

P O L I T E C H N I K A R Z E S Z O W S K A

im. Ignacego Łukasiewicza

WYDZIAŁ MATEMATYKI I FIZYKI STOSOWANEJ

**Aldona Świrad**

Predykcja szeregów czasowych: metody klasyczne i elementy uczenia maszynowego

**Praca dyplomowa inżynierska**

Opiekun pracy:  
dr hab. Liliana Rybarska-Rusinek , prof. PRz

Rzeszów, 2024

Spis treści

[Wstęp 4](#_Toc185626612)

[Rozdział 1. Wprowadzenie teoretyczne 5](#_Toc185626613)

[1.1 Definicja i własności szeregów czasowych 5](#_Toc185626614)

[1.2 Model ARIMA 6](#_Toc185626615)

[Rozdział 2. Metody uczenia maszynowego dla szeregów czasowych 8](#_Toc185626616)

[2.1 Model RNN 15](#_Toc185626617)

[2.2 Model LSTM 19](#_Toc185626618)

[Rozdział 3. Analiza wybranych szeregów czasowych metodami klasycznymi 23](#_Toc185626619)

[3.1 Analiza szeregów czasowych – opis, wykresy, zakres zmienności 24](#_Toc185626620)

[3.2 Dobór modelu 36](#_Toc185626621)

[Rozdział 4. Prognozowanie szeregów czasowych 39](#_Toc185626622)

[4.1 Prognozowanie metodą ARIMA 40](#_Toc185626623)

[4.2 Prognozowanie metodą LSTM 47](#_Toc185626624)

[4.3 Porównanie 61](#_Toc185626625)

[Podsumowanie 63](#_Toc185626626)

[Bibliografia 64](#_Toc185626627)

# Wstęp

Tekst akapitu.

Tekst akapitu.

**Cel pracy**

Celem niniejszej pracy inżynierskiej jest zbadanie i porównanie skuteczności metod analizy szeregów czasowych: metod klasycznych oraz metod opartych na uczeniu maszynowym, ze szczególnym uwzględnieniem modeli LSTM. Głównym celem pracy jest zastosowanie tych modeli do prognozowania szeregów czasowych oraz ocena ich przydatności w różnych scenariuszach analizy danych czasowych.

**Zakres pracy**

W pierwszym rozdziale pracy podane zostaną najważniejsze pojęcia z zakresu modelowania i predykcji szeregów czasowych. Rozdział drugi poświęcony zostanie analizie własności i modelowaniu wybranych szeregów czasowych, z zastosowaniem metod klasycznych. W rozdziale trzecim podane zostaną podstawowe pojęcia i metody uczenia maszynowego (w tym modele RNN i LSTM), stosowane w analizie szeregów czasowych. Porównanie wybranych metod prognozowania szeregów czasowych pod kątem ich skuteczności, wraz z wnioskami dotyczącymi przydatności poszczególnych modeli w różnych scenariuszach analizy szeregów czasowych, będą zawarte w ostatnim rozdziale pracy. W podsumowaniu zawarte będą sugestie dotyczące dalszych badań oraz potencjalnych zastosowań praktycznych uzyskanych wyników.

Praca będzie opierać się głównie na analizie literatury naukowej oraz eksperymentach przeprowadzonych na rzeczywistych danych, co umożliwi kompleksowe zbadanie tematu oraz wyciągnięcie trafnych wniosków dotyczących analizy szeregów czasowych za pomocą metod klasycznych i uczenia maszynowego.

# Rozdział 1. Wprowadzenie teoretyczne

W rozdziale 1. pracy przybliżymy podstawowe pojęcia związane z szeregami czasowymi oraz zaprezentujemy jeden z najczęściej wykorzystywanych modeli do ich analizy – ARIMA [1].Zrozumienie tych pojęć oraz metodologii jest kluczowe do poprawnej analizy i prognozowania szeregów czasowych.

## 1.1 Definicja i własności szeregów czasowych

Szereg czasowy (ang. *time series*) [3] to sekwencja danych gromadzonych w równych odstępach czasu, np. dziennie, miesięcznie, rocznie. Każda obserwacja w szeregu czasowym składa się z dwóch elementów: momentu w czasie, do którego się odnosi, oraz wartości zmiennej mierzonej w tym momencie. Przy czym, kolejność obserwacji ma istotne znaczenie. Przykładami szeregów czasowych mogą być kursy walut, wartości indeksów giełdowych, liczba sprzedanych produktów w danym okresie, czy zmiany temperatury w ciągu roku.

W praktyce, analiza szeregów czasowych ma na celu identyfikację struktury danych, wyodrębnienie trendów, sezonowości oraz innych wzorców, a także prognozowanie przyszłych wartości na podstawie dostępnych danych historycznych, czym będziemy się zajmować w tej pracy.

***Stacjonarność szeregu czasowego***

Szeregi czasowe mogą być stacjonarne lub niestacjonarne [1]. Szereg stacjonarny charakteryzuje się stałą średnią, wariancją i autokorelacją w czasie, natomiast szereg niestacjonarny może zawierać zmieniające się średnie, wariancje lub inne nieliniowe zależności. Sprawdzenie, czy szereg czasowy jest stacjonarny, umożliwia użycie testów m.in.:

**-** *Rozszerzony test Dickeya-Fullera* – test pierwiastka jednostkowego, który wykrywa obecność trendu w szeregu czasowym.

*Hipotezy*: H0 - szereg nie jest stacjonarny, H1 - szereg jest stacjonarny.

**-** *Test Phillipsa-Perrona* – test statystyczny bazujący na estymacji autokorelacji reszt. *Hipotezy*: H0 - szereg nie jest stacjonarny, H1 - szereg jest stacjonarny.

**-** *Test Kwiatkowskiego-Phillipsa-Schmidta-Shina (KPSS)* – test oparty na analizie wariancji między próbkami z różnych okresów szeregu czasowego.

*Hipotezy*: H0 - szereg jest stacjonarny, H1 - szereg nie jest stacjonarny.

Aby ocenić wyniki testów, sprawdzamy wartość *p-value*. Jeżeli wartość ta jest większa od poziomu istotności np. *α = 0,05*, to nie ma podstaw do odrzucenia hipotezy zerowej. Celem skutecznej analizy szeregów czasowych i budowy modeli prognostycznych, często stosuje się metody przekształcania szeregów niestacjonarnych w stacjonarne.

***Trend***

Trend odnosi się do długoterminowego kierunku, w jakim zmienia się wartość danych w szeregu czasowym [1]. Jest to ogólny wzorzec wzrostu, spadku lub stagnacji, który utrzymuje się przez dłuższy okres. Trend może być wynikiem różnych czynników, takich jak zmiany demograficzne, technologiczne, gospodarcze lub społeczne. Przykłady trendów to wzrost liczby wskaźnika sprzedaży na przestrzeni lat, spadek cen określonych towarów w wyniku ulepszeń technologicznych czy stopniowy wzrost średniej temperatury w wyniku zmian klimatycznych. Trend może być liniowy (stałe tempo zmiany) lub nieliniowy (zmienne tempo wzrostu lub spadku).

***Sezonowość***

Sezonowość to powtarzający się wzorzec w danych, który występuje w stałych, regularnych odstępach czasu, najczęściej w ramach jednego roku, miesiąca, tygodnia lub innego cyklu. Jest ona spowodowana zjawiskami kalendarzowymi lub cyklicznymi, takimi jak pory roku, święta, dni tygodnia czy godziny w ciągu dnia. Na przykład, sprzedaż lodów wzrasta latem (sezonowość roczna), a ruch w restauracjach typu fast-food jest wyższy w weekendy (sezonowość tygodniowa). Sezonowość można łatwo zidentyfikować w danych jako powtarzalny wzorzec.

***Różnicowanie niestacjonarnego szeregu czasowego***

Różnicowanie [10] służy do przekształcania niestacjonarnych szeregów czasowych w szeregi stacjonarne. Proces ten eliminuje trend i sezonowość, co ułatwia modelowanie fluktuacji w krótszych okresach.

Proces różnicowania polega na obliczeniu różnicy między kolejnymi obserwacjami:

***P-value***

P-value, czyli wartość p, to miara używana w statystyce do określenia, czy wyniki testu statystycznego są istotne. Wartość p reprezentuje prawdopodobieństwo uzyskania wyników równie ekstremalnych jak zaobserwowane, zakładając, że hipoteza zerowa jest prawdziwa. Niska wartość p (zazwyczaj poniżej 0,05) sugeruje, że można odrzucić hipotezę zerową, wskazując, że dane dostarczają dowodów na rzecz hipotezy alternatywnej.

***Test Ljung-Boxa***

Test Ljung-Boxa to statystyczny test używany do oceny, czy dane są niezależne w czasie, czyli czy nie występuje w nich autokorelacja w wielu opóźnieniach (lagach). Jest szczególnie użyteczny w analizie szeregów czasowych do sprawdzenia, czy model dobrze dopasowuje dane, czyli czy reszty są losowe. Niska wartość p wskazuje, że dane mogą wykazywać autokorelację, co sugeruje, że model jest niedopasowany.

***Test Jarque-Bera***

Test Jarque-Bera jest testem statystycznym sprawdzającym, czy reszty modelu statystycznego mają rozkład normalny. Test ten opiera się na analizie skośności (asymetrii) i kurtozy (spłaszczenia) danych. Jeśli wartość p jest niska, odrzuca się hipotezę zerową o normalności, co oznacza, że reszty mogą nie być normalnie rozłożone.

***Autokorelacja***

Autokorelacja odnosi się do zależności między wartościami danej zmiennej w różnych momentach czasu w ramach szeregu czasowego. Oznacza to, że wcześniejsze wartości mogą wpływać na przyszłe obserwacje. Obecność autokorelacji może sugerować, że dane nie są niezależne, co może wpływać na wyniki modeli statystycznych i wymaga uwzględnienia w analizie.

***Heteroskedastyczność***

Heteroskedastyczność odnosi się do sytuacji, w której wariancja błędów modelu (reszt) nie jest stała w całym zakresie danych. Jest to problematyczne w modelach regresji, ponieważ może prowadzić do nieefektywnych estymacji parametrów i błędnych wniosków statystycznych. Testy takie jak test Breuscha-Pagana mogą pomóc w wykryciu heteroskedastyczności.

***Normalność reszt***

Normalność reszt to założenie w wielu modelach statystycznych, które zakłada, że błędy modelu są normalnie rozłożone. Normalność jest istotna, ponieważ wiele testów statystycznych opiera się na tym założeniu, aby zapewnić poprawne wnioskowanie. Sprawdzenie normalności reszt można przeprowadzić za pomocą testu Jarque-Bera lub analizy graficznej.

## 1.2 Model ARIMA

Model ARIMA (ang. *AutoRegressive Integrated Moving Average*) [11] to jeden z najpopularniejszych modeli stosowanych w analizie szeregów czasowych. Jest to model liniowy, który łączy trzy kluczowe elementy: autoregresję (AR), różnicowanie (I) i średnią ruchomą (MA).

***Model autoregresyjny AR(p)***

Autoregresja polega na przewidywaniu wartości zmiennej w danym momencie na podstawie wcześniejszych obserwacji tej samej zmiennej [13]. Model *AR(p)*, gdzie parametr *p* oznacza liczbę opóźnionych obserwacji (lagów), zakłada, że wartość zmiennej zależy od jej wcześniejszych wartości:

gdzie to wartość zmiennej w czasie *t*, to współczynniki modelu, to błąd losowy w czasie t.

***Model średniej ruchomej MA(q)***

Średnia ruchoma zakłada, że bieżąca wartość zmiennej jest kombinacją poprzednich błędów losowych [13]. Model *MA(q)*, gdzie *q* oznacza liczbę opóźnień błędów losowych, zakłada, że wartość zmiennej w danym momencie zależy od wcześniejszych błędów:

gdzie to to błędy losowe, to współczynniki modelu.

***Model ARIMA(p,d,q)***

Model *ARIMA* łączy w sobie model autoregresji *AR*(*p*), średniej ruchomej *MA(q)* oraz różnicowanie (*d*) [13]. Model ten opisuje szereg czasowy za pomocą kombinacji wcześniejszych wartości zmiennej oraz błędów losowych, przy jednoczesnym uwzględnieniu odpowiedniej liczby różnicowań, w celu zapewnienia stacjonarności. Ogólna postać modelu *ARIMA(p, d, q)* wygląda następująco:

gdzie to przewidywana wartość zmiennej w czasie *t*, to współczynniki autoregresji, to współczynniki średniej ruchomej, błąd losowy w czasie *t*.

Model ARIMA jest szeroko stosowany w prognozowaniu szeregów czasowych ze względu na swoją elastyczność oraz zdolność do modelowania różnych wzorców, takich jak trendy i sezonowość [12]. Jednak, dobór odpowiednich wartości parametrów modelu wymaga analizy danych oraz testów diagnostycznych, takich jak autokorelacja reszt czy testy stacjonarności.

Całość materiału potrzebna do zrozumienia zagadnień tej pracy została omówiona na kursie „Szeregi czasowe” w toku studiów, więc przytoczyłam tylko najbardziej kluczowe aspekty tej wiedzy.

# Rozdział 2. Metody uczenia maszynowego dla szeregów czasowych

Uczenie maszynowe (*ang. machine learning*) [2] to gałąź sztucznej inteligencji (*ang. artificial intellligence*) [4] zajmująca się tworzeniem różnorodnych algorytmów i modeli analizy danych, w której systemy samodzielnie się uczą na podstawie wprowadzonych danych. System uczenia maszynowego jest więc trenowany, a nie programowany w rozumieniu klasycznym. Przedstawia się mu wiele przykładów istotnych dla rozważanego problemu, a następnie znajduje wzorzec statystyczny, który umożliwia systemowi „wymyślenie” reguły poszukiwania rozwiązania. Uczenie maszynowe jest ściśle powiązane ze statystyką matematyczną, ale różni się od niej pod kilkoma ważnymi względami. Ze względu na niestandardowy sposób podejścia do rozwiązywania problemów, uczenie maszynowe umożliwia analizę złożonych, dużych zbiorów danych, których badanie nie byłoby możliwe przy pomocy tradycyjnego programowania i metod statystycznych. W rezultacie uczenie maszynowe wykorzystuje niewiele teorii matematycznych. Jest to praktyczna dyscyplina, w której idee są częściej dowodzone empirycznie, niż teoretycznie.

