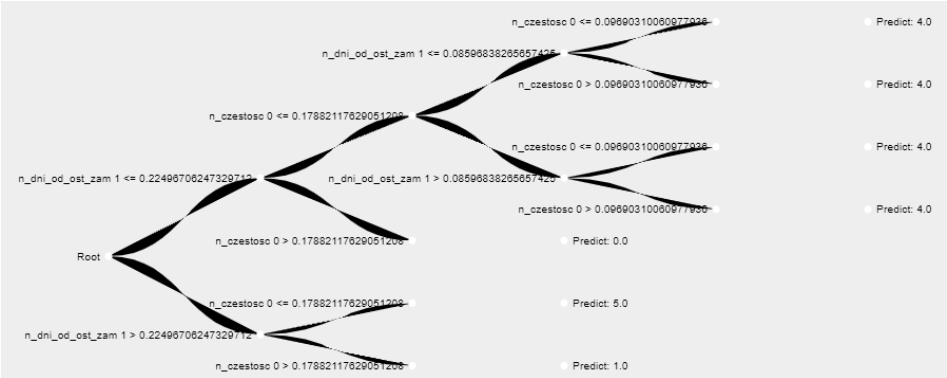
Modeli eksploracji danych jest wiele. Generalnie należą one do trzech kategorii: klasyfikacji, predykcji i klasteryzacji[4]. W projekcie RTOM ograniczono ofertę do modeli dostępnych w bibliotekach Apache Mahout [www 4], MLlib Sparka [www 5], Tensorflow Core [www 6] i Pandas [www 7] (zob. także: Laserson, 2015; Owen i in., 2012).

Dla przykładu realizacji prac wykorzystano modele klasteryzacji dostępne w bibliotekach Apache Mahout, MLlib Sparka i Tensorflow Core [IBM, 2011; Ryza i in., 2015], z których wybrano w przykładzie m.in. algorytm *k-means* [Witten, 2017; Marz, Warren, 2015]. Otrzymane klastry stanowiące grupy klientów zinterpretowano drzewem decyzyjnym (rys. 6), korzystając z biblioteki *pyspark.mllib.tree*.

Obrazując rozważania, można zauważyć, że klaster 5 zawiera klientów, którzy dokonują zakupów często, ale ich ostatni zakup był dokonany relatywnie dawno. Menedżer może wykorzystać tę informację do przygotowania kampanii reklamowej do takich klientów celem zachęcenia ich do powrotu do zakupów

[4] Klasyfikacja i predykcja są bardzo podobne i na ogół wiążą się z typem wykorzystywanych danych do budowy modelu. Jeśli atrybut decyzyjny jest kategoryczny, wówczas problem predykcji wartości takiego atrybutu jest przedstawiany jako problem klasyfikacji. Jeśli atrybut decyzyjny jest ciągły (numeryczny), problem jest zwany problemem predykcji.

w danej sieci handlowej. Kamieniem milowym etapu jest opracowanie dokumentacji modeli eksploracji, zestawu modeli, wraz z określeniem danych wykorzystanych w procesie uczenia, testowania i walidacji.



Rys. 6. Fragment drzewa decyzyjnego

Źródło: Opracowanie własne.

Przedmiotem **czwartego etapu** jest ocena jakości wybranych modeli eksploracji. Warunkiem niezbędnym realizacji zadania jest wyraźne określenie kryteriów ewaluacji. Na ogół zadanie oceny jest problemem wielokryterialnym [Shmueli i in., 2017]. Zdarza się jednak często, że menedżerowie, akcjonariusze i eksperci nakładają inne priorytety na przedstawione formalne kryteria ewaluacji modeli.

Generalnie modele powinny być ocenione pod względem jakości i efektywności jeszcze przed wdrożeniem na próbce danych z *data sandbox*. Zaleca się tu dwustopniowe testowanie modeli, mianowicie: najpierw na pilotowej próbie, później na pełnym materiale informacyjnym. Dzięki temu ogranicza się koszty/czas modyfikacji modelu wynikające czasami z prostych błędów czy niedopatrzeń, tym samym zmniejsza się ryzyko związane testowaniem i walidacją wersji produkcyjnej platformy. Wskazane jest też stopniowe rozszerzanie zakresu oceny, np. do grupy towarów, wybranych kanałów sprzedaży czy obszaru rynku.

