GPW Analizer – badania nad modelem wnioskującym

# Wstęp

GPW Analizer jest systemem, który ma wspomagać decyzję inwestorów giełdowych. Po sparametryzowaniu przez użytkownika zapytania ma dać odpowiedź, czy dana inwestycja się zwróci. Przeznaczony jest dla inwestorów długoterminowych, tj. takich, którzy nastawiają się na mniejsze ryzyko, gdyż wpłacają pieniądze na dłuższe okresy czasu. Obsługiwanym rynkiem jest rynek regulowany Giełdy Papierów Wartościowych w Warszawie. Program wyliczający prognozy jest osobnym podprojektem o nazwie Stock Analyzer, został napisany w języku JAVA z użyciem pakietu WEKA. Niniejszy dokument jest sprawozdaniem z prac podjętych nad znalezieniem modelu, który jak najlepiej przewidzi przyszłe kursy spółek.

## Okres przewidywania

GPW Analizer przewiduje wartości akcji spółek na następujące okresy:

* 30 dni
* 60 dni
* 90 dni
* 180 dni
* 360 dni

## Dane statystyczne

Stock Analyzer jest programem, który dzięki danym statystycznym prognozuje przyszłe notowania. Do analiz zgromadziliśmy następujące dane:

* Giełdy zagraniczne – wartości. Dane pochodzą ze strony http://finance.yahoo.com
  + Dow Jones Industrial Average
  + NASDAQ Composite
  + Eurostoxx50
* Kursy walut – wartości. Dane ze strony http://nbp.pl/.
  + Dolar
  + Euro
* Indeksy GPW – kurs zamknięcia oraz wartość obrotu. Dane ze strony http://www.gpw.pl.
* Akcje Spółek – kurs zamknięcia, wolumen obrotu, liczba transakcji, wartość obrotu. Dane ze strony http://www.gpw.pl.

W bazie danych posiadamy powyższe dane od 01.01.2007r.

## Zasada dokonywania analiz

Akcje spółek są trudne do przewidzenia, gdyż zależą od wielu czynników. Pierwszym problemem była odpowiedź na pytanie, jak wnioskować wartości akcji. Można scalić ze sobą otrzymane dane i na ich podstawie wyliczyć ostateczną wartość akcji. W Stock Analyzer zdecydowaliśmy się przyjąć inny model wnioskujący, bowiem wnioskowanie odbywa się na następującej zasadzie:

* Przewidywanie wartości giełd zagranicznych.
* Na podstawie ww. przewidywań, oraz danych historycznych – waluty.
* Na podstawie ww. przewidywań, oraz danych historycznych – indeksy GPW.
* Na podstawie ww. przewidywań, oraz danych historycznych – wartości akcji.

Przy czym do przewidzenia pierwszych trzech punktów używamy wszystkich wcześniejszych danych, natomiast do wartości akcji bierzemy pod uwagę jedynie dane historyczne i indeksy, do których dana spółka należy.

## Algorytmy użyte do predykcji

Do predykcji wartości używamy następujących algorytmów zaimplementowanych w pakiecie WEKA:

* LinearRegression
* IBk
* AdditiveRegression
* Vote
* M5P
* REPTree

Docelowo ostateczny wynik ma być średnią ważoną wyników poszczególnych algorytmów.

## Od czego zależny jest wynik? Co tak naprawdę chcemy ustalić?

Ostateczny wynik jest zależny przede wszystkim od wyników poszczególnych algorytmów. Jednak aby model był dokładniejszy postanowiliśmy każdemu z algorytmów nadać wagę i to właśnie ustalenie odpowiedniej wagi jest jednym z dwóch celów badań. Drugim jest ustalenie liczby dni, którą należy wziąć do poszczególnej analizy (30, 60, 90, 180, 360). Manipulując tymi parametrami spróbujemy ustalić model, który w przeszłości uzyskałby najlepsze wyniki.

