



PRZYKŁAD ROZWIĄZANIA ZAGADNIENIA PREDYKCYJNEGO ZA POMOCĄ TECHNIK DATA MINING

dr Janusz Wątroba, StatSoft Polska Sp. z o.o.

W przykładzie prezentowane jest zastosowanie technik *data mining* do zagadnienia predykcijnego. Do wykonania analizy zostanie wykorzystany program *STATISTICA Data Miner*.

Wprowadzenie

Autorzy różnych publikacji poświęconych problematyce *data mining* podają różne definicje tego pojęcia. Wiele z nich jest podobnych do tej, którą podają Weiss i Indurkha [7]: **Data mining oznacza proces poszukiwania wartościowych informacji (wiedzy) w sytuacji, gdy mamy do czynienia z dużą ilością danych.** Większość autorów podkreśla również, że rozwój technik *data mining* wiąże się z zastosowaniem trzech dziedzin: baz danych, metod statystycznych oraz uczenia maszynowego (*machine learning*). Reprezentatywnym przykładem *data mining* może być analiza kredytów udzielanych przez bank małym firmom. Banki wykorzystują bazy danych, zawierające informacje o kredytach, przy czym dostępne są zarówno dane z bieżącego okresu, jak i z przeszłych okresów. Kryteria stosowane przy ocenie ryzyka związanego z kredytowaniem małych firm są opisywane w postaci pewnych cech. Obejmują one takie cechy jak: pomiar kondycji finansowej firm, ostatnio osiągnięte zyski lub zaległe długi lub cechy bardziej dotyczące ludzi, np. przebieg kredytowania „osobistego” właściciela firmy. Do dyspozycji jest szereg przypadków, czyli rekordów, dotyczących konkretnego długu danej firmy. Banki analizują te dane w poszukiwaniu poziomu ryzyka kredytowego, który gwarantuje w przyszłości spłacenie kredytu w nadziei zmniejszenia ryzyka niespłacenia kredytu. Do każdego przypadku jest przypisana etykieta, oznaczająca prawidłową odpowiedź, tzn. czy dany kredyt został spłacony czy też nie. Zamiast etykiety mogłaby też występować określona wartość (np. rzeczywiste zyski lub straty w przypadku każdej pożyczki), a celem byłaby minimalizacja strat. Tak więc technicznym celem *data mining* jest „nauczenie się” kryteriów decyzyjnych w celu przypisywania etykiet do nowych przypadków.

Wspomniani wcześniej autorzy, Weiss i Indurkha, dzielą różne typy problemów występujących w *data mining* na dwie ogólne kategorie: *data mining* dla zagadnień predykcyjnych (*predictive data mining*) oraz zagadnienia związane z odkrywaniem wiedzy (*knowledge*



discovery). Zagadnienia predykcyjne są opisywane za pomocą określonych celów, które są powiązane z przeszłymi przypadkami o znanych odpowiedziach. Są one wykorzystywane do konstruowania prognozy dotyczącej odpowiedzi dla nowych przypadków. Natomiast zagadnienia odkrywania wiedzy zwykle opisują stan przed prognozą, gdy informacja nie jest wystarczająca do skonstruowania prognozy. Zagadnienia odkrywania wiedzy są zwykle komplementarne w stosunku do predykcyjnego *data mining*, ale jednocześnie są bliższe raczej procesowi przygotowania decyzji niż jej podejmowaniu.

Dwa podstawowe problemy predykcyjne to klasyfikacja i regresja. W przypadku klasyfikacji odpowiedź (zmienna zależna) jest jakościowa (np. „spłacił kredyt” lub „nie spłacił kredytu”). W przypadku regresji odpowiedź (zmienna zależna) jest liczbą (np. wielkość zysku lub straty). Szczególny przypadek regresji stanowią szeregi czasowe, gdzie pomiary tej samej cechy są wykonywane w kolejnych momentach czasu.

Zagadnienia predykcyjne występują w bardzo wielu problemach poznawczych i decyzyjnych. Mają one szczególne znaczenie zwłaszcza w sytuacjach, gdy zachodzi potrzeba częstego podejmowania decyzji. Główna trudność, która wtedy się pojawia, polega na tym, że w momencie ustalania decyzji nie jest znany przyszły stan różnych zjawisk i procesów. Stąd też nie można jednoznacznie ocenić, jaką korzyść przyniesie nam dana decyzja. Prognozy dostarczają zazwyczaj dodatkowych informacji, które zmniejszają lukę informacyjną, i mogą przyczynić się do zmniejszenia ryzyka związanego z podejmowaniem decyzji.

