Zadanie: chcemy przewidzieć jakie będzie zapotrzebowanie na energię elektryczną na danym obszarze.

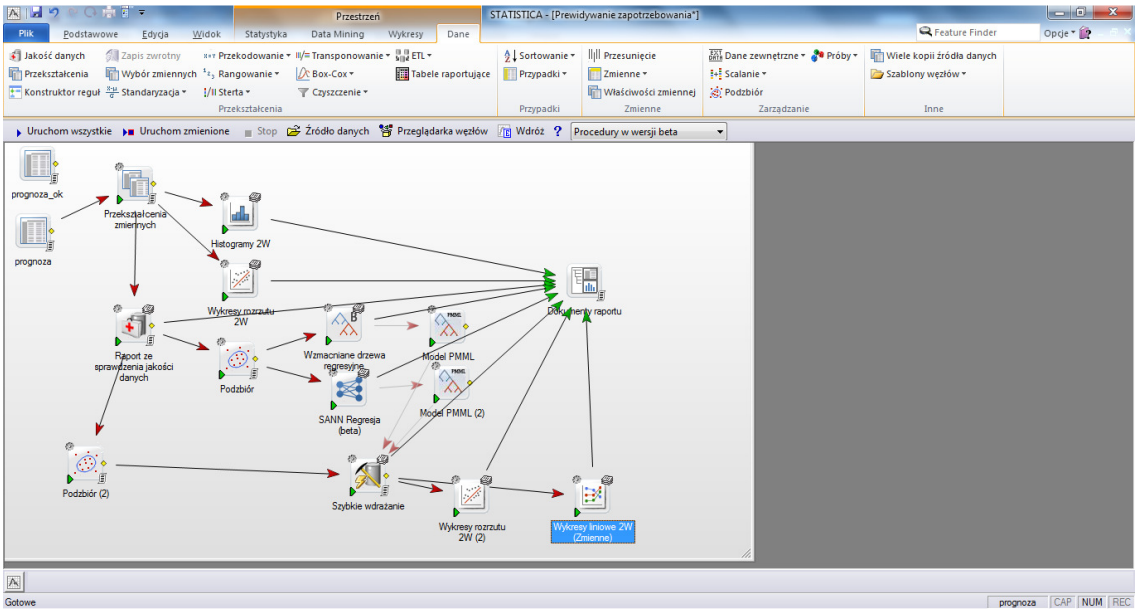
Dane:

- dane z 3 miesięcy począwszy od listopada
- Zmienna „Z" – informuje jakie było zapotrzebowanie (zmienna celu, którą będziemy przewidywać
- Dwie zmienne dotyczące pogody: „Temperatura" (średnia temperatura dobowa – a mamy zapotrzebowanie godzinowe - problem) i „Zachmurzenie"
- Zmienna „Dzień tygodnia" (Niedziele i święta oraz pozostałe)

1. Otwórz plik „prognoza.sta"
2. Otwórz przestrzeń roboczą
3. Wstaw źródło danych
4. Zapotrzebowanie na energię dzisiaj o 12 będzie podobne do tego z wczoraj o tej porze.
5. Tworzenie zmiennej opóźnionej:
    a. Dodaj węzeł Dane -> Przekształcenia -> Przekształcenia zmiennych
    b. Otwórz parametry i wpisz: Z_wczoraj=lag(z;24)
    c. Uruchom węzeł i oglądnij dokument
        i. na początku 24 puste przypadki
6. Sprawdzanie ogólnej charakterystyki danych
    a. Wstaw węzeł Wykresy -> Histogramy
    b. Wybierz zmienne 4-8 i wyłącz dopasowanie rozkładu normalnego – nie będzie potrzebne
    c. Uruchom i oglądnij raport
        i. Wykres „Godziny" – takie samo – ok.
        ii. Wykres „Dzień tygodnia" – Niedziele/Święta jest częstsza (bo Boże Nar, Nowy Rok, itd. zostało przerzucone) – ok.
        iii. Wykres „Temperatura" – dosyć mocno się zmieniała
        iv. Wykres „Zachmurzenie" – problem bo mamy B/D – jest stała – trzeba będzie ją usunąć
        v. Wykres „Z" – na pierwszy rzut oka OK., ale pierwszy słupek wyższy niż następny – podejrzane – trzeba się przyjrzeć
7. Sprawdzanie zależności między zmienną opóźnioną a bieżącą wartością (jak ma się zapotrzebowanie z wczoraj do dzisiaj)
    a. Wstaw węzeł Wykresy -> Rozrzutu
    b. Wybierz zmienne X: 9, Y:8
    c. zakładka „Więcej": „Statystyki" wybrać „R kwadrat" dla dopasowania liniowego
    d. Uruchomić i oglądnąć wykres – coś nie tak
        i. Sprawdzić obszar na początku wykresu
            - zaznaczyć na wykresie Edycja-> Wyróżnianie
            - zaznaczyć obszar – dziwny przebieg
            - zobaczmy, które to przypadki: Etykietuj->Zastosuj
            - w których wierszach arkusza występuje problem: prawy klawisz myszy -> Podzbiór
            - widać, że z jednego dnia – wartości były źle zakodowane – zły separator dziesiętny – błąd przy wprowadzaniu albo błąd przy eksport/import
            - Trzeba dane poprawić i analizy powtórzyć
8. Otwórz poprawiony arkusz „prognoza_ok.sta"
9. Podepnij nowe źródło danych
10. Uruchom wszystkie węzły
11. Sprawdź jak teraz wyglądają histogramy i wykresy
12. Ocena jakości danych
    a. Wstaw węzeł Dane -> Jakość danych (górne menu)

b. Wybierz zmienne 4-8
c. W zakładce „Wynik” wybierz „Utwórz raport diagnostyczny i oczyść dane”
d. Uruchom – sprawdza braki danych, wartości odstające, nadmiarowość i generuje raport)
e. Otwórz „Raport z badania jakości danych” – np. zmienna Zachmurzenie została usunięta

