



PRZYKŁAD PROGNOZOWANIA Z WYKORZYSTANIEM METOD DATA MINING

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

Wprowadzenie

Prognozowanie jest jednym z najczęściej występujących zadań analizy danych – ktoś nie chciałby wiedzieć, co stanie się w przyszłości, a zwłaszcza wykorzystać tę wiedzę przy podejmowaniu decyzji. Ze względu na popularność prognozowania oraz rozmaite dziedziny jego stosowania opracowano bardzo wiele sposobów budowy modeli prognostycznych (przegląd metod prognozowania znajduje się w podręczniku [1]). My zajmiemy się budową modelu z wykorzystaniem podejścia *data mining* i przestrzeni roboczej *STATISTICA Data Miner*.

Krótko o *data mining*

Można się spotkać z wieloma definicjami *data mining*; jedną z rozsądniejszych wydaje się zaproponowana w podręczniku [2]: „*Data mining* jest procesem badania i analizy dużych ilości danych metodami automatycznymi lub półautomatycznymi w celu odkrycia znaczących wzorców i reguł”.

W podejściu *data mining* kluczowe jest uzyskanie odpowiedzi na pytanie nurtujące badacza, rozwiązanie konkretnego problemu, przewidzenie wartości pewnej ważnej z praktycznego punktu widzenia wartości. Zazwyczaj mniej ważne jest sformułowanie ogólnego wniosku czy reguły.

W *data mining* model oceniamy na podstawie trafności jego przewidywań. Jednak podczas dopasowywania parametrów modelu (który to proces nazywamy uczeniem) może wystąpić ten sam błąd, który zdarza się popełnić ludziom, tzn. algorytm nauczy się na pamięć rozwiązywania zadań przedstawionych w czasie uczenia. Nauczony na pamięć model świetnie przewiduje wartości, które i tak znamy, ale całkowicie zawodzi dla nowych przypadków – czyli jest zupełnie bezużyteczny. Takie zjawisko nazywamy przeuczeniem i aby go uniknąć dzielimy dane na co najmniej dwie próby:

- ♦ **uczącą**: te dane pokazujemy algorytmowi na etapie tworzenia modelu,



- ♦ **testową**: dla tych danych wyłącznie stosujemy uzyskany model i oceniamy jego przydatność.

W *data mining* bardzo często stosujemy metody uczące się, dla których nie określamy z góry postaci zależności (np. że średnia wartość zapotrzebowania na energię elektryczną rośnie liniowo albo wykładniczo), oczekujemy raczej, że algorytm znajdzie odpowiednią postać zależności. Dostyc często uzyskane w ten sposób modele są na tyle skomplikowane, iż nie umożliwiają łatwej interpretacji przez człowieka, i stosujemy je jak czarną skrzynkę: tzn. nie wnikamy w to, jak budowana jest prognoza, byleby tylko była trafna. W *data mining* stosuje się również tradycyjną statystykę, jednak zazwyczaj mniej restrykcyjnie sprawdza się założenia dotyczące tych metod, a uzyskane wyniki oceniamy raczej poprzez uzyskaną trafność przewidywań, a nie wyniki testów statystycznych modelu.

Cel i zadania

Naszym celem jest zbudowanie modelu przewidującego godzinowe zapotrzebowanie na energię elektryczną. Będziemy przewidywać zapotrzebowanie w poszczególnych godzinach na następną dobę. Przede wszystkim chcemy uzyskać trafne przewidywania (z małym błędem w próbie testowej) – model nie musi być zrozumiały dla człowieka.

1 Nr	2 Miesiąc	3 DzMies	4 Godzina	5 DTS	6 Temperatura	7 Zachmurzenie	8 Z	9 Próba
1	1	Listopad	1	1	Niedz./Św.	5,0 B/D	305,058	u
2	2	Listopad	1	2	Niedz./Św.	5,0 B/D	285,228	u
3	3	Listopad	1	3	Niedz./Św.	5,0 B/D	275,947	u
4	4	Listopad	1	4	Niedz./Św.	5,0 B/D	277,628	u
5	5	Listopad	1	5	Niedz./Św.	5,0 B/D	275,699	u
6	6	Listopad	1	6	Niedz./Św.	5,0 B/D	264,828	u
7	7	Listopad	1	7	Niedz./Św.	5,0 B/D	251,278	u
8	8	Listopad	1	8	Niedz./Św.	5,0 B/D	251,137	u
9	9	Listopad	1	9	Niedz./Św.	5,0 B/D	256,598	u
10	10	Listopad	1	10	Niedz./Św.	5,0 B/D	261,671	u
11	11	Listopad	1	11	Niedz./Św.	5,0 B/D	255,764	u
12	12	Listopad	1	12	Niedz./Św.	5,0 B/D	253,401	u
13	13	Listopad	1	13	Niedz./Św.	5,0 B/D	254,167	u
14	14	Listopad	1	14	Niedz./Św.	5,0 B/D	247,350	u

Rys. 1. Fragment arkusza danych.

