Wrocław 18.06.2023

Sprawozdanie

Analiza i predykcja kursu akcji CD Projekt

Jakub Piotrowski; 266502

1. Wprowadzenie

Celem tego projektu było zastosowanie zdobytych umiejętności z zakresu analizy danych, uczenia maszynowego oraz przetwarzania szeregów czasowych do analizy i predykcji kursu akcji CD Projekt.

2. Spodziewane wyniki

Spodziewałem się, że wykorzystując różne techniki analizy danych i uczenia maszynowego, będę w stanie zidentyfikować istotne wzorce i zależności w danych, a także skutecznie przewidzieć przyszłe wartości akcji.

3. Czego nauczyłem się podczas ćwiczenia

Wykonując to ćwiczenie, miałem okazję praktycznie zastosować wiele technik i metod, które poznałem na wykładach. Nauczyłem się również, jak istotna jest staranna analiza i przetwarzanie danych przed zastosowaniem modeli predykcyjnych. Nauczyłem się również interpretować wyniki oceny jakości modeli.

4. Wykorzystane narzędzia

Do przeprowadzenia badań użyłem języka programowania Python, popularnego w dziedzinie analizy danych i uczenia maszynowego. Wykorzystałem wiele bibliotek, takich jak pandas do manipulacji danymi, matplotlib i seaborn do wizualizacji, oraz scikit-learn do budowy i oceny modeli predykcyjnych. Dodatkowo, skorzystałem z biblioteki Keras dla modeli sieci neuronowych LSTM oraz ARIMA.

5. Opis przeprowadzonych badań

W pierwszej kolejności dokonałem przetwarzania danych, sprawdzając, czy istnieją puste wiersze lub brakujące dane. Wizualizując dane z pliku CSV, zidentyfikowałem braki w kolumnie "volume". Następnie przeanalizowałem dane, korzystając z biblioteki pandas i seaborn. Obliczyłem statystyczne cechy oraz wymiar fraktalny, wyznacznik Hursta oraz entropię.

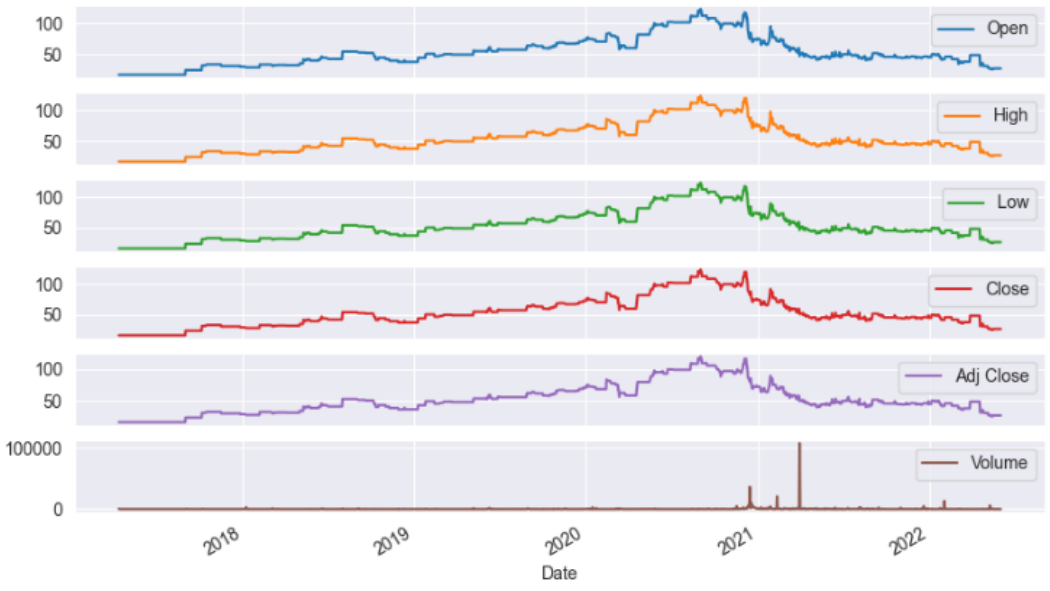
Następnie zastosowałem różne techniki uczenia maszynowego z biblioteki scikit-learn, takie jak regresja liniowa, regresja wielomianowa, sieci neuronowe LSTM oraz model ARIMA. Do oceny jakości modeli skorzystałem z metryk takich jak średni błąd bezwzględny (MAE), średni błąd kwadratowy (MSE), pierwiastek błędu średniokwadratowego (RMSE) oraz współczynnik determinacji (R^2).

