



## TWORZENIE I STOSOWANIE MODELU DATA MINING ZA POMOCĄ PRZEPISÓW STATISTICA DATA MINER NA PRZYKŁADZIE WYKRYWANIA NADUŻYĆ

*Tomasz Demski, StatSoft Polska Sp. z o.o.*

Narzędzia zgłębiania danych (*data mining*) rozwijane są od wielu lat i obecnie dostępnych jest wiele dojrzałych metod dostosowanych do rozmaitych zadań i wymagań dotyczących stosowania uzyskiwanych rozwiązań. Równocześnie postęp w szybkości działania komputerów i pojemności pamięci masowych powoduje, że coraz mniejszą przeszkodę w analizie stanowi wielkość danych. W związku z tym coraz ważniejsze jest ułatwienie wykonywania całego procesu prowadzącego od surowych danych do wiedzy, z uwzględnieniem przygotowania i oczyszczenia danych oraz stosowania modeli dla nowych przypadków. Właśnie z myślą o ułatwieniu wykonywania analiz w praktyce przygotowano *Przepisy STATISTICA Data Miner*.

*Przepisy STATISTICA Data Miner* umożliwiają rozwiązywanie zadań ukierunkowanego data mining poprzez wykonanie określonej sekwencji operacji. Użytkownik jest prowadzony przez całą analizę: od wskazania danych, poprzez ich sprawdzenie, oczyszczenie i przekształcenie, zbudowanie modelu i zastosowanie go dla nowych przypadków. Na koniec można utworzyć raport podsumowujący wszystkie wykonane działania

Przykład poświęcony będzie wykorzystaniu *Przepisów STATISTICA Data Miner* do tworzenia modelu i stosowania go dla nowych danych. Zadaniem będzie wskazanie transakcji w sklepie internetowym, które najprawdopodobniej wiążą się próbą wyłudzenia. Do budowy modelu przewidującego, czy transakcja jest, czy nie jest nadużyciem, wykorzystane zostaną drzewa klasyfikacyjne, drzewa wzmacniane (*boosted trees*) oraz różne architektury sieci neuronowych (informacje o tych metodach można znaleźć w podręcznikach [2] i [3]). Modele uzyskane różnymi metodami zostaną ocenione, a najlepszy z nich zostanie wdrożony w *STATISTICA Enterprise* (dokładniejszy opis tego systemu można znaleźć w [1]).

### Omówienie danych

Będziemy analizować dane o transakcjach zarejestrowanych przez sklep internetowy (dane te służyły pierwotnie jako zbiór uczący dla uczestników konkursu data mining DM Cup 2005; <http://www.data-mining-cup.com>).



Każda transakcja została przydzielona do jednej z dwóch klas:

- ◆ Wysokie ryzyko oszustwa
- ◆ Niskie ryzyko oszustwa

W naszym przykładzie wykorzystamy nieco zmodyfikowane dane. Przede wszystkim wprowadzono do nich zmienne pochodne obliczane na podstawie zmiennych istniejących w danych surowych. W używanym zbiorze danych znajdują się następujące zmienne:

- ◆ *Klient\_ID* – identyfikator klienta (zmienna nieuwzględniana w modelu)
- ◆ *TARGET* – identyfikator poziomu ryzyka: *Tak* oznacza wysokie ryzyko, *Nie* oznacza niski poziom ryzyka.
- ◆ *Email* – informuje, czy wraz z zamówieniem podano adres e-mail.
- ◆ *Telefon* – informuje, czy wraz z zamówieniem podano numer telefonu.
- ◆ *Poadana\_DU* – klient podał datę urodzenia.
- ◆ *Adres\_ten\_sam* – informuje, czy adres dostarczenia i faktury jest ten sam.
- ◆ *Zamówił newsletter* – informuje, czy wraz z zamówieniem został zamówiony newsletter.
- ◆ *Płatność* – określa wybrany przez klienta sposób płatności.
- ◆ *Rodzaj\_karty* – informuje o typie karty, jeśli w zmiennej *Płatność* występuje wartość *Karta kredytowa*.
- ◆ *POW* – pozostały okres ważności karty kredytowej w latach.
- ◆ *Zgodność nazwiska* – informuje, czy nazwisko właściciela konta lub karty kredytowej jest zgodne z nazwiskiem adresata.
- ◆ *Kwota* – wartość zamówienia w euro.
- ◆ *Dzień* – dzień tygodnia, w którym dokonano zamówienie.
- ◆ *Godzina* – godzina złożenia zamówienia.
- ◆ *Liczba artykułów* – liczba zamówionych artykułów.
- ◆ *Długość sesji* – długość trwania sesji, podczas której dokonano zakupu w minutach.
- ◆ *Nowy klient* – informuje, czy kupował dokonał nowy klient.
- ◆ *Liczba poprzednio zamówionych* – liczba artykułów zamówionych w poprzednim zamówieniu.
- ◆ *Wartość poprzedniego zamówienia* – wartość poprzedniego zamówienia.
- ◆ *Bieżący poziom monitu*.
- ◆ *Najwyższy poziom monitu*.
- ◆ *Wiek* – wiek klienta.
- ◆ *LiczbaMiesiący* – liczba miesięcy ważności karty.
- ◆ *Niezgodności* – zmienna informująca o wystąpieniu pewnych rzadkich zdarzeń m.in. niezgodności podanego miasta i kodu, braku miasta dostawy, czy w ciągu ostatnich 3 dni wysłano zamówienie z tego samego komputera (identyfikacja przez IP i plik *cookie*) itp.

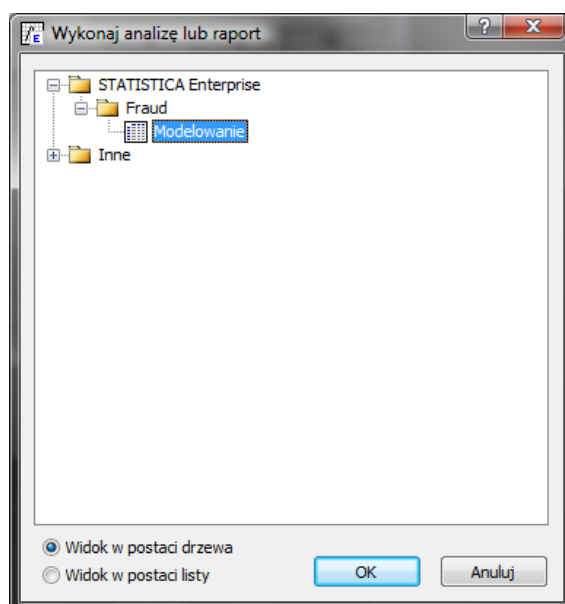


W oryginalnym zbiorze danych około 5,8 % przypadków trafiło do grupy wysokiego ryzyka. Przy tak nie zrównoważonej częstotliwości grup trudno jest zbudować poprawny model. Najprostszym sposobem rozwiązania tej trudności jest wylosowanie z danych próby o zrównoważonych licznosciach. Chociaż można to zrobić w *Przepisach Data Miner*, my skorzystamy z wcześniej przygotowanej próby o zrównoważonych licznosciach.

Dane przygotowane do uczenia zostały zapisane w bazie danych, a odczytamy je z niej, korzystając ze *STATISTICA Enterprise* (zob. [1]).


## Tworzenie modelu

Zaczynamy od pobrania danych. W *STATISTICA Enterprise* istnieje już konfiguracja analizy wykonującej pobranie odpowiednich danych (tworzenie konfiguracji analiz przedstawiono w [1]). Po uruchomieniu programu w oknie *Wykonaj analizę lub raport* (widocznym poniżej) wskazujemy konfigurację *Wykrywanie przyczyn* i klikamy OK.



Rys. 1. Pobieranie danych do modelowania.