Należy wyraźnie podkreślić, że przy podejmowaniu decyzji ważną okazuje się nie tylko świadomość korzyści płynących z prognozowania, lecz również znajomość jego ograniczeń. Nie można traktować prognozy jako sądu stanowczego. Co prawda niespełnienie się prognoz lub występowanie dużego błędu zawsze może się zdarzyć, ale częstość formułowania błędnych prognoz będzie maleć w miarę opanowywania metod przewidywania przyszłych zdarzeń lub stanów, zaś użytkownicy prognoz tym mniej się rozczarują, im lepiej potrafią formułować zadania prognostyczne i lepiej korzystać z informacji, jaką niesie ze sobą prognoza. Systematyczne podejście prognostyczne prowadzi zazwyczaj do uzyskiwania konkretnych korzyści, ale wiara w to, że można w pełni odkryć tajemnice przyszłości, stanowi tylko pobożne życzenie [1].

W przypadku predykcyjnego *data mining* wykorzystywane są m. in. takie techniki analityczne jak: modele regresji liniowej prostej i wielokrotnej, modele regresji nieliniowej, sieci neuronowe, drzewa regresyjne, GAM (Generalized Additive Models) czy też metoda MARS (Multivariate Adaptive Regression by Splines). Warto w tym miejscu podkreślić, że wszystkie te techniki są dostępne w programie *STATISTICA Data Miner*. Część z nich zostanie wykorzystana w trakcie budowy przykładowego projektu *data mining*.

Z zagadnieniami predykcyjnymi możemy się spotkać w wielu dziedzinach. Mogą to być zarówno problemy o charakterze poznawczym (kiedy chodzi nam o poznanie struktury czynników wpływających na określone zjawisko lub proces), jak i w zagadnieniach praktycznych (kiedy chodzi nam np. o wspomaganie procesów decyzyjnych). Przykładowe obszary, w których spotkać można zagadnienia predykcyjne to:



- ◆ analiza sprzedaży (np. ocena czynników wpływających na poziom sprzedaży lub budowa modelu służącego do przewidywania poziomu sprzedaży w przyszłości),
- ◆ analizy finansowe (poszukiwanie czynników determinujących poziom zysku lub poniesionej straty albo próba budowy modelu opisującego kształtowanie się kursów walut),
- ◆ analizy zagadnień ubezpieczeniowych (np. ocena czynników mających wpływ na wielkość szkody w ubezpieczeniach komunikacyjnych),
- ◆ zagadnienia ochrony zdrowia (np. próba określenia modelu opisującego koszty różnych zabiegów medycznych lub koszty pobytu pacjenta w szpitalu na określonym oddziale),
- ◆ zagadnienia medyczne (ocena czynników wpływających na stan kliniczny pacjenta po przebytych leczeniu lub analiza przeżycia po określonym zabiegu).

Opis problemu i przykładowych danych

Celem przeprowadzanych analiz jest budowa modeli wyjaśniających wpływ różnych predyktorów na ceny nieruchomości w pewnym rejonie Kalifornii. Ceny nieruchomości są wyrażone poprzez medianę wartości nieruchomości (**Mediana wartości domów**; pełni ona w analizie rolę zmiennej zależnej). Wśród dostępnych potencjalnych predyktorów występują następujące zmienne:

- ◆ Długość geograficzna i Szerokość geograficzna; określające położenie rejonu w którym znajdują się nieruchomości,
- ◆ Mediana wieku domu; „wiek” nieruchomości, wyrażony w latach,
- ◆ Pokoje ogółem, łączna liczba pokoi w branych pod uwagę nieruchomościach,
- ◆ Sypialnie ogółem, łączna liczba sypialni w branych pod uwagę nieruchomościach,
- ◆ Liczba osób, łączna liczba osób zamieszkujących nieruchomości,
- ◆ Liczba mieszkań, łączna liczba mieszkań w branych pod uwagę nieruchomościach,
- ◆ Mediana dochodu, wartość mediany dochodu osób zamieszkujących nieruchomości wyrażona w tys. dolarów.

Poniżej zamieszczono fragment arkusza danych programu *STATISTICA*, który zawiera wykorzystywane dane.

Oprócz budowy modeli drugim celem niniejszego przykładu jest praktyczna prezentacja sposobu przeprowadzania bardziej złożonych analiz z użyciem narzędzi analitycznych zawartych w programie *STATISTICA Data Miner* [5].



Data: Kalifornia.sta (9v by 18373c)

Dane dotyczą cen nieruchomości w pewnym rejonie Kalifornii oraz zmiennych, które stanowią potencjalne predyktory, umożliwiające przewidywanie cen nieruchomości.

