Dokumentacja projekt pt. „Budowa modelu predykcyjnego dla wskaźnika inflacji”

Marcel Loster, Michał Pobuta

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

22.06.2023

1. Wstęp

Inflacja, będąca jednym z najważniejszych wskaźników ekonomicznych, wpływa na różne aspekty naszego życia codziennego, takie jak ceny towarów, stopy procentowe, a także decyzje inwestycyjne. Przewidywanie inflacji jest kluczowe dla instytucji finansowych, rządu, a także dla indywidualnych konsumentów i inwestorów. W tym projekcie naszym celem jest opracowanie skutecznego modelu do przewidywania inflacji, wykorzystując różnorodne cechy ekonomiczne.

2. Cel i zakres projektu

Celem tego projektu jest stworzenie modelu do przewidywania inflacji, bazując na danych historycznych. Zakres projektu obejmuje zebranie danych, przeprowadzenie analizy danych, wybór odpowiednich cech, które mają wpływ na inflację, oraz opracowanie modelu predykcyjnego.

3. Metodologia

3.1. Przygotowanie danych

Dane, które wykorzystaliśmy do tego projektu, obejmują 26 różnych zmiennych ekonomicznych od 2003 do 2022 roku. Dane te obejmują wskaźniki takie jak stopa bezrobocia, dochód i wydatki budżetu narodowego, indeksy giełdowe (WIG i WIG20), a także ceny różnych towarów i usług.

3.1.1. Analiza korelacji

Zrealizowaliśmy analizę korelacji, której wyniki przedstawiliśmy w postaci heatmapy.

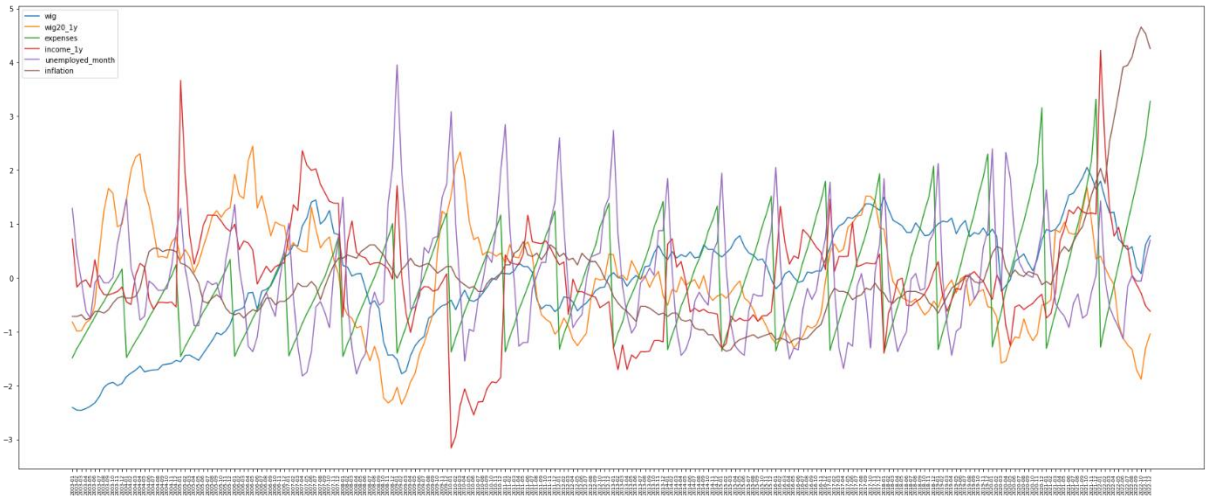
Obraz 1. – Heatmapa zależności pomiędzy danymi

W ten sposób zrozumieliśmy, jak różne zmienne wpływają na siebie nawzajem. Niektóre z naszych odkryć to silna korelacja między bezrobociem a liczbą bezrobotnych, co było spodziewane, biorąc pod uwagę niewielką zmienność populacji w omawianym okresie. Również dochód i wydatki w budżecie narodowym były silnie skorelowane.

Niektóre mniej oczywiste zależności to np. wysoka negatywna korelacja między indeksem WIG a bezrobociem czy wpływ większego budżetu narodowego na redukcję bezrobocia. Zauważono również, że zmienne dotyczące budżetu narodowego, bezrobocia i kursu dolara wydają się wpływać na inflację najmocniej.

3.1.2. Analiza sezonowości

Przeanalizowaliśmy sezonowość dochodów i wydatków budżetu narodowego. Wykresy pokazały, że te dwa wskaźniki mają wyraźny cykl roczny. Rząd zdaje się zbierać i wydawać najmniejszą ilość środków na początku roku, a następnie budżet rośnie w sposób liniowy do grudnia, tylko po to, aby znowu spaść na początku kolejnego roku.



Obraz 2. – Wykres zmiany wybranych danych na przestrzeni lat

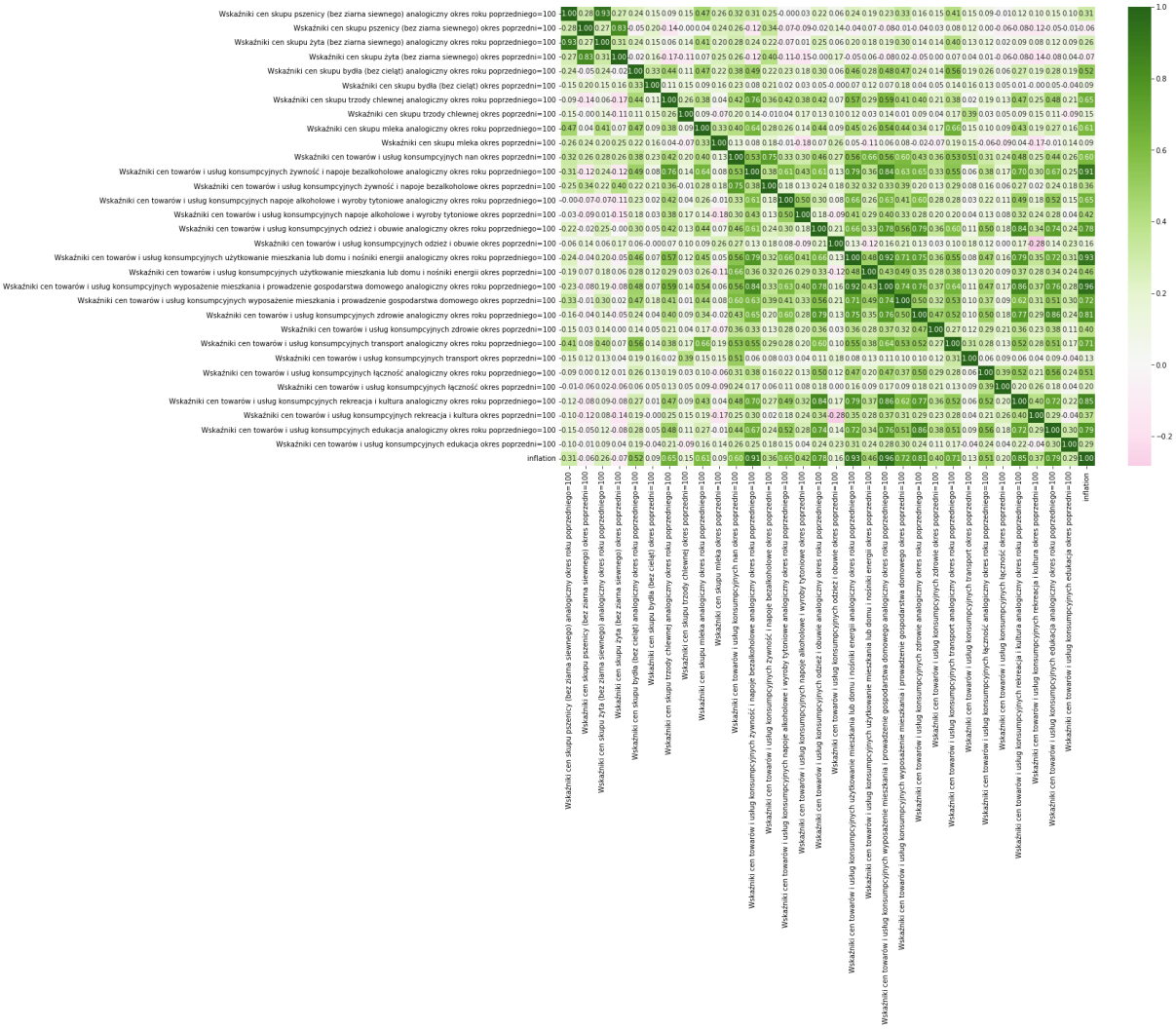
3.1.3. Analiza skupień

Dane zostały podzielone na klastry za pomocą metody "elbow". Wybrano siedem klastrów, które skupiają dane z konkretnych okresów. Na przykład, klaster numer 5 zawiera wszystkie rekordy z stycznia, co wskazuje, że styczeń ma wyraźnie odmienne cechy w porównaniu do innych miesięcy. Co ciekawe, klaster numer 6 skupia wyłącznie dane z 2022 roku, co sugeruje, że ten rok był wyjątkowy pod względem ekonomicznym.

