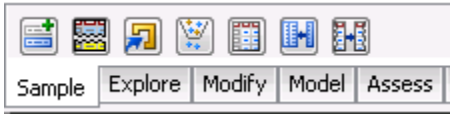
Zaawansowana analityka z SAS Enterprise Miner

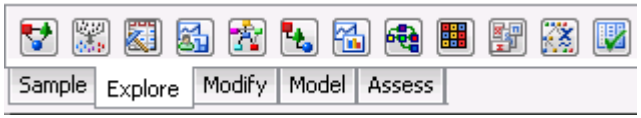
Edycja 6 - 2019/2020

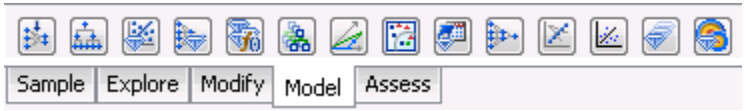
Laboratorium

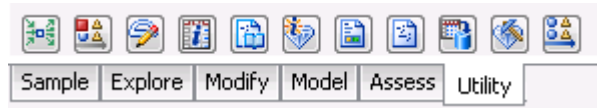
Metodologia SEMMA

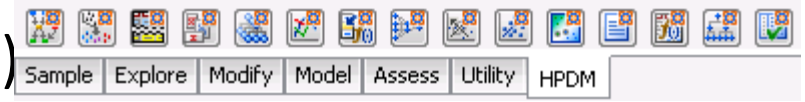
- Próbkowanie (**S**ample)
- Eksploracja (**E**xplore)
- Modyfikacja (**M**odify)
- Modelowanie (**M**odel)
- Ocenianie (**A**ssess)
- Użytkowe (Utility)
- High Performance Data Mining (HPDM)



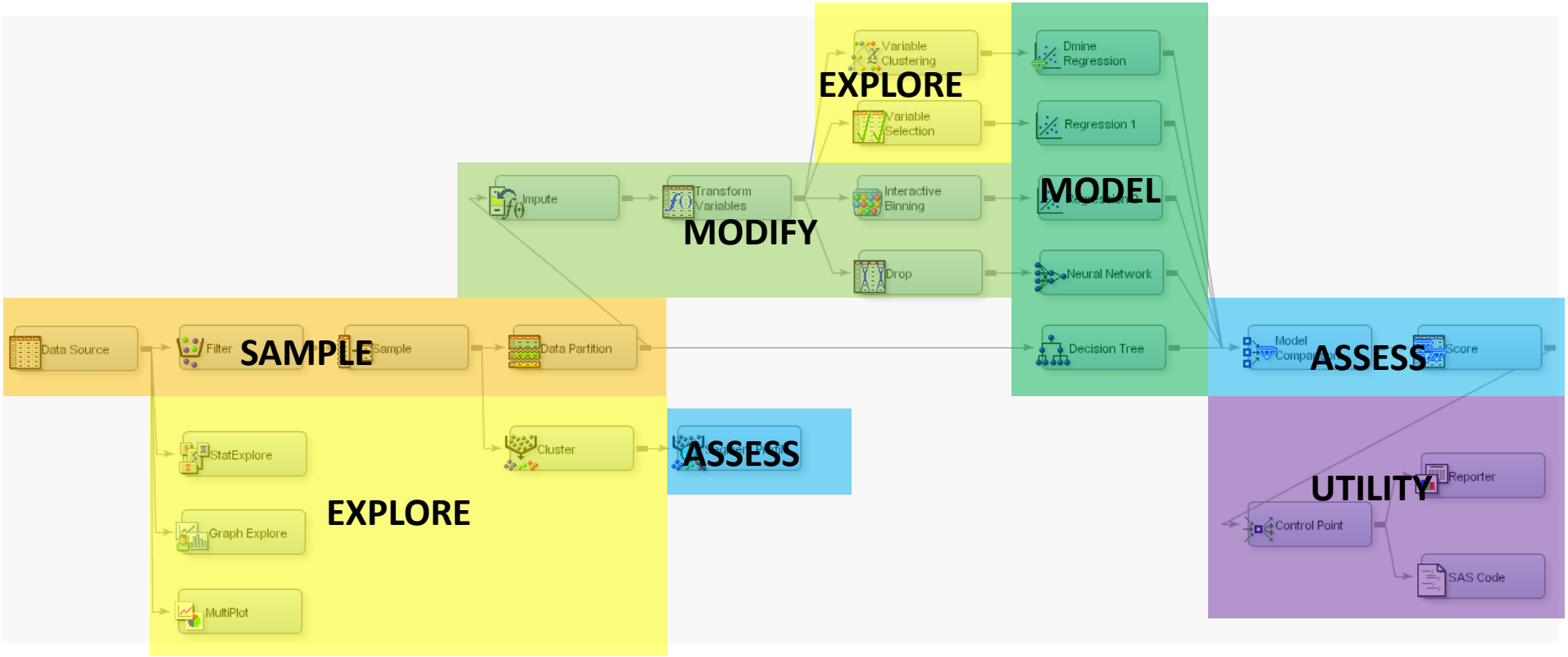








Przykład diagramu analizy danych



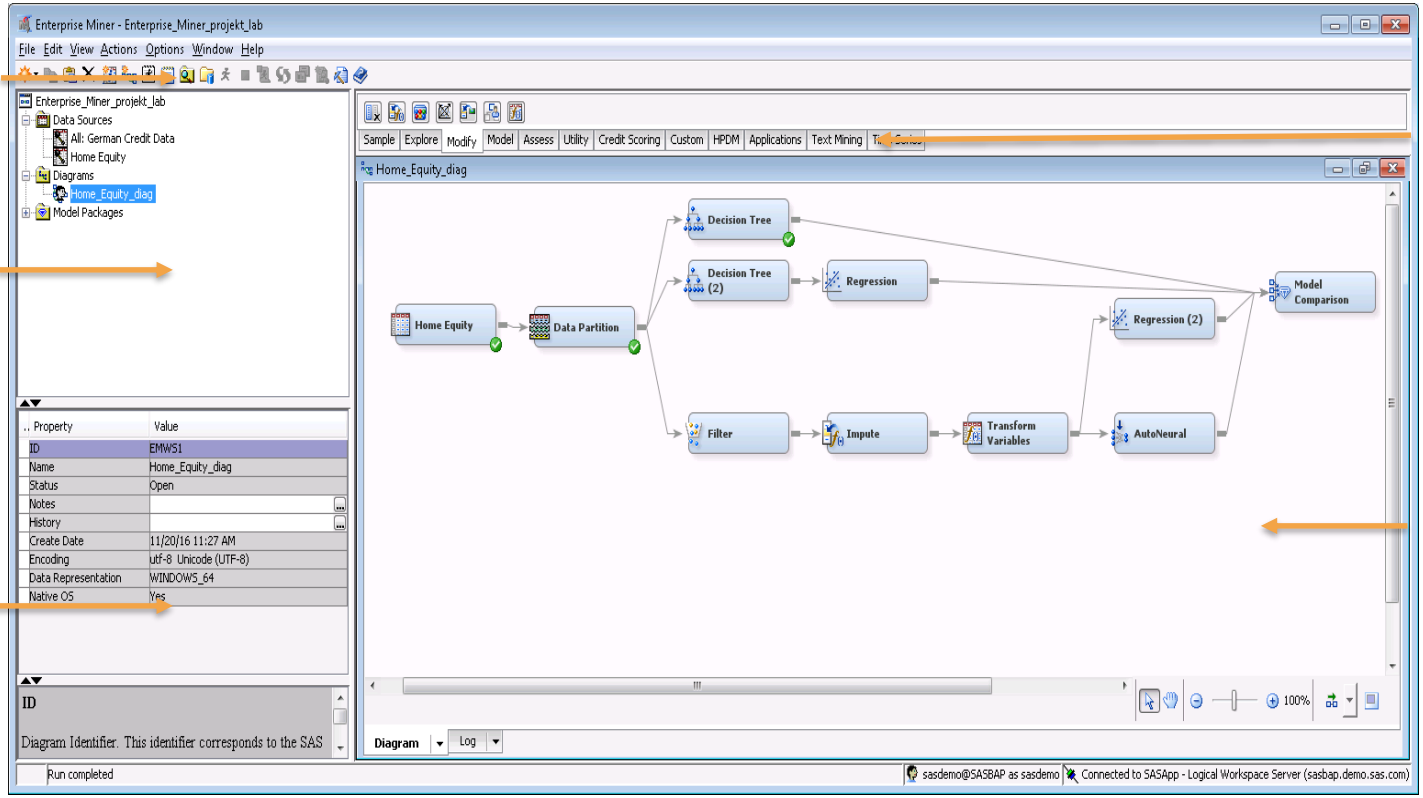
Rozpoczęcie pracy w SAS EM

Budowa interfejsu użytkownika

Pasek skrótów

Okno drzewa projektu

Okno właściwości

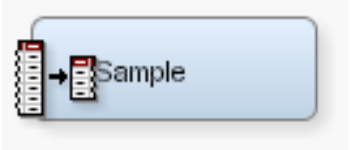


Pasek narzędzi

Okno diagramu

SEMMA

SAMPLE



Podział zbioru

Próbkowanie

KLIENT A	0
KLIENT B	1
KLIENT C	1
KLIENT D	0
.	0
	0
.	0
.	0
.	0
.	1
	0
.	0
.	0
.	1
.	1
.	0
KLIENT Z	1

Identyfikacja wartości odstających

KLIENT B	1	1
KLIENT C	2	1
KLIENT F	3	0
KLIENT H	2	0
.	5	0
	8	0
.	50	0
.	82	0
.	8	0
.	1	1
	2	0
.	4	0
.	6	0
.	0	1
.	7	1
.	1	0
KLIENT W	3	0

SEMMA

SAMPLE SCHEMATY PRÓBKOWANIA

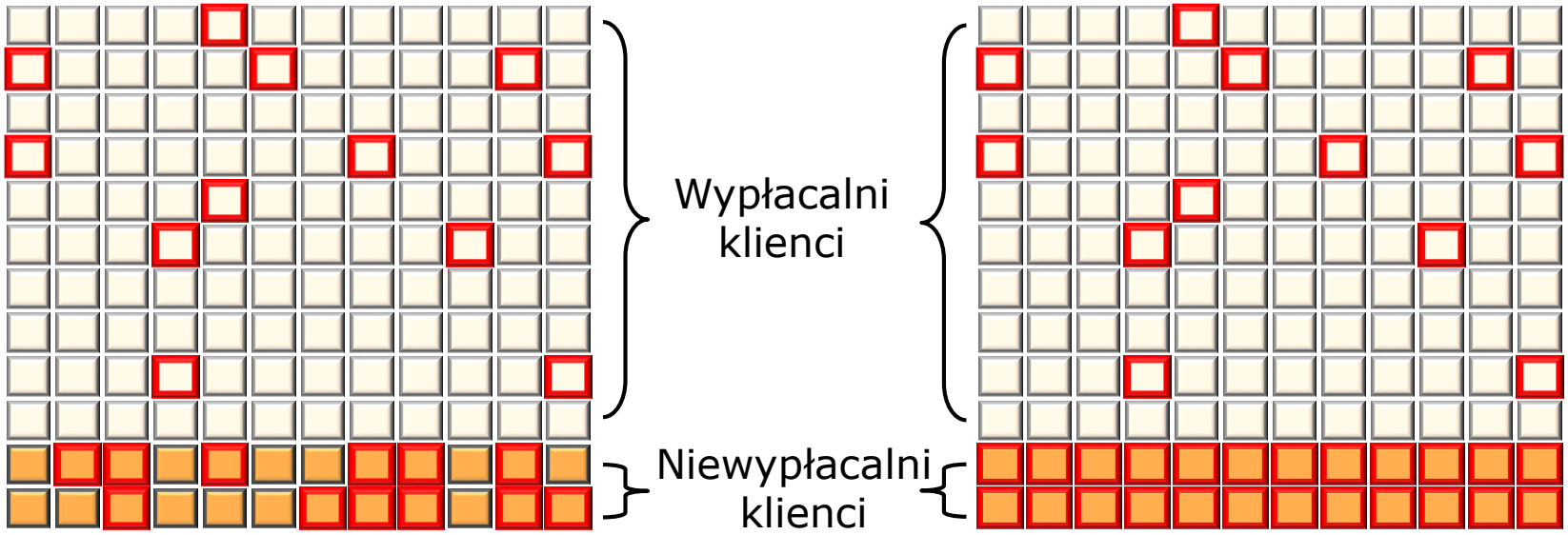
- N-krotne losowanie bez zwracania
  - Każdy element wylosowany do próbki, nie jest już brany pod uwagę w następnym losowaniu.
- Losowanie systematyczne
  - Do próbki brany jest co n-ty element począwszy od pewnego losowo lub nielosowo określonego elementu.
- Wzięcie do próbki N-pierwszych obserwacji.
- Losowanie grupowe/warstwowe.
- Losowanie warstwowe proporcjonalne

Aby próba była reprezentatywna dla całego zbioru, rozmiar próbki w każdej warstwie powinien być proporcjonalny do jej liczności.

SEMMA

SAMPLE - NADRÓBKOWANIE/PRZEPRÓBKOWANIE

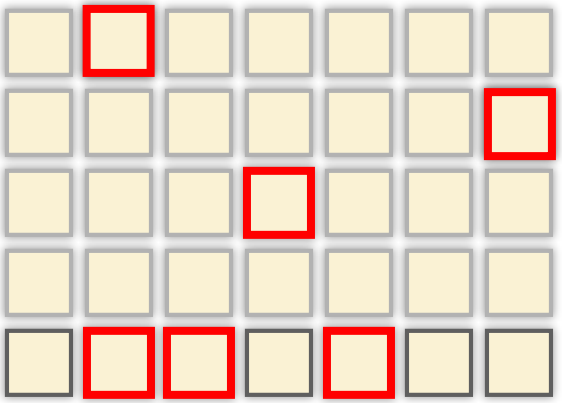
- Czasem stosuje się tzw. przepróbkowanie (oversampling) – losując tak samo dużą próbkę z każdej warstwy, niezależnie od jej rozmiaru.



SEMMA

SAMPLE - KIEDY STOSOWAĆ PRZEPRÓBKOWANIE?

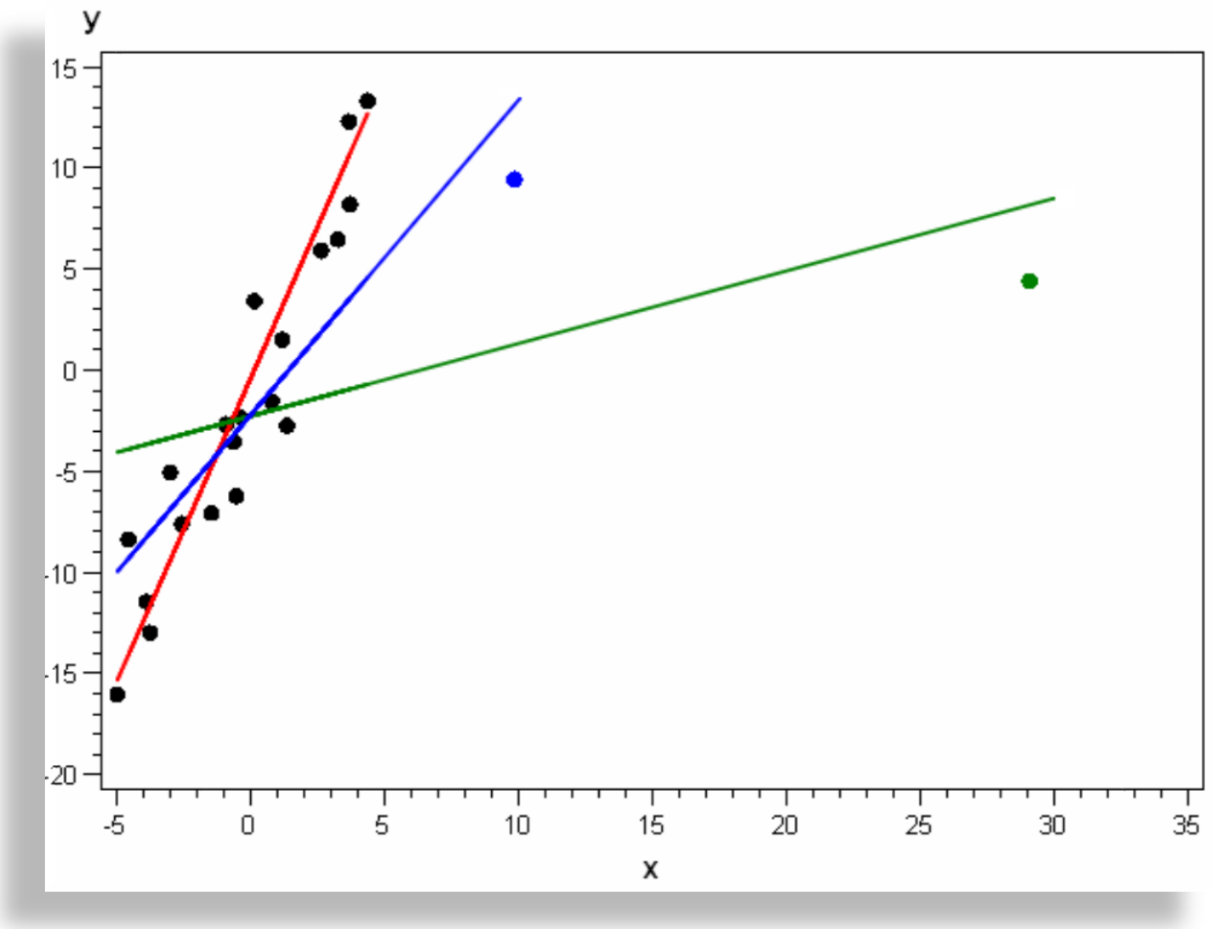
- Kiedy warstwy interesujące nas pod kątem danego zjawiska (np. zbiór niewypłacalnych klientów) są małe w porównaniu z pozostałymi.



- Taka sytuacja zachodzi często przy analizach typu credit scoring lub przy badaniu zjawiska churn. Przepróbkowanie daje wtedy często lepsze wyniki niż zwykłe metody próbkowania.

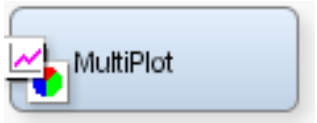
SEMMA

SAMPLE - WARTOŚCI ODSTAJĄCE



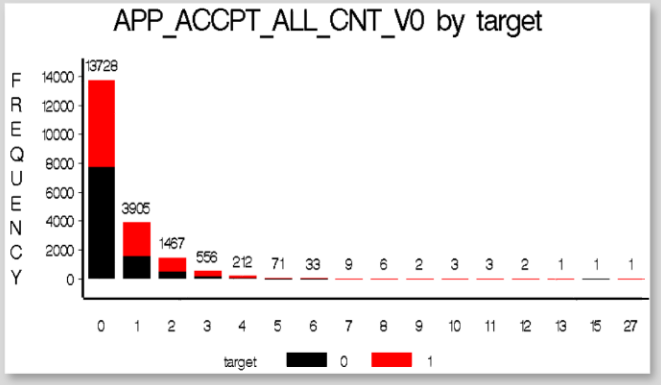
SEMMA

EXPLORE

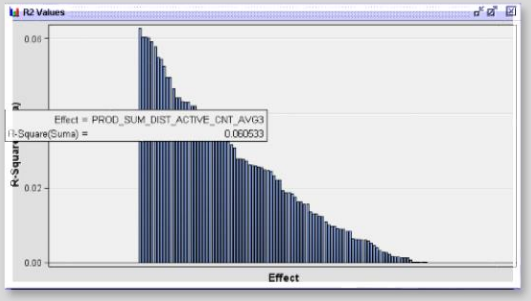


Rozkład zmiennych

Target Level	Variable	Mean	Std. Deviation	Non Missing	Missing ▼	Minimum	Median	Maximum
0	APP_LAST_ACCPT_ALL_DY_V0	58.07618	26.17831	6800	364	0	67	102
1	APP_LAST_ACCPT_ALL_DY_V0	48.43695	27.26472	6836	164	0	67	102
0	CARDCR_ALL_INT_PEN_AMT_V0	1.324134	5.136291	7000	0	0	0	277.24
1	CARDCR_ALL_INT_PEN_AMT_V0	0.045557	0.360619	7000	0	0	0	17.73
0	CARDCR_ALL_INT_DUE_GOT_...	42.90347	180.1206	7000	0	0	0	5097.027
1	CARDCR_ALL_INT_DUE_GOT_...	1.731586	9.50785	7000	0	0	0	349.7633
0	PROD_TRN_ALL_AMT_ZM26	9.24349	495.233	7000	0	0	0	58576.11
1	PROD_TRN_ALL_AMT_ZM26	1.640902	18.4742	7000	0	0	0.400307	1422.769
0	PROD_TRN_ALL_CNT_ZM26	9.24349	495.233	7000	0	0	0	58576.11
1	PROD_TRN_ALL_CNT_ZM26	1.640902	18.4742	7000	0	0	0.400307	1422.769
0	PROD_TRN_OBC_UZN_DF_AM...	107.0252	3869.878	7000	0	-110003	0	92551.17
1	PROD_TRN_OBC_UZN_DF_AM...	348.442	4984.869	7000	0	-98235	0	304182.4
0	PROD_TRN_OBC_AMT_ZM26	4.288364	203.553	7000	0	0	0	24061.47
1	PROD_TRN_OBC_AMT_ZM26	1.458968	12.05099	7000	0	0	0.402031	893.5385
0	PROD_TRN_OBC_AMT_MIN3	388.4454	7023.667	7000	0	0	0	475147.3
1	PROD_TRN_OBC_AMT_MIN3	910.19	7906.646	7000	0	0	0	476042.8
0	PROD_TRN_OBC_AMT_V0	714.5435	9865.953	7000	0	0	0	475147.3
1	PROD_TRN_OBC_AMT_V0	1630.507	11909.09	7000	0	0	68.01	505820.9
0	CARDCR_ALL_USE_LMT_AVG_...	0.013057	0.071832	7000	0	0	0	1
1	CARDCR_ALL_USE_LMT_AVG_...	0.029303	0.085025	7000	0	0	0	1.88078
0	CARDCR_TRN_LUK_ALL_WD_...	2.007286	8.950371	7000	0	0	0	99



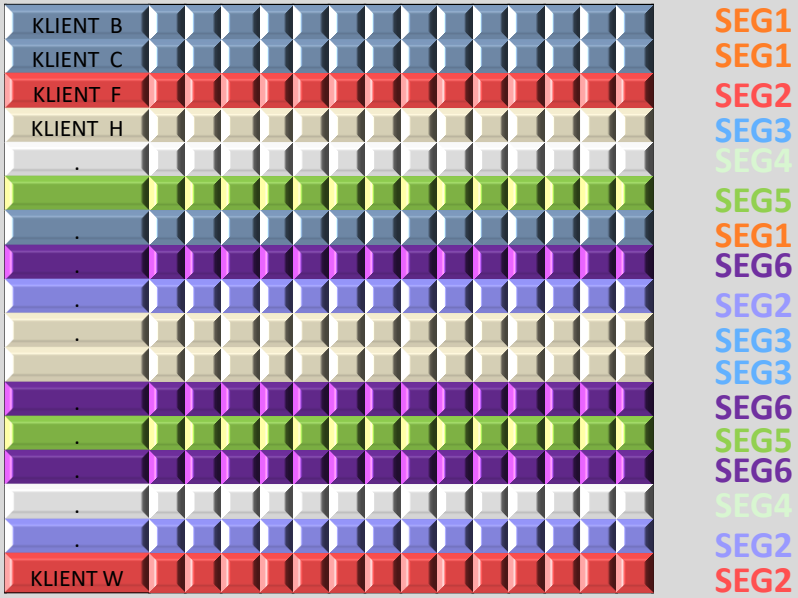
Wybór zmiennych



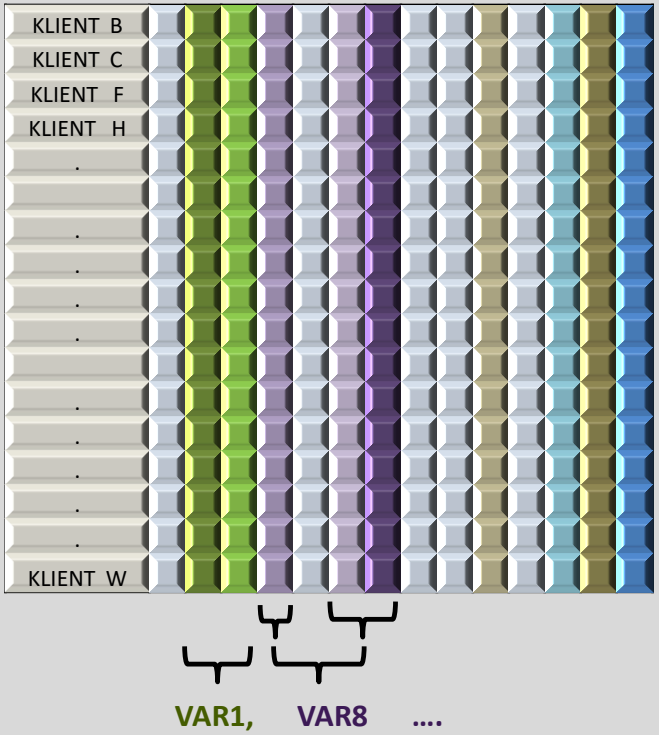
SEMMA

EXPLORE

Segmentacja



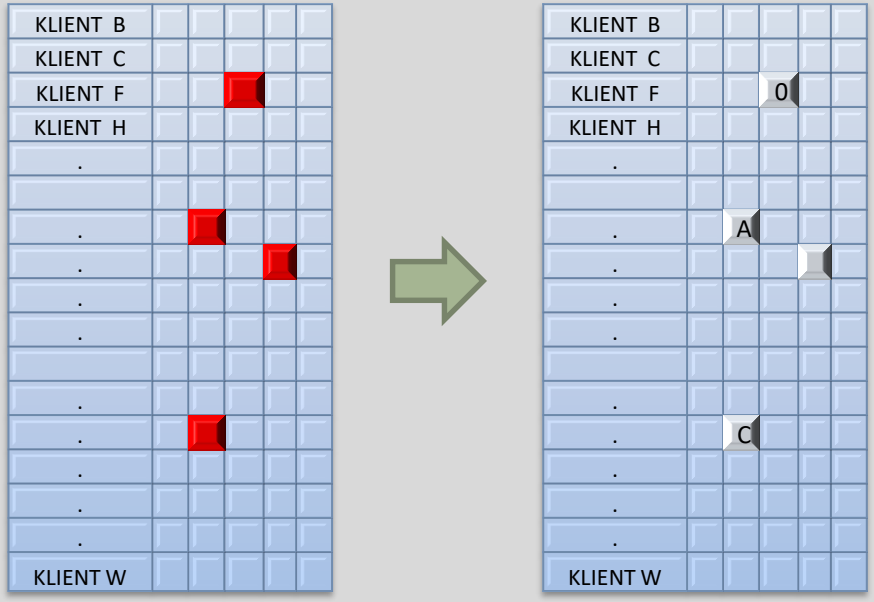
Grupowanie zmiennych



SEMMA

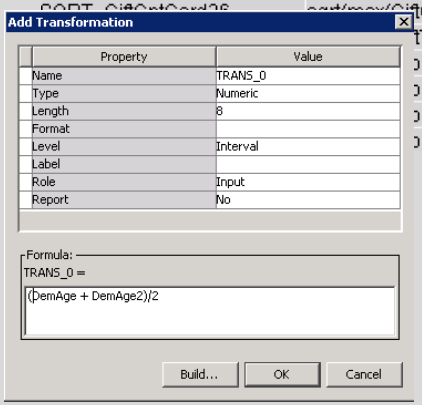
MODIFY

Uzupełnienie wartości brakujących



Przekształcenie zmiennych

Name	Formula ▲
PWR_DemAge	(max(DemAge-0, 0.0)/87)**4
PWR_DemMedIncome	(max(DemMedIncome-2499, 0.0)/197502)**0.25
PWR_GiftAvgCard36	(max(GiftAvgCard36-1.33, 0.0)/258.67)**0.25
SQR_StatusCatStarAll	(max(StatusCatStarAll-0, 0.0))**2
LOG_DemMedHomeValue	log(max(DemMedHomeValue-0, 0.0)/600000 + 1)
LOG_GiftAvg36	log(max(GiftAvg36-0, 0.0)/260 + 1)
LOG_GiftAvgAll	log(max(GiftAvgAll-1.5, 0.0)/448.5 + 1)
LOG_GiftAvgLast	log(max(GiftAvgLast-0, 0.0)/450 + 1)
LOG_GiftCnt36	log(max(GiftCnt36-0, 0.0)/16 + 1)
LOG_GiftCntCardAll	log(max(GiftCntCardAll-0, 0.0)/41 + 1)
LOG_PromCnt12	log(max(PromCnt12-2, 0.0)/57 + 1)
SQRT_GiftCntAll	sqrt(max(GiftCntAll-1, 0.0)/90)