Dysponujemy danymi z trzech miesięcy: listopada, grudnia i stycznia; łącznie mamy 2208 obserwacji. W pliku danych znajdują się następujące wielkości:

- ♦ Numer kolejnej obserwacji – zmienna *Nr*,
- ♦ Czas i data obserwacji: zmienne *Miesiąc*, *DzMiesiąca* (numer kolejny dnia miesiąca) oraz *Godzina*,



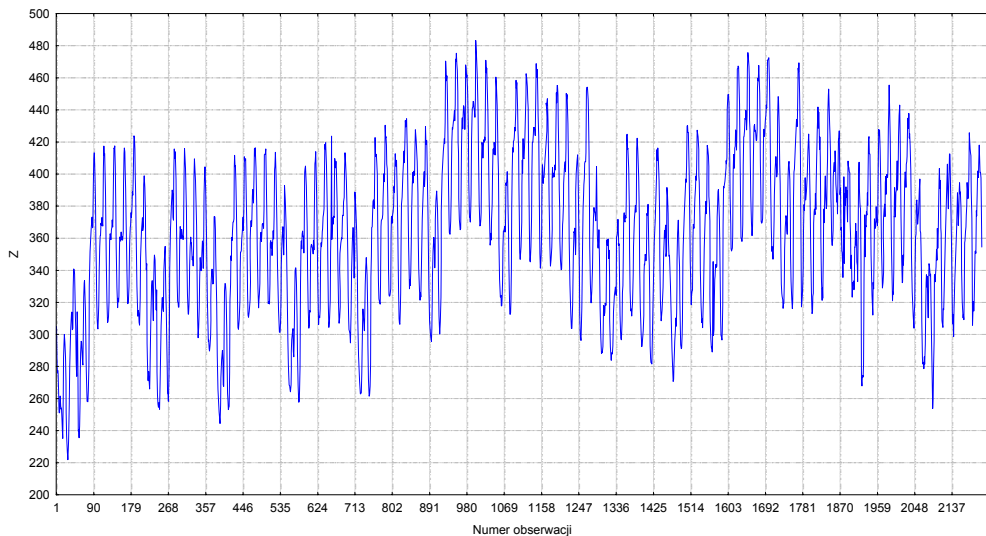
- ◆ Dzień tygodnia: zmienna *DTS* (jeśli dzień był świąteczny, to zmienna przyjmuje wartość *Niedz./Św.*),
- ◆ Informacja o pogodzie w ciągu dnia: zmienne *Zachmurzenie* i *Temperatura* (wartości zostały uśrednione po całej dobie),
- ◆ Zapotrzebowanie na energię elektryczną: zmienna *Z*.

Ponadto w arkuszu danych umieszczono kolumnę *Próba* z informacją, czy obserwacja ma trafić do próby uczącej czy testowej. Próbę testową tworzą dwa ostatnie tygodnie obserwowanego okresu. Na rys. 1 widzimy fragment wejściowego arkusza danych.

Możemy przyjąć, że dane są dobrej jakości, ponieważ większość z nich jest na bieżąco wykorzystywana w analizach i raportach, zbierana automatycznie (nie wpisywana z klawiatury) i są sprawdzane na etapie wpisywania do bazy danych. Niemniej jednak przeprowadzimy wstępną analizę danych: przede wszystkim w celu uzyskania ogólnego wglądu w zależności i reguły dotyczące zapotrzebowania na energię elektryczną, ale również aby wykryć ewentualne „dziwne” obserwacje.

Wstępna analiza i przygotowanie danych

Prognozowanie zazwyczaj zaczynamy od zobaczenia, jak zmienia się w czasie interesująca nas wielkość. Poniżej widzimy wykres liniowy przedstawiający zapotrzebowanie na energię elektryczną w kolejnych godzinach w badanym przez nas okresie.

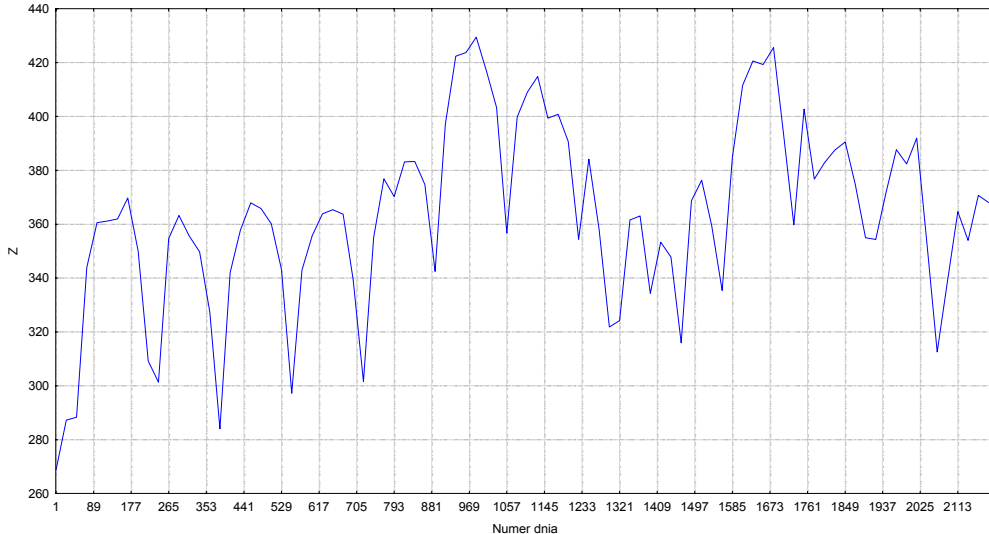


Rys. 2. Wykres przebiegu godzinowego zapotrzebowania na energię elektryczną.

