Zaawansowana analityka z SAS Enterprise Miner

Edycja 6 - 2019/2020

Laboratorium

OPEN SOURCE INTEGRATION

- Ładowanie pakietów R
  - Otwarcie oprogramowania R -> 
  - W konsoli R wprowadzić polecenie

```
install.packages("package_name")
```

  - Wybrać lokalizację repozytorium, w którym pakiet ma zostać zainstalowany
  - Wykorzystanie pakietu R możliwe jest po rozpakowaniu zawartych w nim funkcji

```
library(package_name)
```

- Wykorzystanie R w SAS Enterprise Miner
  - SAS nie dostarcza oprogramowania R
    - Oprogramowanie pobiera się ze strony http://cran.r-project.org.
  - Oprogramowanie R musi być zainstalowane na tym samym serwerze /komputerze, co oprogramowanie SAS
  - Kompatybilność wersji SAS, R, PMML

Enterprise Miner Version	R Version	PMML Version
13.1	2.13.0 – 3.0.2	pmml_1.4.1
13.2	2.15.3 – 3.0.3	pmml_1.4.1
14.1	3.0.1 – 3.1.2	pmml_1.4.2

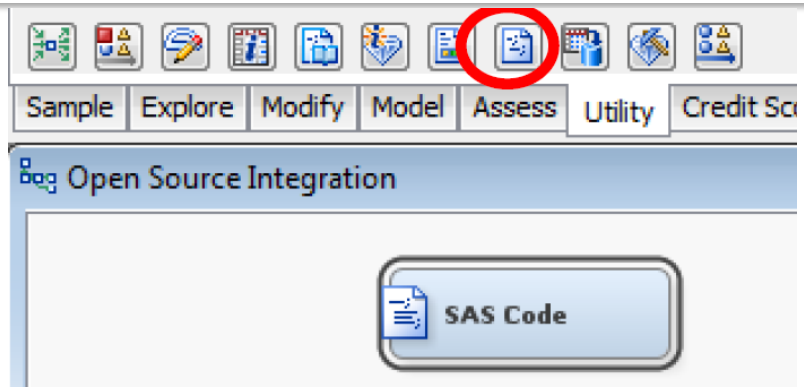
  - Weryfikacja komunikacji SAS i R z poziomu SAS Enterprise Miner
    - W węźle SAS Code użyć polecenia

```
proc options option=rlang;
run;
```

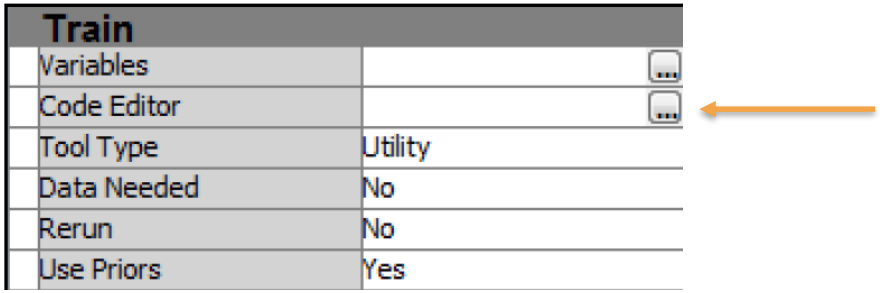
    - Rezultaty: RLANG – komunikacja SAS-R, NORLANG – brak komunikacji

Weryfikacja współpracy SAS Enterprise Miner i R

Użycie węzła SAS Code z zakładki Utility



Wybranie z opcji węzła edytora kodu



Wpisanie polecenia weryfikującego komunikację z SAS i R

```
proc options option=rlang;
run;
```

Weryfikacja wyniku w Rezultatach węzła

```
      SAS (r) Proprietary Software Release 9.4  TS1M3

 RLANG            Enables SAS to execute R language statements.
NOTE: PROCEDURE OPTIONS used (Total process time):
      real time               0.00 seconds
      cpu time                0.00 seconds
```

- Tryb działania węzła
  - Supervised
  - Unsupervised
- Tryb wyniku
  - PMML
  - MERGE
  - NONE

Train	
Variables	...
Code Editor	...
Language	R
Training Mode	Supervised
Output Mode	PMML

	Tryb PMML	Tryb Merge
Kod scoringowy	Kod scoringowy typu Data Step jest generowany jedynie dla trybu wyniku PMML	Modele nie zawierają Kodu scoringowego SAS -> należy użyć funkcjonalności scoringu z R • Wykorzystanie funkcji *predict()*
Kompatybilność z pakietami R	Tryb PMML może zostać użyty jedynie dla poniższych pakietów R: • linear models (lm) {base} • generalized linear models (glm) {base} • multinomial log-linear models (multinom) {nnet} • decision trees (rpart) {rpart} • neural networks (nnet) {nnet} • K-means clustering (kmeans) {base}	Możliwość integracji z pakietami R, dla których PMML nie jest wspierany
Ocena modelu	Ocena modelu jest wykonywana automatycznie dla jednocześnie ustawionych opcji Supervised, PMML	Ocena modelu odbywa się po zastosowaniu po węźle Open Source integration węzła Model Import

- Stosowanie odwołań w SAS Enterprise Miner
  - Cel: wydajne odwołania do zmiennych w zbiorze danych
  - Referencje automatycznie tworzone w SAS EM
    - Makrozmienne przechowujące teksty, np. listę zmiennych typu nominalnego ze zbioru

    - Przydatne zmienne
      - &EMR_MODEL – model R
      - &EMR_NUM_TARGET, &EMR_CLASS_TARGET – zmienna objaśniana/zmienna celu
      - &EMR_NUM_INPUT, &EMR_CLASS_INPUT – zmienne objaśniające
      - &EMR_IMPORT_DATA – zbiór danych użyty w ścieżce modelowania
    - Zmienne wymagane w trybie Merge
      - &EMR_EXPORT_TRAIN – eksportuje wynik scoringu dla danych treningowych
      - &EMR_EXPORT_VALIDATE – eksportuje wynik scoringu dla danych walidacyjnych
      - &EMR_IMPORT_VALIDATE – import danych walidacyjnych do R

- Dostosowanie kodu R do wykorzystania w SAS Enterprise Miner

```
library(rpart)
mytree <- rpart(Y ~ X1 + X2 + X3 +
                C1 + C2 + C3, data=mydata)
```

Tryb PMML

Tryb Merge

```
library(rpart)
&EMR_MODEL <- rpart(&EMR_NUM_TARGET ~
     &EMR_NUM_INPUT + &EMR_CLASS_INPUT,
          data=&EMR_IMPORT_DATA)
```

```
library(rpart)
&EMR_MODEL <- rpart(&EMR_NUM_TARGET ~
&EMR_NUM_INPUT + &EMR_CLASS_INPUT,
data=&EMR_IMPORT_DATA)
&EMR_EXPORT_TRAIN <-
   predict(&EMR_MODEL, &EMR_IMPORT_DATA)
```

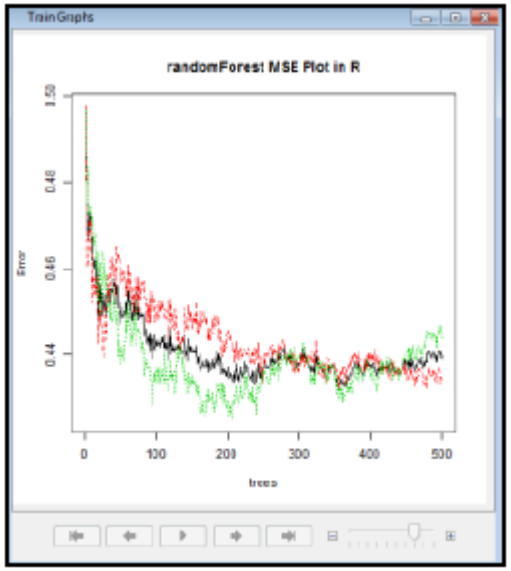
- Czy jest konieczne dostosowywanie kodu R do wykorzystania w SAS Enterprise Miner?
  - Wymagane jest użycie
    - &EMR_MODEL
    - &EMR_IMPORT_DATA
  - Do zmiennych objaśniających można odwoływać się po nazwach (uwaga: R – case sensitive)
  - Zmienna objaśniana, jeżeli jest wymieniana z nazwy, musi być poprzedzona literą ***r***

```
library(rpart)
&EMR_MODEL <- rpart(rY ~ X1 + X2 + X3 +
C1 + C2 + C3, data=&EMR_IMPORT_DATA)
```

