METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XXII/4, 2021, s. 124– 138

PORÓWNANIE SKUTECZNOŚCI DWÓCH KULTUR ANALITYCZNYCH

Borkowski Bolesław https://orcid.org/0000-0001-6073-6173
Instytut Ekonomii i Finansów
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: boleslaw_borkowski@sggw.edu.pl
Karwański Marek https://orcid.org/0000-0001-5192-7920
Szczesny Wiesław https://orcid.org/0000-0002-8083-4624
Instytut Informatyki Technicznej
Szkoła Główna Gospodarstwa Wiejskiego w Warszawie
e-mail: marek_karwanski@sggw.edu.pl; wieslaw_szczesny@sggw.edu.pl

Streszczenie: W pracy dokonaliśmy analizy porównawczej skuteczności dwóch kultur analitycznych: statystycznej analizy i modeli algorytmicznych wykorzystujących metody uczenia maszynowego (sieci neuronowe, lasy losowe, algorytmy wzmocnienia gradientowego). Materiałem badawczym były dane pochodzące z Polskiego Systemu Informacyjnego Lasów Państwowych (SILP) i dotyczyły one pożarów zarejestrowanych w lasach państwowych. Łączna liczba obserwacji wynosiła 36328 a maksymalna liczba cech uwzględnionych w badaniu wynosiła 16. W badanym okresie łączna liczba zarejestrowanych pożarów wynosiła 25660 (od 2007 r. do końca maja 2016 r.) wahała się średnio w roku od 571 w 2016 r. do 3 897 w 2015 r.
Analizę przyczyn wybuchu pożarów przeprowadziliśmy dwoma różnymi metodami, powszechnie znaną analizą statystyczną (model logitowy) i przy wykorzystaniu metod uczenia maszynowego, tj. lasy losowe, wzmocnienie gradientowe i sieci neuronowe). Celem pracy były badania skuteczności tych dwóch kultur analitycznych. Przeprowadzone badania nie wykazały zauważalnych różnic pomiędzy dwoma analizowanymi kulturami analitycznym w precyzji oszacowań modeli.

Słowa kluczowe: modele zmiennych jakościowych, metody uczenia maszynowego

JEL classification: C35, C45

https://doi.org/10.22630/MIBE.2021.22.4.11

WPROWADZENIE

Ostatnio, oprócz tradycyjnej analizy statystycznej, bardzo dynamicznie rozwijają się algorytmiczne metody analizy danych empirycznych, tj. sieci neuronowe, lasy losowe czy modele boostingowe. Mniej więcej w 2000 roku w „świecie statystyki" dokonał się rozłam. (Efron, 2020). Na początku XXI wieku społeczność statystyczna została wyraźnie podzielona na dwa obozy - statystykę (rozkłady prawdopodobieństwa) i uczenie maszynowe (algorytmiczne) . Żywy opis tej polaryzacji kulturowej przedstawił Leo Breiman (2001) w swoim wpływowym eseju zatytułowanym „Modelowanie statystyczne: dwie kultury". Breiman (2001) argumentował, że modele algorytmiczne są znacznie bardziej elastyczne, skalowalne i dokładne w przypadku złożonych problemów związanych z dużymi zbiorami danych. Natomiast tradycyjne metody statystyczne opierają się na założonych a priori modelach, głównie parametrycznych, które są odpowiednie dla małych zbiorów danych.

To stanowisko o wyższości algorytmicznego podejścia do analizy danych było podważane przez kilku wybitnych statystyków. Po pierwsze, pomimo przekonujących wyników prognoz, brak wyraźnego modelu generowania danych (podstawy probabilistycznej) może sprawić, że metody algorytmiczne będą mniej przydatne w badaniach naukowych, por. Cox (2001) i Parzen (2001). Efron (2020) wyartykułował to znakomicie: „Porzucenie modeli matematycznych jest bliskie porzucenia historycznego celu naukowego, jakim jest zrozumienie natury".

Po drugie, natura modeli algorytmicznych w postaci czarnej skrzynki sprawia, że są one nieprzeniknione, nieprzejrzyste i niełatwe do zinterpretowania. W rzeczywistości, jeśli chodzi o modele uczenia maszynowego, sam Breiman zgodził się, że „ich mechanizm tworzenia prognozy jest trudny do zrozumienia ...".

W dużej mierze kulturowy podział między tymi dwoma ramami stale narastał w ciągu ostatnich dziesięcioleci, zwłaszcza w erze deep-learningu.

Aby dokonać postępu, jesteśmy zmuszeni rozważyć pytanie: jak te dwie różne kultury pasują do siebie? Czy możemy zbudować ramy analizy danych, które łączą najlepsze pomysły z obu stron i oferują sposób na połączenie tych dwóch metod w jedno?

W kulturze statystycznej problemem jest wybór odpowiedniego modelu na podstawie całego dostępnego zbioru danych i jego ocena. Model ten wybieramy z rodziny rozkładów, która stanowi odpowiednik hiperparametrów w kulturze algorytmicznej. Parametry takiego modelu szacujemy zazwyczaj metodą OLS lub ML na podstawie próby i wnioskujemy w oparciu o wielkość teoretycznej funkcji straty/wiarogodności czy dokonaliśmy właściwego wyboru. Do oceny jakości parametrów modelu dostępne jest szerokie instrumentarium i wskazówki kiedy taki model możne być podstawowym narzędziem predykcji.

