Gradnja napovednih modelov za klike na oglase v družabnih omrežjih

Vesna Novak, Matej Guid

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
E-pošta: vesna.n@gmail.com, matej.guid@fri.uni-lj.si

Building prediction models for advertisement clicks in social networks

The revenues from large social networks, which today are measured in billions of dollars and increasing year by year, mostly come from online advertising. Advertising on social networks is becoming an increasingly important form of advertising for products and services. A typical social networking ad campaign contains multiple ad groups and each ad group contains multiple ads. When managing advertising campaigns, it is particularly important to decide which ad groups and ads are to be active and for which more advertising money should be earmarked. We compared different prediction models for predicting clicks on ads on social networks to optimize the management of display advertising campaigns. With optimization, decision support for campaign managers and automation of display campaign management can be improved, maximizing the number of clicks.

1 Uvod

Družabna omrežja so internetna spletna mesta družabnih medijev, ki jih uporabljamo z namenom povezovanja s prijatelji, družino, sodelavci ali strankami. Kar 45% celotne svetovne populacije uporablja vsaj eno družabno omrežje, število uporabnikov pa iz dneva v dan strmo narašča [1]. Ker lahko ima družabno mreženje poleg socialnega tudi poslovni namen ni nenavadno, da je postalo pomembna osnova za upravljalce oglaševalskih kampanj. Oglaševanje na družabnih omrežjih je v porastu, namenja se mu vedno več oglaševalskega denarja. Slovenski oglaševalci so v letu 2017 digitalnemu oglaševanju namenili 47,2 milijona evrov, ocenjuje pa se, da bo v letu 2019 obseg investicij v digitalno oglaševanje občutno presegel 60 milijonov evrov. 82% oglaševalcev bo oglaševalo na družabnih omrežjih, med oblikami digitalnega oglaševanja pa naj bi letos največ sredstev namenili ravno prikaznemu oglaševanju [2].

Pri oglaševanju se sprašujemo kaj je namen oglaševanja, kdo je ciljna skupina, kakšno bo sporočilo, katere kanale za oglaševanje uporabiti, koliko denarja nameniti določenemu oglasu, kako meriti in oceniti rezultate [3]. Pri vseh teh vprašanjih bi bilo koristno vnaprej vedeti, če sploh in koliko klikov na oglas pričakovati. S tem bi se upravljalcu oglaševalskih kampanj olajšala odločitev, katere oglase ali oglasne skupine pustiti aktivne in koliko denarja jim nameniti. S pomočjo strojnega učenja na preteklih podatkih poskušamo zgraditi napovedni model, ki z napovedovanjem časovnih vrst kar se da uspešno napove prihodnost klikov na oglas in s tem upravljalcu oglaševalske kampanje pomaga pri upravljalskih odločitvah, prav tako pa je pomembna informacija pri samodejnem upravljanju oglaševalskih kampanj.

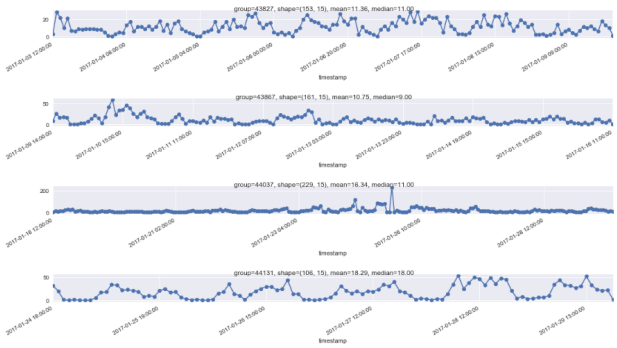
Na realnih podatkih iz družabnih omrežij Facebook in Twitter, ki so agregirani na urnem nivoju in po skupinah oglasov, smo primerjali različne algoritme za napovedovanje časovnih vrst (angl. *time series forecasting*): XGBoost, ARIMA, VAR in LSTM nevronske mreže ter ugotavljali njihovo zmožnost za napovedovanje več časovnih točk (ur) vnaprej.

