Metody Systemowe i Decyzyjne

Raport zaliczeniowy

Krzysztof Głowacz

czerwiec 2023

Spis treści

1 Wstęp

Tematem zadania zaliczeniowego była prezentacja umiejętności nabytych w czasie kursu podczas rozwiązywania realnego problemu. Celem niniejszej pracy było zbadanie, czy brak obecności Polski w strefie Euro może mieć wpływ na stabilność polskich spółek notowanych na giełdzie. Do tego celu zebrałem dane dotyczące wskaźników makroekonomicznych i cen indeksów giełdowych w Polsce i w siedmiu krajach strefy Euro o podobnym poziomie gospodarczym do Polski, następnie na podstawie tych danych wytrenowałem modele regresji, które kolejno miały posłużyć do predykowania zmian cen indeksów giełdowych na podstawie ustalonych wartości parametrów makroekonomicznych.

2 Dane

2.1 Dobór i pozyskanie danych

Mając jasno zdefiniowany problem przystąpiłem do wyboru krajów strefy Euro biorących udział w badaniu. Na podstawie wartości PKB per capita[1] zdecydowałem się wybrać: Grecję, Hiszpanię, Litwę, Łotwę, Portugalię, Słowację i Słowenię. Do analizy zdecydowałem się wybrać dane z lat 2018-2022. Początkowo rozpocząłem poszukiwanie danych na stronach banków centralnych i urzędów statystycznych poszczególnych państw. Mogłoby to jednak prowadzić do niespójności danych, gdyż te same wskaźniki bywają mierzone w nieco inny sposób. Ostatecznie udało mi się znaleźć wszystkie potrzebne odczyty zmian PKB[2], poziomu inflacji[3] oraz stóp bezrobocia[4] na stronie internetowej Organizacji Współpracy Gospodarczej i Rozwoju. Dane dotyczące poziomów stóp procentowych pochodzą ze strony Europejskiego Banku Centralnego[5], natomiast dane o cenach indeksów giełdowych poszczególnych państw pochodzą z trzech źródeł: ze strony internetowej magazynu The Wall Street Journal[6](dla Grecji i Polski), z portalu Investing.com[7](dla Hiszpanii, Litwy, Łotwy, Portugalii i Słowacji) oraz ze strony słoweńskiej giełdy[8]. Wszystkie wymienione źródła oferowały API do pobierania danych w formacie *.csv*.

2.2 Pre-processing danych

Dzięki pobraniu odpowiadających sobie danych z tych samych źródeł (wyjątek stanowią notowania indeksów giełdowych, ale dane dotyczące cen giełdowych są podawane w bardzo podobnym formacie we wszystkich źródłach i nie występują między nimi różnice w sposobach ich liczenia) możliwe było jednolite przygotowane danych dla każdego kraju. Moim celem było takie przekształcenie danych, aby każdy wiersz zawierał wartość miesięcznej zmiany danego wskaźnika. Osiągnąłem to w następujący sposób:

- PKB - dane dotyczące PKB zawierały wartości zmiany PKB w danym kwartale względem kwartału poprzedniego. Założyłem w tym miejscu, że procentowa zmiana PKB względem poprzedniego miesiąca jest taka sama dla każdego miesiąca w danym kwartale. Stąd do obliczenia zmiany PKB w każdym miesiącu danego kwartału skorzystałem ze wzoru:

$$x = 100 \cdot \sqrt[3]{\frac{100+p}{100}} - 100,$$

  gdzie: x - liczona wartość miesięcznej zmiany PKB, p - podana kwartalna wartość zmiany PKB. Przykładowo, w Polsce w trzecim kwartale 2020 roku odnotowano wzrost PKB na poziomie $6,75\%$, zatem obliczone wartości zmiany PKB w Polsce w lipcu, sierpniu i we wrześniu 2020 roku to $2,20\%$.

