Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 375 · 2018

Informatyka i Ekonometria 14

Filip Wójcik

Uniwersytet Ekonomiczny we Wrocławiu
Wydział Zarządzania, Informatyki i Finansów
Katedra Inteligencji Biznesowej w Zarządzaniu
filip.wojcik@ue.wroc.pl

PROGNOZOWANIE DZIENNYCH OBROTÓW PRZEDSIĘBIORSTWA ZA POMOCĄ ALGORYTMU XGBOOST – STUDIUM PRZYPADKU

Streszczenie: Celem niniejszego artykułu było zbadanie możliwości wykorzystania algorytmu Extreme Gradient Boosting (XGBoost) jako narzędzia prognozowania obrotów przedsiębiorstwa. Na studium przypadku wybrano udostępnione przez firmę Rossmann (wraz z prośbą o opracowanie innowacyjnej metody prognozowania) dane, obejmujące informacje z mikro- i makrootoczenia oraz obrotów 1115 oddziałów. Działanie algorytmu porównano z klasycznymi modelami SARIMAX i Holta–Wintersa, wykorzystując walidację krzyżową oraz testy statystycznej istotności różnic trafności predykcji. Badano metryki średniego błędu procentowego, współczynnik Theila oraz skorygowany współczynnik determinacji. Wyniki przekazano do weryfikacji firmie Rossmann. Potwierdzono, iż XGBoost po zastosowaniu odpowiedniej obróbki danych i sposobu uczenia osiąga lepsze rezultaty niż modele klasyczne.

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, prognozowanie.

JEL Classification: C52, C53, C63.

Wprowadzenie

Prognozowanie stanowi jedną z najistotniejszych metod przewidywania zjawisk w działalności organizacji i podmiotów gospodarczych. W literaturze zawarto wiele definicji, zgodnych jednak co do faktu, iż prognoza jest sądem odnoszącym się do zdarzeń przyszłych i niepewnych, których możliwe zaistnienie przewiduje się z wyprzedzeniem [Cieślak, 2005, s. 20]. Prognozy o charakterze naukowym muszą opierać się na dobrze sformułowanych teoriach i poddawać się falsyfikacji poprzez empiryczną weryfikację.

Na przestrzeni lat nauki, takie jak statystyka, ekonometria czy informatyka, wykształciły szereg narzędzi pozwalających na formułowanie prognoz. Zgodnie jednak z zespołem twierdzeń pod wspólną nazwą *No Free Lunch theorem* [Flach, 2012, s. 20] nie można wskazać jednego, uniwersalnego modelu możliwego do zastosowania we wszystkich sytuacjach. Nie ma wątpliwości, iż pewne klasy problemów poddają się łatwiejszemu modelowaniu za pomocą specyficznej grupy algorytmów, inne zaś – grupy odmiennej. Wraz ze wzrostem ilości danych, jakimi dysponują przedsiębiorstwa, coraz trudniejsze staje się takie dobranie modeli, aby z jednej strony zachować ich czytelność dla odbiorcy biznesowego, z drugiej zaś – trafność. Nabiera to szczególnego znaczenia w środowisku dużych wolumenów danych (*big data*), skutecznie utrudniającym kompleksową i wyczerpującą analizę.

Celem niniejszego artykułu jest przedstawienie innowacyjnego zastosowania algorytmu XGBoost [Chen, Guestrin, 2016] do prognozowania obrotów. Został on stworzony jako klasyfikator i narzędzie regresji, nie zaś jako system prognozowania szeregów czasowych. Badania wykazały jednak, iż dzięki odpowiedniej obróbce wstępnej danych i ich przygotowaniu, możliwe jest użycie omawianego algorytmu w przedstawionym kontekście.

1. Studium przypadku – kontekst biznesowy

W 2016 r. firma Rossmann (zwana dalej Firmą lub Organizatorem) upubliczniła na platformie Kaggle.com (zwanej dalej Platformą lub Portalem) częściowo zanonimizowane dane operacyjne dotyczące swojej działalności, począwszy od 2013 r., wraz z uzyskanymi obrotami (waluta oraz jednostka nieujawnione). Ufundowała nagrody dla osób, które znajdą nowy sposób prognozowania obrotów ze sprzedaży [www 1]. Firma chciała otrzymać lepsze wyniki niż klasyczne modele oparte na metodach, takich jak: ARIMA, SARIMA, ARIMAX, ETS.

Udostępnione informacje zawierały szczegóły mikro- i makrootoczenia 1115 poszczególnych oddziałów Firmy (bliskość konkurenta, od ilu lat operuje on w pobliżu, jak duży asortyment posiada dana placówka, ile było aktywnych promocji itd.) oraz charakterystykę okresu kalendarzowego (święta państwowe, ferie szkolne itd.). Dane były miejscami zniekształcone, nieprawidłowe lub w oczywisty sposób sprzeczne ze sobą. Mimo to, zdaniem Organizatora, możliwe było zidentyfikowanie, na podstawie podanych atrybutów, istotnych trendów i korelacji, a następnie opisanie ich za pomocą modelu predykcyjnego, zdolnego ekstrapolować na przyszłość. Organizator określił kryterium trafności, według

którego oceniał dostarczone rozwiązania – pierwiastek zmodyfikowanego [www 2][1] średniego względnego błędu prognozy *ex post*, w przedziale weryfikacji obejmującym daty 1.08.2015-17.09.2015, nazwanego *root mean square percentage error* (RMSPE):

$$RMSPE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right)^2}$$

gdzie:
y_i – wartość zmiennej endogenicznej w *i*-tej jednostce czasu,
$\hat{y}_i$ – wartość prognozowana zmiennej endogenicznej w *i*-tej jednostce czasu.

Weryfikacja trafności predykcji odbywa się za pomocą platformy Kaggle.com, dzięki automatycznym i pozostającym poza kontrolą lub wpływem autora skryptom Organizatora, na zadanym zbiorze danych kontrolnych.

2. Przegląd badań

Wykorzystanie algorytmu XGBoost w ekonomii do prognozowania szeregów czasowych badane było pod kilkoma względami. Podejmowano próby użycia go jako narzędzia przewidywania cen ropy naftowej [Gumus, Kiran, 2017] oraz rentowności portfela akcji [Ghosh, Purkayastha, 2017] na podstawie danych o zmianach kursów walut – *Open-High-Low-Close* (OHLC). W obu przypadkach jednak posługiwano się zbiorem z tylko jednym rodzajem zmiennych – wyliczonymi indeksami sezonowymi oraz uśrednionymi wartościami zmiennej zależnej w różnych przekrojach.

