GDAŃSK UNIVERSITY OF TECHNOLOGY

Eksploracja danych

PROCES EKSPLORACJI DANYCH

Wojciech Waloszek

wowal@eti.pg.gda.pl

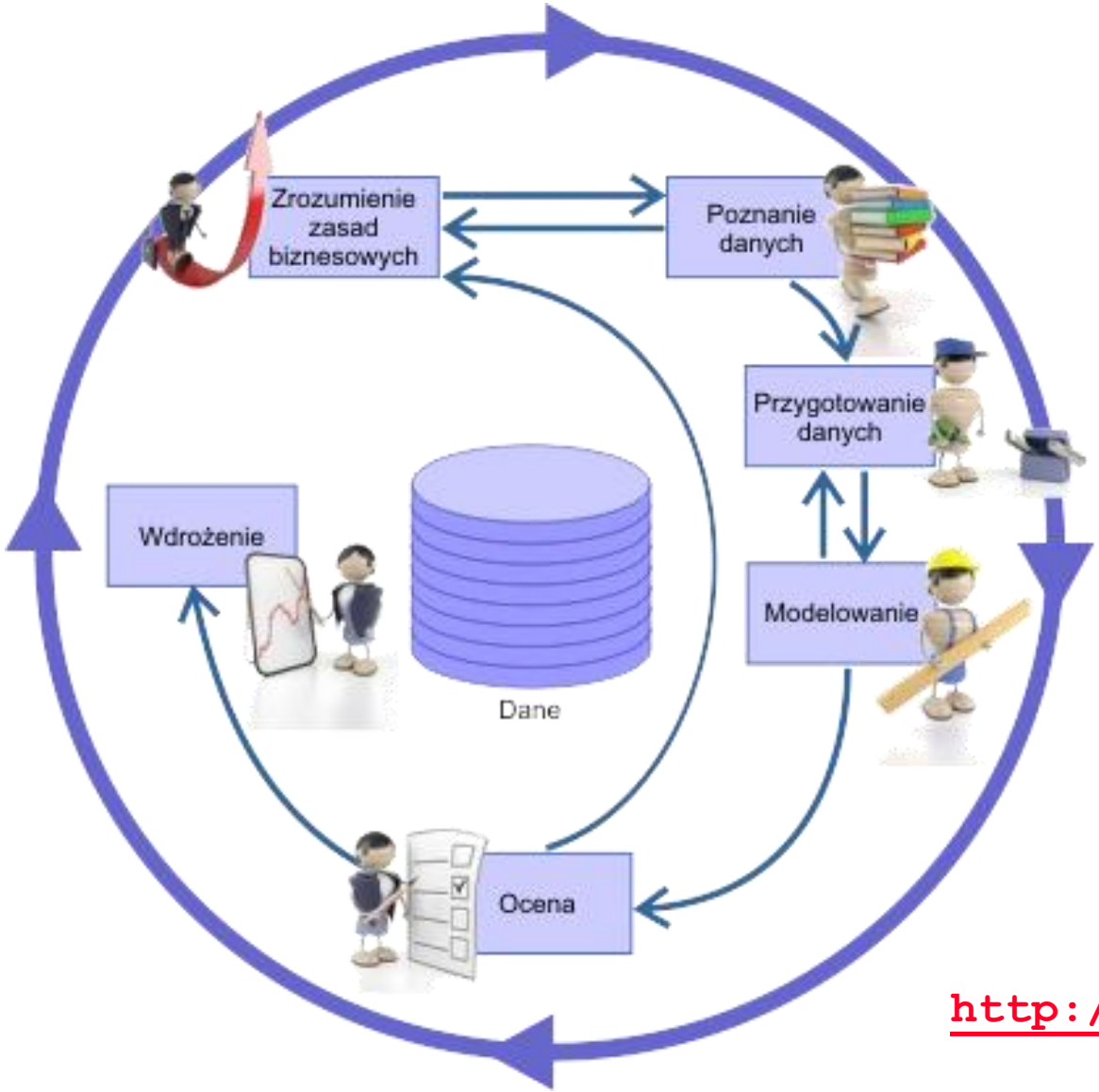
Teresa Zawadzka

tegra@eti.pg.gda.pl

Katedra Inżynierii Oprogramowania
Wydział Elektroniki, Telekomunikacji i Informatyki
Politechnika Gdańska