W przypadku dużych zbiorów danych znacząca część analityków statystycznych używa także metod uczenia maszynowego. W podejściu

algorytmicznym jest dostępnych wiele różnych metod uczenia maszynowego takich jak: k-najbliższych sąsiadów (k-NN), lasy losowe (RF), maszyny wektorów nośnych (SVM), maszyny gradient boostingu (GBM), sieci neuronowe (ANN) itp. Na podstawie zbioru uczącego poszukujemy modeli z bardzo wieloma zmiennymi na podstawie empirycznej funkcji strat. Wielkość tej funkcji będzie zależała od dopasowania rozkładu empirycznego, czyli od wyboru próbki.

Różnice w podejściu zależą od zastosowanego modelu – modele statystyczne wymagają założeń co do rozkładów ale nie wymaga to zbioru testowego, natomiast modele uczenia maszynowego wymagają osobnego oszacowania hiperparametrów na podstawie zbioru testowego. Najczęściej zbiór danych empirycznych dzieli się losowo na 3 części:

1. $D_{trenning}$ – dane treningowe, w dalszej części będziemy używać oznaczenia T1,
2. D_{test} – dane testowe, w dalszej części będziemy używać oznaczenia T2,
3. $D_{validate}$ - dane walidacyjne, w dalszej części będziemy używać oznaczenia V.

W badaniach niezbędny jest kompromis pomiędzy wielkością wolumenu danych dla tych trzech zbiorów, w praktyce najczęściej dzieli się dane w proporcji: 70%, 20%, 10%. Wyniki szkolenia będą lepsze dla większego $D_{trenning}$ ale test i walidacja będą bardziej niezawodne (mniej rozrzucone), jeśli D_{test} i $D_{validate}$ będą większe.

Zbiór uczący (60 – 80% oryginalnego zestawu danych) służy do poszukiwania najlepszych modeli (algorytmów), najczęściej znajdujemy wiele z nich. W fazie testowej wybieramy algorytm, który ma najniższą wartość ERM (Empiryczna Miara Ryzyka). Ostateczną kontrolę wybranego algorytmu przeprowadzamy na wydzielonym zbiorze walidacyjnych.

W literaturze coraz częściej w przypadkach, gdy używa się kilku modeli i w dodatku zbudowanych w oparciu o działania (obowiązujące reguły) w ramach obu kultur (statystyki i metod algorytmicznych), do porównania oceny modeli używa się bardzo często dokładności predykcyjnej oszacowanej na podstawie tzw. walidacyjnego zbioru danych. Czyli stosuje się wydzielenie dodatkowego zbioru danych do oceny modeli predykcyjnych zbudowanych w ramach pojęć i narzędzi w ramach danej kultury modelowania danych. Co więcej, aktualnie, jeśli liczba obserwacji (wolumen danych) jest duża w stosunku do liczby atrybutów (zmiennych) część badaczy budując modele według reguł z kultury statystycznej dodatkowo przeprowadza walidację na wydzielonym zbiorze nie biorącym udziału w tworzeniu modelu, mimo że podział na partycje zmniejsza dokładność estymatorów. Jest to jednak działanie dodatkowe, a nie obowiązkowe. Łączenie obu typów modeli powinno jednak uwzględniać ich specyfikę.

Celem pracy była analiza porównawcza dokładności predykcji prawdopodobieństwa wybuchu pożarów w lasach państwowych przy wykorzystaniu dwóch różnych podejść do analizy danych empirycznych, a mianowicie na podstawie modeli zbudowanych w oparciu o dwie różne kultury analityczne. W

przypadku statystycznej analizy i modelowania danych wykorzystaliśmy model logitowy. Natomiast w podejściu algorytmicznym zastosowaliśmy 3 różne klasy modeli: lasy losowe, gradient boosting i sieci neuronowe (ANN1 i ANN2).

METODY BADAWCZE

Metody statystyczne

- Regresja logistyczna

Jako model statystyczny wykorzystano regresją logistyczną. Regresja logistyczna to model procesu uczenia, który możemy opisać przez funkcję $f: X \rightarrow Y$, przy pomocy której szacujemy rozkład $P(Y|X)$ w przypadku, gdy Y ma wartości dyskretne, a $X = (X_1, \dots, X_k)$ jest wektorem zawierającym zmienne dyskretne lub ciągłe.

Rozważmy przypadek, w którym Y jest zmienną boolowską i regresja logistyczna przyjmuje postać parametryczną rozkładu logistycznego $P(Y|X, w)$. Szacujemy parametry w na podstawie danych uczących. Model parametryczny regresji logistycznej w tym przypadku to:

$$P(Y = 0|X) = \frac{1}{1 + exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}$$

$$P(Y = 1|X) = \frac{exp(w_0 + \sum_{i=1}^{n} w_i X_i)}{1 + exp\left(w_0 + \sum_{i=1}^{n} w_i X_i\right)}$$

gdzie w_i to parametry modelu.

Bardzo przydatną właściwością modelu logistycznego jest fakt, że prowadzi on do liniowego wyrażenia. Liniowy model logistyczny ma postać:

$$\log \frac{Pr(Y = 1|X = x)}{Pr(Y = 0|X = x)} = w_{k0} + w_k^T x$$

Model regresji logistycznej jest bardzo ogólny. Możemy zapisać wspólną gęstość X i Y jako

$$Pr(X, Y = k) = Pr(X) Pr(Y = k|X)$$

gdzie $Pr(X)$ oznacza brzegową gęstość zmiennej losowej X.

Model regresji logistycznej pozostawia brzegową gęstość X jako arbitralną funkcję $Pr(X)$ i dopasowuje parametry $Pr(Y|X)$ przez maksymalizację wiarogodność warunkowej. Prawdopodobieństwo brzegowe $Pr(X)$ jest całkowicie ignorowane.