Jak widać na wykresie (rys. 2), w przebiegu zapotrzebowania nie pojawiają się dziwne, gwałtowne skoki. Widzimy natomiast dosyć wyraźną okresowość (sezonowość). Nie ma

wyraźnego trendu, natomiast wydaje się, że mniej więcej od 1/3 obserwacji nastąpił wzrost średniego zapotrzebowania na energię.

Trend będziemy mogli łatwiej zauważyć, jeśli sporządzimy wykres przedstawiający dane zregulowane (opcja umożliwiająca tworzenie takich wykresów znajduje się w *STATISTICA* na karcie *Więcej* okna *Wykresy liniowe 2W*). Na rys. 3 poniżej widzimy, jak wygląda średnie zapotrzebowanie godzinowe w obrębie kolejnych dob.

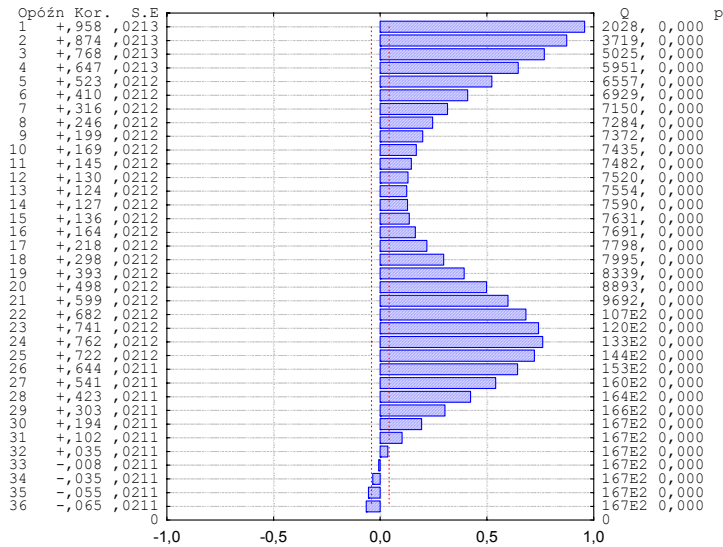


Rys. 3. Wykres przebiegu godzinowego zapotrzebowania uśrednionego po całej dobie.

Na wykresie wyraźniejsza jest okresowość, brak trendu i skok średniego zużycia, o których pisaliśmy powyżej.

W prognozowaniu często wykorzystuje się zmienne opóźnione, tzn. przewidujemy wartość zmiennej na podstawie jej wcześniejszych obserwacji. Związek między bieżącym zapotrzebowaniem a jego poprzednimi wartościami zbadamy za pomocą wykresu autokorelacji (rys. 4). W oczy rzuca się bardzo silny związek między bieżącym zapotrzebowaniem a jego wartością godzinę wcześniej: współczynnik korelacji tych wielkości wynosi 0,958. Tak duża wartość daje nadzieję, iż wykorzystanie w modelu obserwacji opóźnionej o 1 zaowocowałoby trafnymi przewidywaniami. Jednak nasze zadanie polega na przewidywaniu całej przyszłej doby i dlatego nie wykorzystamy zapotrzebowania z poprzedniej godziny w modelu.

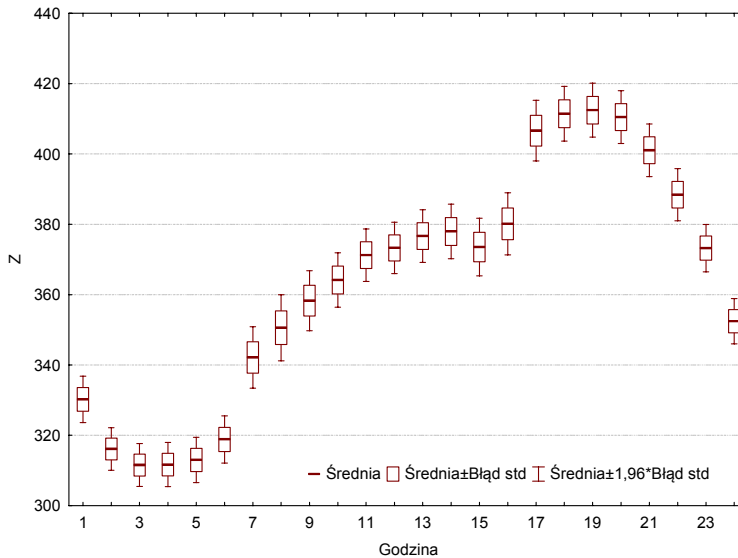
Zauważmy, że bardzo wyraźna jest również autokorelacja z zapotrzebowaniem z poprzedniej doby (obserwacja opóźniona o 24): współczynnik korelacji wynosi 0,762. Tę wielkość będziemy mogli wykorzystać w naszym modelu.



Rys. 4. Autokorelacja dla zapotrzebowania.