GDAŃSK UNIVERSITY OF TECHNOLOGY

Proces eksploracji danych (*Cross Industry Standard Process for Data Mining*)



http://www.crisp-dm.org

GDAŃSK UNIVERSITY OF TECHNOLOGY

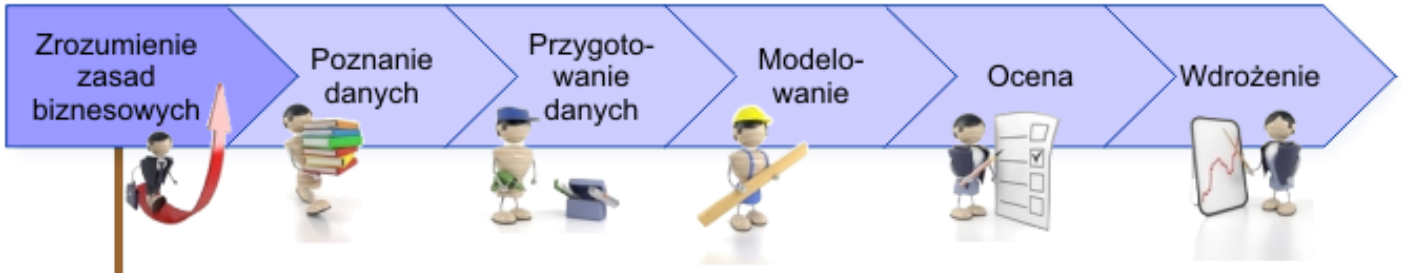
Zrozumienie zasad biznesowych (ang. *Business Understanding*)

Zrozumienie celów i wymagań projektu z perspektywy biznesowej, przekształcenie tej wiedzy w definicje problemów eksploracji danych i wstępne zaplanowanie prac w celu osiągnięcia wytyczonych celów.

1. *Jakie informacje mają być wykryte przez model eksploracji danych?*
2. *Czy otrzymane dane będą używane do przewidywania przyszłości czy wyłącznie do analizowania ukrytych zależności pomiędzy historycznymi danymi?*
3. *Jakie informacje i z jakim wyprzedzeniem model ma przewidywać?*
4. *Jakie zależności między danymi model ma wykrywać?*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zrozumienie zasad biznesowych



- Zidentyfiko-wanie celów bizneso-wych
  - Tło
  - Cele biznesowe
  - Kryteria sukcesu biznesowego
- Zrozumienie sytuacji
  - Dostępność zasobów
  - Wymagania, założenia i ograniczenia
  - Ryzyko
  - Terminologia
  - Koszty i zyski
- Określenie celów eksploracji danych
  - Cele eksploracji danych
  - Kryteria sukcesu eksploracji
- Wytworze-nie planu projektu
  - Plan projektu
  - Wybór narzędzi i technik

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zrozumienie zasad biznesowych

Raport ze zrozumienia zasad biznesowych (ang. *Business Understanding Report*)

GDAŃSK UNIVERSITY OF TECHNOLOGY

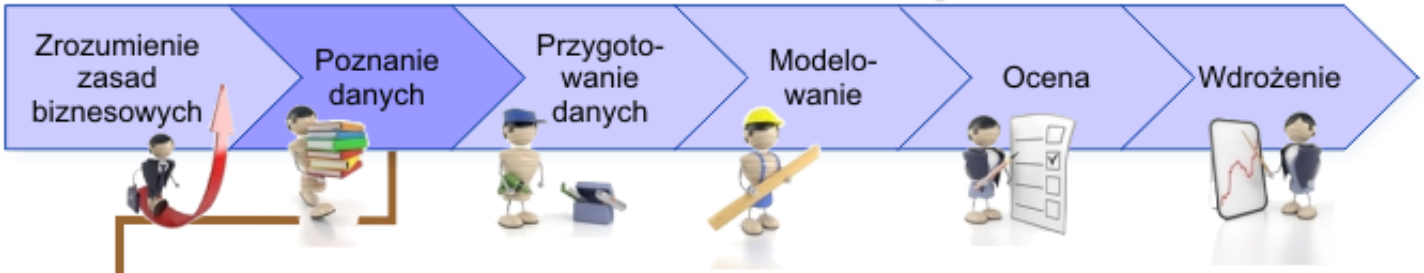
Poznanie danych (ang. *Data understanding*)

Wstępne zgromadzenie danych i wykonanie czynności niezbędnych do zapoznania z danymi, identyfikacji problemów z jakością danych, zebrania pierwszych spostrzeżeń dotyczących danych oraz wykrycia interesujących podzbiorów w celu sformułowania hipotez dotyczących ukrytych informacji.

