

Klasyfikacja reżimów rynkowych dla kontraktu TTF front-month

Gabriela Jeznach, Szymon Pawłowski

29 stycznia 2026

1 Wstęp

Celem projektu jest klasyfikacja reżimów rynkowych dla dziennych notowań kontraktów terminowych na gaz naturalny notowanych na ICE Endex (Dutch TTF Natural Gas Futures). Konkretnie, dla danego dnia t interesuje nas najbliższy (kalendarzowo) kontrakt miesięczny. Taki szereg czasowy jest nazywany „front-month” i pozwala na analizę ceny danej klasy produktów terminowych w sposób ciągły w jednym wymiarze czasowym. Problem ma charakter stricte predykcyjny, dla każdego dnia handlu t chcemy przewidzieć, jaki reżim będzie dominował w kolejnych 20 dniach roboczych, korzystając wyłącznie z informacji dostępnej do dnia t włącznie.

1.1 Opis danych i wnioski z eksploracji szeregu

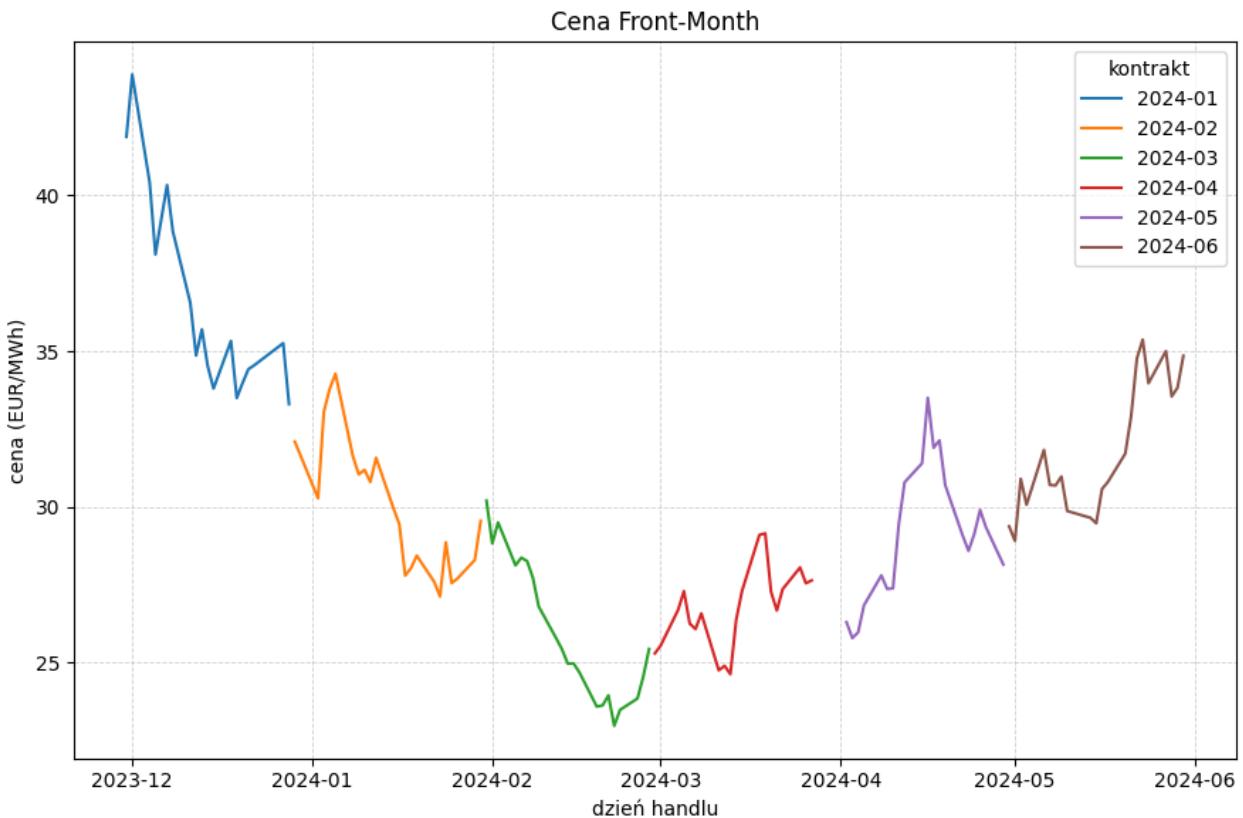
Dane wejściowe stanowią dzienne ceny kontraktów typu front-month, w szczególności ceny otwarcia oraz zamknięcia (to głównie z nich będziemy korzystać). Na rysunku 1 przedstawiliśmy schematycznie sposób łączenia kolejnych kontraktów miesięcznych w jeden ciąg czasowy (krótsze przerwy są wynikiem kolorowania wykresu ceny kontraktami, a dłuższe nałożeniem się zmiany kontraktu z weekendem, pomimo tych przerw wykres jest poprawny). Taka konstrukcja pozwala ograniczyć analizę do jednego wymiaru czasowego, przy jednoczesnym zachowaniu ciągłości ekonomicznej szeregu cenowego. Zakres danych obejmuje notowania wszystkich kontraktów od stycznia 2016 aż do grudnia 2024, co daje dni handlu od 2015-11-30 aż do 2024-11-28.

Zanim przejdziemy do wniosków z analizy eksploracyjnej, wprowadzimy definicję logarytmicznych stóp zwrotu, z których będziemy często korzystać w dalszej części raportu.

