Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Anže Gregorc

Uporaba strojnega učenja za kvantitativne trgovalne strategije

diplomsko delo

univerzitetni študijski program
prve stopnje
računalništvo in informatika

Mentor: prof. dr. Igor Kononenko

Ljubljana, 2018

Besedilo je oblikovano z urejevalnikom besedil LaTeX.

Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Tematika naloge:

Cilj diplomske naloge je izdelava trgovalne strategije, ki je donosna v različnih trendih (npr. v naraščajočem, padajočem ter stranskem). Namen strategije je lahko npr. izkoristiti nihanja cene iz lokalnih ekstremov k povratku srednje vrednosti (ang. mean reversion). V praksi to pomeni, da po veliki rasti cene sama cena tendenčno zaniha proti srednji vrednosti danega časovnega okna. Omenjena strategija lahko deluje v različnih časovnih intervalih z uporabo različnih statističnih metod. Kandidat seveda lahko predlaga in preizkusi tudi druge strategije. V okviru diplomske naloge se preizkusijo algoritmi strojnega učenja za implementacijo in testiranje trgovalne strategije na realnih podatkih.

Zahvaljujem se mentorju prof. dr. Igorju Kononenku za vodenje in strokovno pomoč. Zahvaljujem se tudi prijateljem, ki so mi pomagali. Zahvala gre tudi družini za podporo skozi leta študija.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
ETF	exchange-traded fund	indeksni vzajemni skladi
BAC	Bank of America Corporation	Bank of America Corporation
BTC	Bitcoin	Bitcoin
USD	US Dollar	ameriški dolar
SMA	simple moving average	preprosto drseče povprečje
EMA	exponential moving average	eksponentno drseče povprečje
BB	Bollinger band	Bollingerjev pas
RSI	relative strength index	indeks relativne moči
MACD	moving average convergence divergence	drseča konvergenca divergenca
OBV	on-balance volume	indikator OBV
CLV	close location value	indikator CLV
EDA	exploratory data analysis	raziskovalna analiza podatkov
k-NN	k-nearest neighbors	k-najbližjih sosedov
ANN	artificial neural networks	umetne nevronske mreže
CA	classification accuracy	klasifikacijska točnost

Povzetek

Naslov: Uporaba strojnega učenja za kvantitativne trgovalne strategije

Avtor: Anže Gregorc

V diplomski nalogi smo s pomočjo zgodovinskih podatkov delnice BAC, menjalnega tečaja BTC-USD in tehnične analize trgovanja naučili različne modele strojnega učenja trgovalnih strategij. Podatke iz tehnične analize smo pridobili s pomočjo različnih indikatorjev. Podatke smo označili po strategiji, ki temelji na povratku k srednji vrednosti. Strategija se lahko spreminja s pomočjo parametrov in tako izkoristi nihanja lokalnih ekstremov v različnih dolžinah časovnega okna. Pozorni smo na to, da je strategija donosna v različnih trendih (naraščajočem, padajočem ter stranskem). Podrobno smo opisali uporabljene metode, torej modele nadzorovanega strojnega učenja. Strategije smo preizkusili na omenjenih realnih podatkih in analizirali rezultate. Dobro strategijo, ki je donosna tako v naraščajočem kot tudi padajočem trendu, je prikazal le eden od modelov. To je naključni gozd z 10 drevesi.

Ključne besede: trgovalne strategije, tehnična analiza, strojno učenje, menjalnica, borza.

Abstract

Title: Using machine learning for quantitative trading strategies

Author: Anže Gregorc

In this thesis, we trained different machine learning models the trading strategies with the help of the historical data of BAC shares, the exchange rate of BTC-USD and the technical analysis of trading. The data from the technical analysis were obtained by using different indicators. We marked the data according to the strategy based on the mean reversion. The strategy can be changed with the help of parameters and thus exploiting the fluctuations of the local extremes in different lengths of the time window. We are mindful of the fact that the strategy should be profitable in various trends (the rising, the decreasing, and the lateral). We described in detail the methods used, i.e. the models of supervised machine learning. The strategies were tested on aforementioned real data and the results were analysed. A good strategy that is profitable, both in rising and decreasing trends, was only achieved by one of the models. That is a random forest with 10 trees.

Keywords: trading strategies, technical analysis, exchange, bourse.

Poglavje 1

Uvod

Trgovanje je stara panoga, ki sega vse do trinajstega stoletja, kjer so se v hiši družine Van der Beurze zbirali trgovci z blagom, posredniki in investitorji ter tam sklepali posle [22]. Glavni namen vsakega trgovca je kupiti z namenom nadaljnje prodaje ter pri tem zaslužiti. Zato mora imeti dobro načrtovano trgovalno strategijo, ki mu omogoča prodati dobrino po večji ceni, kot jo je kupil.

Ljudje so skozi obdobja poskusili razviti veliko trgovalnih strategij. Temeljijo predvsem na dveh kategorijah: fundamentalni in tehnični analizi. Fundamentalna analiza temelji na upoštevanju socialnega, političnega in ekonomskega stanja na nekem območju, ki uporablja določeno dobrino. Tehnična analiza pa je študija tržnih podatkov z različnimi statističnimi orodji. Tehnična analiza predvideva, da trgovanje v preteklosti vpliva na gibanje cene v prihodnosti [11].

Borze so se skozi čas razvijale. Sedaj so vse največje borze avtomatizirane ter imajo na voljo tudi trgovanje preko spleta. Tako je trgovanje precej bolj dinamično. Ljudje ne potrebujejo biti fizično prisotni na borzi, kjer svoj denar menjajo za dobrino. Borze sedaj ponujajo spletne platforme, ki omogočajo trgovanje praktično kjerkoli. Posameznik mora imeti le spletni bančni račun ter internetno povezavo in že lahko prične trgovati.

1.1 Motivacija za izbrano diplomsko temo

Živimo v času, ko prav na vseh panogah poizkuša avtomatizirati čim več stvari. Enako velja tudi pri trgovanju. Računalnik veliko hitreje lahko kupi oziroma proda določeno dobrino. Vedno bolj pa so uspešni tudi pri naprednih trgovalnih strategijah. Naučijo se lahko strategij, ki si jih človek prej niti ni uspel zamisliti.

Malo za šalo, slika 1.1 predstavlja uspešni trgovalni stroj. Mi seveda ne bomo izdelali robotske roke, ki bo kupovala in prodajala dobrine, ampak program, ki bo to počel s pomočjo algoritmov strojnega učenja.

Slika 1.1: Trgovalni robot [16].

1.2 Cilj diplomske naloge in glavni koraki

Glavni cilj diplomske naloge je izdelati trgovalno strategijo s pomočjo strojnega učenja, ki je donosna v različnih trendih (na primer v naraščujočem,

padajočem ter stranskem). Uporabili smo le tehnično analizo, ki temelji na podlagi kvantitativnih zgodovinskih podatkov trgovanja. Prvi korak je zagotovo zbiranje podatkov. Potrebno je najti zgodovino trgovanja za določeno dobrino. Naslednji pomemben korak je pregled literature, kaj že obstaja, kateri modeli strojnega učenja so bili v preteklosti uspešni pri napovedovanju trendov ter posledično trgovalnih strategijah. Nato pa z uporabo večih napovednih modelov najti nekaj dobrih trgovalnih strategij.

V poglavju 2 najprej opišemo vrste trgov. Temu sledi opis podatkov, ki smo jih uporabili v tej diplomski nalogi. V poglavju 3 na kratko razložimo, kaj je tehnična analiza. Glavni pripomoček tehnične analize so indikatorji, ki jih razdelimo na več tipov. Nato pa bolj podrobno opišemo indikatorje, ki smo jih uporabili. Lastnosti trgovalnih strategij so v poglavju 4. V razdelku 4.1 pa se osredotočimo na trgovalno strategijo, ki izkoristi nihanja cene iz lokalnih ekstremov k povratku srednje vrednosti. Na koncu poglavja še predstavimo našo implementacijo omenjene strategije. Poglavje 5 je namenjeno opisu modelov strojnega učenja, ki smo jih uporabili. Postopek gradnje modelov z različnimi vrednostmi parametrov in izbiro najboljših je opisan v poglavju 6. Nato sledi prikaz rezultatov in razlaga uspešnosti posameznih učnih modelov. Zadnje poglavje 7 pa govori o možnih izboljšavah za nadaljnje delo.

Poglavje 2

Podatki

Osnoven vir podatkov so predstavljale cene različnih dobrin v preteklosti. Te podatke smo pridobili na platformi Yahoo Finance, ki je eden večjih medijskih hiš. Je del omrežja Yahoo! ter ponuja finančne novice, podakte o borznih kotacijah, finančna poročila in še mnogo drugih stvari. Zgodovinske podatke ponuja iz vseh večjih skupin trgov.

- **Surovine** so osnovne dobrine [5]. Na trgu so uporabljene večinoma za proizvodnjo in predelavo v nove izdelke oziroma storitve. Mednje spadajo zlato, srebro, surova nafta, kakav, sladkor, pšenica ...
- **Delnice** so oblike vrednostnega papirja, ki imetniku predstavlja delež lastništva v podjetju. Imetnik tako pridobi pravico do izplačila dividende ali pravico do glasovanja v podjetju [10].
- **Svetovni indeksi** pri financah so statistično merilo sprememb na trgu vrednostnih papirjev. V primeru finančnih trgov so indeksi trga delnic sestavljeni iz hipotetičnega portfelja vrednostnih papirjev, ki predstavljajo določen trg ali njen segment. Na primer S&P 500 je skupno merilo za Ameriške delnice in obveznice [8].
- **Valute (forex)** so trgi, kjer se trguje z denarnimi valutami. Tak trg je največji, saj povprečno na dan doseže nekaj bilijonov dolarjev trgovane vrednosti. Vključuje vse valute na svetu [13].

- **Kriptovalute** so digitalne oziroma virtualne valute, ki za varnost uporabljajo kriptografijo. Najpomembnejša lastnost kriptovalut je decentraliziranost. To pomeni, da je odporen na vmešavanje ali manipulacijo držav oziroma nekega osrednjega organa [6].

