Smanjenje generalizacijske pogreške klasifikatora kroz augmentaciju trendova u financijskim vremenskim nizovima

Skender, Sven

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 721

SMANJENJE GENERALIZACIJSKE POGREŠKE KLASIFIKATORA KROZ AUGMENTACIJU TRENDOVA U FINANCIJSKIM VREMENSKIM NIZOVIMA

Sven Skender

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 721

SMANJENJE GENERALIZACIJSKE POGREŠKE KLASIFIKATORA KROZ AUGMENTACIJU TRENDOVA U FINANCIJSKIM VREMENSKIM NIZOVIMA

Sven Skender

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 30. rujna 2024.

DIPLOMSKI ZADATAK br. 721

Pristupnik:	**Sven Skender (0036514458)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	prof. dr. sc. Zvonko Kostanjčar

Zadatak: **Smanjenje generalizacijske pogreške klasifikatora kroz augmentaciju trendova u financijskim vremenskim nizovima**

Opis zadatka:

Financijski vremenski nizovi inherentno su zašumljeni i volatilni, što otežava razlikovanje pravih trendova od kratkoročnih oscilacija. Kako bi se to riješilo, razvijeni su različiti algoritmi za označavanje trendova, od kojih je svaki parametriziran i generira različite trendove ovisno o parametrima. U svrhu poboljšanja generalizacije klasifikatora prilikom predviđanja budućih trendova, skup za treniranje potrebno je augmentirati korištenjem oznaka dobivenih iz različitih postavki parametara spomenutih algoritama za označavanje. Ovaj pristup bi trebao olakšati treniranje i povećati robusnost modela. Predloženu metodu potrebno je evaluirati na povijesnim financijskim vremenskim nizovima različitih financijskih imovina.

Rok za predaju rada: 14. veljače 2025.

Sadržaj

1. Uvod

Vrijednosti financijskih vremenskih nizova, poput cijena dionica, valutnih tečajeva i drugih financijskih instrumenata često osciliraju jer su pod utjecajem raznih faktora kao što su tržišni sentiment, ekonomske vijesti, geopolitički događaji, ponašanja investitora, itd. Upravo zbog toga otežano je razlikovanje pravih trendova od kratkoročnih oscilacija. Financijski vremenski nizovi posjeduju specifična svojstva poput niskog omjera signala i šuma te nestacionarnosti zbog čega i s razlogom predstavljaju jedan od zahtjevnijih problema u analizi podataka.

Pravovremeno prepoznavanje dugoročnih trendova može pozitivno utjecati na upravljanje pozicijama određene financijske imovine gdje za vrijeme rastućeg trenda želimo otvoriti poziciju u toj imovini, dok za vrijeme padajućeg trenda želimo zatvoriti tu poziciju. U svrhu rješavanja problema identifikacije dugoročnih trendova predloženo je nekoliko različitih algoritama za definiranje trenda. S obzirom na to da je svaki od tih algoritama za definiranje trenda parametriziran, ovisno o odabiru algoritma i njegovih parametara, kao izlaz mogu se dobiti značajno različito identificirani trendovi. Sukladno tome, postavlja se i pitanje kako te definicije trendova utječu na uspješnost klasifikatora treniranih na oznakama trenda dobivenim na izlazu algoritma za označavanje trenda. Algoritmi za označavanje trenda na ulazu uzimaju cijene financijskih vremenskih nizova i uz odgovarajuće parametre algoritma na izlaz daju niz binarnih oznaka koje identificiraju određeni trend. Pozitivna oznaka trenda, oznaka 1, ukazuje na rastući trend kretanja tržišta i sugerira kupnju pozicije u imovini. Negativna oznaka 0 identificira padajući trend i optimalno vrijeme za zatvaranje pozicije u imovini na tržištima gdje short-selling nije opcija. Problemi toga su što za različite algoritme za označavanje vremenskih nizova za isti vremenski niz dobivaju se različite oznake modela što značajno može utjecati na uspješnost modela predviđanja te njihovi različiti izlazi mogu sugerirati na različite trendove i strategije za trgovanjem pozicijama. Također, tako dobivene oznake

od algoritama za označavanje trendova mogu se koristiti za učenje klasifikatora, kojima je cilj predvidjeti budući smjer kretanja trenda pojedine financijske imovine. Dosadašnji radovi fokusirali su se prvenstveno na kvalitetnu izradu značajki i dobru implementaciju klasifikatora, zanemarujući fokus na označavanju trendova u svrhu što boljeg predviđanja tržišta i nisu uključivali financijske mjere uspješnosti pri evaluaciji modela. Tako na primjer autori u radu [1] predložili su pristup temeljen na SVM-u za predviđanje tržišnih trendova, no njihovo rješenje bazira se na arbitrarnoj definiciji trenda te izostanku ključnih financijskih metrika pri evaluaciji modela. Najčešće se odabire samo jedan algoritam i arbitrarno odabrani parametar, te se klasifikatori treniraju na tako dobivenim oznakama.

Do sada, niti jedan rad nije kombinirao više različitih definicija trenda, kako bi obogatio skup za učenje (augmentirao oznake) i na taj način napravio klasifikatore robusnijim, odnosno smanjio njihovu generalizacijsku pogrešku. U svrhu poboljšanja generalizacije klasifikatora prilikom predviđanja budućih trendova, u ovom radu bavimo se augmentacijom početnog skupa za treniranje korištenjem oznaka dobivenih iz različitih algoritama za označavanje. Sukladno tome, predložena metoda evaluirana je zatim na povijesnim financijskim vremenskim nizovima različitih financijskih imovina.

2. Strojno učenje za vremenske nizove

2.1. Analiza financijskih vremenskih nizova

Svaka imovina ima svoju knjigu ograničenih naloga (eng. Limit Order Book, LOB), koja se sastoji od ograničenih naloga i sadrži sve potrebne informacije za trgovanje, odražavajući namjeru trgovca da kupi ili proda određenu imovinu. Ograničeni nalog definiran je sljedećim komponentama:

- **Strana** (eng. bid/ask) – označava kupovne (eng. bid) ili prodajne (eng. ask) naloge,
- **Količina** – broj jedinica imovine,
- **Ograničena cijena** – unaprijed određena cijena po kojoj se nudi nalog,
- **Vrijeme unosa** – trenutak kada je nalog poslan u knjigu.

Kupovni nalozi (eng. bid) predstavljaju interes za kupnju određene imovine po određenoj cijeni, dok prodajni nalozi (eng. ask) označavaju ponude za prodaju. Kada trgovac unese tržišni nalog (eng. market order), navodeći stranu (kupnja/prodaja) i količinu, nalog se trenutačno izvršava spajanjem s odgovarajućim ograničenim nalozima po najboljoj dostupnoj cijeni, čime dolazi do sklapanja trgovine i formiranja tržišne cijene.

Postoje tri glavne vrste naloga. Tržišni nalozi (eng. market orders) izvršavaju se odmah po najboljoj dostupnoj cijeni. Ograničeni nalozi (eng. limit orders) su za kupnju ili prodaju po određenoj maksimalnoj (za kupnju) ili minimalnoj (za prodaju) cijeni. Nalog za otkazivanje (eng. cancel order) uklanja prethodno uneseni limitirani nalog [2].

Za sažeto prikazivanje tržišnih dinamika, knjiga naloga često se uzorkuje i agregira u OHLCV podatke (Open, High, Low, Close, Volume), koji pružaju pregled kretanja cijena

unutar definiranih vremenskih intervala trajanja Δt. Kraj svakog intervala označen je diskretnim vremenskim korakom t.

- **Open (O)** – prva zabilježena cijena u intervalu (p_t^O),
- **High (H)** – najviša postignuta cijena (p_t^H) tijekom uzorkovanog perioda,
- **Low (L)** – najniža zabilježena cijena (p_t^L),
- **Close (C)** – posljednja zabilježena cijena u intervalu (p_t^C),
- **Volume (V)** – ukupna količina imovine kojom se trgovalo unutar tog vremenskog okvira.

U ovom radu korišten je $\Delta t = 1$ minuta jer takva granularnost najbolje odgovara analizi unutar dnevnog (intraday) trgovanja.

2.2. Strojno učenje i klasifikacija

Strojno učenje jedno je od najbrže rastućih područja u računalnoj znanosti. Razlog tome nije samo kontinuirani rast količine podataka, već i napredak teorije koja omogućuje bolju obradu podataka i njihovog pretvaranja u korisno znanje. U raznim znanstvenim disciplinama, od astronomije do biologije, ali i u svakodnevnom životu, tehnologija sve dublje je integrirana u našu svakodnevicu, dok se naš digitalni otisak neprestano širi, generirajući i prikupljajući sve veće količine podataka.

Bez obzira na to radi li se o znanstvenim ili osobnim podacima, ako oni ostanu neiskorišteni, nemaju nikakvu vrijednost. Zbog toga, pojedinci i organizacije razvijaju nove načine za njihovu obradu i njihovog pretvaranja u korisne proizvode i usluge. U tom procesu, strojno učenje igra sve veću ulogu, postajući ključni alat u izvlačenju vrijednosti i korisnih informacija iz podataka. Generalno, strojno učenje predstavlja programiranje računala kako bi optimiziralo određeni kriterij uspješnosti koristeći primjere iz podataka ili prošlih iskustava. Model je definiran do određenih parametara, a proces učenja sastoji se od izvršavanja programa s ciljem optimizacije tih parametara koristeći određeni skup podataka za treniranje. Model može biti prediktivan, s ciljem predviđanja budućih događaja, deskriptivan, s ciljem stjecanja uvida iz podataka, ili kombinacija oba

pristupa [3]. Napretkom i razvojem umjetne inteligencije i strojnog učenja došlo je do sve veće upotrebe modela strojnog učenja koji su se i pokazali kao najbolji za predviđanje financijskih tržišta [4]. Metode predviđanja vremenskih nizova korištene u strojnom učenju uglavnom uključuju regresiju i klasifikaciju. Regresija za predviđanje koristi apsolutne vrijednosti rasta i pada cijena. Tijekom procesa treniranja modela nije potrebno označavati podatke, dok klasifikacija često zahtijeva označavanje podataka za treniranje klasifikacijskih modela [5]. Za razliku od nekih drugih primjena modela strojnog učenja, gdje je zavisna varijabla odabrana od strane stručnjaka i nedvosmislena, u financijskim vremenskim nizovima može biti definirana na različite načine, a označavanje se provodi algoritamski preko algoritama za označavanje. Oni služe za identifikaciju trendova s obzirom na predefinirane parametre. U kontekstu financijskih nizova algoritmi se koriste za prepoznavanje i označavanje trendova kako bi se olakšala interpretacija podataka ili pripremili podatci za modele strojnog učenja. Označavanje oznaka na opisani način motiviralo je razvoj klasifikacijskih algoritama za predviđanje financijskih trendova u financijskim vremenskim nizovima s ciljem maksimiziranja preferirane mjere financijskog učinka investitora [6]. U ovom radu, za mjeru financijskoga učinka korišten je kumulativni povrat.

2.3. Postupak odabira i izračuna značajki

OHLCV (Open, High, Low, Close, Volume) podatci pružaju samo trenutnu sliku kretanja tržišta. Kako bi model imao nove informacije dodane su nove značajke. Za računanje novih značajki korištene su postojeći, poznati tehnički indikatori koji pružaju dodatnu informaciju o obrascima i trendovima tržišta koje promatramo. Na taj način olakšana je separacija primjera u prostoru značajki, što dovodi do smanjenja generalizacijske pogreške kada se model primjenjuje na prethodno neviđenim podatcima. Također, nove značajke pomažu u filtriranju kratkoročnih oscilacija tržišta i daju bolji uvid u stvarne trendove koje želimo izolirati.

Za izradu značajki korišteni su sljedeći tehnički indikatori:

- **1. Simple Moving Average (SMA)**
  Računa se kao srednja vrijednost povrata cijena kroz rolling window od N minuta

i koristi se za izglađivanje cijena kroz određeni period vremena [7] [8]:

$$\mathrm{SMA}_N = \frac{1}{N} \sum_{i=0}^{N-1} R_i, \tag{2.1}$$

gdje je R_i vrijednost dnevnog povrata cijene u i-tom trenutku, a N je broj minuta u kliznom prozoru.

- **2. Moving Volatility (MV)**

  Pokazuje koliko se cijena mijenjala kroz određeni period vremena [9]:

$$MV_N = \sqrt{\frac{1}{N} \sum_{i=0}^{N-1} (R_i - \bar{R})^2}, \tag{2.2}$$

  gdje je $\bar{R}$ srednja vrijednost dnevnog povrata kroz određeno razdoblje.

- **3. Return On Last N Minutes (RETURN)**

  Mjeri kratkoročni trend i brzinu pomaka cijene u posljednjih N minuta:

$$R_T = \frac{P_T - P_{T-N}}{P_{T-N}} \times 100, \tag{2.3}$$

  gdje je P_T cijena u trenutku T, a P_{T-N} cijena N minuta prije.

- **4. Exponential Moving Average (EMA)**

  Isto kao i kod Moving Average-a računa se kao srednja vrijednost povrata cijena kroz rolling window od N minuta s promjenom da je veća težnja stavljena na zadnji povrat, stoga je i vrijednost EMA bliža pravoj vrijednosti na kojoj se i mjerila [7] [10]:

$$EMA = \alpha R_T + (1 - \alpha) EMA_{T-1}, \tag{2.4}$$

  gdje je $\alpha = \frac{2}{N+1}$ i R_T vrijednost dnevnog povrata u trenutku T.

- **5. High-Low Difference (HIGH_LOW)**

  Razlika između visoke i niske cijene tijekom određenog perioda i daje ideju totalne promjene cijene za taj period:

$$HIGH\_LOW = P_{HIGH} - P_{LOW}, \tag{2.5}$$

gdje je P_{HIGH} najviša cijena, a P_{LOW} najniža cijena tijekom razdoblja.

- **6. Open-Close Difference (OPEN_CLOSE)**
  Razlika između prve i zadnje cijene daje smjer trenda. Ako je razlika negativna, cijena je počela rasti, inače pada:

$$OPEN\_CLOSE = P_{OPEN} - P_{CLOSE}. \tag{2.6}$$

  gdje je P_{OPEN} prva cijena, a P_{CLOSE} zadnja cijena tijekom razdoblja.

- **7. Moving Average Convergence/Divergence (MACD)**
  MACD je indikator momenta koji prati trend i pokazuje odnos između dva *moving average-a* cijena. Otkriva promjene u snazi, smjeru, momentumu i trajanju trenda u cijeni dionice. Pretvara dva indikatora koji prate trend, *moving average-e*, u oscilator momentuma oduzimanjem dulje *EMA* od kraće *EMA*. Kad MACD linija presijeca signalnu liniju s donje strane, to označava bullish period, a kad MACD linija presijeca signalnu liniju s gornje strane, to označava bearish period [7] [11]:

$$MACD = EMA_{12} - EMA_{26}, \tag{2.7}$$

  gdje EMA_{12} je kraći 12-periodni EMA, a EMA_{26} je dulji 26-periodni EMA.

  Signalna linija je eksponencijalni pokretni prosjek MACD linije, računat s periodom N_{signal} gdje je $N_{\text{signal}} = 9$:

$$\text{Signal}_t = EMA_{\text{signal}}(MACD_t). \tag{2.8}$$

- **8. Average Directional Index (ADX)**
  Povećanje potencijala za profit može se postići trgovanjem u smjeru trenda. ADX se koristi za kvantificiranje trendova i mjeri snagu trenda. Vrijednost ADX-a izračunava se pomoću pokretnih srednjih vrijednosti. ADX indikator ima vrijednost između 0 i 100. Općenito se smatra da je vrijednost ADX-a iznad 25 znak jakog trenda. ADX indikator nam govori snazi trenda, ali ne i smjeru trenda[7] [12]:

$$ADX = \frac{\sum_{i=0}^{N} | + DI_i - -DI_i|}{N}, \tag{2.9}$$

gdje je $+DI_i$ i $-DI_i$ pozitivni i negativni smjer trenda u trenutku i.

- **9. Relative Strength Index (RSI)**

  RSI je indikator momenta i mjeri brzinu i promjene u kretanju cijena. RSI vrijednost leži između 0 i 100. Vrijednost iznad 70 označava da je imovina prekupljena, a vrijednost ispod 30 označava da je imovina preprodana [7] [13]:

$$RSI = 100 - \frac{100}{1 + RS}, \tag{2.10}$$

  gdje je RS (Relative Strength) omjer prosječnih dobitaka i prosječnih gubitaka i izračunat kao:

$$RS = \frac{\text{Average Gain}}{\text{Average Loss}}, \tag{2.11}$$

  i gdje prvo izračunavamo dnevnu promjenu cijene ΔP:

$$\Delta P = P_{\text{today}} - P_{\text{yesterday}}, \tag{2.12}$$

  a potom izračunavamo prosječne dobitke i prosječne gubitke:

$$\text{Average Gain} = \frac{1}{N} \sum_{i=1}^{N} \text{Gain}_i, \tag{2.13}$$

$$\text{Average Loss} = \frac{1}{N} \sum_{i=1}^{N} \text{Loss}_i. \tag{2.14}$$

  Na kraju, koristimo RS za izračunavanje RSI-ja.

- **10. On Balance Volume (OBV)**

  OBV je indikator momentuma koji koristi promjene u trgovačkom volumenu kao pokazatelj budućeg kretanja cijena imovine. Formulacija OBV-a temelji se na te-

oriji da volumen prethodi promjeni cijena. Cilj praćenja OBV-a je predvidjeti vjerojatni smjer promjena cijena u bliskoj budućnosti. Rast OBV-a odražava pozitivni pritisak volumena koji može dovesti do viših cijena, dok opadajući OBV predviđa pad cijena [14] [15]:

$$OBV = OBV_{t-1} + V_t \quad \text{ako je cijena rastuća,} \tag{2.15}$$

$$OBV = OBV_{t-1} - V_t \quad \text{ako je cijena opadajuća,} \tag{2.16}$$

i gdje je V_t volumen u trenutnom vremenu t.

- **11. Momentum Index (MTM)**
  MTM mjeri brzinu kojom se trend ubrzava ili usporava. To čini usporedbom najnovije, zadnje cijene sa zadnjom cijenom iz određenog prošlog razdoblja [16] :

$$MTM = P_T - P_{T-N}, \tag{2.17}$$

  gdje je P_T cijena u trenutnom vremenu T, a P_{T-N} cijena prije N minuta.

