Modelowanie Statystyczne w Zarządzaniu Wierzytelnościami Masowymi

Laboratorium 3.

1. Zapoznaj się z podsumowaniem danych.
2. Zakoduj cechy jakościowe `Product` oraz `Gender`.
3. Uzupełnij za pomocą wybranej metody braki danych tych zmiennych aplikacyjnych, dla których jest to możliwe i sensowne.
4. Usuń obserwacje odstające cech `LoanAmount`, `DPD` oraz `LastPaymentAmount`.
5. Zestandaryzuj cechy aplikacyjne.
6. Do zbioru cech aplikacyjnych dodaj cechę wydzielającą klientów, którzy dokonali jakiejkolwiek wpłaty w pierwszych 6 miesiącach obsługi (klienci dobrzy). Jaki jest udział dobrych klientów w zbiorze?
7. Wykonaj analizę skupień za pomocą algorytmu k-średnich z użyciem zmiennych `TOA` oraz `M_LastPaymentToImportDate`.
8. Stwórz macierz kontyngencji rzeczywistej oraz prognozowanej dobroci klienta przyjmując, że w danym skupieniu sprawy są uznawane za dobre jeżeli udział dobrych klientów w skupieniu jest wyższy od udziału dobrych klientów w całym zbiorze danych.
9. Jaka jest jakość klasyfikacji uzyskanej w poprzednim punkcie?
10. Przedstaw wyniki klasyfikacji na trójwymiarowym wykresie (pakiet `ggplot2`).
11. Czy dodanie innych cech aplikacyjnych do modelu poprawia jakość klasyfikacji?
12. Jaka liczba skupień daje najlepsze wyniki klasyfikacji?
13. Dokonaj prognozy dobroci klienta z wykorzystaniem algorytmu k-najbliższych sąsiadów `knn` z pakietu `class` (eksperymentuj z różną liczba najbliższych sąsiadów oraz zakresem cech). Jaka jest optymalna specyfikacja modelu?
14. Wykonaj punkt (13) z użyciem odrębnych zbiorów uczącego i testowego. Jak zmieniają się wnioski?
15. Napisz własną funkcję klasyfikatora najbliższych sąsiadów z następującymi parametrami:

    - liczba najbliższych sąsiadów,
    - zbiór uczący,
    - zbiór testowy,
    - cechy,
    - udział liczby sąsiadów z danej klasy jako próg klasyfikacji.