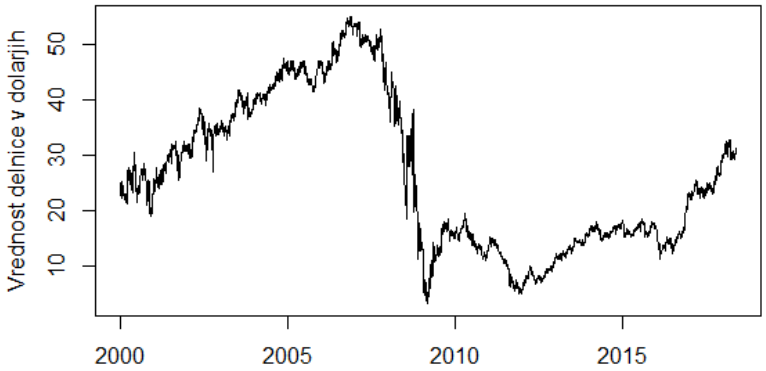
- **ETF** ali borzno trgovalni skladi so vrednosti papirji, ki sledijo indeksom, surovinam, obveznicam ali drugim sredstvom. Od vzajemnih skladov se razlikujejo po tem, da so ETF namenjeni vlaganju in ne ponujajo varčevalnih načrtov [7].

- **Vzajemni skladi** so naložbena sredstva, sestavljena s premoženjem večjega števila investitorjev. Vzajemne sklade upravljajo denarni upravitelji, ki vlagajo kapital sklada in poskušajo ustvariti kapitalske dobičke in dohodke vlagateljev sklada.

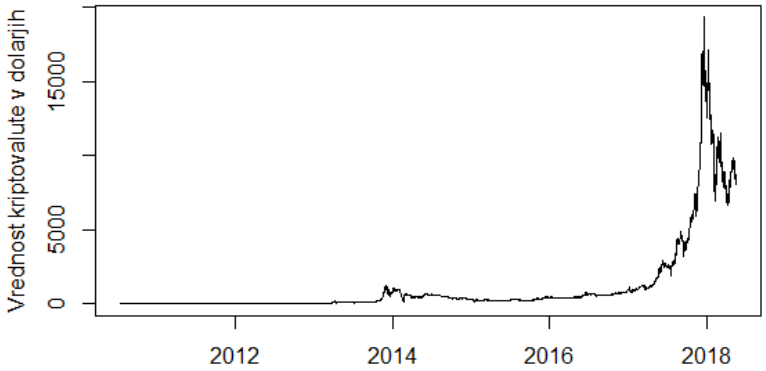
V tej diplomski nalogi smo se omejili na delnice BAC (Bank of America) ter kriptovaluto BTC (Bitcoin) z menjalnim tečajem BTC-USD. Gibanje delnice BAC in kriptovalute BTC je prikazano na slikah 2.1 in 2.2 Pridobili smo dnevne vrednosti omenjenih dobrin. Za delnico BAC smo dobili dnevni interval med 3.1.2000 in 23.5.2018. Za kriptovaluto BTC smo dobili dnevni interval med 17.7.2010 in 23.5.2018.

Podatki vključujejo vrednosti na začetku izbranega intervala, nato sledi najvišja, najnižja, končna vrednost določene dobrine ter volumen. Iz teh podatkov se izračunajo tehnični indikatorji, ki so opisani v poglavju 3. Primer zapisa podatkov:

DATE,OPEN,HIGH,LOW,CLOSE,VOLUME
2018-04-23,30.270000,30.400000,30.120001,30.320000,50686300
2018-04-24,30.459999,30.860001,30.000000,30.190001,81486800
2018-04-25,30.090000,30.340000,29.799999,30.139999,65893500



Slika 2.1: Gibanje vrednosti delnice BAC med 3.1.2000 in 23.5.2018.



Slika 2.2: Gibanje vrednosti kriptovalute BTC med 27.7.2010 in 23.5.2018.

Poglavje 3

Tehnična analiza in indikatorji

Tehnična analiza je študija preteklih tržnih podatkov. Ukvarja se z verjetnostmi in nikoli ni 100% zanesljiva. Tehnični pristop pri investiranju je prevzaprav odraz ideje, da se cene gibljejo v trendih. Na trende vplivajo spreminjajoči se pogledi investitorjev na različne ekonomske, monetarne, politične in psihološke dejavnike. Umetnost tehnične analize je identificirati preobrat trenda v relativno zgodnji fazi in vztrajati po tem trendu, dokler ne se ne pojavi nov preobrat [17].

Glavni pripomoček tehnične analize so indikatorji, ki jih razdelimo na več tipov. Med njimi so tudi [20]:

- **Tradicionalna (Zahodna) tehnična analiza** vključuje širok razpon indikatorjev na podlagi vzorcev cen. Te vsebujejo vsa pomembna testiranja mej trga (odpornost na vrhu in podpora na dnu).

- **Indikatorji preobratov svečnikov** so vzorci v določenem časovnem oknu, ki signalizirajo preobrat v ceni. To so signali, ki na podlagi momenta in spreminjanja nakupne ali prodajne moči prepoznavajo vzorce, ki predvidevajo konec trenutnega trenda in začetek novega.

- **Trendi** so izraženi v več oblikah, tudi s ceno in momentom. Trendu se lahko sledi z indikatorji drsečega povprečja in trendnimi linijami.

- **Indikatorji količine (volumna)** najdemo v številnih oblikah, ki so zasnovani tako, da se s pomočjo sprememb v volumnu ugotovi, kdaj se zgodi preobrat trenda. Količino trgovanja velikokrat imenujemo volumen.

- **Indikatorji momenta** ne merijo samo smeri trenda, temveč tudi njihovo hitrost. Ko se moment začne upočasnjevati, to verjetno pomeni, da kupci ali prodajalci izgubljajo nadzor nad trendom in tako druga stran začne spreminjati ceno in zgodi se preobrat trenda.

Odločili smo se, da uporabimo indikatorje iz večih različih tipov. Indikatorji istega tipa se opirajo na podobne podrobnosti v podatkih in tako skupek indikatorjev istega tipa ne prinese večje dodatne informacije. Z uporabo indikatorjev različnih tipov lahko lažje vidimo stanje trga iz več različnih pogledov in je naša strategija zaradi tega bolj kvalitetna [3].

V naslednjih podpoglavjih so opisani glavni indikatorji, ki smo jih uporabili.

3.1 Preprosto drseče povprečje (SMA)

Preprosto drseče povprečje je osnoven indikator trenda. Izračuna se tako, da se za n časovnih obdobij doda cena ob zaprtju trga in nato to skupno vsoto deli z n.

$$SMA(n) = \frac{\sum_{i=1}^{n} a_i}{n}$$

Od izbire vrednosti n je odvisno, kakšno obdobje nas zanima. Če je n majhen (okoli 15), se SMA bolj odziva na manjše spremembe v ceni in mu rečemo kratkoročno povprečje. Ko je n okoli 50, je graf SMA bolj gladek, saj se manj prilega kratkoročnim spremembam. Takrat merimo dolgoročno povprečje.

Slika 3.1: Graf prikazuje gibanje cene BTC z indikatorji SMA. Modra linija predstavlja 15 dnevni SMA, oranžna pa 50 dnevni SMA [21].

3.2 Eksponentno drseče povprečje (EMA)

Podoben indikator, kot SMA, le da je bolj utežen na podatke bližnje preteklosti. Ta tip drsečih povprečij reagira hitreje na razlike v ceni kot SMA.

$$EMA = (cena\,ob\,zaprtju - EMA_{prejšnja}) * \alpha + EMA_{prejšnja}$$

$$\alpha = \frac{2}{izbrana\,časovna\,perioda + 1}$$

Primer vrednosti α za 10 časovnih obdobij: $(2/(10 + 1)) = 0.1818 = 18.18\%$ Prva vrednost EMA v seriji podatkov pa je SMA.

Slika 3.2: Graf prikazuje gibanje cene BTC, črna linija pa je 10 dnevni EMA [21].

3.3 Bollinger Band® (BB)

Indikator je razvil znan tehnični trgovec John Bollinger. Prikazuje zgornjo in spodnjo vrednost dveh standardnih deviacij stran od indikatorja SMA. Standardna deviacija je mera volatilnosti[1]. Ko je volatilnost trga večja, je tudi pas med zgornjo in spodnjo vrednostjo BB širši. Pri manjši volatilnosti trga pa se pas oži.

[1]Volatilnost ali nihajnost je statistična mera za verjetnost, da cena (delnice ali točke sklada) v kratkem času močno zraste ali pade.

Slika 3.3: Graf prikazuje gibanje cene BTC ter zgornjo in spodnjo mejo indikatorja BB. Srednja linija pa predstavlja 20 dnevni SMA [21].

3.4 Indeks relativne moči (RSI)

Indeks relativne moči je indikator momenta, ki ga je razvil tehnični analitik Welles Wilder. Primerja velikost nedavnih dobičkov in izgub v določenem časovnem obdobju in tako meri hitrost in spremembo gibanja cen trga. Primarno se uporablja za ocenjevanje precenjenosti in podcenjenosti trga.

$$RSI(n) = 100 - \frac{100}{1 + RS}$$

$$RS = \frac{AvgU}{AvgD}$$

AvgU je povprečje vseh premikov cene navzgor v zadnjih n časovnih intervalih. AvgD pa predstavlja povprečje vseh premikov cene navzdol v zadnjih n časovnih intervalih.

Slika 3.4: Graf prikazuje gibanje cene BTC ter indikator RSI [21].

