

Anexa 4

Universitatea "Politehnica" din București

Facultatea de Electronică, Telecomunicații și Tehnologia Informației

***Sistem automatizat de tranzacționare a criptomonedelor
bazat pe algoritmi de machine learning***

Lucrare de disertație

prezentată ca cerință parțială pentru obținerea titlului de

Master în domeniul inginerie electronica si telecomunicatii, programul
de studii de masterat *BIOSINF*

Conducător științific

S.L. Horia Cucu

Absolvent

Ing. Radu Adrian Codreanu

Declarație de onestitate academică

Prin prezența declar că lucrarea cu titlul Sistem automatizat de tranzacționare a criptomonedelor bazat pe algoritmi de machine learning.

, prezentată în cadrul Facultății de Electronică, Telecomunicații și Tehnologia Informației a Universității "Politehnica" din București că cerință parțială pentru obținerea titlului de Master în domeniul inginerie electronica si telecomunicatii, programul de studii BIOSINF este scrisă de mine și nu a mai fost prezentată niciodată la o facultate sau instituție de învățământ superior din țară sau străinătate.

Declar că toate sursele utilizate, inclusiv cele de pe Internet, sunt indicate în lucrare, că referințe bibliografice. Fragmentele de text din alte surse, reproduse exact, chiar și în traducere proprie din altă limbă, sunt scrise între ghilimele și fac referință la sursă. Reformularea în cuvinte proprii a textelor scrise de către alți autori face referință la sursă. Înțeleg că plagiatul constituie infracțiune și se sancționează conform legilor în vigoare.

Declar că toate rezultatele simulărilor, experimentelor și măsurărilor pe care le prezint ca fiind făcute de mine, precum și metodele prin care au fost obținute, sunt reale și provin din respectivele simulări, experimente și măsurători. Înțeleg că falsificarea datelor și rezultatelor constituie fraudă și se sancționează conform regulamentelor în vigoare.

București, *dată*

Absolvent *Prenume NUME*

(semnătura în original)

Cuprins

1. INTRODUCERE	8
2. CRIPTO MONEDE ȘI INVESTIȚIILE DIGITALE.	10
FUNȚIONAREA SISTEMELOR DE TIP BLOCKCHAIN	12
PLATFORMELE DE EXCHANGE ȘI API-UL PUBLIC	13
PIAȚA ACTIUNILOR	15
ANALIZA TEHNICA	16
INDICATORI TEHNICI.....	19
ANALIZA FUNDAMENTALA	21
TEORII LEGATE DE PROBLEMA PREZICERII PIETELOR	22
IPOTEZA PIEȚEI EFICIENTE	22
IPOTEZA MERSULUI ALEATOR.....	23
IPOTEZA STRATEGIILOR AUTO DESTRUCTIVE	24
CONCLUZII	24
3. MACHINE LEARNING.....	24
MODELELE UTILIZATE	25
<i>Rețeaua neuronală simplă</i>	25
<i>Rețeaua neuronală recurentă</i>	26
<i>Rețeaua neuronală convolutională</i>	28
<i>Regresia liniară</i>	29
4. SISTEM AUTOMATIZAT DE TRANZACȚIONARE	30
ALEGEREA BAZEI DE DATE	33
PRELUAREA BAZEI DE DATE	34
LIMITĂRI ALE BAZELOR DE DATE.....	35
TOOL-URI SOFTWARE ȘI BIBLIOTECI UTILIZATE	35
<i>Tensorflow</i>	35
<i>Biblioteci C++</i>	35
<i>Biblioteci Python</i>	36
SEPARAREA DATELOR DE ANTRENARE ȘI TESTARE	36
ANTRENAREA MODELELOR.....	38
<i>Parametri inițiali de antrenare pentru fiecare model</i>	38
<i>Antrenarea și salvarea datelor rezultate</i>	39
REZULTATE EXPERIMENTALE	41
<i>Rezultate obținute pentru primul set de antrenare</i>	41

SISTEMUL AUTOMATIZAT DE TRANZACTIONARE	65
Limbajul de programare.....	65
Diagrama inițială de functionare.....	65
Perioada de cercetare	70
Arhitectura finală și flow-ul de execuție.....	73
Scalare și îmbunătățiri viitoare	76
CONCLUZII	77
BIBLIOGRAFIE	78

FIGURI

FIGURA 1 BLOCURILE UNUI BLOCKCHAIN	11
FIGURA 2 DIAGRAMĂ BLOCKCHAIN	11
FIGURA 3 TOTALUL CAPITALIZĂRII CRYPTO MONEDELOR 2013-2018	13
FIGURA 4 INTERFAȚĂ EXCHANGE-ULUI BITTREX	14
FIGURA 5 GRAFIC AL PREȚULUI ACȚIUNILOR TESLA	16
FIGURA 6 CANDLES (LUMÂNĂRI)	17
FIGURA 7 GRAFIC REPREZENTAT PRIN LUMÂNĂRI ȘI REPREZENTAREA VOLUMULUI.....	17
FIGURA 8 ORDER BOOK.....	18
FIGURA 9 MACD.....	19
FIGURA 10 BOLLINGER BANDS	19
FIGURA 11 RSI	20
FIGURA 12 O SERIE DE NUMERE GENERATE ALEATOR	23
FIGURA 13 EXEMPLU DE BUCLĂ ÎNTR-O REȚEA NEURONALĂ RECURENTĂ.....	26
FIGURA 14 REȚEA NEURONALĂ RECURENTĂ REPREZENTATĂ SECVENTIAL	27
FIGURA 15 LSTM	27
FIGURA 16 GRU	28
FIGURA 17 EXEMPLU DE MODEL DE REGRESIE LINIARĂ PENTRU DETERMINAREA TRENDULUI.....	29
FIGURA 18 SISTEM AUTOMATIZAT DE TRANZACTIONARE	30
FIGURA 19 ARHITECTURA UNUI SISTEM AUTOMATIZAT DE TRANZACTIONARE	31
FIGURA 20 SET ANTRENARE.....	36
FIGURA 21 SET TESTARE	37
FIGURA 22 SET ANTRENARE.....	37
FIGURA 23 SET TESTARE	38
FIGURA 24 EVALUARE RNN / CNN.....	42
FIGURA 25 EVALUARE / ITERATII	43
FIGURA 26 EVALUARE / ITERATII	44
FIGURA 27 ACURATEȚE / SEMN	47
FIGURA 28 ACURATEȚE / SEMN	48
FIGURA 29 ACURATEȚE / SEMN	49
FIGURA 30 ACURATEȚE / SEMN	49
FIGURA 31 ACURATEȚE / SEMN	50
FIGURA 32 EVALUARE / ITERATII.....	52

FIGURA 33 EVALUARE / ITERATII.....	52
FIGURA 34 ACURATEȚE SEMN	56
FIGURA 35 ACURATEȚE SEMN	56
FIGURA 36 ACURATEȚE SEMN	57
FIGURA 37 ACURATEȚE SEMN	57
FIGURA 38 ACURATEȚE SEMN	58
FIGURA 39 PREDICȚIE / REAL	59
FIGURA 40 PREDICȚIE / REAL - ZOOM	59
FIGURA 41 PREDICȚIE / REAL OFFSET	60
FIGURA 42 PREDICȚIE / REAL OFFSET -1	60
FIGURA 43 CROSS CORRELATION (PREZIS / REAL).....	61
FIGURA 44 VALOARE STAȚIONARĂ / NESTATIONARA	61
FIGURA 45 100 ZILE (PREDICȚIE / REAL).....	63
FIGURA 46 RNN / PREDICȚIE 5 ZILE	64
FIGURA 47 RNN / PREDICȚIE 5 ZILE	64
FIGURA 48 ARHITECTURĂ SISTEM.....	67
FIGURA 49 MACD - TEMA	71
FIGURA 50 MACD - TRIX	71
FIGURA 51 MACD - MFI.....	72
FIGURA 52 FEAR GREED INDEX	72
FIGURA 53 ARHITECTURĂ SISTEM.....	74
FIGURA 54 DIAGRAMĂ DE FUNCTIONARE.....	75
FIGURA 55 FUNCȚIONAREA APLICATIEI	76

TABELE

TABEL 1 DATE RETURNATE DE EXCHANGE (PER ZI)	34
TABEL 2 REȚEA NEURONALĂ SIMLPA / REGRESIA LINIARA.....	41
TABEL 3 RNN / CNN.....	42
TABEL 4 NN / LR.....	51
TABEL 5 RNN / CNN.....	51

1.Introducere

Analiză și prezicerea prețurilor din piața de acțiuni a fost un subiect de cercetare atât în industria financiară cât și în industria software. Motivația monetară din spatele prezicerii vânzării sau cumpărării de acțiuni la poziții profitabile este un foarte important conducător în cercetarea acestui domeniu. În mod tradițional metodele utilizate pentru evaluarea acțiunilor și generarea de semnale de tranzacționare sunt împărțite în două mari categorii:

- Analiză fundamentală
- Analiza tehnică

Dar recent, analiza sentimentului a contribuit considerabil la generarea de semnale de tranzacționare.

Analiză fundamentală implică studierea informațiilor fundamentale a unei companii, cum ar fi, politica companiei, știri despre companie, profit anual, etc.

Analiza tehnică pe de altă parte, evaluează mișcarea prețului acțiunii și volumului de tranzacționare. Persoanele care abordează această metodă de analiză consideră că toate informațiile necesare, inclusiv informațiile fundamentale sunt incluse în istoricul prețului. Deși analiza tehnică obișnuia să fie calitativă în natura sa - cu scopul de a găsi regularități prin extragerea diferitelor tipuri de șabloane din date cu zgomot și examinând vizual grafice ale mișcării prețului unei acțiuni, tehnologia modernă din prezent și mai exact algoritmi de machine learning au transformat această metodologie într-o unealtă mai mult cantitativă. Studiile arată că 80-90% din investitori și profesioniști din acest domeniu se bazează pe cel puțin un tip de analiză tehnică.

Analiza sentimentului, care implică cuantificarea reacției pieței la diferite evenimente și noutăți a dovedit de curând valoare predictivă.

O combinație între indicatori de analiză tehnică și indicatori de sentiment al pieței a fost folosită în această lucrare.

Datorită gradului ridicat de competitivitate din acest spațiu există și foarte puține lucrări științifice publicate, neexistând motive pentru a publica un sistem profitabil. Datorită câștigurilor monetare foarte ridicate din vânzarea acestor tipuri de sisteme, există și o stigmă negativă atribuită domeniului în comunitatea academică. Metodele tradiționale de predicție a prețurilor sunt de asemenea și rareori profitabile în mod consistent, datorită schimbării condițiilor pieței.

Piața crypto monedelor este o piață relativ nouă, aceasta a prins popularitate foarte recent motiv pentru care încă nu există bază tehnică adânc cercetată în acest domeniu. Foarte mulți analiști însă compara piața crypto monedelor cu piața acțiunilor, similaritățile fiind evidente.

Datorită creșterii rapide în capitalizare și volum, recent foarte multe companii private și foarte mulți investitori utilizează tehnici de analiză moștenite din piața acțiunilor. Existența exchange-urilor online și accesului foarte ușor la acestea pentru aproape orice persoană, face că crearea și utilizarea unui sistem automatizat de tranzacționare să fie ceva cu adevărat fezabil.

Într-adevăr, piețele nu rămân stabile, în special piața crypto monedelor - fiind renumită pentru volatilitate - iar indicatorii care au valoare predictivă pe o anumită perioadă pot înceta din a mai genera venituri când se schimbă diferiți factori din piață.

În această lucrare vor fi abordate aspectele generale ale pieteleor economice, aspecte tehnologice legate de crypto monede și piața crypto monedelor, vor fi analizate metodele de analiză tehnică din spațiul tranzacționării de acțiuni, vor fi studiați și analizați diferiți indicatori tehnici, pentru a putea evalua posibilitatea utilizării lor în piața crypto monedelor. De asemenea vor fi prezentate aspecte importante ale studiului economic realizat cu algoritmi de machine learning, utilitatea și fezabilitatea utilizării diferitelor tipuri de rețele neuronale pentru obiective diferite.

2. Cripto monede și investițiile digitale.

Cripto monedele reprezintă un mediu alternativ de transfer al unor valori digitale. Esența fiecărei monede cripto se află în fundația sa criptografică. Tranzacțiile între nodurile rețelei sunt realizate prin criptografie într-o rețea decentralizată și securizată. Cripto monedele reprezintă astfel date dintr-o bază de date pe care nici o entitate nu o poate modifica fără a respecta anumite condiții.

În trecut s-a încercat crearea unor monede digitale (în anii 90') folosind diferite sisteme ca Flooz, Beenz, DigiCash - dar toate acestea au fost un eșec. În principiu eșecul se datorează problemelor de fraudă sau financiare între diferite companii.

Toate aceste sisteme folosesc ceea ce se numește "utilizarea unei terțe părți" - însemnând că tranzacțiile vor fi facilitate și verificate de o companie. Datorită eșecurilor unor astfel de companii crearea unei monede digitale nu a putut deveni realizabilă pentru o perioadă lungă de timp.

În anul 2009, un programator anonim (sau un grup de programatori) - sub aliasul **Satoshi Nakamoto** a introdus **Bitcoin**. Satoshi a descris acest sistem ca fiind : "peer to peer electronic cash" . Este complet decentralizat, însemnând că nu există nici un server central care să controleze autoritatea.

O foarte importantă problemă pe care un sistem de plată online trebuie să o rezolve, este problema dublei cheltuieli - (double spending) - tehnică fraudulenta prin care se încearcă utilizarea aceleași sume de două ori. Sistemele tradiționale folosesc un server central care deține o bază de date ce înregistrează toate tranzacțiile. Această soluție însă, implică existența unei autorități care poate controla toate fondurile participanților din rețea.

Într-o rețea descentralizată ca Bitcoin, fiecare participant se va ocupa de păstrarea și scrierea în această bază de date. Acest lucru se realizează prin tehnologia **Blockchain**

Un Blockchain reprezintă un șir de blocuri de date care înregistrează tranzacțiile într-un ledger digital, într-o manieră descentralizată.



Figura 1 Blocurile unui blockchain

Blockchain-ul este un ledger **distribuit** - deschis pentru oricine dorește să participe în rețea, datele înregistrate fiind **nealterabile** (datele sunt toate tranzacțiile realizate în rețea)

O tranzacție reprezintă transferul de valoare de la o entitate la alta. În mod usual o tranzacție conține informații legate de entitatea care trimite, entitatea care primește, și valoarea trimisă.

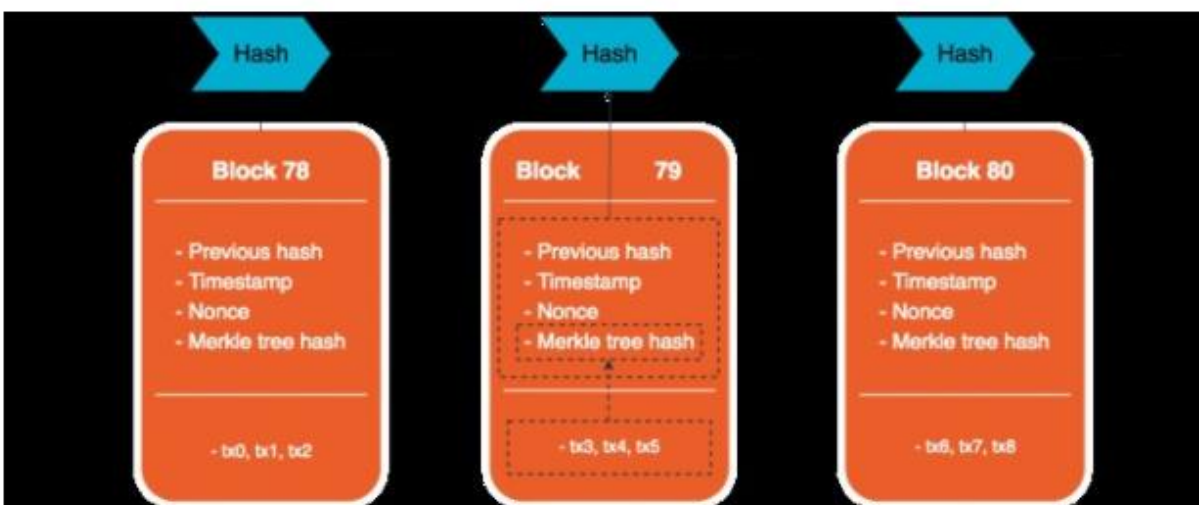


Figura 2 Diagramă blockchain

În imaginea prezentată mai sus se poate observa ce conțin exact blocurile care compun blockchain-ul. După cum se poate observa, fiecare bloc conține un număr de tranzacții, un hash unic, dar de asemenea și hash-ul blocului precedent, astfel creând un lanț (chain). Fiecare bloc conține un checksum al datelor sale proprii, care este apoi pus în header, astfel ajungând în următorul bloc și astfel se păstrează integritatea datelor.

Funcționarea sistemelor de tip Blockchain

Când o tranzacție este creată (conținând expeditorul, datele și recipientul) trebuie trimisă către rețeaua blockchain.

Din păcate tranzacțiile nu sunt executate direct ci trebuie să aștepte ceea ce se numește - procesarea. Aceste tranzacții sunt validate pentru a putea fi sumarizate în blocuri și atașate în istorie.

Este nevoie de validare pentru a ne asigura că participanții nu pretind că au o valoare mai mare decât au cu adevărat (evitând astfel problema double spending).

Pentru că nu există o variabilă pe care să ne arate valoarea expeditorului, această valoare trebuie să fie calculată din întreaga istorie a blockchainului. Istoria ne arată originea valorii respective și traseul său prin rețea.

Minerii (nodurile) sunt entitățile care mențin blockchain-ul funcțional prin validarea tranzacțiilor. Pentru a procesa tranzacțiile și a le adăuga în istoria blocurilor, aceste blocuri necesită rezolvarea unei probleme matematice complexe.

Această metodă de validare se numește Proof of Work (PoW) deoarece dacă o entitate reușește să rezolve această problemă matematică, a dovedit că a realizat muncă. În realitate această muncă este realizată prin computații complexe și hashing - care în esență necesită curent electric.

Minerii sunt recompensați cu un anumit număr de monede pentru fiecare rezolvare a unei astfel de probleme, astfel aceștia sunt plătiți pentru munca lor. Dificultatea de calcul crește în timp, astfel devenind din ce în ce mai greu pentru mineri să găsească blocuri (pentru Bitcoin - la fiecare 2016 blocuri).

Datorită popularității acestei tehnologii, au apărut foarte multe alternative la Bitcoin - denumite în general - alt coins.

Creșterea în valoare a acestor monede în ultimii ani a atras din ce în ce mai mulți participanți. Vânzarea și cumpărarea acestor bunuri digitale se realizează pe diferite platforme online (atât descentralizate cât și centralizate) denumite "crypto exchanges".



Figura 3 Totalul capitalizării crypto monedelor 2013-2018

Tranzacționarea acestor monede se realizează în general asemănător cu tranzacționarea în piața acțiunilor.

Platformele de Exchange și API-ul public

Fiind valori digitale care pot fi deținute de oricine pe glob, tranzacționarea acestor monede se realizează în mod uzual pe exchange-uri online. Aceste platforme de exchange pun la dispoziție utilizatorilor un cadru organizat de tranzacționare, viteze ridicate, și capacitate de volum foarte mare. Datorită faptului că majoritatea exchange-urilor cu volum mare sunt centralizate, acestea aduc performanțe ridicate în ceea ce privește viteza de procesare.

Un exchange pentru crypto monede pune la dispoziție investitorilor o serie de unelte de bază pentru a facilita tranzacțiile realizate, un număr ridicat de opțiuni de crypto monede (existând exchange-uri cu sute de monede tranzacționabile), dar și un API public, pentru a permite investitorilor posibilitatea de a crea sisteme automatizate de tranzacționare.

Interfața grafică a unui exchange arată că în figură prezentată mai jos:



Figura 4 Interfață exchange-ului Bittrex

În mod uzual investitorii tranzacționează folosind interfața grafică a exchange-ului.

Majoritatea exchange-urilor însă pun la dispoziție un API public. Acestea sunt în general API-uri în format JSON.

Conectarea la un API se face prin simpla accesare a unei interfețe pazate pe URL printr-o conexiune de tip HTTPS. Accesare URL-ului va returna un răspuns în format JSON care va conține informațiile căutate.

În aplicația prezentată în această lucrare, s-a realizat un wrapper în C++ pentru a îngloba toate funcțiile API puse la dispoziție de exchange-ul de crypto-monedă Bittrex.