Program wczyta dane z bazy danych i wyświetli je w arkuszu *STATISTICA*. Zauważmy, że w praktyce uzyskanie danych do tworzenia i stosowania modelu często jest złożone. Dlatego przygotowanie i zapisanie szablonu pobierania danych daje bardzo duże korzyści: analityk nie musi wiedzieć, jak wydobyć dane ze źródłowych systemów bazodanowych, a zamiast wykonywać wiele operacji ręcznie, po prostu wybiera potrzebną mu konfigurację i klika OK.

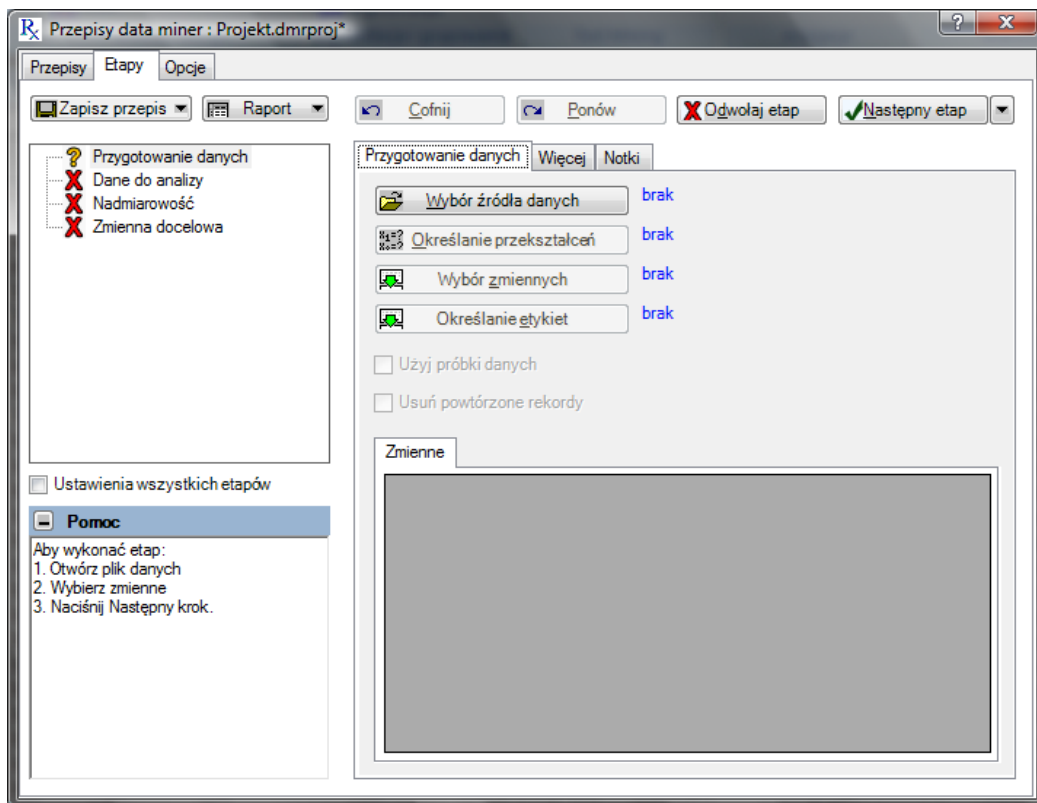
Po wczytaniu danych uruchamiamy *Przepisy Data Miner*. W tym celu przechodzimy na kartę *Data Mining* i naciskamy przycisk  na wstążce.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	Klient_ID	TARGET	Email	Telefon	PodanaDU	Adres_ten_sam	Zamówił	Płatność	Rodzaj_karty	POW	Zgodność	Kwota	Dzień	Godzina	Liczba
1	37	Nie	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	3	Brak danych	43,5	Sro	7	
2	47	Tak	Nie	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	2	Brak danych	68,8	Piat	23	
3	70	Tak	Nie	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	3	Brak danych	46,5	Czwa	8	
4	113	Nie	Tak	Nie	Nie	Nie	Nie	Karta lojalnościowa	Kundenkarte	2	Tak	42,8	Wt	14	
5	164	Nie	Tak	Tak	Nie	Nie	Nie	Polecenie zapłaty	Nie dotyczy	3	Tak	23,5	Sro	23	
6	192	Tak	Tak	Tak	Tak	Tak	Nie	Faktura	Nie dotyczy	2	Brak danych	92,7	Niedz	12	
7	195	Nie	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	3	Brak danych	29,99	Wt	5	
8	229	Tak	Tak	Nie	Nie	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	91,1	Wt	4	
9	233	Nie	Nie	Nie	Nie	Tak	Nie	Faktura	Nie dotyczy	3	Brak danych	23,6	Niedz	11	
10	247	Nie	Tak	Tak	Tak	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	18,5	Sro	21	
11	256	Tak	Nie	Nie	Nie	Nie	Nie	Polecenie zapłaty	Nie dotyczy	3	Tak	8,5	Sob	10	
12	288	Tak	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	12,6	Sro	15	
13	338	Tak	Nie	Nie	Nie	Tak	Nie	Faktura	Nie dotyczy	2	Brak danych	6,5	Pon	11	
14	443	Tak	Tak	Nie	Tak	Tak	Nie	Polecenie zapłaty	Nie dotyczy	2	Tak	19,99	Niedz	8	
15	457	Nie	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	47,7	Piat	23	
16	463	Tak	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	35,55	Sob	7	
17	465	Nie	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	1	Brak danych	82,3	Sro	9	
18	559	Tak	Tak	Nie	Tak	Tak	Nie	Karta kredytowa	Visa	2	Nie	6,5	Pon	15	
19	576	Tak	Tak	Nie	Nie	Tak	Nie	Karta kredytowa	Visa	2	Tak	8,7	Niedz	18	
20	577	Nie	Tak	Nie	Tak	Tak	Tak	Faktura	Nie dotyczy	2	Brak danych	9,99	Sob	2	
21	602	Nie	Tak	Nie	Nie	Nie	Nie	Karta kredytowa	Eurocard	1	Tak	5,2	Pon	3	
22	615	Tak	Tak	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	3	Brak danych	14,5	Pon	5	
23	673	Tak	Nie	Nie	Tak	Tak	Nie	Faktura	Nie dotyczy	2	Brak danych	68,4	Niedz	22	
24	704	Tak	Tak	Nie	Tak	Nie	Nie	Karta kredytowa	Amex	3	Nie	14,5	Piat	21	
25	801	Nie	Tak	Nie	Tak	Tak	Nie	Karta kredytowa	Eurocard	3	Tak	4,1	Niedz	5	
26	810	Nie	Nie	Nie	Nie	Tak	Nie	Faktura	Nie dotyczy	2	Brak danych	18,5	Sob	10	

Rys. 2. Dane pobrane do STATISTICA.

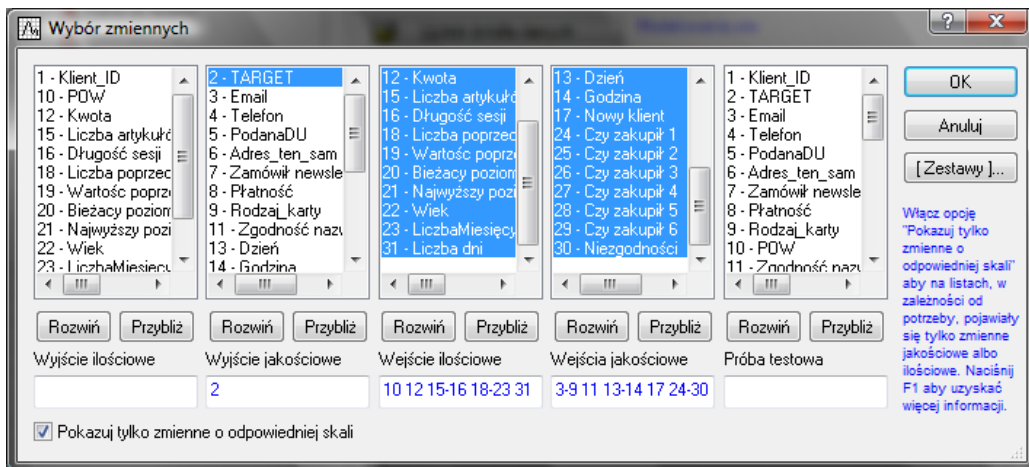
Na ekranie pojawi się okno *Przepisów*, w którym klikamy przycisk *Nowy*.