- Inflacja - pobrane poziomy inflacji były policzone w taki sposób, że dla każdego kraju punktem odniesienia były odpowiednie odczyty z roku 2015 i to one stanowiły bazowe 100%. Wartości te przeliczyłem na potrzeby zadania obliczając różnicę punktów procentowych między aktualnie liczonym miesiącem, a miesiącem go poprzedzającym. Przykładowo, w Polsce w grudniu 2020 roku inflacja podana w zestawieniu danych wynosiła 109.5645%, a w styczniu roku 2021 110.9483%. Zatem na cele dalszej analizy obliczona zmiana inflacji dla stycznia 2021 roku to $1.3838 p.p.$

- Bezrobocie - dane dotyczące bezrobocia pobrane zostały jako miesięczne odczyty stopy bezrobocia w danym kraju. Zatem tu również na potrzeby dalszej analizy policzyłem miesięczną zmianę poziomu stóp bezrobocia wyrażoną w punktach procentowych. Przykładowo, w Polsce w grudniu 2017 roku stopa bezrobocia wynosiła 4.4%, a w styczniu 2018 roku 4.2%. Zatem obliczona wartość zmiany poziomu bezrobocia w styczniu 2018 roku to $-0.2 p.p.$

- Stopy procentowe - dane dotyczące stóp procentowych w danym kraju zostały pobrane jako miesięczne odczyty poziomów stóp procentowych. W tym przypadku, podobnie jak wyżej, liczyłem miesięczne różnice stóp procentowych wyrażone w punktach procentowych. Przykładowo, w Polsce w listopadzie 2018 roku odnotowano wartość 3.19%, a w grudniu 2018 roku 2.94%. Zatem dla grudnia 2018 roku wartość zmiany stóp procentowych została policzona jako $-0.25 p.p.$

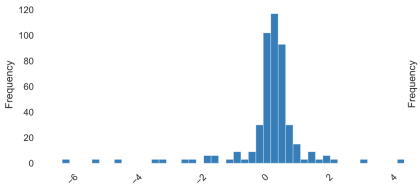
- Indeksy giełdowe - w celu obliczenia miesięcznej zmiany poziomu cen indeksu giełdowego danego kraju pobrałem dzienne odczyty giełdowe indeksu z każdego miesiąca badanego okresu, a następnie obliczyłem wartość zmiany jako procent:

$$\frac{S_{last\_close}}{S_{first\_open}} \cdot 100\%,$$

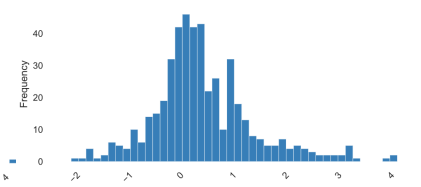
gdzie: S_{last_close} - cena akcji na zamknięciu ostatniej sesji giełdowej danego miesiąca, S_{first_open} - cena akcji na otwarciu pierwszej sesji giełdowej danego miesiąca.

2.3 Analiza wstępna danych (Exploratory Data Analysis)

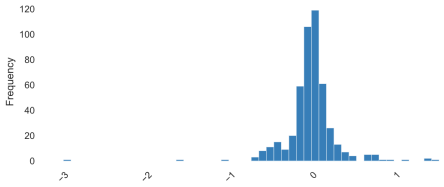
Kolejnym krokiem badania było przeprowadzenie wstępnej analizy danych (EDA), która ma na celu zapoznać nas z danymi ukazując pewne ogólne zależności, jakie można zaobserwować. Skorzystałem na tym etapie z biblioteki *ydata-profiling*[9] tworząc raport dla połączonych danych ze wszystkich krajów. Histogramy poszczególnych cech wyglądają następująco:



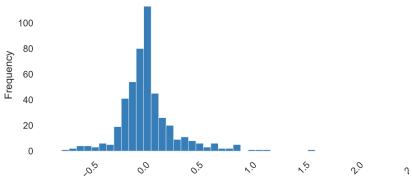
(a) Rozkład zmian PKB.



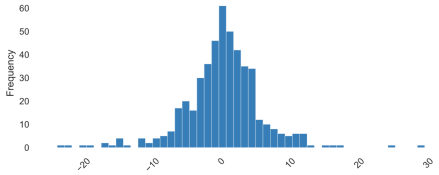
(b) Rozkład zmian poziomu inflacji.



(c) Rozkład zmian bezrobocia.