3.5 Drseča konvergenca divergenca (MACD)

MACD je tako trendni kot momentni indikator, ki prikazuje razmerje med dvema drsečima povprečjema cen. Standardni MACD se izračuna tako, da od 26 dnevnega EMA odštejemo 12 dnevni EMA. 9 dnevni EMA od indikatorja MACD, ki se imenuje "signalna linija", se nato doda prikazu MACD in deluje kot signal za nakup in prodajo dobrine.

$$MACD = EMA(12) - EMA(26)$$

$$signalna\,linija = EMA_{MACD}(9)$$

$$MACD_{histogram} = MACD - signalna\,linija$$

Slika 3.5: Graf prikazuje gibanje cene BTC z indikatorjem MACD. Modra linija je MACD, oranžna predstavlja signalno linijo, izrisan pa je tudi histogram MACD [21].

3.6 Indikator OBV

OBV je indikator momenta, ki uporablja pretok volumna za predvidevanje sprememb v ceni trga. Joseph Granville je razvil metriko OBV v šestdesetih letih prejšnjega stoletja. Verjel je, da se bo cena sčasoma povzpela navzgor, ko se količina trgovanja močno poveča brez znatne spremembe cene trga.

$$OBV = \begin{cases} OBV_{prejšnji} + (Trenutni\,volumen), & \text{če je cena ob zaprtju večja od prejšnje} \\ OBV_{prejšnji} - (Trenutni\,volumen), & \text{če je cena ob zaprtju manjša od prejšnje} \\ OBV_{prejšnji}, & \text{sicer} \end{cases}$$

Slika 3.6: Graf prikazuje gibanje cene BTC ter indikator OBV [21].

3.7 Indikator CLV

CLV je mera, ki oceni, kje se cena trga zapre v primerjavi z najvišjo ter najnižjo dnevno ceno. Giblje se med +1 in -1, kjer vrednost +1 pomeni, da je zaprtje enako najvišji ceni in vrednost -1 pomeni, da je zaprtje enako najnižji ceni.

$$CLV = \frac{(close - low) - (high - close)}{(high - low)}$$

3.8 Stohastični RSI

Stohastični RSI je indikator, ki se giblje med 0 in 100. Podatki indikatorja RSI se uporabijo za izračun stohastičnega oscilatorja. Formula stohastičnega oscilatorja je sledeča:

$$\%K = 100 * \frac{close - low}{high - low}$$

$$\%D = 3 \text{ dnevno drseče povprečje } \%\text{K}$$

kjer vrednost **close** predstavlja zadnjo vrednost ob zaprtju, **low** je najnižja vrednost v določenem časovnem oknu, **high** pa najvišja vrednost v določenem časovnem oknu. Za izračun stohastičnega RSI samo zamenjamo vrednosti:

$$\%K = 100 * \frac{RSI - Najnižji\,RSI}{Najvišji\,RSI - Najnižji\,RSI}$$

Slika 3.7: Graf prikazuje gibanje cene BTC ter indikator Stohastični RSI. Modra črta prikazuje vrednost %K, oranžna pa %D [21].

Poglavje 4

Trgovalne strategije

Trgovalna strategija opisuje specifikacije za trgovanje, vključno s pravili za nakup in prodajo dobrine in upravljanje denarnih sredstev. Ko je pravilno testirana in izvedena, lahko trgovalna strategija zagotovi matematična pričakovanja za določena pravila, kar pomaga trgovcem in vlagateljem ugotoviti, ali je strategija donosna. Trgovalne strategije niso jamstvo za uspeh, vendar pa so lahko učinkovite pri povečevanju donosov, prilagojenih tveganjem [12]. Obstaja veliko različnih tipov strategij, kot pa je že omenjeno, smo se mi omejili na tehnične trgovalne strategije.

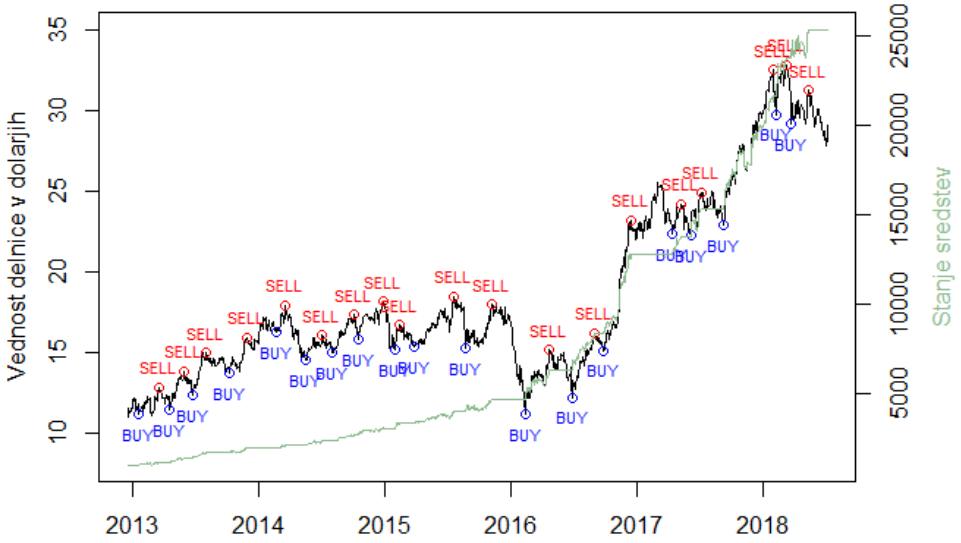
4.1 Povratek k srednji vrednosti

Narava je polna primerov povratka srednje vrednosti. Od spremembe višine vode, ki teče po reki Nil, do spremembe ravni ozona. Psiholog Daniel Kahneman je za ameriško športno-novinarsko franšizo Sports Illustrated izjavil znameniti stavek: "Usoda športnika, katerega slika je v naslovnici te revije, je takšna, da bo naslednjo sezono nastopal slabše" [14]. Znanstveni razlog je, da lahko športnikovo nastopanje ocenimo kot naključno porazdeljeno okoli sredine, torej po izjemno dobri predstavi v eni sezoni (kar postavi športnika na naslovnico revije Sports Illustrated) zelo verjetno sledi bolj povprečna v naslednji.

Ideja strategije pri trgovanju je izkoristiti nihanja cene iz lokalnih ekstremov k povratku srednje vrednosti (ang. mean reversion). Časovne vrste, ki se tako gibljejo, imenujemo *stacionarne časovne vrste*. Vendar pa se cene v večini primerov ne gibljejo tako, ampak kot geometrijski naključni sprehod s trendom [2]. *Nestacionarne časovne vrste* pa teoretično lahko transformiramo v *stacionarne* na veliko načinov:

- **Detrending** oziroma odstranjevanje trenda iz časovne vrste. Eden od načinov je odštevanje regresije časovne vrste od cen iste časovne vrste.
- **Sezonske prilagoditve** so razlike v tehnikah odstranjevanja trendov v različnih časih.
- **Transformacije** podatkov lahko pretvorijo v stacionarno časovno vrsto. Na primer pretvorba razlik cen v odstotke.
- **Tehnike glajenja EDA** lahko razumemo kot neparametrične načine odstranjevanja trendov. Primer je odstranjevanje (odštevanje) drseče mediane. John W. Tukey je v svoji knjigi o raziskovalni analizi podatkov [23] postopek odstranjevanja trendov ponavljal toliko časa, dokler niso končni podatki postali stacionarni in simetrično porazdeljeni okoli 0.

V tej diplomski nalogi smo implementirali strategijo na podoben način. *Kupi* takrat, ko je cena najmanjša v razponu t časovnih enot ter *prodaj*, ko je cena največja v razponu t časovnih enot. Vse preostale podatke pa smo označili kot ne naredi ničesar oziroma drži (angl. *hold*). S to strategijo se nanašamo prav na povratek k srednji vrednosti ter poizkušamo čim bolj izkoristiti nihanje cene iz lokalnih ekstremov. Primer izvedbe strategije na realnih podatkih prikazuje slika 4.1. Seveda je pa to le teoretična zmogljivost strategije, saj smo imeli vse informacije glede gibanja cene in postavili *kupi*/*prodaj* na najdonosnejše mesto. Služi nam kot označevanje podatkov za modele nadzorovanega strojnega učenja.



Slika 4.1: Graf prikazuje gibanje cene delnice BAC. Pri točkah BUY bi po naši strategiji kupili, pri SELL prodali. Zelena črta prikazuje stanje sredstev, če bi na začetku vložili 1000 dolarjev. Kot vidimo, bi s to strategijo v približno šestih letih imeli več kot 25000 dolarjev, torej kar 2500% povečanje stanja sredstev.

Poglavje 5

Uporabljene metode

Za preizkus delovanja trgovalne strategije smo uporabili metode strojnega učenja. V nadaljevanju smo predstavili metode nadzorovanega strojnega učenja, ki smo jih uporabili pri našem delu.

Nadzorovano učenje je poimenovano tako, ker lahko delujemo kot vodnik za učenje algoritma. To pomeni, da so možni izidi algoritma že znani, ter podatki, uporabljeni za učenje algoritma, so že označeni s pravilnimi odgovori.

Vhodna spremenljivka $x \in X$ je v našem primeru podana kot vektor zadnjih odstotkov nihanja cene ter vrednosti indikatorjev, opisanih v poglavju 3. Izhodna spremenljivka $y \in Y$ pa je označen podatek, v našem primeru to opisuje razdelek 4.1. Tehnika nadzorovanega učenja išče funkcijo $f : X \to Y$ tako, da se ta čim bolj prilega podatkom. Ker ima naša izhodna spremenljivka diskretno zalogo vrednosti, bomo naše nadzorovane učne modele uporabljali kot klasifikatorje. Ločimo jih glede na način predstavitve klasifikatorjeve funkcije. V naslednjih razdelkih so opisani klasifikatorji oziroma učni modeli, ki smo jih uporabili.