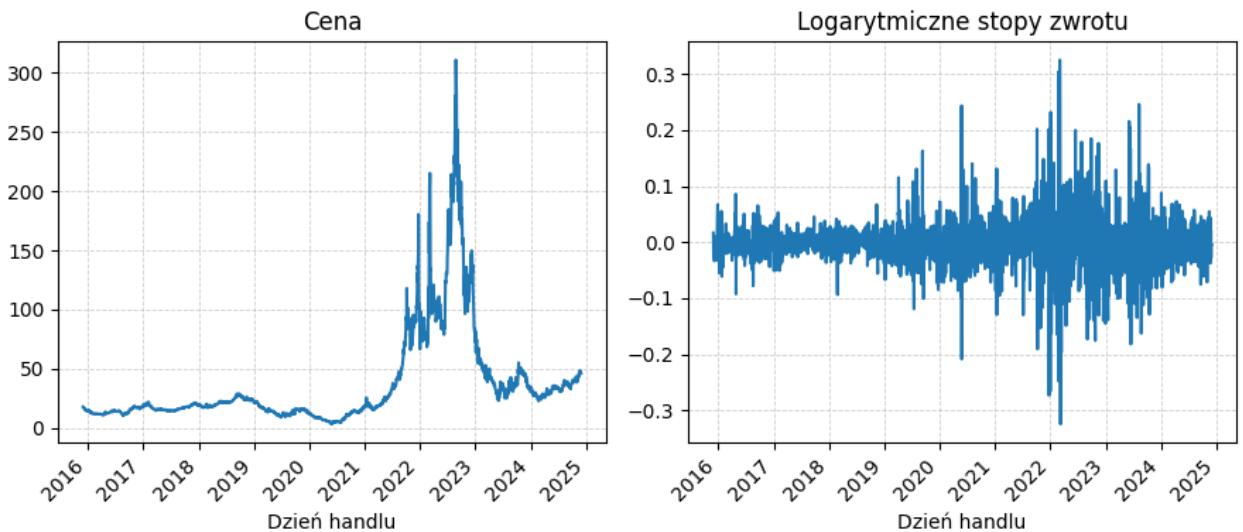
Niech C_t oznacza cenę zamknięcia kontraktu w dniu t , natomiast O_t cenę otwarcia w tym samym dniu. Logarytmiczna stopa zwrotu jest zdefiniowana jako

$$r_t = \begin{cases} \log\left(\frac{C_t}{C_{t-1}}\right), & \text{jeżeli dzień } t \text{ nie jest pierwszym dniem handlu danym kontraktem,} \\ \log\left(\frac{C_t}{O_t}\right), & \text{jeżeli dzień } t \text{ jest pierwszym dniem handlu danym kontraktem.} \end{cases}$$

Takie podejście pozwala uniknąć sztucznego skoku ceny wynikającego wyłącznie ze zmiany kontraktu miesięcznego, który nie ma interpretacji ekonomicznej jako rzeczywisty ruch rynkowy. Na rysunku 2 widoczny jest wykres ceny oraz logarytmicznych stóp zwrotu dla całego badanego okresu.



Rysunek 1: Ceny kontraktów typu front-month po połączeniu w jeden szereg czasowy.

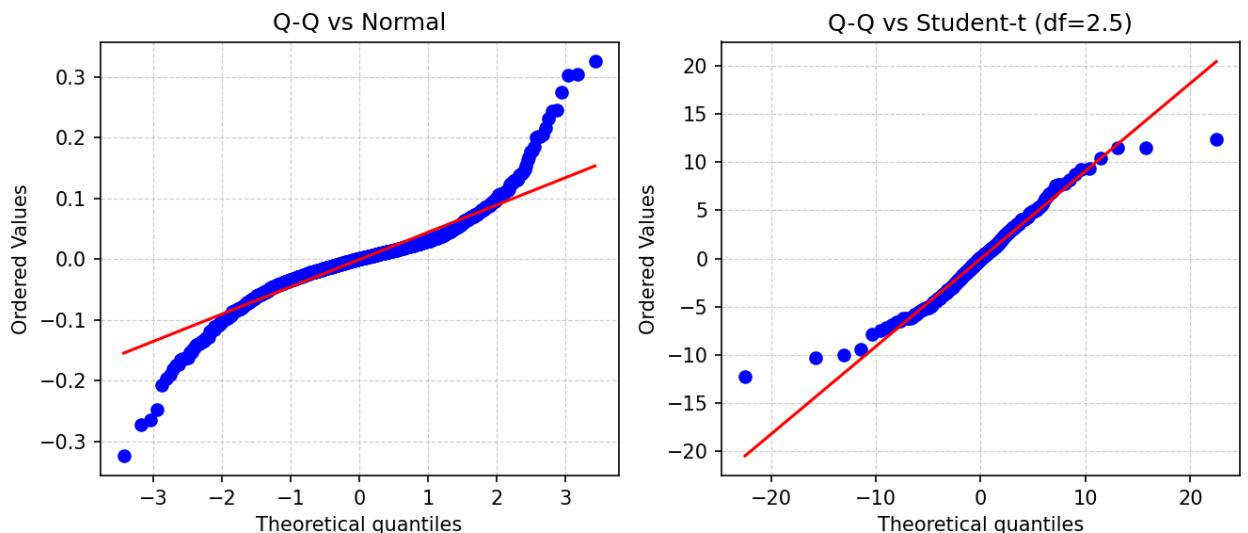


Rysunek 2: Wykresy ceny oraz logarytmicznych stóp zwrotu r_t .