W zbiorze metod uczenia maszynowego można wyróżnić niezwykle popularne sieci neuronowe (*ang. neural networks*) [8] oraz uczenie głębokie (*ang. deep learning*) [14]. Jedną z klas sieci neuronowych, stanowiących metodę uczenia głębokiego, są rekurencyjne sieci neuronowe (RNN, *ang. recurrent neural networks*) [7], a także ich wariant LSTM (*ang. long short-term memory*) [9], które zostaną opisane w niniejszym rozdziale.

Obraz zawierający tekst, krąg

Opis wygenerowany automatycznie

**Rysunek 2.1. Schemat hierarchii sztucznej inteligencji w kontekście LSTM.**

Cechą charakterystyczną odróżniającą klasyczne programowanie od uczenia maszynowego jest to, że w metodzie klasycznej samodzielnie tworzymy algorytm (reguły) i dostarczamy dane wejściowe, które mają być przetwarzane zgodnie z tym algorytmem (regułami). Na ich podstawie otrzymujemy dane wyjściowe, stanowiące rozwiązanie problemu. W przypadku uczenia maszynowego wprowadzamy do systemu dane wejściowe i wyjściowe (przewidywane odpowiedzi), które dzielimy na *dane treningowe i testowe*. Dane treningowe służą do nauki modelu, natomiast dane testowe pozwalają sprawdzić jego skuteczność na nieznanych wcześniej przykładach. Można również wyodrębnić *dane walidacyjne*, które posłużą do wstępnej oceny modelu podczas jego budowy. Dostarczone dane są przez program interpretowane za pomocą neuronów, które są w swej istocie równaniami matematycznymi. Na wyjściu otrzymujemy algorytm (zestaw reguł), który można zastosować do pracy z nowymi danymi, celem uzyskania oryginalnych rozwiązań.

Obraz zawierający tekst, zrzut ekranu, diagram, design

Opis wygenerowany automatycznie

**Rysunek 2.2. Różnica między klasycznym programowaniem, a uczeniem maszynowym**

***Sieć neuronowa***

Sieć neuronowa **[7]** inspirowana jest neuronem biologicznym (rysunek 2.3). Nie jest to nowy wynalazek. Pierwszą sieć neuronową opracowali już w 1943 roku Warren McCulloch i Walter Pitts. Jej reprezentantem w świecie biologicznym jest mózg, zaś w świecie cyfrowym komputer, który interpretuje dane wejściowe i zwraca odpowiedzi. Pojedynczy neuron (węzeł) otrzymuje dane wejściowe poprzez synapsy, modelowane przez pojedynczą liczbę lub wagę (*ang. weight*), określającą siłę znaczenia danego połączenia w neuronie. Dane wejściowe są mnożone przez wagi i sumowane, celem aktywacji węzła. Wartość funkcji aktywacji jest następnie porównywana z wartością progową (odchyleniem, *ang. bias*), celem umożliwienia modelowi przesunięcia funkcji aktywacji i lepszego dopasowania do danych. Jeżeli wartość funkcji aktywacji przekracza wartość progową, na wyjściu otrzymujemy istotną wartość dodatnią (np. 1), w przeciwnym razie dostajemy wartość 0. Otrzymana wartość jest przenoszona do kolejnych neuronów, jako dana wejściowa. W ten sposób tworzy się cała sieć połączeń (sieć neuronowa). Obraz zawierający tekst, kwiat

Opis wygenerowany automatycznie

Sztuczny neuron posiada warstwy, które możemy podzielić na warstwę wejścia, warstwy ukryte i warstwę wyjścia (patrz rysunek 2.4).

**Rysunek 2.3. Przepływ informacji w neuronie naturalnym.**

Obraz zawierający tekst, diagram, linia, design

Opis wygenerowany automatycznie

**Rysunek 2.4. Schemat neuronu i sieci neuronów.**

***Funkcje aktywacji***

*Funkcja aktywacji* [6]to kluczowy element sztucznych sieci neuronowych, wybierany dla warstw ukrytych i warstwy wyjściowej, który decyduje o tym, czy i w jaki sposób neuron "aktywuje się" (przekazuje sygnał). Funkcja aktywacji wprowadza nieliniowość do modelu. Bez niej operacje na danych wejściowych składałyby się wyłącznie z iloczynu skalarnego danych wejściowych i wag, do których dodana zostałaby wartość progowa (odchylenie). Wówczas, bez względu na ilość warstw, mogłyby one uczyć się tylko liniowych transformacji danych wejściowych, co znacznie ograniczyłoby zakres przestrzeni hipotez sieci. Ponadto, niektóre z funkcji aktywacji dokonują normalizacji wyjścia, co oznacza przekształcenie wartości wyjściowej w określony zakres. Pomaga to w stabilnym trenowaniu sieci neuronowej.

Najpopularniejszą funkcją aktywacji jest *funkcja ReLU* (*ang. Rectified Linear Unit*) [15]

gdzie dla ujemnych argumentów wejściowych funkcja zwraca zawsze 0, a dla dodatnich wartości pozostają bez zmian.

Obraz zawierający linia, Wykres, diagram, Równolegle

Opis wygenerowany automatycznie

**Rysunek 2.5 Wykres funkcji ReLU.**

Użycie funkcji ReLU, ze względu na jej postać, jest wydajne, ułatwia obliczenia   
i pozwala na zwiększenie szybkości obliczeń, jednak może prowadzić do sytuacji gdy neurony nie będą się uczyć - wygenerują wartość wyjściową 0. Wartość 0 zawsze jest problematyczna w uczeniu sieci neuronowych, gdyż może powodować „zapominanie” - utratę informacji, przez co uczenie jest nieefektywne.

Chcąc uniknąć takiej sytuacji, można skorzystać z innego typu funkcji aktywacji - *funkcji sigmoidalnej* [15],postaci:

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie**Rysunek 2.6 Wykres funkcji tangensa hiperbolicznego.**

**Funkcja tanh** (tangens hiperboliczny) jest odmianą funkcji sigmoidalnej. Jej wartości są zawarte w przedziale (-1,1).

Podobnie, jak funkcja sigmoidalna, pozwala na modelowanie złożonych zagadnień. Jednak, ze względu na wartości symetryczne względem punktu (0, 0), ma ona szersze zastosowania w warstwach ukrytych sieci neuronowej, co ułatwia naukę kolejnych warstw.

Obraz zawierający linia, Wykres, diagram

Opis wygenerowany automatycznie

**Rysunek 2.7 Wykres funkcji sigmoidalnej.**

Wybór funkcji aktywacji jest bardzo ważny, gdyż ma wpływ na wydajność sieci. Każda z tych funkcji ma swoje wady i zalety, a ich dobór zależy od analizowanego problemu.

***Walidacja***

Walidacja w kontekście uczenia maszynowego to proces oceny wydajności modelu na zbiorze danych, który nie był używany podczas jego trenowania. Celem walidacji jest sprawdzenie, jak dobrze model generalizuje na nowych, niewidzianych wcześniej danych. Jest to kluczowy etap, ponieważ pomaga w identyfikacji problemów, takich jak przeuczenie (overfitting), gdzie model może doskonale dopasować się do danych treningowych, ale słabo radzić sobie z danymi testowymi lub walidacyjnymi.

Zwykle proces walidacji polega na podziale dostępnych danych na dwie części: jedną do trenowania modelu, a drugą do walidacji (oceny wydajności modelu). W trakcie trenowania, model jest regularnie oceniany na danych walidacyjnych, co pozwala monitorować jego zdolność do generalizowania oraz dostosowywać hiperparametry, aby poprawić jego ogólną wydajność.

W praktyce, walidacja jest istotnym krokiem w budowie modelu, który pomaga upewnić się, że model nie tylko "uczy się" danych treningowych, ale także potrafi skutecznie przewidywać na nowych, nieznanych danych.

Strata treningowa i strata walidacyjna to kluczowe miary oceny jakości modelu uczenia maszynowego. **Strata treningowa** mierzy błąd, który model popełnia podczas dopasowywania się do danych treningowych, wskazując, jak dobrze model "uczy się" na podstawie tych danych. Celem trenowania jest minimalizowanie tej straty, aby model jak najlepiej dopasował się do danych treningowych. **Strata walidacyjna** natomiast mierzy błąd modelu na danych walidacyjnych, które nie były używane w trakcie trenowania. Pokazuje to, jak dobrze model generalizuje i jak skutecznie potrafi przewidywać na nowych, niewidzianych wcześniej danych. Strata treningowa zazwyczaj maleje z każdą epoką treningową, ponieważ model stopniowo dopasowuje się do danych treningowych. Natomiast strata walidacyjna może początkowo maleć, ale jeśli model zaczyna przeuczać się na danych treningowych (staje się zbyt dopasowany do tych danych), strata walidacyjna może zacząć rosnąć, podczas gdy strata treningowa nadal będzie maleć. Monitorowanie obu tych miar jest istotne, aby upewnić się, że model nie przeucza się i potrafi skutecznie przewidywać na nowych danych.

***Propagacja wsteczna***

*Propagacja wsteczna* (*ang. backpropagation*) [14]to kluczowy algorytm stosowany podczas uczenia sztucznej sieci neuronowej. Polega on na minimalizacji błędu predykcji (*funkcji kosztu*) poprzez aktualizację wag i progów (odchyleń) w każdej epoce (*ang. epoch*) sieci. *Funkcja kosztu* (*ang. cost function*) [5], nazywana również funkcją straty (*ang.* *loss function*), mierzy jak dobrze model przewiduje oczekiwane wyniki. Inaczej, funkcja ta określa różnicę pomiędzy wartościami przewidywanymi przez model, a wartościami rzeczywistymi z danych treningowych. Jej minimalizacja jest kluczowym celem podczas trenowania modelu. To ona wskazuje, czy model uczy się dobrze, czy też zachodzą trudności. Jednymi z najczęściej stosowanych funkcji kosztu są:

MSE - błąd średniokwadratowy (*ang. mean squared error*).

MAE – średni błąd bezwzględny (*ang. mean absolute error*).

gdzie to wartości rzeczywiste, to wartości przewidywane, to ilość elementów.

Celem minimalizacji funkcji kosztu propagacja wsteczna często wykorzystuje metody optymalizacji. Wagi i odchylenia są aktualizowane zgodnie z regułą *gradientu spadkowego* (*ang. gradient descent*) [11], zmniejszając błąd modelu poprzez przesunięcie parametrów w kierunku przeciwnym do gradientu. Algorytm oblicza gradient, korzystając z reguły łańcuchowej [16], dla każdego parametru (wag i odchyleń) w każdej warstwie sieci, co pozwala określić w jakim stopniu dany parametr wpływa na błąd.

Algorytm propagacji wstecznej składa się z dwóch etapów:

1. *Przejście do przodu* - dane wejściowe są wprowadzane do warstwy wejściowej. Po połączeniu z odpowiednimi wagami i dodaniu odchyleń, są przekazywane do warstw ukrytych sieci. Każda warstwa ukryta używa funkcji aktywacji. Dane wyjściowe   
z ostatniej warstwy ukrytej są przekazywane do warstwy wyjściowej, gdzie stosowana jest kolejna funkcja aktywacji. Na koniec, dla danych wyjściowych obliczana jest wartość funkcji kosztu.

2. *Przejście do tyłu* – obliczona wartość funkcji kosztu jest propagowana z powrotem przez sieć (od warstwy wyjściowej do wejściowej), aby dostosować wagi i odchylenia, a przez to zminimalizować błąd w kolejnej iteracji.

Głównymi zaletami algorytmu propagacji wstecznej są m.in.:

- *łatwość stosowania* - nie wymaga szerokiej wiedzy nt. sieci neuronowych,

- *elastyczność* - może być stosowana dla szerokiego zakresu zagadnień,

- *wydajność* - przyspiesza uczenie się poprzez bieżącą aktualizację parametrów.

Do wad algorytmu zalicza się natomiast:

- *problem zanikającego gradientu* - w bardzo głębokich sieciach, szczególnie przy stosowaniu sigmoidalnych funkcji aktywacji, gradienty mogą stawać się na tyle małe, że utrudnią proces uczenia się sieci,

- *problem wybuchającego gradientu* - sytuacja odwrotna do problemu zanikającego gradientu, może powodować rozbieżność sieci w trakcie uczenia się,

- *przeuczenie modelu* – czyli jego nadmierne dopasowanie, pojawia się zwykle wówczas, gdy sieć jest zbytnio złożona lub źle regulowana.

2.1 Model RNN

*Rekurencyjne sieci neuronowe* (*RNN,* *ang. Recurrent Neural Networks***)** [7]to sieci neuronowe zaprojektowane do przetwarzania danych sekwencyjnych, m.in. szeregów czasowych. Struktura RNN umożliwia rekurencję, co oznacza, że dane wyjściowe otrzymane z jednego kroku czasowego są przekazywane jako dane wejściowe do następnego kroku. Odróżnia je to od tradycyjnych sieci neuronowych, w których  dane wejściowe i wyjściowe są traktowane niezależnie.