6. Analiza danych

Fragment danych:

```
Date,Open,High,Low,Close,Adj Close,Volume
2017-04-12,16.430000,16.430000,16.430000,16.430000,15.956594,600
2017-04-13,16.430000,16.430000,16.430000,16.430000,15.956594,0
2017-04-17,16.430000,16.430000,16.430000,16.430000,15.956594,0
2017-04-18,16.430000,16.430000,16.430000,16.430000,15.956594,0
2017-04-19,16.430000,16.430000,16.430000,16.430000,15.956594,0
2017-04-20,16.430000,16.430000,16.430000,16.430000,15.956594,0
```

Wizualizacja danych:



Entropia:

```
,Entropy
Open,4.927274378613267
High,4.893274391154588
Low,4.899420117678805
Close,4.848489597915166
Adj Close,4.881076906797823
Volume,4.67375331733955
```

Wymiar fraktalny:

```
,Fractal Dimension
Open,1.6116326671440986
High,1.6136488148781107
Low,1.614027932790695
Close,1.6189721189430648
Adj Close,1.6187132410562253
Volume,1.1105971585256926
```

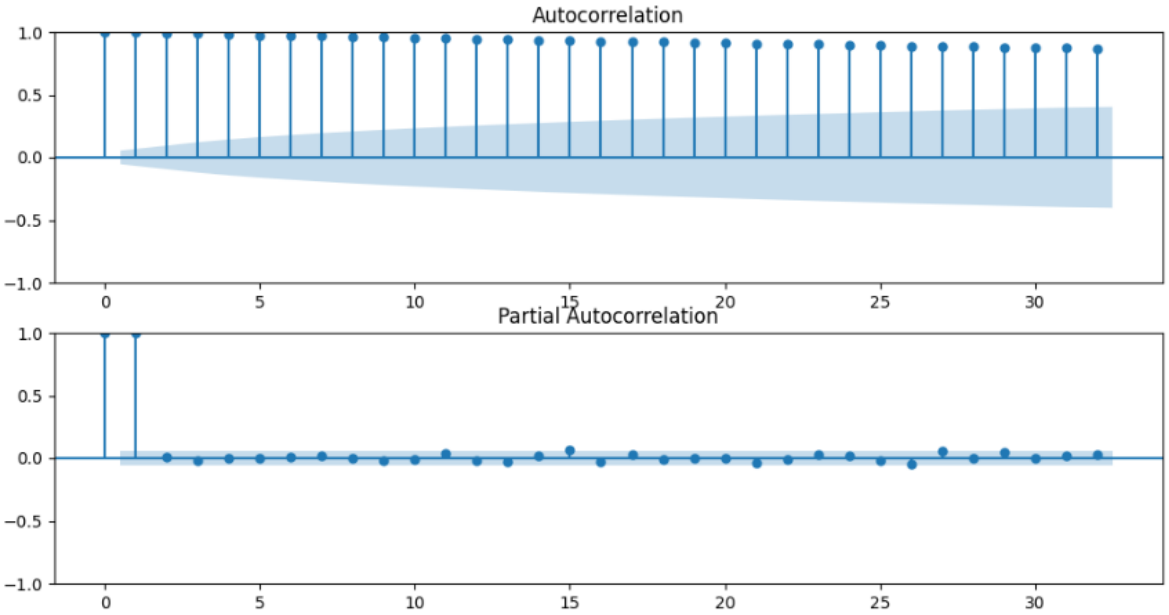
Wykładnik Hursta:

```
,Hurst Exponent
Open,0.905783477368113
High,0.8941875979022065
Low,0.950637414662752
Close,0.8924051152534928
Adj Close,0.9267431791860175
Volume,0.7620676391473428
```

Wartości statystyczne:

	Open	High	Low	Close	Adj Close	Volume
count	1295.0	1295.0	1295.0	1295.0	1295.0	1295.0
mean	52.86866338455599	52.98709431505792	52.80454446332047	52.927489687258685	51.65025770965252	683.6293436293437
std	23.93165152131327	24.023310254786132	23.880836626123838	23.968231803704306	23.197780365411006	3625.1266719376936
min	16.43	16.43	16.43	16.43	15.956594	100.0
25%	37.400002	37.400002	37.400002	37.400002	36.322376	166.66666666666669
50%	48.5	48.5	48.5	48.5	47.830933	300.0
75%	63.5	63.575001	63.5	63.5	61.670341	550.0
max	124.0	124.0	124.0	124.0	120.427124	108200.0
median	48.5	48.5	48.5	48.5	47.830933	300.0
kurtosis	0.2233418201992845	0.20995617444226422	0.2346068071939129	0.21865911539885552	0.2239103898253485	643.6100605487535