1	2	3	4	5	6	7	8	9
Dł. geogr.	Szer. geogr.	Med. wieku domu	Pokoje og.	Sypialnie og.	Liczba osób	Liczba mieszkań	Med. dochodu	Med. wart.
1	-122,23	37,88	41	880	129	322	126	8,33
2	-122,22	37,86	21	7099	1106	2401	1138	8,30
3	-122,26	37,84	42	2555	665	1206	595	2,08
4	-122,27	37,85	40	751	184	409	166	1,36
5	-122,27	37,85	42	1639	367	929	366	1,71
6	-122,28	37,85	41	535	123	317	119	2,40
7	-122,28	37,85	49	1130	244	607	239	2,46
8	-122,28	37,84	49	1916	447	863	378	1,93
9	-122,27	37,84	48	1922	409	1026	335	1,80
10	-122,27	37,83	49	1655	366	754	329	1,38
11	-122,27	37,83	49	1215	282	570	264	1,49
12	-122,27	37,83	48	1798	432	987	374	1,10

Przygotowanie danych do analiz

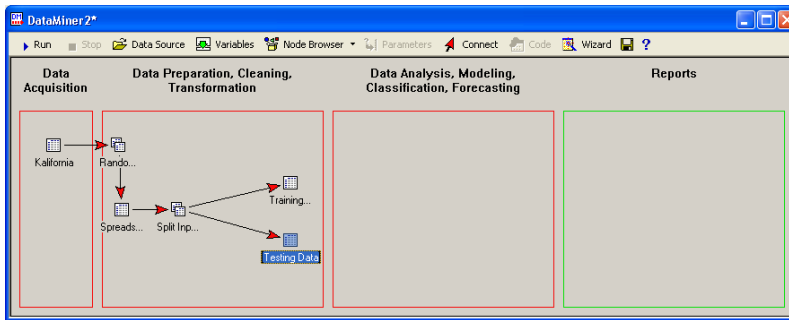
Dane do opisywanego przykładu są zapisane w pliku *Kalifornia.sta*. Rozpoczynając budowanie projektu *data mining*, w panelu *Źródło danych*, wskazujemy plik danych o nazwie *Kalifornia.sta*. W tym celu możemy skorzystać z przycisku *Data Source*. Następnie w oknie *Select dependent variables and predictors*, które pojawi się na ekranie, wskazujemy zmienną zależną i predyktory (zmiennne objaśniające). W naszym przykładzie w charakterze zmiennej zależnej ciąglej użyjemy zmiennej o nazwie **Mediana wartości domu**, natomiast jako predyktory ciągle wskazujemy zmienne: **Długość geograficzna**, **Szerokość geograficzna**, **Mediana wieku domu**, **Pokoje ogółem**, **Sypialnie ogółem**, **Liczba osób**, **Liczba mieszkań** oraz **Mediana dochodu**. Kliknięciem przycisku *OK* akceptujemy dokonane wybory, a następnie w oknie *Select dependent variables and predictors* jeszcze raz klikamy przycisk *OK*. Spowoduje to powrót do obszaru roboczego projektu.

Ze względu na czas obliczeń wyjściowy plik danych ograniczymy do 10% przypadków. W tym celu skorzystamy z przycisku *Node Browser* i w lewym panelu okna, które pojawi się na ekranie, rozwijamy katalog *Data Cleaning and Filtering*, a następnie w prawym panelu klikamy dwukrotnie lewym przyciskiem myszy węzeł analizy o nazwie *Random Sample Filtering*. Za pomocą tego węzła możemy wybrać losowy podzbiór przypadków o żądanej liczebności. Powracamy do obszaru projektu, klikamy prawym przyciskiem myszy ikonkę oznaczającą wstawiony węzeł oraz w podręcznym menu wybieramy opcję *Edit parameters*. W oknie, które pojawi się na ekranie, klikamy kartę *General* i w polu *Percent of Cases* wprowadzamy wartość *10*. Kliknięciem przycisku *OK* akceptujemy dokonany wybór i wracamy ponownie do obszaru roboczego projektu. Aby wykonać żądane zadanie, używamy przycisku paska narzędzi *Run*. Program wykonuje odpowiednie obliczenia i umieszcza wyniki w nowym arkuszu. Możemy je przejrzeć, klikając prawym przyciskiem ikonkę nowego arkusza oraz wybierając w podręcznym menu opcję *View Document*.