Metody Uczenia maszynowego

- Las losowy

Las losowy jest zbiorem – klasyfikatorem zespołowym – opartym na drzewach decyzyjnych. Podstawową ideą jest uśrednienie wielu zaszumionych,

w przybliżeniu, nieobciążonych modeli drzew decyzyjnych, uzyskując tym samym zmniejszenie błędu. Drzewa są idealnymi kandydatami, ponieważ mogą być stosowane w przypadku złożonych relacji w strukturach danych, a przy dobraniu odpowiednich parametrów mają stosunkowo niskie obciążenie. Co więcej, każde drzewo wygenerowane, w lesie losowym ma identyczny rozkład. Wartość średnia uzyskiwana z takich drzew jest taka sama, oznacza to, że obciążenie drzew użytych w lesie losowym jest takie samo jak w przypadku pojedynczego drzewa, ale poprawę uzyskuje się poprzez redukcję wariancji. Tendencja do nadmiernego dopasowania modelu jest korygowana za pomocą agregacji bootstrapowych. Korelacja pomiędzy drzewami, będąca głównym problemem przy korzystaniu z zespołu modeli jest zmniejszana w wyniku użycia losowo wybranych podzbiorów próbek i losowo wybranych cech.

Niech T będzie zbiorem parametrów $\theta \in T$, podczas uczenia, na każdym etapie b, wybieramy do estymacji drzewa f_{rf}^{b} losowo niewielki podzbiór $T(\Theta_b) = T_b \subset T$ takich wartości. Po wybudowaniu B drzew predyktorem lasu losowego jest

$$\hat{f}_{rf}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} f_{rf}^{b}\big(T(\Theta_b)\big)$$

Θ_b – parametr charakteryzujący b-te drzewo w lesie losowym opisuje: wybrane zmienne, punkty cięcia i wartości w węzłach końcowych. Poprzez mechanizm losowości zmniejszamy korelację między dowolną parą drzew f_{rf}^{b} w zespole, a tym samym zmniejszamy wariancję $\hat{f}_{rf}^{B}(x)$.

- Gradient Boosting

Niech funkcja decyzyjna będzie klasyfikatorem zespołowym $H_T(x) = \sum_{t=1}^{T} \alpha_i h_t(x)$ .Niech ℓ oznacza (wypukłą i różniczkowalną) funkcję straty. Z niewielką stratą ogólności możemy napisać

$$\ell(H) = \frac{1}{n}\sum_{i=1}^{n} \ell(H(x_i), y_i) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\sum_{t=1}^{T} \alpha_i h_t(x), y_i\right)$$

gdzie α_i to parametry modelu, a $h_t(\,)$ to funkcje bazowe.

Modelowanie GB polega na znalezieniu rozwiązania, które minimalizuje funkcję strat $\ell(\,)$ poprzez sekwencyjne dodawanie nowych funkcji bazowych do rozwinięcia bez zmiany parametrów dopasowanych wcześniej. W iteracji *t* dodajemy do zespołu klasyfikator $\alpha_i h_t(x)$. Załóżmy, że w iteracji t mamy klasyfikator zespołowy $H_t(x)$. Teraz w iteracji *t* + 1 dodajemy h_{t+1}, który najbardziej minimalizuje stratę,

$$h_{t+1} = arg\min_{h \in H} \ell(H_t + \alpha h_t)$$

Najlepszym sposobem znalezienia minimum jest użycie algorytmu *gradient descent* w przestrzeni funkcji. Niestety gradient jest definiowany tylko w punktach danych treningowych x_i, podczas gdy ostatecznym celem jest uogólnienie $H_T(x)$ na nowe dane, które nie są reprezentowane w zbiorze uczącym. Możliwym rozwiązaniem tego dylematu jest budowa drzewa $T(x;\ \Theta_m)$ w m-tej iteracji, którego przewidywania t_m są jak najbardziej zbliżone do ujemnego gradientu. Prowadzi to do

$$\theta_m = arg\min_{\Theta} \sum_{i=1}^{N} \left(-g_{im} - T(x_i;\theta)\right)^2$$

Oznacza to, że dopasowuje się drzewo T do ujemnych wartości gradientu metodą najmniejszych kwadratów. Istnieją szybkie implementacje realizujące algorytm produkujący drzewa decyzyjne zgodne z metodą najmniejszych kwadratów.

- Sieci neuronowe

Sieci neuronowe są modelowane jako kolekcje przekształceń afinicznych - neuronów, polegające na mnożeniu przez macierze wag w i dodawaniu wektora odchylenia b, naprzemiennie z nieliniową transformacją ϕ: zwaną funkcją aktywacji:

$$f_{w,b}(X) = \phi\left(M_{w,b}^{(n)^T}\phi\left(M_{w,b}^{(n-1)^T}\phi\left(M_{w,b}{}^{(n-2)^T}\ldots\right)\right)\right)$$

$f_{w,b}$ jest modelem mapowania wejścia-wyjścia danej sieci neuronowej.

Model 1 sieci neuronowej.

Rozważmy sieć neuronową, której wejście jest zmienną losową X, a wyjście jest zmienną losową $Y = f_{w,b}(X)$, z parametrami b i w. Załóżmy, że sieć jest używana do aproksymacji docelowej zmiennej losowej Z. Niech funkcja błędu mierzy odległość między Y i docelową zmienną losową Z. Dobrym kandydatem jest oczekiwana wartość ich kwadratowej różnicy

$$C(w,b) = E[(Y-Z)^2] = E[\left(f_{w,b}(X) - Z\right)^2]$$

gdzie $E[\ ]$ oznacza operator wartości oczekiwanej.