- Ilustracja modelu w R

```
png("Rplot.png")
plot(&EMR_MODEL)
```

- Wynik dostępny w Rezultatach węzła Open Source Integration Node



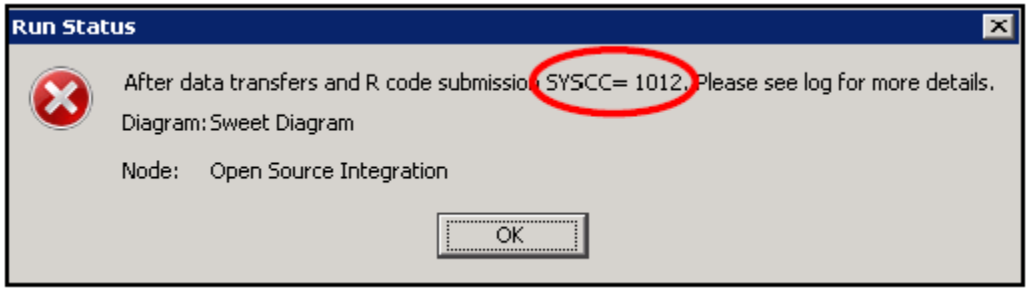
- Przydatne informacje o R
  - Case Sensitive

```
my_lm <- lm(Height ~ Weight, data=mydata)
```

≠

```
my_lm <- LM(height ~ Weight, data=mydata)
```

- Częste błędy



- Powód:
  - ‚literówki' w makrozmiennych
  - Błędne użycie makrozmiennej
  - Błędy składni R

Ćwiczenie 5

Ćwiczenie 5

Rozszerzyć proces modelowania na zbiorze PMAD_PVA **(diagram Donation_Analysis_2)** o model dostępny w R

Zaimportować diagram: **OpenSourceIntegration_DIAG_start.xml**

a) Zweryfikować, czy możliwa jest komunikacja SAS Enterprise Miner – R
b) Dodać model R do przebiegu modelowania
c) Porównać model z utworzonymi wcześniej

Ćwiczenie 5a

Zweryfikować, czy komunikacja SAS Enterprise Miner – R jest dostępna

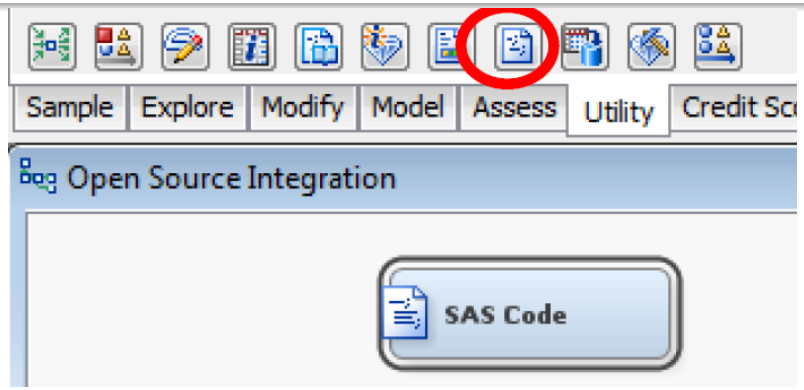
- W węźle SAS Code -> Code Editor wpisać polecenie:

```
proc options option=rlang;
run;
```

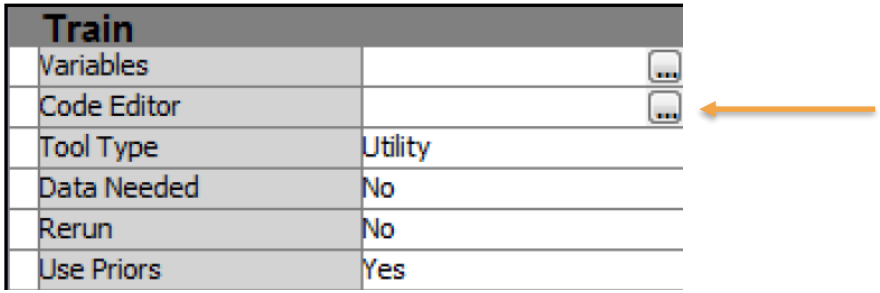
- W rezultatach węzła zweryfikować wynik procedury.

Ćwiczenie 5a

Użycie węzła SAS Code z zakładki Utility



Wybranie z opcji węzła edytora kodu



Wpisanie polecenia weryfikującego komunikację z SAS i R

```
proc options option=rlang;
run;
```

Weryfikacja wyniku w Rezultatach węzła (View->SAS Results->Log)

```
    SAS (r) Proprietary Software Release 9.4  TS1M3

 RLANG            Enables SAS to execute R language statements.
NOTE: PROCEDURE OPTIONS used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
```

Ćwiczenie 5a

Pytania kontrolne:

Wybierz prawidłowe stwierdzenie/stwierdzenia:

- Są 3 tryby wyniku dla węzła Open Source Integration: PMML, Merge, None
- PMML tworzy kod scoringowy automatycznie bez ingerencji użytkownika
- PMML jest wspierany przez wszystkie pakiety R
- Tryb Merge wymaga dodatkowej funkcjonalności R do utworzenia kodu scoringowego

Ćwiczenie 5b

Dodać model R do przebiegu modelowania

Zwrócić uwagę z jakich zmiennych korzysta model i na tej podstawie zidentyfikować odpowiednie miejsce przebiegu modelowania.

Pytania kontrolne:

- Jakie węzły utworzyły zmienne G_?

```
#Random Forest Model
#Using the randomForest R Package and Function

library(randomForest)

set.seed(12345)

myForest <- randomForest(TARGET_B ~ G_DemCluster + G_StatusCat96NK +
GiftAvg36 + GiftCnt36 + GiftCntCard36 + GiftTimeLast + StatusCatStarAll,
data = PVA, ntree = 500, mtry = 5, importance = TRUE)

importance(myForest)

png("RPlot.png")
plot(myForest, main='randomForest MSE Plot in R')
```

Ćwiczenie 5b

Dostosowanie kodu modelu R do wymagań SAS Enterprise Miner

Jaki tryb wyniku należy użyć w węźle Open Source Integration?

Czy w związku z trybem wyniku konieczne jest zastosowanie dodatkowych węzłów?

```
library(randomForest)

set.seed(12345)

&EMR_MODEL <- randomForest(&EMR_CLASS_TARGET ~
&EMR_CLASS_INPUT + &EMR_NUM_INPUT, data =
&EMR_IMPORT_DATA, ntree = 500, mtry = 5, importance = TRUE)

importance(&EMR_MODEL)
```