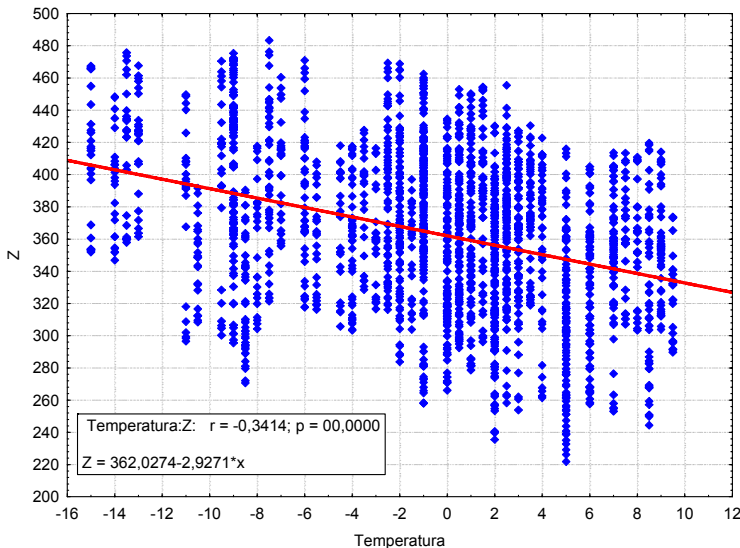
Teraz zajmiemy się wpływem na zapotrzebowanie na energię elektryczną zmiennych, które są zapisane w wejściowym pliku danych i możemy je wykorzystać do przewidywania.

Na wykresie ramka-wąsy przedstawionym na rys. 5 widzimy, jak kształtuje się zapotrzebowanie w poszczególnych godzinach. Zależność jest bardzo wyraźna i silna, a jej przebieg zgodny ze zdrowym rozsądkiem.



Rys. 5. Zapotrzebowanie w różnych godzinach.

W pliku danych zapisana jest średnia temperatura dobowa, która powinna również wpływać na zapotrzebowanie na energię elektryczną. Wpływ temperatury na zapotrzebowanie pokazuje diagram korelacyjny na rys. 6. Jak widać, im niższa temperatura, tym większe zapotrzebowanie na energię – w okresie zimowym jest to jak najbardziej rozsądny wynik. Na wykresie naniesione jest równanie regresji liniowej zapotrzebowania w funkcji temperatury: możemy je zinterpretować tak, że spadek temperatury o jeden stopień powoduje średnie zwiększenie zapotrzebowania o około 3 jednostki, jest to mniej więcej 1% średniej wartości zapotrzebowania na energię.

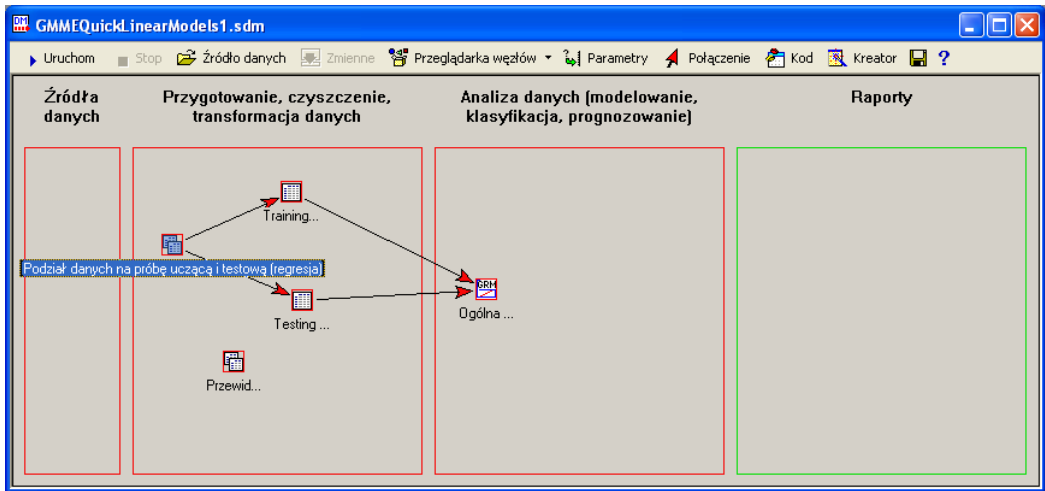


Rys. 6. Zapotrzebowanie na energię a temperatura.

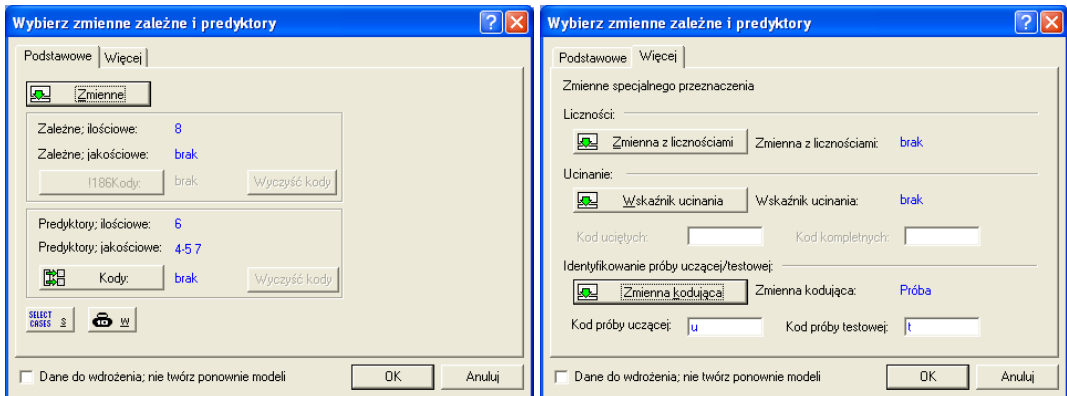
Dane zawierają zmienną *Zachmurzenie*, wydaje się, że ta wielkość powinna wpływać na zapotrzebowanie na energię elektryczną. Moglibyśmy przeanalizować ten wpływ podobnie jak wpływ temperatury. Przy tworzeniu wykresu spotka nas przykra niespodzianka: otóż *Zachmurzenie* przyjmuje wyłącznie jedną wartość: 'B/D' i na nic nam się nie przyda.

