

UNIVERSITATEA “ALEXANDRU IOAN CUZA” IAȘI

FACULTATEA DE INFORMATICĂ



LUCRARE DE LICENȚĂ

Compararea modelelor statistice în predicția comportamentului criptomonedelor

Chelmuș Rareș-Ștefăniță

Sesiunea: iulie, 2018

Coordonator științific:

Daniela Gîfu

DECLARAȚIE DE CONSIMȚĂMÂNT

Prin prezenta declar că sunt de acord ca Lucrarea de licență cu titlul „*Compararea modelelor statistice în predicția comportamentului criptomonedelor*”, codul sursă al programelor și celelalte conținuturi (grafice, multimedia, date de test etc.) care însoțesc această lucrare să fie utilizate în cadrul Facultății de Informatică.

De asemenea, sunt de acord ca Facultatea de Informatică de la Universitatea „Alexandru Ioan Cuza” din Iași, să utilizeze, modifice, reproducă și să distribuie în scopuri necomerciale programele-calculator, format executabil și sursă, realizate de mine în cadrul prezentei lucrări de licență.

Iași, *data*

Absolvent *Prenume Nume*

(semnătura în original)

Cuprins

| | |
|--|----|
| Introducere | 6 |
| Contribuții | 11 |
| 1. Abordare practică | 12 |
| 2. Detalii despre aplicație | 15 |
| 2.1 Set de date | 15 |
| 2.1.1 Date de tip text | 15 |
| 2.1.2 Date de tip numeric | 17 |
| 2.2 Fluxul de lucru | 19 |
| 2.3 Algoritmii de învățare automată..... | 25 |
| 2.3.1 Regresie Liniară | 25 |
| 2.3.2 Random Forests | 26 |
| 2.3.3 Random Forests cu Analiza Sentimentala | 26 |
| 3. Rezultatele comparării modelelor..... | 31 |
| 4. Discuții..... | 34 |
| 5. Concluzii și direcții viitoare de cercetare | 36 |
| 6. Bibliografie | 38 |

Compararea modelelor statistice în predicția comportamentului criptomonedelor

Chelmuș Rareș-Ștefăniță

Lucrare de licență, Iași, 2018

Structura Licenței:

Introducere, Contribuție, Rezulate, Discuții, Concluzii si cercetări viitoare

Cuvinte Cheie:

Criptomonedă, învățare automată, random forests, regresie liniară, analiza sentimentelor, NLP

Aria de studiu:

Informatică

Scopul studiului:

Compararea modelelor de învățare automată în scopul predicției criptomonedelor, optimizarea setului de date pentru antrenarea modelelor și realizarea predicțiilor în timp real prin intermediul aplicației.

Elementele de inovație si cercetare științifică:

Noutatea în această lucrare este dezvoltare unei formule ce ajută la completarea eficientă a golurilor din setul de date de antrenare pentru algoritmi de învățare automată si compararea cu metode de padare descoperite anterior.

Utilitatea practică a lucrării:

Aplicația este destinată amatorilor și persoanelor entuziasmate de subiectul criptomonedelor, oferind predicții orientative pe viitorul apropiat în legătură cu comportamentul lor financiar (creșterile și scăderile viitoare), astfel reducând riscurile personale economice în posibilele tranzacții și investiții.

Introducere

În această lucrare prezentăm o unealtă pentru realizarea de predicții pentru piața de criptomonede, aplicația fiind destinată antreprenorilor și entuziaștilor de criptomonede ce au nevoie de îndrumare în vederea investițiilor. Aplicația folosește algoritmi de învățare automată și de analiză a sentimentelor pe articole de pe canale online de specialitate. Interoperabilitatea celor două tipuri de algoritmi au ca scop creșterea acurateții predicției, deoarece sentimentul creat în jurul unei monede este influențat de ceea ce se publică în media de specialitate și prin urmare modifică decizia celor ce investesc în criptomoneda respectivă.

Istoria umană este plină de tentative de predicții și de profeții [Nelson, 2000], dar abilitatea de anticipa evenimente viitoare în evoluția diferitelor domenii, în special în economia finanțelor [Gifu and Cristea, 2012], este o sarcină complexă ce poate fi rezolvată folosindu-ne de practici ce înglobează tehnici din aria Inteligenței Artificiale (AI). Predicția unei piețe financiare (scopul acestei lucrări) rămâne o provocare importantă cu multe implicații în dezvoltarea unei economii sigure (pentru state, în cadrul companiilor, a organizațiilor). De-a lungul anilor, observând și studiind grafice și date, economiștii dezvoltat tehnici pentru a-și face o idee în vederea creșterii și scăderii prețurilor în piață (ex.: Martingales) [Feller, 1971]. Precizia predicției depinde de înțelegerea factorilor raționali (ex.: investesc când prețul e scăzut, vând când prețul e ridicat) și emoționali (ex.: simt că această companie va fi de succes) ce influențează piața de stocuri. În acest sens, sunt o mulțime de variabile de luat în calcul: impactul mass media, comportamentul investitorilor, istoria prețurilor, calamități naturale, decalajul cauzat de fusuri orare, climatul politic al statelor implicate în piață etc. Acești factori pot afecta sistematic fluctuațiile piețelor financiare, astfel din cauza numărului mare de variabile, specialiștii în finanțe și în big data consideră că încercarea de investi conform predicțiilor și a investi într-o manieră complet aleatorie va conduce la același rezultat. Această perspectivă este cunoscută drept EMH (Efficient Market Hypothesis)¹ sau RWH (Random Walk Hypothesis) [Fama, 1965; Jonathan Clarke et al., 2001; Zunino et al., 2012; Bariviera et al., 2014; Marwala, 2015; Marwala and Hurwitz, 2017]. Începând cu doctoratul lui Eugene Fama (1965), EMH a devenit una dintre cele mai

cunoscute teorii în economia financiară care confirmă conexiunea dintre prețuri și informația publică.

Pentru această lucrare am ales ca piață de studiu și de predicție piața de criptomonede. De ce? Deoarece Bitcoin (BTC), Ethereum (ETH), Ripple (XRP) și restul criptomonedelor și a token-urilor reprezintă viitorul lumii valutare. Mai mult, în ultimul an, piața cryptovalutelor a crescut de la o valoare totală estimată la 16 miliarde de dolari (10 ianuarie 2017) până la vârful înregistrat pe 7 ianuarie 2018 de 829 de miliarde de dolari. Datorită numărului mare de tranzacții, în fiecare zi sunt obținute un volum enorm de date (text și grafice) care în mare pot fi folosite de către publicul larg în scopuri diverse. În vederea predicției pieței de criptomonede, o combinație de modele matematice cu modele psihologice s-a dovedit a fi eficientă în acest sens, rezultatele fiind mult mai precise decât folosirea singulară a modelelor matematice. Deciziile masei de oameni care investesc în astfel de bunuri sunt influențate puternic de ceea ce se scrie în mass media despre fiecare monedă în parte, creîndu-se un sentiment în jurul fiecărei monede virtuale, astfel, componenta NLP se dovedește a fi indispensabilă în procesul de predicție.

Motivație

Bitcoin, Ethereum, Ripple, Monero, Electroneum și lista continuă, pe toate le-am urmărit de la apariția lor până în prezent, visând perioade bune la ferme de GPU-uri și ASIC-uri care minează non-stop aceste monede. Mereu am fost fascinat de aceste cryptovalute, odată ca și concept, datorită faptului că la ora actuală este cea mai performantă metodă de schimb și troc, asemănându-se cu valutele folosite în filmele și romanele din genul science-fiction și odată ca și idee de minat (procesul prin care se securizează rețeaua și se obțin monede noi „practic, recompensa pentru că ți-ai oferit puterea computațională pentru securizarea rețelei”), ceea ce m-a dus mereu cu gândul la a avea mini fabrici ce produc precum în jocurile RTS (real time strategy).

Cu timpul, studiind graficele și fluctuațiile haotice ale pieței de criptomonede, am început să observ un soi de patternuri: monedele au o perioadă de maturizare, „liniștea” de 7-9 luni (perioadă în care valoarea monedei scade sub prețul din primele săptămâni de la lansare și se

menține într-o plajă redusă de preț), exploziile după perioadă de „liniște” (perioadă scurtă în care criptomoneda înregistrează creșteri foarte mari). Toate aceste patternuri m-au condus la întrebarea: *Oare putem prezice comportamentul lor?* De atunci am început să studiez fenomenul și tehnicile de predicție, lucru ce mă fascinează și în prezent.



Figura 1. Logourile monedelor aflate în top trei în iunie 2018: Bitcoin (BTC), Ethereum (ETH) și

Lucrări anterioare

Odată cu dezvoltarea tehnologică în domeniul informatic, o colecție de algoritmi de inteligență artificială au fost utilizați pentru a prezice fluctuațiile piețelor. O parte dintre ele au fost dezvoltate prin observații, precum mean reversion, proces bazat pe presupunerea că stocurile vor atinge valoarea medie în timp [Spierdijk and Bikker 2012; Dai et al., 2013]. Oricum, problema care rămâne este că nu poți prezice când sau cât timp vor rămâne prețurile pe valoarea medie. Cunoaștem faptul că valoarea stocurilor fluctuează într-o manieră aleatorie [Granger, 1992], cu posibilități de a prezice cu probabilități ridicate de succes în anumite momente ale graficului.