3.1.4. Analiza cen

Dane cen zostały zebrane za pomocą podobnych metod dla każdej kolumny, co umożliwiło łatwą analizę bez potrzeby standaryzacji. Najbardziej zaskakujące było to, że ceny odzieży i obuwia pokazały bardzo niewielką korelację z innymi

cenami. Analiza skupień dla danych cenowych ujawniła okresy o najwyższym wzroście cen.



Obraz 3. – Macierz zależności pomiędzy wskaźnikami cen towarów i usług

3.1.4. Podsumowanie

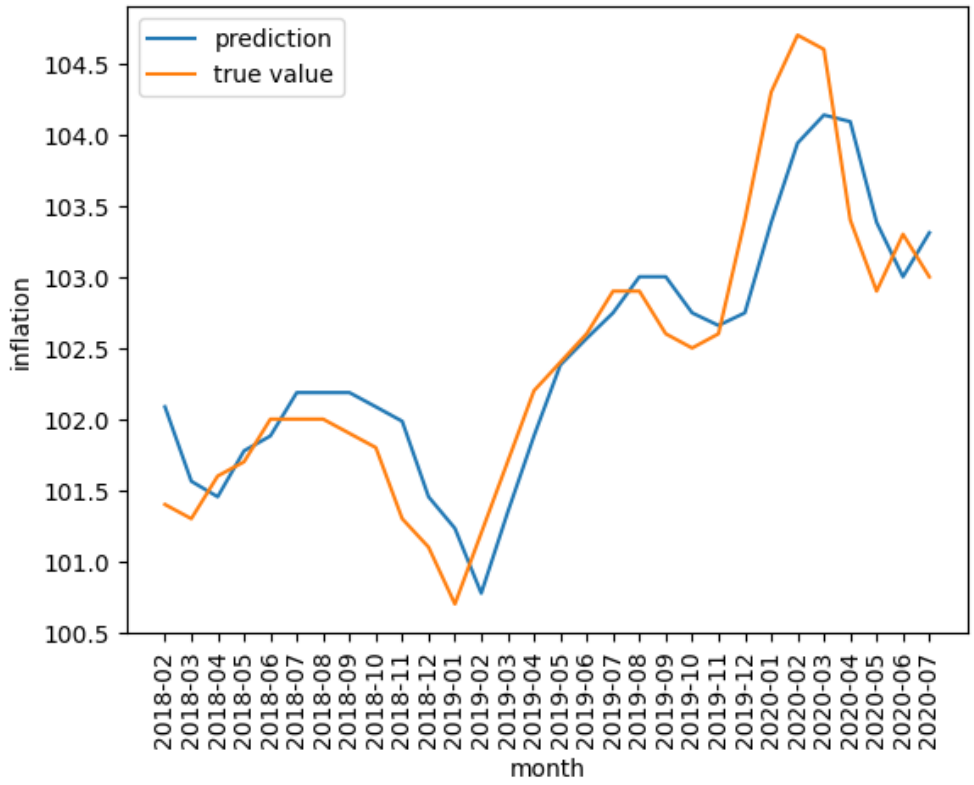
To wszystko stanowiło nasze wstępne przygotowanie danych. Na podstawie tych analiz zidentyfikowaliśmy potencjalnie ważne cechy, które mogą wpływać na inflację.

4. Budowa i ocena modeli

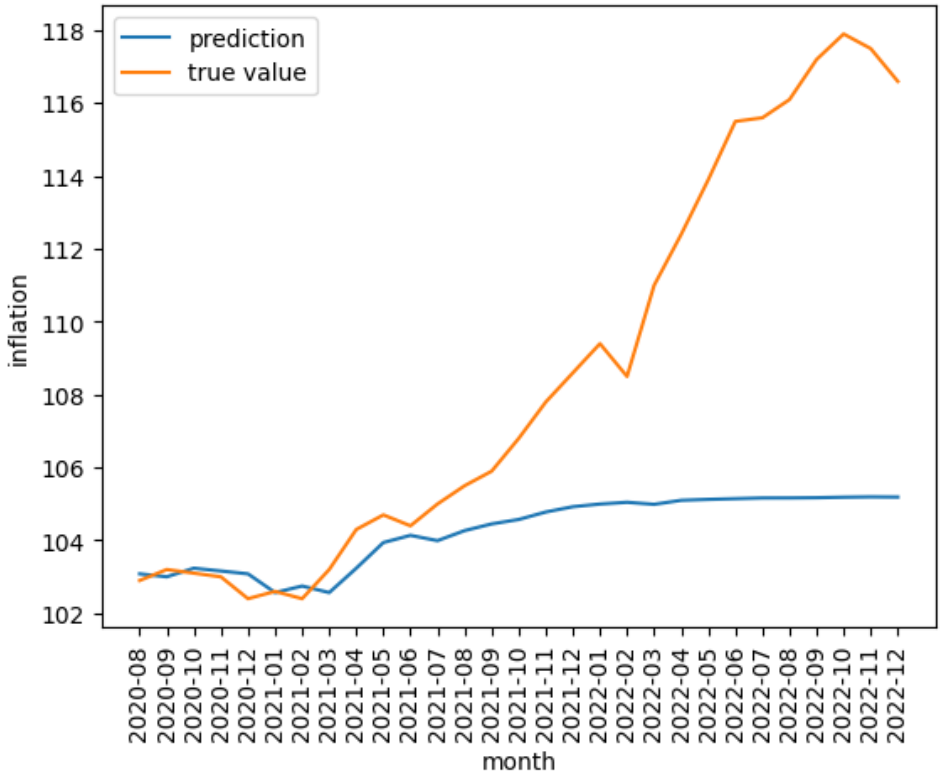
4.1. Modele bazujące na danych historycznych inflacji

4.1.1. LSTM - Przewidywanie na podstawie ostatniego miesiąca inflacji

W tym modelu skorzystaliśmy z sieci rekurencyjnej LSTM, która opiera się tylko na wartości inflacji z ostatniego miesiąca. Przewidywane wartości inflacji zostały porównane z rzeczywistymi wartościami, a następnie obliczono błąd średniokwadratowy (MSE) dla zbioru testowego.