Rys. 3. Okno *Przepisów Data Miner*.

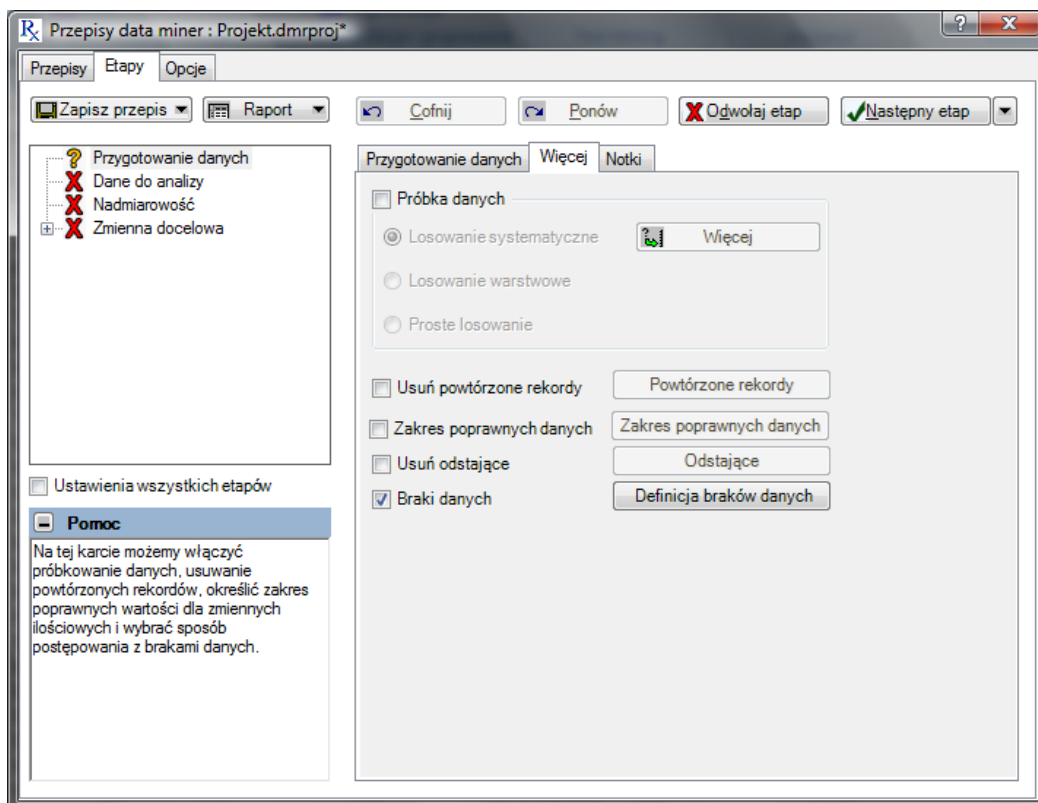


Tworzenie modelu zaczynamy od wskazania źródła danych: klikamy przycisk *Wybór źródła danych* i wskazujemy arkusz *Modelowanie*. Możemy teraz wybrać zmienne do modelowania, w tym celu naciskamy przycisk *Wybór zmiennych*.



Rys. 4. Wybór zmiennych.

Jako wyjście (zmienną zależną) wskazujemy *Target*, a jako wejścia wskazujemy wszystkie zmienne, zgodnie z podpowiedzią programu. Zauważmy, iż program automatycznie przypisał zmienne do klas jakościowa i ilościowa.

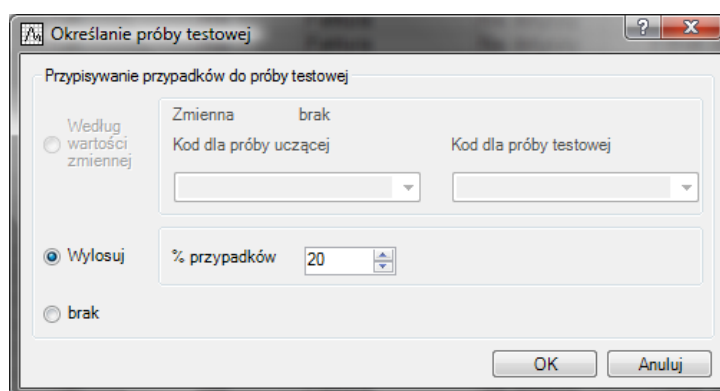


Rys. 5. Czyszczenie danych.

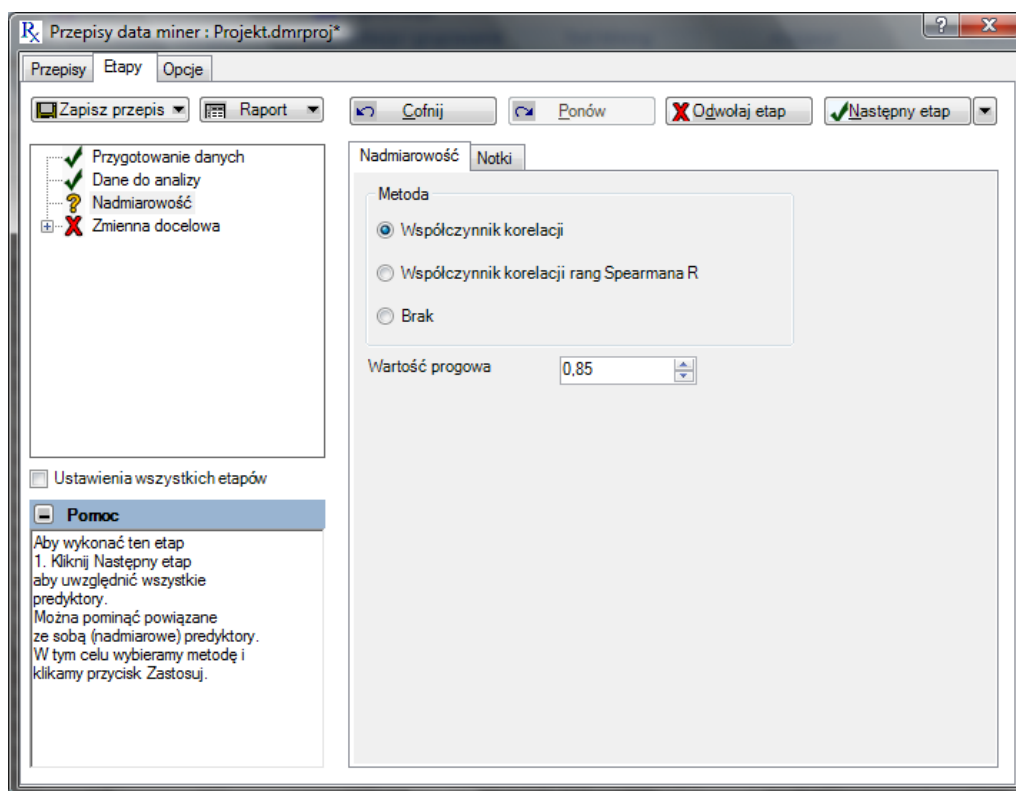


Na karcie *Więcej* (zob. rys. 5) możemy określić sposób czyszczenia danych. Do dyspozycji mamy losowe próbkowanie (możemy tu m.in. wylosować próbę zbilansowaną), usuwanie powtórzonych rekordów, wykrywanie obserwacji poza dopuszczalnym zakresem, obsługę nietypowych obserwacji i braków danych. W naszym projekcie wykorzystamy tylko obsługę braków danych.

