Primena masinskog učenja u modeliranju dinamike knjige limitiranih naloga
Application of machine learning in modeling the dynamics of the limit order book

dr Dragana Radojičić[1]
Univerzitet u Beogradu
Ekonomski fakultet

Sadržaj – Tema ovog rada je proučavanje dinamike i statističkih osobina Knjige graničnih naloga. Da bi se opisala složena dinamika trgovanja prisutna na tržištu, koriste se koncepti mašinskog učenja. Veštačka inteligencija postaje sve više prisutna u raznim oblastima. Potencijalna primenu koncepta mašinskog učenja u finansijama zainteresovala je veliki broj istrazivaca kako iz akademije tako i iz industrije. Pošto je trgovanje doživelo automatizaciju i digitalizaciju, koncept fizičkog trgovanja je zamenjen elektronskim trgovanjem. Zbog toga postoji veliko interesovanje u razvijanju algoritama za automatsko trgovanje zasnovanih na mašinskom učenju. Cilj ovog istraživanja je da proučava distribuciju varijable koja modelira razliku između najmanje cene na strani za kupovinu, i najvece cene na strani za prodaju. S obzirom da su finansijska tržišta informativna, postoji potencijal u analizi istorijskih podataka o akcijama, kao i u razvoju algoritama za analizu tih podataka.

Abstract - The topic of this paper is the study of the dynamics and statistical properties of the Limit Order Book. In order to describe the complex trading dynamics present in the market, machine learning concepts are used. Artificial intelligence is becoming more and more present in various fields. The potential application of the machine learning concepts in finance has interested a large number of researchers from both academia and industry. As trading has undergone automation and digitalization, the concept of physical trading has been replaced by electronic trading. Therefore, there is great interest in developing machine learning algorithms based on machine learning. The aim of this study is to study the distribution of a variable that models the difference between the lowest price on the buying side and the highest price on the selling side. Since financial markets are informative, there is potential in analyzing historical stock data, as well as in developing algorithms for analyzing that data.

1 UVOD I MOTIVACIJA ZA ISTRAŽIVANJE

Kao što je navedeno u članku Washington Post-a "The robots-vs.-robots trading that has hijacked the stock market"[1] otprilike 50% ukupnog obima trgovanja izvršavaju roboti. Berze proizvode ogromnu količinu empirijskih podataka. Ideja je da se proučava informativnost karakteristika ekstrahovanih iz baze podataka koja replicira bazu podataka sa NASDAQ (eng. National Association of Securities Dealers Automated Quotations) berze, kako bi se klasifikovao vektor baze podataka pomoću neuronske mreže.

Posebna pažnja posvećena je proučavanju ponašanja varijable koja modelira razliku između najmanje cene na strani za kupovinu, i najveće cene na strani za prodaju. Empirijska studija pokazuje da je ta razlika najčešće jednaka minimalnom razmaku između dve cene prisutne u knjizi limitiranih naloga. Prema ranijim istraživanja (pogledajte [1], [2] i [3]) finansijska tržišta su informativna, što može biti korisno da se identifikuje i definiše strategija trgovanja. Ideja je da se svaki vektor tržišnih podataka klasifikuje u jednu od oznaka iz skupa S = {kupi, prodaj, neaktivan}.

Prevenstveno, relevantne karakteristike iz baze podataka berze se ekstraktuju, i proučava se njihova informativnost. S obzirom da na berzi ima više miliona događaja u toku samo jednog dana, knjiga limitiranih naloga proizvodi ogromnu bazu padataka. U radu [4] je predložen sistem rekonstrukcije podataka i transformacija baze podataka knjige graničnih naloga kako bi se izvukle karakteristike od interesa. Da bi mogli da sprovedemo istraživanje, u radu [5] predstavljeno je okvirno postolje za obrada podataka iz LOBSTER[2] baze. Nove metode za seleketovanje informativnih karaktersitika iz knjige limitiranih naloga su predstavljene u radu [6], i ustanovljeno je da se performanse modela dugotrajno kratkoročne memorije (eng. Long short-term memory) neuronske mreže poboljšavaju kada izaberemo karakteristike pomoću tih metoda. Štaviše, koristi se višekriterijumska optimizacija da bi se izabrao najbolji model od 7 predloženih modela, uzimajući u obzir više kriterijuma za ocenu modela. U radu [7] primenjuju se Furijeove transformacije da bi se izvukle nove karakteristike iz baze podataka sa berze, i posebna pažnja je usmerena na proučavanje ponašanja signala koji potiču iz različitih izvora, kao što su na primer signali cena različitih nivoa knjige limitiranih naloga. Dosta istraživanja o knjigama limitiranih narudzbina i srodnim oblastima o mikrostrukturama tržišta su sprovedena. Istraživaja se uglavnom odnose na karakterizaciju karakteristike kao što su likvidnost, volatilnost i kotirani raspon.