Ćwiczenie 5b

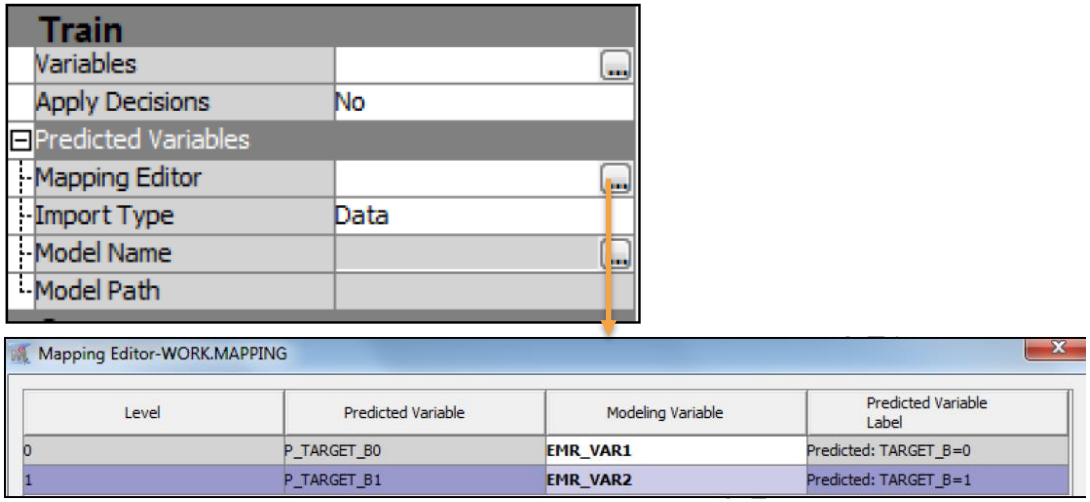
- Dodatkowe instrukcje wynikające z trybu MERGE

```
&EMR_EXPORT_TRAIN <- predict(&EMR_MODEL,
&EMR_IMPORT_DATA, type="prob")

&EMR_EXPORT_VALIDATE <- predict(&EMR_MODEL,
&EMR_IMPORT_VALIDATE, type="prob")

png("RPlot.png")
plot(&EMR_MODEL , main='randomForest MSE Plot in R')
```

- Dodatkowe węzły wynikające z instrukcji MERGE: Model Import
- Zmapowanie wyników scorowania dla obu poziomów zmiennej celu
- Zmiana nazwy węzła na **R Forest**

Pytania kontrolne:

- Czy las losowy jest dobrej jakości?
- Zweryfikować statystyki dla zbioru treningowego i walidacyjnego

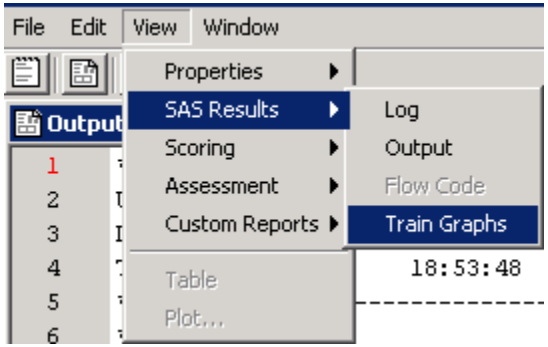


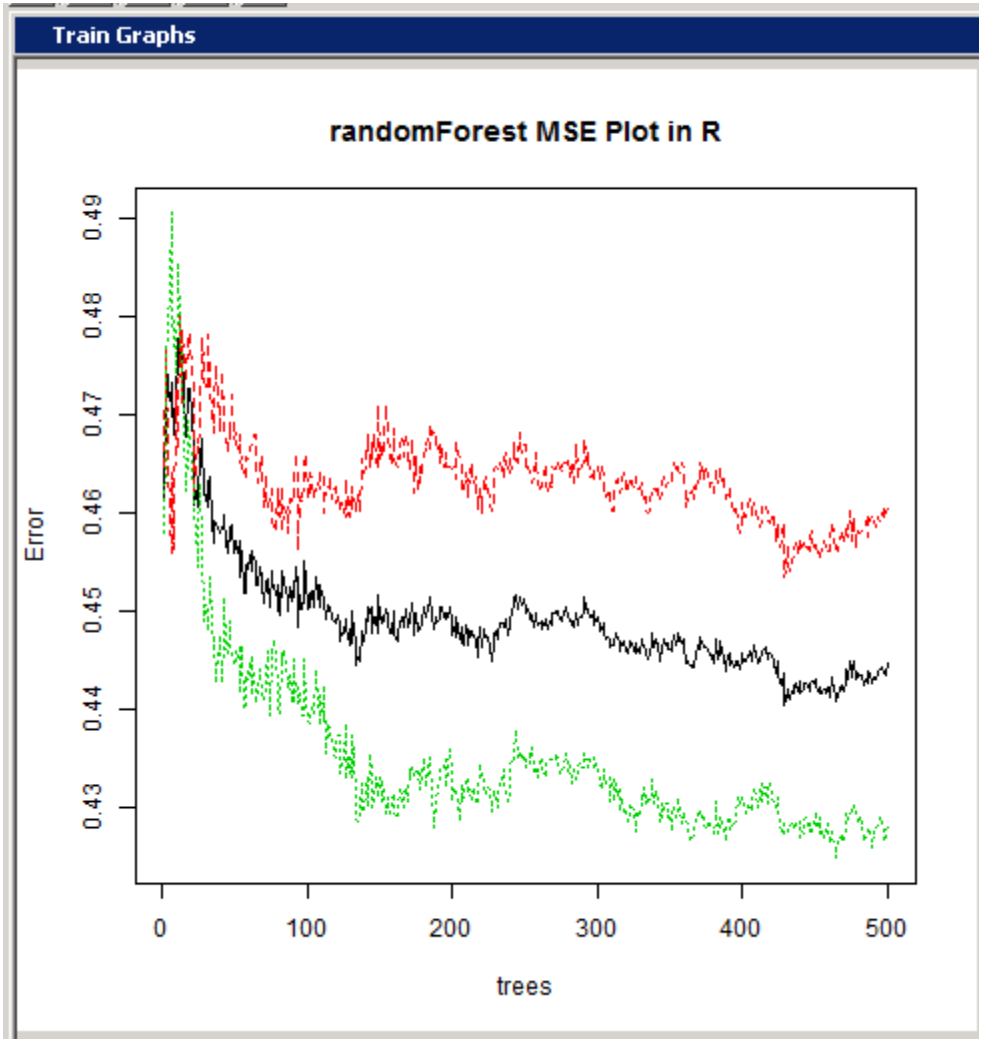
Wynik węzła Open Source Integration

Output

```
Type          Variable     Label

TARGET        TargetB      Target Gift Flag
PREDICTED     P_TargetB1   Predicted: TargetB=1
RESIDUAL      R_TargetB1   Residual: TargetB=1
PREDICTED     P_TargetB0   Predicted: TargetB=0
RESIDUAL      R_TargetB0   Residual: TargetB=0
FROM          F_TargetB    From: TargetB
INTO          I_TargetB    Into: TargetB



R Handle Contents

     Handle          Available    Contents

EMR_NUM_TARGET          Y
EMR_CLASS_TARGET        Y        rTargetB
EMR_NUM_INPUT           Y        GiftAvg36 + GiftCnt36 + GiftCntCard36 + GiftTimeLast + StatusCatStarAll
EMR_CLASS_INPUT         Y        G_DemCluster + G_StatusCat96NK
EMR_CHAR_INPUT          Y




                          0          1 MeanDecreaseAccuracy MeanDecreaseGini
G_DemCluster     13.008471 10.1775063            16.949152         329.9890
G_StatusCat96NK   5.244171 -0.4413477             4.492006         118.6405
GiftAvg36         2.059278 14.3220889            13.028500         920.0162
GiftCnt36         2.791688  9.6082676            12.987595         290.9384
GiftCntCard36     7.694253  8.0474099            14.382452         299.0109
GiftTimeLast     14.479443 16.4188308            22.708730         572.3205
StatusCatStarAll 18.767544 -3.4016246            12.757111         109.9019


*-------------------------------------------------------------*
* Score Output
*-------------------------------------------------------------*
```