Budowa modelu w przestrzeni roboczej **STATISTICA Data Miner**

Po wstępnej analizie danych przechodzimy do budowy w przestrzeni roboczej *STATISTICA Data Miner* modelu przewidującego zapotrzebowanie na energię elektryczną. Z menu *Data Mining* wybieramy pozycję *Przestrzeń robocze – Modelowanie i eksploracja wielowymiarowa – Podręczny projekt dla modeli liniowych*. Na ekranie pojawi się przestrzeń robocza przedstawiona na rys. 7. W przestrzeni roboczej model i sposób jego stosowania dla nowych danych określamy jako schemat (graf) przepływu danych. Dzięki temu mamy przejrzystą prezentację operacji wykonywanych w celu uzyskania modelu, możemy łatwo aktualizować modele i wprowadzać do nich zmiany.

Rys. 7. Przestrzeń robocza *Podręczny projekt dla modeli liniowych*.

Pierwszym krokiem określania modelu jest wskazanie źródła danych: w tym celu klikamy przycisk *Źródło danych* na pasku narzędzi przestrzeni roboczej (zob. rys. 7) i wskazujemy plik z danymi o zapotrzebowaniu na energię (*Prognozowanie.sta*). Po wskazaniu źródła danych program wyświetla okno *Wybierz zmienne zależne i predyktory*, w którym określamy typy zmiennych wykorzystywanych w analizie. W naszym przypadku *Z* będzie zależną zmienną ilościową, *Temperatura* predyktorem ilościowym, a *Godzina*, *DTS* oraz *Zachmurzenie* predyktorami jakościowymi.

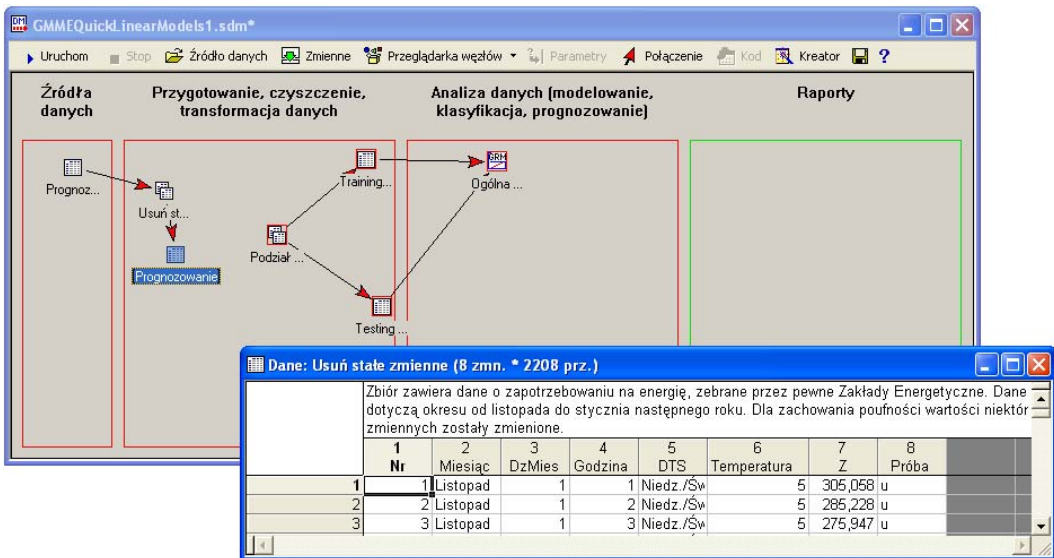


Rys. 8. Wybór zmiennych.

W zbiorze danych mamy zmienną *Próba*, rozróżniającą próby: uczącą i testową. Aby wykonać podział na próby według wartości tej zmiennej (zamiast domyślnego losowego podziału), przechodzimy na kartę *Więcej*, i w grupie *Identyfikacja próby uczącej/testowej* klikamy przycisk *Zmienna kodująca* i wybieramy zmienną *Próba* jako identyfikator próby, po czym jako kod danych uczących wskazujemy 'u', a testowych 't'. Ustawienia na obu kartach okna *Wybierz zmienne zależne i predyktory* widzimy na rys. 8.



Jak wspomnieliśmy wcześniej, zmienna *Zachmurzenie* jest stała i na nic nam się nie przyda. Moglibyśmy po prostu jej nie wybierać do analizy, ale postąpimy inaczej, tak aby zabezpieczyć się przed wystąpieniem takiego problemu dla innych zmiennych w przyszłości. Do eliminacji zmiennych, które przyjmują tylko jedną wartość, zastosujemy węzeł *Usuń stałe zmienne* z foldera *Czyszczenie danych* przeglądarki węzłów. Do przestrzeni roboczej węzeł *Usuń stałe zmienne* wstawiamy, klikając przycisk *Przeglądarka węzłów* na pasku narzędzi przestrzeni roboczej (zob. rys. 7), przechodząc do odpowiedniego foldera i dwukrotnie klikając węzeł – analogicznie jak w eksploratorze Windows. W celu wykonania operacji klikamy przycisk *Uruchom* lub naciskamy klawisz F5. Po zakończeniu przetwarzania w przestrzeni roboczej pojawi się nowe źródło danych, już bez stałej zmiennej (zawartość źródła danych możemy podejrzeć poleceniem *Pokaż dokument* z menu *Węzły*). Wynik wykonania projektu *data mining* pokazano na rys. 9.

