IBM SPSS Modeler

przykład wprowadzający

Dane do analizy - informacje o klientach banku

Tabela (6 zmiennych, 2 464 rekordów) #1

Plik Edycja Generuj

Tabela Adnotacje

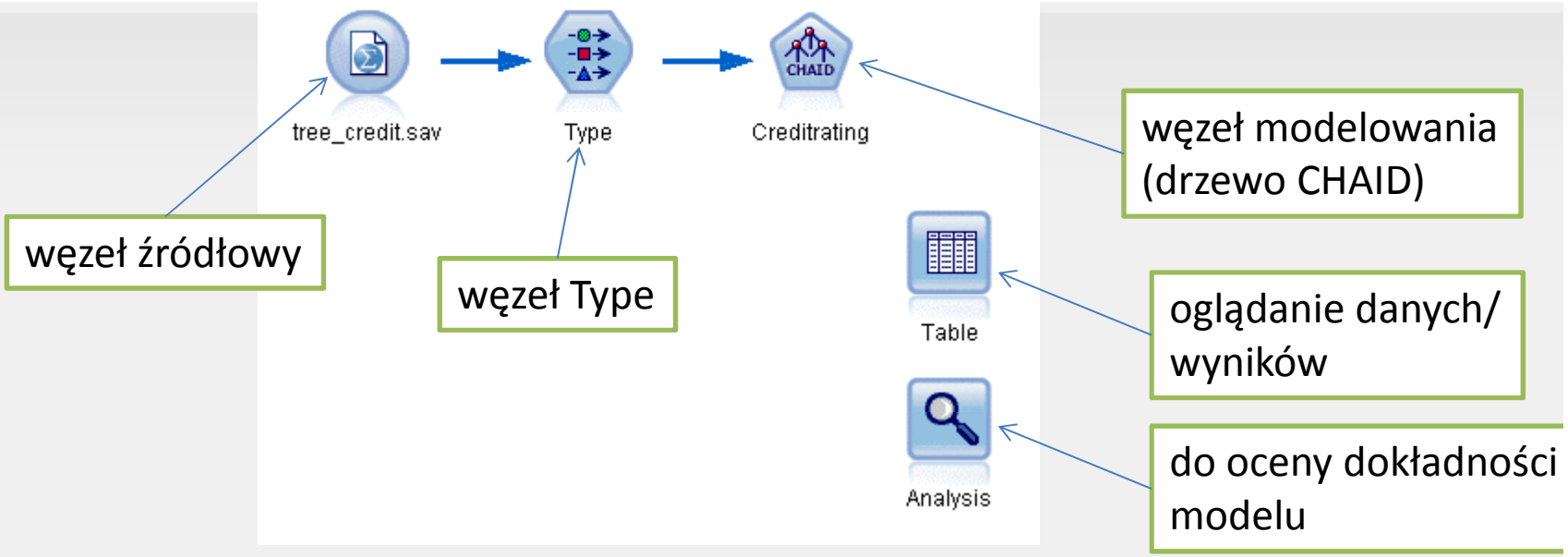
	Credit rating	Age	Income level	Number of credit cards	Education	Car loans
1	Bad	36.220	Medium	5 or more	College	More than 2
2	Bad	21.990	Medium	5 or more	College	More than 2
3	Bad	29.170	Low	5 or more	High school	More than 2
4	Bad	32.753	Low	5 or more	College	None or 1
5	Bad	36.771	Medium	5 or more	College	More than 2
6	Bad	39.325	Medium	5 or more	College	More than 2
7	Bad	31.699	Medium	5 or more	College	More than 2
8	Bad	34.718	Low	5 or more	High school	More than 2
9	Bad	31.531	Low	5 or more	High school	More than 2
10	Bad	24.780	Medium	5 or more	College	More than 2
11	Bad	22.763	Low	5 or more	College	More than 2
12	Bad	45.966	Low	5 or more	High school	More than 2
13	Bad	29.386	Medium	5 or more	High school	More than 2
14	Bad	29.215	Low	5 or more	College	None or 1
15	Bad	39.603	Low	5 or more	High school	More than 2
16	Bad	39.456	Low	5 or more	College	More than 2
17	Bad	34.127	Low	5 or more	College	More than 2
18	Bad	35.818	Medium	5 or more	College	More than 2
19	Bad	35.966	Medium	5 or more	College	More than 2
20	Bad	26.263	High	5 or more	College	More than 2

OK

Field name	Description
Credit_rating	Credit rating: 0=Bad, 1=Good, 9=missing values
Age	Age in years
Income	Income level: 1=Low, 2=Medium, 3=High
Credit_cards	Number of credit cards held: 1=Less than five, 2=Five or more
Education	Level of education: 1=High school, 2=College
Car_loans	Number of car loans taken out: 1=None or one, 2=More than two

- Otwieramy przykładowy strumień (w Demo)
  - Plik -> Otwórz strumień
  - w katalogu Demo wybieramy (Demo->Streams)
    - klikamy na strumień *modelingintro.str*

