Using Decision Trees for Predicting Financial Markets Events

Prof.dr. Ion LUNGU
Catedra de Informatică Economică, A.S.E Bucureşti
Ec. Valentin-Dragoş MILITARU, GRIFCO S.A.

Decision trees represents one of the frequently implemented data mining techniques, perhaps also because, unlike other techniques, this one offers outputs with both predictive and descriptive potential. It thus allows complex analyses of the cause-effect relationships between various attributes. This article is an attempt to prove that, using automated knowledge aquisition techniques (supervised learning to be more precise) one can obtain performant models of predicting the events that characterise the financial markets.
Keywords: *financial markets modelling, decision trees, data mining, data classification.*

Studiul de caz analizat în acest articol a fost instrumentat cu ajutorul versiunii 3.4 a aplicației Discoverer 2000, creată şi introdusă pe piață în 1999 de către firma germană Prudential Systems Software GmbH. Deşi din punct de vedere funcțional acoperă o singură tehnică de *data mining* (arbori decizionali), Discoverer este o aplicație care prezintă un grad de flexibilitate şi o viteză de lucru suficient de ridicate pentru a se constitui într-un instrument de lucru valoros pentru analiştii de date. În plus, interfața cu utilizatorul are o pronunțată orientare vizuală, motiv pentru care şi prezentarea ce urmează va fi însoțită de un număr mare de reprezentări grafice. O deficiență de natură lingvistică a programului este aceea că, în prezent, este disponibil numai în limba germană.

Obiectiv. Prezentul studiu are ca obiectiv demonstrarea utilității arborilor decizionali în construirea de modele previzionale pentru pieţele financiare. Pentru fiecare din cele şase cursuri de schimb a căror evoluție a fost monitorizată, s-a urmărit construirea unui model care să poată oferi estimări corecte asupra sensului de variație al pieței într-o proporție semnificativ mai mare față de estimările eronate.

Datele utilizate. Pentru derularea experimentului s-au utilizat şase seturi de date, reprezentând evoluția în perioada 6-29 aprilie 2005 pentru cursurile: EURUSD, GBPUSD, USDCHF, EURGBP, EURCHF, GBPCHF. Tranzacțiile au fost agregate la fiecare 60 de secunde, rezultând următoarele date aferente prețului: *open, high, low, close*. Din aceste date s-a reținut numai valoarea prețului de închidere (*close*), toate referirile următoare la noțiunea de "curs" vizând-o exclusiv pe aceasta. Transformările aplicate datelor brute sunt descrise în secțiunea următoare.

Preprocesarea datelor. Preprocesarea datelor s-a făcut utilizând modulul *Indicator Builder* din aplicația MetaStock Professional, creată de firma Equis International. Seturile de date folosite în acest studiu au fost concepute astfel încât fiecare moment de timp (fiecare instanță din seturile de date) să poată fi descris atât prin prisma evoluției trecute a cursurilor cât şi prin prisma evoluției viitoare.

Pentru fiecare din cele şase cursuri valutare au fost calculate următoarele variabile:

i) Variabilele independente – sunt cele care reflectă informația cu privire la evoluția din trecut a cursului. Calculul variabilelor independente a fost făcut folosind o combinație de doi indicatori predefiniți în MetaStock: media mobilă simplă şi panta liniei de regresie, ambii aplicați cursului (prețul *close*) din seturile de date inițiale.

Fie *t* orizontul de timp pentru care s-au calculat variabilele independente; *v(t)* variabila independentă calculată pentru *t* momente de timp din trecut; *avg(c,t)* media simplă a cursului (*c*) din ultimele *t* momente de timp; *p(c,t)* panta liniei de regresie ce indică trendul cursului (*c*) pentru ultimele *t* momente de timp; trendul este calculat de MetaStock sub forma unei regresii liniare prin metoda celor

mai mici pătrate.

Atunci fiecare variabilă *v(t)* are formula: *v(t) = p(avg(c,t),t) × 10000* **(1)**

Datorită faptului că variațiile de la un minut la altul ale cursurilor valutare analizate sunt extrem de mici, *p(avg(c,t),t)* înregistrează valori cu ordin de mărime chiar şi de 1×10^{-6}, mai ales pentru valori mari ale lui *t*. Din acest motiv şi pentru a facilita algoritmului de explorare a datelor diferențierea instanțelor, s-a considerat necesară folosirea constantei 10.000 ca multiplicator.