- **12. Average True Range (ATR)**
  ATR je tehnički indikator analize koji razlaže cijeli raspon cijene imovine tijekom određenog vremena kako bi odredio tržišnu volatilnost. Indikator se najčešće koristi za kvantificiranje volatilnosti uzrokovane razmacima i ograničenim usponima ili padovima, a ne za sugeriranje smjera cijene. Dionica s visokim nivoom volatilnosti ima viši ATR, dok dionica s niskim nivoom volatilnosti ima niži ATR [7] [17]:

$$ATR = \frac{1}{N}\sum_{i=1}^{N} TR_i, \tag{2.18}$$

  gdje je:

$$TR = \max\left(P_{high} - P_{low}, |P_{high} - P_{close\_prev}|, |P_{low} - P_{close\_prev}|\right), \tag{2.19}$$

  i gdje su: P_{high} najviša cijena tijekom dana, P_{low} je najniža cijena tijekom dana,

P_{close_prev} je zadnja cijena prethodnog dana, a N je broj razdoblja.

- **13. Money Flow Index (MFI)**

  Indeks novčanog toka (MFI) je tehnički oscilator koji koristi podatke o cijeni i volumenu za prepoznavanje signala prekupljenosti ili preprodanosti određene imovine. Također se može koristiti za uočavanje divergencija koje upozoravaju na promjenu trenda cijene. Oscilator se kreće između 0 i 100 [18] [19]:

$$MFI = 100 - \frac{100}{1 + \frac{\sum_{i=1}^{N}(PM_i \times V_i)}{\sum_{i=1}^{N}(NM_i \times V_i)}}, \tag{2.20}$$

  gdje je PM_i pozitivni novčani tok, a NM_i negativni novčani tok.

- **14. Mean-Deviation-Rate (MDR)**

  Mean-Deviation-Rate (MDR) je mjera koja se koristi za analizu cijena na temelju odstupanja od srednje vrijednosti unutar određenog vremenskog razdoblja. U kontekstu financijskih podataka, MDR se koristi za mjerenje varijacije cijena u odnosu na srednju cijenu unutar određenog vremenskog okvira [20]:

$$f_{ij} = \frac{x_{ij} - M_{\lambda i}}{M_{\lambda i}}, \quad x_{ij} \in X, \tag{2.21}$$

$$M_{\lambda s} = \frac{\sum_{i=s}^{s+\lambda-1} x_i}{\lambda}, \quad x_i \in x, \quad s = 1, 2, \ldots, N - \lambda + 1, \tag{2.22}$$

  gdje: x_{ij} predstavlja zadnju cijenu u matrici podataka X, $M_{\lambda s}$ označava srednju vrijednost zadnjih cijena unutar kliznog prozora veličine λ, f_{ij} je rezultat skaliranja zadnje cijene x_{ij} s obzirom na srednju vrijednost unutar kliznog prozora, λ predstavlja dimenziju kliznog prozora, odnosno broj uzastopnih podataka koji se koriste za izračunavanje srednje vrijednosti, dok suma cijena x_i u formuli za $M_{\lambda s}$ uključuje cijene unutar prozora duljine λ.

- **15. Change Of High Price in last N minutes (HIGH-CHANGE)**

Pokazuje promjenu visoke cijene u zadnjih N minuta:

$$HIGH - CHANGE = P_{HIGH,T} - P_{HIGH,T-N}, \tag{2.23}$$

gdje je $P_{HIGH,T}$ najviša cijena u trenutnom vremenu, a $P_{HIGH,T-N}$ najviša cijena prije N minuta.

- **16. Change Of Volume in last N minutes (VOLUME-CHANGE)**
  Pokazuje promjenu volumena trgovanja u zadnjih N minuta:

$$VOLUME - CHANGE = V_T - V_{T-N}, \tag{2.24}$$

gdje je V_T volumen u trenutnom vremenu, a V_{T-N} volumen prije N minuta.

2.4. Postupak izračuna oznaka

Razmotrimo povijesni niz cijena imovine p_t , gdje $t = 1, 2, 3, \dots, T$, a p_t je cijena imovine u trenutku t. Za svaki vremenski trenutak t, gdje $t \leq T - H + 1$, oznaka y_t može se izračunati kao funkcija budućih cijena imovine u trenutku t s unaprijed gledajućim prozorom veličine H i parametrima θ:

$$y_t = g(p_t, p_{t+1}, p_{t+2}, \dots, p_{t+H-1}; \theta), \tag{2.25}$$

gdje funkcija g preslikava ulazni niz u skup $\{0, 1\}$ i služi za definiranje oznaka trenda. Oznaka 1 označava pozitivni trend (rastući trend), dok oznaka 0 označava negativni trend (padajući trend).

Za treniranje algoritma h za klasifikaciju trendova, u svakom vremenskom trenutku t, gdje $t \geq L$, potreban je odgovarajući vektorski prikaz značajki $x_t \in X$ kao funkcija prethodnih cijena imovine do trenutka t, s prozorom gledanja unatrag veličine L i parametrima φ:

$$x_t = f(p_{t-1}, p_{t-2}, p_{t-3}, \dots, p_{t-L}; \varphi), \tag{2.26}$$

gdje $f : \mathbb{R}_+^L \times \mathbb{R}^M \to \mathbb{R}^N$.

Te oznake i značajke čine skup podataka za treniranje:

$$D = \{(x_t, y_t)\}_{t=L}^{T-H+1}, \tag{2.27}$$

koji se zatim koristi za treniranje algoritma $h : X \to \{0, 1\}$.

Za proizvoljnu funkciju označavanja g i seriju cijena p, imamo seriju trendova:

$$y = [y_1, y_2, y_3, \dots, y_{T-H+1}] \tag{2.28}$$

gdje je $y_t \in \{0, 1\}$. Promjena labele iz 0 u 1 ukazuje na promjenu cijene i definira rastući trend. Takav tip trenda sugerira na kupnju i držanje pozicije u imovini, dok padajući trend (promjena labela iz 1 u 0) sugerira zatvaranje pozicije. Kumulativni povrat od S segmenata rastućeg trenda definiran je kao:

$$R = \prod_{i=1}^{S}(1 + r_i) - 1 \tag{2.29}$$

gdje je r_i povrat pojedinog segmenta rastućeg trenda izračunat prema formuli:

$$r_i = \frac{p_{t+\Delta t} - p_t \cdot (1 + \phi)}{p_t \cdot (1 + \phi)} \tag{2.30}$$

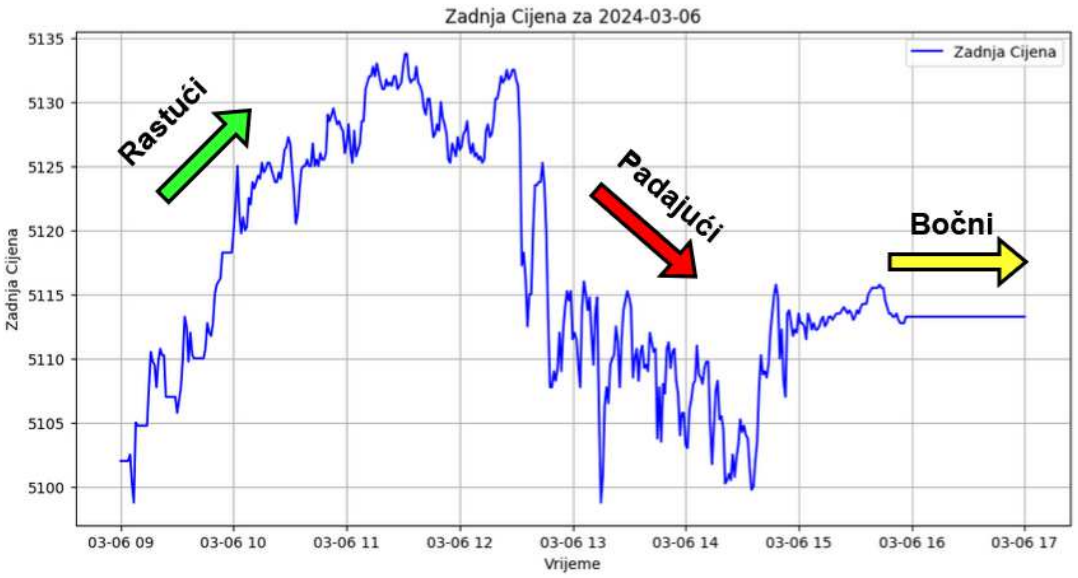
pri čemu ϕ definira transakcijsku naknadu za otvaranje pozicije u imovini. Vremenska oznaka t označava početak rastućeg trenda, dok Δt označava njegovo trajanje [6].

3. Algoritmi za označavanje trenda u financijskim vremenskim nizovima

3.1. Definicija trenda

Generalno, za trend možemo reći da je pokazatelj smjera tržišta. Dugoročni trend na financijskom tržištu može se promatrati kao pravac koji predstavlja dominantni smjer kretanja cijena tijekom dužeg vremenskog razdoblja. Međutim, na taj osnovni pravac superponiran je kratkoročni šum, koji proizlazi iz nasumičnih fluktuacija cijena uzrokovanih tržišnim sentimentom, vijestima, ponašanjem ulagača i drugim čimbenicima.

Ovaj šum manifestira se kroz veliki broj lokalnih minimuma i maksimuma, što znači da se unutar dugoročnog trenda mogu pojaviti česte i nagle promjene smjera cijene koje ne narušavaju njegov dugoročni smjer. Ovisno o tome kreće li se tržišna cijena prema gore, dolje ili bočno, smjer tog kretanja govori nam o trendu tržišta [21]. Kad pričamo o trendovima razlikujemo nekoliko vrsta trendova. Rastući trend definiran je kao kontinuirani niz rastućih vrijednosti cijene neke imovine, dok padajući trend prepoznaje se po kontinuiranom padu cijene. Niz vrijednosti cijene koji ni nakon nekog duljeg perioda nema pomaka gore ili dolje zove se stagnirajući, odnosno bočni trend cijene [22]. Navedene vrste trendova, vizualno se mogu vidjeti na slici ispod:



Slika 3.1. Grafički prikaz tri različita trenda: rastući, padajući i bočni

Osim o vrsti trenda, može se govoriti i o intenzitetu, odnosno stupnju trenda. Naime, percepcija trenda, kako netko označava određeni trend kao rastući ili padajući ovisi o perspektivi investitora koji ga analizira. Primjerice, dugoročni investitor veliku važnost neće pridavati kratkotrajnim promjenama cijena tijekom nekoliko dana ili tjedana, dok će za dnevnog investitora čak i nekoliko uzastopnih dana rasta predstavljati značajan rastući trend koji zahtijeva pravovremenu reakciju i odluku otvaranja ili zatvaranja pozicije u imovini.

Zbog toga važno je razumjeti o čijem interesu i strategiji se radi i pravovremeno je prepoznati i reagirati. Kako bi zadovoljili potrebu za opažanjem različitih stupnjeva trendova uvedeni su i različiti algoritmi za označavanje. Svaki algoritam na svoj način i u ovisnosti o odabiru određenih parametara definira i drukčije trendove zbog čega model s više algoritama za označavanje ima i veću šansu generalizacije i smanjenja prenaučenosti na samo određeni stupanj trenda.

U svrhu razlikovanja pravih trendova od kratkoročnih oscilacija u ovom radu korišteni su i različiti algoritmi za označavanje trendova od kojih je svaki drukčije parametriziran i generira različite trendove ovisno o početnim parametrima. U nastavku je svaki od tih algoritama detaljnije objašnjen.

3.2. Oracle algoritam identifikacije trenda

Algoritam za označavanje financijskih vremenskih nizova Oracle čini bazni algoritam na temelju kojih se može provjeravati izvedba ostalih algoritama za učenje. Algoritam identificira optimalne oznake y^* unutar vremenske serije cijena koje donose teoretski maksimalne kumulativne povrate, uzimajući u obzir troškove provizije povezane s iniciranjem novih pozicija [6]. Ovu optimizaciju može se opisati na sljedeći način:

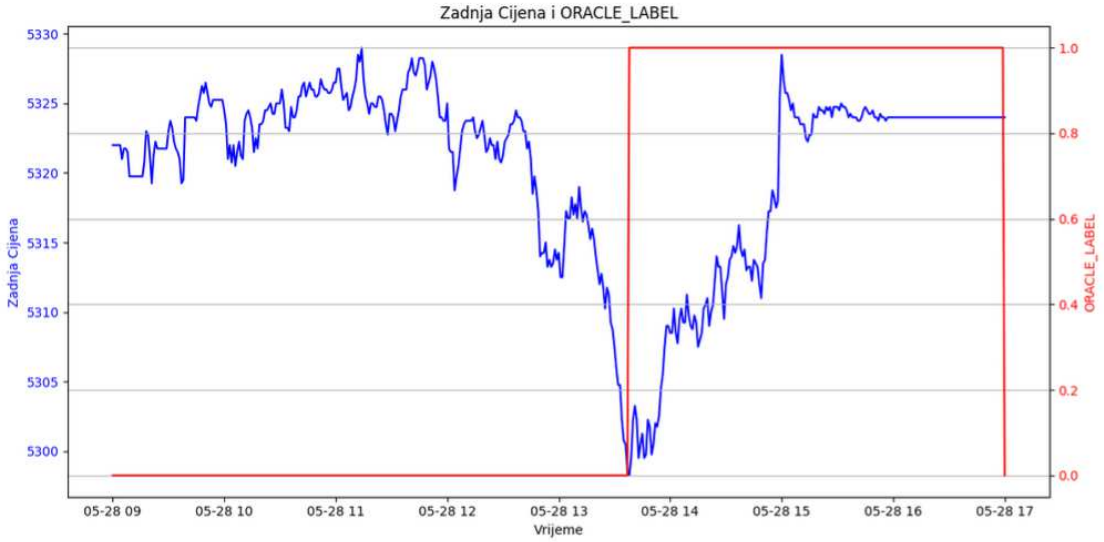
$$y^* = \arg\max_y c(y, \varphi) \tag{3.1}$$

uz uvjet:

$$\varphi = \vartheta, \quad y_T = i, \quad \text{gdje je} \quad i \in \{0, 1\}. \tag{3.2}$$

Treba napomenuti da je transakcijska naknada φ definirana od strane burze i koristi se prilikom izračuna svih kumulativnih povrata u ovom radu, dok je ϑ parametar Oracle algoritma za označavanje.

Primjer oznaka trenda dobivenih pomoću Oracle označavanja prikazan je na slici ispod:



Slika 3.2. Oznake trenda dobivene Oracle algoritmom

3.3. Fixed Time Horizon (FTH) algoritam identifikacije trenda

Za cjenovni niz p_t, H-korak unaprijed povrat (*engl. H-step-ahead return*) u unaprijed gledanom prozoru veličine H definira se kao:

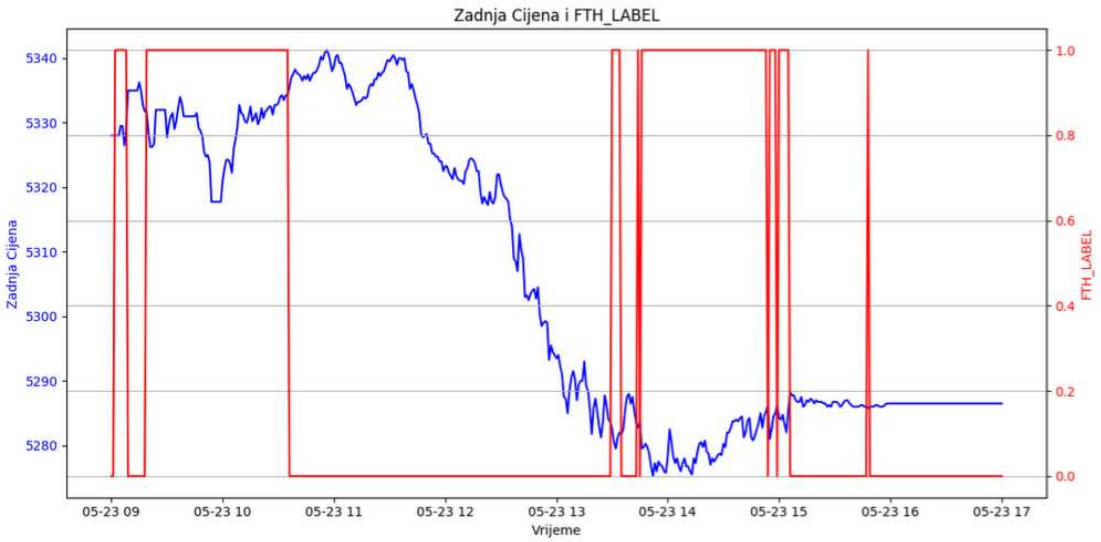
$$r_t = \frac{p_{t+H} - p_t}{p_t} \tag{3.3}$$

Za svaki vremenski trenutak $t < T-H$, dodijeljena je binarna oznaka y_t uspoređujući H-korak unaprijed povrat r_t s unaprijed definiranim pragom τ:

$$y_t = \begin{cases} 1 & \text{ako } r_t \geq \tau \\ 0 & \text{inače} \end{cases} \tag{3.4}$$

Označavanje s fiksnim vremenskim horizontom (FTH) široko se koristi u literaturi financijskih vremenskih nizova zbog svoje očigledne jednostavnosti. S obzirom na to da u izračun uzima samo vrijednost početne cijene p_t i vrijednost završne cijene p_{t+H} ignorira bilo kakvu dinamiku cijene unutar tog vremenskog perioda. Također, može rezultirati pozicijama s negativnim povratima (što nije poželjno za treniranje algoritma klasifikacije) te često dovodi do jako neuravnoteženih skupova podataka [6] [23].

Primjer oznaka trenda dobivenih pomoću FTH označavanja s optimalnim parametrima za $H = 9$ i $\tau = 0.0001$ prikazan je na slici ispod:



Slika 3.3. Oznake trenda dobivene FTH algoritmom

3.4. Algoritam za identifikaciju kontinuiranog trenda (CTL)

Kada se tržište razvija uz kontinuirani trend, može se podijeliti na rastuće i padajuće tržište. Ulagači bi u rastućem tržištu trebali kupiti i zadržati imovinu, dok bi u tržištu s mogućnošću kratke prodaje trebali zauzeti kratku poziciju. Ako kratka prodaja nije moguća, ulagači bi trebali prodati imovinu u padajućem tržištu. Njihova se pozicija ne bi trebala mijenjati sve dok se ne pojavi signal da bi se tržišni trend mogao promijeniti.

