Metody przetwarzania dużych ilości danych

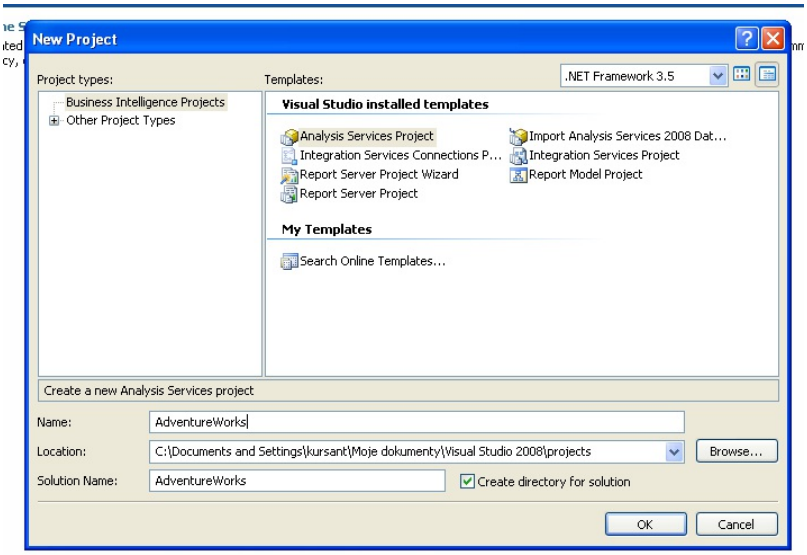
Sprawozdanie z laboratorium

Data	Tytuł zajęć	Uczestnicy
15.12.2020 18:55	Data Mining	Adam Hyjek (234987) Bartosz Rodziewicz (226105)

Streszczenie laboratorium

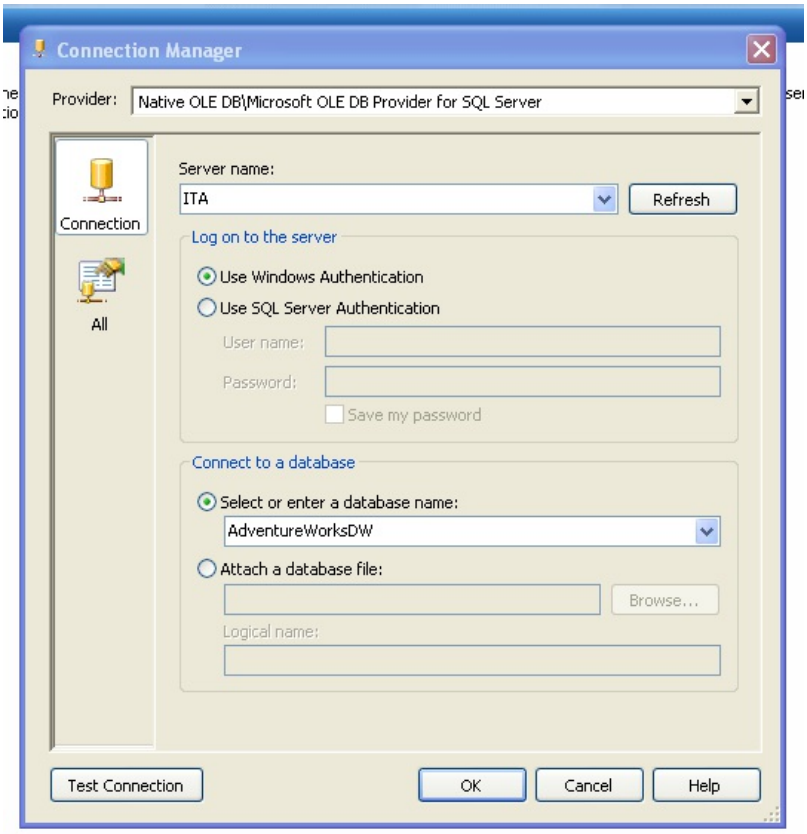
Problem 1

Stworzenie nowego projektu Analysis Services w SS BIDS

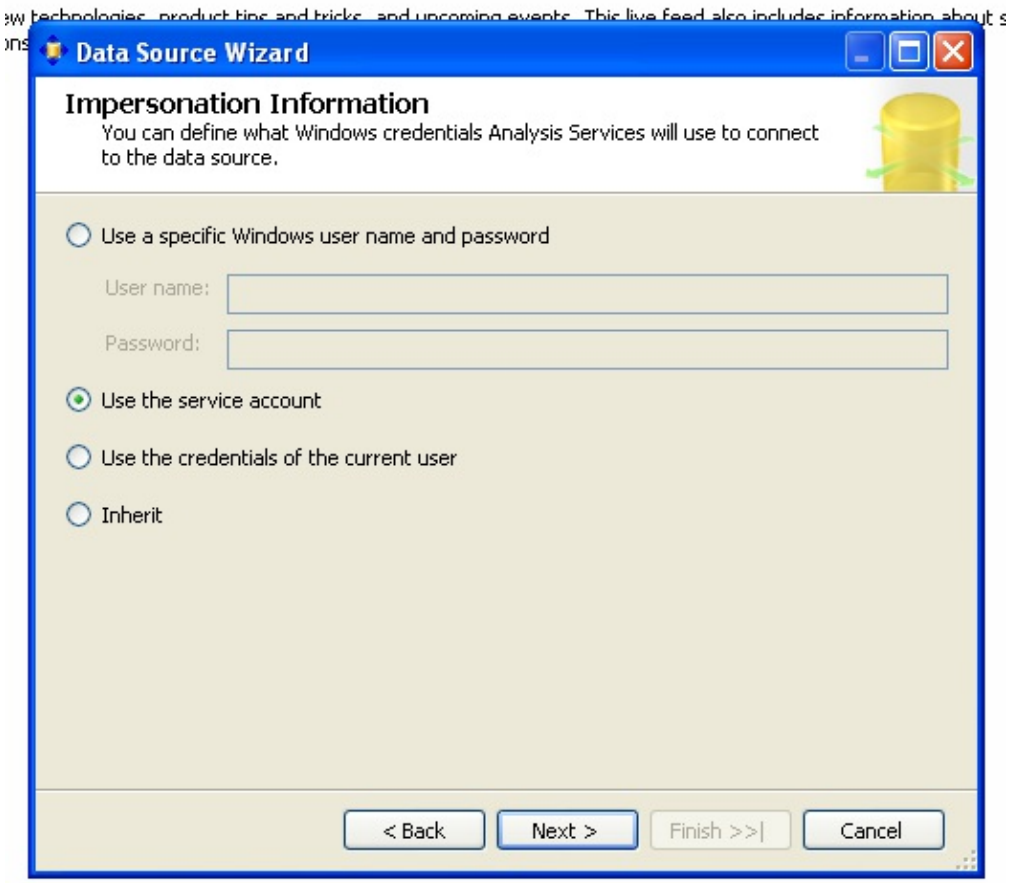


Stworzenie nowego projektu.