5.1 K najbližjih sosedov (k-NN)

Algoritem k-NN kot znanje uporablja kar množico vseh učnih primerov (samo zapomni si vse primere). Tako algoritem ne potrebuje v naprej zgrajenega modela, ampak se napovedi računajo šele ob poizvedbi novega primera. Tej vrsti učenja pravimo *leno učenje*. Pri klasifikaciji novega primera iz učne množice poišče k najbolj podobnih (najbližjih) primerov. Ker moramo nov primer primerjati z vsakim od učnih primerov, je zato časovna zahtevnost precej večja kot pri drugih metodah učenja. Nov primer klasificiramo v razred, ki mu pripada največ bližnjih primerov. Torej med učnimi primeri poiščemo k primerov $u_1, ..., u_k$, ki so najbližji novemu u_x in mu napovemo razred r_x iz množice vseh klasifikacijskih razredov $C \in \{C_1, ..., C_m\}$ po naslednji formuli:

$$r_x = \arg\max_{r \in c} \sum_{n=1}^{k} \delta(r, r^{(i)}) \tag{5.1}$$

kjer je

$$\delta(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases} \tag{5.2}$$

Parameter k običajno nastavimo na neko liho število npr. 1, 5, 15 in se tako izognemo morebitnim neodločenim primerom. S k reguliramo, kako se k-NN prilega podatkom. Manjši kot je k, bolj se bo model prilegal, vendar pa to lahko pomeni, da učni podatki, ki so napačni, pridejo bolj do izraza. V primeru, da je pa k prevelik, pa h klasifikaciji prispevajo tudi učni primeri, ki niso dovolj podobni novemu primeru in tako pripomorejo k napačni klasifikaciji [15].

Na kvaliteto napovedi pa vpliva tudi izbrana metrika pri računanju razdalj med novim primerom in učnimi primeri. Pogosto uporabljeni metriki sta evklidska in manhattanska razdalja. Razdalja para vhodnih spremenljivk u_1 in u_2 dimenzije n se z evklidsko formulo izračuna kot

$$d_E(u_1, u_2) = \sqrt{\sum_{i=1}^{n} (u_1^{(i)} - u_2^{(i)})^2} \tag{5.3}$$

z manhattansko pa

$$d_M(u_1, u_2) = \sum_{i=1}^{n} |u_1^{(i)} - u_2^{(i)}| \tag{5.4}$$

5.2 Naključni gozdovi (RF)

Naključni gozd sestavlja skupina odločitvenih klasifikacijskih dreves. Vsako drevo glasuje za določen razred. Končen razred primera, ki ga naključni gozd napove, je razred, kamor ga uvršča večina klasifikacijskih dreves. Za večjo klasifikacijsko točnost naključnega drevesa je pomembno, da ga sestavljajo drevesa, ki so med seboj čim bolj različna. Tako vsako drevo odkrije kakšen nov koncept, ki se skriva v podatkih [1].

5.3 Umetne nevronske mreže (ANN)

Modeli umetnih nevronskih mrež skušajo oponašati biološke nevronske mreže. Osnovni gradnik ANN je enota, ki jo zaradi podobnih lastnosti, kot v naravi, imenujemo nevron, ker pa si lahko ANN predstavljamo kot usmerjen graf, pa jo imenujemo tudi vozlišče. Vsak nevron je povezan v strukturo celotne mreže z drugimi nevroni. Te povezave predstavljajo sinapse, preko katerih nevron prejme različne signale od povezanih nevronov preko vhodne povezave. Vrednost signala, ki ga nevron izračuna, izračunamo po naslednji enačbi:

$$P_i = \sum_j W_{ji} X_j \tag{5.5}$$

kjer je X_j vrednost j-tega nevrona, W_{ji} utež sinapse med j-tim in i-tim nevronom in P_i nova vrednost i-tega nevrona. Vrednost signala pa v izhod

nevrona preslika aktivacijska (izhodna) funkcija:

$$Y_i = f(P_i) \tag{5.6}$$

Za izhodno funkcijo lahko vzamemo binarno deterministično funkcijo. Izhodne vrednosti so 0 in 1 ali -1 in 1. Lahko je tudi deterministična zvezna. Primer take funkcije, ki se v praksi velikokrat uporablja, je sigmoidna funkcija

$$f(x) = \frac{1}{1 + e^{-x}} \tag{5.7}$$

Nevronske mreže se razlikujejo tudi po tem, kako povežemo vozlišča. V tej diplomski nalogi smo uporabljali večnivojski perceptron. Spada med usmerjene ANN in ga lahko predstavimo v obliki usmerjenega acikličnega grafa, vhodne podatke pa vsako vozlišče dobiva iz prejšnjih nivojev. Pri klasifikaciji pa ima tipično ANN v zadnjem nivoju toliko vozlišč, kot je vseh klasifikacijskih razredov. Vozlišče z največjo izhodno vrednostjo nato določi razred določenem primeru.

Poglavje 6

Metodologija in rezultati

V tem poglavju opišemo postopek gradnje modelov z različnimi vrednostmi parametrov in izbiro najboljših. Nato sledi prikaz rezultatov in razlaga uspešnosti posameznih učnih modelov.

6.1 Gradnja modelov nadzorovanega učenja

Da bi prišli do čim boljših rezultatov, smo preizkusili veliko načinov gradnje modelov. Predvsem smo jim spreminjali različne parametre, ki so se nam zdeli pomembni pri iskanju uspešnega modela.

6.1.1 Priprava podatkov

Vhodni podatki modelov nadzorovanega učenja so 5 odstotkov podatkov zadnjih gibanj končnih cen (angl. *close price*) dobrine $(\frac{cena(t)}{cena(t-1)} - 1)$ skupaj z indikatorji. Te podatke smo tudi normalizirali. S tem smo se izognili morebitnim težavam, kadar imajo atributi različno skalo. Kot želen izhodni podatek smo vzeli označene podatke iz trgovalne strategije, ki smo jo opisali v razdelku 4.1. Podatke smo razdelili na učne, validacijske in testne približno v razmerju 60:20:20. Podatke pred razdelitvijo nismo premešali, saj sta si dva zaporedna podatka v časovni vrsti lahko med seboj odvisna.

6.1.2 Nabor parametrov

Modele smo poskusili graditi tudi z različnimi vrednostmi parametov. Spreminjali smo že samo trgovalno strategijo s tem, ko smo za cela števila t (opisan je v razdelku 4.1) vzeli od 1 do 20. S tem smo regulirali odstotek kupovanja in prodajanja dobrine. Če je bil t manjši, je bilo kupčij več in tako smo simulirali hitrejše trgovanje, ki poskuša ustvariti dobiček tudi z manjšimi premiki cen dobrine (kratkoročno trgovanje). Obratno lahko rečemo za večje vrednosti parametra t. Strategija takrat poišče večja lokalna nihanja cene in se tako osredotoči na bolj dolgoročno trgovanje. Ker je bilo veliko modelov kljub temu zelo nedejavno (v vsakem primeru je napovedal *drži* (angl. *hold*)), smo se odločili vpeljati še parameter agresivnosti. Ker smo modele gradili s pomočjo knjižnice CORElearn, smo si pomagali s parametrom costMatrix (cene napak) [18]. Kot je v tem delu zapisano, en element matrike cen napak predstavlja vrednost napake, ki jo model stori, če napove napačno stanje. Torej, če damo stanjema *kupi* in *prodaj* manjšo vrednost napake, bo z veliko verjetnostjo večkrat napovedal *kupi* in posledično bo model bolj agresiven. Naša matrika cen napak je predstavljena tako:

$$\text{costMatrix} = \begin{array}{c} \\ \textit{kupi} \\ \textit{drži} \\ \textit{prodaj} \end{array} \begin{array}{c} \begin{array}{ccc} \textit{kupi} & \textit{drži} & \textit{prodaj} \end{array} \\ \begin{bmatrix} 0 & 1 & 1 \\ \mathrm{c} & 0 & \mathrm{c} \\ 1 & 1 & 0 \end{bmatrix} \end{array}$$

pri čemer so vrstice pravo stanje in stolpci napovedano stanje primera. V diagonali matrike so vrednosti 0 in pomenijo, da ni škode, če napove stanje, ki je tudi pravilno. Vrednosti 1 so privzete vrednosti iz knjižnice [18]. Spremenljivki c pa spreminjamo vrednost. Ta se giblje od 0 do 1 v korakih po 0,05 in pomeni, če je prava vrednost *drži*, je potem cena napake, da se napove *kupi* oziroma *prodaj*, enaka c. Vrednosti parametrov t in c smo optimizirali na validacijski množici z izčrpno metodo, torej smo preizkusili vse kombinacije možnih vrednosti teh dveh parametrov ($20 * 21 = 420$ kombinacij).

Spreminjali smo tudi parametre, ki so značilni za vsak model posebej.

Pri metodi k-NN smo preizkusili relativno majhne vrednosti k (3 in 5), saj je stanj *kupi* in *prodaj* dosti manj, kot pa stanj *drži* in tako je prostor, kjer se dobrino *kupi*/*prodaj*, večji. RF smo preizkusili z 10 in 100 drevesi, nevronske mreže pa smo preizkusili z različnim številom nevronov z enim skritim slojem. Števila nevronov v skritem sloju smo postavili na 5, 25, 50, 100.

6.1.3 Filtriranje napovedanih izhodnih podatkov

Modeli nadzorovanega učenja ne vedo, ali smo nazadnje kupili ali prodali določeno dobrino. Zato se lahko zgodi, da v zaporedju podatkov model napove stanje *kupi*, ko imamo dobrino že v lasti. V takem primeru ne moramo upoštevati takega stanja. Ker smo predpostavili, da stanje *kupi* pomeni vloži vsa razpoložljiva sredstva v dobrino, se ne moreta zgoditi dve zaporedni stanji *kupi*. Vmes moramo dobrino tudi prodati.

Zaradi takih primerov smo napovedane izhodne podatke filtrirali. Upoštevali smo samo stanja *kupi* in *prodaj*, ki so smiselna. Nesmiselna stanja pa smo pretvorili v stanje *drži*.