1. *Jaka jest charakterystyka danych źródłowych?*
2. *Jakie są zależności między danymi źródłowymi?*
3. *Jaka jest jakość danych?*
4. *Na ile dokładnie dane źródłowe opisują proces biznesowy?*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Poznanie danych



- Zebranie wstępnych danych → Wstępny raport o danych
- Opisanie danych → Raport opisujący dane
- Eksploracja danych → Raport z eksploracji danych
- Weryfikacja jakości danych → Raport o jakości danych

GDAŃSK UNIVERSITY OF TECHNOLOGY

Poznanie danych

- Raport danych początkowych (ang. *Initial Data Collection Report*)
- Raport z opisu danych (ang. *Data Description Report*)
- Raport z eksploracji danych (ang. *Data Exploration Report*)
- Raport jakości danych (ang. *Data Quality Report*)

GDAŃSK UNIVERSITY OF TECHNOLOGY

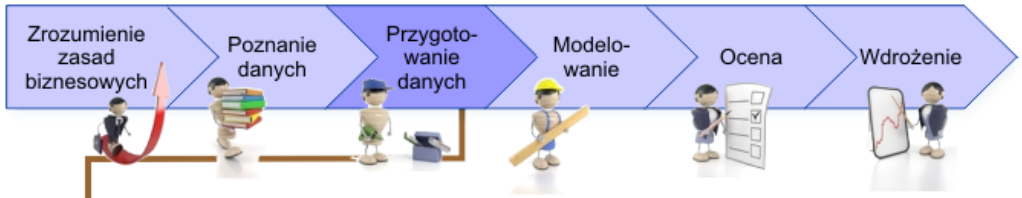
Przygotowanie danych (ang. *Data preparation*)

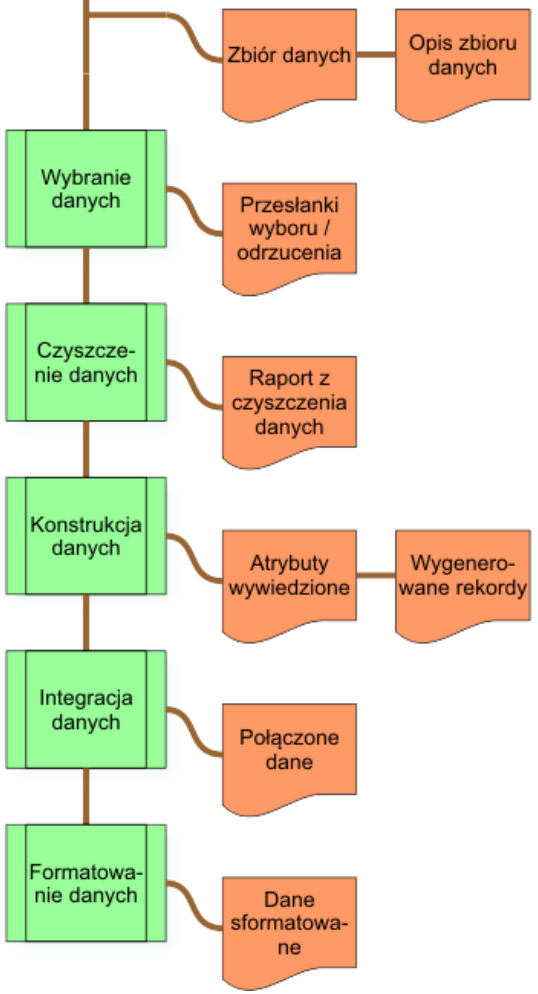
Wszystkie czynności potrzebne do skonstrułowania wynikowego zbioru danych (danych, którymi będzie zasilone narzędzie modelowania) ze zbioru danych źródłowych. Przygotowanie danych jest wykonywane kilkakrotnie.

1. *Przekształcenie danych w celu ich oczyszczenia oraz sformatowania.*
2. *Odizolowanie i oznaczenie nietypowych danych.*
3. *Wyeliminowanie lub uzupełnienie brakujących informacji.*
4. *Dyskretyzacja danych.*
5. *Normalizacja danych rozumiana jako zastąpienie różnych, reprezentujących tę samą sytuację biznesową wartości jedną wartością.*
6. *Spłaszczenie danych.*

GDAŃSK UNIVERSITY OF TECHNOLOGY

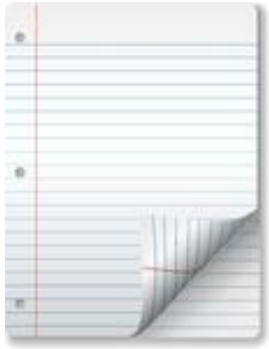
Przygotowanie danych





GDAŃSK UNIVERSITY OF TECHNOLOGY

Przygotowanie danych



Raport opisu zbioru danych
(ang. *Dataset Description Report*)