Obraz 4. Wykres inflacji oraz przewidzianej inflacji w latach 2018-2020

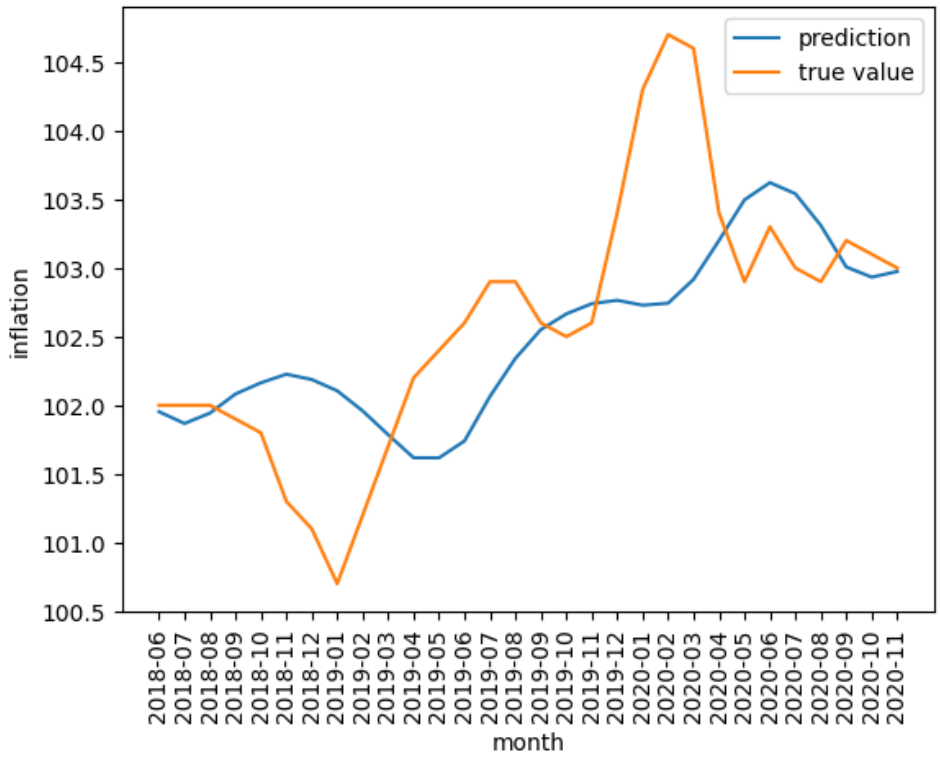


Obraz 5. Wykres inflacji oraz przewidzianej inflacji w latach 2020-2022

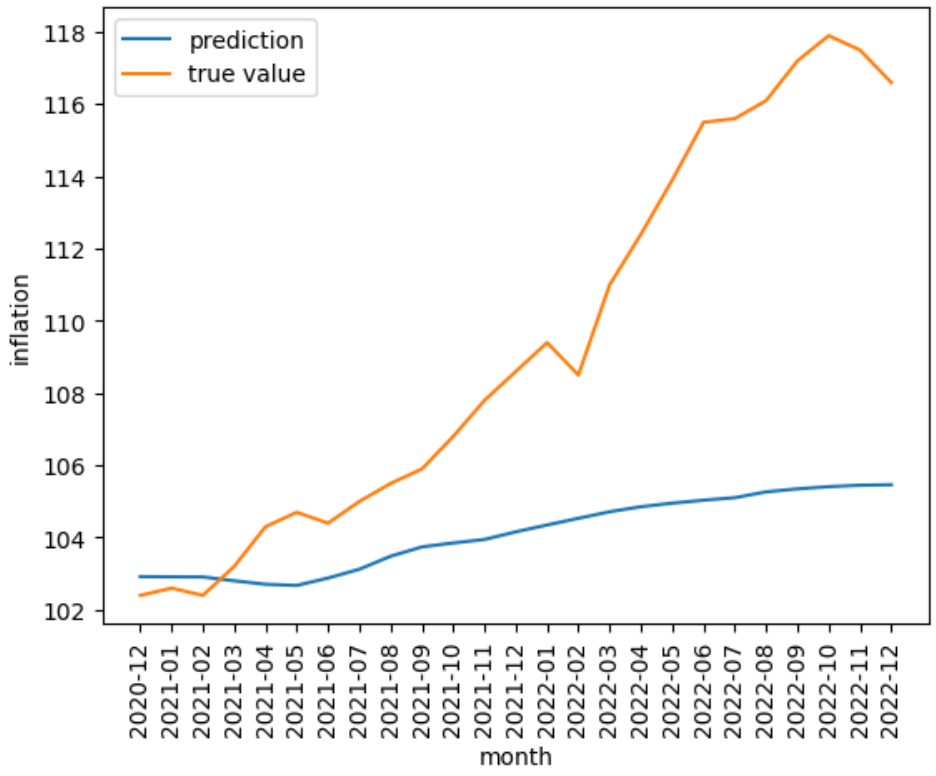
Pierwszy wykres predykcji oparty na ostatnim miesiącu inflacji wydaje się być dość dokładny (MSE ~0.01688). Jednak drugi wykres, który obejmuje okres po 2020 roku, pokazuje niemal płaską predykcję, sugerując, że ten model ma trudności w przewidywaniu inflacji w tym okresie(MSE ~3.79781).

4.1.2. LSTM - Przewidywanie na podstawie ostatnich 5 miesięcy

W tym modelu zastosowaliśmy również sieć rekurencyjną LSTM, która opiera się na wartościach inflacji z ostatnich pięciu miesięcy. Podobnie jak w poprzednim przypadku, przewidywane wartości inflacji zostały porównane z rzeczywistymi wartościami, a następnie obliczono błąd średniokwadratowy (MSE) dla zbioru testowego. Wykresy predykcji i rzeczywistych wartości inflacji zostały przedstawione.



Obraz 6. Wykres inflacji oraz przewidzianej inflacji w latach 2018-2020

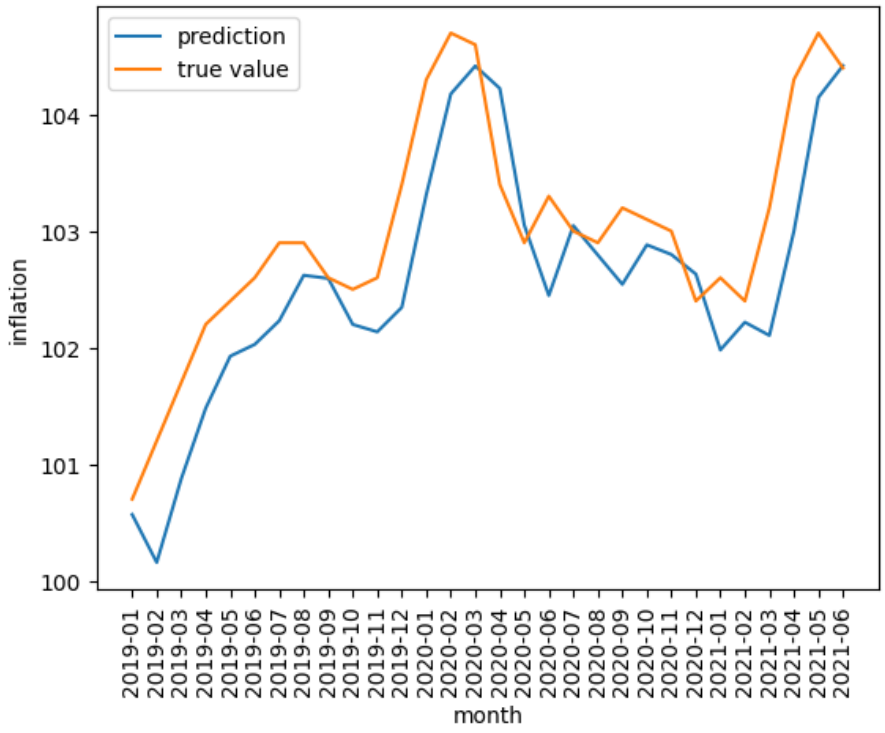


Obraz 7. Wykres inflacji oraz przewidzianej inflacji w latach 2020-2022

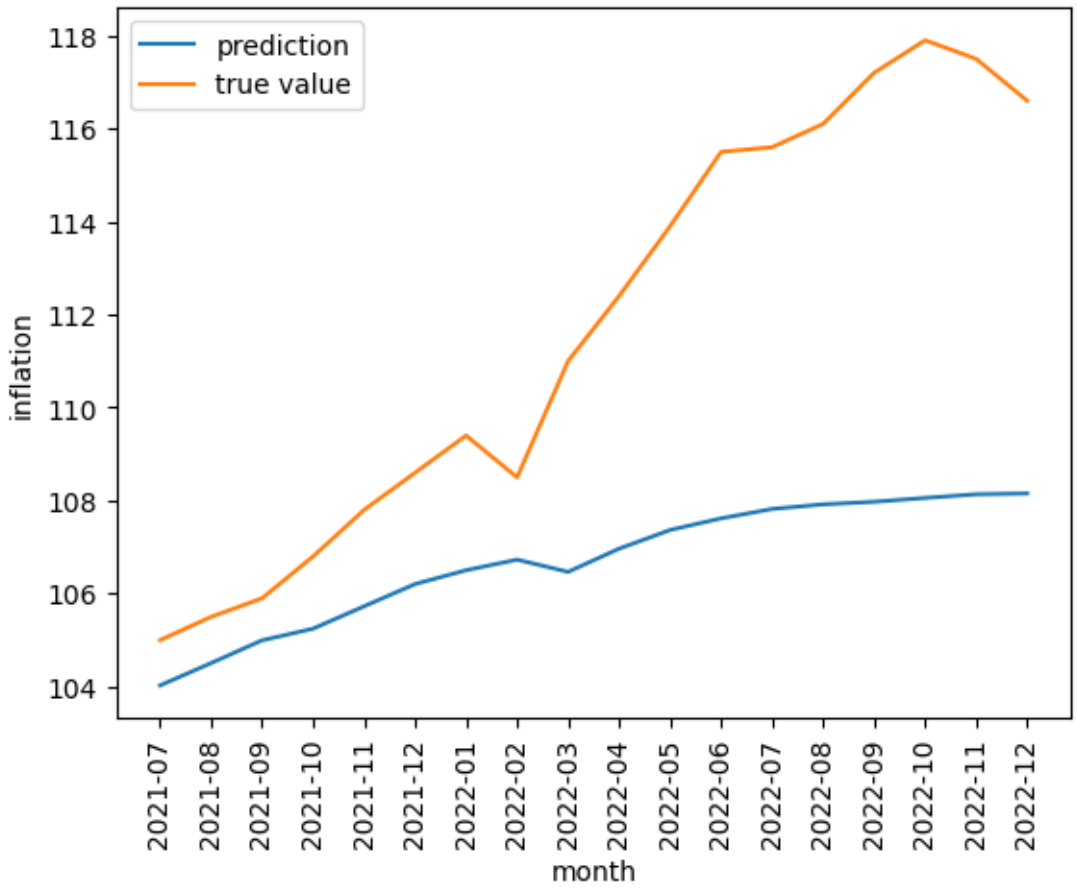
W przypadku tego modelu, pierwszy wykres predykcji na podstawie ostatnich pięciu miesięcy inflacji wydaje się być bardziej zróżnicowany (MSE ~0.05745). Jednak drugi wykres, który obejmuje okres po 2020 roku, jest podobny do wcześniejszego modelu, z niemal płaską predykcją inflacji (MSE ~4.51723).