6.1.4 Izbira najboljših modelov

Parametra t in c, opisana v razdelku 6.1.2, smo spreminjali pri vseh modelih nadzorovanega učenja, ki so opisani v poglavju 5. Samo zaradi teh dveh parametrov je bilo različic enega posameznega modela kar 420. Poleg tega pa ima vsak model še svoje interne parametre, ki smo jih spreminjali, zato smo morali med njimi najti najboljše. Za vsak model smo izbrali najboljšega iz rezultatov na validacijski množici podatkov. Ti modeli so predstavljeni tudi v razdelku 6.2. Da ovrednotimo, kateri model je boljši od drugega, smo se omejili na dva postopka:

- **Donosnost**

  Glavni cilj modelov je donosnost oziroma profitabilnost. Tako smo kot najboljše modele uvrstili tiste, ki so na podlagi validacijskih podatkov dosegli največjo donosnost. Kot začetno razpoložljivo sredstvo smo

vsakemu modelu dali 1000 USD. Nato smo s pomočjo cen dobrine v validacijskih podatkih ter napovedanih stanj modela simulirali trgovanje. Dobiček smo izračunali tako, da smo dobljeni končni vrednosti sredstev odšteli začetnih 1000 USD. Modele smo razvrstili po njihovem dobičku in vzeli tistega, ki ima največji dobiček na validacijski množici. Zatem smo izbran model testirali na testni množici.

- **Tveganje**

  Smiselno nam je bilo preveriti tudi, kolikšno je tveganje izgube naših sredstev, ko izberemo določen model. Tako smo se odločili izbrati še postopek, ki bi izbral najboljši model, pri katerem je najmanj tveganja.

  Sharpovo razmerje (angl. *Sharpe ratio*) je trenutno postala najbolj razširjena metoda za izračun tveganju prilagodljivega donosa [9]. Uporabili smo jo tudi za naše ovrednotenje modelov. Sharpovo razmerje je povprečni donos, ki presega donos dobrine brez tveganja na enoto volatilnosti oziroma celotnega tveganja [19]. Formula razmerja je sledeča:

$$\text{Sharpe ratio} = \frac{\overline{r}_p - r_f}{\sigma_p} \tag{6.1}$$

  kjer je $\overline{r}_p$ povprečna stopnja donosnosti določene dobrine pri danem modelu, r_f stopnja donosnosti dobrine, ki je brez tveganja, σ_p pa standardna deviacija donosa določene dobrine pri danem modelu. Kot r_f smo vzeli 8% letno stopnjo donosnosti. Ker pa so naši podatki dnevni, smo ta odstotek delili s 365 dnevi, saj se BTC trguje vsak dan, za BAC pa smo delili z 252, saj je približno toliko tudi trgovalnih dni na leto. Torej smo pri trgovanju z BTC vzeli $r_f \approx 0,00022$, pri trgovanju z BAC pa $r_f \approx 0,00032$.

  Primer izračuna Sharpovega razmerja: za nakup ameriških zakladnih računov (za katerega je pričakovana donosnost brez tveganja) je razmerje enako 0. Na splošno pa velja, da večja kot je vrednost Sharpovega razmerja, bolj privlačna je tveganju prilagodljiva donosnost [9]. Torej

smo modele razvrstili po Sharpovem razmerju in vzeli tistega, ki ima to vrednost največjo na validacijski množici. Zatem smo izbran model testirali na testni množici.

6.2 Rezultati

6.2.1 Rezultati pri trgovanju z menjalnim tečajem BTC-USD

Modele strojnega učenja smo testirali na obdobju od 15.5.2017 do 1.8.2018. To je približno eno leto in 3 mesece. Cena BTC se je dne 15.5.2017 gibala okoli 1880 USD, 1.8.2018 pa okoli 7542 USD. Najvišja cena v tem obdobju je bila 19870,62 USD, najniža pa 1791,12 USD. Cena je zelo nihala, saj je več kot 10 kratna razlika med najvišjo in najnižjo ceno. Za nas je to obdobje zanimivo, saj vsebuje tako naraščajoče kot tudi padajoče trende. Tako lahko analiziramo donosnost metod v obeh primerih.

Tako kot v razdelku 6.1.4, smo tudi pri rezultatih vsakemu modelu kot začetno razpoložljivo sredstvo dali 1000 USD. Torej s strategijo *kupi* na začetku in *drži* bi v tem obdobju imeli dobiček približno 3011 USD. Ta podatek nam lahko služi kot primerjavo z donosnostmi strategij modelov strojnega učenja. Rezultati na testnih podatkih pri modelih, ki so bili najdonosnejši na validacijski množici, so prikazani v tabeli 6.1, najboljši modeli pri primerjanju s Sharpovim razmerjem pa v tabeli 6.2. V tabeli 6.3 so podane vrednosti parametrov t in c, pri katerih so modeli dosegli najboljše rezultate na validacijski množici. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, 0,10, ..., 1.00). Uporabili smo izčrpno preiskovanje, torej smo preizkusili celoten nabor vrednosti parametrov in izbrali najboljše vrednosti. Na sliki 6.3 je prikazan graf donosnosti modela RF 10 pri vseh kombinacijah vrednosti parametrov t in c, graf vrednosti Sharpovega razmerja pri tem modelu z vsemi kombinacijami parametrov t in c pa na sliki 6.4. V Dodatek 7 smo dodali tabelo 7.1, ki prikazuje primerjave med donosnostjo in Sharpo-

vim razmerjem na validacijski in na testni množici za vse modele, ki so bili najboljši glede donosnosti ali tveganja na validacijski množici podatkov.

Najboljši rezultat pri iskanju najdonosnejših na testnih podatkih izmed modelov, ki so bili najdonosnejši na validacijskih, je dosegel model NN 25. Njegova strategija je prikazana na sliki 6.1. Vendar pa se je NN 25 najbolje odnesel le na testnih podatkih. Na validacijskih je bil najboljši model RF 100.

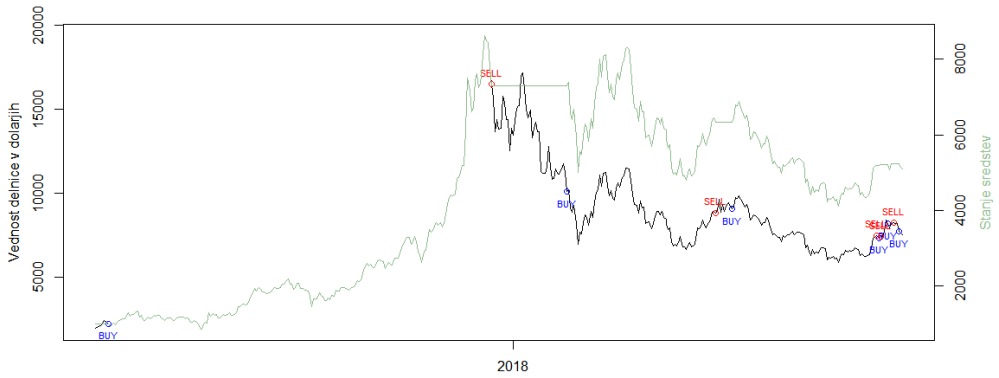
Najboljši rezultat pri iskanju najmanj tvegane strategije na testnih podatkih izmed modelov, ki so bili najdonosnejši na validacijskih, pa je dosegel RF 10. Njegova strategija je prikazana na sliki 6.2. Omenjeni model je imel največjo vrednost Sharpovega razmerja tudi na validacijskih podatkih.

Tabela 6.1: Rezultati modelov na testnih podatkih. Prikazani so modeli, ki so bili pri validacijskih podatkih najdonosnejši. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, 0,10, ..., 1.00). Številka pri RF modelu predstavlja število dreves, pri NN pa število vozlišč v skritem nivoju.

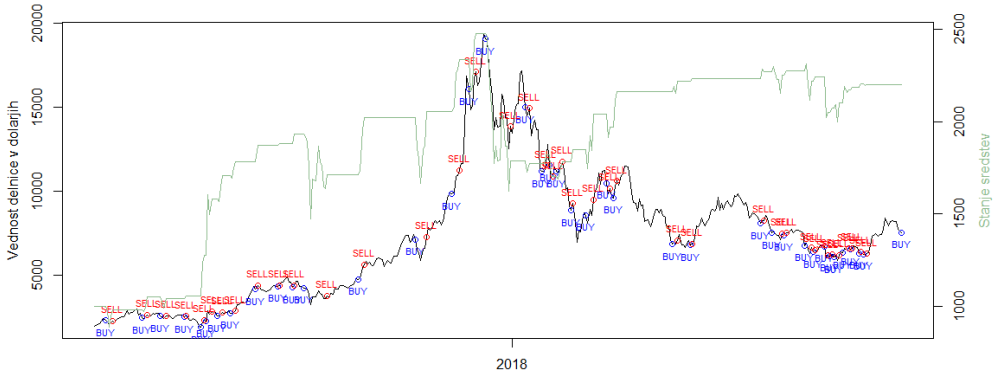
Model	**število kupčij**	**donosnost**	**Sharpovo razmerje**	**Utežena mera F_1**	**CA**
k-NN ($k = 3$)	15	648,8	0,82	0,77	0,82
k-NN ($k = 5$)	13	-164,7	0,05	0,85	0,88
RF 10	6	-304,70	-0,53	0,88	0,91
RF 100	4	1013,6	1,04	0,86	0,90
NN 5	4	1440,6	1,15	0,65	0,75
NN 10	6	688,6	1,06	0,65	0,75
NN 25	11	4101,0	1,85	0,41	0,55
NN 50	2	204,06	0,37	0,92	0,94
NN 100	2	329,18	0,58	0,90	0,93

Tabela 6.2: Rezultati modelov na testnih podatkih. Prikazani so modeli, ki so imeli pri validacijskih podatkih navječjo vrednost Sharpovega razmerja. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, 0,10, ..., 1.00). Številka pri RF modelu predstavlja število dreves, pri NN pa število vozlišč v skritem nivoju.

