Tytuł:
DOPASOWANIE MODELU DO DANYCH

Zespół:
Oktawia Hankus, Zuzanna Nogala

Motywacja:
Stworzenie wygodnego narzędzia, które sprawnie oceni, który model z najpopularniejszych uogólnionych modeli regresji liniowej (*) będzie lepiej dopasowany do danych, w formacie .csv, przekazanych przez użytkownika. Użytkownik na podstawie wybranych przez siebie predyktorów i zmiennej odpowiedzi otrzyma w aplikacji R Shiny pełną analizą zachowania parametrów modeli, informacje czy predyktory są istotne i sugestia, że być może inne będą bardziej odpowiednie i jaki model będzie najlepszy bądź, że żaden z modeli się nie nadaje.

(*) regresja liniowa, logistyczna, Poissona, model proporcjonalnego hazardu Coxa, …

Dane:

- Opis zbioru:
Do przetestowania działania pakietu wykorzystamy zbiór danych z oficjalnej strony rządu Meksyku https://datos.gob.mx/busca/dataset/registro-civil. Baza obejmuję 4923 związków małżeńskich, które w latach 2000 - 2015 wzięły rozwód w mieście
Xalopa w Meksyku. Każda informacja została opisana za pomocą 41 zmiennych, w tym:

Data rozwodu,
data małżeństwa,
data urodzenia (mężczyzny/kobiety),
narodowość (mężczyzny/kobiety),
miesięczne zarobki (mężczyzny/kobiety),
zawód (mężczyzny/kobiety),
poziom edukacji (mężczyzny/kobiety),
status zatrudnienia (mężczyzny/kobiety),
liczba dzieci…

- Format danych: plik .csv

Narzędzia:

Kluczowe pakiety:
data.table/tidyverse,
ggplot2,
kableExtra,
cleaner/cleanr/janitor(?)

Inne:
Github

Planowanie funkcjonalności:

Konieczne funkcjonalności:
1. Na podstawie wybranych kolumn, na których użytkownik chce utworzyć model ocena, który model jest najlepiej dopasowany do danych.
2. Wysłanie feedbacku do użytkownika, który model jest dobry lub że żaden z modeli w pakiecie nie będzie pasował do danych użytkownika.

3. Wybór dowolnego z wyżej wymienionych modeli przez użytkownika, niekoniecznie sugerowanego przez pakiet.
4. Interaktywna wizualizacja danych.
5. Aplikacja R Shiny, gdzie będzie zawarta:
    1. Strona główna gdzie załącza się plik i jeżeli plik jest załączony to wyświetlanie fragmentu tabeli,
    2. Strona do wizualizacji danych - dobór argumentów x, y z danych i narysowanie,
    3. Strona modelu - wybieramy, co jest zmienną odpowiedzi, co jest zmiennymi niezależnymi i za pomocą pakietu jest wybierany najlepszy model i wysyłany feedback:
        1. Model najlepszy to: …
        2. Porównanie z innymi modelami: estymacje parametrów …
    4. Strona z predykcją.
    5. Zakładka „moje modele” (stworzone modele w „zakładce modele” lądują tu)

Opcjonalne funkcjonalności:
1. Czyszczenie zbioru danych za pomocą innych bibliotek (ujednolicenie nazw kolumn, …)
2. Dodanie do feedbacku sugestii, jakie inne zmienne lepiej pasują do predykcji wybranej zmiennej odpowiedzi, bo np: zbyt duże skorelowanie użytych zmiennych

Zarys pracy:

1. Napisanie pakietu:

    - Funkcja - szukanie modelu który najlepiej pasuje do danych
        - Plus mniejsze funkcje pomocnicze,
    - Funkcja - czyszczeniem danych (opcjonalnie),
    - Funkcje - wizualizacja danych (wybór jaki rodzaj wykresów można wyświetlić…),
    - Funkcje - pobieranie danych od użytkownika i odpowiednie przekształcenia.

2. Napisanie dokumentacji do pakietu

3. Stworzenie aplikacji R-Shiny:
    - Pobieranie danych od użytkownika,
    - Zaprogramowanie aplikacji - zakładki, wyświetlanie wykresów, zaprogramowanie dobrze wyświetlanych feedbacków.