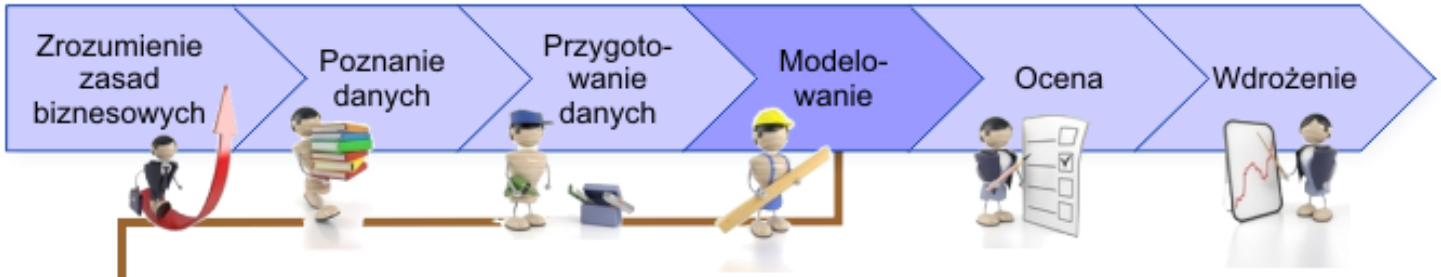
Modelowanie (ang. *Modeling*)

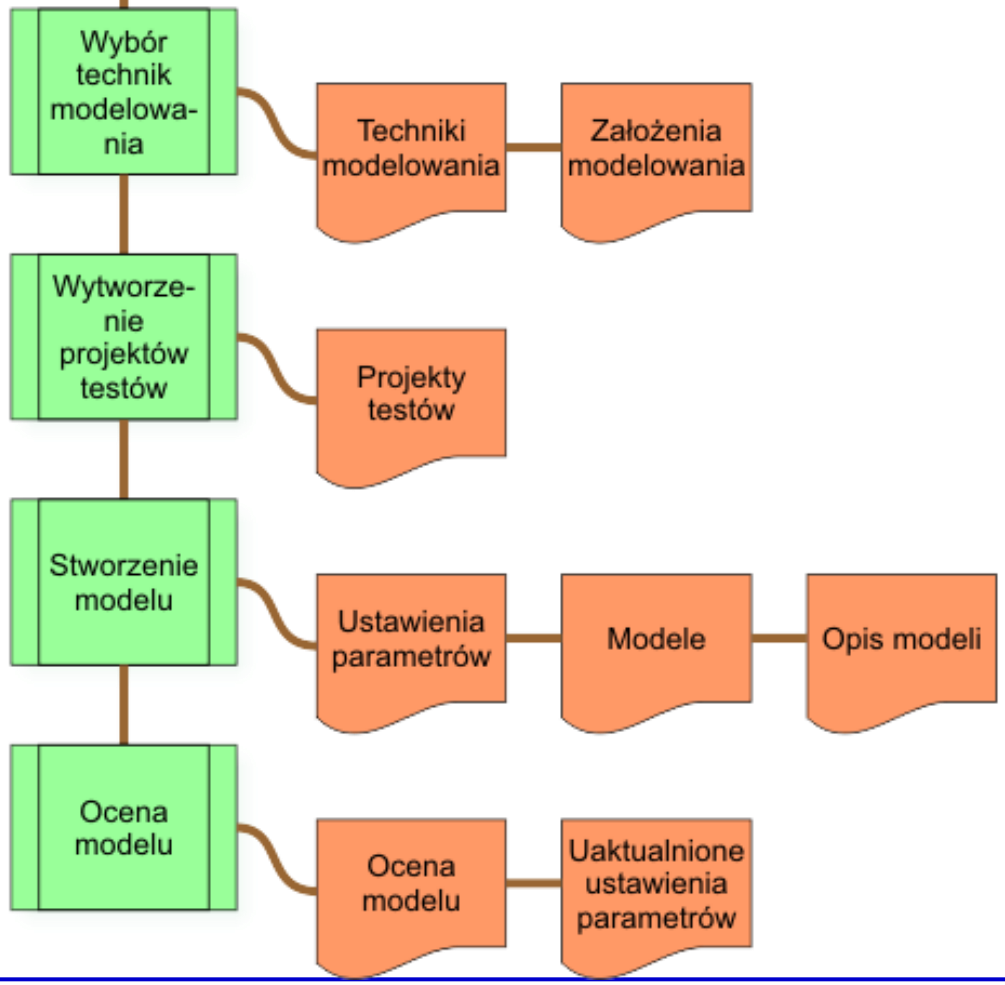
Wybranie i zastosowanie różnych technik modelowania oraz kalibracja ich parametrów do optymalnych wartości. Typowo, dla jednego problemu eksploracji danych jest kilka technik. Niektóre techniki mają specyficzne wymagania dotyczące danych, dlatego często następuje powrót do fazy przygotowania danych.