4.1.3. LSTM - Przewidywanie na podstawie ostatnich 12 miesięcy

Tak samo jak w poprzednich przypadkach zastosowaliśmy sieć rekurencyjną LSTM, tylko tym razem w oparciu o dane z ostatnich 12 miesięcy. Następnie wyliczono błąd średnio kwadratowy (MSE)



Obraz 8. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021

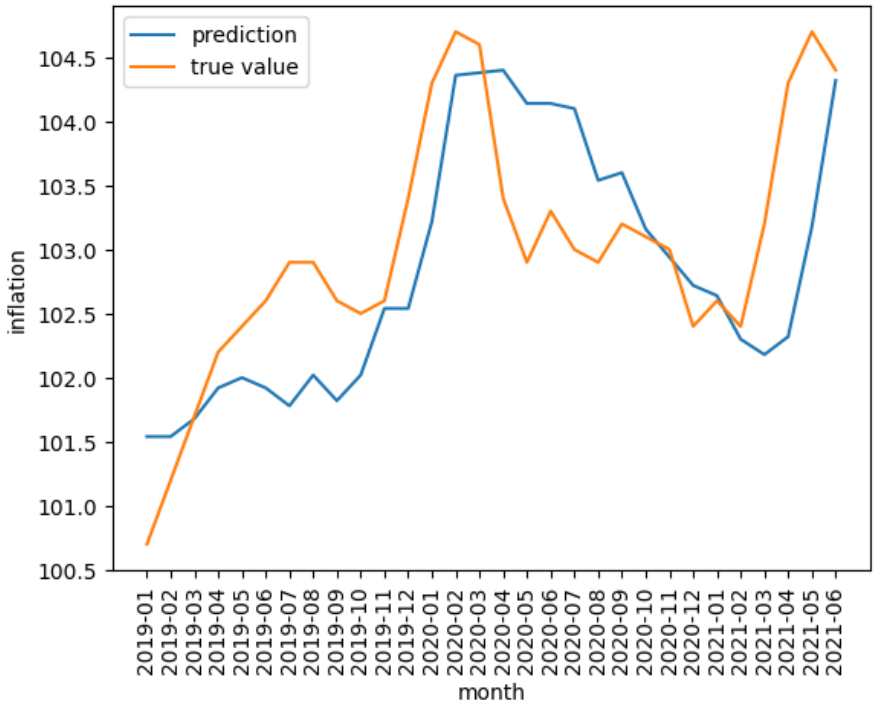


Obraz 9. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

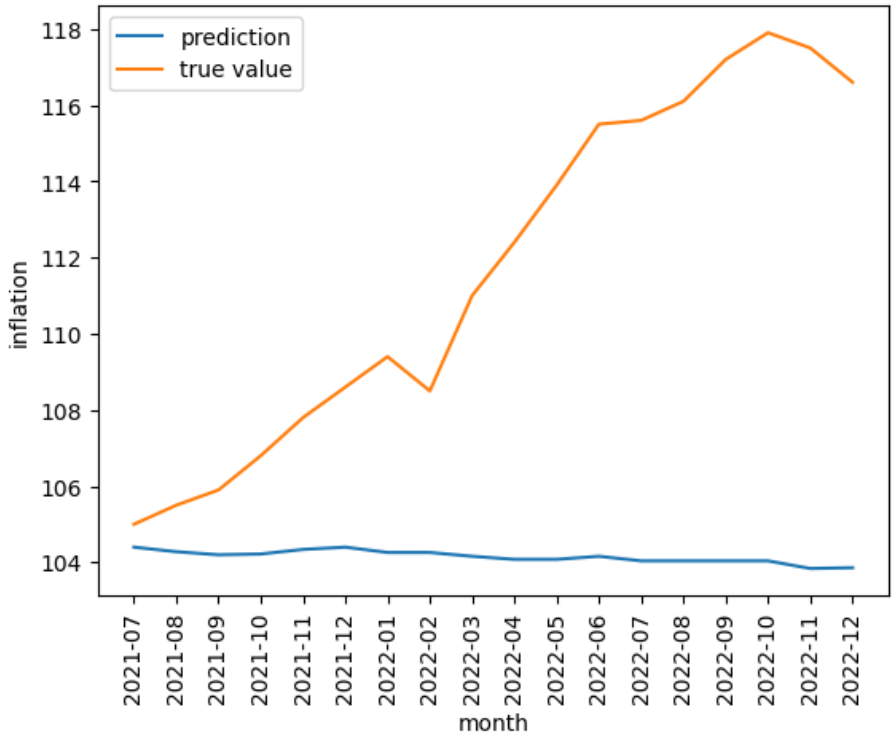
W przypadku tego modelu, pierwszy wykres predykcji na podstawie ostatnich pięciu 12 miesięcy wydaje się być dosyć dokładny (MSE ~0.03717). Jednak drugi wykres, który obejmuje okres po 2021 roku, pod koniec bardzo się „wypłaszcza", co świadczy o jego niedokładności (MSE ~3.43933).

4.1.4. KNN - Przewidywanie na podstawie ostatnich 12 miesięcy

Do kolejnej analizy użyliśmy modelu KNN. Algorytm K-nearest neighbors (KNN) jest metodą uczenia maszynowego, która opiera się na koncepcji "najbliższych sąsiadów". Wykorzystaliśmy dane historyczne inflacji z ostatnich 12 miesięcy. Na koniec policzyliśmy błąd (MSE).



Obraz 10. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



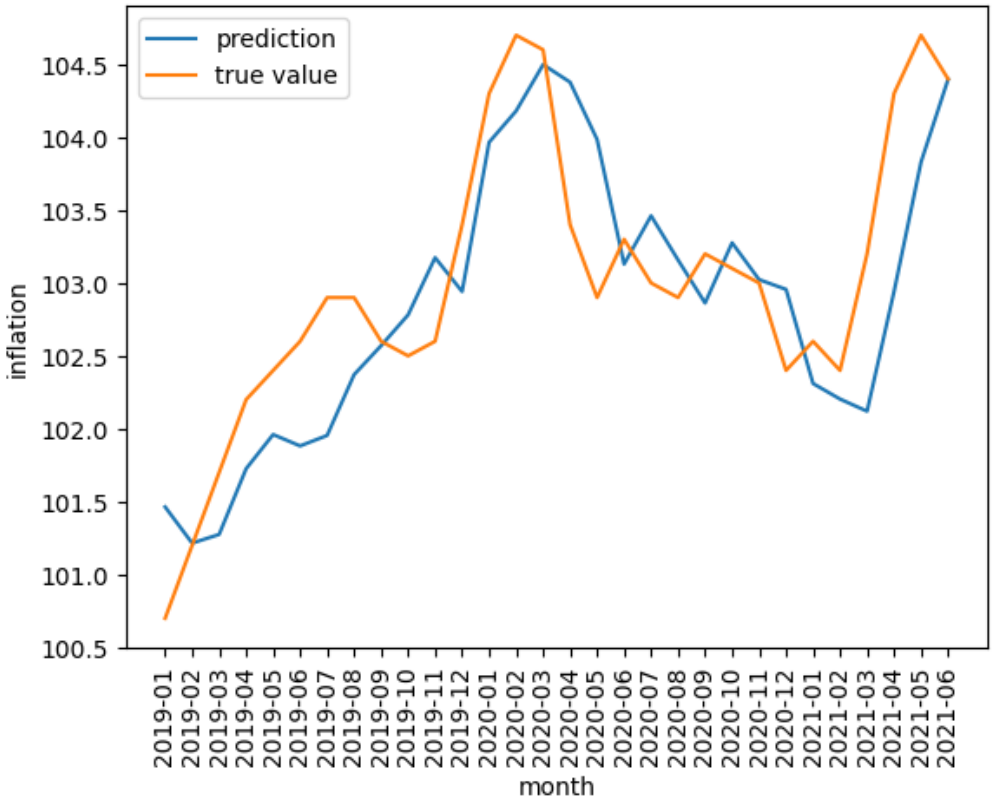
Obraz 11. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

Jak widać na uzyskanych wykresach, ponownie udało nam się uzyskać dosyć podobne wyniki dla wykresu pierwszego, natomiast w drugim standardowo dramat.