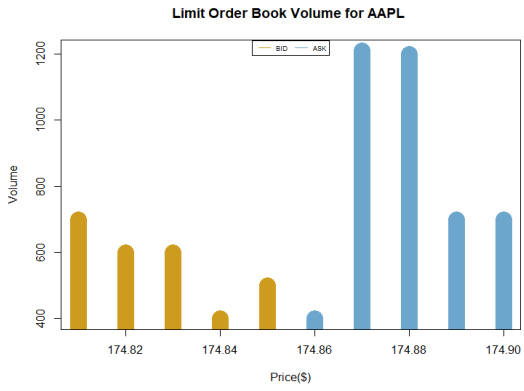
U ovom istraživanju ideja je da se proučava distribucija varijable kotirani raspon koja nam može ukazati na likvidnost marketa, sto je neophodan faktor za dobro funkcionisanje berze.

[1] https://www.washingtonpost.com/news/wonk/wp/2018/02/07/the-robots-v-robots-trading-that-has-hijacked-the-stock-market/. Accessed: 2022-03-10

[2] https://www.https://lobsterdata.com/.Accessed: 2022-03-10

2 KNJIGA LIMITIRANIH NALOGA

Knjiga limitiranih naloga (eng. Limit Order Book) je knjiga naloga u elektronskom obliku u koju se upisuju nalozi za kupovinu i prodaju po unapred određenoj ceni. Nalozi su prikazani sa svojom ukupnom količinom dostupnom za trgovanje po odgovarajućoj ceni. Za razliku od ograničenog naloga, tržišni nalog (eng. market order) je nalog u kojem cena nije unapred definisana, već se nalog izvršava prema trenutno najpovoljnijoj ceni prisutnoj u knjizi naloga. Glavni zadatak knjige limitiranih naloga je da evidentira sve dolazne i odlazne naloga. Knjiga limitiranih naloga je definisana na diskretnoj mreži, gde svaka tačka mreže reprezentuje cenu koja je registrovana na berzi za prodaju ili kupovinu naloga. Minimalna udaljenost između dva nivoa cena naziva se tick (eng.). Pogledajte Sliku 1 za primer snimka knjige limitiranih naloga za akcije kompanije Apple. Knjiga linmitiranih naloga ima dve strane. Na BID strani beleži se cena po kojoj će trejder ili berza prodati hartiju od vrednosti ili cena po kojoj investitor može kupiti hartiju od vrednosti. Sa druge strane, na ASK strani beleži se cena po kojoj će trejder ili berza kupiti hartiju od vrednosti ili cena po kojoj investitor može da proda hartiju od vrednosti. Minimalna cena u trenutku t na strani ASK, označena sa $P_{t,1}^{a}$, naziva se najbolja cena tražnje (eng. the best ask price) i predstavlja najnižu cenu po kojoj će investitor prodati akciju. Maksimalna cena u trenutku t na strani BID, označena sa $P_{t,1}^{b}$, naziva se najbolja ponuđena cena (eng. the best bid price) i predstavlja najvišu cenu po kojoj će investitor kupiti akciju. Prva cena manja od najbolje ponuđene cene postavlja se na nivo 2 na strani ponude, dok je najniža dostupna cena viša od najbolje cene tražnje postavljena na nivo 2 na strani potražnje, itd. Glavni posao knjige limitiranih naloga je da evidentira cene i količine naloga na obe strane do određenog broja nivoa. Za svaku cenu akcije koja je prisutna na berzi, knjiga naloga beleži raspoloživi obim koji predstavljaju broj akcija koje su dostupne po toj ceni.