Kako bi razlikovali kontinuirane trendove, prvo je potrebno odrediti najviše i najniže točke povijesnih podataka unutar određenog vremenskog razdoblja te ih pohraniti u vektore h i l, gdje t označava broj najviših točaka, a m broj najnižih točaka. Za izračun stupnja trenda vremenskih serija koristi se TD indeks, koji pokazuje razinu kontinuiranog kretanja između susjednih najviših i najnižih točaka, a izračun se temelji na sljedećim formulama:

$$h = \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_{t-1} \\ h_t \end{bmatrix}, \quad l = \begin{bmatrix} l_1 \\ l_2 \\ \vdots \\ l_{m-1} \\ l_m \end{bmatrix} \tag{3.5}$$

,

$$TD(h_i, l_{i-1}) = \left| \frac{h_i - l_{i-1}}{l_{i-1}} \right|, \quad i > 1 \tag{3.6}$$

$$TD(l_i, h_{i-1}) = \left| \frac{l_i - h_{i-1} - 1}{h_{i-1}} \right|, \quad i > 1 \tag{3.7}$$

Nadalje, usporedbom parametra fluktuacije ω s indeksom vrijednosti TD, kontinuirani trend definiran je kao fluktuacijska amplituda između dva uzastopna vrha i dna koja premašuje zadani prag parametra ω, dok se u suprotnom fluktuacija smatra normalnom fluktuacijom bez kontinuiranog trenda. Kao osnova za izračun odabrane su najnovije najniže i najviše cijene, a rast tržišta iznad ili pad tržišta ispod ω parametra definiran je kao kontinuirani rastući trend ili kontinuirani opadajući trend. Zatim su svi podaci označeni s 1 u razdoblju rastućeg trenda, a 0 u razdoblju silaznog trenda za treniranje modela [5].

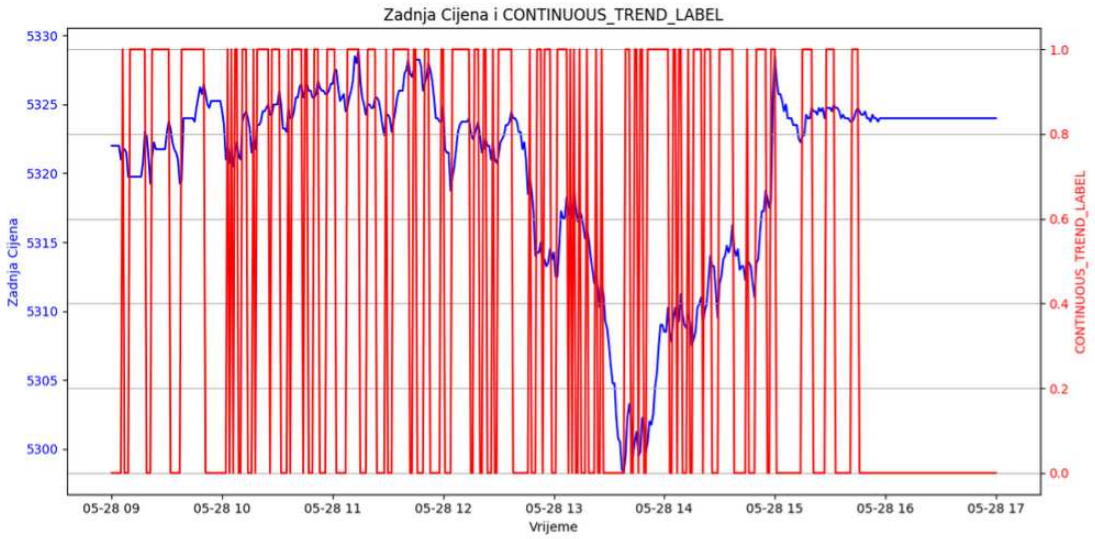
Algoritam možemo vidjeti u nastavku:

Algorithm 1 Continuous Trend Označavanje

Input: p, ω
Output: y
$FP \leftarrow p_1, p_H \leftarrow p_1, HT \leftarrow 1, p_L \leftarrow p_1, LT \leftarrow 1,$
$Cid \leftarrow 0, FP_N \leftarrow 0, N \leftarrow \text{len}(p)$
for $i = 1$ **to** N **do**
  if $p_i > FP + p_1 \cdot \omega$ **then**
    Set $[p_H, HT, FP_N, Cid] \leftarrow [p_i, t_i, i, 1]$
  else if $p_i < FP - p_1 \cdot \omega$ **then**
    Set $[p_L, LT, FP_N, Cid] \leftarrow [p_i, t_i, i, 0]$
  end if
end for
for $i = FP_N + 1$ **to** N **do**
  if $Cid > 0$ **then**
    if $p_i > p_H$ **then**
      Set $[p_H, HT] \leftarrow [p_i, t_i]$
    else if $p_i < p_H - p_H \cdot \omega$ and $LT \leq HT$ **then**
      for $j = 1$ **to** N **do**
        if $t_j > LT$ and $t_j \leq HT$ **then**
          Set $y_j \leftarrow 1$
        end if
      end for
      Set $[p_L, LT, Cid] \leftarrow [p_i, t_i, 0]$
    end if
  else
    if $p_i < p_L$ **then**
      Set $[p_L, LT] \leftarrow [p_i, t_i]$
    else if $p_i > p_L + p_L \cdot \omega$ and $HT \leq LT$ **then**
      for $j = 1$ **to** N **do**
        if $t_j > HT$ and $t_j \leq LT$ **then**
          Set $y_j \leftarrow 0$
        end if
      end for
      Set $[p_H, HT, Cid] \leftarrow [p_i, t_i, 1]$
    end if
  end if
end for

Slika 3.4. Primjer algoritma za identifikaciju kontinuiranog trenda

Primjer oznaka trenda dobivenih pomoću algoritma za identifikaciju kontinuiranog trenda s optimalnim parametrom $\omega = 0.0001$ može se vidjeti na slici ispod:



Slika 3.5. Oznake trenda dobivene algoritmom za kontinuirano označavanje trenda

4. Metodologija

4.1. Programsko okruženje i korištene programske biblioteke

Za potrebe ovog diplomskog rada, programski kod izvodio se na Google Colabu [24], platformi koja omogućuje pokretanje Python koda u "cloud" okruženju bez potrebe za prijašnjom instalacijom bilo kakvog softverskog okruženja. Google Colab nudi besplatnu dostupnost GPU-a i TPU-a što je korisno za rad s većim skupovima podataka i zahtjevnijim algoritmima, kao algoritmima dubokog učenja i analize vremenskih i financijskih nizova.

Za analizu, manipulaciju i pregled podataka, kao i za implementaciju modela korištene su neke od popularnih Python biblioteka: NumPy za numeričke izračune i rad s nizovima [25], Pandas za manipulaciju i analizu tabličnih podataka [26], te Matplotlib za vizualizaciju tih istih podataka [27]. Za izvođenje tehničke analize na financijskim vremenskim nizovima korištena je Pandas TA biblioteka [28]. Ova biblioteka omogućuje jednostavno izračunavanje financijskih pokazatelja koji su korišteni za treniranje modela kao što su pokretni prosjek, ADX (Average Directional Index), RSI (Relative Strength Index), MACD (Moving Average Convergence Divergence), i mnogi drugi koji će bit navedeni kasnije. Za implementaciju modela korišten je model XGBClassifier iz biblioteke XGBoost. Navedena biblioteka omogućuje primjenu naprednih metoda za klasifikaciju financijskih vremenskih nizova [29]. Kasnije, za evaluciju tog istog modela korišten je još i sklearn [30] koji ima mogućnost izračuna različitih metrika kao što su točnost, preciznost, recall-a, F1-score, te AUC (Area Under the Curve).

4.2. Podatci i predobrada podataka

Podatci koji su korišteni za ovaj rad petogodišnje su minutne OHLCV vrijednosti za 3 različite vrste futuresa: E-mini S&P 500 futures, gold futures i crude oil futures. Općenito, futures su financijski ugovori između dvije strane u kojima se dogovaraju o kupnji ili prodaji određene robe ili financijskog instrumenta po unaprijed dogovorenoj cijeni na određeni datum u budućnosti [31].
Podatci o futuresima spremljeni su na Google Drive zbog čega je za početak rada bilo potrebno povezati Google Drive i Google Colab i navedene podatke spremiti u odgovarajući folder. Nakon toga provedena je iteriracija po svim csv tablicama podataka te spajanje i spremanje po odgovarajućim kategorijama futuresa, svaki u svoj DataFrame. Implementacija cijeloga koda iz ovog rada može se vidjeti na ovoj poveznici [32].

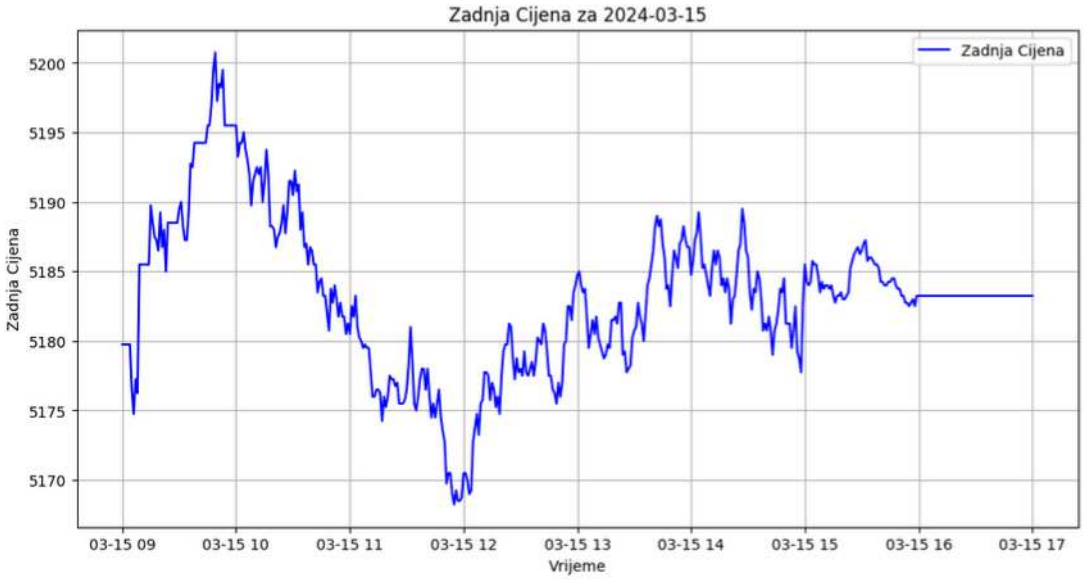
Algoritmi za označavanje također su učitani na sličan način i implementirani sa svojim Python funkcijama za označavanje trendova. Korišteni su intraday podatci tako da su svi podatci očišćeni i ostavljeni su minutni podatci samo u periodu od 9.00 h - 17.00 h. Definirana je i funkcija check_missing_minutes koja je služila za provjeru nedostajućih minuta u danu. Funkcija radi tako što definira i popunjava raspon sa svim mogućim minutama u danu od 9.00 h do 17.00 h. Taj raspon potom uspoređuje s rasponom minuta u našem skupu i kao razliku ta dva skupa izbacuje nedostajuće minute, ako postoje. Pomoću te funkcije potvrđeno je da nema nedostajućih minuta u našem početnome skupu.

Primjer kako izgledaju podatci za E-mini S&P 500 futuresa:

Tablica 4.1. Vrijednosti E-mini S&P 500 futuresa

Vrijeme	Open	High	Low	Close	Volume
2023-10-23 09:00:00	4242.00	4242.25	4240.75	4241.25	5528.0
2023-10-23 09:01:00	4241.50	4243.50	4239.50	4240.50	5527.0
2023-10-23 09:02:00	4240.50	4240.75	4238.00	4239.50	6540.0
2023-10-23 09:03:00	4239.25	4240.00	4238.25	4239.50	5296.0
2023-10-23 09:04:00	4239.50	4241.75	4239.00	4241.25	3788.0
⋮	⋮	⋮	⋮	⋮	⋮
2024-06-02 16:56:00	5299.25	5299.25	5299.25	5299.25	0.0
2024-06-02 16:57:00	5299.25	5299.25	5299.25	5299.25	0.0
2024-06-02 16:58:00	5299.25	5299.25	5299.25	5299.25	0.0
2024-06-02 16:59:00	5299.25	5299.25	5299.25	5299.25	0.0
2024-06-02 17:00:00	5299.50	5304.25	5299.25	5302.75	1766.0

A ovako izgledaju podatci grafički za jedan dan:



Slika 4.1. Graf zadnje vrijednosti cijene za E-mini S&P 500 futures

Prilikom računanja raznih tehničkih pokazatelja poput pokretnog prosjeka, RSI i MACD došlo je do pojave Nan - not a number vrijednosti koje je trebalo interpolirati. Metode koje su korištene su ffil() i bfill(). Metoda ffill() popunjava nedostajuću vrijednost zadnjom važećom vrijednošću koja se pojavila prije te nedostajuće vrijednosti u nizu podataka. Ako su vrijednosti niza poput "[1, 2, np.nan, 7]", korištenjem metode ffill() dobiva se niz "[1, 2, 2, 7]". Za razliku od ffill(), metoda bfill() nedostajuću vrijednost popunjava s prvom važećom vrijednošću nakon nedostajuće. Na primjer, ako je niz definiran kao "[1, 2, np.nan, 7]", korištenjem metode ffill() dobiva se niz "[1, 2, 7, 7]" [33]. Navedene metode su iz biblioteke Pandas [34] [35].

4.3. Odabir značajki

Nove značajke računate su programski pomoću već spomenute Pandas TA biblioteke za računanje tehničkih indikatora. Tako na primjer imamo indikatore poput:

1. Simple Moving Average (MA)
2. Moving Volatility
3. Return On Last N Minutes
4. Exponential Moving Average (EMA)

5. High-Low Difference
6. Open-Close Difference
7. Moving Average Convergence/Divergence (MACD)
8. Average Directional Index (ADX)
9. Relative Strength Index (RSI)
10. On Balance Volume (OBV)
11. Momentum Index (MTM)
12. Average True Range (ATR)
13. Money Flow Index (MFI)
14. Mean-Deviation-Rate (MDR)
15. Change Of High Price in last N minutes
16. Change Of Volume in last N minutes

Dodatno, kreiranje tehničkih značajki implementirano je s dodatnim parametrom N koji služi za odabir perioda ili prozora kroz koji se izračunava određeni indikator i značajka. Tako kod izračuna vektora značajki iterira se kroz listu proizvoljnih vrijednosti N i s njome se računaju značajke koje ulaze u konačan vektor značajki.

Vektor značajki nakon tog procesa sadrži 66 značajki, ovdje zbog nemogućnosti prikaza vidi se samo početnih 7.

Vrijeme	Open	High	Low	Close	Volume	MA_5	EMA_5
2023-10-26 09:14:00	4180.75	4181.00	4178.75	4180.75	2915.0	-0.000006	-0.000092
2023-10-26 09:15:00	4180.75	4182.75	4179.25	4182.25	4707.0	-0.000030	-0.000010
2023-10-26 09:16:00	4182.25	4183.25	4180.00	4180.00	4496.0	-0.000036	-0.000106
2023-10-26 09:17:00	4180.00	4182.25	4179.25	4179.25	4240.0	-0.000072	-0.000119
2023-10-26 09:18:00	4179.25	4183.50	4179.25	4181.00	4088.0	0.000024	-0.000022

Tablica 4.2. Primjer dijela vektora značajki korištenog za treniranje modela

Ono što se može primjetiti je da se zbog načina na koji se računaju neke značajke, dio vremenskih podataka nedostaje. To se očituje u početnim minutama, odmah nakon 9.00 h i u zadnjih nekoliko minuta prije kraja tradea u 17.00 h. Iz tog razloga nedostajuće vrijednosti popunjene su uz pomoć forward filla i backward filla pazeći da su sve vrijednosti značajki osim značajke volumena popunjene s prethodnim ili sljedećim redovima, ovisno o kojem popunjavanju se radi, a da je za vrijednost volumena stavljena vrijednost 0. To je vrlo bitno jer je vrijednost volumena vrlo specifična i stoga vrlo teška za predvidjeti.

4.4. XGBoost model

Za klasifikaciju podataka odabran je Extreme Gradient Boosting (XGBoost) model. XGBoost skalabilni je sustav za učenje s pojačavanjem stabala i dostupan je kao open-source paket [36]. Najvažniji faktor iza uspjeha XGBoost-a je njegova skalabilnost u svim scenarijima. Sustav radi više od deset puta brže od postojećih popularnih rješenja i skalira se na milijarde primjera u distribuiranim ili memorijski ograničenim okruženjima. Skalabilnost XGBoost-a rezultat je nekoliko važnih optimizacija sustava i algoritama.

Algoritam radi kao i svaki nadzirani algoritam učenja kojemu je cilj minimizirati zadanu funkciju gubitka $L(y, h(x))$ nad skupom od N primjera za učenje, gdje su ulazne značajke $x_i \in \mathbb{R}^D$ povezane s pripadajućim oznakama y.

Algoritam konstruira aditivni model od M slabih učenika G_i, koji pripadaju istoj klasi modela, poput plitkih stabala odlučivanja. Hipoteza $h(x)$ definirana je kao ponderirani zbroj izlaza slabih učenika:

$$h(x) = \sum_{i=1}^{M} \beta_i G_i(x), \tag{4.1}$$

gdje je β_i težina i-tog slabog učenika.

Algoritam trenira slabe učenike sekvencijalno, prilagođavajući njihove parametre

kako bi minimizirao funkciju gubitka. U svakoj fazi m, algoritam računa negativni gradijent funkcije gubitka u odnosu na trenutni izlaz modela za svaki primjer i:

$$v_{i,m} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial h(x_i)}\right]_{h(x_i)=h_{m-1}(x_i)}, \tag{4.2}$$

gdje je h_{m-1} izlaz modela nakon prethodne faze.

Zatim se trenira novi slabi učenik G_m koristeći izračunate negativne gradijente $v_{i,m}$ kao oznake. Ovo se postiže minimizacijom funkcije gubitka najmanjih kvadrata:

$$G_m = \arg\min_G \sum_{i=1}^{N} (v_{i,m} - G(x_i))^2. \tag{4.3}$$

Nakon treniranja slabog učenika, algoritam izračunava težinu β_m koja minimizira funkciju gubitka kada se izlaz trenutnog modela kombinira s izlazom novog slabog učenika:

$$\beta_m = \arg\min_\beta \sum_{i=1}^{N} L(y_i, h_{m-1}(x_i) + \beta G_m(x_i)). \tag{4.4}$$

Dobivena težina β_m koristi se za ažuriranje modela tako da uključuje novog slabog učenika G_m:

$$h_m(x) = h_{m-1}(x) + \beta_m G_m(x). \tag{4.5}$$

Ovaj postupak ponavlja se kroz određeni broj iteracija ili dok funkcija gubitka ne konvergira. Na kraju, predikcije slabih učenika kombiniraju se pomoću ponderiranog većinskog glasanja kako bi se dobila konačna predikcija modela.