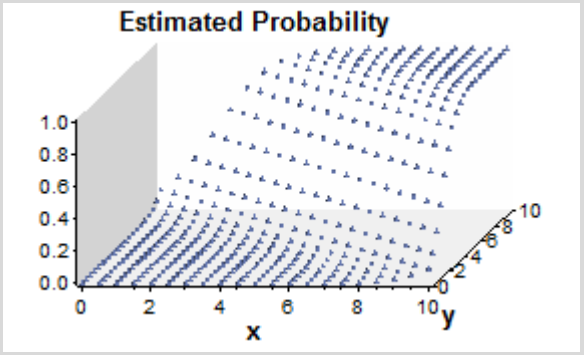
SEMMA

MODEL

Regression

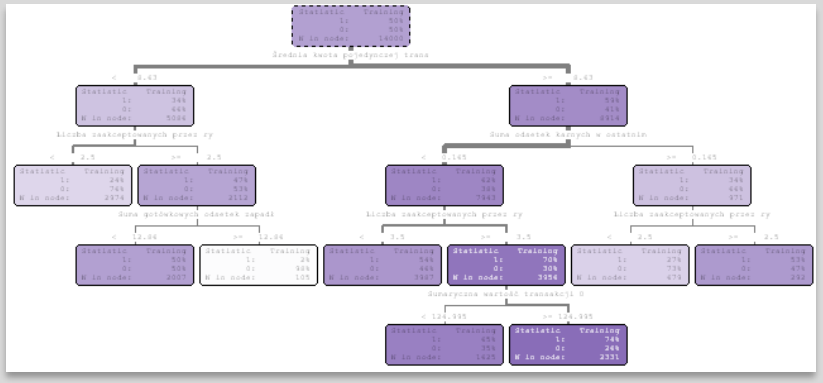
Decision Tree

Analiza regresji




```
Logit(SCORE) =
      a0
      + a1 * M_Var3
      + a2 * M_Var7
      + a3 * M_Var10;
```

Drzewa decyzyjne

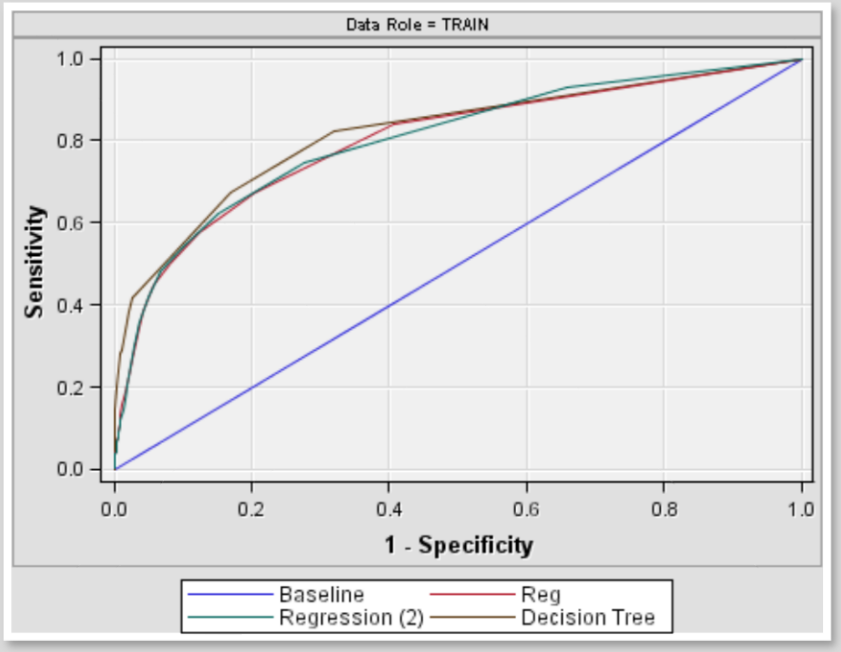


```
if 0 <= Var1 < 0.5 and Var2 > 10
and VAR8 = . then
   SCORE = 0.8;
else if Var1 < 0 and Var6 > 3 then
   SCORE = 0.1;
else ...
```

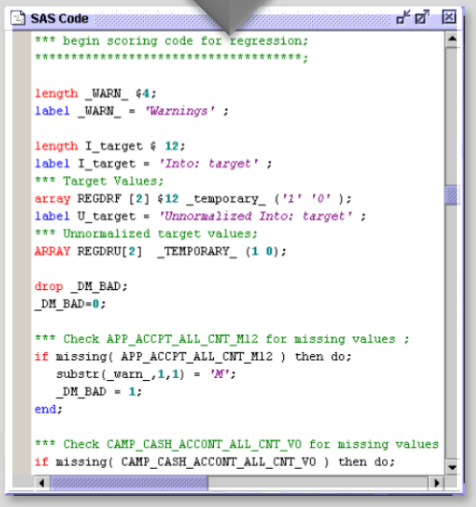
SEMMA

ASSESS

Porównanie modeli



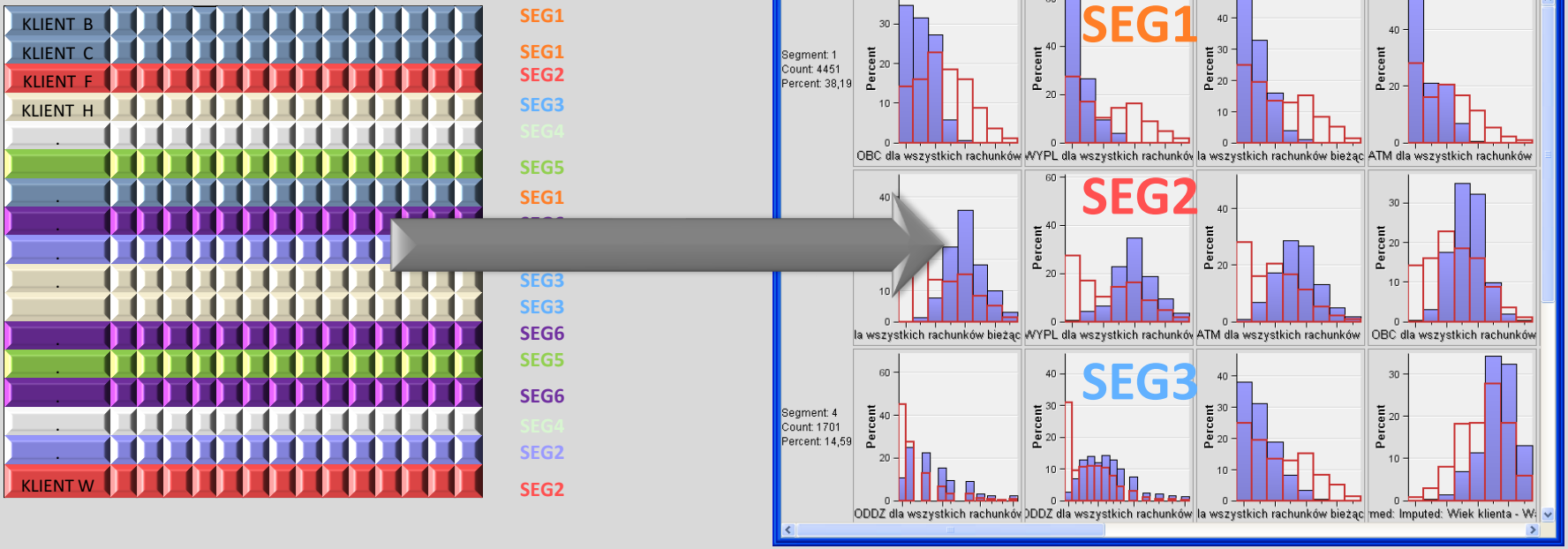
Generacja kodu scoringowego



SEMMA

ASSESS

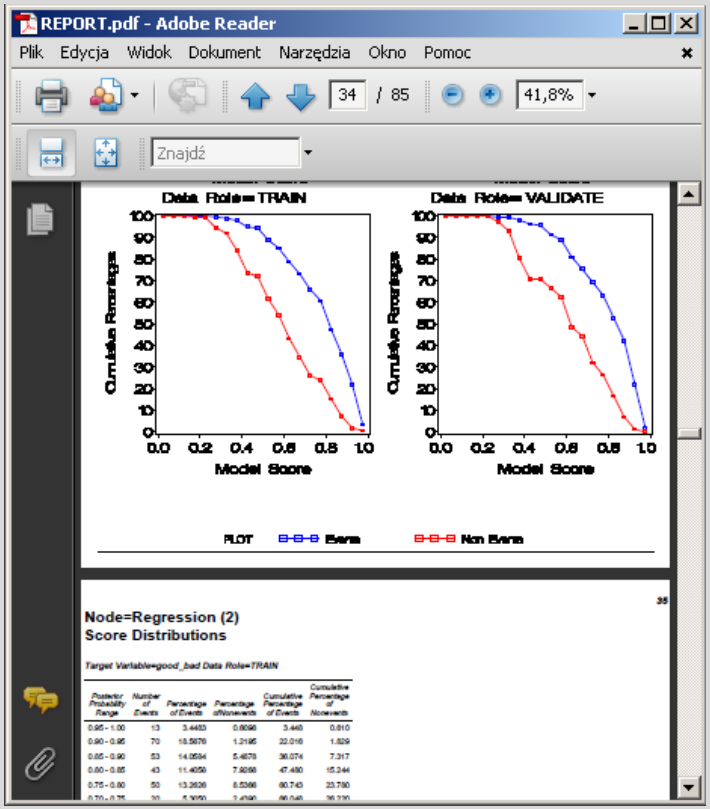
Profilowanie segmentów



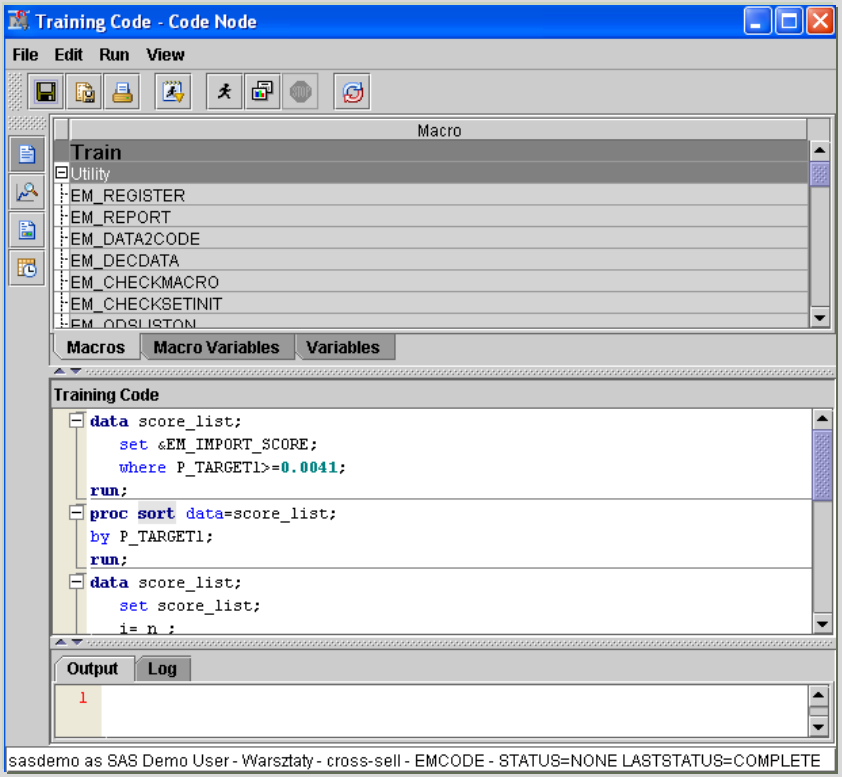
SEMMA

WĘZŁY DODATKOWE

Dokumentacja ścieżki przetwarzania



Własny kod użytkownika



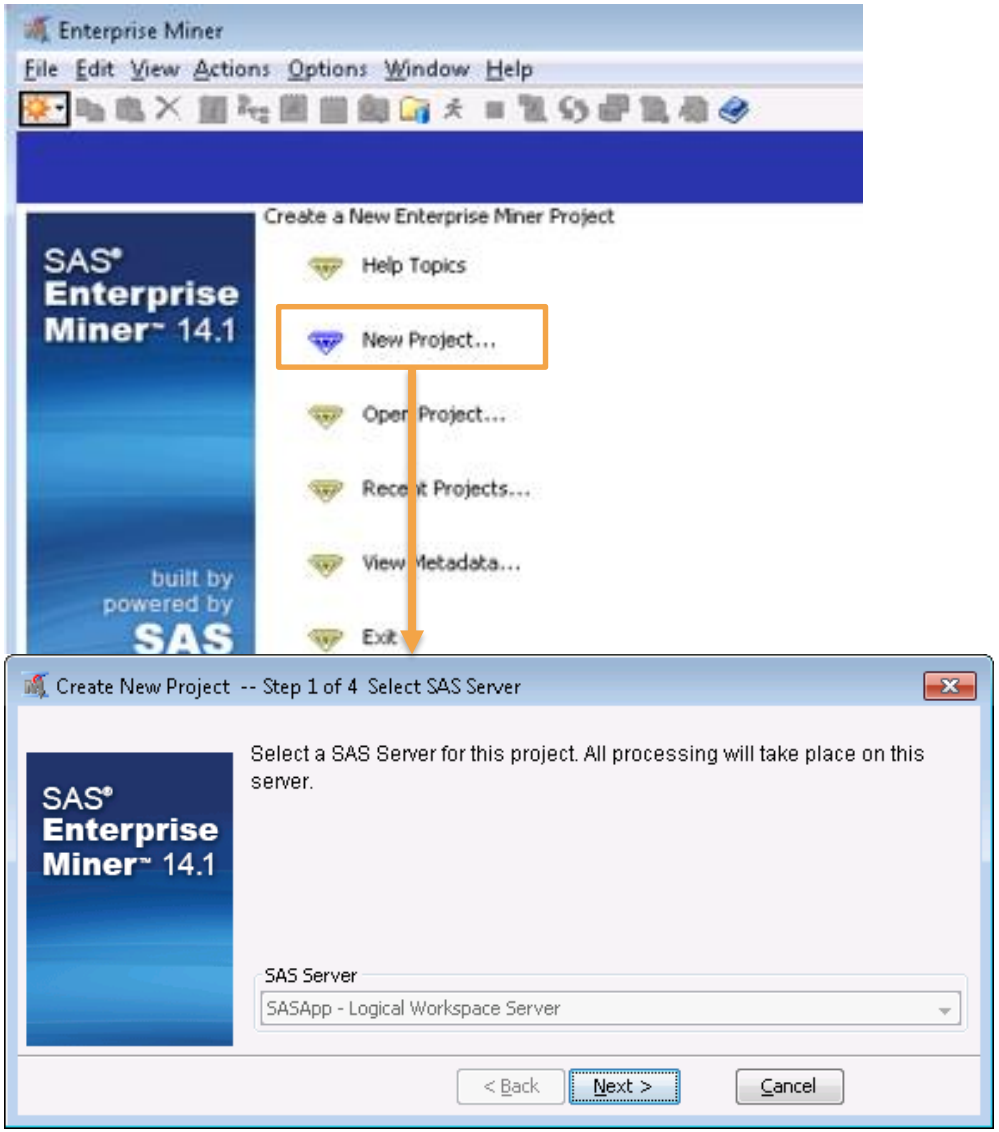
Tworzenie nowego projektu EM

Praca w SAS Enterprise Miner jest zorganizowana w postaci **Projektów.**

Użytkownik po zalogowaniu ma dostęp do istniejących projektów oraz ma możliwość tworzenia nowych projektów.

W przypadku instalacji serwerowej użytkownik może mieś dostęp do projektów innych użytkowników

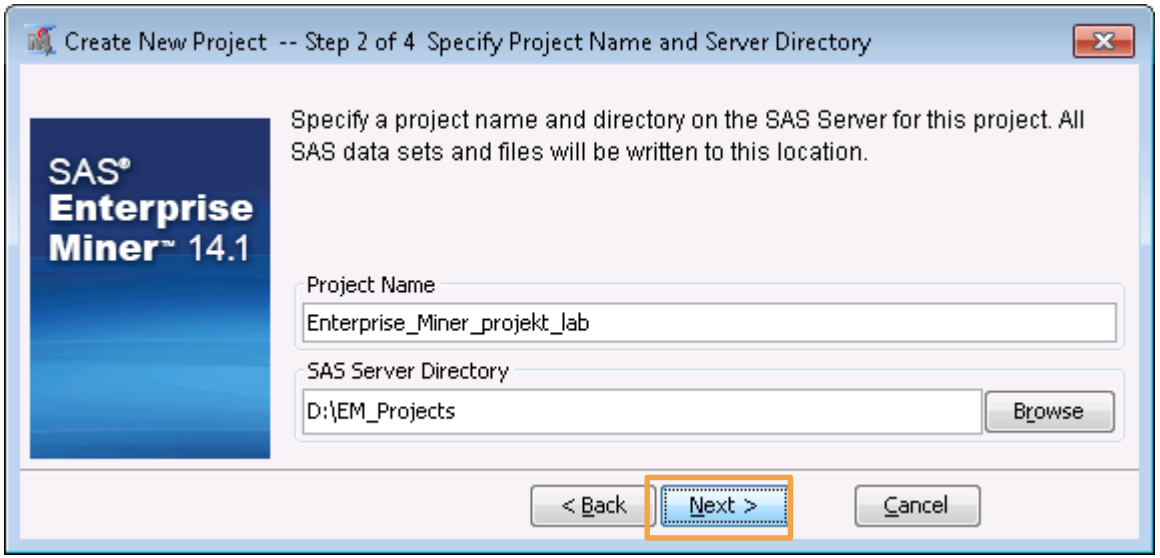
Rozpoczęcie procesu tworzenia projektu



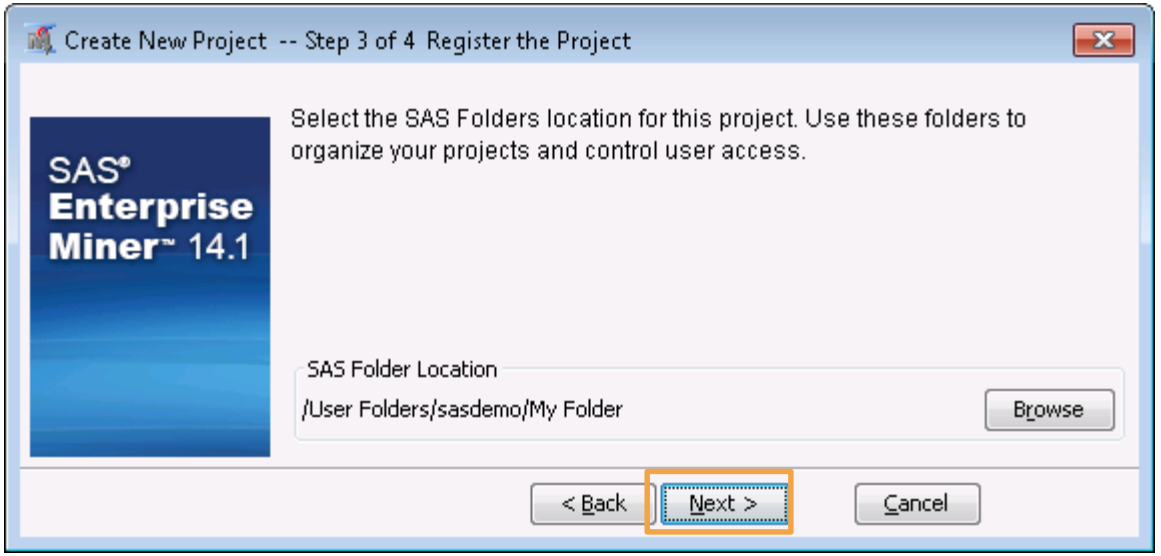
Tworzenie nowego projektu EM

Wskazanie fizycznej lokalizacji i nazwy projektu

Jeżeli użytkownik poda ścieżkę i nazwę już istniejącego projektu – pojawi się stosowny komunikat i zależnie od decyzji użytkownika może on zostać nadpisany.