Prin metoda descrisă anterior s-au construit 11 variabile independente, pe următoarele orizonturi de timp:

Variabila	Orizont de timp
v_2	2 minute
v_5	5 minute
v_10	10 minute
v_15	15 minute
v_20	20 minute
v_30	30 minute
v_60	60 minute
v_90	90 minute
v_120	120 minute
v_150	150 minute
v_240	240 minute

Variabilele independente iau valori în domeniul real şi au următoarea semnficație: dacă *v(t) > 0*, atunci curba mediei cursului calculată pentru ultimele *t* momente de timp a avut o evoluție crescătoare, iar dacă *v(t) < 0*, atunci atunci curba mediei cursului calculată pentru ultimele *t* momente de timp a avut o evoluție descrescătoare. Cu cât valorile se îndepărtează mai mult de zero, cu atât mai pronunțată a fost variația mediei în intervalul de timp studiat.

ii) Variabilele dependente – sunt cele care reflectă informația cu privire la evoluția viitoare a cursului. S-au avut în vedere trei variabile dependente, aferente a trei orizonturi diferite de timp (scurt – 5 minute, mediu – 60 de minute, lung – 240 de minute), fiecare fiind supusă separat predicției în cadrul experimentului. Fie *d(t)* variabila dependentă calculată pentru un orizont de *t* unități de timp şi fie două evenimente a căror apariție s-a dorit a fi anticipată, anume creşterea şi respectiv scăderea cursului pentru o perioadă de timp *t.* Valorile celor trei variabile dependente au fost stabilite folosind următoarea convenție de codificare:

a) Dacă se înregistrează o creştere viitoare, anume media cursului pentru următoarele *t* unități de timp este mai mare decât cursul curent, atunci *d(t) = +1*

b) Dacă se înregistrează o scădere viitoare, anume media cursului pentru următoarele *t* unități de timp este mai mică decât cursul curent, atunci *d(t) = -1*

c) Dacă nu se înregistrează nici o creştere şi nici o scădere a cursului, adică dacă media cursului pentru următoarele *t* unități de timp este egală cu valoarea curentă a cursului (cazuri extrem de rare pentru orizonturi mari de timp însă relativ frecvente pentru orizontul de timp de 5 minute), atunci *d(t) = 0.* Datorită acestei codificări, problema de previziune a evoluției cursului a fost transformată întruna de clasificare.

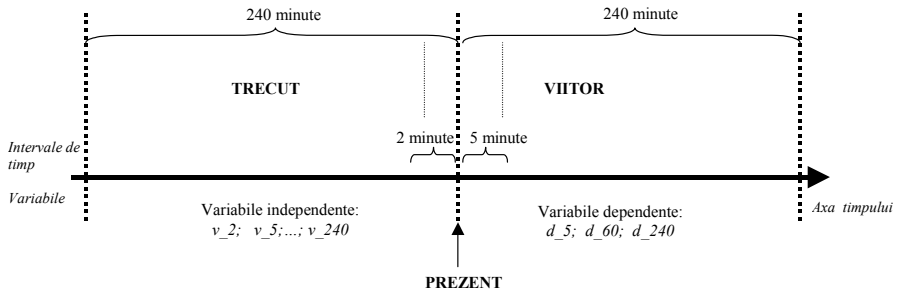

Fig. 1. Reflectarea relației dintre trecut şi viitor în variabilele independente şi dependente

Configurarea aplicației în vederea modelării

Pentru fiecare din cele şase situații analizate s-a construit câte trei modele de clasificare (arbori decizionali): unul pentru previzionarea evoluției pe termen scurt (a variabilei *d_5*), al doilea pentru previzionarea evoluției pe termen mediu (*d_60*) şi ultimul pentru previzi-

onarea evoluției pe termen lung (*d_240*). Pentru a aduce la un numitor comun cele 18 modele şi a putea efectua ulterior o analiză comparativă a lor, s-au definit condiții de lucru comune, detaliate în cele ce urmează.