Nadalje, bitno je i naglasiti da je model vrlo fleksibilan i omogućuje detaljno podešavanje kroz različite hiperparametre. Korištenje pravih hiperparametara može značajno poboljšati performanse modela [37]. Ovo su parametri koji su izabrani za treniranje modela i njihova popratna objašnjenja [38]:

- **lambda** (L2 regularizacija):
  - Definira L2 regularizaciju, koja pomaže u smanjenju složenosti modela.
  - Koristi se raspon `np.logspace(-3, 1, num=3)`, što znači da se vrijednosti lambda uzimaju u logaritamskoj skali između 1×10^{-3} i 10.
- **alpha** (L1 regularizacija):
  - Definira L1 regularizaciju, koja također smanjuje složenost modela i omogućava veći broj nula u parametrima.
  - Koristi se raspon `np.logspace(-3, 1, num=3)`, s logaritamskim rasporedom vrijednosti između 1×10^{-3} i 10.
- **learning_rate** (brzina učenja):
  - Ovaj parametar kontrolira koliko će svako novo stablo ispraviti pogreške prethodnog.
  - Koristi se raspon `np.logspace(-3, -1, num=3)`, s logaritamskim rasporedom od 1×10^{-3} do 0.1, omogućujući podešavanje brzine učenja za bolju konvergenciju.
- **n_estimators** (broj stabala):
  - Definira broj stabala koja će bit trenirana.
  - Koristi se raspon `np.linspace(100, 500, num=5, dtype=int)`, što omogućuje odabir pet vrijednosti između 100 i 500.
- **max_depth** (maksimalna dubina stabla):
  - Definira maksimalnu dubinu svakog stabla. Veće vrijednosti mogu omogućiti složenije modele, ali mogu dovesti do prekomjernog prilagođavanja.
  - Koristi se raspon `np.arange(3, 7)`, što znači da vrijednosti variraju između 3 i 6.
- **min_child_weight** (minimalna težina djeteta):

  - Definira minimalnu težinu svih uzoraka u čvoru. Ako broj uzoraka u čvoru padne ispod ove vrijednosti, podjela čvora neće se dogoditi.
  - Koristi se raspon `np.arange(2, 6)`, što znači da se vrijednosti mogu birati između 2 i 5.
- **colsample_bytree** (uzorak značajki po stablu):
  - Definira proporciju značajki koje će bit nasumično odabrane za svako stablo.
  - Koriste se vrijednosti $[0.5, 0.7, 1.0]$, što znači da će bit odabrana različita veličina skupa značajki, od polovine do svih značajki.
- **subsample** (proporcija uzorka):
  - Ovaj parametar kontrolira koliko podataka će bit korišteno za treniranje svakog stabla. Smanjenje ove vrijednosti može smanjiti overfitting.
  - Koriste se vrijednosti $[0.5, 0.7, 1.0]$, što omogućuje korištenje različitih veličina podskupa podataka za treniranje modela.

4.5. Augmentacija oznaka

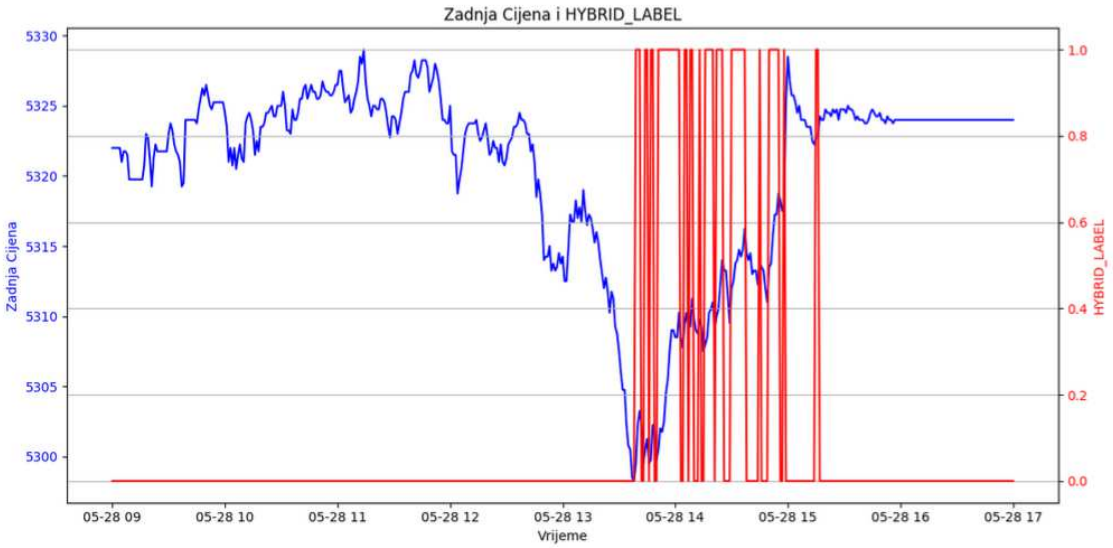
U kontekstu strojnog učenja augmentacija podataka predstavlja tehniku proširivanja trenutačnog skupa podataka s novim, sintetičkim podatcima, sve u svrhu povećavanja generalizacije, poboljšavanja predviđanja modela i proširivanja postojećeg skupa za treniranje uslijed manjka podataka, prisutstva šuma ili nebalansiranosti među klasama. Augmentacija skupa podataka stvaranjem novih, sintetičkih primjera može poboljšati performanse i generaliziranje modela. Ovo je osobito korisno za algoritme za označavanje trendova, gdje je cilj predvidjeti buduće pomake cijena, odnosno predvidjeti buduće financijske trendove [39].

Kako bi ispitali tvrdnju da augmentacija početnog skupa olakšava treniranje i pomaže robusnosti modela početni skup podataka proširen je s dvije nove oznake: HYBRID_LABEL i HYBRID_ORACLE čiji postupci dobivanja su u nastavku detaljnije objašnjeni.

Hibridna Oznaka

Prvi početni skup za treniranje sadržavao je oznake dobivene korištenjem sva tri algoritma za označavanje. Algoritmi su pokrenuti sa svojim optimalnim parametrima, čije treniranje je prethodno izvedeno te unutar kojih je uključena i konstantna provizija za trgovanje. Korištena cijena za računanje oznaka je zadnja cijena (eng. close price). Nakon što su oznake izračunate, nova, hibridnu oznaka, dobivena je tako što joj je vrijednost oznake 1 samo u trenutku kada su sve tri preostale oznake imale vrijednost 1, inače je hibridnoj oznaci dodijeljena vrijednost 0.

Primjer Hybrid_Label oznake trenda na grafu može se vidjeti na slici ispod:



Slika 4.2. Hibridna oznake trenda Hybrid_Label

Hibridna Oracle oznaka

Kako bi dobili drugu hibridnu oznaka, Hybrid Oracle, korišten je isključivo Oracle algoritam za označavanje. Algoritam je primjenjen četiri puta i to svaki puta s drugom vrstom cijene (open, high, low, close). Nova, hibridna oznaka dobila je vrijednost 1 samo u slučaju kada su sve ostale četiri oznake imale vrijednost 1 u tom trenutku, dok u ostalim slučajevima vrijednost hibridne značajke bila je 0.

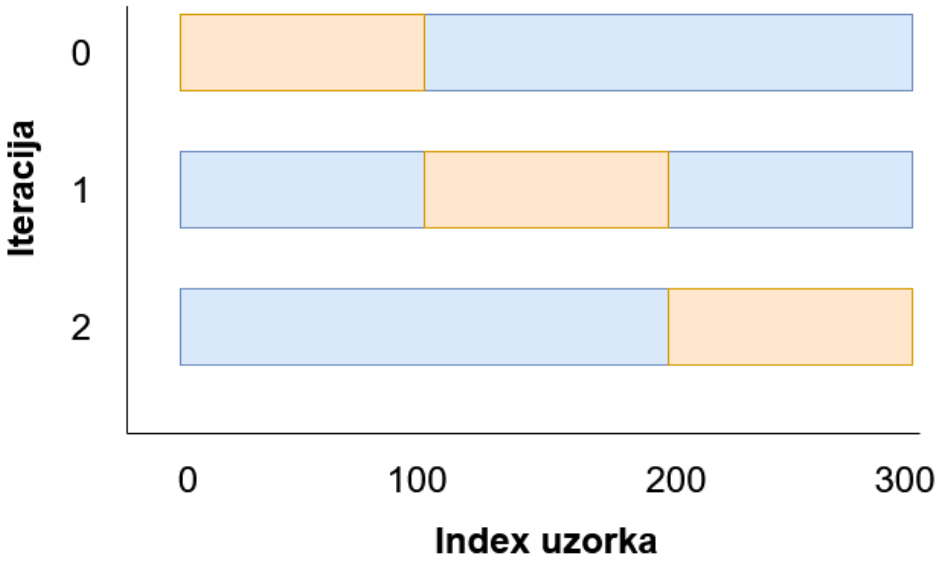
Primjer Hybrid_Oracle oznake trenda prikazana je na slici ispod:

Slika 4.3. Hibridna oznake trenda Hybrid_Oracle

4.6. Postupak treniranja

Unakrsna validacija metoda je za procjenu sposobnosti generalizacije modela. Ideja je da se unutar familije modela odabere onaj model koji najbolje generalizira, odnosno koji dobro radi na neviđenim primjerima. Simulacija neviđenih primjera ostvarena je tako što je dio primjera iz skupa za učenje izdvojen i glumi neviđene primjere [40]. Skupovi da bi bili dobri moraju biti disjunktni, odnosno ne smiju se preklapati. Uobičajena podjela je 70% primjera za učenje i 30% primjera za ispitivanje ili 60:40 i slično, sve ovisno o potrebama i količini početnih podataka.

Jedna od mogućih metoda je i K-fold unakrsna validacija gdje se podatci podijele na k podskupova i onda se trenira model na svim podskupovima, osim jednog koji se koristi za ispitivanje. Taj proces ponavlja se sve dok model nije ispitan na svim podskupovima, odnosno k puta, i onda se za finalnu metriku uzima prosjek rezultata dobivenih u svakoj pojedinačnoj iteraciji. Tako opisani način sprječava prenaučenost modela i evaluira performanse modela na više robustan način nego običnom podjelom na učenje i ispitivanje. Najčešće se nasumično odabiru uzorci iz dostupnih podataka i onda se dijele u skupove za treniranje i ispitivanje. To se može i vidjeti na slici ispod gdje su skupovi podataka korišteni za treniranje prikazani plavom bojom, dok su skupovi podataka za ispitivanje prikazani narančastom bojom:



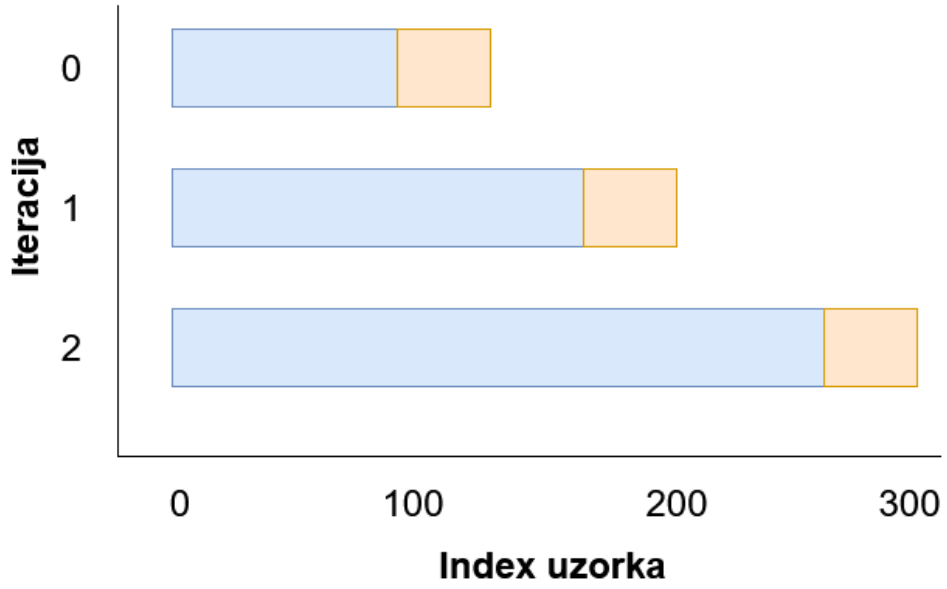
Slika 4.4. Primjer K-fold unakrsne validacije

Problem kod tog pristupa je taj što se kod takve metode podatci nasumično odabiru što ne odgovara radu s vremenskim nizovima. Ne mogu se za predikciju vremenskih nizova koristiti podatci iz budućnosti ili podatci koji nisu vremenski slijedni. Vremenska relacija mora se očuvati.

Dvije tehnike za provođenje unakrsne validacije koje rješavaju gore navedeni problem su: unakrsna validacija s podjelom vremenskih nizova i blokovska unakrsna validacija.

Općenito, metoda koja se može koristiti za unakrsnu validaciju modela vremenskih nizova unakrsna je validacija na osnovi kotrljajućeg (rolling) pristupa. Radi se tako što se započinje s manjim podskupom podataka za treniranje i onda se predviđaju kasniji podatci na kojima se potom provjerava točnost za predviđanje. Ti se onda isti predviđeni podatci uključuju kao dio sljedećeg skupa za treniranje, a ostali podatci se predviđaju.

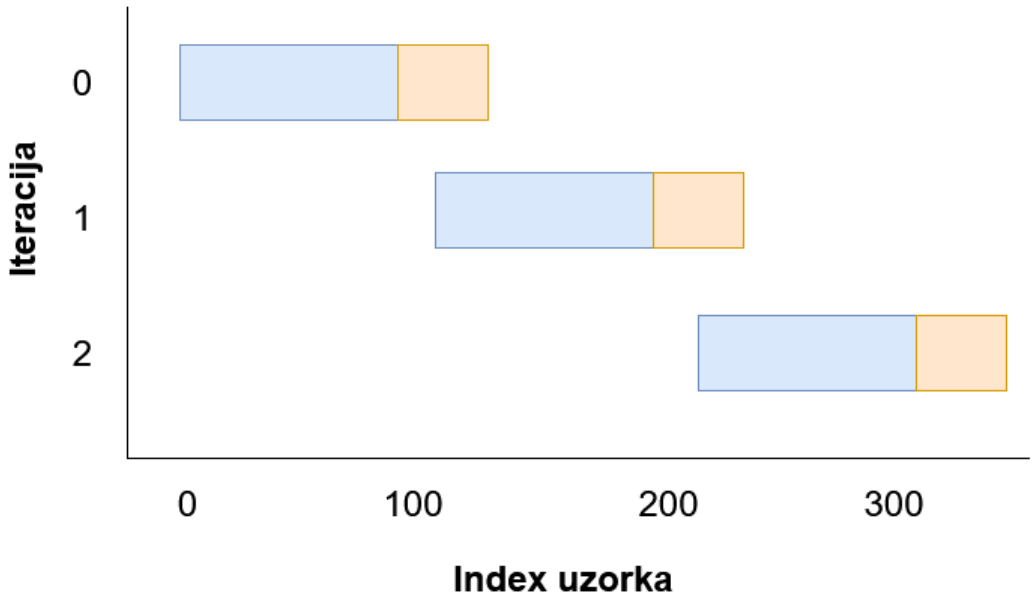
Kod unakrsne validacije s podjelom vremenskih nizova vizualna reprezentacija može se vidjeti na slici ispod:



Slika 4.5. Primjer unakrsne validacije s podjelom vremenskih nizova

Horizontalna os prikazuje vremenski slijed budući da podatci nisu nasumično promiješani, već je očuvan kronološki redoslijed. Ideja kod podjele vremenskih nizova je da se u svakoj iteraciji skup za treniranje podijeli u dva dijela uz uvjet da je skup za ispitivanje uvijek vremenski ispred skupa za treniranje.

Međutim, takav pristup može dovesti do propuštanja informacija iz budućih podataka u model. Model će promatrati buduće obrasce kako bi ih predvidio i pokušati ih zapamtiti. Iz tog razloga uvedena je blokovska unakrsna validacija. Ona funkcionira tako što dodaje na dvije pozicije marginu. Prva pozicija je između skupova za treniranje i ispitivanje kako bi se spriječilo da model promatra vrijednosti zaostajanja koje se koriste dvaput, jednom kao regresor, a drugi put kao odgovor. Druga je između skupova korištenih u svakoj iteraciji kako bi se spriječilo da model pamti obrasce iz jedne iteracije u sljedeću [41]. Vizualizacija se vidi na slici ispod:



Slika 4.6. Primjer blokovske unakrsne validacije vremenske serije

4.7. Traženje optimalnih parametara za algoritme označavanja

Zbog povećanja robusnosti modela i efikasnosti predikcije algoritme za označavanje potrebno je optimizirati nalaženjem najboljih parametara koji maksimiziraju određeni kriterij. Kriterij koji je korišten je maksimizacija balansiranosti oznaka u obje klase, odnosno da je broj pozitivnih (oznaka 1) i negativnih oznaka (oznaka 0) što bliži. Oracle algoritam nije optimiziran jer su parametri njegovog algoritma zadnja cijena i fee koji je konstantan (određena mu je vrijednost 0.001). Optimizacija FTH i CTL algoritma rađena je na sličan način, uz razliku što je FTH, uz zadnju cijenu i fee, sadržavao i parametre τ i H, dok CTL parametar ω. Optimizacija je provedena definiranjem raspona mogućih vrijednosti za svaki parametar po kojem se onda iteriralo i za svaki dan računale oznake dobivene algoritmima označavanja s njihovim odgovarajućim parametrima. Potom, računan je broj dobivenih pozitivnih oznaka i njihov omjer u ukupnom broju oznaka. Ako je omjer pozitivnih oznaka u toj iteraciji bio bliži 50% u usporedbi s ostalim omjerima, navedeni omjer i parametri bili bi pospremljeni u varijable kao optimalni parametri. Nakon što je algoritam prošao kroz sve kombinacije parametara na izlaz bi ispisao najbolji omjer, onaj koji je najbalansiraniji i parametre za koje je on dobiven.

