Robert Kozakiewicz Gdańsk, 26.04.2015

Dawid Piasecki

Eksploracja danych

Zrozumienie warunków biznesowych i identyfikacja źródeł danych

**1. Cele biznesowe organizacji.**

1.1. Krótka charakterystyka organizacji

Nasza organizacja jest firmą ubezpieczeniową, zajmującą się ubezpieczeniami majątkowymi, a dokładnie komunikacyjnymi. Organizacja ta oferuje ubezpieczenia mieszkańcom Stanów Zjednoczonych i Europy .

1.2. Cele biznesowe organizacji

* zwiększenie bazy klientów "bezkolizyjnych" i bezpiecznych
* zwiększenie stawek klientom niebezpiecznym
* znalezienie grupy wiekowej o największej rentowności ubezpieczania

1.3. Kryteria sukcesu organizacji

* dziesięcioprocentowy spadek kosztów wypłacanych odszkodowań
* wzrost liczby klientów o pięć procent

**2. Cele biznesowe eksploracji danych.**

2.1.

a) Jaka jest charakterystyka kierowcy, który uczestniczy w wypadku?

b) Jaka jest charakterystyka warunków pogodowych podczas wypadków?

c) Jaka jest charakterystyka pojazdu, który uczestniczy w wypadku?

d) Jak pora dnia i data w roku wpływa na ilość wypadków?

e) Jak charakterystyka drogi(rodzaj drogi i oświetlenie) wpływa na ilość wypadków?

2.2.

a) Wyznaczenie profilów niebezpiecznego kierowcy ( przedziały wiekowe, płeć).

b) Wyznaczenie charakterystyki warunków pogodowych przyczyniających się do zwiększenia ilości wypadków.

c) Wyznaczenie charakterystyk niebezpiecznych pojazdów.

d) Wyznaczenie charakterystyki czasu w jakim dochodzi do wypadków.

e) Wyznaczenie charakterystyk drogi, na których najczęściej dochodzi do wypadków.

**3. Identyfikacja źródeł danych.**

Będziemy korzystać z danych w formacie csv pobranych ze strony rządowej Stanów Zjednoczonych (http://catalog.data.gov/dataset/traffic-tickets-issued-beginning-2008). Dane te zawierają informacje o mandatach przyznanych kierowcom w stanie Nowy Jork w 2013 roku: rodzaj wykroczenia, kod wykroczenia, rok, miesiąc oraz czas wykroczenia, wiek oraz płeć kierowcy, stan, w którym wydano prawo jazdy kierowcy, oddział policji oraz sąd zajmujący się sprawą wykroczenia.

Drugi zestaw danych to dane wypadków z Leeds z 2014 roku(http://data.gov.uk/dataset/road-traffic-accidents). Zawierają one informację o warunkach pogodowych, rodzaju drogi, wieku kierowcy, pogodzie, dacie wypadku.

**4. Analiza źródeł danych pod kątem ich kompletności względem celów zdefiniowanych celów eksploracji danych.**

a) Potrzebne dane: wiek i płeć karanych kierowców - dostępny w zidentyfikowanym źródle "Nowy Jork".

b) Potrzebne dane: pogoda podczas wypadku, rodzaj nawierzchni - dostępna w zidentyfikowanym źródle "Leeds".

c) Potrzebne dane: rodzaj pojazdu - dostępny w zidentyfikowanym źródle "Leeds".

d) Potrzebne dane: miesiąc, czas w którym popełniono wykroczenie - dostępny w zidentyfikowanym źródle "Nowy Jork".

e) Potrzebne dane: rodzaj drogi oraz światła na niej - dostępny w zidentyfikowanym źródle "Leeds".