Stworzenie źródła danych

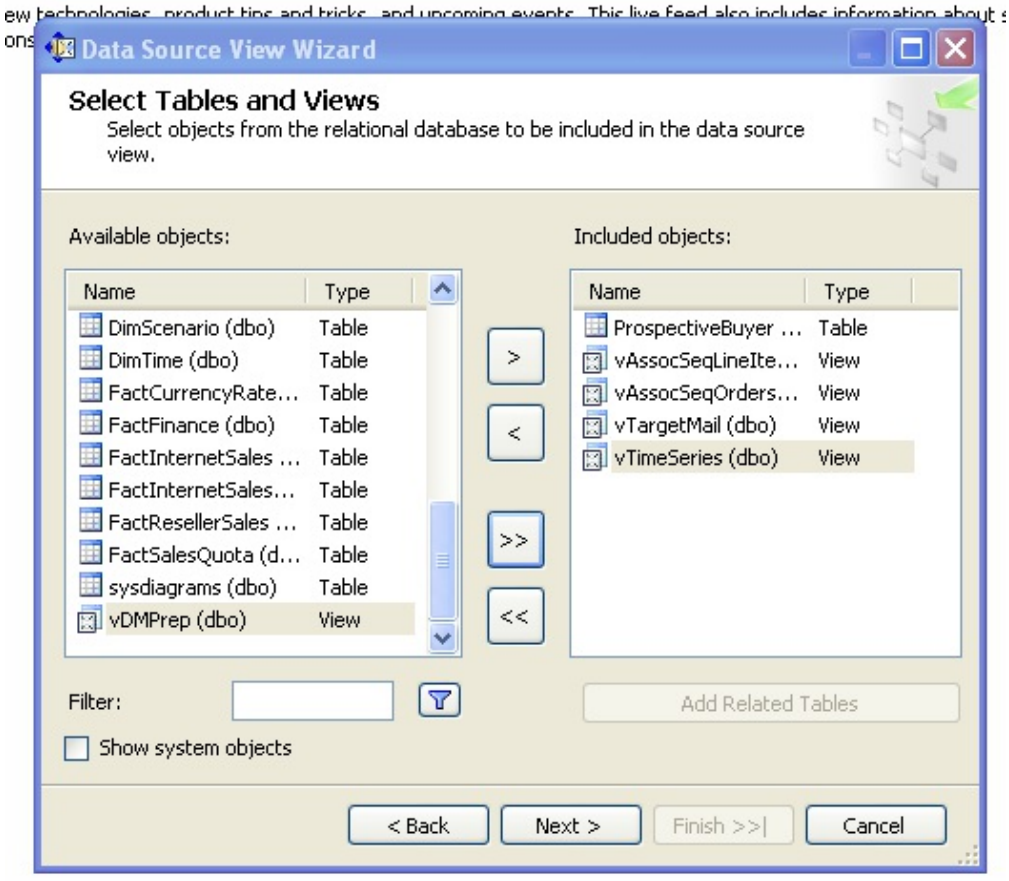


Stworzenie połączenia z bazą danych (różnica do instrukcji, serwer to nie localhost, a ITA).

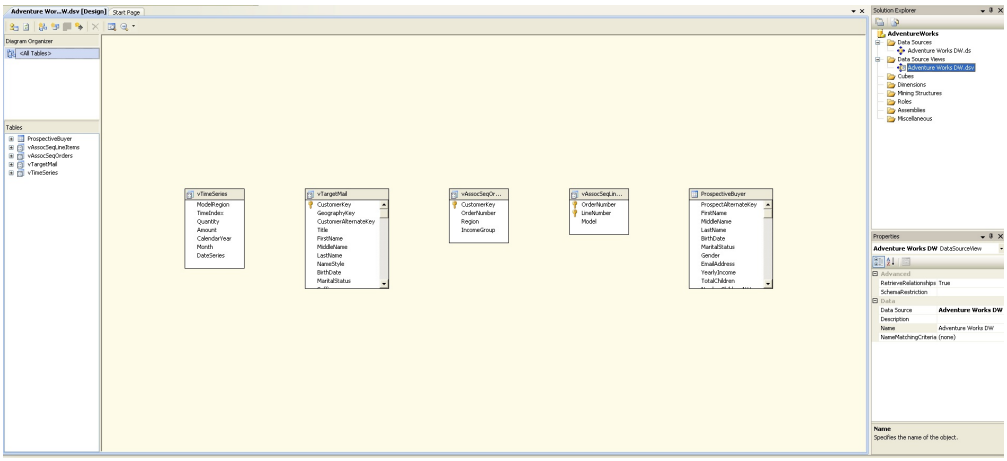


Wybór metody logowania.

Stworzenie widoku źródła danych

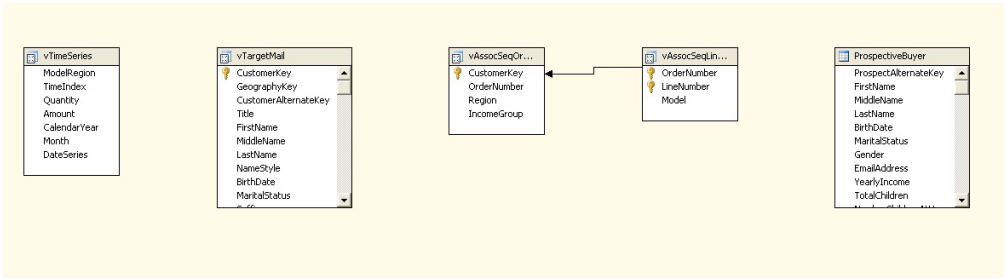


Wybór tabel do widoku źródła danych.



Nowy widok źródła danych.