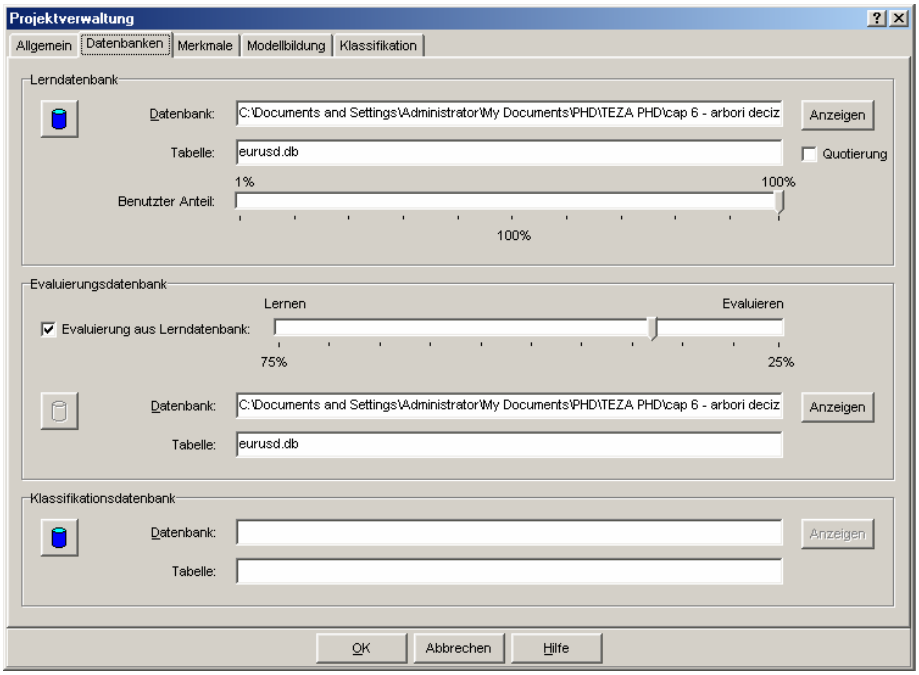

Fig. 2. Stabilirea proporţiilor de utilizare a seturilor de date în DISCOVERER – 75% pentru instruire şi 25% pentru evaluare

Seturile de date – pentru fiecare instrument valutar s-a construit un set de date conținând 25.000 de înregistrări; atributele seturilor de date au fost cele rezultate în etapa de preprocesare.

Construirea modelelor de clasificare s-a făcut utilizând primele 75% din înregistrări din fiecare set de date, iar ultimele 25% din instanțe au fost folosite pentru testarea modelelor[1]. Luând în considerare cronologia înregistrărilor şi unitatea de timp comprimată în fiecare înregistrare (1 minut), rezultă că instruirea s-a făcut folosind date din primele 18 zile din intervalul analizat (24 de zile din luna aprilie 2005), iar testarea modelelor s-a realizat pe durata ultimelor 6 zile.

Alegerea variabilelor – Variabilele dependente au fost tratate ca având valori numerice reale. Pentru toate cele 11 variabile independente utilizate[2] a fost selectată opțiunea normalizării după formula: $x = (x - m) / s$ **(2),** unde m este valoarea medie a atributului x iar s este deviația standard. Variabila independentă[3] a fost aleasă succesiv ca fiind *d_5*, *d_60* şi *d_240*. Deoarece DISCOVERER nu permite organizarea instanțelor în mai mult de două clase, au fost definite cele două situații care făceau obiectul analizei drept:

i) clasa 1: "CRESTERE" conținând instanțele pentru care variabila dependentă avea valoarea +1

ii) clasa 2: "DESCRESTERE" conținând instanțele pentru care variabila dependentă avea valoarea –1

Instanțele pentru care variabila dependentă avea valoarea zero au rămas neutilizate[4], aşa cum se poate observa şi din figura 3.

Configurarea parametrilor de instruire – Principalul parametru de construire a modelului de clasificare îl constituie funcția de separare a claselor; în cazul de față, separarea s-a decis a se face prin axe paralele, variantă mai rapidă decât funcția liniară sau polinomială. Toate celelalte setări au fost alese astfel încât să permită

[1] Germ. *Evaluierung aus Lernendatenbank*
[2] Germ. *Benutzte Merkmale*
[3] Germ. *Zielmerkmal*
[4] Germ. *Unbenutzte Ausprägung*

o dezvoltare cât mai extinsă a arborelui decizional:
- Gradul de segmentare a fost stabilit ca având valoarea 2, însemnând că dezvoltarea arborelui decizional se face astfel încât segmentele-copil să conțină cel puțin două instanțe.
- Gradul de ramificare (Germ. *Verzweigungsgrad*) a fost fixat la valoarea 10, însemnând că algoritmul poate ramifica fiecare nod în maxim 10 segmente-copil.
- Complexitatea arborelui (Germ. *Baumkomplexität*) a fost fixată la 90 (valoare relativă, pe o scală de la 0 la 100, 0 fiind echivalent cu un arbore creat în urma unei unice splitări a nodului inițial şi 100 echivalând cu un arbore pentru care toate nodurile-frunză conțin o singură instanță).
- Limitele de indexare (Germ. *Indexgrenzen*) au rămas nemodificate, la valorile implicite ale aplicației.