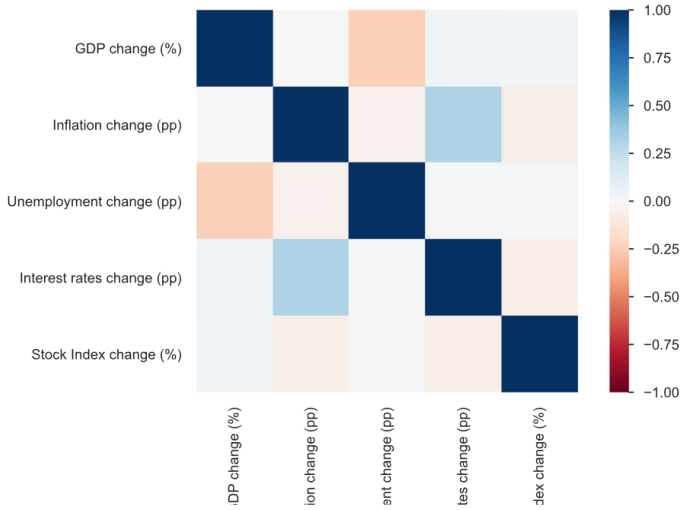
(d) Rozkład zmian poziomu stóp procentowych.



(e) Rozkład zmian cen indeksów giełdowych.

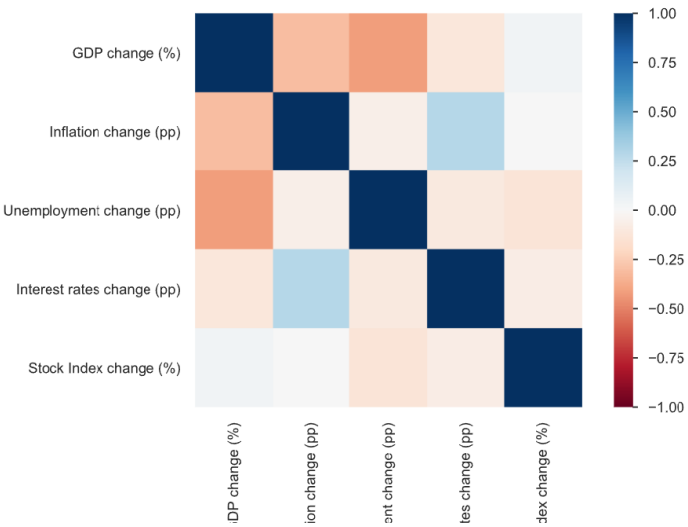
Możemy dostrzec, że wszystkie przypominają rozkład normalny, jednak różnią się 'sigmą' - zmiany poziomu inflacji czy cen indeksów giełdowych są znacznie bardziej rozrzucone względem średniej wartości niż zmiany PKB czy bezrobocia. Widzimy także, że wszędzie średnia wartość jest w pobliżu zera, natomiast w przypadku zmian PKB i zmian poziomu inflacji te wartości średnie są lekko przesunięte na prawo, na podstawie czego możemy wyciągnąć wniosek, że w latach 2018-2022 mogliśmy raczej obserwować wzrostu poziomu PKB, a także rosnącą inflację w grupie badanych państw.

Następnym etapem EDA było w tym przypadku przeanalizowanie macierzy korelacji cech (tzw. heatmapy). Macierz ta dla połączonych danych wszystkich państw wygląda następująco:



Rysunek 1: Macierz korelacji cech - dane ze wszystkich państw.

Jesteśmy więc w stanie wyciągnąć pierwsze wnioski dotyczące danych: niewielkie zależności występują między wartościami zmian bezrobocia i zmian PKB, a także między zmianami poziomów inflacji i stóp procentowych. Jeśli jednak skupimy się na cesze, która w naszej dalszej analizie będzie cechą zależną - na zmianach cen indeksów giełdowych, to łatwo możemy dostrzec, że nie jest ona zbytnio skorelowana z żadną z pozostałych cech, choć można dostrzec, że jest nieznacznie większa korelacja cen indeksów giełdowych z poziomem inflacji oraz poziomem stóp procentowych w porównaniu do dwóch pozostałych cech. Należy jednak pamiętać, że każdy kraj może mieć swoją specyfikę w sensie makroekonomicznym i rzeczywiście, macierze korelacji przygotowane dla każdego kraju z osobna czasem mocno odbiegają wyglądem od tej zaprezentowanej na Rysunku nr 1. Przykładem może być macierz korelacji przygotowana dla Słowacji:

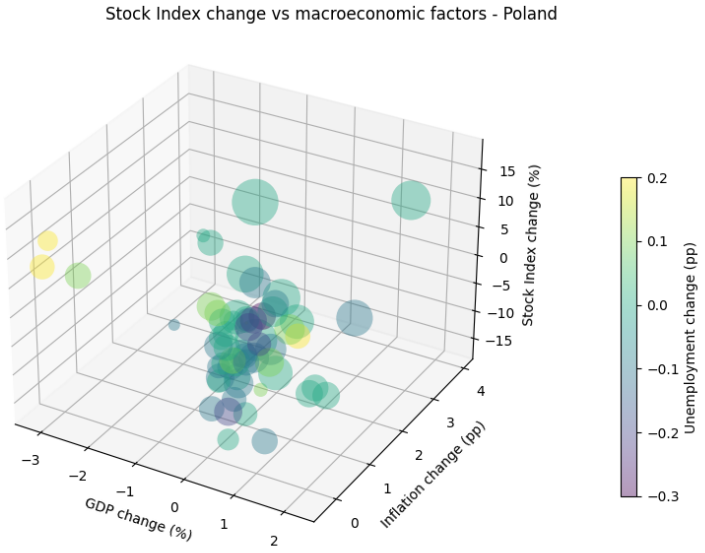


Rysunek 2: Macierz korelacji cech - dane ze Słowacji.

Macierze korelacji przygotowane dla wszystkich pozostałych państw są przedstawione w Dodatku(6.1) na końcu raportu.

2.4 Wizualizacje danych

Przed przystąpieniem do tworzenia modeli regresji postanowiłem dokonać jeszcze wizualizacji danych. W przypadku niniejszego problemu znajdujemy się w przestrzeni pięciowymiarowej. Aby przedstawić na wykresie dane tego typu wprowadziłem czwarty wymiar w postaci rozmiaru znaczników oraz piąty w formie koloru. W ten sposób jesteśmy w stanie przyjrzeć się zebranym danym dla danego kraju w nieco innej postaci. Przykładowo, wizualizacja danych dla Polski wygląda następująco:



Rysunek 3: Wizualizacja danych dla Polski.

Rozmiar znaczników dotyczy wysokości stóp procentowych - im większy znacznik, tym większa była zmiana stóp procentowych w danym miesiącu. Na podstawie Rysunku nr 3 możemy próbować wyciągać pierwsze wnioski z ogólnych obserwacji. Na przykład, że największe zmiany cen indeksu giełdowego (tu: WIG20) były powiązane ze wzrostem stopy bezrobocia. Są to jednak pewne przypuszczenia, które na relatywnie małym zbiorze danych mogą być całkowicie nietrafione. Dlatego też, w celu lepszego zbadania zależności i powiązań między poszczególnymi cechami skorzystałem z bibliotek implementujących różne modele regresji.
Wizualizacje przygotowane dla pozostałych państw znajdują się w Dodatku(6.2).

3 Modele

Przed przystąpieniem do trenowania modeli podzieliłem oba zbiory: wartości zmiennych zależnych i wartości zmiennych niezależnych, na dwie części - zbiór danych przeznaczony do testowania modelu oraz zbiór danych przeznaczony do oceny jakości danego modelu. Dokonałem tego podziału w proporcji $1:9$, ponieważ przy tej naturze problemu i liczbie dostępnych danych zdecydowałem, że kosztem gorszej generalizacji i większym ryzykiem overfittingu, postaram się lepiej dopasować model do zebranych danych. We wszystkich ocenach modelu jako miarę jakości przyjąłem wartość błędu średniokwadratowego (dalej będzie on czasem określany skrótowo jako MSE).

3.1 Model SVR

Mając problem o dużej złożoności, zdecydowałem się od razu rozpocząć trenowanie modeli od modelu SVR. W pierwszym podejściu zdecydowałem się użyć następujących parametrów:

$$kernel = rbf,\ C = 10,\ gamma = auto$$

Otrzymałem następujące wartości błędów średniokwadratowych dla modeli wytrenowanych na poszczególnych państwach:

Polska	30.8
Grecja	43.6
Łotwa	4.06
Portugalia	15.9
Litwa	18.6
Słowacja	15.3
Hiszpania	9.51
Słowenia	1.34

Tabela 1: Wyniki pierwszego eksperymentu.