Gradnja napovednih modelov za napovedovanje dogodkov na družabnih omrežjih, še zlasti v kontekstu oglaševanja, je dandanes izredno aktualna tema, vendar je akademskih člankov s tega področja relativno malo. V članku [4], kjer napovedujejo klike na oglase na družabnem omrežju Facebook so izpostavili, da na končno napoved najbolj vpliva pravilen izbor informacij, torej tistih, ki predstavljajo zgodovinske podatke o uporabniku ali samem oglasu. Tudi v članku [5], kjer napovedujejo razmerje med prikazi in kliki (CTR)[1] na družabnem omrežju Twitter, za bolj uspešne napovedi izpostavljajo pomembnost značilk, povezanih z uporabnikom, njegovimi meta podatki in zgodovino njegovih klikov. Naš nabor podatkov teh informacij ne zajema. S preizkušanjem in primerjavo bomo skušali ugotoviti, kateri parametri in koliko izmed teh, ki jih imamo na voljo, najbolj vplivajo na napovedno točnost pri napovedovanju klikov na oglase, za različna časovna obdobja. S tem se veliko bolj približamo perspektivi upravljalca oglaševalskih akcij, ki ima tipično na voljo statistike oglasov v preteklih dneh, sprejeti pa mora odločitev, katerim oglasom bo v prihodnje namenil več oz. manj oglaševalskega denarja. Večina sorodnih del se osredotoča na napovedovanje samo ene točke vnaprej [6], medtem ko je naš namen ugotoviti, za koliko ur vnaprej so napovedi še smiselne [7]. Prav tako želimo primerjati različne modele, tako univariatne kot multivariatne, in ugotoviti, če in katere zunanje značilke lahko pripomorejo k boljši napovedni točnosti.

[1] Razmerje, ki pove, kako pogosto uporabniki, ki vidijo oglas, tega na koncu tudi kliknejo.

	Povprečje	Min	25%	50%	Max
Facebook	23,01	0	6	15	341
Twitter	10,91	0	0	1	1223

Tabela 1: Statistični podatki klikov.



Slika 1: Prikaz nekaj časovnih vrst, ki prikazujejo število klikov na oglase v družabnem omrežju Facebook.

2 Metode

Podatki, ki smo jih uporabili v tem delu, obsegajo tipične najpomembejše statistike oglaševalskih akcij na družabnem omrežju Facebook (clicks, actions, amount, impressions, link clicks, conversions, page likes, engagements, social clicks, social impressions, video views) in Twitter (clicks, impressions, conversions, engagements, retweets, replies, likes, app clicks, url clicks, follows in 9 atributov povezanih z ogledi videov). Podatki ne zajemajo informacij o samem uporabniku ali oglasu in so agregirani na urni nivo, brez vmesnih manjkajočih ur. Facebook podatki obsegajo 60 časovnih vrst klikov različnih dolžin – med 63 in 305 zaporednih časovnih točk, medtem ko Twitter podatki obsegajo 36 časovnih vrst klikov, tipično precej daljših, kot pri Facebook podatkih - med 281 in 1127 zaporednih časovnih točk. Skupno imamo torej na voljo 96 časovnih vrst različnih dolžin. Statistične podatke klikov obeh družabnih omrežij prikazuje tabela 1. Primer nekaj časovnih vrst, ki prikazujejo število klikov na oglase iste oglasne skupine je razviden iz slike 1.

Za napovedovanje časovnih vrst uporabljamo programski jezik Python in knjižnice *NumPy*, *pandas*, *PyFlux*, *TensorFlow* in *Keras*. V nadaljevanju so predstavljeni algoritmi oz. metode, ki smo jih uporabili za gradnjo napovednih modelov.

2.1 ARIMA

Model ARIMA je splošen model za napovedovanje in eden izmed najpogosteje uporabljenih metod za analizo časovnih vrst. Obstaja veliko različnih ARIMA modelov, osnovni pa je poznan kot ARIMA(p, d, q) in je sestavljen iz avtoregresijskega dela reda p (AR), diferenciranja stopnje d (I) in modela drsečih sredin reda q (MA). Vsi ARIMA modeli temeljijo na univariatnih napovedih, uporabljajo torej podatke le ene časovne vrste in ne vključujejo podatkov soležnih časovnih vrst, ki pogosto lahko vsebujejo pomembne informacije za bolj natančne napovedi. Uporabili smo model ARIMA iz knjižnice PyFlux, ki s funkcijo *predict(h)* omogoča preprosto napovedovanje *h* točk v prihodnost [8].

2.2 VAR

Model vektorske avtoregresije (*angl. vector autoregression, VAR*) je eden izmed bolj uspešnih, fleksibilnih in enostavnih modelov za uporabo pri analizi multivariatnih časovnih vrst [9]. Je razširitev univariatnega avtoregresivnega modela v dinamično multivariatno časovno vrsto. Model VAR se je kot posebej uporaben izkazal za opisovanje dinamičnega obnašanja pri ekonomskih in finančnih časovnih vrstah [10]. Uporabili smo model VAR iz knjižnice StatsModels.