Budowanie strumienia

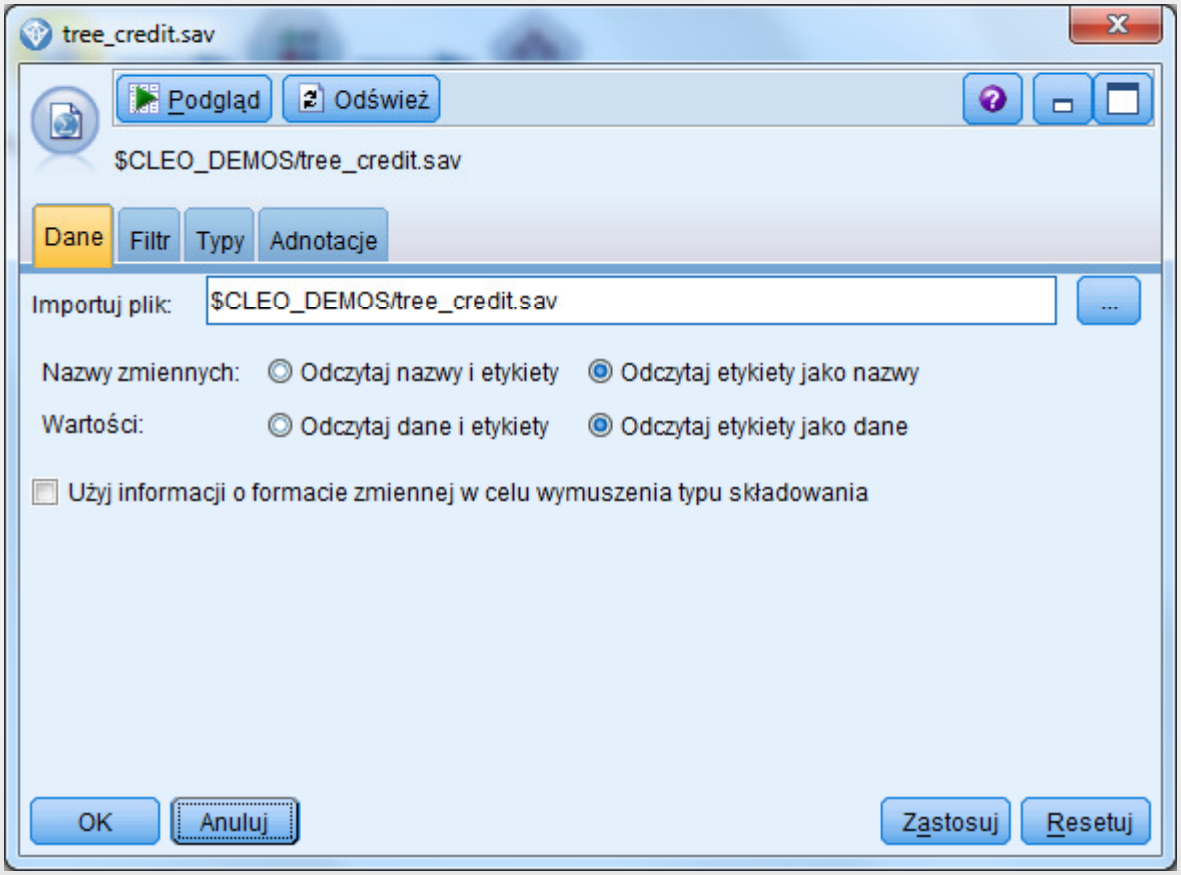


Aby zbudować strumień, który będzie tworzył model, potrzebujemy przynajmniej 3 elementów:

1. węzeł *źródłowy*, który czyta dane z zewn. źródeł (np. IBM SPSS Statistics)
2. węzeł *Type*, który określa własności pól, np.:
    1. poziom pomiarów (np. typ danych)
    2. specyfikację pól (czy pole jest target czy input w modelowaniu)
3. węzeł *modelowania*, który generuje model w czasie uruchomienia strumienia

Węzeł źródłowy - wczytywanie danych

- źródło danych: *tree_credit.sav* (kliknij 2x)



Węzeł *Typy* – poziom pomiaru

- Węzeł *Typy* określa poziom pomiaru dla każdego pola
  - specyfikuje typ danych dla każdego pola
  - w naszym przykładzie są 3 różne poziomy pomiary
    - Pole *Ilościowe* (jak *Age*) przyjmuje ciągłe wartości numeryczne
    - Pole *Nominalne* (jak *Credit rating*) przyjmuje dwie lub więcej różne wartości, np. *Bad*, *Good* lub *No credit* historię kredytów
    - Pole *Porządkowe* (jak *Income level*) określa dane z wieloma różnymi wartościami, w których jest porządek, np. Low, Medium and High.