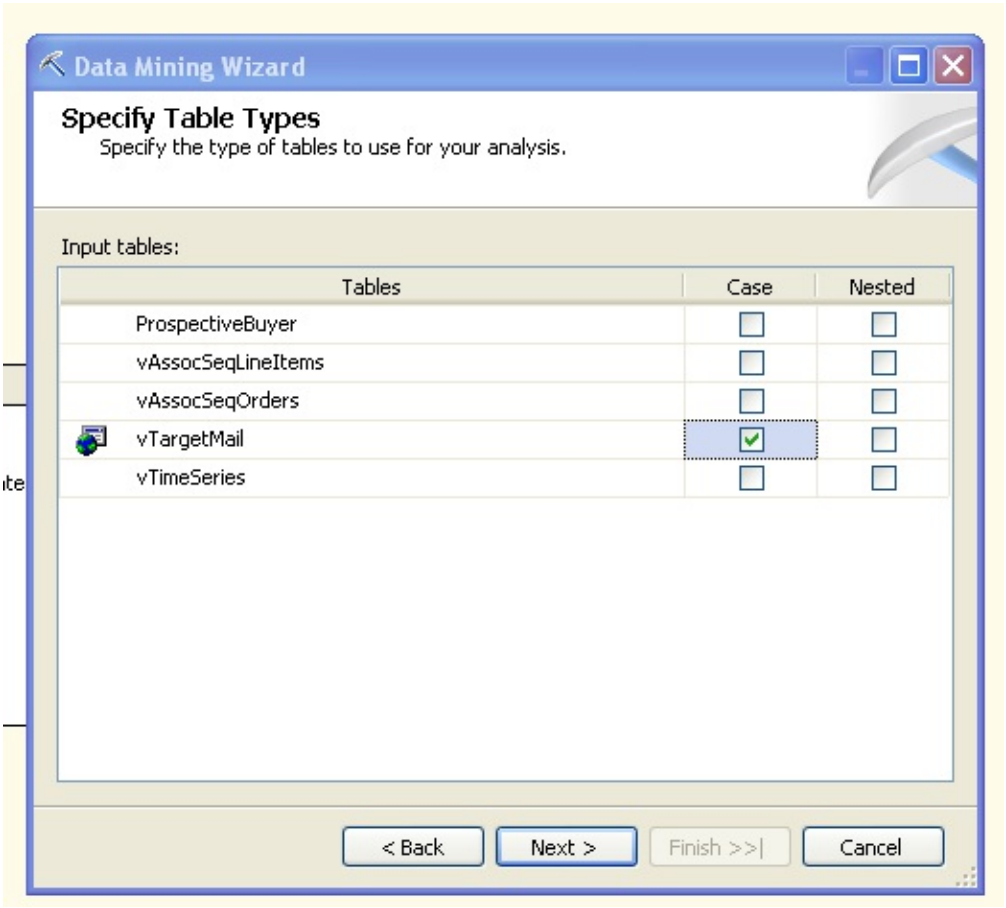
Modyfikacja widoku źródła danych



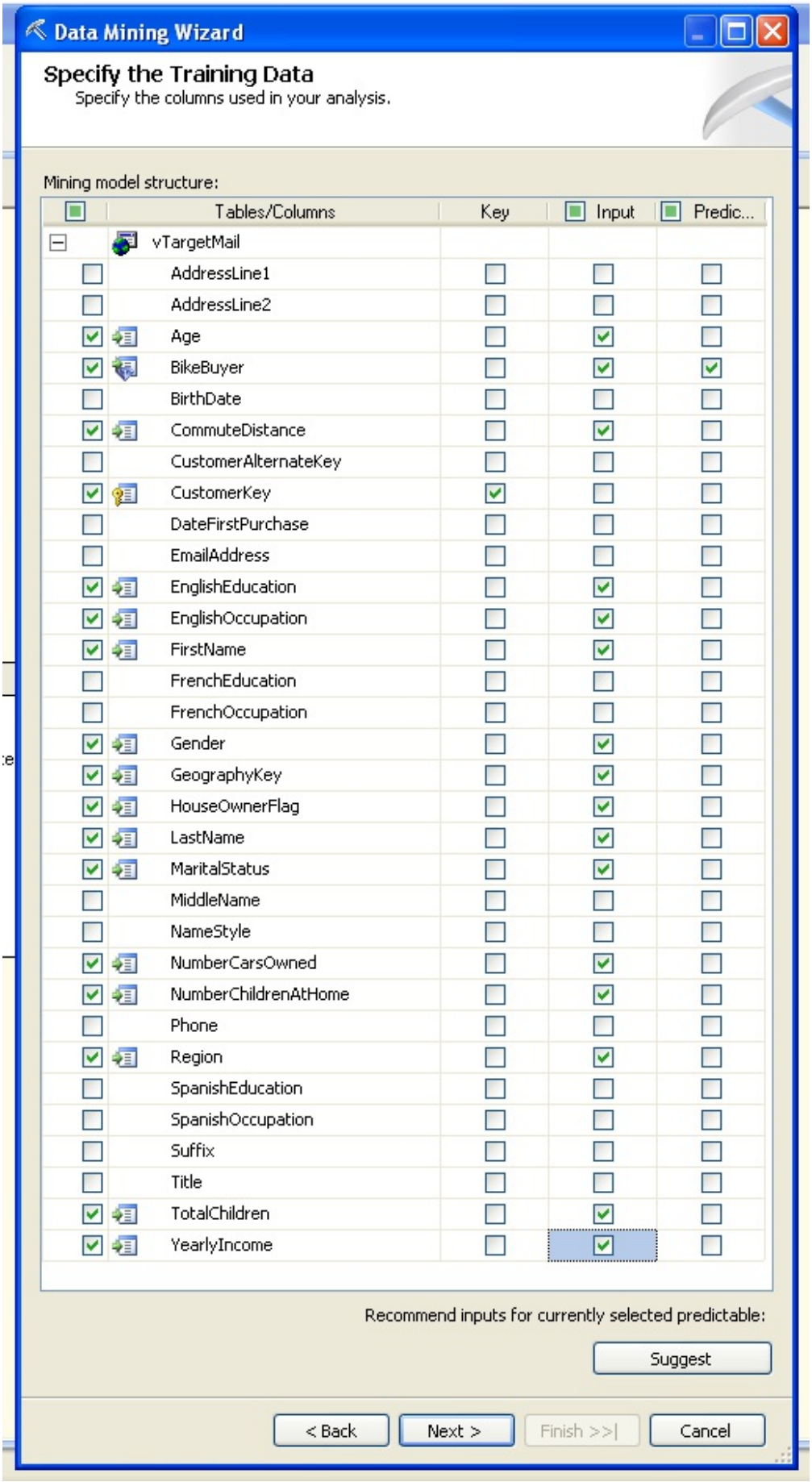
Utworzenie nowego związku między tabelami.

