Robert Kozakiewicz Gdańsk, 12.06.2015

Dawid Piasecki

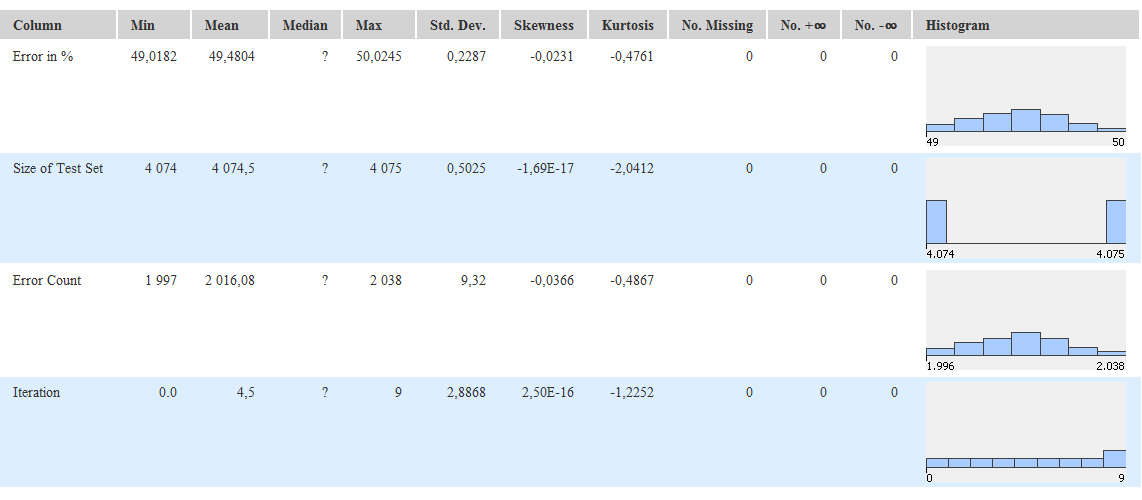
Eksploracja danych

Walidacja modeli eksploracjii danych

**1. Ocena błędów modeli.**

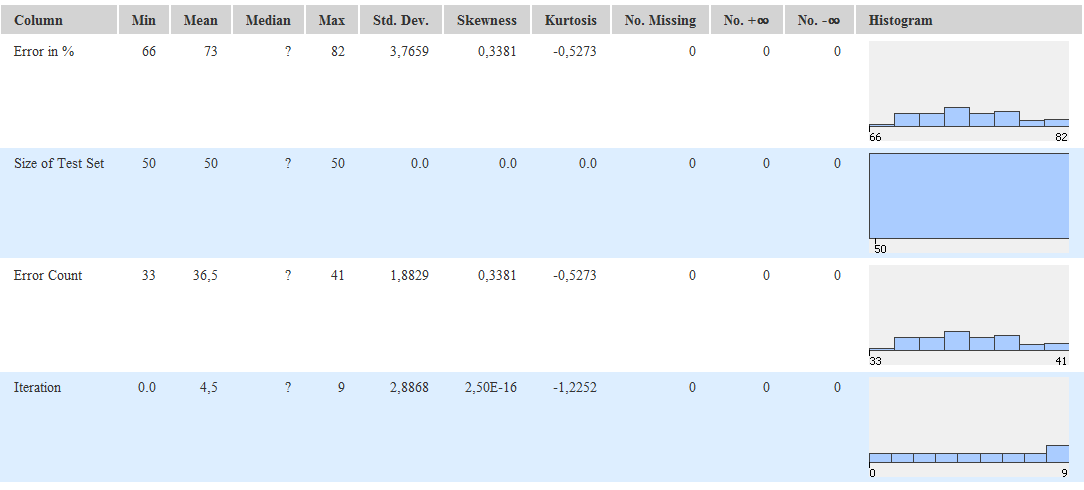
a) Jaka jest charakterystyka kierowcy, który uczestniczy w wypadku?

Do tego celu użyliśmy danych o mandatach, interesujące nas kolumny to 'Gender' oraz 'Age at Violation'. Jako docelowy atrybut ustaliliśmy kolumnę 'Violation Description'. Wartość 0 oznacza płeć męską, natomiast 1 płeć żeńską. Użyliśmy węzła 'Decision tree learner'. Wybór spośród 6 wartości. Błąd obliczony za pomocą walidacji skrośnej wynosi około 50%.