Slika 1. Snimak knjige limita naloga od 5 nivoa akcija kompanije AAPL (Apple Inc.) sa NASDAQ berze (8. januara 2018.u 10:02 časova). Na strani ponude/traživanja postavljaju se nalozi za kupovinu/prodaju (žuto/plavo), i najbolja bid cena je 174.85$ sa obimom od 550 akcija, dok je najbolja cena ponude 174.86$ sa obimom od 440 akcija.

Knjiga limita naloga koja sadrži N nivoa je u trenutku t definisana vektorima

$$P_t^a = \left(P_{t,1}^a, P_{t,2}^a, \ldots, P_{t,N}^a\right) \qquad (1)$$

$$P_t^b = \left(P_{t,1}^b, P_{t,2}^b, \ldots, P_{t,N}^b,\right), \qquad (2)$$

koji predstavljaju najbolje N cene na strani traženja i ponude, respektivno; kao i sa vektorima

$$V_t^a = \left(V_{t,1}^a, V_{t,2}^a, \ldots, V_{t,N}^a\right) \quad \text{i} \quad (3)$$

$$V_t^b = \left(V_{t,1}^b, V_{t,2}^b, \ldots, V_{t,N}^b,\right), \qquad (4)$$

koji predstavljaju broj akcija dostupnih za trgovinu u trenutku t po cenama prisutnim u vektorima P_t^a i P_t^b, respektivno. Sveobuhvatan pregled matematičkog koncepta knjige limitiranih naloga je izložen u odeljku II u [8].

Pri opisivanju oblika knjige limitiranih naloga često se pominju dve promenljive: kotirani raspon (eng. quoted spread) i srednja cena (eng. Mid-price). Kotirani raspon, koji označavamo sa $QuotedSpread_t$, je definisan kao razlika između najniže tražene cene i najviše ponuđene cene:

$$QuotedSpread_t = P_{t,1}^b - P_{t,1}^a. \qquad (5)$$

Razlika u ceni koju plaća kupac i dobija prodavac je trošak likvidnosti. Postoje berze, kao što je na primer NASDAQ, gde trejderi obezbeđuju likvidnost. Razlike u rasponu između najniže tražene i najviše ponuđene cene ukazuju na promene likvidnosti. Što je manji raspon i što je više limitiranih naloga u knjizi, to je veća likvidnost hartije od vrednosti. Likvidnost je mera sposobnosti da se jeftino izvrši transakcija. Dovoljna likvidnost je neophodna i sastavna komponenta tržišta koje dobro funkcioniše.

Srednja cena, koju označavamo sa $MidPrice_t$, je definisana kao aritmetički prosek najbolje cene tražnje i najbolje ponuđene cene:

$$Mid\ Price_t = \frac{1}{2}\left(P_{t,1}^a + P_{t,1}^b\right). \qquad (6)$$

Srednja cena se često koristi kao aproksimacija realne cene akcije. Pogledajte rad [9] gde su izloženi različiti koncepti modeliranja srednje cene u kontekst knjige limitiranih naloga.

3 MODEL

Označimo sa $\dot{T}$ parametarski skup koji predstavlja skup vremena trgovanja kada su registrovani događaji na berzi. Broj naloga dostupnih po ceni p u vreme t označimo sa $V(t,p)$. U knjizi limitiranih naloga cene su zapisane na diskretnoj mreži i minimalnu udaljenost između dva nivoa cena (eng. Tick) ozmačićemo sa δ. Posmatramo dvoparametarski proces koji predstavlja proces obima naloga:

$$\left\{V(t,p), t \in \dot{T}, p \in \delta N \vee p \in \frac{\delta}{2} N\right\}. \quad (7)$$

Primetimo da su cene podeljene po nivoima, i kako je minimalan razmak između cena δ, nivoi u knjizi limitiranih naloga cene uzimaju vrednosti iz skupa δN ili $\delta/2*N$. S obzirom da se cena menja, pogodno je pratiti obim porudžbina na datoj udaljenosti i (tikova) od srednje cene. Dakle, posmatramo centriranu knjigu naloga $U_t(p)=V(t,MidPrice_t+p)$.