În cazul unui grafic liniar, am descoperit că regresia liniară poate fi folosită în mod intuitiv, fiind printre primele încercări în a anticipa creșterile sau scăderile într-un model liniar practic precum piața de stocuri [Xin, 2009]. Acest algoritm, unul dintre cele mai cunoscute în machine learning, a fost folosit în rezolvarea problemelor de predicție datorită naturii setului de date folosit în piața de stocuri (grafic, CSV) [Nunno, 2017].

Alt algoritm de machine learning folosit în reducerea riscului privind investițiile în piața de stocuri este Random Forests. Acest model construiește arbori de decizie pe baza eșantioanelor aleatorii din setul de date de antrenament și formează rezultatul din „votul” majoritar al arborilor. Algoritmul a dat roade în tranzacțiuni datorită faptului că această abordare elimină problema de overfitting al arborilor, însă eficiența a fost înregistrată pe perioade lungi [Khaidem et al., 2009] și pe piața de stocuri. Se cunoaște că piața de stocuri este mult mai stabilă față de piața de criptomonede iar piața de stocuri are o existență mult mai îndelungată față de cea a criptomonedelor (aproximativ 4 ani).

Problema principală a predicției comportamentului financiar al criptomonedelor este dat de perioadă încă incipientă a dezvoltării și existenței lor, perioadă în care mass media are un cuvânt solid de spus în influențarea deciziilor celor implicați în tranzacționarea valutei virtuale. Influența canalelor de media a fost observată și în piața stocurilor [Barber & Odean, 2008], însă în cazul criptomonedelor, mecanismul de tranzacție este foarte simplificat, tranzacțiile putând fi realizate cu ușurință de oriunde și oricând, astfel, utilizatorii pot acționa în virtutea sentimentului.

În general, analiza sentimentelor/opiniilor (SA), metodă din aria NLP (procesarea limbajului natural), este o abordare populară folosită în algoritmi de predicție. Hilbert consideră că știrile negative cresc probabilitatea ca prețurile stocurilor să scadă. Totodată, știrile pozitive ce fac referire la un bun anume au crescut probabilitatea că valoarea bunului respectiv pe piața de stocuri să crească [Hilbert et al., 2014].

Aplicația realizată va folosi trei abordări în vederea predicției criptomonedelor: regresia liniară, random forests și random forests folosind analiza sentimentelor. În capitolele

următoare vor fi descriși algoritmi aleși, motivele pentru care au fost aleși acești algoritmi, rezultatele și concluziile aferente.

Lucrarea este structurată în patru secțiuni:

- **Abordare practică** – prezintă planul în care este aplicată teoria.
- **Detalii despre aplicația** – sunt expuse componentele, fluxul datelor în interiorul aplicației de predicție și utilizarea ei.
- **Rezultate** – rezultatele obținute în urma rulării testelor
- **Discuții** – explicarea rezultatelor prezentate în capitolul anterior
- **Concluzii și direcții viitoare de cercetare** – sunt prezentate concluziile lucrării de licență și subiectele ce doresc să le cercetez și să le aprofundez în viitor.

Contribuții

În acesta secțiune voi prezenta contribuțiile la nivel teoretic și practic aduse de mine în vederea optimizării procesului de predicție și totodată compararea performanțelor în această arie dintre modelele de învățare automată implementate în aplicație.

De asemenea, voi descrie setul de date fosite de către aplicație și metodele utilizate în procesul de predicție a pieții de criptomonede, pentru a folosi cazuri practice, vom exemplifica și discuta metodele folosind trei monede virtuale pe baza cărora vom construi modelele noastre de predicție.

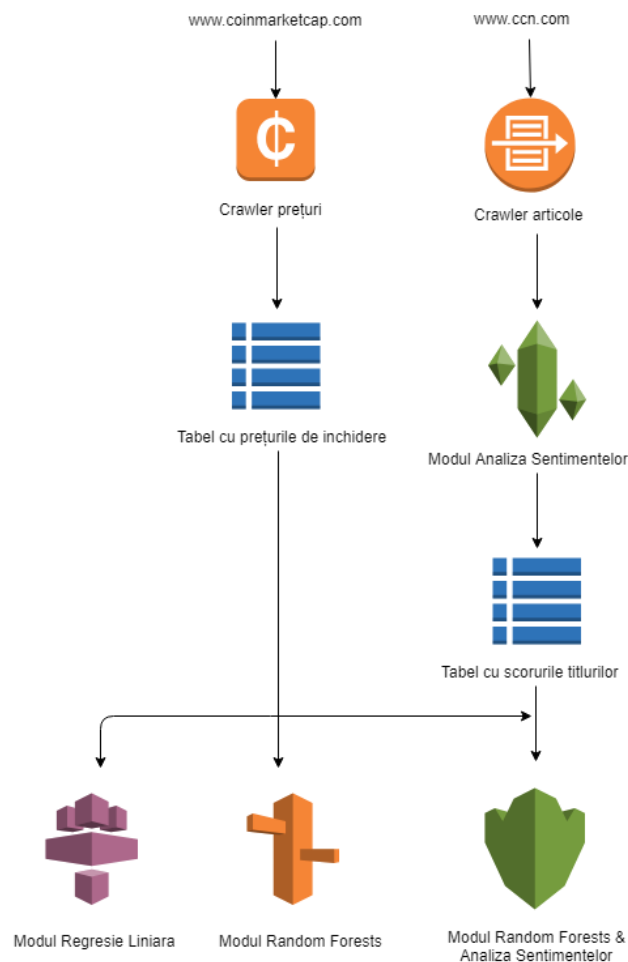


Figura 2. Arhitectura procesului de predicție

1. Abordare practică

Aplicația în sine a fost construită cu scopul de a analiza comparația dintre predicțiile produse de trei algoritmi de învățare automată și de asemenea pentru a realiza predicții în timp real pe o perioadă maximă de 10 zile. Perioada menționată a fost aleasă în acest sens deoarece ne putem crea o imagine orientativă a pieței pe o perioadă scurtă mult mai acurată decât dacă am anticipa prețurile pentru o perioadă de o lună sau mai mult.

Pentru exemple am ales ca monede de studiu criptomonede aflate în top cinci pe piața de criptomonede, deoarece aceste monede sunt considerate ca fiind mature. Matur este folosit în sensul că acestea au o perioadă de existență mai mare de un an și pentru că au o activitate considerabilă în materie de tranzacții (setul de date este diversificat). Volumul de tranzacții și desigur, notorietatea lor, sunt direct proporționale cu locul pe care îl ocupă în clasament. Mai mult, mass media urmărește fidel cum fluctuează aceste monede în timp și orice activitate din partea guvernelor, companiilor și a organizațiilor în legătură cu ele, activități precum regulări ale pieței, taxe și impozite impuse, interzicerea comercializării, închiderea site-urilor de tranzacții, opiniile privind anticiparea financiară a monedelor din partea investitorilor și a influencerilor, etc.

Pentru analiza sentimentala, am folosit biblioteca de Python, NLTK (Natural Language Toolkit), iar perioada din care s-au recoltat date pentru a realiza testele este de 3 luni, din 15 Octombrie 2017 până la 05 Ianuarie 2018. Motivul pentru care s-a ales o perioadă așa de scurtă este din cauza faptului că piața de criptomonede încă se află în curs de maturizare, canalele specializate pe piața criptovalorilor a avut o activitate intensificată începând cu Octombrie 2017 iar pe o perioadă de un an prețurile diferă foarte mult (ex: prețul unui Bitcoin la 1 Ianuarie 2017 era de 998,33 dolari valoarea de încheiere pe ziua respectivă iar la data de 17 Decembrie 2017 – moment în care s-a înregistrat vârful istoric al monedei – un Bitcoin valora 20.089 de dolari), reflectând volatilitatea accentuată a pieței și totodată datele respective pot afecta negativ antrenarea algoritmilor de învățare automată.

Bitcoin Charts



Figura 3. Graficul criptomonedei Bitcoin (BTC) în intervalul 1 ianuarie 2017 – 1 Ianuarie 2018

Ethereum Charts



Figura 4. Graficul pentru Ethereum (ETH) în intervalul 1 ianuarie 2017 – 1 Ianuarie 2018



Figura 5. Graficul criptomonedei Ripple (XRP) în intervalul 1 ianuarie 2017 – 1 Ianuarie 2018

După cum observăm în figurile 3,4 și 5, monedele urmează trenduri în anumite perioade ce nu se mai repetă, astfel, datele anterioare ar „păcăli” algoritmi de învățare automată.

2. Detalii despre aplicație

Aplicația oferă utilizatorului posibilitatea de a testa trei tipuri de algoritmi de predicție: regresie liniară, random forests și random forests ajutat de analiza sentimentală. Primele două abordări folosesc un singur set de date (istoricul prețurilor monedelor) iar a treia metodă folosește două seturi de date (istoricul prețurilor și scorurile obținute în urma analizei de sentimente). Poate compara performanțele anterioare ale predicțiilor, specificând moneda și datele calendaristice între care utilizatorul dorește să facă antrenarea și predicția. Totodată, poate verifica în timp real predicții pe următoarele 10 zile.