1. *Jakie techniki modelowania zastosować?*
2. *Jak zbudować model?*
3. *Jak przetestować model?*
4. *Jak ocenić model?*

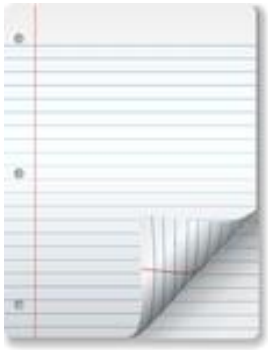
GDAŃSK UNIVERSITY OF TECHNOLOGY

Modelowanie





Modelowanie



Raport z modelowania
(ang. *Modeling Report*)

GDAŃSK UNIVERSITY OF TECHNOLOGY

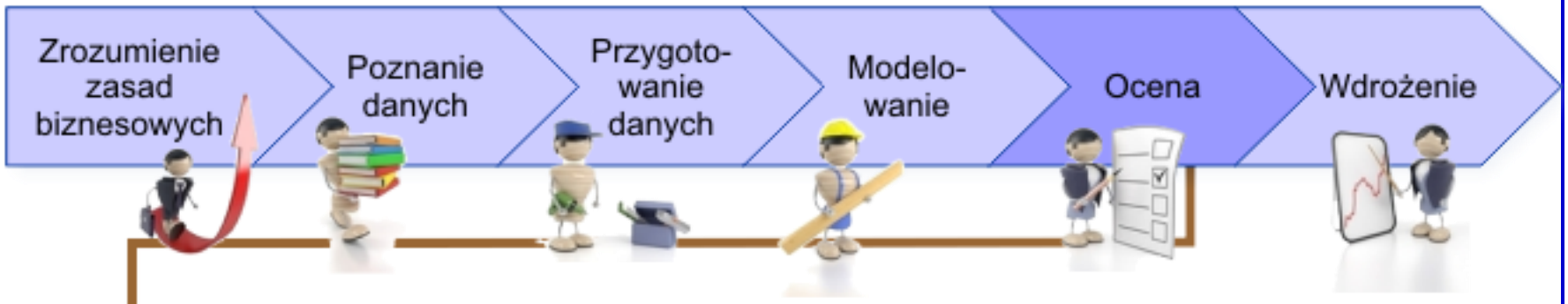
Ocena
(ang. *Evaluation*)

Ocena modelu i przegląd kroków prowadzących do konstrukcji modelu, aby uzyskać pewność, że model odpowiada celom biznesowym. Kluczowym zadaniem jest ustalenie, czy nie została ominięta żadna istotna kwestia biznesowa. W etapie końcowym wymagana jest dyskusja uzyskanych rezultatów eksploracji danych.

1. *Czy model odpowiada zdefiniowanym celom biznesowym?*
2. *Czy otrzymane wyniki eksploracji danych są pozytywnie ocenione przez końcowego użytkownika?*

GDAŃSK UNIVERSITY OF TECHNOLOGY

Ocena



- Ocena wyników
  - Ocena biznesowa wyników
    - Zatwierdzone modele
- Rewizja procesu
  - Rewizja procesu
- Określenie kolejnych kroków
  - Lista możliwych działań
    - Decyzje

GDAŃSK UNIVERSITY OF TECHNOLOGY

Ocena

Raport oceny
(ang. *Evaluation Report*)

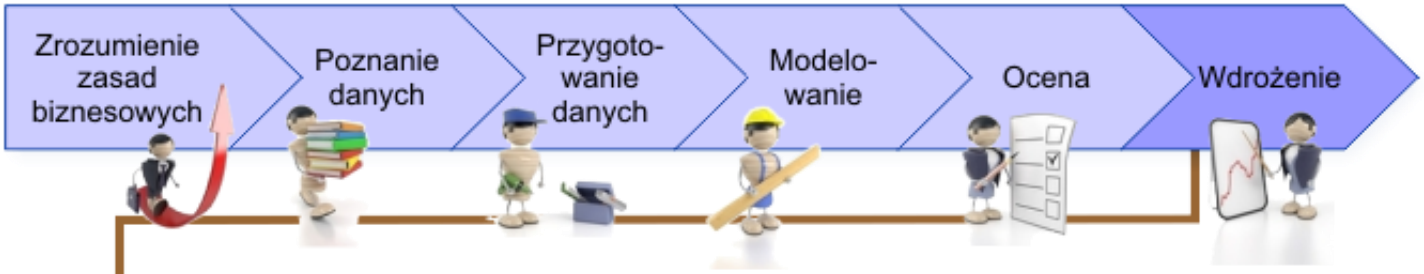
GDAŃSK UNIVERSITY OF TECHNOLOGY

Wdrożenie (ang. *Deployment*)

Odpowiednie zorganizowanie wydobytej wiedzy i jej opublikowanie w formie właściwej dla końcowego użytkownika.