Bardzo ważnym zadaniem jest przygotowanie danych do budowy i oceny modelu (uczenia modelu, testowania i walidacji). Przed uruchomieniem modelu na całym rzeczywistym materiale informacyjnym zaleca się przeprowadzenie oceny anomalii w danych wejściowych w trakcie pobierania ich przez model. Operacja ta pozwala na podniesienie jakości wyników oraz sformułowanie ewentualnych rekomendacji eksploatacyjnych odnośnie funkcjonowania modelu w warunkach rzeczywistych. Działający model ocenia się nie tylko pod względem jakości i efektywności, ale też współpracy z innymi zasobami platformy.

Jakość wybranych modeli oceniana jest według ustalonych kryteriów biznesowych i ogólnie przyjętych metryk oceny dla poszczególnych kategorii modeli eksploracji. W omawianym przykładzie oceniono zaproponowane modele klasteryzacji. Ogólnie, miary ewaluacji można podzielić na dwie kategorie: oceny wewnętrznej wyników klasteryzacji i oceny bazującej na kryteriach zewnętrznych.

W przypadku zastosowania wielu algorytmów klasteryzacji przy kryteriach wewnętrznych ocenia się hierarchię klastrów, biorąc pod uwagę podobieństwo instancji wewnątrz klastrów i podobieństwo pomiędzy klastrami. Wśród miar oceny stosowane są następujące [Witten, 2017; Shmueli, Patel, Bruce, 2010]:

– wskaźnik Daviesa-Boudina:

$$DB = 0.5n \sum max ((\delta_i + \delta_j) / d(c_i,c_j),$$

gdzie n oznacza liczbę klastrów, c_i i c_j centrody klastrów, δ_i i δ_j średnie odległości d między elemetami danego klastra i centroidem.

Algorytm generujący najmniejszą wartość wskaxnika *DB* jest uważany za najlepszy według kryterium oceny wewnętrznej;

– wskaźnik Dunna:

$$D = min (d(i,j) / max\ d'(k),$$

gdzie *d(i,j)* oznacza odległośc między klastrami *i* i *j*, zaś *d'(k)* miarę odległości wewnątrz klastra *k*.

Wskaźnik Dunna koncentruje się na gęstości klastrów i odległości między klastrami. Algorytmy preferowane według wskaźnika Dunna to te, które osiągają wysokie wartości wskaźnika.

W metodach ewaluacji według kryteriów zewnętrznych wyniki klasteryzacji oceniane są przy wykorzystaniu danych zewnętrznych, niebranych pod uwagę w procesie klasteryzacji. Takim danymi są na przykład klienci, których przynależność do klastra jest oznaczona wcześniej przez ekspertów. Wówczas ocena klasteryzacji wynika z porównania zawartości klastrów oznaczonych przez ekspertów z klastrami utworzonymi przez algorytm. Wśród stosowanych miar należy wymienić:

– wskaźnik jednorodności klastrów obliczany według wzoru:

$$WJK = 1/N \sum max\ |m \cup d|,$$

gdzie *M* oznacza liczbę klastrów utworzonych przez algorytm, *D* liczbę klas eksperta;

– wskaźnik Jaccarda, który mierzy podobieństwo między dwoma zbiorami obserwacji według nastepującego wzoru:

$$WJ = TP / (TP + FP + FN).$$

W przypadku dwóch identycznych zbiorów $WJ = 1$;

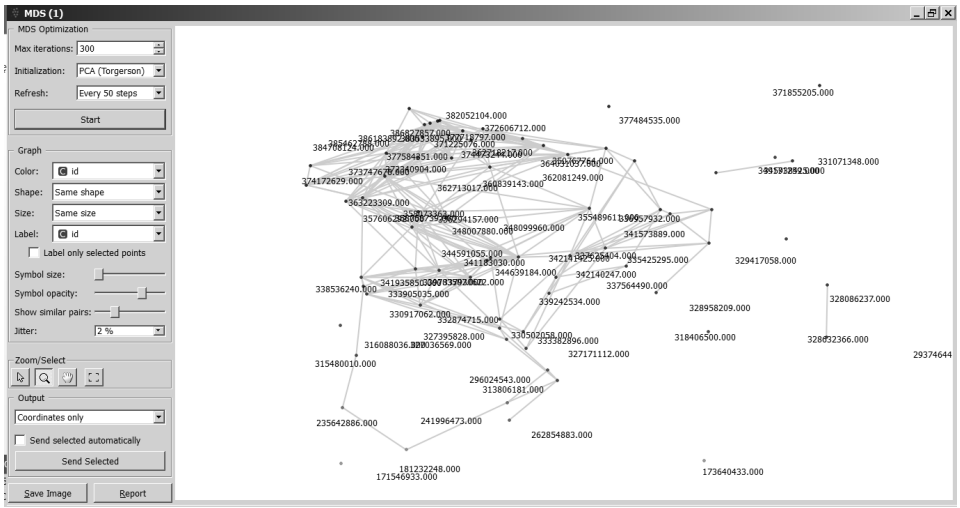
– wskaźnik Randa, obliczany według wzoru:

$$WR = (TP + TN) / (TP + FP + FN + FN.$$

Wskaźnik Randa, jak i poprzednie, oparty jest na porównaniu z benchmarkiem podanym przez eksperta. Generalnie, informuje on o podobieństwie oceny prawidłowych decyzji między wynikami algorytmu klasteryzacji a benchmarkiem.