2.3 XGBoost

XGBoost (angl. eXtreme Gradient Boosting) je metoda, ki omogoča vzporedno gradnjo večjega števila odločitvenih dreves na eni sami napravi, kar ji omogoča tudi do več kot 10-krat višjo časovno učinkovitost od obstoječih implementacij Gradient Boosting metode [11]. Z algoritmom XGBoost smo ugotavljali pomembnost zunanjih značilk. Po pričakovanjih sta poleg klikov v prejšnjih časovnih točkah najpomembnejši značilki “impressionns“ in “social_clicks“.

2.4 LSTM

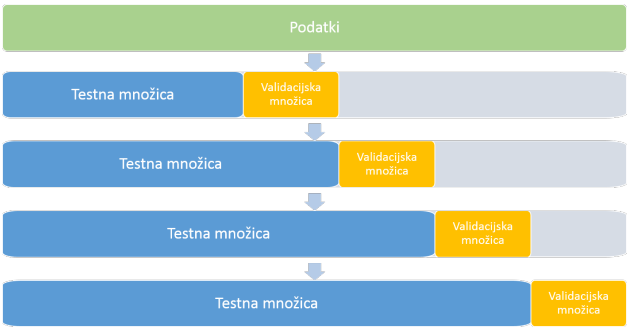
Za napovedovanje časovnih vrst se vse bolj uporabljajo povratne nevronske mreže (angl. *Recurrent neural networks, RNN*), predvsem njihova različica nevronske mreže z dolgim kratkoročnim spominom (angl. *Long Short-Term memory neural networks, LSTM*). Uporabljajo se za multivariatne napovedi pri napovedovanju časovnih vrst [12]. LSTM nevronske mreže so dandanes sposobne preseči univariatne modele za napovedovanje časovnih vrst [13], zato jih želimo poleg klasičnih modelov preizkusiti tudi v naši raziskavi.

2.5 Vztrajnostni model

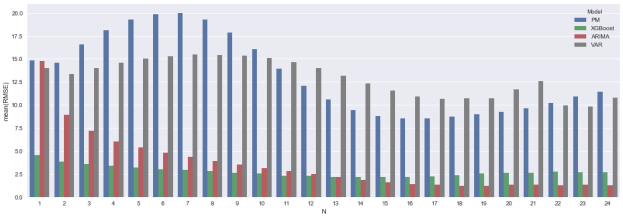
Za ovrednotenje modelov in določitev uspešnosti napovednih vrednosti uporabljamo urni vztrajnostni model, kjer predpostavimo, da bo število klikov v naslednji uri enako številu klikov v prejšnji uri.

3 Izvedba poskusov

Da bi dobili čim bolj primerljive rezultate med modeli, se pri vseh modelih učimo na isti učni množici in napovedujemo na isti testni množici za vse časovne vrste. Za napovedovanje uporabljamo sprehod naprej (angl. *walk-forward validation*), prikazan na sliki 2, kjer se pri napovedovanju v vsakem koraku učna množica povečuje za eno točko, testna pa se za eno točko pomanjša, ne glede na to, koliko točk vnaprej napovedujemo. Korakov je toliko, kolikor je dolga testna množica, zmanjšana za 24 ur. Končamo takrat, ko je seštevek učne množice, koraka in števila želenih napovednih točk večji ali enak velikosti časovne vrste.



Slika 2: Sprehod naprej.



Slika 3: Povprečni RMSE za vseh 60 skupin oglasov družabnega omrežja Facebook, za napovedi od 1 do 24 napovednih točk vnaprej.

Za ocenjevanje napovedne točnosti uporabljamo koren povprečne kvadratne napake (angl. *root mean squared error*, RMSE). Napovedujemo od 1 do 24 točk vnaprej. V primeru napovedovanja ene točke vnaprej dobimo 24 vrednosti, ki jih nato uporabimo za izračun RMSE. V primeru napovedovanja 24 ur vnaprej dobimo samo 1 vrednost, ta pa predstavlja seštevek vseh 24 napovedanih vrednosti. Testna množica pri napovedovanju je dolga 24 točk (en dan), dolžina učne množice pa je odvisna od velikosti časovne vrste. Zmanjšana je le za velikost testne množice. Psevdokoda 1 prikazuje postopek obdelave podatkov in napovedovanja prihodnjih vrednosti. Postopek je enak za vse modele, le pri LSTM smo zaradi stohastične lastnosti nevronskih mrež model testirali deset krat za isto napoved in pri tem uporabili povprečne vrednosti.