Inne publikacje podejmowały próby opracowania hybrydowego modelu łączącego wartości resztowe ARIMA z gradientowymi predykcjami XGBoost [Gurnani i in., 2017]. W takim kontekście wyniki ARIMA stają się zmienną egzogeniczną. Metodologicznie najbliższą niniejszemu opracowaniu jest praca porównująca model XGBoost, ARIMA i naiwny klasyfikator Bayesowski [Pavlyshenko, 2016] – autor zastosował podejście autoregresyjne do szkolenia modelu XGBoost*,* używając dwóch zmiennych niezależnych, informujących o zyskach przedsiębiorstwa w dniu poprzedzającym oraz ilości transakcji.

[1] Należy w tym miejscu zaznaczyć, iż wzór ewaluacyjny dostarczony przez Organizatora jest odmienny od definicji [m.in. Cieślak, 2005, s. 51], jednak przedstawiona wyżej postać zgodna jest ze stanem faktycznym. W dalszej części tekstu autor będzie się odwoływał do RMSPE jako funkcji zdefiniowanej przez Organizatora, natomiast *względny błąd prognozy ex post w przedziale weryfikacji* rozumiany będzie zgodnie z podręcznikową definicją.

Żadna z wyżej wymienionych publikacji nie skupia się szczegółowo na sposobie uzyskania najlepszej predykcji w algorytmie XGBoost, innymi słowy nie odnosi się do kwestii, w jaki sposób systematycznie przygotować dane dla modelu. Cytowane podejścia nie wskazują też, które zmienne okazały się finalnie najistotniejsze w procesie predykcji. Ich autorzy nie korzystali także z dodatkowej kategorii zmiennych egzogenicznych, którymi są tzw. dane przekrojowe, czyli informacje niezmienne w czasie, pochodzące np. z mikro- i makrootoczenia podmiotu.

Celem niniejszego artykułu jest przybliżenie procesu szkolenia XGBoost jako narzędzia prognozowania oraz wskazanie najistotniejszych, pod względem trafności, cech, a także ogólnej metodologii przygotowania wolumenu dla modelu.

3. Problem badawczy

Firma Rossmann, wyrażając zainteresowanie nowymi algorytmami prognozowania obrotów, wyraziła oczekiwanie, iż zostaną odnalezione skuteczniejsze metody, dające lepsze niż dotąd rezultaty, na wysoce zaszumionym i zniekształconym zbiorze uczącym. Jest to konkretny przypadek pojedynczego podmiotu, będący jednak, zdaniem autora, egzemplifikacją szerszego zagadnienia – prognozowania szeregów czasowych na podstawie dużych wolumenów danych o zróżnicowanej jakości. Rodzi się zatem pytanie, czy możliwe jest trafne prognozowanie obrotów przedsiębiorstwa, jeśli dysponuje się niepełnymi informacjami z jego mikro- i makrootoczenia, z wykorzystaniem metod innych niż klasyczne modele – (S)ARIMA(X), ETS lub metoda Holta–Wintersa. Trafność w tym kontekście rozumiana jest szerzej niż zostało to opisane przez Organizatora i obejmuje dodatkowe kryteria – obok błędu względnego prognozy także skorygowany współczynnik determinacji oraz wartość współczynnika Theila.

W niniejszym artykule została podjęta próba zastosowania algorytmu XGBoost należącego do rodziny tzw. estymatorów złożonych (*ensemble model*), w charakterze narzędzia prognozowania obrotów przedsiębiorstwa, na podstawie danych dotyczących jego bieżącej działalności i mikrootoczenia.

Sformułowano następujące hipotezy badawcze:

1. **Algorytm regresyjno-klasyfikacyjny XGBoost może służyć do prognozowania szeregów czasowych i uzyskuje lepsze wyniki trafności od metod (S)ARIMA(X) oraz Holta–Wintersa na przedmiotowym zbiorze danych** według kryterium RMSPE[2], współczynnika Theila dla przedziału weryfikacji oraz skorygowanego współczynnika determinacji[3] R^2.

[2] W rozumieniu zdefiniowanym przez Organizatora.

[3] Pozostałe miary według definicji za [Cieślak, 2005, s. 49-53].

2. **Algorytm XGBoost potrafi dostarczyć odbiorcy końcowemu informacji wyjaśniających, które czynniki w największym stopniu wpływają na jakość prognozy, podobnie jak w przypadku analizy ocen parametrów modeli (S)ARIMA(X).**

W kolejnych sekcjach zostaną wskazane użyte metody przygotowania i obróbki danych oraz opisany proces modelowania wraz ze specyfikacją innowacyjnego algorytmu XGBoost. Hipotezy badawcze zostaną poddane weryfikacji względem wybranych kryteriów – zarówno przez autora, jak i niezależną platformę Kaggle.com, z wykorzystaniem definicji podanej przez Organizatora.

4. Metody

Poniższe podpunkty prezentują kolejno charakterystykę wybranych klasycznych metod prognozowania szeregów czasowych, porównywanych z algorytmem XGBoost. Opisano w nich także działanie nowej metody i założenia przyjęte w trakcie jej testowania, a na końcu – metodologię porównywania wyników uzyskanych przez modele oraz oceny ich przydatności.

4.1. Modele SARIMA i SARIMAX

Seasonal Autoregressive Integrated Moving Average (SARIMA) należy do szerszej rodziny liniowych modeli autoregresyjnych i średnich ruchomych. Odznacza się zdolnością do modelowania czterech rodzajów zjawisk:

- **autoregresji rzędu p AR(p)**, gdzie bieżąca wartość szeregu czasowego jest sumą skończonej liczby p kombinacji liniowych jego poprzednich wartości [Zeliaś, Pawełek, Wanat, 2013, s. 234, 238];
- **średniej ruchomej q MA(q)**, gdzie bieżąca wartość szeregu czasowego jest liniową kombinacją q poprzednich, niezależnych od siebie zakłóceń losowych ϵ_t [Zeliaś, Pawełek, Wanat, 2013, s. 234, 238];
- **procesów d-krotnie zintegrowanych**, gdzie doprowadza się do przekształcenia niestacjonarnego szeregu czasowego do postaci stacjonarnej, poprzez wykorzystanie operacji różnicowania – czyli odejmowania od wartości w danej chwili t, wartości poprzedniej, co powtarzane jest d-razy [Zeliaś, Pawełek, Wanat, 2013, s. 239];
- **wyżej wymienionych procesów w ujęciu sezonowym P_s, D_s, Q_s**, gdzie po zidentyfikowaniu sezonowości (tygodniowej, miesięcznej, rocznej) można wyznaczyć własności charakterystyczne autoregresji, średniej ruchomej i róż-

nicowania w ujęciu sezonowym. Dzięki temu model SARIMA funkcjonuje na dwóch poziomach: bieżącym (biorącym pod uwagę autokorelację ostatnich obserwacji) oraz sezonowym (badającym wpływ poprzednich okresów na okres bieżący).