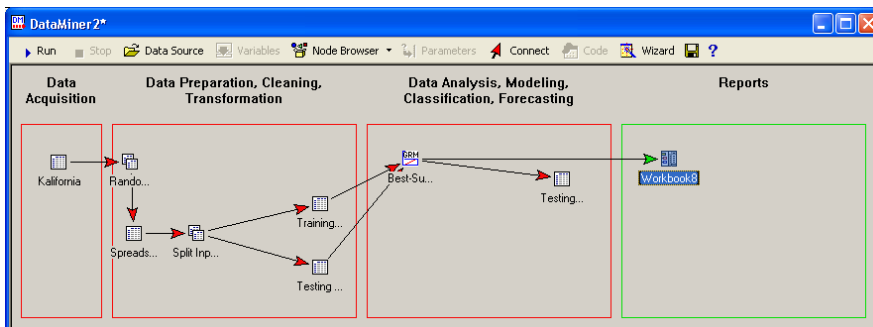
Kolejną czynnością, którą przeprowadzimy, będzie podział uzyskanego podzbioru danych na dwie części: próbę uczącą (training sample) i próbę testową (testing sample). Próba ucząca będzie wykorzystywana do szacowania parametrów tworzonych modeli, natomiast próba testowa będzie wykorzystywana do testowania uzyskanych rozwiązań. Do próby testowej weźmiemy 25% przypadków (spośród wybranych wcześniej 10%). Odpowiedni węzeł analizy: *Split data into Training and Testing Samples* przywołujemy z katalogu *Regression Modeling and Multivariate Exploration* w *Przeglądarce węzłów*. Po powrocie do obszaru roboczego projektu klikamy prawym przyciskiem myszy ikonkę oznaczającą wstawiony węzeł oraz w podręcznym menu wybieramy opcję *Edit parameters*. W oknie, które pojawi się na ekranie, klikamy kartę *General* i w polu *Approximate percent of cases for testing*: wprowadzamy wartość 25.

Poniżej przedstawiono wygląd obszaru roboczego projektu po wstawieniu odpowiednich węzłów analitycznych.



Budowa modeli predykcyjnych

Po etapie wstępnego przygotowania danych przystąpimy do zasadniczej części analizy. Przy budowie pierwszego modelu wykorzystamy regresję wieloraką. W tym celu do próby uczącej i testowej podpinamy węzeł analizy o nazwie *Best-Subset and Stepwise ANCOVA with Deployment*, który znajduje się w katalogu *Regression Modeling and Multivariate Exploration*. Klikając przycisk *Run*, umieszczony na pasku narzędzi, uruchamiamy odpowiednią analizę.



Aby obejrzeć otrzymane wyniki, klikamy dwukrotnie ikonkę symbolizującą skoroszyt z wynikami analizy. Na ekranie pojawi się okno skoroszytu, pokazane na poniższym rysunku. Zaznaczając odpowiednią pozycję w lewym panelu skoroszytu, możemy obejrzeć odpowiednie wyniki.

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	8,806606E+11	1	8,806606E+11	247,5791	0,000000
Dł. geogr.	9,807646E+11	1	9,807646E+11	275,7212	0,000000
Szer. geogr.	1,128623E+12	1	1,128623E+12	317,2885	0,000000
Med. wieku domu	7,240752E+10	1	7,240752E+10	20,3658	0,000007
Pokoje og.	2,249708E+10	1	2,249708E+10	6,3246	0,012019
Sypialnie og.	6,774596E+10	1	6,774596E+10	19,0453	0,000014
Liczba osób	3,043402E+11	1	3,043402E+11	85,5688	0,000000
Liczba mieszkań	1,295233E+10	1	1,295233E+10	3,6413	0,056569
Med. dochodu	2,644224E+12	1	2,644224E+12	743,3678	0,000000
Error	4,965694E+12	1396	3,557088E+09		

I tak wyniki zamieszczone w arkuszu *Univariate Tests of Significance* sugerują, że wszystkie dostępne w pliku danych predyktory powinny zostać uwzględnione w modelu. Poniżej przedstawiono okno z wybranymi miarami dobroci dopasowania modelu. Wynika z nich, że zbudowany model wyjaśnia nieco ponad 60% oryginalnej zmienności zmiennej zależnej.

Dependent Variable	Multiple R	Multiple R ²	Adjusted R ²	SS Model	df Model	MS Model	SS Residual	df Residual	MS Residual	F	p
Med. wart. domu	0,778334	0,605804	0,603545	7,631327E+12	8	9,539159E+11	4,965694E+12	1396	3,557088E+09	268,1733	0,00

Kolejne okno zawiera oceny parametrów strukturalnych modelu, które są wykorzystywane przy tworzeniu prognozy. Są w nim również podawane standaryzowane oceny współczynników regresji dla każdego z predyktorów. Z ich wartości wynika, że stosunkowo największy wpływ na wartości zmiennej zależnej wykazują zmiany długości i szerokości geograficznej rejonu, w którym były ulokowane nieruchomości.