4 KOTIRANI RASPON

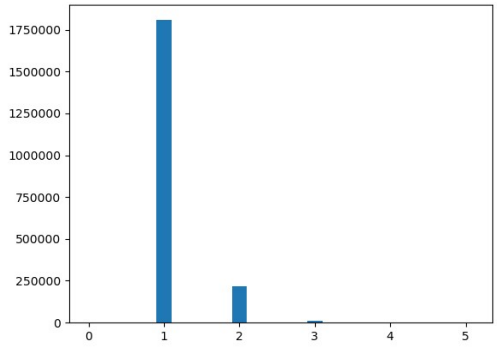
Centralna tačka ovog istraživanja je da se posmatra ponašanje kotiranog raspona u toku jednog trgovačkog dana, preciznije njegova distribucija. Studija koja je sprovedena u [10] pokazuje da je kotirani raspon najčešće jednak minimalnom razmaku između dve cene prisutne u knjizi limitiranih naloga. Promene u kotiranom rasponu nam mogu ukazati na promene u likvidnosti, pogledajte rad [11]. Studija prikazana u radu [12] rad predlaže procenu efektivne razlike između ponude i potražnje, i proučava kako takva procena utiče na likvidnost. Rezultati prikazani u radu [13] pokazuju da za procenu efektivnog ili kotiranog raspona pod normalnim tržištem uslovima Istanbulske berze, tick daje najbolje rezultate u poredejenu sa jos 5 posmatranih različitih metoda.

5 KAHANOV ALGORITAM ZA SUMIRANJE

Prosečan raspon tokom svih vremena trgovanja se izračunava korsicenjem Kahanov algoritam za sumiranje (eng. Kahan Summation algoritam), pogledajte [14]. Sličan algoritam za sumiranje predstavljen je još ranije u [15] i taj algoritam posmatra numerički proces kao propisano ponavljanje elementarnih operacija koje stvaraju mapiranje posmatranog matematičkog problema. Slična tehnika koja prati akumuliranu grešku u celobrojnim operacijama je prikazana u [16]. Kahanovo sumiranje smanjuje numeričku grešku u konačnom rezultatu pri sabiranju niza velikih brojeva ili brojeva sa decimalnim zarezom u poređenju sa klasičnim pristupom. Ideja Kahanovog algoritma za sumiranje je da kompenzuje grešku pri sabiranju brojeva tako što se uvodi posebna promenljiva za čuvanje greške u svakom koraku dok se izvršavaju aritmetičke operacije. Kahanovo sumiranje minimizira grešku pri sabiranju brojeva sa decimalnim zarezom i smanjuje značajan gubitak preciznosti pri aritmetičkim operacijama. Kahanov algoritam za sumiranje se može ilustrovati pseudokodom prikazanom na Slici 2.

```
function KahanSumation(NizBrojeva)
    var suma = 0.0
    var kompenzator = 0.0
    for i = 1 to NizBrojeva.length do

            #Primena ispravku
        var y = NizBrojeva[i] - kompenzator
            #Inkrement zbir
        var t = suma + y
            #Računamo kompenzator
        kompenzator = (t - suma) - y

        suma = t

    next i

    return sum
```

Slika 2. Pseudokod Kahanovog algoritma za sumiranja.



Slika 3. Histogram koji prikazuje distribuciju kotiranog raspona u broju tikova za izabrani trgovački dan (2019-07-01) za MU akcije.

6 STATISTIČKA ANALIZA

Manji kotirani rasponi ponude i traženja su znak veće likvidnosti, dok su veći kotirani rasponi znak manje likvidnosti ili veoma promenljivih akcija. U situacijama kada je kotirani raspon veliki, znatno je teže trgovati u i van pozicije po fer ceni jer je manja likvidnost.