1. Zbiór wszystkich całkowalnych do kwadratu zmiennych losowych na przestrzeni prawdopodobieństwa tworzy przestrzeń Hilberta z iloczynem skalarnym $\langle X,Y\rangle = E[XY]$. Definiuje to normę $\|X\|^2 = E[X^2]$, która indukuje odległość $d(X,Y) = \langle X — Y\rangle$. Jeżeli funkcja kosztu reprezentuje kwadrat odległości, $C(w,b) = d(Y,Z)^2$ wówczas minimalizacja kosztów jest równoznaczna ze znalezieniem parametrów (w,b) minimalizujących odległość między outputem Y i celem Z.
2. Sieć neuronowa przekształca wejściową zmienną losową X na zmienną losową $Y = f_{w,b}(X)$, która jest sparametryzowana przez w i b. Informacja generowana

przez zmienną losową $f_{w,b}(X)$ jest sigma algebrą $\mathcal{E}_{w,b} = \mathfrak{S}\left(f_{w,b}(X)\right)$. Wszystkie takie sigma algebry generują przestrzeń $\mathcal{E} = \bigvee_{w,b} \mathcal{E}_{w,b}$, , która jest sigma-algebrą generowaną przez sumę $\cup_{w,b} \mathcal{E}_{w,b}$ .. Problem można teraz sformułować w następujący sposób: Biorąc pod uwagę informację $\mathcal{E}$, znajdź najlepszą prognozę Z. Jest to zmienna losowa, oznaczona przez $Y = E[Z|\mathcal{E}]$ warunkową oczekiwaną wartością Z przy danej $\mathcal{E}$. Najlepszy predyktor, Y , jest określony przez $\mathcal{E}$ (tj. jest $\mathcal{E}$ -mierzalny) i znajduje się w najmniejszej możliwej odległości od Z.

Do zmniejszenie nadmiernego dopasowania używa się strategii regularyzacji. Jeden z najpowszechniejszych rodzajów regularyzacji polega na wprowadzeniu kary do normy parametru: Często używana kara L^2 polega na dodaniu członu regularyzacji $U(\theta) = \frac{1}{2}\|w\|_2^2$ do funkcji celu. Niestety często powoduje zjawisko znane jako zanik gradientu wagi i dlatego do symulacji użyto również innego modelu sieci.

Model 2 sieci neuronowej.

Niech p i q będą dwiema funkcjami gęstości prawdopodobieństwa przy zadanej mierze dx. Odległość Kullbacka-Leiblera DKL definiowana jest przez

$$DKL(p||q) = -\int p(x) \ln \frac{q(x)}{p(x)} dx \qquad (*)$$

Odległość Kullbacka-Leiblera może być uważane za funkcje kosztu dla sieci neuronowych.

Niech X będzie wejściową zmienną losową dla danej sieci neuronowej i niech $Y = f_\theta(X, \xi)$, gdzie $\theta = (w, b)$ i ξ jest zmienną losową oznaczającą szum w sieci. Oznaczmy docelową zmienną losową przez Z. Naszym zadaniem jest minimalizacja funkcji straty. Gęstość warunkową Y przy danym X, oznaczmy przez $q = p_\theta(y\,|x)$ i jest to warunkowa funkcja gęstości modelu. Wspólna gęstość na zbiorze treningowym (X,Z) oznaczmy przez $p = p_{X,Z}(x, z)$.

Wartość $\theta^* = \arg\min_\theta S\big(p_{X,Z}, p_\theta(Z|X)\big)\big)$ jest minimum funkcji kosztu (*).

Porzucenie (dropout) to metoda na zmniejszenie nadmiernego dopasowania, która sprawdza się w przypadku dużych rodzin sieci neuronowych. Główną ideą jest tymczasowe, losowe usunięcie neuronów (ukrytych) z sieci wprowadzające tylko liniowe zaburzenia.

Rozważmy N neuronów w ukrytej warstwie i usuńmy (losowo) jeden neuron, uzyskamy na wyjściu z neuronu i wartość $y_i = y - \lambda_i \sigma(w_i x + b_i)$ gdzie $y = \sum_{i=1}^N \lambda_i \sigma(w_i x + b_i)$ oznacza początkowy wynik. Korekta y_i jest podejmowana z prawdopodobieństwem $q_i = q = \frac{1}{N}$ co prowadzi do wyrażenia końcowego jako średniej z $\sum_{i=1}^n q_i y_i = \left(1 - \frac{1}{N}\right) y = (1 - q) y$, która jest proporcjonalna do wyniku początkowego sieci, y.

Miary użyte do porównywania modeli

Do porównywania skuteczności metod zastosowaliśmy następujące miary:

- krzywą ROC (Receiver Operating Characteristic) jako narzędzie do oceny poprawności klasyfikatora, która informuje o jego czułości i specyficzności. Jakość klasyfikacji za pomocą krzywej ROC można ocenić wyliczając różne wskaźniki, spośród których najpopularniejszy jest pole pod krzywą (AUC) (Area Under ROC Curve).
- AUC - Metryka oceny AUC jest obliczana jako pole pod krzywą ROC (ocena poprawności klasyfikatora) i jest skalarną reprezentacją oczekiwanej wydajności klasyfikatora. Współczynnik AUC zawsze ma wartość z przedziału od 0 do 1, przy czym wyższe wartości reprezentują lepszy klasyfikator.
- zliczeniowy R^2, definiowany jako R_{Square} (precision), w kontekście proporcji trafnych prognoz:

$$R^2 = \left\{\frac{liczba\ trafnych\ prognoz}{łączna\ liczba\ obserwacji}\right\}$$

- $R^2_{C\&S}$ (Cox and Snell 1989) definiowany następującym wyrażeniem:

$$R^2_{C\&S} = 1 - exp\left\{\frac{2[\log L(M) - \log L(0)]}{liczba\ obserwacji}\right\}$$

gdzie $\log L(M)$ i $\log L(0)$ to logarytmy zmaksymalizowanej wiarygodności dla dopasowanego (bieżącego) modelu i modelu „zerowego" zawierającego tylko wyraz wolny.