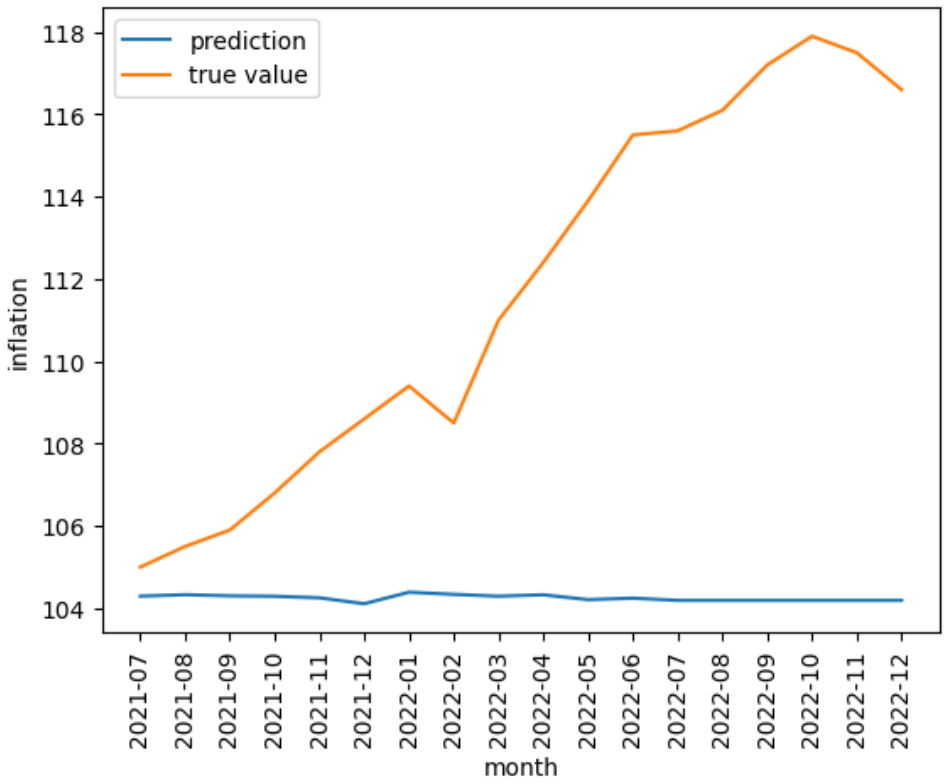
Dla okresu 2019-2021 uzyskano MSE ~0.05995, co jest dobrym wynikiem, a dla okresu 2021-2022 MSE ~7.51178, czyli znowu słabo.

4.1.5. CatBoost - Przewidywanie na podstawie ostatnich 12 miesięcy

CatBoost to zaawansowany algorytm gradient boosting, który został zaprojektowany do pracy z danymi kategorycznymi. Proces trenowania modelu CatBoost polega na dostosowaniu gradientu funkcji straty do zbioru danych treningowych. Algorytm automatycznie obsługuje zmienne kategoryczne i automatycznie buduje drzewa decyzyjne, optymalizując proces predykcji. W trakcie trenowania, model CatBoost tworzy wiele słabszych modeli, które następnie są połączone w silny model predykcyjny. W tym przypadku jak poprzednio podzieliliśmy dane na dwie części, a następnie narysowano wykresy oraz wyliczono MSE.



Obraz 12. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021

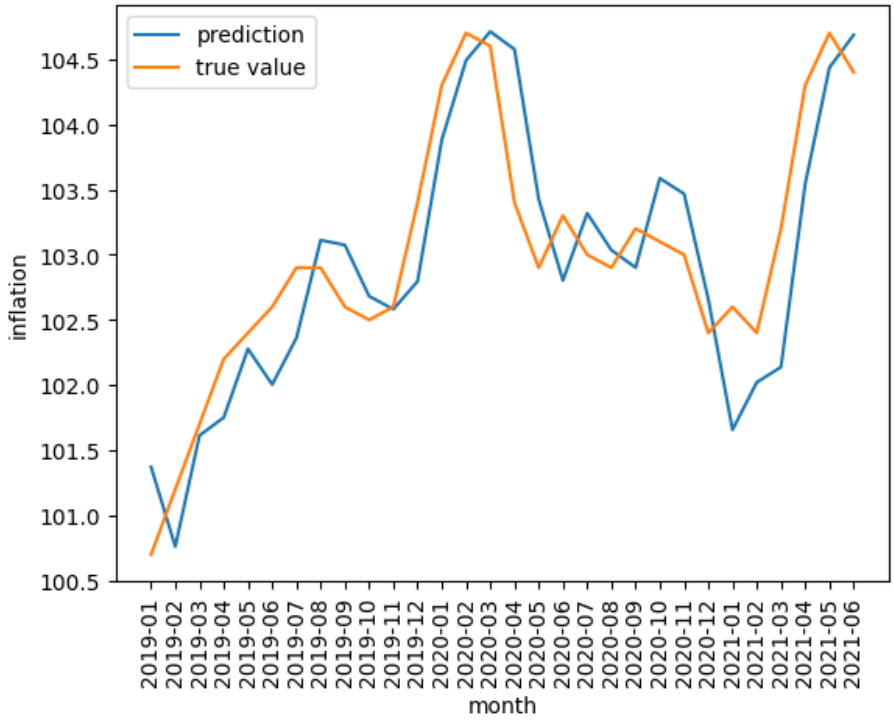


Obraz 13. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

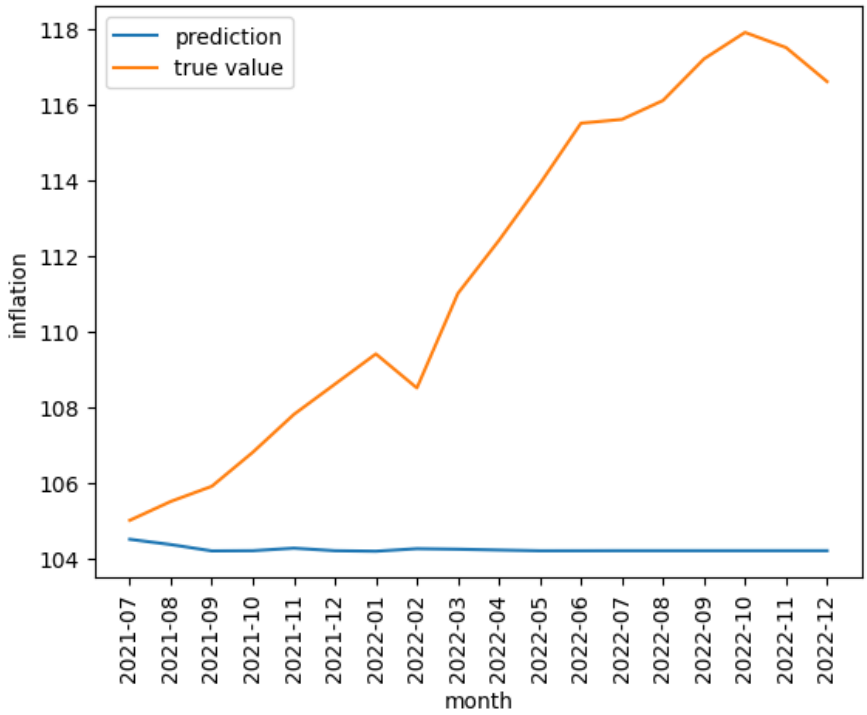
Analogicznie, pierwszy wykres bardzo podobny do rzeczywistego (MSE~0.03387), natomiast drugi praktycznie płaski (MSE ~7.28625).

4.1.6. XGBRegressor- Przewidywanie na podstawie ostatnich 12 miesięcy

Podobnie jak CatBoost, XGBRegressor wykorzystuje technikę gradnient boosting. Jest on oparty na drzewach decyzyjnych i charakteryzuje się wysoką skalowalnością, szybkością działania oraz zdolnością do radzenia sobie z złożonymi zależnościami między zmiennymi. Analiza dokładnie taka sama jak poprzednio – wykresy z podziałem na dwa okresy oraz obliczony dla nich błąd.