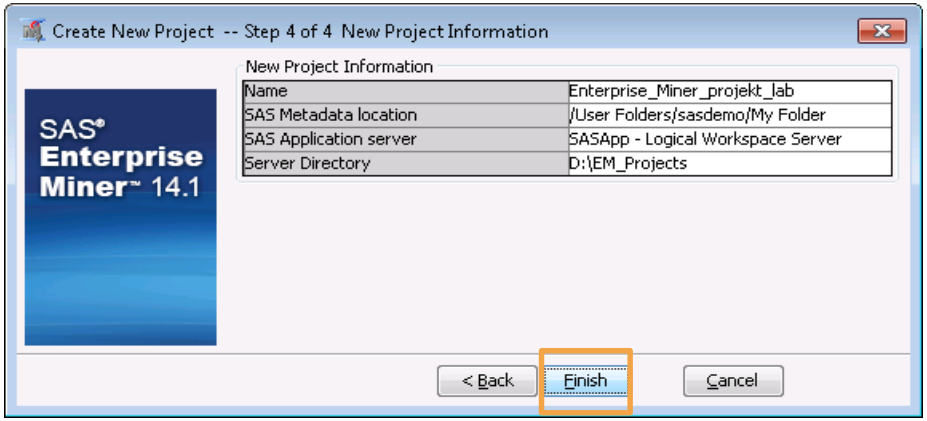


Wskazanie lokalizacji metadanowej projektu

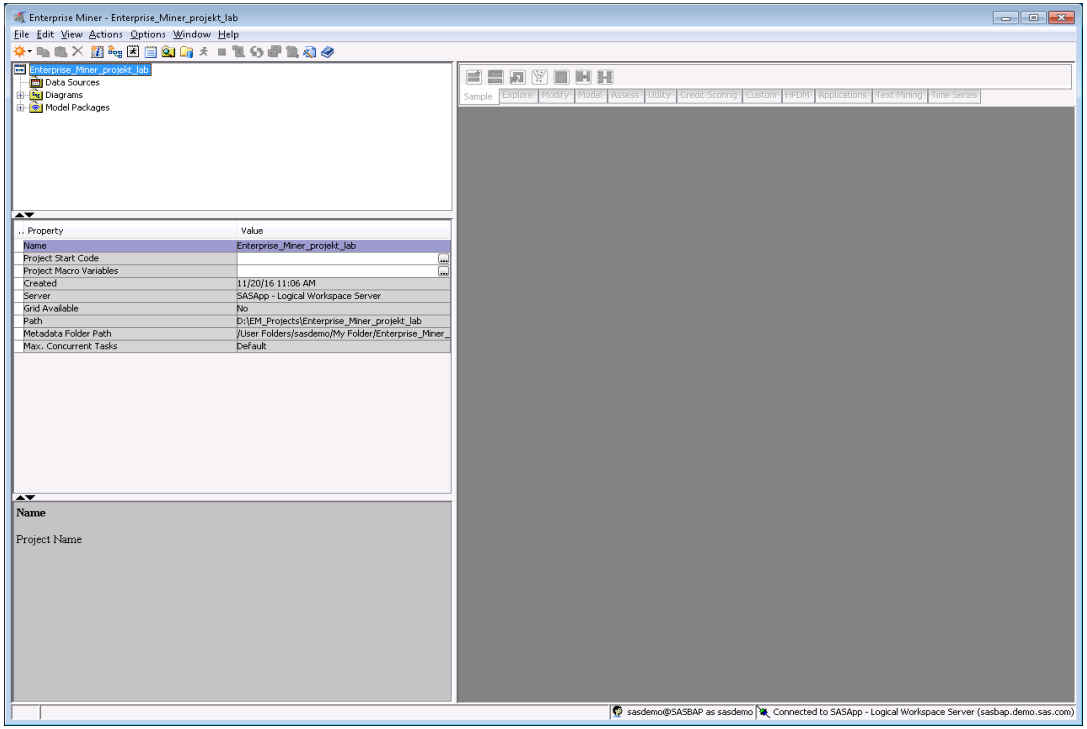


Tworzenie nowego projektu EM

Podsumowanie procesu tworzenia projektu

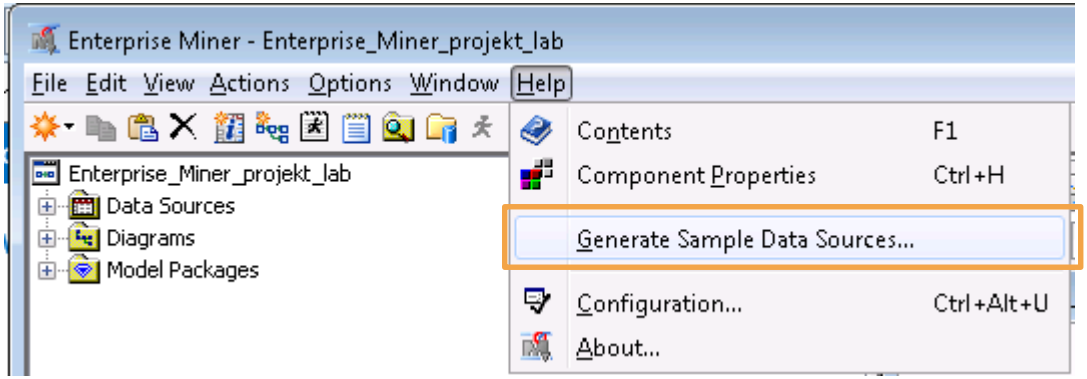


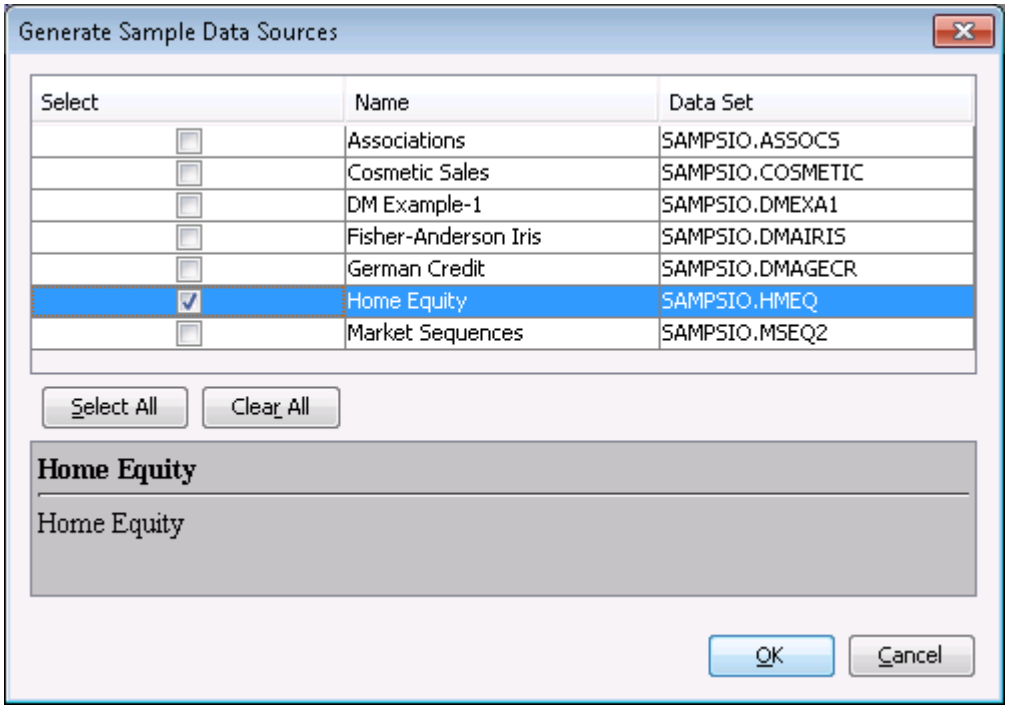
Nowy projekt



Dodanie źródła danych

W SAS Enterprise miner dostępne są gotowe przykładowe źródła danych

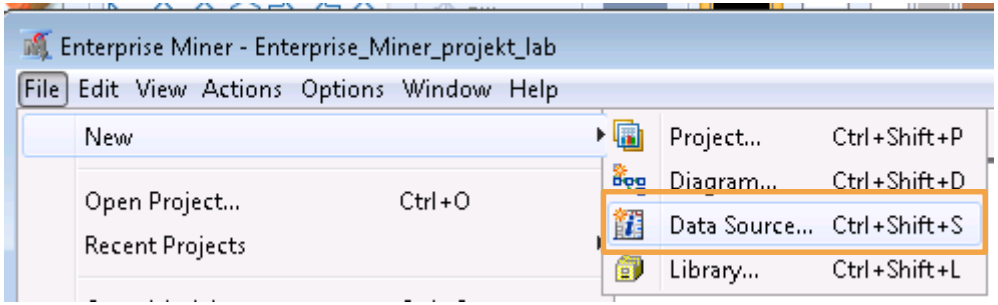


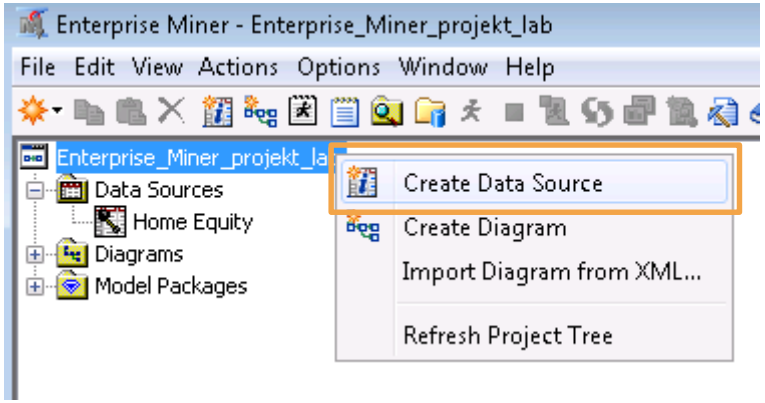


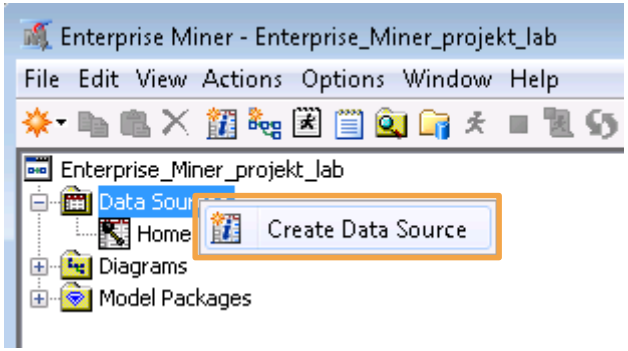
Dodanie źródła danych

Standardowo użytkownik definiuje źródło danych na podstawie własnego zbioru.

Jest to możliwe na kilka sposobów

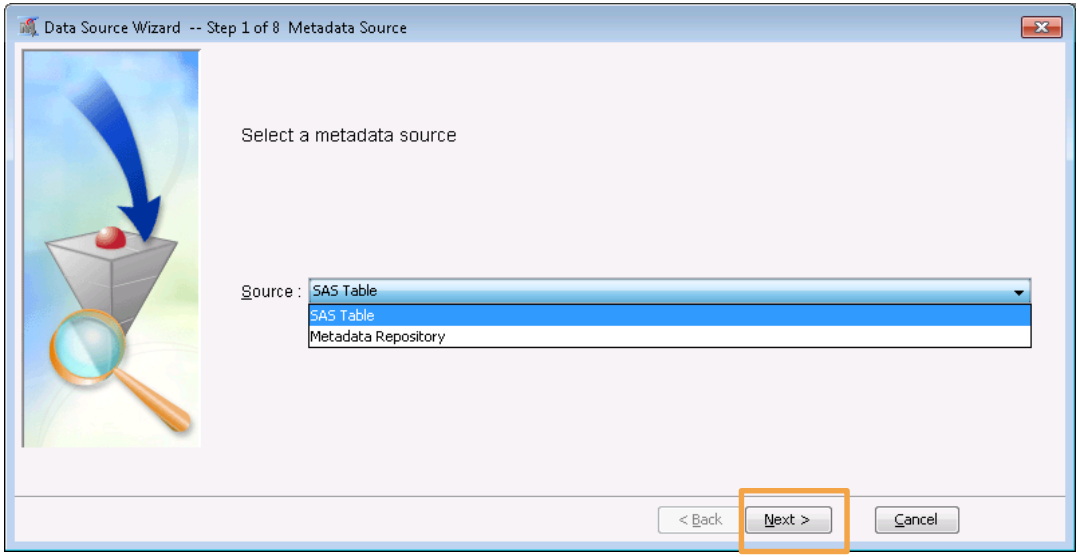






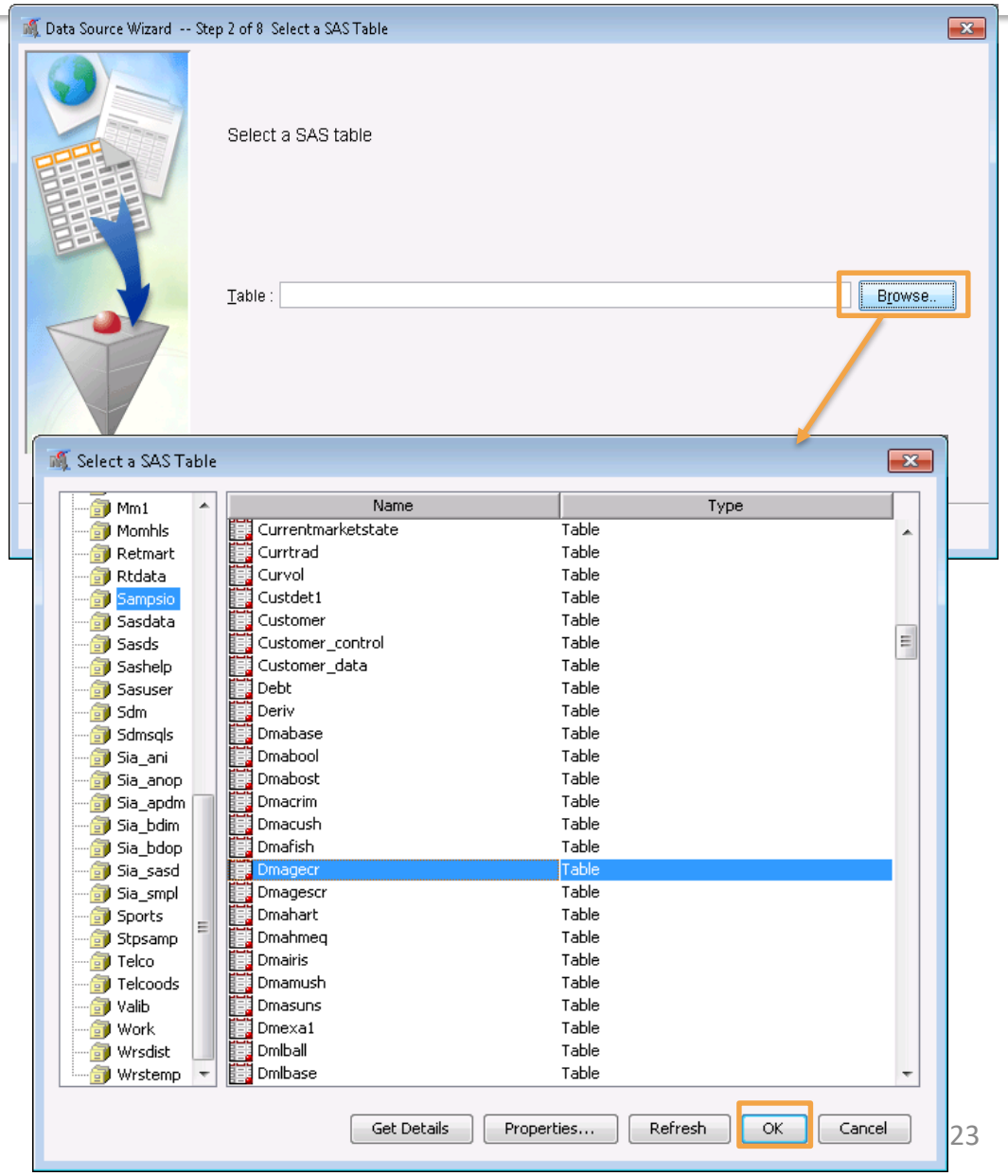
Dodanie źródła danych

Użytkownik określa, czy źródło danych jest już zarejestrowane w metadanych, czy jest dostępne w jednej z bibliotek



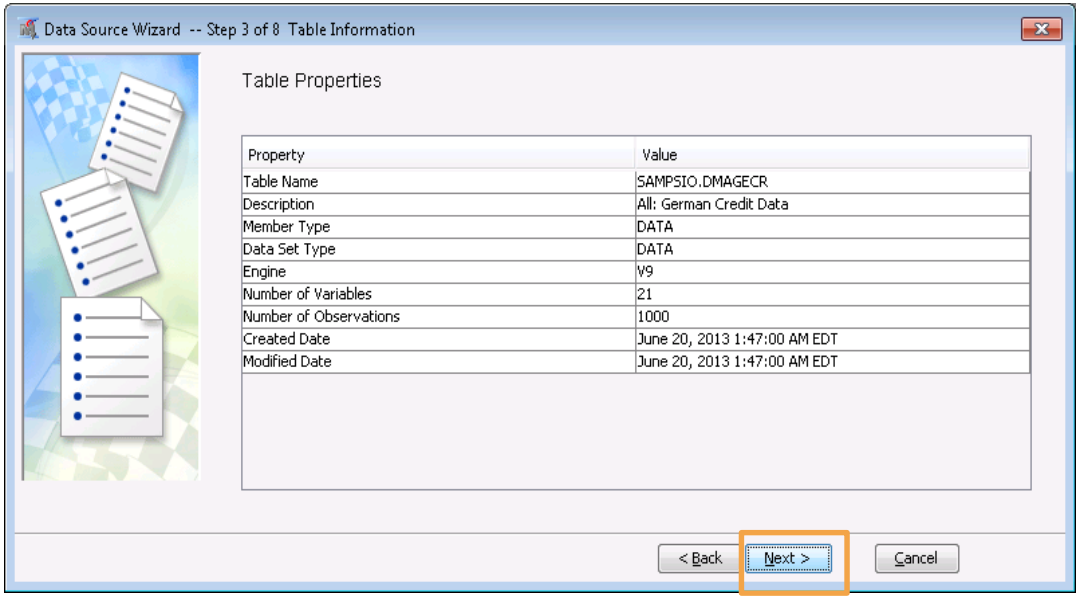
Dodanie źródła danych

Określenie szczegółowej lokalizacji zbioru danych, tj. biblioteki i nazwy tabeli



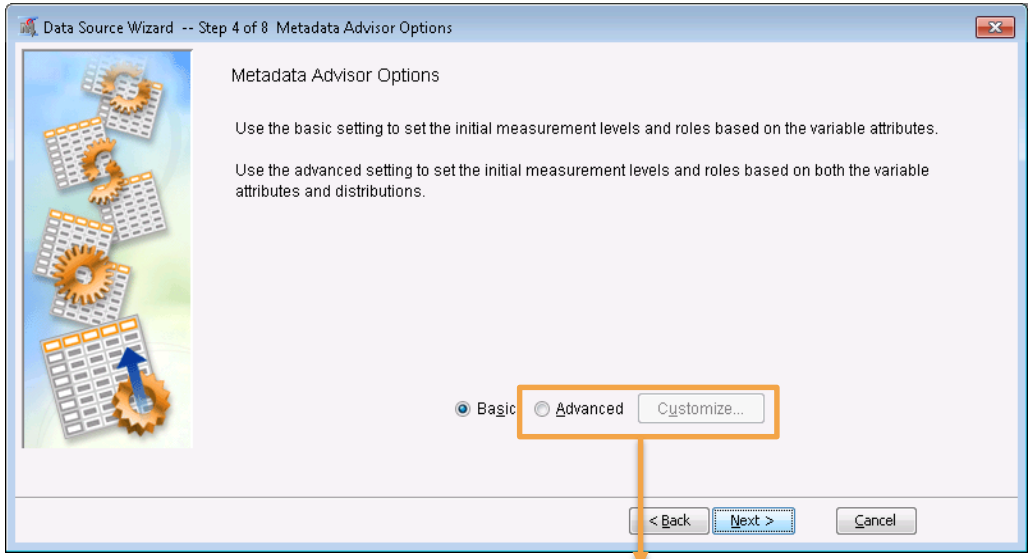
Dodanie źródła danych

Podsumowanie podstawowych metadanych źródła

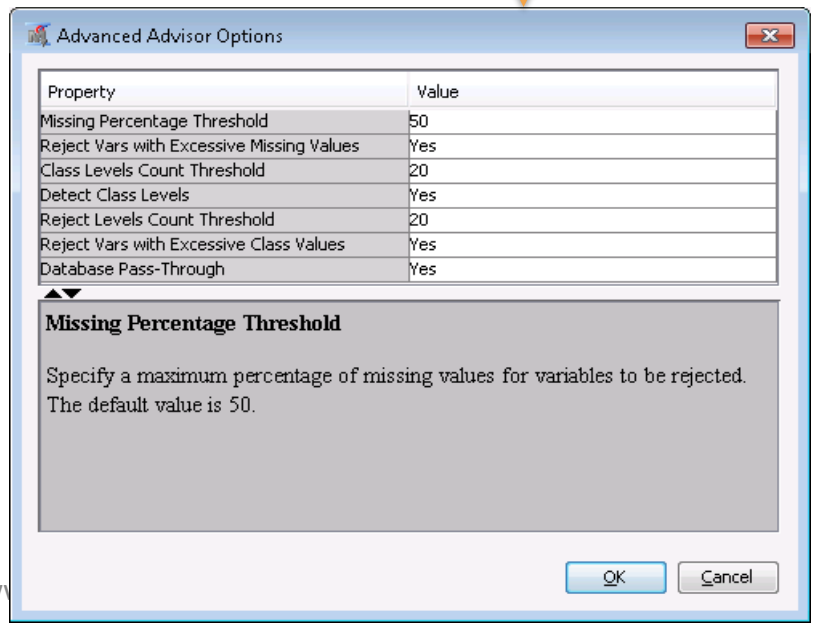


Dodanie źródła danych

Inicjalizacja kreatora poziomów i ról zmiennych



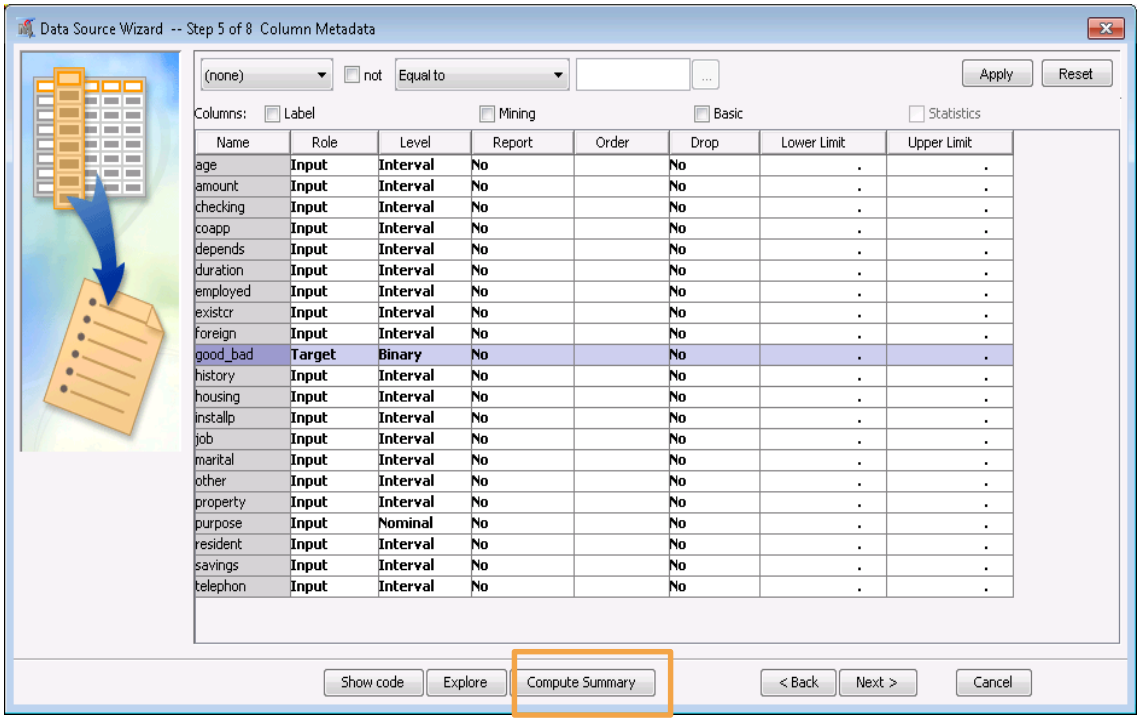
W przypadku kreatora zaawansowanego, role i poziomy zmiennych inicjalizowane są na podstawie analizy wartości zmiennych



Dodanie źródła danych

Dla każdej zmiennej można określić jej poziom i rolę

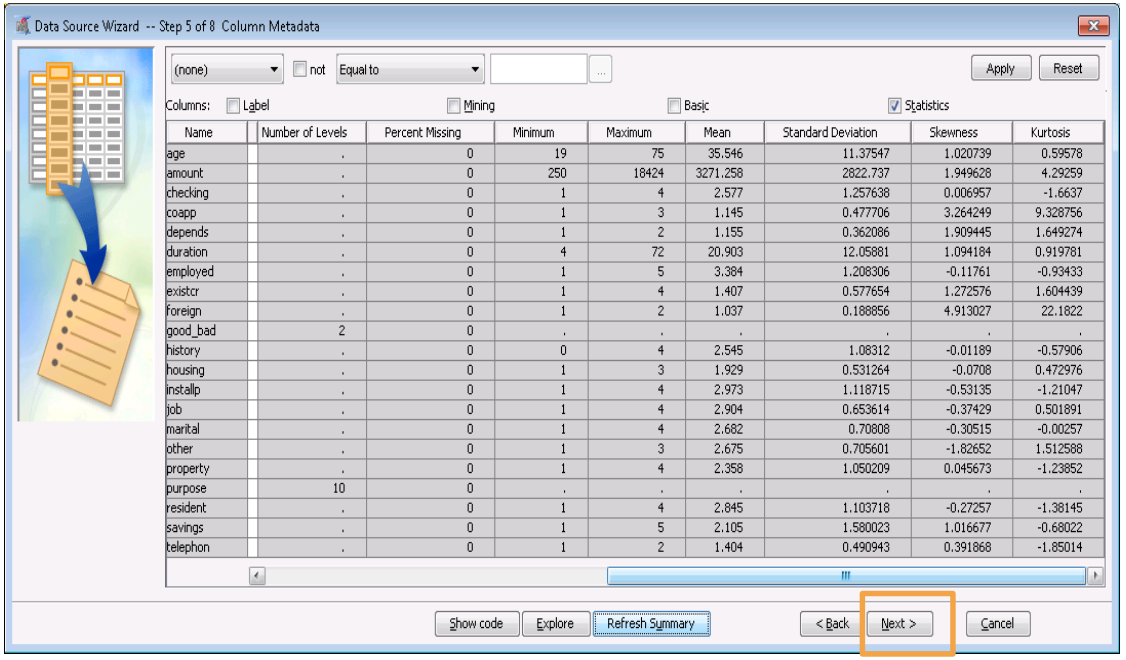
Zadanie może zostać zrealizowane również za pomocą kodu użytkownika



Dodanie źródła danych

Dla każdej zmiennej można wygenerować statystyki, które ułatwią określenie roli i poziomu zmiennych.