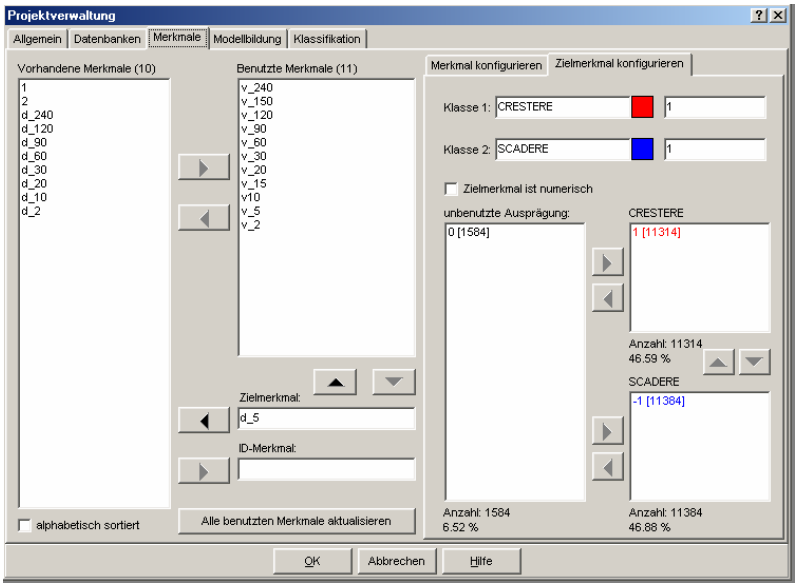

Fig. 3. Stabilirea variabilei independente şi definirea claselor

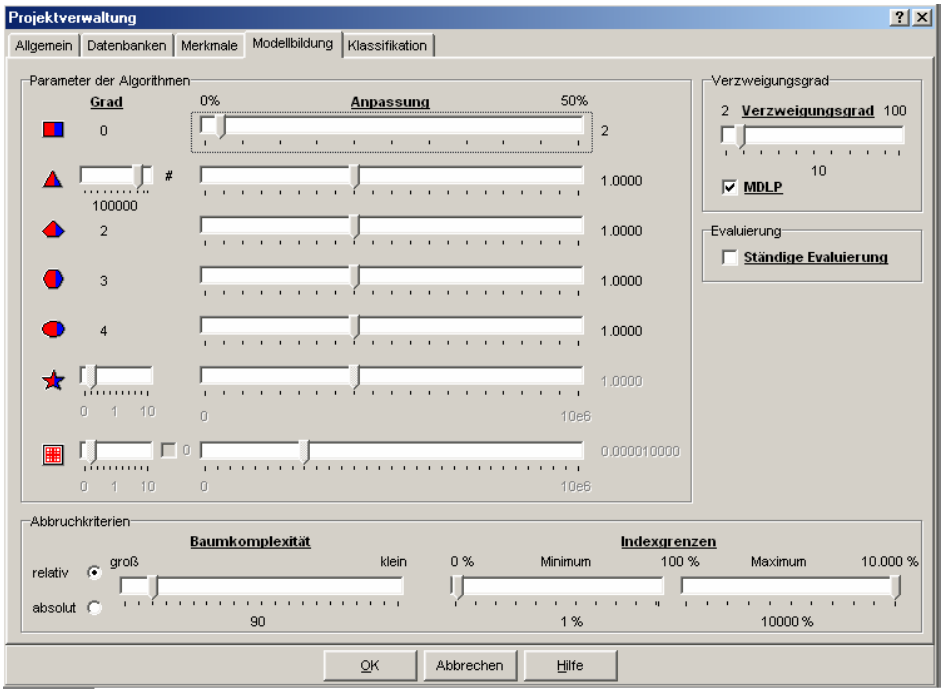

Fig. 4. Setarea parametrilor de instruire

Cu setările detaliate anterior, au fost efectuate 18 explorări ale celor 6 seturi de date, câte una pentru fiecare orizont de timp vizat. Pentru obținerea unor modele optime de clasificare, a fost utilzată următoarea matrice de costuri:

	CRESTERE	SCADERE
CRESTERE	1	-1
SCADERE	-1	1

Fiecare instanță corect clasificată a fost notată cu un punct, în timp ce fiecare instanță incorect clasificată a fost depunctată cu un punct.