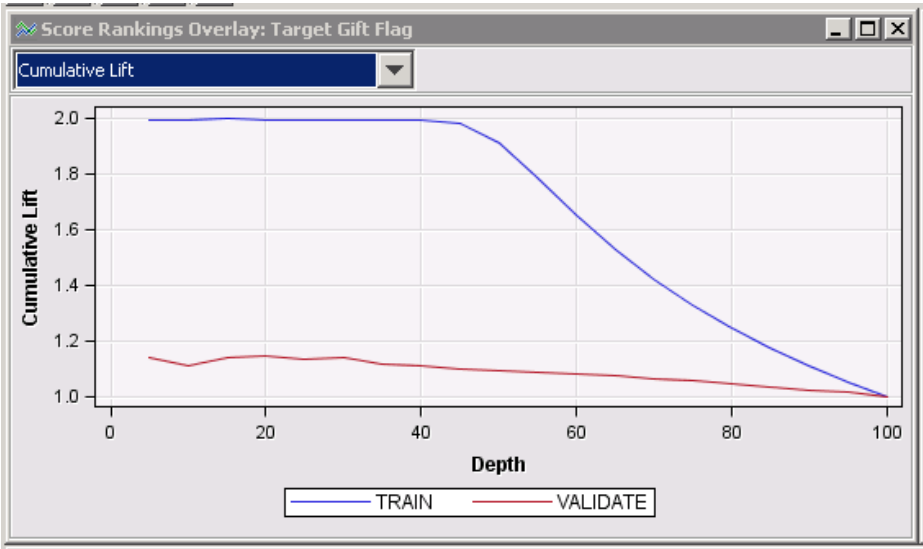
Wynik węzła Open Source Integration

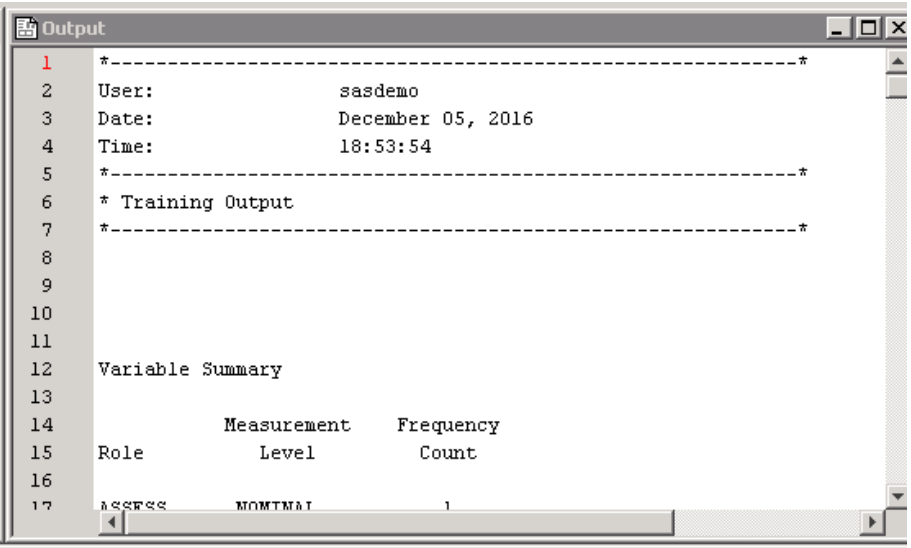




Wyniki węzła Model Import

Zastosować węzeł Model Import po węźle Open Source Integration





Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetB	Target Gift Flag	_ASE_	Average Squared Error	0.058529	0.272076	.
TargetB	Target Gift Flag	_DIV_	Divisor for ASE	11620	7752	.
TargetB	Target Gift Flag	_MAX_	Maximum Absolute Error	0.96	0.996	.
TargetB	Target Gift Flag	_NOBS_	Sum of Frequencies	5810	3876	.
TargetB	Target Gift Flag	_RASE_	Root Average Squared Error	0.241927	0.521609	.
TargetB	Target Gift Flag	_SSE_	Sum of Squared Errors	680.1016	2109.133	.
TargetB	Target Gift Flag	_DISF_	Frequency of Classified Ca...	5810	3876	.
TargetB	Target Gift Flag	_MISC_	Misclassification Rate	0.044923	0.456656	.
TargetB	Target Gift Flag	_WRONG_	Number of Wrong Classific...	261	1770	.