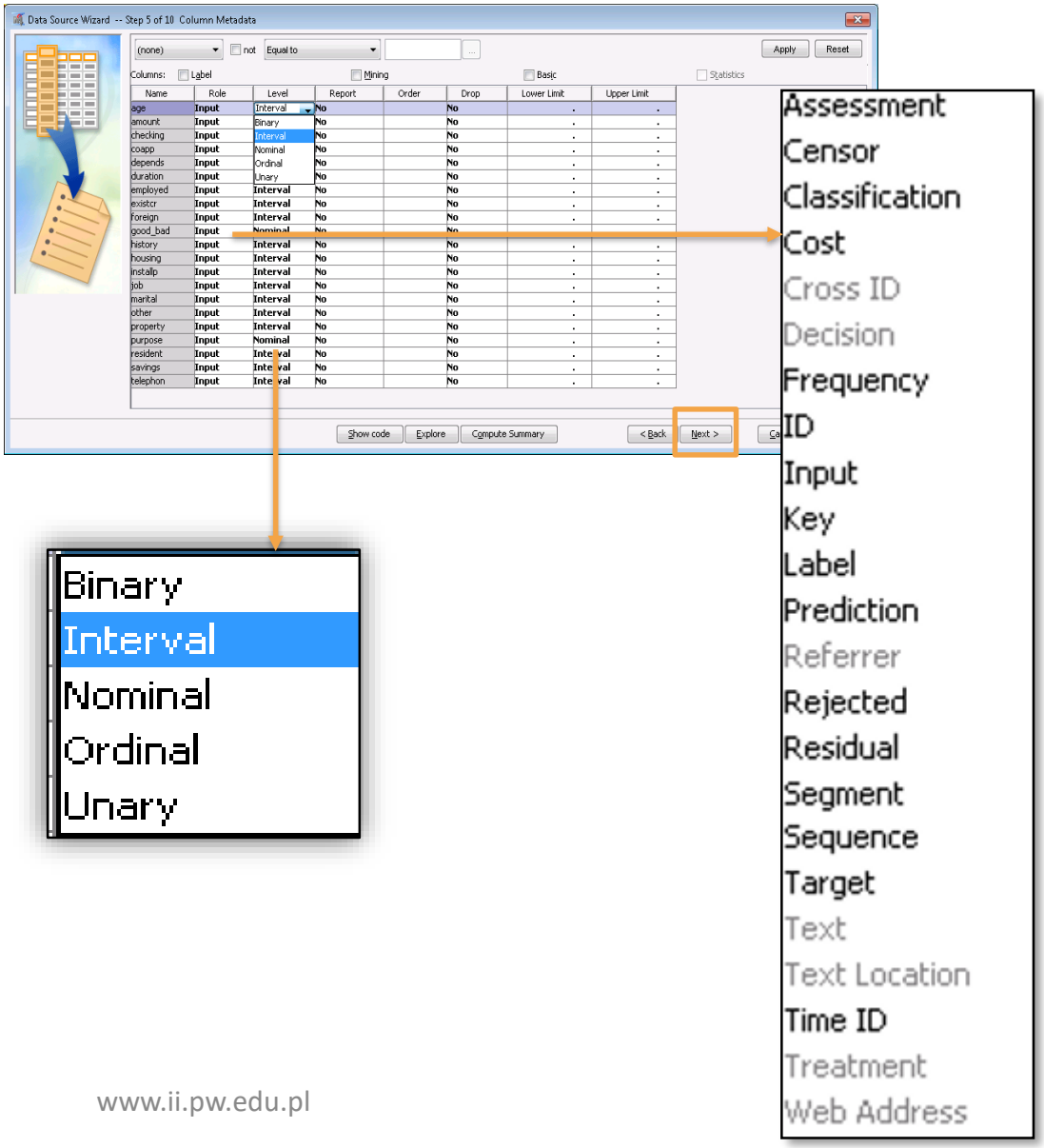
Przy dużych zbiorach danych zadanie to może być czasochłonne



Dodanie źródła danych

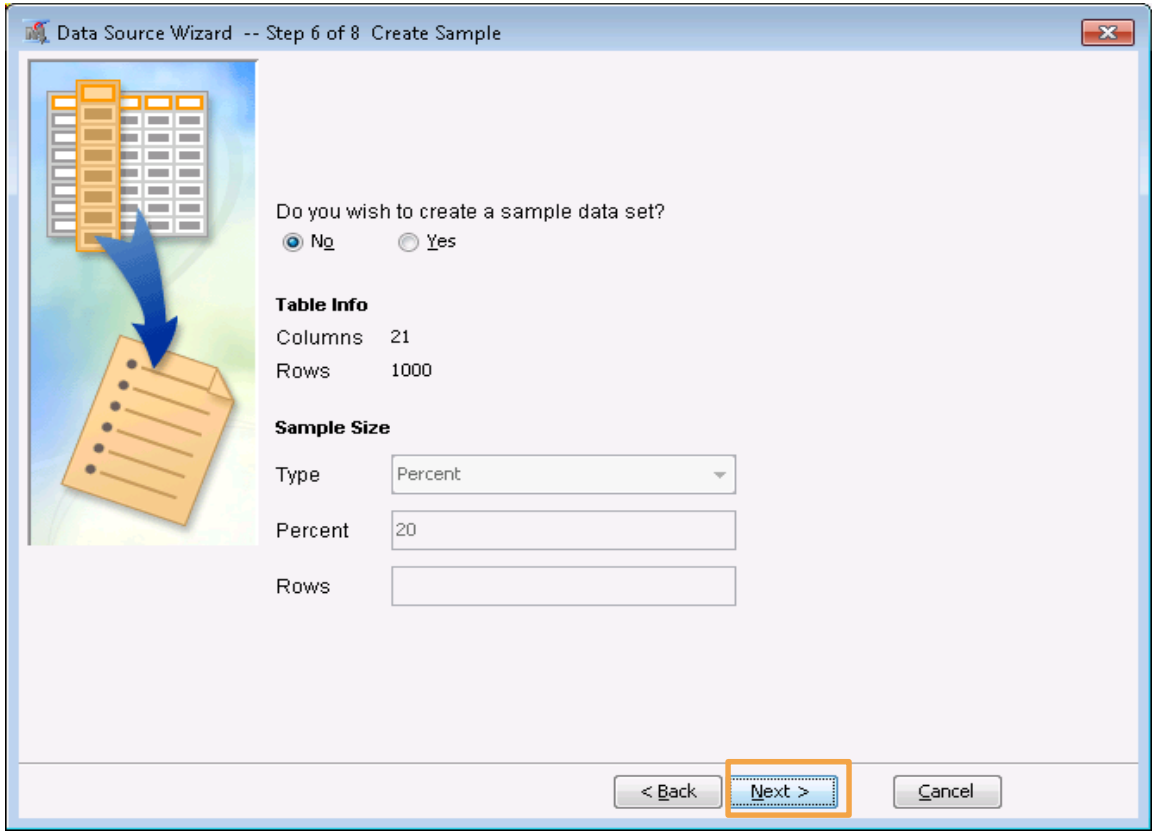
Najpopularniejsze role zmiennych:

- Input – zmienna wejściowa
- ID - identyfikator
- Target – zmienna celu



Dodanie źródła danych

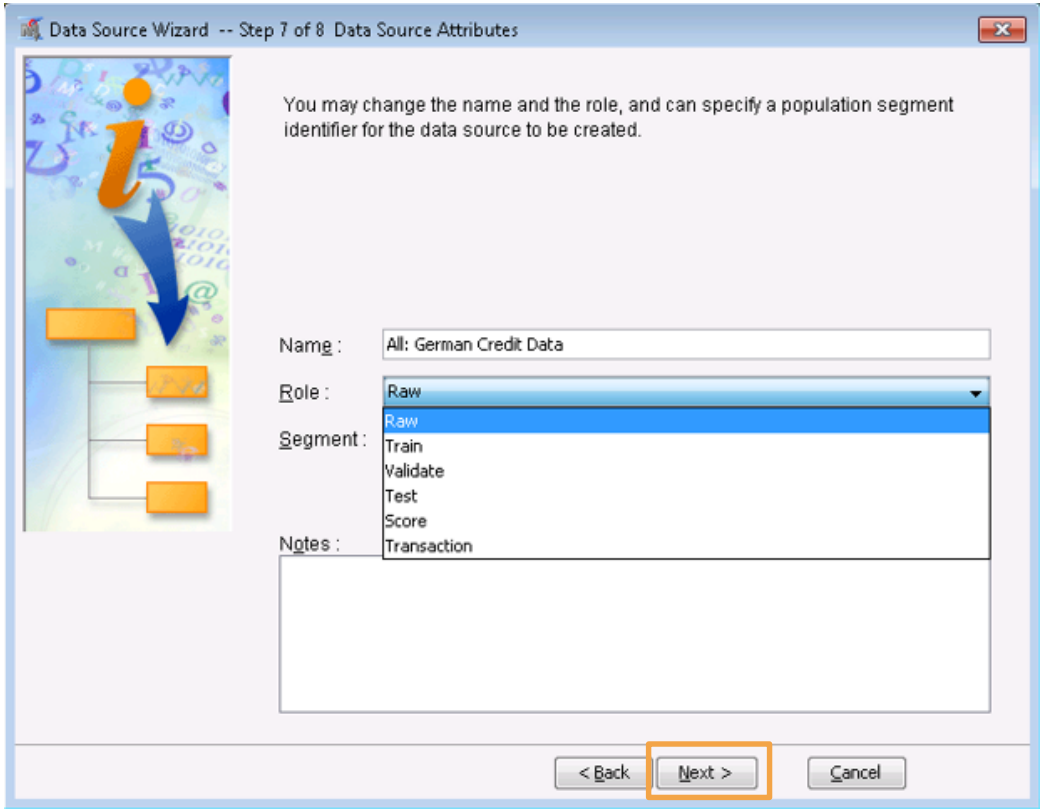
Podczas tworzenia zbioru użytkownik może wykonać próbkowanie (losowanie proste bez zwracania)



Dodanie źródła danych

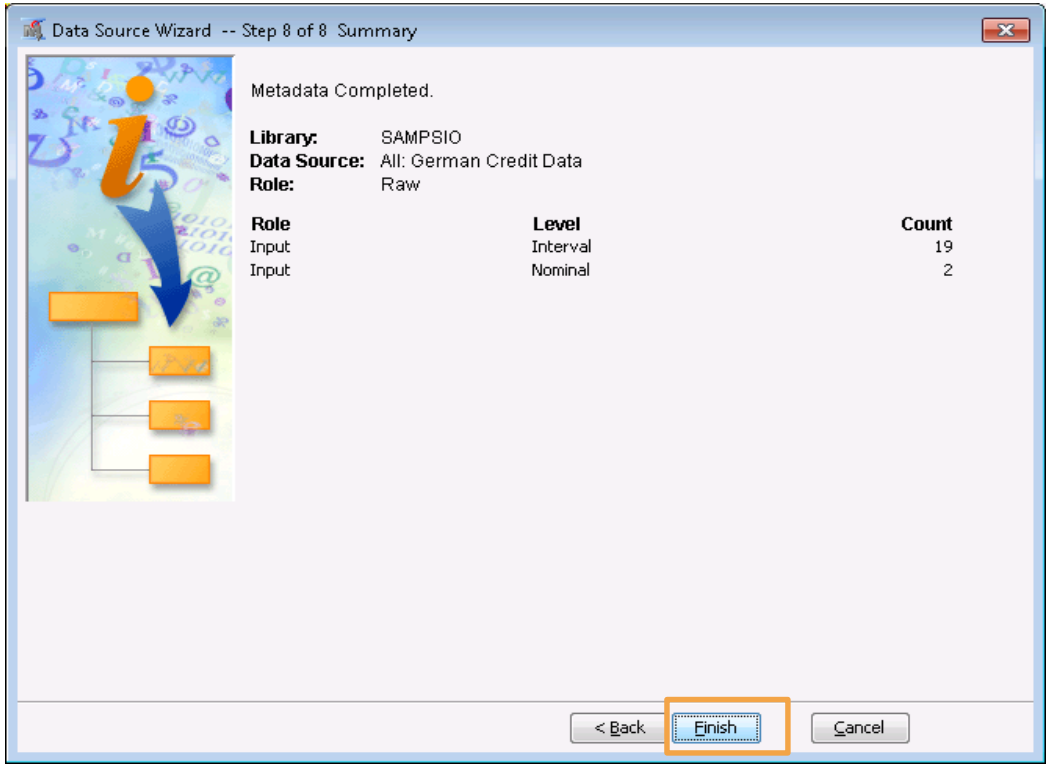
Określenie roli zbioru danych:

- Raw – najczęstsza rola
- Train – zbiór treningowy
- Validate – zbiór walidacyjny
- Test – zbiór testowy
- Score – zbiór do skorowania
- Transaction – zbiór o strukturze transakcyjnej (wiele wierszy opisujących obiekt modelowania, np. klienta)



Dodanie źródła danych

Podsumowanie procesu tworzenia źródła danych

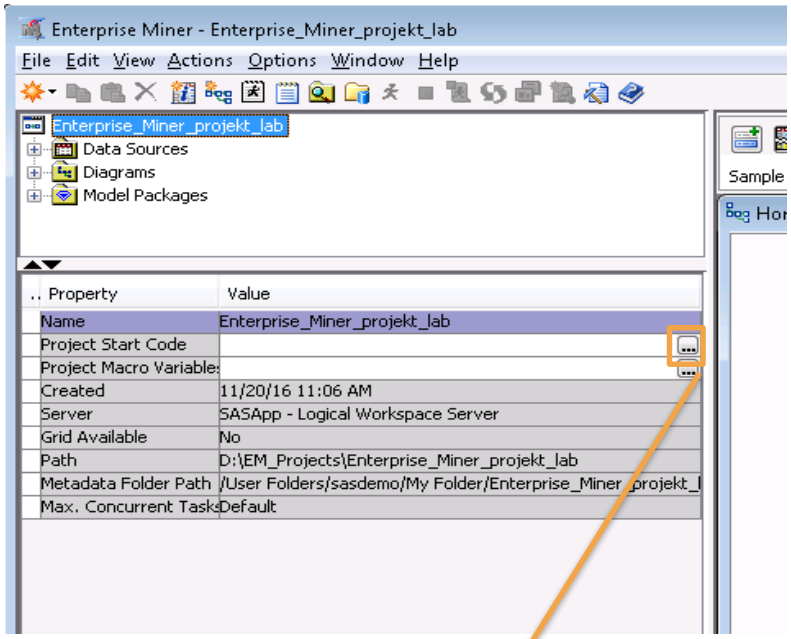


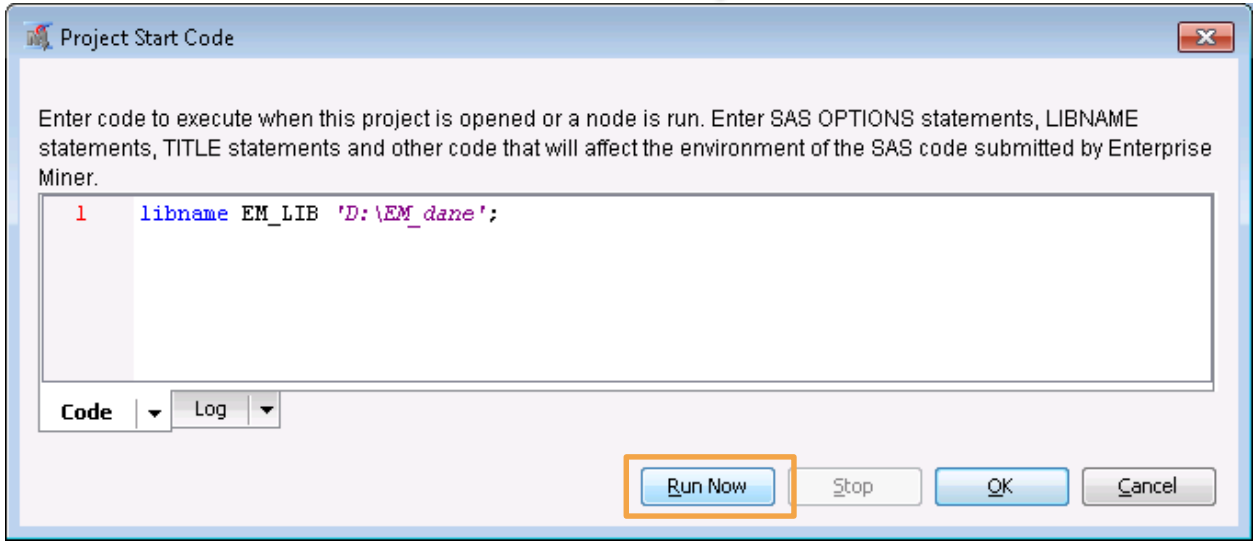
Dodanie źródła danych

Jeżeli źródło danych nie znajduje się w bibliotece dostępnej w środowisku, użytkownik powinien zdefiniować bibliotekę w kodzie startowym projektu.

Po zdefiniowaniu kodu, należy go uruchomić.

Od tego momentu nowa biblioteka będzie widoczna dla SAS Enterprise Miner.

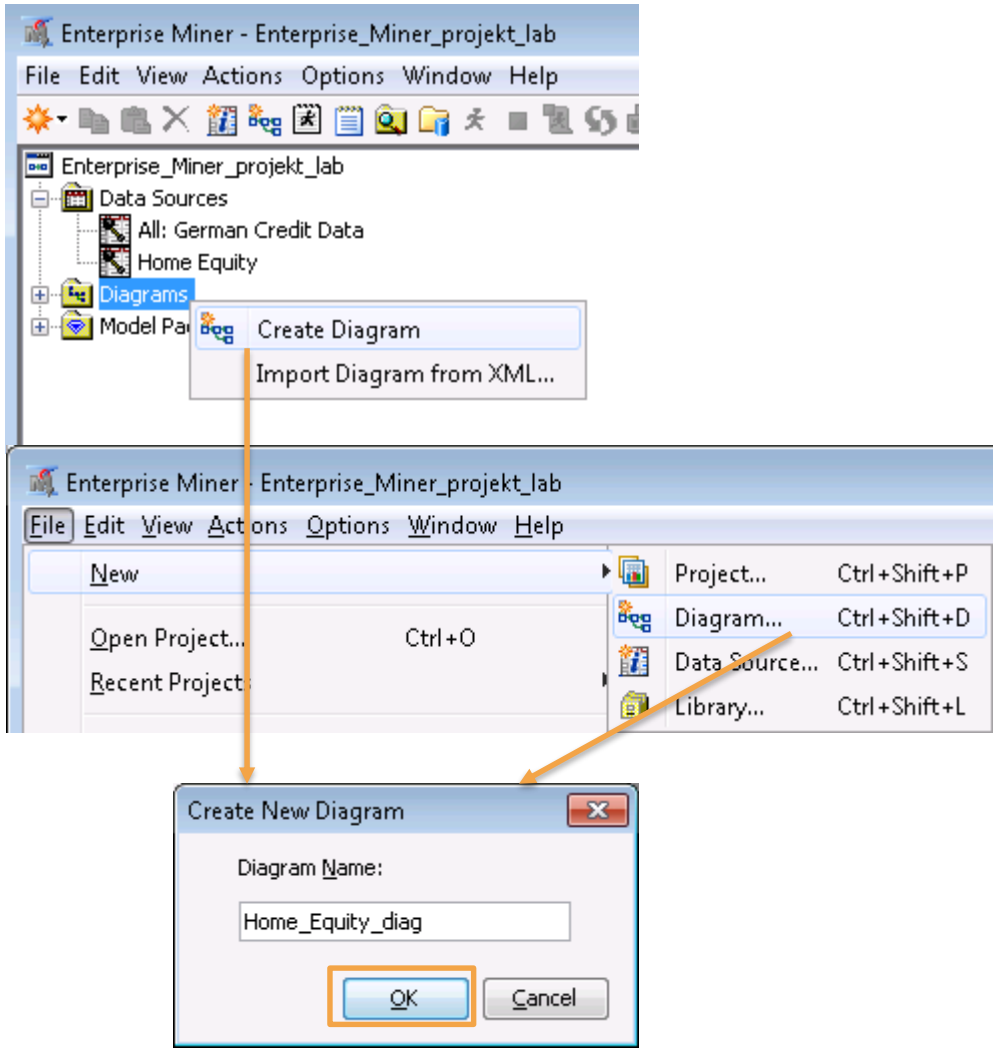




Tworzenie diagramu

W ramach projektu użytkownik tworzy diagramy w których konstruuje logikę modelowania.

Istnieje kilka możliwości utworzenia diagramu.

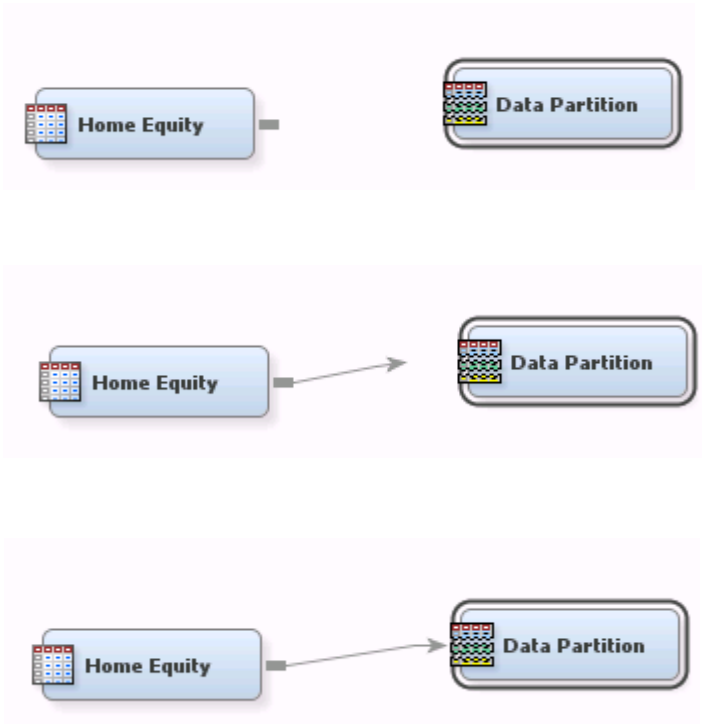


Praca z diagramem

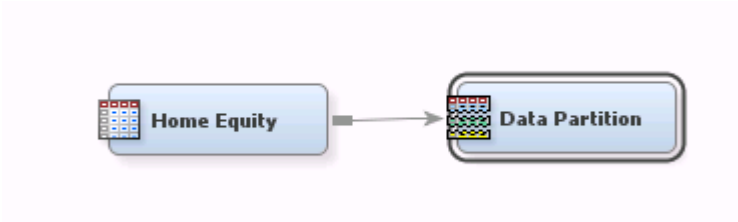
Na diagram przerzuca się poszczególne elementy:

- Źródła danych dodane wcześniej do projektu
- Węzły dostępne w zakładkach SAS EM

Połączenia między węzłami wykonuje się poprzez najechanie na końcówkę jednego węzła (wówczas pojawia się ikona ołówka) i przeciągnięcie powstającej strzałki na początek kolejnego węzła



Praca z diagramem

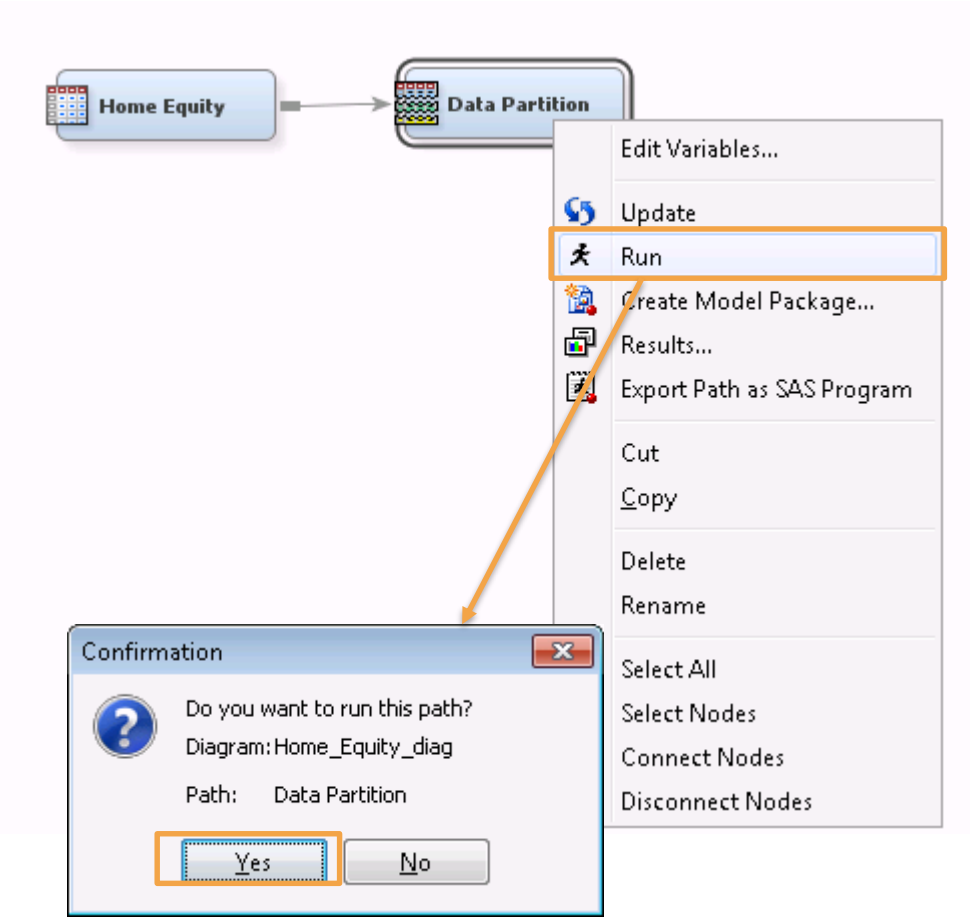


Każdy węzeł ma dedykowane opcje parametryzujące go, które użytkownik może zmieniać.

General	
Node ID	Part
Imported Data	
Exported Data	
Notes	
Train	
Variables	
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0
Report	
Interval Targets	Yes
Class Targets	Yes

Uruchomienie diagramu lub jego fragmentu

Po kliknięciu prawym przyciskiem myszy na na wybrany węzeł pojawia się menu, które umożliwia uruchomienie diagramu do zaznaczonego węzła włącznie.