Problem 2

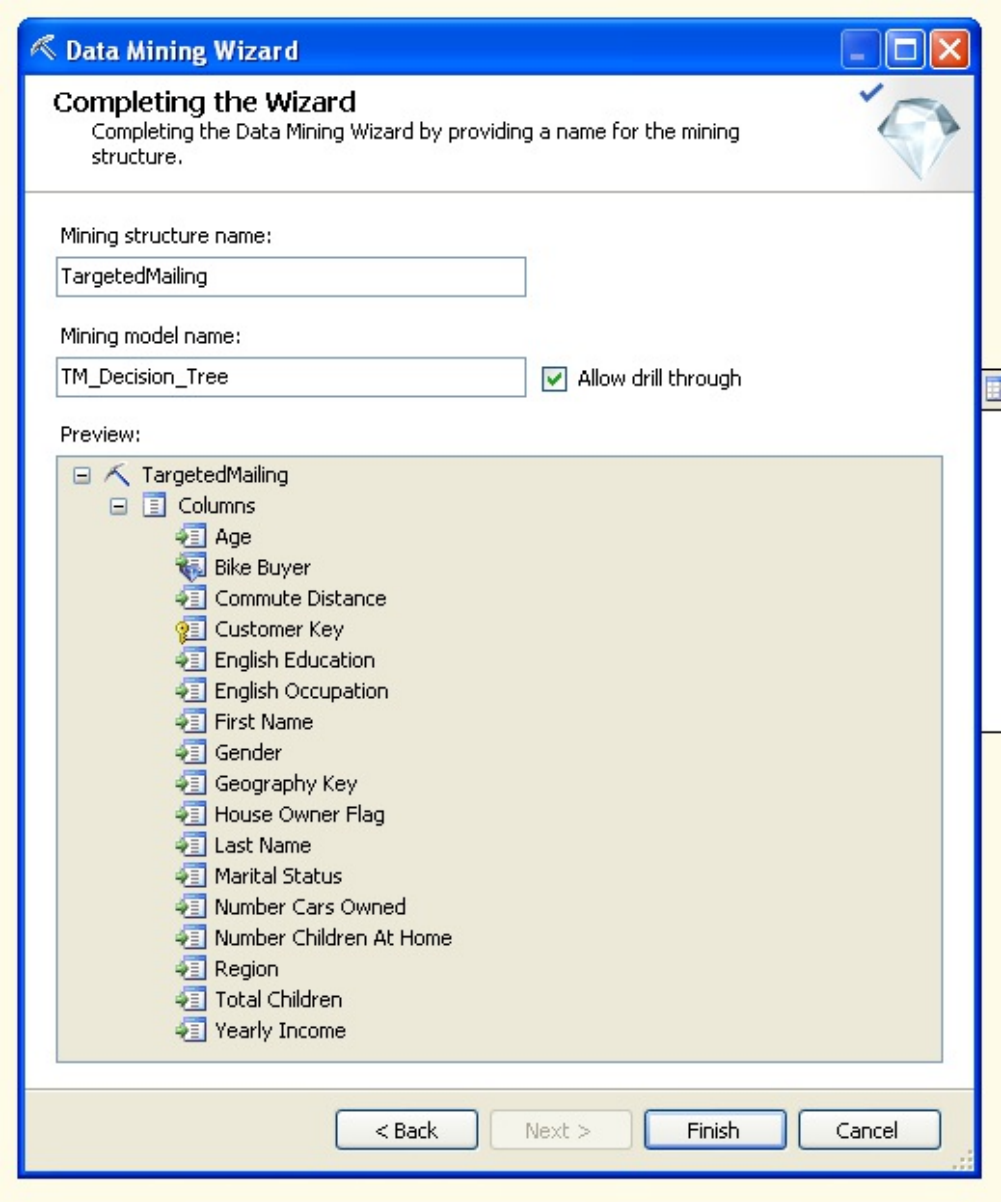
Stworzenie struktury eksploracji danych dla modelu Targeted Mailing



Wybór tabeli do eksploracji.

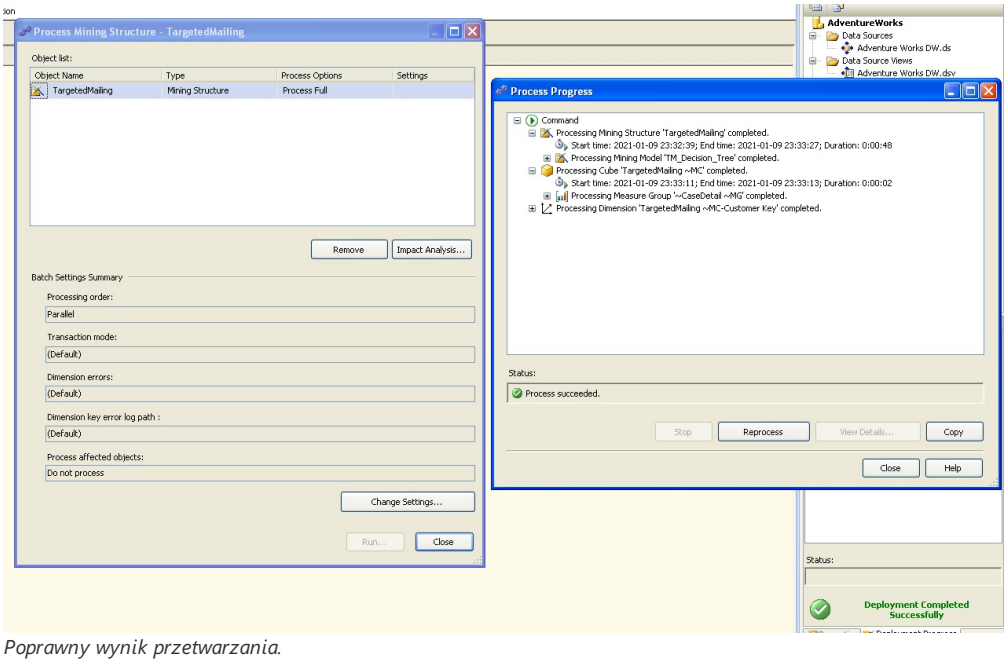


Wybór danych treningowych.



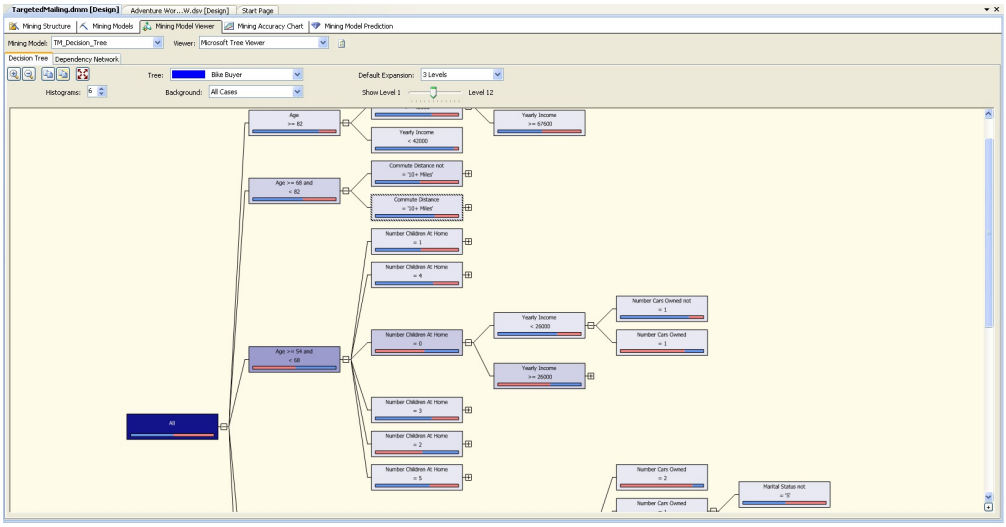
Podsumowanie tworzenia struktury eksploracji danych.