Psevdokoda 1 Obdelava podatkov in napovedovanje

```
for model ∈ {vsi modeli} do
    Priprava prejetih podatkov
    for obdelava ∈ {odstranitev trenda, normalizacija, standardizacija, logaritemska preslikava} do
        Obdelava podatkov
        Iskanje parametrov z mrežnim iskanjem
        Shranjevanje najustreznejših parametrov
        Razdelitev na učno in testno množico
        for vsak T iz testne množice do
            Učenje na učni množici
            Napovedovanje testne množice
            Inverz obdelave / vrnitev trenda, sezonskosti
            Izračun RMSE
            Beleženje rezultatov
        end for
    end for
end for
```

3.1 Optimizacija parametrov in obdelava podatkov

Preden se časovno vrsto lahko uporabi na modelih za napovedovanje, jo je potrebno ustrezno pripraviti in poiskati ustrezne parametre, ki jih model zahteva. Ker večina uporabljenih algoritmov predpostavlja, da je časovna vrsta stacionarna, ji s predhodno obdelavo odstranimo trend in sezonskost. Predhodno časovno vrsto obdelalamo še s standardizacijo, normalizacijo in logaritemsko preslikavo ter primerjamo rezulate za različne kombinacije parametrov in obdelav.

Pri multivariatnih modelih zunanje značilke poiščemo s pomočjo matrike intenzitete (angl. *heat map*), ki prikazuje, kako močno so atributi med seboj korelirani. Pri ARIMA modelu najustreznejše parametre p, d in q poiščemo z mrežnim iskanjem (angl. *grid search*). Ravno tako za iskanje zamika in kriterijske funkcije pri modelu VAR uporabimo mrežno iskanje.

4 Rezultati

Ker smo poskuse izvajali na dveh podmnožicah časovnih vrst, za vsako družabno omrežje posebej, rezultate predstavljamo v ločenih podpoglavjih. Prav tako smo ločeno izvajali poskuse z LSTM nevronskimi mrežami, zato so primerjave LSTM nevronskih mrež z ostalimi modeli predstavljene posebej.

4.1 Facebook

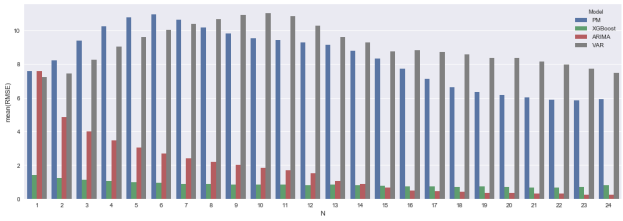
Primerjava modelov XGBoost, ARIMA in VAR za napovedi od 1 do 24 časovnih točk vnaprej je prikazana s sliko 3. Slika prikazuje izračunane povprečne RMSE, za vseh 60 skupin oglasov družabnega omrežja Facebook. Poskusi so pokazali dobro delovanje modela XGBoost, sledi mu model ARIMA. Rezultati nakazujejo, da se pri napovedovanju nekaj točk vnaprej multivariatni modeli XGBoost odlično obnesejo, vendar pa z oddaljenostjo dogodkov, ki so predmet napovedovanja, napovedna točnost XGBoost modelov upada v primerjavi z univariatnimi modeli ARIMA.

4.2 Twitter

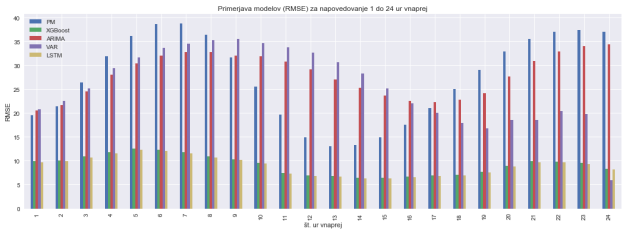
Časovne vrste družabnega omrežja Twitter so daljše, posledično so daljše učne množice, kar je pri večini napovednih modelov prednost. Tudi poskusi na časovnih vrstah družabnega omrežja Twitter so pokazali najboljše delovanje modela XGBoost. Sledi mu model ARIMA. Tudi pri časovnih vrstah družabnega omrežja Twitter se izkaže, da pri napovedih 13 časovnih točk in več vnaprej modela ARIMA preseže uspešnost modela XGBoost.