Każdy krok czasowy w RNN można opisać jako:

gdzie to macierz wymiaru zawierająca dane wyjściowe w kroku czasowym *t*, to ilość próbek, liczba neuronów, to funkcja aktywacji, to macierz wymiaru zawierająca dane wejściowe, gdzie to liczba wejść, to macierz wymiaru zawierająca wagi dla wejść w kroku czasowym *t*, to macierz wymiaru zawierająca wagi dla wyjść z poprzedniej iteracji, to wektor   
(o wymiarze ) odchyleń dla każdego neuronu. W kroku czasowym nie istnieją żadne dane wyjściowe, zatem zakłada się, że ich wartość równa jest 0.

Neuron RNN potrafi zachować informacje o wartości stanu w poszczególnych krokach czasowych w *komórkach pamięci* (*ang.* *memory cells*). Pojedynczy neuron RNN może zapamiętywać sekwencje z około 10 lub więcej kroków czasowych, zależnie od rozważanego zagadnienia. Jest to kluczowe w przewidywaniu zdarzeń i podejmowaniu decyzji, np.: podczas przetwarzania języka naturalnego, rozpoznawania mowy, analizy wideo, czy analizy szeregów czasowych. Neuron RNN można sobie wyobrazić jako zapętlony sam w sobie (rysunek 2.8), czyli działający rekurencyjnie. Można go też przedstawić w postaci rozwijającego na kolejne neurony (rysunek 2.9) - taka wizualizacja jest bardziej przejrzysta dla dalszych rozumowań.

Obraz zawierający diagram, Czcionka, linia, tekst

Opis wygenerowany automatycznie

**Rysunek 2.8 Pojedynczy neuron RNN.**

Obraz zawierający diagram, tekst, szkic, linia

Opis wygenerowany automatycznie

**Rysunek 2.9 Rozwinięty neuron RNN.**

***Rodzaje sieci RNN***

Modele RNN możemy podzielić na różne rodzaje:

- *Sieć sekwencyjna* (*ang. sequence-to-sequence network*) - otrzymuje dane wejściowe   
i dokonuje predykcji na ciąg kolejnych kroków czasowych, np.: podając ceny produktu z *N* poprzednich dni sieć może zostać wytrenowana celem predykcji ceny produktu na 7 kolejnych dni do przodu, czyli np. od *N-7* dnia do przyszłego tygodnia.

- *Sieć sekwencyjno-wektorowa* (*ang. sequence-to-vector network*) - otrzymuje dane wejściowe i generuje wyłącznie ostateczny wynik (lub pojedynczą wartość), np.: zbierając sekwencję zdań sieć określi, czy mają one zabarwienie pozytywne - 1, czy negatywne - 0.

- *Sieć wektorowo-sekwencyjna* (*ang. vector-to-sequence network*) – w każdym kroku czasowym podawana jest taka sama informacja, na postawie której otrzymujemy wynik, np.: przedstawiamy sygnał dźwiękowy i dostajemy jego opis.

- *Koder-dekoder (ang. encoder-decoder)* – połączenie sieci sekwencyjno-wektorowej oraz wektorowo-sekwencyjnej, stosowana przykładowo do przetwarzania języka naturalnego na inny język.

***Propagacja wsteczna w czasie***

W rekurencyjnych sieciach neuronowych stosowany jest szczególny typ algorytmu propagacji wstecznej, tj. *propagacja wsteczna w czasie* *(BPTT, ang. Backpropagation Through Time)* [17]*.* Algorytm ten rozszerza propagację wsteczną na dane, które mają strukturę czasową, umożliwiając propagowanie błędów przez wiele kroków czasowych. W sieciach rekurencyjnych, BPTT oblicza gradienty nie tylko dla wag, ale również dla "ukrytych stanów" w każdym kroku czasowym. Błąd jest propagowany zarówno wstecz przez warstwy, jak i przez czas, co oznacza, że gradienty muszą być obliczane na podstawie całej sekwencji danych. Ten rodzaj propagacji wstecznej będzie wykorzystywany w niniejszej pracy, w części praktycznej – predykcji za pomocą LSTM.

Podobnie, jak w przypadku klasycznego algorytmu propagacji wstecznej, na etapie propagacji wstecznej w czasie, może pojawić się kilka istotnych problemów. Dokonując pierwszych obliczeń przy dłuższych sekwencjach można zauważyć, że im dłuższa sekwencja tym mniej dokładne wyniki, co wydaje się sprzeczne z ideą sieci neuronowych, gdyż teoretycznie ilość danych do nauki zwiększa się. Problem ten można powiązać   
z problemem *zanikającego gradientu* (*ang. vanishing gradient*)lub *wybuchającego gradientu* (*ang. exploding gradient*) w algorytmie propagacji wstecznej.

***Problem zanikającego gradientu w algorytmie BPTT***

W algorytmie BPTT na poszczególnych krokach czasowych dochodzi wielokrotnego mnożenia wag. W przypadku, gdy wartości wag (elementy macierzy ) są z przedziału   
(-1, 1), w kolejnych krokach czasowych dochodzi do coraz mniejszej aktualizacji wag   
z poprzedniego kroku czasowego, a zatem zanikania gradientu. Efekt zanikania gradientu powoduje, że początkowy gradient w długim okresie czasu ma minimalny wpływ na końcowy wynik, utrudniając modelowi naukę i przewidywanie długoterminowych zależności.

***Problem wybuchającego gradientu w algorytmie BPTT***

Odwrotne zachowanie wykaże model, gdy wartości wag (elementy macierzy ) są   
z przedziałów . W wyniku wielokrotnego mnożenia wag, w kolejnych krokach czasowych wartości wag staną się bardzo duże, co doprowadzi do błędów numerycznych. Funkcja kosztu zacznie gwałtownie rosnąć, uniemożliwiając dalsze uczenie sieci. A czym dłuższa sekwencja tym wyniki będą bardziej odbiegały od prawdy.

Ze względu na wyżej wymienione ograniczenia, RNN nie jest odpowiednim modelem do zadań wymagających przetwarzania skomplikowanych lub długich zależności czasowych. Dlatego, w niniejszej pracy, do analizy szeregów czasowych zostanie wykorzystany *model LSTM* (*ang. Long Short-Term Memory*) **[numer(y) pozycji literatury dot. LSTM]**, który stanowi ulepszoną wersję RNN. LSTM, dzięki swojej architekturze, opierającej się na komórkach pamięci i mechaniźmie bramek, skutecznie radzi sobie zarówno z problemem zanikającego, jak i wybuchającego gradientu, co pozwala na modelowanie zależności długoterminowych.

LSTM nie tylko eliminuje główne wady klasycznych RNN, ale także cechuje się większą stabilnością i skutecznością w uczeniu sieci neuronowej. Model LSTM będzie zatem kluczowym narzędziem w realizacji celów opisanych w pracy.

## 2.2 Model LSTM

Jedną z bardziej rozbudowanych rekurencyjnych sieci neuronowych jest model LSTM (*ang. long short-term memory*) [11]. Jest on zaprojektowany tak, aby efektywnie gospodarować „pamięcią”. Jest też odpowiedzią na problem zanikającego i wybuchającego gradientu. Struktura LSTM jest o wiele bardziej skomplikowana od konwencjonalnej RNN, lecz dzięki temu model ten może swobodnie działać na długich sekwencjach danych i ich długoterminowych zależności.

***Struktura sieci LSTM***

Podstawowym zadaniem sieci LSTM jest uczenie się jakie informacje należy zapamiętywać przez dłuższy okres czasu, a jakie odrzucać. W celu ograniczenia zakresu przekazywanych informacji, w sieci LSTM wprowadzono trzy bramki (zapominającą, wejściową i wyjściową), patrz rysunek 4.10. W odróżnieniu od RNN, składającej się   
z jednej warstwy sieci neuronowej (z funkcją aktywacji sigmoidalną, tanh lub ReLU), sieć LSTM składa się z trzech warstw sigmoidalnych oraz warstwy z funkcją aktywacji­ tanh. Stan pamięci długotrwałej (*ang. cell*) oznaczono na rysunku linią zieloną, zaś stan pamięci krótkotrwałej linią czerwoną. Stan został pozbawiony niewygodnego mnożenia wag, dzięki czemu usunięto problem zanikającego lub wybuchającego gradientu. Stan posiada wagi dla każdej kolejnej bramki, dzięki czemu w wygodny sposób będzie je modyfikował, nie tracąc istotnych informacji.

Obraz zawierający tekst, diagram, zrzut ekranu, Plan

Opis wygenerowany automatycznie

**Rysunek 2.10 Schemat budowy modelu LSTM.**

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

**Rysunek 2.11 Schemat budowy bramki zapominającej modelu LSTM.**

***Bramka zapominająca* *(ang. forget gate)***pełni głownie rolę zapominania (usuwania informacji, które nie są przydatne), lecz co ciekawe ma ona też wpływ na zapamiętywanie poprzez dodanie nowych „wspomnień”. Do bramki podawane są dane wejściowe oraz , które są mnożone przez macierze wag odpowiednio i . Do sumy iloczynów danych wejściowych i wag dodawane jest odchylenie (bias). Wynik podlega działaniu sigmoidalnej funkcji aktywacji

Jeżeli wyjście z funkcji aktywacji wynosi 0, to część informacji jest zapominana, gdy zaś wynosi 1, to informacja jest przechowywana do dalszego użycia.

Obraz zawierający tekst, zrzut ekranu, diagram, Równolegle

Opis wygenerowany automatycznie

**Rysunek 4.12 Schemat budowy bramki wejściowej modelu LSTM.**

***Bramka wejściowa*** ***(ang. input gate)***decyduje o dodawaniu nowych przydatnych informacji do aktualnej pamięci długotrwałej. W tym celu dane wejściowe są najpierw poddawane działaniu sigmoidalnej funkcji aktywacji

Spośród tych danych wybierane są informacje, które mają zostać zapamiętane. Korzystając następnie z funkcji aktywacji tanh, przekazującej wartości z zakresu [-1, 1]

otrzymujemy wektor wartości nowego potencjalnego stanu długotrwałego

Jest to zatem główna warstwa generująca pamięć długotrwałą, a więc tą która w większym stopniu wpływa na wynik końcowy i jest kluczowa w strukturze całego modelu.

Uzyskany nowy stan długotrwały wykorzystamy w ***bramce wyjściowej (ang. output gate).*** Bramka ta odpowiada za wyodrębnienie z aktualnego stanu komórki istotnych informacji, które zostaną wyprowadzone na wyjściu. W pierwszej kolejności nowy wektor stanu długotrwałego poddawany jest działaniu funkcji , po czym dane wejściowe regulowane są za pomocą funkcji sigmoidalnej

Na koniec, wartości regulowane przez funkcję sigmoidalną oraz wartości wektora , są mnożone celem wyodrębnienia danych wyjściowych, a za razem danych wejściowych do kolejnej komórki

Obraz zawierający tekst, zrzut ekranu, diagram, Czcionka

Opis wygenerowany automatycznie

**Rysunek 4.13 Schemat budowy bramki wejściowej modelu LSTM.**

# Rozdział 3. Analiza wybranych szeregów czasowych metodami klasycznymi

Dane do analizy, która zostanie przeprowadzona w niniejszym rozdziale, pobrano ze strony <https://fred.stlouisfed.org/>. *Federal Reserve Economic Data (FRED),* to platforma internetowa Federal Reserve Bank of St. Louis, która oferuje otwarty dostęp do szerokiej bazy danych ekonomicznych, dotyczących m.in. inflacji, zatrudnienia, produkcji, stóp procentowych, bilansów handlowych. Jest to jedno z najważniejszych źródeł danych gospodarczych w Stanach Zjednoczonych i na świecie.

Wybrano trzy typy szeregów czasowych: bez trendu i sezonowości, z trendem, a także z trendem i sezonowością, zawierające dane rzeczywiste zgromadzone w okresach miesięcznych od 2014.01.01 do 2024.01.01. Obróbkę danych, analizę i predykcję przeprowadzono przy użyciu języka programowania Python w środowisku Microsoft Visual Studio. Do tego celu wykorzystano biblioteki: **numpy**, **pandas** i **matplotlib**.

## 3.1 Analiza szeregów czasowych – opis, wykresy, zakres zmienności

***Szereg czasowy bez trendu i sezonowości***

Jako pierwszy do analizy wybrano szereg czasowy bez widocznego trendu   
i sezonowości <https://fred.stlouisfed.org/series/REAINTRATREARAT1MO>. Przedstawia on jednomiesięczną realną stopę procentową (ang. 1-Month Real Interest Rate). Fragment danych przedstawiono w tabeli 3.1, a wizualizację całego szeregu czasowego pokazano na rysunku 3.1. Kod służący do pobrania danych podano poniżej:

**Listing 3.1 Wczytanie danych do środowiska Visual Studio Code za pomocą języka Python.**

|  |
| --- |
| data\_rate = pd.read\_csv("C:/Users/Aldona/Documents/GitHub/Machine-learning-in-time-series-analysis/data/interest\_rate.csv", sep=',', encoding='utf-8', index\_col = 'DATE', parse\_dates = True) # upewnienie się, że daty będą rozpoznawane jako daty  df\_rate = pd.DataFrame(data\_rate)  df\_rate.columns.values[0] = 'REAL\_INTEREST\_RATE'  df\_rate.index.freq = 'MS' # zaznaczam, że dane są miesięczne (zazwyczaj powinno się i tak ustawić automatycznie)  print(df\_rate.head(10)) |

Za pomocą analogicznego kodu zaimportowano także pozostałe dwa szeregi czasowe. Dane nie potrzebowały na tym etapie żadnej szczególnej obróbki. Nie było wartości pustych (null), ani innych komplikacji.

|  |  |
| --- | --- |
| **DATE** | **REAL\_INTEREST\_RATE** |
| 2014-01-01 | -1.547831 |
| 2014-02-01 | -1.563561 |
| 2014-03-01 | -0.425359 |
| 2014-04-01 | -1.576272 |
| 2014-05-01 | -1.945522 |
| 2014-06-01 | -1.248905 |
| 2014-07-01 | -2.327665 |
| 2014-08-01 | -1.696280 |
| 2014-09-01 | -2.775060 |
| 2014-10-01 | -1.153638 |

**Obraz zawierający tekst, linia, Wykres, pismo odręczne

Opis wygenerowany automatycznieTabela 3.1 Początkowe 10 wierszy danych pierwszego szeregu.**

**Rysunek 3.1 Wykres 1-miesięcznej rzeczywistej stopy procentowej.**

Można zauważyć, że dane charakteryzuje brak trendu, czy sezonowości. Widoczna jest natomiast cykliczność danych.