- 95% przedziały ufności WCL wyliczane z modelu Wald`a Confidence Limits)

Dane empiryczne

W badaniu wykorzystano materiał empiryczny dotyczący zarejestrowanych pożarów w lasach państwowych w okresie od 2007 roku do końca maja 2016 roku. Liczba obserwacji w próbie wynosiła 36328 a maksymalna liczba obserwacji 16. Do porównania skuteczności dwóch kultur analitycznych wykorzystaliśmy wyniki estymacji modeli logitowych, sieci neuronowych (ANN1 i ANN2), lasów losowych (RF) i gradient boostnig (BS) na próbach o różnej liczbie obserwacji.

WYNIKI ANALIZY EMPIRYCZNEJ

Zgodnie z metodologią używaną dla metod uczenia maszynowego badane zbiory (tak w pierwszej próbie jak i w drugiej) podzieliliśmy na losowo trzy rozłączne części: zbiór treningowy T (uczący) – 60% liczby obserwacji, zbiór testowy T – 30% liczby obserwacji i zbiór walidacyjny V – 10% liczby obserwacji. W celu zobrazowania różnic pomiędzy dwiema kulturami analitycznymi, na tych samych danych empirycznych, przeprowadziliśmy estymację parametrów wszystkich rozpatrywanych modeli na dwóch próbach o różnej wielkości obserwacji: na małej próbie (374) losowo wybranych obserwacji z badanej

populacji, co dało: (T1,T2,V) =(217, 122, 35) oraz na całej populacji (36328) obserwacji, czyli: (T1,T2,V) =(21885, 10759, 3684). Dodatkowo dla kontroli przeprowadziliśmy analogiczne obliczenia (zaprezentowane tylko na rysunkach) na próbie o liczności 122, podzielonej na (T1,T2,V)= (71, 22, 29). Każdorazowo (dla każdej z prób) wszystkie obliczenia realizowaliśmy na tym samym podziale na podzbiory T1,T2,V. Analizę porównawczą dla wszystkich metod, także dla modeli logistycznych, przeprowadziliśmy na zbiorze walidacyjnym, zgodnie z zasadą w podejściu algorytmicznym (nie jest to obligatoryjne w analizie statystycznej). Dodatkowo w przypadku modelu logistycznego analizę przeprowadzono na całej próbie ze wszystkimi potencjalnymi (dostępnymi) zmiennymi objaśniającymi (w tabeli jest to pierwszy model **logistic all var (T1+T1+V)**). Następnie dokonaliśmy walidacji tego modelu na zbiorze walidacyjnym V (na części zbioru, który był częścią zbioru użytego do budowy modelu), a wynik walidacji w wierszu Tabeli 1 i 2 oznaczyliśmy symbolem model **logistic all var V***. Dokonaliśmy także budowy modelu logistycznego ze wszystkimi zmiennymi na części obserwacji T1+T2 oraz jego walidacji na zbiorze V, wynik jest w wierszu oznaczony symbolem **logistic all var V**. Chcemy w tym miejscu wyraźnie zaznaczyć, że taki sposób postępowania nie jest wskazany w analizach statystycznych, w których dobór cech do modelu jest pierwszym bardzo ważnym elementem analizy. Z tego powodu, w przypadku modelu logistycznego jako pierwszy krok dokonano selekcji zmiennych, metodą regresji krokowej (tylko 6 zmiennych statystycznie istotnych w przypadku próby o liczności 374 , a 14 w przypadku tej o liczności 36328) – w tabeli oznaczony on jest jako logistic (T+T+V). W dalszej kolejności oszacowaliśmy model logistyczny z uwzględnieniem tylko wybranych cech. Analizę porównawczą skuteczności metod analizy przeprowadziliśmy na zbiorach walidacyjnym V, który brał udział w estymacji modelu, który został zbudowany zgodnie z praktyką statystyczną (czyli przeprowadzono wszystkie niezbędne testy). Wartość tej walidacji oznaczono w Tabeli symbolem **logistic V*.** Natomiast dodatkowo w celach porównawczych dokonaliśmy budowy według tej zasady (kultura statystyczna) na zbiorze T1+T2 i dokonaliśmy walidacji na zbiorze V. Wynik walidacji znajduje się w wierszu **logistic V.**

Wiersze zawierające w nazwie (T1+T2) oraz (T1+T2+V) oznaczają odpowiednio walidację modeli na zbiorach (T1+T2) oraz (T1+T2+V) . Natomiast te, które mają w nazwie tylko symbol V dotyczą walidacji modeli na zbiorze (V). Modele mające w nazwie (T1+T2) lub (V) zbudowane zostały na zbiorze (T1+T2), co w przypadku kultury „statystycznej" jest niepotrzebnym pomniejszeniem materiału służącego do budowy modelu.

