Metody Wyjaśnialnej Sztucznej Inteligencji (XAI), jako kluczowy element do klarownego zrozumienia złożonych modeli uczenia maszynowego w scoringu kredytowym.

Marcin Chlebus, PhD
Katedra Data Science, Wydział Nauk Ekonomicznych, Uniwerystet Warszawski
Data Juice Lab sp. z o.o.

Modelowanie Ryzyka Kredytowego

Większość nowych badań w zakresie modelowania ryzyka kredytowego rekomenduje stosownie **coraz bardziej złożonych modeli** (Leo et al., 2019).

W modelowaniu ryzyka kredytowego proponowane są:

1. Drzewa decyzyjne (Davis et al., 1992; Frydman et al., 1985;)
2. Algorytm k-najbliższych sąsiadów (KNN) (Brown & Mues, 2012; Henley & Hand, 1996),
3. Maszyna wektorów nośnych (Vapnik, 1995);
4. **Regresja Logistyczna** (Campbell et al., 2008; Siddiqi, 2005) & probit regression (Tsaih et al., 2004),
5. Sieci neuronowe (ANNs) (Malhotra & Malhotra, 2002; Thomas et al., 2002),
6. Lasy losowe (Kruppa et al., 2013),
7. Heterogeniczne i homogeniczne modele zespolone (Lessmann et al., 2015).

Ogólna zasada: im **bardziej elastyczny/złożony** model tym **lepsza zdolność prognostyczna**.

Uwaga: Rudin, C. (2019). *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*.

Modelowanie Ryzyka Kredytowego w dobie AI – podstawowe regulacje

1. W finansach i finansach ilościowych **stosowane są coraz bardziej złożone systemy ML/AI**.
2. Bardzo **często bez głębszego zrozumienia**, w jaki sposób **podjęto decyzję**.
3. W wysoce uregulowanej dziedzinie aplikacji **przepisy i regulacje warunkują potrzebę stosowania narzędzi XAI**, ponieważ systemy AI często nie są zgodne z prawem (Weber et al. 2020).
   - Amerykańska ustawa o przejrzystości finansowej z 2021 r. (FTA) (Maloney 2021)
   - Unijna ustawa o sztucznej inteligencji (AIA) (Komisja Europejska 2021) zaleca bardzo wysoki poziom przejrzystości przy stosowaniu procesu decyzyjnego wspieranego przez sztuczną inteligencję w praktyce (Hoofnagle 2013; Elliott i in. 2021).
   - RODO wymaga, aby decyzje zawsze należały do ludzi (art. 22 RODO), co wymaga pewnego zaufania ze strony pracowników i ich zdolności do śledzenia decyzji podejmowanych przez sztuczną inteligencję.

Fig. 5 Publication trend of XAI research in Finance ($n = 60$)