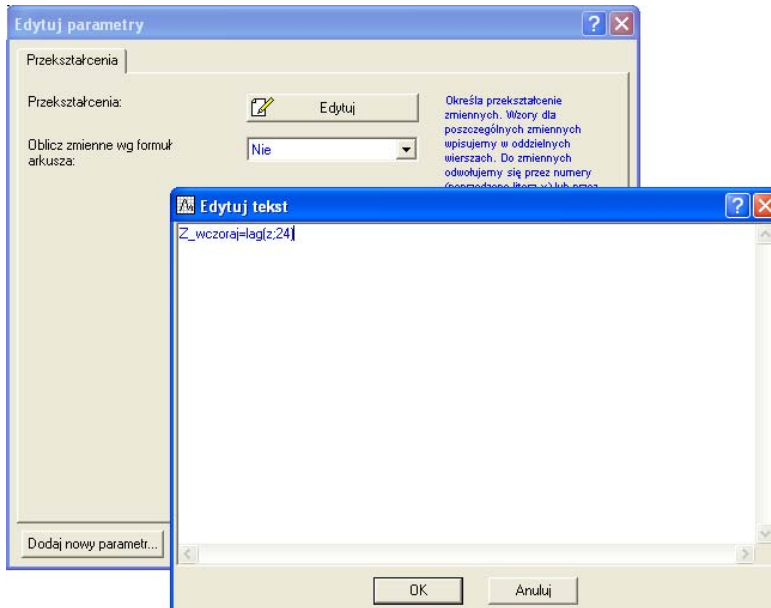


Rys. 9. Przestrzeń robocza po uruchomieniu węzła usuwającego stałe zmienne i wynikowy arkusz danych.

Do prognozowania chcemy wykorzystać zmienną, której nie ma w pliku wejściowym: wartość zapotrzebowania z poprzedniego dnia. Użyjemy do tego celu węzła *Przekształcenia zmiennych* (z foldera *Przekształcenia danych*), obliczającego nowe wartości zmiennej na podstawie wzoru podanego przez użytkownika. Po wstawieniu tego węzła do przestrzeni roboczej klikamy go dwukrotnie, po czym w oknie *Edytuj parametry* naciskamy przycisk *Edytuj* i podajemy formułę tak jak na rys. 10. W formule korzystamy ze standardowej funkcji arkusza *Lag(x,op)*, zwracającej wartość x sprzed op obserwacji. Po lewej stronie znaku równości wpisaliśmy zmienną *Z_wczoraj*, której do tej pory nie było w arkuszu – w takim przypadku program doda nową zmienną o tej nazwie do arkusza i obliczy jej wartości według podanego wzoru.

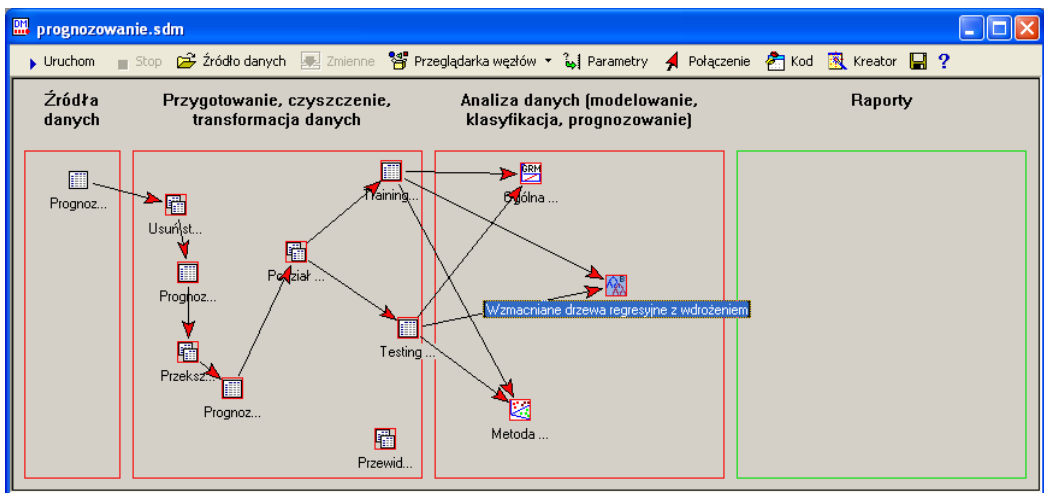


Po dodaniu nowej zmiennej do arkusza powinniśmy uwzględnić ją na liście predyktorów. W tym celu dwukrotnie klikamy nowe źródło danych, a potem w oknie *Wybierz zmienne zależne i predyktory* klikamy przycisk *Zmienne* i dołączamy zmienną *Z_wczoraj* do listy predyktorów ilościowych (np. klikając jej nazwę na tej liście przy wciśniętym klawiszu Ctrl).



Rys. 10. Określanie wzoru dla obliczenia zmiennej z wczorajszym zapotrzebowaniem na energię.

Po dokonaniu wyboru zmiennych łączymy strzałką wynikowe źródło danych z węzłem dzielącym dane na część uczącą i testową.