Wyniki modeli zamieszczono w tabelach 1 i 2 jako **Logistic V**, **ANN1 V**, **ANN2 V**, **Gradient Boosting V** i **Random Forest V**. Należy jednak podkreślić, że do porównań zamiast **Logistic V** powinno się wziąć **Logistic V*,** gdyż postępowanie według zasad kultury statystycznej zapewnia „powtarzalność/stabilność" rozwiązania. Jednak w tym przypadku praktycznym różnice pomiędzy V* i V są nieistotne.

Tabela 1. Wyniki estymacji na całej populacji

	AUC	Standard Error	95% Wald Confidence Limits		R-Square (precision)	R-Square
Logistic all var (T+T+V)	0,8883	0,00169	0,8850	0,8916	1,0000	0,4176
Logistic all var V*	0,8950	0,00514	0,8849	0,9051	0,8170	0,4289
Logistic all var V	0,8946	0,00515	0,8845	0,9047	0,8157	0,4279
Logistic (T+T+V)	0,8883	0,00169	0,8850	0,8916	1,0000	0,4176
Logistic V*	0,8949	0,00514	0,8849	0,905	0,8168	0,4288
Logistic V	0,8946	0,00515	0,8845	0,9047	0,8165	0,4280
ANN1 T+T	0,8966	0,00171	0,8932	0,8999	0,8123	0,4287
ANN1 V	0,9039	0,0049	0,8943	0,9135	0,8293	0,4427
ANN2 T+T	0,8872	0,0018	0,8836	0,8907	0,7805	0,308
ANN2 V	0,8973	0,00509	0,8873	0,9072	0,7883	0,3302
Gradient Boosting T+T	0,8474	0,00214	0,8432	0,8516	0,7951	0,3308
Gradient Boosting V	0,8508	0,0063	0,8384	0,8631	0,7975	0,3371
Random Forest (T+T)	0,8941	0,00176	0,8907	0,8975	0,8190	0,4274
Random Forest V	0,8969	0,00518	0,8867	0,907	0,8244	0,4347

Źródło: Obliczenia własne na podstawie danych empirycznych z populacji o 36328 obserwacjach, (w podziale na T = 21885; T = 10759; V= 3684) gdzie: T+T+V oznacza estymację modelu na całej próbie, T+T – estymacja modelu na próbie treningowej plus testowej, V – walidacja modelu na wydzielonej części próby.

Dla dużych prób oszacowania parametrów modeli tak przy wykorzystaniu kultury algorytmicznej jak i kultury statystycznej są porównywalne. Nieznacznie lepszą precyzję dopasowania, mierzoną wielkością AUC, zaobserwowaliśmy w przypadku oszacowania parametrów metodą sieci neuronowych ANN1 V. Przedziały ufności, dla wskaźnika AUC oceniającego proces klasyfikacyjny, nie są w większości przypadków rozłączne, poza jednym wyjątkiem dotyczącym metody Gradient Boosting (por. rysunek 1).

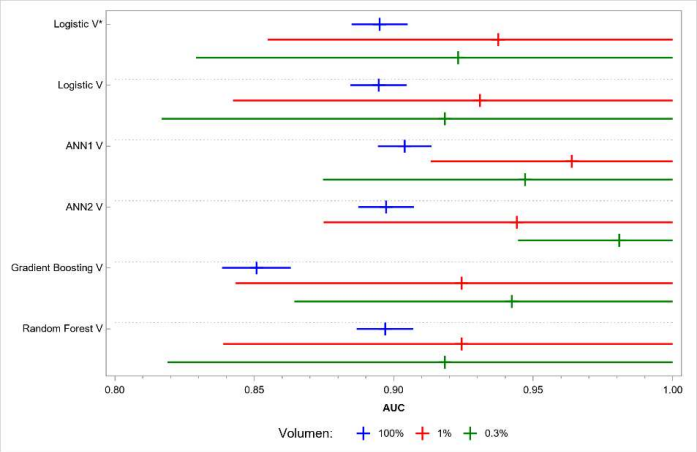
W przypadku wybrania progu uwzględniającego jednakową ważność błędów R^2zliczeniowy także dla modelu logit jest on nieznacznie mniejszy. Jest to zgodne z intuicją, bo model logit jest znacznie węższą rodziną modeli niż te które są wykorzystywane w kulturze algorytmicznej w rozpatrywanych tu technikach. Widać to wyraźnie gdy wykorzystamy wszystkie zmienne do budowy modelu logitowego nie usuwając parametrów (współczynników) nieistotnych, wówczas dopasowanie modelu do danych empirycznych jest najwyższe (R_{Square} precision).

Jeżeli natomiast postępujemy zgodnie z metodologią badań statystycznych, oprócz doboru postaci analitycznej modeli, dokonujemy doboru cech łącznie z estymacją i weryfikacją statystycznej istotności parametrów, precyzja otrzymanych wyników jest porównywalna z wynikami uzyskanymi metodami z tzw. grupy

uczenia maszynowego (por. oszacowania na podstawie walidacyjnych modeli). Tylko nieznacznie mniejsze uzyskaliśmy dopasowanie modeli logitowego (logistic selected variables) niż metodami uczenia maszynowego (por.

