Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
63 tomas, 2022, 21–27
https://doi.org/10.15388/LMR.2022.29755

Vilnius University Press

Akcijų kainų ARIMA ir LSTM prognozavimo metodų lyginamoji analizė

Aivaras Bielskis, Igoris Belovas

Duomenų mokslo ir skaitmeninių technologijų institutas, Vilniaus universitetas
Akademijos g. 4, LT-08412 Vilnius
El. paštas: aivaras.bielskis@mif.vu.lt; igoris.belovas@mif.vu.lt

Įteiktas 2022 birželio 30; publikuotas 2022 gruodžio 10

Santrauka. Darbe yra pritaikomi ir palyginami akcijų kainų prognozavimo metodai: statistiniai laiko eilučių metodai (ARIMA, SARIMA) bei neuroniniais tinklais grįstas metodas (LSTM). Bendrovių *Amazon*, *Apple*, *Google*, *Netflix* ir *Tesla* akcijų kainų modeliavimo rezultatai yra vertinami pasitelkiant MAE ir MRE matus. Darbe gautos išvados leido nustatyti taikytų metodų trūkumus ir apsibrėžti patobulinimų ir tolimesnių tyrimų gaires.

Raktiniai žodžiai: laiko eilutės; neuroniniai tinklai; prognozavimas; ARIMA; SARIMA; LSTM

AMS: 62M10, 68T01

1 Įvadas

Akcijų kainų prognozavimas yra plačiai nagrinėjama tema pasaulyje. Egzistuojantys metodai yra nuolat tobulinami, yra ieškoma naujų kelių, vykdomi aktualių įvykių ir procesų tyrimai. Aktualus klausimas yra laiko eilučių metodų efektyvumo palyginimas su giliuoju mokymusi grįstomis prognozavimo priemonėmis. Iš naujausių (2021–2022 m.m.) publikacijų paskelbtuose leidiniuose, įtrauktuose į *Clarivate Analytics Web of Science* DB, kuriuose yra lyginami stochastiniai ARIMA (angl. *Autoregressive Integrated Moving Average*) ir neuroninių tinklų (pvz., LSTM, angl. *Long Short-Term Memory*) modeliai, galima pažymėti Menculini *et al.* (nagrinėjusius didmenines maisto kainas) [9], Mbah *et al.* (nagrinėjusius kalkakmenio kainų svyravimus) [8], Vuong *et al.* (nagrinėjusius svyravimus valiutų rinkoje) [16] ir Dou *et al.* (nagrinėjusius pu-erh arbatos kainas) [5].

Menculini *et al.* rezultatai parodė, kad ir ARIMA modelis, ir LSTM modelis veikia panašiai atliekant nagrinėjamą prognozavimo užduotį. Mbah *et al.* gavo, kad

ARIMA modelio tikslumas yra 95,7%, o rekurentinio neuroninio tinklo, naudojančio ilgos laikinosios atminties (LSTM) sluoksnius, tikslumas yra 91,8%, t.y., ARIMA modelis pranoko RNN modelį nagrinėjamam uždaviniui (palyginimui buvo naudota simetrinė vidurkio absoliuti procentinė paklaida (SMAPE angl. *Symmetric Mean Absolute Percentage Error*)). Vuong *et al.* pasiūlytas hibridinis XGBoost+LSTM modelis pasirodė geriau už bazinį ARIMA. Dou *et al.* rezultatai parodė, kad ARIMA labiau tinka trumpalaikiam prognozavimui. Reikia pažymėti, kad 2018 m. (tačiau labai intensyviai cituojamo) Siami-Namini *et al.* (Texas Tech) tyrimo, palyginančio ARIMA ir LSTM modelių adekvatumą, rezultatai parodė neuroniniais tinklais grįsto modelio pranašumą [13].