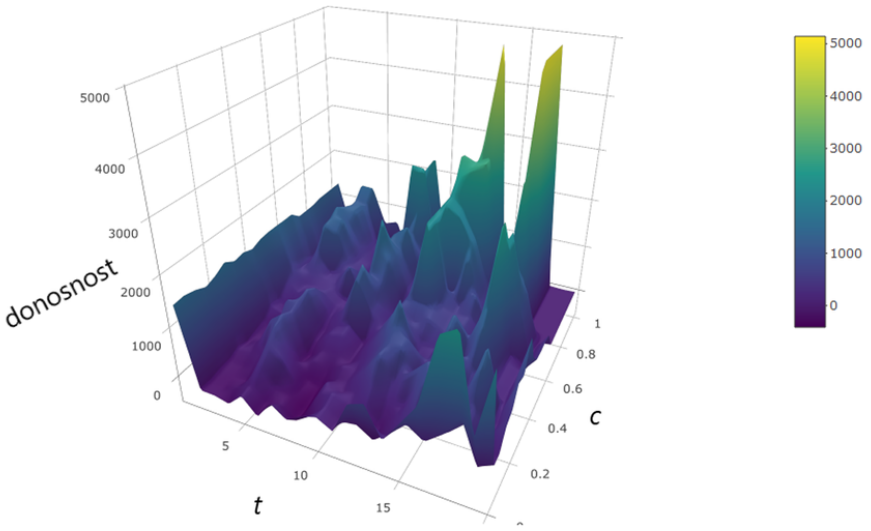
Model	**število kupčij**	**donosnost**	**Sharpovo razmerje**	**Utežena mera F_1**	**CA**
k-NN ($k = 3$)	15	648,8	0,82	0,77	0,82
k-NN ($k = 5$)	4	-351,5	-0,17	0,66	0,74
RF 10	79	1197,8	1,23	0,51	0,56
RF 100	4	1013,6	1,04	0,86	0,90
NN 5	4	1440,6	1,15	0,65	0,75
NN 10	6	688,6	1,06	0,65	0,75
NN 25	11	4101,0	1,85	0,41	0,55
NN 50	2	204,06	0,37	0,92	0,94
NN 100	2	329,18	0,58	0,90	0,93



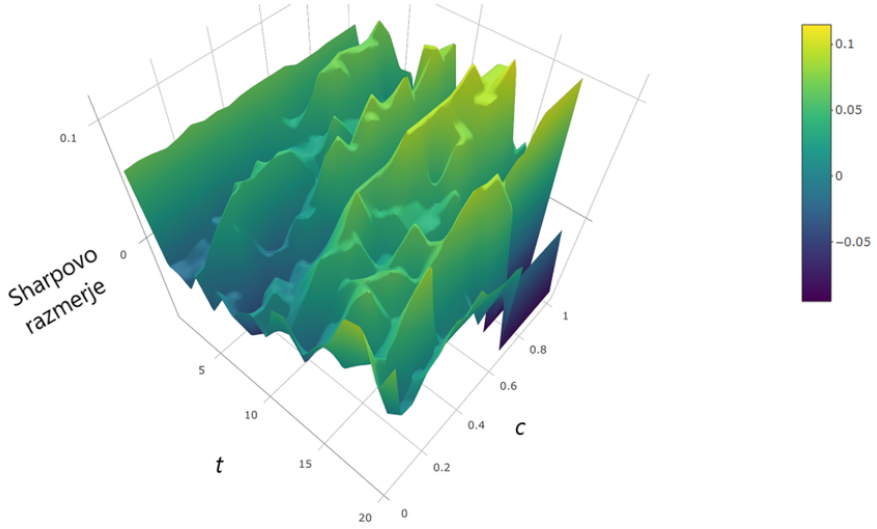
Slika 6.1: Graf prikazuje gibanje cene menjalnega tečaja BTC-USD in strategijo, ki jo je naredila NN s 25 vozlišči v skritem nivoju. Pri točkah BUY bi po strategiji kupili, pri SELL prodali. Zelena črta prikazuje stanje sredstev, če bi na začetku vložili 1000 dolarjev.



Slika 6.2: Graf prikazuje gibanje cene menjalnega tečaja BTC-USD in strategijo, ki jo je naredil RF z 10 drevesi. Pri točkah BUY bi po strategiji kupili, pri SELL prodali. Zelena črta prikazuje stanje sredstev, če bi na začetku vložili 1000 dolarjev.



Slika 6.3: 3D graf prikazuje donosnost modela RF 10 na validacijskih podatkih pri različnih parametrih t in c. Legenda nam pove, da bolj kot je rumena barva, večja je donosnost. Temno modra pa pomeni, da je bila donosnost okoli 0.



Slika 6.4: 3D graf prikazuje vrednost Sharpovega razmerja modela RF 10 na validacijskih podatkih pri različnih parametrih t in c. Legenda nam pove, da bolj kot je rumena barva, večja je vrednost Sharpovega razmerja. Temno modra pa pomeni, da je bila vrednost majhna.

Tabela 6.3: Leva tabela predstavlja parametre modelov iz tabele 6.1. Desna pa parametre modelov iz tabele 6.2.

Model	t	c
k-NN ($k = 3$)	6	0,5
k-NN ($k = 5$)	10	0,3
RF 10	14	0,9
RF 100	11	0,7
NN 5	5	0,5
NN 10	5	0,8
NN 25	2	0,6
NN 50	18	0,9
NN 100	16	0,6

Model	t	c
k-NN ($k = 3$)	6	0,5
k-NN ($k = 5$)	4	1
RF 10	2	0,9
RF 100	11	0,7
NN 5	5	0,5
NN 10	5	0,8
NN 25	2	0,6
NN 50	18	0,9
NN 100	16	0,6

6.2.2 Rezultati pri trgovanju z BAC

Modele strojnega učenja smo testirali na obdobju od 1.1.2014 do 31.12.2015. Testno obdobje je trajalo 2 leti. Cena delnice BAC se je dne 1.1.2014 gibala okoli 16,4 USD, 31.12.2015 pa okoli 16,8 USD. Najvišja cena v tem obdobju je bila 18,45 USD, najniža pa 14,51 USD. Torej je cena v primerjavi z BTC občutno manj nihala. Tudi začetna in končna cena sta skoraj enaki in lahko rečemo, da je v tem obdobju stranski trend.

Kot že omenjeno, smo tudi tukaj vsakemu modelu nastavili začetno razpoložljivao sredstvo na 1000 USD. S strategijo *kupi* na začetku in *drži* v tem obdobju ne bi prišli do kakšnega večjega donosa. Zaradi tega je bil cilj modelov v tem primeru le imeti donos pozitiven.

Rezultati na testnih podatkih pri modelih, ki so bili najdonosnejši na validacijski množici, so prikazani v tabeli 6.4, najboljši modeli pri primerjanju s Sharpovim razmerjem pa v tabeli 6.5. V tabeli 6.6 so podane vrednosti parametrov t in c, pri katerih so vsi ti modeli dosegli najboljše rezultate. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, 0,10, ..., 1.00). Uporabili smo izčrpno preiskovanje, torej smo preizkusili celoten nabor vrednosti parametrov in izbrali najboljše vrednosti. Na sliki 6.7 je prikazan graf donosnosti modela NN 25 pri vseh kombinacijah vrednosti parametrov t in c, graf vrednosti Sharpovega razmerja pri tem modelu z vsemi kombinacijami parametrov t in c pa na sliki 6.8. V Dodatek 7 smo dodali tabelo 7.2, ki prikazuje primerjave med donosnostjo in Sharpovim razmerjem na validacijski in na testni množici za vse modele, ki so bili najboljši glede donosnosti ali tveganja na validacijski množici podatkov.

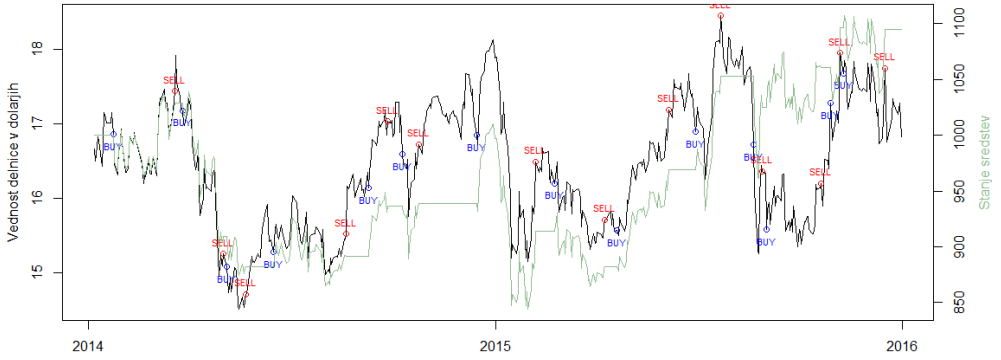
Najboljši rezultat pri iskanju najdonosnejših na testnih podatkih je dosegel model NN 25. Njegova strategija je prikazana na sliki 6.5. Ampak na validacijskih se NN 25 ni dobro odnesel, presegel ga je model RF 10. Slika 6.6 pa prikazuje model NN 50, ki je na validacijskih podatkih izgledal obetavno, a se mu trgovanje na testnih podatkih ni posrečilo.

Tabela 6.4: Rezultati modelov na testnih podatkih. Prikazani so modeli, ki so bili pri validacijskih najdonosnejši. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, ..., 1.00). Številka pri modelu RF predstavlja število dreves, pri NN pa število vozlišč v skritem nivoju.