Formalnie zapisujemy model SARIMA(p,d,q)(P,D,Q) zgodnie z równaniem [Zagdański, Suchwałko, 2016, s. 189]:

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)Z_t, Z_t \sim WN(0, \sigma^2)$$

gdzie:
$\phi(z) = 1 - \phi_1 z - \phi_2 z - \cdots - \phi_p z^p$ – parametry autoregresji,
$\Phi(z) = 1 - \Phi_1 z - \Phi_2 z - \cdots - \Phi_p z^p$ – parametry autoregresji sezonowej,
$\theta(z) = \theta_1 z + \theta_2 z + \cdots + \theta_q z^q$ – parametry średniej ruchomej,
$\Theta(z) = \Theta_1 z + \Theta_2 z + \cdots + \Theta_q z^q$ – parametry sezonowej średniej ruchomej,
$B \; i \; B^s$ – operatory różnicowania (w tym sezonowego – S):

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t,$$

Z_t – zjawisko białego szumu w chwili *t*,
$Y_t = (1 - B)^d(1 - B^s)^D X_t$ – *d*-krotnie (*D*-krotnie dla sezonowości) zróżnicowany szereg czasowy.

SARIMA operuje tylko na opóźnionych wartościach zmiennej endogenicznej [Cieślak, 2005, s. 235] traktowanej jako szereg czasowy. Wariantem tego rozwiązania, operującym dodatkowo na zmiennych egzogenicznych, jest tzw. model SARIMAX, zwany przez niektórych autorów Regresją Dynamiczną [Hyndman, Athanasopoulos, 2018, rozdz 9]. Przyjmuje ona postać liniowego równania charakterystycznego dla regresji, w której statycznymi zmiennymi niezależnymi jest macierz X, natomiast ostatni człon, opisujący składnik losowy modelu, jest w istocie procesem SARIMA zgodnym z przywołaną wcześniej definicją. Taka konstrukcja umożliwia włączenie do modelu czynników zewnętrznych, które kształtują obserwowane zjawisko.

4.2. Modele z rodziny wygładzania wykładniczego

Drugą rodziną klasycznych modeli prognozowania, którą zdecydowano się wykorzystać w analizie porównawczej, są tzw. modele wygładzania wykładniczego. Zakładają one, iż wartość trendu w bieżącej chwili *t* uzależniona jest od wartości w chwilach poprzedzających, wpływ ten jednak maleje wraz z upływem czasu [Zagdański, Suchwałko, 2016, s. 278-279, 295].

Jednym z najbardziej uniwersalnych spośród modeli tego typu jest sezonowa metoda Holta–Wintersa. Jest ona w stanie obsłużyć zarówno szeregi charakteryzowane przez trend, jak i wahania sezonowe. Kolejne wartości trendu, coraz bardziej oddalone w czasie, wywierają stopniowo mniejszy wpływ na wartość bieżącą [Zagdański, Suchwałko, 2016, s. 294-295]. Model addytywny można formalnie opisać na trzy sposoby [Zagdański, Suchwałko, 2016, s. 296]:

– jako równanie poziomu szeregu w chwili *t*:

$$L_t = \alpha(X_t - S_{t-s}) + (1-\alpha)(L_{t-s} + b_{t-1}),$$

– jako równanie tendencji rozwojowej (trendu) w chwili *t*:

$$b_t = \beta(L_t - L_{t-s}) + (1-\beta)b_{t-1}$$

– jako równanie indeksu sezonowego w chwili *t*:

$$S_t = \gamma\,(X_t - L_t) + (1-\gamma)S_{t-s}\,,$$

gdzie α, β, γ są parametrami wygładzającymi przyjmującymi wartości w zakresie 0,1.

Prognoza addytywna dla punktu przesuniętego w czasie o *h* jednostek wyznaczana jest na podstawie równania:

$$F_{n+h} = L_n + hb_n + S_{n+h-s}$$

Model Holta–Wintersa, obok rodziny ARIMA, jest zaliczany do tzw. klasycznych metod prognozowania i wielokrotnie wykorzystywany w praktyce badań ekonomicznych [Cieślak, 2005, s. 148-151].

4.3. Charakterystyka algorytmu XGBoost

Algorytm Extreme Gradient Boosting (XGBoost), będący pochodną klasycznych drzew decyzyjnych i lasów losowych, został opracowany w 2014 r., upowszechniony zaś w 2016 r. [Chen, Guestrin, 2016]. Poniżej zostanie przedstawiona jego skrócona charakterystyka.

W taksonomii metod uczenia maszynowego, rozumianego zgodnie z definicją jako nauka o tworzeniu automatów zdolnych do uczenia się wraz z nabywanym doświadczeniem [Mitchell, 1997, s. 1-3], algorytm ten znajduje się w gronie tzw. procedur uczenia nadzorowanego. Oznacza to, iż otrzymuje on dwa zbiory danych: treningowy (wraz z oczekiwanymi wartościami zmiennej endogenicznej) oraz testowy. Dana jest także funkcja oceniająca, której zadaniem jest informowanie o stopniu osiągniętego dopasowania predykcji do wartości oczekiwanych [Cichosz, 2007, s. 43-44]. System uczy się znajdować istotne korela-

cje i połączenia pomiędzy atrybutami w kolejnych iteracjach, kierując się informacją zwrotną, pochodzącą z funkcji oceniającej. Następnie trafność sprawdzana jest na zbiorze testowym.

Szczegółowa charakterystyka algorytmu XGBoost wykracza poza ramy niniejszej publikacji ze względu na swą obszerność i liczne odwołania do literatury. Wszystkie właściwości omawianej metody oraz wyprowadzenia modelu matematycznego opisano w artykule [Chen, Guestrin, 2016]. Poniżej zostaną przywołane zatem jedynie najistotniejsze, w świetle dalszych rozważań, cechy:

1. **Model XGB nie ma postaci analitycznej** – jest uzyskiwany poprzez numeryczną optymalizację funkcji błędu w kolejnych iteracjach.
2. **Algorytm ma postać tzw. modelu złożonego (*ensemble*)** – jest oparty na regresyjnych drzewach decyzyjnych (CART) partycjonujących przestrzeń przykładów metodą dziel-i-rządź, w dążeniu do wprowadzenia maksymalnego porządku danych (według zadanych kryteriów)[4]. XGBoost szkoli szereg drzew, z których każde kolejne rozpoczyna naukę od nowa, z uwzględnieniem poprawki na wartości resztowe poprzedniego.
3. **Każde kolejne drzewo poddawane jest regularyzacji**, tj. ograniczeniu ilości parametrów metodą L1 bądź L2.
4. **Wartość błędu (wartość resztowa w rozumieniu klasycznej regresji) popełnionego przez drzewo dla każdego elementu badanej populacji jest zapisywana i wykorzystywana przez drzewo następne w kolejności**, by poprawić dotychczasowe wyniki. Formalnie funkcja błędu drzewa numer *t* względem wartości oczekiwanych zmiennej endogenicznej wygląda następująco [Chen, Guestrin, 2016]:

$$\mathcal{L}^t \approx \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{t-1} + f(x_i)^t + \Omega(f^t)\right),$$

gdzie:

t – numer iteracji (numer kolejny drzewa),

$l(x, y)$ – funkcja penalizująca błąd, np. RMSPE,

y_i – wartość zmiennej endogenicznej w i-tej obserwacji treningowej,

$\hat{y}_i^{t-1}$ – prognozowana przez drzewo numer t–1 (poprzednie w kolejności) wartość i-tej zmiennej endogenicznej,

$f(x_i)^t$ – prognoza wartości i-tej zmiennej endogenicznej, uzyskana przez drzewo t,

$\Omega(f^t)$ – funkcja regularyzacji drzewa numer t.

[4] Dokładny opis i specyfikacja algorytmu CART, wprowadzonego oryginalnie w 1984 r., znajduje się w publikacji profesora Leo Breimana i in.: *Classification and Regression Trees* [2017].

Z powyższego wynika, iż algorytm XGBoost jest nie tylko algorytmem złożonym, ale także iteracyjną procedurą typu *boosting* (stąd nazwa), której głównym zadaniem jest korekcja predykcji przy wykorzystaniu błędów poprzednich iteracji[5].

5. **Przykłady treningowe w modelu wybierane są na zasadzie losowania ze zwracaniem (tzw. *bootstrap*) z oryginalnego zbioru D**, z założeniem rozkładu jednostajnego – tj. każda próbka ma jednakowe szanse znaleźć się w podzbiorze [Morzy, 2013, s. 318-319]. Obserwacje, które nie zostały wybrane do treningu, służą jako zbiór testowy w danej iteracji (tzw. *out of bag error*) i wartości resztowe uzyskane w ten sposób są estymatorami błędu rzeczywistego [Breiman, 2001].

Podstawowym zastosowaniem dla algorytmu XGBoost w opisanej wyżej postaci jest klasyfikacja obiektów oraz regresja, przy założeniu statyczności danych, tj. ich niezmienności w czasie.

4.4. Metodologia badań

W literaturze przedmiotu opisanych jest wiele działań wymaganych do prawidłowego sformułowania prognozy. Wśród nich można wymienić:

a) znajomość i ocenę kształtowania się zmiennej prognozowanej,
b) ocenę stabilności prawidłowości ekonomicznej, poddawanej ocenie w miarę upływu czasu,
c) stabilność i rodzaj rozkładu składnika losowego modelu,
d) znajomość wartości zmiennych objaśniających w czasie, dla którego prognoza będzie formułowana,
e) sprawdzenie dopuszczalności ekstrapolacji prognozy poza dostępną próbkę [Zeliaś, Pawełek, Wanat, 2013, s. 40].

Proces modelowania danych i predykcji omówiony jest także w literaturze poświęconej uczeniu maszynowemu i sztucznej inteligencji. Klasyczna metodologia CRISP-DM (*Cross Industry Process for Data Mining*), opracowana w 2003 r. przez Colina Shearera [2000], definiuje następujące kroki konieczne do prawidłowego przeprowadzenia analizy:

a) zrozumienie procesu biznesowego,
b) zrozumienie danych,
c) obróbkę danych,
d) modelowanie,
e) ewaluację,
f) wdrożenie.

[5] Więcej na temat ogólnych zasad działania systemów typu *boosting* w: [Morzy, 2013, s. 318-322].

Biorąc pod uwagę powyższe i znając kryteria oceny sformułowane przez Organizatora, przeprowadzono analizę, łącząc obie wspomniane praktyki:

1. W ramach kroków 1 i 2 CRISP-DM oraz w zgodzie z wytycznymi nr 1, 2 i 4, dotyczącymi dopuszczalności prognozowania, przeprowadzono eksploracyjną analizę szeregów czasowych, aby ustalić ich właściwości, a także zbadano rozkład i strukturę zmiennych egzogenicznych.
2. W ramach kroku 3 CRISP-DM naniesiono niezbędne poprawki wymagane w dalszym modelowaniu (uzupełnienie wartości brakujących, agregacja itd.).
3. W ramach kroku 4 CRISP-DM przeprowadzono proces modelowania:
   a) z wykorzystaniem klasycznych metod,
   b) z wykorzystaniem innowacyjnego algorytmu XGBoost.
4. W ramach kroków 5 CRISP-DM oraz 3 metodologii prognozowania zbadano trafność dopasowania modeli i przeprowadzono testy statystycznej istotności różnic pomiędzy nimi.

Ze względu na obszerność udostępnionego wolumenu danych oraz stopień skomplikowania całego procesu w kolejnych podpunktach umieszczono podsumowanie przeprowadzonych działań i ich zbiorcze wyniki.

4.5. Zastosowane metody obróbki danych

Zgodnie z opisaną wyżej metodologią proces modelowania rozpoczęto od eksploracji danych oraz ich przygotowania do prognozowania. Niniejszy podpunkt przedstawia podjęte działania w opisywanym zakresie.

Wolumen dostarczony przez Organizatora obejmuje zarówno zmienną endogeniczną (obrót ze sprzedaży w poszczególnych placówkach w wybranych dniach), jak i informacje z mikro- i makrootoczenia [www 3], w szczególności te przedstawione w tabeli 1.

Tabela 1. Struktura danych

Zmienna	Typ	Opis
1	*2*	*3*
Date	data	znacznik czasu dla danego rekordu
StoreID	numeryczny	identyfikator sklepu
Sales	numeryczny/szereg czasowy	obrót ze sprzedaży **zmienna endogeniczna** – cel prognozy
Customers	numeryczny/szereg czasowy	ilość klientów w placówce danego dnia; formuje szereg czasowy, ale jest zmienną egzogeniczną/objaśniającą
Open	binarny	informuje, czy sklep był otwarty danego dnia
State/School holiday	dyskretny	wskazuje, czy dany dzień był dniem wolnym lub feriami szkolnymi
StoreType	dyskretny	typ sklepu

cd. tabeli 1

1	*2*	*3*
Assortment	dyskretny	typ asortymentu
CompetitionDistance	numeryczny	odległość w metrach do najbliższego konkurenta
CompetitionOpenSince	data	wskazuje czas otwarcia lokalu konkurenta
Promo/Promo2/Promo2Since/ PromoInterval	binarny	dane dotyczące promocji w czasie

Źródło: Na podstawie: [www 3].

