

Praca domowa 2

Zaawansowane Metody Uczenia Maszynowego

Termin oddania: 09.05.2024

1 Cel

Celem pracy domowej jest wykorzystanie metod wyjaśnialnego uczenia maszynowego do oceny przygotowanych modeli.

2 Dane

W tej pracy domowej wykorzystaj jeden zbiór danych dla dowolnego problemu (regresja lub klasyfikacja). Warto skorzystać z większego zbioru danych z większą liczbą kolumn/cech, aby otrzymywne wyniki były interesujące.

Uwaga

Z możliwej listy zbiorów danych usuwamy zbiór FICO oraz zbiór o cenach nieruchomości w Seattle (zbiór danych z zajęć).

3 Modele

Aby wykonać kolejne części zadania domowego należy przygotować co najmniej 3 modele różnego typu (np. model regresji, drzewo decyzyjne, las losowy lub inny komitet modeli).

4 Analiza modeli

Wykorzystując zbudowane modele w Sekcji 3 przygotuj analizę globalną i lokalną modeli posługując się metodami wyjaśnialnego uczenia maszynowego.

Analiza lokalna

- Wybierz co najmniej 2 obserwacje ze zbioru danych i oblicz dla nich predykcję modeli oraz oblicz dekompozycję predykcji modeli za pomocą metody SHAP lub Break Down.
- Wybierz co najmniej 2 obserwacje ze zbioru danych i oblicz dla nich predykcję modeli oraz oblicz rozkład predykcji modeli za pomocą profili Ceteris Paribus / ICE.

Analiza globalna

- Dla przygotowanych modeli oblicz permutacyjną istotność zmiennych. Jakie cechy są istotne dla tych modeli? Czy mają podobne czy różne istotne zmienne? Przygotuj model na zbiorze danych ograniczonym tylko do najistotniejszych cech i sprawdź czy nadal utrzymuje podobną jakość predykcyjną.
- Dla pewnych wybranych zmiennych (można na przykład bazować na permutacyjnej istotności zmiennych) przedstaw profile częściowej zależności (PDP, ALE). Czy są zauważalne jakieś zachowania pomiędzy różnymi modelami?

5 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- raport w pliku `NUMERINDEKSU.ipynb` (lub `NUMERINDEKSU.Rmd`, gdy praca domowa jest wykonana w R) zawierający kod oraz opisy wyników wymaganych w Sekcji 2, 3 i 4,
- dane, jeżeli nie da się ich zaczytać używając kodu i muszą być jako oddzielny plik, np. `.csv`.

6 Ocena

Łączna liczba punktów do zdobycia jest równa 10, w tym:

- modele - 1 punkt,
- analiza globalna - 4 punkty,
- analiza lokalna - 4 punkty,
- jakość kodu i raportu - 1 punkt.

7 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_PD2`. Tak przygotowany katalog należy przesłać na adres `anna.kozak@pw.edu.pl` do dnia 09.05.2024 do godziny 23:59. Tytuł wiadomości: `[ZMUM][PD2] Nazwisko Imię`.