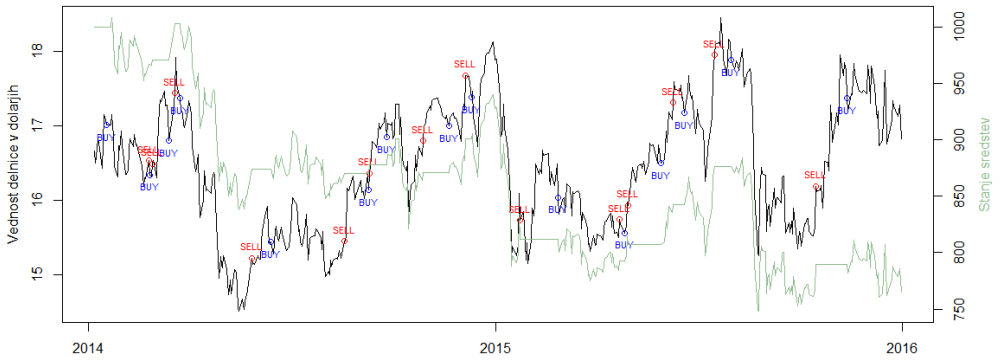
Model	**število kupčij**	**donosnost**	**Sharpovo razmerje**	**Utežena mera F_1**	**CA**
k-NN ($k = 3$)	57	17,8	-0,17	0,53	0,60
k-NN ($k = 5$)	8	-42,8	-0,31	0,84	0,86
RF 10	2	-17,8	-6,12	0,90	0,93
RF 100	5	-110,5	-0,57	0,90	0,93
NN 5	49	92,7	0,05	0,76	0,78
NN 10	7	25,9	-0,08	0,79	0,84
NN 25	28	95,2	0,06	0,89	0,89
NN 50	29	-234,8	-1,02	0,86	0,87
NN 100	29	56,3	-0,33	0,86	0,87

Tabela 6.5: Rezultati modelov na testnih podatkih. Prikazani so modeli, ki so imeli pri validacijskih podatkih največjo vrednost Sharpovega razmerja. Optimizacija je potekala po vrednostih parametrov t (1..20) in c (0,00, 0,05, 0,10, ..., 1.00). Številka pri modelu RF predstavlja število dreves, pri NN pa število vozlišč v skritem nivoju.

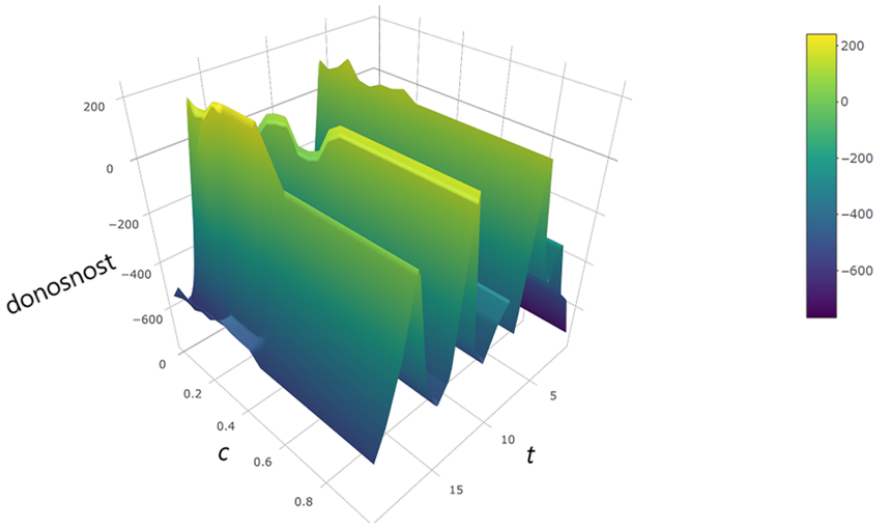
Model	**število kupčij**	**donosnost**	**Sharpovo razmerje**	**Utežena mera F_1**	**CA**
k-NN ($k = 3$)	57	17,8	-0,17	0,53	0,60
k-NN ($k = 5$)	8	-42,8	-0,31	0,84	0,86
RF 10	2	-17,8	-6,12	0,90	0,93
RF 100	2	-17,8	-6,12	0,89	0,92
NN 5	49	92,7	0,05	0,76	0,78
NN 10	7	25,9	-0,08	0,79	0,84
NN 25	28	95,2	0,06	0,89	0,89
NN 50	3	-20,7	-0,20	0,90	0,93
NN 100	29	56,3	-0,33	0,86	0,87



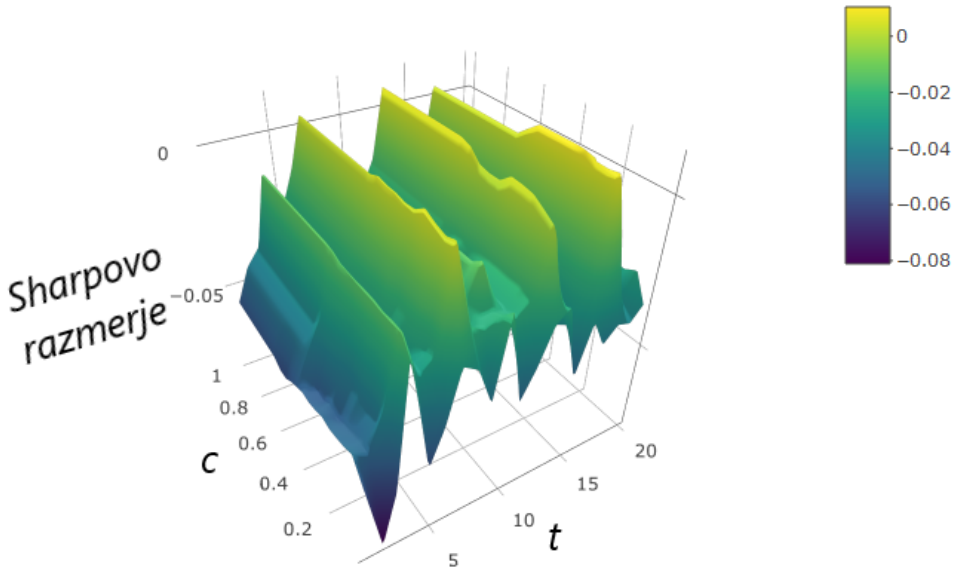
Slika 6.5: Graf prikazuje gibanje cene delnice BAC in strategijo, ki jo je naredila NN s 25 vozlišči v skritem nivoju. Pri točkah BUY bi po strategiji kupili, pri SELL prodali. Zelena črta prikazuje stanje sredstev, če bi na začetku vložili 1000 dolarjev.



Slika 6.6: Graf prikazuje gibanje cene delnice BAC in strategijo, ki jo je naredila NN s 50 vozlišči v skritem nivoju. Pri točkah BUY bi po strategiji kupili, pri SELL prodali. Zelena črta prikazuje stanje sredstev, če bi na začetku vložili 1000 dolarjev.



Slika 6.7: 3D graf prikazuje donosnost modela NN 25 na validacijskih podatkih pri različnih parametrih t in c. Legenda nam pove, da bolj kot je rumena barva, večja je donosnost. Temno modra pa pomeni, da je bila donosnost okoli -700.



Slika 6.8: 3D graf prikazuje vrednost Sharpovega razmerja modela RF 10 na validacijskih podatkih pri različnih parametrih t in c. Legenda nam pove, da bolj kot je rumena barva, večja je vrednost Sharpovega razmerja. Temno modra pa pomeni, da je bila vrednost majhna.

6.2.3 Analiza rezultatov

Pri trgovanju z BTC je le eden od modelov presegel donosnost s strategijo kupi na začetku in drži do konca. To je kar malce razočaranje, vendar pa moramo vedeti, da je cena kar precej narasla, zadnja cena v testnem obdobju pri BTC-USD je bila kar štirikrat večja kot prva. Kljub takim rezultatom pa bi vseeno raje priporočili strategijo kakšnega od modelov, kot pa strategijo kupi na začetku in drži do konca, saj slednja lahko predstavlja veliko tveganje. Če cena kar precej pade, nudi model vsaj malo verjetnosti, da stanje sredstev ne pade z njim, ampak pri strategiji kupi na začetku in drži temu ni tako.

Kot lahko vidimo, sta si tabeli 6.1 in 6.2 zelo podobni. Večina modelov je

Tabela 6.6: Leva tabela predstavlja parametre modelov iz tabele 6.4. Desna pa parametre modelov iz tabele 6.5.

Model	t	c
k-NN ($k = 3$)	2	0,55
k-NN ($k = 5$)	10	0,25
RF 10	16	1
RF 100	16	0,55
NN 5	6	0,25
NN 10	9	0,8
NN 25	16	0,45
NN 50	13	0,3
NN 100	13	0,4

Model	t	c
k-NN ($k = 3$)	2	0,55
k-NN ($k = 5$)	10	0,25
RF 10	16	1
RF 100	14	0,95
NN 5	6	0,25
NN 10	9	0,8
NN 25	16	0,45
NN 50	16	0,7
NN 100	13	0,4

najboljša z enakimi parametri tako pri donosnosti kot tudi pri tveganju. Iz tega lahko sklepamo, da je cena pri modelih z različnimi parametri približno enako nihala in je zato donosnost igrala pomembno vlogo tudi pri izračunu Sharpovega razmerja.

Zanimivo se rezultati donosnosti pri različnem številom dreves pri RF in različnem številom vozlišč pri NN kar precej razlikujejo. Zaradi tega predvidevamo, da že majhna sprememba v strategiji trgovanja lahko pomeni veliko razliko v donosnosti. Na primer, pri sliki 6.1 lahko vidimo, da je model prodal skoraj pri vrhu, kar je bila tudi najpomembnejša in najdonosnejša odločitev. Če se model ne bi odločil takrat prodati, bi bila donosnost dosti manjša, torej lahko tudi rečemo, da je imel model v tem primeru kar malce sreče. Dejstvo, ki potrjuje, da je imel model srečo in je bil najdonosnejši izmed vseh modelov pri testnih podatkih, je to, da je bil pri validacijskih podatkih šele nekje na sredini lestvice najdonosnejših.

Utežena mera F_1 in klasifikacijska natančnost (CA) sta zelo odvisni od izbire parametrov t in c. Večja kot sta parametra t in c, več je stanj *drži* v označeni strategiji in tudi v strategiji, ki jo določi model, to pomeni tudi, da je večja klasifikacijska točnost, kot tudi utežena mera F_1.