Wyniki te nie są satysfakcjonujące. Poza wyjątkami, takimi jak Słowenia czy Łotwa, otrzymane wyniki są nieakceptowalnie wysokie, ponieważ te wartości błędów oznaczają, że model podczas predykcji może zwrócić wartość zmiany cen indeksu giełdowego z dokładnością niemalże równą rozpiętości przedziału wartości realnych, możliwych do osiągnięcia w rzeczywistości. W przypadku zmian cen indeksów giełdowych za taki przedział uznaję w przybliżeniu $[-15, 15]$. Zatem, przykładowo dla Grecji, pierwiastek z błędu kwadratowego wynosi w przybliżeniu 6.6, więc jest to dokładność dyskwalifikująca taki model. W celu poprawy doboru modelu zdecydowałem się skorzystać z metody GridSearch, która sprawdza wszystkie możliwe kombinacje parametrów zadanych poprzez argument $param\_grid$ dla podanego estymatora. Jako ocenę jakości modelu wybrałem, zgodnie z założeniem wspomnianym wyżej, wartość błędu średniokwadratowego. Wejściowa siatka parametrów wyglądała następująco:

$$kernel = [rbf,\ poly],\ C = [0.001,\ 0.01,\ 0.5,\ 1,\ 5,\ 10,\ 25,\ 50,\ 100],\ gamma = [scale,\ auto]$$

Wytrenowany model podczas predykowania wartości zmian cen indeksów giełdowych na tym samym zbiorze danych testowych, na którym pracował podstawowy model SVR, osiągnął następujące wartości MSE:

Polska	7.93
Grecja	48.1
Łotwa	4.2
Portugalia	12.8
Litwa	20.5
Słowacja	9.7
Hiszpania	9.45
Słowenia	1.71

Tabela 2: Wyniki eksperymentu z użyciem GridSearch.

W większości przypadków udało się poprawić wynik względem podstawowego modelu SVR, chociaż doszło także do pogorszenia miary jakości, jak np. w przypadku Grecji. Problem ten może wynikać z faktu, że GridSearch dokonuje pewnych optymalizacji podczas dostrajania hiperparametrów (zapobiegających m.in. przeuczeniu) i przy tak małym zbiorze danych metody te mogą realnie pogorszyć ogólny wynik modelu w przyjętym sposobie oceny.

3.2 Model RFR

Aby poprawić jakość modelu zdecydowałem się skorzystać z zupełnie innego rodzaju estymatora, którego nazwa to: Random Forest Regressor. Wykorzystuje on technikę zwaną lasem losowym do rozwiązywania problemów regresji, używając w tym celu drzew decyzyjnych. Wejściowe parametry, których użyłem to:

$$n\_estimators = 10,\ max\_features = sqrt,\ max\_depth = 3$$

Następnie, podobnie jak w przypadku modelu SVR, użyłem metody GridSearch do znalezienia możliwie najlepszych hiperparametrów estrymatora RFR. Jako siatkę przeszukiwanych parametrów przyjąłem:

$$n\_estimators = [10,\ 20,\ 30],\ max\_features = [sqrt],\ max\_depth = [2,\ 3,\ 5]$$

Otrzymane wartości MSE będące ocenami tychże modeli prezentuje tabela nr 3:

Kraj	MSE(RFR)	MSE(GridSearch[RFR])
Polska	19.4	12.2
Grecja	80.3	70.2
Łotwa	10.0	6.61
Portugalia	11.0	12.3
Litwa	19.0	17.8
Słowacja	13.1	7.85
Hiszpania	16.0	15.4
Słowenia	6.08	5.55