Oprócz podanych miar oceny klasteryzacji stosuje się też inne wskaźniki, takie jak F-score, wskaźnik Fowkesa-Mallowsa i inne.

Analitycy marketingu często obrazują wyniki klasteryzycji w postaci projekcji przestrzeni wielowymiarowej MDS (ang. *Multi-Dimensional Scaling*), co pokazano na rys. 7. Diagram MDS pozwala na nie tylko na łatwą ocenę wizualną skupisk i ich rozproszenie, ale też wskazuje obiekty nietypowe.



Rys. 7. Wielowymiarowa reprezentacja klastrów

Źródło: Opracowanie własne.

Wymienione miary pozwalają na ustalenie, czy wybrane modele spełniają wszystkie założone wymagania biznesowe i postawione hipotezy, zdefiniowane w pierwszym etapie metodyki. W przypadku dokonania pozytywnej oceny przez menedżera, akcjonariuszy i analityków można podjąć decyzję o wdrożeniu i upowszechnienia modelu. Kamieniem milowym jest opracowanie raportu ewaluacji modeli eksploracji zawierającego podane powyżej wartości miar i wskaźników.

Ostatnim etapem metodyki jest wdrożenie pozytywnie ocenionych modeli eksploracji. Realizacja tego etapu przebiega dwustopniowo. Najpierw wdrożona jest tzw. pilotowa wersja platformy w rzeczywistym środowisku produkcyjnym i oceniane są wyniki pod względem merytorycznym, użytkowym i efektywnościowym. Generowane raporty oceniane są przez menedżerów i analityków biznesowych pod kątem ich poprawności, kompletności oraz użyteczności w podejmowaniu decyzji. Równocześnie działanie platformy jest monitorowane przez projektantów i przyszłych administratorów systemu. Monitorowanie dotyczy głównie sprawności obliczeniowej i stopnia wykorzystania zasobów pamięci. Wcześniejsza walidacja wersji pilotowej aplikacji pozwala na ograniczenie ryzyka niepowodzenia przy uruchomieniu pełnej wersji, a także na współdziałanie ze wszystkimi komponentami systemu informacyjnego firmy. Umożliwia też dokonanie drobnych adjustacji i dostrojenia przed wdrożeniem pełnej wersji platformy.

W drugim stopniu tego etapu uruchomiona jest aplikacja w pełnym środowisku produkcyjnym. Wyniki działania są upowszechniane użytkownikom, przy czym często równocześnie wymagane są dodatkowe szkolenia, zdefiniowanie nowych ról organizacyjnych czy zatrudnienie nowych specjalistów. Zaznaczyć należy, że nowe rozwiązania biznesowe i technologiczne rewolucjonizują na ogół dotychczasowe praktyki i procesy podejmowania decyzji.

Proces doskonalenia systemów i procesów decyzyjnych nigdy się nie kończy. Wraz z postępem rozwijają się nowe technologie informacyjno-komunikacyjne, doskonalą się metody eksploracji danych, zmieniają się dane i źródła informacji. Stąd już po wdrożeniu należy myśleć o rozwoju i planować przyszłe aktualizacje i rozszerzenia platformy. Na rys. 1 dalszy rozwój systemu ilustruje przerywana strzałka prowadząca do pierwszego etapu procesu.

Główne kamienie milowe etapu to:
- plan wdrożenia aplikacji i upowszechniania wyników,
- plan monitorowania i utrzymania aplikacji,
- opracowanie ostatecznego raportu i dokumentacji systemu.

3. Kilka uwag o Big Data w kontekście zaproponowanej metodyki eksploracji danych

Zaproponowana metodyka została przedstawiona w kontekście zadań projektu RTOM z uwzględnieniem technologii Big Data i przetwarzania w czasie rzeczywistym danych biznesowych. Opisane etapy eksploracji pokazały, że nie jest to podejście stosowane w rozwiązaniach analityki biznesowej systemów ty-

pu Business Intelligence. W tych systemach, mimo pozornego podobieństwa, nie ma się do czynienia z ogromnymi strumieniami danych napływającymi w czasie rzeczywistym [Marz, Warren, 2015]. Nie ma się też do rozwiązania problemów technologicznych związanych ze skalowalnością oraz heterogenicznością źródeł i danych. Problem integracji różnych komponentów oprogramowania i sprawności procesów eksploracji jest również mniej istotny. Te aspekty starano się wyeksponować w metodyce przyjętej dla realizacji platformy RTOM.