2.1 Set de date

Se folosesc două tipuri de colecții de date: tabela de prețuri cu care se încheie fiecare zi. Primul set reprezintă baza pe care se antrenează fiecare algoritm, iar al doilea set ne ajută să mărim performanța predicției. Datele sunt minate de pe internet de pe siteuri specializate în aria de cryptocurrency.

2.1.1 Date de tip text

Setul de date de tip text sunt formate din titluri ale articolelor ce fac referire la criptomoneda pe care dorim să îi construim graficul de predicție. Titlurile articolelor sunt preluate de pe www.ccn.com folosind un crawler propriu, articolele sunt alese dacă conțin referiri la moneda aleasă. Am ales acest site deoarece are un trafic mare de utilizator, influent astfel un segment larg din aria entuziaștilor de criptomonede. De ce am ales să procesam date de acest tip doar dintr-o singură sursă? Odată că aceleași știri vor apărea și pe alte site-uri de specialitate, în acest fel datele procesate nu vor fi redundanțe și totodată site-ul postează știrile dintr-o perspectivă neutră.

Am evitat să preluăm titluri de la surse precum Youtube, Reddit sau Twitter deoarece informațiile vin în majoritatea cazurilor de la surse nesigure și biasate. Chiar dacă cantitatea de informații este abundentă pentru fiecare monedă (știrile se răresc pentru monezile ce coboară în clasament sau prea noi) riscăm să procesam informații false, destinate să provoace FUD (Fear,

Uncertainty and Doubt – termen folosit în argoul lumii criptovalutelor, făcând referire la știri sau informații publicate cu scopul de a răspândi nesiguranță și frică referitor la investirea într-o anumită criptomonedă).

În procesul de analiza sentimentală intră doar titlul articolului deoarece în titlu se află sentimentul principal (ca articolul e unul negativ, pozitiv sau neutru), analiza întregului text fiind problematică și costisitoare din punct de vedere al spațiului și al timpului de procesare.

Articolele ce privesc moneda căreia vrem să îi construim graficul de predicție nu apar în fiecare zi, sunt zile în care nu se scrie nimic despre moneda respectivă, din această cauză rămân goluri în setul de date. Problema va fi rezolvată folosindu-se o formulă dezvoltată de mine ce va aproxima pentru zilele cu goluri (le vom denumi blancuri). Totodată, există zile în care apar mai mult de un articol, problemă ce va fi rezolvată prin altă formulă ce va aproxima sentimentul predominant pentru ziua .

| Data | Title | N words per title |
|----------------|---|------------------------------|
| 15/10/2017 | Hours to go: How to Watch Ethereum's Fork as It Happens | 11 |
| 16/10/2017 | Bulls Take Breather? Bitcoin Slows as Price Struggles to Breach \$6,000 | 11 |
| ... | ... | ... |
| 1/01/2018 | What Will the Bitcoin Price Be in 2017? | 8 |
| 02/01/2018 | Is It Too Late To Buy Bitcoin? Video: \$1 Million? Bitcoin Sign Guy on Why It's Not Too Late to Buy | 21 |
| 03/01/2018 | RSK Beta Brings Ethereum-Style Smart Contracts Closer to Bitcoin | 10 |
| 04/01/2018 | Ripple Fever? Other Crypto Assets Are Outpacing Its 2018 Gain | 10 |
| 05/01/2018 | Ethereum Price Highs Overshadow New Wave of Tech Issues | 9 |
| Total | | 4473 |
| Average | | 8.89 |

Figura 6. Exemplu de set de date extrase de pe www.ccn.com

2.1.2 Date de tip numeric

Tabelul cu prețuri este luat de pe www.coinmarketcap.com cu ajutorul unui web crawler. Site-ul menționat este cel mai frecventat site de statistică din aria criptomonedelor, are indexate toate criptomonedele existente și de asemenea este site-ul cu cele mai multe informații istorice ale evoluției monedelor. Informațiile extrase de către crawler sunt data și cu prețul de încheiere pe ziua respectivă. Pe site sunt găsite pentru fiecare zi prețul cu care începe în ziua respectivă, minimumul și maximumul zilei și cu prețul de încheiere al zilei.

Am ales doar ultimul din șir deoarece alte articole pot apărea până la sfârșitul zilei, astfel să putem corela toate articolele apărute din ziua respectivă cu prețul final. Totodată, e mult mai dificil să prezicem dacă avem prea multe atribute la antrenament și evităm overfittingul.

Market Cap ▾ Trade Volume ▾ Trending ▾ Tools ▾

Search Currencies

Q

All ▾ Coins ▾ Tokens ▾ USD ▾

Next 100 → View All







| ▲# | Name | Market Cap | Price | Volume (24h) | Circulating Supply | Change (24h) | Price Graph (7d) |
|----|--|-------------------|-------------|------------------|----------------------|--------------|---|
| 1 |  Bitcoin | \$196,978,854,505 | \$11,717.80 | \$19,004,600,000 | 16,810,225 BTC | 13.45% |  |
| 2 |  Ethereum | \$101,696,586,391 | \$1,047.81 | \$8,251,620,000 | 97,056,324 ETH | 16.55% |  |
| 3 |  Ripple | \$60,248,664,465 | \$1.56 | \$9,118,600,000 | 38,739,142,811 XRP * | 46.66% |  |

Figura 7. Topul celor 3 mai tranzacționate monede virtuale

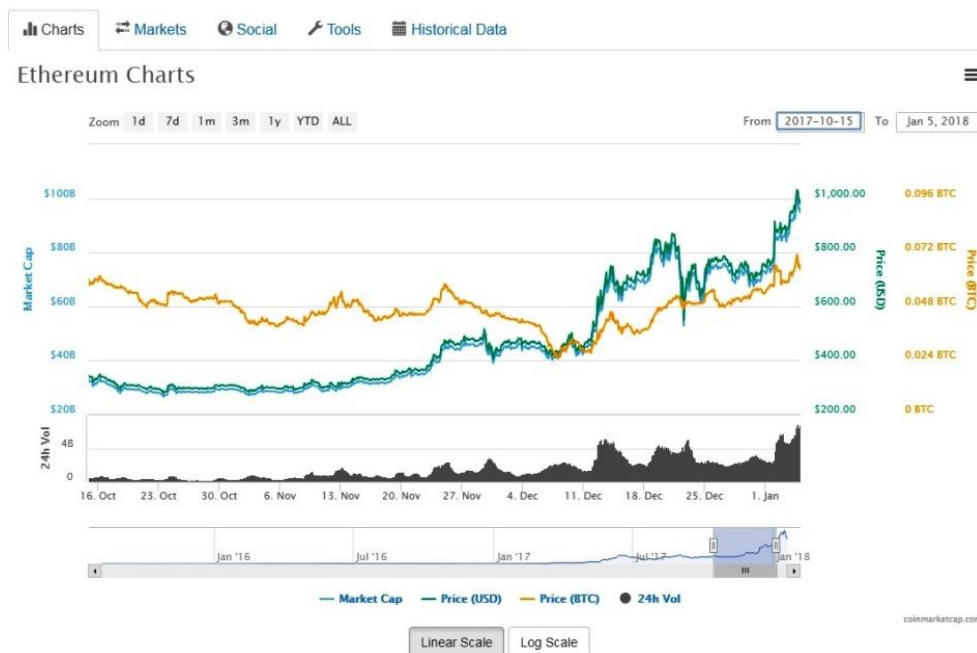
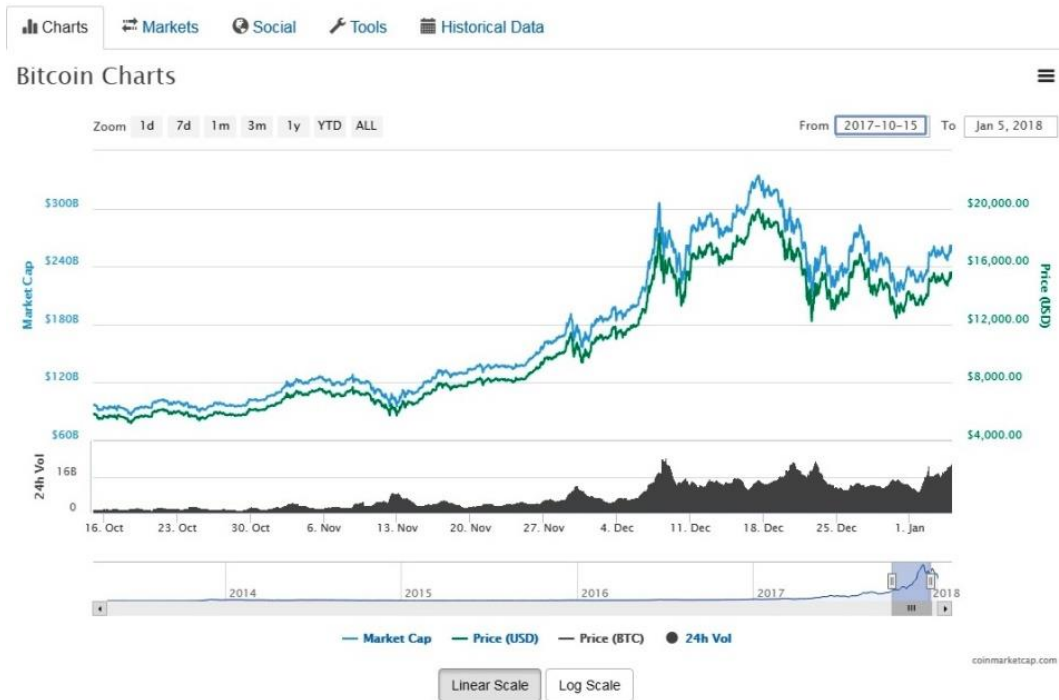