Do kolejnego etapu przechodzimy, naciskając przycisk *Następny etap*. Program sprawdzi dane i, jeśli nie wykryje żadnych problemów, przejdziemy od kolejnego etapu, w którym określamy podział na próbę uczącą i testową. Próba ucząca zostanie użyta do znalezienia modelu, a testowa do oceny jego działania. Na karcie *Dane do analizy* klikamy przycisk *Określanie próby testowej*. Na ekranie otworzy się przedstawione niżej okno, w którym wybieramy tworzenie losowej próby testowej złożonej z 20% przypadków.



Rys. 6. Określanie próby testowej.

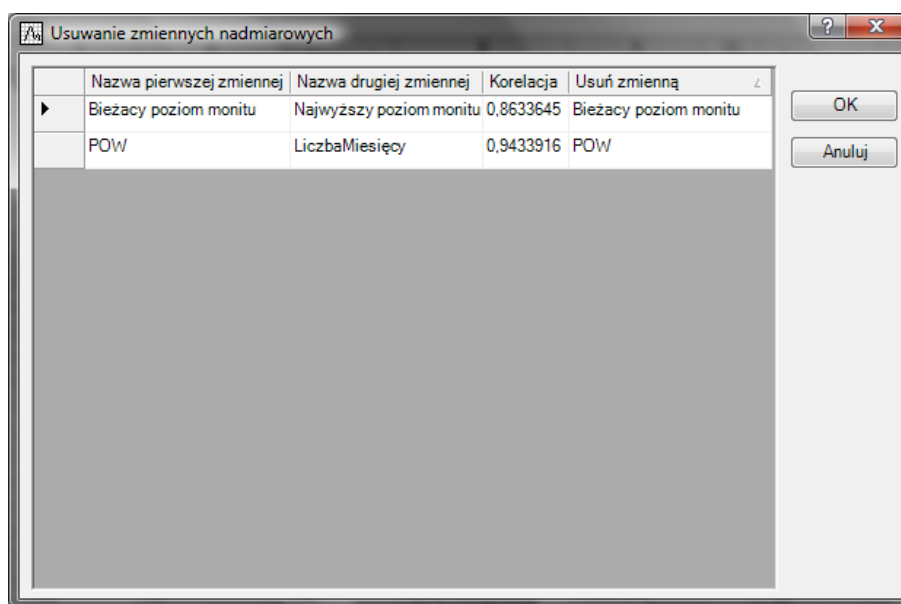


Rys. 7. Nadmiarowość.



Po naciśnięciu przycisku *Następny etap* przechodzimy do kolejnej fazy modelowania: usuwania zmiennych nadmiarowych. W praktycznych zastosowaniach dosyć często zdarza się, że dane zawierają zmienne przenoszące tę samą informację. Na etapie *Nadmiarowość* możemy wyeliminować takie zmienne z modelu. W naszym przypadku jako zbędne uznamy zmienne których współczynnik korelacji wynosi co najmniej 0,85.

Po kliknięciu *OK* program znajdzie pary zmiennych o współczynniku korelacji co najmniej 0,85 i zaproponuje usunięcie zbędnych zmiennych.



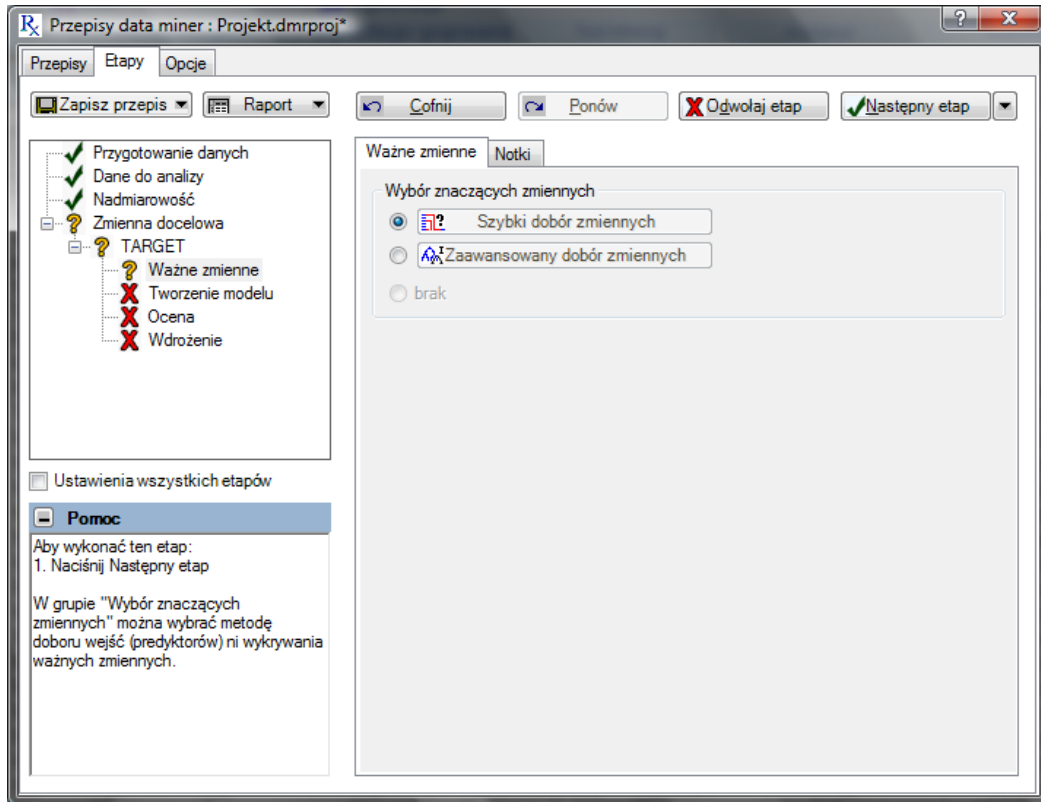
Rys. 8. Wybór działań wobec nadmiarowych zmiennych.

W naszym przypadku nadmiarowe zmienne to:

- ◆ *POW* i *LiczbaMiesiący*.
- ◆ *Bieżący poziom monitu* i *Najwyższy poziom monitu*.

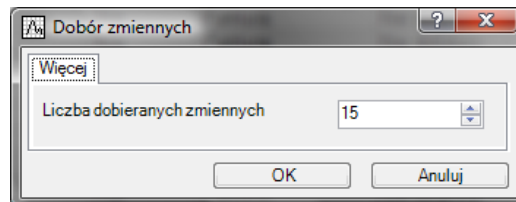
Pierwsza para przenosi tę samą informację, z tym, że raz wyrażoną w latach (z zaokrągleniem do pełnego roku), a raz w miesiącach. Rozsądne jest zachowanie tylko jednej z tych zmiennych, tej która jest dokładniejsza, czyli zmiennej *LiczbaMiesiący*. Również w przypadku drugiej pary zmiennych rozsądnym wydaje się zachowanie tylko jednej z nich: *Najwyższy poziom monitu*. Oznacza to przyjęcie propozycji programu, a więc klikamy *OK*.

Kolejny etap to odrzucenie zmiennych nie wpływających na wielkość, którą chcemy przewidywać. Zauważamy, iż w wielu zastosowaniach jest to kluczowy problem: zdarza się, że mamy dosłownie tysiące zmiennych, z których tylko kilka jest naprawdę istotnych.



Rys. 9. Dobór ważnych zmiennych.