Ćwiczenie 1

Opis zbioru danych - HMEQ

Zbiór zawiera informacje o osobach wnioskujących o kredyty hipoteczne

Nazwa zmiennej	Etykieta	Opis
BAD	Loan Default Status	Status Default Kredytu (1- klient nie spłacił kredytu w terminie, 0 – klient spłacił kredyt w terminie)
LOAN	Amount of this Loan	Wartość kredytu z wniosku
MORTDUE	Amount Due on First Mortgage	Wartość pozostała do spłaty z tytułu kredytu hipotecznego
VALUE	Property Value	Wartość nieruchomości
REASON	Reason for this Loan	Powód kredytu (DebtCon – konsolidacja, HomeImp – remont)
JOB	Job Category	Typ zatrudnienia
YOJ	Years at Current Job	Liczba lat zatrudnienia u aktualnego pracodawcy
DEROG	Number of Derogatory Reports	Liczba odnotowanych naruszeń prawa
DELINQ	Number of Delinquent Trade Lines	Liczba linii kredytowych, których nie spłacono
CLAGE	Age of Oldest Trade Line (months)	Liczba miesięcy od otwarcia pierwszej linii kredytowej
NINQ	Number of Recent Credit Inquiries	Liczba zapytań kredytowych w ostatnim okresie czasu
CLNO	Number of Trade Lines	Liczba linii kredytowych
DEBTINC	Debt to Income Ratio	Stosunek długu do przychodu

Ćwiczenie 1

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Home Equity
    - Zmienna celu: BAD
- Diagram
  - Home Equity_diag1
- Model
  - Drzewo decyzyjne

a) Przeanalizować modelowany poziom (BAD = 1) dla zmiennej celu
b) Wykonać partycjonowanie i zbudować model drzewa decyzyjnego z domyślnymi parametrami
c) Dokonać modyfikacji parametrów drzewa
d) Zbudować drzewo CHAID

Ćwiczenie 1a

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Home Equity
    - Zmienna celu: BAD
- Diagram
  - Home Equity_diag1
- Model
  - Drzewo decyzyjne

- Utworzyć projekt i diagram
- Dodać zbiór z listy przykładowych zbiorów EM do projektu
- Określić poziom zmiennej celu
- Zweryfikować poziom modelowanego zjawiska w zbiorze źródłowym
  - Wykres słupkowy
- Dla zmiennej **Job**, zweryfikować poziomy zmiennej celu
  - Wykres kołowy nakładany

Ćwiczenie 1a - odpowiedzi

Ustawienie roli i poziomu dla zmiennej celu BAD

Variables - Ids

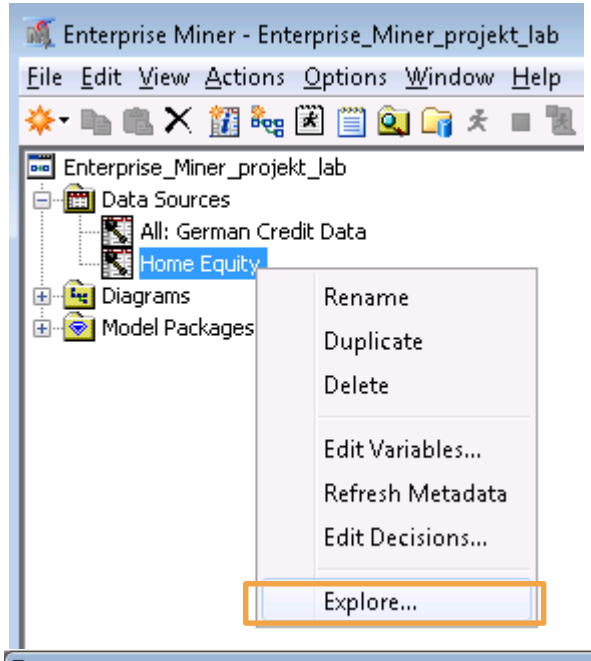
(none) not Equal to

Columns: Label

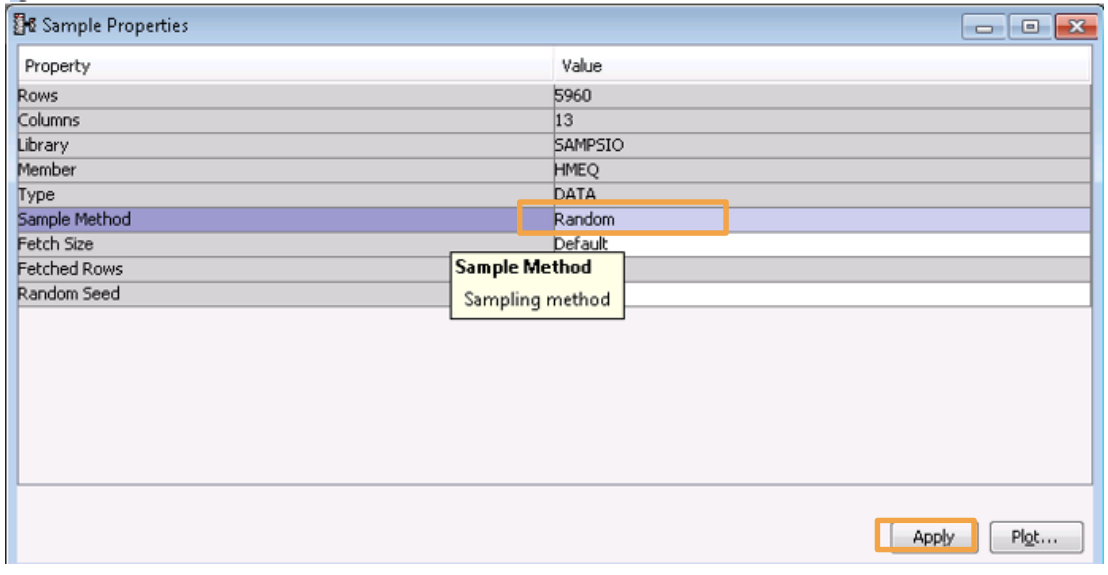
Name	Role	Level	Report	Or
BAD	Target	Binary	No	
CLAGE	Input	Interval	No	
CLNO	Input	Interval	No	
DEBTINC	Input	Interval	Yes	
DELINQ	Input	Interval	No	
DEROG	Input	Interval	No	
JOB	Input	Nominal	No	
LOAN	Input	Interval	No	
MORTDUE	Input	Interval	No	
NINQ	Input	Interval	No	
REASON	Input	Nominal	No	
VALUE	Input	Interval	No	
YOJ	Input	Interval	No	

Ćwiczenie 1a - odpowiedzi

Zweryfikować poziom modelowanego zjawiska w zbiorze źródłowym



1. Otworzenie Eksploratora (Explore…) z poziomu zbioru w drzewie projektu

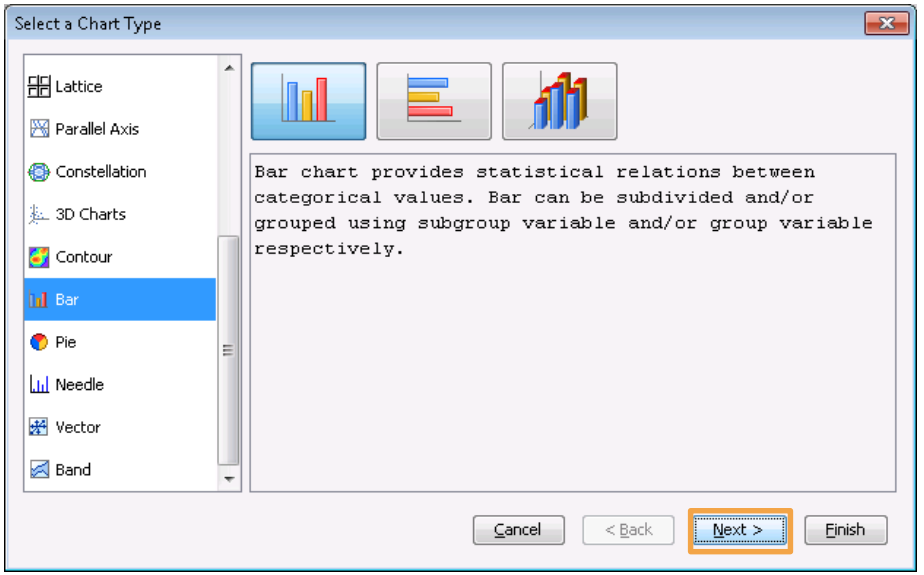


2. Zmiana ustawienia próbkownia z *Top* na *Random*
3. Aplikacja zmienionych ustawień
4. Przejście do polecenia *Plot…*

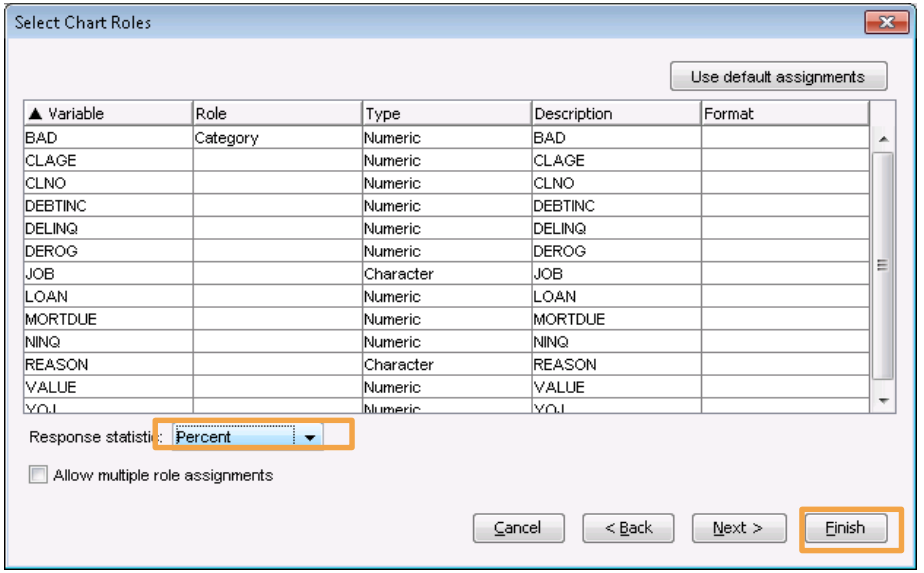
Ćwiczenie 1a - odpowiedzi

Zweryfikować poziom modelowanego zjawiska w zbiorze źródłowym

1. Wybór wykresu słupkowego: *Bar*



2. Określenie roli *Category* dla zmiennej celu *Bad*
3. Wybór statystyki *Percent*



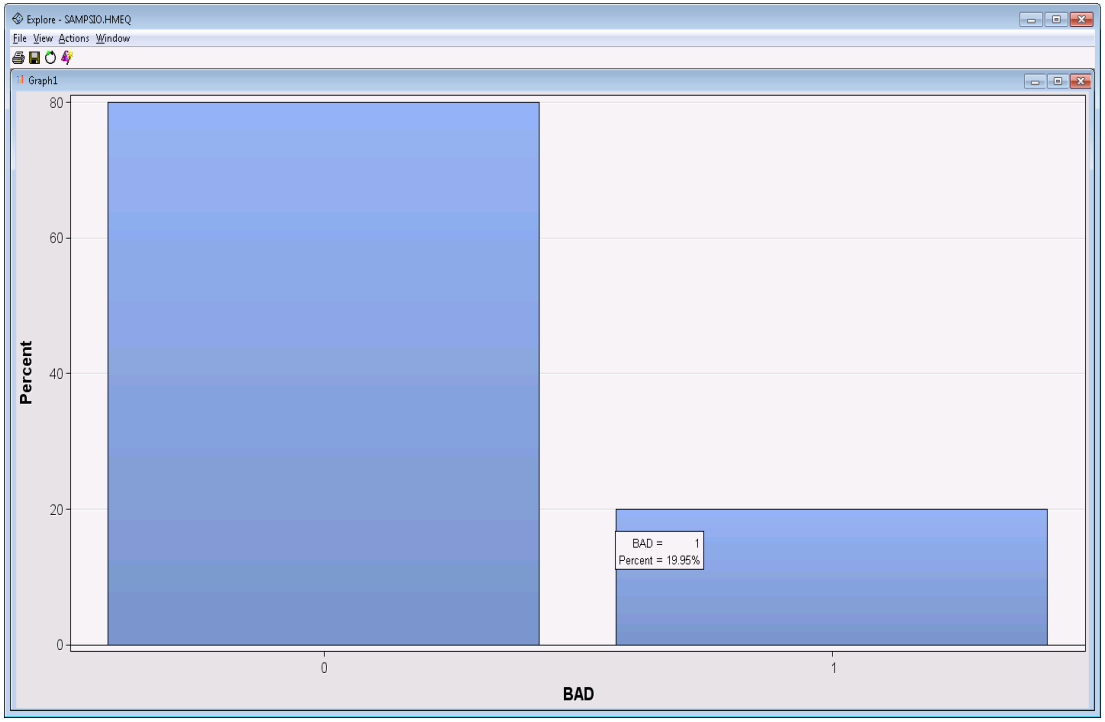
Pytanie kontrolne:

Dlaczego wybrano statystykę Percent zamiast Frequency?

Ćwiczenie 1a - odpowiedzi

Zweryfikować poziom modelowanego zjawiska w zbiorze źródłowym

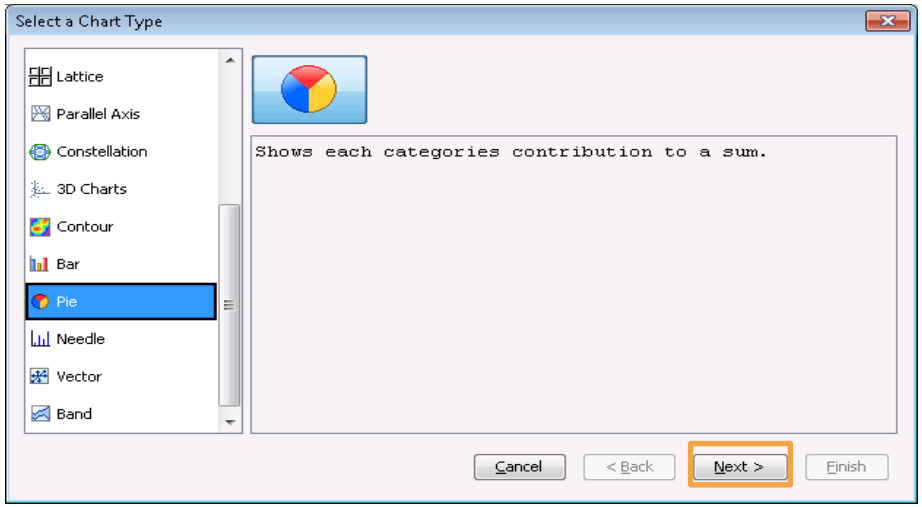
Pytanie kontrolne: jaki jest poziom modelowanego zjawiska?



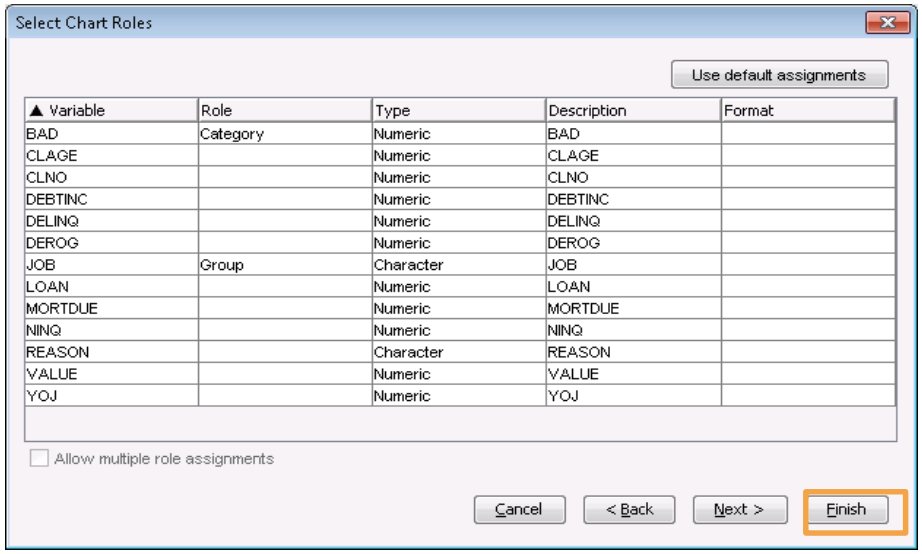
Ćwiczenie 1a - odpowiedzi

Dla zmiennej **Job**, zweryfikować poziomy zmiennej celu

1. Wybór wykresu kołowego: *Pie*



2. Określenie roli:
   1. *Category* dla zmiennej celu *Bad*
   2. *Group* dla zmiennej *Job*

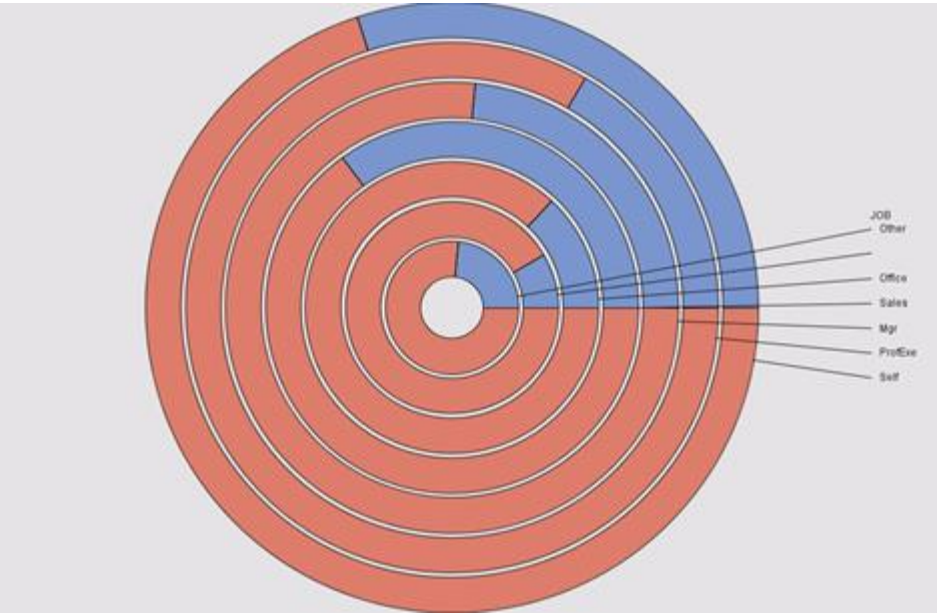


▲ Variable	Role	Type	Description	Format
BAD	Category	Numeric	BAD	
CLAGE		Numeric	CLAGE	
CLNO		Numeric	CLNO	
DEBTINC		Numeric	DEBTINC	
DELINQ		Numeric	DELINQ	
DEROG		Numeric	DEROG	
JOB	Group	Character	JOB	
LOAN		Numeric	LOAN	
MORTDUE		Numeric	MORTDUE	
NINQ		Numeric	NINQ	
REASON		Character	REASON	
VALUE		Numeric	VALUE	
YOJ		Numeric	YOJ	

Ćwiczenie 1a - odpowiedzi

Dla zmiennej **Job**, zweryfikować poziomy zmiennej celu

Pytanie kontrolne: dla jakiego zawodu poziom modelowanego zjawiska jest najmniejszy, a dla którego największy?



Ćwiczenie 1b

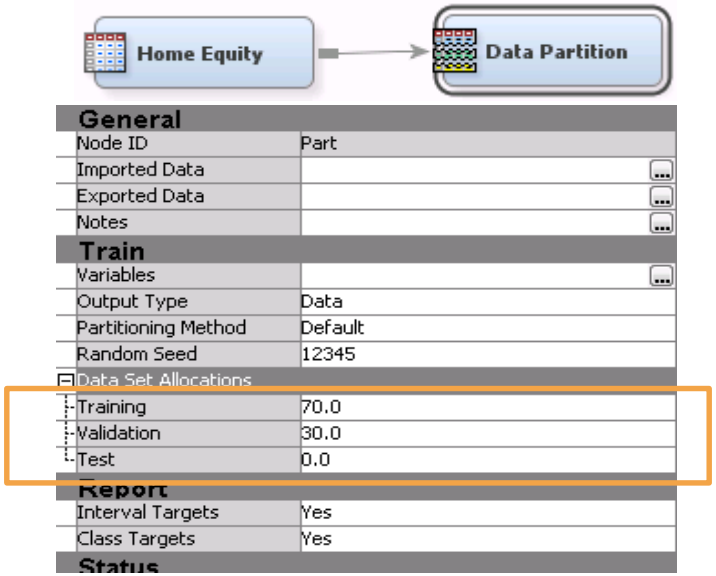
- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Home Equity
    - Zmienna celu: good_bad
- Diagram
  - Home Equity_diag1
- Model
  - Drzewo decyzyjne

- Wykonać partycjonowanie
  - 70% część treningowa
  - 30% część walidacyjna
- Uruchomić algorytm drzewa decyzyjnego z domyślnymi ustawieniami
- Zinterpretować wyniki
  - Podjąć decyzję o ewentualnej zmianie parametrów drzewa

Ćwiczenie 1b

Wykonać partycjonowanie

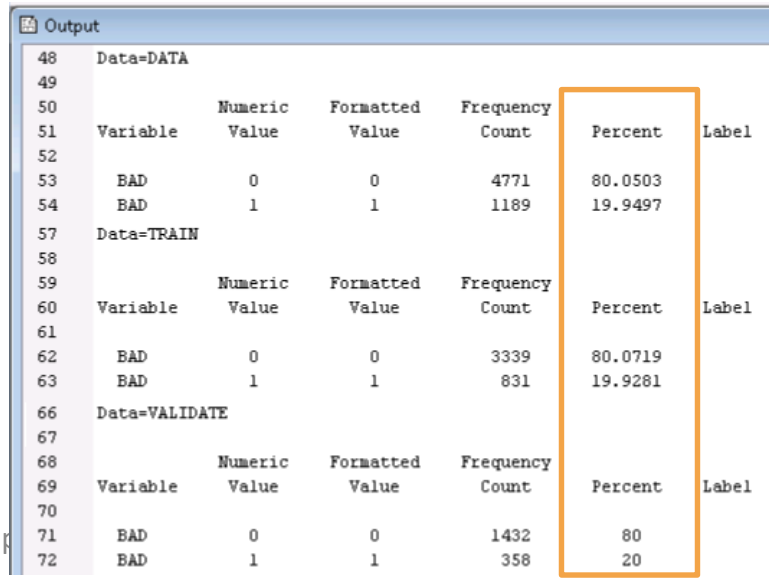
1. Na diagramie umieścić dane wejściowe
2. Dołączyć do nich węzeł *Data Partition* z zakładki *Input*
3. Zmodyfikować ustawienia węzła



Pytanie kontrolne:

Jaka metoda podziału zbioru została użyta?

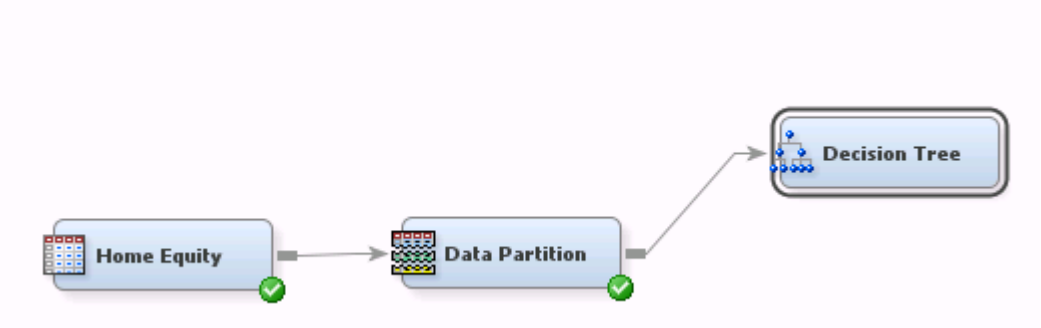
1. Zweryfikować wyniki węzła:
2. Jaki poziom modelowanego zjawiska występuje w zbiorze treningowym, a jaki w zbiorze walidacyjnym?

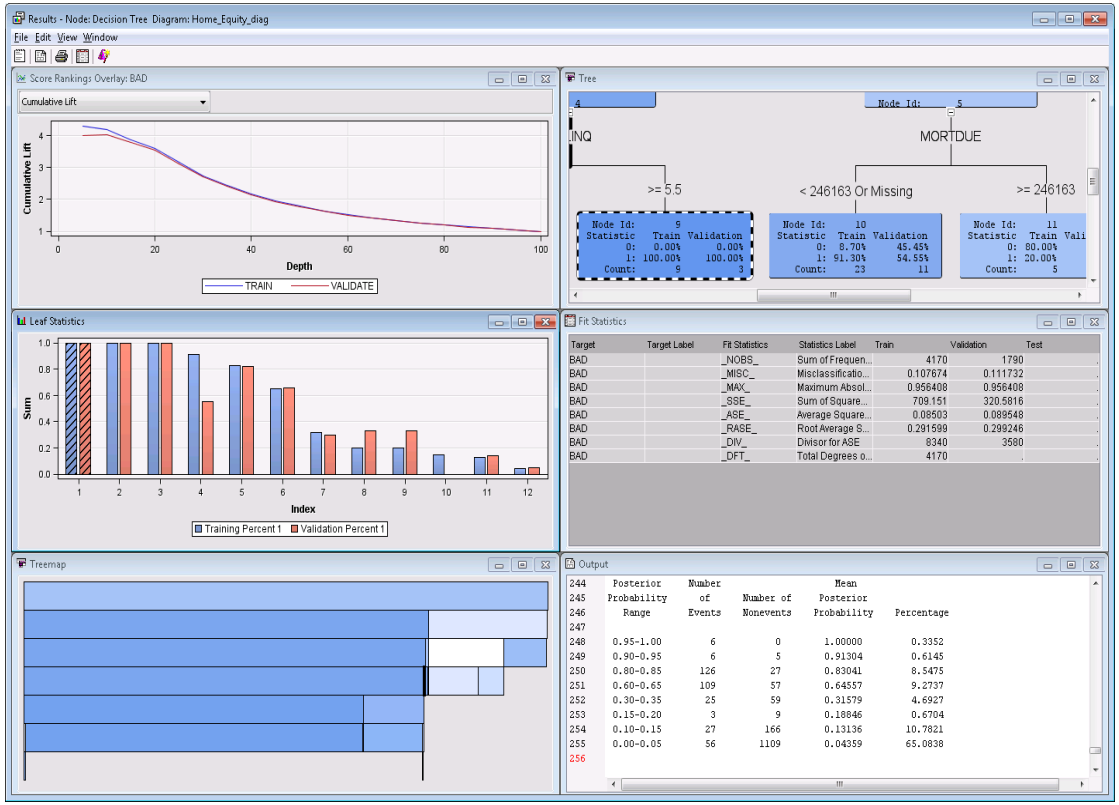


Ćwiczenie 1b

Uruchomić algorytm drzewa decyzyjnego z domyślnymi ustawieniami

1. Do węzła *Data Partition* dołączyć węzeł *Decision Tree* z zakładki *Model*
2. Uruchomić diagram





Pytania kontrolne:

1. Jaka reguła prowadzi do liścia 1?
2. Ile obserwacji znajduje się w części walidacyjnej i testowej liścia?
3. Czy model jest stabilny?