Figura 8. În cele 3 imagini se poate observa volatilitatea monedelor din topul clasamentului, pentru Bitcoin (BTC), Ethereum (ETH) și Ripple (XRP), în ordinea menționată, pe perioadă 15 Octombrie 2017 – 5 Ianuarie 2018

2.2 Fluxul de lucru

Fluxul de lucru începe cu colectarea de date care este realizată de două crawlere ce minează datele de pe site-urile specificate anterior. Un crawler extrage informațiile de tip text iar celălalt își adună informațiile privind prețurile. Informațiile extrase sunt filtrate pentru a putea fi procesate de către algoritmi de predicție. Informațiile extrase sunt stocate sub formă de dicționar pentru a putea fi identificate și accesate facil.

Următorul pas este să trecem titlurile articolelor extrase ce fac referire la moneda aleasă prin modulul `SentimentIntensityAnalyzer` din biblioteca NLTK în Python, aceasta fiind componenta de analiză de sentimente. Titlu va fi anotat automat și va fi trecut printr-un algoritm de detecție a polarității din care selectăm atributele de pozitiv (`sentiment_scores["pos"]`) și negativ (`sentiment_scores["neg"]`) conform codului de mai jos:

```

sentiment_module=SentimentIntensityAnalyzer()
pos = 0
for i, item in data_set.iterrows():
    try:
        sentiment_scores = sentiment_module.polarity_scores(item["title"])
        score = sentiment_scores["pos"] - sentiment_scores["neg"]
        scores.set_value(pos, 'score', score)
    except TypeError:
        pass
    pos += 1

```

Filtrul seta o valoare în intervalul [0,1] pentru scorul pozitiv (în ce proporție titlu pare să confere un sentiment pozitiv) și o valoare în același interval pentru cel negativ (cât de mult pare a fi sentimentul negativ extras din text). Scorul final pentru fiecare titlu extras va fi calculat după formula:

$$final_score = (sum(positive_score) - sum(negative_score)) / number_of_articles$$

(1)

În formula (1) funcția sum() este folosită pentru rezolva problema articolelor multiple apărute într-o singură zi, răspunzând la întrebarea: *Care sentiment este dominant?* . Media sumei scorurilor negative minus scorurile pozitive împărțit la numărul de articole ne conduce spre sentimentul predominant din ziua respectivă.

| Data | Scores |
|-------------|---------------|
| 15/10/2017 | 0 |
| 16/10/2017 | -0.157 |
| 16/10/2017 | -0.149 |
| 16/10/2017 | 0 |
| 17/10/2017 | -0.192 |
| 18/10/2017 | 0 |
| 18/10/2017 | -0.208 |
| 18/10/2017 | 0 |
| 26/10/2017 | 0.426 |
| 26/10/2017 | -0.216 |
| ... | ... |

Figura 9. Tebelul rezultat în urma procesării datelor prin modulul NLTK de analiză a sentimentelor

Distingem în tabelul de la Figura 9. două situații distincte: perioade de timp în care nu există informații, din cauza faptului că nu s-a scris nimic în ziua respectivă despre monedă (ex.: perioada dintre 18/10/2017 – 26/10/2017) și intrări multiple pentru o singură dată calendaristică (ex.: pentru data 16/10/2017 există trei intrări cu scoruri diferite, semnificând apariția a trei articole pe site ce fac referire la moneda căutată).

Din moment ce articole nu apar zilnic în legătură cu moneda vizată, în tabelul de date apar goluri ce necesită a fi approximate cu valori. Goel de la Universitatea Stanford [Mittal and Goel, 2012] a umplut aceste goluri folosind următoarea formulă (2):

$$gap = (x + y) / 2 \quad (2)$$

În formulă, „gap” este data calendaristică ce nu are intrări în tabel (aflată în intervalul căutat), x și y sunt scorurile analizei sentimentale dintre două zile între care există goluri de date, x aparținând zilei de început și y e valoarea scorului pentru ziua de final. Între x și y sunt un număr n de zile în care există goluri (ex.: pentru perioada 18/10/2017 – 26/10/2017,

x este valoarea scorului de pe data 18/10/2017, y este pentru 26/10/2017 iar n este egal cu 7, semnificând numărul de zile în care nu s-au publicat articole despre moneda vizată).

Pentru a realiza o aproximare pentru aceste goluri mult mai precisă, am modificat formula lui Goel astfel:

$$gap = (x + y) / (\sqrt{n + 1} + 1) \quad (3)$$

Alterarea produsă schimbă modul în care se comportă graficul scorurilor analizei sentimentale, reflectând o evoluție mai naturală de la x la y. Din prismă empirică, algoritmul de predicție răspunde mai bine la această schimbare în formulă. Remarcăm trei cazuri în care se poate comporta graficul formulei. În grafice vom lua x și y în diferite cazuri în timp ce n (numărul de zile fără scor) va fi egal cu 5 (Figurile 10,11 și 12):

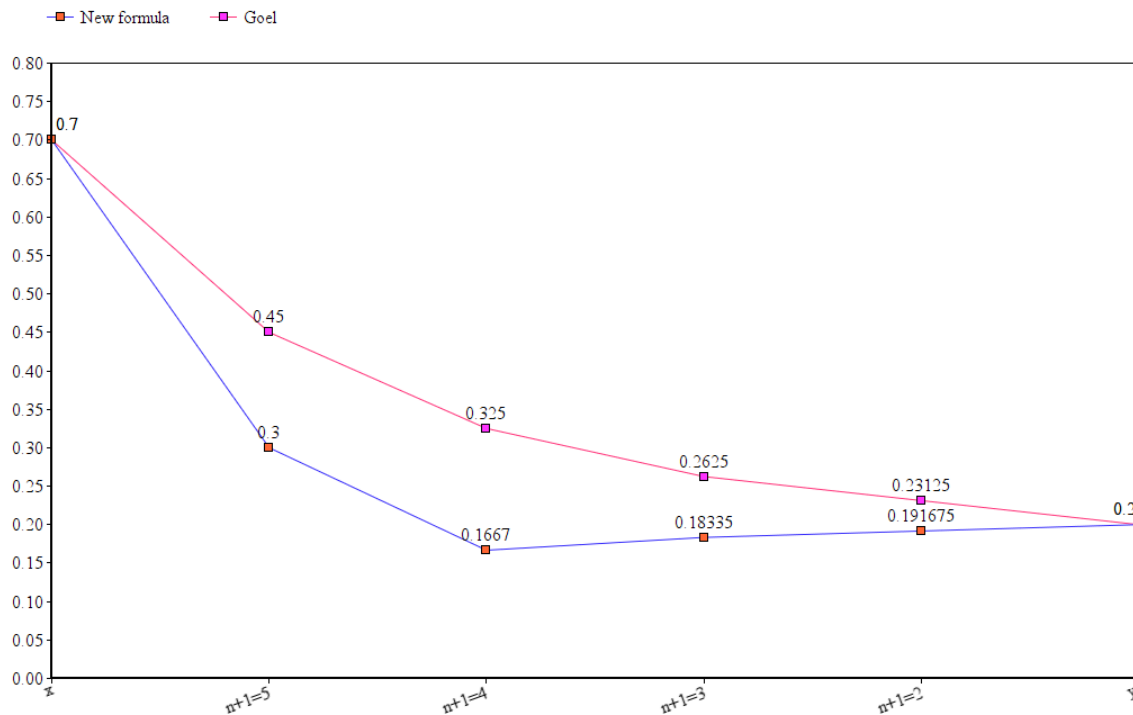


Figura 10. Cazul 1: $x > y$

În figurile 10,11 și 12 se conturează diferența dintre graficul funcției lui Goel (linia roșie) și graficul funcției noastre (linia albastră) [Chelmuș R., Gîfu D. and Iftene S., 2018]. În figura 10 avem analiza primului caz în care x (0,7) este mai mare decât y (0,2). Formula lui Goel scade gradual fără să fie sub valoarea lui y . În schimb, în formula noastră, valorile dintre x și y sunt mai mari decât y ($0,3 > y$) și totodată mai mici ($0,1667 < y$), notând o dinamică a sentimentului mai naturală (există și creșteri și scăderi în valori, nu doar scăderi precum în prima formulă).