Astfel, toate funcțiile API au fost asociate unui set de funcții în aplicația C++, aceste funcții pot fi apelate în timp real iar informațiile returnate sunt apoi parsate și păstrate în diferite variabile.

A fost de asemenea utilizată și o aplicație de scraping pentru colectarea tuturor datelor istorice pentru toate monedele puse la dispoziție pe acest exchange. Date care ulterior vor fi folosite pentru antrenarea diferitelor modele.

Piața acțiunilor

Piața acțiunilor (Stock market) - se referă colecția piețelor și a exchange-urilor unde se realizează tranzacționarea acțiunilor de la companiile publice sau alte clase de securități. Această piață reprezintă una dintre cele mai vitale componente ale economiei pieței libere. Aduce companiilor acces la capital în schimbul unei anumite părți date investitorilor.

Această piață este formată din 2 secții: Piața principală și piața secundară. Piața principală este aceea unde noi acțiuni sunt vândute pentru prima oară în ceea ce se numește IPO (Initial Public Offering). Investitorii instituționali de obicei cumpără majoritatea acestor acțiuni.

Astfel valoarea companiei devine publică și numărul de acțiuni eliberat va determina prețul de deschidere al acțiunii în IPO. Toate tranzacțiile ulterioare se realizează în piața secundară.

În piața secundară se află atât investitori instituționali cât și investitori individuali. O companie folosește banii strânși în IPO pentru a crește, dar o dată ce acțiunile încep să fie tranzacționate, aceasta nu mai primește fonduri din vânzarea sau cumpărarea acțiunilor sale.

Acțiunile companiilor mai mari sunt în general tranzacționate pe platforme de exchange. Astfel de exchange-uri există în marile orașe ale lumii, incluzând: Londra, Tokyo, etc. Exchange-urile sunt etichetate care adună atât vânzători cât și cumpărători într-o manieră organizată. Acțiunile sunt listate și tranzacționate. Execuția tranzacțiilor se realizează electronic. Aceste tranzacții ale acțiunilor determină prețul pentru fiecare moment de timp (prețul la care s-a realizat ultima tranzacție reprezintă valoarea acelei acțiuni). Un grafic al acțiunii prețului unei astfel de acțiuni se poate observa în imaginea prezentată mai jos.



Figura 5 Grafic al prețului acțiunilor TESLA [9]

În imaginea prezentată mai sus se pot observa toate datele disponibile în fiecare moment de timp despre acțiunea prețului.

Analiza tehnică

Fiecare fereastră de timp este determinată de ceea ce se numește o "lumânare" - această fereastră conține 5 parametri:

- Prețul de deschidere (valoarea inițială la intrarea în fereastră de timp - lumânare)
- Prețul de închidere (ultima valoare a prețului în momentul trecerii în următoarea fereastră de timp)
- Prețul maxim (valoarea maximă a prețului din acea fereastră de timp)
- Prețul minim (valoarea minimă a prețului din acea fereastră de timp)
- Culoarea (verde/alb - dacă prețul de închidere este mai mare ca cel de deschidere sau roșu/negru în caz contrar)

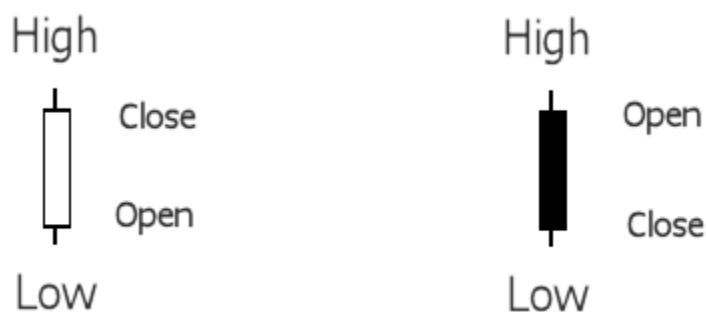


Figura 6 Candles (lumânări) [9]

De asemenea în fiecare fereastră de timp se înregistrează și volumul - cantitatea totală tranzacționată în acea perioadă de timp.

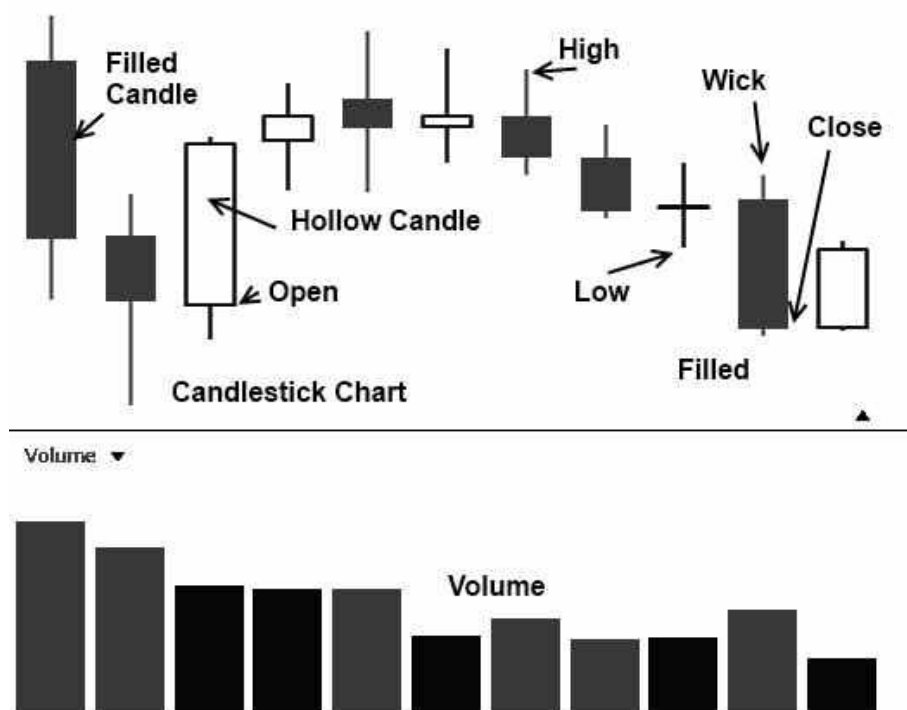


Figura 7 Grafic reprezentat prin lumânări și reprezentarea volumului [9]

De asemenea, platformele de exchange pun la dispoziție vizualizarea oridenelor de cumpărare și a ordinelor de vânzare - Order book.

▼ ORDER BOOK BTC/USD

COUNT	PRICE	AMOUNT	TOTAL	TOTAL	AMOUNT	PRICE	COUNT
5	2,810	983	983	46	46	2,820	39
12	2,800	2,919	3,903	109	62	2,830	28
12	2,790	33	3,937	200	90	2,840	21
6	2,780	16	3,953	237	37	2,850	22
11	2,770	28	3,982	283	45	2,860	24
10	2,760	50	4,033	698	415	2,870	26
16	2,750	99	4,132	844	145	2,880	37
10	2,740	19	4,152	980	135	2,890	64
18	2,730	141	4,293	1,139	158	2,900	88
15	2,720	14	4,308	1,261	122	2,910	43
36	2,710	31	4,339	1,328	66	2,920	39
72	2,700	187	4,526	1,429	101	2,930	38
21	2,690	51	4,578	1,508	79	2,940	33
33	2,680	54	4,633	1,661	152	2,950	66
27	2,670	30	4,663	1,789	128	2,960	26
33	2,660	73	4,737	1,860	70	2,970	18
56	2,650	230	4,967	1,921	60	2,980	24
19	2,640	22	4,990	2,034	113	2,990	29

Figura 8 Order book

Se pot extrage informații utile chiar și din numărul sau valoarea ordinelor.

În continuarea acestei lucrări cei care realizează tranzacții pe platforme de exchange vor fi denumiți generic "traderi".

Traderii în general folosesc diferite tehnici de tranzacționare pentru a realiza profit - încercând să speculeze dacă prețul va crește sau va scădea.

În literatura de specialitate s-a demonstrat că există o corelare foarte strânsă între emoțiile trader-ului și acțiunile pe care acesta le va face în funcție de acțiunea prețului. Astfel prețul este strâns corelat de sentimentul general al pieței.

Prețul unei acțiuni poate de asemenea să fie influențat de diferite evenimente din cadrul companiei. Aceste evenimente pot afecta în mod negativ sau pozitiv prețul acțiunilor (spre exemplu - o companie cu acțiuni publice anunță un parteneriat cu o companie foarte cunoscută, o astfel de știre poate influența pozitiv prețul acțiunii).

În general Traderii folosesc algoritmi personali bazați pe diferiți indicatori economici care încerca să determine caracteristicile cele mai importante ale pieței la fiecare moment de timp.

Indicatori tehnici

Indicatorii economici reprezintă în general calcule matematice bazate pe preț, volum sau alte valori publice ale unei acțiuni. Analizând date istorice, analiștii folosesc indicatori pentru a prezice viitoarele mișcări în prețuri. Exemple de astfel de indicatori sunt: MACD, Bollinger Bands, RSI, etc.



Figura 9 MACD [9]

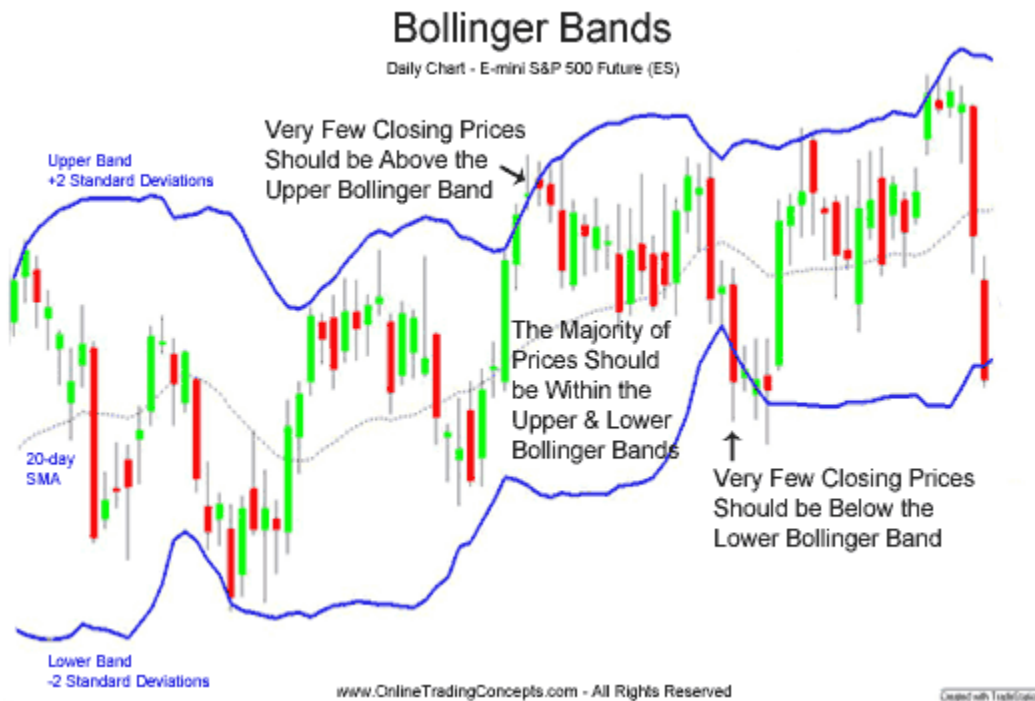


Figura 10 Bollinger bands [9]



Figura 11 RSI [9]

Acești indicatori în general pentru determinarea prețului pe un termen scurt de timp. Sunt foarte utilizați de traderi, aceștia folosind diferite strategii personale pentru a maximiza profitul și a minimiza riscul.

Există două tipuri de astfel de indicatori:

- Suprapuși - indicatori care folosesc aceeași scală ca prețul și sunt afișați peste graficul acțiunii. (Bollinger Bands)
- Oscilatori - indicatori care oscilează între un minim local și un maxim, sunt afișate în mod usual sub graficul acțiunii.

Fiecare trader își creează un set de astfel de indicatori pentru a își crea propria strategie. De asemenea astfel de indicatori sunt în general integrați în sisteme automate de tranzacționare.

Analiză fundamentală

Când vine vorba de acțiuni sau alte instrumente economice asemănătoare (criptomonedele), analiză fundamentală încearcă să determine valoarea prin observarea factorilor de fundal care afectează afacerea unei companii și succesul ei. Analiză fundamentală se poate aplica, mai larg, unei industrii sau economiei ca întreg.

Analiză fundamentală încearcă să răspundă la întrebări ca:

- Face profit compania?
- Este într-o poziție suficient de puternică încât să învingă viitori competitori?
- Își poate plăti datoriile?
- Este compania bine manageriată?

În esență toate aceste întrebări duc către o întrebare de ansamblu - Merită această companie o investiție în acțiunile ei?

Fundamente: Cantitativ și calitativ

Factorii fundamentali se pot împărți în două mari categorii:

- Cantitativi - capabili de a fi măsurați sau exprimați în numere
- Calitativi - legat de sau bazat pe calitatea unui anumit lucru, de multe ori în opoziție cu cantitatea. Spre deosebire de factorii cantitativi, factorii calitativi sunt mai puțin tangibili.

Conceptul de valoare intrinsecă

Una dintre cele mai importante asumări ale analizei fundamentale constă în faptul că prețul de piață al unei acțiuni / criptomonede nu reflectă valoarea sa "reală". În jargonul financiar, această valoare "reală" este cunoscută ca valoare intrinsecă. [10]

Fiecare trader în parte, bazat pe analiza sa fundamentală proprie și cercetarea asupra companiei sau monedei respective, își poate crea o estimare a valorii intrinseci. Bazat pe această valoare, un trader urmărește să cumpere acea acțiune sau moneda când prețul de piață este mult sub valoarea sa intrinsecă.

Toată această analiză se bazează pe faptul că, după o durată de timp, prețul va ajunge la valoarea sa intrinsecă.

Există de asemenea câteva necunoscute:

- Trader-ul nu poate ști dacă estimarea sa a valorii intrinseci este corectă
- Trader-ul nu poate ști când prețul de piață a acțiunii/monedei va atinge valoarea intrinsecă

Cele mai mare critici legate de analiză fundamentală vin în principal din două mari grupuri:

- Cei care propun analiza tehnică
- Cei care cred în ipoteza pieței eficiente

Teorii legate de problema prezicerii piețelor

Există câteva teorii din economia financiară care susțin că demersul în crearea unui predictor de piață este inutil, vor fi prezentate cele mai mari dintre acestea, iar pe parcursul lucrării vom analiza rezultate experimentale în conformitate cu informațiile din aceste ipoteze.

Ipoteza pieței eficiente

Efficient-market hypothesis (EMF) este o teorie în economia financiară care spune că prețurile unui bun reflectă în totalitate toată informația disponibilă. O implicare directă a acestei teorii este că este imposibil să "învingi piață" în mod consistent, deoarece prețul pieței poate reacționa doar la informație nouă. [10]

Există trei variante a acestei ipoteze:

- Weak (slabă): Informațiile de tranzacționare (nivele și schimbări în prețuri și volume) a bunurilor tranzacționate (acțiuni, proprietăți, etc.) sunt deja încorporate în prețuri. Dacă această formă de eficiență este adevărată atunci analiză tehnică este inutilă pentru a genera profit.
- Semi-strong (semi-puternică): Prețurile conțin toate informațiile (conține chiar și informațiile din statuturile financiare, date de la SEC, etc.). Dacă această formă a ipotezei este adevărată atunci ambele tipuri de analiză (analiza tehnică și analiză fundamentală) sunt inutile în generarea de profit. (în acest caz doar informații "din interior" pot fi utile în generarea de profit - insider trading)
- Strong (puternică): Prețurile conțin absolut toate informațiile (atât cele publice cât și cele non-publice (din interior) - insider) și astfel nici măcar cei din interior nu pot aștepta profit superior față de publicul neinformați.

Ipoteza mersului aleator

Această teorie financiară spune că prețurile pieței evoluează conform unui mers aleator (astfel, schimbările de preț sunt aleatoare), astfel acestea nu pot fi prezise. Această teorie este consistentă cu EMF (Efficient-Market hypothesis).

Burton G. Malkiel, un profesor de economie de la Universitatea Princeton, și autorul cărții: *A Random Walk Down Wall Street*, a realizat un test unde studenții au primit un stock ipotetic cu valoare inițială de 50\$. Prețul de închidere a fiecărei zile a fost determinat de o aruncare de monedă. Dacă rezultatul era cap - prețul se închidea cu jumătate de punct mai sus, dacă rezultatul era pajura - prețul se închidea cu jumătate de punct mai jos. Astfel de fiecare dată, prețul avea 50% șansă să se închidă mai jos sau mai sus de ziua anterioară. Ciclii și trenduri au fost determinate din teste. Malkiel a pus apoi rezultatele într-un grafic și le-a dat mai departe unui trader (chartist) - care interpretează pattern-uri din trecut pentru a face o predicție în viitor, pe baza faptului că "istoria tinde să se repete". Trader-ul i-a spus lui Malkiel că trebuie neapărat să cumpere acel stock. Din moment ce aruncările de moneda sunt complet aleatorii, stockul fictiv nu avea nici un trend. Malkiel spune că acest lucru indică faptul că prețul piețelor poate fi atât de aleator că o aruncare de monedă.

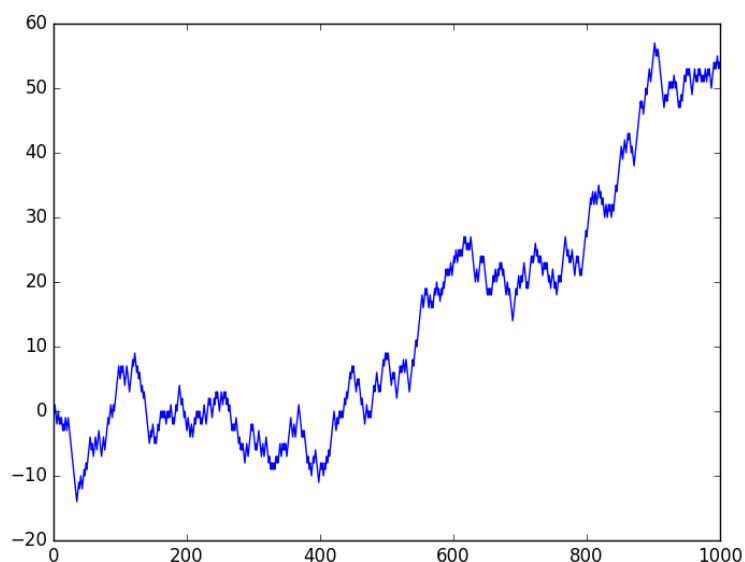


Figura 12 O serie de numere generate aleator

Se poate observa în graficul de mai sus, cum o serie de numere generată aleator poate semăna cu prețul unei acțiuni/ criptomonedă.

Ipoteza strategiilor auto destructive

Această ipoteză nu se bazează pe cele anterioare, ba chiar dimpotrivă, asumându-ne că teoriile anterioare sunt false - se poate afirma că putem prezice prețuri viitoare pe baza informațiilor din prezent și trecut folosind o strategie algoritmică.

Ipoteza spune că un model de succes poate duce către propria distrugere. Dacă ar exista un model simplu de predicție care poate fi aplicat de oricine, atunci, în timp, tot avantajul va fi tranzacționat și erodat.

Acesta este și unul din motivele pentru care nu există multe lucrări academice pe acest subiect. Dacă un model de succes ar fi făcut public, acesta într-un timp nu foarte lung nu va mai avea succes.

Concluzii

Aceste teorii ne învață că trebuie să păstrăm un grad relativ de scepticism relativ la rezultatele obținute. Realizarea unui model de succes, profitabil, se poate dovedi a fi foarte dificilă.

3. Machine learning

Recent se pune din ce în ce mai multă importanță pe utilizarea algoritmilor de machine learning în diferite domenii, printre care, și cel economic. Cercetarea curentă în domeniu încearcă găsirea celor mai bune metode de predicție a prețurilor sau recunoașterea de pattern-uri. Majoritatea trader-ilor din ziua de astăzi folosesc sisteme inteligente de tranzacționare care îi ajută în prezicerea prețurilor, astfel ajutându-I în luarea unor decizii de investiție.

Având în vedere explicațiile anterioare cu privire la metodele de analiză a piețelor (analiză fundamentală și analiză tehnică) lucrarea se va înclina către analiză tehnică, având cel mai mult sens - aceasta fiind bazată pe pattern-uri și variații a prețurilor din trecut.

Cu toate acestea, există sisteme care încearcă predicții bazate pe analiza fundamentală, și chiar utilizează ambele metode de analiză, într-un sistem hibrid, pentru a putea face o predicție. În această lucrare însă se va analiza performanța algoritmilor bazată pe analiza tehnică.

