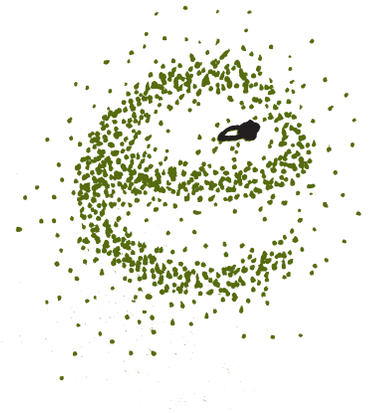
Skoro wszystkie modele są błędne, to które są użyteczne?

*Przemysław BIECEK**

* MI2.AI, Uniwersytet Warszawski, Politechnika Warszawska

Jedno z najbardziej znanych powiedzeń związanych z modelowaniem statystycznym, ukute przez George'a Boxa, brzmi: „Wszystkie modele są błędne, ale niektóre są użyteczne". To krótkie zdanie stanowi głęboką refleksję nad zastosowaniami statystyki. Z definicji modele stanowią uproszczony opis rzeczywistości. Gdy analizujemy złożone zjawiska (np. ekonomiczne, przyrodnicze, medyczne...), to niemożliwe jest stworzenie modelu uwzględniającego wszystkie istotne czynniki. Pomimo tego nawet niedoskonałe modele potrafią być przydatne. Używano ich przez dziesięciolecia, by formułować pytania i weryfikować hipotezy dotyczące otaczającego nas świata, takie jak: Czy analizowana terapia przynosi pozytywne efekty medyczne?, Czy inwestycje w edukację przekładają się na wyniki uczniów? To tylko dwa przykłady zagadnień badawczych, na które można odpowiedzieć za pomocą modeli statystycznych. Takie podejście jest podstawą wszystkich nauk empirycznych.

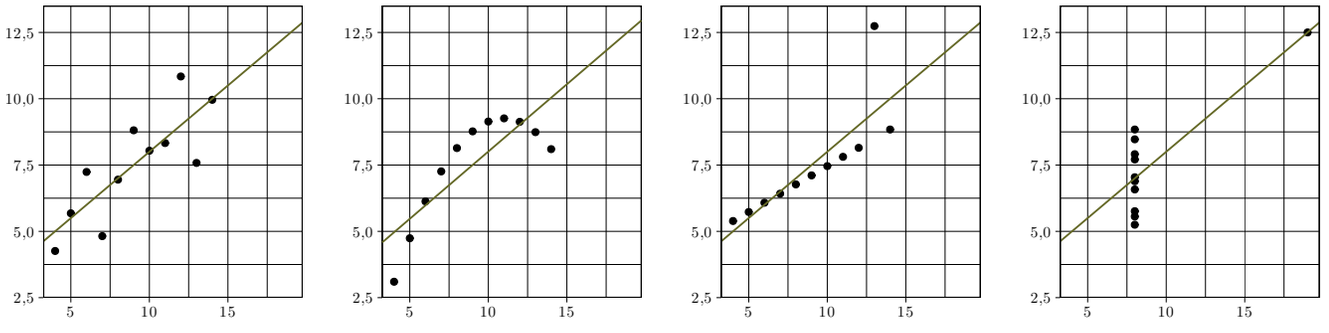
G. Box, „Science and statistics", Journal of the American Statistical Association (1976).



Nawet jeżeli zgodzimy się z powiedzeniem George'a Boxa, to pozostaje ważne pytanie: jak stwierdzić, które modele są przydatne? *Niektóre*, ale nie wszystkie, więc wybór właściwego modelu jest kluczowy dla poprawnego wnioskowania. Standardowe podejście do tego zagadnienia polega na ustaleniu pewnego kryterium jakości dopasowania do analizowanych danych, a następnie wybraniu modelu, który ma najlepszy wynik w sensie wybranego kryterium. Istnieje kilka popularnych kryteriów używanych przez statystyków, w tym: RMSE, R^2, AIC, BIC (wybacz, Drogi Czytelniku, że nie rozwijam skrótów ani nie zdradzam technicznych szczegółów, jednak dla tego tekstu nie mają one istotnego znaczenia); kilka innych stosowanych jest przez osoby zajmujące się uczeniem maszynowym. Ogólna procedura pozostaje taka sama: zaczynamy od grupy kandydujących modeli, wybieramy najlepszy według określonego kryterium i uznajemy go za najlepszy opis rzeczywistości.

