Prognozowanie zwrotów z inwestycji giełdowych – prezentacja wyników

Raport projektu z uczenia maszynowego w finansach

Kasia Górczyńska

Jacek Jankowiak

Mikołaj Szymczak

1. Wprowadzenie

Niniejszy raport prezentuje wyniki końcowe analizy polegającej na prognozowaniu zwrotów z inwestycji giełdowych (ROI) metodami uczenia maszynowego. Etapy doboru zmiennych oraz czyszczenia zbioru danych zostały opisane w notatnikach ipynb oraz w prezentacji.

W projekcie zaimplementowano 5 modeli:

- Dummy Regressor – model, który zawsze zwracał średnią wartość ze wszystkich ROI w zbiorze treningowym. Ten model służył tylko do wstępnego wdrożenia całego pipeline'u eksperymentu, żeby przetestować, czy napisane metody działają. Z tego powodu Dummy Regressor został pominięty przy prezentacji wyników w tym raporcie.

- Regresja Liniowa – Wybór zmiennych objaśniających do modelu regresji liniowej odbył się poprzez analizę korelacji liniowych zmiennych ze zmienną objaśnianą, czyli ROI. Mapa cieplna korelacji została zawarta w prezentacji. Ostatecznie wybrano 4 zmienne:
    - score - conviction score, czyli ocena inwestycji dokonana przez ekspertów
    - close – cena zamknięcia danego dnia
    - year – rok
    - econ_branch – branża danej firmy. Ta zmienna została rozbita na 18 zmiennych binarnych, które określały, czy firma jest z danej branży, czy nie jest.
- Sieć LSTM

- Sieć ANN
- XGboost

Parametry sieci neuronowych oraz algorytmu XGBoost również zostały opisane w prezentacji.

Regresja liniowa stanowiła w badaniu tak zwany model bazowy. Oznacza to, że wszystkie inne modele były z nią porównywane, a celem było przebicie jej wyników. Jeżeli bardziej złożone modele okazałyby się gorsze, albo tylko niewiele lepsze od regresji liniowej, to w środowisku biznesowym należałoby się zastanowić, czy opłaca się wdrażać coś bardziej skomplikowanego i kosztowniejszego niż regresja lub modele naiwne.

Wyjściem modeli były prognozowane wartości ROI dla danej spółki w danym okresie. Stworzone zostały cztery warianty tej zmiennej objaśnianej – ROI za miesiąc od dnia obserwacji, ROI za kwartał, za pół roku oraz za rok. Każdy z tych wariantów był rozpatrywany osobno i trenowane były dla nich osobne modele, choć o niezmiennych parametrach.

roi_month	roi_quarter	roi_halfyear	roi_year
0.015807	-0.085058	0.048927	0.406473
-0.048047	-0.010888	0.087811	0.348402
-0.002717	-0.068954	-0.070992	0.178668
0.006640	-0.023935	0.124923	0.361334
0.008034	0.025860	0.082350	0.442882

Podział na zbiór treningowy i testowy odbywał się poprzez losowanie obserwacji z następującymi proporcjami: 67% obserwacji trafiło do zbioru treningowego, a 33% do testowego. Mimo daty stanowiącej jedną ze zmiennych dane nie zostały potraktowane jako szeregi czasowe ze względu na zbyt duże luki w sekwencjach.

Dla każdego modelu odbywało się testowanie jego konsekwentności i umiejętności generalizacji poprzez etap walidacji krzyżowej. W przypadku regresji liniowej zbiór danych był dzielony na 10 segmentów, które były następnie przypisywane po kolei do zbioru treningowego i testowego (parametr n_splits = 10). Podział ten wykonywany był 3 razy, za każdym razem

w inny sposób (parametr n_repeats =3). Oznacza to, że model regresji liniowej tworzony był 30 razy dla każdego wariantu ROI. Wartości błędów były następnie uśredniane. Pozostałe modele, czyli XGBoost, LSTM i ANN były dzielone tylko na 5 segmentów i nie odbywało się powtórzenie tego podziału. Było to spowodowane znacznie dłuższym czasem uczenia się tych modeli – walidacja krzyżowa o takich parametrach jak przy regresji liniowej zajęłaby dziesiątki godzin.

Do oceny jakości modeli wybrane zostały trzy miary błędów – błąd średniokwadratowy MSE (mean square error), MAPE (Mean Absolute Percentage Error) oraz R^2, czyli współczynnik determinacji.

Poniższe wyniki zostały posortowane malejąco według wartości R^2.

2. Roi w horyzoncie miesięcznym

Nazwa modelu	MSE	MAPE	R^2
XGBoost	0,05	619 274 041 232,48	0,3
LSTM	0,01	705 941 500 777,65	0,19
ANN	0,01	368 011 572 565,68	0,04
Reg. Liniowa	0,06	291 491 207 293,52	0

W przypadku prognozy zwrotu z inwestycji z miesięcznym wyprzedzeniem największym R^2 cechował się model XGBoost. Mimo największego dopasowania modelu do danych XGBoost miał najwyższe wartości błędów MSE i MAPE. Sieć ANN miała R^2 równy 0,04, ale cechowała się najniższymi wartościami błędów MSE i MAPE.