Fakt, iż dostarczone dane są indeksowane czasem w jednakowych odstępach (dziennych) i opisują zmienność pewnej mierzalnej wielkości (obrót, liczba klientów), sprawia, że można je zakwalifikować jako szeregi czasowe – uporządkowany ciąg obserwacji $y_1, y_2, \dots, y_n$ [Zeliaś, Pawełek, Wanat, 2013, s. 70-71]. Pozostałe zmienne są zestawem danych objaśniających, mających dostarczyć dodatkowych informacji do modelu. Wolumen zawiera 1115 unikalnych sklepów, z których każdy opisany jest poprzez zaprezentowane wyżej atrybuty oraz posiada własny szereg czasowy obrotu.

W przedstawionym kontekście podstawową strukturą danych dla pojedynczego sklepu jest macierz zmiennych egzogenicznych, w której każda kolumna opisuje atrybut (o określonej przeciwdziedzinie), wiersz zaś jest wektorem reprezentującym element badanej populacji (sklep) w chwili *t* poprzez wartości jego poszczególnych atrybutów. Formalnie można zapisać [Zaki, Meira, 2014, s. 1]:

$$\mathcal{X} = \begin{bmatrix} A_1 & A_2 & \dots & A_m \\ x_{1a_1} & x_{1a_2} & \dots & x_{1a_m} \\ x_{2a_1} & x_{2a_2} & \dots & x_{2a_m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{na_1} & x_{na_2} & \dots & x_{na_m} \end{bmatrix}$$

Zmienna endogeniczna przyjmuje postać indeksowanego czasem wektora liczbowego. Formalnie [Zaki, Meira, 2014, s. 33]:

$$\mathcal{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Cały zbiór danych jest zatem ujmowany jako zbiór trójek uporządkowanych – identyfikatora sklepu, wartości atrybutów elementu badanej populacji (oznaczonego jako *x*) oraz zmiennej endogenicznej *y* [Zaki, Meira, 2014, s. 480]:

$$\mathcal{D} = \{id, x_i, y_i\}_i^n, \text{gdzie } id \in \mathcal{S}, |\mathcal{D}| = n, x_i \in \mathcal{X} \wedge y_i \in \mathcal{Y}$$

Szereg czasowy sumy obrotu w danych treningowych dla wszystkich oddziałów obejmuje daty z zakresu 1.01.2013-31.07.2015, natomiast dla danych testowych (względem których Organizator ocenia trafność prognozy) od 1.08.2015-17.09.2015.

W sposób automatyczny dla każdego sklepu przeprowadzono testy sprawdzające własności szeregu czasowego. Ustalono, co następuje:

1. **Stacjonarność szeregu** została sprawdzona testem Dickeya–Fullera [Hyndman, Athanasopoulos, 2018, rozdz. 8.1]. Dla wszystkich sklepów **stwierdzono, iż szereg jest niestacjonarny**.
2. **Obecność autokorelacji** od 1 do 20 obserwacji sprawdzono testem Ljunga––Boxa [Hyndman, Athanasopoulos, 2018, rozdz. 8.1], **potwierdzając obecność autokorelacji o zróżnicowanych rzędach (3-7)**.
3. Obie składowe zbadano za pomocą algorytmu klasycznej dekompozycji szeregu czasowego na trend i sezonowość. Dla wszystkich sklepów wykazano (za pomocą rozszerzonego testu Dickeya–Fullera) obecność tendencji rozwojowej szeregu czasowego. Podobnie **tendencja sezonowa** – trend tygodniowy (7 dni) oraz miesięczny **są istotne statystycznie**.

Z powyższej analizy wynika, iż do modelowania należy użyć metod zdolnych do przetwarzania szeregów niestacjonarnych o zróżnicowanej sezonowości, z tendencją rozwojową oraz posiadających autokorelację. Dodatkowo wykorzystane zostały modele obsługujące zmienne egzogeniczne (SARIMAX).

4.6. Szkolenie modeli klasycznych

Ze zbioru danych treningowych wyłączono ostatnie dwa miesiące (czerwiec i lipiec 2015 r. – 3000 obserwacji) do ostatecznego testu prowadzonego lokalnie. Pozostałe dane wykorzystano do szkolenia klasycznych modeli, wzmiankowanych w podpunktach 4.1 oraz 4.2, za pomocą walidacji krzyżowej szeregu czasowego[6].

Szkolono osobny model dla każdego sklepu, a zatem 1115 modeli SARIMAX. Z tego też względu nie są prezentowane szczegółowe analizy każdego z osobna, a jedynie zbiorcze wyniki. Wykorzystano algorytm *auto arima*, optymalizujący model SARIMAX poprzez iteracyjne przeszukiwanie przestrzeni parametrów w celu minimalizacji zarówno RMSPE, jak i współczynnika zło-

[6] Metoda polegająca na stopniowym dzieleniu zbioru na treningowy i testowy poprzez przesuwanie okna czasowego w przód. Z każdą kolejną iteracją okno treningowe się rozszerza, a testowe zawęża. Dokładny opis metody oraz jej implementacja znajdują się w podręczniku *Forecasting: Principles and Practice* [Hyndman, Athanasopoulos, 2018, rozdz. 3.4].

żoności AIC oraz BICC. Użyto implementacji autora metody, R.J. Hyndmana, opisanej w publikacji z 2008 r. [Hyndman, Yeasmin, 2007].

4.7. Szkolenie modelu XGBoost

Dla modelu XGBoost zastosowano zbliżoną metodę, jak dla modeli klasycznych – dwa ostatnie miesiące (3000 obserwacji) wyłączono ze zbioru treningowego jako zbiór walidacyjny. Na reszcie przeprowadzono proces uczenia walidacją krzyżową [Morzy, 2013, s. 334-335] szeregu czasowego z użyciem zmiennych egzogenicznych.

Chcąc dostosować algorytm XGBoost do wymogów predykcji czasowych, zastosowano następujące środki:

- wyznaczono **indeksy sezonowe** dla wysokości sprzedaży i ilości klientów: kwartalne, miesięczne i tygodniowe średnie, podobnie jak w algorytmie dekompozycji szeregu czasowego [Hyndman, Athanasopoulos, 2013, rozdz. 6.3]; dla każdej jednostki czasu (miesiąc/kwartał) wyliczono średnią poprzednich wartości w ubiegłych latach w danej jednostce jako wartość indeksu;
- **wyliczono średnie ruchome** obrotu dla ostatnich 3, 5, 7 dni, 1, 2, 3, 6 ostatnich miesięcy oraz półrocza;
- jako zmienne egzogeniczne, dla zachowania relacji czasowych, dodano **wskaźniki czasu**: numer dnia tygodnia, numer miesiąca, kwartału, półrocza i roku;
- dodano **zmienne binarne** informujące, czy dany dzień to sobota lub niedziela bądź święto.