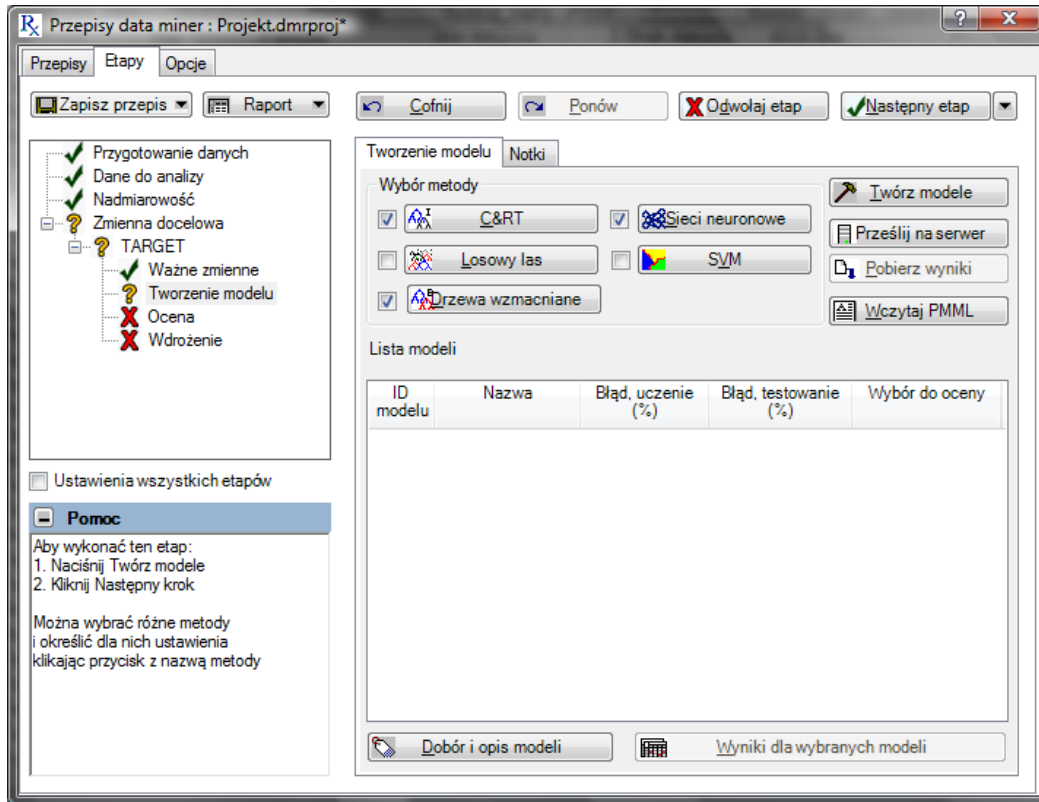
Zastosujemy metodę domyślną: *Szybki dobór zmiennych*. Klikamy przycisk *Szybki dobór zmiennych* i ustawiamy wybór 15 najlepszych predyktorów.



Rys. 10. Dobór ważnych zmiennych.

Następny etap *Przepisu* to tworzenie modeli. Najpierw utworzymy modele przy domyślnych ustawieniach. Po kliknięciu *Twórz model* program znajdzie najlepsze modele metodami drzew decyzyjnych (*C&RT*), drzewami wzmacnianymi (*boosted trees*) oraz różnymi architekturami sieci neuronowych (opis metod można znaleźć w [2] i [3]).

Program znajdzie najlepsze modele i wyznaczy stopę błędów w próbach uczącej i testowej. Obliczenia najdłużej trwają dla sieci neuronowych, warto jednak zauważyć, że w tym wypadku sprawdzane jest kilka architektur sieci neuronowych (domyślnie 5) i wybierana jest najlepsza z nich.



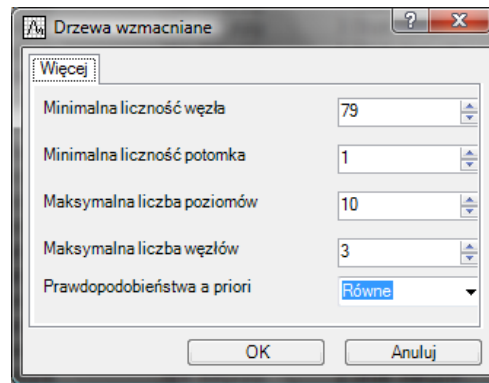
Rys. 11. Tworzenie modeli.

W tabeli poniżej znajduje się zestawienie stóp błędów dla różnych metod. Model możemy uznać za poprawny, gdy trafność przewidywania w obu próbach jest podobna, a najlepszy model to ten, który ma najmniejszy udział błędnych przewidywań w próbie testowej. W naszym przypadku najlepszy wydaje się model sieci neuronowej.

Metoda	Stopa błędów (%)	
	Próba ucząca	Próba testowa
C&RT	28,80	41,18
Drzewa wzmacniane	29,99	38,11
Sieci neuronowe	33,31	34,78

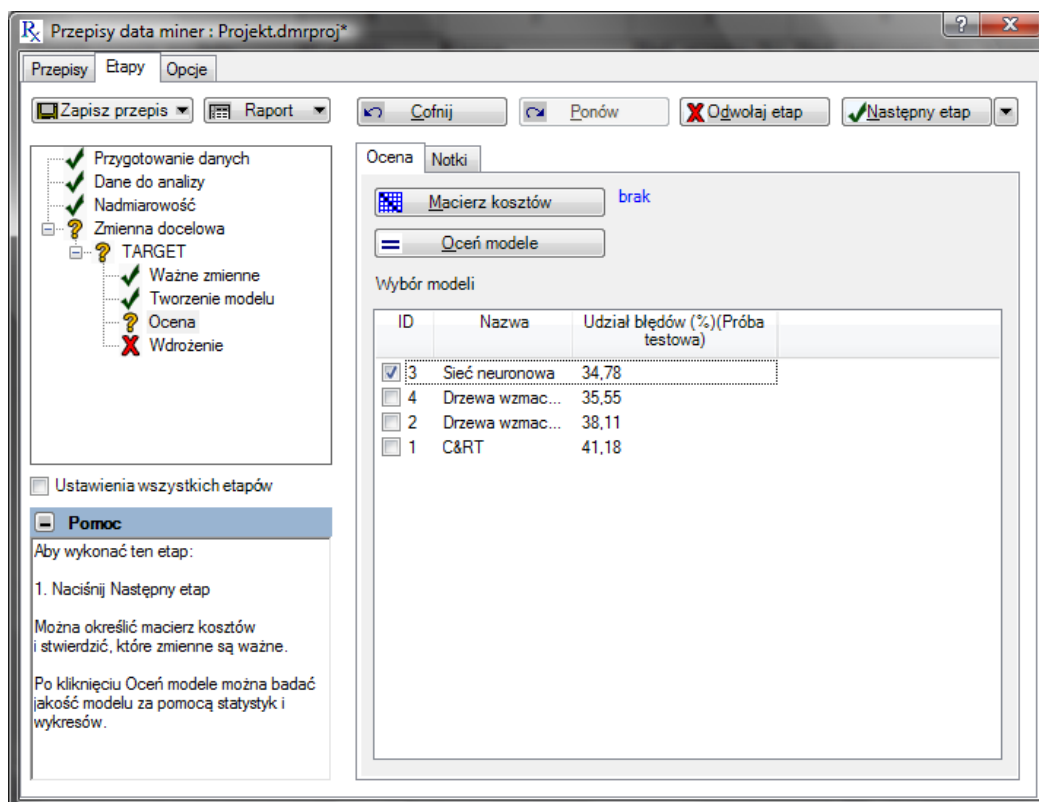
Zauważmy, że modele są uczone na próbach tworzonych losowo. Ponadto podczas tworzenia pewne ustawienia dobierane są losowo (np. początkowe wagi sieci neuronowych). W związku z tym wyniki uzyskiwane przy wielokrotnym tworzeniu modeli mogą być nieco różne.