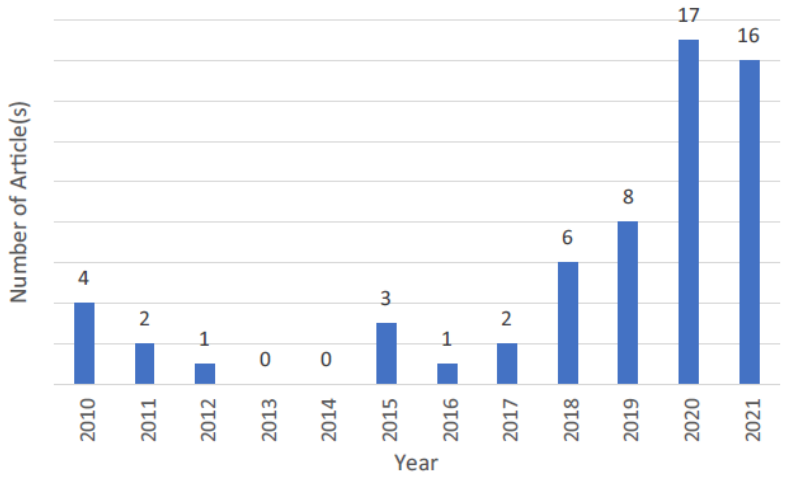
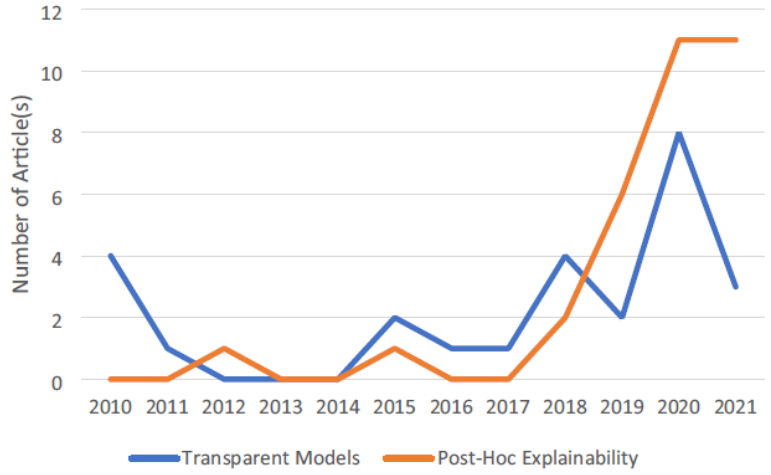
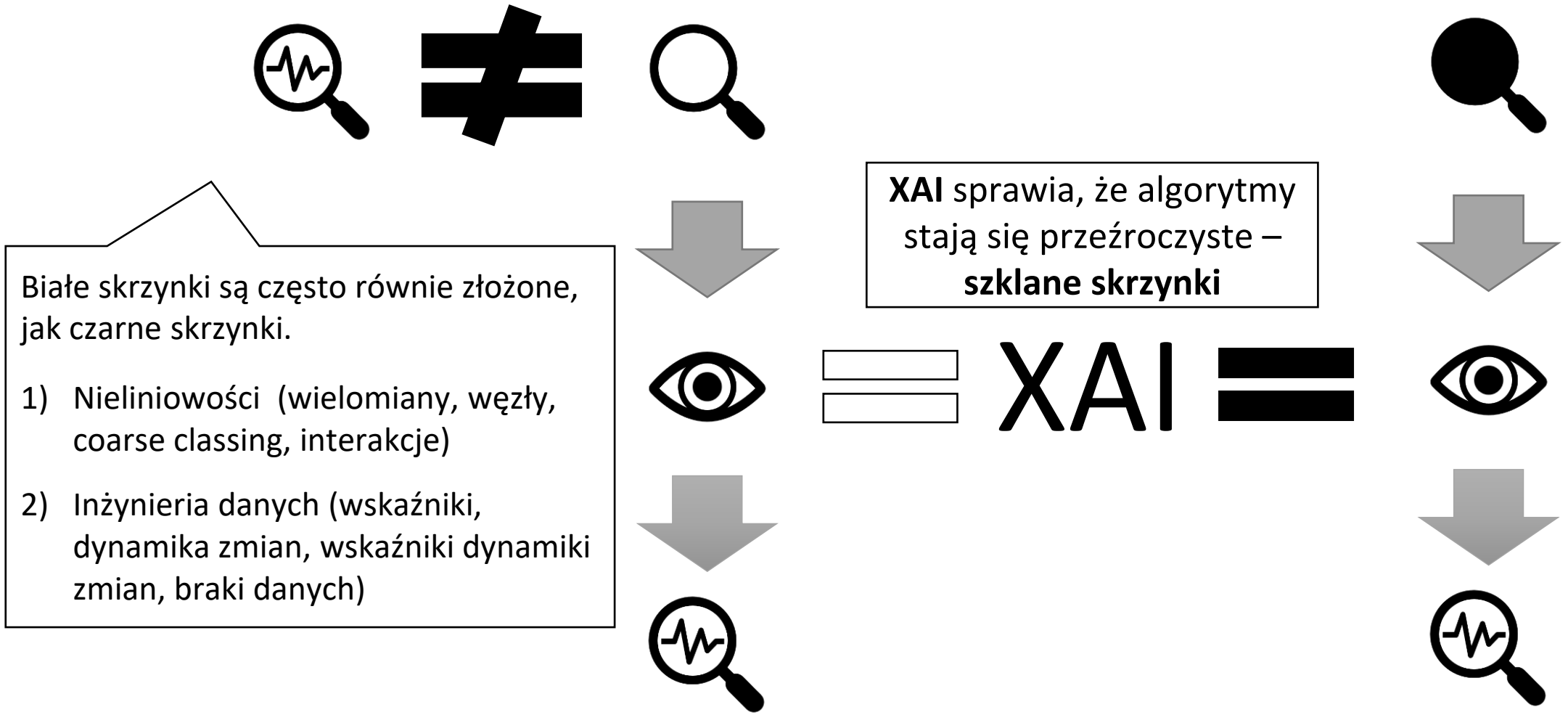


Fig. 8 Publication trend of XAI research in Finance regarding transparent models and post-hoc explainability ($n = 60$)



Weber, P., Carl, K.V. & Hinz, O. Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. *Manag Rev Q* (2023). https://doi.org/10.1007/s11301-023-00320-0