Ćwiczenie 1c

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Home Equity
    - Zmienna celu: good_bad
- Diagram
  - Home Equity_diag1
- Model
  - Drzewo decyzyjne

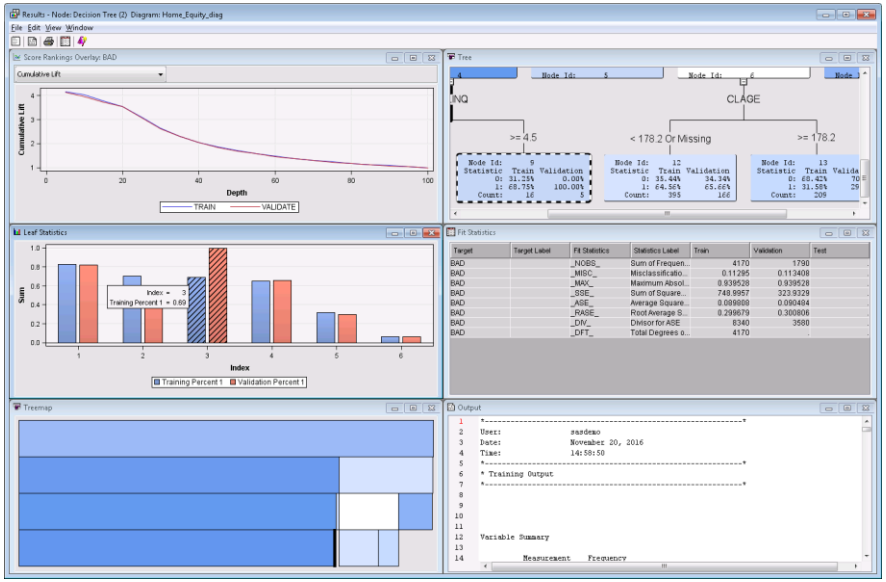
- Zwiększyć rozmiar liścia do 15
- Zinterpretować wyniki

Ćwiczenie 1c

Zwiększyć rozmiar liścia do 15

Property	Value
Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.2
Missing Values	Use in search
Use Input Once	No
Maximum Branch	2
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	15
Number of Rules	5
Number of Surrogate Rule	0
Split Size	.

1. Do węzła *Data Partition* dołączyć kolejny węzeł *Decision Tree* z zakładki *Model*
2. Zmienić ustawienia węzła dla opcji Leaf Size na 15



Pytania kontrolne:

1. Jaka jest liczba liści?
2. Czy model jest stabilny?
3. Która zmienna jest najbardziej istotna w modelu?

Ćwiczenie 1c

Zwiększyć rozmiar liścia do 15

Która zmienna jest najbardziej istotna w modelu?

Results - Node: Decision Tree (2) Diagram: Home_Equity_diag

File Edit View Window

Output

Variable Importance

	Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
67	DEBTINC		1	1.0000	1.0000	1.0000
68	DELINQ		2	0.3297	0.3377	1.0244
69	CLAGE		1	0.2504	0.2633	1.0518
70	VALUE		1	0.2350	0.1552	0.6604

Ćwiczenie 1d

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - Home Equity
    - Zmienna celu: good_bad
- Diagram
  - Home Equity_diag1
- Model
  - Drzewo decyzyjne

- Zbudować drzewo CHAID

CHAID w SAS Enterprise Miner

Opcja	Ustawienie	Komentarz
Nominal Targets -> Nominal Criterion	PROBCHISQ	
Interval Targets -> Interval Criterion	PROBF	
Method	Largest	Uniknięcie automatycznego przycinania
Significance Level	Np. 0.05	Ustawienie poziomu istotności testu F lub Chi2
Maximum Branch	Np. 25	Wartość równa maksymalnej liczbie kategorii w zmiennych nominalnych
Number of Surrogate Rules	0	Brak reguł zastępczych
Exhaustive	0	Wyszukiwanie heurystyczne
Leaf Size	1	
Split Size	2	
Bonferroni Adjustment	Yes	
Time of Bonferroni Adjustment	After	

Ćwiczenie 1d

Zbudować drzewo CHAID

Ustawić opcje wg tabeli na poprzednim slajdzie

Results - Node: Decision Tree (2) Diagram: Home_Equity_diag

File Edit View Window

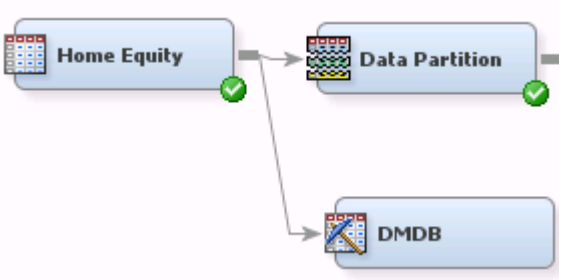
Output

Variable Importance

	Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
67	DEBTINC		1	1.0000	1.0000	1.0000
68	DELINQ		2	0.3297	0.3377	1.0244
69	CLAGE		1	0.2504	0.2633	1.0518
70	VALUE		1	0.2350	0.1552	0.6604

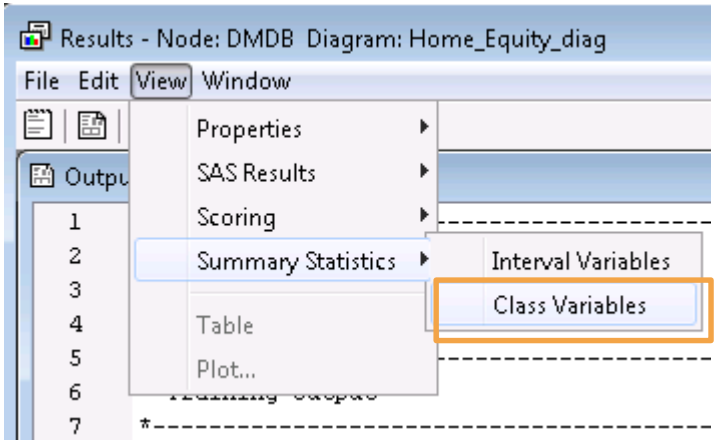
Ćwiczenie 1d

Zbudować drzewo CHAID

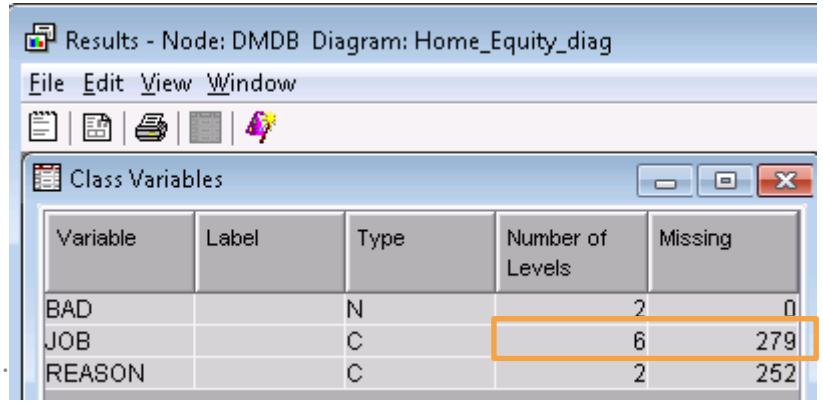


Zweryfikować maksymalną liczbę poziomów zmiennych wejściowych: nominalnych/porządkowych

1. Zastosować węzeł *DMDB* po węźle danych źródłowych
2. Zweryfikować statystyki zmiennych klasyfikujących



3. Jaka jest maksymalna liczba poziomów dla zmiennych objaśniających klasyfikujących?

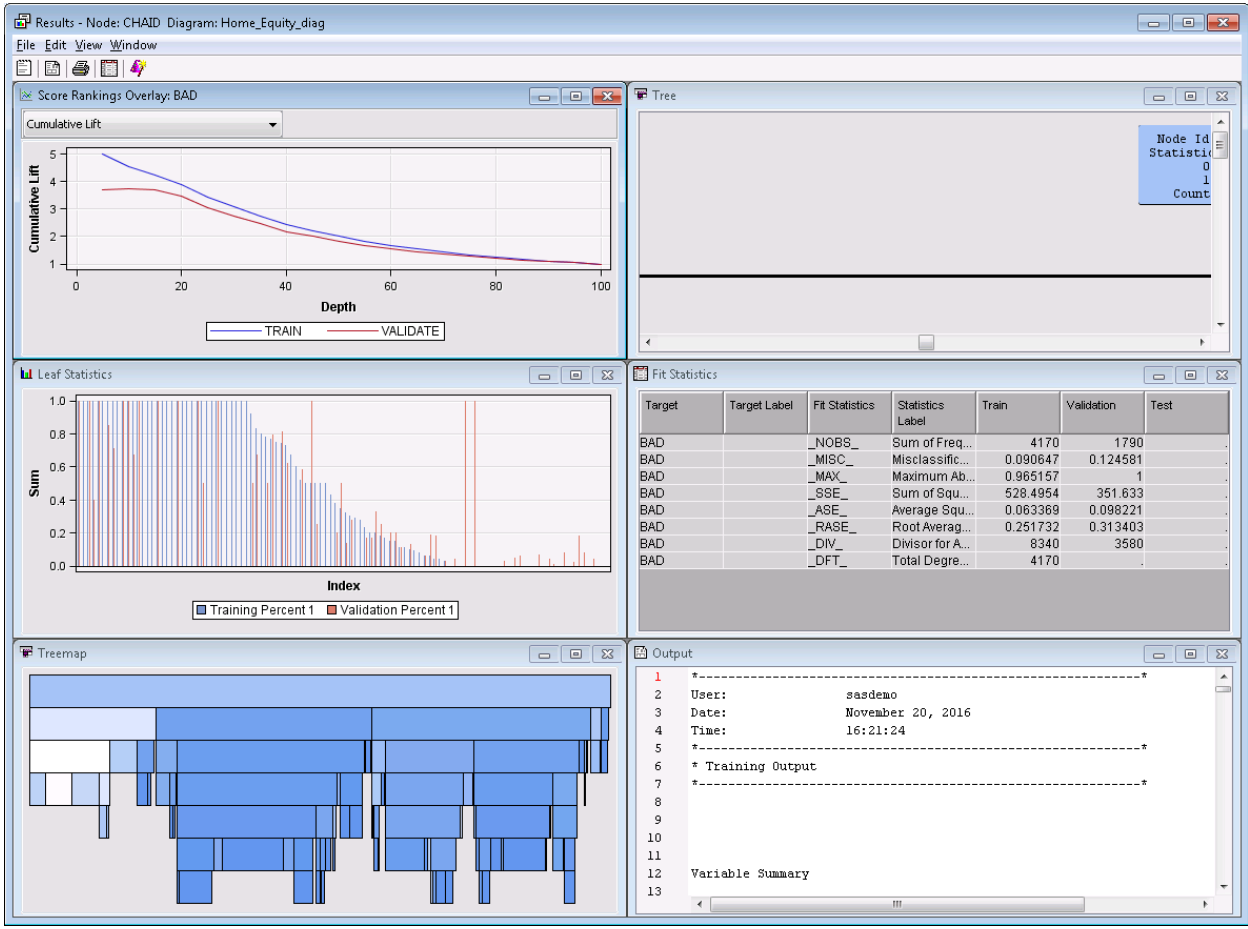


Variable	Label	Type	Number of Levels	Missing
BAD		N	2	0
JOB		C	6	279
REASON		C	2	252

Ćwiczenie 1d

Zbudować drzewo CHAID

Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.05
Missing Values	Use in search
Use Input Once	No
Maximum Branch	7
Maximum Depth	6
Minimum Categorical Size	5
Node	
Leaf Size	1
Number of Rules	5
Number of Surrogate Rules	0
Split Size	2
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	0
Node Sample	20000
Subtree	
Method	Largest
Number of Leaves	1
Assessment Measure	Decision
Assessment Fraction	0.25
Cross Validation	
Perform Cross Validation	No
Number of Subsets	10
Number of Repeats	1
Seed	12345
Observation Based Importance	
Observation Based Importance	No
Number Single Var Importance	5
P-Value Adjustment	
Bonferroni Adjustment	Yes
Time of Bonferroni Adjustment	After
Inputs	No
Number of Inputs	1
Depth Adjustment	Yes



Ćwiczenie 2

Opis zbioru danych - PMAD_PVA

Zbiór zawiera informacje opisujące potencjalnych darczyńców oraz zmienną celu – flagę przekazania darowizny.

Nazwa zmiennej	Etykieta	Opis
ID	Control Number	
TargetB	Target Gift Flag	Zmienna celu – flaga przekazania darowizny
TargetD	Target Gift Amount	Zmienna celu – kwota darowizny
GiftCnt36	Gift Count 36 Months	Cechy opisujące historię przekazywania darowizn
GiftCntAll	Gift Count All Months	
GiftCntCard36	Gift Count Card 36 Months	
GiftCntCardAll	Gift Count Card All Months	
GiftAvgLast	Gift Amount Last	
GiftAvg36	Gift Amount Average 36 Months	
GiftAvgAll	Gift Amount Average All Months	
GiftAvgCard36	Gift Amount Average Card 36 Months	
GiftTimeLast	Time Since Last Gift	
GiftTimeFirst	Time Since First Gift	

Opis zbioru danych - PMAD_PVA

Nazwa zmiennej	Etykieta	Opis
PromCnt12	Promotion Count 12 Months	Cechy opisujące fakt otrzymywania materiałów promocyjnych przez darczyńcę
PromCnt36	Promotion Count 36 Months	
PromCntAll	Promotion Count All Months	
PromCntCard12	Promotion Count Card 12 Months	
PromCntCard36	Promotion Count Card 36 Months	
PromCntCardAll	Promotion Count Card All Months	
StatusCat96NK	Status Category 96NK	Opis statusu darczyńcy
StatusCatStarAll	Status Category Star All Months	
DemCluster	Demographic Cluster	Cechy demograficzne
DemAge	Age	
DemGender	Gender	
DemHomeOwner	Home Owner	
DemMedHomeValue	Median Home Value Region	
DemPctVeterans	Percent Veterans Region	
DemMedIncome	Median Income Region	

Ćwiczenie 2

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - PMAD_PVA
    - Zmienna celu: TargetB
- Diagram
  - Donation_Analysis_1
- Modele
  - Drzewo decyzyjne
  - Regresja logistyczna
  - Sieć neuronowa

a) Zbudować drzewo decyzyjne

b) Zbudować model regresji logistycznej

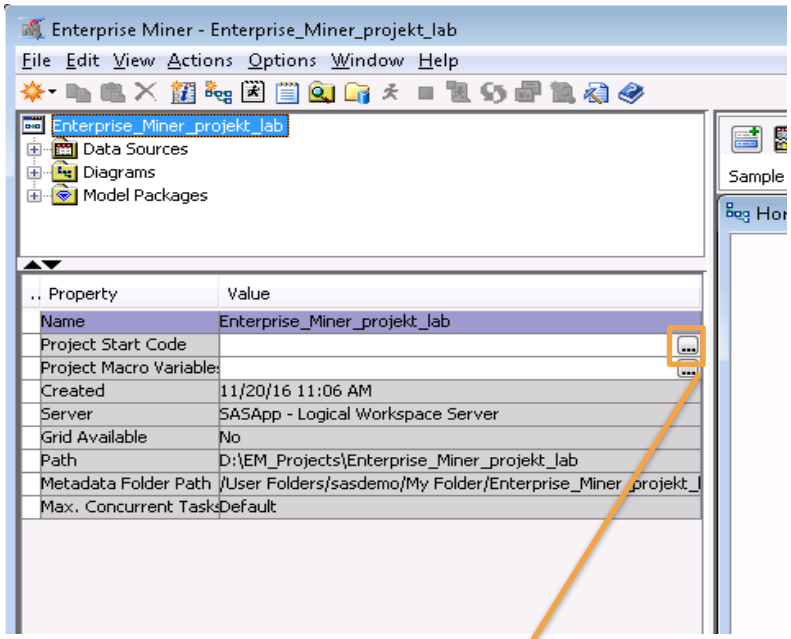
c) Zbudować model sieci neuronowej

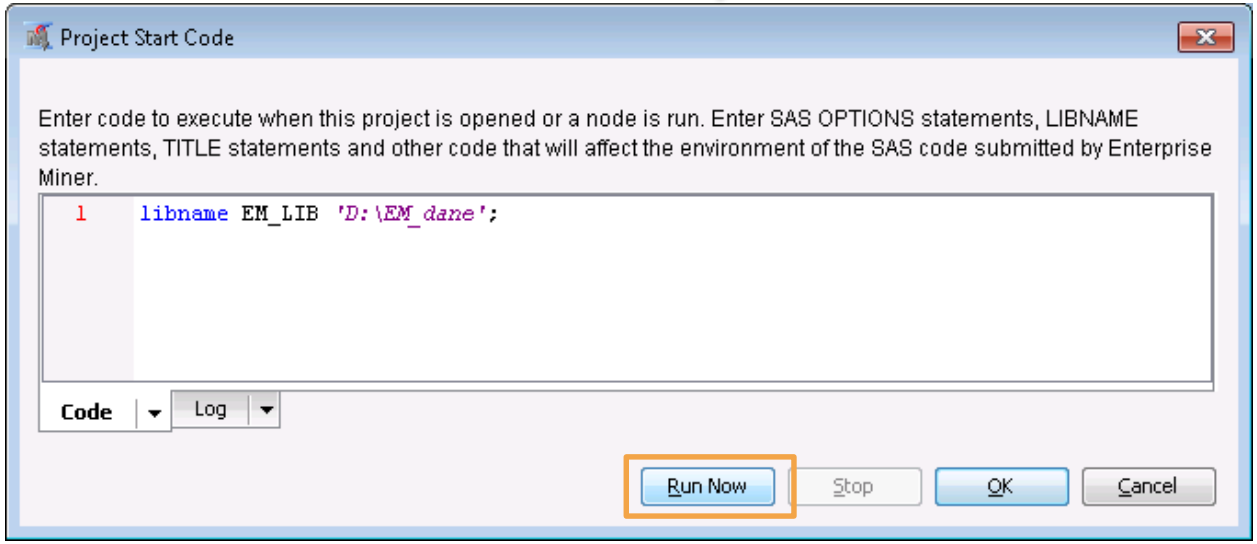
Dodanie źródła danych

Jeżeli źródło danych nie znajduje się w bibliotece dostępnej w środowisku, użytkownik powinien zdefiniować bibliotekę w kodzie startowym projektu.

Po zdefiniowaniu kodu, należy go uruchomić.

Od tego momentu nowa biblioteka będzie widoczna dla SAS Enterprise Miner.





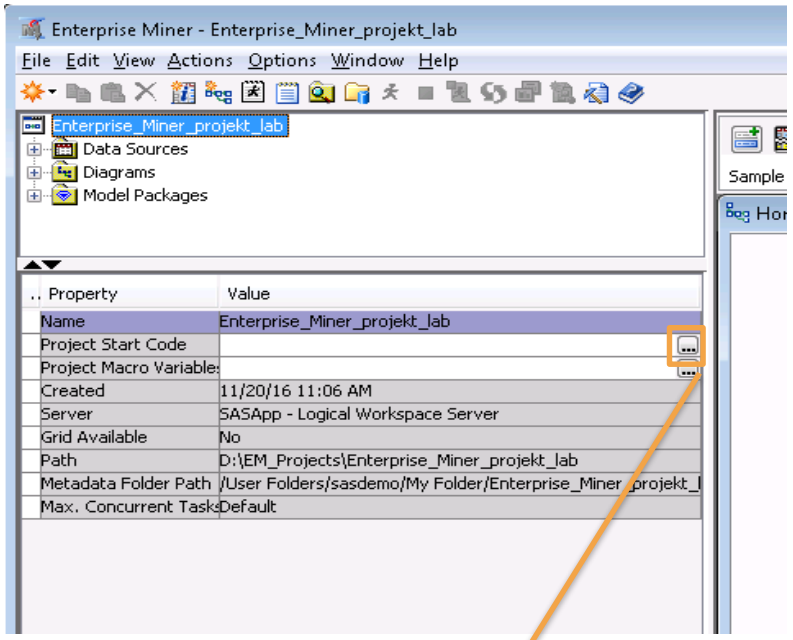
Dodanie źródła danych

Jeżeli źródło danych nie znajduje się w bibliotece dostępnej w środowisku, użytkownik powinien zdefiniować bibliotekę w kodzie startowym projektu.

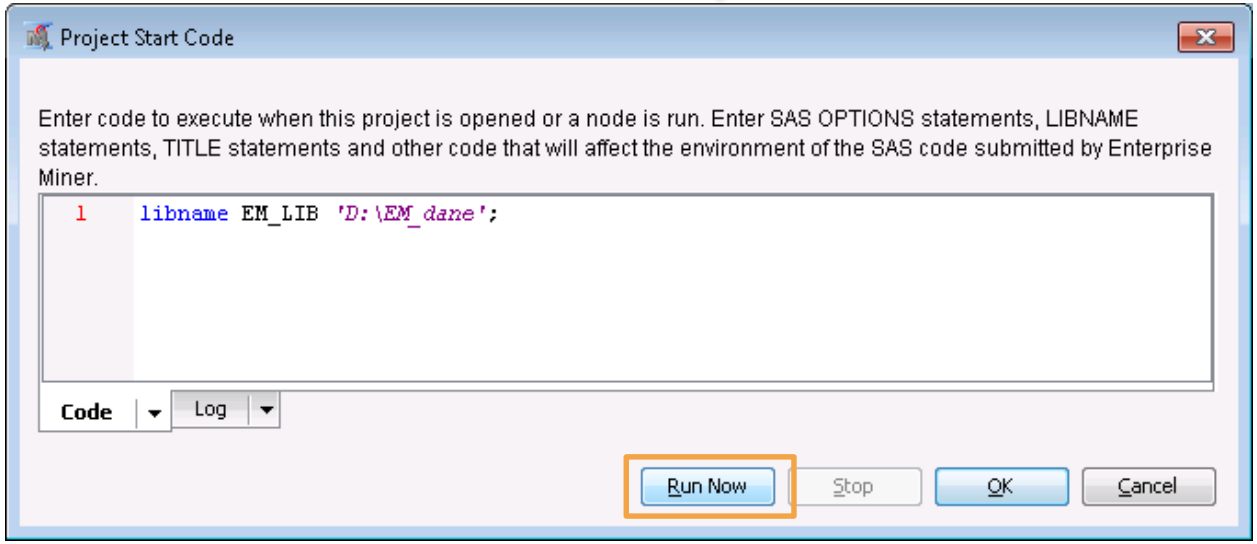
Po zdefiniowaniu kodu, należy go uruchomić.

Od tego momentu nowa biblioteka będzie widoczna dla SAS Enterprise Miner.

1



2

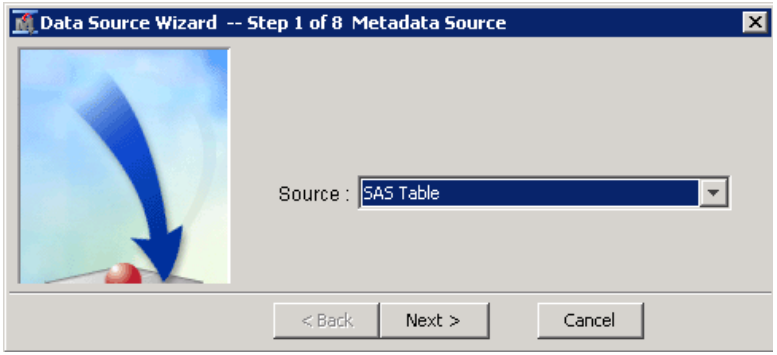


Dodanie źródła danych

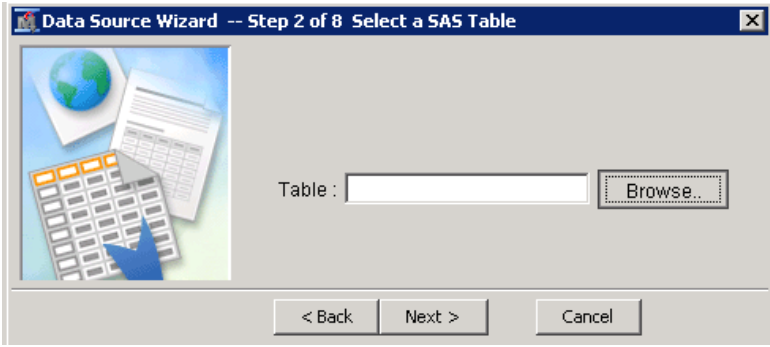
Wybór źródła danych z biblioteki

1

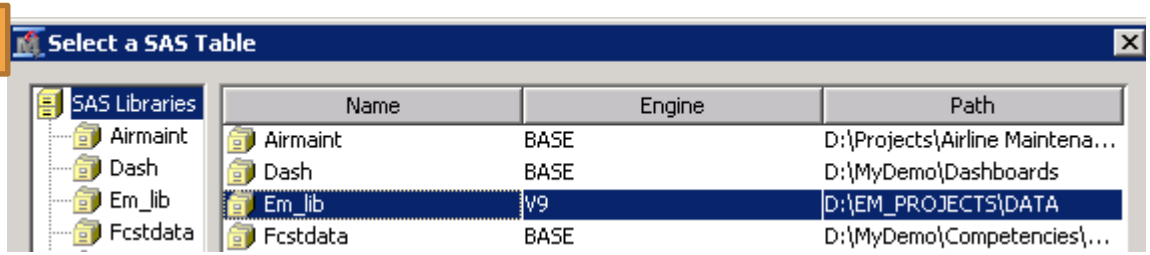
2



3



4



Ćwiczenie 2a

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - PMAD_PVA
    - Zmienna celu: TargetB
- Diagram
  - Donation Analysis
- Model
  - Drzewo decyzyjne