Takie podejście może prowadzić do ciekawych niespodzianek, jak również paradoksów. Jednym z nich jest opracowany 50 lat temu (dokładnie!) *kwartet Anscombe'a*. Francis Anscombe przedstawił cztery sztucznie utworzone zestawy danych, każdy z nich składał się z jedenastu par liczb (można sobie wyobrazić, że były to wzrost i waga jedenastu noworodków). Dane były tak skonstruowane, że dla każdego zestawu najlepiej dopasowana liniowa zależność (w sensie kryterium R^2) była taka sama (i dawała tę samą wartość R^2). Jednak każdy z tych zestawów danych opowiada zupełnie inną historię! Aby to odkryć, najlepiej przedstawić dane w sposób graficzny z użyciem wykresu kropkowego, jak na rysunku 1.

F. Anscombe, „Graphs in Statistical Analysis", American Statistician (1973).



Rys 1. Kwartet Anscombe'a. Dla każdego zbioru danych model $y = x/2 + 3$ ma najlepsze wśród modeli liniowych dopasowanie do danych ze współczynnikiem determinacji $R^2 = 0{,}66$

Nawet najlepiej dopasowany do danych model liniowy może prowadzić do błędnych wniosków dotyczących występujących w nich zależności. Analiza wizualna może w takich przypadkach uzupełniać wnioskowanie statystyczne, dlatego też wielu znanych statystyków przez dekady proponowało nowe metody wizualizacji danych, które dziś określa się mianem narzędzi *eksploracyjnej analizy danych*.

J. Tukey, „Exploratory Data Analysis", Pearson (1977).

Anscombe pokazał, że różne zestawy danych opisujące różne historie mogą dawać ten sam model. Czy jednak może być odwrotnie? Czy jeden zbiór danych może prowadzić do kilku różnych modeli przedstawiających różne historie, a jednak tak samo dopasowanych do danych? Twierdzącej odpowiedzi na to pytanie udzielił Leo Breiman w opiniotwórczym artykule „The Two Cultures". Opisane przez niego zjawisko znane jest dzisiaj pod nazwą „perspektywa Rashomona" lub „mnogość dobrych modeli". Nazwa „Rashomon" odnosi się do filmu Akiry Kurosawy z 1950 roku, opisującego pewne wydarzenie z perspektywy czterech świadków, z których każdy przedstawia inną relację tego, co się wydarzyło. Relacje są tak różne, że nie sposób odgadnąć, jak było naprawdę. Breiman użył tego terminu, aby opisać hipotetyczny scenariusz, w którym kilka modeli ma równie dobre dopasowanie do danych, ale modele w różny sposób je *objaśniają*. Taka sytuacja stawia pod znakiem zapytania wszelkie wnioskowania oparte na „jednym najlepszym" modelu. Na przykład, co należy zrobić w obliczu dwóch modeli, które skutkują różnymi wnioskami co do skuteczności danej terapii medycznej? Któremu z tych modeli należy zaufać, jeśli oba są równie dobrze dopasowane do danych?

L. Breiman, „Statistical Modeling: The Two Cultures", Statistical Science (2001).

Aby zilustrować problem mnogości dobrych modeli, Breiman przedstawił przykład kilku modeli liniowych prowadzących do różnych wniosków na temat danych, które jednocześnie były do nich w tym samym stopniu dopasowane, a stopień tego dopasowania różnił się nieznacznie od najlepszego możliwego dopasowania. Aby uczynić to zjawisko jeszcze bardziej wyraźnym, skonstruowaliśmy niedawno *kwartet Rashomona*. W naszym artykule przedstawiamy drzewo regresji, las losowy, sieć neuronową i model liniowy (szczegóły dotyczące tych modeli nie są w tym tekście dla nas istotne). Wszystkie zostały, w miarę numerycznych możliwości, najlepiej dopasowane do tych samych danych i miały takie samo dopasowanie, a jednocześnie opisywały *całkowicie różne historie*.

P. Biecek, H. Baniecki, M. Krzyziński, and D. Cook, „Performance is not enough: a story of the Rashomon's quartet" arxiv (2023).

. . . Chwileczkę, ale skąd mamy wiedzieć, jakie historie pokazują tak złożone modele jak sieć neuronowa czy las losowy złożony z setek drzew? Z pomocą przychodzą nam techniki wizualizacji rozwijane pod nazwą eXplainable Artificial Intelligence (XAI) lub Explanatory Model Analysis (EMA). Jedną z nich jest analiza zależności częściowej modelu (Partial Dependence, PD), technika zaproponowana przez Jerome'a Friedmana w jego słynnej pracy o metodzie *boosting*. Profile PD można wyznaczyć dla modeli o dowolnej strukturze. Ze względu na tę uniwersalność metoda ta szybko znalazła wiele zastosowań.