Dla modelu ARIMA wygenerowałem wykresy ACF i PACF w celu ustalenia parametrów: autoregresyjny (p), rząd różnicowania (d), parametr średniej ruchomej (q):



7. Użyte modele predykcyjne:

W ramach tego projektu zastosowałem różne modele predykcyjne do analizy i predykcji kursu akcji CD Projekt. Oto krótki opis użytych modeli:

- **Regresja liniowa** to jedna z podstawowych technik uczenia maszynowego. Wykorzystałem ją do znalezienia liniowej zależności między zmiennymi niezależnymi a ceną akcji. Model ten estymuje wartość docelową (ceny akcji) na podstawie liniowej kombinacji cech (np. objętości obrotu). Regresja liniowa pozwoliła mi zrozumieć, czy występuje zależność liniowa między zmiennymi a ceną akcji.

- **Regresja wielomianowa** rozszerza regresję liniową, umożliwiając modelowanie nieliniowych zależności. W tym przypadku wykorzystałem wielomianowy stopień jako rozszerzenie liniowej regresji, aby uwzględnić bardziej skomplikowane wzorce w danych. Model ten pozwolił mi uwzględnić nieliniowe relacje między cechami a ceną akcji.

- **LSTM (Long Short-Term Memory)** to rodzaj rekurencyjnej sieci neuronowej, która jest szczególnie skuteczna w analizie i predykcji szeregów czasowych. Wykorzystałem model LSTM do analizy i predykcji kursu akcji CD Projekt. LSTM potrafi uwzględnić zależności czasowe i długoterminowe zależności w danych. Przeszkoliłem ten model na historycznych danych, a następnie użyłem go do przewidywania przyszłych wartości akcji.

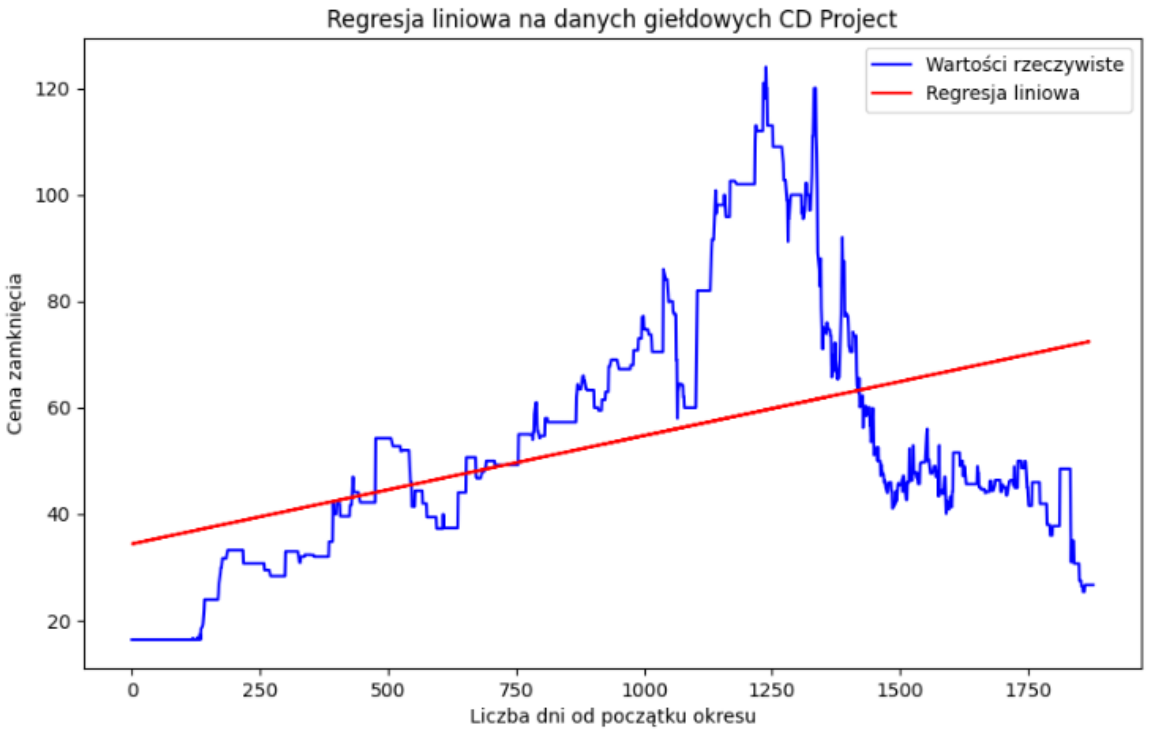
- **Model ARIMA (Autoregressive Integrated Moving Average)** jest popularnym modelem do analizy i predykcji szeregów czasowych. Składa się z trzech składowych: autoregresji (AR), różnicy (I) i średniej ruchomej (MA). Wykorzystałem model ARIMA do analizy i predykcji kursu akcji CD Projekt. Ten model uwzględnia trend, sezonowość i szum w danych, umożliwiając prognozowanie przyszłych wartości akcji na podstawie ich historycznych wzorców.