Interpretarea rezultatelor

Performanțele fiecărui model au fost evaluate cu ajutorul modulului *Analyse* al aplicației DISCOVERER. Pentru fiecare instrument valutar şi pentru fiecare din cele trei variabile dependente previzionate, s-au preluat din aplicația de *data mining* matricile de analiză care prezintă situația clasificărilor pentru sub-seturile de date utilizate la crearea şi respectiv la validarea modelelor. O situație simplificată a performanțelor modelelor este centralizată în tabelul 1.

Tabelul 1 - Performanțele modelelor de clasificare

		Sub-setul de date pentru instruire			Sub-setul de date pentru test		
Instrumentul valutar	**Orizontul de timp pentru predicție**	**Total instanțe**	**Instanțe corect clasificate**	**Precizie (%)**	**Total instanțe**	**Instanțe corect clasificate**	**Precizie (%)**
EURUSD	d_5	17,023	16,990	99.81%	5,675	4,580	80.70%
	d_60	18,180	18,152	99.85%	6,063	5,640	93.02%
	d_240	18,206	18,179	99.85%	6,068	5,780	95.25%
EURCHF	d_5	16,808	16,783	99.85%	5,567	4,409	79.20%
	d_60	18,166	18,132	99.81%	6,058	5,429	89.62%
	d_240	18,203	18,176	99.85%	6,070	5,722	94.27%
USDCHF	d_5	16,989	16,951	99.78%	5,632	4,574	81.21%
	d_60	18,198	18,181	99.91%	6,067	5,687	93.74%
	d_240	18,208	18,184	99.87%	6,071	5,809	95.68%
GBPUSD	d_5	17,225	17,190	99.80%	5,757	4,661	80.96%
	d_60	18,195	18,164	99.83%	6,065	5,673	93.54%
	d_240	18,209	18,187	99.88%	6,070	5,828	96.01%
GBPCHF	d_5	17,658	17,642	99.91%	5,889	4,655	79.05%
	d_60	18,200	18,168	99.82%	6,067	5,598	92.27%
	d_240	18,209	18,191	99.90%	6,071	5,749	94.70%
EURGBP	d_5	15,486	15,466	99.87%	5,195	4,171	80.29%
	d_60	18,140	18,110	99.83%	6,051	5,566	91.98%
	d_240	18,202	18,187	99.92%	6,067	5,732	94.48%

Concluziile care se desprind din analiza tabelului 1 sunt următoarele:

(i) Ca apreciere generală, metoda utilizată a condus la anticiparea evenimentelor din cele două clase (CRESTERE respectiv SCADERE) cu un nivel de precizie bun şi foarte bun.

(ii) Calitatea predicției creşte direct proporțional cu lungimea orizontului de timp asupra căruia se efectuează analiza: în subseturile de date utilizate pentru validarea modelelor, precizia a fost de cca. 80% pentru un orizont de timp de 5 minute, cca. 90% pentru un orizont de timp de o oră (60 minute) şi de cca. 95% pentru un orizont de timp de 240 de minute.

(iii) Complexitatea modelului de clasificare scade odată cu creşterea orizontului de timp; dacă pentru clasificarea variabilei *d_5* modelul are în jur de 3.700-3.800 de noduri (ceea ce în-

seamnă tot atâtea reguli de separare), pentru *d_60* dimensiunea arborelui scade în jurul valorii de 1700-1900 de noduri, iar pentru *d_240* la 1200-1500 de noduri). Dimensiunea arborilor face ca redarea lor grafică în lucrarea de față să fie imposibilă.

Pentru comparație, în figura 5 este prezentat un arbore cu numai 86 de noduri (construit prin reducerea parametrului de complexitate de la valoarea de 90 la 35). Tot pentru comparație, arborele din figură reuşeşte să clasifice corect numai 67% din cele 6068 instanțe ale subsetului de test în cazul previziunii variabilei *d_240* pentru EURUSD.