Korišteni rasponi za traženje optimalnih parametara prikazani su u tablici ispod:

Tablica 4.3. Rasponi parametara korištenih u analizi

Parametar	Raspon	Broj Vrijednosti
τ	$[0.0001, 0.001]$	200
H	$\{1, 2, 3, \ldots, 9\}$	9
ω	$[0.0001, 0.001]$	200

Nakon što je program prošao kroz sve navedene kombinacije parametara, dobivene su navedene vrijednosti za optimalne parametre koji maksimiziraju balansiranost među klasama i njihovi odgovarajući omjeri:

Tablica 4.4. Optimalni Parametri za FTH i CTL Označavanje

Algoritam	Najbolji Parametri	Najbolji Omjer Pozitivnih Oznaka
FTH	$\tau = 0.0001, H = 66$	0.4196
CTL	$\omega = 0.0001$	0.4328

4.8. Podjela podataka

S obzirom da je treniranje modela rađeno blokovskom unakrsnom podjelom, za treniranje modela uzeti su podatci od zadnjih 22 tjedna na koje su implementirane 66 značajki.

Na taj skup dodatno su augmentirane još dvije oznake, HYBRID_LABEL i HYBRID_ORACLE čiji je postupak dobivanja maloprije objašnjen. Navedeni skup zatim je podijeljen na vektor značajki X i vektor oznaka Y koji su korišteni za treniranje modela.

4.9. Treniranje i validacija modela

Treniranje modela provedeno je koristeći Extreme Gradient Boosting (XGBoost), vrlo učinkovit i široko korišten model gradijentno pojačanih stabala [35]. Na početku definirane su sve kombinacije parametara modela kroz koje se iterira i čiji raspon vrijednosti može se vidjeti u tablici ispod:

Tablica 4.5. Rasponi parametara za pretragu hiperparametara

Parametar	**Raspon Vrijednosti**
lambda	$\{10^{-3}, 10^{-1}, 10^{1}\}$
alpha	$\{10^{-3}, 10^{-1}, 10^{1}\}$
learning_rate	$\{10^{-3}, 10^{-2}, 10^{-1}\}$
n_estimators	$\{100, 200, 300, 400, 500\}$
max_depth	$\{3, 4, 5, 6\}$
min_child_weight	$\{2, 3, 4, 5\}$
colsample_bytree	$\{0.5, 0.7, 1.0\}$
subsample	$\{0.5, 0.7, 1.0\}$

Validaciju je provedena već spomenutom blokovskom unakrsnom validacijom i to tako da su za prvi skup za učenje korišteni podatci od prvih 3 mjeseca na kojima se trenirao model, a provjera se radila na skupu za ispitivanje koji je bio veličine 2 tjedna. Na tim podatcima računata je ROC AUC mjera modela koja je potom spremljena u listu. Proces je ponovljen 5 puta za istu kombinaciju parametara i uz pomicanje skupova za 2 tjedna unaprijed. Nakon 5 iteracija izračunata je prosječna vrijednost ROC AUC-a za koju se zatim provjerava je li to najveća dobivena vrijednost do tada. Ako je, vrijednosti modela, parametara i ROC AUC mjera spremljeni su u varijable. Nakon što je algoritam prošao kroz sve iteracije ispisuju se parametri za koje su dobivene najveće vrijednosti ROC AUC mjere.

Nakon završenog treniranja, možemo vidjeti vrijednosti optimalnih parametara za model treniran na različitim oznakama.

Tablica 4.6. Optimalni parametri za oznake FTH, Oracle, CTL, Hybrid oznaka i Hybrid Oracle.

Model	lambda	alpha	learning_rate	n_estimators
FTH	0.001	10.0	0.1	100
Oracle	0.001	0.1	0.01	500
CTL	0.1	10.0	0.1	100
Hybrid Oznaka	0.1	10.0	0.01	200
Hybrid_Oracle	0.1	0.1	0.1	100

max_depth	min_child_weight	colsample_bytree	subsample
3	2	1.0	0.5
4	2	0.7	0.7
3	3	1.0	0.5
4	2	0.5	0.5
3	2	1.0	0.7

4.10. Vrednovanje modela

Kako bi se znalo koji je algoritam najbolji za naš problem potrebno je usporediti predikcije algoritama. Kriterij koji je tu bitan je točnost algoritma na neviđenim podatcima. A da bi to znalo potrebno je znati kako ispravno vrednovati algoritam. Ove tri stvari bitne su za vrednovanje algoritama strojnog učenja:

1. Koju mjeru vrednovanja koristiti.

2. Kako realno procijeniti pogrešku modela.

3. Kako napraviti statističku analizu rezultata.

Mjere vrednovanja su metode koje kvantificiraju točnost ili pogrešku klasifikatora i u teoriji temelje se na usporedbi vektora y_{pred} (predviđene oznake dobivene predikcijom klasifikatora) i y_{true} (stvarne oznake).

Koje se sve mjere mogu izračunati postaje jasnije ako se iz vektora y_{pred} i y_{true} izgradi matrica zabune (kontingencije). To je kvadratna matrica koja na sažet način prikazuje broj podudaranja i nepodudaranja predikcija i stvarnih oznaka. Za binarni klasifikator ona je dimenzija 2x2 i izgleda ovako:

Tablica 4.7. Matrica zabune

	ytrue = 1	**ytrue = 0**
ypred = 1	TP	FN
ypred = 0	FP	TN

Ovdje retci odgovaraju predviđenim oznakama (`pred`), a stupci stvarnim oznakama (`true`). Razlikujemo četiri vrste slučajeva, koji odgovaraju elementima matrice:

- **Stvarno pozitivni** (engl. *true positive*, TP): predikcija je 1 i stvarna oznaka je 1;
- **Lažno pozitivni** (engl. *false positive*, FP): predikcija je 1, a stvarna oznaka je 0;
- **Lažno negativni** (engl. *false negative*, FN): predikcija je 0, a stvarna oznaka je 1;
- **Stvarno negativni** (engl. *true negative*, TN): predikcija je 0 i stvarna oznaka je 0.

Mjera koja se prva može izračunati je točnost (engl. accuracy) i ona je udio točno klasificiranih primjera u skupu svih primjera:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{4.6}$$

Problem s točnošću javlja se kada je skup podataka jako neuravnotežen pa ima mnogo više pozitivnih primjera od negativnih ili obrnuto. Onda je najlakše klasifikator definirati da uvijek predviđa većinsku klasu i tim odabirom točnost klasifikatora bit će vrlo visoka. Iz tog razloga osmišljene su alternativne mjere vrednovanja kao što su:

1. **Preciznost** (engl. *precision*):

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.7}$$

   Preciznost se definira kao udio stvarno pozitivnih primjera (TP) u skupu svih primjera koje je klasifikator označio pozitivno (TP + FP). Idealno, $P = 1$, tj. svi primjeri koje je klasifikator označio pozitivno doista i jesu pozitivni.

2. **Odziv** (engl. *recall*):

$$R = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.8}$$

   Odziv je udio stvarno pozitivnih primjera (TP) u skupu svih pozitivnih primjera (TP + FN). Ova se mjera naziva "odziv" jer nam govori koliko se pozitivnih primjera "odazvalo" klasifikatoru. Idealno, $R = 1$, tj. sve pozitivne primjere klasifikator će označit kao takve. Alternativni nazivi za odziv su stopa stvarnih pozitiva (engl. *true positive rate*, TPR) i osjetljivost (engl. *sensitivity*).

3. **Ispadanje** (engl. *fall-out*, *false positive rate*):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{4.9}$$

   Ispadanje (ili stopa lažnog alarma) je udio lažno pozitivnih primjera (FP) u skupu svih negativnih primjera (FP + TN). Idealno, FPR = 0, tj. klasifikator niti jedan negativni primjer neće lažno proglasit pozitivnim.

4. **Specifičnost** (engl. *specificity*, *true negative rate*):

$$S = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4.10}$$

   Specifičnost je udio stvarno negativnih primjera (TN) u skupu svih negativnih primjera (TN + FP). Idealno, $S = 1$, tj. klasifikator će sve negativne primjere klasificirati kao takve.

Preciznost i odziv daju različite informacije. Ako se klasifikator modelira tako da ima visok odziv, onda je uobičajeno da će imati nižu preciznost i obrnuto. Zato, ako se priča o rezultatu klasifikatora trebale bi se navesti obje mjere. Međutim, često je potrebno točnost klasifikatora prikazati samo jednim brojem. Mjera vrednovanja koja radi upravo to je mjera F1 (engl. F1 score). Mjera F1 definirana je kao harmonijska sredina preciznosti i odziva:

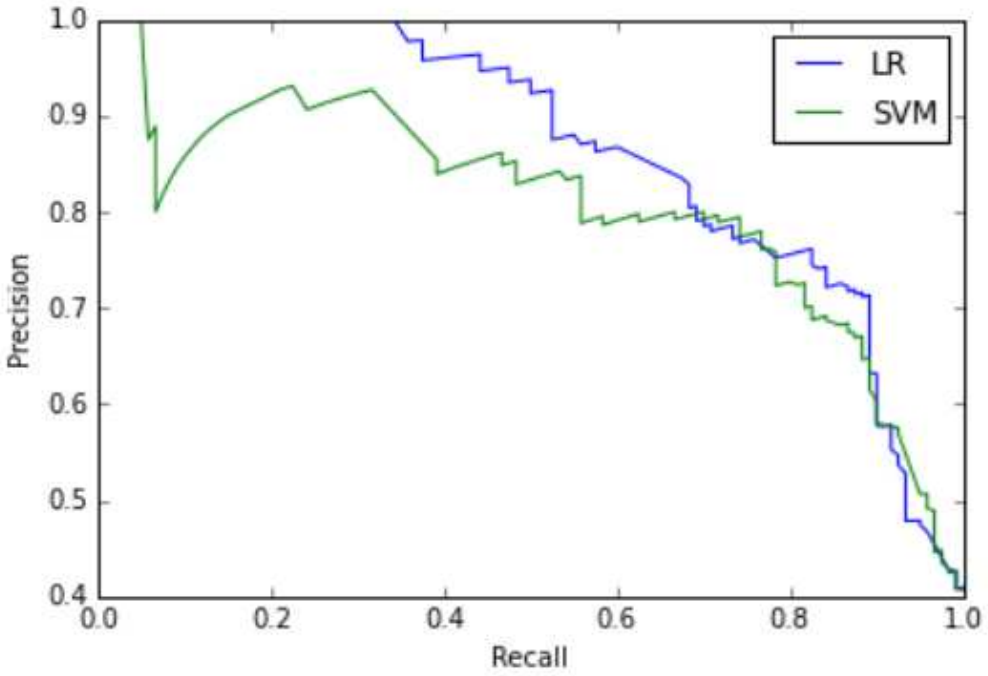
$$F_1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{4.11}$$

4.10.1. Vrednovanje klasifikatora s pragom

Do sada smo pričali o vrednovanju klasifikatora koji na izlazu daju oznaku klase. Mnogi klasifikatori kao što su logistička regresija, naivan Bayesov klasifikator na izlazu daju vjerojatnost klasifikacije, a ne konkretnu oznaku klase. Uobičajena je vrijednost klasifikatorskog praga 0.5. Pitanje je što ako se prag postavi na vrijednost višu od 0.5. U tom slučaju primjeri bi bili klasificirani u pozitivnu klasu samo kada je vjerojatnost da je primjer pozitivan jako velika. Moguća posljedica toga je i da će bit manji broj lažno pozitivnih primjera čime preciznost raste. Također ako se prag smanji na vrijednosti manje od 0.5 time se i potencijalno smanjuje broj lažno negativnih primjera i povećava odziv modela. Ono što se može zaključiti je da se manipuliranjem klasifikatorskog praga direktno utječe na preciznost i odziv klasifikatora.

4.10.2. Krivulja preciznost-odziv

Ako želimo vrednovati klasifikator s obzirom na sve moguće vrijednosti klasifikacijskog praga, možem se skicirati krivulja preciznost-odziv. Krivulja se dobije tako što se vrijednosti praga smanjuju od 1 do 0 te za svaku vrijednost praga potom izračunaju odziv i preciznost klasifikatora. Dobivena krivulja može izgledati ovako:

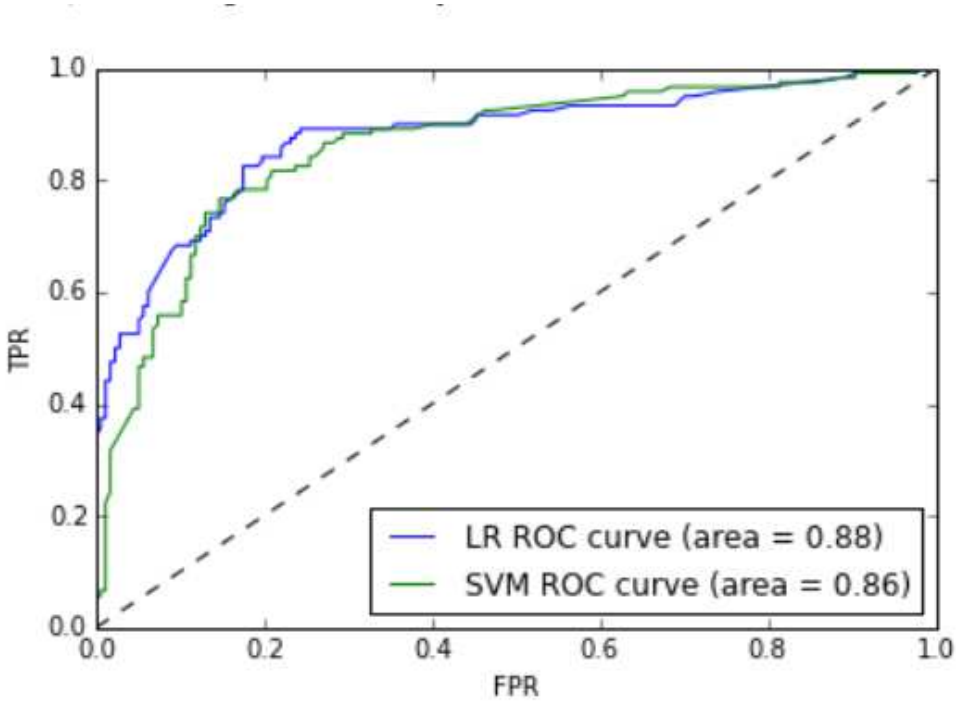


Slika 4.7. Krivulja preciznost-odaziv

Na slici vide se dvije krivulje, plava koja predstavlja model logističke regresije i zelena koja predstavlja model SVM-a, obje dobivene na istom skupu podataka. Drugim riječima, preferiraju se krivulje koje su što bliže točki (R = 1, P = 1), odnosno što bliže gornjem desnom rubu. Prema tome vidi se da krivulja logističke regresije se nalazi bliže tom rubu i da je ona na većini skupa iznad krivulje SVM-a. Stoga, može se zaključiti da je klasifikator LR bolji od klasifkatora SVM-a. Ako se želi proglasiti jedan klasifikator boljim od drugoga njegova krivulja na cijelom području mora biti iznad krivulje lošijeg klasifikatora.

4.10.3. Krivulja ROC

Krivulja ROC (engl. area under ROC curve), skraćeno AUC predstavlja vrijednost stope stvarnih pozitiva (TPR), isto što i odziv, kao funkcije stope lažnog alarma (FPR), odnosno ispadanja. FPR mjeri koliki je udio primjera koje je klasifikator pogrešno proglasio pozitivnima, a TPR koliko je pozitivnih primjera klasifikator detektirao. Ovako izgleda ROC krivulja:



Slika 4.8. Krivulja ROC

Klasifikator je to bolji što njegova krivulja ROC prolazi bliže točki (FPR = 0, TPR = 1), odnosno što je bliže sada gornjem lijevom kutu.

Velika prednost krivulje ROC jest što ona za nasumični klasifikator (klasifikator koji primjere klasificira u nasumično odabrane klase) odgovara pravcu od $(0, 0)$ do $(1, 1)$. Ono što je pogotovo dobro je to da to vrijedi neovisno o tome je li broj pozitivnih i negativnih primjera uravnotežen. Mjera AUC jednostavno je definirana kao površina ispod krivulje ROC. AUC je u intervalu [0,1], što više, to bolje. Za nasumični klasifikator, AUC = 0.5 [42].

5. Rezultati

Nakon što je model istreniran i dobiveni su njegovi optimalni parametri, potrebno je usporediti koliko dobro model predviđa klase, odnosno trendove na neviđenim primjerima. U mjerenju uspješnosti modela korisna je mjera financijskoga učinka, za koju je u ovom radu odabran kumulativni povrat.

Naime, nakon nalaženja optimalnih parametara modela, uzet je skup podataka od zadnjih 6 mjeseci i podijeljen na skup za učenje (22 tjedna) i skup za ispitivanje (2 tjedna). Model je potom inicijaliziran s optimalnim parametrima i treniran na skupu za učenje, pri čemu je za ciljnu varijablu uzeta oznaka koja je dobivena jednim od različitih algoritmima za označavanje i augmentacijom postojećih oznaka. Nakon toga model je predvidio vjerojatnosti za svaku klasu iz čega je izdvojena vjerojatnost za pozitivom klasom za koju se provjerava je li njena predviđena vjerojatnost veća od 0.5. Ako je, uzorku se pridjeljuje pozitivna oznaka 1, inače 0.