Dodatkowo zestaw danych uzupełniono o charakterystykę samych sklepów oraz ich otoczenia – odległość od konkurencji, typ placówki, promocję, asortyment.

XGBoost szkolono globalnie dla wszystkich sklepów, aby algorytm nauczył się rozpoznawać także prawidłowości niespecyficzne w przypadku poszczególnych oddziałów Firmy.

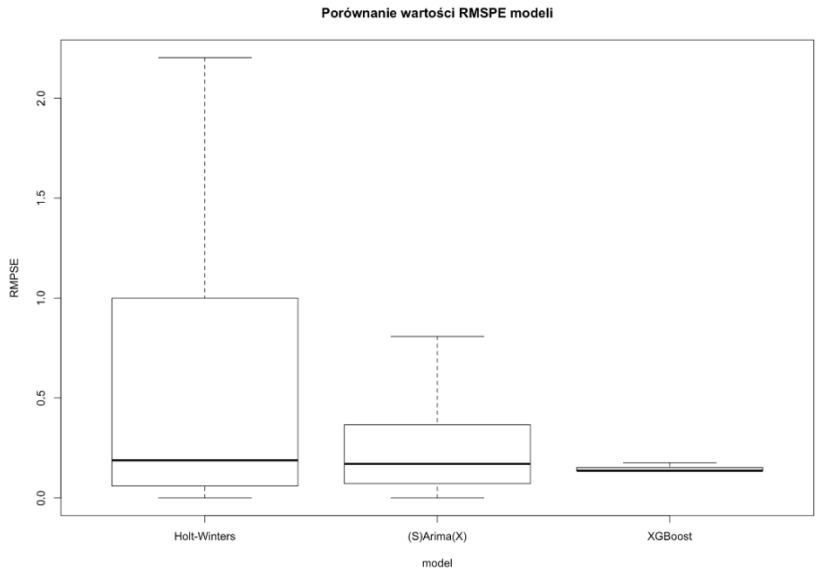
5. Wyniki

Wyniki (zbiorcze dla wszystkich sklepów) prezentuje tabela 2.

Tabela 2. Rozkład wyników dla modeli klasycznych oraz XGBoost

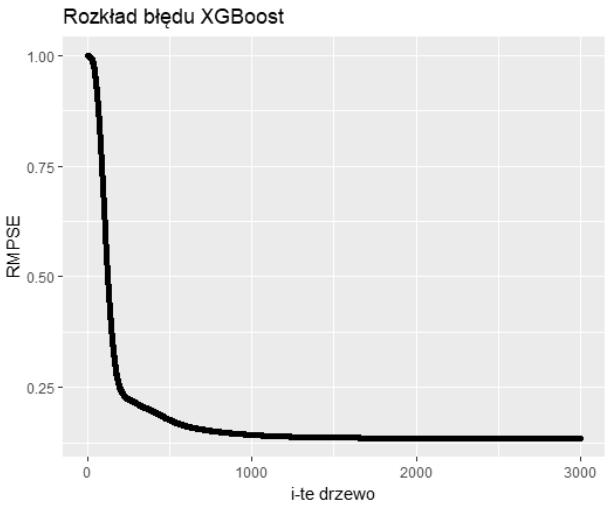
Wartość\Model	SARIMAX	Holt-Winters	XGBoost
Średni RMSPE na zbiorze walidacyjnym	0,306	0,398	0,180
1 kwantyl/mediana/3 kwantyl RMSPE na zbiorze walidacyjnym	0,071/ 0,1705/ 0,366	0,059/ 0,1886/ 1,0	0,1350/ 0,1364/ 0,1517
Współczynnik Theila	0,061	0,169	0,021
Współczynnik R^2	0,838	0,54	0,9
Wynik według Organizatora	0,16	0,367	0,128
Wielkość próbki	3000	3000	3000

Poniższe rysunki prezentują rozkład wartości błędów w zbiorze walidacyjnym dla poszczególnych modeli oraz kształtowanie się RMSPE w miarę nauki XGBoost.



Rys. 1. Porównanie rozrzutu błędów modeli

Analiza rozrzutu wartości miernika RMSPE przedstawiona na rys. 1 wskazuje, iż wartości tego miernika dla XGBoost wyróżniają się najmniejszą średnią i wariancją spośród wszystkich zastosowanych metod, a więc prognozy stawiane metodą XGBoost charakteryzuje większa dokładność *ex post* i większa stabilność. Na rys. 2 wyraźnie widoczna jest także tendencja spadkowa wartości RMSPE w kolejnych iteracjach, stabilizująca się w okolicach tysięcznego powtórzenia. Wykres nie wskazuje na oscylacje ani nagłe wzrosty w kolejnych stadiach, jakie występują w mniej stabilnych algorytmach.



Rys. 2. Rozkład błędu w trakcie nauki modelu XGBoost

W przypadku testów lokalnych, na wyznaczonym zbiorze walidacyjnym, przeprowadzono porównanie różnicy średnich trafności metod zgodnie z rekomendacjami dla algorytmów uczenia maszynowego, stosowanych w regresji i prognozowaniu [Flach, 2012, s. 352-354]. Poszczególne obserwacje (będące wynikami trafności modeli na próbce walidacyjnej) pobierane były niezależnie od siebie, co jest spełnieniem wymogu niezależności. Sprawdzenie normalności rozkładu miernika RMSPE testem Shapiro–Wilka na poziomie istotności $\alpha = 0{,}05$ zakończyło się odrzuceniem hipotezy zerowej dla każdego z modeli, co świadczy o istotnych odchyleniach od normalności. Podobnie wyniki uzyskano testem Levene'a dla homogeniczności wariancji. Jednak ze względu na dużą wielkość próbki (3000 obserwacji dla każdego z modeli) oraz odporność t-testu różnicy średnich na odstępstwo od normalności rozkładu próbek [Boneau, 1960; Lumley i in., 2002] przeprowadzono test T-Welcha [Welch, 1947], uwzględniający nierówną wariancję. Dostosowano także poziom istotności oraz przedziały ufności ze względu na porównania wielokrotne [Kutner i in., 2013, s. 746-754], przyjmując bazowy poziom istotności $\alpha = 0{,}05$. Wyniki prezentuje tabela 3:

1. **Hipoteza zerowa:** różnica średnich pomiędzy grupami jest równa zeru.
2. **Hipoteza alternatywna:** różnica średnich pomiędzy grupami jest różna od zera.