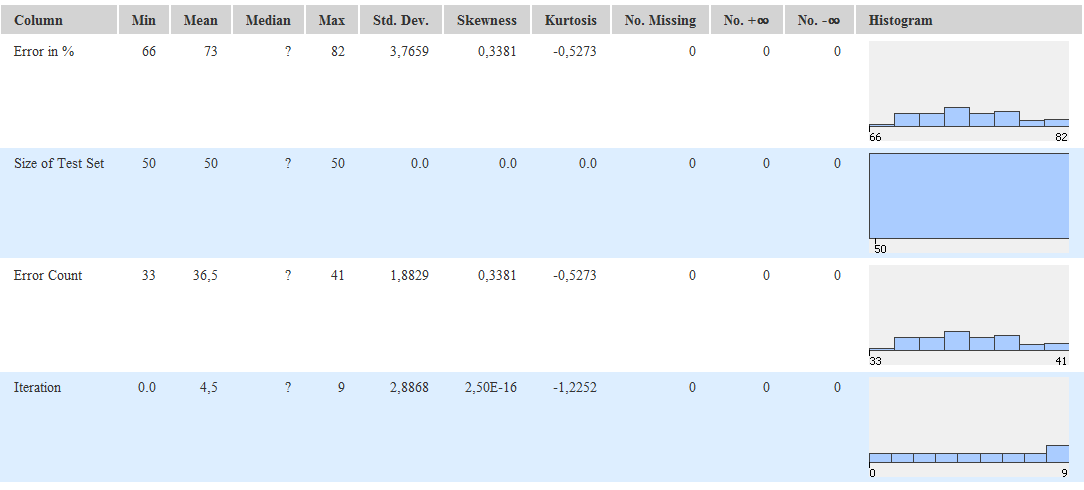
b) Jaka charakterystyka warunków pogodowych podczas wypadków wpływa na ilość pojazdów biorących udział w wypadku?

W rozwiązaniu tego celu wykorzystaliśmy dane ze źródła o wypadkach. Kolumny "Road Surface" i "Weather Conditions". Wybrany model to drzewo decyzyjne z atrybutem celu "Number of Vehicles". Możliwe 1-5 pojazdów w wypadku. Błąd wysoki wynoszący około 70%. Model ten należałoby odrzucić.



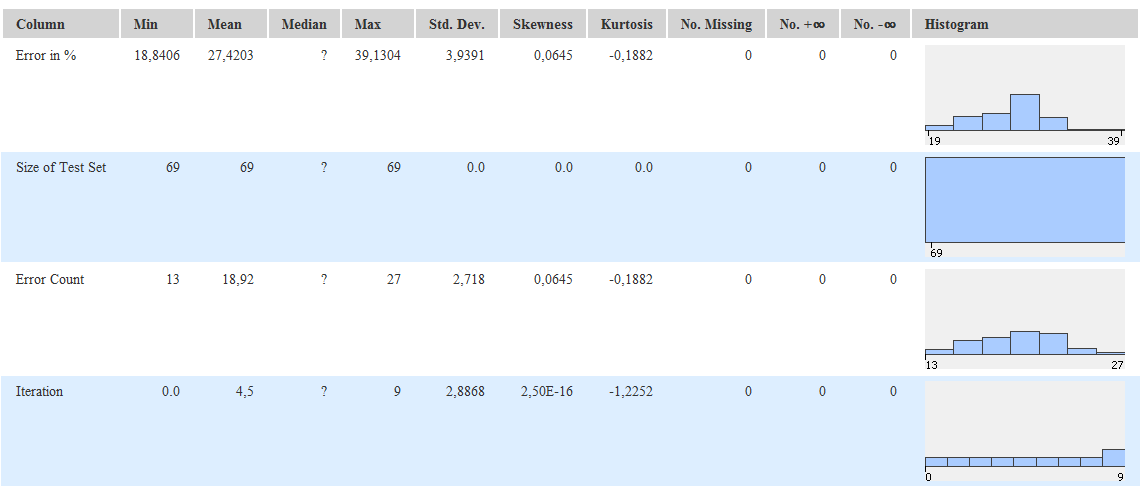
c) Jaka jest charakterystyka pojazdu, który uczestniczy w wypadku?