## Błąd modelu

Aby móc określić, który model jest lepszy, musimy ustalić, co przyjmiemy za jego błąd. Badanie dla określonych parametrów da nam wynik w postaci tablicy z składającej się z n elementów, w którym każdy oznacza pomyłkę algorytmu w stosunku do realnej wartości wyrażony w procentach. Następnie każdy element tej tablicy mnożymy przez współczynnik, który przewartościuje nam błędy. Współczynnik ten wyraża się wzorem:

gdzie:

* z – nasz współczynnik
* w – wartość wyliczona przez model
* maxW – największa wartość spośród wszystkich wyliczonych

Na sam koniec wyliczamy statystykę, tj. funkcję elementów tablicy zdefiniowaną jako średnią arytmetyczną:

Tym samym otrzymujemy błąd naszego pomiaru.

## Oczyszczenie danych

Spółki na giełdzie zmieniają się dynamicznie, co utrudnia nieco przetestowanie modeli na historycznych danych. Testowanie odbyło się na latach 2011 – 2013, jednak pod uwagę zostały wzięte wyłącznie spółki i indeksy, które znajdowały się na giełdzie 30.06.2011 r. W ten sposób unikamy sytuacji, w której algorytmy próbują przewidzieć notowanie dla indeksów/spółek bez danych historycznych. Takie założenie z pewnością wypacza poniekąd wyniki, jednak aplikacja jest ciągle w fazie rozwoju, przez co model powinien być uaktualniany na bieżąco. Z tak przygotowanego zestawu danych będą tworzone kolejne pliki arff.

## Ustalenie wag

Jako, że korzystamy z 5 algorytmów postanowiliśmy, by suma ich wag wynosiła 30. Wagi do badań przydzielaliśmy w następujący sposób:

* algorytm dominujący – 5 algorytmów z wagą 3, jeden z wagą 15.
* średnia arytmetyczna
* przydzielanie na podstawie odchylenia od oczekiwanego rezultatu, odbywa się na następującej zasadzie:

Dla każdego algorytmu przy średniej arytmetycznej liczony jest x, będący różnicą między jego wynikiem, a oczekiwanym rezultatem.

Sumujemy wszystkie x, co daje nam sumę różnic (sx), następnie liczymy trafność dla danego algorytmu ze wzoru:

Następnie na podstawie trafności przydzielamy proporcjonalnie wagi.

# Badania nad modelem 360 dniowym

## Ustalenie liczby dni przyjętych do modelu

Okres 360 dniowy jest największym, jaki bierzemy pod uwagę w GPW Analizer, przez co założyliśmy, że do wnioskowania przyjmiemy wszystkie zgromadzone dane, a więc od 2007 roku.

## Ustalenie wag dla algorytmów

Początkowo przeprowadzone analizę dla modelu ze średnią arytmetyczną, następnie na tej podstawie ustalono wagi dla modelu z odchyleniem od oczekiwanego rezultatu. Badania przeprowadzono na 3 datach:

* 01.01.2012 r.
* 25.05.2012 r.
* 01.10.2012 r.

Wagi od 1 do 6 to kolejne algorytmy dominujące, 7 to średnia, 8 to policzone odchylenie.

W tabeli przedstawiono wyniki badań:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
| 01.01.2012 | 8,172224478 | 8,570744265 | 8,366161368 | 8,865891084 | 8,579868946 | 8,658106296 | 8,577850694 | 8,577891397 |
| 25.05.2012 | 8,948640676 | 9,515484972 | 10,00231863 | 9,600358579 | 12,07859159 | 10,38403065 | 10,12127338 | 10,1220218 |
| 01.10.2012 | 10,3751265 | 10,75765007 | 11,90311779 | 10,40799546 | 13,02775716 | 12,31731947 | 11,42806404 | 11,42880502 |
| Średnia | 9,1653306 | 9,614626 | 10,09053 | 9,624748 | 11,22874 | 10,45315 | 10,0424 | 10,04291 |

Jak widać z danych oraz wykresu, w okresie 360 dniowym najlepiej sprawdził się algorytm, w którym dominujący był LinearRegression, zatem na takie rozłożenie wag zdecydowaliśmy się dla modelu 360 dniowego.