Można zauważyć, że dane charakteryzuje brak wyraźnego trendu, czy sezonowości. Widoczna jest natomiast cykliczność danych.

***Testy stacjonarności szeregu czasowego***

Sprawdzimy, czy powyższy szereg jest stacjonarny. W tym celu przeprowadzimy testy stacjonarności. Należy pamiętać, aby przed wykonaniem testów usunąć z danych puste wartości i ewentualne braki danych.

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

**Rysunek 3.2 Testy stacjonarności dla szeregu pierwszego przed różnicowaniem.**

Testy Phillipsa-Perrona i KPSS wskazują na stacjonarność szeregu. Natomiast test Dickeya-Fullera zaprzecza stacjonarności (rysunek 3.2). Szereg czasowy nie jest więc stacjonarny. Na rysunku 3.3 podano wyniki różnicowania szeregu.

Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

Opis wygenerowany automatycznie

**Rysunek 3.3 Wykresy przed i po różnicowaniu pierwszego szeregu**

Doprowadzenie szeregu do postaci stacjonarnej wymagało jednokrotnego różnicowania. Po zróżnicowaniu szereg jest już stacjonarny (rysunek 3.4).

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

**Rysunek 3.4 Testy stacjonarności dla szeregu pierwszego po różnicowaniu.**

***Dekompozycja szeregu czasowego***

Po zastosowaniu dekompozycji dla oryginalnego szeregu czasowego otrzymujemy

**Listing 3.2 Kod wywołujący wykresy po dekopozycji.**

|  |
| --- |
| results = seasonal\_decompose(df\_rate['REAL\_INTEREST\_RATE']) |

Po zastosowaniu dekompozycji szeregu otrzymujemy wykresy trendu, sezonowości i reszt. Analogicznie wywołane zostały wykresy dla pozostałych dwóch szeregów.

Obraz zawierający tekst, linia, Czcionka, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.5 Wykresy po dekompozycji 1-miesięcznej rzeczywistej stopy procentowej.**

Dekompozycja szeregu czasowego dotyczącego realnej stopy procentowej ujawnia kilka istotnych wniosków. Trend pokazuje, że realna stopa procentowa przez większość okresu była względnie stabilna lub spadała, osiągając najniższe wartości w latach 2020–2021, po czym nastąpił wyraźny wzrost od 2021 roku, który utrzymuje się do 2023 roku. Składnik sezonowy wykazuje regularne, cykliczne wahania, co wskazuje na występowanie powtarzalnych rocznych wzorców, prawdopodobnie związanych z cyklami gospodarki lub kwartalnymi decyzjami banków centralnych. Reszty są losowe i oscylują wokół zera, co świadczy o braku systematycznych błędów modelu, choć w niektórych okresach widać większe odchylenia, które mogą być efektem niespodziewanych zdarzeń, takich jak kryzysy gospodarcze. Ogólnie, dekompozycja pozwala zauważyć zarówno długoterminowy wzrost stóp procentowych, jak i wyraźną sezonowość, co może być przydatne przy analizie i prognozowaniu przyszłych wartości.

***Szereg czasowy z trendem***

Kolejnym szeregiem czasowym, wybranym do analizy, jest szereg opisujący indeks średnich cen konsumpcyjnych w miastach **(Proszę podać link do źródła danych)** (ang. Consumer Price Index (CPI) for all urban consumers). https://fred.stlouisfed.org/series/CPIAUCSL Fragment zbioru danych podano   
w tabeli 3.2, zaś wizualizację szeregu czasowego przedstawiono na rysunku 3.6

Dane nie potrzebowały na tym etapie żadnej szczególnej obróbki. Nie było wartości pustych (null), ani innych komplikacji.

|  |  |
| --- | --- |
| **DATE** | **CPI** |
| 2014-01-01 | 235.288 |
| 2014-02-01 | 235.547 |
| 2014-03-01 | 236.028 |
| 2014-04-01 | 236.468 |
| 2014-05-01 | 236.918 |
| 2014-06-01 | 237.231 |
| 2014-07-01 | 237.498 |
| 2014-08-01 | 237.460 |
| 2014-09-01 | 237.477 |
| 2014-10-01 | 237.430 |

**Tabela 3.2 Początkowe 10 wierszy danych drugiego szeregu.**

Obraz zawierający linia, Wykres, tekst, diagram

Opis wygenerowany automatycznie**Rysunek 3.6 Wykres wskaźnika CPI dla wszystkich konsumentów miejskich.**

Widzimy wyraźnie zaznaczający się trend. Nie zauważamy natomiast sezonowości, ani cykliczności w danych. Przeprowadzamy testy stacjonarności szeregu czasowego.

***Testy stacjonarności szeregu czasowego***

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

**Rysunek 3.7 Testy stacjonarności dla szeregu drugiego przed różnicowaniem.**

Zgodnie z oczekiwaniami wszystkie testy wskazały na niestacjonarność analizowanego szeregu czasowego (rysunek 3.7). Po pierwszym różnicowaniu testy nadal wykazywały niestacjonarność szeregu. Dopiero drugie różnicowanie doprowadziło szereg do postaci stacjonarnej (patrz rysunek 3.8 i 3.9).

Obraz zawierający tekst, Czcionka, linia, zrzut ekranu

Opis wygenerowany automatycznie

**Rysunek 3.8 Wykresy przed i po różnicowaniu drugiego szeregu.**

**Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie**

**Rysunek 3.9 Testy stacjonarności dla szeregu drugiego po różnicowaniu.**

***Dekompozycja szeregu czasowego***

Po zastosowaniu dekompozycji dla oryginalnego szeregu czasowego otrzymujemy. Widzimy wyraźnie zaznaczający się trend. Dekompozycja szeregu (rysunek 3.4) potwierdza istnienie trendu w analizowanym szeregu czasowym.

Obraz zawierający tekst, linia, Czcionka, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.10 Wykresy po dekompozycji wskaźnika CPI dla wszystkich konsumentów miejskich.**

Dekompozycja szeregu czasowego dotyczącego indeksu cen konsumpcyjnych (CPI) wskazuje na kilka istotnych wniosków. Trend wykazuje jednoznaczny wzrost na przestrzeni całego okresu, przyspieszający w latach 2021–2023, co odzwierciedla długoterminowy wzrost poziomu cen i potencjalny wpływ inflacji. Składnik sezonowy ujawnia powtarzalne, regularne wahania w cyklu rocznym, co sugeruje istnienie typowych sezonowych zmian cen, np. związanych z okresem świątecznym, popytem konsumentów lub specyficznymi wydarzeniami w gospodarce. Składnik resztowy (Resid) oscyluje wokół zera, co świadczy o losowym charakterze pozostałych odchyleń. Niemniej jednak, w niektórych okresach, zwłaszcza w latach 2022–2023, widoczne są większe odchylenia, co może wynikać z nieoczekiwanych wydarzeń, takich jak zakłócenia w łańcuchach dostaw czy globalne kryzysy. Podsumowując, CPI charakteryzuje się silnym, przyspieszającym wzrostem trendu oraz wyraźną sezonowością, co może być użyteczne przy analizie inflacji i prognozowaniu przyszłych zmian cen.

***Szereg z trendem i sezonowością***

Ostatni szereg czasowy poddany analizie dotyczy sprzedaży detalicznej materiałów budowlanych i sprzętu ogrodniczego (retail sales: building materials and garden equipment and supplies dealers) [https://fred.stlouisfed.org/series/ MRTSSM444USN](https://fred.stlouisfed.org/series/%20MRTSSM444USN). Jest to szereg z widoczną sezonowością oraz delikatnie zaznaczonym trendem. Fragment zbioru danych podano w tabeli 3.3, zaś wizualizację szeregu czasowego przedstawiono na rysunku 3.11.

Dane nie potrzebowały na tym etapie żadnej szczególnej obróbki. Nie było wartości pustych (null), ani innych komplikacji.

|  |  |
| --- | --- |
| **DATE** | **CPI** |
| 2014-01-01 | 235.288 |
| 2014-02-01 | 235.547 |
| 2014-03-01 | 236.028 |
| 2014-04-01 | 236.468 |
| 2014-05-01 | 236.918 |
| 2014-06-01 | 237.231 |
| 2014-07-01 | 237.498 |
| 2014-08-01 | 237.460 |
| 2014-09-01 | 237.477 |
| 2014-10-01 | 237.430 |

**Tabela 3.3 Początkowe 10 wierszy danych trzeciego szeregu.**

Obraz zawierający tekst, linia, pismo odręczne, Wykres

Opis wygenerowany automatycznie**Rysunek 3.11 Wykres sprzedaży detalicznej materiałów oraz sprzętu i artykułów ogrodniczych.**

Widać wyraźnie zaznaczoną sezonowość, a także delikatny trend na przestrzeni 10-ciu lat, przyśpieszający na przestrzeni ostatnich 5-ciu lat. Przeprowadzamy testy stacjonarności szeregu czasowego.

***Testy stacjonarności szeregu czasowego***

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

**Rysunek 3.12 Testy stacjonarności dla szeregu trzeciego przed różnicowaniem.**

Podobnie, jak w przypadku drugiego szeregu czasowego, wszystkie testy wskazały na niestacjonarność analizowanego szeregu (rysunek 3.12). Po pierwszym różnicowaniu test Dickeya-Fullera wykazał niestacjonarność szeregu. Drugie różnicowanie doprowadziło szereg do postaci stacjonarnej (patrz rysunek 3.13 i 3.14).

Obraz zawierający tekst, Czcionka, zrzut ekranu, pismo odręczne

Opis wygenerowany automatycznie

**Rysunek 3.13 Wykresy przed i po różnicowaniu trzeciego szeregu.**

Obraz zawierający tekst, zrzut ekranu, Czcionka, dokument

Opis wygenerowany automatycznie

**Rysunek 3.14 Testy stacjonarności dla szeregu trzeciego po różnicowaniu.**

***Dekompozycja szeregu czasowego***

1. Obraz zawierający tekst, linia, Czcionka, Wykres

   Opis wygenerowany automatycznie

**Rysunek 3.15 Wykresy po dekompozycji sprzedaży detalicznej materiałów oraz sprzętu i artykułów ogrodniczych.**

Dekompozycja dla trzeciego szeregu pokazuje kilka istotnych cech. Trend wykazuje stabilny wzrost na przestrzeni analizowanego okresu, z wyraźnym przyspieszeniem od około 2020 roku do 2022 roku, po czym obserwuje się delikatne spowolnienie w 2023 roku. Składnik sezonowy charakteryzuje się regularnymi, powtarzalnymi wahaniami w cyklu rocznym, co wskazuje na sezonowość wynikającą z cyklicznych czynników, takich jak zmiany popytu czy specyficzne wydarzenia w analizowanej dziedzinie. Reszty oscylują wokół zera i są losowe, co świadczy o braku systematycznych błędów w modelu, choć w niektórych momentach widoczne są większe odchylenia, które mogą wynikać z nagłych, nieoczekiwanych zdarzeń. Ogólnie rzecz biorąc, MAT cechuje się silnym, wzrostowym trendem i wyraźnym wpływem sezonowości, co może być przydatne przy analizie danych i przewidywaniu przyszłych wartości.

## 3.2 Dobór modelu

Funkcja **auto\_arima** z biblioteki **pmdarima** automatycznie przeszukuje przestrzeń parametrów ARIMA, wybierając te, które minimalizują kryterium informacyjne, takie jak AIC (Akaike Information Criterion) lub BIC (Bayesian Information Criterion). Ułatwia to optymalizację modelu bez potrzeby ręcznego testowania wielu kombinacji parametrów. Nie ma też potrzeby ręcznego różnicowania danych przed użyciem **auto\_arima**, ponieważ model ten automatyzuje proces wyboru najlepszego poziomu różnicowania. Poniżej przedstawiono fragment kodu oraz wyniki działania kodu (rysunek 3.16), w którym zastosowano **auto\_arima** celem doboru modelu dla analizowanych szeregów czasowych.

|  |
| --- |
| from pmdarima import auto\_arima  rate\_fit = auto\_arima(df\_rate["REAL\_INTEREST\_RATE"], trace = True)  cpi\_fit = auto\_arima(df\_cpi['CPI'], trace = True)  mat\_fit = auto\_arima(df\_materials['MAT'].dropna(), trace = True) |

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

**Rysunek 3.16 Wykres ACF dla pierwszego szeregu czasowego przed zastosowaniem modelu.**

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie**Rysunek 3.17 Wykres PACF dla pierwszego szeregu czasowego przed zastosowaniem modelu.**

Te dwa wykresy przedstawiają funkcję autokorelacji (ACF) i funkcję częściowej autokorelacji (PACF), które są kluczowe przy analizie szeregów czasowych i wyborze odpowiednich parametrów dla modelu ARIMA. Wykres ACF pokazuje, jak bieżące wartości szeregu czasowego są powiązane z wcześniejszymi wartościami (autokorelacja) na różnych opóźnieniach (lagach). Wysokie wartości na wykresie ACF wskazują na obecność trendu, sezonowości lub cykliczności. Z kolei wykres PACF eliminuje wpływ pośrednich opóźnień, co pozwala lepiej ocenić bezpośrednią zależność pomiędzy bieżącą wartością a konkretnymi opóźnieniami. Analizując te wykresy, można określić parametry p (dla PACF) i q (dla ACF), które reprezentują odpowiednio rząd autoregresji (AR) i rząd średniej ruchomej (MA) w modelu ARIMA. Na podstawie prezentowanych wykresów można zauważyć, że PACF sugeruje szybkie wygasanie po 2-3 lagach, co wskazuje na możliwy p wynoszący w okolicach 1, a ACF wykazuje dłuższe wygasanie, co może wskazywać na konieczność rozważenia mniejszego q.