Potom, iterira se po skupu za ispitivanje po svakom danu i računaju dnevni kumulativni povrati, prvo na stvarnim oznakama iz skupa za ispitivanje, a onda i na oznakama koje su predviđene preko modela strojnog učenja. Nakon što je završeno računanje za sve dane u skupu, izračunaju se prosječne kumulativni povrati za stvarne i predviđene oznake i ispisuju dobiveni rezultate. Rezultati za svaku pojedinu oznaku prikazani su u tablici ispod:

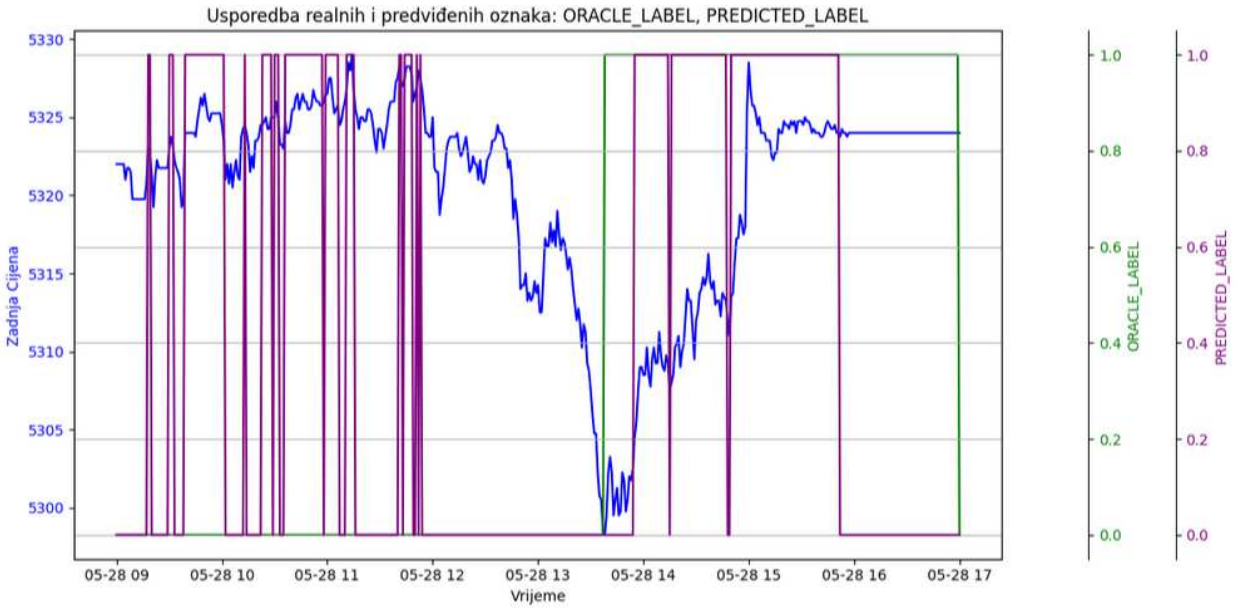
Tablica 5.1. Prosječni kumulativni povrati za različite algoritme označavanja.

Oznake	Povrat stvarnih oznaka	Povrat predviđenih oznaka
ORACLE	0.00408	-0.00870
FTH	-0.01553	-0.00512
CONTINUOUS_TREND	-0.03621	-0.06931
HYBRID_LABEL	-0.00924	X
HYBRID_ORACLE	0.00337	-0.01139

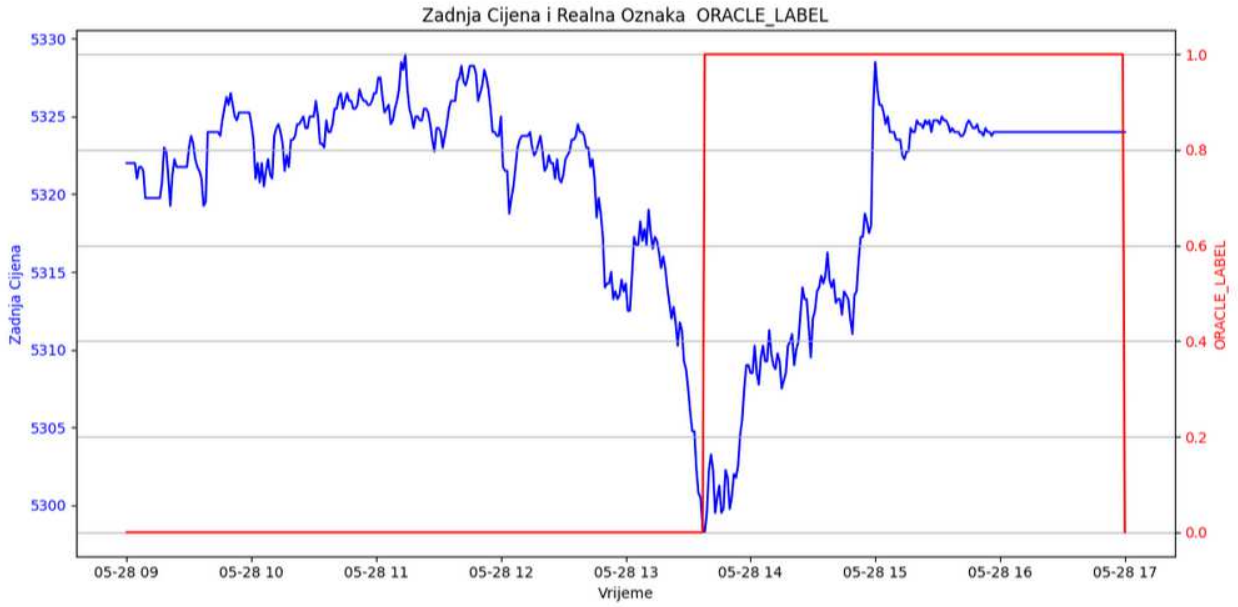
Gornja tablica prikazuje rezultate prosječnih kumulativnih povrata računatih na stvarnim i predviđenim oznakama. Ovi očekivani kumulativni povrati izračunati su uzimajući u obzir transakcijske troškove $\phi = 0.001$. Vidi da je za sve oznake povrat predviđenih oznaka manji od stvarnih što je i očekivano. Za Hybrid Label oznaku, dobivenu prvom augmentacijom, zbog loše balansiranosti klasa predviđena oznaka nema niti jednu pozitivnu oznaku te stoga se ne može izračunati njen kumulativni povrat. Najveći kumulativni povrat računat na stvarnim oznaka dobiven je za oznake dobivene Oracle algritmom za označavanje i čiji povrat iznosi 0.00408. Pozitivni kumulativni povrat dobiven je još i za augmentaciju Oracle algoritma koje je u odnosu na originalni Oracle algoritam, bez augmentacije, manji i iznosi 0.00337. Ostale oznake dobivene FTH, CTL algoritmom i prvom, Hybrid Label, augmenacijom vraćaju negativne povrate, isto tako kao i povrati računani na svim prevdiđenim oznakama od svakog algoritma za označavanje i obje augmentacije što sugerira da strategije temeljene na tim oznakama nisu uspjele generirati pozitivne prinose u promatranom razdoblju.

Grafička usporedba stvarnih i predviđenih trendova prikazana je na sljedećim grafovima, redom za Oracle, FTH i CTL algoritme, te na kraju za hibridna označavanja. Prvi graf prikazuje zajedničku usporedbu oznaka trenda, dok su u nastavku prikazane pojedinačne oznake zasebno na svojim grafovima zbog preglednosti:

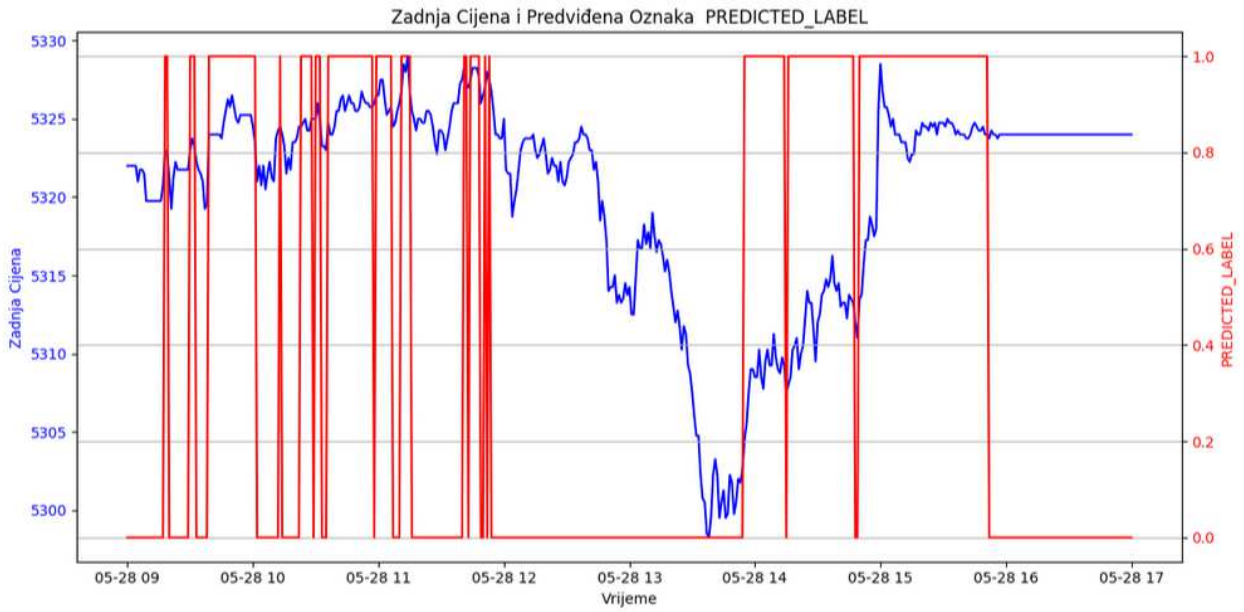
Oracle algoritam označavanja



Slika 5.1. Primjer stvarnih i predviđenih oznaka trenda dobivenih Oracle algoritmom