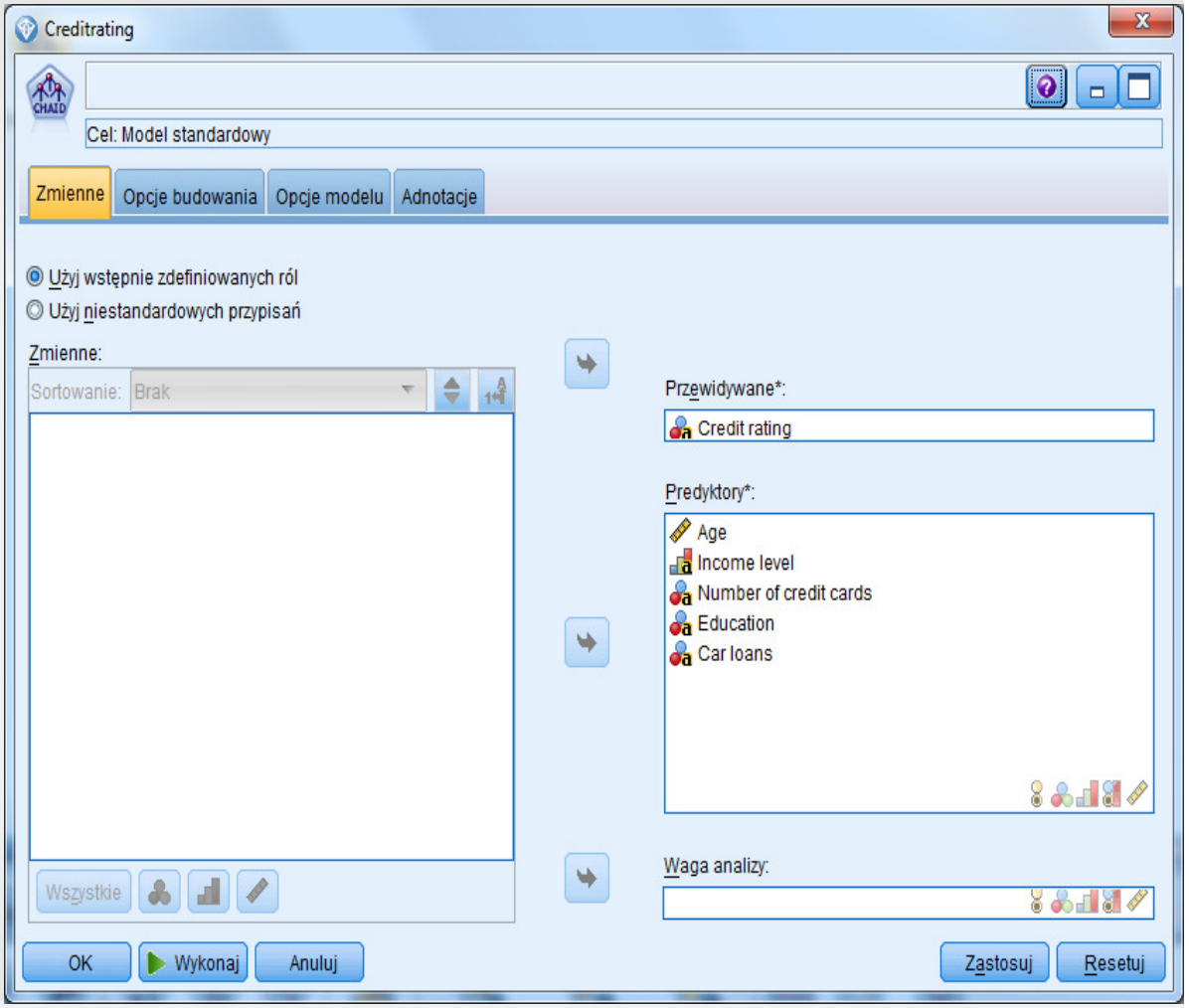
Węzeł *Typy - role*

- Węzeł *Typy* określa również rolę, jaką każde pole pełni w modelowaniu
- Dla pola *Credit rating* rola jest ustawiona jako *Przewidywana*
- Dla pozostałych pól rola jest ustawiona jako *Wejściowa* (predyktor)

Węzeł modelowania - zmienne

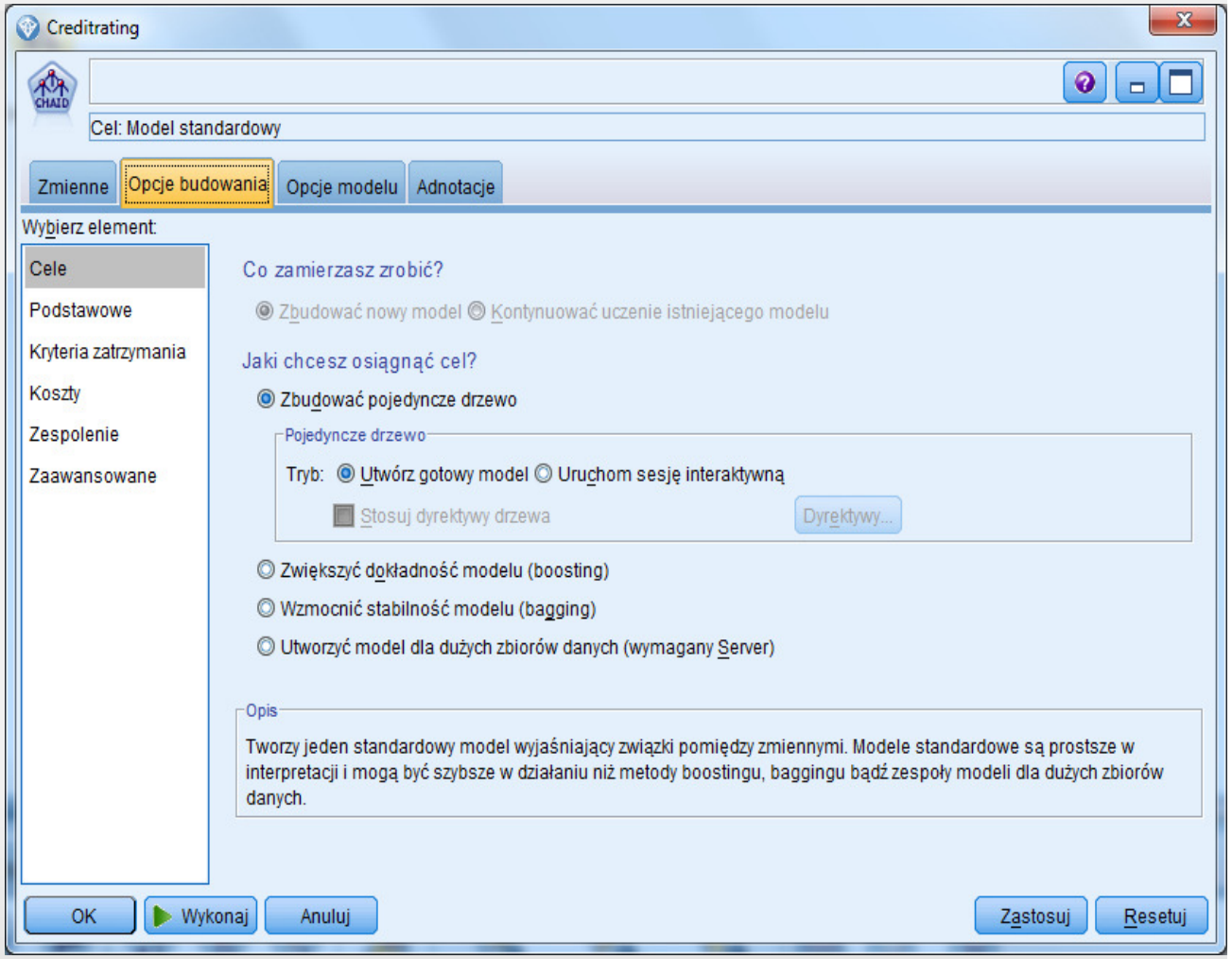
- Węzeł modelowania CHAID generuje model
- W zakładce *Zmienne* wybrana jest opcja *Użyj wstępnie zdefiniowanych ról*
  - *tzn Przewidywana i Wejściowa* będą takie jak zdefiniowane w węźle *Typy* (można to tutaj zmienić)