Każdy z tych modeli miał swoje zalety i ograniczenia, dlatego zastosowałem je równolegle, aby uzyskać różne perspektywy i wyniki predykcji.

8. Wyniki

Różne modele były stosowane w celu przewidywania przyszłych wartości akcji CD Projekt. Zastosowane modele skutecznie identyfikowały wzorce w danych i były w stanie przewidzieć przyszłe wartości akcji z pewną dokładnością. Wśród nich, modele regresji wielomianowej okazały się szczególnie skuteczne.

Model regresji liniowej:



Parametry oceny jakości modelu

mean_absolute_error: 17.412784786351843

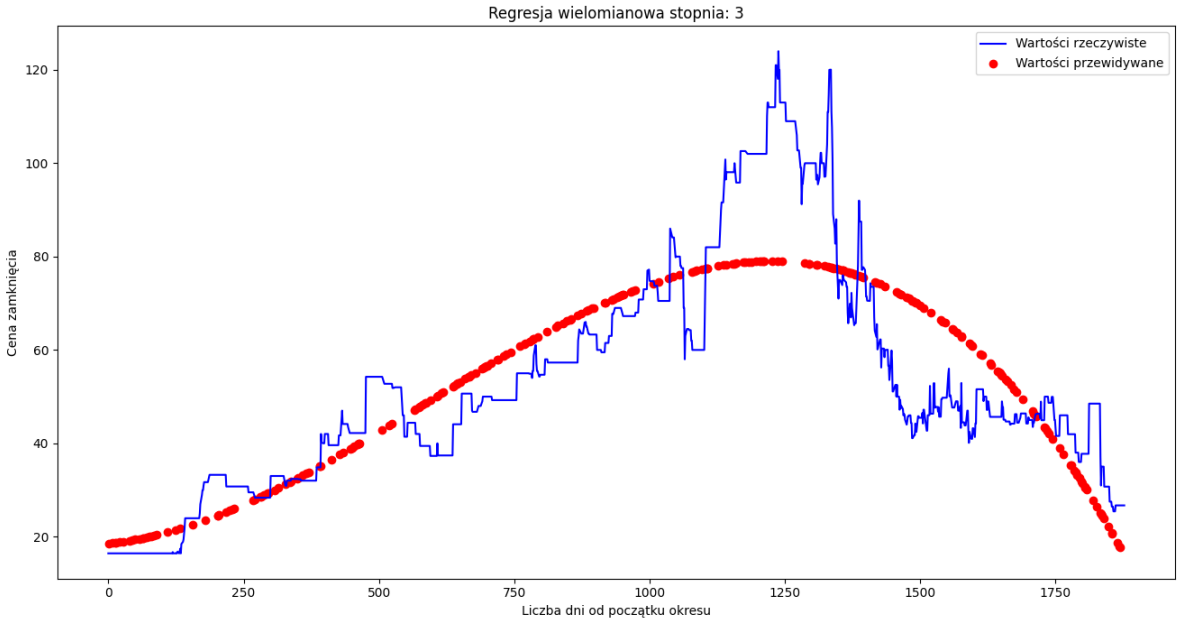
mean_squared_error: 480.3042437294925

rmse: 21.915844581706008

r2_score: 0.10484081775900844

Regresja wielomianowa:

- stopnia 3



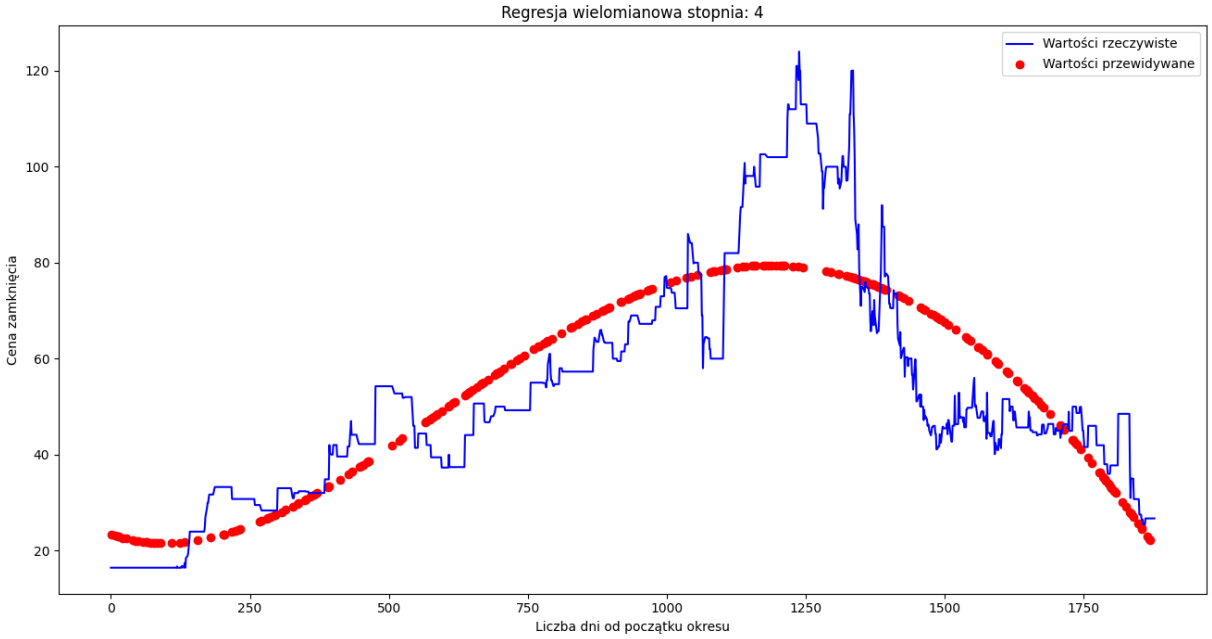
Parametry oceny jakości modelu

mean_absolute_error: 9.340061068498256

mean_squared_error: 150.14098166543928

rmse : 12.253202914562351

r2_score: 0.7201771999245785



- stopnia 4

  Parametry oceny jakości modelu:

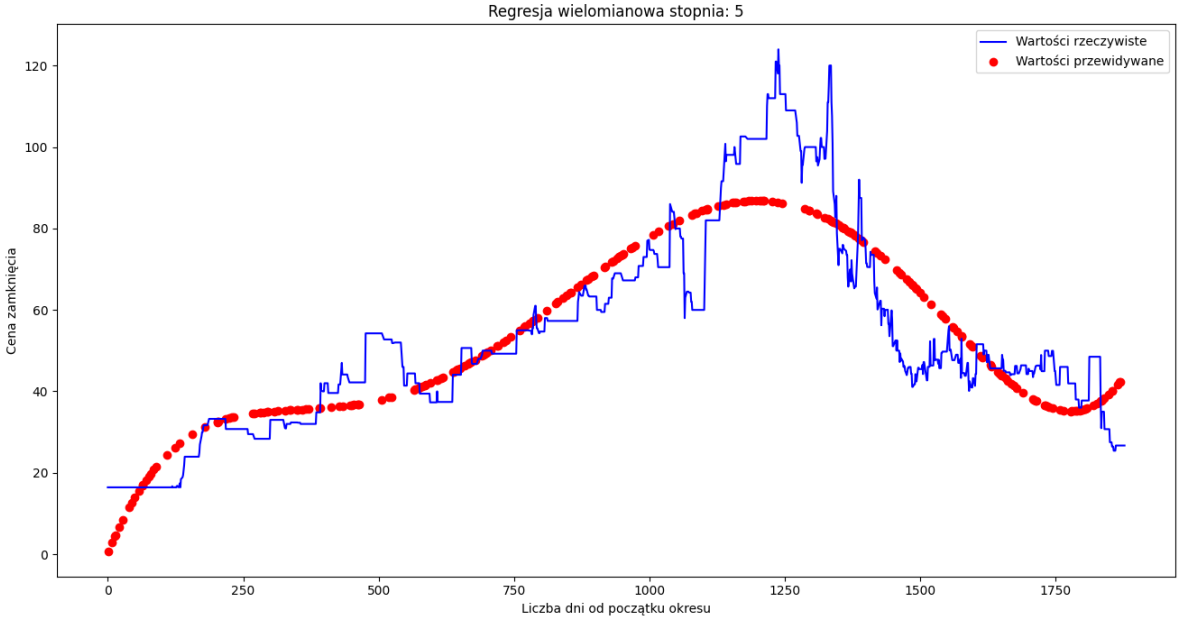
  mean_absolute_error: 9.406553809525555

  mean_squared_error: 144.55168114474768

  rmse : 12.022964740227248

  r2_score: 0.7305941673961782

- stopnia 5



Parametry oceny jakości modelu:

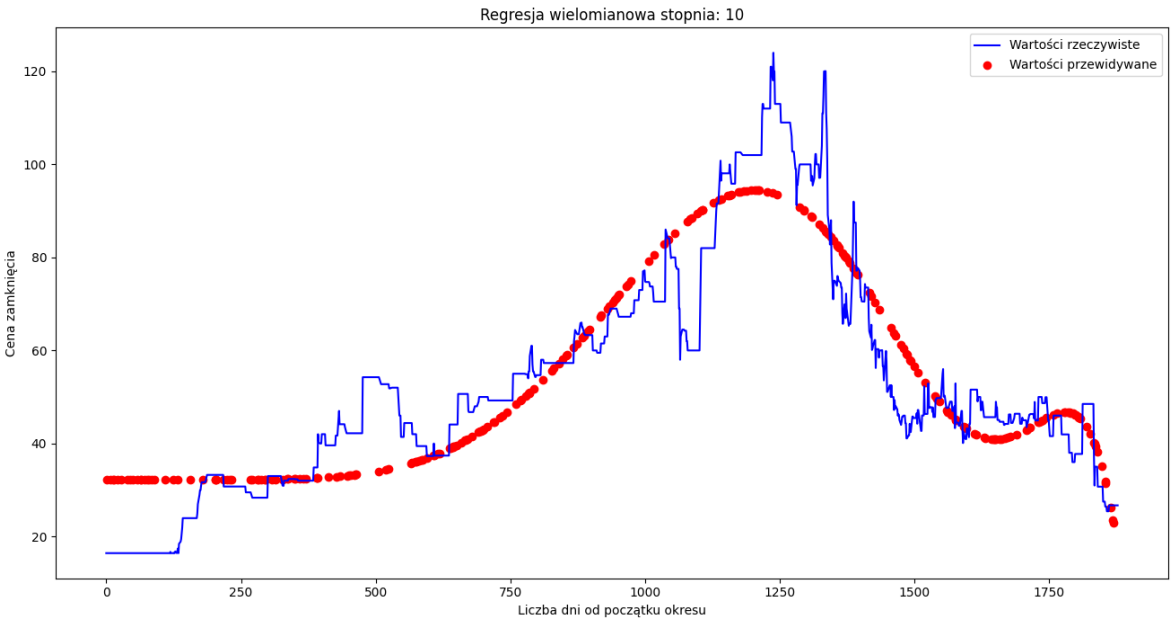
mean_absolute_error: 7.566669938168057

mean_squared_error: 102.76917135888017

rmse : 10.137513075645337

r2_score: 0.8084656369494603

- stopnia 10



Parametry oceny jakości modelu:

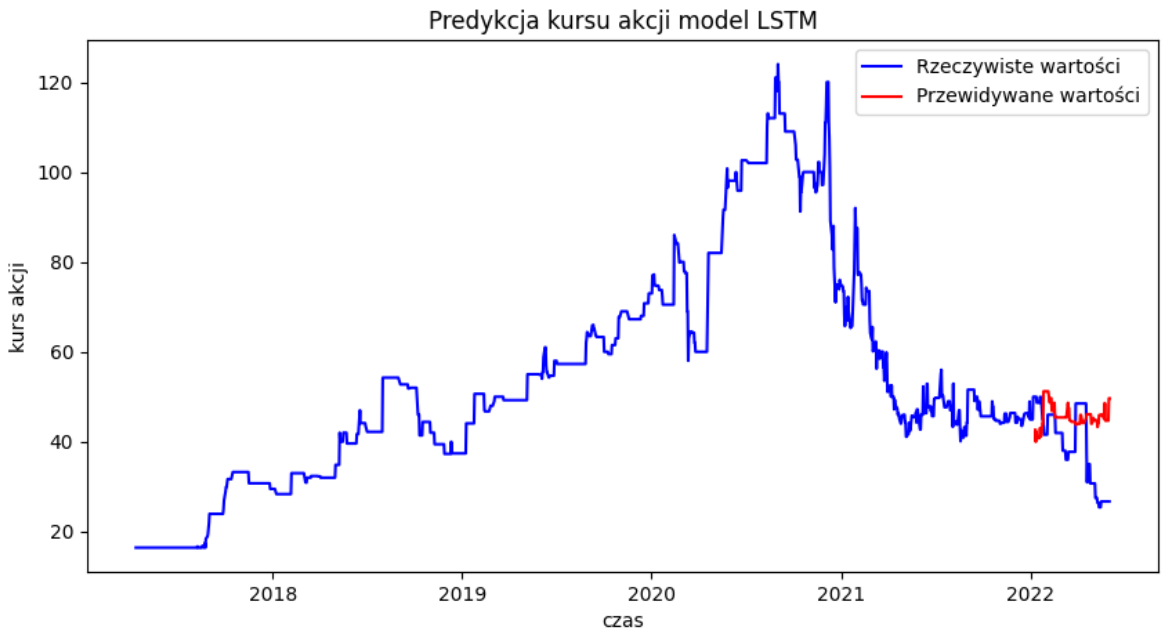
mean_absolute_error: 7.22899099552106

mean_squared_error: 89.18082338689705

rmse : 9.443559889517143

r2_score: 0.8337906983400424

Model LSTM:



Parametry oceny jakości modelu:

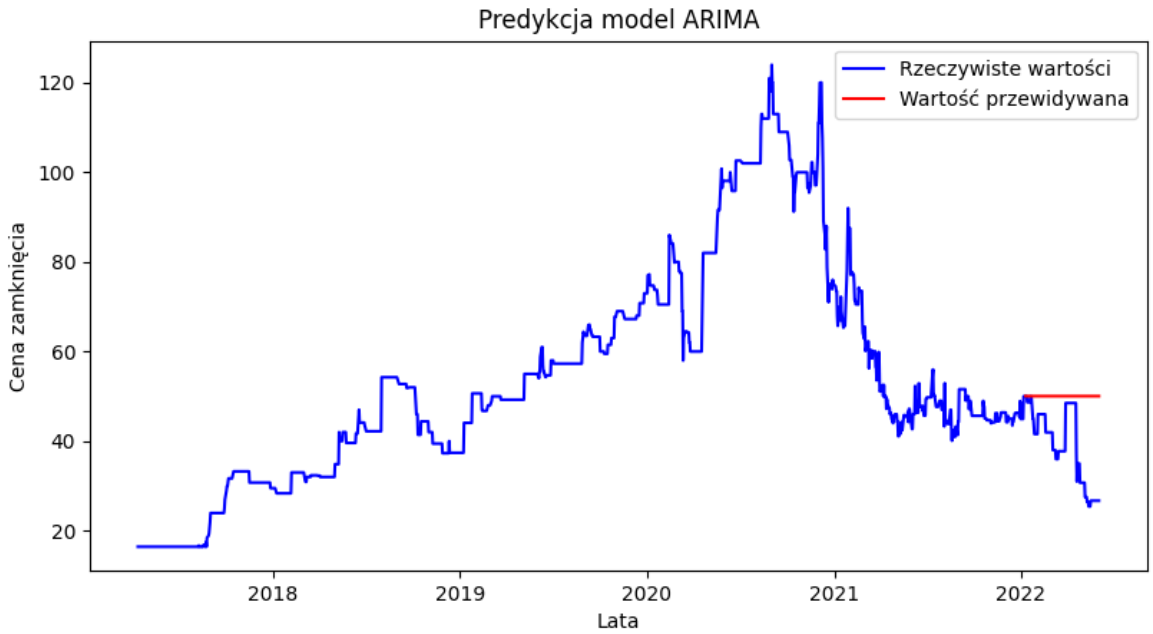
mean_absolute_error: 9.189228123293455

mean_squared_error: 119.92655349248797

rmse : 10.951098277912036

r2_score: -0.8061966915338092

Model ARIMA:



Parametry oceny jakości modelu:

mean_absolute_error: 10.928926576571103

mean_squared_error: 185.8356508769687

rmse: 13.632155034218497

r2_score: -1.798844192616925

9. Analiza uzyskanych wyników

Zastosowane modele przewidywały przyszłe wartości akcji CD Projekt z różnym stopniem dokładności, co można ocenić na podstawie parametrów takich jak średni błąd absolutny (mean_absolute_error), błąd kwadratowy (mean_squared_error), pierwiastek z błędu kwadratowego (rmse) i współczynnik determinacji (r2_score).

Model regresji liniowej zwrócił wysoki średni błąd absolutny (17.41) i błąd kwadratowy (480.30), a także duży pierwiastek z błędu kwadratowego (21.92). Współczynnik determinacji wynoszący 0.10 sugeruje, że tylko 10% zmienności ceny akcji można wyjaśnić tym modelem. Z tego powodu, model regresji liniowej nie jest najbardziej odpowiedni do przewidywania cen akcji CD Projekt.

Modele regresji wielomianowej okazały się znacznie skuteczniejsze. Zwróciły one niższe wartości średniego błędu absolutnego (od 9.34 dla modelu 3 stopnia do 7.23 dla modelu 10 stopnia) i błędu kwadratowego (od 150.14 dla modelu 3 stopnia do 89.18 dla modelu 10 stopnia). Dodatkowo, ich pierwiastki z błędu kwadratowego były znacznie niższe (od 12.25 dla modelu 3 stopnia do 9.44 dla modelu 10 stopnia), a współczynniki determinacji były wyższe (od 0.72 dla modelu 3 stopnia do 0.83 dla modelu 10 stopnia). Wyniki te sugerują, że modele regresji wielomianowej są bardziej skuteczne w przewidywaniu cen akcji CD Projekt.

