**VERJETNOST PREKOMERNEGA PRILEGANJA**

**Motivacija:** Ena od glavnih težav, s katero se srečujemo pri testiranju kvantitativnih trgovalnih strategij, je nevarnost prekomernega prileganja (overfitting) modela na izbrani učni množici. Izvajanje strategije na testnih množicah predstavlja le delno rešitev problema.

**Moja naloga**: za dani nabor trgovalnih strategij izračunaj verjetnost prekomernega prileganja z novo metodo kombinatoričnega simetričnega navzkrižnega preverjanja.

**Opis dela:**

**1. podatki:**

Najprej sem pridobila podatke za delnice, ki so trenutno v indeksu S&P 500.

S&P 500 (oz. Standard & Poor's 500) je borzni indeks, ki temelji na tržni kapitalizaciji 500 velikih podjetji, katerih navadne delnice kotirajo na NYSE (Newyorška borza) ali NASDAQ (National Association of Securities Dealers Automated Quotations).

Za vseh 500 delnic sem vzela zadnjo ceno (Close vrednost). Podatke sem pridobila iz spletne strani yahoo finance s pomočjo funkcije v R. vzeča sem dnevne podatke za čas od 1. januarja 2000 do 1. novembra 2013 (približno 13 let).

Tiste delnice, ki ne obstajajo dovolj dolgo, sem izločila. Izločila sem tudi tiste delnice, ki imajo več kot tri manjkajoče vrednosti skupaj. Ostale manjkajoče vrednosti sem nadomestila z vrednostjo zadnje cene prejšnjega trgovalnega dne.

**2. strategije:**

Nato sem sestavila nekaj enostavnih trgovalnih strategij:

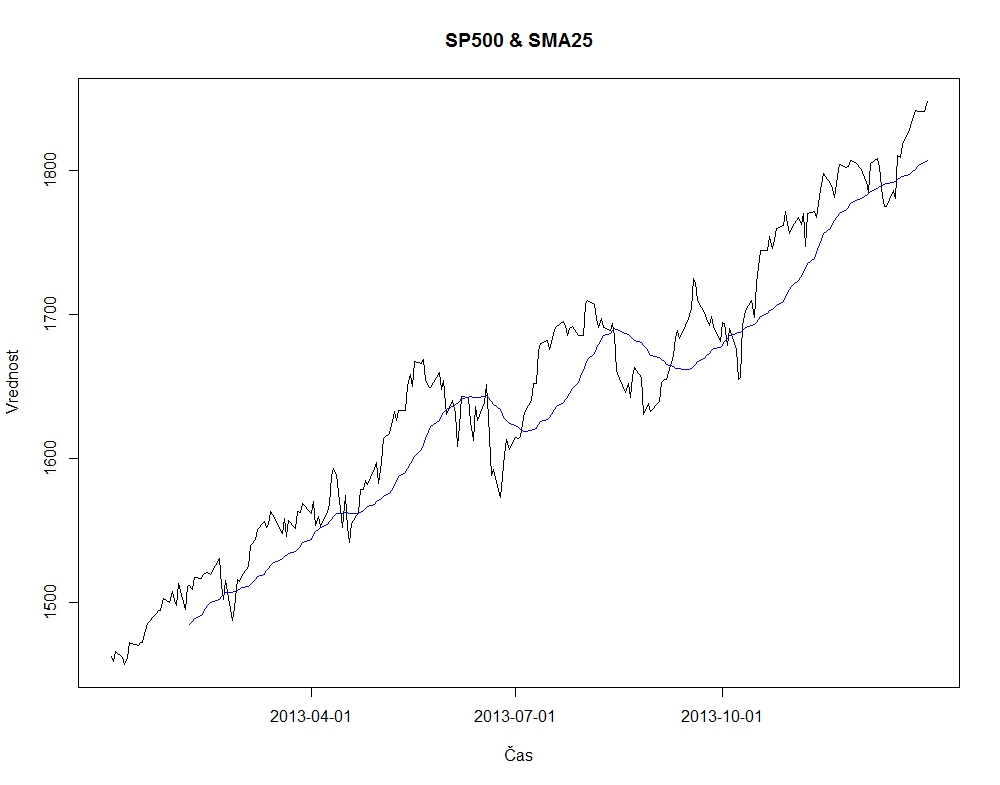
- Random: naključno sem izbrala, ali vstopim v trgovalno pozicijo ali ne