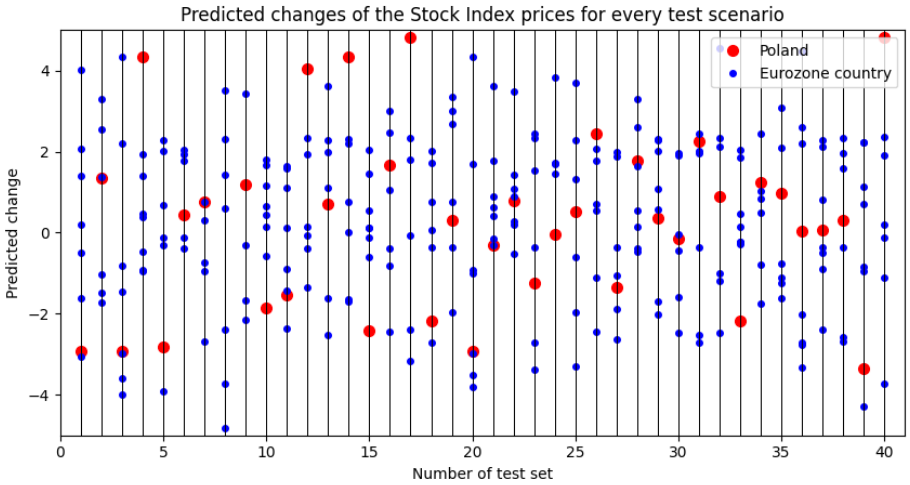
Tabela 3: Wyniki eksperymentu z użyciem GridSearch.

3.3 Wybór odpowiedniego modelu

Na podstawie powyższych eksperymentów widzimy, że nie ma jednej reguły dotyczącej tego, który model popełnia mniejszy błąd na przygotowanych danych. Zależności występujące w danych każdego państwa nie są wspólne, więc lepszym pomysłem od wybrania konkretnego modelu dla wszystkich krajów jest wytrenowanie czterech modeli dla każdego państwa: podstawowego SVR, GridSearch[SVR], podstawowego Random Forest Regressor oraz GridSearch[RFR], a następnie wybranie tego, który w danym przypadku popełnił najmniejszy MSE. Inny problem jaki wystąpił na tym etapie był taki, że w niektórych przypadkach predykowane wartości zmian cen indeksów giełdowych na podstawie testowych wartości cech niezależnych były tak naprawdę identyczne. Aby obejść ten problem, podczas wyboru modelu o najmniejszej osiągniętej wartośći MSE wprowadzam dodatkowe założenie, że predykowane wartości na testowym zbiorze danych nie mogą być identyczne (przy pewnym zadanym poziomie tolerancji). W ten sposób otrzymujemy finalne modele dla wszyskich analizowanych państw, które posłużą nam w ostatniej części badania.

4 Predykcje

Ostatnia część badania dotyczyła sprawdzenia, jak potencjalnie mogłyby zachować się ceny indeksów giełdowych państw podczas fikcyjnych scenariuszy. Ich tworzenie polegało na losowaniu czterech parametrów z ustalonych przedziałów. Parametry te oznaczały odpowiednio miesięczne zmiany: PKB, poziomu inflacji, poziomu bezrobocia i poziomu stóp procentowych w danym scenariuszu. Takich scenariuszy stworzyłem czterdzieści, a następnie na jednym wykresie naniosłem punkty odpowiadające przewidywanym wartościom zwróconym przez uprzednio wytrenowane modele. Wszystkie wartości dotyczące państw ze strefy Euro oznaczyłem jako niebieskie znaczniki, jedynie punkty odnoszące się do Polski zostały oznaczone czerwonym znacznikiem. Otrzymany wykres wygląda następująco:



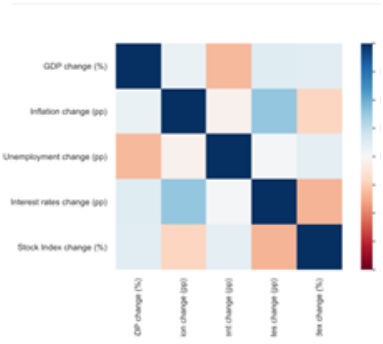
Rysunek 4: Porównanie przewidywanych zmian cen indeksów giełdowych w Polsce i pozostałych krajach.