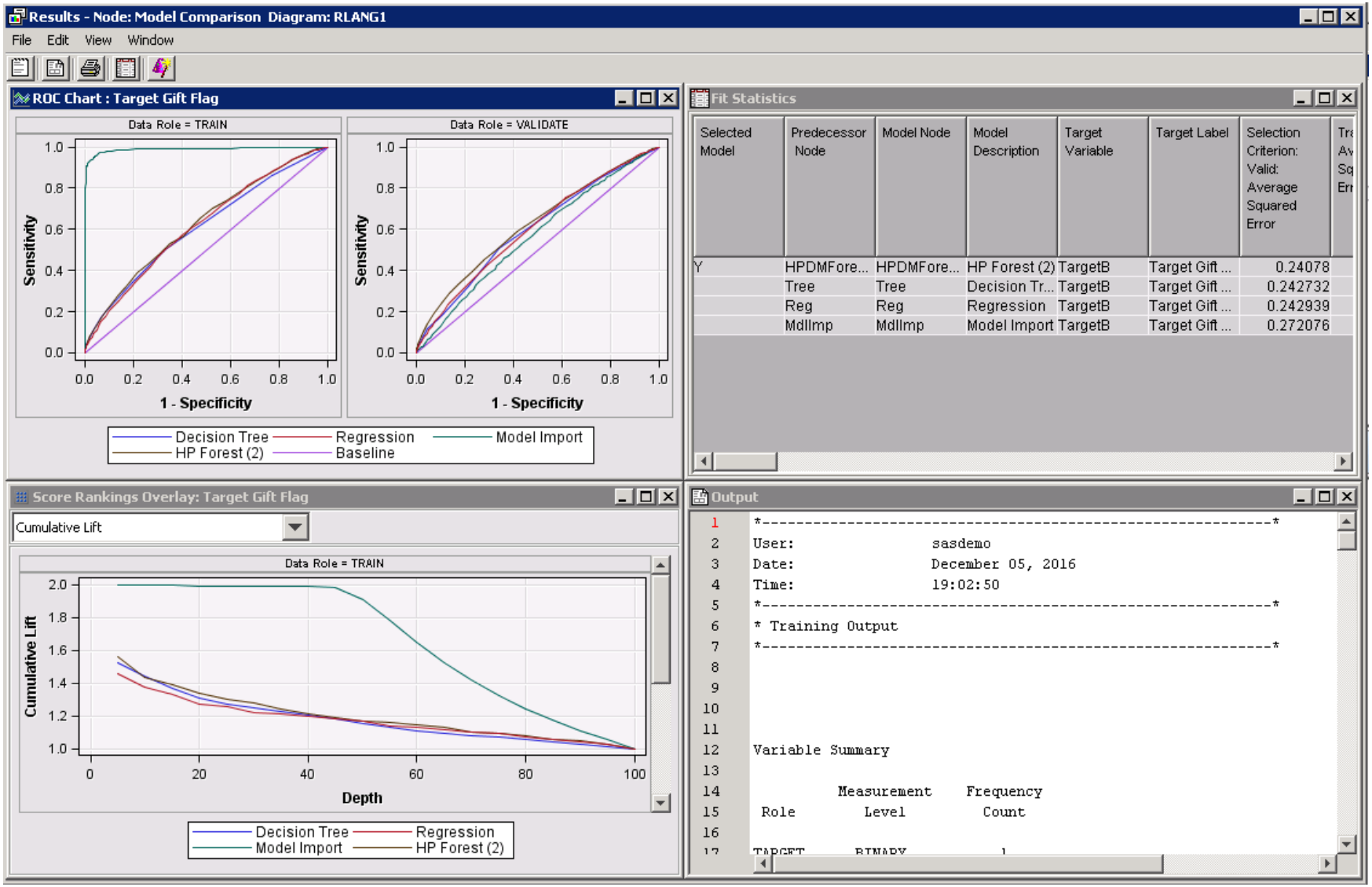
Ćwiczenie 5c

Porównać nowy model z pozostałymi

- Kryterium:
  - Średni błąd kwadratowy (ASE)
- Tabela:
  - Walidacyjna

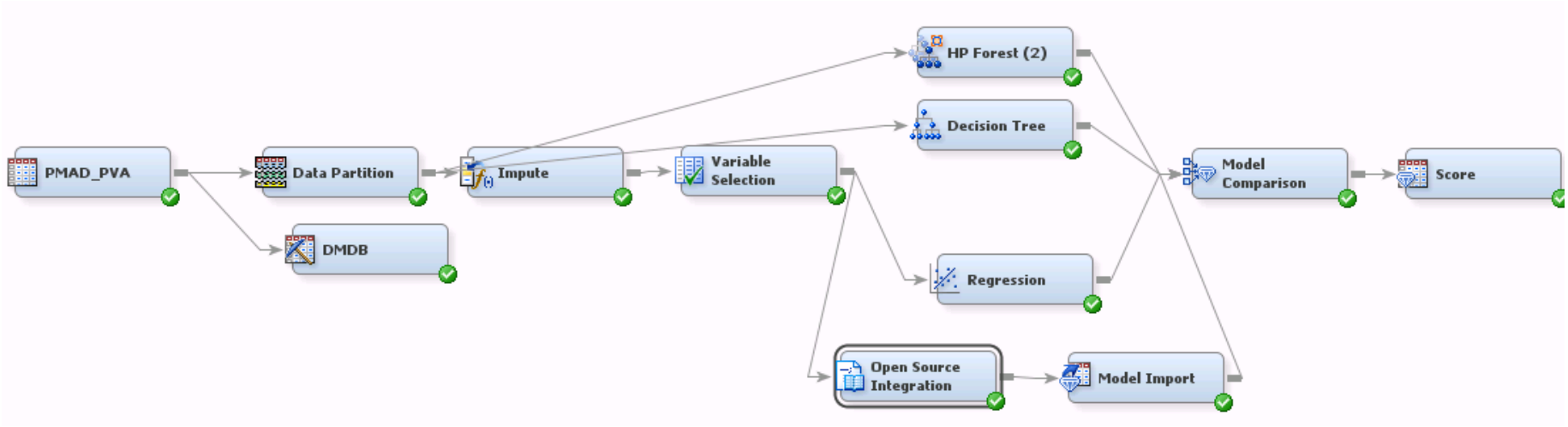
- Zinterpretować wyniki porównania

- Wykorzystać węzeł Model Comparison


Results - Node: Model Comparison Diagram: RLANG1
File Edit View Window
ROC Chart : Target Gift Flag
Data Role = TRAIN
Data Role = VALIDATE
Sensitivity
1 - Specificity
Decision Tree
Regression
Model Import
HP Forest (2)
Baseline
Fit Statistics
Selected Model
Predecessor Node
Model Node
Model Description
Target Variable
Target Label
Selection Criterion: Valid: Average Squared Error
Y HPDMFore... HPDMFore... HP Forest (2) TargetB Target Gift ... 0.24078
Tree Tree Decision Tr... TargetB Target Gift ... 0.242732
Reg Reg Regression TargetB Target Gift ... 0.242939
MdlImp MdlImp Model Import TargetB Target Gift ... 0.272076
Score Rankings Overlay: Target Gift Flag
Cumulative Lift
Data Role = TRAIN
Cumulative Lift
Depth
Output
User: sasdemo
Date: December 05, 2016
Time: 19:02:50
* Training Output
Variable Summary
Role
Measurement Level
Frequency Count

Ćwiczenie 5b

- Diagram wynikowy



Ćwiczenie 6

Opis zbioru danych - Organics

Supermarket wprowadza nową linię produktów Organicznych.
Oczekuje informacji, którzy klienci będą skłonni do zakupu nowych produktów.
Sklep posiada program lojalnościowy, w którym klienci otrzymują kupony na zakup produktów Organicznych.

Nazwa zmiennej	Etykieta	Rola, Poziom
AFFL	Ocena zamożności	INPUT, INTERVAL
AGE	Wiek	INPUT, INTERVAL
AGEGRP1	Grupa wiekowa 1	REJECTED
AGEGRP2	Grupa wiekowa 2	REJECTED
BILL	Kwota płatności	INPUT, INTERVAL
CLASS	Status lojalnościowy klienta	INPUT, NOMINAL
CUSTID	Identyfikator klienta	ID
DOB	Data urodzenia	REJECTED
EDATE	Data danych	REJECTED
GENDER	Płeć (F,M,U)	INPUT, NOMINAL
LCDATE	Data wniosku o kartę lojalnościową	REJECTED
LTIME	Liczba lat w programie lojalnościowym	INPUT, INTERVAL
NEIGHBORHOOD	Typ sąsiedztwa	REJECTED
NGROUP	Grupa sąsiedztwa	INPUT, NOMINAL
ORGANICS	Liczba zakupionych produktów Organicznych	REJECTED
ORGYN	Flaga zakupu produktów Organicznych (1- tak, 0 – Nie)	TARGET, BINARY
REGION	Region geograficzny	INPUT, NOMINAL
TV_REG	Region telewizyjny	INPUT, NOMINAL

Ćwiczenie 6

W zespole tworzone są modele za pomocą narzędzi:

- SAS Enterprise Miner
- R

Należy porównać ich jakość z użyciem SAS Enterprise Miner.

a) Zbudować drzewo decyzyjne
   a) Zastosować metodę imputacji – wymaganą dla modeli R
b) Zaimportować model zbudowany w języku R
c) Porównać modele

Ćwiczenie 6a

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Organics
    - Zmienna celu: Orgyn
- Diagram
  - Organics

Pytania kontrolne:

- Jaki % osób kupuje produkty z linii Organicznej?

- Budowa drzewa decyzyjnego:
  - Dodać do projektu zbiór **Organics**
    - Ustawić role i poziomy zmiennych zgodnie z tabelą opisu zbioru
  - Wykonać eksplorację zbioru
    - Zmienić sposób losowania próby z Top na Random
  - Dokonać podziału zbioru na część treningową i walidacyjną (w proporcji 70/30)
  - Zastosować metodę uzupełniania braków danych (średnia dla zmiennych ciągłych)
    - Utworzyć zmienne (indykatory) dla zmiennych zawierających braki danych, dopuścić je jako zmienne wejściowe w dalszej części modelowania
  - Zbudować model drzewa decyzyjnego z domyślnymi parametrami
  - Zmienić metodę wyboru najlepszego drzewa z domyślnej na Assessment

Ćwiczenie 6a

Pytania kontrolne:

- Jaki % osób kupuje produkty z linii Organicznej?
- Które zmienne zostały użyte do budowy drzewa
- W jak wielu podziałach zostały użyte poszczególne zmienne?

- Impute
  - Indykatory dla zmiennych z brakami danych

Score	
Hide Original Variables	Yes
⊟ Indicator Variables	
├ Type	Unique
├ Source	Imputed Variables
└ Role	Input

Ćwiczenie 6b

- Dodać do projektu kod modelu utworzony w R
  - Model został utworzony z wykorzystaniem pakietu rpart wspieranego przez PMML
- Przekształcić skrypt, tak aby był czytelny dla SAS Enterprise Miner
- Dodać instrukcje wymagane do zilustrowania drzewa

Oryginalna postać skryptu:

```
#Create an R Decision Tree
#Use the rpart R Package

library(rpart)

mytree <- rpart(ORGYN ~ AFFL + AGE + BILL + LTIME + NGROUP +
GENDER + REGION + TV_REG + CLASS, data=organics)
```

Dodatkowe instrukcje:

```
png("RPlot.png")
plot(mytree , main='R Object Plot')
```