# Badania nad modelem 180 dniowym

## Ustalenie liczby dni przyjętych do modelu

Badania zostały przeprowadzone dla 4 różnych liczb dni przyjętych do modelu:

* 360 dni
* 240 dni
* 180 dni
* 120 dni

Badania przeprowadzono dla 3 różnych dat, a następnie wyciągnięto średnią, wyniki przedstawiono w tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dni | Okres 1. | Okres 2. | Okres 3. | Średnia |
| 360 | 8,28993 | 11,4237 | 18,96929 | 12,89431 |
| 240 | 8,161206 | 11,83765 | 19,50505 | 13,16797 |
| 180 | 8,150708 | 11,91225 | 19,95227 | 13,33841 |
| 120 | 8,157131 | 12,16221 | 18,55221 | 12,95719 |

Zarówno w tym, jak i w dalszych badaniach okresy nie zostały wybrane przypadkowo. Podczas pierwszego z nich giełda była stosunkowo stabilna. Podczas drugiego wartości zmieniały się bardziej nieprzewidywalnie, natomiast ostatni z nich zmieniał te wartości mocno.

Jak widać, wartości nie różnią się znacznie od siebie, jednak najlepsze rezultaty otrzymano dla 360 dni ciągu uczącego i to dla takiego modelu zostały ustalone wagi.

## Ustalenie wag dla algorytmu

Testy nad wagami dla poszczególnych algorytmów zostały przeprowadzone identycznie, jak w przypadku modelu 360 dniowego, wyniki przedstawione w tabeli:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
| Okres 1. | 8,077885 | 8,421746 | 8,102299 | 8,688477 | 8,352888 | 8,126216 | 8,28993 | 8,283676 |
| Okres 2. | 11,99966 | 11,10589 | 11,20902 | 10,58854 | 11,68472 | 11,66675 | 11,4237 | 11,44647 |
| Okres 3. | 19,39451 | 17,76339 | 18,71891 | 16,46617 | 22,38119 | 19,09355 | 18,96929 | 19,0006 |
| Średnia | 13,15735 | 12,43034 | 12,67674 | 11,9144 | 14,1396 | 12,96217 | 12,89431 | 12,91025 |

Tutaj najlepsze rezultaty otrzymaliśmy, gdy algorytmem dominującym był Vote.

# Badania nad modelem 90 dniowym

## Ustalenie liczby dni przyjętych do modelu

Badania zostały przeprowadzone dla 4 różnych liczb dni przyjętych do modelu:

* 180 dni
* 120 dni
* 90 dni
* 60 dni

Badania przeprowadzono dla 3 różnych dat, a następnie wyciągnięto średnią, wyniki przedstawiono w tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dni | Okres 1. | Okres 2. | Okres 3. | Średnia |
| 180 | 7,833818 | 11,7271 | 16,84923 | 12,13672 |
| 120 | 7,886722 | 11,80735 | 17,06287 | 12,25231 |
| 90 | 7,971124 | 11,63461 | 18,32932 | 12,64502 |
| 60 | 7,874593 | 11,60261 | 17,08189 | 12,18636 |

Najlepsze rezultaty otrzymaliśmy dla 180 dni, i to one będą testowane dalej.