Model LSTM zwrócił wyniki porównywalne do modelu regresji liniowej, z wysokim średnim błędem absolutnym (9.18), błędem kwadratowym (119.93), pierwiastkiem z błędu kwadratowego (10.95) i ujemnym współczynnikiem determinacji (-0.80). Z tego powodu, model ten nie jest odpowiedni do przewidywania cen akcji CD Projekt.

Podobnie, model ARIMA zwrócił wysokie wartości błędu (średni błąd absolutny 10.93, błąd kwadratowy 185.83, pierwiastek z błędu kwadratowego 13.63) i ujemny współczynnik determinacji (-1.80), co sugeruje, że również nie jest on odpowiedni do przewidywania cen akcji CD Projekt.

Na podstawie powyższej analizy, można stwierdzić, że modele regresji wielomianowej są najbardziej odpowiednie do przewidywania przyszłych wartości akcji CD Projekt.

10. Wnioski

Przeprowadzone badanie dostarczyło licznych istotnych wniosków dotyczących analizy i prognozowania kursu akcji CD Projekt. Najważniejsza obserwacja dotyczyła różnej skuteczności zastosowanych modeli predykcyjnych. Model regresji liniowej, mimo swojej prostoty, nie okazał się skuteczny do przewidywania przyszłych wartości kursu akcji. Niski współczynnik determinacji (R2) oraz wysokie wartości błędu

średniokwadratowego i błędu bezwzględnego wskazują, że ten model nie jest odpowiedni dla tak złożonego zestawu danych, jakim są notowania giełdowe.

Model regresji wielomianowej, z drugiej strony, osiągnął znacznie lepsze rezultaty, co było widoczne zarówno pod kątem wielkości błędów, jak i wartości R2. Wzrost stopnia wielomianu przekładał się na poprawę wyników, co sugeruje, że zwiększenie złożoności modelu pozwala na lepsze dopasowanie do analizowanych danych. W tym kontekście, model wielomianowy 10 stopnia okazał się najskuteczniejszy, prezentując najniższe wartości błędów i najwyższy współczynnik determinacji.

Modele LSTM i ARIMA, mimo ich zaawansowanej struktury i teoretycznej zdolności do skutecznego modelowania sekwencji czasowych, nie były w stanie efektywnie przewidzieć przyszłych wartości akcji. To może wynikać z faktu, że fluktuacje kursów giełdowych są efektem wielu czynników, które są trudne do uwzględnienia w modelach.

Ćwiczenie to było dla mnie cennym doświadczeniem, które pozwoliło mi lepiej zrozumieć, jak różne modele działają na praktycznych danych. Uświadomiło mi to, jak ważne jest dobranie odpowiedniego modelu do analizowanego zestawu danych i określonego celu. Niezależnie od wyboru metody, kluczowe jest zrozumienie jej ograniczeń i możliwości, a także umiejętność poprawnej interpretacji wyników.

Powyższa analiza ujawniła, że techniki analizy danych i uczenia maszynowego mogą okazać się przydatne do przewidywania przyszłych wartości akcji CD Projekt, ale nie wszystkie modele są równie skuteczne. Modele regresji wielomianowej, które potrafiły uwzględniać złożoność i nieliniowość danych, okazały się najbardziej efektywne. Z drugiej strony, proste modele regresji liniowej i bardziej zaawansowane modele takie jak LSTM i ARIMA nie spełniły oczekiwań.

W kontekście decyzji inwestycyjnych, te wnioski są wartościowym przypomnieniem, że choć modele predykcyjne mogą dostarczyć przydatnych

informacji, nie są one niezawodne i nie powinny stanowić jedynego kryterium. To jest szczególnie istotne w przypadku rynku giełdowego, gdzie liczne czynniki mogą wpływać na ceny akcji, a wiele z nich jest trudnych do przewidzenia czy modelowania. Wykorzystanie technik uczenia maszynowego może zatem zwiększyć nasze zrozumienie rynków finansowych i pomóc w identyfikacji trendów, ale nie zastąpi holistycznego podejścia do inwestowania, które uwzględnia również aspekty fundamentalne i makroekonomiczne.

Dalsze badania mogą skupić się na zastosowaniu innych modeli do przewidywania kursu akcji, a także na wykorzystaniu technik uczenia maszynowego do identyfikacji najważniejszych czynników wpływających na zmienność kursu.