Zastosować finalny kod w węźle Open Source Integration

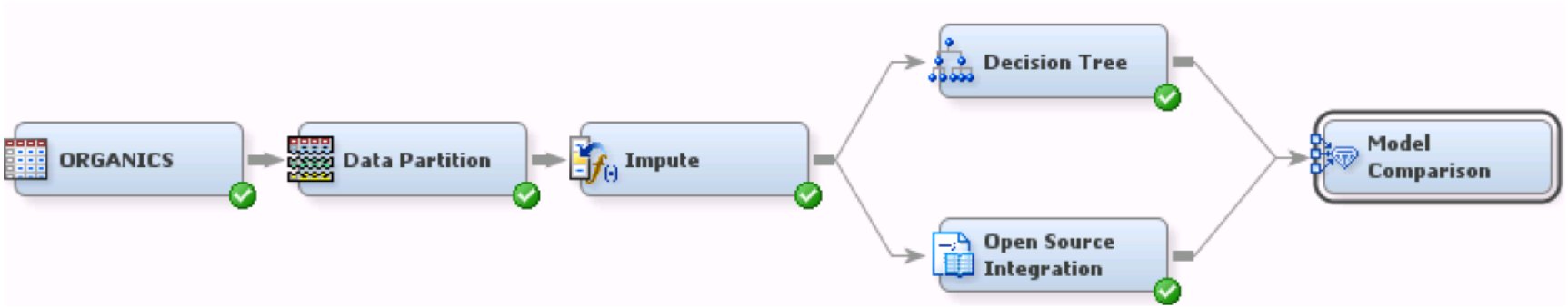
Ćwiczenie 6c

Porównać utworzone modele

Pytania kontrolne:

- Który model ma lepszą jakość?
  - Zinterpretować miary ASE i Liftu skumulowanego

- Wykorzystać węzeł Model Comparison



TECHNIKI GRUPOWANIA DANYCH

Metody określania liczby skupień

- Clustering Cubic Criterion - CCC
- pseudo-F
- pseudo-T^2

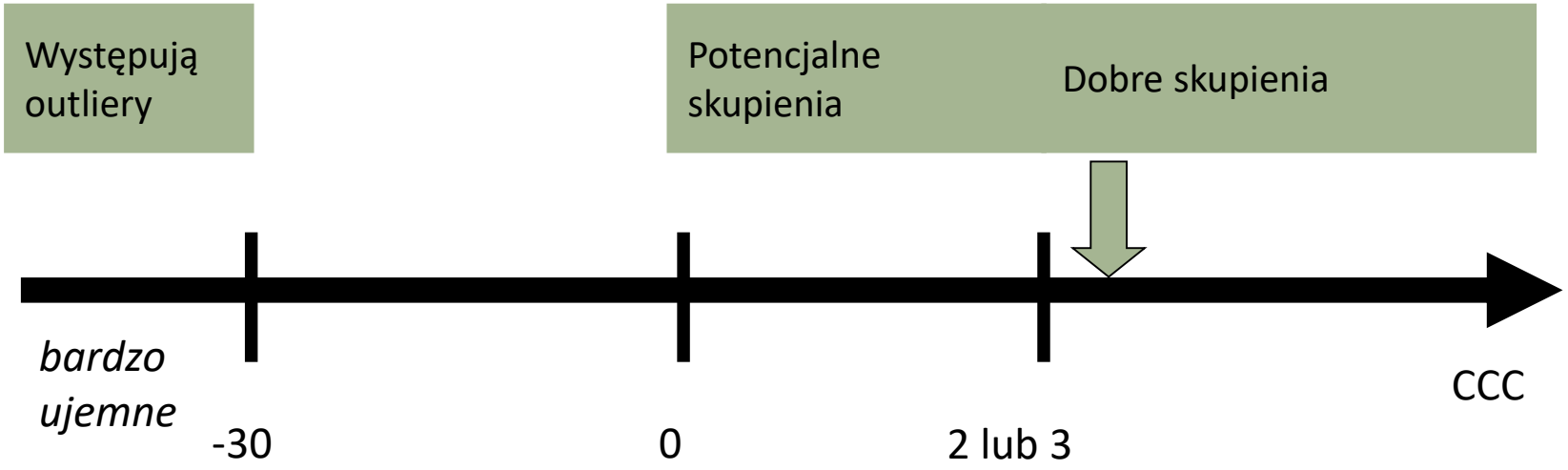
Clustering Cubic Criterion - CCC

- Kryterium wyboru optymalnej liczby skupień
- Kryterium bazuje na porównaniu oczekiwanej i obserwowanej sumy kwadratów odległości obserwacji wewnątrz skupień
- Formuła liczenia kryterium CCC została ustalona eksperymentalnie

$$CCC = ln\left[\frac{1 - E(R^2)}{1 - R^2}\right]\frac{\sqrt{\frac{np^*}{2}}}{(.001 + E(R^2))^{1.2}}$$

- Wizualizowane jest na wykresie:
  - Oś X przedstawia liczbę skupień (w zakresie od 1 do N/10, gdzie N to liczba obserwacji)
  - Oś Y przedstawia wartość CCC
- Kryterium nie nadaje się do oceny skupień jeśli są:
  - Mało liczne (mniej niż 10 obserwacji w skupieniu)
  - Nieregularne lub wydłużone

Interpretacja kryterium CCC



Należy szukać lokalnych maksimów dla CCC ≥ 2(3)

Statystyka pseudo-F

- Statystyka pseudo-F (PSF) mierzy różnorodność między grupami na zadanym poziomie hierarchii
- Wysokie wartości statystyki oznaczają, że występują różnice między zidentyfikowanymi grupami (średnie wartości zmiennych różnią się)
- Statystyka nie ma rozkładu F-Snedecora

Statystyka pseudo-F

Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
15	CL16	OB94	22	0.0015	.921	.975	-11	68.4	1.4	
14	CL19	OB49	28	0.0021	.919	.972	-11	72.4	1.8	
13	CL15	OB52	23	0.0024	.917	.969	-10	76.9	2.3	
12	CL13	OB96	24	0.0018	.915	.966	-9.3	83.0	1.6	
11	CL12	OB93	25	0.0025	.912	.962	-8.5	89.5	2.2	
10	CL11	OB78	26	0.0031	.909	.957	-7.7	96.9	2.5	
9	CL10	OB76	27	0.0026	.907	.951	-6.7	107	2.1	
8	CL9	OB77	28	0.0023	.904	.943	-5.5	120	1.7	
7	CL8	OB43	29	0.0022	.902	.933	-4.1	138	1.6	
6	CL7	OB87	30	0.0043	.898	.920	-2.7	160	3.1	
5	CL6	OB82	31	0.0055	.892	.902	-1.1	191	3.7	
4	CL22	OB61	37	0.0079	.884	.875	0.93	237	10.6	
3	CL14	OB23	29	0.0126	.872	.827	2.60	320	10.4	
2	CL4	CL3	66	0.2129	.659	.697	-1.3	183	172	
1	CL2	CL5	97	0.6588	.000	.000	0.00	.	183	

Potencjalne liczby segmentów

Statystyka pseudo-T^2

- Statystyka pseudo-T^2 jest odmianą testu T^2 Hotellinga
- Duża wartość statystyki oznacza, że średnie wartości zmiennych w grupach różnią się istotnie -> grupy nie powinny być łączone

Statystyka pseudo-T^2

Cluster History

NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Tie
15	CL16	OB94	22	0.0015	.921	.975	-11	68.4	1.4	
14	CL19	OB49	28	0.0021	.919	.972	-11	72.4	1.8	
13	CL15	OB52	23	0.0024	.917	.969	-10	76.9	2.3	
12	CL13	OB96	24	0.0018	.915	.966	-9.3	83.0	1.6	
11	CL12	OB93	25	0.0025	.912	.962	-8.5	89.5	2.2	
10	CL11	OB78	26	0.0031	.909	.957	-7.7	96.9	2.5	
9	CL10	OB76	27	0.0026	.907	.951	-6.7	107	2.1	
8	CL9	OB77	28	0.0023	.904	.943	-5.5	120	1.7	
7	CL8	OB43	29	0.0022	.902	.933	-4.1	138	1.6	
6	CL7	OB87	30	0.0043	.898	.920	-2.7	160	3.1	
5	CL6	OB82	31	0.0055	.892	.902	-1.1	191	3.7	
4	CL22	OB61	37	0.0079	.884	.875	0.93	237	10.6	
3	CL14	OB23	29	0.0126	.872	.827	2.60	320	10.4	
2	CL4	CL3	66	0.2129	.659	.697	-1.3	183	172	
1	CL2	CL5	97	0.6588	.000	.000	0.00	.	183	

Potencjalne liczby klastrów

Ćwiczenie 8

DODATKOWE

Opis zbioru danych - Hotel

Dane o średnich wydatkach klientów w sieci hoteli.

