Uczenie maszynowe w finansach

Case 3

Michał Orzołek

1. Opis modelu

W przedstawionym rozwiązaniu zastosowano model XGBoost w wersji regresyjnej (XGBRegressor), którego celem jest prognoza jednodniowej przyszłej log-stopy zwrotu instrumentu AMD.

Zmienna docelowa została zdefiniowana jako y_next= ln(P_t+1)-ln(P_t) na podstawie ceny AdjClose, a zbiór cech obejmuje 32 zmienne techniczne skonstruowane z opóźnień log-zwrotów, miar kroczących (sumy, średnie, odchylenia), relacji ceny do SMA i EMA, wskaźników RSI i MACD, szerokości pasm Bollingera oraz komponentów wolumenowych i prostych miar intraday.

Dane wejściowe zostały pobrane z yfinance dla interwału dziennego od 2019-01-01, a po usunięciu braków wynikających z konstrukcji cech i targetu zakres danych użytych w modelu wynosi 2019-03-14 do 2025-12-04, przy czym liczba obserwacji po dropna to 1693. Podział czasowy został wykonany w sposób sekwencyjny na trzy zbiory: treningowy od 2019-03-14 do 2022-12-30 (959 obserwacji), walidacyjny od 2023-01-03 do 2023-12-29 (250 obserwacji) oraz testowy od 2024-01-02 do 2025-12-04 (484 obserwacje).

Strojenie hiperparametrów przeprowadzono metodą RandomizedSearchCV z TimeSeriesSplit (4 foldy), a funkcją celu była minimalizacja MAE. Najlepszy zestaw hiperparametrów obejmuje m.in. 600 estymatorów, maksymalną głębokość 6, learning rate 0.1 oraz colsample_bytree 0.6.

Istotnym elementem konstrukcji strategii jest sposób generowania sygnału inwestycyjnego. Model produkuje prognozę y_pred, która jest porównywana z progiem θ. Sygnał surowy przyjmuje wartość 1, gdy prognoza przekracza próg, i 0 w przeciwnym razie. Następnie sygnał wykonywany jest z opóźnieniem o jeden dzień poprzez przesunięcie o 1 okres, co w kodzie realizuje kolumna Signal.

Wybór theta został przeprowadzony na zbiorze walidacyjnym w prostym backteście log-zwrotów, a jako kryterium wyboru użyto najwyższej wartości cum_return_% w tabeli walidacyjnej. Najlepszy próg w tym porównaniu wyniósł 0.0005. Dodatkowo z raportu ekspozycji wynika, że średni udział dni z pozycją long na podstawie kolumny Signal wynosi 72.05% w treningu, 74.0% w walidacji oraz 85.33% w teście.

2. Miary jakości predykcji

Jakość predykcji modelu została oceniona na walidacji i teście przy użyciu MAE oraz RMSE. Na zbiorze walidacyjnym uzyskano MAE równe 0.02183 i RMSE równe 0.02983,

natomiast na zbiorze testowym MAE wyniosło 0.02391, a RMSE 0.03407. Po doborze hiperparametrów model został ponownie wytrenowany na połączonym zbiorze treningowym i walidacyjnym, a następnie wygenerowano predykcje dla całego zbioru data w celu konstrukcji sygnału inwestycyjnego.

3. Miary efektywności strategii

Backtest strategii został wykonany z użyciem biblioteki backtesting.py na wydzielonym okresie testowym 2024-01-02 do 2025-12-04, z kapitałem początkowym 100 000 oraz prowizją 0.1%.

Zbudowano dwie strategie: MLAllInLongCash, która otwiera pozycję long, gdy Signal wynosi 1 i zamyka ją, gdy Signal wynosi 0, oraz BuyHold, która otwiera pojedynczą pozycję long na początku okresu i utrzymuje ją do końca.

Wyniki strategii ML wskazują na końcową wartość kapitału 191 508.53 oraz stopę zwrotu 91.51% w analizowanym okresie, przy ekspozycji 90.91%. Dla tej samej próbki benchmark kup i trzymaj osiągnął końcową wartość 161 096.01 oraz stopę zwrotu 61.10% przy ekspozycji 99.59%. Roczna stopa zwrotu według raportu backtesting.py wynosi 40.26% dla strategii ML i 28.18% dla kup i trzymaj, a CAGR odpowiednio 26.27% i 18.67%.

Zmienność roczna jest bardzo zbliżona w obu podejściach i wynosi 76.37% dla ML oraz 75.96% dla benchmarku. Współczynnik Sharpe'a dla ML to 0.527 przy wartości 0.371 dla kup i trzymaj, a wskaźnik Sortino wynosi 1.271 dla ML i 0.814 dla benchmarku.

Maksymalne obsunięcie kapitału w strategii ML to -56.44%, przy średnim obsunięciu -11.09%, natomiast w kup i trzymaj maksymalne obsunięcie wynosi -62.95%, przy średnim -10.85%. Czas trwania najdłuższego obsunięcia to 453 dni dla ML oraz 579 dni dla benchmarku.

Statystyki transakcyjne strategii ML obejmują 29 transakcji, win rate 44.83%, najlepszą transakcję na poziomie 56.28%, najgorszą -16.03% oraz średni wynik na transakcję 2.27%. Profit factor wynosi 2.68, expectancy 2.83%, a średni czas trwania transakcji to około 20 dni, przy medianie 3 dni i maksymalnym czasie 325 dni. Dla kup i trzymaj raport wskazuje pojedynczą transakcję trwającą praktycznie cały okres testowy.

W ramach weryfikacji poprawności środowiska i danych wykonano także sanity backtest na uproszczonych danych OHLC, którego wyniki pokrywają się z parametrami strategii kup i trzymaj w okresie testowym, co potwierdza spójność feedu danych wykorzystanych w backtestach.

4. Podsumowanie

Zaimplementowany pipeline obejmuje pełną ścieżkę od pobrania danych, przez inżynierię cech, estymację i strojenie modelu XGBoost z walidacją szeregową, aż po

konstrukcję sygnału long/flat z doborem progu theta na walidacji oraz finalny backtest na wydzielonej próbie testowej z użyciem backtesting.py. W ramach okresu testowego strategia oparta o sygnał ML osiąga wyższy zwrot końcowy i wyższy CAGR niż strategia kup i trzymaj, przy zbliżonej zmienności rocznej, nieco wyższym Sharpe i mniejszym maksymalnym obsunięciu. Wyniki te są przedstawione liczbowo w raportach backtestu oraz zwizualizowane w wygenerowanych plikach HTML zgodnie z logiką kodu.