Przetwarzanie modelu ekstrakcji danych



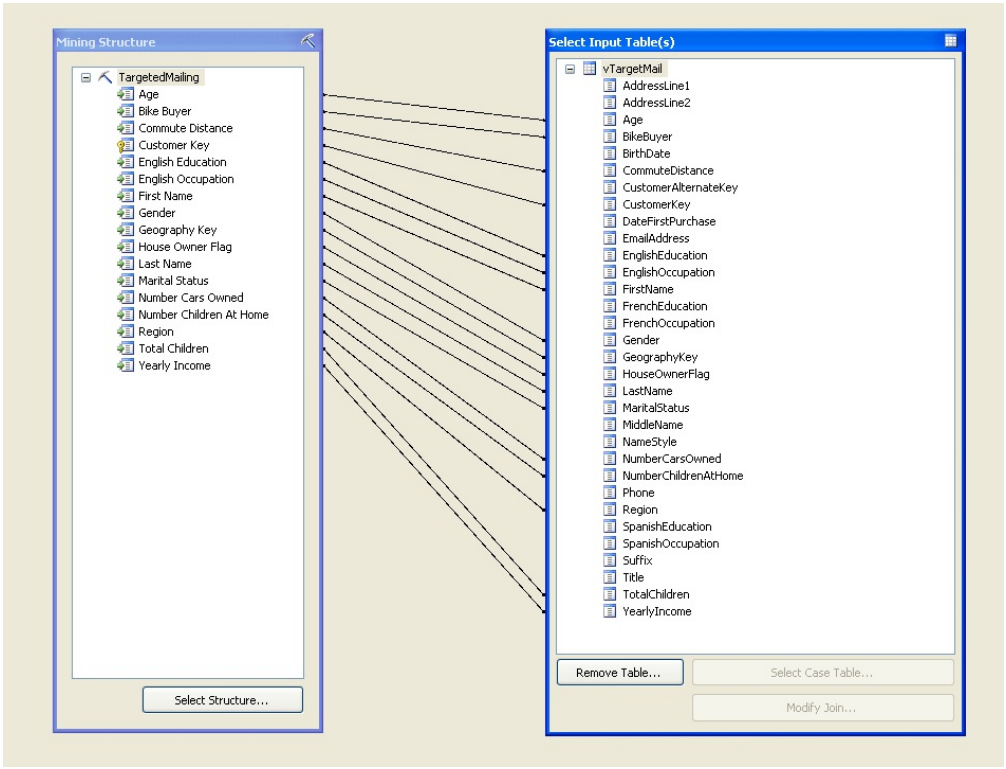
Poprawny wynik przetwarzania.

Analiza modelu opartego o drzewa decyzyjne



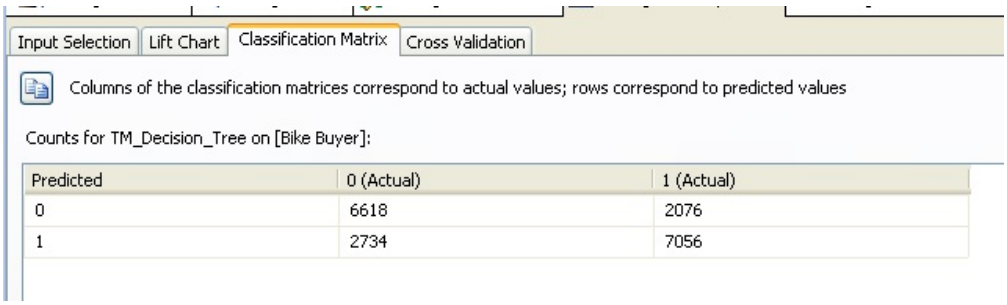
Wygląd widoku analizy modelu.

Mapowanie kolumn



Mapowanie kolumn.

Macierz klasyfikacji

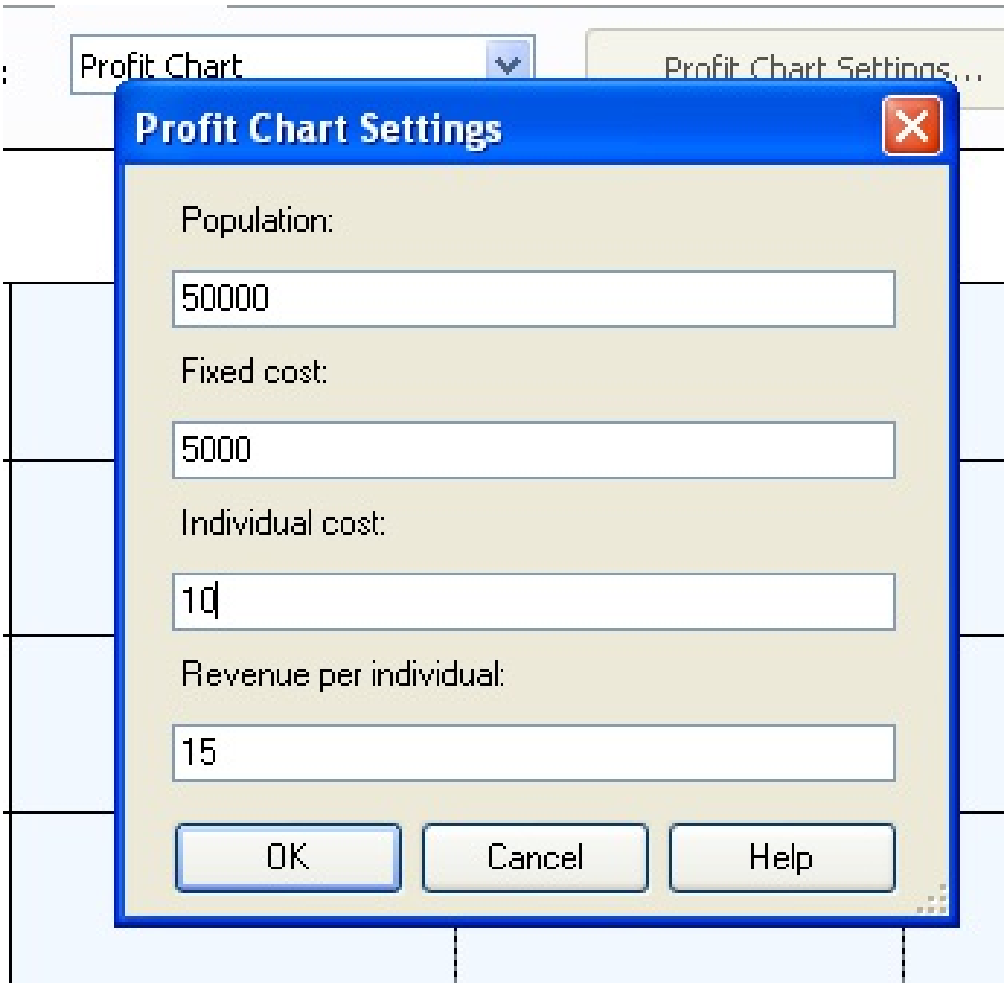


Wyniki macierzy klasyfikacji.