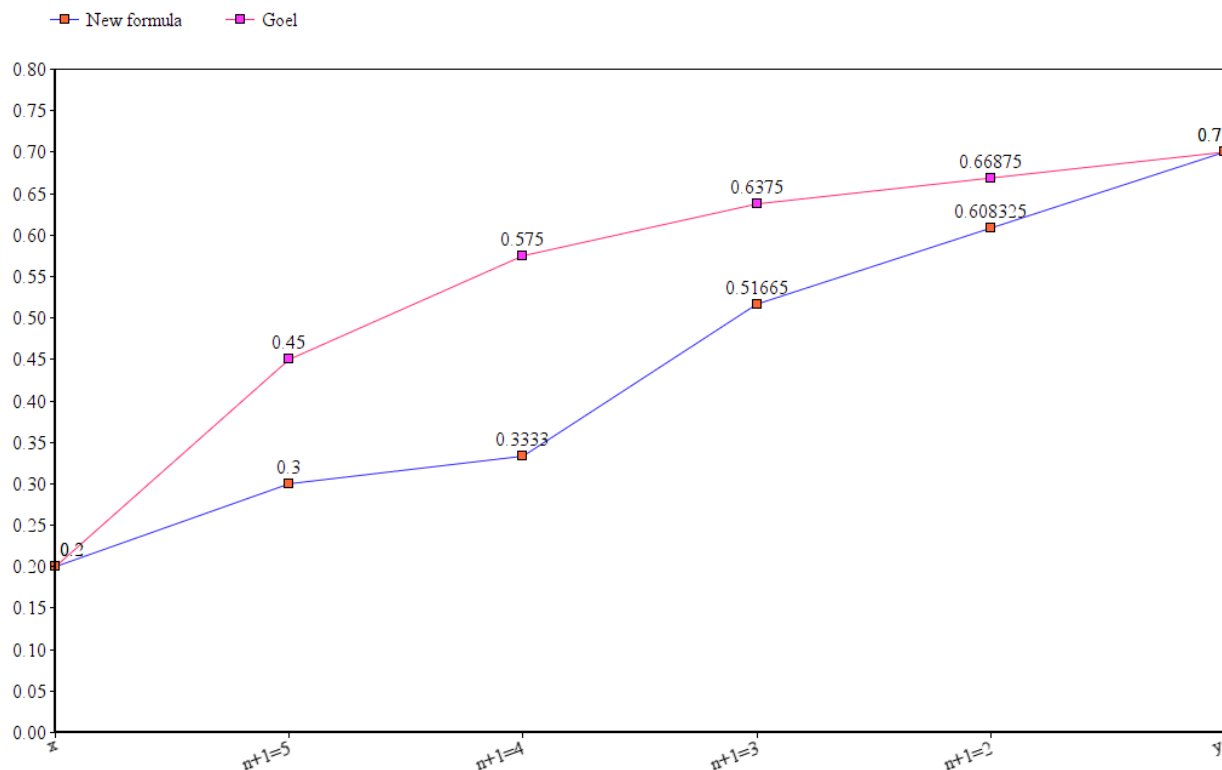


Figura 11. Cazul 2: $x < y$

În cazul numărul doi (figura 11.), notăm faptul că formula dezvoltată nu înregistrează creșteri abrupte precum funcția lui Goel ($0,45 > 2 \cdot x$), funcția noastră generând mai multă diversitate în setul de date.

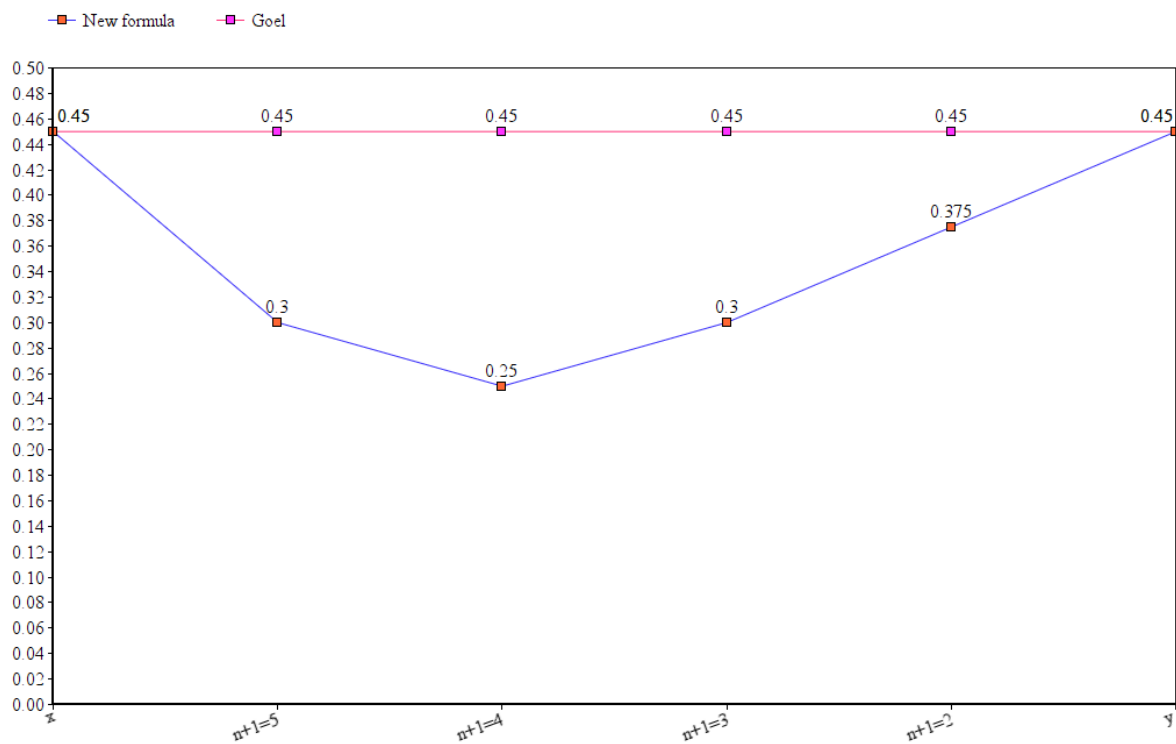


Figura 12. Cazul 3: $x = y$

În ultimul caz (figura 12.) observăm că pentru x egal cu y , formula lui Goel nu înregistrează nicio schimbare, o consecință al acestui fenomen este că antrenarea cu un set de date generat în acest mod va produce performanțe scăzute în predicția monedei virtuale.

| Data | Scores | Close |
|---------------------|---------------|--------------|
| 15/10/2017 00:00:00 | 0 | 336.6 |
| 16/10/2017 00:00:00 | -0.153 | 333.38 |
| 17/10/2017 00:00:00 | -0.192 | 317.08 |
| 18/10/2017 00:00:00 | -0.208 | 314.32 |
| 19/10/2017 00:00:00 | -0.0206 | 308.09 |
| 20/10/2017 00:00:00 | 0.00094 | 304.01 |
| 21/10/2017 00:00:00 | 0.0067 | 300.19 |
| 22/10/2017 00:00:00 | 0.00658 | 295.45 |
| 23/10/2017 00:00:00 | 0.00656 | 286.95 |
| 24/10/2017 00:00:00 | 0.00656 | 298.33 |
| 25/10/2017 00:00:00 | 0.02231 | 297.93 |
| 26/10/2017 00:00:00 | 0.105 | 0.105 |
| ... | ... | ... |

Figura 13. Setul de date rezultat în urma procesării titlurilor, scorurilor și completarea golurilor cu formula dezvoltată

2.3 Algoritmii de învățare automată

Am testat trei algoritmi de machine learning prin care punctăm eficiența folosirii componentei NLP în elaborarea predicțiilor pe piața criptomonetara. Algoritmii de învățare automată folosiți (linear regression, random forests) sunt implementați cu ajutorul librăriei de Python scikit-learn iar seturile de date sunt stocate și manipulate folosind librăriile NumPy și Pandas.

2.3.1 Regresie Liniară

În aria predicțiilor, regresia lineară este o metodă primitivă utilizată în anticiparea prețurilor în piața de stocuri, însă s-a dovedit a fi inefficientă din punct de vedere al performanței din cauza volatilității bunurilor. Formula (4) regresiei are în general următoarea formă:

$$y = b_0 + b_1 \times x \quad (4)$$

Unde y este valoarea dependentă (prețul pe care vrem să-l anticipăm), b_0 se numește termen liber, b_1 este panta graficului și x este variabila independentă.

În secvența de cod de mai jos, este prezentat modulul de antrenare al algoritmului de regresie liniară folosind funcții din biblioteca SciKit.

```
linear_mod = linear_model.LinearRegression()
#-----
dates = np.reshape(range(1, len(self.data_training["date"])+1),
                    (len(self.data_training["date"]), 1))
prices = np.reshape(self.data_training["close"],
                    (len(self.data_training["close"]), 1))
self.linear_mod.fit(dates, prices)
```

2.3.2 Random Forests

Algoritmul Random Forests antrenează un număr de arbori pe eșantioane selectate aleatoriu din setul de date. Rezultatul se obține din votul majoritar al predicțiilor arborilor de decizie. În acest caz, vom folosi doar prețul de încheiere pentru a antrena algoritmul, tot setul de date mai puțin ultimele 10 intrări (linii din tabel ce vor fi folosite pentru testare).