13. Podział danych na **zbiór uczący i testowy**
    a. Do węzła raportu jakości danych podłącz węzeł Dane -> Podzbiór – w prognozowaniu nie losujemy, ale wybieramy ostatnie operacje jako testowe – bo wraz z upływem czasu wzorce mogą się zmieniać
        i. jest to zbiór uczący, w którym ustaw zmienne 1-8
        ii. w „określone przez” wpisz: v0 <= (2208-2*7*24)
        iii. skopiuj ten węzeł (będzie to próba testowa)
        iv. w „określone przez” wpisz: v0 > (2208-2*7*24)
    b. Uruchom węzeł uczący
14. **Modelowanie**
    a. Wybierz węzeł „Podzbiór”
    b. Wstaw węzeł Data Mining -> Wzmacniane drzewa -> Wzmacniane drzewa regresyjne
        i. ustaw zmienne: 7 4-5 6 8 ___ generacja kodu tylko w PMML
    c. Wstaw węzeł Data Mining -> Sieci neuronowe -> Sieci neuronowe regresja
        i. ustaw zmienne: 7 6 8 4-5 generacja kodu tylko w PMML
    d. Sprawdzanie dokładności
        i. do węzła „Szybkie wdrażanie” podepnij na wejście zbiór testowy („Podzbiór(2)”)
        ii. w węźle „Szybkie wdrażanie” wybierz „Zapisz wartości lub klasy przewidywane i reszty” i zmienną 7 (wygenerujemy arkusz, w którym oprócz zmiennej przewidywanej będzie także wartość obserwowana), w zakładce „Dane wyjściu” wybierz „Dane wejściowe, predykcje i reszty”
        iii. Uruchom węzeł, oglądnij wyniki („Szybkie wdrożenie” -> Podsumowanie wdrożenia (frakcja błędów))
            - widać duży błąd dla drzew
    e. Poprawiamy drzewa – ważne z ilu elementów budowany jest finalny model, złożoność składowego drzewka – domyślnie z 3 węzłów, czyli 1 podział – warto zwiększyć
    f. Zmiana parametrów w węźle „Wzmacniane drzewa regresyjne”
        i. otworzyć "Parametry" i w zakładce "Generacja kodu" zaznaczyć tylko PMML
        ii. zakładka Specyfikacja -> Więcej
            - liczba drzew=900
            - maksymalna liczba węzłów=7
            - Uruchom zmienione
            - Sprawdź poprawę błędu
    g. Sprawdzenie zależności między wartościami obserwowanymi a przewidywanymi
        i. Wstaw węzeł Wykresy -> Rozrzutu (po węźle „Szybkie wdrażanie”)
        ii. Wybierz zmienne 6 1 (wartość przewidywana)
        iii. W „Więcej” zaznacz Statystyki -> R kwadrat
        iv. Uruchom
        v. Wydaje się, że jest ok – widać rozrzut ale nie widać braku symetrii
    h. Sprawdzenie czy model oddał wzorce przebiegu szeregu
        i. do węzła „Szybkie wdrożenie” podpiąć Wykresy -> Liniowy
            - ustaw zmienne 1 (przewidywana) 6 (rzeczywista)
            - zaznacz Rodzaj wykresu=wielokrotny
            - Uruchomić
                a. widać, że wykres jest ok., problem z „pikami”


STATISTICA - [Prewidywanie zapotrzebowania*]
Przestrzeń
Plik
Podstawowe
Edycja
Widok
Statystyka
Data Mining
Wykresy
Dane
Feature Finder
Opcje
Jakość danych
Przekształcenia
Konstruktor reguł
Zapis zwrotny
Wybór zmiennych
Standaryzacja
Przekodowanie
Rangowanie
Sterta
Transponowanie
Box-Cox
Czyszczenie
ETL
Tabele raportujące
Przekształcenia
Sortowanie
Przypadki
Przypadki
Przesunięcie
Zmienne
Właściwości zmiennej
Zmienne
Dane zewnętrzne
Próby
Scalanie
Podzbiór
Zarządzanie
Wiele kopii źródła danych
Szablony węzłów
Inne
Uruchom wszystkie
Uruchom zmienione
Stop
Źródło danych
Przeglądarka węzłów
Wdróż
Procedury w wersji beta
prognoza_ok
prognoza
Przekształcenia zmiennych
Histogramy 2W
Wykresy rozrzutu 2W
Raport ze sprawdzenia jakości danych
Podzbiór
Wzmacniane drzewa regresyjne
SANN Regresja (beta)
Model PMML
Model PMML (2)
Dokumenty raportu
Podzbiór (2)
Szybkie wdrażanie
Wykresy rozrzutu 2W (2)
Wykresy liniowe 2W (Zmienne)
Gotowe
prognoza
CAP
NUM
REC