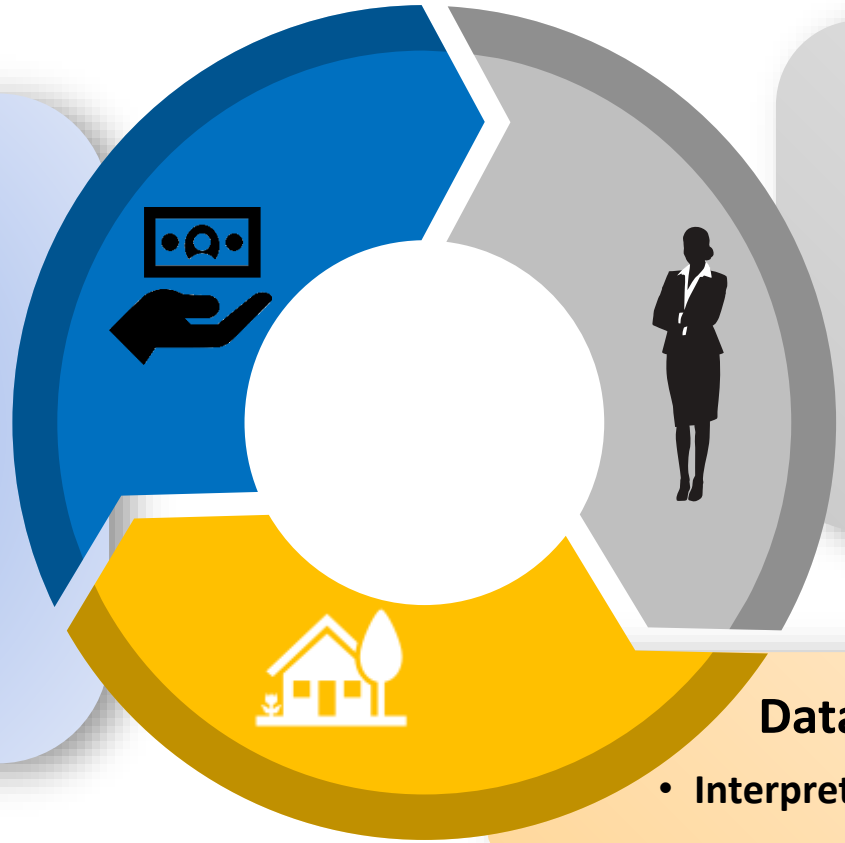
Białe i czarne skrzynki vs przeźroczyste skrzynki



Potrzeba wyjaśnienia sposobu działania algorytmów - interesariusze

Właściciele biznesowi

- Zrozumienie jak model podejmuje decyzje
- Analizy typu „What-If”
- Mitygacja ryzyka reputacji i operacyjnego
- Zarządzanie ryzykiem nadużycia



Regulatorzy

- GDPR
- ETHICS GUIDELINES FOR TRUSTWORTHY AI
- EBA REPORT ON BIG DATA AND ADVANCED ANALYTICS
- WHITE PAPER: On Artificial Intelligence - A European approach to excellence and trust
- Know your model

Data Scientists (Zespół deweloperski)

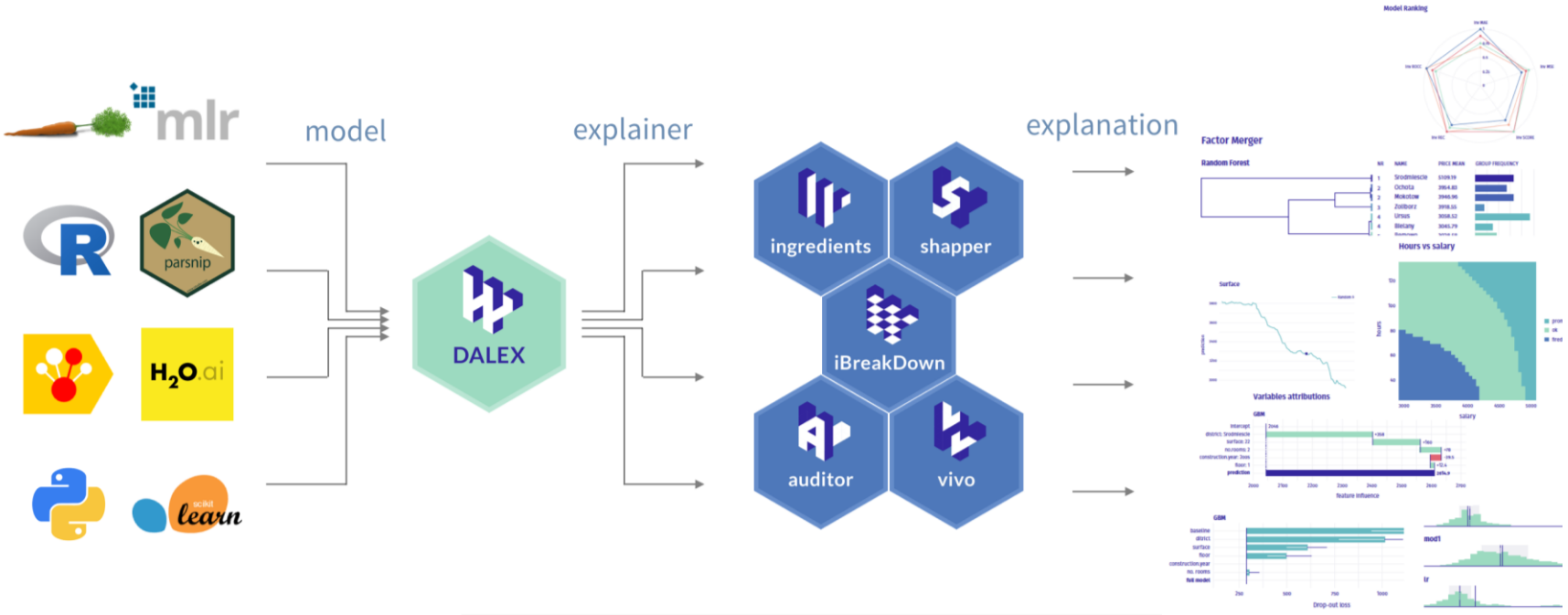
- Interpretowalność, stabilność i zaufanie do modelu

XAI jako element najlepszych praktyk (1/2)

1. **Metody agnostyczne** - **te same metody mogą być aplikowane na (praktycznie) wszystkie algorytmy predykcyjne**: LIME (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), DALEX (Biecek, 2018) lub InterpretML (Nori et al., 2019).