► Kliknij zakładkę *Opcje budowania*



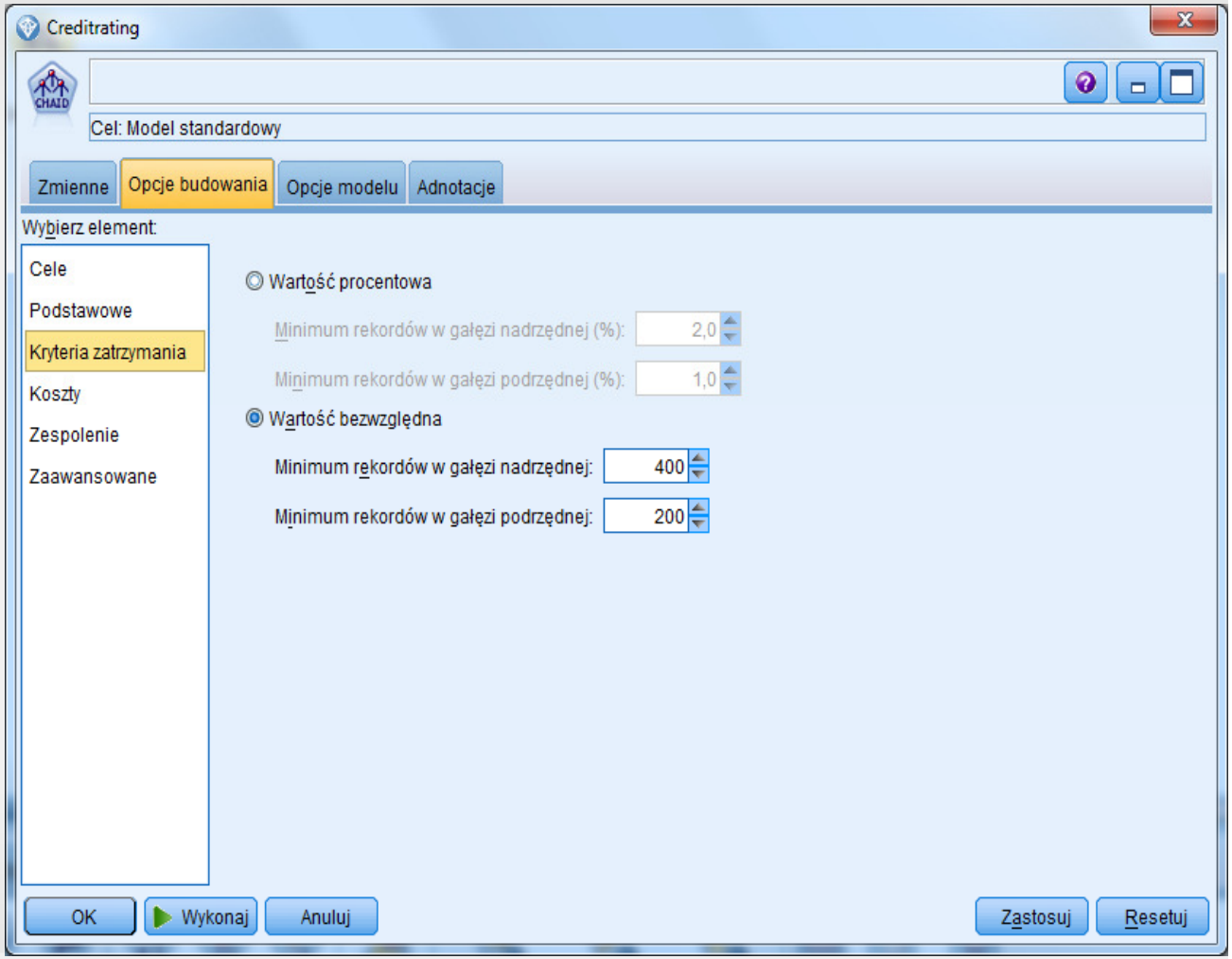
Węzeł modelowania – opcje budowania

- Specyfikowanie opcji rodzaju modelu, który chcemy zbudować
- Budujemy nowy model, więc wybieramy domyślne ustawienia
- Chcemy pojedynczy, standardowy model drzew decyzyjnego bez żadnych rozszerzeń, tak więc wybieramy domyślnie opcję *Zbudować pojedyncze drzewo*



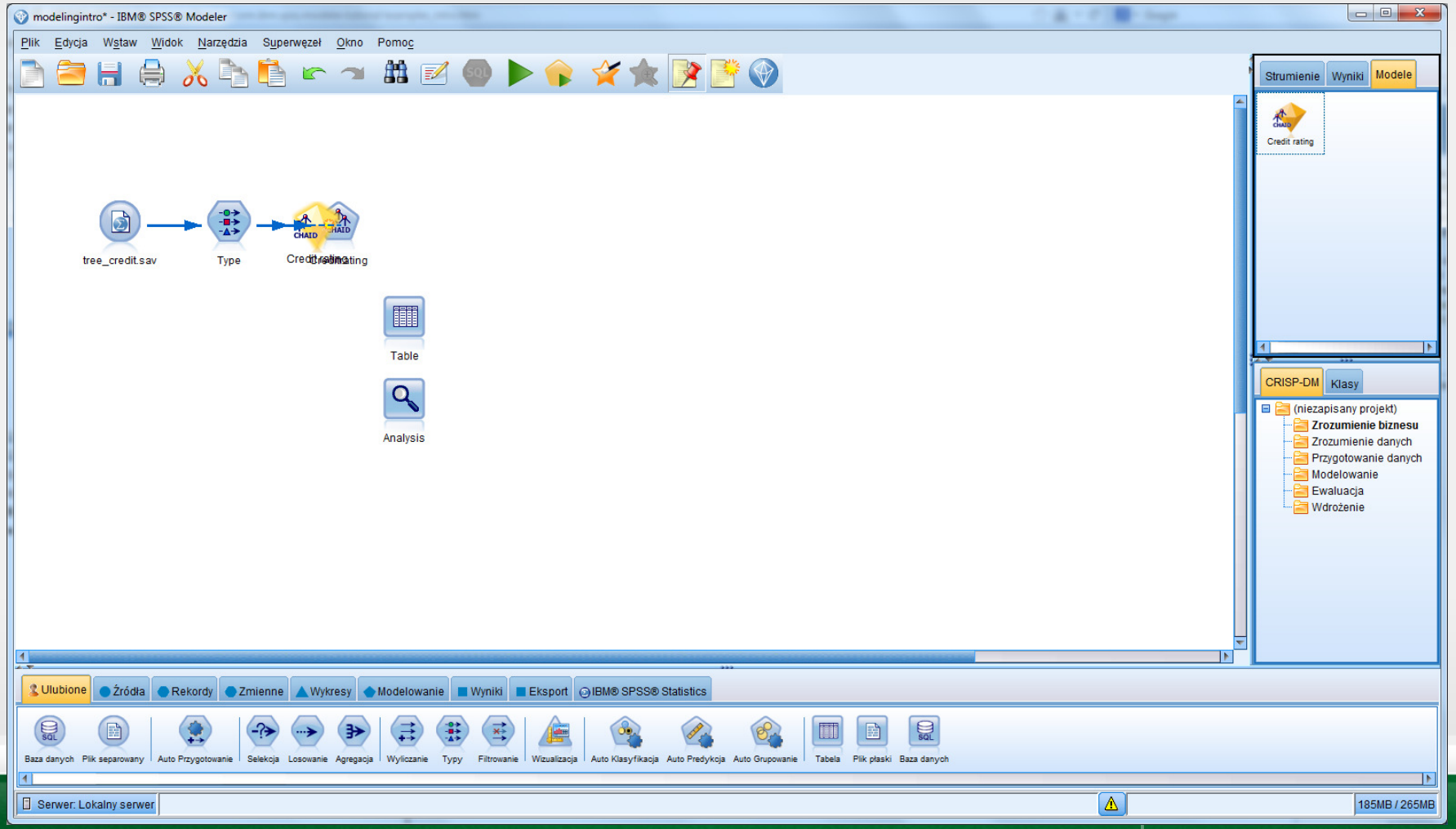
Węzeł modelowania – opcje budowania – kryterium zatrzymania