Nazwa zmiennej	Etykieta
CUST_ID	Identyfikator klienta
SEX	Płeć
AVG_AGE	Średni wiek
AVG_DURATION	Średni czas pobytu
SUM_Babysitting_childcare	Wartość usług opieki nad dzieckiem
SUM_Beauty_services	Wartość usług salunu piekności
SUM_Business_center	Wartość usług centrum biznesowego
SUM_Business_services	Wartość usług biznesowych
SUM_Bar_lounge	Wartość wydatków w barze
SUM_Casino	Wartość wydatków w kasynie
SUM_Hair_salon	Wartość wydatków fryzjerskich
SUM_Full_service_health_spa	Wartość wydatków spa
SUM_Sauna	Wartość wydatków na saunę
SUM_Room	Wartość wydatków na pokój
SUM_Spa_services_on_site	Wartość wydatków spa w hotelu

Sortowanie zbiorów danych

```
proc sort data = zbiór_wejściowy
          out = zbiór_wejściowy ;
        by zmienna;
run;
```

```
proc sort data = zbiór_wejściowy (keep = lista_zmiennych_oddzielonych_spacją)
          out = zbiór_wejściowy (rename=(orig_nazwa_zmiennej=nowa_nazwa_zmiennej));
        by zmienna;
run;
```

Nazwa zbioru wejściowego jest zapisywana w konwencji: **nazwa_biblioteki.nazwa_zbioru**
Jeżeli zbiór znajduje się w bibliotece WORK, to poniższe zapisy są sobie równoważne:
WORK.nazwa_zbioru = nazwa_zbioru

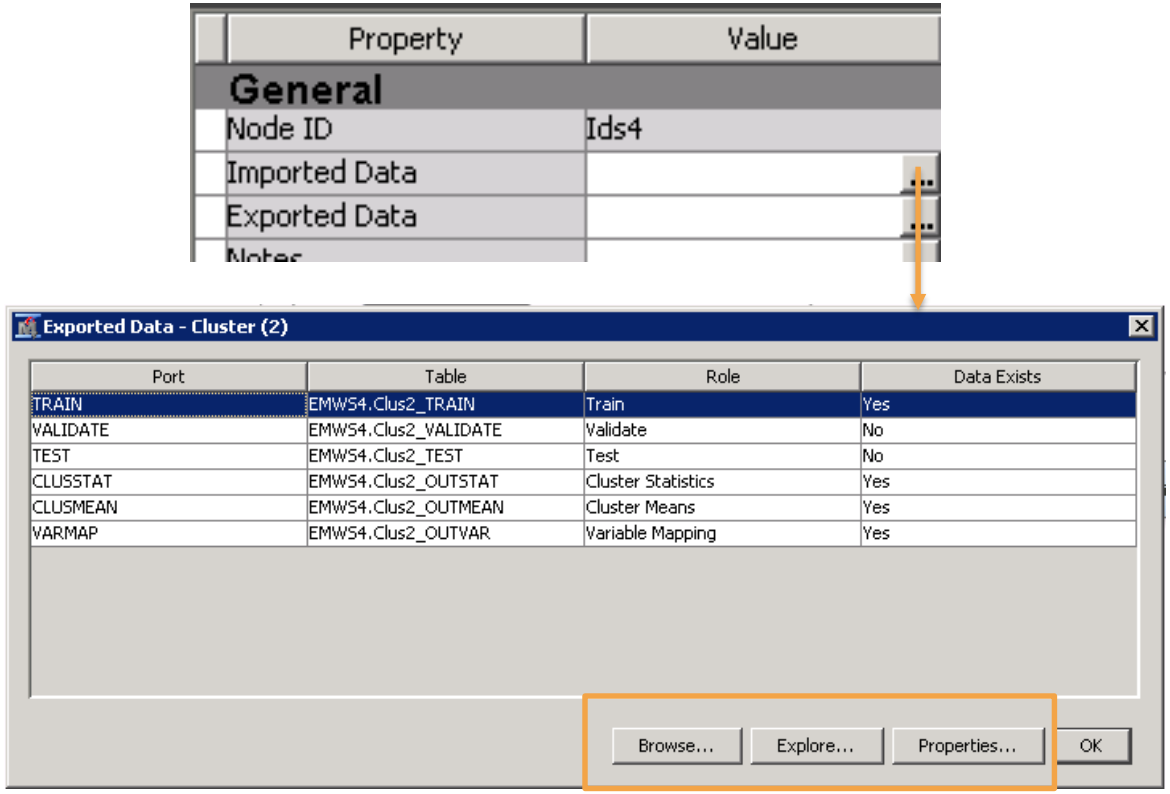
Łączenie posortowanych zbiorów danych

```
data nazwa_zbioru_wyjściowego;
     merge nazwa_zbioru_wejściowego1
           nazwa_zbioru_wejściowego2;
     by zmienna;
run;
```

Obliczanie liczebności przecięć wartości 2 zmiennych

```
proc freq data=nazwa_zbioru_wejściowego;
        tables nazwa_zmiennej1*nazwa_zmiennej2/ out=nazwa_zbioru_wyjściowego;
run;
```

Lokalizacja zbiorów wynikowych z analiz



Ćwiczenie 8

Sieć hoteli chce dopasować ofertę do profilu swoich klientów

Jakie segmenty klientów można wyróżnić na podstawie danych zebranych przez sieć?

Wykonać segmentację klientów

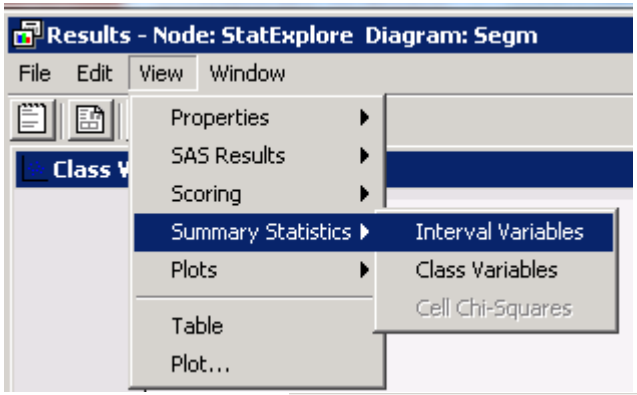
- Projekt
  - Enterprise_Miner_unsup_lab_*nazwisko*
- Zbiór źródłowy
  - Hotel – do nauki modeli
  - Hotel_score – do scoringu
- Diagram
  - SEGM_Hotel
- Model
  - Cluster

a) Zdecydować, czy/jak należy uzupełnić wartości brakujące
b) Zdecydować, czy występują wartości odstające
c) Wykonać standaryzację zmiennych wejściowych
d) Wykonać segmentację
e) Wykonać profilowanie
f) Dla 3 klientów, którzy nie brali udziału w segmentacji, zweryfikować, do których segmentów należą

Ćwiczenie 8a

Zdecydować, czy i jak należy uzupełnić wartości brakujące

- Wczytać zbiór z domyślnymi ustawieniami
- Zastosować węzeł StatExplore lub DMDB i zweryfikować wartości statystyk