Obraz zawierający tekst, Czcionka, zrzut ekranu, czarne i białe

Opis wygenerowany automatycznie

**Rysunek 3.18 Dobór parametrów modelu ARIMA dla pierwszego szeregu czasowego.**

W przypadku dopasowania modelu ARIMA do pierwszego szeregu czasowego, najlepszym modelem okazał się ARIMA(1,0,1), osiągając wartość AIC = 429.732. Model ten uwzględnia proces autoregresji rzędu pierwszego (*p = 1*) oraz proces średniej ruchomej rzędu pierwszego (*q = 1*). Obecność procesu średniej ruchomej wskazuje na silną autokorelację dla sąsiadujących (*lag = 1*) obserwacji. Inne modele, takie jak ARIMA(0,0,1), czy ARIMA(2,0,1), miały wyższe błędy dopasowania, co oznacza, że zmiana w strukturze modelu, czy zwiększenie wartości parametrów nie poprawiło znacząco dopasowania. Ponadto, model ARIMA(2,0,2) uzyskał wartość AIC = inf (rysunek 3.16), co oznacza, że nie był w stanie dopasować się do danych. Może to sugerować problemy z konwergencją lub nadmierną złożonością modelu. Zatem model ARIMA(1,0,1) oferuje dobre dopasowanie przy jednoczesnym zachowaniu prostoty   
i efektywności w modelowaniu pierwszego szeregu.

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.19 Wykres ACF dla pierwszego szeregu czasowego po zastosowaniu modelu.**

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie **Rysunek 3.20 Wykres PACF dla pierwszego szeregu czasowego po zastosowaniu modelu.**

Po zastosowaniu modelu zarówno na wykresie ACF, jak i PACF, wartości dla większości lagów znalazły się w obrębie przedziałów ufności, co oznacza, że model skutecznie uchwycił zależności czasowe w danych. Brak wyraźnych pików w ACF i PACF po zastosowaniu modelu sugeruje, że pozostałości są zbliżone do białego szumu, a struktura danych została prawidłowo odwzorowana. Szczególnie warto zwrócić uwagę, że na wykresie PACF potwierdza się wcześniejsze przypuszczenie – znaczące wartości tylko dla dwóch pierwszych lagów wskazują, że wyższe opóźnienia nie mają istotnego wpływu na modelowanie, co potwierdza zasadność zastosowanego podejścia. Model dobrze wyjaśnił strukturę danych, co wskazuje na jego poprawne dopasowanie.

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.21 Wykres ACF dla drugiego szeregu czasowego przed zastosowaniem modelu.**

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.22 Wykres PACF dla drugiego szeregu czasowego przed zastosowaniem modelu.**

Wysokie wartości na wykresie ACF wskazują na obecność trendu. Z kolei wykres PACF eliminuje wpływ pośrednich opóźnień, co pozwala lepiej ocenić bezpośrednią zależność pomiędzy bieżącą wartością, a konkretnymi opóźnieniami. Analizując te wykresy, można określić parametry p (dla PACF) i q (dla ACF), które reprezentują odpowiednio rząd autoregresji (AR) i rząd średniej ruchomej (MA) w modelu ARIMA. Na podstawie prezentowanych wykresów można zauważyć, że PACF sugeruje konsekwentne, ale łagodne wygasanie. Wartości mieszczą się w przedziale ufności około 10 laga, lecz wcześniejsze nie tworzą drastycznej różnicy, dlatego możliwe, że model z p wynoszący w okolicach 2 również da dobre prognozy. Wykres ACF wykazuje wysokie wartości w okolicy 2. laga, co może wskazywać na konieczność rozważenia q wynoszącego 2.

Obraz zawierający tekst, Czcionka, zrzut ekranu, dokument

Opis wygenerowany automatycznie

**Rysunek 3.23 Dobór parametrów modelu ARIMA dla drugiego szeregu czasowego.**

W przypadku drugiego szeregu czasowego, najlepszym modelem okazał się ARIMA(0,2,2), osiągając wartość AIC = 229.137 (rysunek 3.17). Model ten uwzględnia proces średniej ruchomej rzędu drugiego (*q = 2*) oraz dwa różnicowania (*d = 2*). Jest to zgodne ze spostrzeżeniem, że dane wymagają podwójnego różnicowania, aby sprowadzić szereg do postaci stacjonarnej. Modele takie jak ARIMA(1,2,2), ARIMA(0,2,3), czy ARIMA(0,2,1) dawały wyższe wartości AIC. Zatem zwiększenie wartości parametrów lub wprowadzenie zmian w strukturze modelu nie poprawiło dopasowania.

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

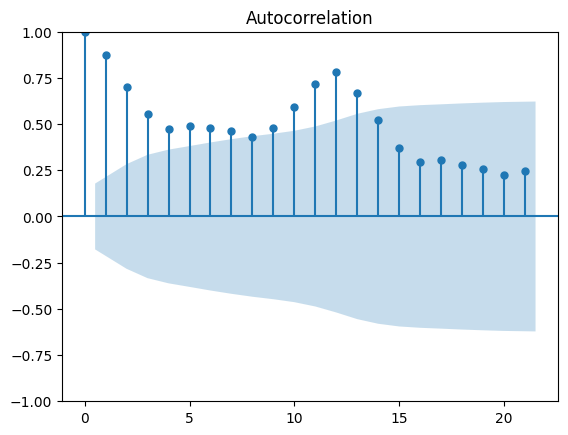
Opis wygenerowany automatycznie

**Rysunek 3.24 Wykres ACF dla drugiego szeregu czasowego po zastosowaniu modelu.**

Obraz zawierający tekst, zrzut ekranu, linia, Wykres

Opis wygenerowany automatycznie **Rysunek 3.25 Wykres PACF dla drugiego szeregu czasowego po zastosowaniu modelu.**

Po zastosowaniu modelu zarówno na wykresie ACF, jak i PACF, wartości dla większości lagów znalazły się w obrębie przedziałów ufności, co oznacza, że model skutecznie uchwycił zależności czasowe w danych. Brak wyraźnych pików w ACF i PACF po zastosowaniu modelu sugeruje, że pozostałości są zbliżone do białego szumu, a struktura danych została prawidłowo odwzorowana. Szczególnie warto zwrócić uwagę, że na wykresie PACF potwierdza się wcześniejsze przypuszczenie – znaczące wartości tylko dla dwóch pierwszych lagów wskazują, że wyższe opóźnienia nie mają istotnego wpływu na modelowanie, co potwierdza zasadność zastosowanego podejścia. Model dobrze wyjaśnił strukturę danych, co wskazuje na jego poprawne dopasowanie.



**Rysunek 3.26 Wykres ACF dla trzeciego szeregu czasowego przed zastosowaniem modelu.**

Obraz zawierający tekst, zrzut ekranu, diagram, linia

Opis wygenerowany automatycznie

**Rysunek 3.27 Wykres PACF dla trzeciego szeregu czasowego przed zastosowaniem modelu.**

Wysokie wartości na wykresie ACF wskazują na obecność trendu i sezonowości. Z kolei wykres PACF eliminuje wpływ pośrednich opóźnień, co pozwala lepiej ocenić bezpośrednią zależność pomiędzy bieżącą wartością, a konkretnymi opóźnieniami. Analizując te wykresy, można określić parametry p (dla PACF) i q (dla ACF). Na podstawie prezentowanych wykresów można zauważyć, że PACF wygasa, w okolicach laga 12. Rośnie, ale nadal największe wartości wynosi w okolicach 2, stąd można przypuszczać, że p wynoszące 2 da najlepsze wyniki. Wykres ACF wykazuje wysokie wartości w okolicy 2. laga, co może wskazywać na konieczność rozważenia q wynoszącego 2.

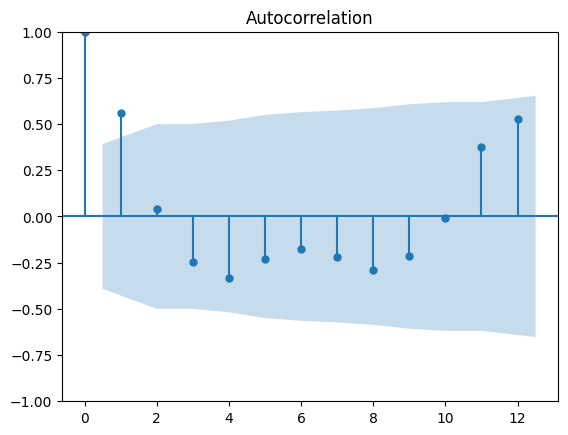
Ponadto dodaję składniki sezonowe (2,2,2,12) w modelu ARIMA, aby uchwycić powtarzające się wzorce sezonowe w danych. Parametr P=2 oznacza, że model uwzględnia dwa sezonowe opóźnienia autoregresyjne, co pozwala opisać zależności między bieżącymi wartościami a wartościami z dwóch poprzednich sezonów. Parametr D=2 wskazuje, że dane są różnicowane sezonowo dwa razy, co eliminuje powtarzalne wzorce i stabilizuje sezonową komponentę. Parametr Q=2 oznacza dwa sezonowe opóźnienia w składniku średniej ruchomej, co uwzględnia sezonowe błędy prognozy w dwóch poprzednich sezonach. Okres s=12 określa, że sezonowość w danych występuje co 12 jednostek czasu, czyli co 12 miesięcy w przypadku danych miesięcznych. Te parametry razem pozwalają modelowi dokładnie odwzorować powtarzające się zmiany w danych wynikające z cyklu sezonowego.

Dla trzeciego szeregu czasowego, najlepszym modelem okazał się ARIMA(2,1,2) osiągając wartość AIC = 2277.976, co czyni go najskuteczniejszym wyborem spośród testowanych modeli. Model ten uwzględnia proces autoregresji rzędu drugiego (*p = 2*), jedno różnicowanie (*d = 1*) oraz proces średniej ruchomej rzędu drugiego (*q = 2*). Dane cechują się zatem zarówno zależnościami autoregresyjnymi, jak i silną autokorelacją dla sąsiadujących obserwacji (*lag = 2*).

Obraz zawierający tekst, Czcionka, zrzut ekranu, papier

Opis wygenerowany automatycznie

**Rysunek 3.28 Dobór parametrów modelu ARIMA dla trzeciego szeregu czasowego.**



**Rysunek 3.29 Wykres ACF dla trzeciego szeregu czasowego po zastosowaniu modelu.**

Obraz zawierający tekst, zrzut ekranu, diagram, Wykres

Opis wygenerowany automatycznie

**Rysunek 3.26 Wykres PACF dla trzeciego szeregu czasowego po zastosowaniu modelu.**

Po zastosowaniu modelu zarówno na wykresie ACF, jak i PACF, wartości dla większości lagów znalazły się w obrębie przedziałów ufności, co oznacza, że model skutecznie uchwycił zależności czasowe w danych. Brak wyraźnych pików w ACF i PACF po zastosowaniu modelu sugeruje, że pozostałości są zbliżone do białego szumu, a struktura danych została prawidłowo odwzorowana. Szczególnie warto zwrócić uwagę, że na wykresie PACF potwierdza się wcześniejsze przypuszczenie – znaczące wartości tylko dla dwóch pierwszych lagów wskazują, że wyższe opóźnienia nie mają istotnego wpływu na modelowanie, co potwierdza zasadność zastosowanego podejścia. Model dobrze wyjaśnił strukturę danych, co wskazuje na jego poprawne dopasowanie.

Użycie funkcji **auto\_arima** umożliwiło szybkie i skuteczne znalezienie najbardziej dostosowanych parametrów modelu ARIMA dla każdego spośród analizowanych szeregów czasowych, minimalizując wartości kryterium informacyjnego AIC. Wyniki pokazują, że szeregi te charakteryzują się odmienną strukturą parametrów.

# Rozdział 4 Prognozowanie szeregów czasowych

Modele LSTM (a także inne modele RNN) nie wymagają badania stacjonarności szeregu czasowego, czy sprowadzania szeregu czasowego do postaci stacjonarnej, gdyż potrafią efektywnie pracować na danych niestacjonarnych. Jest to jedna z ich kluczowych zalet, w porównaniu do klasycznych metod analizy szeregów czasowych, w których niejednokrotnie konieczna jest stacjonaryzacja danych. Jeżeli jednak model LSTM lub RNN nie dawałby satysfakcjonujących wyników, warto rozważyć stacjonaryzację szeregu czasowego. Proces ten, polegający na usunięciu trendu lub sezonowości, może pomóc modelowi lepiej nauczyć się wzorców w danych, co często przekłada się na poprawę jakości prognoz. Aby umożliwić efektywne uczenie modelu oraz jego weryfikację, ważnym elementem obróbki danych jest również analiza długości szeregu czasowego oraz odpowiedni podział danych na zbiór treningowy i testowy.

4.1 Prognozowanie metodą ARIMA

Metodą klasyczną, wykorzystywaną w tej pracy do prognozowania wszystkich trzech analizowanych szeregów czasowych, jest model ARIMA. Dobór modelu oraz testy stacjonarności zostały przeprowadzone we wcześniejszym rozdziale. Dokładne działanie kodu, krok po kroku, zostanie opisane dla pierwszego szeregu (jednomiesięcznej rzeczywistej stopy procentowej). Dla pozostałych dwóch szeregów przedstawione zostaną tylko wyniki, gdyż działanie kodu jest analogiczne.