Datorită recente creșteri în popularitate și a numărului relativ mic de lucrări științifice în acest sens, s-a hotărât experimentarea folosind rețele neuronale - se vor realiza experimente folosind rețele neuronale simple, recurente, convolutionale și pentru comparație se va utiliza și regresia liniară. Se vor parcurge în continuare modelele menționate și principalele motive din spatele alegerii acestora.

Modelele utilizate

În vederea realizării unui experiment cât mai riguros, și datorită lipsei de experimente publicate în acest sens, se va încerca experimentarea, pe același set de date cu mai multe tipuri de modele pentru realizarea predicției prețurilor viitoare. Rezultatele experimentale vor fi apoi analizate pentru a putea extrage o concluzie.

Rețeaua neuronală simplă

O problemă importantă în domeniul machine learning este cum se poate face un algoritm care se va comporta bine atât pe datele de antrenare dar și la intrări noi. Foarte multe strategii în machine learning sunt construite astfel încât să reducă eroarea de testare dar cu costul unei erori mai mari de antrenare. Aceste strategii sunt cunoscute ca regularizări. Un regularizator eficient este unul care reduce semnificativ varianța, dar nu mărește foarte mult bias-ul.

Regularizarea este folosită de foarte mult timp (cu mult înainte de deep learning). Modelele liniare cum ar fi regresia liniară permit strategii de regularizare simple și eficiente (după cum vom observa în subcapitolele ulterioare)

Majoritatea regularizărilor sunt bazate pe limitarea capacității modelelelor prin adăugarea unui parametru de penalizare $\omega(\theta)$ funcției obiectiv J^* :

$$J^*(\theta; X, y) = J(\theta; X, y) + \alpha\omega(\theta)$$

Unde α este un hiper parametru care ponderează contribuția relativă a parametrului ω relativ la funcția obiectiv standard J .

Având $\alpha=0$ rezultă în eliminarea regularizării.

Unul dintre cei mai comuni parametri de penalizare este L2, numit de asemenea și weight decay.

Această strategie de regularizare aduce ponderile mai aproape de origine prin adăugarea unui termen de regularizare $\omega(\theta) = \frac{1}{2}\|w\|^2$

Acest tip de regularizare încearcă reducerea erorii pătratice medii prin adăugarea de bias, dar în același timp, reducerea semnificativă a variantei, astfel se evită fenomenul de over-fitting.

De asemenea, se va utiliza pentru comparație și regularizarea de tip L1 (regresie lasou) - diferența dintre cele două tipuri de regularizare constă în parametrul de penalizare, în loc de ridicarea la pătrat a coeficientului se folosește valoarea absolută.

Diferența cea mai importantă dintre cele două constă în faptul că L1 micșorează coeficientul trăsăturilor mai puțin importante până la zero, astfel, eliminând complet unele trăsături.

Rețeaua neuronală recurentă

Acesta este un tip de rețea neuronală unde conexiunile dintr-un noduri formează un graf direcționat de-a lungul unei secvențe temporale. Acest lucru îi permite să aibă un caracter dinamic. Rețelele recurente (prescurtate de acum RNN) își pot folosi memoria internă pentru a procesa secvențe de intrare (o buclă). Una dintre cele mai importante caracteristici ale acestui tip de rețea este că pot fi folosite în conectarea informațiilor din trecut cu informațiile din prezent. Este evident din aceste caracteristici că acest tip de rețea neuronală poate reprezenta un candidat bun pentru predicția prețurilor de piață, având în vedere caracterul secvențial al datelor.

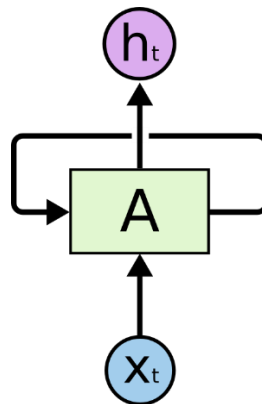


Figura 13 Exemplu de buclă într-o rețea neuronală recurentă [7]

Buclă permite trecerea informației dintr-un pas în altul al rețelei. O astfel de rețea se poate reprezenta și secvențial astfel:

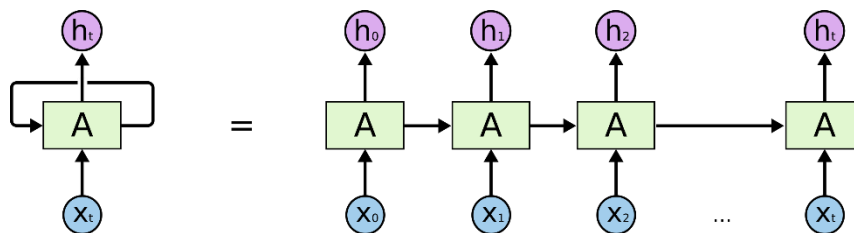


Figura 14 Rețea neuronală recurentă reprezentată secvențial [7]

Se poate observa chiar din modul acesta de reprezentare că este cel mai potrivit model de rețea pentru a fi folosită la date cu caracter secvențial.

RNN-urile cele mai utilizate (datorită eficienței lor ridicate) în special în această arie de cercetare sunt rețelele de tip LSTM și cele de tip GRU (variație de LSTM).

Rețelele de tip LSTM au caracteristică importantă de a învăța dependențe pe o perioadă mare de timp.

Ca și rețele neuronale clasice, acestea au o structură repetitivă, dar blocul repetitiv este diferit. În loc de utilizarea unui singur strat, se folosesc 4 straturi.

LSTM are abilitatea de a adăuga sau elimina informație stării celulei, această caracteristică este regularizată de structuri denumite porți (gates) - acestea sunt compuse dintr-un strat sigmoid și o operație de multiplicare. Stratul sigmoid scoate la ieșire numere între 0 și 1, astfel determinând cât din fiecare componentă va trece mai departe - într-un modul LSTM există 3 astfel de porți.

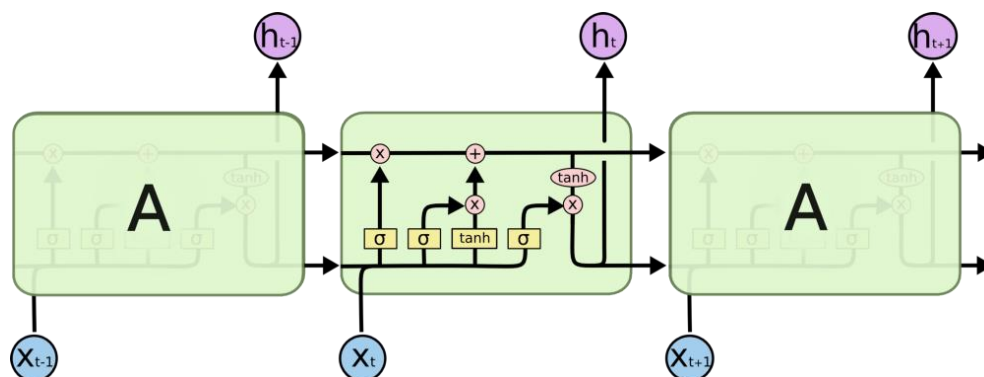


Figura 15 LSTM [7]

GRU (Gated Recurrent Unit) este o variație populară a rețelelor LSTM, introdusă de Cho, et al. (2014). Această unește starea celulei cu starea ascunsă, combină porțile de uitare și porțile de intrare într-o singură poartă, modelul rezultat este mai simplu decât modelul standard LSTM, după cum se poate observa în imaginea de mai jos:

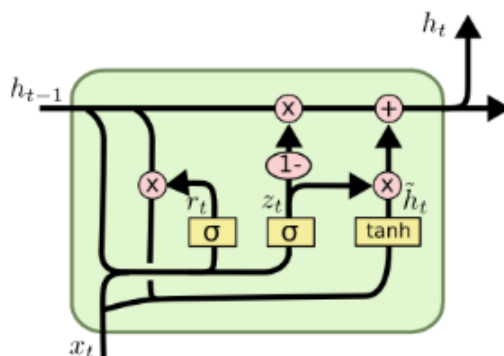


Figura 16 GRU [7]

După cum se poate observa în imaginile prezentate mai sus [7] rețelele de acest tip realizează o procesare secvențială în timp, astfel informația trebuie să parcurgă toate celulele până să ajungă la celulă curentă. Acest lucru înseamnă că informația poate fi foarte ușor coruptă datorită multiplicării de multe ori cu numere foarte mici (<0). Aceasta lucru cauzează fenomenul de dispariție a gradientilor (vanishing gradients). În vederea rezolvării acestor probleme se utilizează module LSTM, astfel se poate trece peste unele unități și astfel rețeaua poate reține pe o perioadă mai lungă de timp - se elimină o parte din problema de dispariție a gradientilor, însă nu în totalitate, deoarece încă există o cale secvențială de la celule mai vechi la celulă curentă. Din această cauză, astfel de rețele neuronale nu pot reține decât secvențe de ordinul sutelor.

Rețeaua neuronală convolutională

Rețelele neuronale convolutionale sunt utilizate în mod uzual în aplicații ce necesită recunoaștere de pattern-uri pentru date aranjate în matrici bidimensionale. Din această motiv ele sunt foarte populare în aplicații de imagistică digitală, reprezentând o arie foarte amplă de cercetare.

Motivul pentru care am ales acest tip de rețea neuronală ca posibil candidat pentru predicția de prețuri este capacitatea rețelelor convolutionale de a găsi pattern-uri locale, deși această aplicație are datele sub forma unei serii temporale.

O rețea neuronală recurentă poate folosi doar date din trecut, cât timp o rețea convolutională se poate "uita" și în viitor.

În general rețelele convolutionale sunt mult mai bune la găsirea de pattern-uri locale deoarece principala asumare a rețelelor de acest tip este că un astfel de pattern rămâne relevant peste tot.

De asemenea, rețelele convolutionale sunt foarte bune pentru extracția de trăsături, din același motiv prezentat mai sus.

După cum am prezentat în capitolul Analiză tehnică, foarte mulți traderi își bazează predicțiile pe pattern-uri bine determinate, acestea rămânând consistente în permanență.

Rezultatele obținute folosind acest model pot demonstra dacă această ramură a analizei tehnice poate sau nu să reprezente un bun indicator pentru o predicție.

Regresia liniară

Aceasta încearcă să modeleze relația dintre două variabile prin potrivirea unei ecuații liniare cu datele observate. Una din aceste date reprezintă caracteristica urmărită, iar cea de a doua variabilă, este variabilă de dependență. În cazul nostru prețul și timpul. În mod evident înainte de a încerca această potrivire, trebuie determinat dacă există o relație între cele două variabile de interes. Potrivit bazelor analizei tehnice se poate deduce pe baza acestora că există o corelație între timp și preț.

Acest model se poate considera un candidat bun pentru realizarea de predicții datorită capacității sale de a determina ceea ce economiștii numesc "trend".

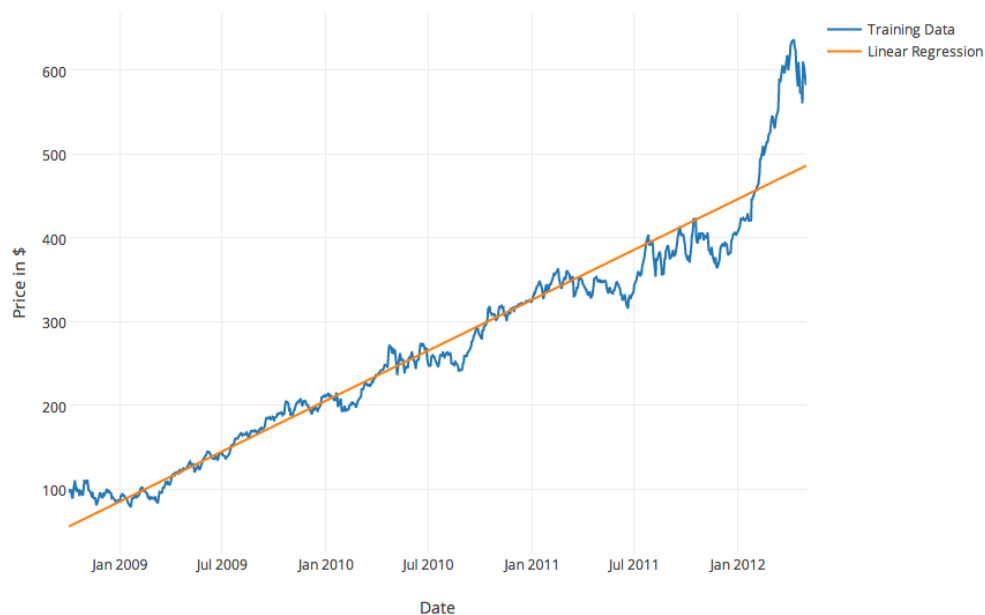


Figura 17 Exemplu de model de regresie liniară pentru determinarea trendului

Având la dispoziție aceste modele, în continuare se vor realiza experimente și comparații pe diferite seturi de date, în vederea construirii unui sistem automatizat de tranzacționare.

4. Sistem automatizat de tranzacționare

ATS (Automated trading systems) reprezintă aplicații software care creează ordine de vânzare și cumpărare în mod automat. Acest program va decide în mod automat când trebuie să pună un ordin de cumpărare sau un ordin de vânzare. Decizia sistemului este bazată pe o anumită strategie de tranzacționare. Strategiile sunt în general determinate de analiză tehnică a pieței (prezentat în subcapitolul anterior). Flow-ul de date dintre exchange și sistem este bidirecțional, după cum se poate observa în imaginea de mai jos:

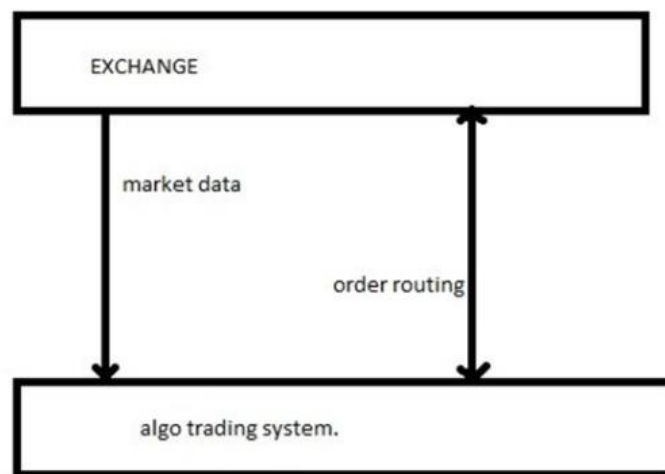


Figura 18 Sistem automatizat de tranzacționare

Arhitectura generală a unui astfel de sistem se poate observa în imaginea de mai jos

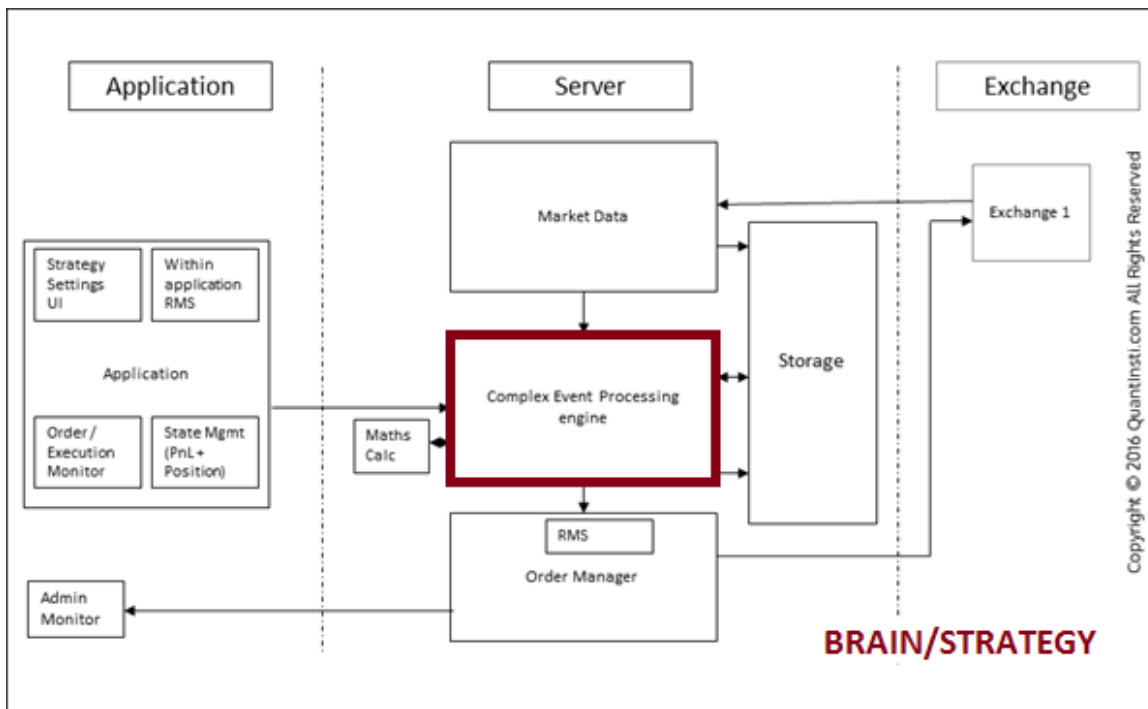


Figura 19 Arhitectura unui sistem automatizat de tranzacționare

Un astfel de sistem este împărțit în mai multe nivele (nivelul aplicației - nivelul de server - și nivelul exchange-ului) - mulți layered system

În primul nivel (Application)

- În acest nivel aplicația software reprezintă interfața om-masina:
 - o Primește input care va decide strategia
 - o Afișează date importante - execuția ordinelor, statusul mașinii, etc.

Nivelul serverului

- În acest nivel se adună datele importante de la platforma de exchange
- Datele sunt procesate în funcție de strategia aleasă
- Se realizează decizia de plasare a ordinelor, trimisă mai departe către exchange

Nivelul exchange-ului

- În general platformele de exchange pun la dispoziție un API public prin care utilizatorii pot apela funcții predefinite pentru a accesa datele publice de pe platformă
- De asemenea pot fi apelate funcții predefinite pentru a plasa sau anula ordine de vânzare/cumpărare.

Avantajele utilizării unui sistem automatizat de tranzacționare

- **Minimizarea emoțiilor.** Sistemele automatizate de tranzacționare minimizează emoțiile în timpul tranzacționării. Prin acest lucru, persoanele care tranzacționează au posibilitatea de a urma un anumit plan. Datorită faptului că ordinele se setează în mod automat, utilizatorul nu va avea posibilitatea să ezite în executarea tranzacției.
- **Consistența.** una dintre cele mai mari provocări în tranzacționare este *să planifici o tranzacție și să tranzacționezi planul*. Chiar dacă planul de tranzacționare are posibilitatea să fie profitabil, utilizatorii care nu respectă regulile alterează orice așteptare ar fi avut sistemul. Nu există un plan de tranzacționare care să câștige 100% din timp, pierderile sunt o parte din sistem. Dar aceste pierderi pot fi traumatizante din punct de vedere psihologic, astfel un trader care are două sau trei tranzacții nereușite poate decide să nu mai facă încă o tranzacție. Dacă următoarea tranzacție ar fi fost profitabilă, trader-ul deja a distrus orice așteptare ar fi avut sistemul.
- **Viteza ridicată** de setare a ordinelor. Din moment ce mașinile pot răspunde imediat la schimbarea condițiilor de piață, sistemele automatizate sunt capabile să genereze ordine imediat ce un anumit criteriu este îndeplinit. Intrarea sau ieșirea dintr-o tranzacție chiar și într-un interval de câteva secunde poate face o diferență mare în profitabilitatea tranzacției. Imediat cum o poziție a fost determinată, toate celelalte tranzacții sunt automat generate, incluzând prețuri de protecție pentru minimizarea pierderilor, etc.
- **Cantitatea de informație**, sistemele automatizate de tranzacționare pot lua în considerare o cantitate uriașă de informație înainte de setarea unei anumite poziții. Diferite modele de inteligență artificială pot lucra în același timp pentru minimizarea riscului și maximizarea profitului. Foarte multe date pot fi procesate într-o perioadă foarte scurtă de timp.

Dezavantajele utilizării unui astfel de sistem:

- **Eșecuri tehnice** - Teoria din spatele tranzacționării automatizate este simplă: Crearea unui software, programarea unor reguli, apoi utilizarea sa pentru tranzacționare. În realitate, însă, sistemele automatizate sunt sofisticate însă nu infailibile. Depinzând de structura sistemului sau a aplicației (client - server) pot exista defecțiuni în rețea ce

pot altera capabilitatea mașinii de a procesa corect datele sau chiar de a eșua în setarea ordinelor, orice alt tip de eroare din punct de vedere tehnic, erori care nu pot fi prevăzute anterior, pot determina pierderi financiare majore.

