Wybór cech (*feature selection*) pozwala zidentyfikować atrybuty, które mają największy wpływ na predykcję atrybutów wyjściowych.

Dane *customer_dbsave.sav* zawierają informacje o odpowiedziach na pytania ankietowe udzielonych przez 5000 klientów firmy telefonicznej. Dane zawierają informacje o wieku klienta, zatrudnieniu, dochodach i statystyki wykorzystania telefonu. Trzy „docelowe" pola wskazują, czy klient odpowiedział na zapytania ofertowe. Firma chce wykorzystać te dane do wspomagania predykcji którzy klienci odpowiedzą w przyszłości na podobne oferty.

Przykład wykorzystuje tylko jedną ofertę jako cel. Wykorzystany zostanie model drzewa decyzyjnego CHAID: bez wyboru cech i z wyborem cech (10 najlepszych predykatorów). W drugim przypadku otrzymane wyniki okażą się efektywniejsze.

Budowanie strumienia

W pustym oknie wstaw źródło „Plik Statistica" i jako źródło wskaż plik customer_dbsave.sav.
Wstaw węzeł „Typy" z zakładki „Zmienne" i zmień kierunek dla „response_01" na „Przewidywana". Dla „custid", „response_02", „response_03" ustaw kierunek na „Brak" .
Wstaw węzeł „Dobór predyktorów" z zakładki „Modelowanie" wskazując wcześniej węzeł „Typy" jako źródło. Uruchom strumień.
Kliknij otrzymany model (w prawym górnym rogu) i prawym przyciskiem myszy wybierz „Przeglądaj". Górny panel pokazuje pola, które zostały uznane za użyteczne w predykcji.

Wstaw zbudowany model do strumienia wskazując wcześniej węzeł „Typy" i kliknij na niego dwukrotnie. Wybierz 10 pierwszych predyktorów.
Aby porównać wyniki bez wyboru predykatorów trzeba wygenerować dwa modele: jeden który użyje wybór i drugi, który tego wyboru cech nie wykorzysta. Wstaw dwa węzły „CHAID'" z zakładki „Modelowanie": jeden połącz do węzła „Typy" (zmień jego nazwę na „All input fields", drugi do wygenerowanego „response_01" (zmień jego nazwę na „Top 10 fields") .
Uruchom węzeł „All input fields" – zauważ jak długo się wykonuje. Wykonaj to samo dla drugiego węzła.
Drugi model wykonał się szybciej (różnice te na pewno będą bardziej zauważalne dla większej liczby danych). Drugie drzewo zawiera również mniej węzłów niż pierwsze. Jest łatwiejsze do interpretacji i zrozumienia.