Analiza eksploracyjna wskazuje kilka typowych własności danych finansowych, istotnych z punktu widzenia dalszego doboru modeli:

- **Niestacjonarność cen:** test ADF dla szeregu cen nie pozwala odrzucić hipotezy o pierwiastku jednostkowym ($p \approx 0.119$), a test KPSS odrzuca hipotezę stacjonarności trendowej ($p \approx 0.01$), co sugeruje, że bezpośrednie modelowanie poziomu ceny może być problematyczne. Z tego względu większość analizy prowadzimy na logarytmicznych stopach zwrotu.

- **Stacjonarność r_t :** dla zwrotów logarytmicznych ADF odrzuca hipotezę o pierwiastku jednostkowym ($p \approx 0$), a KPSS nie odrzuca hipotezy stacjonarności ($p \approx 0.1$). To uzasadnia przejście na reprezentację w postaci zwrotów.
- **Rozkład r_t :** analiza rozkładu logarytmicznych stóp zwrotu wskazuje na istotne odstępstwa od założenia normalności. Test D'Agostino–Pearsona odrzuca hipotezę o normalnym rozkładzie zwrotów ($p \approx 0$). Obserwacja wykresów Q-Q względem rozkładu normalnego oraz rozkładu t Studenta (rys. 3) wskazuje na wyraźnie grubsze ogony empirycznego rozkładu w porównaniu do rozkładu normalnego. Jednocześnie dopasowanie rozkładu Studenta pozwala znacznie lepiej opisać zachowanie ogonów. Wartość parametru liczby stopni swobody jest niewielka, co potwierdza silną leptokuryczność danych (dodatnią kurtozę, czyli grubsze ogony). Test Kołmogorowa–Smirnowa nie daje podstaw do odrzucenia hipotezy, że zwroty pochodzą z rozkładu t Studenta o dopasowanych parametrach ($df = 2.5$, $loc = -0.00045$, $scale = 0.0264$) ($p \approx 0.34$).
- **Słaba, ale niezerowa struktura zależności w czasie:** autokorelacja r_t jest ograniczona i trudna do bezpośredniego wykorzystania liniowymi zależnościami, natomiast może istnieć informacja nieliniowa (co motywuje metody nieliniowe i bogatsze cechy).
- **Heteroskedastyczność:** test ARCH Engle'a wskazuje silną obecność efektu ARCH ($p \ll 10^{-6}$), a zależności w kwadratach zwrotów są istotne (Ljung–Box dla r_t^2). To zmotywowało nas do przetestowania modeli ARCH/GARCH.



Rysunek 3: Wykresy Q-Q logarytmicznych stóp zwrotu r_t .

1.2 Konstrukcja etykiety i interpretacja reżimu

Wszystkie porównywane metody uczą się tego samego celu, dla każdej daty t etykieta opisuje „trend” w przyszłym oknie $t+1, \dots, t+H$. Dla ustalonego horyzontu H budujemy miarę dla przyszłego okna

$$S_t^{(H)} = \frac{\sum_{i=1}^H r_{t+i}}{\sigma(r_{t+1:t+H}) \sqrt{H}}, \quad (1)$$

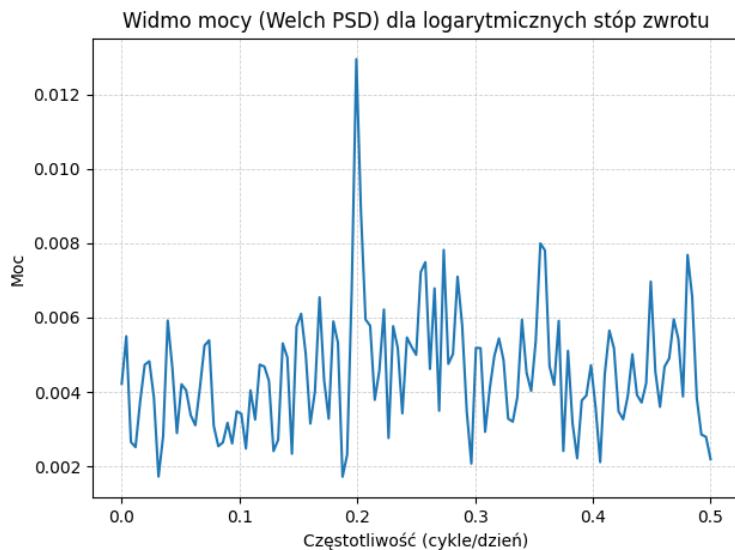
gdzie $\sigma(\cdot)$ jest odchyleniem standardowym w przyszłym oknie.

Etykiety są trójklasowe

$$y_t = \begin{cases} 1, & \text{gdy } S_t^{(H)} \geq k, \\ -1, & \text{gdy } S_t^{(H)} \leq -k, \\ 0, & \text{w przeciwnym razie.} \end{cases} \quad (2)$$

Próg k dobierany jest tak, aby empiryczne udziały trzech klas były możliwie zblżone do $\frac{1}{3}$, co stabilizuje uczenie klasyfikatorów i sprawia, że metryki wieloklasowe (np. macro-F1) są bardziej miarodajne. W praktyce rozkład klas jest bliski zbalansowanemu (znormalizowana entropia etykiety ≈ 0.997).