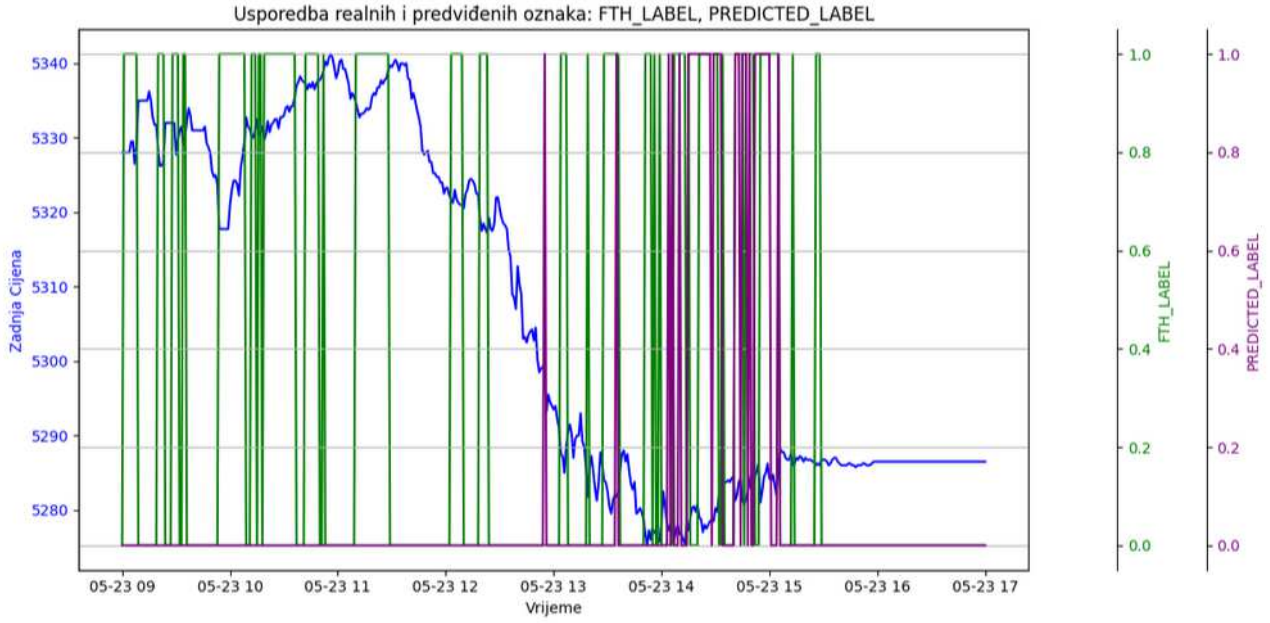


Slika 5.2. Primjer stvarnih oznaka trenda dobivenih Oracle algoritmom

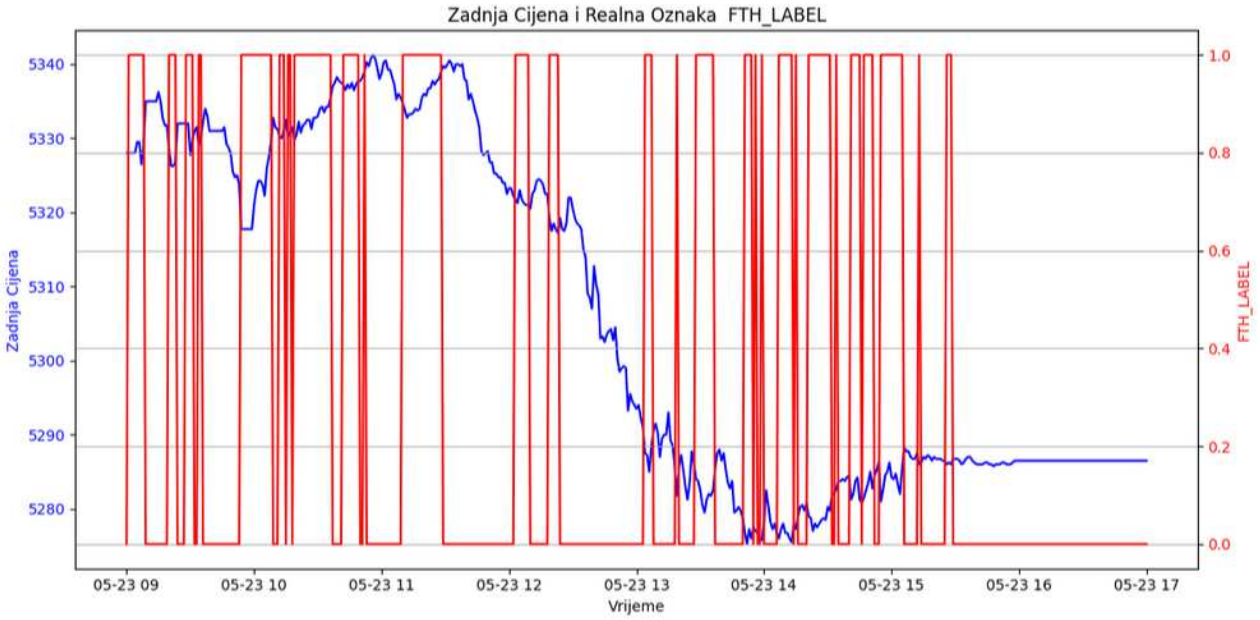


Slika 5.3. Primjer predviđenih oznaka trenda dobivenih Oracle algoritmom

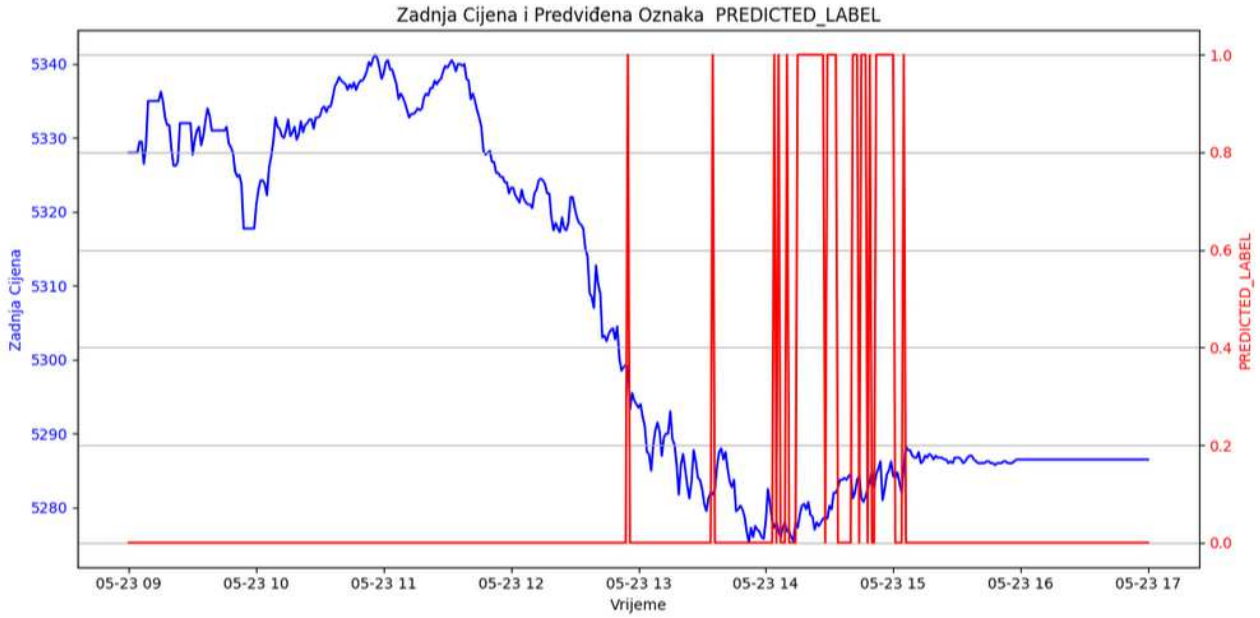
FTH algoritam označavanja



Slika 5.4. Primjer stvarnih i predviđenih oznaka trenda dobivenih FTH algoritmom

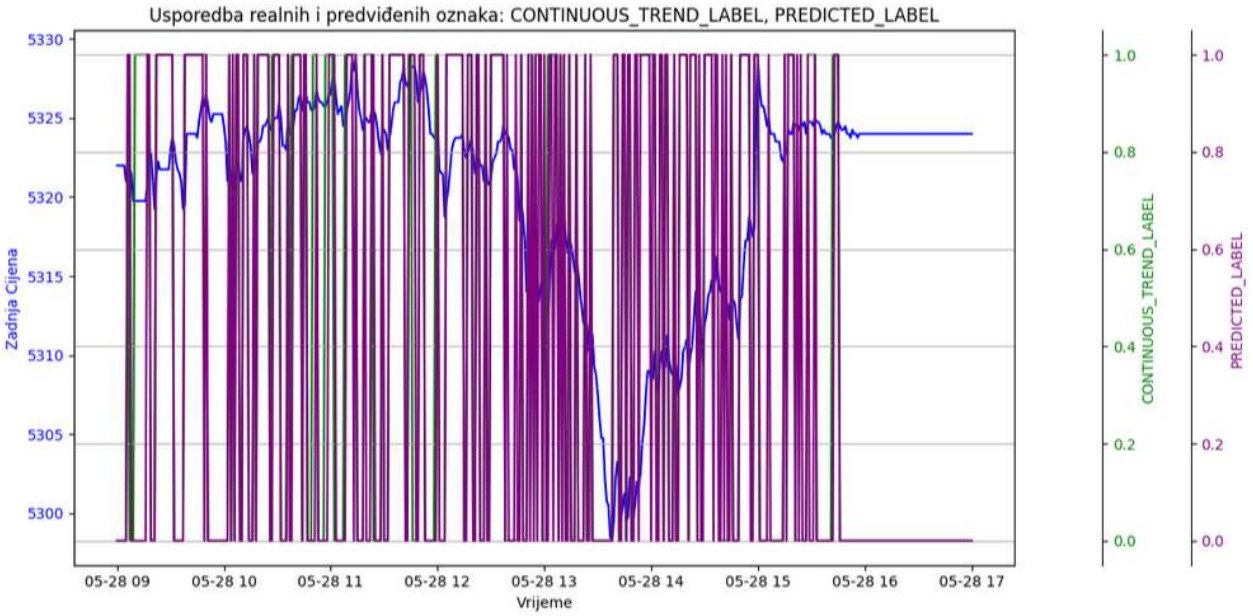


Slika 5.5. Primjer stvarnih oznaka trenda dobivenih FTH algoritmom

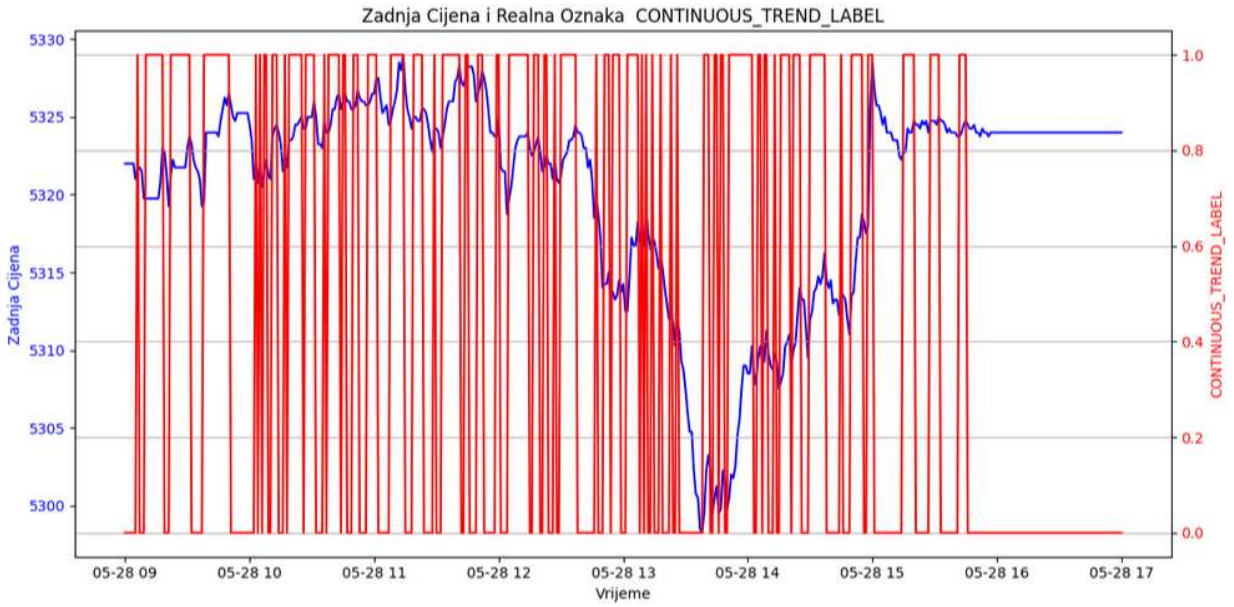


Slika 5.6. Primjer predviđenih oznaka trenda dobivenih FTH algoritmom

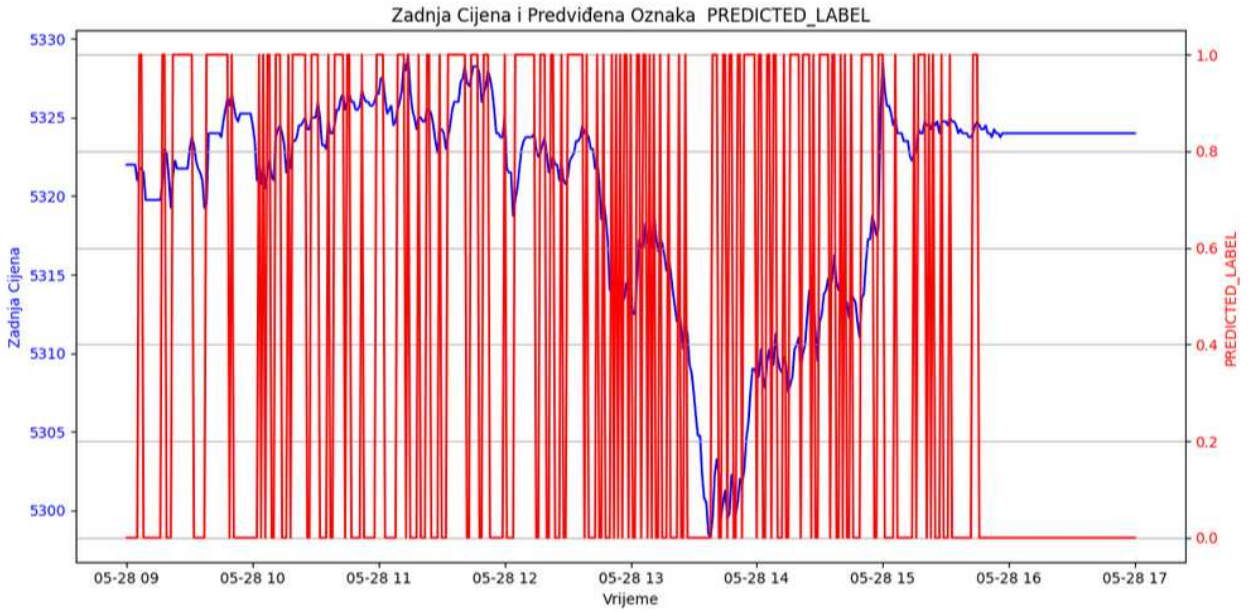
CTL algoritam označavanja



Slika 5.7. Primjer stvarnih i predviđenih oznaka trenda dobivenih CTL algoritmom

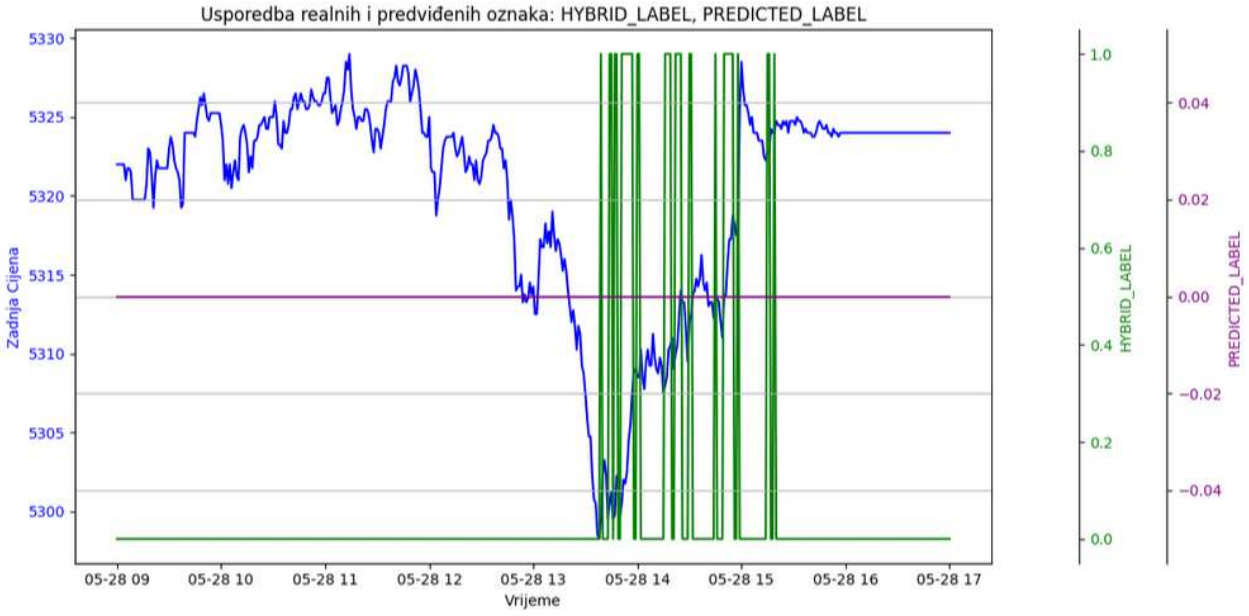


Slika 5.8. Primjer stvarnih oznaka trenda dobivenih CTL algoritmom



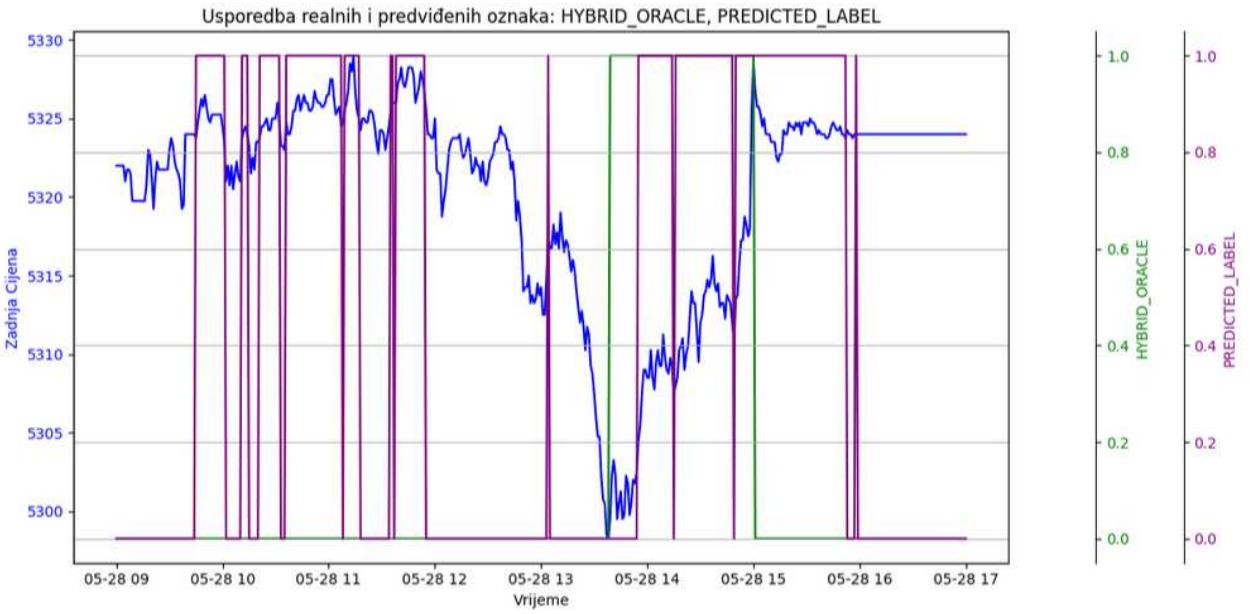
Slika 5.9. Primjer predviđenih oznaka trenda dobivenih CTL algoritmom

Hybrid Label augmentirano označavanje

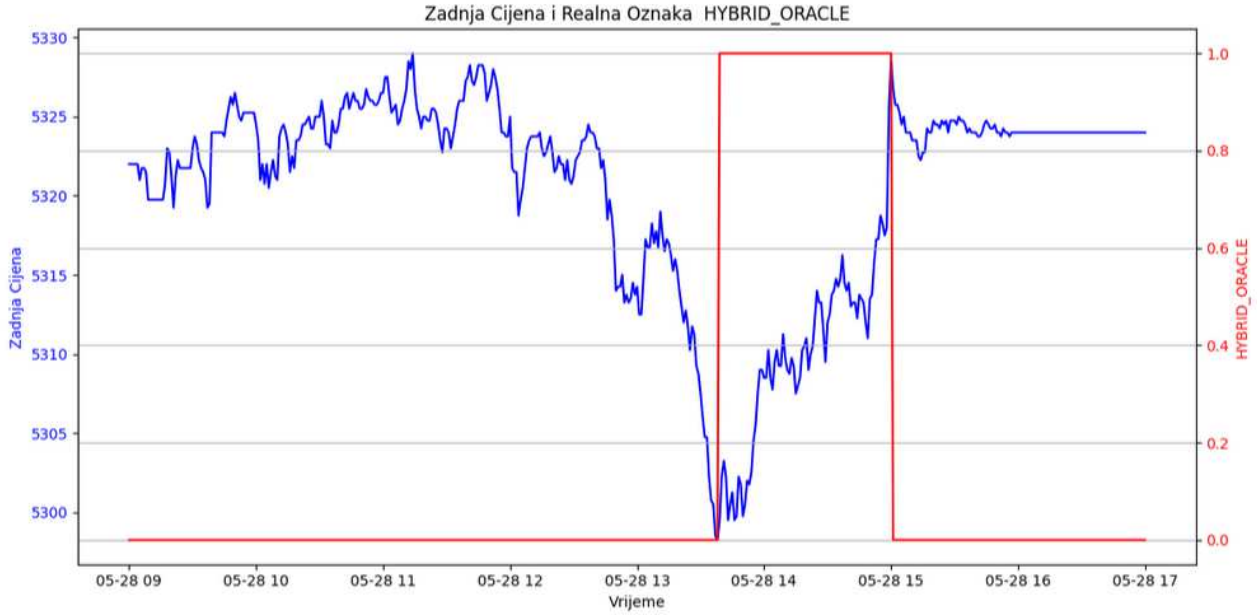


Slika 5.10. Primjer stvarnih i predviđenih Hybrid Label oznaka dobivenih augmentacijom

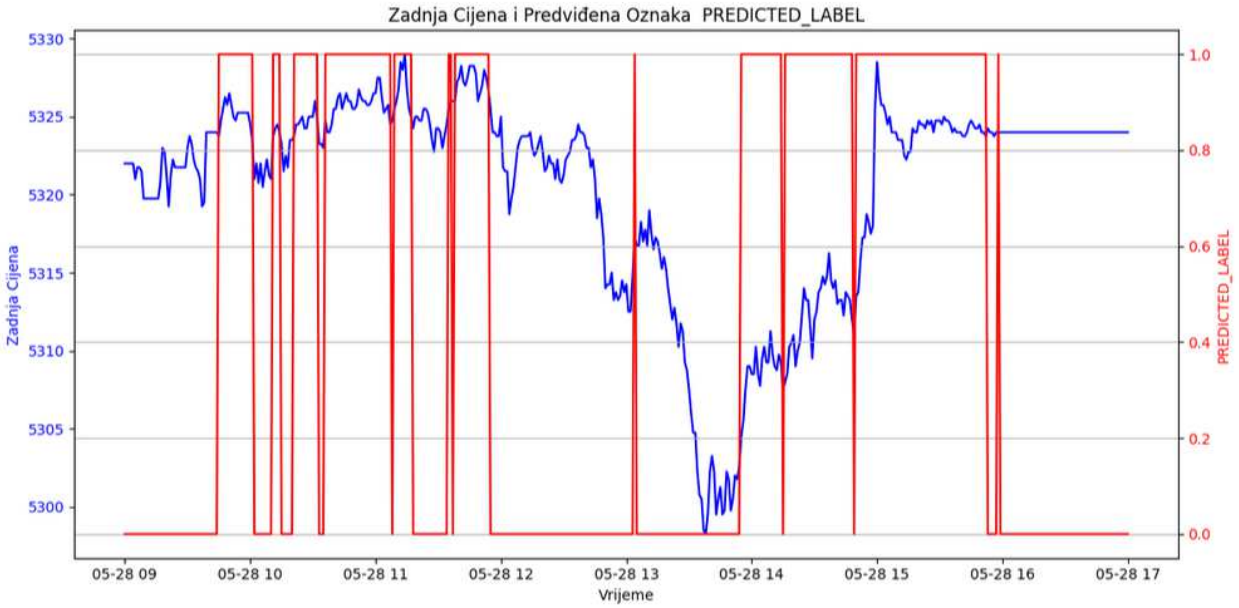
Hybrid Oracle augmentirano označavanje



Slika 5.11. Primjer stvarnih i predviđenih Hybrid Oracle oznaka dobivenih augmentacijom

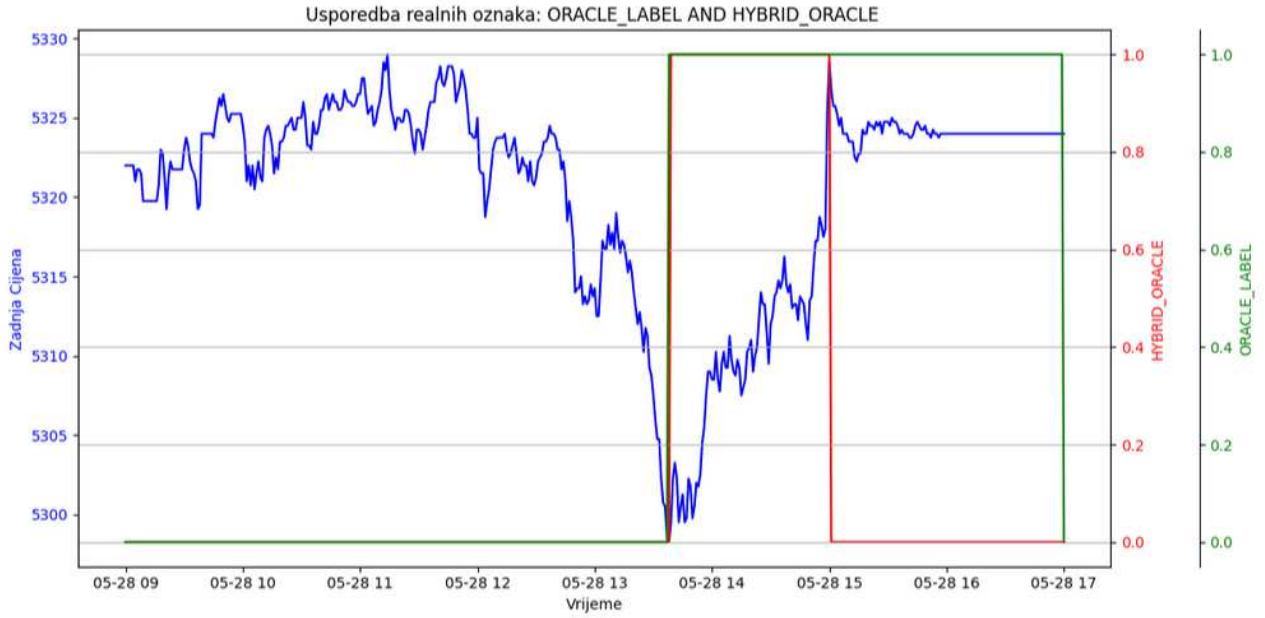


Slika 5.12. Primjer stvarnih Hybrid Oracle oznaka dobivenih augmentacijom



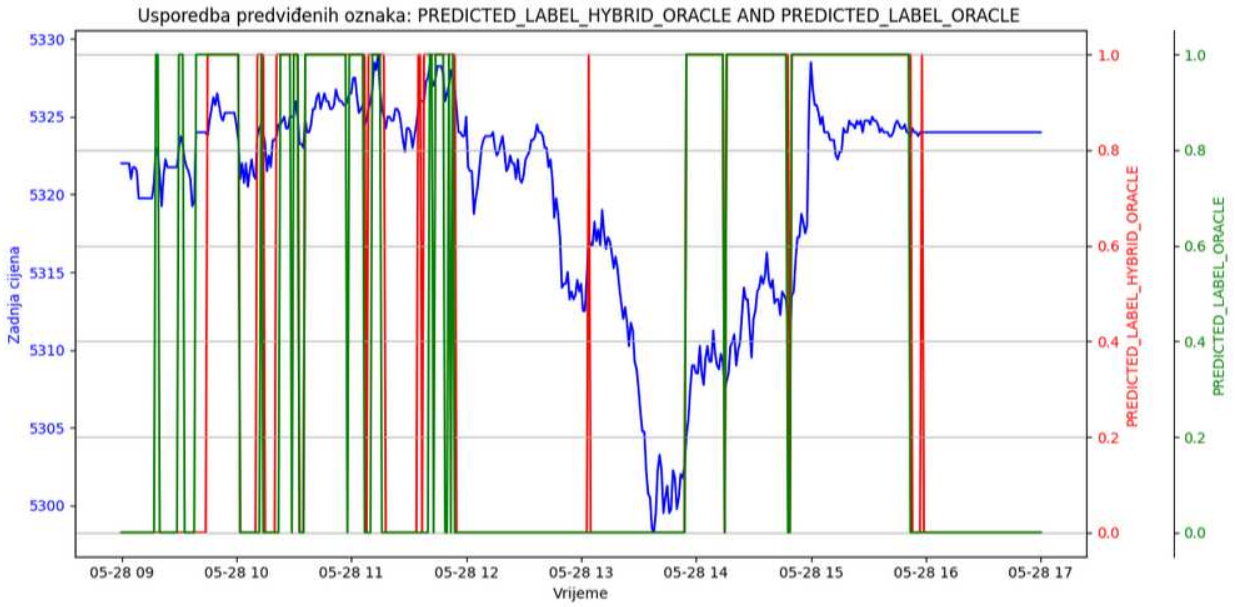
Slika 5.13. Primjer predviđenih Hybrid Oracle oznaka dobivenih augmentacijom