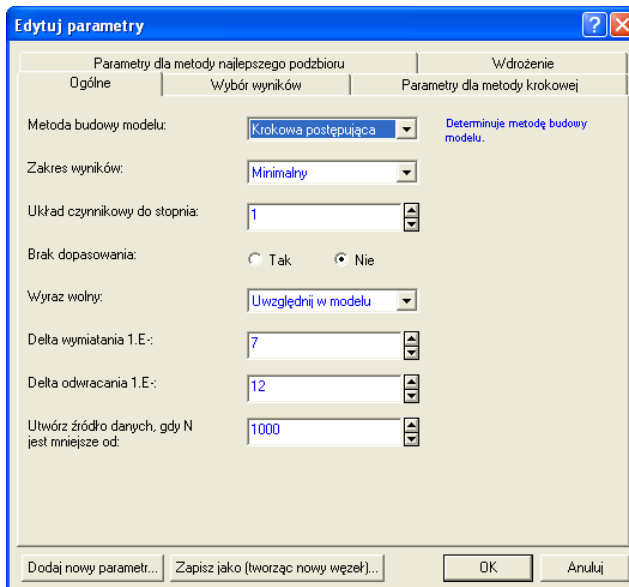
Rys. 11. Przestrzeń robocza po wstawieniu węzłów analitycznych.



W zastosowanym przez nas szablonie projektu znajduje się jeden węzeł budujący model. Węzeł ten wykorzystuje uogólnioną regresję liniową. Oprócz tego węzła do przestrzeni roboczej wstawimy węzły stosujące dwie techniki typowe dla *data mining*: wzmacniane drzewa regresyjne (ang. *boosted regression trees*) oraz metodę wektorów nośnych (wspierających, ang. *support vector machines*). Finalny projekt *data mining* przedstawiony jest na rys. 11. Obie te metody umożliwiają modelowanie bardzo złożonych, nieliniowych zależności i są oceniane jako jedne z najsilniejszych technik *data mining*. Opis wzmacnianych drzew regresyjnych i metody wektorów nośnych znajduje się w podręcznikach [3] i [4]. Aby użyć tych metod, do przestrzeni roboczej wstawiamy węzły *Wzmacniane drzewa regresyjne z wdrożeniem* i *Metoda wektorów wspierających z wdrożeniem (regresja)*.

Warto zauważyć jedną ważną zaletę tworzenia projektów w przestrzeni roboczej. Otóż jeśli pojawią się nowe dane (np. za kolejny miesiąc), to wystarczy je podpiąć do projektu i reszta wykona się automatycznie.

Przed uruchomieniem projektu zmienimy jeszcze ustawienia węzła regresji: otóż dla uzyskania dobrego modelu, bez zbędnych zmiennych, dobrze jest włączyć automatyczny dobór zmiennych. Zastosujemy metodę krokową postępującą: polega ona na tym, że zaczynamy od modelu zawierającego wyłącznie stałą (wyraz wolny), następnie wstawiamy do modelu najsilniejszy predyktor (tzn. taki, dla którego poziom p^1 jest najmniejszy) i tak dalej, aż do osiągnięcia takiej sytuacji, że dla każdego dostępnego predyktora poziom p jest większy od wartości progowej (zazwyczaj 0,05).

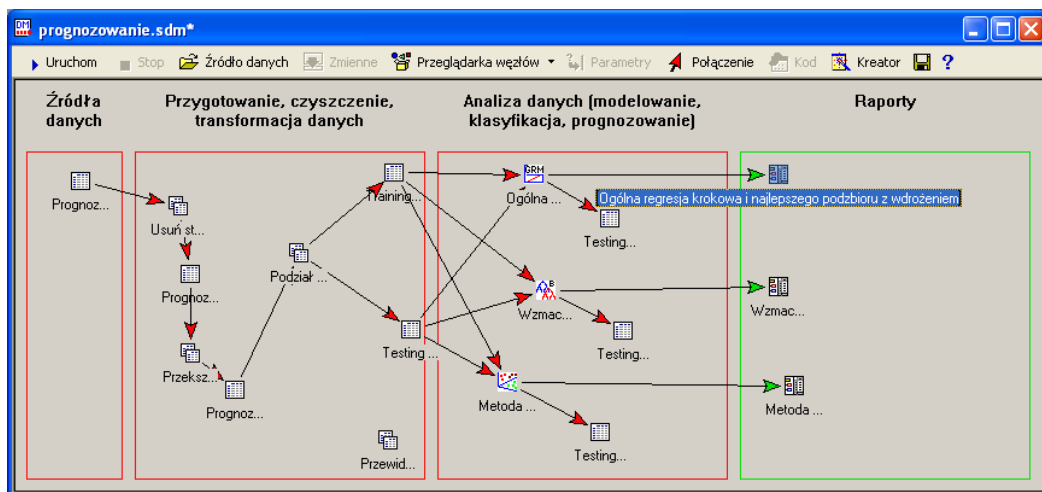


Rys. 12. Ustawienia dla węzła wykonującego regresję.

¹ W modelach regresyjnych poziom p dla zmiennej jest obliczamy jako prawdopodobieństwo tego, że współczynnik dla danej zmiennej w równaniu regresji przyjmie obserwowaną lub większą wartość, gdy w rzeczywistości (mówiąc bardziej ściśle w populacji generalnej) współczynnik ten jest równy zeru.