- Buy&Hold: delnico na začetku kupim in jo imam do konca

Ostale trgovalne strategije sem sestavila na podlagi indikatorjev. Uporabila sem indikatorje SMA, RSI in Bollinger.

- SMA (simple moving average) oz. preprosto drseče povprečje je aritmetična sredina gibanja cene delnice. Izračuna se tako, da se seštejejo vse vrednosti v željenem obdobju, nato pa delimo z dolžino obdobja. Jaz sem uporabila dolžine 5, 25, 50 in 150 dni. Običajno je nakupni znak, ko zadnja cena zraste čez krivuljo drsečega povprečja, prodajni pa, ko pade pod njo. Jaz sem uporabila samo en signal: Vstopila sem v pozicijo, ko je bila zadnja cena nad SMA.

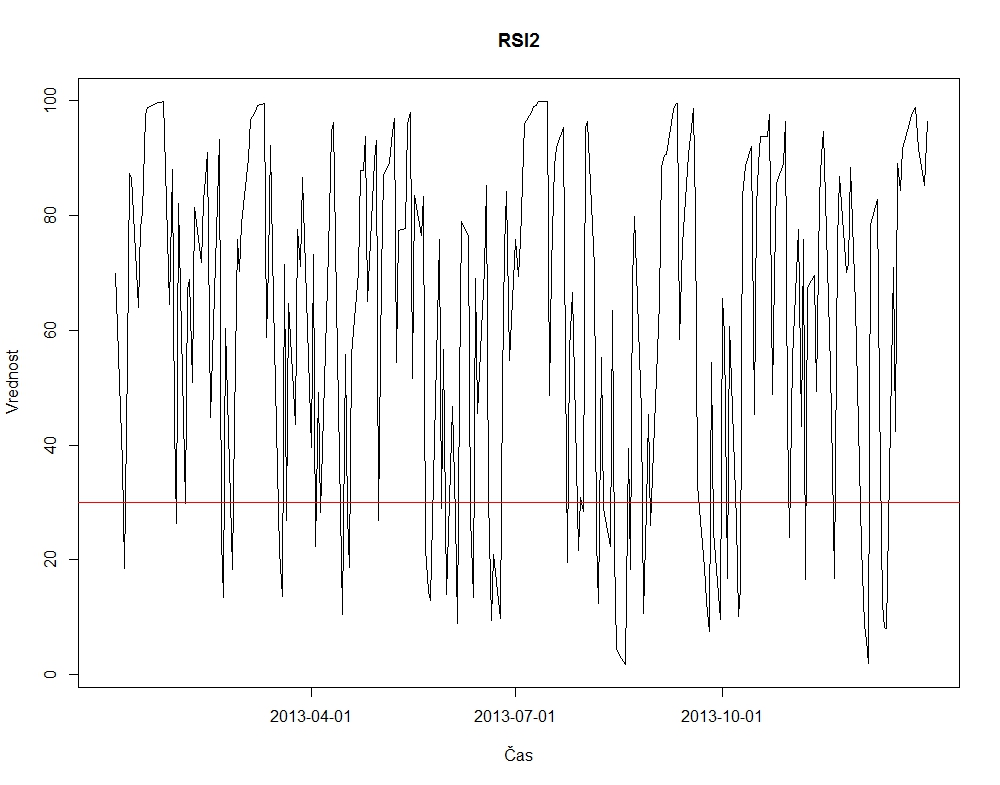
Primer zadnje cene delnice indeksa SP500 v zadnjem letu in indikatorja SMA:



- RSI (Relative Strength Indeks) oz. indeks relativne moči primerja relativno moč vsote dobičkov v dneh preučevanega obdobja, ko je cena naraščala (ko je bila končna cena višja od končne cene prejšnjega dne) v primerjavi z vsoto izgub v dneh proučevanega obdobja, ko se je cena nižala ( v dneh, ko je bila končna cena nižja od končne cene prejšnjega dne). Indikator primerja velikosti rasti z velikostmi padcev vrednosti papirja. Indeks se giblje med vrednostma 0 in 100. Če vrednost indeksa pade pod 30, je to znak za podcenjenost vrednostnega papirja in bo verjetno sledil porast cene (znak za nakup vrednostnega papirja), če pa zraste nad 70, je to znak za precenjenost vrednostnega papirja in sledil naj bi padec cene (znak za prodajo vrednostnega papirja). Jaz sem uporabila samo znak za nakup: vstopila sem v pozicijo, če je RSI padel pod 30. Uporabila sem RSI na obdobju 2 in 14 dni.

Formula za izračun RSI: RSI = 100 – 100/(1-RS), kjer je RS povprečna vrednost pozitivnih period v opazovanem obdobju deljeno s povprečno vrednostjo negativnih period. Pozitivne in negativne periode so izračunane kot absolutna sprememba tečaja od zaključnega do zaključnega tečaja.

Primer RSI2 za indeks SP500 v zadnjem letu in meja 30:



- Bollingerjevi trakovi ustvarjajo trak cenovnega razpona. Sestavljeni so iz treh krivulj: srednja krivulja predstavlja srednjeročni trend. Običajno je to preprosto drseče povprečje (SMA). Zgornja krivulja je za faktor krat standardni odklon premaknjena navzgor, spodnja krivulja pa je za faktor krat standardni odklon premaknjena navzdol. Običajno je faktor 2 ali 1 ali 0.5. za izračun SMA in standardnega odklona se običajno uporablja enako časovno obdobje.