5 Wnioski

Wykres przedstawiony na Rysunku nr 4 może sugerować, że nie istnieje żadna zależność między obecnością państwa w strefie Euro, a stabilnością cen jego indeksu giełdowego. Należy jednak pamiętać, że wykres ten został stworzony na podstawie modeli, które nawet po odpowiednim dostrojeniu parametrów osiągały wysoce niezadowalające wyniki przy ocenie ich jakości. Z tego powodu, dalsza analiza i wyciąganie jakichkolwiek wniosków przy modelach o tak małej dokładności wydają się bezcelowe. Przyczyną takiego stanu rzeczy może być bardzo mały zestaw danych. Chcąc sprawdzić, czy faktycznie tak jest, poszukałem i pobrałem znacznie większy zestaw danych tych samych cech. Zebrałem dane z lat 2011-2022 dla Polski i spróbowałem zrobić analogiczną analizę. Jednak również i w tym przypadku modele nie były w stanie się odpowiednio dopasować do danych. Zasadnym więc jest w takiej sytuacji postawić hipotezę, że nawet przy dużo większym zestawie danych, nie jesteśmy w stanie w zadowalający sposób przewidywać zmian cen na giełdzie na podstawie zmian czynników makroekonomicznych danego państwa.

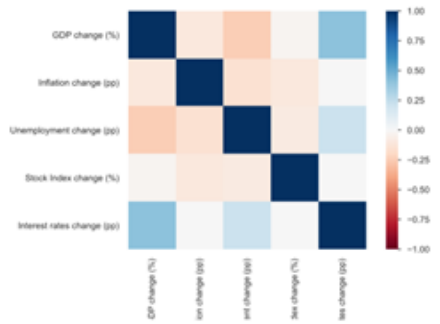
6 Dodatek

6.1 Macierze korelacji cech

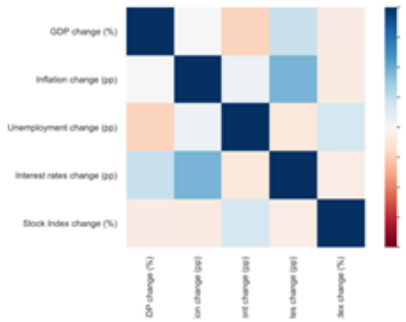


(a) Grecja.

(b) Łotwa.

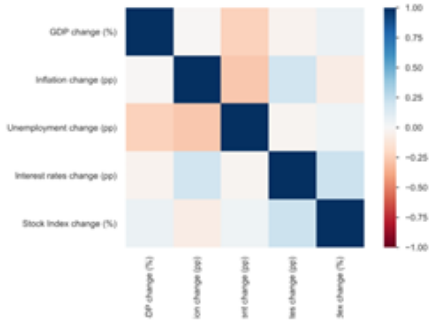


(c) Litwa.

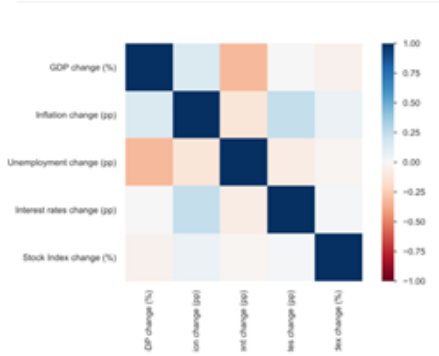


(d) Polska.

(e) Portugalia.

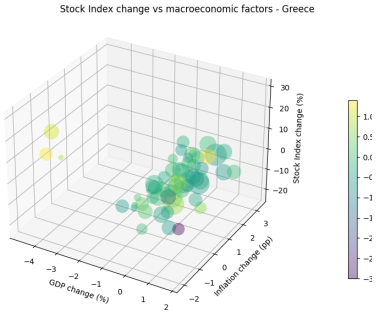


(f) Słowenia.

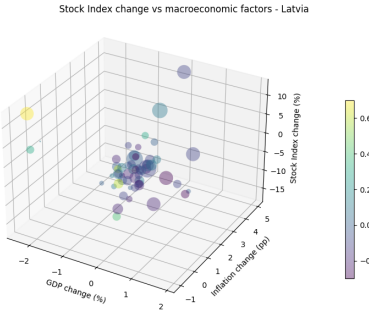


(g) Hiszpania.

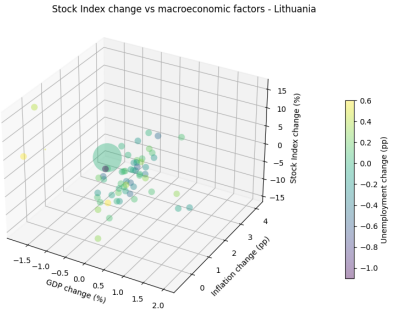
6.2 Wizualizacje danych



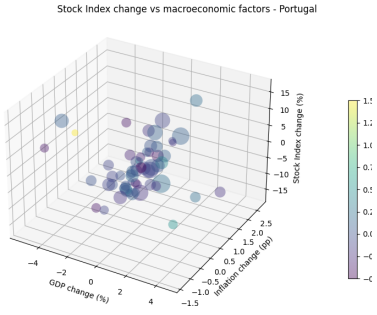
(a) Grecja.