Obraz 14. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021

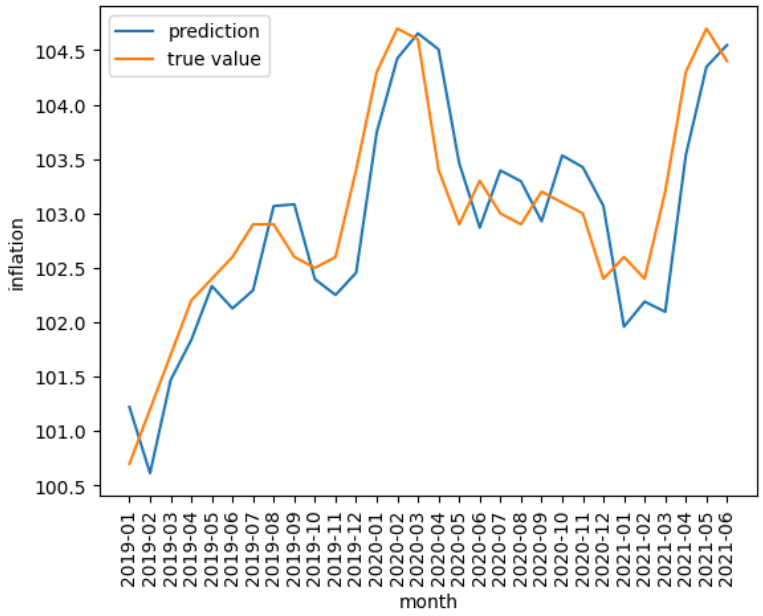


Obraz 15. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

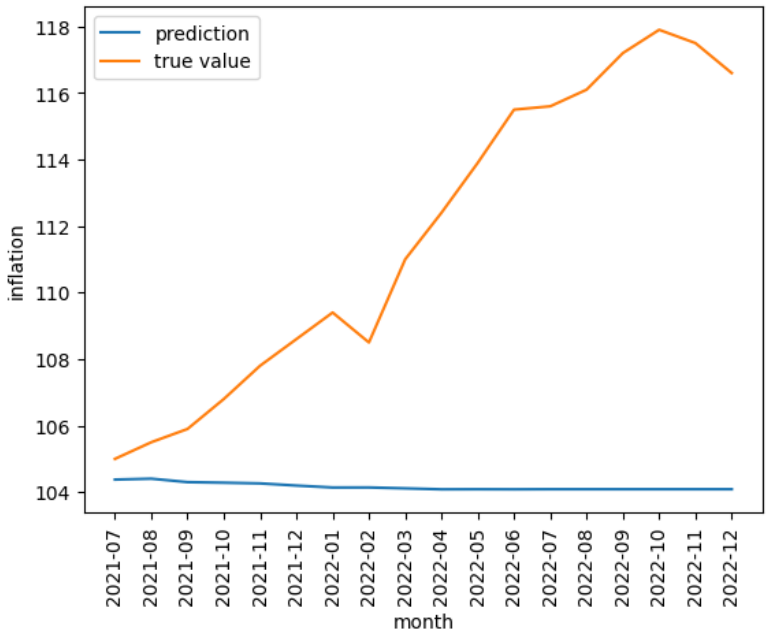
Analizując wykresy można ponownie zauważyć, że dla lat 2019-2021, wynik jest bardo blisko dokładnego (MSE ~0.025163), natomiast drugi ponownie, prawie płaski (MSE ~7.31401).

4.1.7. Random forest - Przewidywanie na podstawie ostatnich 12 miesięcy

Ostatni z analizowanych algorytmów – Random Forest. Random Forest to algorytm uczenia maszynowego, który działa na zasadzie zespołu słabszych modeli, czyli drzew decyzyjnych. Każde drzewo w lesie jest trenowane na różnych podzbiorach danych treningowych i zmiennych. Ostateczna predykcja jest wykonywana na podstawie głosowania lub uśredniania wyników z poszczególnych drzew. Analiza dokładnie taka sama jak poprzednio – wykresy z podziałem na dwa okresy oraz obliczony dla nich błąd.



Obraz 16. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



Obraz 17. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

Tak samo jak w poprzednich wypadkach, pierwszy dosyć dokładny(MSE ~0.02669), natomiast drugi znowu płaski(MSE ~7.450415).

4.1.8. Podsumowanie

Jak łatwo zauważyć, dla każdego z algorytmów dla lat 2019-2021, wyniki były całkiem dokładnie, co świadczy o tym, że modele zostały dobrze wytrenowane i w tych latach łatwo było by określić inflację w następnych latach. Natomiast żaden z algorytmów nie był w stanie dobrze przewidywać inflacji od roku 2021. Co może mieć sens, ponieważ w tych latach, duży wpływ na inflacje miały inne czynniki niż tylko inflacja z poprzedniego roku. Od roku 2021 na podstawie samej inflacji z zeszłego roku nie jesteśmy w stanie przewidywać przyszłej inflacji. Plusem naszego modelu jest to, że nawet od 2021, jest w stanie lepiej przewidzieć inflacje niż modele NBP.

4.2. Modele wielozmiennowe

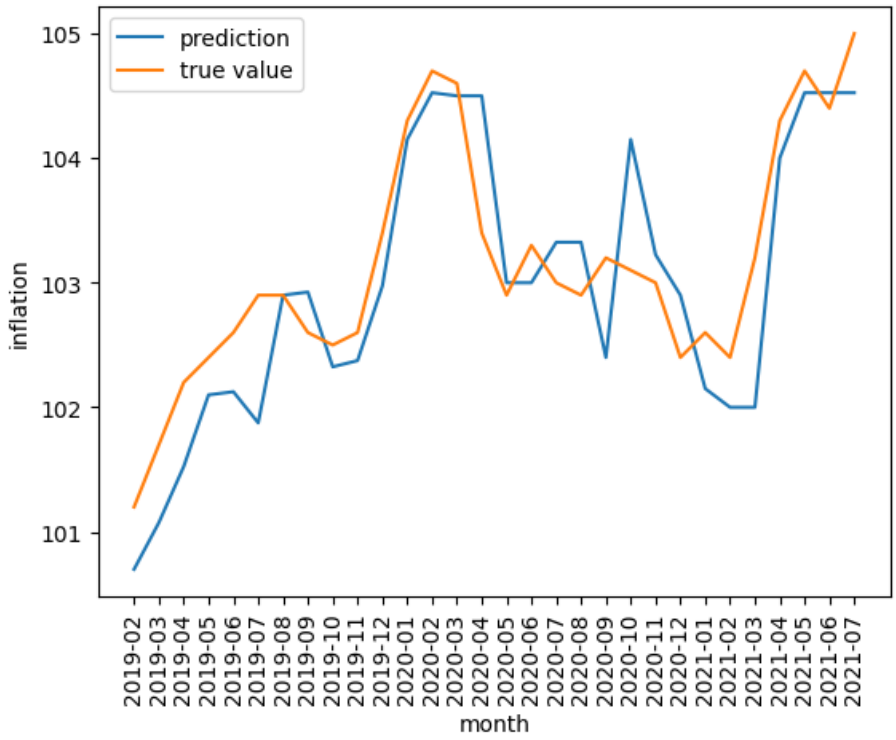
4.2.1 Wprowadzenie

W tej części, będziemy porównywać kilka wybranych przez nas algorytmów, ale tym razem dla dwóch najbardziej znaczących zmiennych wpływających na inflacje: kursu dolara [usd] oraz wskaźnika inflacji.

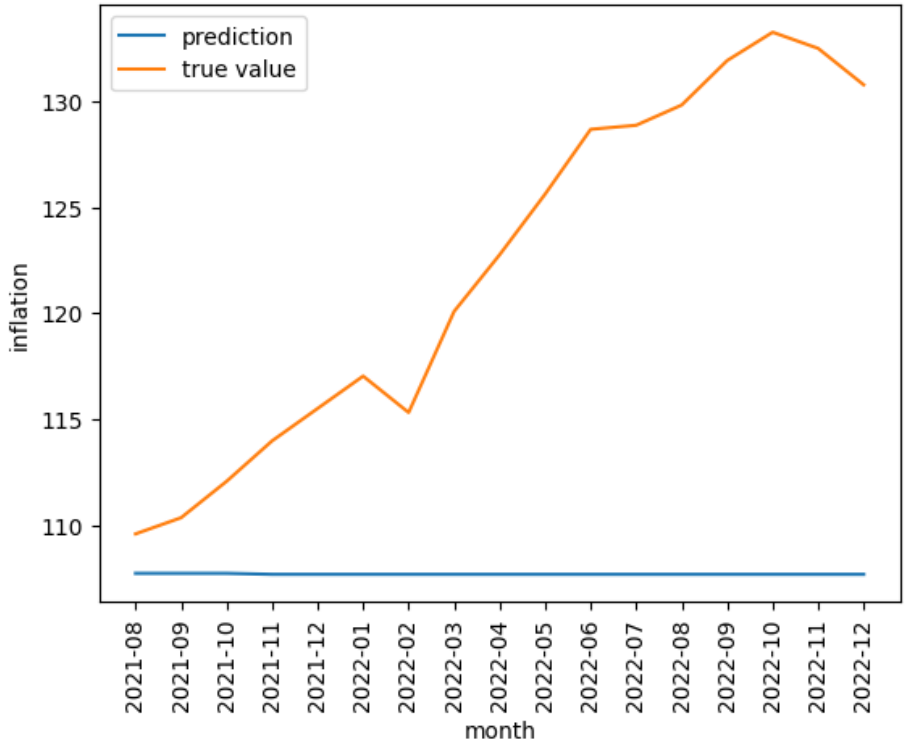
Te dwie zmienne wybraliśmy korzystając z algorytmu SelectKBest z biblioteki sklearn. Testowaliśmy kombinacje większych ilości zmiennych, ale wyniki nie różniły się znacząco więc zdecydowaliśmy się użyć jak najmniejszej, zadowalającej liczby wymiarów.