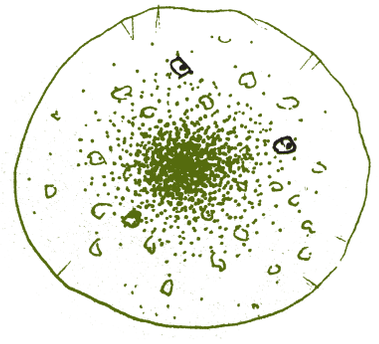
J. Friedman, „Greedy Function Approximation: A Gradient Boosting Machine", Annals of Statistics (2000).

Przedstawmy intuicję stojącą za profilem zależności częściowej. Wartość PD dla wybranej zmiennej S i jej potencjalnej wartości t to średnie przewidywanie modelu dla dostępnych danych, w których wartość tej zmiennej została „na siłę" ustalona wszędzie jako t.

Aby opisać rzecz bardziej formalnie, załóżmy, że mamy N obserwacji p *zmiennych wejściowych* (np. dla N noworodków rejestrujemy wzrost, wagę, kolor oczu. . .), gdzie i-ta obserwacja to $(x_{i,1}, \ldots, x_{i,p})$. Na podstawie tych zmiennych chcemy *coś* przewidywać (to *coś* zwykło się nazywać *zmienną docelową*) i w tym celu używamy funkcji $f(x_1, \ldots, x_p)$, która jest naszym *modelem*. *Zależność częściowa* s-tej zmiennej to funkcja określona wzorem

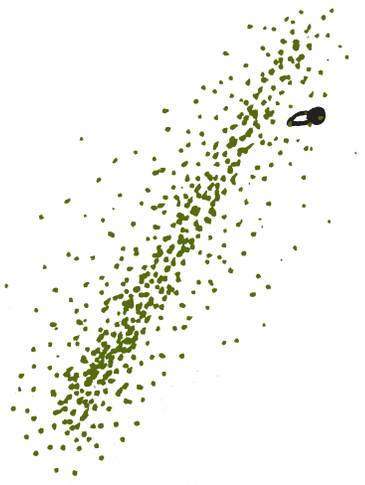
$$PD^s(t) = \frac{1}{N}\sum_{i=1}^{N} f(x_{i,1}, \ldots, x_{i,s-1}, t, x_{i,s+1}, \ldots, x_{i,p}).$$

P. Biecek, T. Burzykowski, „Explanatory Model Analysis", CRC Press (2021). `https://ema.drwhy.ai/`



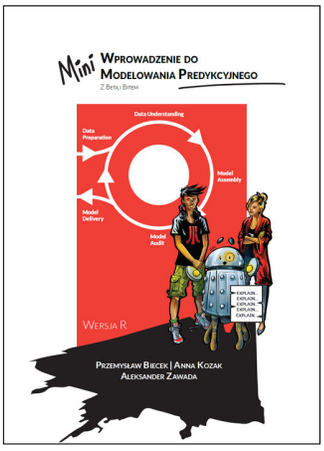
Profil PD mówi nam zatem, jak zmienia się przewidywana wartość zmiennej docelowej, gdy zmienia się wartość interesującej nas zmiennej wejściowej, przy jednoczesnym utrzymaniu wszystkich innych zmiennych wejściowych na wartościach obserwowanych w analizowanych danych.

Profile PD każdego modelu w kwartecie Rashomona przedstawiono na rysunku 2, sugerując wyraźne i różne zależności między zmiennymi, szczególnie dla zmiennych X_2 i X_3. Na przykład zmienna X_2 jest nieistotna w pierwszym modelu (tzn. średnio jej zmiana nie wpływa na przewidywania modelu), ale ma dodatni efekt w pozostałych modelach (tzn. średnio jej zwiększenie powoduje wzrost przewidywań modelu). Z kolei zmienna X_3 jest nieistotna w pierwszym modelu, ma ujemny efekt w drugim modelu i dodatni w trzecim i czwartym. Jednak skoro