Prosečna razlika tokom svih vremena trgovanja je izračunata korišćenjem algoritma Kahan Sumation (da bi se izbegle numeričke greške kod sabiranje velikih brojeva). Simulacija programa nam govori da je prosečan raspon jednak 0,0111745351512686445 za akcije kompanije MU u toku izabranog trgovačkog dana (01.07.2019).

Na Slici 3 je prikazan histogram koji prikazuje distribuciju kotiranog raspona u odnosu na broj tikova za izabrani trgovački dan. Ovaj histpgram sumira rezultate dobijene programskom simulacijom, korišćenjem Kahanovog algoritma za sumiranje i potvđuje pretpostavku da je kotirani raspon najčešće jednak upravo

vličini jednog ticka. To nam ukazuje da postoji zadovoljavajuća likvidnost i da berza dobro funkcioniše.

S obzirom da radimo sa velikom bazom podataka, tehnike masinskog ucenja nam mogu pomoci da lakse obradimo podatke i izdvojimo relevantne karakteristike kako bi u daljem istrazivanju uz pomoc neuronske mreze klasifikovali svaki vektor posmatrane baze podataka.

NAPOMENA (ZAHVALNICA)

Autorka je zahvalna Profesoru Dr. Thorsten Rheinländer na komentarima i konstruktivnim sugestijama, kao i na pomoći u analizi LOBSTER baze podataka.

LITERATURA

1 Cont, R., Kukanov, A., and Stoikov, S. The price impact of order book events. Journal of financial econometrics, 12(1):47–88, 2014.

2 Palguna, D. and Pollak, I. Mid-price prediction in a limit order book. IEEE Journal of Selected Topics in Signal Processing, 10(6):1083–1092, 2016.

3 Zheng B., Moulines E., and Abergel, F. Price jump prediction in limit order book. arXiv preprint arXiv:1204.1381, 2012.

4 Radojičić, D., Kredatus, S. and Rheinländer, T., 2018, November. An approach to reconstruction of data set via supervised and unsupervised learning. In 2018 IEEE 18th International Symposium on Computational Intelligence and Informatics (CINTI) (pp. 000053-000058). IEEE.

5 Radojičić, D. and Kredatus, S., 2020. An approach for processing data from NASDAQ stock exchange database. In Proceedings of the 10th International Conference on Information Society and Technology (ICIST 2020). Society for Information Systems and Computer Networks.

6 Radojičić D, Radojičić N, Kredatus S. A multicriteria optimization approach for the stock market feature selection. Computer Science and Information Systems. 2021;18(3):749-69.

7 Radojičić D, Kredatus S. The impact of stock market price fourier transform analysis on the gated recurrent unit classifier model. Expert Systems with Applications. 2020 Nov 30;159:113565.

8 Gould, M.D., Porter, M.A., Williams, S., McDonald, M., Fenn, D.J., Howison, S.D.: Limit order books. Quantitative Finance 13(11), pp 1709–1742, 2013

9 Delattre S, Robert CY, Rosenbaum M. Estimating the efficient price from the order flow: a Brownian Cox process approach. Stochastic Processes and their Applications. 2013 Jul 1;123(7):2603-19.

10 Dayri K, Rosenbaum M. Large tick assets: implicit spread and optimal tick size. Market Microstructure and Liquidity. 2015 Jun 4;1(01):1550003.

11 Muranaga J, Ohsawa M. Measurement of liquidity risk in the context of market risk calculation. a BIS volume entitled The Measurement of Aggregate Market Risk. 1997.

12 Hagströmer B. Bias in the effective bid-ask spread. Journal of Financial Economics. 2021 Oct 1;142(1):314-37.

13 Guloglu ZC, Ekinci C. A comparison of bid-ask spread proxies: Evidence from Borsa Istanbul futures. Journal of Economics Finance and Accounting. 2016;3(3):244-54.

14 Kahan W. Pracniques: further remarks on reducing truncation errors. Communications of the ACM. 1965 Jan 1;8(1):40.

15 Babuska I. Numerical stability in mathematical analysis. InIFIP Congress (1) 1968 Aug (Vol. 68, pp. 11-23).

16 Bresenham JE. Algorithm for computer control of a digital plotter. IBM Systems journal. 1965;4(1):25-30.