Największym błędem MSE i najmniejszym współczynnikiem determinacji charakteryzowała się regresja liniowa. Oznacza to, że wszystkie modele zdołały pokonać model bazowy pod względem tych miar. Co ciekawe regresja liniowa miała najniższy procent błędu MAPE, chociaż nadal był on bardzo wysoki.

3. Roi w horyzoncie kwartalnym

Nazwa modelu	MSE	MAPE	R^2
XGBoost	0,09	250 739 376 984,75	0,19
LSTM	0,02	282 388 887 074,43	0,17
ANN	0,02	308 098 796 086,99	0,17

Reg. Liniowa	0,1	202 136 145 490,01	0,01

Również w horyzoncie kwartalnym najbardziej dopasowany okazał się XGBoost – jego współczynnik determinacji wyniósł 19%. Mimo to ponownie miał on najwyższy błąd MSE, jednak tym razem jego błąd MAPE był drugi najniższy ze wszystkich modeli zaraz po regresji liniowej.

Zarówno sieć ANN jak i LSTM miały takie same MSE oraz R^2. LSTM osiągnął trochę niższy błąd MAPE. Z tego powodu uplasował się na drugim miejscu.

Model bazowy ponownie cechował się najwyższym MSE i najniższym R^2, i ponownie wypadł najlepiej pod względem miary MAPE.

4. Roi w horyzoncie półrocznym

Nazwa modelu	MSE	MAPE	R^2
XGBoost	0,12	186 034 237 810,27	0,36
ANN	0,03	211 277 354 626,58	0,3
LSTM	0,05	143 931 291 247,74	0,05
Reg. Liniowa	0,15	180 460 677 781,09	0,01

Wnioski z prognozy z półrocznym wyprzedzeniem nie różniły się wiele od prognoz w krótszych horyzontach. Kolejność w rankingu była taka sama jak w horyzoncie kwartalnym.

Wartości współczynnika determinacji okazały się większe niż przy poprzednich wariantach ROI. Wynika z tego, że modele dopasowywały się do danych tym lepiej, im dłuższy był horyzont prognozy.

Sieć LSTM osiągnęła w tym wariancie błąd MAPE niższy niż regresja liniowa, tak więc model bazowy został pokonany również pod tym względem.

5. Roi w horyzoncie rocznym

Nazwa modelu	MSE	MAPE	R^2
ANN	0,06	284 356 554 522,30	0,42
XGBoost	0,18	333 229 490 246,91	0,42
LSTM	0,1	539 952 214 709,77	0,04
Reg. Liniowa	0,22	286 588 457 232,89	0,02

W horyzoncie rocznym zmienił się lider rankingu. Pod względem współczynnika determinacji R^2 obydwa modele – XGBoost i sieć ANN – osiągnęły dopasowanie 42%, jednak ANN cechowało się przy tym znacznie niższymi wartościami błędów MSE i MAPE.

6. Interpretacja wyników

Pomimo przetestowania pięciu modeli uczenia maszynowego dopasowanie modelu, czyli współczynnik R^2, ani razu nie przebiło bariery 50%. Zaistniała natomiast zależność, że R^2 miało tym większą wartość, im dalszy był horyzont prognozy. Największe dopasowanie zostało osiągnięte podczas prognozy ROI z wyprzedzeniem rocznym – sieć ANN i model XGBoost osiągnęły po 42% R^2. Możliwym wyjaśnieniem tej zależności jest, że modele lepiej uczyły się przewidywać długoterminowe zależności. Być może wybrane zmienne objaśniające – w tym przede wszystkim conviction score ustalany przez ekspertów – były przydatniejsze przy długookresowej prognozie.

Błędy MSE na pierwszy rzut oka mogą się wydawać niskie, jednak jest to spowodowane niewielkimi wartościami zmiennej objaśnianej. Więcej informacji o jakości prognoz dostarczył procentowy błąd MAPE. Osiągał on wartości procentowe wyrażane w milionach. Oznacza to, że prognozowane wartości bardzo istotnie różniły się od rzeczywistych.

Podsumowując, głównym wnioskiem z przeprowadzonej analizy jest uznanie złożoności badanego zagadnienia. Predykcja dokładnych wartości zwrotów inwestycji w odległym horyzoncie czasowym to zadanie wymagające dużego nakładu kosztów i czasu. Wszystkie badane modele pokonały model bazowy, jednak żaden z nich nie osiągnął wyników, które mogłyby wspomóc podejmowanie decyzji w środowisku biznesowym – ich błędy były zbyt duże.

Dalsze kroki przy analizowaniu tematu prognozowania zwrotów z inwestycji giełdowych powinny obejmować redefinicję zadania i ustalenia innej zmiennej celu. Przykładowym kierunkiem rozwoju może być podział wartości ROI na przedziały i utworzenie klas, jak na przykład Słaba Inwestycja (-nieskończoność; 0,2) , Średnia Inwestycja (0,2; 0,6), Dobra Inwestycja (0,6; +nieskończoność). Wtedy analiza przekształciłaby się w zadanie klasyfikacji. Zmniejszyłoby to znacząco wymiarowość zadania, a ponadto umożliwiłoby wykorzystanie algorytmów klasyfikacji, które mogłyby okazać się skuteczniejsze dla tego zagadnienia.