Podsumowując dotychczasowe doświadczenia, należy zwrócić uwagę na kilka kwestii kluczowych dla rozwoju platformy RTOM, mianowicie:

- **Jakość danych a wielkość wolumenu danych.** Badania pokazują, że wraz ze zwiększeniem strumienia danych pochodzących z różnych źródeł pogarsza się ich jakość. Dlatego w projekcie RTOM niesłychanie ważne są procesy zbierania i przygotowania danych. Jakość danych decyduje o jakości modeli eksploracji, o użyteczności generowanych wyników. W szczególności dotyczy to procesów czyszczenia, filtracji zakłóceń i szumów oraz algorytmów uzupełniania brakujących informacji danych zgromadzonych w *data sandboxie*.
- **Dostępność modeli.** Większość algorytmów i modeli eksploracji jest dzisiaj ogólnie dostępna w bibliotekach; podaliśmy w opracowaniu kilka referencji. Nie ma zatem potrzeby przedstawienia pełnej specyfikacji modeli. Ważniejsze są zatem dla użytkownika opisy profili algorytmów z ich parametryzacją oraz interfejsem do innych użytecznych komponentów platformy RTOM, na przykład związanych z oceną modeli czy wizualizacją danych i wyników.
- **Hadoop**, open source'owy produkt Apache, nie jest platformą eksploracji danych; jest jednym z narzędzi zarządzania i operowania na bardzo dużych zbiorach danych [White, 2015]. Niewątpliwie komponenty Hadoop, MapReduce czy HDFS system usprawniają proces działania na dużych, rozproszonych zbiorach danych [www 4]. Zaznaczyć należy jednakże, że Hadoop sprawdza się na problemach liniowych, lecz *gros* aplikacji biznesowych to problemy nieliniowe. Dlatego w metodyce sięgnięto do m.in. do Apache Mahout oraz Apache Spark MLlib, które zapewniają efektywną eksplorację danych z wykorzystaniem Hadoop.
- **Interpretacja wyników i ich wykorzystanie w podejmowaniu decyzji.** Wiele z podanych modeli jest ocenianych pod kątem jakości, dokładności i sprawności działania. Natomiast w aplikacjach biznesowych zwraca się główną uwagę na kryteria ekonomiczne związane z kosztem oraz konkretnymi efektami mierzalnymi i niemierzalnymi danego modelu. Oprócz wymienionych, dla menedżera ważne są też takie cechy modeli, jak łatwość zrozumienia ich działania i interpretowalność wyników.

Podsumowanie

W artykule przedstawiono propozycję rozszerzenia metodyki CRISP-DM na aplikację analityczno-decyzyjną w obszarze marketingu, operującą w czasie rzeczywistym na dużych heterogenicznych zbiorach danych. Metodyka ta została wykorzystana w realizacji projektu naukowo-badawczego RTOM. RTOM umożliwia zautomatyzowaną, personalizowaną analizę danych, klasyfikację i predykcję, opartą na gromadzeniu i przetwarzaniu danych empirycznych o kliencie i produktach w modelu wielokanałowej sprzedaży, z wykorzystaniem algorytmów sztucznej inteligencji i geotargetowania.

W celu zilustrowania metodyki pokazano przykład klasteryzacji danych o kliencie. W projekcie zaimplementowano wiele algorytmów klasyfikacji, klasteryzacji oraz wyszukiwania wzorców. Z uwagi na przeznaczenie systemu dokonano adaptacji podstawowych modeli analizy i eksploracji danych, które zostały uznane przez ekspertów jako użyteczne i interesujące w obszarze marketingu. Większość modeli pochodzi z bibliotek Apache Mahout, MLlib Sparka, Tensorflow Core i Pandas. RTOM jest platformą otwartą, tym samym może być funkcjonalnie rozbudowana pod kątem potrzeb konkretnego podmiotu handlowego stosującego podejście *omnichannel*. Dotychczasowe doświadczenia praktykowania metodyki są pozytywne – sprzyja ona efektywnej i innowacyjnej współpracy pracy menedżerów marketingu, analityków i informatyków.

Literatura

Azevedo A., Santos M.F. (2008), *KDD, SEMMA and CRISP-DM: A Parallel Overview* [w:] Proceedings of the IADIS European Conference on Data Mining, s. 182-185.