- **Monitorizarea** - Deși ar fi ideal să putem lăsa o mașină să execute tranzacții în mod nesupervizat, acest lucru poate duce la pierderi majore. Orice defecțiune sau eroare de sistem (software bugs) care pot genera comportamente neașteptate ale sistemului, poate însemna diferența între profit și pierdere. Din această cauză, majoritatea sistemelor automatizate necesită monitorizare atentă.
- **Optimizarea** - după cum am menționat în partea introductivă, condițiile pieței se schimbă, mai ales în piața crypto monedelor, adaptarea unor modele la schimbări radicale poate fi un lucru imposibil, astfel, existența unui sistem perfect optim este ceva imposibil. Optimizarea sistemelor actuale de tranzacționare automatizată se realizează constant, aceste sisteme necesită cercetare constantă și re-evaluare a modului de funcționare relativ la condițiile pieței.

Alegerea bazei de date

Datorită rarității lucrărilor academice pe acest subiect, nu există (sau sunt foarte rare) lucrări care să utilizeze baze de date standardizate pentru efectuarea experimentelor, majoritatea folosind pur și simplu date istorice preluate de la diferite exchange-uri.

În aceste lucrări am ales ca bază de date, datele istorice ale monedei Bitcoin, aceasta fiind prima crypto moneda tranzacționată, are de asemenea cel mai mare volum și cea mai mare capitalizare, această alegere a fost făcută în vederea obținerii celor mai consistente rezultate și reducerea erorilor de metodologie.

Piața crypto monedelor fiind încă relativ nouă, multe dintre acestea sunt tranzacționate pe exchange-uri cu lichiditate foarte scăzută, astfel prețul poate fi foarte ușor manipulat de diferite entități din piață.

Bitcoin însă este moneda cu cea mai mare lichiditate și suferă cel mai puțin de manipulare rău voită în piață, deși chiar pentru aceasta există uneori semne evidente de manipulare. Aceste creșteri sau scăderi de preț realizate intenționat de o singură entitate pot induce evident erori în analiza noastră, trebuie să avem în vedere și acest aspect.

De asemenea în această piață, există foarte multe exchange-uri online care permit tranzacționarea de crypto monede. Multe dintre acestea însă suferă de probleme mari de

lichiditate. Vom extrage datele istorice, pentru compunerea bazei de date, de la exchange-urile cu cele mai mari valori de lichiditate publicate.

O altă problemă în această piață o reprezintă faptul că multe exchange-uri fac publice valori de volum false, pentru a atrage tranzacții. Un studiu recent efectuat, face evident acest lucru, și de asemenea, arată care dintre exchange-urile analizate au cu adevărat volumul publicat corect.

Astfel, se va utiliza că baza de date, datele istorice de la moneda Bitcoin, preluate de pe Binance, un exchange cu lichiditate mare și dovedit că fiind valid.

Preluarea bazei de date

Pentru realizarea mai multor tipuri de experimente se vor prelua mai multe tipuri de baze de date, variind fereastră de timp a unei lumânări (candle).

Se vor prelua astfel baze de date cu fereastră de timp de 2 ore și o zi. Ideea principală a acestei lucrări se va baza pe predicția cu fereastră de o zi, dar vom analiza rezultatele modelelor și pe celelaltă baza de date pentru a observa dacă există o corelație între aceste tipuri de date.

Datele sunt preluate direct de la exchange folosind API-ul public al acestuia, exchange-ul returnează datele în format JSON. Datele vor fi apoi transformate în format csv pentru o mai ușoară utilizare a acestora.

În tabelul de mai jos se pot observa datele returnate:

date	high	low	open	close	volume	quoteVolume	weightedAverage
1.42E+09	244	225	225	244	46.27631	0.193117	239.6278
1.42E+09	245	240.25	240.2501	240.25	55.8949	0.230429	242.5685
1.42E+09	245	245	245	245	14.72224	0.060091	245
1.42E+09	249	235	245	235	129.1212	0.539055	239.5326
1.42E+09	235.001	235	235	235	96.56756	0.410926	235.0001

Tabel 1 Date returnate de exchange (per zi)

Un alt aspect foarte important ce trebuie avut în vedere este modul în care vrem să folosim datele pentru antrenare. În loc de valoarea prețului, se va folosi **diferența procentuală de la prețul anterior** - pentru **stationarizarea** datelor. Acest aspect este legat de modul în care o rețea neuronală va interpreta datele și va fi explicat în capitolele următoare.

Limitări ale bazelor de date

După cum am precizat anterior, nu vom folosi doar datele per zi ci și per 2 ore, acest aspect este foarte important deoarece ne poate mări rezoluția temporală și se poate dovedi util pentru realizarea de tranzacții în timpul unei zile (intra day). În general exchange-urile de acțiuni nu pun la dispoziția publicului valori istorice pe ferestre de timp mai mici de o zi (acestea fiind disponibile doar contra cost). Însa în piață crypto monedelor aceste date sunt publice chiar și până la ferestre de 5 minute. Se va putea realiza în viitor o analiză mai complexă chiar și pentru ferestre de timp și mai mici.

De asemenea, în datele istorice lipsesc date despre order book (Fig. 7). Aceste date pot fi utile în predicția prețurilor deoarece pot reda informații relevante despre evoluția prețurilor pe perioade mici de timp. Poate chiar există o corelație strânsă între magnitudinea order book -ului și variația prețurilor pe o fereastră scurtă de timp (Exemplu: Un order book încărcat cu foarte multe ordine de cumpărare la un moment de timp și foarte puțin ordine de vânzare la același moment de timp, poate fi indicator al unei creșteri a prețului în viitorul apropiat).

Tool-uri software și biblioteci utilizate

Tensorflow

Această bibliotecă se va utiliza pentru crearea unei rețele neuronale, permite realizarea cu ușurință a unui model de rețea neuronală, antrenarea, testarea dar și servirea de predicții pentru acestea. Folosește python pentru a aduce un front-end API ușor de utilizat, și execută aplicațiile în high-performance C++. Se vor crea multiple modele de rețele neuronale.

Biblioteci C++

Acestea se vor folosi pentru construcția sistemului final de tranzacționare, realizând conectarea la exchange și execuția task-urilor.

Biblioteci Python

Acestea vor fi utilizate pentru colectarea datelor de la exchange, construirea bazelor de date de antrenare/testare, integrarea cu tensorflow pentru antrenarea și rularea diferitelor modele de rețele neuronale, cros-validare cât și calcularea rezultatelor finale ale modelelor.

Separarea datelor de antrenare și testare

Având în vedere caracterul volatil al pieței crypto monedelor, este foarte importantă determinarea seturilor de antrenare și testare, datele fiind deasemenea și limitate. Se va avea în vedere faptul că în 2017-2018 a avut loc o creștere masivă a capitalizării pieței crypto monedelor, de aceea se poate dovedi util să împărțim experimentul în două sub-seturi:

- Antrenare 2013-2017, testare 2017 - prezent (~ 70% - 30%)



Figura 20 Set antrenare

Sistem automatizat de tranzacționare a criptomonedelor bazat pe algoritmi de machine learning



Figura 21 Set testare

- Antrenare 2013-2018, testare 2018 - prezent (~ 85% - 15%)



Figura 22 Set antrenare



Figura 23 Set testare

Se vor utiliza datele BTC / USDT - perioadă eșantion de 1 zi (datorită motivațiilor din capitolele anterioare) .

Antrenarea Modelelor

Fiecare model din cele prezentate anterior (regresie liniară, rețea neuronală simplă, rețea recurentă, rețea convolutională) va fi antrenat iterativ folosind hiper parametri aleatori. Unele modele vor folosi parametri aleatori dintr-un set dat (datorită caracteristicilor rețelei).

Parametri inițiali de antrenare pentru fiecare model

- Regresie liniară:
 - o Fereastra de analiză (k) : 1 - 25

- Rețea neuronală simplă:
 - o Fereastra de analiză (k): 1 - 25
 - o Batch size: ([500, 1000, 2000, 4000])
 - o Nr. Epoci: 100

- Funcții cost: L1, L2
- Lambda: $10 \cdot \text{rând}(-8, -2)$
- Rețea neuronală recurentă:
 - Fereastra de analiză (k) : ([24, 32, 48, 64])
 - Batch size: ([1000, 2000, 4000])
 - Tip de celulă: LSTM, GRU
 - Nr epoci: 100
 - Dropout: (0, 1)
 - Funcții de cost: L1, L2
 - Lambda: $10 \cdot \text{rând}(-10, -6)$
- Rețea neuronală convolutională:
 - Fereastra de analiză (k) : ([24, 32, 48, 64])
 - Batch size: ([1000, 2000, 4000])
 - Nr. Epoci: 150
 - Dropout: (0, 1)
 - Funcții de cost: L1, L2
 - Lambda: $10 \cdot \text{rând}(-10, -6)$

Parametrii inițiali au fost aleși în funcție de cele mai uzuale valori utilizate în scopul predicției pe bază de date de tip serie sau ajustate pe baza rezultatelor experimentale.

Antrenarea și salvarea datelor rezultate

Antrenarea se realizează folosind tensorflow (toate modelele prezentate mai sus, sunt modele standard de rețele neuronale create în tensorflow), se iterează prin fiecare model folosind setul de antrenare determinat anterior, și se utilizează parametri aleatori din intervalele setate. După fiecare iterație (în funcție de numărul de epoci - în cazul rețelelor neuronale) se realizează și testarea, folosind setul de testare determinat anterior. Rezultatele sunt stocate în următorul format:

```
# Train results:
Mean absolute error: 0.027572
SD absolute error: 0.033307
Sign accuracy: 0.565259
Mean squared error: 0.001870
Sqrt of MSE: 0.043239
Mean error: 0.000302
```

```
# Test results:  
Mean absolute error: 0.024018  
SD absolute error: 0.024877  
Sign accuracy: 0.552573  
Mean squared error: 0.001196  
Sqrt of MSE: 0.034579  
Mean error: 0.000244
```

Din exemplul prezentat mai sus se pot observa valorile urmărite pentru caracterizarea fiecărui model în parte:

Toate aceste valori se calculează pentru tot șirul de valori prezise și reale.

Eroarea absolută medie: 0.024018 - arată că modelul este la 2.4018% distanță de valoarea reală procentuală de schimbare a prețului. - $\text{mean}(\text{abs}(\text{predicted} - \text{truth}))$

Deviația standard a erorii absolute : 0.024877 - $\text{std}(\text{abs}(\text{predicted} - \text{truth}))$

Acuratețea semnului: 0.552573 - arată de câte ori se realizează o predicție corectă a semnului (creșterea sau scăderea prețului) - 55.2573% - $\text{mean}(\text{equal}(\text{sign}(\text{prediction}), \text{sign}(\text{truth})))$

Eroarea medie - 0.00024 - $\text{mean}(\text{predicted} - \text{truth})$

În final se calculează o valoare de evaluare a modelului: eroarea absolută medie + eroarea absolută a deviației standard. O valoare mică reprezintă un model bun.

Antrenarea se face într-o buclă infinită, datorită faptului că există un număr mare de combinații posibile de hiper parametri, este foarte important că antrenarea să se desfășoare pe o perioadă cât mai mare de timp.

Rezultate experimentale

Rezultate obținute pentru primul set de antrenare

Rețeaua neuronală simplă și regresia liniară:

k	NN eval		LR eval
1	0.0508		0.0506
2	0.0521		0.0519
3	0.0494		0.0501
4	0.0547		0.0535
5	0.0527		0.0538
6	0.0515		0.0524
7	0.0528		0.0511
8	0.0491		0.0553
9	0.051		0.0605
10	0.0519		0.0633
11	0.0528		0.91
12	0.0543		0.0581
13	0.0512		1.129
14	0.0501		0.0601
15	0.0526		3.75
16	0.052		2.09
17	0.0529		1.15
18	0.052		2.064
19	0.0521		1.598
20	0.0517		2.85

Tabel 2 Rețea neuronală simplă / regresia liniară

În tabelul de mai sus, se poate observa efectul modificării ferestrei de analiză "k" asupra scorului de evaluare a celor două modele.

Cele mai bune scoruri au fost obținute la k=8 pentru NN și k=3 pentru LR. Aceste valori au fost obținute după o serie de iterații asupra parametrilor fiecărui model.

În ambele cazuri, se obțin valori bune ale scorului (cât mai aproape de 0) însă în cazul regresiei liniare se observa erori foarte mari pentru o parte din modele în cazul ferestrelor mari de analiză, lucru care era de așteptat având în vedere variațiile prețului pe perioade mai mari.

Rețeaua neuronală recurentă și Rețeaua neuronală convolutională:

k	RNN eval	CNN eval
24	0.05	0.0561
32	0.051	0.092
48	0.0509	0.2368
64	0.0489	0.514

Tabel 3 RNN / CNN

Din tabelul de mai sus se pot observa scorurile de evaluare pentru cele două modele, în funcție de dimensiunea ferestrei de analiză.

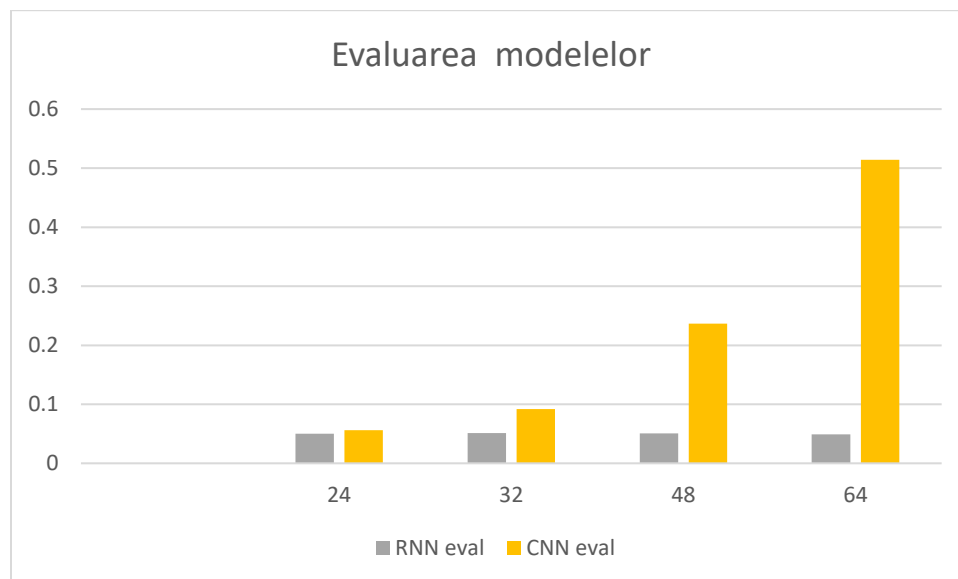


Figura 24 Evaluare RNN / CNN

Se poate observa din aceste valori că rețeaua neuronală recurentă are rezultate din ce în ce mai bune cu creșterea ferestrei de analiză. În timp ce rețeaua neuronală convolutională are rezultate din ce în ce mai rele.

O cauză posibilă a acestui efect (relativ la rețeaua convolutională) poate fi dificultatea în descoperirea pattern-urilor locale din setul de date, în mod evident, creșterea ferestrei de analiză

va determina rezultate din ce în ce mai eronate datorită dimensiunii mici a setului de date de antrenare.

Folosind valorile ferestrelor pentru cele mai bune rezultate a fiecărei rețele, se va observa parcurgerea hiper parametrilor în vederea găsirii unui model cât mai bun.

Pentru rețeaua neuronală recurentă:

În graficul de mai jos se pot observa rezultatele pentru antrenarea cu $k = 64$

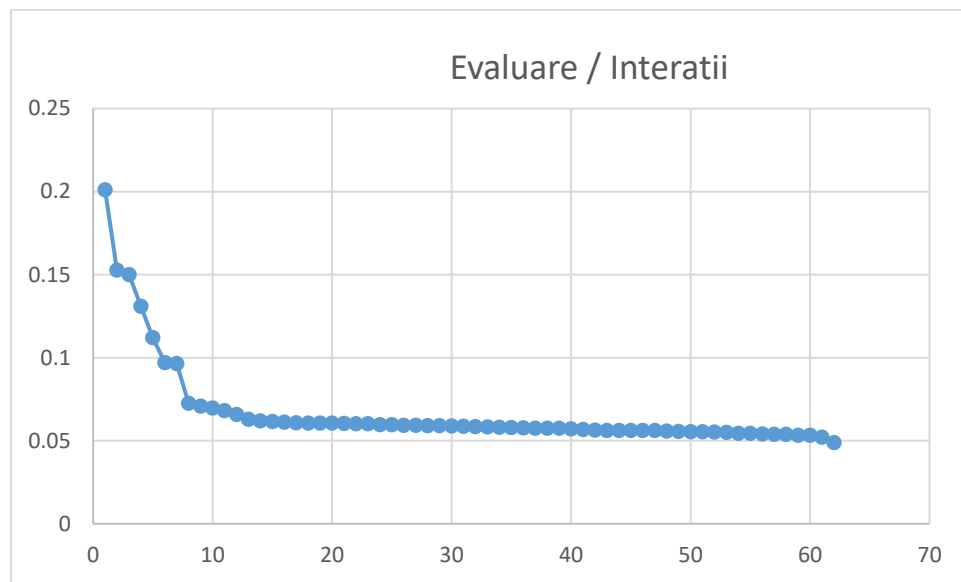


Figura 25 Evaluare / iterații

Se poate observa îmbunătățirea rezultatelor datorită numărului ridicat de iterații prin hiper parametri.

Pentru rețeaua neuronală convolutională:

În graficul de mai jos se pot observa rezultatele pentru antrenarea cu $k = 24$

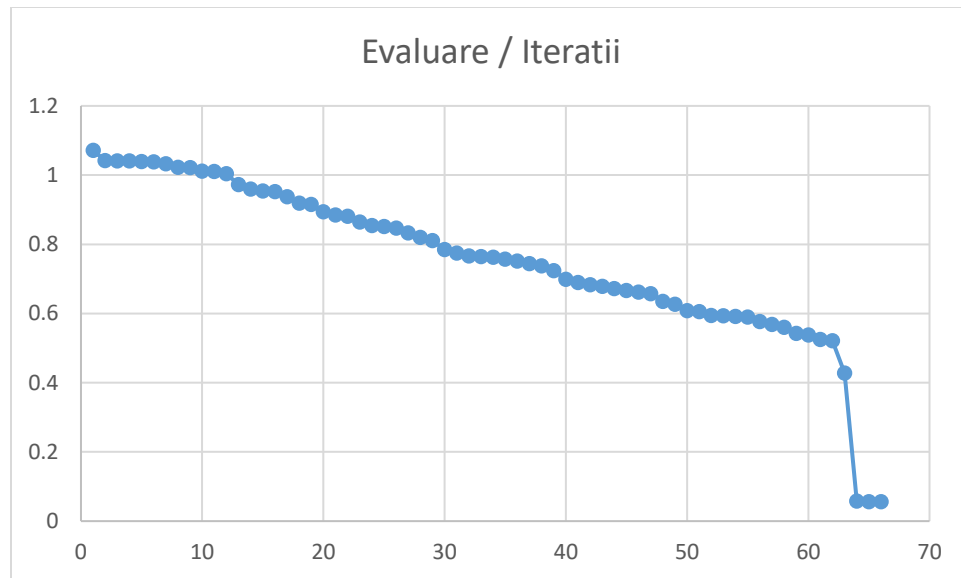


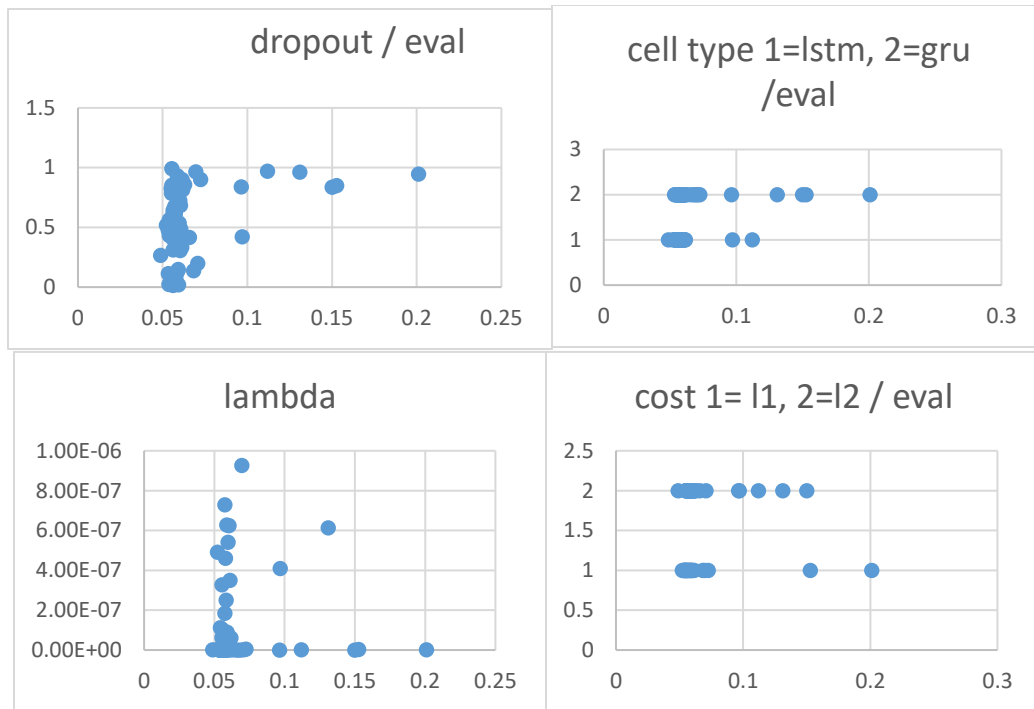
Figura 26 Evaluare / iterații

În cazul acestui model, se poate observa îmbunătățirea datorată modificării hiperparametrilor, însă setările cu rezultatele cele mai bune sunt puține la număr și par a fi cazuri izolate, rămâne în continuare de analizat de ce anume, acest model, întâmpină dificultăți în recunoașterea de pattern-uri locale. Este necesară experimentarea cu ferestre de analiză mai mici, sau utilizarea unor seturi mai ample de date (analiza pe perioade mai mici de timp - ex: de la o zi la 2 ore).