- W tym przykładzie chcemy, aby drzewo było stosunkowo proste, więc będziemy ograniczać wzrost drzewa poprzez ustawienie minimalnej liczby przypadków dla rodzica i węzłów potomnych.
- ▶ W zakładce *Opcje budowania*, wybierz *Kryteria zatrzymania*
- ▶ Wybierz opcję *Wartość bezwzględna*
- ▶ Ustaw *Minimum rekordów w gałęzi nadrzędnej* na 400
- ▶ Ustaw *Minimum rekordów w gałęzi podrzędnej* na 200.
- Uruchom budowę modelu
- ▶ *Wykonaj*



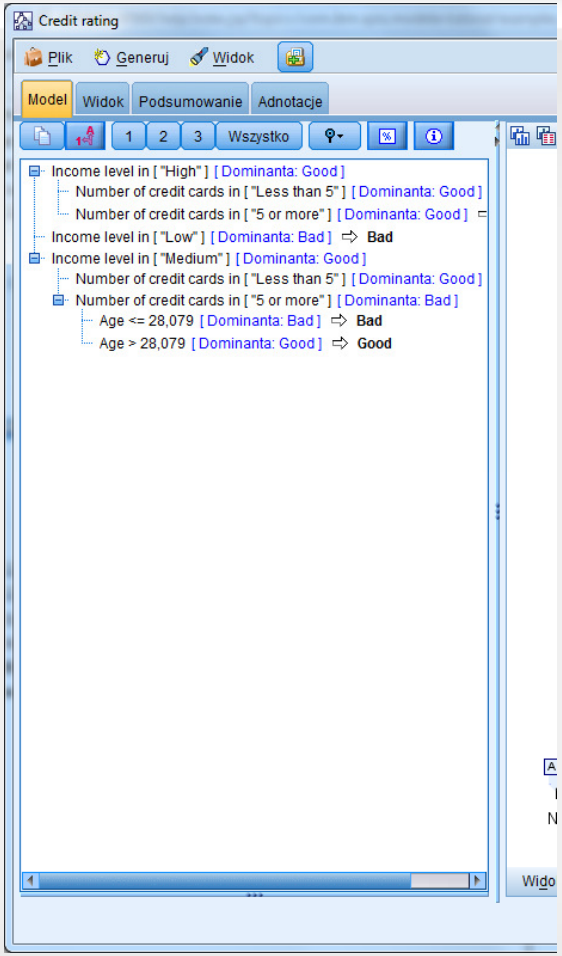
Przeglądanie modelu

- Zbudowany model pojawia się w zakładce *Model* w prawym górnym rogu i zostaje dodany do strumienia – *Przeglądaj* lub *Edytuj*



