Nr rejestracyjny: 2019/34/E/ST6/00052; Kierownik projektu: dr hab. inż. Przemysław Biecek

HOMER: Human Oriented autoMated machinE leaRning

Modele uczenia maszynowego sa używane wszedzie. Modelowanie predykcyjne całkowicie zmieniło dyscypliny oparte na danych, takie jak medycyna, biologia molekularna, finanse, transport i wiele innych. Rosnaca dostepność dużych źródeł danych w połaczeniu z najnowszymi osiagnieciami w obszarze uczenia maszynowego prowadzi do kolejnej rewolucji przemysłowej.

ALE: Modele predykcyjne sa wciaż tworzone recznie przez analityków w żmudnym i pracochłonnym procesie. Wiekszość czasu poświecanego na eksploracje danych i trenowanie modeli to zestaw prób i błedów. Modele sa coraz bardziej złożone. Brak zrozumienia złożonych modeli i słaba automatyzacja powoduja problemy z odtwarzalnościa i jakościa modeli. Prowadzi to do groźnych sytuacji.

- Modele nie działaja poprawnie i sa trudne do debugowania. Na przykład *Watson for Oncology* jest krytykowany przez onkologów za *niebezpieczne i niedokładne* zalecenia (Ross and Swetliz, 2018).
- Wyniki sa systematycznie tendencyjne. Na przykład Amazon porzucił prace nad system do sprawdzania CV, ponieważ był tendencyjny wobec kobiety (Dastin, 2018) a model COMPAS do predykcji recydywy dyskryminował ze wzgledu na rase (Larson et al., 2016).
- Dryf danych prowadzi do pogorszenia wydajności modeli. Na przykład bardzo popularny model *Google Flu* po dwóch latach dał gorsze prognozy niż poziom bazowy (Salzberg, 2014).
- Prognozy modelu jest błedne, ale nikt nie może wyjaśnić, które dane wejściowe steruja ta konkretna prognoza. Wiele przykładów takich problemów można znaleźć w ksiażce (O'Neil, 2016) z wyrazistym podtytułem *,,Jak duże zbiory danych zwiekszaja nierówności i zagrażaja demokracji"*.

Wiekszość z tych problemów zostałaby zauważona dzieki lepszym zorientowanym na człowieka metodom automatycznego debugowania, eksploracji i wyjaśniania modeli uczenia maszynowego.

Głównym celem tego projektu jest opracowanie nowych metod eksploracji modeli zorientowanych na człowieka, interpretowalnych audytów modeli predykcyjnych i automatycznego trenowania modeli. Nowo powołany zespół badawczy stworzy gramatyke dla rozwoju tematyki Model Orietned Machine Learnign.

Jest to ogólny długoterminowy cel zespołu badawczego. Przygotowaliśmy również siedem szczegółowych pytań badawczych i odpowiadajacych im siedem zadań dla tego projektu.

Q1 Czy możemy użyć złożonego modelu uczenia maszynowego jako wzorca do wyodrebnienia możliwych do interpretacji cech i modelu interpretowalnego?

Q2 Czy interpretowalne miary oceny modelu, takie jak ranking Elo, usprawniłyby proces wyboru modelu?

Q3 Czy miary oparte na rankingu Elo zwiekszyłoby efektywność meta-uczenia sie w optymalizacji hiperparametrów?

Q4 Czy interpretowalne wyniki Elo pomogłyby w wykryciu dryfu w modelu?

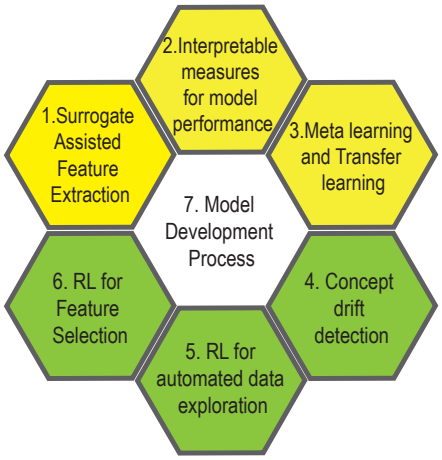

Figure 1: Żółte zadania dotycza interakcji człowiek-model, zielone automatyzacji budowy modeli.

Q5 Czy uczenie ze wzmocnieniem pomoże w automatycznej eksploracji danych?

Q6 Czy uczenie ze wzmocnieniem usprawnił proces wyboru funkcji?

Q7 W jaki sposób metodyki opracowane dla inżynierii oprogramowania pasuja do Data Science?

Te pytania inspirowane sa wstepnymi wynikami opublikowanymi w pracach Biecek (2018), Gosiewska and Biecek (2019), Staniak and Biecek (2019) and Biecek (2019).

References

Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19:1–5.

Biecek, P. (2019). Model Development Process. *https://arxiv.org/abs/1907.04461*.

Dastin, J. (2018). Amazon scraps secret ai recruiting tool that showed bias against women. *Reuters*.

Gosiewska, A. and Biecek, P. (2019). ibreakdown: Uncertainty of model explanations for non-additive predictive models. *CoRR*, abs/1903.11420.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, New York, NY, USA.

Ross, C. and Swetliz, I. (2018). Ibm's watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. *Statnews*.

Salzberg, S. (2014). Why google flu is a failure. *Forbes*.

Staniak, M. and Biecek, P. (2019). The Landscape of R Packages for Automated Exploratory Data Analysis. *https://arxiv.org/abs/1904.02101*.