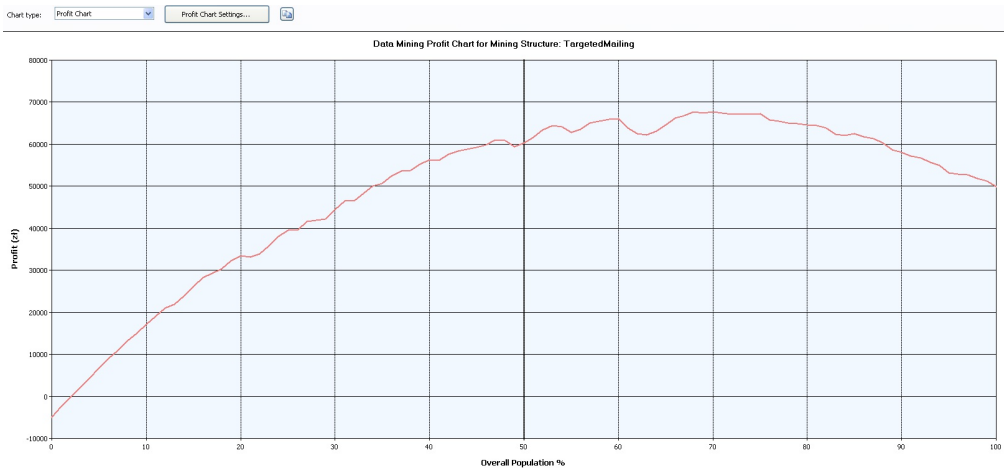
Jaki jest całkowity błąd predykcji drzewa decyzyjnego, a także jaka część osób odrzuci ofertę z kampanii?

Wskaźnik	Działanie	Wynik
Całkowity błąd predykcji	(2734+2076)/(6618+2734+2076+7056) = 4810/18484 = ~0.26	26%
Jaka część osób odrzuci ofertę	2734/(2734+7056) = 1367/4895 = ~0.279	27.9%

Wykres przewidywanego zysku



Wybór parametrów wykresu.

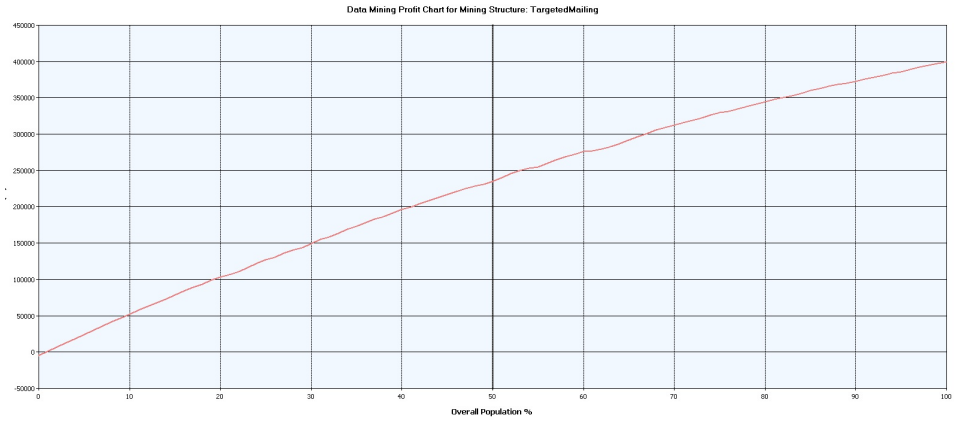


Wykres dla indywidualnego kosztu równego 10.

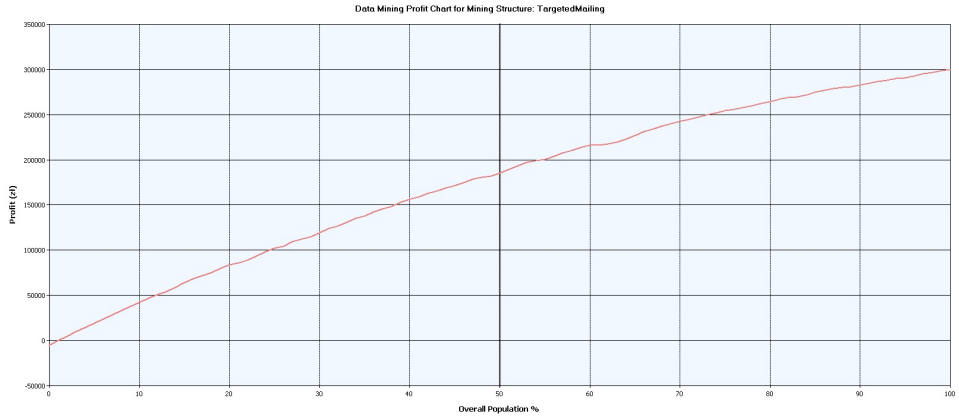
Jak można zinterpretować otrzymany wykres?

Wykres mówi o tym jakiego zysku można się spodziewać wysyłając kampanie reklamową do wybranego procentu populacji uwzględniając całkowity koszt kampanii, koszt kampanii na osobę, możliwy zysk od osoby oraz prawdopodobieństwo, czy dana osoba zdecyduje się na zakupy.

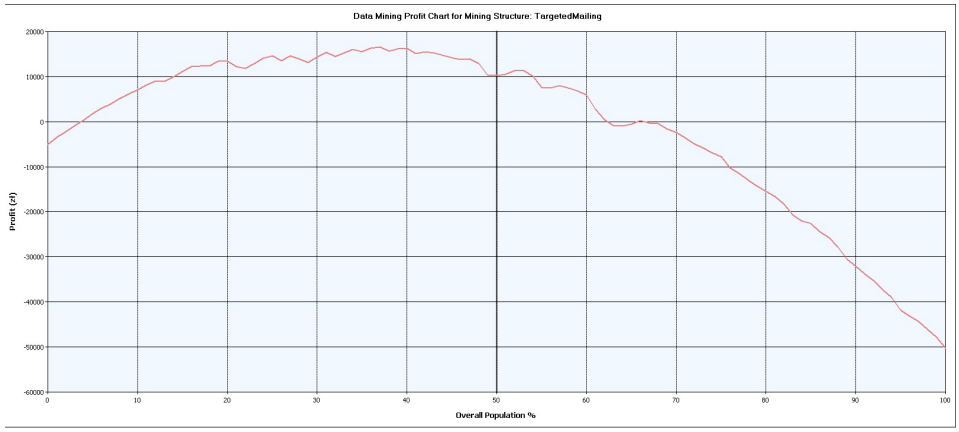
Zbadaj kształt wykresu dla różnych wartości



Wykres dla indywidualnego kosztu równego 3.



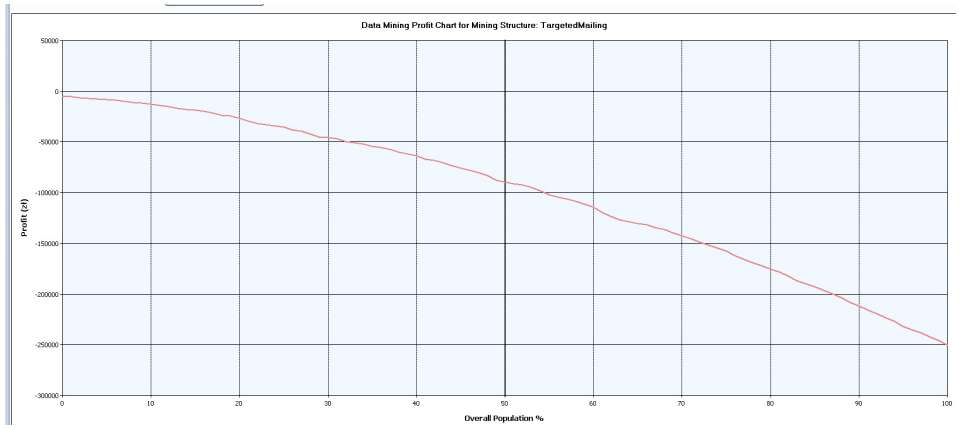
Wykres dla indywidualnego kosztu równego 5.