1. *W jaki sposób włączyć eksplorację danych do procesu podejmowania decyzji?*
2. *W jakiej formie udostępnić wyniki końcowemu użytkownikowi?*
3. *W jaki sposób i z jaką częstotliwością aktualizować model?*

Wdrożenie



- Planowanie wdrożenia
  - Plan wdrożenia
- Planowanie monitoringu i utrzymania
  - Plan monitoringu i utrzymania
- Wytworze-nie raportu końcowego
  - Raport końcowy
    - Prezentacja końcowa
- Rewizja projektu
  - Dokumenta-cja doświad-czeń

Wdrożenie

- Plan wdrożenia (ang. *Deployment Plan*)
- Plan monitorowania i utrzymania (ang. *Monitoring and Maintanance Plan*)
- Raport końcowy (ang. *Final Report*)

GDAŃSK UNIVERSITY OF TECHNOLOGY

„Czarna skrzynka” czy „Biała skrzynka”

- Czarna skrzynka
  - dostępność oprogramowania typu „czarna skrzynka”
  - łatwość manipulowania danymi
  - potężna moc algorytmów eksploracji danych
  - analizy przygotowane na nieprzygotowanych danych
  - modele oparte na błędnych założeniach
- Biała skrzynka
  - zrozumienie algorytmicznych i statystycznych struktur leżących u podłoża oprogramowania

„Wyobraź sobie czarną skrzynkę zdolną do odpowiedzi na każde pytanie, które jej zadano. Każde pytanie. Czy wyeliminuje to potrzebę uczestnictwa ludzi, jak wielu sugeruje? Wprost przeciwnie. Zasadniczy problem sprowadza się nadal do kwestii ludzkiej. Jak mam poprawnie sformułować pytanie? Jak mam ustalić parametry, aby otrzymać rozwiązanie właściwe dla konkretnego przypadku, którym się interesuję? Jak mam otrzymać wyniki w rozsądnym czasie i w formie, którą będę mógł zrozumieć? Zauważ, że te wszystkie pytania łączą proces odkrywania ze mną, z moim ludzkim użyciem.”

Georges Grinstein

GDAŃSK UNIVERSITY OF TECHNOLOGY

Zastosowanie eksploracji danych

- Klasyfikacja
- Regresja
- Segmentacja
- Prognozowanie
- Asocjacja
- Analiza sekwencyjna
- Analiza tekstów
- Inteligentne sprawdzanie poprawności danych
- …

GDAŃSK UNIVERSITY OF TECHNOLOGY

Klasyfikacja (1)

Klasyfikacja jest formą analizy predykcyjnej polegającej na przewidywaniu jednej lub więcej podanych wartości. Wynikiem klasyfikacji może być prosta odpowiedź *Tak* lub *Nie*, bądź *Prawda* lub *Fałsz*, ale może być również jedna z wielu wartości, na przykład nazwa towaru czy nazwisko klienta.

1. *Problem kredytodawcy, który chce wiedzieć, czy udzielić danemu klientowi kredytu. Jeżeli tak, to na jakich warunkach? Jakie jest ryzyko niespłacenia kredytu w przypadku tego klienta?*
2. *Problem handlowca, który zastanawia się, czy utraci danego klienta? Jeżeli tak, co spowodowało, że wybrał on konkurencję?*

GDAŃSK UNIVERSITY OF TECHNOLOGY

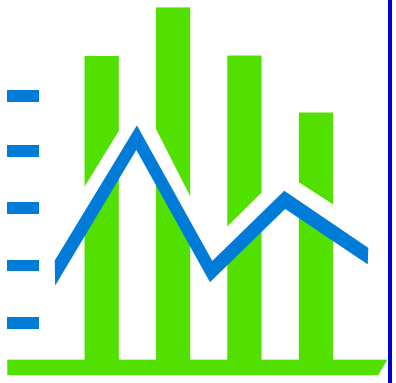
Klasyfikacja (2)

3. *Problem marketingowca zastanawiacącego się, kim są klienci firmy? Czy istnieją jakieś zależności między danymi demograficznymi klientów a ich chęcią kupowania w tej a nie innej firmie? Na których klientach należy się bardziej skoncentrować, ponieważ mogą oni przynieść firmie największe zyski?*
4. *Problem brygadzisty, który chce dowiedzieć się, dlaczego pewne serie produkowanych towarów mają więcej usterek niż inne? Jakie zmiany w produkcji mogą spowodować, że towary będą lepszej jakości?*

Algorytmy:
drzewa decyzyjne, regresja logistyczna, naiwny klasyfikator Bayesa, sieci neuronowe, …

GDAŃSK UNIVERSITY OF TECHNOLOGY

Regresja



Regresja jest podobną do klasyfikacji formą analizy predykcyjnej, ale w jej przypadku przewidywane wartości nie muszą należeć do określonego zbioru. Wynikiem regresji może więc być np. wartość przyszłej sprzedaży lub przewidywany czas trwałości pewnego produktu.