***Analiza dokładności dopasowania modelu ARIMA dla pierwszego szeregu***

Do stworzenia modelu korzystamy z funkcji **ARIMA** z biblioteki **statsmodels**. Na wejściu wymaga ona podania danych treningowych szeregu czasowego oraz dobranych wcześniej parametrów modelu.

**Listing 4.1 Kod generujący model ARIMA dla pierwszego szeregu.**

|  |
| --- |
| from statsmodels.tsa.arima.model import ARIMA  model\_arima\_rate=ARIMA(train\_rate, order=(1,0,1))  model\_arima\_rate = model\_arima\_rate.fit()  model\_arima\_rate.summary() |

Model jest dopasowywany do danych treningowych za pomocą metody **.fit()**, która optymalizuje parametry modelu tak, aby jak najlepiej odwzorowywały dane.

***Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie***

**Rysunek 4.1 Model ARIMA dla pierwszego szeregu.**

W modelu uwzględniono jeden składnik autoregresyjny (AR) oraz jeden składnik średniej ruchomej (MA). Wyniki wskazują, że składnik autoregresyjny (AR.L1) jest istotny statystycznie (*p = 0.000 < 0.05*) i ma silny wpływ na zmienną zależną (współczynnik 0.7266), podczas gdy składnik średniej ruchomej (MA.L1) jest nieistotny (*p = 0.793*). Jego wartość powinna być równa zero, a wyznaczona wartość (-0.0305) jest przypadkowa. Stała modelu jest na granicy istotności (*p = 0.073 > 0.05*). Kryteria informacyjne AIC (323.239) i BIC (333.497) wskazują na wystarczające dopasowanie modelu. Testy diagnostyczne potwierdzają brak autokorelacji reszt, dobre dopasowanie modelu (Ljung-Box, *p = 0.95 > 0.05*) oraz brak heteroskedastyczności (*p = 0.86*). Test Jarque-Bera (*p = 0.02 < 0.05*) sugeruje pewne odchylenia od normalności reszt. Ogólnie model dobrze opisuje dane.

***Prognozowanie dla pierwszego szeregu czasowego***

W tym kroku generowana jest prognoza na danych testowych za pomocą modelu ARIMA na poziomie ufności 95%. Prognoza jest gotowa do dalszej analizy i porównania   
z danymi rzeczywistymi.

**Listing 4.2 Kod generujacy predykcje dla szeregu pierwszego.**

|  |
| --- |
| rate\_pred = model\_arima\_rate.get\_forecast(len(test\_rate.index))  rate\_pred\_df = rate\_pred.conf\_int(alpha = 0.05)  rate\_pred\_df["Predictions"] = model\_arima\_rate.predict(start = rate\_pred\_df.index[0], end = rate\_pred\_df.index[-1])  rate\_pred\_df.index = test\_rate.index  rate\_pred\_out = rate\_pred\_df["Predictions"]  rate\_pred\_out |

Obraz zawierający tekst, Wykres, diagram, linia

Opis wygenerowany automatycznie**Rysunek 4.2 Prognozy dla szeregu czasowego jednomiesięcznej rzeczywistej stopy procentowej generowane za pomocą modelu ARIMA.**

Zgodnie z otrzymanymi rezultatami (rysunek 4.2) widzimy, że model przewidział  
początkową tendencję wzrostową danych. Jednak, ze względu na cykliczno-szumowy charakter szeregu, model w dalszym przedziale czasowym nie był w stanie uchwycić wzorców. W zamian dąży do wartości średniej obserwacji. Warto zauważyć, że przedział ufności zawiera większość danych testowych, nie rozszerza się, a zatem nie przewiduje dalszego trendu wzrostowego (lub spadkowego) stóp procentowych.

***Analiza dokładności dopasowania modelu ARIMA dla drugiego szeregu***

Analogicznie do pierwszego szeregu czasowego tworzymy model ARIMA dla szeregu drugiego (rysunek 4.3). Model uwzględnia dwa składniki średniej ruchomej (MA)   
i różnicowanie rzędu drugiego, wskazujące na potrzebę eliminacji trendu w danych. Oba składniki MA (ma.L1 = -0.3181 i ma.L2 = -0.4176) są istotne statystycznie (*p = 0.000 < 0.05*), co oznacza, że wpływają na zmienność CPI. Wartość sigma2 = 0.2764 jest również istotna (*p = 0.000 < 0.05*) i wskazuje na wariancję reszt. Kryteria AIC = 152.614, BIC = 160.244 sugerują dobre dopasowanie modelu do danych.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie

**Rysunek 4.3 Model ARIMA dla drugiego szeregu.**

**Rysunek 4.4 Prognozy dla szeregu czasowego opisującego indeks średnich cen konsumpcyjnych w miastach generowane za pomocą modelu ARIMA.Obraz zawierający tekst, zrzut ekranu, Wykres, linia

Opis wygenerowany automatycznie**

Testy diagnostyczne wykazały brak autokorelacji reszt (Ljung-Box, *p = 0.73 > 0.05*) oraz istotnej heteroskedastyczności reszt (*p = 0.17 > 0.05*). Test Jarque-Bera (*p = 0.12 > 0.05*) pokazuje, że reszty mogą być zgodne z rozkładem normalnym. Model jest dobrze dopasowany do danych.

***Prognozowanie dla drugiego szeregu czasowego***

Na rysunku 4.4 widzimy, że model przewidział istniejący trend w danych. Dane testowe w pełni zawierają się w obszarze ufności, co oznacza, że predykcje są prawidłowe.

***Analiza dokładności dopasowania modelu ARIMA dla trzeciego szeregu czasowego***

Model ARIMA w tym szeregu, z racji na jego sezonowość o okresie 12 miesięcy, przyjmuje postać SARIMA(2,1,2)×(2,2,2)12.

**Listing 4.3 Kod generujący model ARIMA dla trzeciego szeregu.**

|  |
| --- |
| model\_arima\_mat=ARIMA(train\_mat, order=(2,1,2), seasonal\_order=(2,2,2,12)) |

Obraz zawierający tekst, menu, zrzut ekranu, Czcionka

Opis wygenerowany automatycznie

**Rysunek 4.5 Model ARIMA dla trzeciego szeregu.**

Składniki niesezonowe modelu AR (ar.L1 = -1.0727, ar.L2 = -0.9067) i MA (ma.L1 = 1.0665, ma.L2 = 0.7610) są istotne statystycznie (*p = 0.000 < 0.05*), co oznacza, że wpływają znacząco na modelowanie zmiennej zależnej. W części sezonowej tylko składnik ma.S.L12 (-1.5492) jest istotny (*p = 0.035 < 0.05*). Wartości kryteriów AIC (1292.761) i BIC (1313.125) wskazują na relatywnie skomplikowaną strukturę modelu. Test Ljung-Boxa (*p = 0.51 > 0.05*) sugeruje brak autokorelacji reszt, test Jarque-Bera   
(*p = 0.00 < 0.05*) wskazuje, że reszty nie mają rozkładu normalnego, zaś test na heteroskedastyczność (*p = 0.00 < 0.05*) ujawnia zmienność wariancji reszt. Pomimo problemów z resztami, model jest dobrze dopasowany.

Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie***Prognozowanie dla trzeciego szeregu czasowego***

**Rysunek 4.6 Prognozy dla szeregu czasowego opisującego wartość sprzedaży detalicznej materiałów budowlanych oraz sprzętu i artykułów ogrodniczych generowane za pomocą modelu ARIMA.**

Model prognozuje trend i sezonowość występujące w danych. Dane testowe znajdują się w obrębie obszaru zmienności, co oznacza, że predykcje są prawidłowe. Kształt przedziału ufności wskazuje na dalszą sezonowość i trend wzrostowy danych. Możemy stwierdzić, że model dokonał wystarczająco precyzyjnej prognozy.

## 4.2 Prognozowanie metodą LSTM

Prognozowanie szeregów czasowych można również przeprowadzić za pomocą metod uczenia maszynowego. Modelem, który zostanie użyty do wykonania prognoz, będzie opisany wcześniej LSTM. Dokładne działanie kodu, krok po kroku, zostanie opisane dla pierwszego szeregu czasowego, zawierającego dane o jednomiesięcznej rzeczywistej stopie procentowej. Dla pozostałych dwóch szeregów czasowych przedstawione zostaną jedynie wyniki, gdyż działanie kodu jest w ich przypadku analogiczne.

***Podział danych na dane treningowe i testowe. Skalowanie danych***

Z racji tego, że dane obejmują okres od 2014-01 do 2024-01, czyli 121 miesięcy, dane podzielono mniej więcej w stosunku 2:1. Dane z początkowych 96 miesięcy potraktowano jako zbiór treningowy, a pozostałe jako zbiór testowy.

**Listing 4.4 Podział danych na dane treningowe i testowe dla pierwszego szeregu.**

|  |
| --- |
| # Podział danych na zbiór treningowy i testowy  train\_size\_rate = 96  # Pierwsze 96 miesięcy to zbiór treningowy  train\_data\_rate = scaled\_data\_rate[:train\_size\_rate]  test\_data\_rate = scaled\_data\_rate[train\_size\_rate - sequence\_length:] |

Aby proces uczenia przebiegał sprawniej, za pomocą biblioteki **MinMaxScaler** dane zostały następnie zeskalowane do przedziału (0, 1).

**Listing 4.5 Normalizacja danych dla pierwszego szeregu.**

|  |
| --- |
| # Normalizacja danych  scaler\_rate = MinMaxScaler(feature\_range=(0, 1))  scaled\_data\_rate = scaler\_rate.fit\_transform(df\_rate) |

|  |  |  |
| --- | --- | --- |
| **DATE** | **REAL\_INTEREST\_RATE** | **SCALED\_RATE** |
| 2014-01-01 | -1.54783 | 0.367893 |
| 2014-02-01 | -1.56356 | 0.366291 |
| 2014-03-01 | -0.42536 | 0.482203 |
| 2014-04-01 | -1.57627 | 0.364997 |
| 2014-05-01 | -1.94552 | 0.327394 |
| 2014-06-01 | -1.2489 | 0.398335 |
| 2014-07-01 | -2.32766 | 0.288477 |
| 2014-08-01 | -1.69628 | 0.352776 |
| 2014-09-01 | -2.77506 | 0.242916 |
| 2014-10-01 | -1.15364 | 0.408037 |
| 2014-11-01 | -0.0394 | 0.521508 |
| 2014-12-01 | -2.86519 | 0.233737 |

**Tabela 4.1 Zestawienie fragmentu niezeskalowanych i zeskalowanych danych  
dla zbioru reprezentującego pierwszy szereg czasowy.**

|  |  |  |
| --- | --- | --- |
| **DATE** | **CPI** | **SCALED\_CPI** |
| 2014-01-01 | 235.288 | 0.011745 |
| 2014-02-01 | 235.547 | 0.017368 |
| 2014-03-01 | 236.028 | 0.027811 |
| 2014-04-01 | 236.468 | 0.037363 |
| 2014-05-01 | 236.918 | 0.047133 |
| 2014-06-01 | 237.231 | 0.053928 |
| 2014-07-01 | 237.498 | 0.059725 |
| 2014-08-01 | 237.460 | 0.058900 |
| 2014-09-01 | 237.477 | 0.059269 |
| 2014-10-01 | 237.430 | 0.058249 |
| 2014-11-01 | 236.983 | 0.048544 |
| 2014-12-01 | 236.252 | 0.032674 |

**Tabela 4.2 Zestawienie fragmentu niezeskalowanych i zeskalowanych danych  
dla zbioru reprezentującego drugi szereg czasowy.**

|  |  |  |
| --- | --- | --- |
| **DATE** | **MATERIALS** | **SCALED\_MATERIALS** |
| 2014-01-01 | 19688 | 0.031619 |
| 2014-02-01 | 18801 | 0.000000 |
| 2014-03-01 | 24103 | 0.188999 |
| 2014-04-01 | 30137 | 0.404092 |
| 2014-05-01 | 33416 | 0.520978 |
| 2014-06-01 | 30072 | 0.401775 |
| 2014-07-01 | 28642 | 0.350800 |
| 2014-08-01 | 26446 | 0.272520 |
| 2014-09-01 | 26195 | 0.263573 |
| 2014-10-01 | 27329 | 0.303996 |
| 2014-11-01 | 24821 | 0.214594 |
| 2014-12-01 | 24056 | 0.187324 |

**Tabela 4.3 Zestawienie fragmentu niezeskalowanych i zeskalowanych danych  
dla zbioru reprezentującego trzeci szereg czasowy.**

***Sekwencje dla modelu LSTM***

Za pomocą funkcji **create\_sequences** dane treningowe zostały podzielone na sekwencje o zadanej długości (**seq\_length**). Każdej sekwencji wejściowej (tworzonej na podstawie kolejnych zestawów danych) funkcja przypisuje wartość wyjściową, która jest obserwacją w czasie bezpośrednio następującą po tej sekwencji. W ten sposób tworzony jest zbiór danych, na którym model LSTM będzie mógł uczyć się zależności czasowych   
i na tej podstawie prognozować przyszłe wartości szeregu czasowego. Parametr **sequence\_length** określa, ile obserwacji ma zawierać każda sekwencja wejściowa. Dzięki tej funkcji, dane są odpowiednio przygotowywane do trenowania modelu LSTM, który analizuje wzorce w szeregach czasowych.