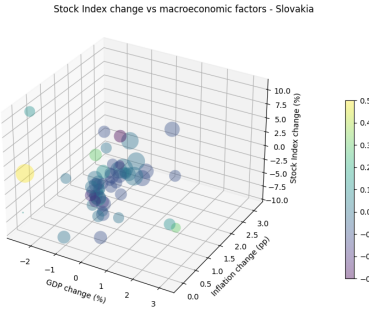
(b) Łotwa.



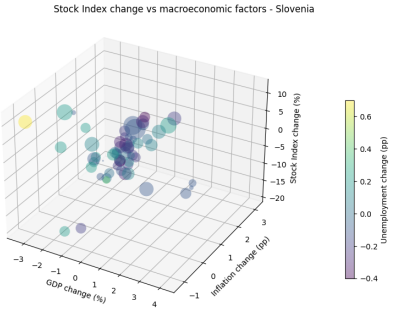
(c) Litwa.



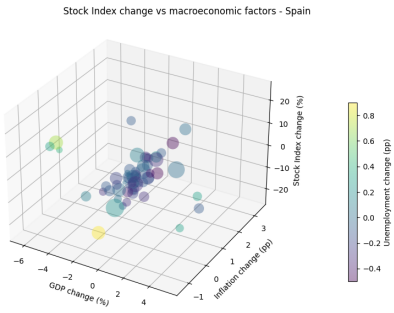
(d) Portugalia.



(e) Słowacja.



(f) Słowenia.



(g) Hiszpania.

6.3 Korzystanie z plików źródłowych programu

W celu poprawnego uruchomienia programu, który wczytuje dane, przetwarza je, trenuje modele oraz dokonuje predykcji, a także przygotowuje wykresy należy:

1. Sprawdzić, czy w folderze *Code* znajdują się pliki: *ChartsCreator.py, DataPreprocessing.py, EDA.py, Predictor.py, Start.py, SVRmodelCreator.py.*
2. W pliku *Start.py* w linijce 15 ustawić ścieżkę bezwzględną do katalogu z pobranymi danymi (katalog *Data*).
3. Uruchomić skrypt *Start.py.*

Bibliografia

[1] Eurostat. Gdp per capita, consumption per capita and price level indices. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=GDP_per_capita,_consumption_per_capita_and_price_level_indices, 2021. Data dostępu: 03.06.2023.

[2] OECD. Quarterly gdp. https://data.oecd.org/gdp/quarterly-gdp.htm?fbclid=IwAR1hZ_lkQwxmsCFggAmZbSV48rVFww_rHpHDB3PJ8QYB-VcGLTHy0ivdBrE. Data dostępu: 31.05.2023.

[3] OECD. Inflation (cpi). https://data.oecd.org/price/inflation-cpi.htm?fbclid=IwAR3WuAsEid4XBH5QMGcya2gLFx6EyGkgHWymOKO-DudbMWte4jOCv4NiQfs. Data dostępu: 31.05.2023.

[4] OECD. Unemployment rates. https://data.oecd.org/unemp/unemployment-rate.htm?fbclid=IwAR23Dcu1jN52UBTrjPVQWpFLRU8Dno01TnhS5En8VTRMq7IXeBcPt-lWE1s. Data dostępu: 31.05.2023.

[5] European Central Bank. Interest rates. https://sdw.ecb.europa.eu/home.do. Data dostępu: 31.05.2023.

[6] The Wall Street Journal. Market data. https://www.wsj.com/market-data. Data dostępu: 31.05.2023.

[7] Investing.com. Indieces. https://www.investing.com/indices/. Data dostępu: 31.05.2023.

[8] Ljubljana Stock Exchange. Slovenia stock market. https://ljse.si/en. Data dostępu: 31.05.2023.

[9] ydata-profiling. https://ydata-profiling.ydata.ai/docs/master/index.html.