

Program szkolenia

Credit scoring dla zaawansowanych

Prowadzący: dr inż. Artur Suchwałko

1. Krótkie wprowadzenie do R

- wprowadzenie do środowiska R
 - specyfika i przegląd możliwości systemu
 - instalacja i konfiguracja
 - użytkowanie R
 - system pomocy
 - graficzny interfejs użytkownika
 - korzystanie z wbudowanych funkcji
- R od podstaw: typy i struktury danych
 - typy zmiennych
 - obiekty i ich podstawowe własności (wektory, macierze, zmienne tekstowe, listy i ramki danych)
 - podstawowe operacje na obiektach
- elementy programowania w R
 - podstawy języka R
 - instrukcje sterujące przepływem kodu
 - tworzenie własnych skryptów i funkcji

2. Wstęp

- ocena ryzyka kredytowego przed powstaniem systemów scoringowych
- zalety i obszary stosowania systemów scoringowych
- idea działania: wykorzystanie danych historycznych do przewidywania przyszłego zachowania
- zalety i wady punktowej oceny ryzyka
- korzyści wynikające z zastosowania systemów scoringowych
- rodzaje systemów scoringowych
 - podział ze względu na zastosowanie
 - podział ze względu na sposób budowy
 - łączenie różnych rodzajów scoringu

3. Przegląd etapów kompletnego procesu budowy systemu scoringowego

- organizacja projektu (w tym definicja celu biznesowego dla systemu scoringowego)
- wstępna analiza danych

- definicja parametrów
 - definicja dobrego i złego klienta: przekształcenie celu biznesowego na cel statystyczny
 - application window i performance window
 - wykluczenia
 - segmentacja
- przygotowanie danych
 - cechy stosowane w scoringu
 - wybór próby konstrukcyjnej
 - zgromadzenie i oczyszczenie danych
- budowa karty scoringowej w skrócie
 - analiza i przekształcenia cech do budowy systemu scoringowego
 - regresja logistyczna: podstawy teoretyczne i praktyka
 - metody wyboru cech do budowy modelu
 - sposoby oceny jakości systemu scoringowego
 - uwzględnienie wniosków odrzuconych (reject inference)
- wykorzystywanie systemu scoringowego w praktyce
 - raporty podsumowujące proces budowy karty scoringowej
 - wdrożenie karty (w tym dobór punktu odcięcia: iso-risk, iso-acceptance)
 - monitorowanie karty

4. Analiza i przekształcenia cech do budowy systemu scoringowego

- analiza pojedynczych cech
 - Weight of Evidence, odds
 - rozkłady cech (tablice kontyngencji, histogramy)
 - obserwacje brakujące oraz obserwacje odstające
 - kontrola jakości i czyszczenie danych
 - wstępny wybór cech do konstrukcji modelu — analiza zdolności dyskryminacyjnej cech
- przedziałowanie zmiennych ciągłych (fine classing i coarse classing)
 - rola przedziałowania
 - metody przedziałowania
 - * weight of evidence (WoE)
 - * maksymalizacja entropii
 - * zastosowanie drzew klasyfikacyjnych
- analiza zależności między cechami i konstrukcja cech pochodnych (generated characteristics, cross characteristics)

5. Regresja logistyczna: podstawy teoretyczne i praktyka

- wprowadzenie do regresji logistycznej
 - zalety regresji logistycznej
 - modele regresyjne - wprowadzenie na przykładzie regresji liniowej
 - czym jest scoring: model liniowy, model logit i probit
- idea modelu (trzy podejścia: dummy variables, przekodowanie WoE, modele dla zmiennych ciągłych)
- podstawy statystyczne
- budowa modelu, własności praktyczne i teoretyczne
- dwa podejścia do budowy systemów scoringowych metodą regresji logistycznej
- interpretacja wyników
- diagnostyka modelu: testy statystyczne i wykresy
- wnioskowanie statystyczne dla regresji logistycznej