Wykres dla indywidualnego kosztu równego 12.

Im wyższy koszt wykonania kampanii na osobę (przy takim samym potencjalnym zysku), tym bardziej się opłaca wysłać kampanie do mniejszej ilości, lepiej wyselekcjonowanych ludzi, przy których prawdopodobieństwo sukcesu jest większe.

Dlaczego przy ustawieniu Individual Cost powyżej wartości 15 wykres zysku nie posiada wartości dodatnich?



Wykres dla indywidualnego kosztu równego 16.

Wykres nie posiada wartości dodatnich, ponieważ potencjalny zysk z osoby, decydującej się na zakupy, jest niższy niż koszt wykonania kampanii.

Zmiany kolumn w modelu

Counts for TM_Decision_Tree on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	5938	2239
1	3414	6893

Nowa wartość macierzy klasyfikacji.

Oblicz na nowo całkowity błąd predykcji i błąd dla kupna roweru. Czy otrzymane wyniki są lepsze czy gorsze od poprzednio uzyskanych?

Wskaźnik	Działanie	Wynik
Całkowity błąd predykcji	(3414+2239)/(5938+3414+2239+6893) = 5653/18484 = ~0.3058	30.6%
Jaka część osób odrzuci ofertę	3414/(3414+6893) = 3414/10307 = ~0.33	33%

Wyniki są gorsze od poprzednich.

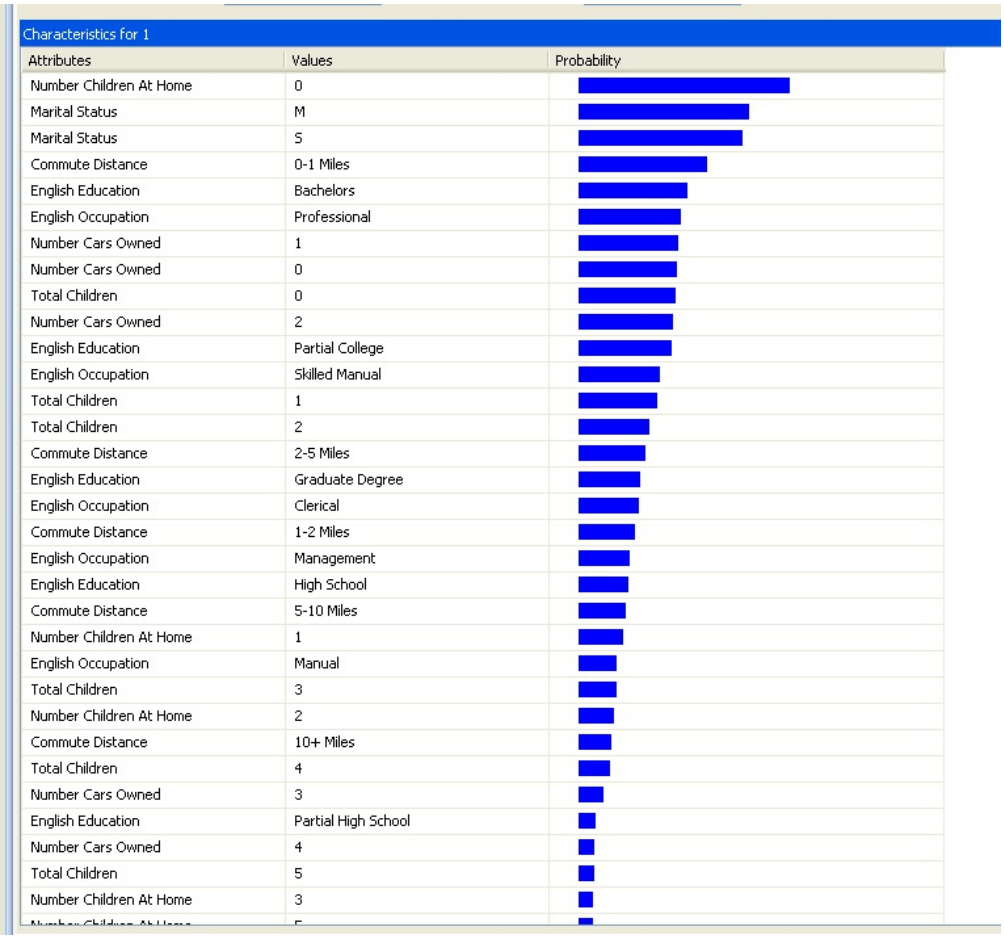
Problem 2 – kontynuacja modułu 1

Algorytm naiwny Bayesowski

Structure	TM_Decision_Tree	TM_Bayes
	Microsoft_Decision_Trees	Microsoft_Naive_Bayes
Age	Input	Ignore
Bike Buyer	Predict	Predict
Commute Distance	Input	Input
Customer Key	Key	Key
English Education	Input	Input
English Occupation	Input	Input
First Name	Ignore	Ignore
Gender	Input	Input
Geography Key	Input	Ignore
House Owner Flag	Input	Input
Last Name	Ignore	Ignore
Marital Status	Input	Input
Number Cars Owned	Input	Input
Number Children At Home	Input	Input
Region	Ignore	Ignore
Total Children	Input	Input
Yearly Income	Input	Ignore

Nowy algorytm Bayesowski.

Wpływ poszczególnych atrybutów na wyniki.



Wpływ poszczególnych wartości atrybutów na wyniki.

Czy aktualnie stworzony model jest lepszy czy gorszy od drzew decyzyjnych?

Counts for TM_Bayes on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	5912	3621
1	3440	5511

Wynik macierzy klasyfikacji.

Algorytm jest gorszy.

Zmienić kolumny wejściowe dla tego modelu, czy można w ten sposób poprawić wyniki dla tego algorytmu?

Counts for TM_Bayes on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6073	3484
1	3279	5648

Wynik macierzy klasyfikacji po dodaniu poprzednio wyłączonych atrybutów (imię, nazwisko, region).

Wynik delikatnie się poprawił, jednak wciąż jest słabszy niż algorytm na podstawie drzew decyzyjnych.

