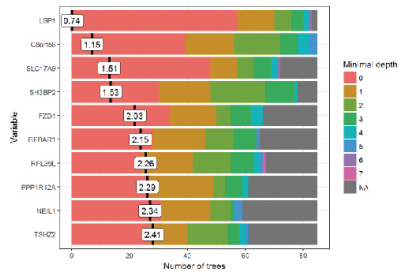
Nr rejestracyjny: 2017/27/B/ST6/01307; Kierownik projektu: dr hab. inż. Przemysław Biecek

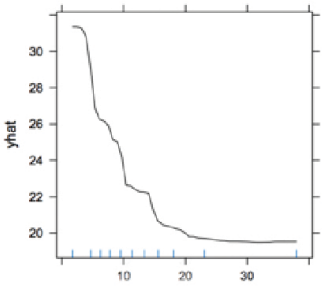
DALEX: Lokalne, brzegowe i globalne objaśnienia złożonych modeli uczenia maszynowego

Global



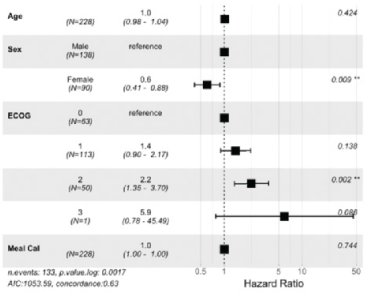
Objaśnienia struktury modelu, globalna istotność zmiennych, globalne trasformacje

Marginal



Warunkowa odpowiedź modelu brzegowe efekty dla jednej lub dwóch zmiennych

Local



Lokalna struktura modelu, odpowiedź dla jednej obserwacji struktura w otoczeniu punktu

Celem projektu jest opracowanie metod badania czarnych skrzynek, czyli złożonych modeli uczenia maszynowego, np. głębokich sieci, komitetów predyktorów, wysokowymiarowych modeli regresyjnych. Metody te: (1) zbudują wyjaśnienia - kluczowe zmienne determinujące odpowiedź modelu, (2) uzasadnią odpowiedź modelu dla pojedynczej predykcji, (3) przeprowadzą diagnostykę modelu.

Metoda

Praca nad metodami eksploracji modeli predykcyjnych podzielona jest na trzy zadania badawcze: wyjaśnienia lokalne (część LIVE), wyjaśnienia predykcji (EXPLAIN) i wyjaśnienia warunkowe (CONDA).

Planowane prace są kontynuacją badań pilotażowych [Biecek2017, Dąbrowska2017, Paluszyńska2017, Staniak2017, Sitko2017]. W przypadku wyjaśnień lasów losowych punktem wyjścia jest metoda randomForestExplainer [Paluszyńska2017]. To rozwiązanie pozwala lepiej zrozumieć które zmienne odgrywają główną rolę w podejmowaniu decyzji przez las (patrz rysunek). Istniejąca implementacja wspiera klasyczne lasy losowe. Będzie ona rozszerzona aby wspierać inne komitety predyktorów. W przypadku badań nad lokalnymi wyjaśnieniami, prace oparte będą o rozszerzenia pakietu live - Local Interpretable (Model-agnostic) Visual Explanations, który jest rozszerzeniem metody LIME [Ribeiro2016] na przypadek danych mieszanych. W przypadku badań nad brzegowymi wyjaśnieniami, prace będą prowadzone w oparciu o takie narzędzia jak krzywe partial dependence plots [Greenwell 2017] oraz ich adaptacje dla danych jakościowych.

Wpływ rezultatów

Projekt ma fundamentalne znaczenie dla dalszego rozwoju modelowania predykcyjnego z użyciem złożonych modeli. Czarne skrzynki są coraz chętniej wykorzystywane, jednakże nawet najwyższa skuteczność uzyskana podczas budowy modelu nie gwarantuje jego poprawności.

Metody i narzędzia do wizualnej eksploracji modeli uczenia maszynowego pozwolą na szybką identyfikację błędów w strukturze czarnej skrzynki. Dodatkowo, zwiększą interpretowalność modelu, która jest istotna między innymi w spersonalizowanej medycynie czy regulowanym scoringu kredytowym. Stworzone przez mój zespół metody ułatwią ocenę istotności zmiennych w modelu oraz przeniesienie wyuczonych reguł na nowe modele.

Literatura

[Biecek2017a] Przemysław Biecek, Marcin Kosiński (2017). archivist: An R Package for Managing, Recording and Restoring Data Analysis Results Journal of Statistical Software. 82 (22)

[Dąbrowska2017] Aleksandra Dąbrowska, Alicja Gosiewska, Przemysław Biecek "MLExpResso: Integrative analyses and visualization of gene expression and DNA methylation data" (2017) https://github.com/geneticsMiNIng/MLGenSig/

[Paluszyńska2017] Aleksandra Paluszyńska, Przemyslaw Biecek "BlackBoxOpener: an R package" (2017) https://github.com/geneticsMiNIng/BlackBoxOpener

[Sitko2017] Agnieszka Sitko, Przemysław Biecek "factorMerger: an R package" (2017). Submitted to Journal of Computations and Graphical Statistics https://arxiv.org/abs/1709.04412

[Staniak2017] Mateusz Staniak, Przemysław Biecek. "Local Interpretable (Model-agnostic) Visual Explanations" (2017) https://github.com/MI2DataLab/live

[Ribeiro2016] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. "Why Should I Trust You?: Explaining the Predictions of Any Classifier." (2016). https://arxiv.org/abs/1602.04938

[Greenwell2017] Brandon Greenwell. "pdp: An R Package for Constructing Partial Dependence Plots." (2017) https://journal.r-project.org/archive/2017/RJ-2017-016/index.html