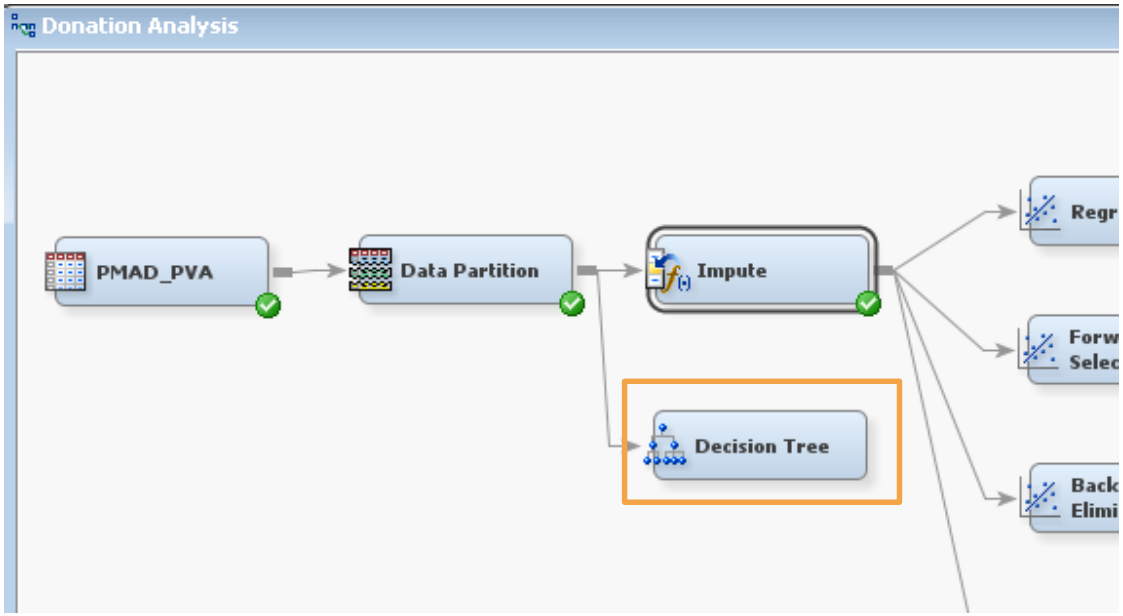
Budowa drzewa decyzyjnego:

- Dodać do projektu zbiór PMAD_PVA
  - TargetB – rola TARGET, poziom Binary
  - TargetD – rola REJECTED
- Dokonać podziału zbioru na część treningową i walidacyjną (w proporcji 70/30)
- Zbudować model drzewa decyzyjnego z domyślnymi parametrami
- Zmienić metodę wyboru najlepszego drzewa z domyślnej na Assessment

Ćwiczenie 2a

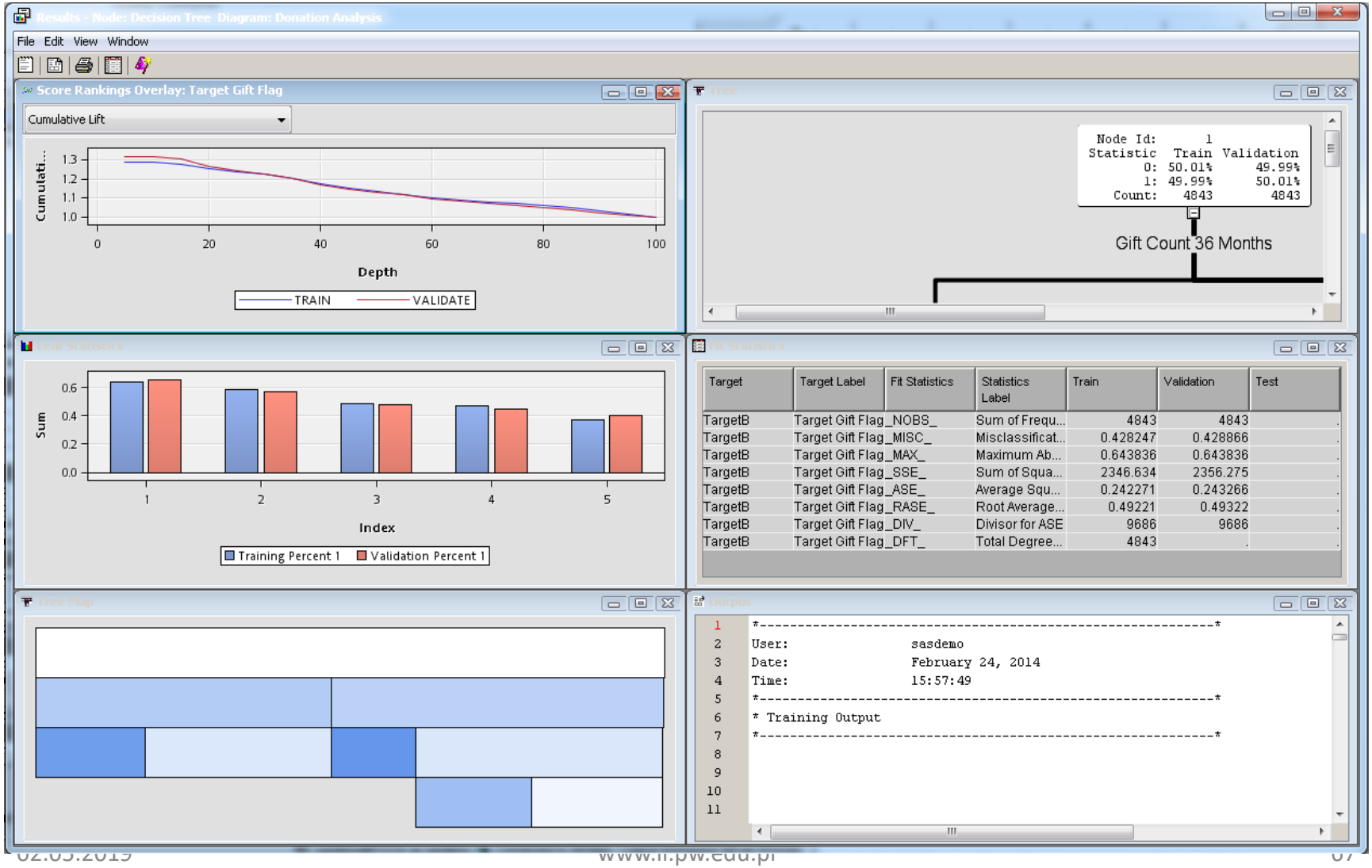
- Pytania kontrolne:
  - Jak drzewa decyzyjne radzą sobie z brakami danych?
  - Czy drzewa decyzyjne wymagają przekształceń zmiennych?
  - Czy drzewa decyzyjne wymagają wstępnej selekcji zmiennych?
  - Jaką metodę wyboru najlepszego drzewa należy zastosować?
  - Jaka jest liczba liści w wybranym modelu?
  - Jaka jest minimalna liczebność liścia?
  - Która zmienna jest najbardziej istotna?

- Budowa drzewa decyzyjnego:



Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Average Square Error
Assessment Fraction	0.25

Ćwiczenie 2a



Ćwiczenie 2b

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - PMAD_PVA
    - Zmienna celu: TargetB
- Diagram
  - Donation Analysis
- Model
  - Regresja logistyczna

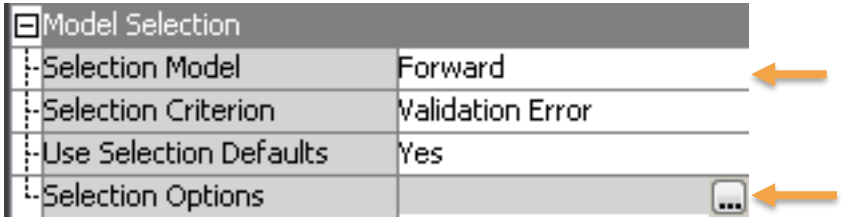
Budowa modelu regresji logistycznej:

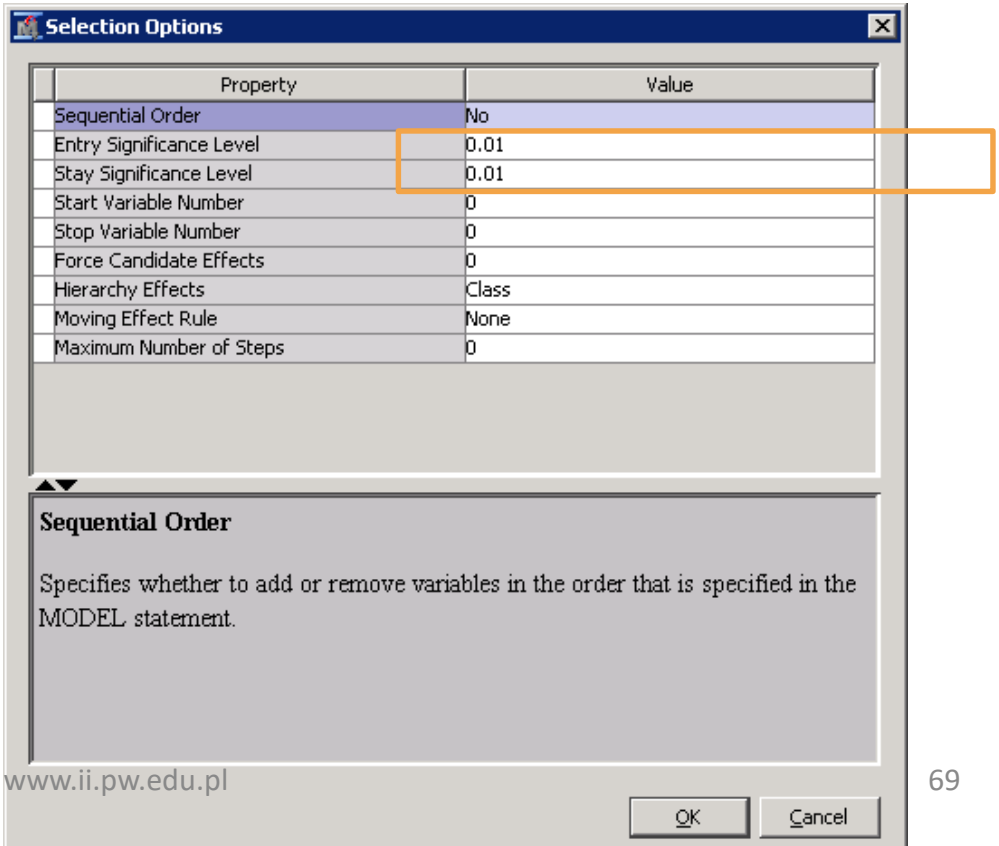
- Dokonać podziału zbioru PMAD_PVA na część treningową i walidacyjną (w proporcji 70/30)
- Uzupełnić braki danych dla zmiennych objaśniających
- Zbudować 3 modele regresji logistycznej z metodą doboru zmiennych Stepwise, Forward oraz Backward
- Zmienić poziom istotności dla wejścia oraz pozostania zmiennych w modelu

Ćwiczenie 2b

- Pytania kontrolne:
  - Jak regresja logistyczna radzi sobie z brakami danych?
  - Czy regresja logistyczna wymaga przekształceń zmiennych?
  - Czy regresja logistyczna wymaga wstępnej selekcji zmiennych?
  - Jaka jest liczba zmiennych objaśniających w każdym modelu?
  - Które zmienne powtarzają się w więcej niż jednym modelu? Jaka jest ich istotność?
  - Czy zmiana poziomu istotności dla wejścia/pozostania zmiennej w modelu spowodowała poprawę jakości modelu?

- Budowa modelu regresji logistycznej:





Ćwiczenie 2c

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - PMAD_PVA
    - Zmienna celu: TargetB
- Diagram
  - Donation Analysis
- Model
  - Sieć neuronowa

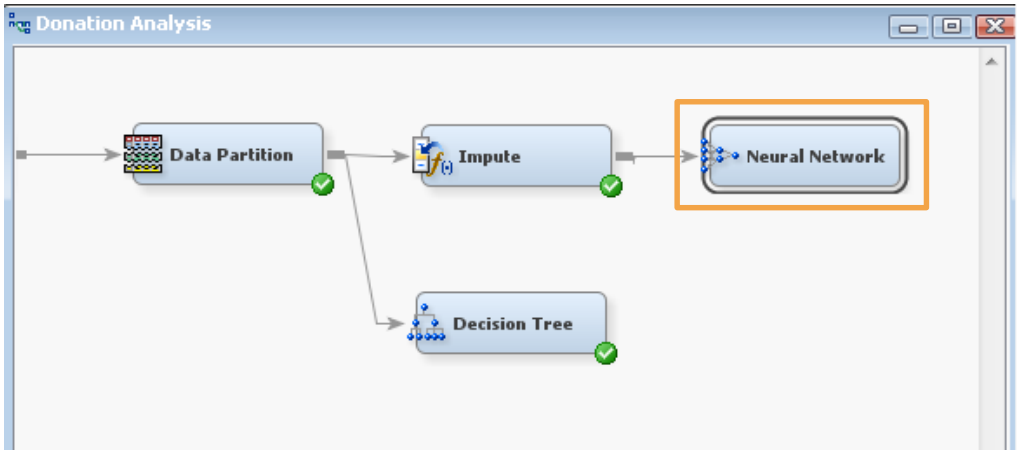
Budowa modelu sieci neuronowej:

- Dokonać podziału zbioru PMAD_PVA na część treningową i walidacyjną (w proporcji 70/30)
- Uzupełnić braki danych dla zmiennych objaśniających
- Zbudować model sieci neuronowej z domyślnymi parametrami
- Zmienić metodę wyboru najlepszego modelu na Average Error

Ćwiczenie 2c

- Pytania kontrolne:
  - Jak sieci neuronowe radzą sobie z brakami danych?
  - Czy sieci neuronowe wymagają przekształceń zmiennych?
  - Czy sieci neuronowe wymagają wstępnej selekcji zmiennych?
  - Jaka jest liczba parametrów w modelu?
  - Jaką wartość błędu średniokwadratowego ma najlepszy model?
  - Z ilu warstw składa się wybrana sieć?

- Budowa modelu sieci neuronowej:



Train	
Variables	
Continue Training	No
Network	
Optimization	
Initialization Seed	12345
Model Selection Criterion	Average Error
Suppress Output	No

DODATKOWE WĘZŁY WYKORZYSTYWANE W PROCESIE MODELOWANIA

Ćwiczenie 3

Ćwiczenie 3

- Projekt
  - Enterprise_Miner_projekt_lab_*nazwisko*
- Zbiór źródłowy
  - PMAD_PVA
    - Zmienna celu: TargetB
- Diagram
  - Donation_Analysis_2
- Modele
  - Drzewo decyzyjne
  - Regresja logistyczna
  - Las losowy

W poprzednim procesie dodać nowe węzły EM

a) Zbudować model drzewa decyzyjnego
b) Zbudować model regresji
c) Zbudować model lasu losowego
d) Porównać zbudowane modele
e) Wygenerować kod scoringowy
f) Utworzyć pakiet modelu
g) Zarejestrować model w metadanych

Ćwiczenie 3 a)

Zbudować model drzewa decyzyjnego

Pytania kontrolne

- Zdecydować jakie kroki (węzły EM) są wymagane do zbudowania drzewa decyzyjnego

- Jaka maksymalna liczba gałęzi została zastosowana w drzewie? Która zmienna została użyta do tego podziału?
- Ile zmiennych okazało się istotnych w drzewie?
- Jaka jest liczba liści w drzewie?
- Która ze zmiennych jest najbardziej istotna?
- Czy model jest stabilny?
- Jaka jest wartość statystyki ASE dla zbioru walidacyjnego?

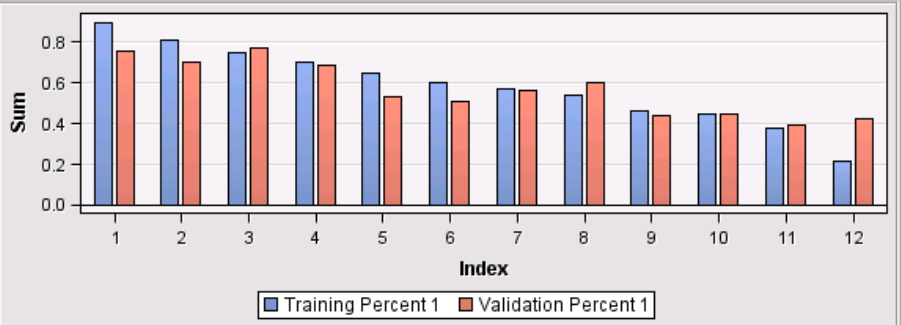
- Do zbioru wejściowego podłączyć węzeł Data Partition
  - Zastosować proporcję 60:40 (zbiór treningowy: zbiór walidacyjny)
- Jeśli jest to wymagane, zastosować dodatkowe węzły
- Zastosować model drzewa decyzyjnego
  - Maksymalna liczba gałęzi: 6
  - Maksymalna głębokość drzewa: 10
  - Rozmiar liścia: 10

File Edit View Window

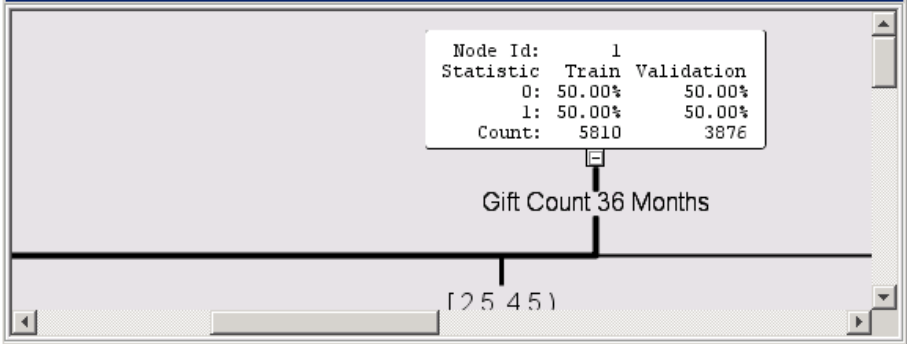
Output

```
43   PREDICTED    P_TargetB1    Predicted: TargetB=1
44   RESIDUAL     R_TargetB1    Residual: TargetB=1
45   PREDICTED    P_TargetB0    Predicted: TargetB=0
46   RESIDUAL     R_TargetB0    Residual: TargetB=0
47   FROM         F_TargetB     From: TargetB
48   INTO         I_TargetB     Into: TargetB
49
50
51   *-------------------------------------------------------------*
52   * Score Output
53   *-------------------------------------------------------------*
```

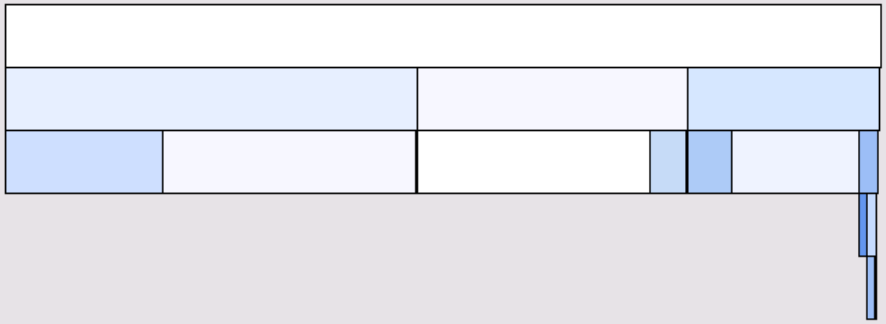
Leaf Statistics



Tree

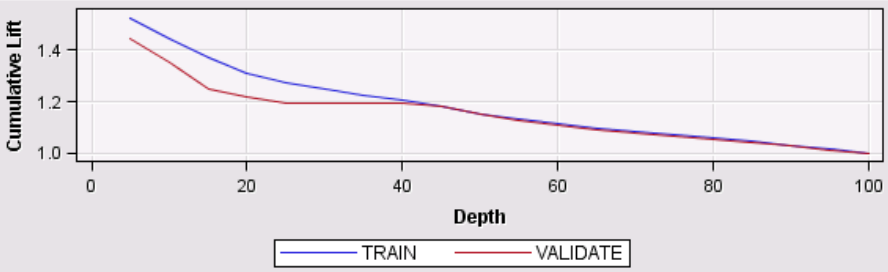


Treemap



Score Rankings Overlay: Target Gift Flag

Cumulative Lift



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetB	Target Gift Fl...	_NOBS_	Sum of Freq...	5810	3876	.
TargetB	Target Gift Fl...	_MISC_	Misclassifica...	0.416523	0.417183	.
TargetB	Target Gift Fl...	_MAX_	Maximum Ab...	0.898305	0.898305	.
TargetB	Target Gift Fl...	_SSE_	Sum of Squa...	2781.949	1881.657	.
TargetB	Target Gift Fl...	_ASE_	Average Squ...	0.23941	0.242732	.
TargetB	Target Gift Fl...	_RASE_	Root Averag...	0.489296	0.492678	.
TargetB	Target Gift Fl...	_DIV_	Divisor for A...	11620	7752	.
TargetB	Target Gift Fl...	_DFT_	Total Degree...	5810	.	.

Ćwiczenie 3 b)

Zbudować model regresji

Pytania kontrolne

- Czy regresja wymaga dodatkowych kroków przygotowujących dane do modelowania? Czy należy zastosować dodatkowe węzły EM? Jeśli jest to wymagane, zastosować dodatkowe węzły

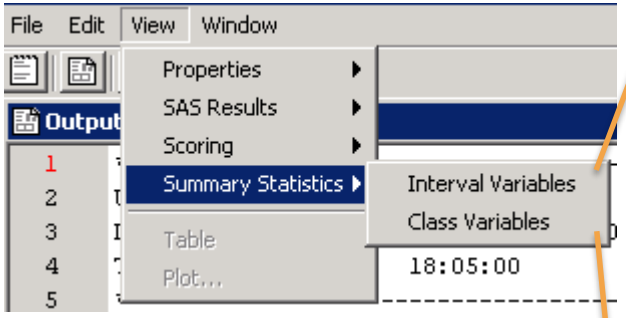
- Jeśli jest to wymagane, zastosować dodatkowe węzły
- Wybrać typ regresji (liniowa/logistyczna)

Ćwiczenie 3 b)

Zbudować model regresji

- Zweryfikować, czy w zbiorze występują braki danych wykorzystując węzeł DMDB
- Podłączyć węzeł bezpośrednio po zbiorze wejściowym

Ćwiczenie 3 b)



Interval Variables

Variable	Label	Missing	N	Minimum	Maximum	Mean	Standard Deviation	Skewness	Kurtosis
DemAge	Age	2407	7279	0	87	59.15084	16.5164	-0.38791	-0.47761
DemMedH...	Median Ho...	0	9686	0	600000	110986.3	98670.86	2.378211	6.451365
DemMedIn...	Median Inc...	2357	7329	2499	200001	53513.46	19805.17	1.739964	5.240484
DemPctVet...	Percent Vet...	0	9686	0	85	30.60427	11.39499	-0.20706	1.27441
GiftAvg36	Gift Amount...	0	9686	0	260	14.8762	10.05701	5.627792	77.09997
GiftAvgAll	Gift Amount...	0	9686	1.5	450	12.48932	9.209297	14.48649	561.7552
GiftAvgCard...	Gift Amount...	1780	7906	1.33	260	14.22443	10.02271	6.051455	87.12627
GiftAvgLast	Gift Amount...	0	9686	0	450	16.01774	12.0418	9.918893	246.0504
GiftCnt36	Gift Count 3...	0	9686	0	16	3.205451	2.133421	1.288353	2.047415
GiftCntAll	Gift Count A...	0	9686	1	91	10.50764	8.993401	1.863109	6.047766
GiftCntCard...	Gift Count ...	0	9686	0	9	1.856597	1.595419	1.172452	1.494867
GiftCntCard...	Gift Count ...	0	9686	0	41	5.58249	4.736894	1.331353	2.024864
GiftTimeFirst	Time Since ...	0	9686	15	260	71.10035	37.69198	0.195399	-1.24787
GiftTimeLast	Time Since ...	0	9686	4	27	18.00217	4.073549	-0.77805	2.469076
PromCnt12	Promotion ...	0	9686	2	59	12.98885	4.823458	2.873723	11.99538
PromCnt36	Promotion ...	0	9686	4	78	29.34823	7.809743	0.261958	2.174341
PromCntAll	Promotion ...	0	9686	5	174	48.48348	23.06148	0.460765	0.216596
PromCntCa...	Promotion ...	0	9686	0	17	5.392009	1.323648	0.684994	5.798685
PromCntCa...	Promotion ...	0	9686	2	28	11.95468	4.571568	-0.4266	-0.98685
PromCntCa...	Promotion ...	0	9686	2	56	19.00712	8.562193	0.142856	-0.78032
StatusCatSt...	Status Cate...	0	9686	0	1	0.540574	0.498377	-0.16286	-1.97388

Class Variables

Variable	Label	Type	Number of Levels	Missing
DemCluster	Demographic Cluster	C	26	0
DemGender	Gender	C	3	0
DemHomeOwner	Home Owner	C	2	0
StatusCat96NK	Status Category 96NK	C	6	0
TargetB	Target Gift Flag	N	2	0

Ćwiczenie 3 b)

Zbudować model regresji

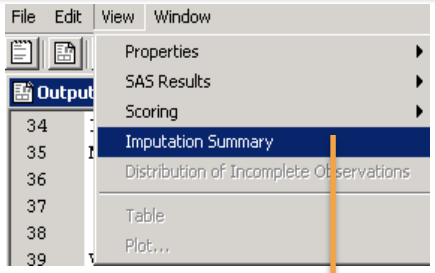
Pytania kontrolne:

- Jaką wartością zostały uzupełnione braki danych dla zmiennej DemAge
- Jak dużo zmiennych zostało wybranych przez węzeł Variable selection

- Uzupełnić braki danych
  - W węźle Impute:
    - dla zmiennych ciągłych zastosować metodę średniej
- Wykonać grupowanie wartości zmiennych nominalnych
  - W węźle Variable Selection wybrać:
    - Target Model: R-Square
    - Use group Variables: Yes

Ćwiczenie 3 b)

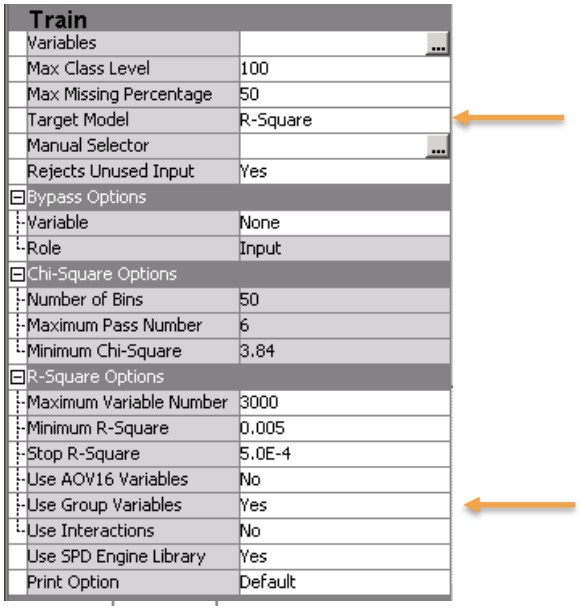
Rezultaty węzła Impute



Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
DemAge	MEAN	IMP_DemAge	59.30309	INPUT	INTERVAL	Age	1435
DemMedIncome	MEAN	IMP_DemMedIncome	53580.18	INPUT	INTERVAL	Median Income Region	1427
GiftAvgCard36	MEAN	IMP_GiftAvgCard36	14.28659	INPUT	INTERVAL	Gift Amount Average Car...	1096