Lietuvoje, kaip rodo Lietuvos akademinė elektroninė biblioteka eLABa, šios srities tyrimai vykdomi mažesniu intensyvumu. Iš naujesnių publikacijų galima pažymėti Gasparėnienės *et al.* [6] (nagrinėjamos aukso kainos tendencijų prognozės; skaičiavimai parodė, kad ARIMA modelis tinka tik trumpalaikėms aukso kainų prognozėms (ne daugiau kaip 1 metai)), Česnavičiaus [4] (darbe pristatoma ARIMA modeliu paremta Lietuvos elektros energijos kainos prognozė), Belovo *et al.* [1, 3]. Nors nuo 2015 m. kainų prognozavimo tematikoje buvo apginta nemažai (dešimtys) bakalauro ir magistro baigiamųjų darbų, daktaro disertacijos buvo apgintos tik dvi [7, 15].

2 Nagrinėjami prognozavimo metodai

Kainų prognozavimas yra netriviali problema, glaudžiai susijusi su beprecedenčiais ekonominių tendencijų ir sąlygų pokyčiais iš vienos pusės, ir nepilna informacija – iš kitos. Šiame straipsnyje ARIMA ir LSTM modeliai lyginami efektyvumo mažinant paklaidos lygį atžvilgiu. Tradicinių prognozavimo metodų atstovas, ARIMA, pasirinktas dėl jo gebėjimo apibūdinti nestacionarios prigimties empirinius duomenis. Giliojo mokymosi pagrįstų algoritmų atstovas, LSTM, naudojamas dėl jo gebėjimo išsaugoti ir modeliuoti specifines duomenų savybes ilgesniuose laiko intervaluose.

2.1 Autoregresiniai integruoti slenkančio vidurkio modeliai (ARIMA ir SARIMA)

ARIMA modeliai yra vieni paprasčiausių ir dažniausiai naudojamų ekonometrinių metodų, skirtų vieno kintamojo laiko eilučių modeliavimui [10]. ARIMA yra autoregresinio (AR, angl. *Autoregressive*) ir slenkančio vidurkio (MA, angl. *Moving-Average*) modelių junginys, žymimas ARIMA(p, d, q), su pagrindiniais parametrais $p, d, q \in \mathbb{N}_0$, atitinkančiais koreliuotų vėlavimų bei diferencijavimo eiles. Parametras p nusako autoregresinę priklausomybę su p vėlavimais (AR(p)), d – diferencijavimo eilę, o q – slenkančio vidurkio vėlavimo eilę (MA(q)). ARIMA procesas aprašomas lygtimi

$$x_t = c + \sum_{j=1}^{p} \alpha_j x_{t-j} + \epsilon_t + \sum_{j=1}^{q} \theta_j \epsilon_{t-j},$$

čia $\alpha_j \neq 0$ ir $\theta \neq 0$ yra autoregresinio ir slenkančio vidurkio modelių koeficientai, c yra konstanta, o ϵ_t žymi baltąjį triukšmą su nulinių vidurkiu ir $\sigma^2 > 0$ dispersija.

Tiriant laiko eilutes neretai pasitaiko, kad jos pasižymi periodiškumu, t.y. laike pasikartojančiais dėsniais. Todėl pravartu naudoti ARIMA modelio išplėtimą, pridedant

sezoninę dalį, vertinančią per sezoną perstumtų vėlavimų koreliaciją [10]. SARIMA modelis žymimas SARIMA$(p,d,q)(P,D,Q)s$. Čia s yra parametras, apibūdinantis sezono trukmę, o P, Q ir D – sezonines SAR(P), SMA(Q) modelių eiles ir sezoninio diferencijavimo eilę.