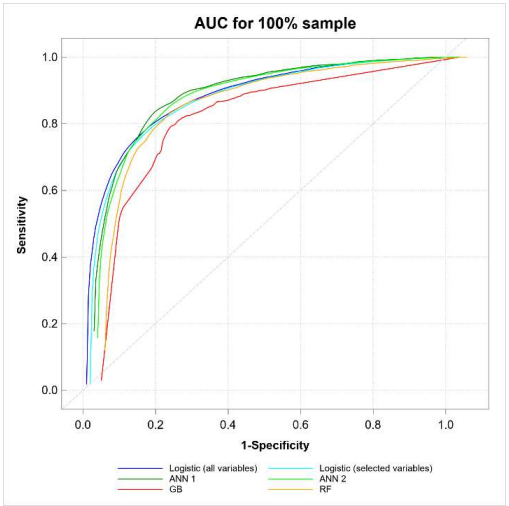
Gdyż szukamy najlepszego przybliżenia danych, a mimo to otrzymane wyniki są praktycznie takie same (nieznacząca poprawa)

Rysunek 1. Wielkości przedziałów ufności dla AUC na zbiorach walidacyjnych.



Źródło: Obliczenia własne na podstawie danych empirycznych (obliczenia i rysunki wykonane w systemie SAS/STATrel15.2)

Rysunek 2. Krzywe ROC dla zbioru walidacyjnego dla poszczególnych technik klasyfikacyjnych w przypadku całego zbioru danych



Źródło: Obliczenia własne na podstawie danych empirycznych z populacji o 36328 obserwacjach (obliczenia i rysunki wykonane w systemie SAS/STATrel15.2)

W drugiej kolejności wykonaliśmy analizę porównawczą na losowo wybranych próbach o różnej liczebności. W tabeli 2 prezentujemy wyniki uzyskane na podstawie próby o 374 obserwacjach. Uzyskane wyniki nie wskazują jednoznacznie metody o wyraźnie większej skuteczności predykcyjnej. Najlepsze oszacowanie na zbiorze walidacyjnym, mierzone wartością AUC, zliczeniowym R^2 i R_{Square} precision uzyskaliśmy z modelu logitowego uwzględniającego wszystkie cechy (wszystkie zmienne objaśniające) na podpróbie walidacyjnej. Co więcej było one zauważalnie wyższe niż oszacowania uzyskane z modeli sieci neuronowych, random forest czy gradient boosting. Jednakże precyzja oszacowania parametrów na modelach logitowych, uwzgledniających tylko cechy statystycznie istotne była porównywalna z wynikami uzyskanymi metodami uczenia maszynowego.

W przypadku tej próby wszystkie oszacowania AUC mieszczą w podobnych przedziałach ufności, tylko, że w modelach uczenia maszynowego oszacowania są po prawej stronie przedziału, a w pozostałych modelach po lewej ich stronie (por. rysunek 1).

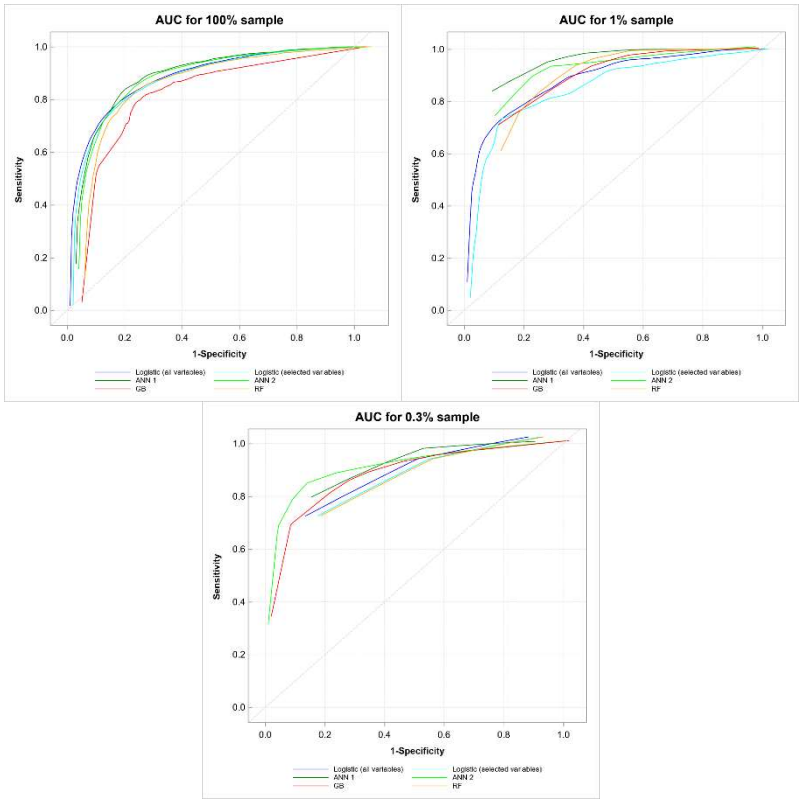
Tabela 2. Wyniki estymacji na próbie 374 elementowej

	AUC	Standard Error	95% Wald Confidence Limits		R-Square (precision)	R-Square
Logistic all var (T+T+V)	0,9285	0,0235	0,8824	0,9746	0,8607	0,5094
Logistic all var V*	1	0	1	1	0,9655	0,731
Logistic all var V	0,9327	0,0431	0,8482	1	0,8276	0,4591
Logistic (T+T+V)	0,8457	0,0359	0,7753	0,9162	0,7623	0,364
Logistic V*	0,9231	0,0479	0,8291	1	0,8621	0,4841
Logistic V	0,9183	0,0518	0,8167	1	0,7931	0,4828
ANN1 T+T	0,9083	0,0324	0,8447	0,9719	0,8495	0,4837
ANN1 V	0,9471	0,037	0,8746	1	0,8966	0,542
ANN2 T+T	0,8569	0,0387	0,781	0,9329	1,0000	0,3615
ANN2 V	0,9808	0,0185	0,9446	1	1,0000	0,6306
Gradient Boosting T+T	0,9731	0,0125	0,9486	0,9977	0,8817	0,6127
Gradient Boosting V	0,9423	0,0398	0,8643	1	0,8621	0,5519
Random Forest (T+T)	0,9968	0,0028	0,9913	1	1,0000	0,7096
Random Forest V	0,9183	0,0507	0,8189	1	1,0000	0,4847