2. Kluczowe metody analizy pomagające w wytłumaczeniu algorytmów:
   - **Poziom modelu (globalny)**:
     - **Agnostyczne metody analizy ważności zmiennych: LOCO** (Lehmann & Romano, 2006) lub **PFI** (Fisher et al., 2018). Ważność zmiennej jest mierzona jako **krańcowa utrata jakości modelu w wyniku pominięcia wybranej zmiennej. Ważność jest liczona wtedy na podstawie**: AUROC, Gini, K-S, IV czy AUPRC measures.
     - **Analizy scenariuszowe** (**What-if)** takie jak **PDP** (Friedman, 2001), **ALE** (Apley & Zhou, 2016) lub wykres zależności **SHAP dependence plot** (Lundberg et al., 2020). Wykresy tego typu pokazują krańcowy wpływ jednej lub kliku zmiennych na prognozę z modelu. Zależność może być liniowa, nieliniowa monotoniczna, a nawet silnie niemonotoniczna.
   - **Poziom obserwacji (lokalny):**
     - **Shapley values (Strumbelj & Kononenko, 2014), SHAP values (Lundberg & Lee, 2017) lub profile Break Down (Gosiewska & Biecek, 2020). Wykresy tego typu przestawiają wpływ zmiennych na prognozę dla wybranej instancji i dzięki temu pozwalają zrozumieć, dlaczego model podjął określoną decyzję.**
     - **Analizy scenariuszowe (What-if) na poziomie pojedynczej obserwacji to wykresy Ceteris Paribus (Goldstein et al., 2015).**

XAI jako element najlepszych praktyk (2/2)



https://github.com/ModelOriented/DrWhy/blob/master/README.md

Shap

https://github.com/slundberg/shap

https://github.com/interpretml

Studium przypadku:

Enabling Machine Learning Algorithms for Credit Scoring - Explainable Artificial Intelligence (XAI) methods for clear understanding complex predictive models.

Przemysław Biecek, PhD, Marcin Chlebus, PhD, Janusz Gajda, PhD, Alicja Gosiewska, Anna Kozak, Dominik Ogonowski, Jakub Sztachelski, Piotr Wojewnik, PhD;

Cel badania

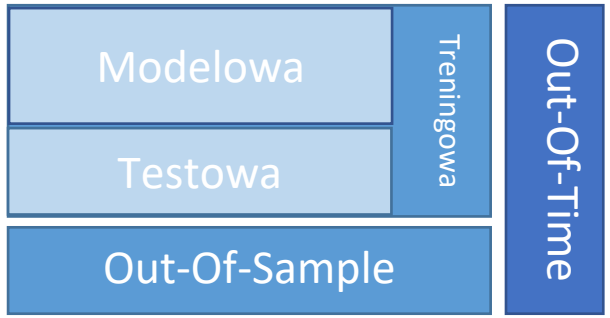
1. **Porównanie algorytmów stosowanych zgodnie z najlepszymi praktykami w modelowaniu ryzyka kredytowego (regresja logistyczna, regresja logistyczna z przekształceniami WoE) z nowoczesnymi algorytmami uczenia maszynowego (Lasy losowe, Extreme Gradient Boosting Machine, Gradient Boosting Machine, logistyczna regresja regularyzowana, sieci neuronowe i inne) na bardzo dużym zbiorze danych otrzymanych od Biura Informacji Kredytowej S.A. zawierającymi dane behawioralne klientów firm pożyczkowych.**

2. **Propozycja poszerzenia najlepszych praktyk o narzędzia XAI, które pozwalają tłumaczyć działanie algorytmów decyzyjnych, a dzięki temu ułatwiają stosowanie tych algorytmów w rzeczywistości.**