2.2 Ilgos laikinosios atminties neuroniniai tinklai (LSTM)

Dirbtinio intelekto, ypač mašininio mokymosi, atsiradimas paskatino sukurti metodų rinkinį, kuris pasirodė esąs labai naudingas daugelyje skirtingų krypčių. Vienas proveržis, be abejo, buvo gilusis mokymasis, kuris iš esmės pakeitė mūsų būdą tvarkyti ir naudoti duomenyse esančią informaciją. Gilusis mokymasis gali veiksmingai aptikti ir modeliuoti paslėptus dėsningumus, automatiškai išgaudamas funkcijas, kurios kitu atveju turėtų būti išgaunamos rankiniu būdu tikrinant duomenų rinkinį. Standartinis pasirinkimas susidūrus su laiko eilučių problemomis yra LSTM neuroninių tinklų panaudojimas. Ilgos laikinosios atminties neuroniniai tinklai (LSTM, angl. *Long Short-Term Memory*) yra specialus rekurentinių neuroninių tinklų (RNN, angl. *Recurrent Neural Network*) tipas, galintis prisiminti ankstesnių etapų reikšmes naudojimui ateityje [11]. LSTM pagrindinis pranašumas lyginant su RNN yra tas, kad LSTM įvertina, ar perduodama informacija yra svarbi, ar ne, ir jeigu informacija nėra svarbi, ji pašalinama.

2.3 Prognozių tikslumo vertinimas

Vidutinė absoliučioji paklaida (MAE, ang. *Mean Absolute Error*) ir vidutinė santykinė paklaida (MRE, ang. *Mean Relative Error*) yra matai naudojami šiame darbe prognozės paklaidai įvertinti,

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|x_t - \hat{x}_t|, \qquad MRE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - \hat{x}_t}{x_t}\right|,$$

čia n yra laiko eilutės elementų skaičius, x_t – stebėta laiko eilutės reikšmė laiko momentu t, o $\hat{x}_t$ – prognozė laiko momentu t.

3 Tyrimo objektas

Buvo tiriamos istorinės *Tesla, Amazon, Netflix, Google, Apple* įmonių akcijų kainos (https://finance.yahoo.com/). Empiriniai duomenys (akcijų kaina dienos pradžioje) buvo išskirstyti keturiais skirtingais periodais: 6 metų, nuo 2016-05-31 iki 2022-05-31 dienos, kuriuos sudarė 1511 elementų, 4 metų, nuo 2018-05-31 iki 2022-05-31 dienos, kuriuos sudarė 1008 elementai, 2 metų, nuo 2020-05-31 iki 2022-05-31, dienos kuriuos sudarė 504 elementai ir 1 metų, nuo 2021-05-31 iki 2022-05-31 dienos, kuriuos sudarė 252 elementai. Visiems keturiems periodams testavimai buvo atlikti dienų periodui nuo 2022-03-01 iki 2022-05-31 dienos.

4 Metodologija

Darbe ARIMA ir SARIMA modeliai skaičiuojami pasitelkiant *Python* kalboje esančios bibliotekos *statsmodels.tsa.statespace.sarimax* paketo SARIMAX pagalba. Prieš

1 lentelė. ARIMA ir SARIMA modelių parametrai.

Duomenys	ARIMA	SARIMA
Google, 6 metai	(6, 1, 4)	(6, 1, 4)(6, 1, 4, 12)
Google, 4 metai	(6, 1, 4)	(6, 1, 4)(6, 1, 4, 12)
Google, 2 metai	(2, 1, 2)	(2, 1, 2)(2, 1, 2, 12)
Google, 1 metai	(3, 1, 2)	(3, 1, 2)(3, 1, 2, 12)
Apple, 6 metai	(2, 1, 0)	(2, 1, 0)(2, 1, 0, 12)
Apple, 4 metai	(2, 1, 0)	(2, 1, 0)(2, 1, 0, 12)
Apple, 2 metai	(2, 1, 0)	(2, 1, 0)(2, 1, 0, 12)
Apple, 1 metai	(1, 1, 0)	(1, 1, 0)(1, 1, 0, 12)
Amazon, 6 metai	(6, 1, 6)	(6, 1, 6)(6, 1, 6, 12)
Amazon, 4 metai	(1, 1, 3)	(1, 1, 3)(1, 1, 3, 12)
Amazon, 2 metai	(3, 1, 1)	(3, 1, 1)(3, 1, 1, 12)
Amazon, 1 metai	(2, 1, 2)	(2, 1, 2)(2, 1, 2, 12)
Tesla, 6 metai	(4, 1, 3)	(4, 1, 3)(4, 1, 3, 12)
Tesla, 4 metai	(3, 1, 2)	(3, 1, 2)(3, 1, 2, 12)
Tesla, 2 metai	(3, 1, 2)	(3, 1, 2)(3, 1, 2, 12)
Tesla, 1 metai	(2, 1, 6)	(2, 1, 6)(2, 1, 6, 12)
Netflix, 6 metai	(4, 1, 5)	(4, 1, 5)(4, 1, 5, 12)
Netflix, 4 metai	(1, 1, 0)	(1, 1, 0)(1, 1, 0, 12)
Netflix, 2 metai	(2, 1, 2)	(2, 1, 2)(2, 1, 2, 12)
Netflix, 1 metai	(1, 1, 0)	(1, 1, 0)(1, 1, 0, 12)