Dodatkowo oceniliśmy przewidywalność szeregu logarytmicznych stóp zwrotu r_t przy użyciu miary forecastability opartej na entropii spektralnej. Widmo mocy estymowane metodą Welch'a jest w dużej mierze płaskie (rys. 4), co wskazuje na brak dominujących częstotliwości i silnych struktur cyklicznych w danych. Znormalizowana entropia spektralna wynosi 0.987, a odpowiadający jej wynik forecastability jedynie 0.013, co oznacza, że badany szereg jest bardzo bliski białemu szumowi.



Rysunek 4: Widmo mocy logarytmicznych stóp zwrotu r_t .

W konsekwencji historia r_t niesie niewielką ilość informacji predykcyjnej o przyszłym zachowaniu rynku. Implikuje to, że zadanie prognozowania reżimu, a w szczególności klasyfikacji etykiety y_t , jest z natury trudne i nie należy oczekwać bardzo wysokiej skuteczności modeli opartych wyłącznie na przeszłych zwrotach. Celem dalszej analizy jest zatem identyfikacja modeli osiągających stabilną, choć umiarkowaną, przewagę nad losowym zgadywaniem, a nie konstrukcja modelu o wysokiej, deterministycznej trafności.

1.3 Zakres projektu i porównywane podejścia

Projekt porównuje metody klasyczne dla szeregów czasowych oraz nowoczesne podejścia uczenia maszynowego w ramach jednego, spójnego protokołu ewaluacyjnego. Wszystkie modele uczone są na tym samym zbiorze danych wejściowych, względem tej samej zmiennej objaśnianej, a ich jakość oceniana jest przy użyciu identycznej procedury walidacji typu walk-forward z rozszerzającym się oknem uczącym. Dla każdej daty t , model trenowany jest na danych dostępnych do $t-1$ i

generuje predykcję dla obserwacji t , co minimalizuje ryzyko wystąpienia „look-ahead bias” i zapewnia wiarygodną ocenę out-of-sample.

Pierwsze trzy lata danych (768 dni handlu) są wyłączone z części walidacyjnej (1537 dni handlu) i służą wyłącznie do inicjalnego uczenia modeli oraz doboru hiperparametrów tam, gdzie jest to wymagane. Takie założenia pozwalają nam na bezpośrednie porównanie poszczególnych podejść.

2 Modele i uzasadnienie wyboru

2.1 Model ARCH/GARCH do prognozowania zmienności i miary typu Sharpe

Model z rodziny ARCH/GARCH (z częścią autoregresyjną w równaniu średniej) jest dopasowywany do szeregu r_t , a następnie generuje prognozy średniej i wariancji. Z prognoz budowana jest prognoza miary $S_t^{(H)}$, po czym stosujemy regułę progową $\pm k$ do mapowania na klasy $\{-1, 0, 1\}$. Przetestowaliśmy model w dwóch wariantach, w pierwszym k jest wyznaczane na podstawie własnej prognozy in-sample (na danych treningowych) (analogicznie jak ma to miejsce przy tworzeniu etykiet), a w drugim jest hiperparametrem. Zdecydowaliśmy się na ten model, ponieważ

- w danych występuje heteroskedastyczność i klastrowanie zmienności, które są naturalnym obszarem zastosowań modeli ARCH/GARCH,
- etykieta jest skonstruowana jako relacja „zwrot do ryzyka” w przyszłym oknie, więc podejście jest spójne z definicją celu,
- rozkład t-Studenta pozwala lepiej uwzględnić grube ogony rozkładu stóp zwrotu.

2.2 Klasyfikator wieloklasowy XGBoost na cechach inżynieryjnych

Budujemy wektor cech opisujących zachowanie ceny i stóp zwrotu, w szczególności opóźnienia r_t , statystyki kroczące (sumy stóp zwrotu i miary zmienności), miary położenia ceny w przedziale min–max w oknie, wygładzania wykładnicze (EWMA), cechy kalendarzowe (dzień tygodnia, miesiąc, kwartał) oraz wskaźnik zmiany kontraktu. Następnie uczymy trójklasowy model XGBoost. Zdecydowaliśmy się na ten model, ponieważ

- metoda dobrze modeluje nieliniowości i interakcje między cechami,
- stanowi punkt odniesienia dla podejścia opartego o inżynierię cech,
- zapewnia kontrolę nad przeuczeniem poprzez regularyzację i próbkowanie obserwacji/cech.

2.3 Sieć konwolucyjna z wejściem postaci reprezentacji obrazowych (RP i GAF)

Szereg ceny w przesuwającym się oknie (60 dni) przekształcamy do postaci dwóch obrazów, wykresu rekurencji (RP) oraz pola Grama w postaci kątowej (GAF). Otrzymana para obrazów stanowi dwukanałowe wejście do klasyfikatora opartego o sieć konwolucyjną. Zdecydowaliśmy się na ten model, ponieważ

- reprezentacje RP i GAF kodują strukturę dynamiki w oknie czasowym bez ręcznego projektowania złożonych cech,

- sieci konwolucyjne potrafią wykrywać powtarzalne wzorce w danych 2D, co może odpowiadać sygnaturom reżimów (trend, konsolidacja, powrót do średniej),
- podejście jest komplementarne do metod tablicowych, nacisk jest przeniesiony z inżynierii cech na uczenie reprezentacji (Representation Learning).