Usporedba Oracle i Hybrid Oracle označavanja



Slika 5.14. Primjer stvarnih Oracle i Hybrid Oracle oznaka trenda

Na sljedećem grafu možemo vidjeti i usporedbu oznaka trenda dobivenih Oracle algoritmom za označavanje i oznaka dobivenih drugom augmentacijom sa različitim ulazima cijena u Oracle algoritam:



Slika 5.15. Primjer predviđenih Oracle i Hybrid Oracle oznaka trenda

Zadnji eksperiment u ovom radu je i treniranje modela na skupu koji je dobiven na malo drukčiji način. Obično, za svaki uzorak podataka X, koristi se jedan algoritam označavanja za generiranje izlaznih oznaka Y. To znači da je model treniran na jednom skupu podataka gdje je svaki primjer imao samo jednu oznaku. Međutim, sada želimo proširiti dataset tako da za svaki ulazni uzorak X dobijemo tri različite oznake, koristeći tri različita algoritma označavanja. To znači da ćemo od originalnog skupa podataka, koji je imao N uzoraka, stvoriti novi dataset sa $3N$ uzoraka, gdje svaki ulazni podatak X sada ima tri različite verzije izlaznih oznaka Y_1, Y_2, Y_3.

Konkretno, umjesto da imamo:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_N, Y_N) \tag{5.1}$$

Sada ćemo imati:

$$(X_1, Y_{1,1}), (X_1, Y_{1,2}), (X_1, Y_{1,3}), (X_2, Y_{2,1}), (X_2, Y_{2,2}), (X_2, Y_{2,3}), \ldots, (X_N, Y_{N,3}) \tag{5.2}$$

Gdje su $Y_{i,1}$, $Y_{i,2}$ i $Y_{i,3}$ oznake dobivene različitim algoritmima za označavanje trendova.

Ovakvim pristupom povećavamo veličinu dataseta i omogućavamo modelu da uči iz različitih perspektiva označavanja trenda, što može poboljšati njegovu sposobnost gene-

ralizacije.

Model treniran na maloprije objašnjen način potom je testiran na testnom skupu računajući kumulativne povrate dobivene na oznakama dobivenih na izlazu tako treniranog model. Rezultate takvog modela, zajedno sa rezultatima kumulativnih povrata dobivenih od različitih algoritama za označavanje, radi usporedbe, mogu se vidjeti na sljedećoj tablici:

Tablica 5.2. Prosječni kumulativni povrati za različite algoritme označavanja.

Oznake	**Povrat predviđenih oznaka**
ORACLE	-0.00870
FTH	-0.00512
CONTINUOUS_TREND	-0.06931
HYBRID_LABEL	X
HYBRID_ORACLE	-0.01139
MJEŠOVITi PRISTUP	-0.06818

Iz tablice vidimo da su kumulativni povrati dobiveni na navedeni način vrlo slični povratu dobivenom algoritmom za kontinuirano označavanje trendova, no i dalje manji nego povrati FTH i Oracle algoritma.

Kao klasifikacijsku mjeru upješnosti modela koristila se mjera ROC AUC koju smo ispitivali na trening i test skupu podataka. Dobiveni rezultati ROC AUC mjere vide se u tablici ispod:

Tablica 5.3. ROC AUC rezultati za različite oznake na trening i test skupu

Oznaka	Train ROC AUC	Test ROC AUC
FTH_LABEL	0.9027	0.6761
ORACLE_LABEL	0.8283	0.7968
CONTINUOUS_TREND_LABEL	0.8388	0.9424
HYBRID_LABEL	0.8976	0.8099
HYBRID_ORACLE	0.8050	0.7902
MJEŠANI PRISTUP	0.7504	0.6872

Prema tablici vidimo da se model treniran na FTH oznakama prenaučio i da na testnom skupu daje najmanje rezultate. Najboljim modelom za generalizaciju pokazao se model treniran na oznakama od algoritma za kontinuirano označavanje, dok modeli trenirani na hibridnim oznakama se pokazuju najboljima nakon njega.

6. Zaključak

U ovom radu naglašavaju se izazovi prepoznavanja trendova u financijskim vremenskim nizovima, koji su prirodno zašumljeni i podložni velikoj volatilnosti. U radu je primijenjeno nekoliko različitih algoritama za označavanje trendova, Oracle označavanje, označavanje kontinuiranog trenda (CTL) te označavanje trenda s fiksnim vremenskim horizontom (FTH), pri čemu svaki od tih algoritama generira različite oznake ovisno o odabranim parametrima. Cilj rada bio je poboljšati generalizaciju klasifikatora proširivanjem oznaka dobivenih iz različitih postavki algoritama.

Pristup augmentaciji izveden je u dva različita oblika. U prvom pristupu, koristili su se sva tri algoritma za označavanje oznaka (od oracle algoritma, algoritma za kontinuirani trend te algoritma za trend s fiksnim vremenskim horizontom) za zadnju cijenu u minuti, a augmentirana, hibridna oznaka poprimala je pozitivnu vrijednost samo u slučaju kada su sve tri oznake sugerirale rastući trend, inače je poprimila negativnu oznaku. Iako ovaj pristup gubi određenu sekvencijalnost podataka, evaluacija je provedena mjerenjem klasifikacijskih mjera na treniranom skupu te financijskih mjera na testnom skupu. U drugom pristupu, oracle označavanje primijenjeno je na četiri ključne točke u minuti – prvu, najveću, najmanju i zadnju cijenu – čime je kreirana druga hibridna oznaka. Ovdje se rastući trend identificira samo u trenucima kada sve četiri serije pokazuju usklađen rast, čime se postiže veća konzistentnost oznaka.

Rezultati pokazuju da augmentirani modeli, iako značajno poboljšavaju robusnost i klasifikacijsku učinkovitost u odnosu na modele trenirane isključivo na neaugmentiranom skupu podataka, ne nadmašuju model treniran isključivo na oracle oznakama. Ovaj nalaz sugerira da oracle označavanje pruža najkvalitetniju informaciju za treniranje klasifikatora, dok hibridni i miješani pristupi, unatoč svojim prednostima, zahtijevaju daljnju optimizaciju i dodatna istraživanja.

Ovo istraživanje doprinosi boljem razumijevanju utjecaja augmentacije skupa za treni-

ranje na performanse modela u složenom i dinamičnom okruženju financijskih tržišta. Rezultati sugeriraju da bi buduća istraživanja trebala usmjeriti pažnju na razvoj novih metoda i kombinacija augmentacije, kako bi se dodatno unaprijedila preciznost prepoznavanja trendova. Nadalje, potrebna su dodatna eksperimentalna istraživanja za evaluaciju novih ideja u svrhu poboljšanja prediktivne moći klasifikatora, što bi u konačnici moglo dovesti do robusnijih modela za predviđanje kretanja na financijskim tržištima.

Literatura

[1] Y. Lin, H. Guo, i J. Hu, "An svm-based approach for stock market trend prediction", u *2013 International Joint Conference on Neural Networks, IJCNN 2013*, ser. Proceedings of the International Joint Conference on Neural Networks, prosinac 2013., 2013 International Joint Conference on Neural Networks, IJCNN 2013 ; Conference date: 04-08-2013 Through 09-08-2013. https://doi.org/10.1109/IJCNN.2013.6706743

[2] M. Prata, G. Masi, L. Berti *et al.*, "Lob-based deep learning models for stock price trend prediction: A benchmark study", *Artificial Intelligence Review*, sv. 57, str. 116, 2024. https://doi.org/10.1007/s10462-024-10715-4

[3] E. Alpaydin, *Introduction to Machine Learning, fourth edition*, ser. Adaptive Computation and Machine Learning series. MIT Press, 2020. [Mrežno]. Adresa: https://books.google.hr/books?id=uZnSDwAAQBAJ

[4] M. Dixon, I. Halperin, i P. Bilokon, *Machine Learning in Finance*. Springer, 2020. [Mrežno]. Adresa: https://link.springer.com/book/10.1007/978-3-030-27461-0

[5] S. J. T. B. W. S. Wu D, Wang X, "A labeling method for financial time series prediction based on trends", *Entropy*, sv. 22, br. 10, str. 1162, 2020. https://doi.org/10.3390/e22101162

[6] T. Kovačević, A. Merćep, S. Begušić, i Z. Kostanjčar, "Optimal trend labeling in financial time series", *IEEE Access*, sv. 11, str. 83 822–83 832, 2023. https://doi.org/10.1109/ACCESS.2023.3303283

[7] C. Koonitz, *Technical Analysis for Beginners: A Practical Guide for Charting*

(second edition). Tripod Solutions inc., 2018. [Mrežno]. Adresa: https://books.google.hr/books?id=Ahj_EAAAQBAJ

[8] Investopedia, "Simple moving average (sma)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/s/sma.asp

[9] ——, "Volatility", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/v/volatility.asp

[10] ——, "Exponential moving average (ema)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/e/ema.asp

[11] ——, "Moving average convergence divergence (macd)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/m/macd.asp

[12] ——, "Average directional index (adx)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/a/adx.asp

[13] ——, "Relative strength index (rsi)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/r/rsi.asp

[14] T. Turner i P. Headley, *Short-Term Trading in the New Stock Market*. St. Martin's Publishing Group, 2006. [Mrežno]. Adresa: https://books.google.hr/books?id=zLfxpidUSnIC

[15] Investopedia, "On-balance volume (obv)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/o/onbalancevolume.asp

[16] ——, "Mark to market (mtm)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/m/marktomarket.asp

[17] ——, "Average true range (atr)", 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/a/atr.asp

[18] T. Edits, *70+ Technical Indicators - Mastering Intraday Trading*. by Mocktime Publication, 2023. [Mrežno]. Adresa: https://books.google.hr/books?id=U_WrEAAAQBAJ

[19] Investopedia, “Money flow index (mfi)”, 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/m/mfi.asp

[20] D. Wu, X. Wang, J. Su, B. Tang, i S. Wu, “A labeling method for financial time series prediction based on trends”, *Entropy*, sv. 22, str. 1162, 10 2020. https://doi.org/10.3390/e22101162

[21] J. J. Murphy, *Technical Analysis of the Futures Markets*. Hoboken, NJ: Wiley, 2019. [Mrežno]. Adresa: https://py98.ir/blog/wp-content/uploads/2019/08/Technical-Analysis-of-the-Futures-Markets-John-J.-Murphy.pdf

[22] M. Aamir, “Introduction to trend line”, 2021. [Mrežno]. Adresa: https://www.linkedin.com/pulse/introduction-trend-line-muhammad-aamir-crypto-analyst-wzqwc

[23] M. de Prado, *Advances in Financial Machine Learning*. Wiley, 2018. [Mrežno]. Adresa: https://books.google.hr/books?id=oU9KDwAAQBAJ

[24] G. Colab, “Google colab”, 2021. [Mrežno]. Adresa: https://colab.research.google.com/

[25] NumPy, “Numpy”, 2021. [Mrežno]. Adresa: https://numpy.org/

[26] Pandas, “Pandas”, 2021. [Mrežno]. Adresa: https://pandas.pydata.org/

[27] Matplotlib, “Matplotlib”, 2021. [Mrežno]. Adresa: https://matplotlib.org/

[28] P. TA, “Pandas ta”, 2021. [Mrežno]. Adresa: https://github.com/twopirllc/pandas-ta

[29] XGBoost, “Xgboost python api”, 2021. [Mrežno]. Adresa: https://xgboost.readthedocs.io/en/stable/python/

[30] Scikit-learn, “Scikit-learn”, 2021. [Mrežno]. Adresa: https://scikit-learn.org/stable/

[31] Investopedia, “Futures”, 2021. [Mrežno]. Adresa: https://www.investopedia.com/terms/f/futures.asp

[32] Skender, “Diplomski rad”, https://github.com/sven8599/Diplomski-rad/blob/main/Diplomski_Skender_FINAL.ipynb, 2025.

[33] F. Yid, "'ffill' and 'bfill' in data cleaning process", 2021. [Mrežno]. Adresa: https://medium.com/@farisyid/penggunaan-ffill-dan-bfill-pada-proses-data-cleaning-b4f3bfec9767

[34] Pandas, "pandas.dataframe.ffill", 2021. [Mrežno]. Adresa: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.ffill.html

[35] ——, "pandas.dataframe.bfill", 2021. [Mrežno]. Adresa: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.bfill.html

[36] D. X. Team, "Xgboost github repository", 2024. [Mrežno]. Adresa: https://github.com/dmlc/xgboost

[37] T. Chen i C. Guestrin, "Xgboost: A scalable tree boosting system", u *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, kolovoz 2016., str. 785–794. https://doi.org/10.1145/2939672.2939785

[38] X. Developers, "Xgboost parameter documentation", 2024. [Mrežno]. Adresa: https://xgboost.readthedocs.io/en/stable/parameter.html#general-parameters

[39] A. L. Guennec, S. Malinowski, i R. Tavenard, "Data augmentation for time series classification using convolutional neural networks", u *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, Riva Del Garda, Italy, September 2016.

[40] F. o. E. E. University of Zagreb i Computing, "Osnovni koncepti", 2022. [Mrežno]. Adresa: https://www.fer.unizg.hr/-download/repository/SU1-2022-P02-OsnovniKoncepti.pdf

[41] S. Chatterjee, "Cross validation in time series", 2021. [Mrežno]. Adresa: https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4

[42] F. o. E. E. University of Zagreb i Computing, "Vrednovanje modela", 2022. [Mrežno]. Adresa: https://www.fer.unizg.hr/-download/repository/SU1-2022-P21-VrednovanjeModela.pdf

Sažetak

Smanjenje generalizacijske pogreške klasifikatora kroz augmentaciju trendova u financijskim vremenskim nizovima

Sven Skender

Predviđanje trendova cijena financijske imovine često je postavljeno kao klasifikacijski problem, gdje se trendovi označavaju kao pozitivni ili negativni. Zbog zašumljenosti financijskih vremenskih nizova, teško je razlikovati stvarne trendove od kratkoročnih fluktuacija. U literaturi je predloženo nekoliko definicija trendova, te je istraženo kako specifična definicija utječe na performanse klasifikatora koji uče tako definirane trendove na temelju povijesnih podataka. Međutim, do sada nije bilo poznato kako augmentacija skupa za učenje utječe na performanse tih klasifikatora. U ovom radu, kako bi se poboljšala generalizacija klasifikatora pri predviđanju budućih trendova, skup za učenje nadopunjuje se oznakama dobivenim pomoću tri različita algoritma označavanja: Oracle algoritma označavanja, algoritma za označavanje kontinuiranog trenda (CTL) te algoritma za označavanje trenda s fiksnim vremenskim horizontom (FTH). Rezultati pokazuju da augmentirani modeli značajno poboljšavaju robusnost i klasifikacijsku učinkovitost u odnosu na modele trenirane isključivo na neaugmentiranom skupu podataka (dobivenim CTL i FTH algoritmima), no ne nadmašuju modele trenirane isključivo na Oracle oznakama.

Ključne riječi: financijski vremenski nizovi; strojno učenje; algoritma idenfitikacije trenda; augmentacija

Abstract

Reducing classifier generalization error with trend augmentation in financial time series

Sven Skender

Predicting asset price trends is often approached as a classification task, where trends are labeled as either positive or negative. Due to the inherent noise and volatility in asset price series, distinguishing true trends from short-term fluctuations is challenging. In the literature, various trend definitions have been proposed, and their impact on the performance of classifiers trained on historical data has been explored. However, the effect of training data augmentation on these classifiers has remained unclear. In this thesis, the training set is augmented using labels obtained from three different labeling algorithms—namely the Oracle labeling algorithm, the Continuous Trend Labeling (CTL) algorithm, and the Fixed Time Horizon (FTH) trend labeling algorithm—to improve classifier generalization when predicting future trends. The results demonstrate that while the augmented models significantly enhance robustness and classification performance compared to models trained solely on unaugmented data (using CTL and FTH algorithms), they do not outperform the model trained exclusively on Oracle labels.

Keywords: financial time series; machine learning; trend labeling algorithm; augmentation