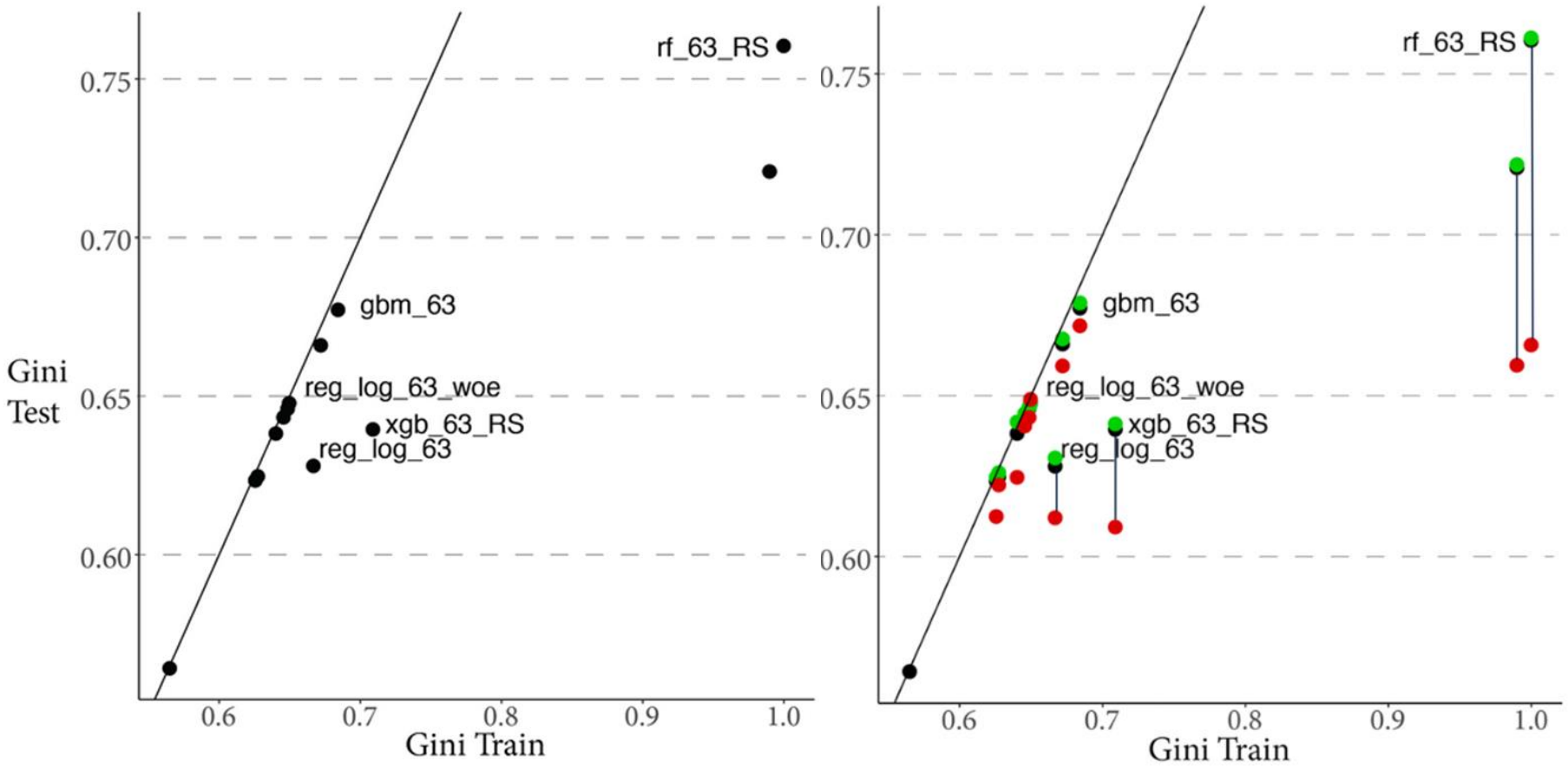
Zbiór danych – BIK S.A.

Źródło	# Obserwacji	# Zmienne	Okres obserwacji	Flaga default	Próbki: Modelowa Testowa Out-of-sample	Próbka: Out-of-time
BIK S.A. Firmy pożyczkowe online	> 5 mio	1 729	10.2017-05.2019	6 M	10.2017-08.2018	09.2018-11.2018

Struktura danych:



Porównanie modeli– zdolność dyskryminacyjna



Las losowy:
1. **Train: silnie douczony (głębokie drzewa)**
2. **Test & OOS: 1-szy**
3. **OOT: 2-gi i widoczny spadek – dodatkowa niepewność**

GBM:
1. **Train: podobne do Test, OOS & OOT**
2. **Test & OOS: 2-gi**
3. **OOT: 1-szy i stabilny**

Regresja logistyczna z WoE:
1. **Train: podobne do Test, OOS & OOT**
2. **Test & OOS: 3-ci**
3. **OOT: 3-ci i stabilny**

- Testing sample
- Out-of-sample
- Out-of-time

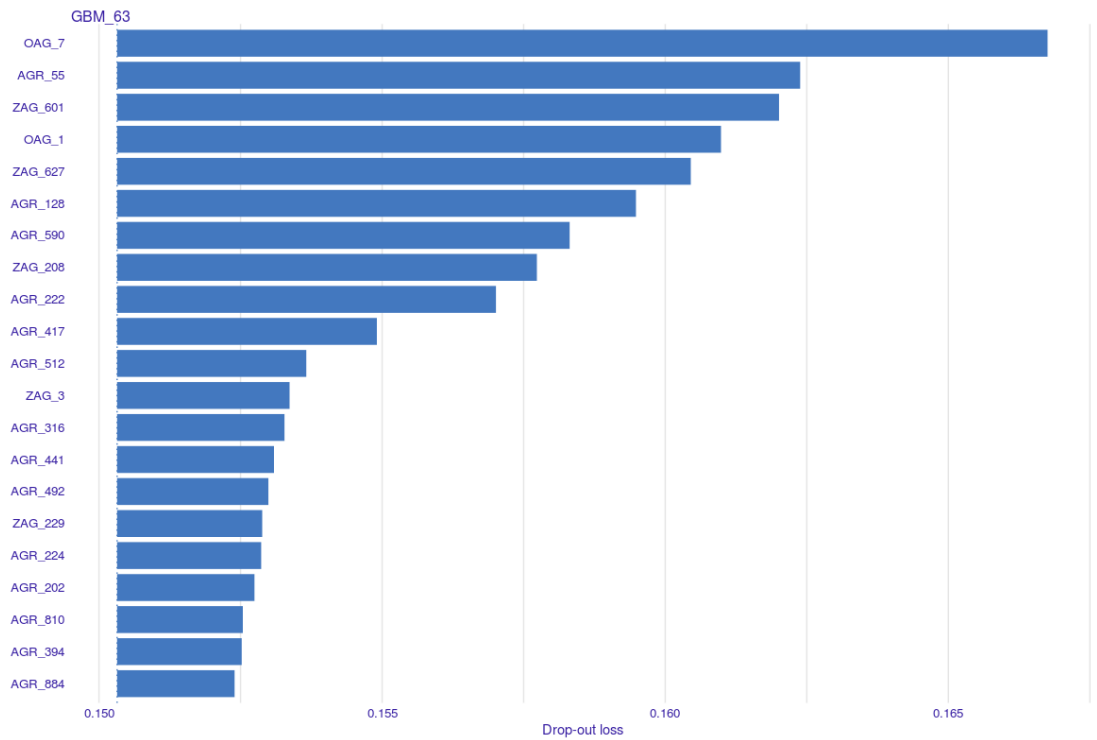
Porównanie modeli– zdolność dyskryminacyjna i czas trenowania/prognozy