Porțiunea de cod următoare arată modul de antrenare a algoritmului Random Forests folosind SciKit. x_{train} și y_{train} sunt seturile de date folosite pentru a antrena arborii, în acest caz, x_{train} este format din zilele calendaristice iar y_{train} sunt prețurile ce corespund zilelor din x_{train} .

```
random_forest = RandomForestRegressor()
random_forest.fit(x_train, y_train)
```

2.3.3 Random Forests cu Analiza Sentimentelor

În acest caz, un modul Random Forests va fi antrenat cu setul textual de date iar al doilea modul Random Forests va fi antrenat cu setul de date numeric și cu datele de ieșire de la primul modul.

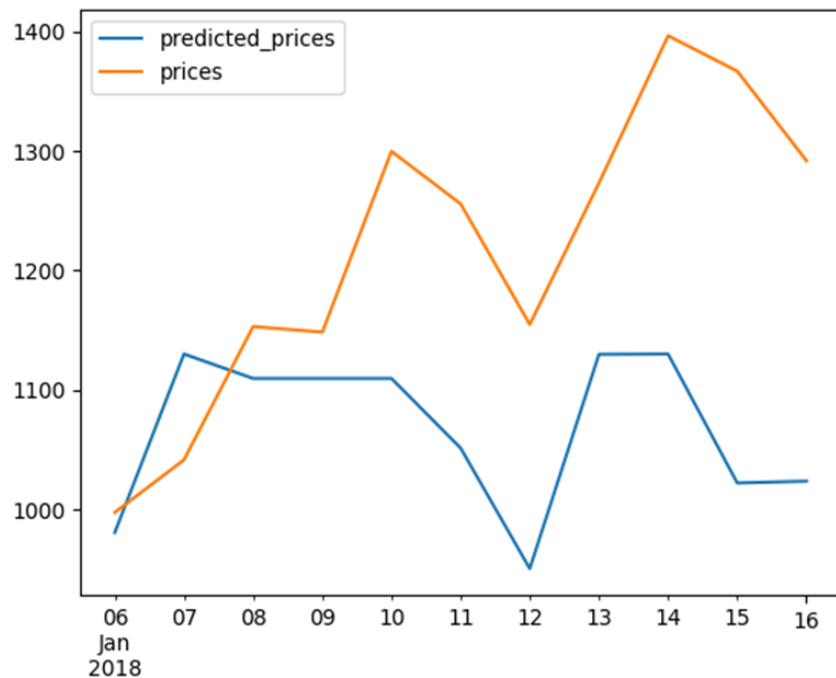


Figura 14. Graficul predicției pentru Ethereum (ETH) folosind Random Forests (RF) cu Analiza de Sentimentala (SA), utilizând formula dezvoltată

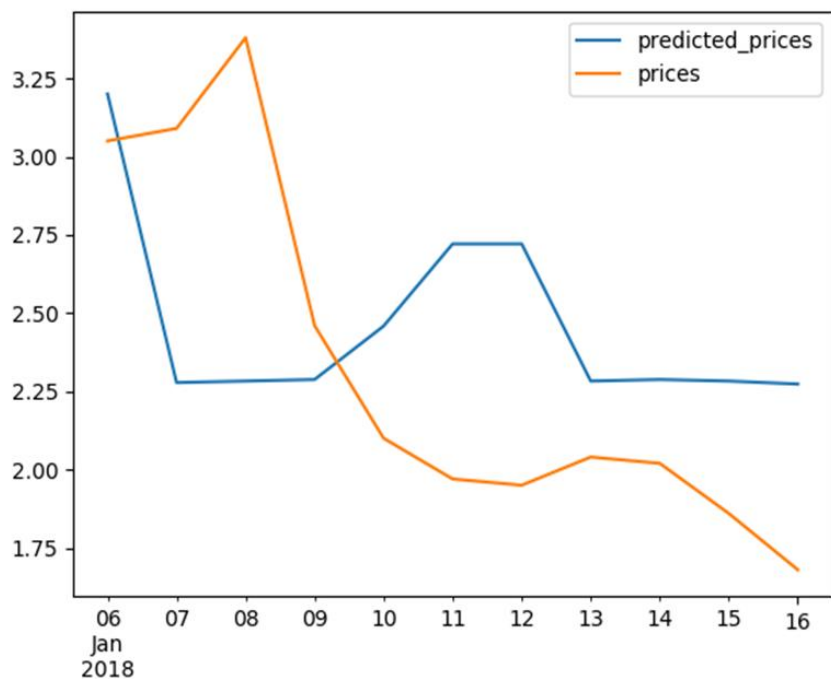


Figura 15. Graficul predicției pentru Ripple (XRP) folosind aceeași metodă din figura anterioară.

Graficele de la figurile 14 si 15 reprezintă outputul de la algoritmul de Random Forest antrenat cu cele două seturi de date. Linia portocalie din ambele părți reprezintă prețul real al monedei la data respectiva iar linia albastră este prețul prezis în intervalul precizat.

Fiecare rulare al modulului Random Forests va oferi un output ușor diferit, datorită proprietății nedeterminate al algoritmului de învățare automată.

2.4 Utilizarea aplicației

În această secțiune va fi prezentată interfața aplicației și instrucțiuni în utilizarea ei. Interfața aplicației a fost dezvoltată folosind biblioteca appJar (de pe appjar.info) pentru Python.

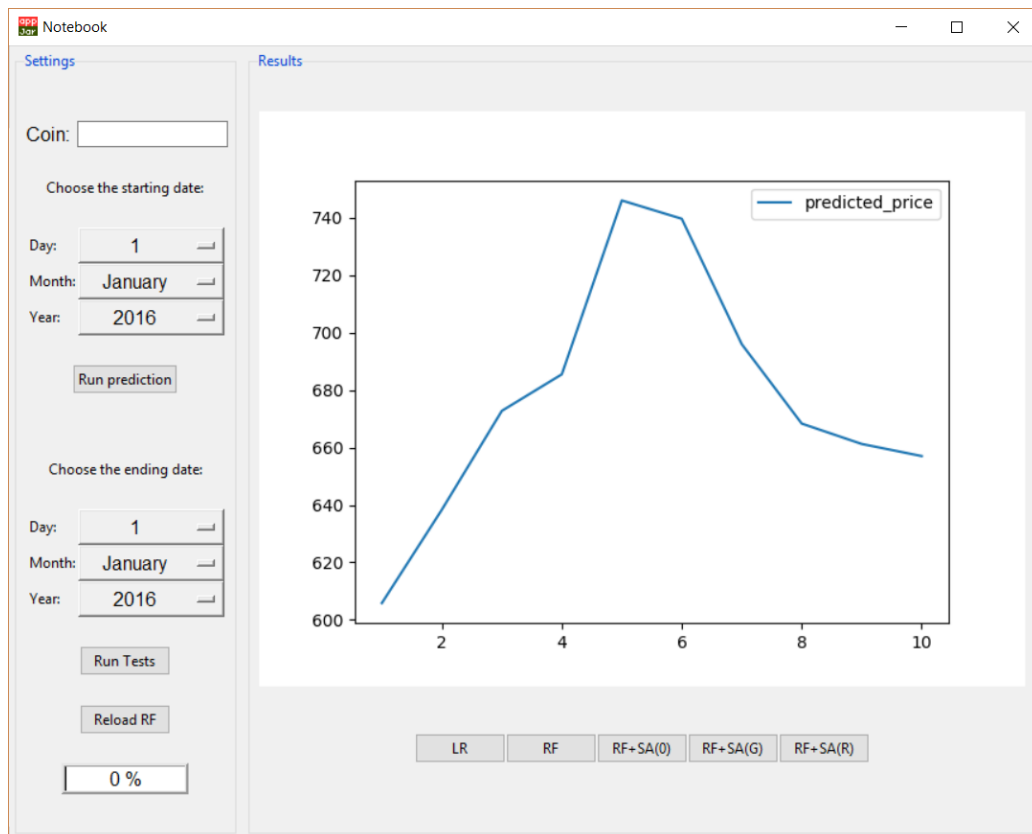


Figura 16. Meniul principal al aplicației

În figura 16 se află meniul aplicației și totodată singura pagină a sa. Tabloul aplicației se împarte în două containere: Settings și Results.

2.4.1 Settings

În „Settings” se stabilesc valorile de input pentru algoritmi de predicție. Câmpul „Coin” se va completa cu moneda pe care dorim să îi prezicem comportamentul financiar. Pentru a obține predicții viitoare de la data curentă, se va seta doar data de start de sub „Choose the starting date:” și se apasă pe butonul „Run prediction”. Starting date face referire la data de la care cawlere încep să adune date textuale și numerice pentru antrenare.

Pentru a rula verificări și teste ale algoritmului (compararea performanțelor trecute cu valori reale), se va seta și data de la „Choose the ending date:” și se va apăsa pe butonul „Run tests”.

Graficul predicțiilor pentru lgoritmi de tip Random Forests pot varia, astfel, pentru a rula din nou modulurile de Random Forests pe același set de date obținute, se va apăsa pe butonul „Reload RF”.

Fiecare rulare durează câteva secunde din cauza euristicilor implementate în cawlere pentru a nu fi respinși de siteuri (unele siteuri au mecanisme de detecție și respingere ale crawlerelor, pentru a nu îngreuna traficul) și de antrenarea arborilor. Finalizarea procesului de predicție va fi anunțat prin încărcarea completă a barei de încărcare din josul secțiunii de setări.