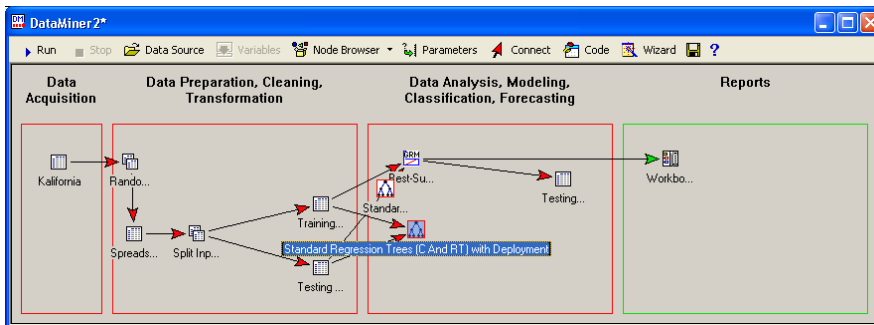
Effect	Med. wart. domu Param.	Med. wart. domu Std. Err	Med. wart. domu t	Med. wart. domu p	-95,00% Cnf Lmt	+95,00% Cnf Lmt	Med. wart. domu Beta (β)
Intercept	-3279795	208444,2	-15,7346	0,000000	-3688693	-2870898	
Dł. geogr.	-39375	2371,3	-16,6049	0,000000	-44026	-34723	-0,811032
Szer. geogr.	-39479	2216,4	-17,8126	0,000000	-43827	-35132	-0,880706
Med. wieku domu	758	167,9	4,5117	0,000007	428	1087	0,068448
Pokoje og.	-7	2,8	-2,5149	0,012019	-12	-2	-0,157283
Sypialnie og.	101	23,1	4,3641	0,000014	56	146	0,447987
Liczba osób	-37	4,0	-9,2498	0,000000	-45	-29	-0,430272
Liczba mieszkań	47	24,7	1,9062	0,056569	-1	96	0,188258
Med. dochodu	38556	1414,1	27,2648	0,000000	35782	41330	0,629790



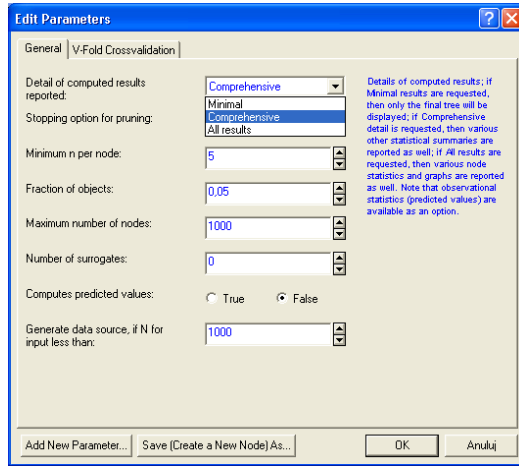
Wśród otrzymywanych wyników jest także arkusz zawierający wartości obserwowane, wartości przewidywane i reszty dla przypadków zaliczonych do próby przeznaczonej do testowania. Jego fragment przedstawia zamieszczony poniżej zrzut.

	1	2	3
	Predicted 1	Residuals 1	Med. wart. domu
1	396059	-37559	358500
8	208041	41959	250000
25	154008	-67508	86500
26	135970	-45470	90500
33	308989	35011	344000
36	136131	138869	275000
39	255637	10763	266400
43	197604	-52704	144900
46	224078	-32978	191100
47	167161	-12861	154300

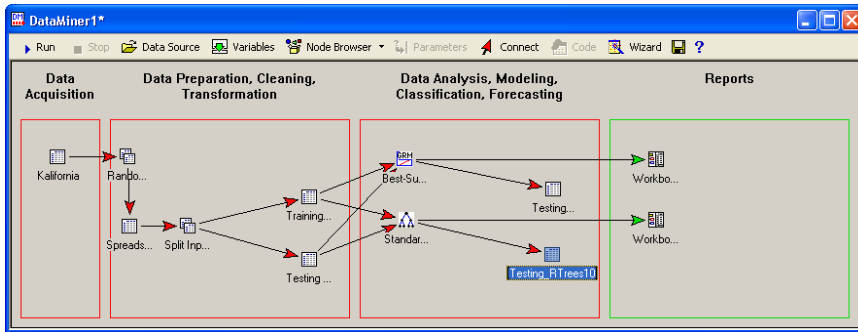
Kolejny model zostanie przygotowany za pomocą modułu *Standard Regression Trees (C and RT) with Deployment*. Zostanie wykorzystana procedura drzew regresyjnych. Jest ona również dostępna w katalogu *Regression Modeling and Multivariate Exploration*. Tak jak poprzednio, aby zastosować ten model do naszych danych, „podpinamy” w obszarze roboczym projektu *data mining* zbiór uczący i testowy do ikony oznaczającej żadaną analizę. Projekt analizy wygląda teraz tak, jak na zamieszczonym poniżej rysunku.