Ustawienia węzła Variable Selection



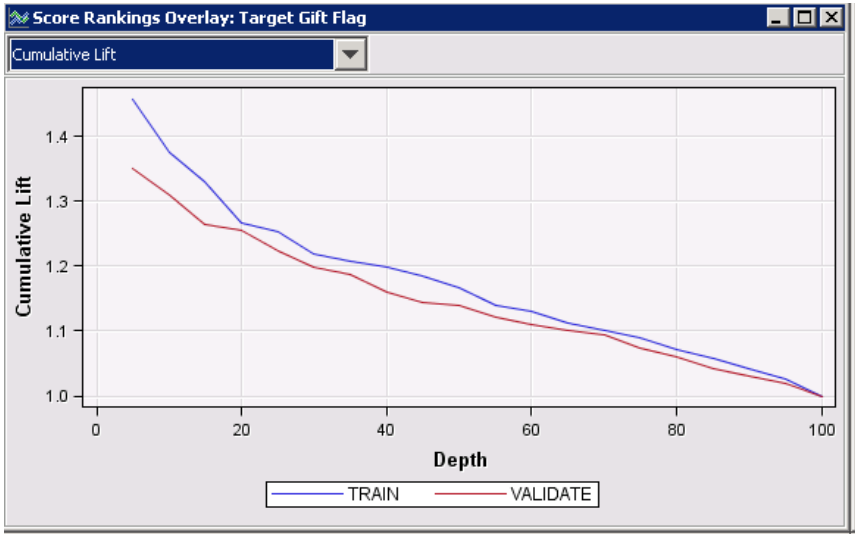
Ćwiczenie 3 b)

Zbudować model regresji

Pytania kontrolne:

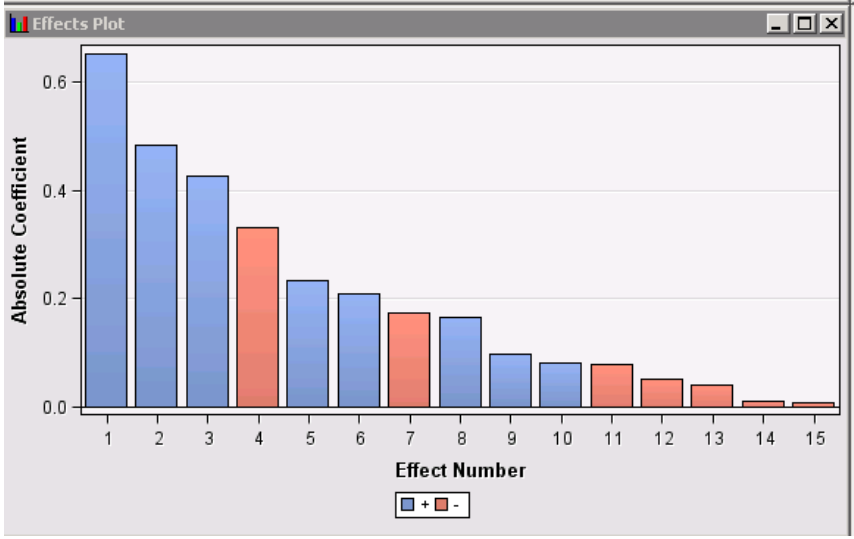
- Ile zmiennych zostało wybranych w finalnym modelu regresji?
- Czy w finalnym modelu została użyta zmienna nominalna?
- Jaka jest wartość statystyki lift skumulowany na 5% listy?
- Które ze zmiennych wpływają na obniżenie prawdopodobieństwa przekazania darowizny?

- Określić parametry węzła Regression:
  - Typ: Logistic Regression
  - Model selekcji: W tył
  - Kryterium selekcji: Kryterium informacyjne Akaike



Fit Statistics

Target	Target Label	Fit Statistics	Statistics Label	Train	Validation	Test
TargetB	Target Gift F...	_AIC_	Akaike's Info...	7833.56	.	.
TargetB	Target Gift F...	_ASE_	Average Sq...	0.239439	0.242939	.
TargetB	Target Gift F...	_AVERR_	Average Err...	0.671563	0.678913	.
TargetB	Target Gift F...	_DFE_	Degrees of ...	5795	.	.
TargetB	Target Gift F...	_DFM_	Model Degr...	15	.	.
TargetB	Target Gift F...	_DFT_	Total Degre...	5810	.	.
TargetB	Target Gift F...	_DIV_	Divisor for A...	11620	7752	.
TargetB	Target Gift F...	_ERR_	Error Function	7803.56	5262.936	.
TargetB	Target Gift F...	_FPE_	Final Predict...	0.240679	.	.
TargetB	Target Gift F...	_MAX_	Maximum A...	0.915231	0.816484	.
TargetB	Target Gift F...	_MSE_	Mean Squar...	0.240059	0.242939	.
TargetB	Target Gift F...	_NOBS_	Sum of Freq...	5810	3876	.
TargetB	Target Gift F...	_NW_	Number of ...	15	.	.
TargetB	Target Gift F...	_RASE_	Root Averag...	0.489325	0.492889	.
TargetB	Target Gift F...	_RFPE_	Root Final P...	0.49059	.	.
TargetB	Target Gift F...	_RMSE_	Root Mean ...	0.489958	0.492889	.
TargetB	Target Gift F...	_SBC_	Schwarz's B...	7933.57	.	.
TargetB	Target Gift F...	_SSE_	Sum of Squ...	2782.283	1883.264	.



Output

```
1   *-------------------------------------------------------------*
2   User:                  sasdemo
3   Date:                  December 05, 2016
4   Time:                  18:18:37
5   *-------------------------------------------------------------*
6   * Training Output
7   *-------------------------------------------------------------*
8
9
10
11
12  Variable Summary
13
14              Measurement     Frequency
15   Role         Level          Count
16
17  INPUT       INTERVAL           5
18  INPUT       NOMINAL            2
19  TARGET      BINARY             1
```

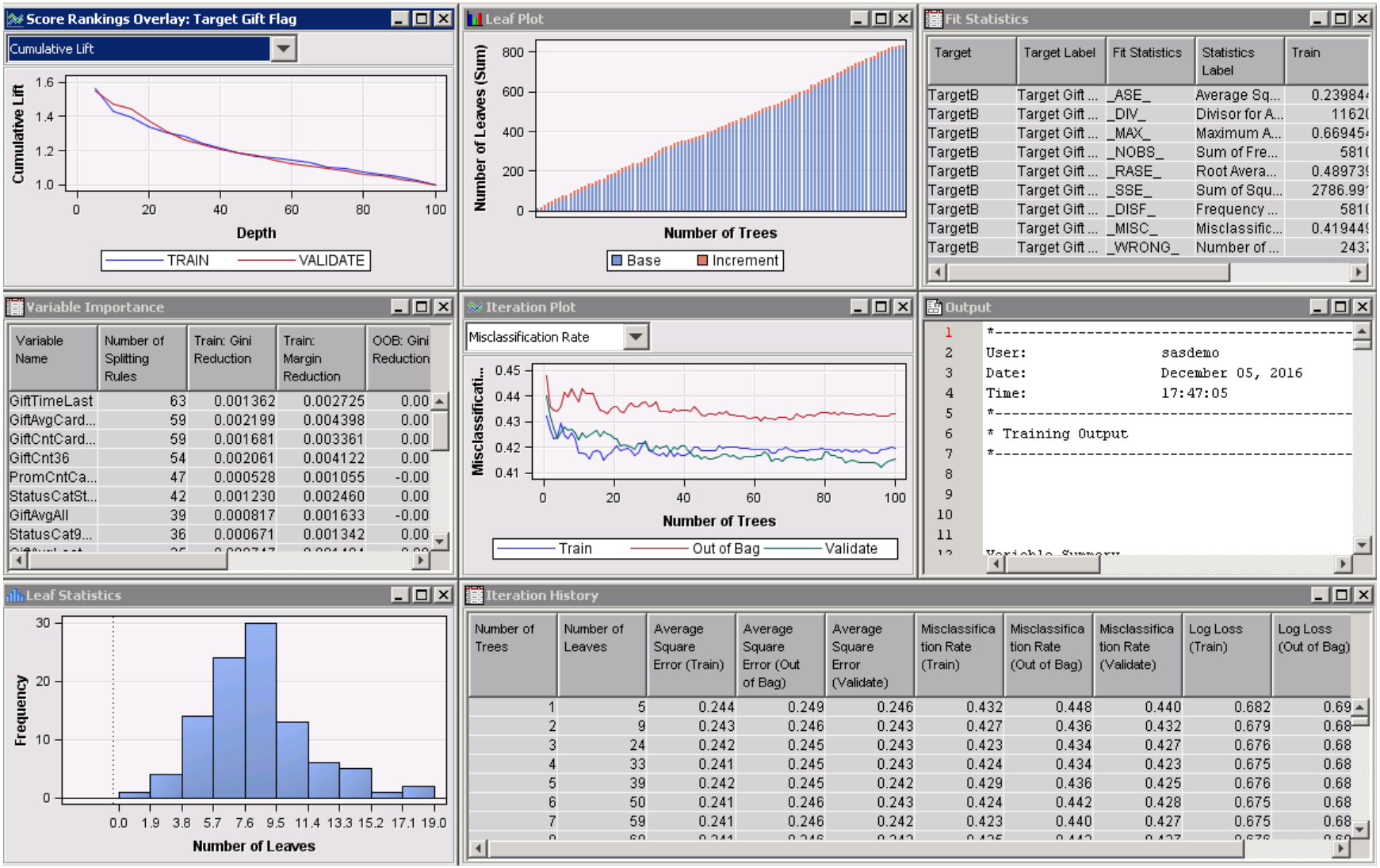
Ćwiczenie 3 c)

Zbudować model lasu losowego

Pytania kontrolne:

- Jak dużo z utworzonych drzew ma 10 lub 11 liści?
- Która ze zmiennych była najczęściej wykorzystywana w podziałach drzew?

- Zdecydować, w którym miejscu diagramu umieścić model lasu losowego
  - Nie zmieniać parametrów węzła Forest


Score Rankings Overlay: Target Gift Flag
Cumulative Lift
Cumulative Lift
Depth
TRAIN
VALIDATE
Leaf Plot
Number of Leaves (Sum)
Number of Trees
Base
Increment
Fit Statistics
Target
Target Label
Fit Statistics
Statistics Label
Train
TargetB
Target Gift ...
ASE
Average Sq...
DIV
Divisor for A...
MAX
Maximum A...
NOBS
Sum of Fre...
RASE
Root Avera...
SSE
Sum of Squ...
DISF
Frequency ...
MISC
Misclassific...
WRONG
Number of ...
Variable Importance
Variable Name
Number of Splitting Rules
Train: Gini Reduction
Train: Margin Reduction
OOB: Gini Reduction
GiftTimeLast
GiftAvgCard...
GiftCntCard...
GiftCnt36
PromCntCa...
StatusCatSt...
GiftAvgAll
StatusCat9...
Iteration Plot
Misclassification Rate
Misclassificati...
Number of Trees
Train
Out of Bag
Validate
Output
User: sasdemo
Date: December 05, 2016
Time: 17:47:05
* Training Output
Leaf Statistics
Frequency
Number of Leaves
Iteration History
Number of Trees
Number of Leaves
Average Square Error (Train)
Average Square Error (Out of Bag)
Average Square Error (Validate)
Misclassification Rate (Train)
Misclassification Rate (Out of Bag)
Misclassification Rate (Validate)
Log Loss (Train)
Log Loss (Out of Bag)

Ćwiczenie 3 d)

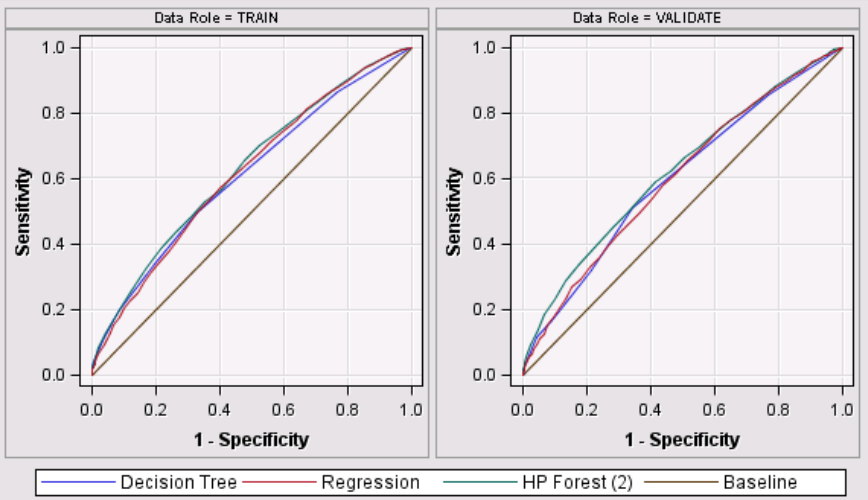
Porównać zbudowane modele

Pytania kontrolne:

- Który z modeli jest najlepszy wg zdefiniowanego kryterium?
- Jakie inne 2 statystyki wskazują, że model jest najlepszy?

- Zastosować węzeł Model Comparison
- Podłączyć do niego wszystkie 3 utworzone modele
  - Określić kryterium wyboru modelu:
    - Lift skumulowany na 10% listy zbioru walidacyjnego

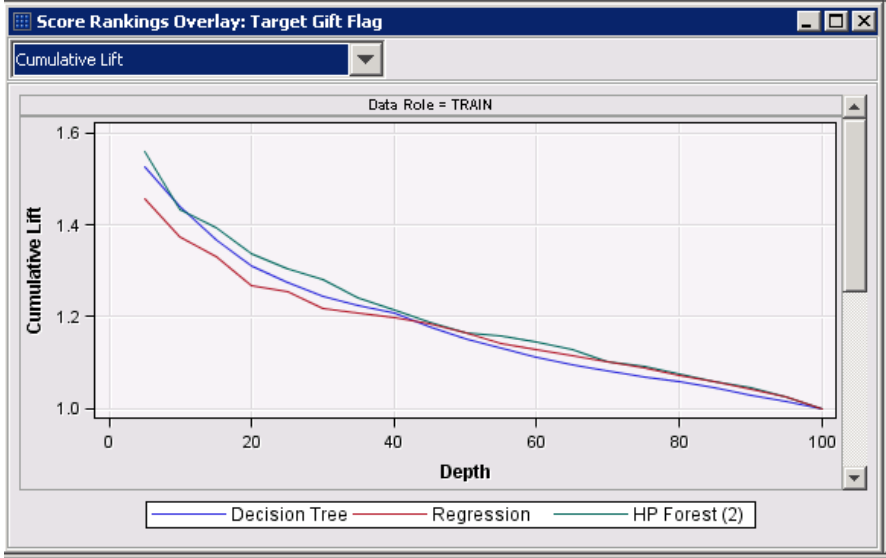
ROC Chart : Target Gift Flag



Fit Statistics

Selected Model	Predecessor Node	Model Node	Model Description	Target Variable	Target Label	Selection Criterion: Valid: Cumulative Lift	Train: Averag Square Error
Y	HPDMFore...	HPDMFore...	HP Forest (2)	TargetB	Target Gift ...	1.469072	0.2
	Tree	Tree	Decision Tr...	TargetB	Target Gift ...	1.349234	0.
	Reg	Reg	Regression	TargetB	Target Gift ...	1.309278	0.2

Score Rankings Overlay: Target Gift Flag



Output

```
*------------------------------------------------------------*
User:                  sasdemo
Date:                  December 05, 2016
Time:                  18:36:09
*------------------------------------------------------------*
* Training Output
*------------------------------------------------------------*



Variable Summary

          Measurement     Frequency
 Role        Level          Count

TARGET      BINARY            1
```

Ćwiczenie 3 e)

Wygenerować kod scoringowy

- Zastosować węzeł Score dla najlepszego modelu zidentyfikowanego w poprzednim węźle

SAS Code

```
1  *-------------------------------------------------------------*;
2  * EM SCORE CODE;
3  *-------------------------------------------------------------*;
4  *-------------------------------------------------------------*;
5  * TOOL: Input Data Source;
6  * TYPE: SAMPLE;
7  * NODE: Ids2;
8  *-------------------------------------------------------------*;
9  *-------------------------------------------------------------*;
10 * TOOL: Partition Class;
11 * TYPE: SAMPLE;
12 * NODE: Part2;
13 *-------------------------------------------------------------*;
14 *-------------------------------------------------------------*;
15 * TOOL: Extension Class;
16 * TYPE: MODEL;
17 * NODE: HPDMForest4;
18 *-------------------------------------------------------------*;
19 %macro em_hpfst_score;
```

Output Variables

Variable Name	Creator	Variable Label	Function	Type
EM_CLASSIFICATI...	Score	Prediction for Targ...	CLASSIFICATION	C
EM_EVENTPROB...	Score	Probability for level...	PREDICT	N
EM_PROBABILITY	Score	Probability of Clas...	PREDICT	N
EM_SEGMENT	Score	Segment	TRANSFORM	N
I_TargetB	HPDMForest4	Into: TargetB	CLASSIFICATION	C
P_TargetB0	HPDMForest4	Predicted: TargetB...	PREDICT	N
P_TargetB1	HPDMForest4	Predicted: TargetB...	PREDICT	N
U_TargetB	HPDMForest4	Unnormalized Into...	CLASSIFICATION	N
WARN	HPDMForest4	Warnings	ASSESS	C
b_TargetB	MdlComp2		TRANSFORM	N

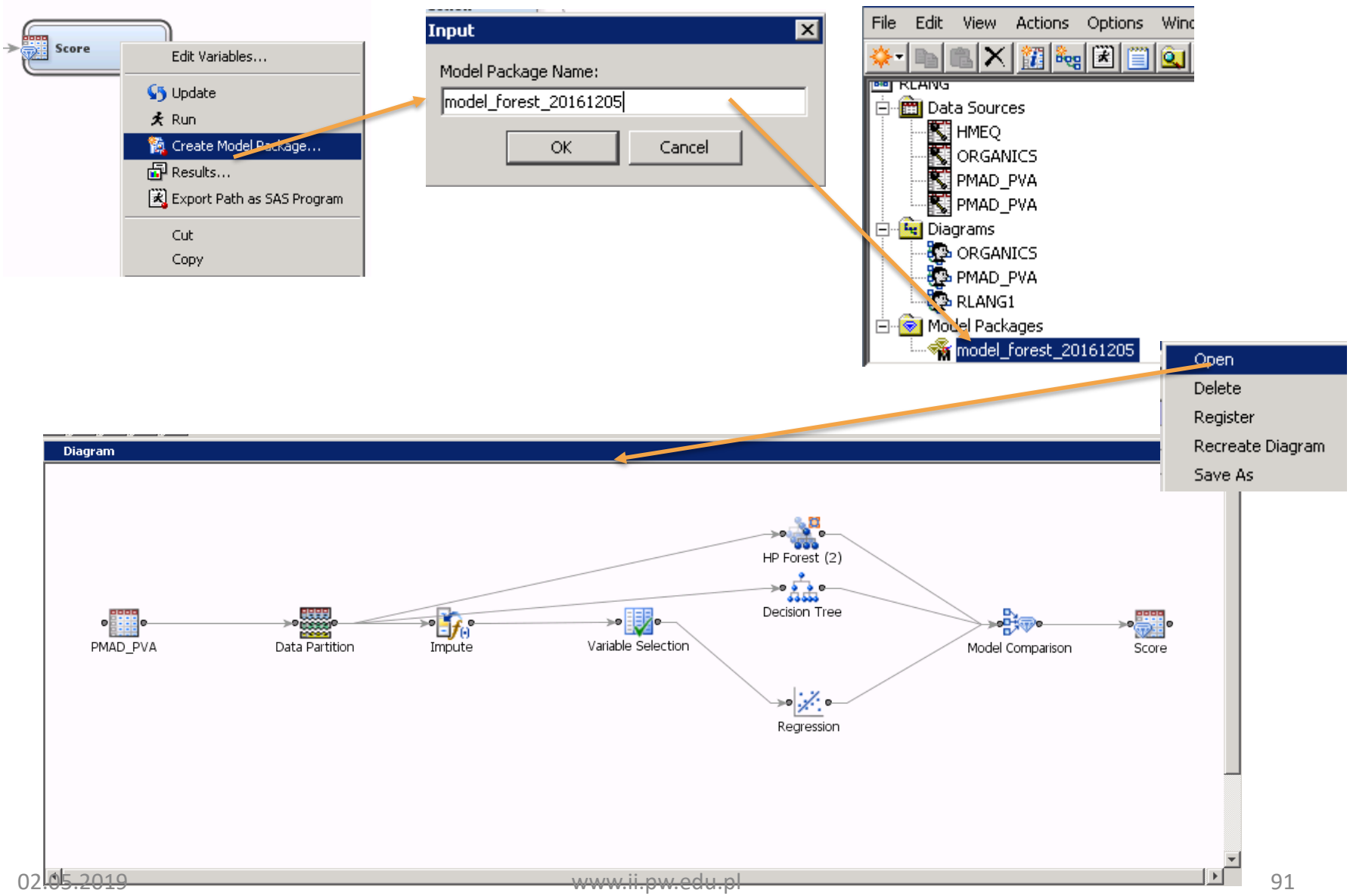
Output

```
1  *-------------------------------------------------------------*
2  User:                  sasdemo
3  Date:                  December 05, 2016
4  Time:                  18:39:18
5  *-------------------------------------------------------------*
6  * Training Output
7  *-------------------------------------------------------------*
8
9
10
11
12 Variable Summary
13
14              Measurement     Frequency
15  Role          Level           Count
16
17 SEGMENT       NOMINAL             1
18 TARGET        BINARY              1
19
```

Ćwiczenie 3 f)

Utworzyć pakiet modelu

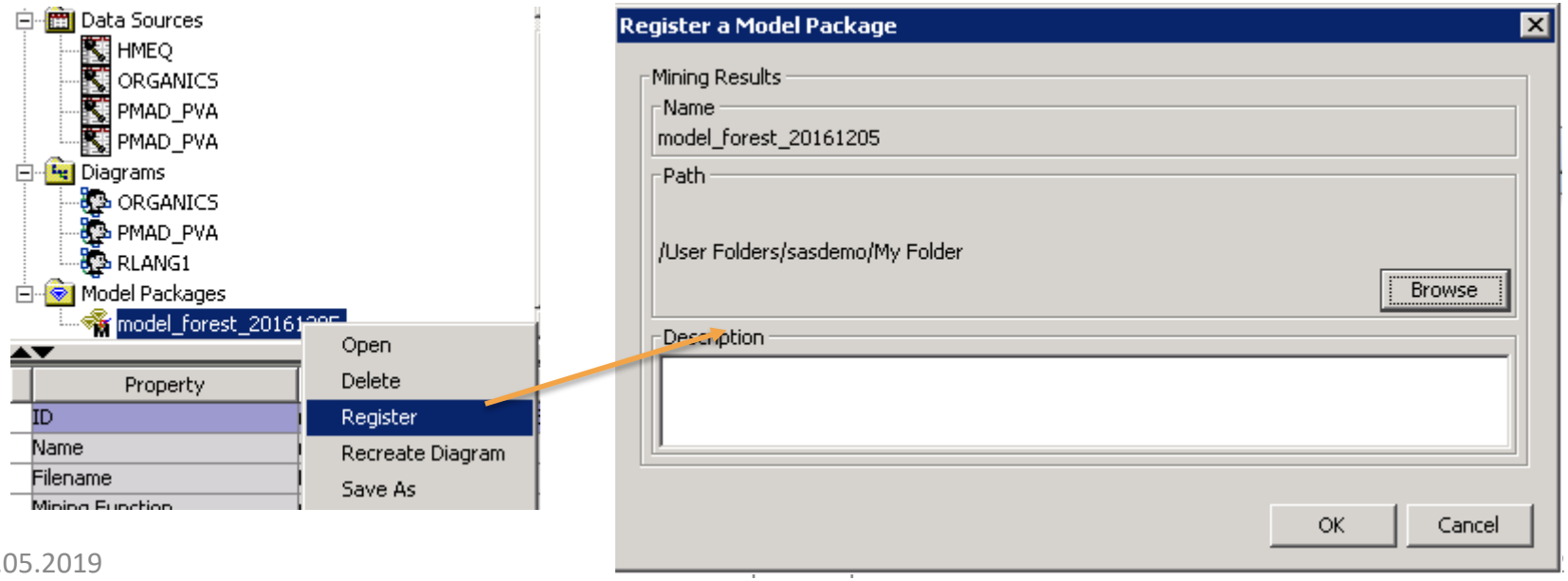
- Pakiet modeli pozwala na zachowanie ścieżki modelowania
- Zabezpiecza przed przypadkową zmianą parametrów lub ponownym przeliczeniem
- Tak przygotowany pakiet może być przenoszony między środowiskami


Score
Edit Variables...
Update
Run
Create Model Package...
Results...
Export Path as SAS Program
Cut
Copy
Input
Model Package Name:
model_forest_20161205
OK
Cancel
File Edit View Actions Options
Data Sources
HMEQ
ORGANICS
PMAD_PVA
PMAD_PVA
Diagrams
ORGANICS
PMAD_PVA
RLANG1
Model Packages
model_forest_20161205
Open
Delete
Register
Recreate Diagram
Save As
Diagram
HP Forest (2)
Decision Tree
PMAD_PVA
Data Partition
Impute
Variable Selection
Model Comparison
Score
Regression

Ćwiczenie 3 g)

Zarejestrować model w metadanych

- Rejestracja w metadanych pozwala na współdzielenie modelu między różnymi aplikacjami SAS



Ćwiczenie 3

- Diagram wynikowy