Catley C., Smith K., McGregor C., Tracy M. (2009), *Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study* [w:] Computer-Based Medical Systems, 22nd IEEE International Symposion on CBMS, s. 1-5.

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

Frazer M., Stiehler B.E. (2014), *Omnichannel Retailing: The Merging of the Online and Offline Environment* [w:] Proceedings of the Global Conference on Business and Finance, Vol. 9, No. 1, s. 655-657.

IBM (2011), *Introducing Apache Mahout*, www.ibm.com (dostęp: 15.02.2017).

Karau H., Konwinski A., Wendell P., Zaharia M. (2015), *Learning Spark: Lightning--Fast Big Data Analysis*, O'Reilly, Sebastopol.

Laserson U., Owen S., Wills J. (2015), *Analytics with Spark: Patterns for Learning from Data at Scale*, O'Reilly, Sebastopol.

Marz N., Warren J. (2015), *Big Data: Principles and Best Practices of Scalable Real-time Data Systems*, Manning Publishing, New York.

Masterson M., Tribby M. (2009), *Changing the Channel: 12 Easy Ways to Make Millions for Your Business*, John Wiley & Sons, Chichester.

Moro S., Laureano R., Cortez P. (2011), *Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology* [w:] Proceedings of European Simulation and Modelling Conference ESM'2011, s. 117-121.

Moutinho L., Huarng K. (2015), *Quantitative Modelling in Marketing and Management*, World Scientific Publishing, Singapore.

Owen S., Anik R., Dunning T., Friedman E. (2012), *Mahout in Action*, Manning Publishing, New York.

Piatetsky-Shapiro G. (2014), *KDNuggets Poll: Data Mining Methodology*, http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html (dostęp: 20.02.2017).

Rigby D. (2011), *The Future of Shopping*, Harvard Business Review, https://hbr.org/2011/12/the-future-of-shopping (dostęp: 18.02.2017).

Rohanizadeh S.S., Moghadam M.B. (2009), *A Proposed Data Mining Methodology and its Application to Industrial Procedures*, "Journal of Industrial Engineering", Vol. 4(1), s. 37-50.

Ryza S., Laserson U., Owen S., Wills J. (2015), *Advanced Analytics with Spark: Patterns for Learning from Data at Scale*, O'Reilly, Sebastopol.

Shearer C. (2000), *The CRISP-DM Model: The New Blueprint for Data Mining*, "Journal of Data Warehousing", Vol. 5, s. 13-22.

Shmueli G., Bruce P., Stephens M., Patel N. (2017), *Data Mining for Business Analytics*, John Wiley & Sons, Chichester.

Shmueli G., Patel N., Bruce P. (2010) *Data Mining for Business Intelligence*, John Wiley & Sons, Chichester.

Wheeler S.R. (2016) *Architecting Experience: A Marketing Science and Digital Analytics Handbook*, World Scientific Publishing, Singapore.

White T. (2015), *Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale*, O'Reilly, Sebastopol.

Witten I., Frank E., Hall M., Pal C. (2017), *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Amsterdam.

[www 1] http://www.kdnuggets.com/2014/10/new-poll-methodology-analytics-data-mining-data-science.html (dostęp: 21.02.2017).

[www 2] https://www.slideshare.net/MicrosoftAT/digital-transformation-book-of-dreams (dostęp: 10.02.2017).

[www 3] http://orange.biolab.sl (dostęp: 21.02.2017).

[www 4] http://mahout.apache.org/users/basics/algorithms.html (dostęp: 21.02.2017).

[www 5] http://spark.apache.org/docs/latest/ml-guide.html (dostęp: 21.02.2017).

[www 6] http//www.tensorflow.org/ (dostęp: 21.02.2017).

[www 7] http://pandas.pydata.org/ (dostęp: 21.02.2017).

A METHODOLOGICAL APPROACH TO ANALYSIS AND EXPLORATION OF MARKETING DATA

Summary: The article proposes a methodology for development of a marketing Decision Support System using data mining methods and Big Data technologies. The main research findings focus on the analysis and exploration of very large, heterogeneous sets of highly volatile marketing data. The approach is inspired by the CRISP-DM methodology which is not oriented towards Big Data applications. The article describes in detail the stages of the project development according to the extended CRISP-DM methodology, taking into account the specificity of the analysis and exploration processes of large marketing databases processed in real time. In order to illustrate the approach, the examples based on real data about customers, transactions and products of the Internet store were discussed.

Keywords: methodology of development of IT applications, data mining, Big Data, marketing.