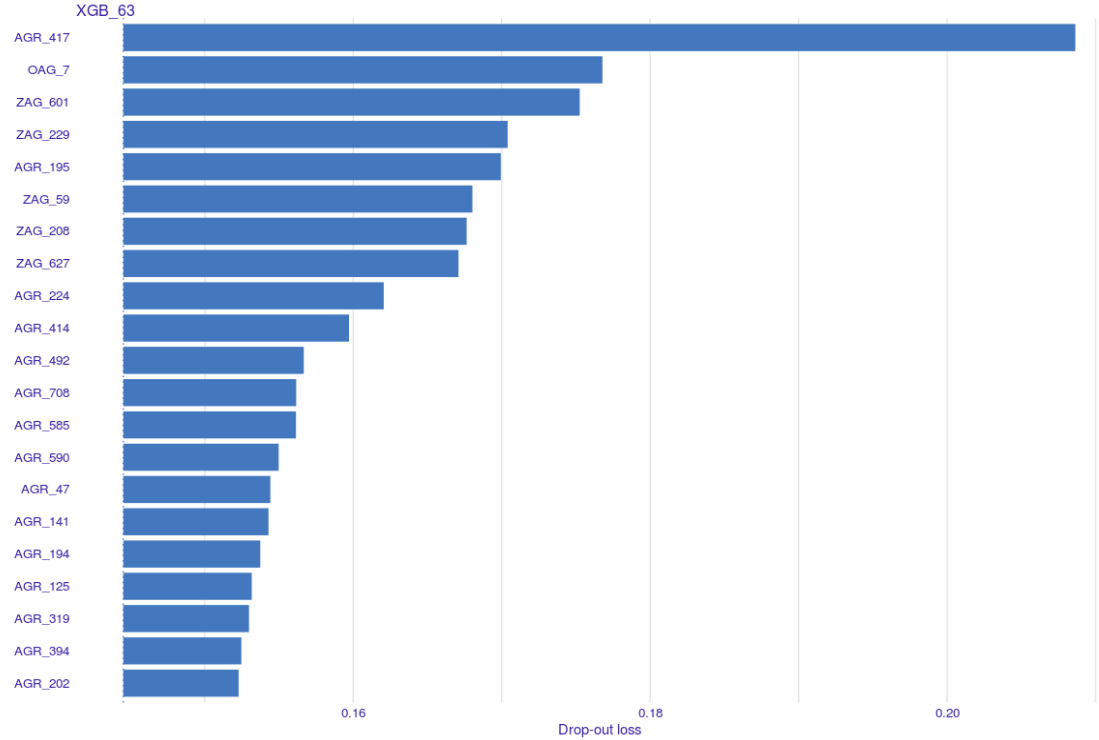
Model	Gini Train	Gini Test	Gni out-of-sample	Gini out-of-time	K-S out-of-time	Czas trenowania (min)	Czas predykcji (sec)	Kometarz
rf_63_RS	1,00	**0,76**	**0,76**	**0,67**	0,51	40	**7**	**1-1000 obs the same**
gbm_63	0,68	**0,68**	**0,68**	**0,67**	0,52	30	**0,01**	**1-1000 obs the same**
reg_log_63_woe	0,65	**0,65**	**0,65**	**0,64**	0,49	0,17	<0,01	
reg_log_63	0,65	0,65	0,65	0,62	0,49	1,50	<0,01	
xgb_63_RS	0,71	0,64	0,64	0,61	0,46	0,20	0,01	

Ze względy na stabilność wyników **GBM został wybrany jako najlepszy**.