Możemy spróbować poprawić działanie modeli. W szczególności drzewa wzmacniane na ogół dają bardzo dobre wyniki, a my dostaliśmy przeuczony model. Aby utworzyć dodatkowy model drzew wzmacnianych ze zmienionymi ustawieniami klikamy przycisk *Drzewa wzmacniane* i w oknie ustawień w polu *Maksymalna liczba węzłów* wpisujemy 3. Ponieważ chcemy ponownie dopasować tylko model drzew wzmacnianych, to odznaczamy pozostałe metody, po czym naciskamy przycisk *Twórz modele*.



Rys. 12. Zmodyfikowane ustawienia dla drzew wzmacnianych.

Utworzony zostanie nowy model o numerze 4 (obok trzech zbudowanych wcześniej przy domyślnych ustawieniach). Nowy model ma stopę błędów w próbie uczącej równą 33,31% (tyle samo co sieci), a w próbie testowej 35,55% i jest to praktycznie równoważne wynikowi uzyskanemu dla sieci neuronowych. Jeśli modelowanie powtórzymy kilka razy (za każdym razem losowo dzieląc dane na uczące i testowe), to uzyskamy podobną stopę błędów, przy czym czasem nieco lepsze są sieci, a czasem drzewa.



Rys. 13. Ocena modeli.

W celu podjęcia finalnej decyzji możemy sprawdzić, jak wyglądają błędy dla poszczególnych klas. I tak w próbie testowej wśród przypadków wskazanych przez drzewa wzmacniane jako oszustwo 62,0% faktycznie należało do tej klasy, natomiast dla sieci frakcja ta wynosi minimalnie mniej, bo 61,7%. Gdybyśmy do wyboru modelu zastosowali kryterium

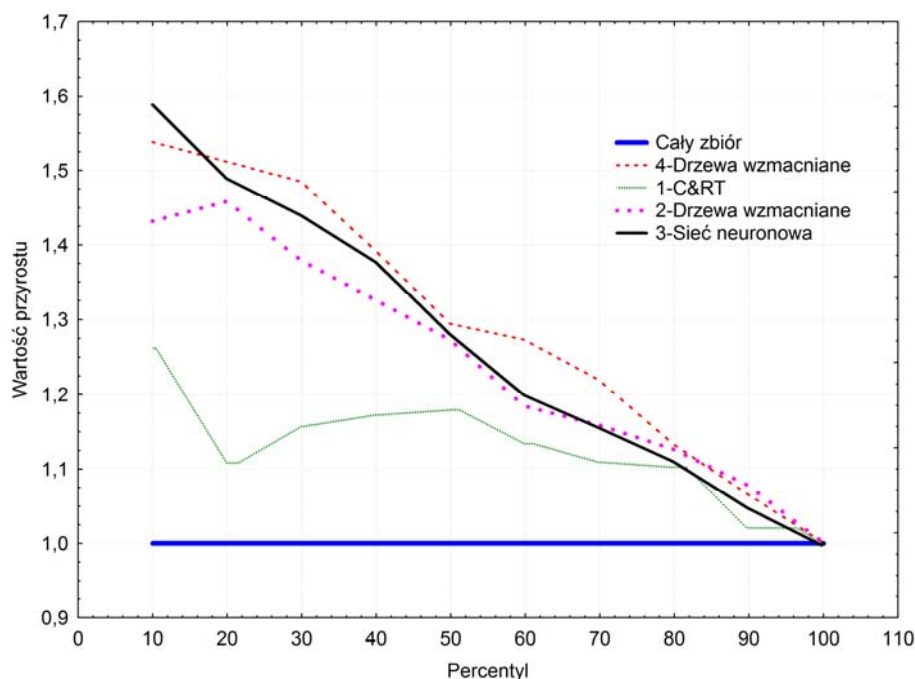


maksymalnej frakcji faktycznych oszustw w zbiorowości wskazanej przez model, to należałoby wybrać drzewa wzmacniane, a nie sieci.

Dokładniejszą ocenę działania modeli możemy wykonać na kolejnym etapie *Przepisu*. Przechodzimy do niego po naciśnięciu przycisku *Następny etap*. Następnie klikamy przycisk *Oceń modele*.

Wygodnym sposobem oceny modeli jest wykres przyrostu (ang. *lift chart*). Widzimy go dla naszych modeli poniżej. Wykres przyrostu pokazuje, o ile częściej niż w danych źródłowych przewidywana klasa występuje w próbie wskazanej przez dany model. Wartość ta jest wyznaczana dla różnego stopnia pewności przewidywań modelu: dla 10% przypadków, które wg modelu najprawdopodobniej należą do klasy, 20% takich przypadków itd. Uzyskana w ten sposób krzywa powinna w miarę gładko spadać od największej wartości do 1: gwałtowne skoki w górę sugerują, że model jest nieodpowiedni (oznaczają one, że model niezgodnie z rzeczywistością przewiduje szansę przynależności do klasy: tam gdzie wg modelu jest ona mniejsza, w rzeczywistości jest większa).

Wszystkie modele oprócz metody C&RT zachowują się poprawnie. Drzewa wzmacniane z poprawionymi ustawieniami (model nr 4) i sieci mają bardzo podobny przebieg krzywej przyrostu, gdybyśmy np. chcieli wybrać do sprawdzenia 30% przypadków o największym wg modelu prawdopodobieństwie przynależności do klasy *Tak*, to lepiej działają drzewa.



Rys. 14. Wykres przyrostu.

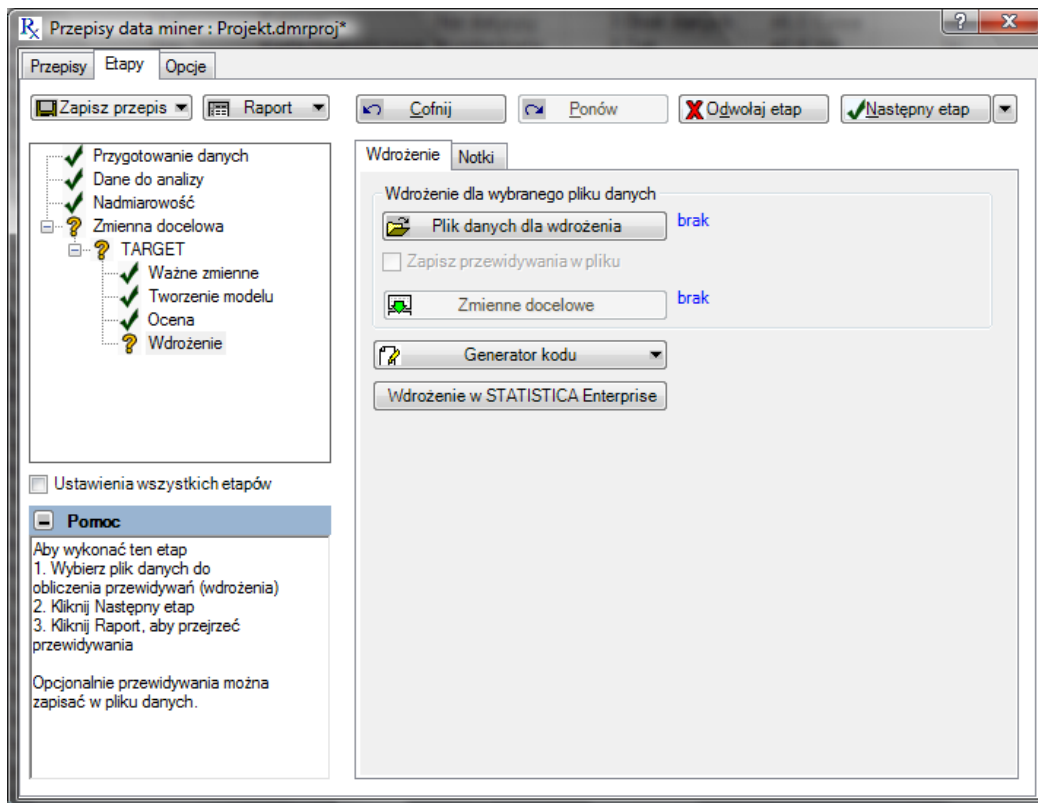
Zgodnie z podpowiedzią programu do stosowania wybierzmy model o mniejszej stopie błędu w próbie testowej. Pozostawiamy zaznaczenie pozycji *Sieci neuronowe* (por. rys. 13) i naciskamy przycisk *Następny etap*. Na ekranie otworzy się okno etapu *Wdrożenie* (zob. rys. 15).