W tym problemie posłużyliśmy się źródłem danych o wypadkach. Interesująca nas kolumna to 'Type of Vehicle'. Wybrany model to drzewo decyzyjne z atrybutem celu "Casualty Severity". Wybór spośród 3 wartości. Błąd obliczony za pomocą walidacji skrośnej wynosi około 54%.



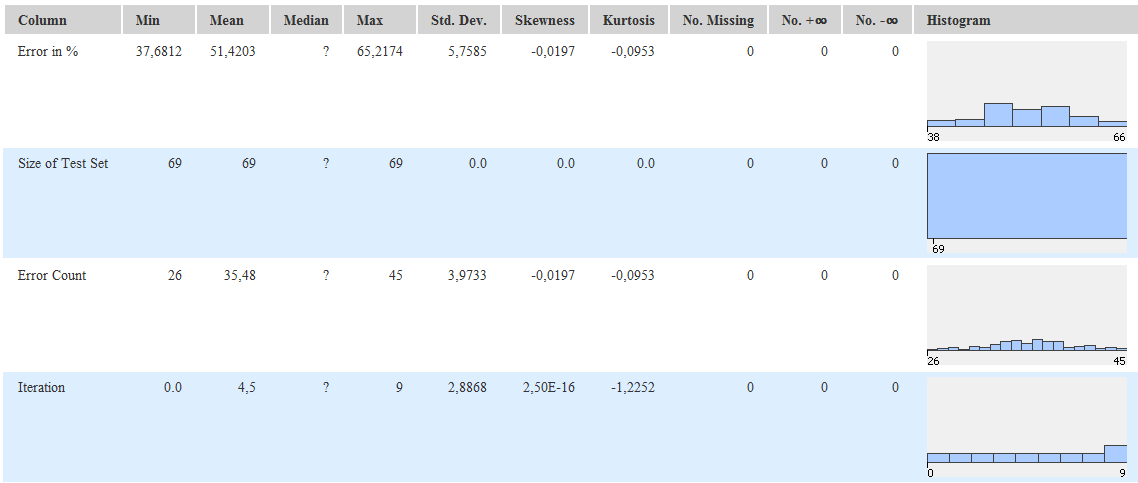
d) Jak pora dnia i data w roku wpływa na rodzaj wypadku?

Danę, które przyczyniły się do rozwiązania tego problemu to: kwartał roku, otrzymany za pomocą "Date Field Extractor" z kolumny "Accident Date" oraz godzina w której miał miejsce wypadek z kolumny "Time (24hr)" ze źródła o wypadkach. Wybrany model to drzewo decyzyjne z atrybutem celu "Casualty Severity". Wybór spośród 3 wartości. Błąd około 27%.



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

Interesujące nas dane znajdują się w źródle z wypadkami, a konkretnie w kolumnach "1st Road Class" oraz "Lighting Conditions". Wybrany model to drzewo decyzyjne z atrybutem celu "Casualty Severity". Wybór spośród 3 wartości. Błąd około 51%. Wysoki.



**2. Czy model odpowiada zdefiniowanym celom biznesowym?**

Zdefiniowane cele biznesowe, wraz z modelami które pomogą w ocenie:

a) zwiększenie bazy klientów "bezkolizyjnych" i bezpiecznych:

- 1a, dzięki określeniu jakie mandaty dostają jacy kierowcy

- 1c, dzięki określeniu jaki rodzaj wypadków powodują jakie pojazdy

-1e, jeśli oszacujemy po jakich drogach jeździ klient

-1d, jeśli określimy w jakich okresach dnia i roku jeździ klient

b) zwiększenie stawek klientom niebezpiecznym

- 1a, dzięki określeniu jakie mandaty dostają jacy kierowcy

- 1c, dzięki określeniu jaki rodzaj wypadków powodują jakie pojazdy

-1d, jeśli określimy w jakich okresach dnia i roku jeździ klient

-1e, jeśli oszacujemy po jakich drogach jeździ klient

c) znalezienie grupy wiekowej o największej rentowności ubezpieczania - 1a, dzięki określeniu jakie mandaty dostają jacy kierowcy