Având în vedere modul de antrenare (prin varierea hiperparametrilor) se va încerca observarea parametrilor care aduc cele mai bune rezultate, sau dacă există vreo corelație între scorul de evaluare și valoarea parametrilor.

Se vor analiza rezultatele pentru ferestrele de analiză 'k' cu cele mai bune scoruri de evaluare. În cazul de față, pentru rețeaua neuronală recurentă - cel mai bun rezultat s-a obținut pentru k=64

În graficele de mai jos se poate observa corelația dintre principalii parametri variați și scorul de evaluare pentru RNN.



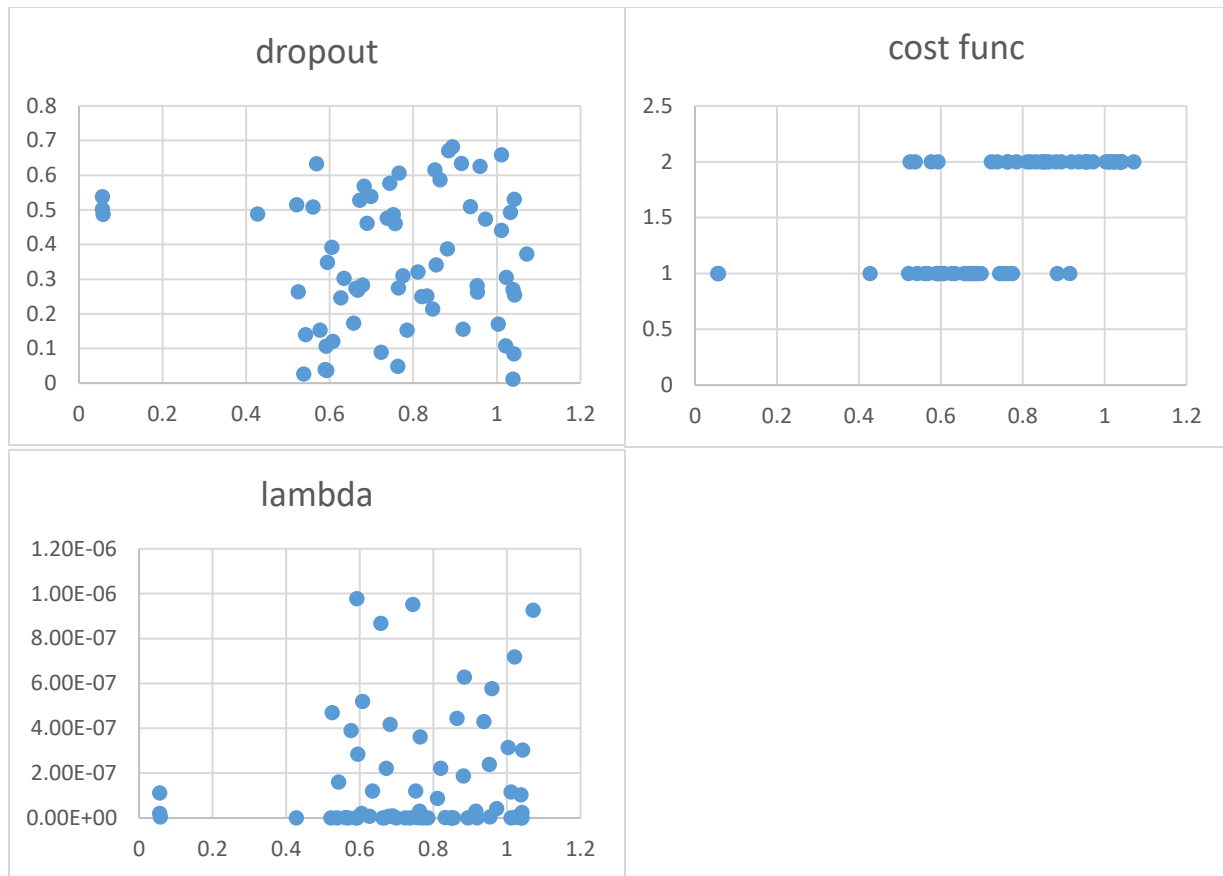
Din aceste rezultate inițiale se poate observa că un lambda foarte mic (apropiat de 0) aduce cele mai bune rezultate - lambda este coeficientul termenului de regularizare, o valoare mică pentru acest coeficient slăbește efectul regularizării. De asemenea, regularizarea cea mai eficientă s-a obținut folosind regularizarea de tip l2.

Tipul celulei cu cel mai bun scor este lstm, însă observăm rezultate mai consistente în zona minimului erorii pentru celula de tip gru.

Variația dropout-ului nu pare să fi afectat în mod consistent rezultatul evaluării.

Datorită dimensiunii mici a bazei de date (- din cauza ferestrei de preț de 1 zi (candle)) variația loturilor de antrenare nu este relevantă.

Pentru CNN:



În cazul rețelei de tip CNN, cele mai bune rezultate au fost obținute pentru valori ~ 0.5 ale dropout-ului, cea mai mică eroare a fost obținută cu o funcție de cost de tip L1 și valoare mică a coeficientului lambda.

Dintre toate modelele antrenate, cel mai bun rezultat din punctul de vedere al erorii absolute minime a fost obținut de rețeaua neuronală recurentă, urmată foarte aproape de modelul bazat pe regresia liniară.

Rezultatele pentru antrenarea și testarea acestui model RNN sunt următoarele:

```
# Train results:
Mean absolute error: 0.027572
SD absolute error:   0.033307
Sign accuracy:      0.565259
Mean squared error: 0.001870
Sqrt of MSE:        0.043239
Mean error:         0.000302
```

```
# Test results:
Mean absolute error: 0.024018
SD absolute error:   0.024877
Sign accuracy:      0.552573
Mean squared error: 0.001196
Sqrt of MSE:        0.034579
Mean error:         0.000244
```

Se poate observa din aceste rezultate faptul că eroarea poate ajunge la valori foarte mici - media eroarii absolute fiind de aproximativ 2.7%

Însă observăm că predicția semnului pentru următoarea valoare are o acuratețe de 55%, aparent aleator.

Având în vedere că procentul mic de modificare a prețurilor de la o zi la alta, deși acuratețea semnului este 55% predicția nu este atât de diferită de prețul real, motiv pentru care eroarea medie absolută este foarte scăzută.

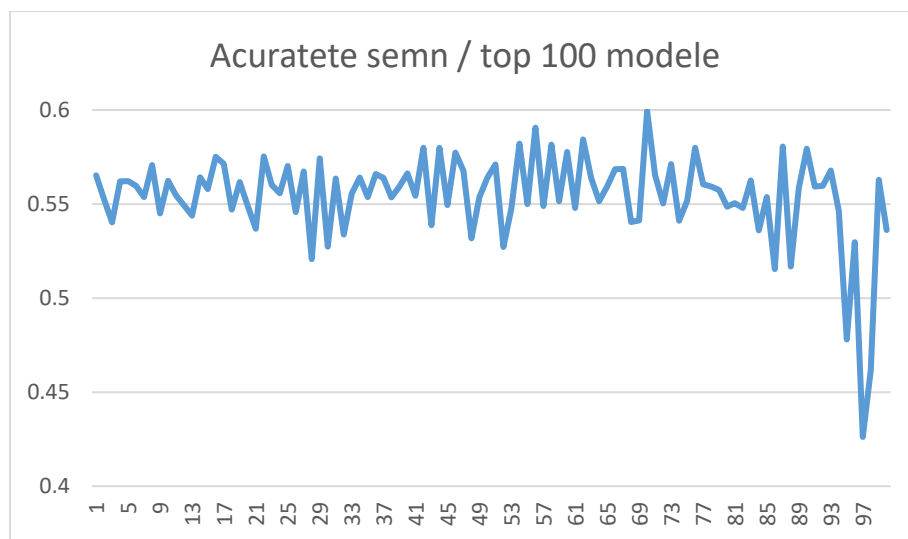


Figura 27 acuratețe / semn

Putem observa în graficul de mai sus pentru top 100 modele de rețea recurentă (sortate după scorul de evaluare, 1 fiind cel mai bun scor) - acuratețea semnului se învâрте în jurul valorii de 50%

În continuare se va analiza această metrică și pentru celelalte modele (rețeaua neuronală simplă, rețeaua neuronală convolutională și regresia liniară)

Rețeaua neuronală simplă:

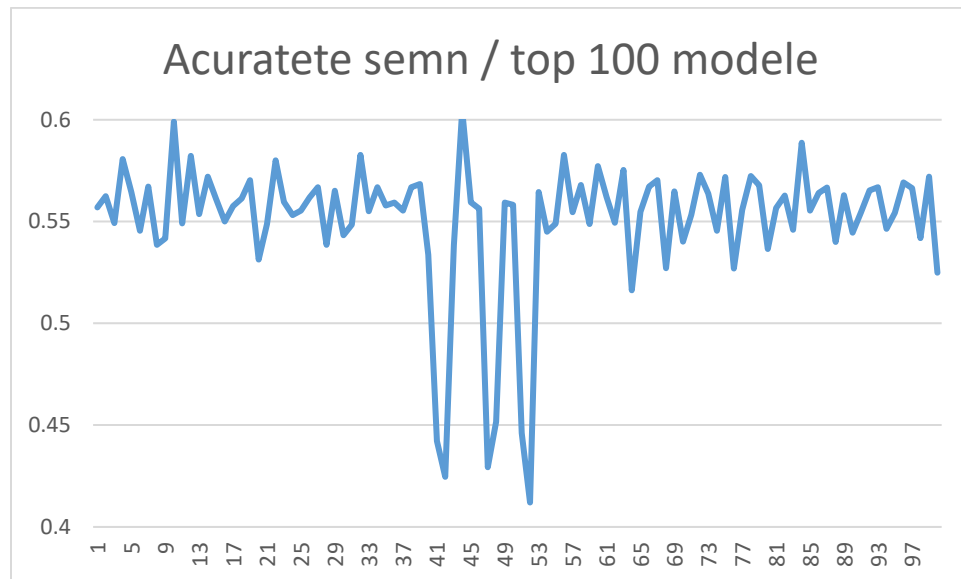


Figura 28 acuratețe / semn

Observăm că acest model are puțini outlieri care ajung aproape de 40% acuratețe, însă, asemănător cu rezultatele precedente, toate valori se învârt în jurul lui 50%. În topul modelelor însă, observăm valori peste 55% uneori apropiat chiar de 60%.

Rețeaua convolutionala:

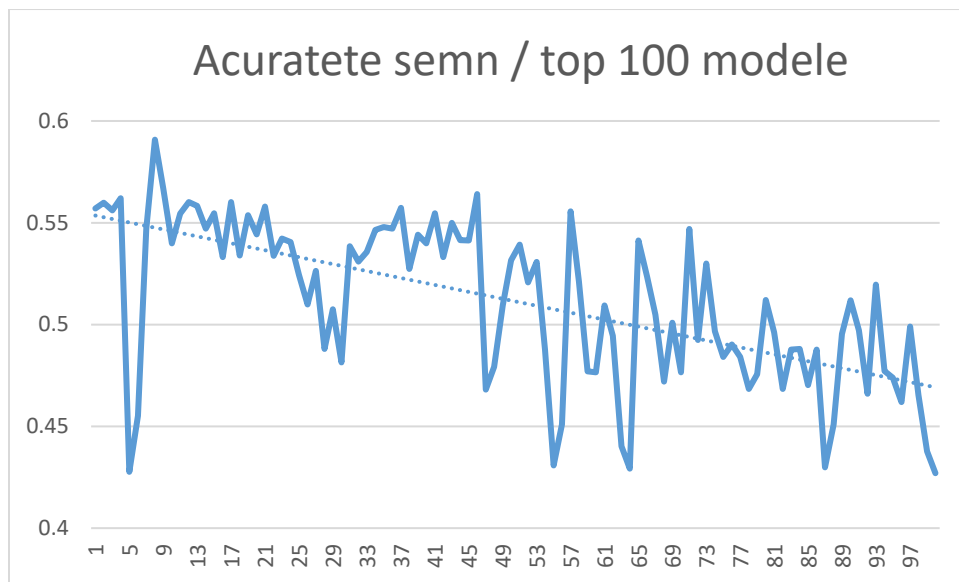


Figura 29 acuratețe / semn

În cazul acestui tip de model, se observă variații mai puternice, însă se poate observa și o linie de trend a acurateții semnului în funcție de scorul de evaluare, lucru care devine și mai evident pentru un set de date mai mare:

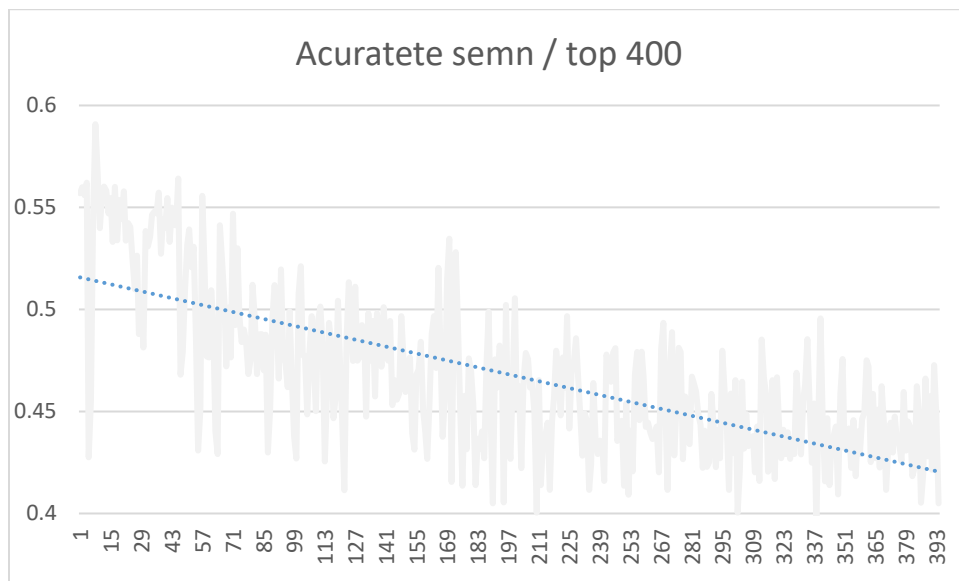


Figura 30 acuratețe / semn

Acest lucru înseamnă că acest tip de model își îmbunătățește scorul o dată cu îmbunătățirea acurateții semnului lucru care nu este observabil pentru celelalte modele.

Regresia liniară:

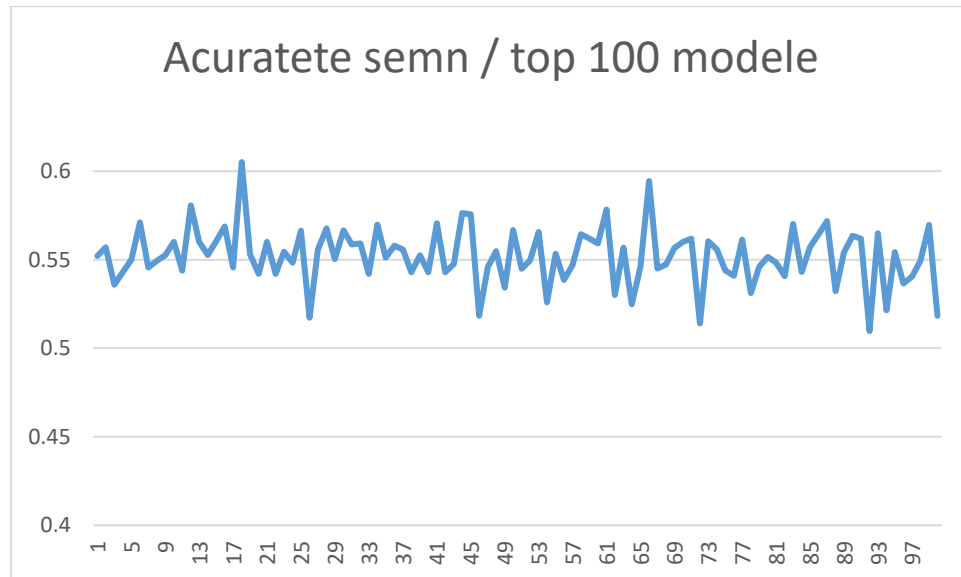


Figura 31 acuratețe / semn

În cazul acestui model se poate observa că variațiile sunt relativ mai mici decât restul modelelor precedente, de asemenea obține consistent acuratețe mai bună - în jurul valorii de 55% (și peste 50% în 100% din cazuri) - și este singurul model care a atins acuratețe de peste 60%

Se va analiza în continuare în același mod folosind un alt set de date pentru antrenare și testare - **pentru fereastră de timp al prețului de 2 ore** - în acest caz, pentru aceeași perioadă totală de timp vom avea mai multe eșantioane, se va încerca predicția pentru următorul eșantion (la fiecare 2 ore).

Rețeaua neuronală simplă și regresia liniară

k	NN	LR
1	0.0174	0.0181
2	0.0192	0.0187
3	0.0184	0.0173
4	0.0176	0.0178
5	0.0189	0.0181
6	0.02	0.0187
7	0.0191	0.018
8	0.018	0.0175
9	0.0186	0.0178
10	0.0199	0.019
11	0.0183	0.018
12	0.0179	0.0183
13	0.0179	0.0188
14	0.0191	0.0184
15	0.0182	0.018
16	0.0186	0.0182
17	0.0196	0.0182
18	0.019	0.0181
19	0.0178	0.0181
20	0.0179	0.0184

Tabel 4 NN / LR

O diferență evidentă între rezultatele din acest experiment și cel prezentat anterior este faptul că eroarea pare să fie mică pentru oricare din valorile lui "k". Toate valorile fiind mult mai bune decât oricare din modelele anterioare.

Rețeaua recurentă și rețeaua convolutionala:

k	RNN	CNN
24	0.0185	0.6545
32	0.0178	0.624
48	0.0186	0.6541
64	0.0184	0.0194

Tabel 5 RNN / CNN

Pentru rețeaua neuronală recurentă observăm același comportament - valori foarte mici ale erorii, net mai bune decât orice rezultat anterior.

Pentru rețeaua convolutională însă observăm valori relativ mari ale erorii însă am obținut și un model cu rezultate bune pentru $k=64$, contrar cu rezultatele anterioare unde cel maim bun model a fost obținut pentru $k=24$.