Co więcej, czas prognozy w GBM wyraźnie niższy niż dla RF

Porównanie modeli– ważność zmiennych





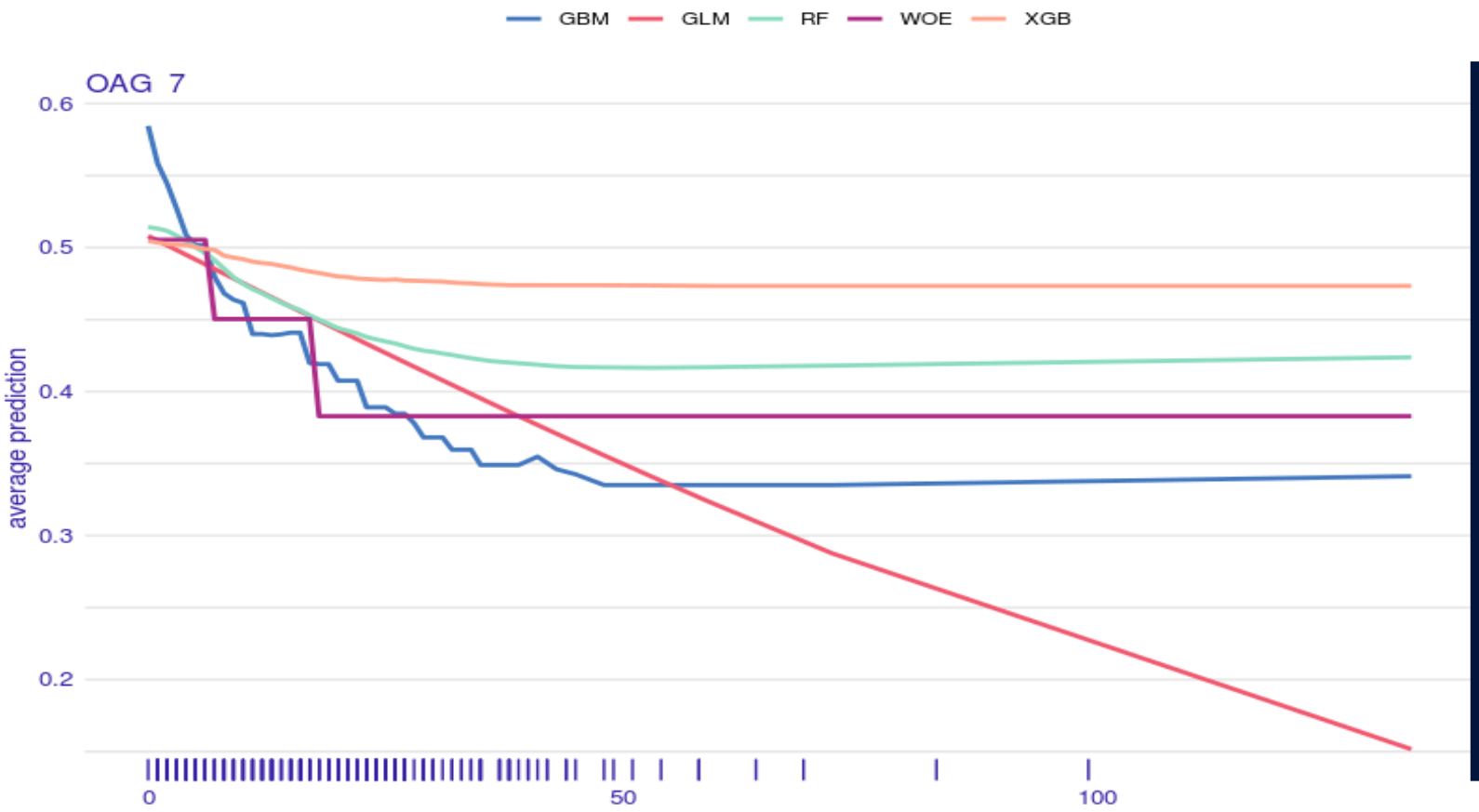
PFI jest mierzony jako krańcowy spadek wartości **AUROC/Gini** → wpływ poszczególnych zmiennych jest podobny między modelami

Wyniki:

1. **OAG_7 spadek o ~3,5 p.p.** w skali **Gini** (GBM i XGBoost)
2. Wiele tych samych zmiennych jest podobnie ważna (OAG_7, ZAG_601, ZAG_627)
3. **W XGB wyraźny wpływ zmiennej AGR_417**, potencjalna przyczyna przeuczenia się algorytmu