- inne metody budowy systemów scoringowych oraz ich wady i zalety
6. Metody wyboru cech do budowy modelu
- wprowadzenie do oceny jakości modelu
 - kryteria zastosowania cech w scoringu (statystyczne, biznesowe, operacyjne)
 - wartość informacyjna cechy (information value)
 - przegląd zupełny zbioru cech
 - inne sposoby wyboru cech
 - metody jednokrokowe (filtry)
 - metody forward, backward bazujące na jakości klasyfikacji
 - metody krokowe bazujące na kryterium AIC (forward, backward, forward-backward)
 - metoda wyboru stabilnych cech bazująca na wielokrotnej budowie modeli klasyfikacyjnych i wyborze cech (komitety modeli)
 - metoda bazująca na random forest (rankingi cech)
 - marginal Chi^2
 - wady metod krokowych wyboru cech
7. Sposoby oceny jakości systemu scoringowego
- kryteria dobroci dopasowania modelu (AIC, R^2)
 - analiza zdolności predykcyjnej modelu
 - rozkłady punktów scoringowych
 - ocena jakości klasyfikacji: confusion matrix
 - ocena zdolności dyskryminacyjnej: krzywa ROC, miary AR, KS i dywergencja
 - strategie oceny jakości: learning / test, cross validation, bootstrap
8. Uwzględnienie wniosków odrzuconych (reject inference)
- idea reject inference
 - przegląd metod reject inference
 - define as bad
 - extrapolation
 - augmentation
 - iterative reclassification
 - mieszanina rozkładów normalnych
9. Budowa finalnej karty scoringowej
- wykrywanie interakcji
 - skalowanie karty
 - metody kalibracji karty do skali wzorcowej (master scale)
 - score banding: statystyka Calinskiego-Harabasza
 - benchmarking: dopasowanie granic kategorii ratingowych (dopasowanie do master scale)
 - przesunięcie liniowe i skalowanie: kalibracja do skali wzorcowej
 - kalibracja dla Low Default Portfolios
10. Analiza segmentacji
- kryteria wykonywania segmentacji
 - metody segmentacji:
 - drzewa klasyfikacyjne
 - analiza skupień
 - metody bazujące na wiedzy eksperckiej

11. Bardzo ważne praktyczne aspekty modelowania

- budowa modeli dla małych zbiorów danych
- budowa modeli dla cech numerycznych (ilościowych) bez przedziałowania
- zależność cech (numerycznych i kategoriowych) — jak sobie z nią poradzić
- nierówne proporcje grup dobrych i złych klientów i konsekwencje
- porównanie podejść do budowy modeli: dummy variables, przekodowanie WoE, modele dla zmiennych ciągłych

12. Wykorzystywanie systemu scoringowego w praktyce

- dobór optymalnego punktu odcięcia
- wdrożenie systemu scoringowego
- monitorowanie skuteczności działania systemu i raportowanie:
 - monitorowanie siły predykcyjnej systemu
 - monitorowanie siły predykcyjnej cech
 - monitorowanie stabilności populacji
- strategia champion-challenger

13. Krótki przegląd innych metod budowy systemów scoringowych oraz ich wady i zalety

- LDA, QDA
- drzewa klasyfikacyjne i pokrewne modele (rozwińcie w dalszej części)
- sieci neuronowe (rozwińcie w dalszej części)
- analiza przeżycia (rozwińcie w dalszej części)
- programowanie matematyczne (skrótowo)

14. Modele oparte na drzewach

- specyfika modeli opartych na drzewach
- przegląd zastosowań modeli opartych na drzewach
- wizualizacja i interpretacja wyników
- praktyczne aspekty związane z budową modeli opartych na drzewach:
 - kryteria wyboru zmiennych
 - kryteria podziału
 - kryteria zatrzymania
 - ocena złożoności struktury drzewa
- drzewa klasyfikacyjne
- postprocessing drzew: upraszczanie i modyfikacje struktury drzew (pruning), analiza ekspercka
- zalety i wady modeli opartych na drzewach

15. Sieci neuronowe

- idea sieci neuronowych
- preprocessing i postprocessing danych
- budowa modeli sieci neuronowych
- dobór parametrów sieci

16. Złożoność a dokładność modeli

- porównanie metod budowy modeli w praktyce
 - regresja logistyczna
 - drzewa klasyfikacyjne
 - sieci neuronowe

- konieczny kompromis między złożonością oraz jakością i uogólnianiem modeli

17. Dodatkowe metody budowy modeli scoringowych

- lasy losowe
- bagging, boosting
- analiza przeżycia