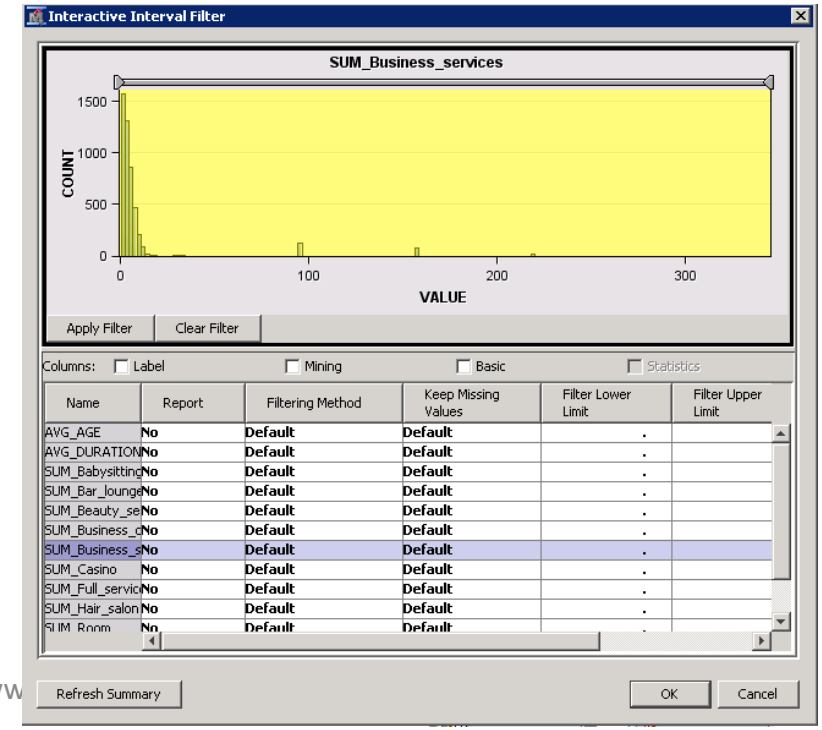
Interval Variables

Ordered Inputs	Data Role	Variable	Median	Missing	Non Missing	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis	Role	Label
1	TRAIN	SUM_Busin...	0			0	98.25	1.58982	6.950135	5.625829	36.52547	INPUT	Wartosc usl...
2	TRAIN	SUM_Baby...	0			0	914.9	20.88847	77.2036	5.066172	28.61411	INPUT	Wartosc usl...
3	TRAIN	SUM_Bar_I...	5.369			0	3438.75	73.91309	241.0405	5.559344	36.27598	INPUT	Wartosc wy...
4	TRAIN	SUM_Hair_...	0			0	3151.05	126.5566	363.2727	3.319181	11.94421	INPUT	Wartosc wy...
5	TRAIN	SUM_Busin...	3.388			0	343.875	8.701193	23.79675	5.620522	37.085	INPUT	Wartosc usl...
6	TRAIN	SUM_Spa_...	0			0	272.7	29.96305	47.44435	1.607941	2.106739	INPUT	Wartosc wy...
7	TRAIN	SUM_Beaut...	68.88			0	1050.35	112.2314	137.2387	1.546082	3.184935	INPUT	Wartosc usl...
8	TRAIN	SUM_Full_...	249.9			0	2863.35	416.3481	496.4386	1.22662	1.235098	INPUT	Wartosc wy...
9	TRAIN	SUM_Casino	933.45			0.574	9235.8	1303.55	1362.35	1.542833	2.50822	INPUT	Wartosc wy...
10	TRAIN	SUM_Room	426.006			7.455	6267.065	592.1835	564.7322	2.759209	11.78415	INPUT	Wartosc wy...
11	TRAIN	SUM_Sauna	3.92			0	19.54667	4.077237	3.591057	0.723448	0.156311	INPUT	Wartosc wy...
12	TRAIN	AVG_AGE	47			19	97	50.25194	18.9724	0.365937	-0.90215	INPUT	Sredni wiek
13	TRAIN	AVG_DURA...	8.666667			1	14.6	8.020647	2.315407	-0.53896	-0.46375	INPUT	Sredni czas...

Ćwiczenie 8b

Zdecydować, czy występują wartości odstające

- Obejrzeć rozkłady zmiennych (po jednokrotnym uruchomieniu) w węźle Filter



Ćwiczenie 8b

Wykonać segmentację bez i z filtrowaniem

- W metodzie filtrowania wybrać metodę percentyli

Ćwiczenie 8c

Wykonać standaryzację zmiennych wejściowych

- Zastosować metodę standaryzacji dostępną w węźle Cluster

Ćwiczenie 8d

Wykonać segmentację

Pytania kontrolne:

- Który z węzłów najbardziej jednoznacznie wskazuje liczbę segmentów?
  - Które kryteria należy zweryfikować?
- Która zmienna jest najbardziej istotna w poszczególnych segmentach najlepszej segmentacji?
- Które z segmentów są najmniej do siebie podobne?
- Który z segmentów jest najbardziej liczny?
- W których segmentach przeważają mężczyźni, a w których kobiety?

- Zastosować węzły Cluster
- We wszystkich przypadkach określić liczbę potencjalnych skupień

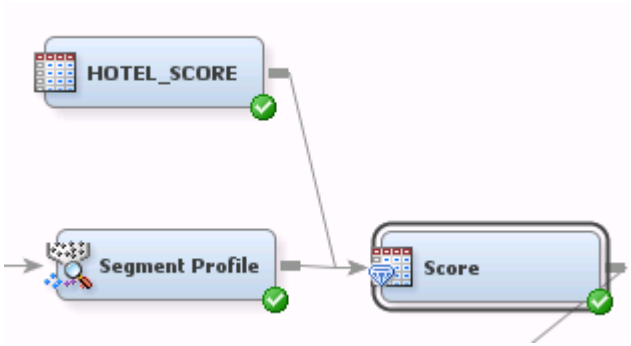
Ćwiczenie 8e

Wykonać profilowanie

- Zastosować węzeł Profile do scharakteryzowania poszczególnych segmentów wybranej segmetacji
- Opcjonalnie: W dokumencie .doc opisać poszczególne segmenty, ilustrując swe wnioski wykresami i statystykami dostępnymi w węźle Profile

Ćwiczenie 8f

Dla 3 klientów, którzy nie brali udziału w segmentacji, zweryfikować, do których segmentów należą



a) Wykorzystać węzeł Score oraz zbiór Hotel Score

b) Opcjonalnie: Zapisać w dokumencie, do których segmentów należą klienci

Ćwiczenie 8*

Wykonać segmentację po wcześniejszym zgrupowaniu zmiennych ciągłych

Zweryfikować, czy statystyki CCC, pseudo F, pseudo T2 jednoznacznie wskazują na liczbę segmentów

Pytania kontrolne:

- Czy uzyskane segmenty są zbliżone do poprzednio uzyskanych
- Czy zmieniły się zmienne najbardziej istotne w poszczególnych segmentach?

- Połączyć zbiory wynikowe z 2 segmentacji i porównać przypisanie poszczególnych klientów do segmentów

Ćwiczenie 8*

Użyć edytora kodu w SAS EM:

- Przed połączeniem zbiorów danych należy je posortować po identyfikatorze klienta
- Wykorzystać procedurę SORT
- Do połączenia zbiorów wykorzystać DATA STEP z instrukcję MERGE.
- Do wyliczenia statystyk wykorzystać procedurę FREQ.
- Wszystkie nowoutworzone zbiory zapisać w bibliotece tymczasowej WORK

```
proc sort data = biblioteka.nazwa_zbioru_wejściowego (keep =
zmienna_w_której_odbywa_sie_sortowanie
nazwa_zmiennej_wynikowej_z_segmentacji)
                    out = biblioteka.nazwa_zbioru_wyjściowego
(rename=(nazwa_zmiennej_wynikowej_z_segmentacji=nowa_nazwa_zmiennej
_wynikowej_z_segmentacji));
        by zmienna_w_której_odbywa_sie_sortowanie;
run;


data biblioteka.nazwa_zbioru_wyjściowego;
        merge biblioteka.nazwa_zbioru_wejściowego1
                    biblioteka.nazwa_zbioru_wejściowego2;
        by zmienna_w_której_odbywa_sie_łączenie;
run;

proc freq data=biblioteka.nazwa_zbioru_wejściowego;
        tables
nowa_nazwa_zmiennej_wynikowej_z_segmentacji1*nowa_nazwa_zmiennej_wy
nikowej_z_segmentacj2/out= biblioteka.nazwa_zbioru_wyjściowego;
run;
```