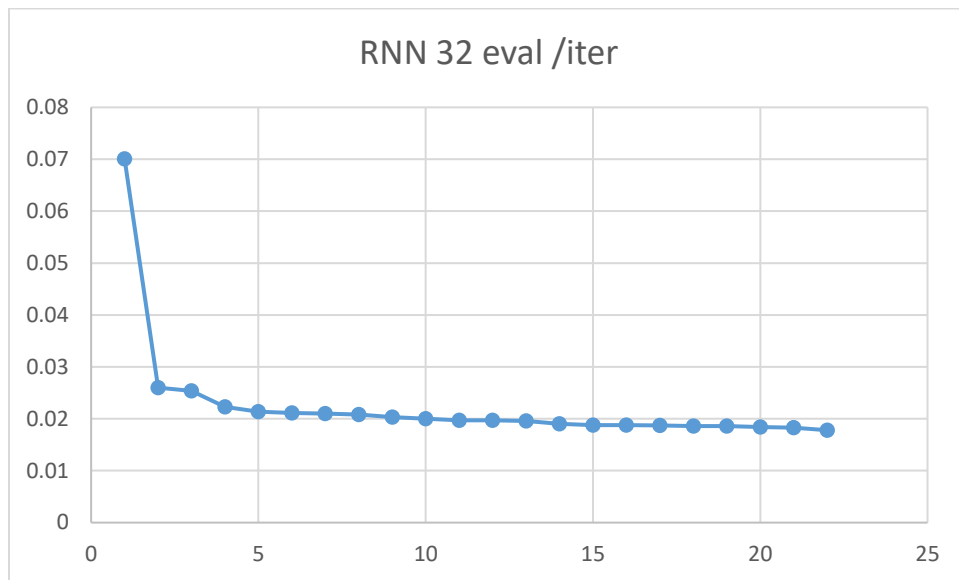


Figura 32 evaluare / iterații

În cazul rețelei neuronale recurente se găsesc rapid hiperparametrii cu erori foarte mici. Însă, că și în experimentul trecut, se găsesc relativ greu hiper parametri corespunzători pentru minimizarea erorii în cazul rețelei convolutionale.

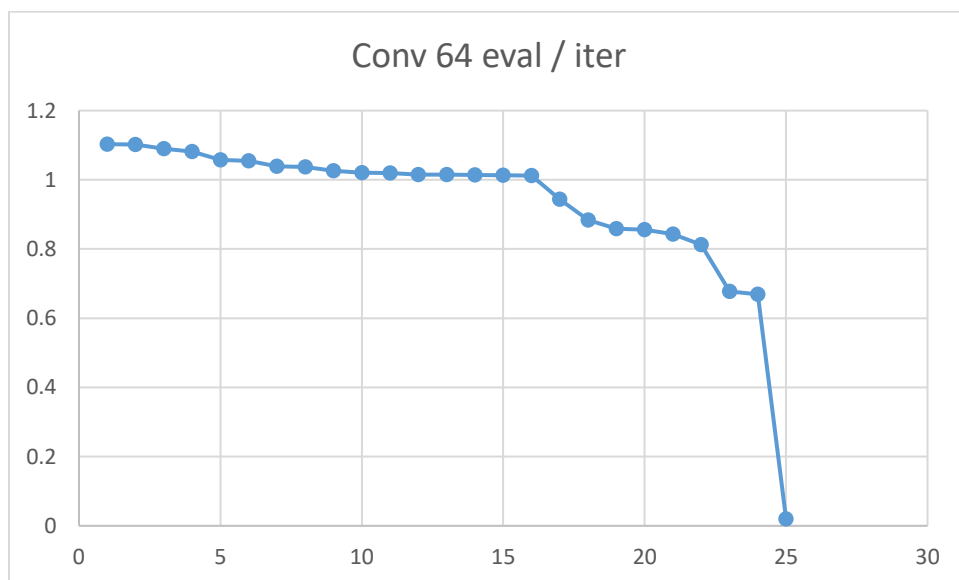
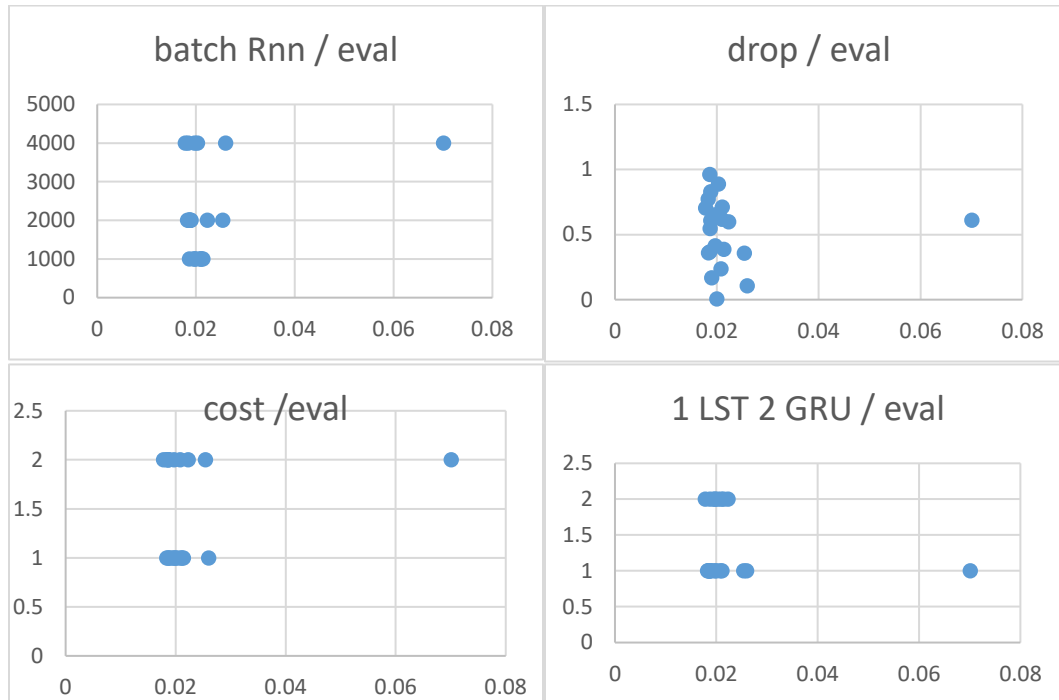


Figura 33 evaluare / iterații

În continuare se vor analiza valorile și corelarea lor față de scorul de evaluare pentru a observa dacă există sau nu un set de parametri care ar funcționa bine pentru ambele tipuri de date.

În cazul RNN:



A fost adăugat în plus față de experimentul trecut - parametrul de batching (lot) datorită faptului că baza de date este acum de 12 ori mai mare - se vor testa folosind loturi de 1000 - 4000 (cu increment de 1000).

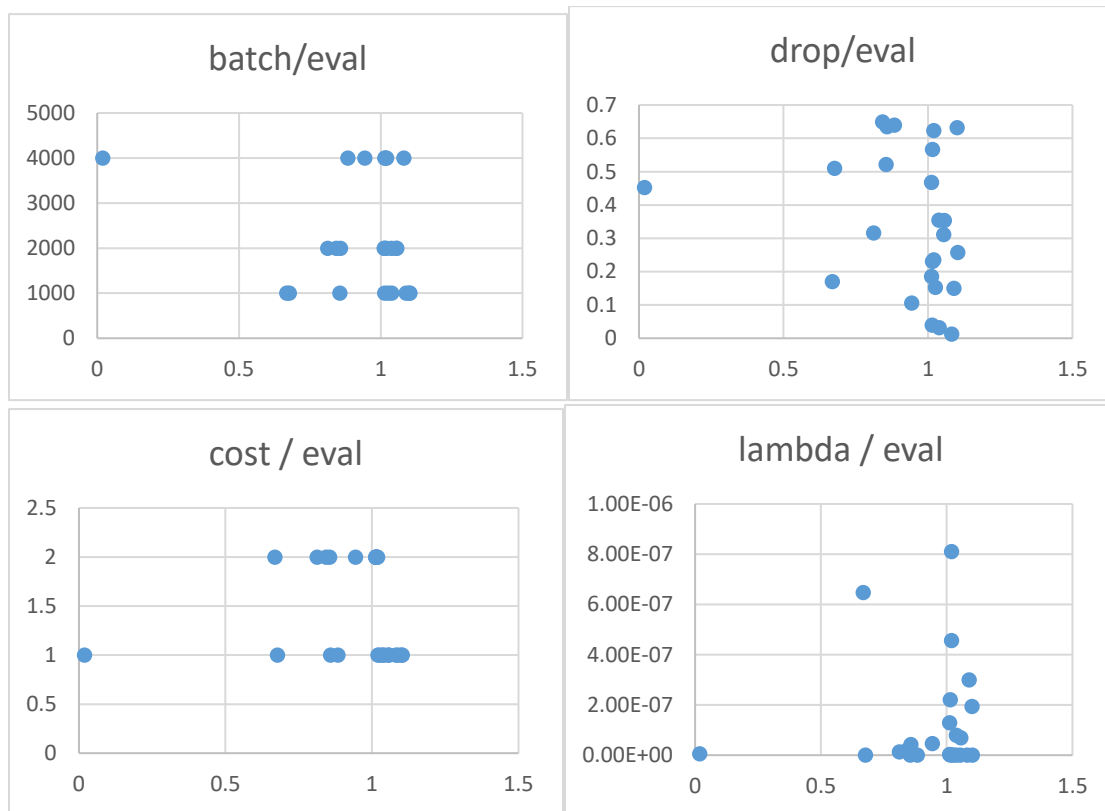
Dimensiunile loturilor cu cele mai bune scoruri, reprezentate de asemenea și de cei mai mult candidați sunt cele de 1000 și 2000.

Ca și în experimentul trecut, valoarea dropout-ului nu pare foarte relevantă obținându-se scoruri bune pentru aproape toate valorile.

Funcția cost cu cel mai bun scor a fost în acest caz L2 (în primul experiment fiind L1), o observație importantă însă este că în majoritatea cazurilor s-au obținut scoruri mult mai bune ca în experimentul trecut.

Funcția cost cu cel mai bun în acest rezultat în acest caz este GRU, cu rezultate foarte apropiate însă de LSTM.

În cazul CNN:



În cazul rețelei convoluționale putem observa că parametrii modelului cu cea mai mică eroare sunt asemănători cu cei din primul experiment:

- Dropout = ~ 0.5
- Funcția cost L1
- Lambda apropiat de 0

Pentru acest experiment, modelul cu cel mai bun scor este modelul bazat pe regresie liniară, însă în jurul aceluși scor cu diferențe foarte mici se află modelul bazat pe rețea neuronală simplă urmat de rețeaua recurență.

Modelul bazat pe regresie liniară:

```
# Train results:
Mean absolute error: 0.007950
SD absolute error: 0.011954
Sign accuracy: 0.450009
Mean squared error: 0.000206
Sqrt of MSE: 0.014357
Mean error: -0.000341

# Test results:
Mean absolute error: 0.007223
SD absolute error: 0.010072
Sign accuracy: 0.461538
Mean squared error: 0.000154
Sqrt of MSE: 0.012394
Mean error: 0.000024
```

Se poate observa din aceste rezultate ca acuratețea semnul este <50% deși valoarea erorii este mult mult mai mică că în experimentul trecut - 0.7%

Având în vedere faptul că am mărit dimensiunea bazei de date de 12 ori prin adăugarea de valori la fiecare 2 ore în loc de 24, variațiile de preț de la un eșantion la altul sunt foarte mici, astfel modelele rezultate prezic variații mici ale prețului și în majoritatea cazurilor sunt foarte aproape de adevăr, din același motiv se degradează și acuratețea semnelui, variațiile de semn fiind foarte dese.

Se pot observa aceste efecte în valorile acurateții semnelui pentru toate modelele antrenate:

-Rețeaua recurență:

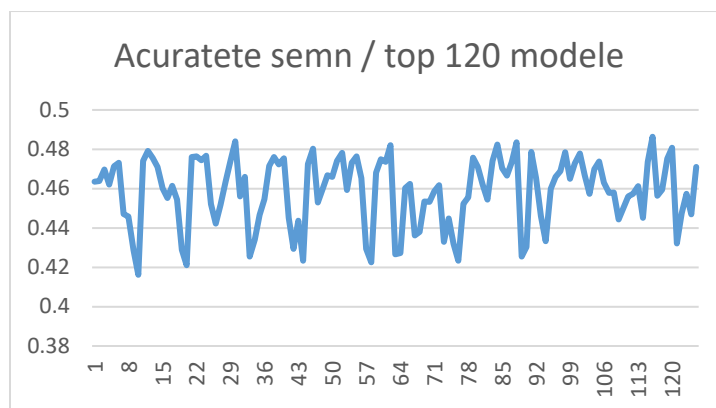


Figura 34 acuratețe semn

- Rețeaua neuronală simplă:

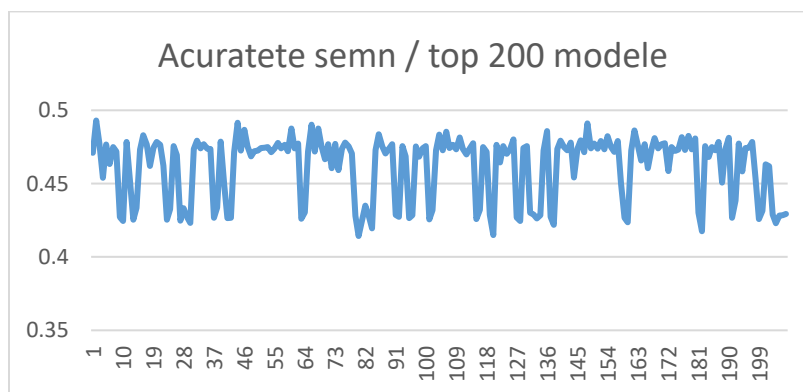


Figura 35 acuratețe semn

- Rețeaua convolutionala:

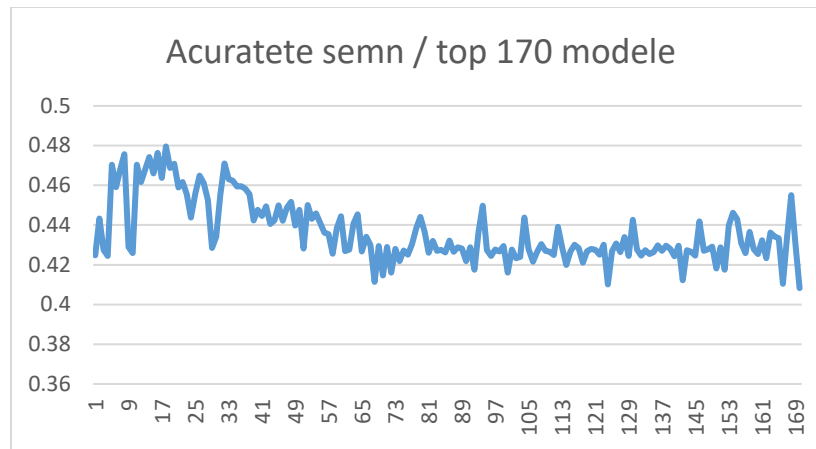


Figura 36 acuratețe semn

- Regresia liniară:

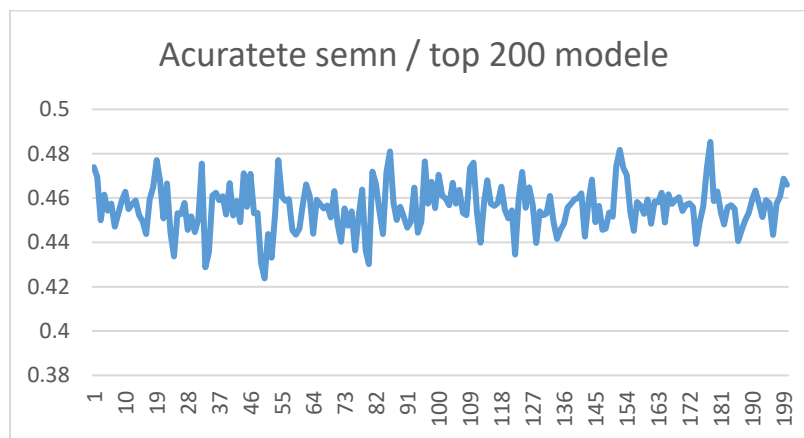


Figura 37 acuratețe semn

În toate aceste cazuri observăm că toate valorile de acuratețe sunt sub 50% în 100% din cazuri. O altă observație ar fi faptul că pentru rețeaua convolutională observăm aceeași caracteristică a trendului - acuratețea care crește o dată cu minimizarea erorii:

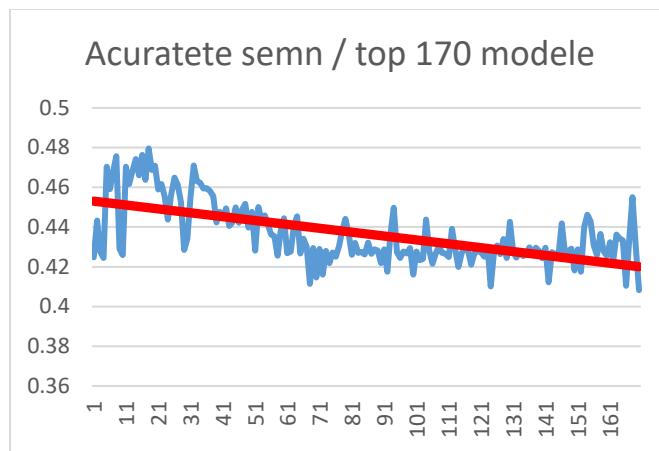


Figura 38 acuratețe semn

Din toate aceste experimente putem face următoarele observații importante ce privesc predicția prețurilor folosind modele de rețele neuronale:

- Predicțiile cu erorile cele mai mici au fost pentru antrenarea pe bază de date cu fereastră de timp de 2h - însă acest lucru se datorează și variațiilor mai mici de preț de la un eșantion la altul.
- Predicțiile cu cea mai mare acuratețe a semnelui au fost făcute pentru antrenarea pe bază de date cu fereastră de timp de 24h - însă valori mai mari ale erorilor.
- Acuratețea de predicție a semnelui este apropiată de 50% în majoritatea cazurilor, cu toate acestea există modele care ajung la 60%
- Rețelele recurente tind să facă o predicție a prețului foarte apropiată de valoarea anterioară (lucru care minimizează eroarea)
- Rețelele convoluționale au obținut scoruri defavorizante însă există cazuri în care se pot apropia de erori foarte mici.

Aceste observații sunt consistente cu majoritatea lucrărilor publice în acest sens, unde acuratețea de predicție a semnelui este în jurul valorii de 50%.

De asemenea, în multe lucrări de specialitate și articole online care studiază posibilitatea predicției folosind rețele neuronale recurente cu celule de tip LSTM, având valori foarte mici ale erorii de predicție [6], clasifică modelele ca fiind bune (cu acuratețe ridicată), însă acest aspect poate fi înșelător.

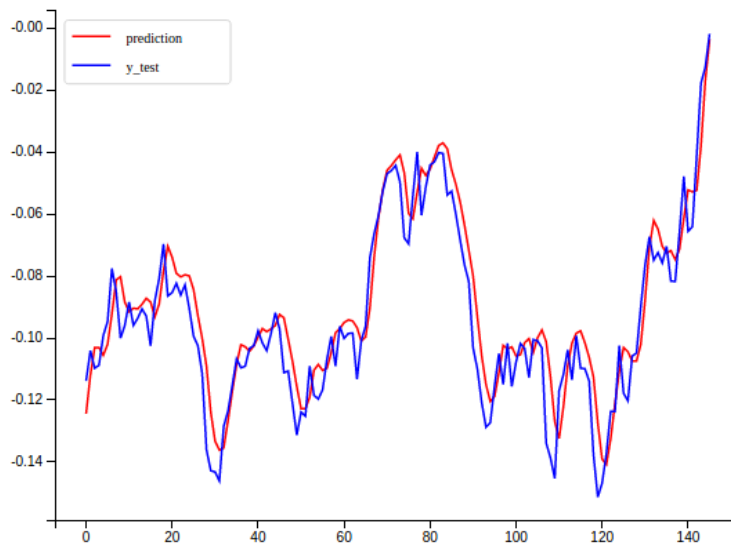


Figura 39 Predicție / real [6]

În imaginea de mai sus, se poate observa predicția - cu roșu, și prețul real cu albastru. Aparent, se poate spune că la prima vedere acest model este foarte precis - autorul ajungând la o valoare a erorii de "0.00859" , însă în realitate, nu este deloc așa.