3 Proces trenowania i konstrukcji modelu

3.1 Protokół walidacji

Ewaluację modeli przeprowadziliśmy w schemacie walk-forward z rozszerzającym się oknem uczącym. Dla każdej daty handlowej t model trenujemy na danych historycznych, a następnie generujemy predykcję dla obserwacji z dnia t . Wszystkie elementy obliczane w procedurze (etykiety i cechy) konstruujemy tak, aby w momencie predykcji nie wykorzystywać informacji z przeszłości.

Etykieta y_t jest definiowana na podstawie miary zależnej od przyszłego okna $t+1, \dots, t+H$ (w eksperymencie $H = 20$ dni roboczych). W związku z tym, w kroku predykcji dla dnia t zbiór treningowy może zawierać jedynie obserwacje, dla których etykieta jest w pełni określona na podstawie danych sprzed dnia t . Oznacza to, że ostatnia obserwacja treningowa spełnia warunek

$$t_{\text{train}} \leq t - (H + 1).$$

Zapewnia nas to, że przyszłe okno wykorzystywane do konstrukcji etykiet w treningu nie nakłada się na dzień predykcji.

Część modeli wykorzystuje cechy obliczane na podstawie historii cen (opóźnienia, statystyki kroczące, reprezentacje obrazowe RP i GAF). W dniu predykcji t do ekstrakcji cech przekazywaliśmy dane z przedziału $1, \dots, t$, natomiast uczenie odbywało się wyłącznie na podzbiorze $1, \dots, t_{\text{train}}$. Następnie z wektora predykcji wybieraliśmy wartość odpowiadającą dniu t .

3.2 Metryki jakości i diagnostyka

Jako miary jakości przyjęto macro-F1 oraz balanced accuracy. Miary te są raportowane zarówno na podziale walidacyjnym (dla diagnostyki procesu uczenia), jak i w ewaluacji out-of-sample w protokole walk-forward.

Macro-F1 traktuje klasy symetrycznie i jest wrażliwa na degradację jakości predykcji klasy neutralnej. Balanced accuracy jest średnią czułości (recall) liczoną osobno dla każdej z klas. Dodatkowo analizowaliśmy globalne macierze pomyłek w klasach $\{-1, 0, 1\}$, co pozwala ocenić charakter błędów, w szczególności zdolność modeli do rozróżniania klasy neutralnej 0 od klas trendowych -1 i 1 .

3.2.1 Model ARCH/GARCH

Przetestowaliśmy dwa warianty podejścia ARCH/GARCH, w którym najpierw prognozowana jest ścieżka logarytmicznych stóp zwrotu w horyzoncie $H = 20$, następnie agregowana do prognozy sumy zwrotów oraz prognozy zmienności w tym samym horyzoncie, a na końcu wyznaczana jest prognoza miary typu Sharpe i mapowana na klasy przez regułę progową $\pm k$. Warianty różniły się sposobem traktowania parametru k : k estymowane endogenicznie na danych treningowych na podstawie prognoz in-sample oraz k traktowane jako hiperparametr.

W praktyce oba warianty okazały się nieskuteczne w tym zadaniu. Model niemal zawsze przewidywał klasę neutralną (0). Dla predykcji out-of-sample udział klasy 0 w predykcjach wyniósł 0.9967, co prowadziło do bardzo niskiej jakości klasyfikacji klas trendowych, macro-F1 = 0.1617 oraz balanced accuracy = 0.3327. Macierz pomyłek pokazuje, że obserwacje klas -1 i 1 były prawie zawsze przypisywane do 0, co skutkowało zerową czułością dla klas trendowych.

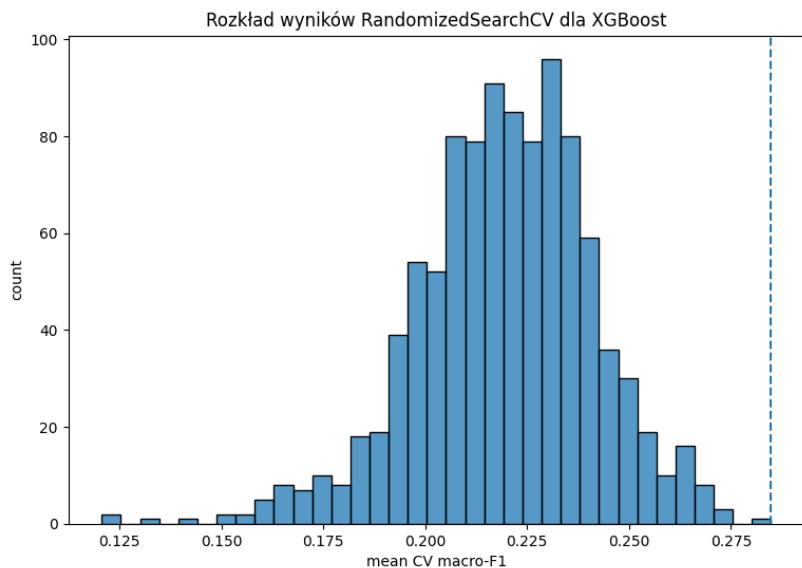
Źródłem problemu jest to, że model ARCH/GARCH w tej konfiguracji koncentruje się głównie na modelowaniu i prognozowaniu zmienności, natomiast prognoza części średniej (mean equation) jest zbyt słaba, aby wygenerować stabilny sygnał kierunkowy w horyzoncie 20 dni. W konsekwencji prognozowana miara pozostaje blisko zera, a reguła progowa mapuje ją niemal zawsze do klasy neutralnej.