W celu włączenia automatycznego doboru zmiennych dwukrotnie klikamy węzeł *Ogólna regresja krokowa i najlepszego podzbioru z wdrożeniem* i w oknie jego parametrów na liście *Metoda budowy modelu* wybieramy *Krokowa postępująca*, tak jak na rys. 12.

Po uruchomieniu projektu program znajdzie model trzema wybranymi przez nas metodami. Uzyskane modele zostaną zastosowane dla danych z próby testowej, a w przestrzeni roboczej pojawią się dla każdego modelu: nowe źródło danych z wynikami stosowania modelu oraz węzeł ze skoroszytem podsumowującym proces budowy modelu (w części *Raporty*). Przestrzeń roboczą po wykonaniu modeli przedstawia rys. 13.



Rys. 13. Przestrzeń robocza po uruchomieniu węzłów analitycznych.

Po uzyskaniu modeli należy ocenić trafność ich przewidywań. Do tego celu użyjemy węzła *Dobroć dopasowania* (znajduje się on w przeglądarce węzłów w folderze *Data mining – Dobroć dopasowania*). Podłączamy ten węzeł do każdego ze źródeł danych powstałych w wyniku stosowania modeli dla danych testowych. Wcześniej wybieramy zmienne dla tych źródeł: jako zmienną zależną ilościową wskazujemy obserwowane wartości (*Z*), a jako predyktor wartości przewidywanej (zmienne te noszą nazwę tworzoną z identyfikatora metody i przyrostka *Przew*). W tabeli poniżej zestawiono miary jakości dopasowania dla stosowanych przez nas metod.

Miara \ Metoda	Regresja	Wzmacniane drzewa regresyjne	Wektory wspierające
Średnia kwadratów reszt	385	428	365
Średni błąd bezwzględny	14,9	15,9	14,8
Względne odchylenie przeciętne	0,0417	0,0445	0,0411



Najlepsze wskaźniki ma metoda wektorów wspierających, aczkolwiek zwykła regresja daje przewidywania porównywalnej jakości. Dla obu tych metod przeciętnie mylimy się o około 4%. Pomimo uzyskiwania lepszych wyników metodą wektorów wspierających, w praktyce być może lepiej byłoby zastosować regresję, ze względu na jej prostotę, łatwość wdrożenia i możliwości interpretacji wyników.

Możemy jeszcze spróbować poeksperymentować z parametrami analiz, aby uzyskać lepsze modele. W przypadku wektorów wspierających (nośnych) warto spróbować zmienić wykorzystywane jądro i parametry modelu *Pojemność*, *Epsilon* i *Ni* (na karcie SVM parametrów węzła). Pamiętajmy, że nie musimy dokładnie wiedzieć, co oznaczają te parametry: wystarczy metodą prób i błędów sprawdzić, czy ich zmiana poprawia uzyskiwane prognozy. W naszym przypadku pozytywny efekt daje zastosowanie jądra wielomianowego zamiast domyślnego RBF: uzyskujemy średnią kwadratów reszt 348 zamiast 365.

W przypadku wzmacnianych drzew regresyjnych zazwyczaj najlepsze wyniki dają ustawienia domyślne. Czasami jednak model polepsza się, gdy zwiększymy parametr *Maksymalna liczba węzłów* – tak jest właśnie w naszym przypadku. Domyślnie *Maksymalna liczba węzłów* wynosi 3, jeżeli zwiększymy ją do 7, to uzyskamy zauważalną poprawę trafności przewidywań, na tyle dużą, że wzmacniane drzewa wyprzedzą wektory nośne. W poniższej tabeli mamy zestawienie miar jakości przewidywań po wprowadzeniu zmian.

Miara \ Metoda	Regresja	Wzmacniane drzewa regresyjne	Wektory wspierające
Średnia kwadratów reszt	385	313	348
Średni błąd bezwzględny	14,9	13,7	14,6
Względne odchylenie przeciętne	0,0417	0,0382	0,04058

Po statystycznej ocenie modeli, należy ocenić je pod kątem ich planowanego zastosowania. W zależności do tego, jaki jest nasz cel, może się okazać, że względne odchylenie przeciętne na poziomie 3,8% jest świetnym wynikiem, albo nie do przyjęcia – to już zależy od praktycznych uwarunkowań.

Jeśli uznamy, że model jest odpowiedni, to zazwyczaj będziemy stosować go dla nowych danych. Możemy wyznaczać przewidywane wartości w przestrzeni roboczej *STATISTICA Data Miner*: w tym celu wystarczy podpiąć nowe dane do węzła *Przewidywania wszystkich modeli (regresja)* i uruchomić projekt. Inny sposób to zapisanie formuły modelu w postaci kodu C lub XML (mówiąc dokładniej specjalnego dialektu XML o nazwie PMML przeznaczonych do stosowania modeli *data mining*) i stosowanie go we własnych programach. Ponadto to w skład systemu *STATISTICA Data Miner* wchodzi specjalne narzędzie do zapisywania przewidywanych wartości w bazach danych.



Literatura

1. Dittman P., *Prognozowanie w przedsiębiorstwie*. 2004, Kraków, Oficyna Ekonomiczna.
2. Berry M. J. A., Linoff G., *Data mining techniques: for marketing, sales, and customer support* 1997, John Willey & Sons.
3. Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, 2005, Wydawnictwo Naukowo-Techniczne.
4. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning*, 2002, Springer-Verlag.