atliekant modeliavimą, yra patikrinamas laiko eilutės stacionarumas; stacionarumui nustatyti naudojamas Dickey–Fuller testas [10]. Nestacionarias eilutes paversti stacionariomis yra naudojama skirtuminė transformacija (diferencijavimas), atliekama pradinių duomenų laiko eilutę keičiant skirtumų laiko eilute. Diferencijavimo eilę nusako parametras d,

$$\Delta^d x_t = \Delta^{d-1} x_t - \Delta^{d-1} x_{t-1}.$$

Reikia pažymėti, kad visoms laiko eilutėms užteko pirmos eilės diferencijavimo, taigi ARIMA ir SARIMA modeliuose $d = 1$. Stochastinių modelių parametrai buvo vertinami pasitelkiant Akaike informacijos kriterijų (AIC) [10]. Gautos optimalios parametrų reikšmės buvo naudojamos prognozavimui (žr. 1 lentelę).

LSTM modelis kuriamas pasitelkiant *Python* kalboje esančią biblioteką *keras*. Prieš pradedant darbą su LSTM, treniravimui ir testavimui skirti duomenys buvo normalizuoti (tiesinė transformacija $[x_{\min}, x_{\max}] \to [-1, 1]$), siekiant, kad duomenų sklaida nebūtų pervertinta. Išbandžius kelis skirtingus paslėptų sluoksnių skaičius, šiame darbe parinktas paslėptų sluoksnių skaičius yra 10. Modelio optimizavimo algoritmas naudojamas *Adam*, o praradimams skaičiuoti naudojama MAE funkcija, praradimai skaičiuojami MSE. Atlikus bandymus buvo nuspręsta, kad optimali mokymosi trukmė yra 20 epochų, todėl šiame darbe pasirinkta naudoti tokią mokymosi trukmę.

5 Rezultatai

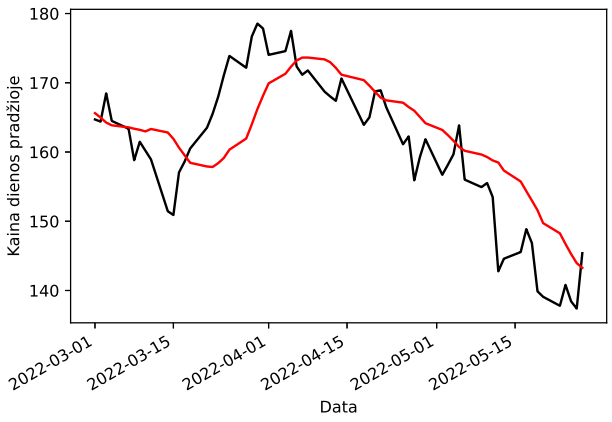
Gauti rezultatai rodo (žr. 2 lentelę), kad ARIMA modelis prognozavimo užduotį geriau atlieka su trumpalaikėmis eilutėmis, kas visiškai atitinka Dou *et al.* išvadą: "*Rezultatai rodo, kad ARIMA labiau tinka trumpalaikiam prognozavimui ir metodo*

2 lentelė. Prognozių paklaidų rezultatai (MRE).