Przed wykonaniem odpowiednich obliczeń zmienimy jeszcze zakres uzyskiwanych wyników. W tym celu klikamy prawym przyciskiem myszy ikonę oznaczającą wybraną przez nas analizę i wybieramy opcję *Edit parameters*. Na karcie *General* tego okna w polu *Detail of computed results reported* wybieramy pozycję *Comprehensive*. Oprócz tego na karcie *V-Fold Crossvalidation* zaznaczamy opcję *v-krotnej oceny krzyżowej* (pozostawiając domyślne parametry). Tak jak poprzednio, aby wykonać odpowiednie obliczenia dotyczące tylko nowo wstawionego węzła, klikamy prawym przyciskiem myszy w obrębie obszaru roboczego projektu i w podręcznym menu wybieramy opcję *Run dirty nodes*.



Obszar roboczy tworzonego przez nas projektu *data mining* wygląda teraz tak, jak na poniższym rysunku.

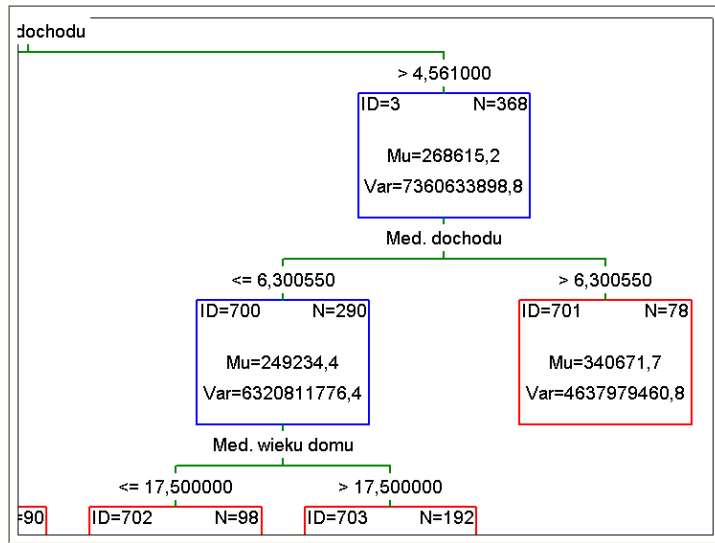


Podobnie jak przy poprzedniej analizie, aby obejrzeć otrzymane wyniki, klikamy dwukrotnie ikonkę symbolizującą skoroszyt z wynikami analizy. Na ekranie pojawi się okno skoroszytu, pokazane na poniższym zrzucie. Zaznaczając odpowiednią pozycję w lewym panelu skoroszytu możemy obejrzeć żądane wyniki.

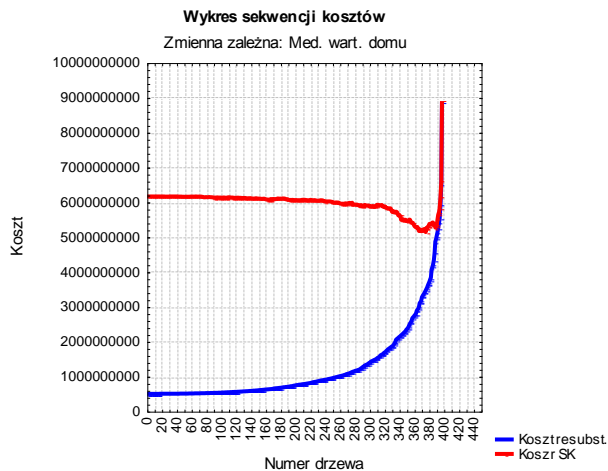
Node #	Left branch	Right branch	Size of node	Node mean	Node variance	Split variable	Split constant
1	2	3	1371	168336,5	8,666775E+09	Med. dochodu	4,5610
2		4	1003	158882,4	6,187264E+09	Med. dochodu	3,0248
4	6	7	530	131919,8	5,415101E+09	Liczba mieszkań	510,0000
6		8	341	117774,5	4,347315E+09	Med. dochodu	2,2275
8			168	94220,2	2,440345E+09		
9			173	140648,0	5,137206E+09		
7			189	157441,3	6,329282E+09		
5	384	385	473	189094,1	5,325187E+09	Med. wieku domu	37,5000
384			383	180652,0	4,243389E+09		
385			90	225020,0	8,334878E+09		
3	700	701	368	268615,2	7,360634E+09	Med. dochodu	6,3006
700	702	703	290	249234,5	6,320812E+09	Med. wieku domu	17,5000
702			98	215182,7	4,819018E+09		
703			192	266615,1	6,193424E+09		