4.2.2 KNN

Zbudowaliśmy model KNN (K-Nearest Neighbors) do predykcji inflacji na podstawie dwóch zmiennych: USD i inflacji. Użyliśmy klasy KNeighborsRegressor z parametrem n_neighbors=4.



Obraz 18. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



Obraz 19. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

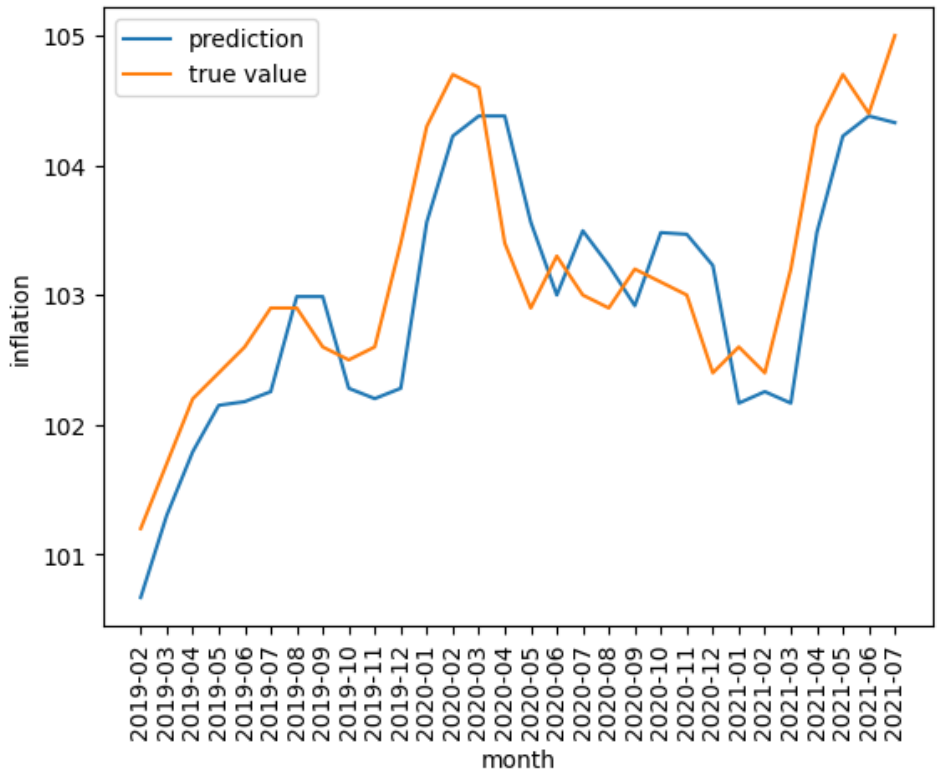
Ocena modelu:

- MSE dla danych testowych: 0.1005
- MSE dla drugiego zestawu testowego: 26.5948

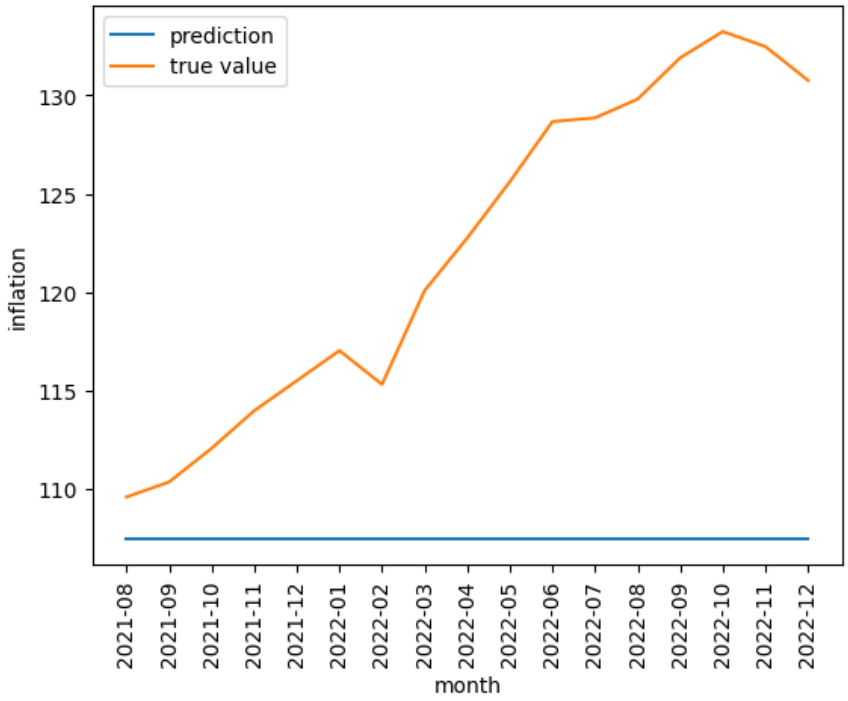
Model KNN osiągnął stosunkowo niskie wartości MSE dla danych testowych, sugerując, że potrafił dokonywać dokładnych predykcji inflacji na podstawie zmiennych USD i inflacji. Jednak wartość MSE dla drugiego zestawu testowego była znacznie wyższa, co sugeruje pewne trudności z generalizacją modelu na nowe dane spoza zakresu trenowania.

4.2.3 Model Random Forest

Zbudowaliśmy model Random Forest dla predykcji inflacji na podstawie dwóch zmiennych: USD i inflacji. Użyliśmy klasy RandomForestRegressor z parametrami min_samples_split=5 i max_depth=3.



Obraz 20. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



Obraz 21. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

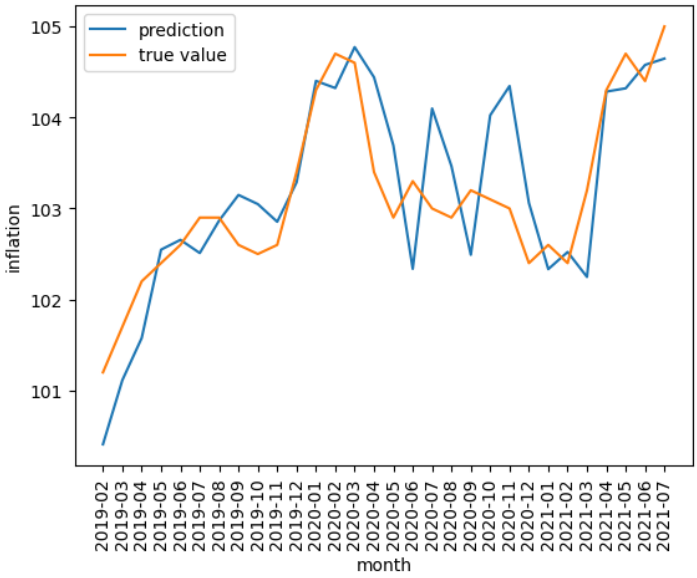
Ocena modelu:

- MSE dla danych treningowych: 0.0377
- MSE dla danych testowych: 0.1074
- MSE dla drugiego zestawu testowego: 27.2398

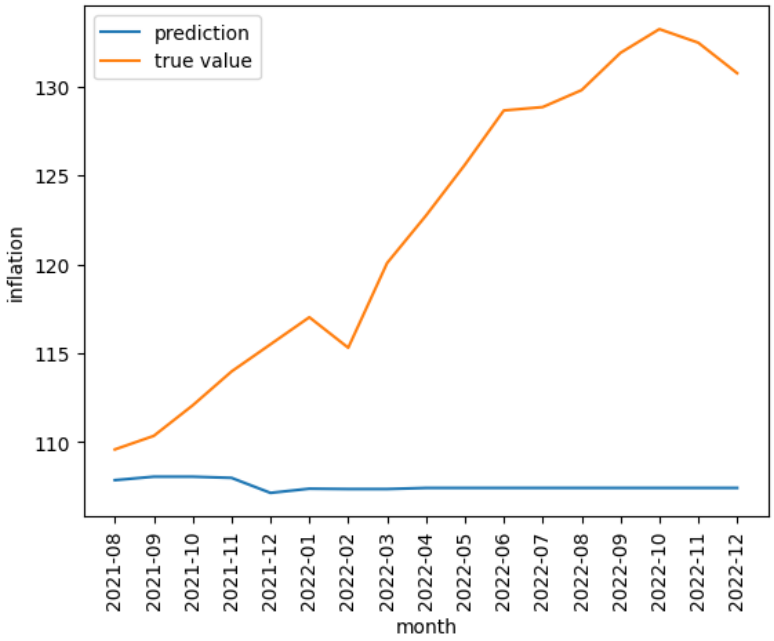
Model Random Forest osiągnął niskie wartości MSE dla danych treningowych i testowych, sugerując, że potrafił dokonywać dokładnych predykcji inflacji na podstawie zmiennych USD i inflacji. Jednak wartość MSE dla drugiego zestawu testowego była znacznie wyższa, co wskazuje na pewne trudności w generalizacji modelu na nowe dane spoza zakresu trenowania.