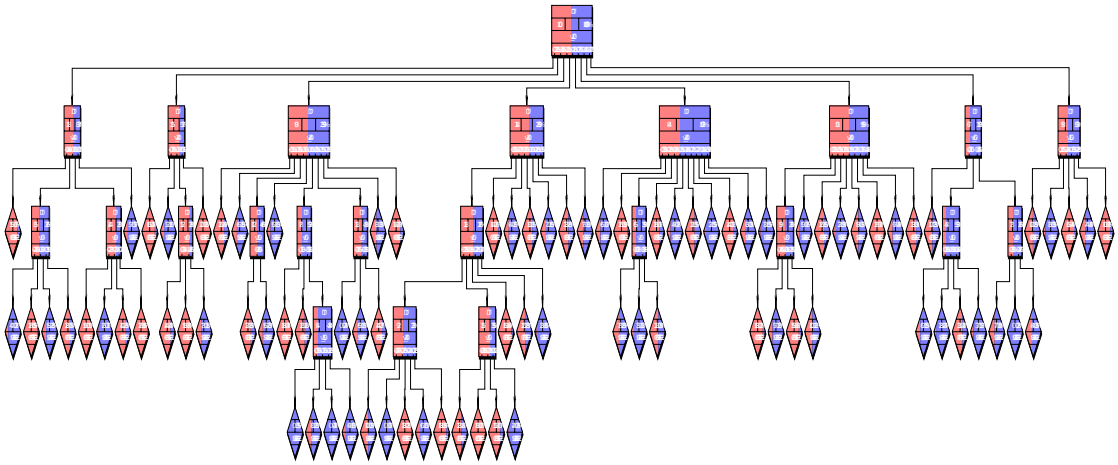

Fig. 5. Arbore decizional cu 86 de noduri şi precizie de 67%

Concluzia finală a acestui studiu este că prin utilizarea tehnologiei de explorare a datelor pusă la punct de Prudential Systems Software GmbH s-a reuşit identificarea unor modele care să previzioneze două categorii de evenimente pe piața FOREX: creşterea, respectiv scăderea cursului valutar. Valoarea experimentului prezentat aici este – cel puțin din punct de vedere teoretic – cu atât mai mare cu cât precizia predicțiilor atinge nivele care surclasează net alte rezultate raportate în articolele ştiințifice care tratează subiecte asemănătoare. Totuşi, deşi din punct de vedere teoretic valoarea experimentului apare drept evidentă, utilitatea practică a modelelor generate în acest caz de DISCOVERER este diminuată datorită absenței din modele a oricărei indicații privind amplitudinea mişcărilor pieței. În condițiile în care investitorii de pe piața reală suportă comisioane de tranzacționare, variația cursului într-o direcție sau alta devine subiect de interes abia după ce amplitudinea variației ajunge să acopere cel puțin valoarea comisionului[1]. Cu toate acestea, experimentul prezentat aici poate fi considerat un bun început pentru o cercetare mai amănunțită a potențialului de utilizare a arborilor decizionali în probleme de optimizare a deciziilor de investiții.

Bibliografie

Introduction to Data mining and Knowledge Discovery - Two Crows Corporation, Potomac (U.S.A.) 1999, ediție electronică: www.twocrows.com

Achelis, Steven B. (1995) - *Technical Analysis from A to Z*- Chicago, IL, Probus Publishing Company

Brand, Estelle; Gerritsen, Rob - *Decision Trees* - DBMS Magazine, Data Mining Solutions Supplement, Feb. 1998

Bodea, Constanța - Nicoleta (1998) - *Inteligență Artificială şi Sisteme Expert* - Editura Inforec, Bucureşti

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. *Step-by-step data mining guide*, CRISP-DM Consortium, 2000

Khabaza, Tom - *Hard Hats for Data Miners: Myths and Pitfalls of Data Mining,* DM Direct Special Report, May 3, 2005 Issue

Kimball, Ralph - *Preparing for Data Mining* - DBMS Magazine, Nov. 1997

Militaru, Valentin D.– *Studiu comparat asupra tehnicilor de data mining utilizate în rezolvarea problemelor de regresie şi clasificare* - Informatica Economica, vol. VII, Nr. 3/2003, p. 105-109

[1] Pentru tranzacțiile pe EURUSD de exemplu, există firme de intermediere care percep comisioane de 4 PIPS pentru un ciclu complet de tranzacții vânzare – cumpărare. Forex Capital Markets este un astfel de broker, specializat în tranzacții spot pe piața FOREX. (http://www.fxcm.com)