Do dyspozycji mamy:

- ◆ zastosowanie modelu dla pliku danych,
- ◆ utworzenie kodu modelu w językach C i PMML,
- ◆ zastosowanie modelu w *STATISTICA Enterprise*.

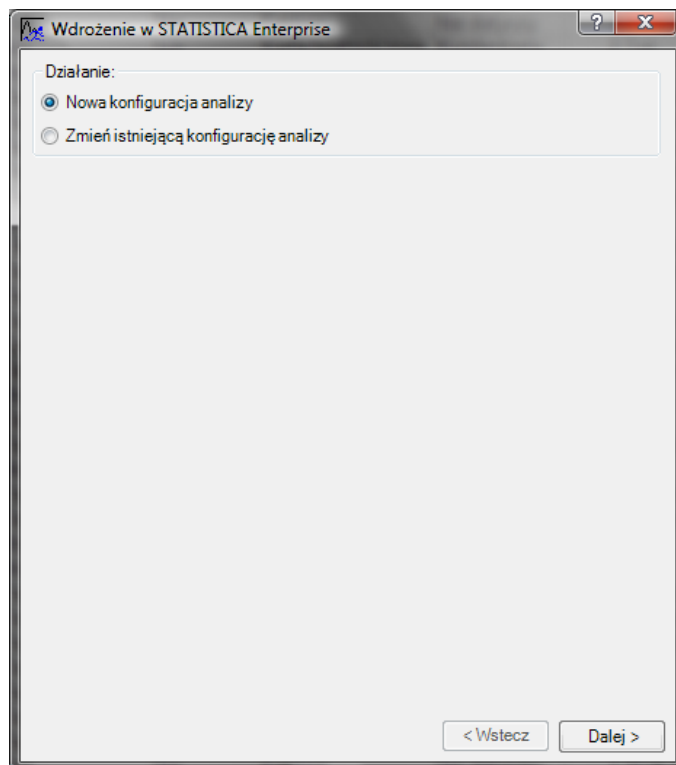
My skorzystamy z ostatniej możliwości. Taki sposób wdrożenia jest korzystny i wygodny, ponieważ wszyscy uprawnieni użytkownicy systemu mają łatwo dostępne przewidywania modelu: aby je uzyskać, wystarczy po prostu wybrać odpowiedni szablon analizy i uruchomić go.



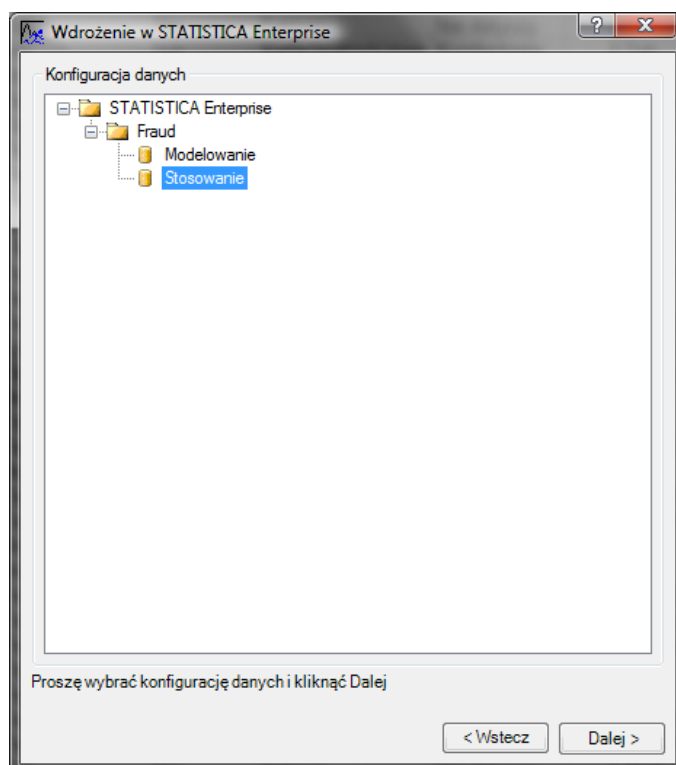
Rys. 15. Stosowanie modelu.

Naciskamy przycisk *Wdrożenie w STATISTICA Enterprise*. Na ekranie otworzy się okno, w którym decydujemy, czy tworzymy nową konfigurację, czy zmieniamy już istniejącą. Utworzymy nową konfigurację analizy.

W następnym etapie wskazujemy źródło danych, dla których stosowany będzie model, tzw. konfigurację danych. W naszym przypadku dla wdrożenia modeli przygotowano konfigurację danych *Stosowanie*.

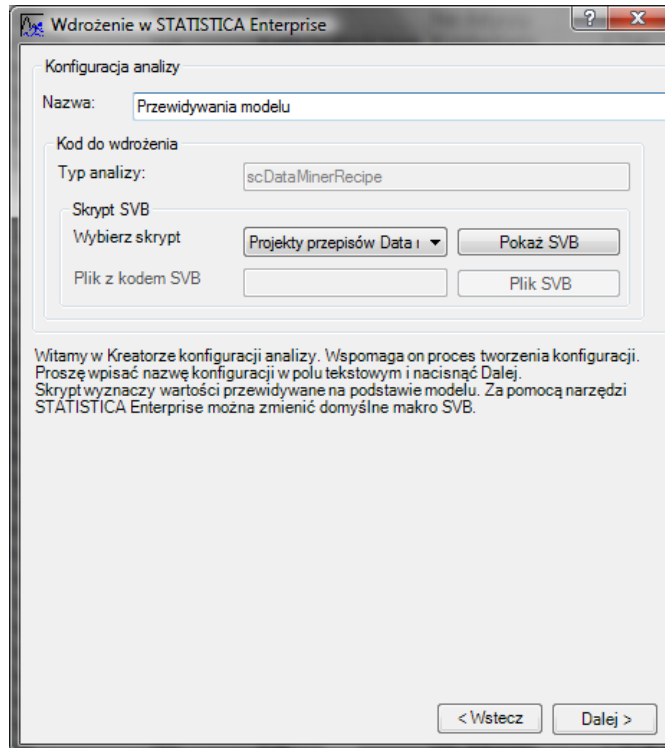


Rys. 16. Pierwszy etap wdrożenia.