4.2.4 Model XGBRegressor

Zbudowaliśmy model XGBRegressor dla predykcji inflacji na podstawie dwóch zmiennych: USD i inflacji. Użyliśmy klasy XGBRegressor z parametrami n_estimators=500 i max_depth=5.



Obraz 22. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



Obraz 23. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

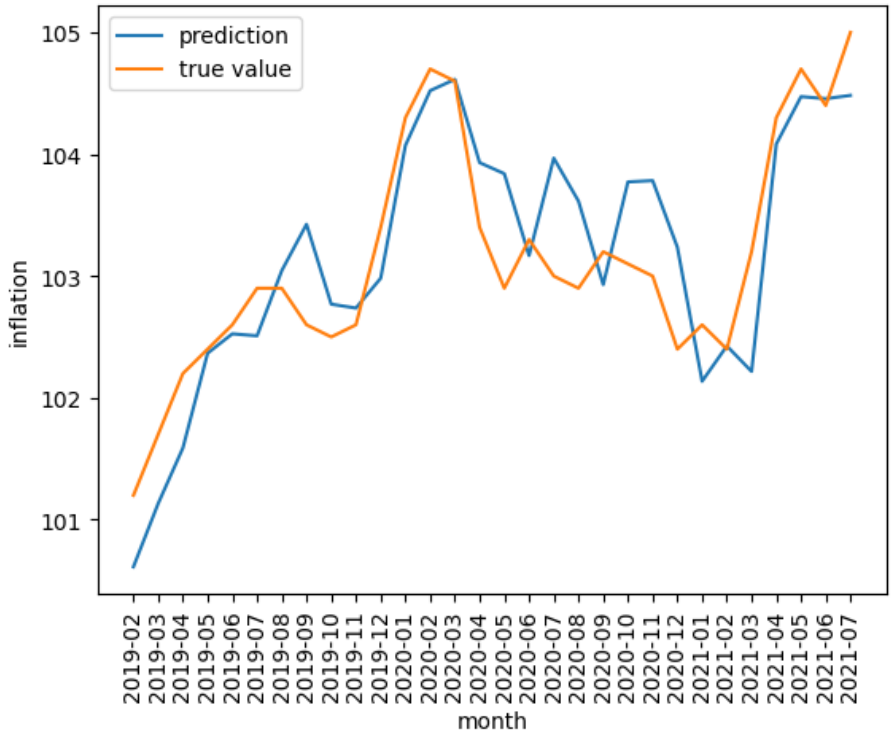
Ocena modelu:

- MSE dla danych treningowych: 5.8427e-06
- MSE dla danych testowych: 0.1324
- MSE dla drugiego zestawu testowego: 27.2695

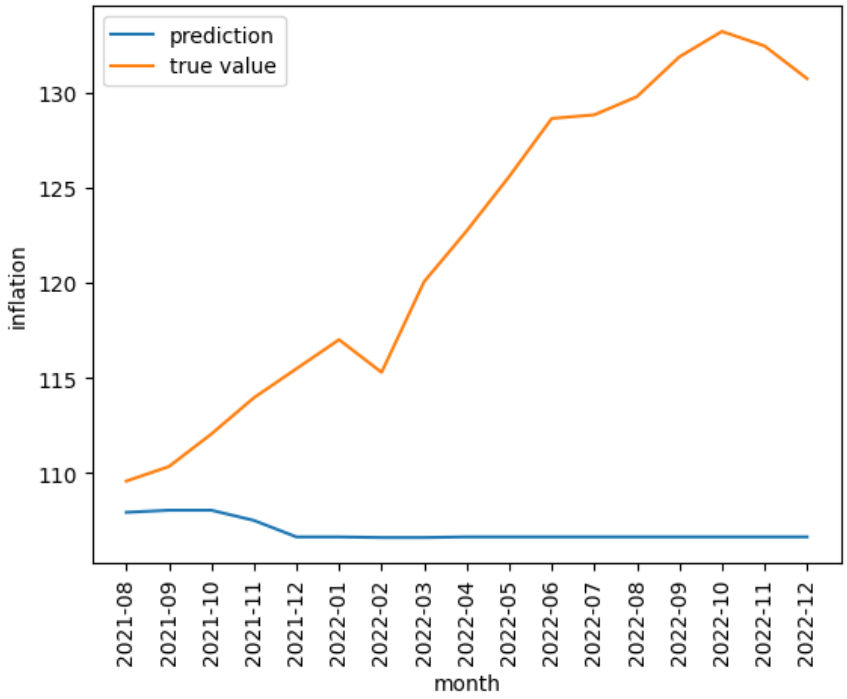
Model XGBRegressor osiągnął bardzo niską wartość MSE dla danych treningowych, sugerując bardzo dobrą dopasowanie modelu do danych treningowych. Jednak wartości MSE dla danych testowych i drugiego zestawu testowego są wyższe, co wskazuje na pewne trudności w generalizacji modelu na nowe dane spoza zakresu trenowania.

4.2.5 Model CatBoostRegressor

Zbudowaliśmy model CatBoostRegressor dla predykcji inflacji na podstawie dwóch zmiennych: USD i inflacji. Użyliśmy klasy CatBoostRegressor z parametrami depth=6 i n_estimators=1000.



Obraz 24. Wykres inflacji oraz przewidzianej inflacji w latach 2019-2021



Obraz 25. Wykres inflacji oraz przewidzianej inflacji w latach 2021-2022

Ocena modelu:

- MSE dla danych treningowych: 0.0067
- MSE dla danych testowych: 0.0956
- MSE dla drugiego zestawu testowego: 29.3383

Model CatBoostRegressor osiągnął niskie wartości MSE dla danych treningowych i testowych, sugerując dobrą zdolność modelu do predykcji inflacji na podstawie zmiennych USD i inflacji. Jednak wartość MSE dla drugiego zestawu testowego jest wyższa, co wskazuje na pewne trudności w generalizacji modelu na nowe dane spoza zakresu trenowania.

4.2.6 Podsumowanie

W tej części projektu testowaliśmy różne modele do przewidywania inflacji na podstawie dwóch zmiennych: USD i inflacji. Modele LSTM i KNN osiągnęły najlepsze wyniki, przewidując inflację z wysoką dokładnością. Modele oparte na Random Forest, XGBRegressor i CatBoostRegressor również były w stanie przewidywać inflację, choć z nieco niższą precyzją.

4.3. Podsumowanie

W ramach tego projektu przeprowadziliśmy analizę i budowę modeli do przewidywania inflacji. Testowaliśmy różne modele, takie jak LSTM, KNN, Random Forest, XGBRegressor i CatBoostRegressor, oraz ocenialiśmy ich skuteczność na podstawie wyników predykcji.

Nasze wyniki wskazują, że modele oparte na historii inflacji (LSTM, KNN) osiągnęły najwyższą dokładność predykcji inflacji. Modele te uwzględniały ostatnie dane dotyczące inflacji i były w stanie przewidzieć jej wartość z dobrą precyzją. Natomiast modele oparte na dwóch zmiennych, USD i inflacji, (Random Forest, XGBRegressor, CatBoostRegressor) osiągnęły nieco niższą dokładność predykcji.

Podsumowując, modele oparte na historii inflacji (LSTM, KNN) okazały się najbardziej efektywne w przewidywaniu inflacji na podstawie dostępnych danych. Te modele mogą być przydatne w prognozowaniu inflacji i wspomaganiu podejmowania decyzji związanych z polityką gospodarczą.