Iz slik 6.1, 6.2, 6.5, 6.6 lahko vidimo, da so se modeli nekaj naučili. Večina stanj *kupi* je postavljenih tako, da je prej cena nekaj časa padala, kar je bil prvi del cilja naše strategije. Drugi del strategije, ki temelji na povratku k srednji vrednosti, torej da po stanju *kupi* cena narašča, pa se velikokrat ni zgodil. To je tudi težji del, saj do uspešnosti pridemo le z napovedovanjem v prihodnost. Podobno lahko rečemo tudi za stanja *prodaj*.

V tabeli 6.7 so prikazani rezultati najboljših modelov na validacijski množici. Pri trgovanju z menjalnim tečajem BTC-USD je bil najdonosnejši model RF 100, največjo vrednost Sharpovega razmerja pa je imel model RF 10. Torej, če bi se v realnosti odločili za najboljši model glede donosnosti na BTC-USD in modelu na začetku dali 1000 USD, bi imeli pri testnih podatkih donosnost 1013,6 USD, če bi pa vzeli model, ki je imel največjo vrednost Sharpovega razmerja pa bi bila donosnost 1197,8 USD. Pri BAC pa je bil najboljši v obeh primerih RF 10. V tem primeru bi imeli v realnosti pri testnih podatkih izgubo -17,8 USD.

Kot najboljšo strategijo bi lahko izpostavili strategijo modela RF z 10 drevesi na podatkih menjalnega tečaja BTC-USD, ki je prikazan na sliki 6.2. Imela je največjo vrednost Sharpovega razmerja na validacijskih podatkih in bila je med najuspešnejšimi tudi na testnih. Prikazala je sposobnost donosnosti v naraščujočem in padajočem trendu, kar je bil tudi cilj te diplomske naloge.

Tabela 6.7: Rezultati donosnosti in vrednosti Sharpovega razmerja pri najboljših modelih na validacijski množici.

		Validacijska množica		**Testna množica**	
Trgovanje s podatki	**Model**	**donosnost**	**Sharpovo razmerje**	**donosnost**	**Sharpovo razmerje**
BTC	RF 100	7322,9	1,72	1013,6	1,04
	RF 10	4982,1	2,48	1197,8	1,23
BAC	RF 10	3567,5	1,52	-17,8	-6,13

Poglavje 7

Zaključek in nadaljnje delo

Na podlagi pridobljenih rezultatov lahko sklepamo, da je izdelava trgovalne strategije, ki je donosna v različnih trendih, zelo zahtevna. Kljub velikim številom trgovalnih indikatorjev se iz podatkov ni dalo dobro ločiti stanja *kupi, drži, prodaj* med seboj. Opazili smo tudi, da imajo strategije modelov v različnem časovnem obdobju tudi različno donosnost. Take strategije bi bilo dobro imeti le nekaj časa, nato pa najti drugo. Zato bi lahko bila izboljšava strategije takšna, da določen model napoveduje le v določenem obdobju, nato pa ga zamenja drug, za tisto obdobje bolj primeren model.

Za nadaljnje delo bi lahko spremenili veliko stvari. Preizkusili bi lahko še kakšne druge trgovalne indikatorje. V naši diplomski nalogi smo se omejili samo na tehnično analizo. Dodali bi lahko tudi fundamentalno ter sentimentalno analizo in tako pridobili še dodatne informacije glede stanja trga. Lahko bi preizkusili še druge modele strojnega učenja, kot na primer metodo podpornih vektorjev. Pri umetnih nevronskih mrežah bi lahko dodali več skritih nivojev. Za napovedovanje časovnih vrst (v to kategorijo spada tudi trgovanje) pa so primerne povratne nevronske mreže. Slednje pri napovedih znajo izkoriščati medsebojne odvisnosti zaporednih podatkov. Najboljše rezultate pri nalogah časovne obdelave pa dosega arhitektura nevronskih mrež z dolgim kratkoročnim spominom. Zanimivo bi bilo videti tudi rezultate modelov spodbujevanega učenja, ki jih v tej diplomi nismo izdelali. V delu [4]

so zapisali, da je globoko spodbujevalno učenje obetavno, da bi odkrili optimalne strategije za delovanje na časovnih vrstah. Vendar pa so v tem delu testirali modele na umetno generiranih podatkih, na katerih se rezultati lahko razlikujejo od testiranja modelov na realno pridobljenih podatkih.

Literatura

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Ernest P Chan. *Algorithmic Trading: Winning Strategies and Their Rationale*. John Wiley & Sons, Inc., 2013.

[3] Robert D Edwards, John Magee, and WH Charles Bassetti. *Technical analysis of stock trends*. CRC press, 2007.

[4] Xiang Gao. Deep reinforcement learning for time series: playing idealized trading games. *arXiv preprint arXiv:1803.03916*, 2018.

[5] Investopedia. Commodity. Dosegljivo: `https://www.investopedia.com/terms/c/commodity.asp`, 2018. [Dostopano: 17. 5. 2018].

[6] Investopedia. Cryptocurrency. Dosegljivo: `https://www.investopedia.com/terms/c/cryptocurrency.asp`, 2018. [Dostopano: 17. 5. 2018].

[7] Investopedia. Exchange-traded fund (etf). Dosegljivo: `https://www.investopedia.com/terms/e/etf.asp`, 2018. [Dostopano: 17. 5. 2018].

[8] Investopedia. Index. Dosegljivo: `https://www.investopedia.com/terms/i/index.asp`, 2018. [Dostopano: 17. 5. 2018].

[9] Investopedia. Sharpe ratio. Dosegljivo: `https://www.investopedia.com/terms/s/sharperatio.asp`, 2018. [Dostopano: 30. 7. 2018].

[10] Investopedia. Stock. Dosegljivo: https://www.investopedia.com/terms/s/stock.asp, 2018. [Dostopano: 17. 5. 2018].

[11] Investopedia. Technical analysis. Dosegljivo: https://www.investopedia.com/terms/t/technicalanalysis.asp, 2018. [Dostopano: 16. 5. 2018].

[12] Investopedia. Trading strategy. Dosegljivo: https://www.investopedia.com/terms/t/trading-strategy.asp, 2018. [Dostopano: 31. 5. 2018].

[13] Investopedia. What is 'forex - fx'. Dosegljivo: https://www.investopedia.com/terms/f/forex.asp, 2018. [Dostopano: 17. 5. 2018].

[14] Daniel Kahneman and Patrick Egan. *Thinking, fast and slow*, volume 1. Farrar, Straus and Giroux New York, 2011.

[15] Igor Kononenko and Marko Robnik Šikonja. *Inteligentni sistemi*. Založba FE in FRI, 2010.

[16] nanalyze. Can artificial intelligence be used for stock trading? Dosegljivo: https://www.nanalyze.com/2017/02/artificial-intelligence-stock-trading/, 2018. [Dostopano: 17. 5. 2018].

[17] Martin J Pring. *Technical analysis explained: The successful investor's guide to spotting investment trends and turning points*, volume 4. McGraw-Hill New-York, 1991.

[18] Marko Robnik-Šikonja and Petr Savicky. Package 'corelearn', 2018.

[19] William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.

[20] Michael Thomsett. Technical analysis of stock trends explained. an easy-to-understand system for trading successfully. *Ethan Hathaway Co LTD*, 2012.

[21] TradingView. Bitcoin / Dollar, Bitstamp. Dosegljivo: `https://www.tradingview.com/chart`, 2018. [Dostopano: 26. 6. 2018].

[22] tsweb. The stock market: from the 'ter buerse' in to wall street. Dosegljivo: `https://www.nbbmuseum.be/en/2010/01/stockmarket.htm`, 2018. [Dostopano: 16. 5. 2018].

[23] John W Tukey. *Exploratory data analysis*, volume 2. Reading, Mass., 1977.

Dodatek

Tabela 7.1: Rezultati donosnosti in vrednosti Sharpovega razmerja pri najboljših modelih, ki so predstavljeni v razdelku 6.2.1. Pod zadnjo horizontalno črto sta modela, ki sta imela največjo vrednost Sharpovega razmerja in imela različna parametra t in c od najdonosnejših modelov.

	Validacijska množica		**Testna množica**	
Model	**donosnost**	**Sharpovo razmerje**	**donosnost**	**Sharpovo razmerje**
k-NN ($k = 3$)	7176,8	2,1	648,8	0,82
k-NN ($k = 5$)	2493,8	1,26	-164,7	0,05
RF 10	5142,1	1,97	-304,7	-0,53
RF 100	7322,9	1,72	1013,6	1,04
NN 5	6821,8	1,91	1440,6	1,15
NN 10	7301,5	2,11	688,6	1,06
NN 25	4030,5	1,87	4101,0	1,04
NN 50	4266,9	2,08	204,06	0,37
NN 100	6439,2	2,05	329,18	0,58
k-NN ($k = 5$)	2349,6	1,53	-351,53	-0,17
RF 10	4982,1	2,48	1197,8	1,23

Tabela 7.2: Rezultati donosnosti in vrednosti Sharpovega razmerja pri najboljših modelih, ki so predstavljeni v razdelku 6.2.2. Pod zadnjo horizontalno črto je model, ki je imel pri največji vrednosti Sharpovega razmerja različna parametra t in c od najdonosnejšega parametra.

	Validacijska množica		**Testna množica**	
Model	**donosnost**	**Sharpovo razmerje**	**donosnost**	**Sharpovo razmerje**
k-NN ($k = 3$)	418,0	0,19	17,9	-0,17
k-NN ($k = 5$)	351,2	0,19	-42,8	-0,31
RF 10	3567,5	1,52	-17,8	-6,13
RF 100	1087,7	0,43	-110,5	-0,57
NN 5	956,2	0,38	92,7	0,05
NN 10	190,0	0,16	25,9	-0,08
NN 25	341,9	0,18	95,2	0,06
NN 50	216,4	0,14	-234,8	-1,02
NN 100	146,4	0,11	-56,3	-0,33
RF 100	869,2	0,57	-17,83	-6,12