Duomenys	ARIMA	SARIMA	LSTM
Google, 6 metai	0,0997	0,0913	0,0373
Google, 4 metai	0,0981	0,0866	0,0430
Google, 2 metai	0,0838	0,0496	0,0491
Google, 1 metai	0,0502	0,0569	0,0528
Apple, 6 metai	0,0705	0,0827	0,0318
Apple, 4 metai	0,0712	0,0858	0,0423
Apple, 2 metai	0,0610	0,1751	0,0382
Apple, 1 metai	0,0512	0,1012	0,0367
Amazon, 6 metai	0,1473	0,1165	0,0575
Amazon, 4 metai	0,1398	0,1095	0,0580
Amazon, 2 metai	0,1169	0,1740	0,0798
Amazon, 1 metai	0,0981	0,0892	0,1030
Tesla, 6 metai	0,1422	0,1529	0,0639
Tesla, 4 metai	0,1411	0,1785	0,0837
Tesla, 2 metai	0,1429	0,1950	0,0744
Tesla, 1 metai	0,1621	0,2509	0,0823
Netflix, 6 metai	0,4479	0,2910	0,1028
Netflix, 4 metai	0,3986	0,3797	0,1341
Netflix, 2 metai	0,2801	0,1579	0,2224
Netflix, 1 metai	0,1637	0,5422	0,3448

prognozavimo paklaida didės, kai prognozavimo laikotarpis ilgėja" [5]. SARIMA modelis tokios tendencijos nerodo.

LSTM prognozės yra tiksliausios visoms tirtoms laiko eilutėms, kas patvirtina Siami-Namini *et al.* išvadą: "*giliuoju mokymusi pagrįsti algoritmai, tokie kaip LSTM, pranoksta tradicinius algoritmus, tokius kaip ARIMA modelis*" [13]. Priešingai ARIMA rezultatams, LSTM prognozės yra geriausios ilgiausiuose intervaluose (6 metai, žr. 1 pav.).



1 pav. LSTM prognozavimo pavyzdys (Apple, 6 metai).

6 Išvados

Darbe nagrinėtų akcijų kainų elgesys pasirodė gana panašus, todėl labai didelio skirtumo prognozuojant jų duomenis nebuvo pastebėta. LSTM modelis prognozavimo užduotį atliko geriau už tradicinius (ARIMA, SARIMA) stochastinius modelius, kurie nedavė patenkinamų prognozavimo rezultatų. Kadangi rezultatai (nors ir patvirtinę ankstesnių tyrėjų išvadas) palieka vietos darbo tobulinimui, planuojamas jo tęsinys. Ateities tyrimo planas:

- pritaikyti alternatyvius informacinius kriterijus, vertinant ARIMA ir SARIMA parametrus (pvz., Bajeso informacijos kriterijų (BIC, ang. *Bayesian Information Criteria*) arba Apibendrintą informacijos kriterijų (GIC, ang. *Generalized Information Criteria*) vietoje AIC, žr. [14]), išbandyti daugiau skirtingų LSTM modelio struktūrų;
- optimizuoti ir išlygiagretinti kodą, nes *Python* SARIMAX paketo reikalaujamas ilgų laiko eilučių apdorojimo kompiuterinis laikas yra nepatenkinamas.
- išbandyti hibridinį (ARIMA ir neuroninių tinklų) modelį (pl. [12, 17]);
- atliekant eksperimentus, naudoti daugiau įvairesnių akcijų; taip pat naudoti ir išvestines finansines priemones (opcionai, uždirbimą nuo akcijos kritimo, varantai); paliesti aktualią elektros kainų sprogimo problemą ir prognozavimo galimybes;
- sujungti prognozavimo metodus su portfelio optimizavimo metodais [2] (kuriant portfelį pridėti išvestines finansines priemones bei pridėti daugiau galimų investavimo priemonių, tokių kaip tarpusavio skolinimas, nekilnojamas turtas, biržos prekės);
- akcijų kainas vertinti ne tik pagal jų istorinius duomenis, bet ir pagal kitus kriterijus: įmonės finansinius rezultatus, įmonės teigiamas/neigiamas naujienas, politinę bei ekonominę padėtį pasaulyje bei valstybėje, kurioje yra įmonės centrinė būstinė.