4.3 Primerjava z LSTM nevronskimi mrežami

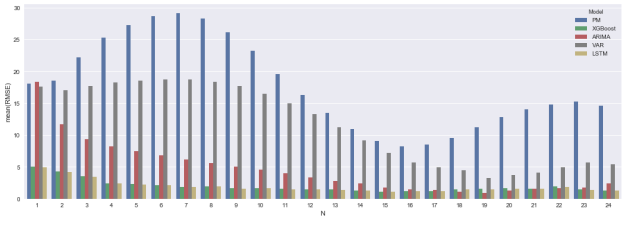
LSTM nevronske mreže dosegajo najboljšo napovedno točnost, ne glede na velikost napovednega okna, njihova slabost pa je v časovni potratnosti. Za napoved, denimo



Slika 4: Povprečni RMSE za vseh 36 skupin oglasov družabnega omrežja Twitter, za napovedi od 1 do 24 napovednih točk vnaprej.



Slika 5: Primerjava vseh modelov za napovedi od 1 do 24 točk vnaprej za eno časovno vrsto.



Slika 6: Povprečni RMSE za 3 skupine oglasov družabnega omrežja Facebook, za napovedi od 1 do 24 napovednih točk vnaprej.

ene točke vnaprej, je model povprečno potreboval 448 sekund, medtem ko so ostali modeli napoved izvedli povprečno v 3 sekundah. Primer rezultatov na oglasni skupini, kjer smo primerjali uspešnost LSTM mrež z ostalimi metodami, je prikazan na sliki 5, medtem ko slika 6 s povprečnimi RMSE prikazuje primerjavo LSTM nevronskih mrež z ostalimi modeli za tri skupine oglasov. LSTM nevronske mreže so bile v povprečju boljše za 1,98% v primerjavi z XGBoost modelom.

5 Zaključek

Z gradnjo napovednih modelov smo primerjali več metod za napovedovanje klikov na oglase v družabnih omrežjih. Osredotočili smo se na družabni omrežji Facebook in Twitter, za kateri smo imeli na voljo realne podatke. S kriterijsko funkcijo RMSE smo pokazali, da so napovedi v skladu s pričakovanji. Sledijo si v pričakovanem vrstnem redu: model LSTM za multivariatne napovedi dosega najboljšo napovedno točnost, sledijo mu regresijski model XGBoost, model ARIMA in model VAR v navedenem vrstnem redu. Zanimali so nas tudi rezultati napovedi več časovnih točk vnaprej. Opažamo, da uporabljene metode, še zlasti LSTM nevronske mreže in XGBoost v uporabljeni domeni omogočajo smiselne napovedi do 24 ur vnaprej. Slabost LSTM nevronskih mrež je v njihovi časovni potratnosti. Po izračunih je čas izvajanja LSTM nevronskih mrež 150 krat počasnejši v primerjavi z ostalimi metodami.

Literatura

[1] Christina Newberry. 130 Social Media Statistics that Matter to Marketers in 2019. `https://blog.hootsuite.com/social-media-statistics-for-social-media-managers/`, Mar 2019.

[2] Tanja Fon. V sloveniji investicije v digitalno oglaševanje v letošnjem letu višje za 25 odstotkov. `https://iprom.si/v-sloveniji-investicije-v-digitalno-oglasevanje-v-letosnjem-letu-visje-za-25-odstotkov/`, Mar 2019.

[3] P Barwise. Harvard Business Essentials: Marketers Toolkit, 2006.

[4] Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. Practical lessons from predicting clicks on ads at Facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9. ACM, 2014.

[5] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1959–1968. ACM, 2015.

[6] Shuai Yuan, Jun Wang, and Xiaoxue Zhao. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, page 3. ACM, 2013.

[7] Jason Brownlee. 4 Strategies for Multi-Step Time Series Forecasting - Machine Learning Masteryn. `https://machinelearningmastery.com/multi-step-time-series-forecasting/`, 2017.

[8] G Box and G Jenkins. Time Series Analysis-Forecasting and Control. San Francisco: Holden Day. 553 p. 1970.

[9] Helmut Lütkepohl, Markus Krätzig, and Peter CB Phillips. *Applied time series econometrics*. Cambridge university press, 2004.

[10] Eric Zivot and Jiahui Wang. Vector autoregressive models for multivariate time series. *Modeling Financial Time Series with S-PLUS®*, pages 385–429, 2006.

[11] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, and Yuan Tang. Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4, 2015.

[12] Jason Brownlee. *Deep Learning for Time Series Forecasting*. 2018.

[13] Kasun Bandara, Christoph Bergmeir, and Slawek Smyl. Forecasting Across Time Series Databases using Long Short-Term Memory Networks on Groups of Similar Series. *arXiv preprint arXiv:1710.03222*, 2017.