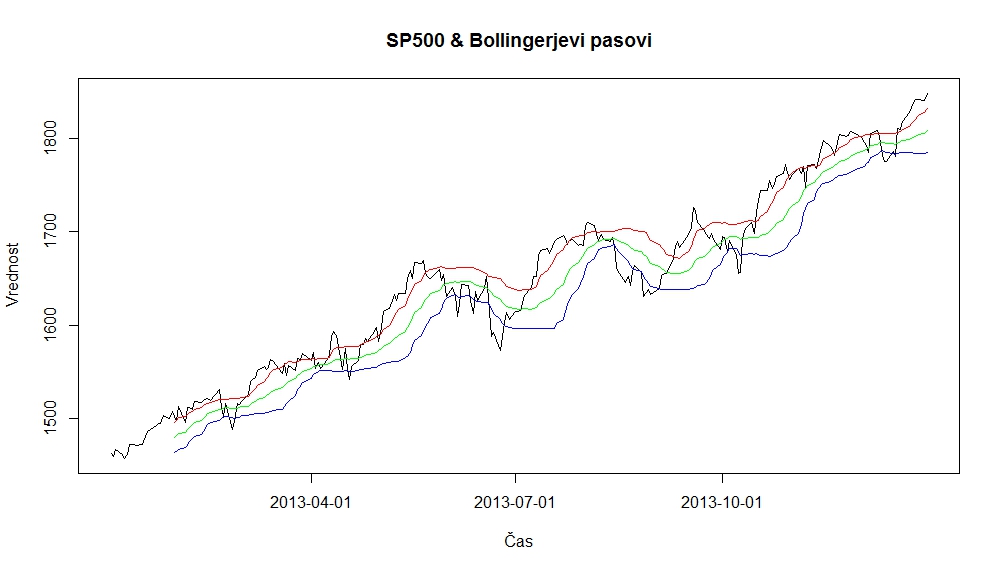
Srednja = SMA20

Zgornja = SMA20 + 1 std

Spodnja = SMA20 – 1 std

Jaz sem za izračun SMA in standardnega odklona uporabila obdobje 20 dni. V času visoke razpršenosti cen se bollingerjevi trakovi razširijo v času nizke razpršenosti cen pa zožijo. Če so cene stalno blizu zgornje meje bollingerjevega traku, pomeni, da je vrednostni papir precenjen in to pomeni signal za prodajo. Če pa so cene stalno blizu spodnje meje traku, to pomeni, da je vrednostni papir podcenjen, kar je signal za nakup. Jaz sem uporabila samo nakupni signal: ko je vrednost delnice padla pod spodnji bollingerjev pas, sem vstopila v pozicijo.

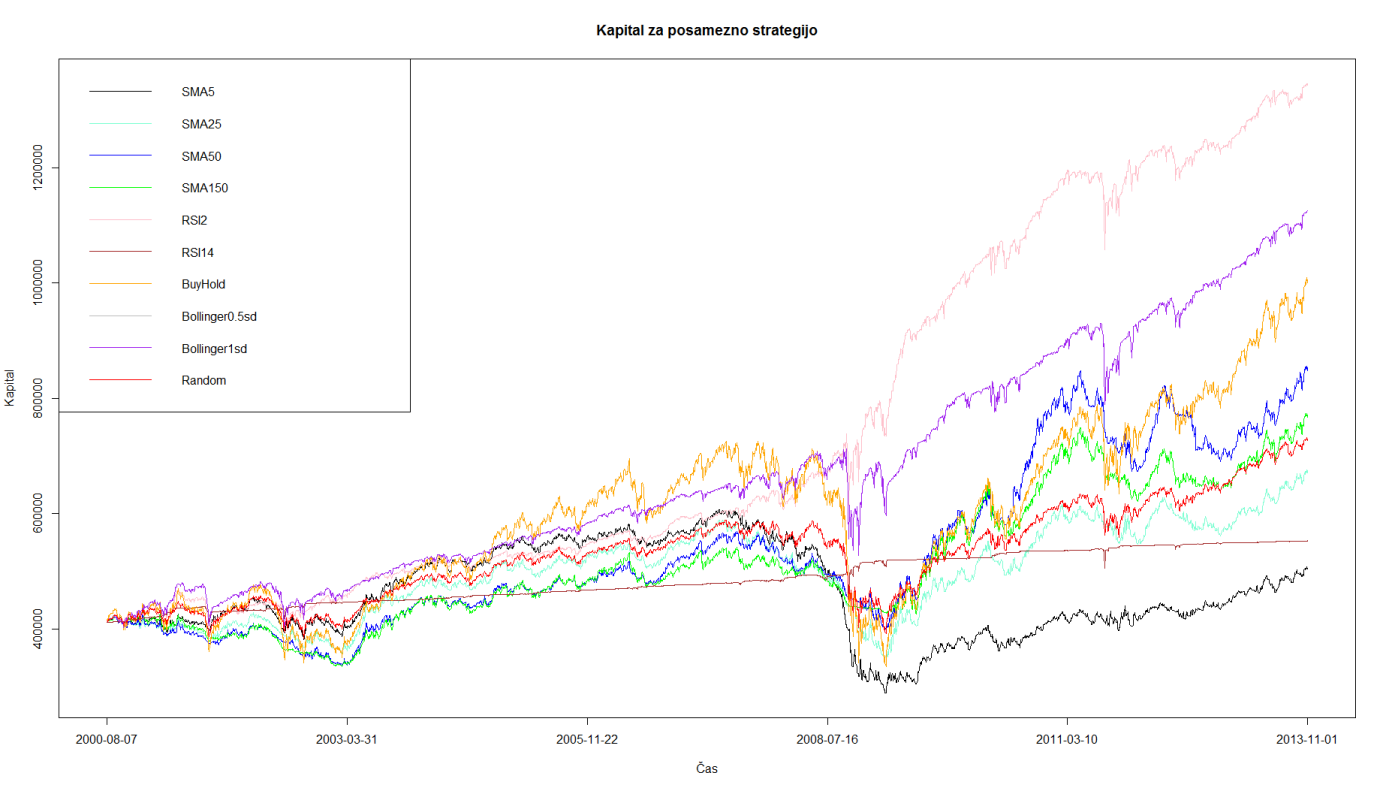
Primer bollingejevega pasu za indeks SP500 v zadnjem letu. SMA20 in 1 standardni odklon.