Tabela 3. Analiza różnicy średnich wartości RMSPE pomiędzy modelami

Modele	Różnica średnich RMSPE	Przedział ufności od	Przedział ufności do	p-wart.
Holt-Winters – SARIMAX	0,095	0,07	0,115	<< 0,001
XGBoost – SARIMAX	–0,126	–0,139	–0,101	<< 0,001
XGBoost – Holt-Winters	–0,216	–0,235	–0,190	<< 0,001

Wyniki te wskazują, iż zachodzi statystycznie istotna różnica pomiędzy modelem XGBoost a pozostałymi zastosowanymi metodami, na korzyść tego pierwszego. Błąd popełniany przez omawiany algorytm był o około 10% do 14% niższy od (S)ARIMA(X) oraz 19% do 23,5% niższy niż w przypadku metody Holta–Wintersa. Wyniki te potwierdza również niezależna analiza przeprowadzona przez Organizatora na wydzielonym zbiorze danych, a także wartości skorygowanego współczynnika R^2 oraz współczynnik Theila (tabela 2).

Po zakończeniu prognozowania można wyznaczyć istotność atrybutów w największym stopniu wpływających na trafność predykcji [Chen, Guestrin, 2016]. Wyliczenie to stanowi wprost implementację tzw. permutacyjnej istotności atrybutów dla lasów losowych, po raz pierwszy zastosowanej przez Leo Breimana [2001]. Skrócone uśrednione wyniki (10 najlepszych atrybutów w 1000 iteracji) przedstawia tabela 4. Kolumna „wzrost trafności" odpowiada uśrednionemu procentowemu wzrostowi trafności predykcji, gdy drzewa decyzyjne składające się na XGBoost dzielą wolumen według danej cechy. Pokrycie wskazuje na relatywną ilość wystąpień cechy w elementach populacji. Częstość użycia opisuje, jak wiele drzew spośród wyszkolonego lasu wykorzystuje tę cechę.

Tabela 4. Uśredniona istotność atrybutów

Atrybut	Znaczenie	Wzrost trafności	Pokrycie	Częstość użycia
1	*2*	*3*	*4*	*5*
CompetitionDistance	odległość od sklepu konkurencji	18,64%	22,99%	8,28%
Promo	promocja tak/nie	16,60%	1,54%	3,27%
Store	typ sklepu	16,19%	29,33%	10,17%
meanSalesDow	średnia obrotu dla danego dnia tygodnia	6,55%	3,66%	5,58%
CompetitionOpen SinceMonth	wyrażony w miesiącach czas od otwarcia konkurencyjnego sklepu	5,27%	3,09%	3,32%
CompetitionOpen SinceYear	wyrażony w latach czas od otwarcia konkurencyjnego sklepu	5,01%	3,23%	3,27%

cd. tabeli 4

1	*2*	*3*	*4*	*5*
Day	numer dnia tygodnia	4,04%	7,18%	15,21%
meanSalesMonth	średnia obrotu dla danego miesiąca	4,04%	3,55%	6,34%
meanQuarterlySales	średnia obrotu dla kwartału	3,70%	1,98%	1,69%
Promo2SinceMonth	wyrażony w miesiącach czas obowiązywania promocji	3,29%	2,58%	2,59%

Z powyższego wynika, iż najsilniejszy wpływ na trafność predykcji mają atrybuty dotyczące obecności konkurencji – odległość i czas operowania na rynku. Dużym wpływem odznaczają się także promocje. Najsilniejsza sezonowość pojawia się dla miesięcy oraz dnia tygodnia (ale nie weekendu), co może sugerować, iż zachowania konsumentów różnią się w zależności od bieżącego dnia roboczego. Pozostałe wyliczone indeksy sezonowe i ruchome średnie zostały zignorowane.

Zaprezentowane wyniki istotności atrybutów różnią się nieznacznie (na poziomie tysięcznych i dziesięciotysięcznych) w zależności od iteracji algorytmu ze względu na jego mocno zrandomizowany charakter.

6. Dyskusja

Wyliczone współczynniki istotności atrybutów stanowią kombinację zmiennych egzogenicznych (z makrootoczenia sklepu) oraz indeksów sezonowych, charakterystycznych dla prognozowania szeregów czasowych. Oznacza to, iż algorytm XGBoost był w stanie wykryć i uwzględnić zarówno statyczne czynniki wpływające na wysokość obrotu, jak też ogólne trendy oraz zachowanie szeregu w czasie. Wyniki te potwierdzają badania innych autorów. Tak w przypadku przewidywania cen ropy naftowej [Gumus, Kiran, 2017], jak rentowności portfeli akcji [Ghosh, Purkayastha, 2017] i wspomagania klasycznego modelu ARIMA [Gurnani i in., 2017], indeksy sezonowe okazały się skutecznym narzędziem. Novum jest jednoczesne uwzględnienie w ramach tego samego modelu zmiennych statycznych z mikrootoczenia przedsiębiorstwa oraz tych opisujących zmienność szeregu w czasie. W podrozdziale 4.5 poświęconemu przygotowaniu danych i ich obróbce wykazano obecność tendencji rozwojowej oraz sezonowości tygodniowej i miesięcznej dla sklepów. Wysoka pozycja odpowiadających im indeksów na liście atrybutów XGBoost potwierdza, iż był on

w stanie zidentyfikować je jako najważniejsze. Równocześnie zignorował wszystkie pozostałe (redundantne) indeksy i średnie ruchome, których istotności nie potwierdziła manualna analiza. Należy to uznać za zachowanie prawidłowe – algorytm wskazał te same składowe sezonowości, co klasyczne metody analizy.

Przedmiotem dalszych badań na innych zbiorach danych powinno być określenie stopnia komplikacji relacji sezonowych, które może wykorzystać procedura XGBoost. W przypadku modeli (S)ARIMA(X) możliwe jest wykonywanie różnicowania, a także wyznaczanie parametrów ruchomej średniej i autoregresji rozmaitych rzędów – zasadne wydaje się przebadanie pod tym względem algorytmu XGBoost w celu systematycznego określenia granic jego możliwości, np. zdolności do wykrywania maksymalnego rzędu autokorelacji.