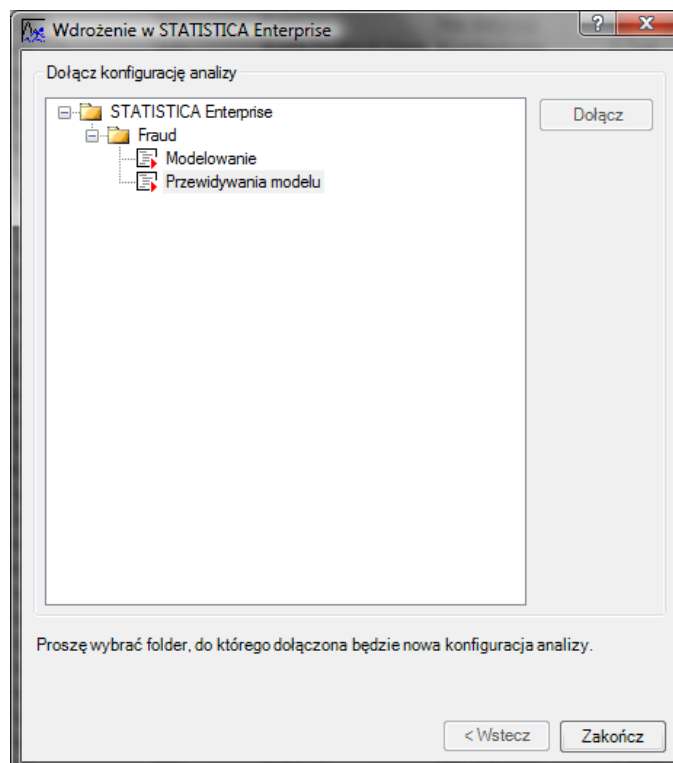
Rys. 17. Wybór konfiguracji danych.

Następny krok to nazwanie konfiguracji wdrażającej model, nazwijmy ją *Przewidywania modelu*.



Rys. 18. Nazwanie konfiguracji analizy.

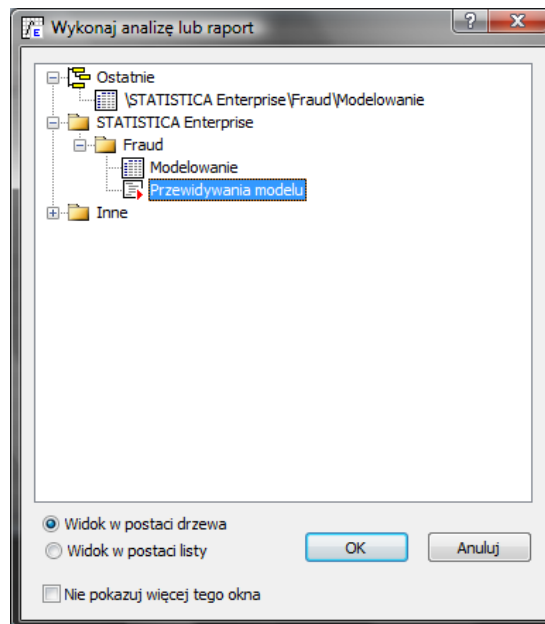
Po podaniu nazwy konfiguracji określamy prawa dostępu do niej i na koniec jej lokalizację w systemie. Po wskazaniu foldera klikamy *Zakończ* – od tego momentu użytkownicy systemu mogą stosować model dla nowych danych „jednym kliknięciem”.



Rys. 19. Umieszczenie konfiguracji analizy.

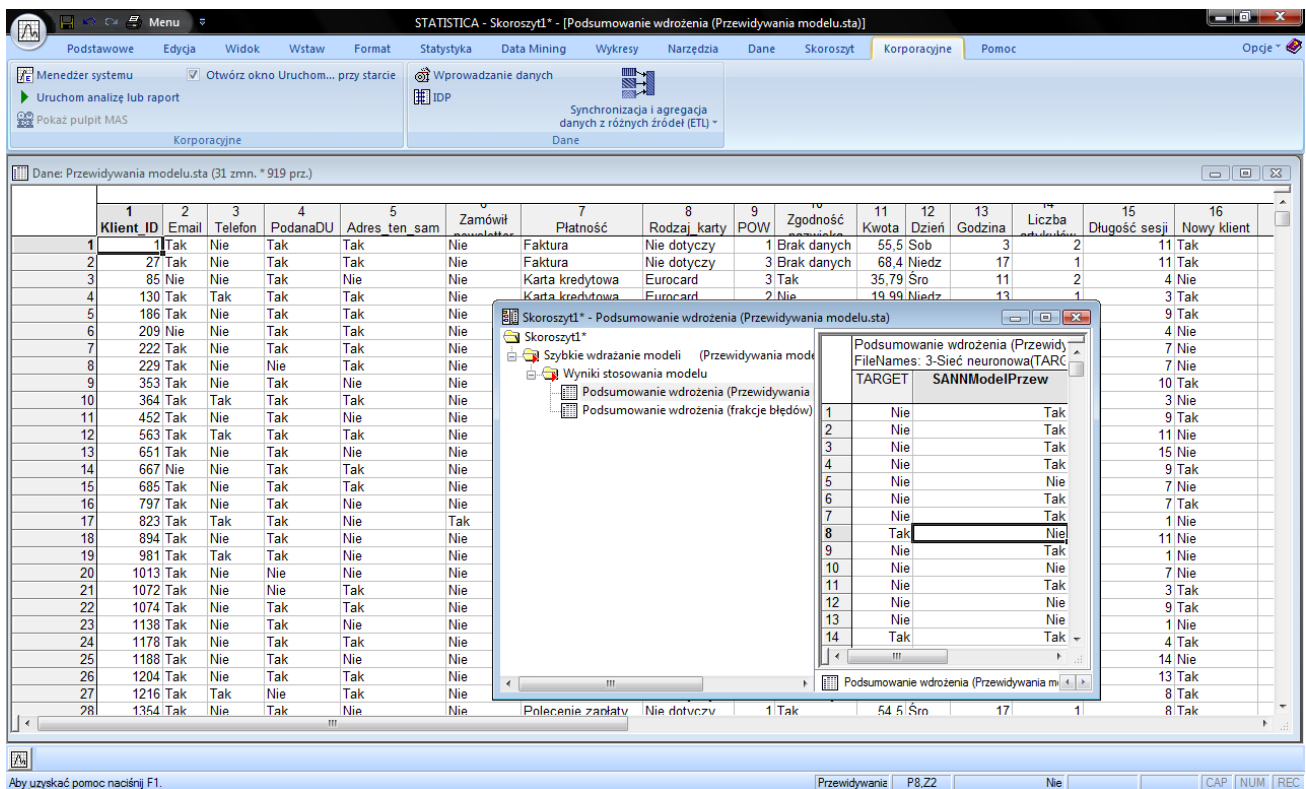


Dla przykładu zobaczmy, jak wygląda stosowanie modelu z punktu widzenia zwykłego użytkownika. Po uruchomieniu *STATISTICA* wybieramy polecenie *Korporacyjne – Uruchom analizę lub raport* i w oknie *Wykonaj analizę lub raport* wybieramy *Przewidywania modelu*.



Rys. 20. Uruchamianie wdrożenia.

Program automatycznie pobierze dane, wykona odpowiednie przekształcenia, zastosuje model, wyznaczy miary trafności i wyświetli wyniki w *STATISTICA*.



Rys. 21. Wdrożenie modelu w *STATISTICA*.



## Podsumowanie

W niniejszym przykładzie przedstawiliśmy, jak korzystając z *Przepisów Data Miner* można stworzyć model, ocenić jego jakość, a potem zastosować go w środowisku *STATISTICA Enterprise*. *Przepisy* prowadzą użytkownika przez całą drogę od danych do wiedzy, wspierając nie tylko właściwe modelowanie, ale również przygotowanie danych do analizy, ocenę modelu i jego stosowanie. Warto też zwrócić uwagę na ułatwienie pracy uzyskiwane dzięki *STATISTICA Enterprise*, zwłaszcza w sferze pobierania danych, przygotowania procedury stosowania modelu i uruchamiania go przez użytkowników.

## Literatura

1. Demski T, *STATISTICA Enterprise jako platforma analityczna dla całej organizacji*, artykuł dostępny na stronie: <http://www.statsoft.pl/czytelnia/jakosc/wprowadzenie.html>
2. Tadeusiewicz R., *Wprowadzenie do sieci neuronowych*, StatSoft Polska, 2001.
3. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, Wydawnictwo Naukowo-Techniczne, 2005.