Samo podejście jest spójne z definicją etykiety (zwrot do ryzyka w przyszłym oknie), jednak w tej postaci nie dostarcza użytecznego modelu reżimu. Naturalnym kierunkiem dalszych prac byłoby wykorzystanie ARCH/GARCH wyłącznie do prognozowania zmienności oraz połączenie tej prognozy z oddzielnym, silniejszym modelem regresyjnym dla prognozowania średniej (np. nieliniowym regresorem lub modelem opartym o cechy), a następnie budowanie miary z dwóch niezależnych komponentów.

3.2.2 Model XGBoost

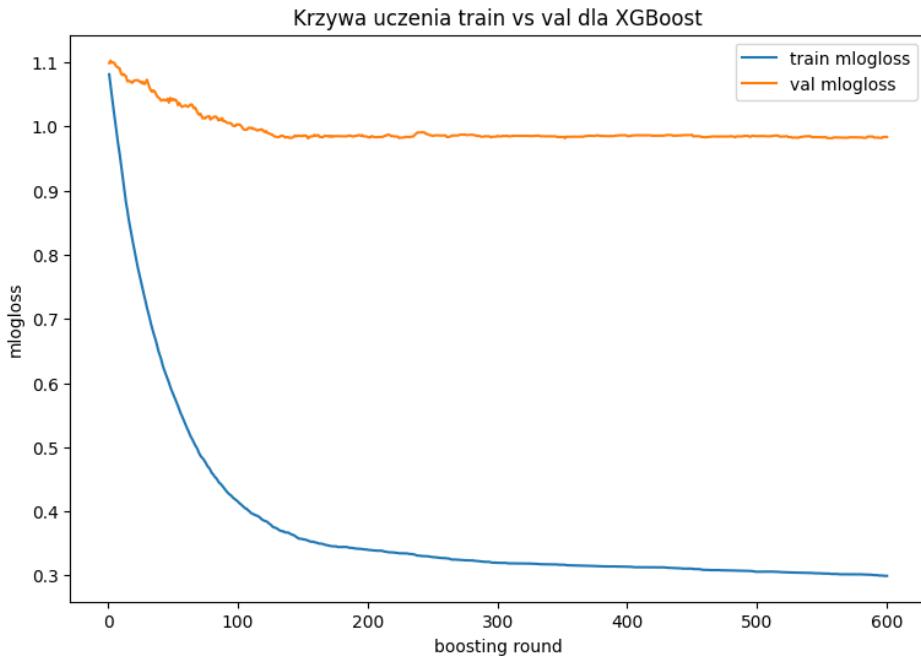
Dobór hiperparametrów klasyfikatora XGBoost wykonaliśmy metodą RandomizedSearchCV z walidacją TimeSeriesSplit (5 podziałów). Zastosowaliśmy lukę (gap) równą $H = 20$ dniom roboczym, aby ograniczyć przeciek informacji wynikający z faktu, że etykieta zależy od przyszłego okna $t+1:t+H$. Kryterium optymalizacji w przeszukiwaniu było macro-F1.

Przeszukiwaliśmy przestrzeń hiperparametrów obejmującą m.in. głębokość drzew, minimalną wagę liścia, tempo uczenia, liczbę estymatorów, parametry próbkowania obserwacji i cech oraz regularyzację. Najlepsza konfiguracja uzyskała średni wynik walidacji krzyżowej macro F1 równy około 0.285. Na rysunku 5 pokazaliśmy rozkład wartości mean_test_score (macro-F1) dla wszystkich próbkiwanych konfiguracji.



Rysunek 5: Rozkład wyników RandomizedSearchCV dla XGBoosta

Dodatkowo dla najlepszych hiperparametrów wyznaczyliśmy krzywą uczenia opartą o wieloklasową log-stratę (mlogloss) na zbiorze treningowym i walidacyjnym, z zachowaniem luki $\text{gap}=H$ (rysunek 6). Na wykresie widać spadek straty walidacyjnej w początkowych rundach boostingu i późniejsze wypłaszczenie, podczas gdy strata treningowa maleje dalej. Oznacza to, że kolejne iteracje boostingu poprawiają dopasowanie do danych treningowych, ale nie przekładają się istotnie na poprawę generalizacji.



Rysunek 6: Przebieg funkcji straty dla najlepszych parametrów dla XGBoosta

3.2.3 Diagnostyka cech, feature importance dla XGBoost

Po dopasowaniu finalnego modelu XGBoost obliczyliśmy wbudowane miary ważności cech. Raportowaliśmy trzy statystyki, gain (średnia poprawa funkcji celu uzyskiwana przez podziały oparte o daną cechę), weight (liczba użyć cechy w podziałach) oraz cover (średnia liczność obserwacji przechodzących przez węzły, w których użyto danej cechy).

Wyniki są spójne, najwyższą ważność według gain mają cechy opisujące poziom ceny względem lokalnych ekstremów oraz miary zmienności zwrotów w krótkich i średnich horyzontach. W top cechach dominują m.in. close_max_60, close_max_40, close_max_10 oraz lr_std_5, lr_std_10, lr_std_20, lr_std_40. Sugeruje to, że model rozpoznaje reżimy głównie poprzez informację o bieżącym poziomie ryzyka (zmienność) oraz poprzez położenie ceny blisko lokalnych maksimów/minimów (kontekst trendu w oknie).

