LABORATORIUM 6 – Oracle Data Miner –Klasyfikacja c.d. i Regresja

Cel ćwiczenia:

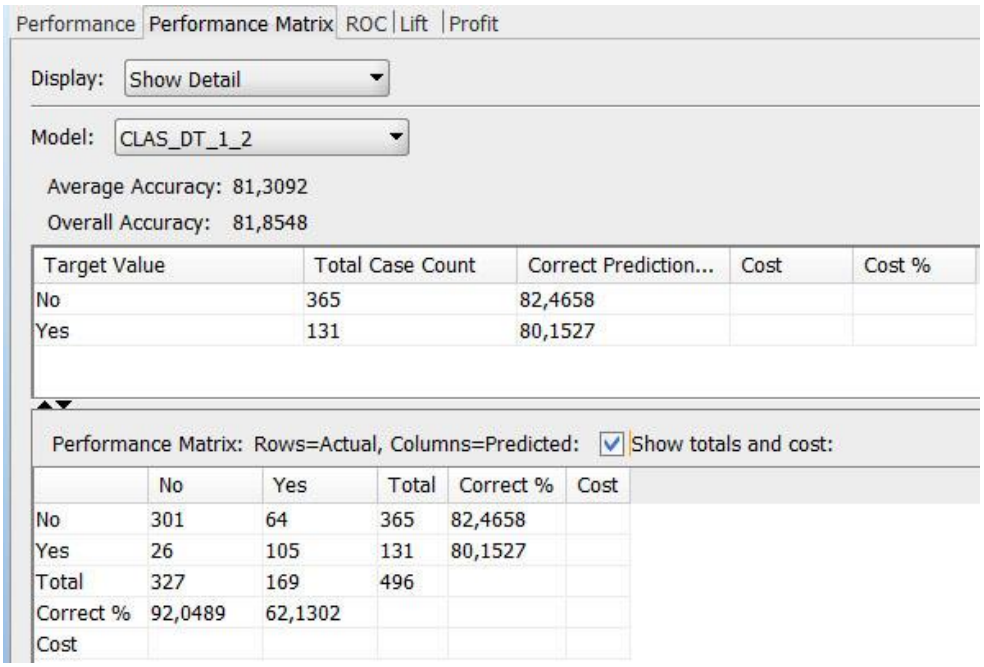
1. Wykorzystania Oracle Data Miner do analizy danych w oparciu o Klasyfikację
2. Wykorzystania Oracle Data Miner do analizy danych w oparciu o Regresję
3. Ocena jakości klasyfikacji i regresji

Środki:

1. Serwer baz danych Oracle Eneterprise Edition 11g.
2. Aplikacja SQL Devloper.
3. Strona internetowa:
    a. http://www.oracle.com/webfolder/technetwork/tutorials/obe/db/11g/r2/prod/bidw/datamining/ODM11gR2.htm

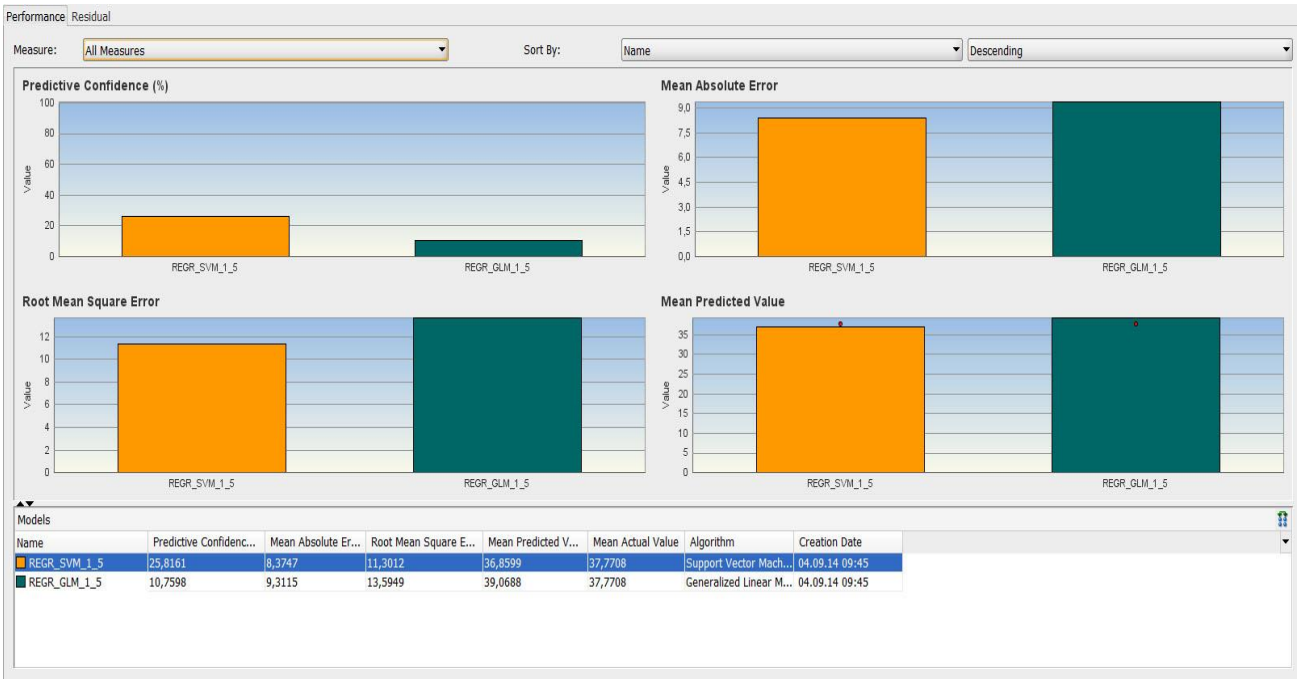
Przebieg:

1. Uruchom SQL Developer.
2. Korzystając z wytycznych na stronie WWW, przeprowadź budowę diagramu przepływu zawierającego model klasyfikujący dane ubezpieczeniowe.
3. Wykorzystaj opcję Compare Test Result do weryfikacji jakości klasyfikacji. Sprawdź w szczególności zakładki Performance i Performance Matrix.
4. Czy potrafisz wyjaśnić i omówić wszystkie współczynniki (poza parametrami Cost) na tych dwóch zakładkach? Sprawdź, czy Twoje obliczenia w oparciu o dolną tabelkę z poniższego screenu (poszczególne dane mogą się różnić w stosunku do rozpatrywanego przykładu – to tylko rysunek poglądowy) zgadzają się z wartościami współczynników w zakładce Performance dla poszczególnych algorytmów (przycisk Display / Show Detail)?



Zadanie 2

1. Zbuduj nowy model – Regresję i przeprowadź proces analizy danych:
    a. Wykorzystaj tabelę: INSUR_CUST_LTV_SAMPLE
    b. Nie wykluczaj żadnych kolumn w analizie
    c. Target: AGE
    d. Case Id: CUSTOMER_ID – nie jest to niezbędne, ale to dobra praktyka. Pozwala w przyszłości na powtarzalność działań.
    e. Po przeprowadzeniu regresji sprawdź opcję Compare Test Results:



    f. Można zweryfikować kilka miar dokładności – pozwalające ocenić różnicę pomiędzy wartościami przewidywanymi a wartościami aktualnie obserwowanymi:
        i. Mean Absolute error - średni błąd bezwzględny
        ii. Mean Square Error – średni błąd kwadratowy
        iii. Root Mean Square Error – pierwiastek błędu średniokwadratowego.
        iv. Oraz Predictive Confidence – dokładność przewidywań (dokładność klasyfikatora) liczona względem losowego wybierania.
    g. Przejdź na zakładkę Residual (górny pasek zakładek). Można zaobserwować wskazania różnic pomiędzy wartością aktualną a przewidywaną (na podstawie zbioru testowego).
    h. Wciąż będąc w zakładce Residual, kliknij ikonę lupy w prawym górnym rogu, aby z formy wykresu przejść na formę tabeli. Widać wówczas dla każdego przypadku (każdego klienta wg customer_id). W kolumnie Residual znajdują się wartości różnic, które brakują wartości przewidywanej, aby dokładnie odpowiadała wartości rzeczywistej (pamiętaj, że wciąż rozważamy zbiór testowy, dla którego znamy wartość rzeczywistą).

i. Wróć na formę wykresu (ikona wykresu słupkowego w prawym górnym rogu okna). Teraz bardziej zrozumiały jest wykres: na osi X podane są wartości przewidywane, a na osi Y informacja, ile brakowało dla danej wartości przewidywanej. Widać więc, dla których wartości przewidywanych algorytm popełniał większe błędy (dalej od poziomu 0 w górę lub dół) a dla których mniejsze błędy (bliżej poziomu 0). Odpowiadamy w ten sposób na pytanie, którym przewidywaniom (predykcją) możemy ufać najbardziej.
j. Korzystając z górnego menu zmień X Axis z Predicted na Actual. Kliknij przycisk Query (prawy górny róg). Ponownie przeliczony zostanie wykres. Odpowiadamy w ten sposób na pytanie: które z wartości rzeczywistych nasz model najtrafniej przewiduje (popełnia najmniejszy błąd, czyli z najmniejszą dodatnią lub ujemną wartością Residual).
k. Korzystając ponownie z menu, można ustawić, z którym innym modelem nasz aktualnie oceniany model będzie porównany. Wybierz model SVM. Ponownie kliknij przycisk Query.
l. Obecnie na ekranie prezentowane są dwa wykresy – dla obu modeli.
m. Zmień X Axis z powrotem na Predicted i ponownie kliknij Query. Widać, że w przypadku GML jest kilka predykcji, które znajdują się daleko po prawej, a ich Residual jest duże i ujemne. Dla SVM nie obserwujemy takich przypadków. Poza tym wyniki prezentują się dość podobnie.
n. Zmień Y Axis na Actual i ponownie kliknij Query. Również w tej formie prezentacji wyraźnie widać kilka odstających przypadków dla GLM.
o. Dla opcji Compare wybierz None i ponownie kliknij Query.
p. Zamknij zakładkę Regress Build i wróć do diagramu przepływu.

5. Przygotuj sprawozdanie (dla własnego użytku, nie wysyłaj go), w którym opiszesz swoje uwagi dotyczące procesu budowy modelu:
    a. Dobrane parametry
    b. Uzyskane wyniki
    c. Uwagi z różnych wariantów konfiguracji algorytmów