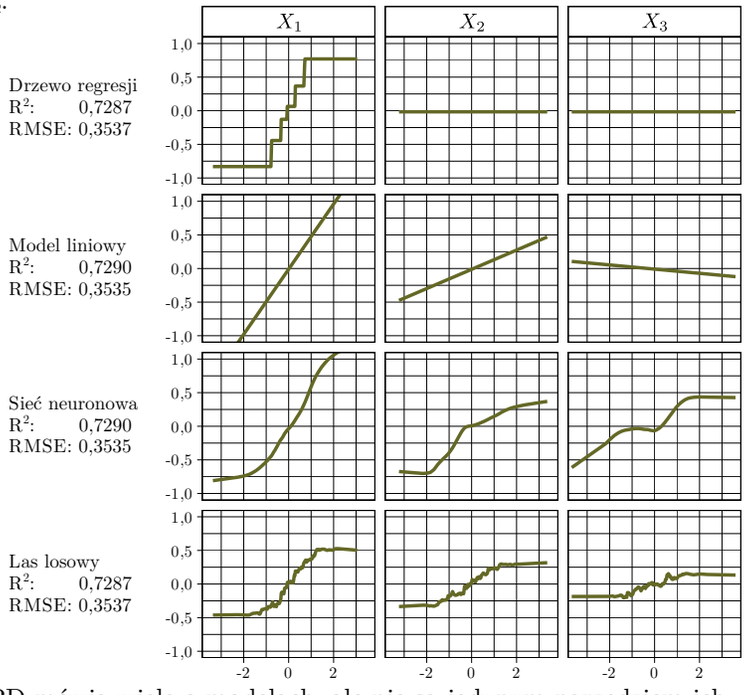
Rys. 2. Kwartet Rashomona. Każdy rząd opisuje inny model, a kolumny kolejne zmienne wejściowe. Panele pokazują profile zależności częściowej. Te cztery modele zostały dopasowane do jednego i tego samego zbioru danych. Przedstawione współczynniki dopasowania R^2 i RMSE pokazują, że wszystkie te modele są równie dobrze dopasowane do analizowanych danych

C. Rudin, Ch. Chen, Z. Chen, H. Huang, L. Semenova, Ch. Zhong, „Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges", arxiv (2021).



wszystkie modele mają takie samo dopasowanie do danych, to jak określić, który z tych opisów jest poprawny?

Któremu modelowi powinniśmy zaufać? A jeśli nie wiemy, to dlaczego mamy ufać któremukolwiek z nich? Ponieważ ze względu na losowe fluktuacje każdy z tych modeli potencjalnie mógłby zostać uznany za najlepszy, analizowanie tylko jednego, najlepszego modelu z pominięciem tych nieco gorszych może być złą strategią.



Profile PD mówią wiele o modelach, ale nie są jedynym narzędziem ich wizualizacji. Mimo swojej uniwersalności mają pewne ograniczenia, np. korelacje między zmiennymi lub interakcje mogą zniekształcić prezentowany obraz. Co więcej, analiza wpływu każdej zmiennej będzie trudna, jeśli model wykorzystuje setki lub tysiące zmiennych. Charakteryzacja zbioru najlepszych modeli jest wciąż otwartym problemem badawczym. Aby zrozumieć świat, nie możemy polegać wyłącznie na perspektywie jednego modelu, nawet jeśli ma on najlepsze wyniki w odpowiednich kryteriach. Musimy zestawiać ze sobą perspektywy różnych modeli, po to aby odróżnić hipotezy poparte danymi od hipotez, które są artefaktem wybranej struktury modelu.

Narzędzia do pełnego rozpoznawania sytuacji, w których możemy mieć do czynienia z perspektywą Rashomona, nie zostały jeszcze opracowane. Do czasu, aż to nastąpi, pozostaje nam eksploracja tych modeli z wykorzystaniem szeregu nowych technik wizualizacji. Kwartet Anscombe'a pokazał, że techniki wizualizacji danych są niezbędne do wnioskowania o naturze relacji między zmiennymi. Podobnie kwartet Rashomona pokazuje, że techniki wizualizacji modeli są równie przydatne.

Czytelniku, jeśli chcesz poznać kolejne metody wizualizacji i eksploracji modeli, sięgnij po komiks statystyczny „Mini Wprowadzenie do Modelowania Predykcyjnego" dostępny na stronie `https://betaandbit.github.io/MiniML/`.

Na przykładzie predykcji śmiertelności wirusa SARS-COV-2 omawia on proces trenowania modeli predykcyjnych, a także metody eksploracji tych modeli z wykorzystaniem profili zależności częściowej, wartości Shapleya czy profili Ceteris Paribus. Przedstawione powyżej przykłady oparte były na sztucznie wygenerowanych danych, jednak podobne wyzwania można napotkać, analizując problemy ze świata rzeczywistego.

Podsumowując, wszystkie modele są błędne, a my nie możemy być pewni, które z nich są przydatne. Możemy jednak przyjrzeć się wielu modelom jednocześnie, a spojrzenie na świat przez taką perspektywę jest niezbędne do oddzielenia zależności prawdziwie wspieranych przez dane od takich relacji, które są artefaktami wybranej techniki modelowania.