Podsumowanie

Analiza wyników badań wskazuje, iż algorytm XGBoost uzyskał na przedmiotowym zbiorze danych statystycznie istotnie lepsze rezultaty w porównaniu do metod klasycznych – SARIMAX oraz modelu Holta–Wintersa. Uczenie z pomocą szczegółowych indeksów sezonowych oraz zmiennych niezależnych pozwoliło skutecznie wykorzystać w charakterze narzędzia prognozowania algorytm przeznaczony pierwotnie do pracy z danymi statycznymi. Indeksy sezonowe zidentyfikowane przez niego jako istotne pokrywają się z manualną analizą własności szeregów.

Ponadto, mimo iż *Extreme Gradient Boosting* nie ma czytelnej postaci parametrycznej, dzięki wykorzystaniu metody permutacyjnej istotności w drzewach decyzyjnych Breimana pozwala na ustalenie wpływu poszczególnych zmiennych egzogenicznych na końcową prognozę, co ma dużą wartość informacyjną dla odbiorców końcowych.

Należy podkreślić, że zgodnie z twierdzeniem *No Free Lunch theorem* nie można wskazać jednego, uniwersalnie lepszego modelu predykcyjnego. Z tego też względu przedstawionych badań nie należy traktować jako dowodu na to, że w każdej sytuacji można zastąpić metody klasyczne nowym podejściem. Powinno być ono traktowane jako kolejne narzędzie, potencjalnie mogące dostarczyć trafniejszych odpowiedzi niż dotychczas stosowane, pod warunkiem zastosowania odpowiedniej obróbki danych i sposobu trenowania.

Literatura

Boneau C.A. (1960), *The Effects of Violations of Assumptions Underlying the T Test*, "Psychological Bulletin", Vol. 57(1), s. 49-64.

Breiman L. (2001), *Random Forests*, "Machine Learning", Vol. 45(1), s. 5-32.

Breiman L., Friedman J., Stone Ch.J., Olshen R.A. (2017), *Classification and Regression Trees*, CRC Press, Boca Raton, FL.

Chen T., Guestrin C. (2016), *XGBoost: A Scalable Tree Boosting System* [w:] *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, ACM Press, New York, s. 785-794.

Cichosz P. (2007), *Systemy uczące się*, Wyd. 2. Wydawnictwa Naukowo-Techniczne, Warszawa.

Cieślak M. (2005), *Prognozowanie gospodarcze: metody i zastosowania*, Wyd. 4, Wydawnictwo Naukowe PWN, Warszawa.

De Livera A.M., Hyndman R.J., Snyder R.D. (2011), *Forecasting Time Series with Complex Seasonal Patterns Using Exponential Smoothing*, "Journal of the American Statistical Association", Vol. 106, s. 1513-1527.

Flach P.A. (2012), *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*, Cambridge University Press, Cambridge.

Ghosh R., Purkayastha P. (2017), *Forecasting Profitability in Equity Trades Using Random Forest, Support Vector Machine and XgBoost* [w:] *10th International Conference on Recent Trades in Engineering Science and Management*, s. 473-486.

Gumus M., Kiran M.S. (2017), *Crude Oil Price Forecasting Using XGBoost* [w:] *2017 International Conference on Computer Science and Engineering (UBMK)*, IEEE, Piscataway Township, NJ, s. 1100-1103.

Gurnani M., Korke Y., Shah P., Udmale S., Sambhe V., Bhirud S. (2017), *Forecasting of Sales by Using Fusion of Machine Learning Techniques* [w:] *2017 International Conference on Data Management, Analytics and Innovation (ICDMAI)*, IEEE, Piscataway Township, NJ, s. 93-101.

Hyndman R., Athanasopoulos G. (2018), *Forecasting: Principles and Practice*, Otexts, Melbourne, http://otexts.org/fpp2/ (dostęp: 13.05.2019).

Hyndman R., Yeasmin K. (2007), *Automatic Time Series Forecasting: The Forecast Package for R*, "Journal of Statistical Software", Vol. 27(9), s. 1-23.

Kutner M.H., Neter J., Nachtsheim C.J., Li W. (2013), *Applied Linear Statistical Models*, McGraw-Hill, Boston.

Lumley T., Diehr P., Emerson S., Chen L. (2002), *The Importance of the Normality Assumption in Large Public Health Data Sets*, "Annual Review of Public Health", Vol. 23(1), s. 151-169.

Mitchell T. (1997), *Machine Learning*, McGraw-Hill, New York.

Morzy T. (2013), *Eksploracja danych: metody i algorytmy*, Wydawnictwo Naukowe PWN, Warszawa.

Pavlyshenko B.M. (2016), *Linear, Machine Learning and Probabilistic Approaches for Time Series Analysis* [w:] *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, IEEE, s. 377-381.

Shearer C. (2000), *The CRISP-DM Model: The New Blueprint for Data Mining*, "Journal of Data Warehousing", Vol. 5, No. 4, s. 13-22.

Welch B.L. (1947), *The Generalisation of 'Student's' Problem when Several Different Population Variances are Involved*, "Biometrika", Vol. 34, No. 1/2, s. 28-35.

Zagdański A., Suchwałko A. (2016), *Analiza i prognozowanie szeregów czasowych: praktyczne wprowadzenie na podstawie środowiska R*, Wydawnictwo Naukowe PWN, Warszawa.

Zaki M.J., Meira W. (2014), *Data Mining and Analysis: Fundamental Concepts and Algorithms*, Cambridge University Press, Cambridge.

Zeliaś A., Pawełek B., Wanat S. (2013), *Prognozowanie ekonomiczne: teoria, przykłady, zadania*, Wyd. 3, Wydawnictwo Naukowe PWN, Warszawa.

[www 1] https://www.kaggle.com/c/rossmann-store-sales#description (dostęp: 9.01.2018).

[www 2] https://www.kaggle.com/c/rossmann-store-sales#evaluation (dostęp: 9.01.2018).

[www 3] https://www.kaggle.com/c/rossmann-store-sales/data (dostęp: 11.01.2018).

FORECASTING DAILY TURNOVER USING XGBOOST ALGORITHM – A CASE STUDY

Summary: The goal of this paper was to investigate use of the *Extreme Gradient Boosting XGBoost* algorithm as a forecasting tool. The data provided by the Rossman Company, with a request to design an innovative prediction method, has been used as a base for this case study. The data contains details about micro- and macro-environment, as well as turnover of 1115 stores. Performance of the algorithm was compared to classical forecasting models SARIMAX and Holt–Winters, using time-series cross validation and tests for statistical importance in prediction quality differences. Metrics of root mean squared percentage error (RMSPE), Theil's coefficient and adjusted correlation coefficient were analyzed. Results where then passed to Rossman for verification on a separate validation set, via Kaggle.com platform. Study results confirmed, that XGBoost, after using proper data preparation and training method, achieves better results than classical models.

Keywords: artificial intelligence, machine learning, forecasting.