**3. uporaba strategij:**

Na vseh 500 (minus nekaj izločenih) delnicah sem uporabila predstavljene strategije. Vsaki delnici v portfelju sem na začetku namenila 1000 eurov kapitala. Ta kapital sem potem uporabila za trgovanje s to delnico. Če je bil znak za nakup sem kupila eno delnico, sicer nisem storila ničesar. Kapital celotnega portfelja sem izračunala kot vsoto kapitalov za posamezno delnico v portfelju.

Slika prikazuje kapital portfelja za posamezno strategijo tekom 13. let:



Kot je razvidno iz grafa je najboljša strategija za dani portfelj na koncu danega obdobja strategija RSI14. Sledi ji Boillinger, ki temelji na SMA20 in SD 1 oz. 0.5. na tretjem mestu je pasivna strategija Buy&Hold, kjer na začetku vložimo 1000 eurov v vsako delnico v portfelju in nato ne počnemo nič.

Kot najslabša se je izkazala strategija SMA5. Najmanj volatilna je strategija RSI14 (saj komaj kdaj vstopimo v pozicijo zaradi visokih vrednosti indikatorja RSI14), vendar ne ravno donosna.

Opazimo očiten padec vseh strategij v koncu leta 2008, zaradi začetka finančne krize, kjer se tudi pomeša vrstni red uspešnosti strategij.

**4. metoda kombinatoričnega simetričnega preverjanja:**

Kakšna je verjetnost, da se optimalna strategija prekomerno prilega, je overfittana? Da imamo prekomerno prileganje mora biti, da je strategija, ki je najboljša na učni množici, zelo slaba na testni množici. Razlog za to je, da je strategija, ki je optimalna na učni množici preveč povezana z učno množico, tako da optimizacija postane škodljiva.

Bolj formalno, definiramo verjetnostni prostor:

* Naj bo R: RT -> R funkcija, ki za dano vrsto uspešnosti n-te strategije (velikosti T) vrne realno mero uspešnosti Rn. Na primer, R je lahko funkcija Sharpe ratio, ki jo uporabimo na vektorju dobička in izgub mt za posamezno strategijo.
* SHARPE RATIO JE ……
* Za dano množico N strategij naj bo Omega vzorčni prostor (sample space) vseh možnih rangiranj strategij glede na Rn (v naraščajočem redu). Omega vsebuje izide, ki predstavljajo N! permutacij rangiranj.
* Naj bo F množica dogodkov, ki vsebuje sigma algebro Omege. Natančneje, eden izmed elementov je podmnožica vseh izidov v Omega kjer je n-ta strategija rangirana v prvi polovici (pod mediano). Temu dogotku rečemo Fn element F, za vsak n = 1,…,N.
* Naj bo Prob verjetnostna mera na F. Na primer, verjetnost, da je uspešnost n-te strategije pod mediano vseh uspešnosti, ustreza relativni frekvenci pri kateri se zgodi katerikoli izid v Fn.

Glede na verjetnostni prostor (Omega, F, Prob) lahko definiramo prekomerno prileganje zgodovinskega preverjanja (backtest overfitting) v povezavi z izbiro investicijskih strategij.

