Nr rejestracyjny: 2021/43/O/ST6/00347; Kierownik projektu: dr hab. inż. Przemysław Biecek

ARES: Wyjaśnienia odporne na ataki w kierunku bezpiecznej i godnej zaufania SI

Wyjaśnialność, dyskryminacja, odporność i bezpieczeństwo są głównymi składowymi godnej zaufania Sztucznej Inteligencji, która jest strategicznie ważnym obszarem rozwoju metod SI. W tym kontekście, głównymi celami projektu ARES są: (1) Opracowanie **ataków** na współczesne wyjaśnienia w celu zbadania słabości oraz ograniczeń istniejących metod i oceny dyskryminacji w uczeniu maszynowym. (2) Wprowadzenie nowych **wyjaśnień odpornych** na ataki. Projekt ma na celu rozwinąć wyjaśnialne uczenie maszynowe w kierunku bezpiecznej i godnej zaufania adopcji rozwiązań SI.

Pytania i hipotezy: W szczególności, staramy się odpowiedzieć na następujące pytania i hipotezy badawcze: Jakie są ograniczenia obecnie stosowanych wyjaśnień? Jak wykryć potencjalną manipulację wyjaśnieniami? Współczesne wyjaśnienia dla modeli uczenia maszynowego trenowanych na danych tabelarycznych nie są odporne ani wiarygodne w kontekście ataków na wyjaśnienia. Co poprawić w kierunku stworzenia solidnych wyjaśnień? Jak ewaluować wyjaśnienia w kontekście adwersarza? Odporne wyjaśnienia wprowadzają nową jakość i są niewrażliwe na potencjalne ataki.

Znaczenie: Pomimo szybkiego rozwoju, obszar badań nad wyjaśnialnym uczeniem maszynowym jest hamowany przez brak uwzględnienia bezpieczeństwa i odpowniednio głębokiej ewaluacji nowych metod i narzędzi. Historycznie, było tak w przypadku scenariuszy zajmujących się atakami i bezpieczeństwem w uczeniu maszynowym. Osiągnięcie pierwszego celu ma przede wszystkim wpływ na różne inne dziedziny badań, które obecnie wykorzystują (i wyjaśniają) modele typu black-box w celu analizy danych i podejmowania decyzji, poprzez wskazanie słabych punktów i ograniczeń ich wyjaśnień. Osiągnięcie drugiego celu ma wpływ na całą dziedzinę uczenia maszynowego, ponieważ ma na celu poprawę obecnie stosowanych narzędzi poprzez wprowadzenie odpornych wyjaśnień niezbędnych do bezpiecznej i godnej zaufania AI.

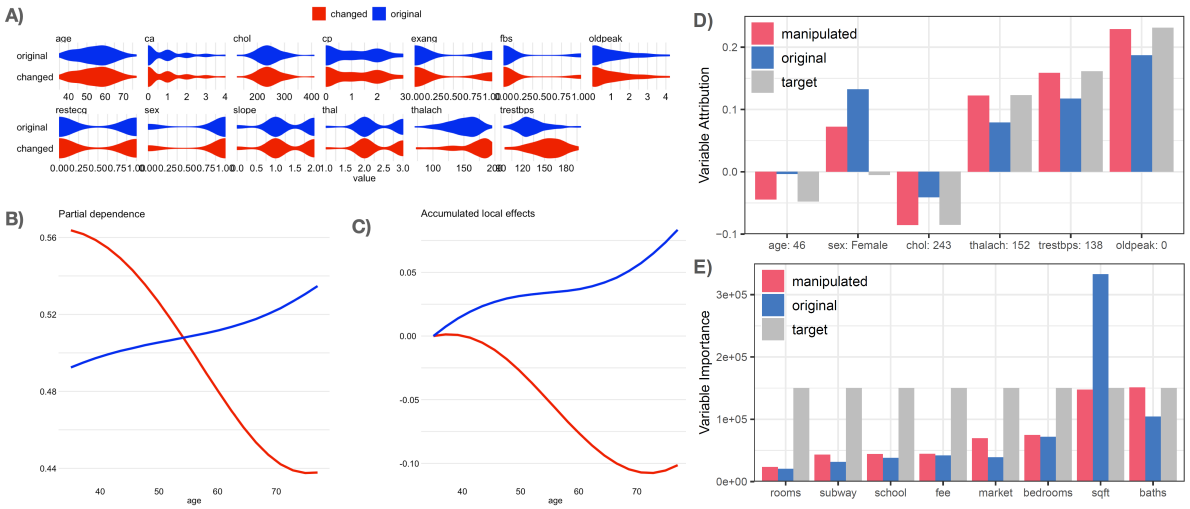

Figure 1: W badaniach wstępnych [1–3], udaje nam się efektywnie manipulować (zmieniać) wyjaśnienia dla danych tabularycznych w rzeczywistych problemach predykcyjnych, wykorzystując podatność wyjaśnień na **A)** zatrucie danych, na przykład metod **B)** partial dependence, **C)** accumulated local effects, **D)** SHAP attribution oraz **E)** SHAP importance, który jest skonstruowany dla problemu finansowego. Niebieskie linie i słupki oznaczają oryginalne wyjaśnienie, podczas gdy zmanipulowane wyjaśnienia są na czerwono.

References

[1] H. Baniecki and P. Biecek. Manipulating SHAP via Adversarial Data Perturbations (Student Abstract). In *(to appear) AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[2] H. Baniecki, W. Kretowicz, and P. Biecek. Fooling Partial Dependence via Data Poisoning. *arXiv preprint arXiv:2105.12837*, 2021.

[3] K. Woźnica, K. Pękala, H. Baniecki, W. Kretowicz, E. Sienkiewicz, and P. Biecek. Do not explain without context: addressing the blind spot of model explanations. *arXiv preprint arXiv:2105.13787*, 2021.