Pentru a demonstra acest lucru se va folosi un set de date în serie temporală - și modelând după un proces aleator (Random walk - după cum a fost explicat în capitolul anterior) -un proces 100% stohastic. Astfel folosirea datelor din trecut pentru antrenare, și utilizarea lor pentru predicție, pur și simplu nu are sens. Rezultatele însă arată în felul următor:

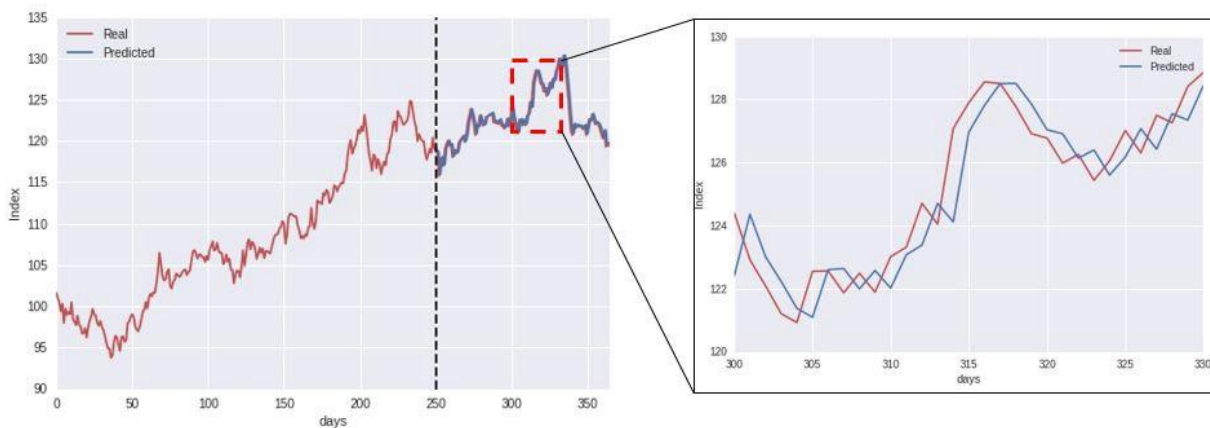


Figura 40 predicție / real - zoom

În imaginea din stânga se poate observa baza de date de antrenare (până la ziua 250) și baza de date de testare (restul)

În imaginea din dreapta însă se poate observa prețul real - cu roșu și prețul prezis cu albastru. Întrebarea normală în acest caz este, cum a putut rețeaua neuronală să prezică ceva complet aleator?

Răspunsul este că nu a putut



Figura 41 predicție / real offset



Figura 42 predicție / real offset -1

Se poate observa în imaginea de deasupra, dacă shiftam cu 1 zi la stânga prețul prezis, prețurile sunt aproape identice.

Datele temporale tind să fie corelate în timp, în acest caz prețul de la "t+1" este foarte probabil apropiat de prețul de la momentul "t".

În graficul de mai jos vom afișa cross-corelarea dintre prețul prezis și prețul real:

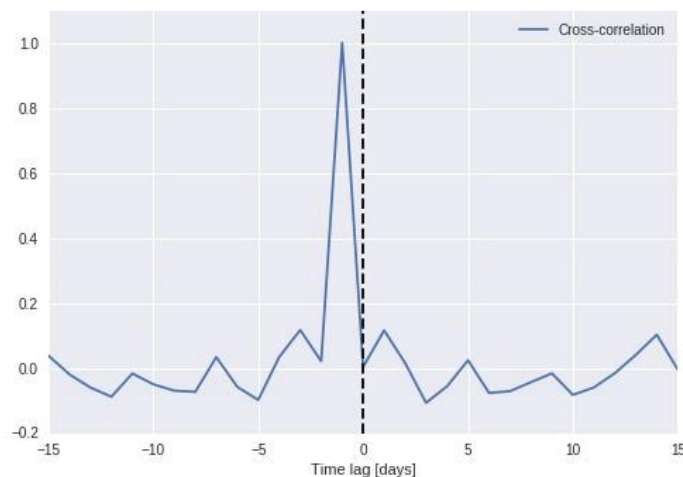


Figura 43 cross correlation (prezis / real)

Putem observa în mod evident un vârf la o întârziere de exact 1 zi.

Asta înseamnă că evaluând modelul în funcție de abilitatea sa de a prezice valoarea prețului de mâine, folosind metrici de tipul RMSE (rădăcină erorii pătratice medii), pot indica predicții cu acuratețe foarte mare. Dar după cum am menționat anterior, datele sunt generate folosind un proces aleator (random walk) astfel este imposibil pentru model să prezică viitorul.

Cum am adresat această eroare în lucrarea de față?

Folosind tehnica de stationarizare (menționată în prelucrarea bazei de date) - o serie de date staționară este aceea pentru care varianța, autocorelația sunt constante în timp. Majoritatea metodelor de prezicere se bazează pe stationarizare prin transformate matematice.

O astfel de metodă o reprezintă - transformarea datelor în diferențe procentuale în timp

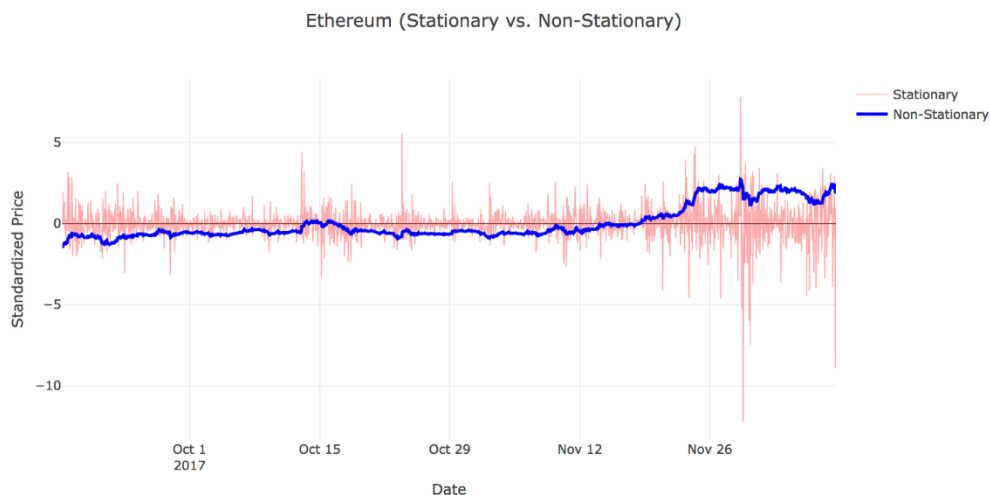


Figura 44 Valoare staționară / nestationara

Definind un model care prezice diferența procentuală dintre valori între momentele de timp în loc de valoarea propriu zisă, este un test mult mai puternic pentru capabilitățile de predicție ale modelelor de rețele neuronale.

Având în vedere că hiper parametrii rețelelor pot fi ajustați astfel încât să obținem valori mai mici ale erorii medii absolute dar și o acuratețe mai mare a prezicerii semnului, de asemenea, având în vedere și că modelele folosite sunt modele relativ simple, se poate observa că există posibilitatea predicției prețului viitor. De asemenea, utilizarea unui set mai mare de caracteristici (care ar putea include un set de indicatori tehnici - prezentați anterior) ar putea aduce îmbunătățiri modelelor actuale.

Folosind modelele cele mai bune din modelele rezultate, am realizat un simplu experiment cu o încercare de predicție a prețului pe o fereastră de timp în **avans** de 100 de zile (nu predicție de la 1 zi la alta). Astfel, am folosit o tehnică utilizată deseori în predicții pe durate lungi, predicțiile se fac în același mod, având inputul "k" (de la $t-k$ la t) se prezice valoarea de la momentul $t+1$, această valoare este apoi adăugată în setul de valori - se shiftează cu +1 apoi se retrimite datele în model, repetând acest proces de 100 de ori. Putem observa astfel tendința de modificare a prețului prezis de rețea:

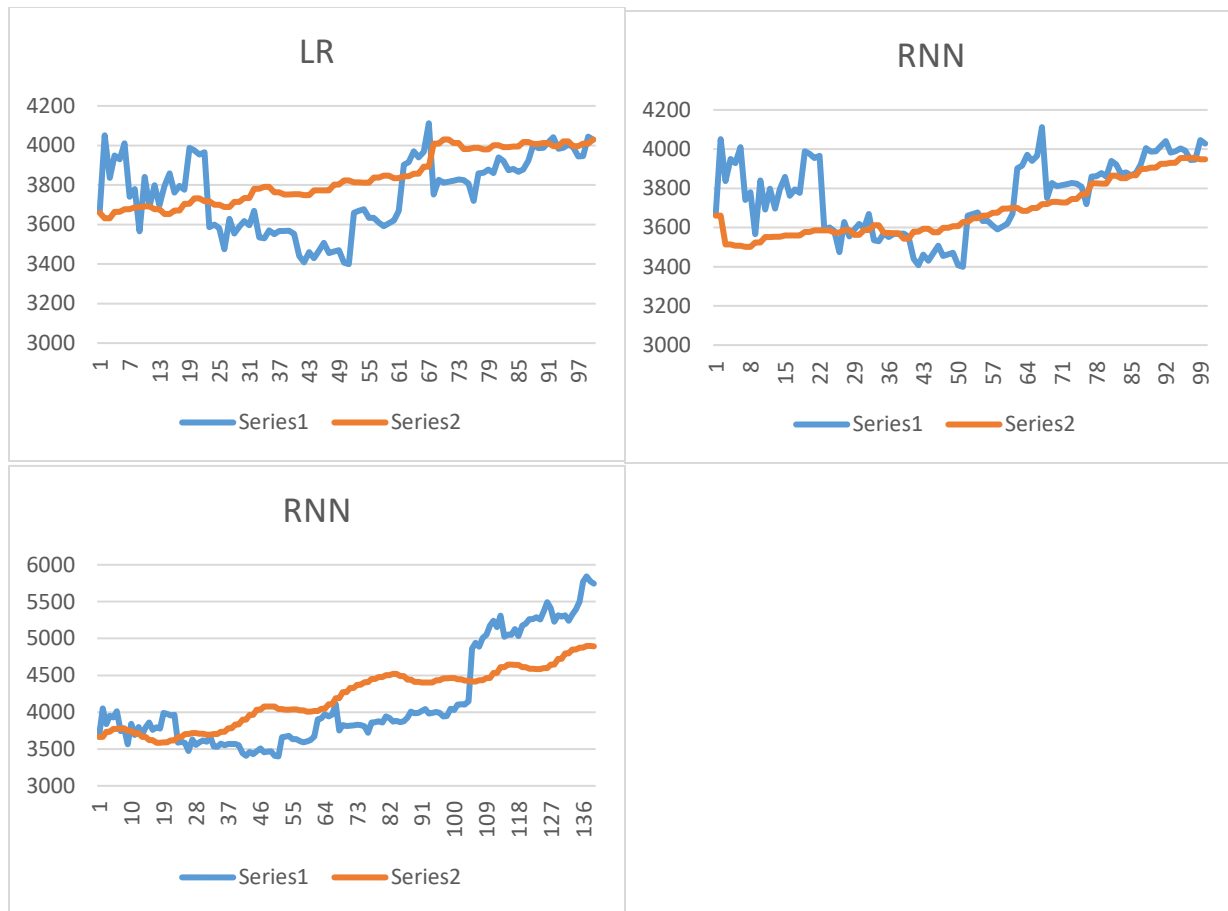


Figura 45 100 zile (predicție / real)

Albastru- prețul real / Portocaliu - prețul prezis

După cum se poate observa, rezultatele sunt foarte interesante, rețeaua reușind să prezică trendul dar chiar și unele variații de preț - având în vedere că este o prognoză pe durata foarte mare, aceste rezultate sunt foarte încurajatoare.

Am testat cel mai bun model din punct de vedere al erorii minime (RNN) pe o fereastră de 5 zile

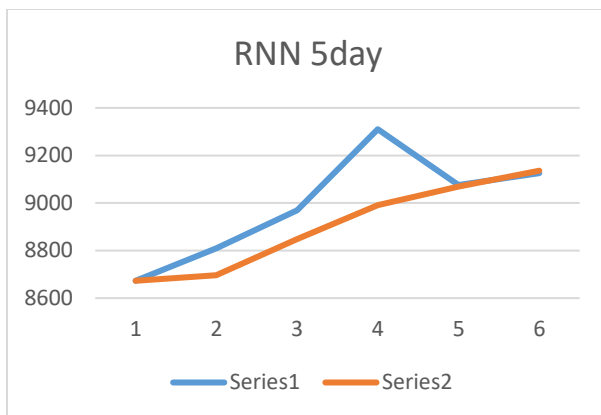


Figura 46 RNN / predicție 5 zile

Din nou, trendul este aproximat corect (albastru - prețul real, portocaliu - prețul prezis), variațiile puternice nu. Folosind aceeași rețea am căutat zile în care prețul era în scădere și am pornit predicția în momentul în care prețul a schimbat trendul, rezultatul poate fi ibservat în graficul de mai jos (cu portocaliu este reprezentat prețul real - cu gri este prețul prezis):

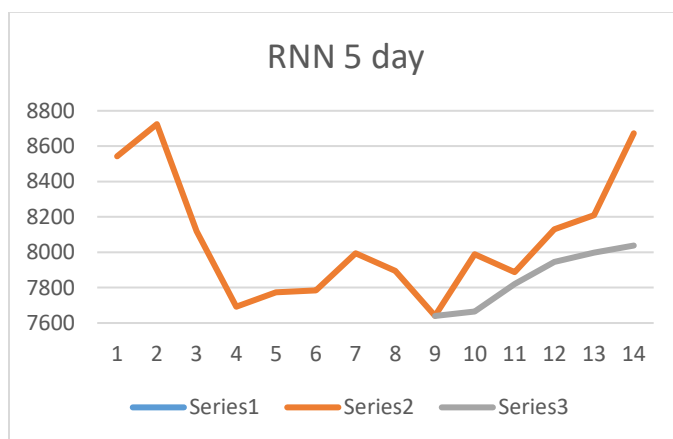


Figura 47 RNN / predicție 5 zile

Se poate observa că rețeaua a reușit prezicerea schimbării de trend.

Rezultatele sunt nu sunt totuși atât de apropiate de realitate încât să putem lua decizii de investiții, și astfel nu putem spune că modelele pot fi utilizate cu încredere, cel puțin nu în mod individual.

O variantă de abordare a problemei ar putea fi combinarea predicției unui model cu alți factori sau algoritmi. Majoritatea trader-ilor profesioniști își creează algoritmi determinați bazați pe combinații de indicatori tehnici pentru a putea determina prețul de cumpărare sau vânzare. De asemenea, există și sisteme de cuantizare a sentimentului pieței - bazate în principal pe date Google trends, sau recunoaștere de sentimente din Tweet-uri cu diferite tag-uri.

Un sistem complex de tranzacționare nu este necesar să fie bazat în totalitate pe un model de machine learning ci pe o combinație de astfel de modele, sau chiar o combinație cu mai multe tipuri de algoritmi.

Această abordare se va discuta în sub-capitolul următor. Se va încerca integrarea celui mai performant model după metricile prezentate anterior într-un sistem care poate utiliza mai multe informații în vederea determinării unei decizii. Acest sistem reprezintă aplicația concretă a lucrării de față.

Sistemul automatizat de tranzacționare

După cum am menționat în capitolele anterioare, se va încerca construirea unui sistem automatizat de tranzacționare, acest sistem va utiliza un model de rețea neuronală din cele prezentate în capitolul anterior. Inițial, în perioada start a proiectului, am pornit de la ideea că se poate crea un sistem complet automatizat și bazat în totalitate pe decizia unei rețele neuronale. În timp însă, după perioada de cercetare și experimentare, am observat că este nevoie de mai mult de atât pentru a putea crea un astfel de sistem, voi detalia procesul și pașii urmăți în subcapitolele următoare.

Limbajul de programare

În mod evident că prima decizie în conceperea unui sistem software, este decizia asupra limbajului de programare. În acest stadiu trebuie luate în considerare tipul aplicației, modul în care aplicația trebuie utilizată, și evident funcționalitatea aplicației. Astfel s-a decis utilizarea unui limbaj de programare orientat pe obiecte, deoarece poate reduce complexitatea prin modularizare.

C++ și Java au fost principalele limbaje alese, în final luând decizia ca C++ să fie limbajul de bază datorită librărilor puse la dispoziție și a modului în care acesta poate integra cu ușurință diferite module software.

Diagrama inițială de funcționare

După cum am menționat anterior, inițial sistemul avea o arhitectură relativ simplă, bazându-se în principal pe un bloc decizional bazat pe rețeaua neuronală cu cele mai bune rezultate.

Astfel, sistemul trebuie să conțină în primul un bloc de înglobare a funcțiilor de API ale unui exchange, pentru a asigura conectarea cât mai sigură și rapidă. C++ permite integrarea librăriei cURL (libcurl) care face posibilă comunicarea cu un server printr-un anumit protocol (ex. HTTP, HTTPS, FTP, FTPS, etc).

De asemenea este nevoie de un bloc care să se ocupe de maparea funcțiilor API în metode C++ parametrizate corect. Există două tipuri de astfel de metode:

- Metode publice - acestea sunt metode prin care se pot cere informații publice de la exchange, acestea nu necesită autentificare (Ex: Prețul curent, Order book, prețuri trecute, etc.).
- Metode private - acestea sunt metode prin care se pot executa doar funcții private și necesită autentificare(plasarea de ordine de vânzare/cumpărare).

Toate aceste blocuri vor fi conectate între ele printr-un bloc principal de procesare, în acest bloc se vor utiliza metodele C++ de comunicare cu serverul exchange-ului și se vor prelucra datele care apoi vor fi trimise către rețeaua neuronală pentru a primi o predicție. Predicția este apoi utilizată pentru a se un ordin de vânzare sau cumpărare, care este apoi din nou exchange-ului.

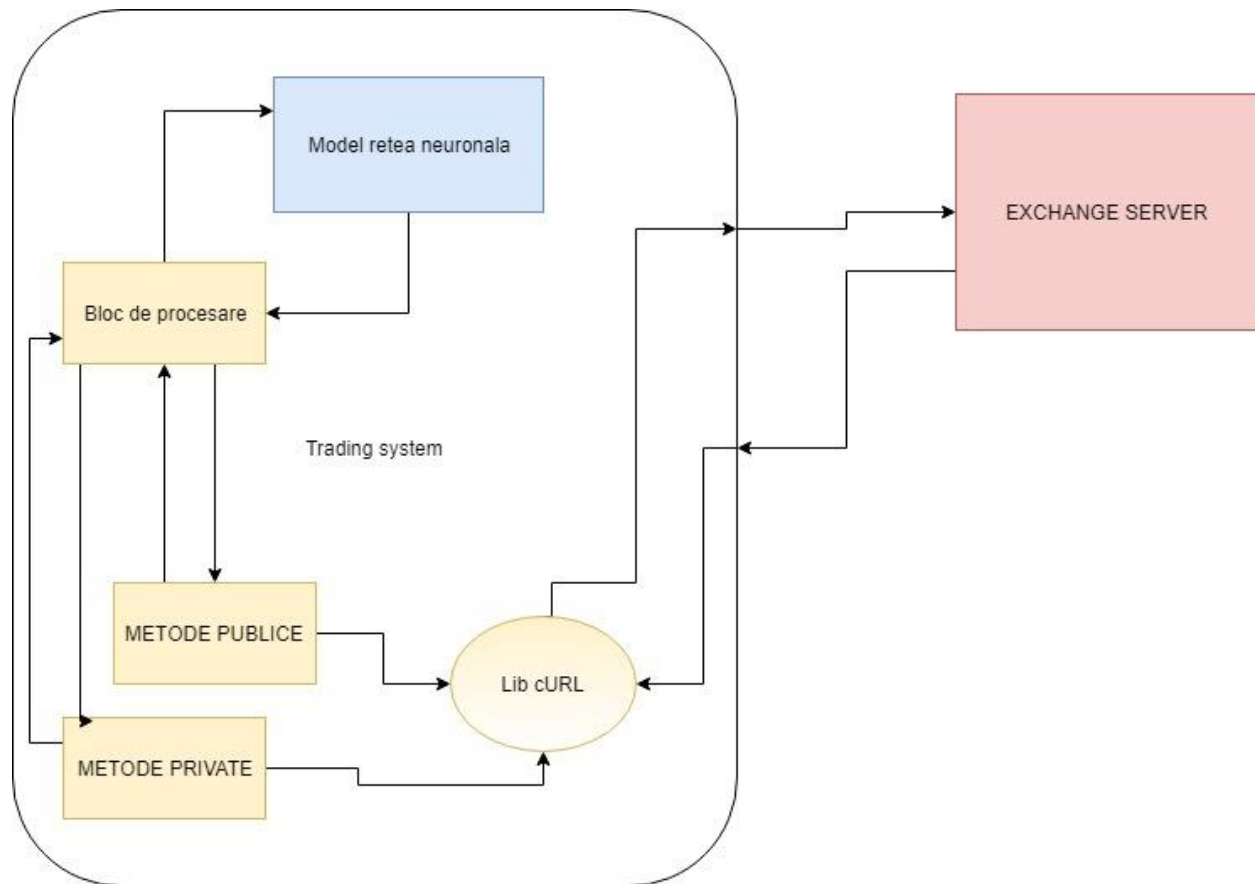


Figura 48 Arhitectură sistem

După cum am menționat, API-ul cu informațiile publice se poate accesa fără să necesite o autentificare, restul funcțiilor, cum ar fi plasarea de ordine de vânzare/cumpărare sau trimiterea de monede, se face doar prin utilizarea unei chei (API Key). Request-ul se face folosind cheia privată într-un mesaj către server.