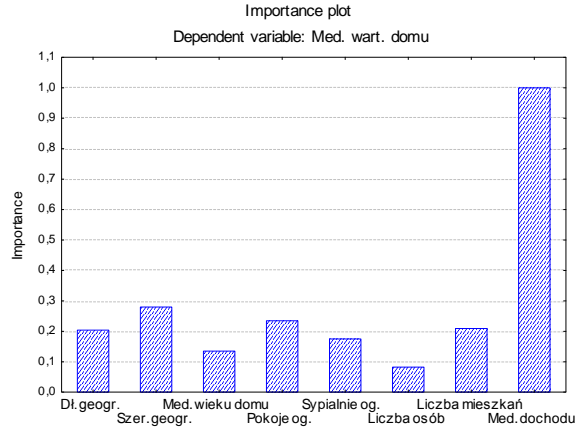
Przeglądanie wyników rozpoczniemy od drzewa regresyjnego. Przedstawia ono reguły decyzyjne występujące przy podziale przypadków. Fragment uzyskanego drzewa przedstawia poniższy zrzut.



Jako kolejny wynik analizy obejrzymy wykres liniowy sekwencji kosztów. Na wykresie tym znajduje się koszt sprawdzianu krzyżowego oraz koszt resubstytucji dla każdego przeciętego drzewa. Punkt przecięcia wykreślonych linii wskazuje optymalne drzewo. W naszym przykładzie jest to drzewo o numerze 408, które posiada 11 węzłów końcowych. Na zamieszczonym poniżej wykresie przedstawiono wykres sekwencji kosztów.



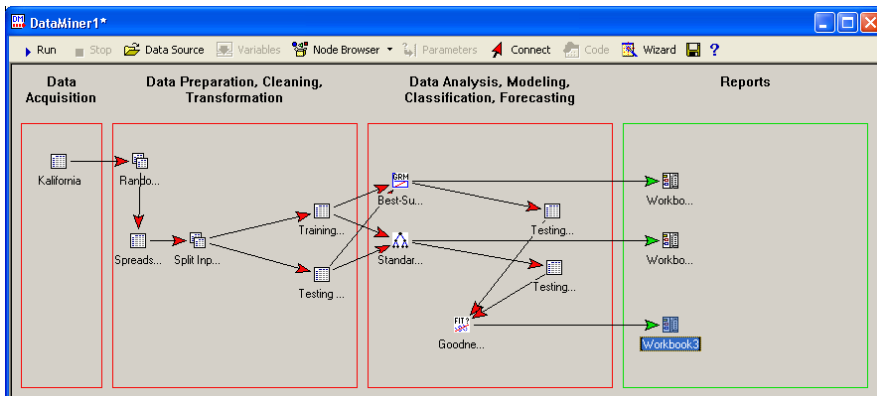
Możemy także obejrzeć wykres przedstawiający wielkość względnego wpływu poszczególnych zmiennych. Jak widać, z tego punktu widzenia najważniejszym predyktorem okazuje się zmienna **Mediana dochodu**. Wykres ten został pokazany poniżej.



W kolejnej części przeprowadzimy ocenę rozwiązań uzyskanych w wyniku zastosowania regresji wielorakiej oraz drzew regresyjnych.

Ocena uzyskanych rozwiązań

W kolejnej części analizy ocenimy dobroć dopasowania uzyskanych wcześniej modeli. W tym celu wykorzystamy węzeł o nazwie *Goodness of Fit for Multiple Inputs*, który znajduje się w katalogu *Statistics/Data Mining/Goodness of Fit*. Z węzłem tym połączymy kolejno arkusze, zawierające wartości obserwowane i prognozowane dla każdego ze stosowanych modeli. Przed połączeniem musimy wskazać dla każdego arkusza wybór zmiennych. Jako zmienną zależną wskazujemy wartość obserwowaną, a jako zmienną niezależną wartość prognozowaną. Możemy także określić zakres uzyskiwanych wyników. Możemy go ustalić po kliknięciu prawym przyciskiem myszy na ikonie symbolizującej odpowiedni węzeł i wybraniu opcji *Edit parameters* oraz karty *Continuous*. Na karcie tej zaznaczamy wszystkie dostępne przyciski opcji. Nasz projekt analizy wygląda teraz tak, jak na poniższym zrzucie.