DEFINICIJA (backtest overfitting): Naj bo n\* strategija z najboljšo uspešnostjo na učni množici, to je Rn\* >= Rn, za vsak n = 1,…,N. označimo z bar(Rn\*) uspešnost strategije n\* na testni množici. Naj bo Me[bar(R)] mediana uspešnosti vseh strategij na testni množici. Potem rečemo, da se proces izbire strategije prekomerno prilega, če za strategijo n\* z najvišjim rangom na učni množici velja: E[bar(Rn\*)] < Me[bar(R)]

DEFINICIJA (verjetnost prekomernega prileganja): Naj bo n\* strategija z optimalno uspešnostjo na učni množici. Ker n\* ni nujno najboljša strategija na testni množici, obstaja ne ničelna verjetnost, da je bar(Rn\*) < Me[bar(R)]. Verjetnost, da se izbrana strategija n\* prekomerno prilega, definiramo kot: PBO == Prob[bar(Rn\* < Me[bar(R)]]

Z drugimi besedami, rečemo, da se proces izbire strategije prekomerno prilega, če je pričakovana uspešnost strategij izbranih na učni množici manjša kot mediana uspešnosti vseh strategij na testni množici. Učna množica je podmnožica opazovanj mt, na katerih smo izbrali optimalno strategijo n\* izmed N alternativ.

POSTOPEK:

1. Konstruiramo matriko M iz vrste uspešnosti N poskusov. Bolj natančno, vsak stolpec n = 1,…, N predstavlja vektor dobičkov in izgub na t=1,…., T opažanjih posamezne konfiguracije modela. M je realna matrika velikosti TxN. Edina pogoja sta, da je M prava matrika (ima enako število vrstic za vsak stolpec, kjer so opažanja sinhrona za vsako vrsti za vseh N poskusov) in da metriko uspešnosti, s katero določimo optimalno strategijo, lahko uporabimo na podvzorcu vsakega stolpca.
2. M razdelimo po vrsticah na S disjunktnih podmatrik enakih velikosti. Vsaka od teh matrik Ms, s=1,…,S, je velikosti T/S x N.
3. Tvorimo kombinacije Cs velikosti S/2 podmatrik Ms, kar nam da S nad S/2 kombinacij. Vsaka kombinacija c element Cs je sestavljena iz S/2 podmatrik Ms
4. Za vsako kombinacijo c element Cs naredimo naslednje:
   1. Konstruiramo učno množico J tako da združimo S/2 podmatrik Ms, ki tvorijo c. j je matrika reda T/S S/2 x N = T/2 x N
   2. Konstruiramo testno množico bar(J) kot komplement J v M. z drugimi besedami bar(J) je matrika velikosti T/2 x N iz vseh vrstic matrike M, ki niso v J.
   3. Konstruiramo vektor R statistike uspešnosti reda N, kjer n-ti element vektorja R pove uspešnost n-tega stolpca matrike J (učne množice)
   4. Določimo element n\* tako, da je Rn <= Rn\* za vsak n = 1,…,N. z drugimi besedami n\* = argmax\_n{Rn}
   5. Konstruiramo vektor bar(R) statistik uspešnosi reda N, kjer n-ti element vektorja bar(R) predstavlja uspešnost n-tega stolpca matrike bar(J) (testne množice)
   6. Določimo relativni rank bar(Rn\*) v bar(R). označimo ga z bar(omega\_c), kjer je bar(omega\_c) element (0,1). To je relativni rank uspešnosti na testni množici povezan z poskusom izbranim na učni množici. Če postopek optimizacije strategij ni everfittan, bi moralo biti, da bar(Rn\*) sistematično prekaša bar(R), tako kot Rn\* prekaša R.
   7. Definiramo logit lambda\_c = log(bar(omega\_c)/(1-bar(omega\_c)). To predstavlja lastnost, da je lambda\_c = 0, ko je bar(Rn\*) enak mediani bar(R). visoke vrednosti logita pomenijo konsistenco med uspešnostji na učni in testni množici, kar kaže na nizko raven prekomernega prileganja backtesta
5. Izračunamo porazdelitev rankov testne množice z zbiranjem vseh lambda\_c, za c element Cs. f(lambda) je potem relativna frekvenca pri kateri se lambda zgodi na vseh Cs, z integralom = 1.