Exemplu de mesaj:

```
https://api.bittrex.com/api/v1.1/path/call?param=value&apikey=0123456789&nonce=1542020339856
```

Exemplu de funcții API:

```
void setGetMarketSummary(strâng market)
{
    ssURL.str("");
    ssURL <<
    "https://bittrex.com/api/v1.1/public/getmarketsummary?market=" <<
    market;
}

void setGetOrdersBook(strâng market, int depth)
```

```
{      ssURL.str("");

      ssURL <<
"https://bittrex.com/api/v1.1/public/getorderbook?market=" << market
<< "&type=both&depth=" << depth;

}
```

Modulele de metode publice și private, au fost denumite `tradingInfo` respectiv `tradingMarket`. În aceste module/clase, definim metode parametrizabile, pentru a manipula datele de la server.

Exemplu de metodă din `tradingInfo`:

```
strâng tradingInfo::getTicks(strâng marketName){
    bittrex.getTicks(marketName);
    bool print1 = false;
    bool sign1 = true;
    bool exportReply1 = false;
    InstOptions instOpt(print1, sign1, exportReply1);
    //Executing instruction
    strâng buffer = bittrex.callCurlPlataform(instOpt);
    marketSum = json::parse(buffer);
    return marketSum["price"];
}
```

Având în vedere că informațiile returnate sunt în format json, acestea trebuie procesate corespunzător.

Exemplu de răspuns de la server:

```
{
  "success": true,
  "message": "",
  "result": [
    {
      "MarketName": "BTC-LTC",
```

```
"High": 0.0135,  
"Low": 0.012,  
"Volume": 3833.97619253,  
"Last": 0.01349998,  
"BaseVolume": 47.03987026,  
"TimeStamp": "2014-07-09T07:22:16.72",  
"Bid": 0.01271001,  
"Ask": 0.012911,  
"OpenBuyOrders": 45,  
"OpenSellOrders": 45,  
"PrevDay": 0.01229501,  
"Created": "2014-02-13T00:00:00",  
"DisplayMarketName": "strâng"  
}  
]  
}
```

Aplicația a fost construită astfel având în permanență în vedere integrarea modelului de rețea neuronală. Până la integrarea ultimului modul s-au realizat însă o serie de teste pentru a determina limitările arhitecturii curente.

S-au descoperit astfel primele probleme ale sistemului:

- Imposibilitatea de a trimite mai multe request-uri
- Imposibilitatea de a prelucra simultan mai multe input-uri

În mod evident sistemul trebuie să funcționeze rapid și astfel trebuie să permită utilizarea în paralel a mai multor metode. S-a decis astfel implementarea arhitecturii având la bază ideea de multi-threading.

C++11 pune la dispoziție threading, o librărie ușor de folosit, care va rezolva problemele menționate și va îmbunătăți astfel performanțele sistemului, foarte multe metode putând fi folosite în paralel, iar cele care nu pot, vor fi semaforizate.

Perioada de cercetare

În perioada de cercetare, am descoperit foarte multe informații interesante și idei pe care nu le-am considerat suficient înainte. Foarte multe detalii interesante despre indicatorii tehnici folosiți de traderi pentru a tranzacționa, de asemenea multe tipuri de arhitecturi de rețele neuronale pe care nici nu le-am avut în vedere inițial.

Astfel, după perioada de cercetare, în care am adunat informații din multe lucrări științifice și experimentat cu diferite modele (după cum a fost prezentat în capitolele anterioare), am ajuns la concluzia că este foarte dificil să obțin un model care să aducă rezultate suficient de bune pentru a îl utiliza individual și complet automatizat în tranzacționare. - (Ar necesita mult mai multe teste și experimente pentru a obține rezultate mult mai precise - ca să confere siguranță necesară de a utiliza fonduri reale pentru tranzacționare). Ideea mea inițială a sistemului trebuie modificată corespunzător, având acum aceste rezultate experimentale.

Am încercat o serie de variante de implementare, însă într-un final am ajuns la următoarea formă:

- Voi utiliza rețeaua neuronală pentru a obține o predicție a prețului
 - o Inițial, predicția va fi pentru ziua următoare (eroarea fiind minimă)
 - o Apoi, se vor face o serie de predicții pe durata mai mare (având în considerare creșterea erorii)
 - o Aceste informații le voi folosi pentru a stabili un trend (prețul va crește sau va scădea - cu probabilitatea X, determinată din numărul de predicții)
 - o Putem folosi și modele antrenare pe perioade mai mici pentru a prezice prețuri intra day
- Voi utiliza algoritmi simpli bazați pe indicatori tehnici
 - o Vor fi create seturi de combinații de indicatori care vor trimite semnale de vânzare sau cumpărare
- Toate acestea vor fi combinate și cu un index care exprimă sentimentul actual al marketului

Având în vedere experimentele prezentate, s-a observat că rețelele neuronale se descurcă bine cu datele per zi și pot prezice variații de preț, de asemenea, modelele cu cele mai mici erori au fost pentru datele per 2 ore, însă după cum am menționat, acest lucru se datorează și variațiilor mult mai mici ale prețului în acele intervale.

Astfel, s-a început implementarea algoritmilor bazați pe indicatorii tehnici:

Inițial s-a pornit prin a studia care sunt cei mai populari / performanți indicatori, multe lucrări de specialitate au menționat indicatori ca MACD, RSI, TEMA, TRIX, MFI - deși au un puternic semnal predictiv în mod individual, acești indicatori sunt mult mai siguri de folosiți într-

un sistem de tip semnal - validare. Se va folosi un indicator pentru a semnala o posibilă tranzacție, ca apoi celălalt indicator să valideze sau nu semnalul. Această metodă reduce riscul, însă reduce și din câștig.

Perechile propuse sunt:

- MACD - TEMĂ



Figura 49 MACD – TEMĂ

- MACD - TRIX



Figura 50 MACD – TRIX

- MACD - MFI



Figura 51 MACD - MFI

Integrarea indexului de sentiment se face printr-un API public, acest index folosește diferite metrice pentru a determina un scor de la 0 la 100 (0 - extreme fear , 100 - extreme greed).



Figura 52 Fear Greed Index

Metricile folosite sunt următoarele:

- Volatilitate (25%) - Creșterea ridicată a volatilității este un bun indicator al sentimentului de frică

- Momentumul marketului (25%) - Creșterea volumului de cumpărare este un bun indicator al sentimentului de lăcomie
- Social media (15%) - Spre exemplu analiză de sentiment asupra mesajelor cu un anumit tag pe platforma Twitter - ex: creșterea bruscă a numărului de taguri cu BTC - corespunde unui sentiment de lăcomie (datorat creșterii în popularitate)
- Studii prin votare (15%)
- Dominantă (10%) - reprezintă ce procent din capitalizarea totală este acoperită de o anumită monedă (în cazul nostru BTC) - creșterea dominantei reprezintă o scădere a capitalizării celorlalte monede (alt - coins) astfel este considerat un semn indicativ al fricii
- Google Trends (10%) - date relative la căutările pe google a diferitelor cuvinte cheie

Arhitectura finală și flow-ul de execuție

În final, toate aceste componente au fost integrate sau implementate în sistemul inițial în module / clase separate.

Având insuficiente date experimentale pentru a putea automatiza în totalitate sistemul de tranzacționare - se propune o implementare care să permită atât automatizarea de tranzacționare, dar și doar semnalarea unor posibile tranzacționări, astfel, trader-ul sau utilizatorul aplicației, să folosească sistemul ca o unealtă ajutătoare în luarea unei decizii.

Se vor trimite semnale de cumpărare sau vânzare de la blocurile algoritmice bazate pe indicatori tehnici, în același timp se vor realiza 10 sau mai multe predicții cu modelul selectat. Din aceste predicții se vor folosi ponderat:

- Predicțiile pentru $t+1$
- Predicțiile pentru $t+5$
- Predicțiile pentru $t+10$

Cea mai mare pondere fiind pentru $t+1$, apoi din ce în ce mai mică pentru următoarele (având în vedere că eroarea va crește de la predicțiile succesive).

Se va determina un trend și un scor de la 1 la 10 (ce procent din predicții sunt peste prețul curent în cele 3 momente).

Se va utiliza indexul de sentiment astfel:

- Apropiat de 100 - (lăcomie) mărește scorul de vânzare
- Apropiat de 0 - (frică) mărește scorul de cumpărare

Scorul final se compune din cele două:

$$80\% * \text{scorul predicției} + 20\% * \text{scorul de sentiment}$$

Arhitectura sistemului rezultat va arăta că în figură de mai jos:

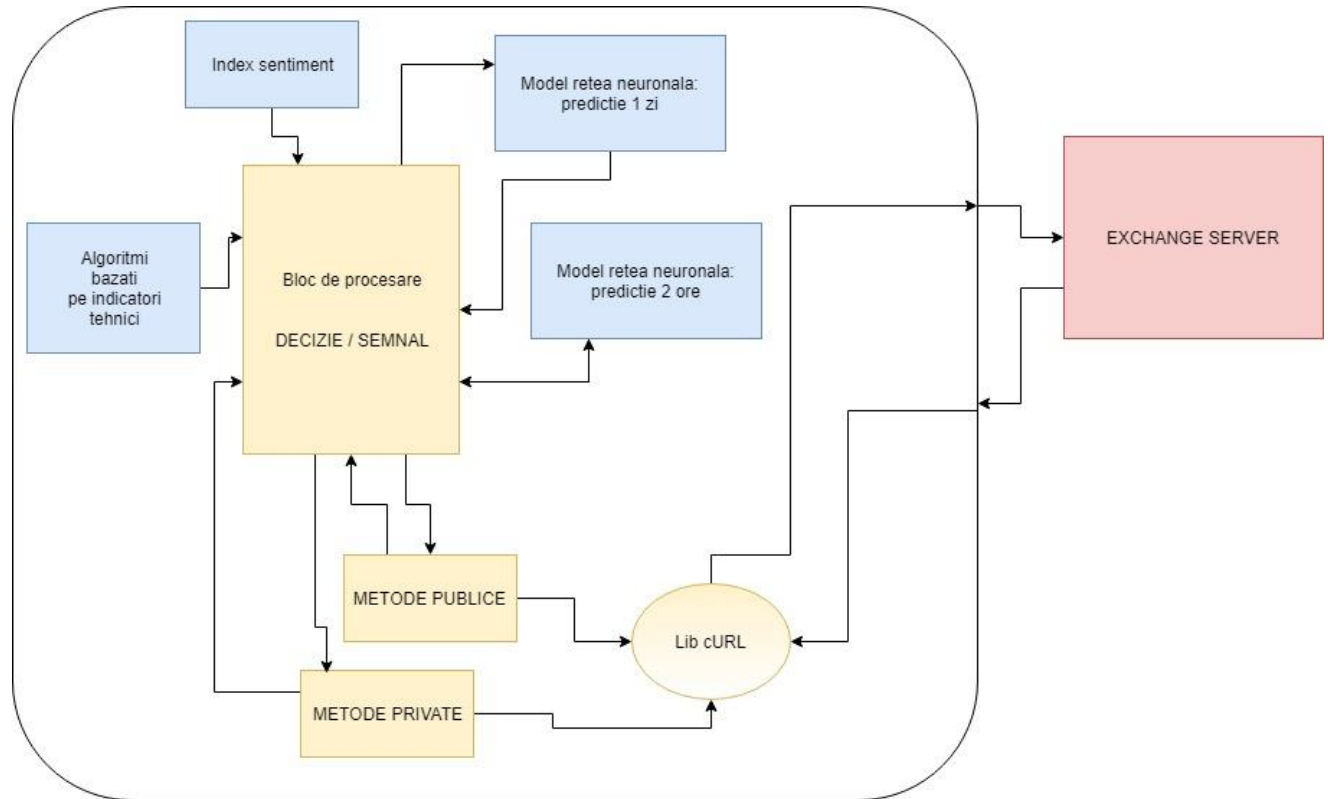


Figura 53 Arhitectură sistem

Se observă că la bază, nu diferă de cea inițială, foarte multe din blocurile inițiale au putut fi utilizate fără modificări în noul sistem.

De asemenea, utilizarea thread-urilor s-a dovedit a fi și mai utilă în implementarea actuală datorită fluxului constant de informații transmise și primite de la diferitele module.

Aplicația în sine, în forma actuală se poate utiliza doar printr-o interfață shell, o interfață grafică nu a putut fi dezvoltată în timp util, însă este în continuare în plan pe parcursul dezvoltării.

Ca fir de execuție, sistemul va funcționa ca în diagrama de mai jos:

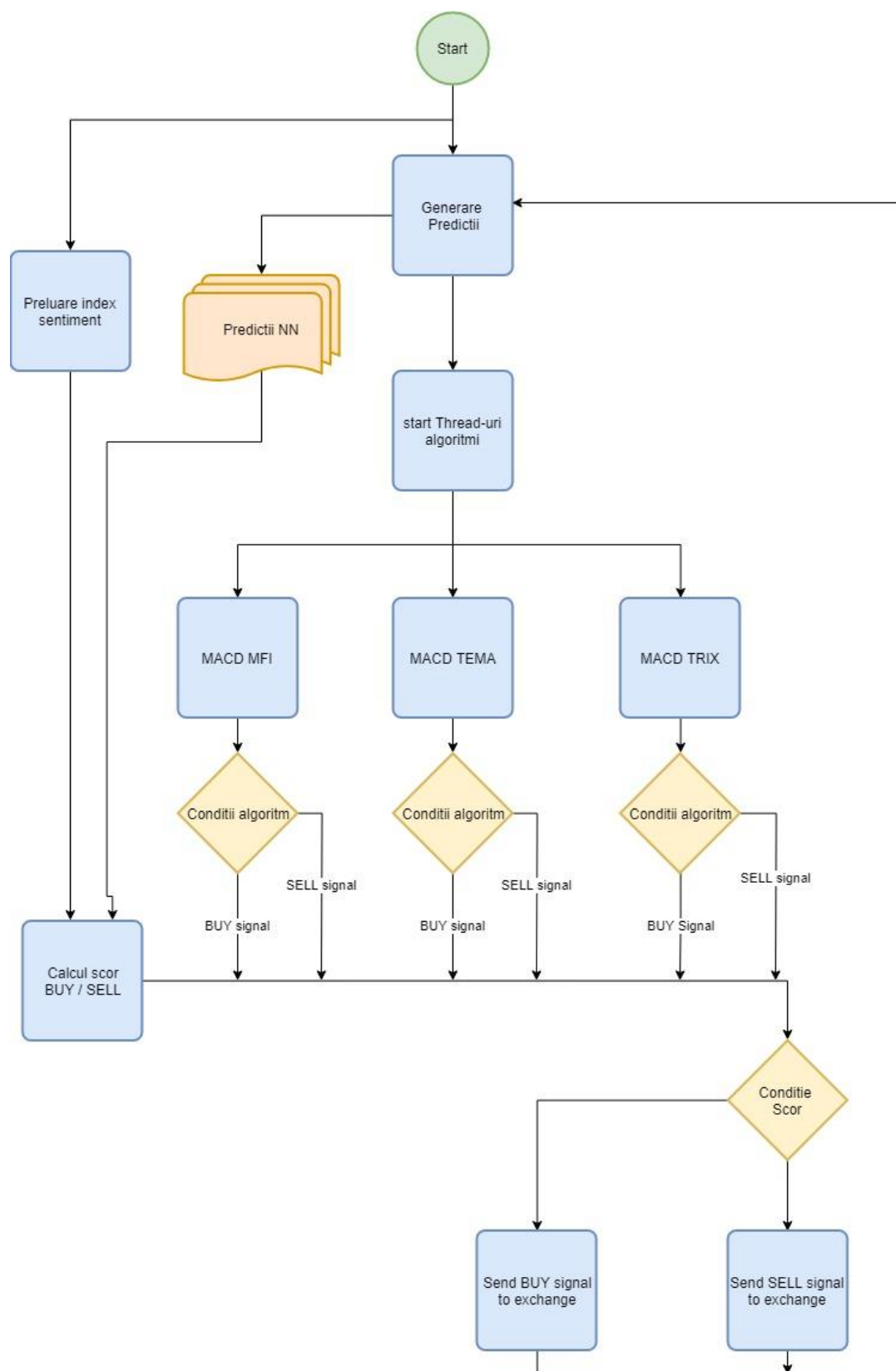


Figura 54 Diagramă de funcționare

În cazul în care se dorește ca aplicația doar să indice semnale, fără a lua decizii de cumpărare sau vânzare, ultimul bloc decizional va fi eliminat, în locul său, se va folosi doar un proces de afișare, se va afișa semnalul de cumpărare/vânzare împreună cu scorul asociat.

```
Main thread started...

Algo thread started for market USDT-BTC
Market sentiment is MIN(Extreme fear)-----84-----MAX (Extreme Greed)
starting second thread...
starting third thread...
  Using MACD MFI ALGORITHM ---starting price is 9710.74000000
  Awaiting signals
  Using MACD TEMA ALGORITHM ---starting price is 9720.60000000
  ***Awaiting signals
  --Using MACD Trix ALGORITHM ---starting price is 9726.90000000
  ---Awaiting signals
```

Figura 55 Funcționarea aplicației

Fiecare tranzacție este apoi înregistrată și scrisă într-un fișier de logging.

Scalare și îmbunătățiri viitoare

Aplicația va necesita o interfață grafică, pentru a permite într-un mod mult mai prietenos ajustarea, configurarea sau urmărirea proceselor.

Există bug-uri legate de error handling care necesită atenție, de multe ori răspunsurile de la server pot veni fie întârziate, fie goale, astfel rezultând în segmentation fault la încercarea convertirii datelor.

Ajustarea ponderilor în momentul de față se face doar în codul sursă, este necesară importarea unui fișier de configurare pentru a putea ajusta parametrii mai ușor.

În ceea ce privește cercetarea legată de rețelele neuronale și predicții, este foarte important de adăugat că există mult mai multe idei ce necesită atenție, cum ar fi, adăugarea altor trăsături în setul de antrenare - cum ar fi valori ale indicatorilor tehnici, sau utilizarea de algoritmi genetici în selecția de trăsături, sau chiar utilizarea lor în selecția de indicatori tehnici. Este un domeniu interesant de aplicabilitate a acestor algoritmi, și cu siguranță, în viitor se pot face demersuri mai complexe în încercarea găsirii unor metode mult mai eficiente.

Concluzii

Având la dispoziție toate datele experimentale cât și diferitele modalități de a utiliza sistemul final, se poate spune că rezultatele sunt într-adevăr mulțumitoare, sistemul reușind să facă tranzacții profitabile pe perioade scurte de timp. Este adevărat însă, că performanțele nu sunt cele mai bune, și există mult spațiu de îmbunătățire, dar consider că s-au făcut pași într-o direcție bună de cercetare, și această lucrare poate constitui baza unor proiecte interesante în viitor.

Bibliografie

- [1] Kai Chen, Yi Zhou and FangyanDai —A LSTM-based method for stock returns prediction: A case study of China stock market, IEEE International Conference on Big Data,2015.
- [2] Stock Market's Price Movement Prediction with LSTM Neural Networks: David M. Q. Nelson, Adriano C. M. Pereira, Renato A. de Oliveira
- [3] Stock Price Prediction Using Attention-based Multi-Input LSTM: Hao Li, Yanyan Shen, Yanmin Zhu
- [4] Stock Market Prediction based on Deep Long Short Term Memory Neural Network Xiongwen Pang1, Yanqiang Zhou1, Pan Wang2, Weiwei Lin3 and Victor Chang4
- [5] Stock Price Prediction on Daily Stock Data using Deep Neural Networks Sneh Jain, Dr. Roopam Gupta, Dr. Asmita A. Moghe
- [6] Predicting Stock Prices using LSTM - Murtaza Roondiwala, Harshal Patel, Shraddha Varma
- [7] <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [8] <https://adventuresinmachinelearning.com/recurrent-neural-networks-lstm-tutorial-tensorflow/>
- [9] "All About Market Indicators": Michael Sincere
- [10] The Fundamental Analysis: An Overview Sónia R. Bentes, Raúl Navas