Część cech ma wysoką wartość weight, ale niekoniecznie najwyższy gain (np. niektóre opóźnione statystyki kroczące), co wskazuje, że są używane często jako stabilne predyktory w wielu miejscach drzewa, nawet jeśli pojedynczy podział nie daje dużej poprawy jakości. Miara cover sugeruje natomiast, że pewne proste cechy związane z ruchem sesyjnym i zwrotami (np. gap, lr_lag_k, lr2_lag_k) bywają wykorzystywane w podziałach obejmujących duże fragmenty danych.

Należy podkreślić, że ważność cech nie określa kierunku wpływu na prawdopodobieństwo klas i nie ma interpretacji przyczynowej. Jest to diagnostyka pokazująca, które grupy cech są faktycznie wykorzystywane przez model.

3.2.4 CNN na reprezentacjach RP i GAF

W trakcie wstępnych eksperymentów z siecią konwolucyjną zaobserwowaliśmy problem degeneracji predykcji. Przy standardowym wczesnym zatrzymaniu opartym na minimalizacji straty walidacyjnej (cross-entropy), model dążył do rozwiązań, w których przewidywana była niemal wyłącznie jedna z klas trendowych, a klasa neutralna 0 nie była przewidywana w ogóle. Zjawisko to skutkowało zerową liczbą predykcji klasy 0 oraz niskimi wartościami macro-F1.

Aby ograniczyć ten efekt, zastosowaliśmy dwie modyfikacje procedury trenowania

- zrównoważone próbkowanie obserwacji w mini-batchach treningowych (balanced batches),
- zmianę kryterium wczesnego zatrzymania z minimalizacji straty walidacyjnej na maksymalizację macro-F1 na zbiorze walidacyjnym.

Po zastosowaniu tych zmian model zaczął przewidywać wszystkie klasy, w tym klasę neutralną.

Ze względu na niedeterministyczny charakter uczenia sieci neuronowych, jakość CNN oceniliśmy na podstawie pięciu niezależnych treningów z różnymi ziarnami losowymi. Dla ustalonej konfiguracji hiperparametrów raportowaliśmy średnie oraz odchylenia standardowe miar jakości na zbiorze walidacyjnym. Uzyskaliśmy macro-F1 0.333 ± 0.024 oraz balanced accuracy 0.375 ± 0.028 , co wskazuje na umiarkowaną stabilność wyniku względem inicjalizacji losowej.

Nie wykonaliśmy pełnej czasowej walidacji krzyżowej hiperparametrów CNN z luką $H = 20$. Decyzja wynikała z kosztu obliczeniowego wielokrotnego trenowania sieci w reżimie czasowym oraz z ograniczonego spodziewanego zysku informacyjnego przy słabym sygnale predykcyjnym w danych. Zamiast tego skupiliśmy się na diagnostyce procedury uczenia oraz na powtórzeniach losowych dla wybranej, stabilnej konfiguracji.

4 Zakończenie i wnioski

4.1 Zakres i protokół oceny końcowej

W tej sekcji porównaliśmy modele na wspólnym oknie predykcji out-of-sample (OOS), wyznaczonym od pierwszej daty, dla której dostępne są kontrakty z $\text{delivery_date} \geq 2019-01-01$. Ocena obejmuje $n = 1537$ obserwacji. Dla zachowania porównywalności wszystkie wyniki raportowane są na tym samym zbiorze dat.

Wszystkie modele predykcyjne działają w reżimie walk-forward, dla każdej daty d model jest dopasowywany wyłącznie na danych historycznych dostępnych przed d , z uwzględnieniem horyzontu etykiety $H = 20$ (ostatnia etykieta, która może być znana w momencie predykcji dla d , pochodzi z dnia $d - (H + 1)$).

4.2 Zestaw porównywanych podejść

W porównaniu uwzględniliśmy

- baseline_prev, predykcję etykiety jako ostatniej etykiety znanej w momencie predykcji (etykieta przesunięta o $H + 1$),
- baseline_freq, losowanie etykiety z rozkładu częstości klas obserwowanych do tej pory (start z rozkładu wyuczonego na początkowym okresie treningowym, aktualizacja online),

- XGBoost, klasyfikator wieloklasowy z cechami ręcznie skonstruowanymi z cen i zwrotów,
- CNN (RP+GAF), sieć konwolucyjna operującą na obrazowej reprezentacji krótkich okien czasowych,
- ARCH/GARCH (dwa warianty), model oparty o prognozę Sharpe'a, w wariantie z k estymowanym endogenicznie oraz w wariantie z k jako hiperparametr.

4.3 Podsumowanie wyników ilościowych

Tabela 1 prezentuje wyniki OOS dla wszystkich porównywanych modeli.

model	n	macro-F1	balanced accuracy
baseline_prev	1537	0.4230	0.4227
cnn	1537	0.3716	0.3905
xgb	1537	0.3437	0.3426
baseline_freq	1537	0.3135	0.3137
arch_no_k	1537	0.1617	0.3327
arch_with_k	1537	0.1617	0.3333

Tabela 1: Wyniki OOS na wspólnym oknie porównawczym.

4.4 Wnioski jakościowe z macierzy pomyłek

4.4.1 baseline_prev

Wynik baseline_prev (macro-F1 = 0.4230, balanced accuracy = 0.4227) pozostaje najwyższy spośród wszystkich rozważanych podejść. Rozkład predykcji jest bardzo zbliżony do empirycznego rozkładu klas, a wartości precision i recall są względnie wyrównane pomiędzy klasami. Oznacza to, że w danych występuje silna trwałość reżimu w czasie, a inercja etykiety stanowi mocny punkt odniesienia.