Marcin Chlebus, PhD, WNE UW & Data Juice Lab, m.chlebus@uw.edu.pl

Porównanie modeli – Partial Dependency Profile

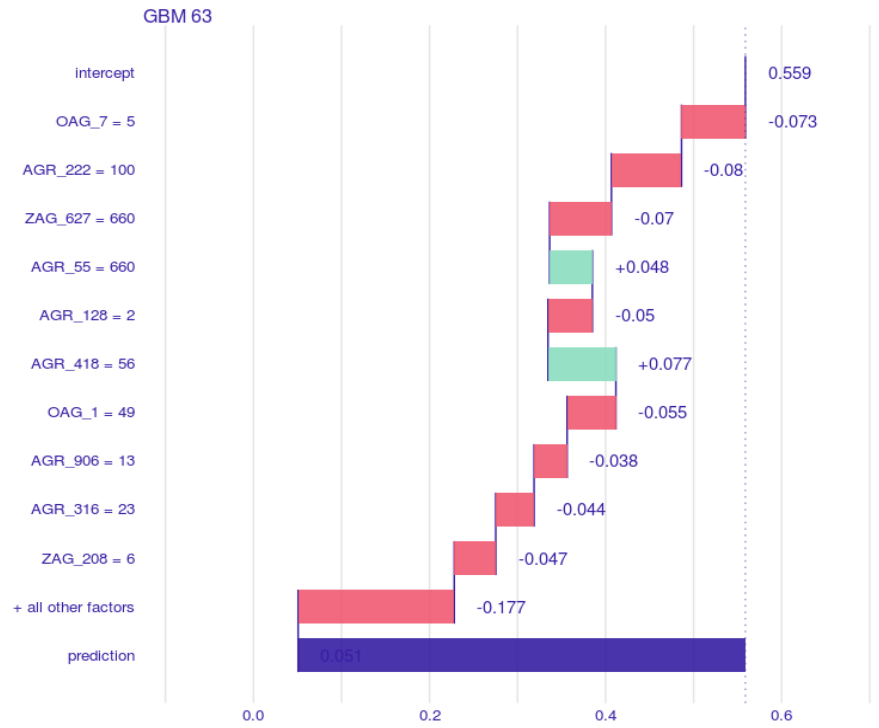


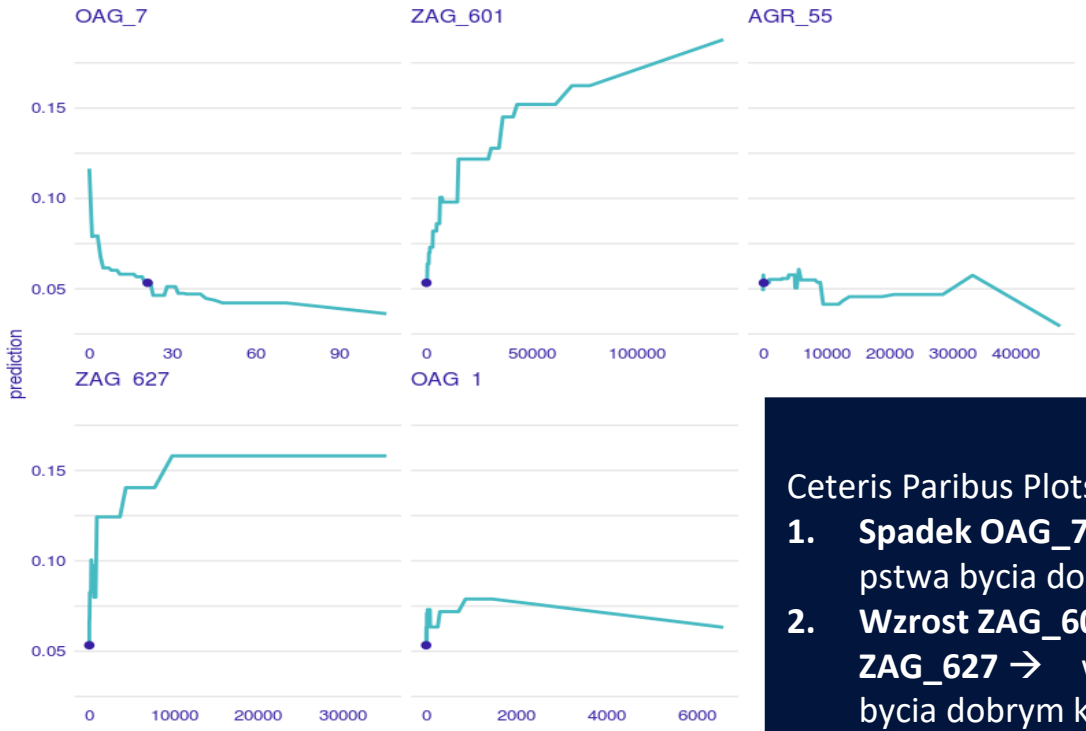
Wszystkie algorytmy prezentują podobną relację między zmienną OAG_7 **i średnią prognozą modelu**

Regresja logistyczna z WoE i **GBM** opisują tę zależność w bardzo podobny sposób

GMB jest bardziej elastyczne, więc opisać tę relację dokładniej.

Interpretacja na poziomie obserwacji – Break Down i Ceteris Paribus Plot





Ceteris Paribus Plots:

1. **Spadek OAG_7** → wzrost pstwa bycia dobrym klientem
2. **Wzrost ZAG_601 lub ZAG_627** → wzrost pstwa bycia dobrym klientem
3. **AGR_55 oraz OAG_1 nie ma wyraźnego wpływy na prognozę pstwa**

Profil Break Down dekomponuje prognozę na wszystkie zmienne niezależne i pokazuje ich wpływ **dla konkretnej obserwacji**:

1. **Średnie pstwo bycia dobrym klientem wynosi ~56%**
2. **OAG_7=5** obniża pstwo o **7,3 p.p.** do **~48%**
3. **AGR_222** obniża pstwo o **8 p.p.** do **~40%**
4. Finalnie, **pstwo bycia dobrym klientem spada do 5% (PD=95%)**

Marcin Chlebus, PhD, WNE UW & Data Juice Lab, m.chlebus@uw.edu.pl

Dziękuję za uwagę!

Marcin Chlebus, PhD

Koordynator programu Data Science & Business Analytics,
Katedra Data Science, Wydział Nauk Ekonomicznych, Uniwersytet Warszawski
m.chlebus@uw.edu.pl