Algorytm oparty o Sztuczne Sieci Neuronowe

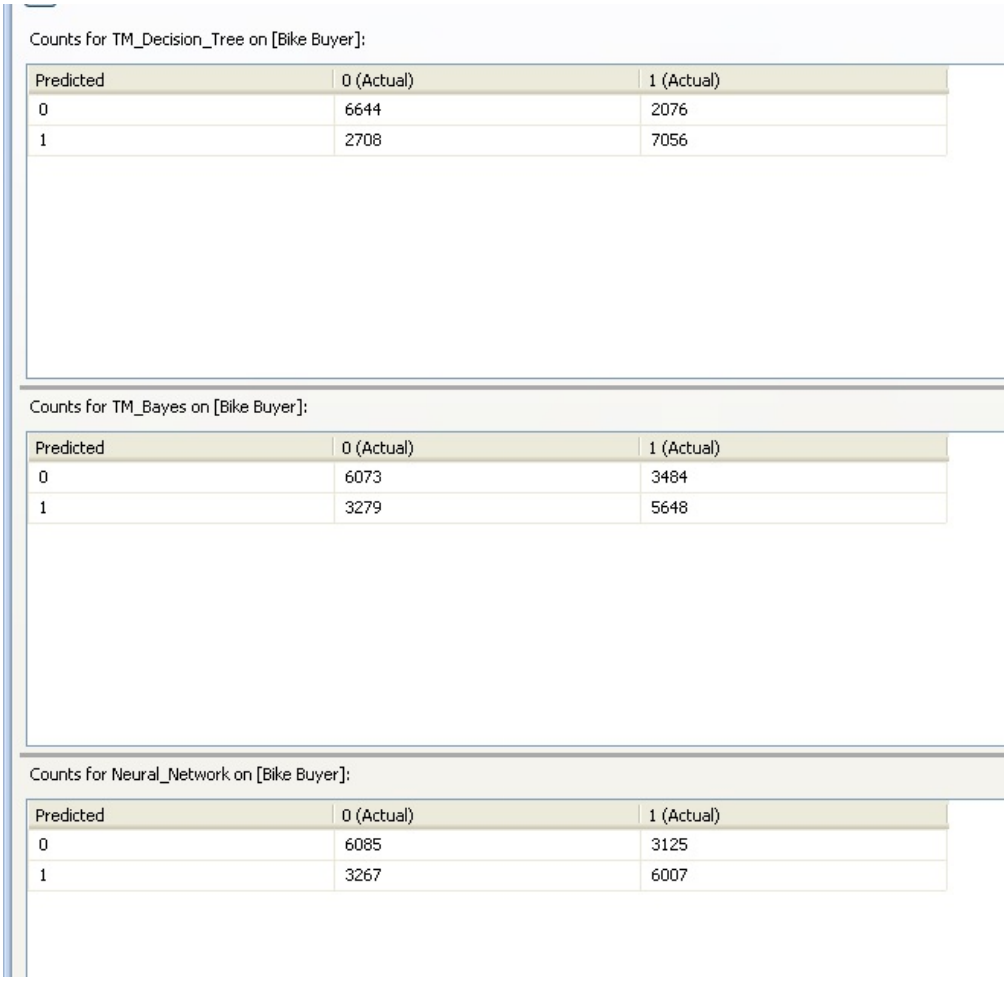
Structure	TM_Decision_Tree	TM_Bayes	Neural_Network
	Microsoft_Decision_Trees	Microsoft_Naive_Bayes	Microsoft_Neural_Network
Age	Input	Ignore	Input
Bike Buyer	Predict	Predict	Predict
Commute Distance	Input	Input	Input
Customer Key	Key	Key	Key
English Education	Input	Input	Input
English Occupation	Input	Input	Input
First Name	Input	Input	Input
Gender	Input	Input	Input
Geography Key	Input	Ignore	Input
House Owner Flag	Input	Input	Input
Last Name	Input	Input	Input
Marital Status	Input	Input	Input
Number Cars Owned	Input	Input	Input
Number Children At Home	Input	Input	Input
Region	Input	Input	Input
Total Children	Input	Input	Input
Yearly Income	Input	Ignore	Input

Nowy algorytm SSN.

Jakie kolumny nie mogą być użyte przez sieć neuronową?

Wszystkie dostępne kolumny mogą być użyte poprzez sieć neuronową.

Porównaj opracowane modele. Które kolumny można zignorować?



Counts for TM_Decision_Tree on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6644	2076
1	2708	7056

Counts for TM_Bayes on [Bike Buyer]:

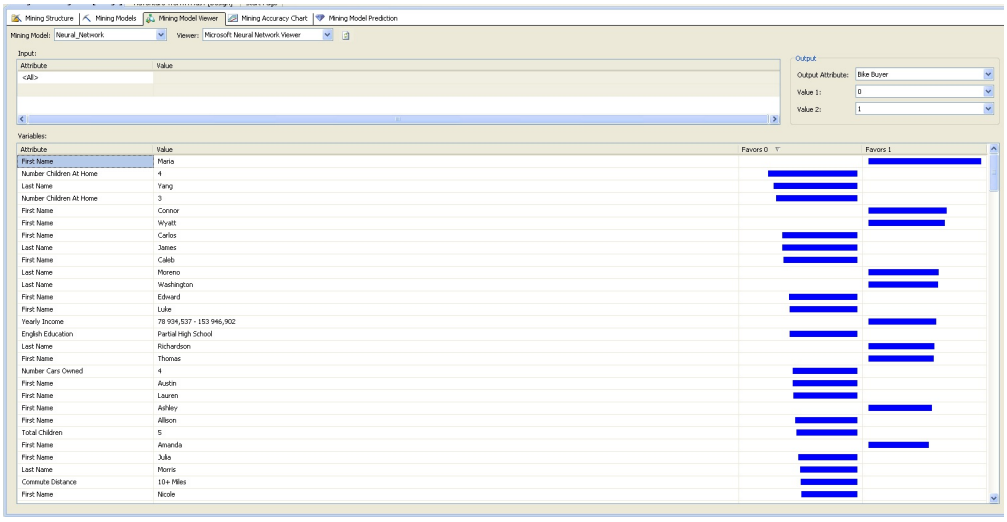
Predicted	0 (Actual)	1 (Actual)
0	6073	3484
1	3279	5648

Counts for Neural_Network on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6085	3125
1	3267	6007

Wynik macierzy klasyfikacji dla wszystkich argumentów.

Powyżej widać, że algorytm SSN jest trochę lepszy niż Bayesowski, jednak gorszy niż drzewa decyzyjne.



Analiza modelu.

Analiza modelu pokazuje, że duży wpływ na dane mają kolumny First and Last Name. Oczywistym jest, że te kolumny nie prezentują żadnych użytecznych informacji, więc można je wyłączyć.

Counts for TM_Decision_Tree on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6644	2076
1	2708	7056

Counts for TM_Bayes on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6014	3476
1	3338	5656

Counts for Neural_Network on [Bike Buyer]:

Predicted	0 (Actual)	1 (Actual)
0	6521	3092
1	2831	6040

Wynik macierzy klasyfikacji po wyłączeniu parametrów First and Last Name.

Powyżej widać, że wyniki delikatnie uległy poprawie, jednak nie jest to znaczący wynik.

Zmień kolumny wejściowe dla tego modelu, czy można w ten sposób poprawić wyniki dla tego algorytmu?

Kilkukrotna próba zmiany kolumn wejściowych nie spowodowała poprawy wyników.