4.4.2 baseline_freq

Baseline losujący etykiety z częstości klas (macro-F1 = 0.3135) jest słabszy od baseline_prev. Model ten w przybliżeniu odtwarza proporcje klas, jednak nie wykorzystuje informacji o lokalnej kontynuacji reżimu, co skutkuje częstszymi pomyłkami, szczególnie w klasach trendowych.

4.4.3 XGBoost

XGBoost (macro-F1 = 0.3437, balanced accuracy = 0.3426) przewiduje wszystkie trzy klasy i poprawia wynik względem baseline_freq. W macierzy pomyłek widzimy jednak, że największym problemem pozostaje klasa neutralna, dla której zarówno precision, jak i recall wynoszą około 0.28. Wyniki sugerują, że ręcznie konstruowane cechy cenowe nie przechwytują w pełni informacji zawartej w lagach etykiety.

4.4.4 CNN (RP+GAF)

Sieć konwolucyjna osiąga macro-F1 = 0.3716 oraz balanced accuracy = 0.3905, co plasuje ją pomiędzy baseline_prev a modelem XGBoost. Obserwujemy wyraźnie lepszą identyfikację reżimów trendowych, szczególnie klasy -1, kosztem słabszej identyfikacji klasy neutralnej. Oznacza to, że reprezentacja

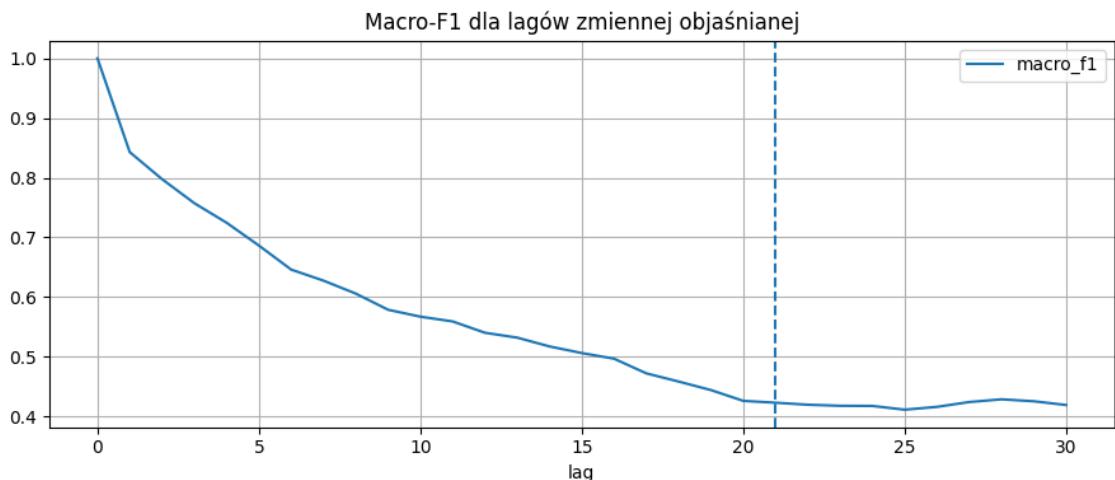
obrazowa skutecznie wychwytuje struktury trendowe i zmiennościowe w krótkich oknach czasowych, jednak ma tendencję do nadmiernej polaryzacji predykcji (model dzieli wyniki na skrajne wartości ignorując te pośrednie). Wynik CNN pokazuje, że nieliniowe modele sekwencyjne są w stanie zbliżyć się do jakości najlepszego baseline'u, choć nadal go nie przewyższają.

4.4.5 ARCH/GARCH (warianty z i bez k)

Oba warianty ARCH/GARCH osiągają macro-F1 ≈ 0.1617 przy balanced accuracy bliskiej $\frac{1}{3}$. Jest to konsekwencja niemal całkowitej degeneracji predykcji do klasy neutralnej. Modele te skutecznie prognozują zmienność, jednak prognoza średniej jest zbyt słaba, aby generować użyteczne sygnały kierunkowe na horyzoncie 20 dni.

4.5 Komentarz do trwałości reżimu

Rysunek 7 przedstawia macro-F1 w funkcji opóźnienia etykiety wykorzystanej jako predyktor. Obserwujemy silną degradację jakości wraz ze wzrostem opóźnienia, jednak nawet dla laga $\ell = 21$, odpowiadającego ograniczeniu informacyjnemu wynikającemu z definicji etykiety ($H = 20$), macro-F1 pozostaje na poziomie około 0.42. Potwierdza to wysoką trwałość reżimów rynkowych w analizowanych danych.



Rysunek 7: Macro-F1 dla baseline opartego o opóźnioną etykietę w funkcji laga ℓ .

4.6 Ograniczenia i dalsze kierunki prac

Wyniki wskazują, że w obecnym ustawieniu zadania dominującą informacją predykcyjną są lag etykiety, a modele oparte wyłącznie o cechy cenowe nie przełamują tej bariery. W dalszych pracach rozważylibyśmy rozszerzenie zbioru zmiennych objaśniających poza same ceny (np. wolumen, zmienne fundamentalne, dane pogodowe, opóźnienia zmiennej objaśnianej, policzone bez wycieku, przesunięcie musi być odpowiednio duże i k inaczej policzone) oraz dalszą stabilizację modeli sekwencyjnych, w szczególności sieci konwolucyjnych.