**Listing 4.6 Kod służący do przygotowania danych pierwszego szeregu do trenowania modelu LSTM.**

|  |
| --- |
| # Funkcja tworząca sekwencje dla LSTM  def create\_sequences(data, seq\_length):      X, y = [], []      for i in range(len(data) - seq\_length):          X.append(data[i:i+seq\_length, 0])          y.append(data[i+seq\_length, 0])      return np.array(X), np.array(y)  # Parametry  sequence\_length = 12  # Długość sekwencji wejściowych  # Generowanie sekwencji  X\_train\_rate, y\_train\_rate = create\_sequences(train\_data\_rate, sequence\_length)  X\_test\_rate, y\_test\_rate = create\_sequences(test\_data\_rate, sequence\_length)  # Zmiana formatu danych dla LSTM: (samples, timesteps, features)  X\_train\_rate = X\_train\_rate.reshape((X\_train\_rate.shape[0], X\_train\_rate.shape[1], 1))  X\_test\_rate = X\_test\_rate.reshape((X\_test\_rate.shape[0], X\_test\_rate.shape[1], 1)) |

Powyższy fragment kodu służy do przygotowania danych do trenowania modelu LSTM, który jest wykorzystywany w analizie szeregów czasowych. Parametr *sequence\_ length = 12* oznacza, że każda sekwencja wejściowa będzie składała się z 12 obserwacji (np. wartości szeregu czasowego z 12 kolejnych miesięcy). Następnie, za pomocą funkcji **create\_sequences**, generowane są sekwencje dla zbiorów treningowego i testowego. Funkcja ta, na podstawie zadanego rozmiaru sekwencji, tworzy dane wejściowe *X\_train\_rate*, *X\_test\_rate* oraz odpowiadające im wartości wyjściowe *y\_train\_rate*, *y\_test\_rate*. Każda sekwencja wejściowa składa się z 12 obserwacji, a odpowiadająca jej wartość wyjściowa to wartość obserwacji w kolejnym punkcie czasowym. Ostatecznie, dane wejściowe są przekształcane do formatu (*samples, timesteps, features*) wymaganego przez model LSTM, gdzie *samples* to liczba próbek, *timesteps* to długość sekwencji,   
a *features* to liczba cech (w tym przypadku 1, ponieważ mamy jedną zmienną wejściową).

***Definiowanie modelu LSTM***

Kojelnym krokiem jest zdefiniowanie modelu sieci neuronowej LSTM (Long Short-Term Memory) przy użyciu klasy **Sequential** z biblioteki **Keras**. Model jest używany do przewidywania wartości szeregów czasowych.

**Listing 4.7 Budowanie** **modelu LSTM dla pierwszego szeregu.**

|  |
| --- |
| # Budowa modelu LSTM  model\_rate = Sequential()  model\_rate.add(LSTM(100, activation='tanh', return\_sequences=False, input\_shape=(sequence\_length, 1)))  model\_rate.add(Dense(1))  # Warstwa wyjściowa  model\_rate.compile(optimizer='adam', loss='mse')  model\_rate.summary() |

Pierwsza linia kodu *model\_rate = Sequential()* inicjalizuje model sekwencyjny, co oznacza, że warstwy sieci neuronowej są dodawane jedna po drugiej. Następnie, za pomocą linii kodu *model\_rate.add(LSTM(100, activation = 'tanh', return\_sequences = False, input\_shape = (sequence\_length, 1)))*, dodawana jest warstwa LSTM ze 100 jednostkami (neuronami), które analizują dane wejściowe oraz funkcją aktywacji *tanh* wprowadzającą nieliniowość. Parametr *return\_sequences = False* oznacza, że model zwróci tylko ostatnią wartość z sekwencji, co jest typowe w zadaniach regresyjnych. *input\_shape = (sequence \_length, 1)* wskazuje, że wejście ma postać sekwencji o długości *sequence\_length* i jednej cesze (np. stopa procentowa). Kolejna warstwa, *model\_rate.add (Dense(1))*, to warstwa wyjściowa składająca się z pojedynczego neuronu. Oznacza to, że model ma przewidywać jedną wartość na podstawie przetworzonych danych. Na koniec, model jest kompilowany za pomocą optymalizatora Adam oraz funkcji straty MSE (średni błąd kwadratowy), co pozwala na jego efektywne trenowanie w zadaniu regresji.

Obraz zawierający tekst, zrzut ekranu, Czcionka, numer

Opis wygenerowany automatycznie **Rysunek 4.7 Modelu LSTM dla pierwszego szeregu.**

Model LSTM składa się z dwóch warstw: warstwy LSTM o 100 jednostkach, która jest odpowiedzialna za uchwycenie zależności sekwencyjnych, oraz warstwy Dense z jednym neuronem do predykcji wyjściowej. Model zawiera łącznie 40,901 trenowalnych parametrów, co oznacza, że wszystkie parametry mogą być optymalizowane podczas treningu. Brak parametrów nietrenowalnych wskazuje, że nie użyto np. zamrożonych wag lub warstw o stałych wartościach. Model jest relatywnie prosty i kompaktowy, co może być korzystne dla wydajności, ale jego efektywność w predykcji będzie zależeć od jakości danych i poprawności hiperparametrów, takich jak liczba jednostek LSTM czy funkcje aktywacji.

***Trenowanie modelu***

Na tym etapie kodu model LSTM jest trenowany na danych treningowych, aby nauczyć się wzorców i zależności w szeregach czasowych.

**Listing 4.8 Trenowanie modelu LSTM dla pierwszego szeregu.**

|  |
| --- |
| # Trenowanie modelu  history\_rate = model\_rate.fit(X\_train\_rate, y\_train\_rate,                                epochs=250,                                batch\_size=12,                                validation\_data=(X\_test\_rate, y\_test\_rate),                                verbose=1) |

Funkcja **model\_rate.fit** jest używana do przeprowadzenia procesu treningowego, trwającego przez określoną liczbę epok (iteracji). W tym przypadku model będzie trenowany przez 250 epok. W epoce dane treningowe są przetwarzane, a model aktualizuje swoje wagi po każdej partii danych (mini-batch) o rozmiarze 12 próbek, co pomaga   
w efektywnym uczeniu się i minimalizacji funkcji straty. Dodatkowo, przy pomocy danych walidacyjnych (testowych) dokonywana jest ocena, jak dobrze model generalizuje, czyli jak dobrze przewiduje wyniki na nowych, niewidzianych wcześniej danych. Dzięki temu możliwe jest monitorowanie zarówno postępów w nauce na danych treningowych, jak   
i ocena modelu na danych, które nie były używane do jego trenowania. Wartości funkcji straty są wyświetlane po każdej epoce, co umożliwia śledzenie procesu uczenia.

***Tworzenie prognozy***

Za pomocą funkcji **model\_rate.predict** wykonywana jest prognoza na dwóch różnych zbiorach danych: treningowym (*X\_train\_rate*) i testowym (*X\_test\_rate*). Pierwsza linia kodu, *train\_predict\_rate = model\_rate.predict(X\_train\_rate)* generuje prognozę na danych treningowych, na których model był wcześniej uczony. Prognoza te odzwierciedla, jak dobrze model dopasował się do danych, na których był trenowany. Linia *test\_predict\_rate = model\_rate.predict(X\_test\_rate)* wykonuje prognozę na danych testowych, które nie były używane w procesie trenowania modelu. Dzięki temu możemy ocenić, jak dobrze model potrafi generalizować swoje przewidywania na nieznanych danych. Uzyskane prognozy są analizowane w celu oceny wydajności i jakości modelu.

**Listing 4.8 Generowanie predykcji modelu LSTM dla pierwszego szeregu.**

|  |
| --- |
| # Prognoza  train\_predict\_rate = model\_rate.predict(X\_train\_rate)  test\_predict\_rate = model\_rate.predict(X\_test\_rate) |

***Odskalowanie danych***

Na tym etapie dokonujemy odskalowania (ang. *inverse scaling*) danych, wcześniej przekształconych do skali innej niż oryginalna, aby model mógł je łatwiej przetworzyć. Do odtworzenia pierwotnych wartości, czyli przekształcenia prognoz i rzeczywistych wyników z powrotem do oryginalnej postaci, używamy skalera **scaler\_rate**.

**Listing 4.8 Odskalowanie danych dla pierwszego szeregu.**

|  |
| --- |
| # Odskalowanie danych  train\_predict\_rate = scaler\_rate.inverse\_transform(np.concatenate((train\_predict\_rate, np.zeros((train\_predict\_rate.shape[0], 1))), axis=1))[:, 0]  test\_predict\_rate = scaler\_rate.inverse\_transform(np.concatenate((test\_predict\_rate, np.zeros((test\_predict\_rate.shape[0], 1))), axis=1))[:, 0]  y\_train\_rescaled\_rate = scaler\_rate.inverse\_transform(np.concatenate((y\_train\_rate.reshape(-1, 1), np.zeros((y\_train\_rate.shape[0], 1))), axis=1))[:, 0]  y\_test\_rescaled\_rate = scaler\_rate.inverse\_transform(np.concatenate((y\_test\_rate.reshape(-1, 1), np.zeros((y\_test\_rate.shape[0], 1))), axis=1))[:, 0] |

Pierwsza część kodu dotyczy prognoz modelu. Do prognoz dla danych treningowych (*train\_predict\_rate*) i testowych (*test\_predict\_rate*) dodawana jest kolumna wypełniona zerami, aby dostosować dane do formatu oczekiwanego przez funkcję **inverse\_transform**. Następnie funkcja ta przekształca dane z powrotem do ich oryginalnej skali, a [:, 0] służy do wyodrębnienia pierwszej kolumny, czyli odwróconych prognoz. Analogicznie postępujemy z wartościami rzeczywistymi *y\_train\_rate* i *y\_test\_rate*. Dzięki temu, zarówno prognozy, jak i dane rzeczywiste, są teraz w tej samej skali, co pozwala na ich porównanie w kontekście wartości rzeczywistych. Ten krok jest kluczowy dla interpretacji wyników i oceny skuteczności modelu.

***Funkcja straty dla pierwszego szeregu czasowego***

Na rysunku 4.8, przedstawiającym funkcję straty dla 250 epok, widać wyraźną różnicę pomiędzy stratą treningową, a walidacyjną. W pierwszych epokach procesu treningowego obydwie straty są duże. Po czym, strata treningowa spada w tempie znacznie szybszym od straty walidacyjnej. Z kolei strata walidacyjna, choć również spada, utrzymuje się na wyższym poziomie. Oznacza to, że model dobrze dopasowuje się do danych treningowych, ale jego zdolność generalizacji do danych testowych rozwija się wolniej. Można także zauważyć, że obydwie funkcje straty szybko dążą do konkretnej, minimalnej wartości. Zatem, aby uniknąć przeuczenia modelu, nie ma sensu ustalać zbyt dużej liczby epok.

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie

**Rysunek 4.8 Funkcja straty podczas trenowanie do 250. epoki dla pierwszego szeregu.**

Moment rozpoczynającego się przeuczenia (*ang. overfitting*) modelu zilustrowano na rysunku 4.9. Widzimy, że w okolicach 350 epoki, pomimo malejącej straty treningowej, strata walidacyjna zaczyna rosnąć.

Obraz zawierający tekst, Wykres, linia, Czcionka

Opis wygenerowany automatycznie

**Rysunek 4.9 Funkcja straty podczas trenowanie od 250. Do 500. epoki dla pierwszego szeregu.**

Oznacza to, że po około 400 epokach, model zbytnio dopasowuje się do danych treningowych, a jego predykcje będą odbiegać od rzeczywistości dla danych testowych.

Analizując dane na rysunkach 4.8 i 4.9 można przyjąć, że liczba 250 epok dla pierwszego szeregu czasowego jest odpowiednia. Funkcja straty jest wówczas bliska minimum ale jednocześnie nie dochodzi do zjawiska przeuczenia modelu.

***Wyniki prognoz dla pierwszego szeregu czasowego***

Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie

**Rysunek 4.10 Prognozy dla szeregu czasowego jednomiesięcznej rzeczywistej stopy procentowej generowane za pomocą modelu LSTM.**

Jak łatwo zauważyć na rysunku 4.10 predykcja odwzorowała dane testowe w miarę dokładnie. Szereg ten jest cykliczny, przypomina błądzenie losowe i nie ma typowych wzorców, dlatego też jego prognozowanie jest utrudnione. Niemniej jednak, model LSTM dosyć dobrze poradził sobie z tym zadaniem. Jednak, wartość błędu średniokwadratowego RMSE (ang. Root Mean Squared Error), będącego pierwiastkiem ze średniej kwadratów błędów (różnic pomiędzy rzeczywistymi, a przewidywanymi wartościami szeregu), to:



RMSE to miara określająca, jak dobrze model prognozuje wartości na podstawie danych rzeczywistych. Miara ta jest bardzo wrażliwa na wartości odstające, znajdujące się w analizowanym szeregu. Stąd wartość RMSE, szczególnie dla zbioru testowego, jest stosunkowo duża w porównaniu ze skalą zagadnienia (wartości stopy procentowej wahają się od około -8 do 6). W takim przypadku RMSE może mylnie wskazywać na słabą wydajności modelu.

***Funkcja straty dla drugiego szeregu czasowego***

Na rysunku 4.11 niemalże od 1 do około 80 epoki widoczny jest wyraźny wzrost straty walidacyjnej. Może zachodzić zjawisko niedouczenia (*ang. underfitting*), co oznacza, że model może być zbyt prosty, przez co uczy się niewłaściwych wzorców.   
W takim przypadku, aby zbadać, czy jesteśmy w stanie uzyskać możliwie najmniejsze straty, zwiększamy liczbę epok (rysunek 4.12).

Obraz zawierający tekst, Wykres, zrzut ekranu, linia

Opis wygenerowany automatycznie

**Rysunek 4.11 Funkcja straty podczas trenowanie do 250. epoki dla drugiego szeregu.**

Obraz zawierający tekst, Wykres, zrzut ekranu, linia

Opis wygenerowany automatycznie

**Rysunek 4.12 Funkcja straty podczas trenowanie do 1000. epoki dla drugiego szeregu.**

Widzimy, że około 900 epoki zaczęło występować zjawisko przeuczenia modelu. Na podstawie analizy wykresów umieszczonych na rysunkach 4.11 i 4.12 można wnioskować, że najbardziej optymalną liczbą epok dla tego modelu będzie 650-700.