1. *Problem przedstawiciela handlowego, który chce wiedzieć, ile zysku przyniesie mu współpraca z danym klientem?*
2. *Problem klienta serwisu, którego interesuje, jak długo urządzenie będzie w naprawie?*
3. *Problem szefa działu PR, którego interesują czynniki wpływające na opinie klientów o firmie.*

Algorytmy:
drzewa decyzyjne, regresja logistyczna, naiwny klasyfikator Bayesa, sieci neuronowe, ...

GDAŃSK UNIVERSITY OF TECHNOLOGY

Segmentacja

Segmentacja polega na grupowaniu rekordów i w przeciwieństwie do klasyfikacji i regresji nie jest formą analizy predykcyjnej. Pozwala ona łączyć w klastry przypadki mające podobną charakterystykę.

1. *Problem pracownika działu kontrolingu, którego zadaniem jest wykrywanie podejrzanych transakcji finansowych.*
2. *Problem osoby odpowiedzialnej za bezpieczeństwo systemu komputerowego, która w dziennikach zdarzeń musi znaleźć podejrzane (nietypowe), mogące świadczyć o ataku wpisy.*
3. *Problem marketingowca, który musi zakwalifikować klientów do najlepiej opisujących ich kategorii.*
4. *Problem asystentki, która chce optymalnie uporządkować dokumenty.*

Algorytmy:
algortym klastrowania, algorytm klastrowania sekwencyjnego, ...

GDAŃSK UNIVERSITY OF TECHNOLOGY

Asocjacja

Asocjacja jest techniką wykrywania istniejących pomiędzy poszczególnymi przypadkami zależności i najczęściej jest stosowana do analizy koszyka zakupów, czyli wyszukiwania najczęściej razem kupowanych towarów.

1. *Problem sprzedawcy chcącego wiedzieć, które towary są kupowane w ramach jednej transakcji, w celu zarekomendowania ich klientom i zwiększenia w ten sposób szansy na sprzedaż krzyżową.*
2. *Problem kierownika sklepu, który musi właściwie zaplanować rozmieszczenie towarów.*

Algorytmy:
reguły asocjacyjne, drzewa decyzyjne, …

GDAŃSK UNIVERSITY OF TECHNOLOGY

Analiza sekwencyjna

Analiza sekwencyjna wykrywa często powtarzające się sekwencje zdarzeń.

1. *Problem projektanta witryny WWW, który chce wiedzieć, w jakiej kolejności odwiedzające jego witrynę osoby wyświetlają poszczególne podstrony. Dysponując taką informacją, może on dostosować witrynę do potrzeb użytkowników i zmniejszyć liczbę osób rezygnujących z poszukiwania interesujących je podstron na niewłaściwie zaprojektowanej witrynie.*
2. *Problem naukowca chcącego przewidzieć dalszy rozwój przeprowadzonych badań (np. wyniki przyszłych eksperymentów) lub chcącego zweryfikować hipotezę statystyczną.*

Algorytmy:
algorytm klastrowania sekwencyjnego, …

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (1)

Analiza samochodowych skarg gwarancyjnych: Przykład działania przemysłowego procesu standaryzacji.

1. Zrozumienie uwarunkowań biznesowych

Celem DaimlerChrysler jest **zmniejszenie kosztów związanych ze skargami gwarancyjnymi i poprawa satysfakcji klientów**. Poprzez rozmowę z zakładowymi inżynierami, którzy są technicznymi ekspertami z konstrukcji samochodów, badacze mogli sformułować konkretne problemy biznesowe, takie jak następujące:

- **Czy są współzależności pomiędzy skargami gwarancyjnymi?**
- **Czy reklamacje z przeszłości są powiązane z podobnymi skargami w przyszłości?**
- **Czy jest związek między konkretnym typem skargi a konkretnym warsztatem?**

Plan to zastosowanie odpowiednich technik eksploracji danych, aby spróbować odkryć te i inne możliwe zależności.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (2)

2. Zrozumienie danych

Badacze wykorzystują system informacji o jakości samochodów firmy DaimlerChrysler (Quality Information System – QUIS), który zawiera informację o ponad **7 milionach samochodów i zajmuje około 40 gigabajtów**. QUIS zawiera szczegóły produkcyjne o tym, **jak i gdzie konkretny samochód został zbudowany, oraz średnio około 30 lub więcej kodów sprzedaży dla każdego samochodu**. QUIS również przechowuje informacje o **skargach gwarancyjnych, które są dostarczane przez warsztaty, podając jedną z co najmniej 5000 możliwych potencjalnych przyczyn**.