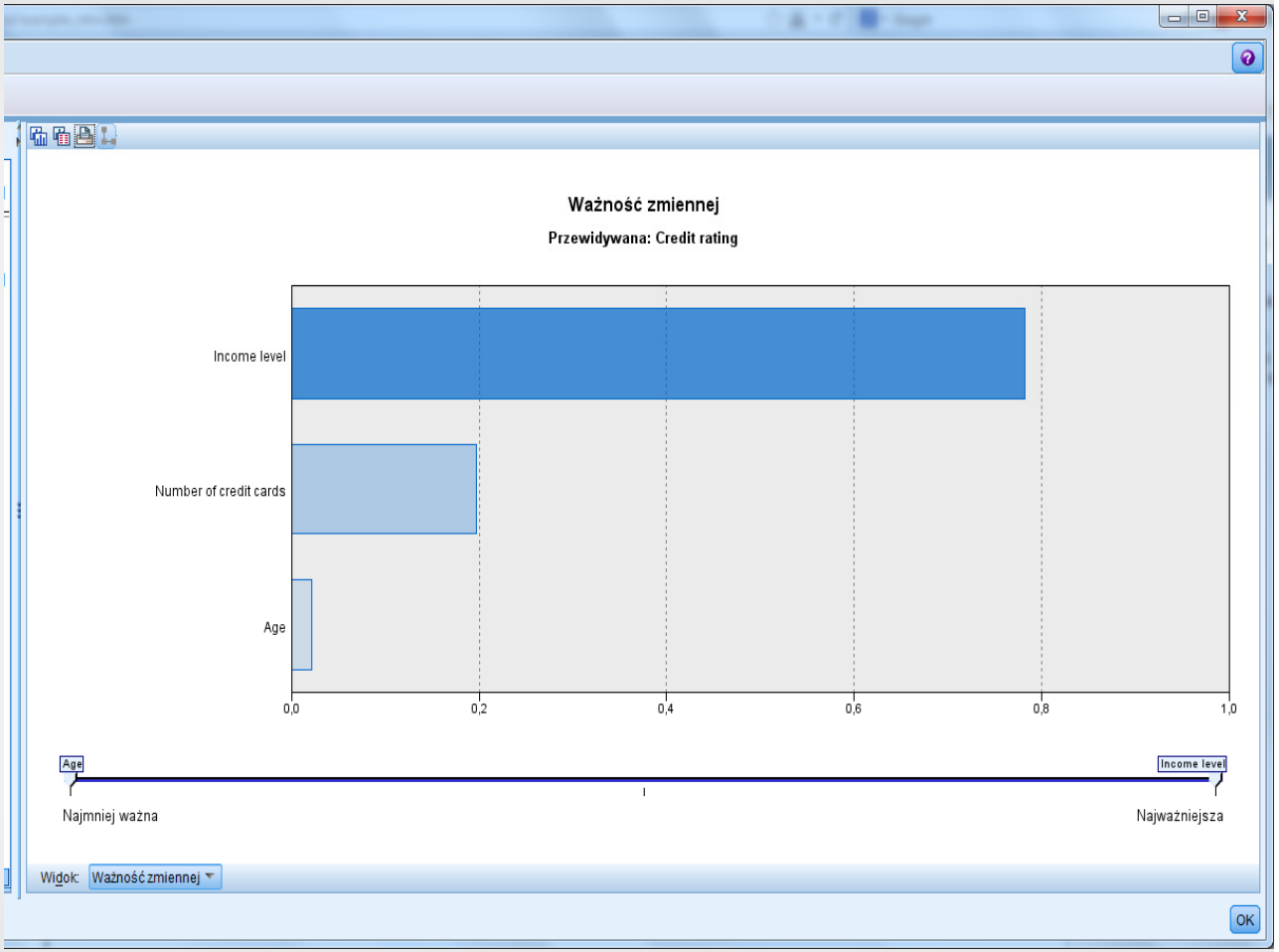
Przeglądanie modelu

- Zakładka *Model* wyświetla detale w postaci zbioru reguł, które będą wykorzystywane do przypisania poszczególnych rekordów do węzłów w oparciu o wartości poszczególnych pól wejściowych.



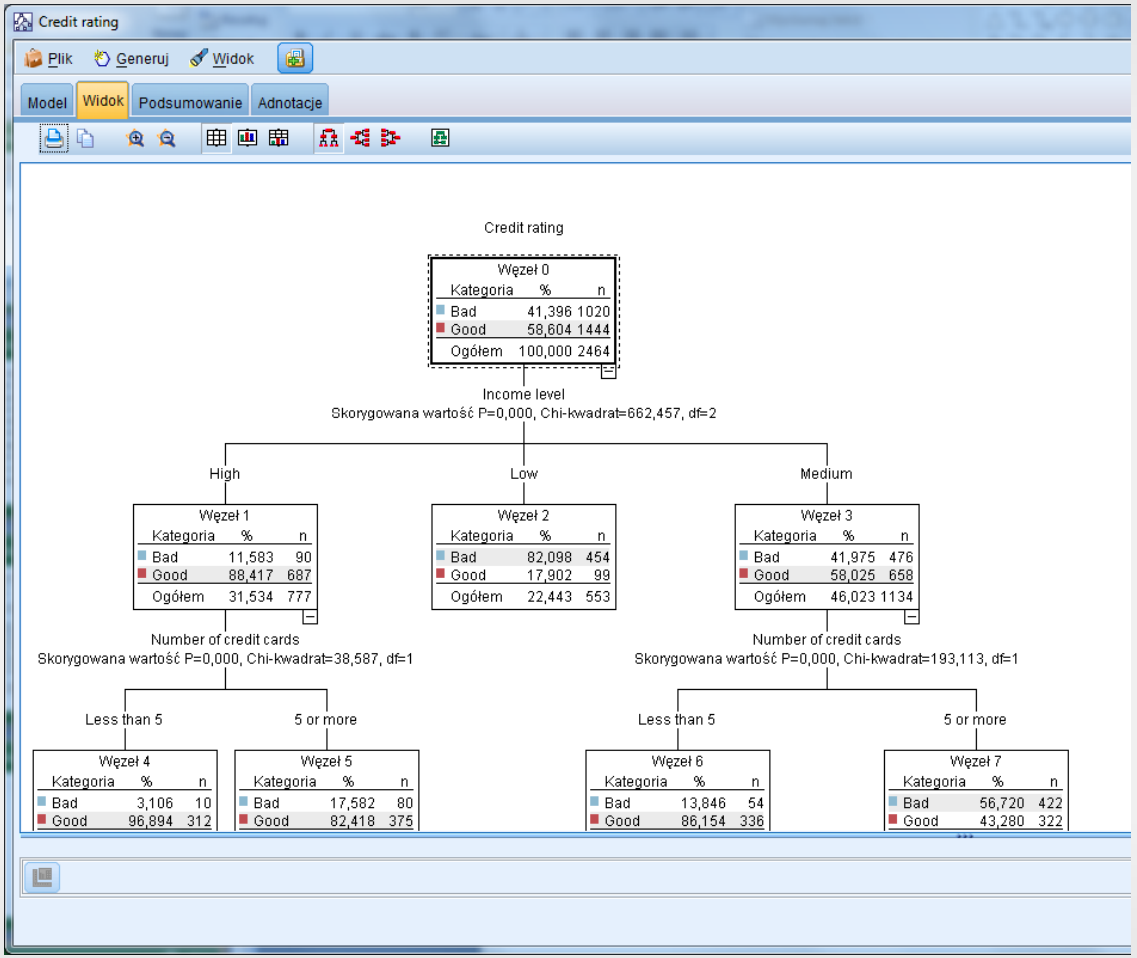
Przeglądanie modelu

- Na prawo – wykres ważności zmiennych - *Income level* jest najbardziej znaczący, a z innych tylko *Number of credit cards*.