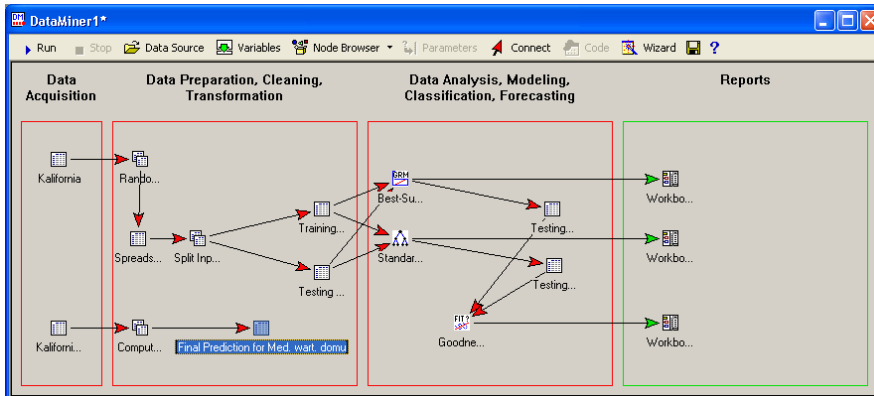


Teraz możemy już uruchomić odpowiednie obliczenia. Wszystkie wyniki obliczeń zawarte są w jednym arkuszu. Jego fragment przedstawiamy poniżej. Biorąc pod uwagę wartość średniego względnego błędu kwadratowego lub wartość średniego względnego błędu absolutnego, możemy stwierdzić, że stosunkowo najlepsze prognozy uzyskujemy w przypadku zastosowania drzew regresyjnych.

	3	4
	Mean relative squared error	Mean relative absolute error
Testing_GLM7(Predicted 1)	1,798607	0,394762
Testing_RTrees10(Predicted 1)	0,207468	0,290235

Zastosowanie uzyskanych modeli do prognozowania wartości nowych przypadków

Na koniec analizy zbudujemy prognozę, opierając się na wynikach wszystkich zastosowanych modeli. W tym celu wykorzystamy węzeł o nazwie *Compute Best Prediction from All Modules*, który znajduje się w katalogu *Regression Modeling and Multivariate Exploration*. Prognozy zbudujemy w oparciu o dane zawarte w pliku *Kalifornia_pred*. Plik ten zawiera tylko dane dla wszystkich występujących w naszej analizie predyktorów. Najpierw musimy wstawić ten plik do obszaru roboczego projektu. Następnie łączymy go strzałką z węzłem *Compute Best Prediction from All Modules*. Przed wykonaniem odpowiednich obliczeń klikamy ikonkę wstawionego pliku prawym przyciskiem myszy i wybieramy opcję *Variable selection*. Wybór zmiennych pozostaje taki sam jak poprzednio. Musimy tylko pamiętać, aby koniecznie zaznaczyć pole wyboru *Data for deployment project; do not re-estimate model* umieszczone w dolnej części okna. Analizę uruchamiamy za pomocą opcji *Run dirty nodes*. Ostateczny wygląd naszego projektu analizy przedstawia poniższy rysunek.





Klikając prawym przyciskiem ikonkę *Final Prediction for Med. wart. domu* i wybierając opcję *View document*, możemy obejrzeć średnią wartość prognozy wyliczoną w oparciu o wyniki uzyskane przez wszystkie trzy zastosowane modele. Wartości te dla wybranych pięciu przypadków przedstawia poniższy zrzut.

	14
	AveragePrediction for Med. wart. domu
1	131289,217
2	112417,287
3	244165,067
4	200540,136
5	125657,461

Literatura

1. Dittmann P., 2000, Metody prognozowania sprzedaży w przedsiębiorstwie, wyd. V, AE we Wrocławiu.
2. Gatnar E., 2001, Nieparametryczna metoda dyskryminacji i regresji, PWN Warszawa.
3. Hastie T. J., Tibshirani R. J., 1990, Generalized Additive Models, Chapman & Hall/CRC.
4. Hastie T. J., Tibshirani R. J., 2001, The Elements of Statistical Learning, Springer.
5. *STATISTICA Data Miner* Manual, StatSoft, Inc., 2002.
6. Weiss S.M., Indurkha N., 1998, Predictive data mining. A Practical Guide, Morgan Kaufmann Publishers, Inc.