Badacze podkreślili fakt, że **baza danych była całkowicie niezrozumiała dla nie-ekspertów**: „Należało zatem znaleźć ekspertów z różnych działów i porozumieć się z nimi; krótko mówiąc, zadanie okazało się raczej kosztowne”. Podkreślili, że **analitycy powinni doceniać wagę, trudność i potencjalne koszty tego wczesnego etapu procesu eksploracji danych**, gdyż podejście tutaj na skróty może prowadzić do kosztownych powtórzeń procesu.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (3)

3. Przygotowanie danych

Badacze odkryli, że chociaż baza danych QUIS jest relacyjna, to ma ograniczony dostęp przez SQL. Musieli **ręcznie wybierać interesujące przypadki i zmienne**, a następnie ręcznie otrzymywać nowe zmienne, które mogłyby zostać użyte podczas modelowania. Na przykład zmienna **liczba dni od daty sprzedaży do pierwszej reklamacji** musiała być otrzymana z odpowiednich dat.

Następnie użyli zastrzeżonego oprogramowania do eksploracji danych, które było używane przez DaimlerChrysler we wcześniejszych projektach. Tutaj napotkali pospolite przeszkody – **wymagania formatu danych różniły się pomiędzy algorytmami**. Skutkiem była dalsza drobiazgowa obróbka danych, aby przekształcić atrybuty w formę przydatną dla algorytmów modelu. Badacze wspomnieli, że **etap przygotowania danych zajął im dużo więcej czasu, niż planowali**.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (4)

4. Modelowanie

Ponieważ w sumie problemem biznesowym z etapu 1 było zbadanie zależności pomiędzy skargami gwarancyjnymi, badacze postanowili zastosować następujące techniki: **1) sieci bayesowskie i 2) reguły asocjacyjne**. Sieci bayesowskie modelują niewiadome przez wyraźne przedstawienie warunkowych zależności pomiędzy różnymi składnikami, dostarczając zatem graficzną wizualizację zależności pomiędzy składnikami. Jako takie, sieci bayesowskie reprezentują naturalny wybór dla modelowania zależności pomiędzy reklamacjami. Reguły asocjacyjne są również naturalnym sposobem badania zależności pomiędzy reklamacjami, ponieważ miara pewności reprezentuje rodzaj prawdopodobieństwa warunkowego, podobnego do sieci bayesowskich.

Szczegóły wyników są poufne, ale możemy przedstawić ogólne wnioski o typie zależności odkrytych przez modele. Jednym ze spostrzeżeń odkrytych przez badaczy było to, że **konkretna kombinacja specyfikacji konstrukcyjnej podwajała prawdopodobieństwo napotkania problemów z samochodowymi przewodami elektrycznymi**. Inżynierowie DaimlerChrysler zaczęli badać, jak ta kombinacja może wpływać na zwiększenie problemów z przewodami.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (5)

Badacze badali, czy pewne warsztaty nie mają więcej skarg gwarancyjnych określonego rodzaju niż inne warsztaty. Ich wynikowe reguły asocjacyjne pokazały, że rzeczywiście, poziom pewności dla reguły **„Jeżeli warsztat X, to problem z przewodami", różnił się znacząco w zależności od warsztatu**. Stwierdzili, że dalsze badania są uzasadnione, aby odkryć przyczyny tych różnic.

5. Ewaluacja

Badacze byli rozczarowani, że wpływ sekwencyjnych reguł asocjacyjnych był stosunkowo mały, zatem w ich opinii wykluczający uogólnienie wyników. W sumie stwierdzili: „**Faktycznie nie odkryliśmy żadnych reguł, które naszym ekspertom mogłyby się wydać interesujące, przynajmniej na pierwszy rzut oka**." Zgodnie z tym kryterium, modele okazały się nieefektywne i nie spełniły wymagań postawionych podczas etapu zrozumienia problemów biznesowych. Badacze tłumaczą to **strukturą bazy danych otrzymaną w „spadku", w której z historycznych lub technicznych powodów części samochodowe były sklasyfikowane przez warsztaty i fabryki, a która nie była projektowana do eksploracji danych**. Badacze sugerują dostosowanie i przeprojektowanie bazy danych, aby stała się bardziej otwarta na odkrywanie wiedzy.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Przypadek użycia (6)

6. Wdrożenie

Badacze zidentyfikowali powyższy projekt jako **projekt pilotażowy i dlatego nie zamierzali wdrażać żadnych dużych modeli z tej pierwszej iteracji**. Po tym projekcie pilotażowym, jednak, zastosowali wyciągnięte z niego wnioski w celu zintegrowania ich metod z istniejącym środowiskiem technologii informacyjnej DaimlerChrysler. Aby dalej dążyć do pierwotnego celu obniżania kosztów skarg, zamierzają rozwinąć wewnętrzną sieć komputerową, zapewniającą możliwość eksploracji QUIS wszystkim pracownikom spółki.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Dziękujemy za uwagę

Zapraszamy na wykład:

PODSTAWOWE ZAGADNIENIA I TERMINY