Źródło: Obliczenia własne na podstawie 374 elementowej próby (T=71: T=22; V=29) wylosowanej losowo z populacji, gdzie: T+T+V oznacza estymację modelu na całej próbie, T+T – estymacja modelu na próbie treningowej plus testowej, V – walidacja modelu na wydzielonej części próby

Analiza krzywych ROC pokazuje zauważalne zróżnicowanie skuteczności podejścia algebraicznego nad tradycyjnym statystycznym (por. rysunek 3). Generalnie im mniejsza próba krzywe ROC są bardziej zróżnicowane, co obrazuje różnicowanie się skuteczności tych dwóch podejść do analizy danych. Na podstawie przeprowadzonych badań wyniki nie wskazują jednak wyraźnej różnicy w skuteczności dwóch analizowanych kultur analitycznych.

Rysunek 3. Krzywe ROC dla zbioru walidacyjnego dla poszczególnych technik klasyfikacyjnych dla populacji i w przypadku losowo wybranej n - elementowej próby



Źródło: Obliczenia własne na podstawie danych empirycznych z prób o zróżnicowanej liczbie obserwacji (obliczenia i rysunki wykonane w systemie SAS/STATrel15.2)

PODSUMOWANIE I WNIOSKI

Przeprowadzone badania porównawcze skuteczności dwóch kultur analitycznych przeprowadzone zostały na zbiorze danych empirycznych dotyczącego pożarów w lasach w okresie 2007 – 2016. Zbiór ten zawierał 36328 obserwacji i 16 zmiennych objaśniających na różnych skalach pomiarowych i

binarną zmienną objaśnianą Y. Zbiór danych empirycznych był zbiorem zrównoważonym (50% rekordów = 1). W analizie porównawczej ograniczyliśmy się do najbardziej popularnych klas modeli, tj. model logitowy, sieci neuronowe, lasy losowe i gradient boosting. Wyniki estymacji parametrów modeli uzyskaliśmy przy zachowaniu standardów obowiązujących w dwóch kulturach analitycznych (podejście statystyczne i algorytmiczne). Przeprowadzone badania na próbach o różnej liczebności nie wykazały znaczących różnic w skuteczności poszczególnych technik obliczeniowych. Wnioski te odnoszą się tylko do tych danych empirycznych. W generalizacji wysnutych wniosków należy dokonać dalszych analiz, szczególnie na danych symulowanych na różnych modelach danych.

Należy jednak podkreślić, że w przypadku modeli zbudowanych według zasad kultury statystycznej dysponujemy ukształtowaną tradycją dość szerokiej bezpośredniej interpretacji uzyskanych wyników. Natomiast w przypadku kultury algorytmicznej interpretacja wyników nie jest dopracowana - jest oparta o budowę i analizę scenariuszy co prowadzi do wielu problemów. Zatem w przypadku gdy uzyskujemy podobną „efektywność" modeli uzyskanych według obu kultur, to odbiorcom analizy danych słuszną wydaje się rekomendacja modelu uzyskanego według zasad kultury statystycznej.

BIBLIOGRAFIA

Berk R. (2012) Criminal Justice Forecasts of Risk, a Machine Learning Approach. Springer.

Breiman L. (2001) Statistical Modeling: The Two Cultures. Statistical Science,16(3), 199-231

Clarke B., Fokoue E., Zhang H. (2009) Principles and Theory for Data Mining and Machine Learning. Springer.

Dinsmore T. (2016) Disruptive Analytics: Charting Your Strategy for Next-Generation Business Analytics. Apress.

Goodfellow I., Bengio Y., Courville A. (2018) Deep Learning. MIT Press.

Hastie T., Tibshirani R., Friedman J. (2009) The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer.

MacKay D. (2005) Information Theory, Inference, and Learning Algorithms. Cambridge.

Mukhopadhyay S., Wang K. (2020) Breiman's "Two Cultures" Revisited and Reconciled. arXiv:2005.13596v1 [stat.ML], https://doi.org/10.48550/arXiv.2005.13596

Murphy K. (2015) Machine Learning: A Probabilistic Perspective. MIT Press Cambridge.

Ovidiu C. (2020) Deep Learning Architectures: A Mathematical Approach. Springer.

Raschka S., Mirjalili V. (2017) Python Machine Learning. Packt Publishing.

COMPARATIVE ANALYSIS OF THE EFFECTIVENESS OF TWO ANALYTICAL CULTURES

Abstract: In this paper we made a comparative analysis of the effectiveness of two analytical cultures: statistical analysis and algorithmic models using machine learning methods (neural networks, random forests, gradient boosting). The research material was data from the Polish State Forestry Information System (SILP) and concerned fires registered in state forests. The total number of observations was 36328 and the maximum number of features included in the study was 16. During the study period, the total number of registered fires was 25660 (from 2007 to the end of May 2016) fluctuated on average per year from 571 in 2016 to 3,897 in 2015.
We analyzed the causes of fire outbreaks using two different methods, the well-known statistical analysis (logit model) and using machine learning methods, i.e. random forests, gradient boosting and neural networks). The purpose of the study was to test the effectiveness of these two analytical cultures. The research conducted showed no noticeable differences between the two analytical cultures analyzed in the precision of model estimates.

Keywords: qualitative variable models, machine learning methods

JEL classification: C35, C45