2.4.2 Results

În acest container se vor afișa rezultatele rularilor menționate mai devreme. Butoanele din partea inferioară a containerului sunt destinate pentru a schimba graficul predicției prezentat în cele cinci opțiuni:

- LR - Linear Regression;
- RF - Random Forests

- RF+SA(0) – Random Forests & Sentiment Analysis (cu padarea golurilor din setul textual cu 0)
- RF+SA(G) – Random Forests & Sentiment Analysis (cu padarea golurilor din setul textual cu formula lui Goel)
- RF+SA(R) – Random Forests & Sentiment Analysis (cu padarea golurilor din setul textual cu formula dezvoltată de mine)

3. Rezultatele comparării modelelor

Am evaluat modelele testate pe perioada de trei luni folosind măsurătorile de performanță: Precision, Recall și F-measure, astfel evaluând acuratețea predicțiilor. Algoritmii au fost rulați de 30 de ori și am considerat ca o predicție fiind bună dacă între două puncte consecutive ale graficului se respectă creșterea sau scăderea prețului prezentă în datele de test. Nu am luat în considerare dacă sunt atinse exact valorile din datele de test sau dacă se află într-un interval de toleranță dat deoarece măsurătorile ar fi mult mai mici iar scopul principal este să ne facem o idee dacă în următoarele zile vor exista scăderi sau creșteri, nu și prețul exact.

#aici pun grafice cu modelele

| Coin | Model | P | R | F-measure |
|------|------------------------------------|--------|--------|-----------|
| BTC | Random Forest | 75.40% | 65.37% | 66.37% |
| | Linear Regression | 62.45% | 60.08% | 61.24% |
| | SA with Random Forests Development | 75.23% | 73.67% | 74.44% |
| ETH | Random Forest | 72.11% | 60.33% | 61.69% |
| | Linear Regression | 62.71% | 55.11% | 56.85% |
| | SA with Random Forests Development | 72.12% | 71.55% | 71.83% |
| XRP | Random Forest | 68.23% | 61.42% | 61.32% |
| | Linear Regression | 59.89% | 54.78% | 55.82% |
| | SA with Random Forests Development | 68.45% | 67.01% | 67.72% |

În tabelul de mai sus am analizat performanțele celor trei modele prezente în aplicație.

În tabelul de mai jos am realizat un număr de teste pe același set de date pentru algoritmul Random Forest cu Analiza Sentimentelor în care setul de date textuale a fost padat în etapa de procesare folosind următoarele metode: umplerea golurilor cu 0, completarea cu formula lui Goel si formula dezvoltată de mine.

| Coin | Metoda de completare | P | R | F-measure |
|------|----------------------|--------|---------|-----------|
| ETH | 0 | 0.62% | 0.601% | 0.61035% |
| | Goel's formula | 0.703% | 0.686% | 0.7054% |
| | Our formula | 0.73% | 0.708% | 0.7188% |
| XRP | 0 | 0.584% | 0.559% | 0.5712% |
| | Goel's formula | 0.678% | 0.651% | 0.6657% |
| | Our formula | 0.681% | 0.656% | 0.6671% |
| BTC | 0 | 0.763% | 0.7709% | 0.7669% |
| | Goel's formula | 0.786% | 0.773% | 0.7794% |
| | Our formula | 0.786% | 0.773% | 0.7794% |

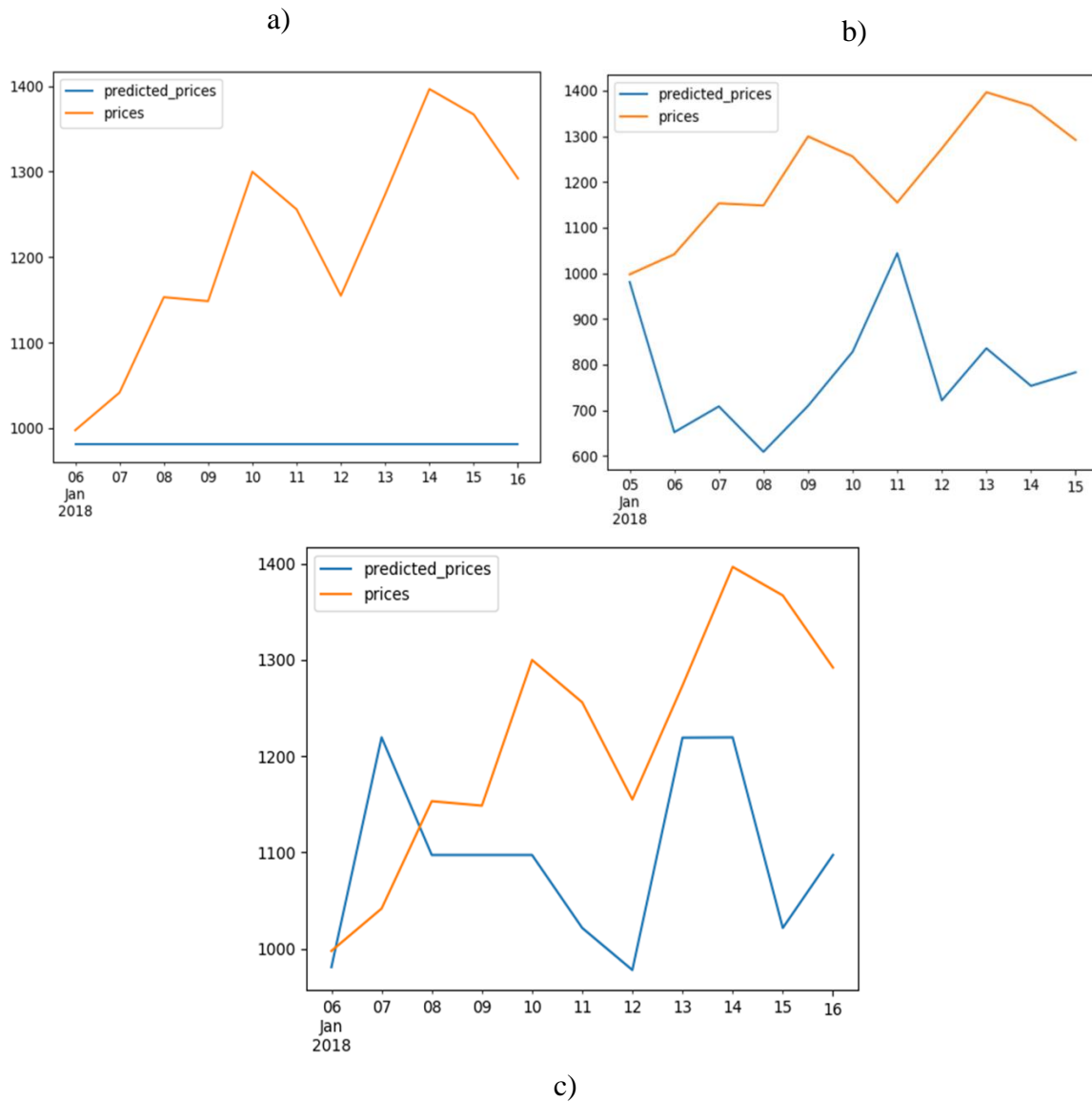


Figura 17. Grafice de predicție pentru Ethereum (ETH) utilizând următoarele metode de completare a golurilor: padarea cu 0 (a), completarea folosind funcția lui Goel (b) și formula dezvoltată de mine (c)

4. Discuții

În această secțiune discutăm rezultatele prezentate în capitolul anterior. Observăm că în primul tabel, în toate cazurile modelul în care se îmbină algoritmul de învățare automată (Random Forests) cu componenta de NLP (Analiză Sentimentelor) obține cele mai bune predicții în comparație cu Random Forests antrenat doar cu setul numeric, și cu Regresia Liniară. Metodă cea din urmă a obținut cel mai slab rezultat, astfel, după cum am menționat în capitolul 2, Regresia Liniară este o metodă primitivă și inefficientă în studiul predicțiilor din aria piețelor financiare.

În al doilea tabel notăm compararea dintre metodele de umplere a golurilor (padare) în setul de date textual. Padarea cu 0 se dovedește a fi inefficientă, observație ce reiese atât din tabel (deține cele mai mici scoruri) cât și în figura 17. a) unde linia albastră de predicție este formată dintr-o linie dreaptă orizontală, semnificând că algoritmul Random Forests a fost antrenat în mod deficitar în cazul acestui tip de padare. În schimb, celelalte două opțiuni de completare a blancurilor se comportă considerabil mai bine, formula lui Goel însă având o evoluție puțin mai slabă față de funcția mea, cel puțin, în scăderea bruscă de pe data de 12 ianuarie 2018 a fost prezisă ca fiind un punct culminant de către algoritm (Figura 17. b)). În cazul setului de date în care padarea a fost realizată cu formula mea, același punct de minim local a fost prezis corect de către algoritm (Figura 17. c)). De asemenea, în al doilea tabel identificăm faptul că pentru moneda Bitcoin (BTC) rezultatele sunt asemenea în cazul ambelor formule, asta se explică pe baza faptului că pentru Bitcoin au existat puține zile în care nu s-au publicat articole (fiind cea mai faimoasă monedă, toată presa este cu ochii pe evoluția sa), în cazul apariției golurilor, acestea sunt izolate și singulare, nefiind multe zile consecutive care nu-publicat nimic. Din acest punct, ambele formule vor același rezultat pentru $n=1$ (n fiind numărul de goluri consecutive).