***Wyniki prognoz dla drugiego szeregu czasowego***

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie

**Rysunek 4.13 Prognozy dla szeregu czasowego opisującego indeks średnich cen konsumpcyjnych w miastach generowane za pomocą modelu LSTM.**

Jak można zauważyć, model LSTM dosyć dokładnie przewidział wartości danych treningowych i testowych. Jest to szereg z wyraźnym trendem, który został dobrze uchwycony. Obserwacje te potwierdzają wartości błędu RMSE:



W drugim szeregu czasowym nie ma istotnych wartości odstających. Stąd błąd RMSE dla zbioru testowego na poziomie 3.0, w porównaniu ze skalą zagadnienia - wartości indeksu średnich cen konsumpcyjnych wahają się od około 230 do 320, jest niewielki. W ten sposób można potwierdzić dobrą wydajność modelu.

***Funkcja straty dla trzeciego szeregu czasowego***

W przypadku trzeciego szeregu czasowego funkcje straty dla zbiorów treningowego   
i testowego mają trend malejący (rysunek 4.14). Około 230 epoki można zauważyć stagnację funkcji straty. Zwiększając liczbę epok (rysunek 4.15), około 350 epoki zauważamy wzrost straty walidacyjnej, przy nieznacznie malejącej stracie treningowej. Takie zachowanie wskazuje na zjawisko przeuczenia się modelu. Analizując wyniki wnioskujemy, że najbardziej optymalną liczbą epok dla trzeciego szeregu czasowego jest 300 epok.

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie

**Rysunek 4.14 Funkcja straty podczas trenowanie do 250. epoki dla trzeciego szeregu.**

Obraz zawierający tekst, Wykres, linia, zrzut ekranu

Opis wygenerowany automatycznie

**Rysunek 4.15 Funkcja straty podczas trenowanie do 5000. epoki dla trzeciego szeregu.**

***Wyniki prognoz dla szeregu trzeciego***

Obraz zawierający tekst, linia, Wykres, diagram

Opis wygenerowany automatycznie

**Rysunek 4.16 Prognozy dla szeregu czasowego opisującego wartość sprzedaży detalicznej materiałów budowlanych oraz sprzętu i artykułów ogrodniczych generowane za pomocą modelu LSTM.**

Jak można zauważyć na rysunku 4.16 predykcja obrazuje przyszłe wartości danych testowych w dobry sposób. Model LSTM poradził sobie z wychwyceniem trendu   
i sezonowości, charakteryzujących trzeci szereg czasowy. Warto też zwrócić uwagę na to, że w predykcji danych testowych po głównym skoku sezonowym uwidoczniono poprawnie dwa mniejsze skoki (rysunek 4.17). Skoki te nie były wyraźnie widoczne we wcześniejszych sezonach, a jednak model je przewidział. To potwierdza dokładne   
i wydajne działanie modeli LSTM.

Obraz zawierający tekst, linia, Wykres, Czcionka

Opis wygenerowany automatycznie

**Rysunek 4.17 Prognozy dla szeregu czasowego opisującego wartość sprzedaży detalicznej materiałów budowlanych oraz sprzętu i artykułów ogrodniczych generowane za pomocą modelu ARIMA z zaznaczonymi charakterystycznymi elementami.**

Na dobrą jakość prognozy dla trzeciego szeregu czasowego i wydajność modelu LSTM wskazuje wartość błędu RMSE dla zbioru treningowego i testowego.



Błąd RMSE dla zbioru testowego na poziomie 2.0, w porównaniu ze skalą zagadnienia - wartości sprzedaży detalicznej materiałów budowlanych wahają się od około 1800 do 5200, jest niewielki. Można zatem potwierdzić dobrą wydajność modelu.

## 4.3 Porównanie prognoz uzyskanych przy pomocy modeli ARIMA i LSTM

W przypadku pierwszego szeregu czasowego model ARIMA dobrze radzi sobie   
z uśrednieniem szumów oraz wyodrębnieniem trendu. Jednakże, prognozy są zbytnio wygładzone, co oznacza, że model nie reaguje na nagłe zmiany w danych.   
W przeciwieństwie do ARIMA, model LSTM jest w stanie uchwycić złożone, nieliniowe wzorce i zależności w danych czasowych. Predykcje są bardziej dynamiczne i lepiej odwzorowują krótkoterminowe fluktuacje, co jest widoczne szczególnie na danych treningowych. Model LSTM trafniej prognozuje dane.

Obraz zawierający tekst, diagram, Wykres, linia

Opis wygenerowany automatycznie

**Rysunek 4.18 Prognozy dla szeregu czasowego jednomiesięcznej rzeczywistej stopy procentowej generowane za pomocą modelu ARIMA i LSTM.**

Porównując obie metody (rysunek 4.18), możemy wnioskować, że ARIMA jest bardziej odpowiednia dla danych z wyraźnym trendem i umiarkowaną zmiennością. Natomiast LSTM dobrze sprawdza się w analizie skomplikowanych i nieregularnych danych. Wybór pomiędzy tymi metodami zależy od specyfiki danych.

Dla drugiego szeregu czasowego (rysunek 4.19) model ARIMA dobrze odwzorowuje trend wzrostowy w danych. Jego predykcje są jednak mocno wygładzone, przez co nie uwzględniają niewielkich fluktuacji. Przedział ufności wyraźnie poszerza się w czasie, co odzwierciedla rosnącą niepewność prognozy. Model LSTM bardzo dobrze dopasowuje się do danych treningowych i w miarę skutecznie odwzorowuje dane testowe. Brak wizualizacji przedziału ufności oznacza, że trudniej ocenić wiarygodność predykcji   
w kontekście niepewności. Niemniej jednak, model LSTM trafniej prognozuje dane.

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie

**Rysunek 4.19 Prognozy dla szeregu czasowego opisującego indeks średnich cen konsumpcyjnych w miastach generowane za pomocą modelu ARIMA i LSTM.**

Model ARIMA wygenerował prognozę skupiając się na trendzie długoterminowym   
w danych i uwzględniając niepewność w formie przedziału ufności. Z kolei LSTM lepiej odwzorowuje szczegóły, co pozwala na dokładniejsze prognozy krótkoterminowe. Jednak, bez wizualizacji niepewności wyniki mogą być mniej przejrzyste. Wybór odpowiedniego modelu zależy od potrzeby.

W przypadku trzeciego szeregu czasowego (rysunek 4.20) model ARIMA uchwycił trend i sezonowość danych. Obszar ufności odzwierciedla większą niepewność danych   
w dłuższym horyzoncie czasowym. Model LSTM doskonale odwzorowuje szczegóły, uwzględniając zarówno trend i sezonowość, jak i krótkoterminowe fluktuacje w danych. Model ten trafniej prognozuje dane.

Obraz zawierający tekst, Wykres, linia, diagram

Opis wygenerowany automatycznie

**Rysunek 4.20 Prognozy dla szeregu czasowego opisującego wartość sprzedaży detalicznej materiałów budowlanych oraz sprzętu i artykułów ogrodniczych generowane za pomocą modelu ARIMA i LSTM.**

Model ARIMA uwzględnia trend długoterminowy i ilustruje niepewność za pomocą przedziału ufności. Jest to szczególnie przydatne przy analizie stabilnych i bardziej przewidywalnych danych. Z kolei LSTM lepiej odwzorowuje złożone wzorce i zmiany krótkoterminowe, przez co jest adekwatny w przypadku danych o dużej zmienności.

# Podsumowanie

W analizie przedstawiono dwa podejścia do prognozowania szeregu czasowego: klasyczny model ARIMA oraz nowoczesny model oparty na sieci neuronowej LSTM. Obie metody mają swoje zalety , a ich wybór zależy od charakterystyki danych i celu analizy.

Model ARIMA wykazuje swoją wartość w analizie danych z wyraźnym trendem i umiarkowaną zmiennością. ARIMA skutecznie odwzorowuje ogólny trend i sezonowość danych, generując wygładzone prognozy, które są konserwatywne i przewidywalne. Przedział ufności wizualizuje rosnącą niepewność w miarę wydłużania horyzontu prognozy, co jest istotnym aspektem w kontekście podejmowania decyzji na podstawie przewidywań. Dzięki tym cechom ARIMA jest odpowiednim wyborem w przypadku analiz, gdzie stabilność i prostota prognozowania mają kluczowe znaczenie.

Model LSTM wyróżnia się zdolnością do uchwycenia złożonych, nieliniowych wzorców i zależności w danych. Dzięki temu lepiej radzi sobie z analizą krótkoterminowych fluktuacji oraz nieregularnych danych, co czyni go bardziej elastycznym narzędziem. LSTM, choć dynamiczniejszy i bardziej precyzyjny w krótkoterminowych prognozach, może być trudniejszy w interpretacji i mniej stabilny w przypadku danych testowych. Niemniej jednak jego zdolność do adaptacji sprawia, że jest wyjątkowo skuteczny w analizie niestandardowych danych, które nie wpisują się w proste wzorce trendów i sezonowości.

Modele klasyczne, takie jak ARIMA, nadal są wartościowym narzędziem i warto z nich korzystać w przypadkach, gdy dane charakteryzują się przewidywalnymi wzorcami. Ich prostota oraz zdolność do uśredniania szumów zapewniają stabilność prognoz.

Modele uczenia maszynowego, takie jak LSTM, oferują ogromny potencjał w analizie bardziej złożonych danych, które mogą wykazywać nieregularności lub niespodziewane zmiany. Są one szczególnie skuteczne w prognozowaniu krótkoterminowym i analizie danych o złożonych wzorcach.

Dzięki swojej elastyczności i zdolności do uchwycenia niestandardowych zależności, modele uczenia maszynowego, takie jak LSTM, stanowią obiecującą ścieżkę rozwoju analizy danych czasowych. Rozwijanie tej ścieżki jest kluczowe dla tworzenia bardziej precyzyjnych i dostosowanych do realnych warunków modeli prognostycznych, co czyni je niezwykle wartościowym narzędziem w szybko zmieniającym się świecie analityki danych.

# Bibliografia

[1] Szeregi czasowe. Praktyczna analiza i predykcja z wykorzystaniem statystyki i uczenia maszynowego, Aileen Nielsen, 2020, Helion

[2] Uczenie maszynowe z użyciem Scikit-Learn, Keras i TensorFlow. Wydanie III, Aurélien Géron, 2023, Helion

[3] Analiza i prognozowanie szeregów czasowych, Warszawa, 2015, Artur Suchwałko, Adam Zagdański, Wydawnictwo Naukowe PWN

[4] Predictive Data Analysis: Leveraging RNN and LSTM Techniques

for Time Series Dataset, Harsh Agarwala, Ginika Mahajana, Anita Shrotriyaa, Deepika Shekhawata, Manipal University Jaipur, India

[5] LONG SHORTTERM MEMORY Neural Computation, Sepp Hochreiter, Fakultat fur Informatik, Technische Universitat Munchen, Munchen, Germany

[6] TimeSeries Prediction Method Based onVariant LSTM Recurrent Neural Network, Jiaojiao Hu1,· Xiaofeng Wang1,·Ying Zhang,·Depeng Zhang,·Meng Zhang,· Jianru Xue, Springer Science+Business Media, LLC, part of Springer Nature 2020

[7] Recurrent Neural Networks for Learning Sequential Data, Adrian Horzyk, Akademia Górniczo-Hutnicza w Krakowie, Kraków

[8] The StatQuestIllustrated Guideto MachineLearning, Josh Starmer, StatQuest Publications, 2022

[9] Long Short-Term Memory Networks With Python, Jason Brownlee, Machine Learning Mastery, 2017

[10] Peter J. Brockwell Richard A. Davis, Introduction to Time Series and Forecasting, Springer, 2002

[11] Khulood Albeladi, Bassam Zafar, Ahmed Mueen, Time Series Forecasting using LSTM and ARIMA, (IJACSA) International Journal of Advanced Computer Science and Applications, 2023

[12] Robert H. Shumway, David S. StofferTime, Series Analysis and Its Applications, Springer, 2011

[13] Mingda Zhang , Time Series: Autoregressive models AR, MA, ARMA, ARIMA,University of Pittsburgh, 2018

[14] Marcin Wolter, Deep Learning, IFJ PAN, 2020

[15] A Review of Activation Function for Artificial Neural Network, Andrinandrasana David Rasamoelina; Fouzia Adjailia; Peter Sinčák, IEEE, 2020

|  |  |
| --- | --- |
| POLITECHNIKA RZESZOWSKA im. I. Łukasiewicza | Rzeszów, Rok |
| Wydział Elektrotechniki i Informatyki |  |
|  |  |

**STRESZCZENIE PRACY DYPLOMOWEJ INŻYNIERSKIEJ**

**TYTUŁ PRACY**

Autor: Imię Nazwisko, nr albumu: (wybierz symbol studiów)-123456 Opiekun: (tytuł naukowy przed) Imię Nazwisko(tytuł naukowy po)

Słowa kluczowe: (max. 5 słów kluczowych w 2 wierszach, oddzielanych przecinkami)

(tekst streszczenia - max. 10 wierszy)

RZESZOW UNIVERSITY OF TECHNOLOGY Rzeszow, Rok

Faculty of Electrical and Computer Engineering

**DIPLOMA THESIS (BS) ABSTRACT**

**TYTUŁ PRACY W WERSJI ANGIELSKIEJ**

Author: Imię Nazwisko, code: (wybierz symbol studiów) -123456 Supervisor: (tytuł naukowy przed) Imię Nazwisko (tytuł naukowy po)

Key words: (max. 5 słów kluczowych w 2 wierszach, oddzielanych przecinkami)

(tekst streszczenia w jęz. angielskim - max. 10 wierszy)