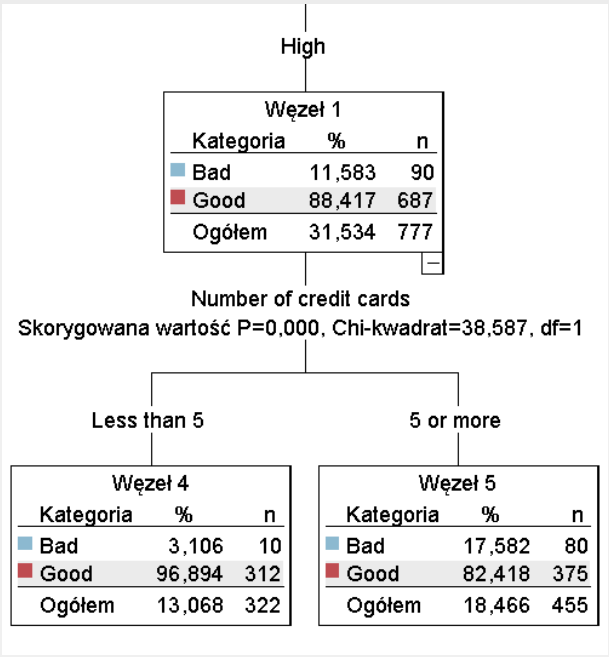
Przeglądanie modelu

- Zakładka *Widok* pokazuje ten sam model ale w postaci drzewa – użyj *Zoom*
- Pierwszy węzeł (*Węzeł 0*) daje podsumowanie dla wszystkich rekordów w zbiorze danych. Ponad 40% przypadków jest zaklasyfikowanych jako *bad risk*. Jest to dość duży odsetek, więc zobaczymy, czy drzewo może dać nam jakieś wskazówki co do tego, jakie czynniki mogą być odpowiedzialne.
- Widać, że pierwszy podział jest w oparciu o *Income level*. Rekordy, gdzie poziom dochodów jest w najniższej kategorii są przypisane do węzła 2, i nie jest zaskoczeniem, kategoria ta zawiera najwyższy procent niespłacanych kredytów. Pożyczki dla tych klientów niosą wysokie ryzyko.
- Ale ponad 17% klientów w tej kategorii nie zachowuje się domyślnie, tak więc predykcja nie zawsze będzie poprawna. A dobry model powinien pozwolić nam przewidzieć najbardziej prawdopodobną odpowiedź dla każdego rekordu na podstawie dostępnych danych.



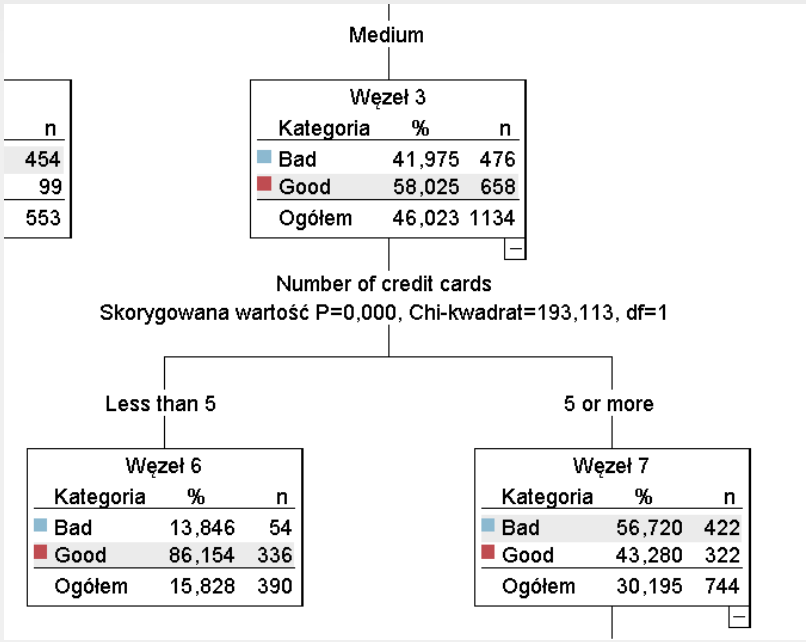
Przeglądanie modelu

- W podobny sposób analizujemy klientów o wysokich dochodach (*Węzeł 1*)
  - zdecydowana większość (88%) to dobrzy klienci, ale więcej niż 1 na 10 to również ci niewypłacalni
    - pytanie: czy możemy ulepszyć nasze kryteria udzielania kredytów w celu zminimalizowania ryzyka?
- Przyjrzyjmy się, jak model dzieli dalej tych klientów na podkategorie (*Węzeł* 4 oraz 5), w oparciu o posiadaną liczbę kart kredytowych. Dla klientów o wysokich dochodach, posiadających mniej niż 5 kart kredytowych, można zwiększyć wskaźnik powodzenia z 88% do 96%.



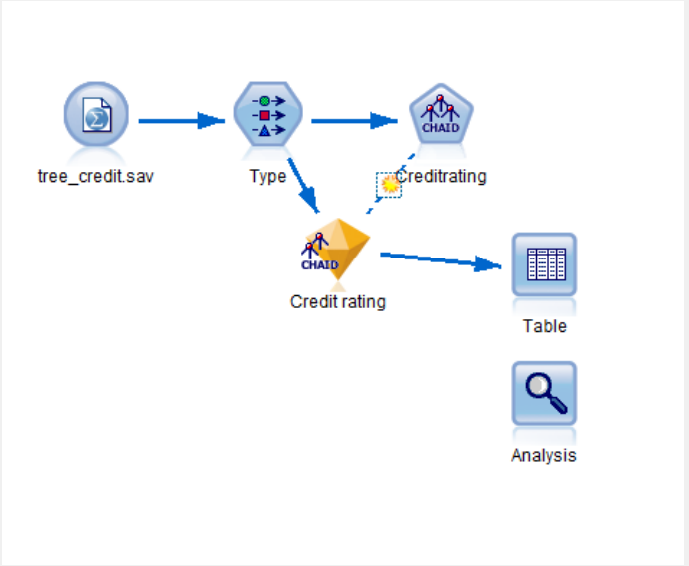
Przeglądanie modelu

- Klienci o średnich dochodach (*Węzeł* 3)? Są jeszcze bardziej równomiernie rozłożeni między oceną dobrą i złą.
- Podobnie, analiza podkategorii (*Węzły* 6 i 7) może tutaj pomóc.
  - wzrost dobrych z 58% do 86%



Ocena modelu

- Do oceny dokładności zbudowanego modelu należy porównać wynik otrzymany z predykcji modelem do rzeczywistych wartości
  - połącz węzeł *Table* do węzła wynikowego modelu i uruchom