Pentru Ripple, ambele tabele se poate vedea cele slabe rezultate cadrul ei, motivul pentru acest fenomen este raritatea cu care articolele referitoare la criptomoneda, astfel, ambele formule nu dau rezultatele scontate din corpusului de dimensiuni mici golurilor consecutive foarte .

Pentru o precum Ethereum (ETH), utilitatea padarii cu o își necesitatea. Se scrie relativ moderat de specialitate la intervale regulate, astfel, completarea acestor goluri se face mod eficient benefic antrenării algoritmului Random Forests.

5. Concluzii și direcții viitoare de cercetare

Scopul lucrării de licență este de a oferi un suport software prin care poate ușura deciziile privind tranzacțiile sau achizițiile de criptomonede de către utilizatorul amator și de a oferi o perspectivă în acest sens. Obiect al lucrării îl face de asemenea identificarea tehnicilor cele mai eficiente în elaborarea de predicții în domeniul criptovalutelor și optimizarea setului de date de antrenament al algoritmilor de învățare automată. Instrumentul și baza teoretică să servească un punct de plecare pentru perosanele ce sunt noi în acest domeniu și doresc să-și reducă riscurile financiare.

Să prezici o piață emergentă nu este o sarcină ușoară. Cantitatea mică de date (pentru majoritatea monedelor) din cauza vârstei pieței, numărul limitat de surse de încredere și volatilitatea crescută ne împinge spre cercetări mai ample. De ce să nu îi oferim timp de maturizare pieței? Deoarece câștigurile sunt foarte mari în acest moment și riscurile de asemenea. EMH (Efficient Market Hypothesis) pare să își spună cuvântul în anumite perioade de timp din piața criptomonedelor, asta datorită comportamentului haotic menționat.

Regresia liniară este într-adevăr o abordare slabă în acest sens, un început pentru predicții, dar nu o variantă viabilă de încredere. Al doilea algoritm de învățare automată folosit își arată eficiența prin rezultatele de la capitolul 3, înregistrând o acuratețe de predicție de 70+%, mai ales în cazul în care acest algoritm (Random Forests) a fost antrenat folosind două seturi de date diferite (unul de natură reală și unul de facțiune psihologică). Precizia ultimului algoritm este accentuată în momentul în care datele de tip textual sunt procesate, approximate valori unde se află goluri în setul de date. Simplă completare cu 0 ar fi un dezastru, Random Forests simplu dând un randament mai bun decât același algoritm antrenat cu două seturi de date iar cel de tip textual să fie padat cu 0. Folosirea unei formule de aproximare reprezintă o soluție simplă la problemă, atingând cea mai mare performanță în cazul formulei noastre.

Pentru aproximarea acestor goluri există loc de îmbunătățire, niciuna din formule nu e perfectă. O metodă de aproximare a golurilor folosind tehnici de învățare automată ar fi o idee în această direcție, însă cantitatea mică de date se dovedește a fi încă un impediment în acest

domeniu. O altă metodă ar fi o euristică, o metodă nedeterministă, o hibridizare între învățare automată și formula dezvoltată (sau una îmbunătățită).

Chiar dacă algoritmi prezentați au atins o performanță bună în teste, eu tot nu aș investi în predicția unui algoritm cu o acuratețe mai joasă de 85% deoarece este încă destul de riscant. Într-un astfel de algoritm, mai ales pentru o piață așa tânără și agitată, este necesar mai mult timp pentru cercetare și de luat în considerare foarte multe variabile. Zilele de sărbătoare reprezintă o variabilă importantă de luat în considerare, deoarece mulți vând în zilele respective sau cu câteva zile în urmă pentru a sărbătorii sau pentru a se pregăti pentru vacanță. Zilele de Crăciun în 2017 au fost marcate de scăderi, iar anul nou Chinezesc care a fost sărbătorit pe 16 februarie 2018, cu 10 zile în urmă au fost înregistrate cele mai mari scăderi din perioada respectivă, asta din cauză că țara cu cel mai mare volum de tranzacții în criptomonede este China, majoritatea platformelor de trading fiind localizate acolo. Există totuși variabile care nu pot fi prevăzute (îmi place să cred că pot, dar cer foarte multe date) sunt de natură politică, practic este greu de prevăzut când guvernele doresc să reglementeze anumite tranzacții (până în 2017 nu au existat reglementări pentru criptomonede) sau să le oprească pentru o perioadă, cum a procedat China și India în mai multe rânduri. Factorul climatic este totodată greu de prezis, calamitățile naturale pot influența volumul de tranzacții din zona afectată.

Identificarea etapelor de maturizare ale unei monede virtuale folosind tehnici din învățare automată este subiectul de cercetare în viitorul apropiat. După cum am descris în introducere la secțiunea de motivație, am identificat patternuri pe care majoritatea criptomonedelor în primele 12 luni le respectă într-o anumită măsură, concretizarea presupunerilor mele fiind următorul obiectiv în aria de predicții realizate cu ajutorul inteligenței artificiale: învățare automată, rețele neuronale și algoritmi genetici.

6. Bibliografie

1. Barber, B. M., and Odean, T.: All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. In: *Review of Financial Studies*, 21, 785–818 (2008).
2. Bariviera, A. F., Zunino, L. Guercio, M. B., Martinez, L. B., and Rosso, O. A.: Revisiting the European sovereign bonds with a permutation-information- theory approach. *Eur. Phys. J. B.* 86: 509. doi:10.1140/epjb/e2013-40660-7 (2014).
3. Clarke, J., Jandik, T., and Mandelker, G.: The efficient markets hypothesis. In: *Robert C. ARFFA, ed. Expert Financial Planning: Investment Strategies from Industry Leaders*. New York: Wiley, Chapter 9, pp. 126-141 (2001).
4. Dai, Y., and Zhang, Y.: *Machine Learning in Stock Price Trend Forecasting*. Stanford University (2013).
5. Daniel, K., Hirshleifer, D., and Subrahmanyam, A.: Investor psychology and security market under- and overreactions. In: *Journal of Finance*, 53, 1839–1885 (1998).
6. Fama, E.: The Behavior of Stock Market Prices. *Journal of Business*. 38: 34–105. doi:10.1086/294743 (1965).
7. Feller, W.: Martingales. In: *An Introduction to Probability Theory and Its Applications, Vol. 2*, New York: Wiley, pp. 210-215 (1971).
8. Gifu, D. and Cristea, D.: Public discourse semantics. A method of anticipating economic crisis presented at the Exploratory Workshop on Intelligent Decision Support Systems for Crisis Management, 8-12 May 2012, Oradea, Romania. In: *International Journal of Computers, Communications and Control*, see, I. Dzitac, F.G. Filip, M.-J. Manolescu (eds.), vol. 7/5, Agora University Editing House, pp. 829-836 (2012).
9. Granger, C. W. J.: Forecasting stock market prices: Lessons for forecasters. In: *International Journal of Forecasting* 8, North-Holland, 3-13 (1992).
10. Hilbert, A., Jacobs, H., and Müller, S.: Media Makes Momentum. *Review of Financial Studies*, 27(12), 3467–3501 (2014).
11. Hou, K., Peng, L., and Xiong, W.: A Tale of Two Anomalies: The Implications of Investor Attention for Price and Earnings Momentum, SSRN 976394 (2009).
12. Khaidem, L., Saha, S., and Dey, S. R.: Predicting the direction of stock market prices using random forest. In: *Applied Mathematical Finance*, 1-20 (2016).
13. Marwala, T.: Impact of Artificial Intelligence on Economic Theory – via arXiv.org (2015).
14. Marwala, T., and Hurwitz, E.: *Artificial Intelligence and Economic Theory: Skynet in the Market*. London: Springer (2017).
15. Mittal, A. and Goel, A.: *Stock prediction using twitter sentiment analysis*. Stanford University, CS229 (2012).
16. Nelson, R.: *Prophecy: A History of the Future - The Rex Research Civilization Kit*, <http://www.rexresearch.com/prophist/phfcon.htm> (2000).

17. Nunno, L.: *Stock Market Price Prediction Using Linear and Polynomial Regression Models* (2017).
18. Spierdijk, L., and Bikker, J. A.: *Mean Reversion in Stock Prices: Implications for Long-Term Investors* (2012).
19. Zunino, L., Bariviera, A. F., Guercio, M. B., Martinez, L. B., and Rosso, O. A.: On the efficiency of sovereign bond markets. *Phys. A Stat. Mech. Appl.*, 391: 4342–4349. doi:10.1016/j.physa.2012.04.009 (2012).
20. Xin, Y.: *Linear Regression Analysis: Theory and Computing* (2009).
21. J. Carrick - Bitcoin as a complement to emerging market currencies *Emerg. Markets Finance Trade*, 52 (2016), pp. 2321-2334