Projekt 1

SAR 2020/2021

Informacje ogólne

- Projekty wykonujemy w zespołach 1-osobowych
- Termin wysłania projektu: 12 stycznia 2021 godzina 23:59:59
- Proszę wysłać projekt na adres e-mail prowadzącego grupę
- Wysyłamy 2 pliki: kod programu (plik imie_nazwisko.R) oraz plik w wynikami (plik: imie_nazwisko.txt)

Dane

1. Projekt dotyczy analizy zbioru danych „Communities and crime" które dotyczą przestępczości w poszczególnych hrabstwach w USA.
2. Dane można pobrać ze strony: https://archive.ics.uci.edu/ml/datasets/communities+and+crime
3. Celem jest prognozowanie zmiennej ViolentCrimesPerPop (liczba przestępstw na 100 K mieszkańców)- ostatnia kolumna w zbiorze danych.
4. Dokładny opis poszczególnych zmiennych znajduje się w pliku: communities.names

Polecenia:

1. Wczytaj dane do programu R.
2. Usuń pierwsze 5 kolumn (zmienne: state, county, community, communityname, fold).
3. Usuń wszystkie kolumny które zawierają braki danych (znak „?").
4. Dokonaj podziału danych na dwa podzbiory: zbiór uczący U (obserwacje o parzystych indeksach) oraz zbiór testowy T (obserwacje o nieparzystych indeksach).
5. Dopasuj model regresji liniowej na danych U (model m).
6. Wykonaj diagnostykę modelu (na danych U), zaproponuj odpowiednie procedury diagnostyczne. Na ich podstawie dokonaj ewentualnie modyfikacji zbioru danych (usunięcie obserwacji, transformacje, itp.). Dopasuj nowy model, nazwijmy go m_mod.
7. Dokonaj selekcji zmiennych bazując na modelu m oraz m_mod. Zastosuje 2 podejścia: BIC BIC backward, AIC backward, nazwijmy otrzymane modele m1, m2 oraz m_mod1, m_mod2).
8. Dokonaj prognozy zmiennej celu na danych testowych T używając modeli:
    a. m, m_mod
    b. m1, m2, m_mod1, m_mod2
    c. m_empty (model zawierający tylko wyraz wolny)
9. Jakość prognozy oceń obliczając RMSE (root of mean squared error): $\sqrt{\sum_{i \in T}[yi - \hat{y}i]^2}$.
10. **BONUS:** zaproponuj własny model (model nie omawiany na zajęciach) do prognozy.
11. Zapisz wyniki do pliku o nazwie: imie_nazwisko.txt. Wyniki powinny być zapisane do tabeli, poniżej przykład:

Model	**RMSE**	**Liczba zmiennych w modelu**
m_empty	2.65	0
m1	1.34	5
m2	1.21	12
…	…	…