## Ustalenie wag dla algorytmu

Testy nad wagami dla poszczególnych algorytmów zostały przeprowadzone identycznie, jak w przypadku modelu 360 dniowego, wyniki przedstawione w tabeli:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
| Okres 1. | 7,10644 | 8,235055 | 7,807168 | 8,155472 | 7,574808 | 7,856703 | 7,833818 | 7,744172 |
| Okres 2. | 12,56218 | 11,15549 | 11,72678 | 11,29039 | 11,25892 | 11,70873 | 11,7271 | 11,85652 |
| Okres 3. | 15,51602 | 17,54653 | 17,66461 | 16,51057 | 16,44502 | 18,37526 | 16,84923 | 16,65578 |
| Średnia | 11,72821 | 12,31236 | 12,39952 | 11,98548 | 11,75958 | 12,6469 | 12,13672 | 12,08549 |

Jak widać, najlepsze wyniki dla dominującego LinearRegression.

# Badania nad modelem 60 dniowym

## Ustalenie liczby dni

Badania zostały przeprowadzone dla 4 różnych liczb dni przyjętych do modelu:

* 120 dni
* 90 dni
* 60 dni
* 45 dni

Badania przeprowadzono dla 3 różnych dat, a następnie wyciągnięto średnią, wyniki przedstawiono w tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dni | Okres 1. | Okres 2. | Okres 3. | Średnia |
| 120 | 7,102607 | 10,23723 | 10,76133 | 9,367055 |
| 90 | 7,557578 | 9,408845 | 11,27754 | 9,414653 |
| 60 | 7,656716 | 10,41989 | 13,1254 | 10,40067 |
| 45 | 6,644687 | 11,80829 | 12,01726 | 10,15674 |

Okresem wziętym do dalszych badań było 120 dni.

## Ustalenie wag dla algorytmów

Testy nad wagami dla poszczególnych algorytmów zostały przeprowadzone identycznie, jak w przypadku modelu 360 dniowego, wyniki przedstawione w tabeli:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
| Okres 1. | 10,30694 | 7,280716 | 7,279334 | 7,240577 | 6,99743 | 7,228019 | 7,102607 | 7,245311 |
| Okres 2. | 9,39126 | 10,71315 | 10,85679 | 10,59954 | 9,043774 | 10,90605 | 10,23723 | 10,08208 |
| Okres 3. | 6,085974 | 13,76551 | 14,11408 | 13,1429 | 9,043774 | 13,9706 | 10,76133 | 9,644304 |
| Średnia | 8,594725 | 10,58646 | 10,75007 | 10,32767 | 8,36166 | 10,70156 | 9,367055 | 8,990564 |

Tym razem najlepszy okazał się algorytm z dominującym M5P.

# Badania nad modelem 30 dniowym

## Ustalenie liczby dni

Badania zostały przeprowadzone dla 4 różnych liczb dni przyjętych do modelu:

* 60 dni
* 45 dni
* 30 dni
* 20 dni

Badania przeprowadzono dla 3 różnych dat, a następnie wyciągnięto średnią, wyniki przedstawiono w tabeli:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dni | Okres 1. | Okres 2. | Okres 3. | Średnia |
| 60 | 7,919874 | 11,54583 | 17,74245 | 12,40272 |
| 45 | 8,134545 | 11,83772 | 17,05664 | 12,34297 |
| 30 | 7,998841 | 11,41849 | 19,08233 | 12,83322 |
| 20 | 7,975387 | 11,44994 | 18,63919 | 12,68817 |

Różnice były niewielkie, jednak najtrafniejszym okresem okazało się 45 dni.

## Ustalenie wag dla algorytmów

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Data | 1. | 2. | 3. | 4. | 5. | 6. | 7. | 8. |
| Okres 1. | 9,452527 | 8,005983 | 7,995217 | 7,996037 | 8,077044 | 7,956899 | 8,134545 | 8,17104 |
| Okres 2. | 11,92478 | 11,69933 | 11,82238 | 11,882 | 11,96807 | 11,80982 | 11,83772 | 11,84179 |
| Okres 3. | 13,04906 | 18,22092 | 17,99298 | 17,59038 | 13,77094 | 18,2287 | 17,05664 | 16,89286 |
| Średnia | 11,47546 | 12,64208 | 12,60353 | 12,48947 | 11,27202 | 12,66514 | 12,34297 | 12,3019 |