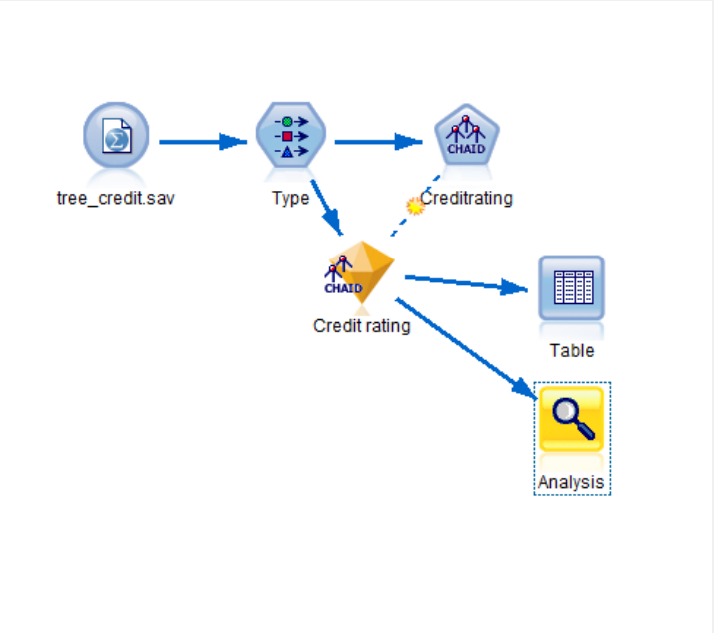
Ocena modelu

- W *$R-Credit rating*, znajduje się przypisanie wartości poczynione prze model. Można je porównać z oryginalnymi wartościami *Credit rating*.
  - nazwa oryginalnego pola jest prefiksowana $R- dla predyktorów oraz $RC- dla wartości zaufania (confidence values). Różne modele używają różne zbiory prefiksów. Wartość zaufania, to wewnętrzne szacowanie modelu w zakresie od 0.0 do 1.0 jak bardzo dokładna jest przewidziana wartość.

Number of credit cards	Education	Car loans	$R-Credit rating	$RC-Credit rating
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Bad	0.806
5 or more	High school	More than 2	Bad	0.820
5 or more	College	None or 1	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.806
5 or more	College	More than 2	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	High school	More than 2	Good	0.563
5 or more	College	None or 1	Bad	0.820
5 or more	High school	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Bad	0.820
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.563
5 or more	College	More than 2	Good	0.823
5 or more	College	More than 2	Bad	0.820

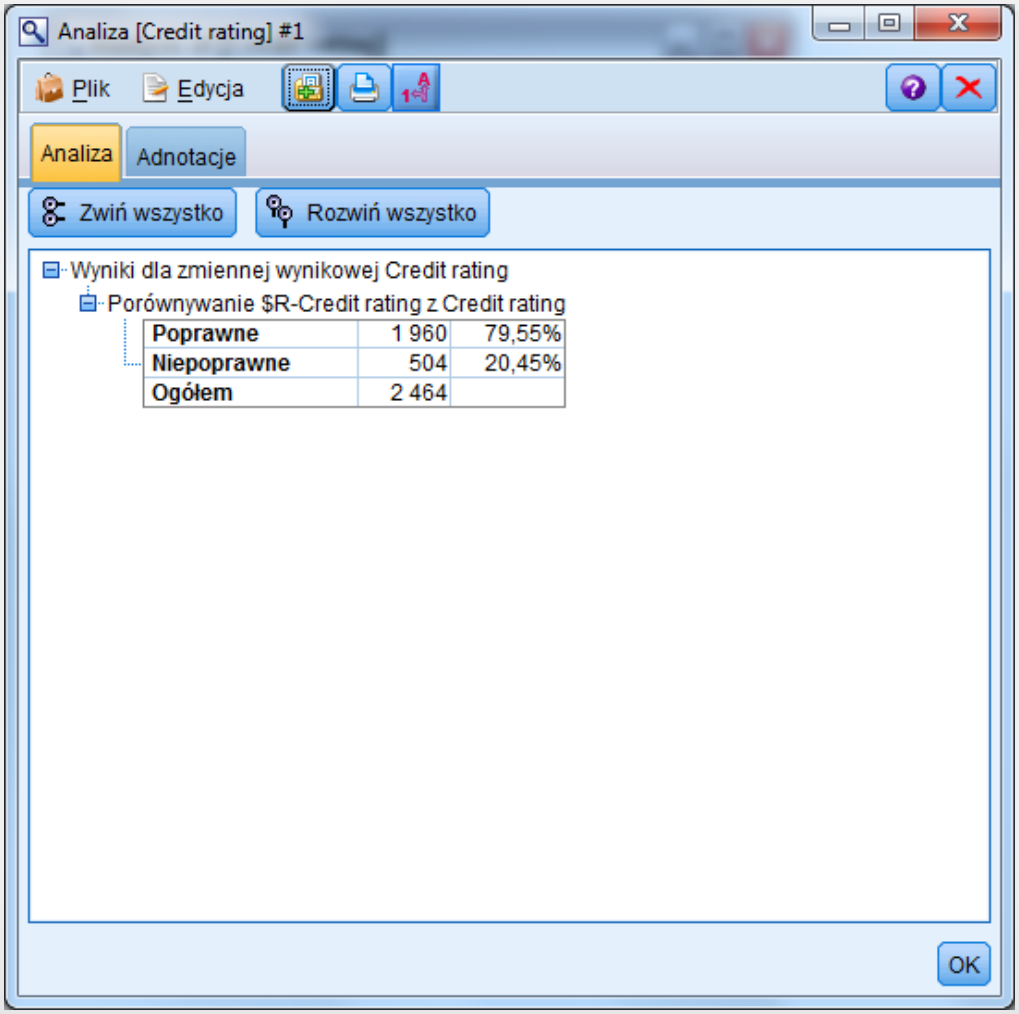
Ocena modelu

- Aby dowiedzieć się dokładnie, ile przewidywań jest poprawnych, możemy przeczytać całą tabelę zliczyć liczbę rekordów, gdzie wartość przewidywanego pola *$R-Credit rating* jest zgodna z wartością *Credit rating*. To samo zrobi węzeł *Analysis* automatycznie.
- ► Połącz ikonkę modelu z węzłem *Analysis*.
- ► Kliknij 2-krotnie węzeł *Analysis* i uruchom *Run*.



Ocena modelu

- Analiza pokazuje, że dla 1960 z 2464 rekordów - ponad 79% - wartości przewidywane przez model odpowiadają rzeczywistym wartością.
- Wynik ten jest ograniczona przez fakt, że rekordy, która użyto do budowy modelu, zostały użyte również do szacowania jego dokładności.
  - w rzeczywistych sytuacjach, trzeba by było użyć węzeł *Partition* do otrzymania rozłącznych zbiorów treningowych i testowych.



Użycie modelu

- Teraz można zmienić węzeł źródłowy *Statistics File* aby wskazywał na inny plik danych, albo można dodać nowe źródło
  - nowy zbiór danych musi zawierać te same pola wejściowe (*Age, Income level, Education* itd.) ale nie pole docelowe *Credit rating*.