Literatūra

[1] I. Belovas. Baltijos šalių akcijų lyginamojo indekso OMX Baltic Benchmark modelių tyrimas. *Liet. matem. rink. LMD darbai, ser. B*, **60**:6–10, 2019.

[2] I. Belovas, L. Sakalauskas, V. Starikovičius. A mixed-stable approach to the management of the portfolio using high-frequency financial data. *Inf. Technol. Control.*, **66**(3):293–307, 2017.

[3] I. Belovas, L. Sakalauskas, V. Starikovičius, E.W. Sun. Mixed-stable models: an application to high-frequency financial data. *Entropy*, **23**(6):1–12, 2021.

[4] M. Česnavičius. Lithuanian electricity market price forecasting model based on univariate time series analysis. *Energetika*, **66**(1):39–46, 2020.

[5] Z. Dou, M. Ji, M. Wang, Y. Shao. Price prediction of pu'er tea based on ARIMA and BP models. *Neural Comput. Appl.*, **34**(5):3495–3511, 2022.

[6] L. Gasparėnienė, R. Remeikienė, A. Sadeckas, R. Ginevičius. The main gold price determinants and the forecast of gold price future trends. *Econ. Sociol.*, **11**(3):248–264, 2018.

[7] J. Katina. *Prognozavimo problemų tyrimas virtualioje akcijų biržoje.* Daktaro disertacija. Vilniaus universitetas, 2015.

[8] T.J. Mbah, H. Ye, J. Zhang, M. Long. Using LSTM and ARIMA to simulate and predict limestone price variations. *Min. Metall. Explor.*, **40**(1):237–246, 2022.

[9] L. Menculini, A. Marini, M. Proietti, A. Garinei, A. Bozza, C. Moretti, M. Marconi. Comparing prophet and deep learning to ARIMA in forecasting wholesale food prices. *Forecasting*, **3**(3):644–662, 2021.

[10] A. Nielsen. *Practical Time Series Analysis.* O'Reilly Media, 2019.

[11] J. Patterson. *Deep Learning: A Practitioner's Approach.* O'Reilly Media, 2017.

[12] N. Pooniwala, R. Sutar. Time series forecasting using a hybrid ARIMA and neural network model. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pp. 1–6, 2021.

[13] S. Siami-Namini, N. Tavakoli, A.S. Namin. A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1394–1401, 2018.

[14] P. Stoica, Y. Selen. Model-order selection: a review of information criterion rules. *IEEE Signal Process. Mag.*, **21**(4):36–47, 2004.

[15] M. Vaitonis. *Didelio dažnio kompiuterizuotų prekybos strategijų inžinerija finansinėse rinkose.* Daktaro disertacija. Vilniaus universitetas, 2020.

[16] P.H. Vuong, T.T. Dat, T.K. Mai, P.H. Uyen, P.T. Bao. Stock-price forecasting based on XGBoost and LSTM. *Comput. Syst. Sci. Eng.*, **38**(2):913–926, 2021.

[17] D. Xu, Q. Zhang, Y. Ding, D. Zhang. Application of a hybrid ARIMA-LSTM model based on the SPEI for drought forecasting. *Environ. Sci. Pollut. Res.*, **29**(3):4128—4144, 2022.

SUMMARY

Comparative analysis of stock price ARIMA and LSTM forecasting methods

A. Bielskis, I. Belovas

In the work, relevant methods of stock price forecasting are applied and compared: statistical time series (ARIMA, SARIMA) and neural network-based (LSTM). The results of stock price (Amazon, Apple, Google, Netflix, and Tesla companies) simulations are evaluated using MAE and MRE measures. The conclusions obtained in the work made it possible to identify shortcomings of the approaches and specify guidelines for improvements and further research.

Keywords: time series; neural networks; forecasting; ARIMA; SARIMA; LSTM