Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.896
DOI: https://doi.org/10.24867/07BE10Becejac

PREDIKCIJA CENE PROIZVODA NA OSNOVU OPISA NJEGOVIH KARAKTERISTIKA

PRODUCT PRICE PREDICTION BASED ON DESCRIPTION OF ITS CHARACTERISTICS

Milana Bečejac, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Tema ovog rada je rešavanje problema predikcije cene proizvoda na osnovu opisa njegovih karakteristika. Specificirani su, implementirani i verifikovani modeli, koji rešavaju ovaj problem. Namena ovih modela je pomoć kupcima u proceni da li je cena proizvoda adekvatna, kao i pomoć prodavcima u određivanju prikladne cene za svoj proizvod.*

Ključne reči: *Predikcija cene, Neuronske mreže, Karakteristike proizvoda, Word2Vector, BagOfWords, LSTM*

Abstract – *The goal of this paper is solving the problem of prediction of price based on products' characteristics. It describes the specification, implementation, and verification of such a system. The purpose of this system is to help buyers find a product with a fair price and to assist sellers in determining a suitable price for their products.*

Keywords: *Price Prediction Neural networks, Product characteristics, Word2Vector, BagOfWords, LSTM*

1. UVOD

U današnje vreme može biti veoma teško odrediti da li neki proizvod na tržištu stvarno vredi toliko kolika mu je cena i detalji mogu dovesti do velike razlike. Na primer, na cenu odeće mogu uticati sezona prodaje ili brend, dok na elektroniku najviše utiče specifikacija proizvoda [1]. Prodavcima je teško da odrede cenu svom proizvodu i da ta cena bude odgovarajuća kupcima. Da bi odredili cenu, prodavci mogu istraživati tržište i tražiti slične proizvode ili pitati druge trgovce za sugestije. Međutim, ove metode oduzimaju puno truda i vremena i mogu dovesti do troškova koji su veći od same vrednosti proizvoda [1].

S druge strane, zbog razvoja tehnologije, sve više je zastupljena kupovina preko interneta i iz tog razloga mnoge *online* prodavnice istražuju načine da kupcima predlože svoje proizvode kroz svoje sajtove koristeći sisteme za preporuku proizvoda [2]. Takođe, pretragom sajtova kupci mogu i samostalno da uporede cene različitih proizvođača za isti proizvod i na taj način odluče gde da izvrše kupovinu.

Ovaj način odlučivanja o kupovini može biti neefikasan ako kupac ne uloži dovoljno vremena u traženju najadekvatnije ponude. Uvođenje adekvatnog metoda za procenu cene proizvoda bi sa jedne strane pomoglo potencijalnim kupcima tako što bi im potvrdio da je zahtevana cena prikladna, dok bi prodavcima pomogla pri odabiru odgovarajuće cene.

Na sajtu *kaggle.com* [3] postoji takmičenje pod nazivom "Mercari Price Suggestion Challenge" i u ovom radu će biti predstavljeno više modela za rešavanje problema predikcije cene proizvoda na osnovu opisa njegovih karakteristika: opisa proizvoda u tekstualnoj formi, naziva brenda, stanja i kategorije proizvoda. Sva rešenja za objektivno određivanje cene proizvoda u ovom radu su realizovana kao modeli za predviđanje cene obučeni na osnovu podataka preuzetog sa *kaggle.com* sajta [3]. Pri implementaciji modela, velika pažnja je posvećena analizi i obradi skupa podataka. Za pretprocesiranje korišćene su različite tehnike obrade teksta kako bi se dobili što adekvatniji rezultati.

Na Kaggle-ovom takmičenju zahtevano je korišćenje RMSLE i iz tog razloga, ovaj metod je korićen kao glavna metrika evaluacije rezultata sitema opisanog u ovom radu. Skup podataka je podeljen na 70% trening, 20% test i 10% validacioni skup podataka.

2. METODOLOGIJA

Sva programska rešenja implementirana za potrebe rada su izrađena u programskom jeziku *Python*.
Za potrebe ovog rada je obučeno 14 modela nad skupovima podataka različitih veličina u cilju utvrđivanja balansa između ostvarenih performansi i vremena utrošenog na treniranje. Pre pokretanja svakog modela izvršeno je predprocesiranje skupa podataka

2.1. Predprocesiranje skupa podataka

Nakon analize podataka, odlučeno je da nad skupom podataka, koji će biti korišćen za treniranje modela bude izvršeno pretprocesiranje kako bi se dobili što bolji rezultati predikcije cene proizvoda.
Metode obrade teksta koji su izvršeni nad obeležjima naziv proizvoda i opis proizvoda kao što je predloženo u radu [4]:

- Pretvaranje svih slova reči iz velikih u mala slova
- Uklanjanje znakova interpukcije
- Uklanjanje stop-reči reči koje se smatraju nebitnim u engleskom jeziku, kao što su: and, the, for, a, in, to,...
- Pretvaranje svih reči u koren (*stem*) te reči
- Dopuna nedostajućih vrednosti, zamena sa stringom "missing". Dopuna je izvršena nad tekstualnim obeležjima opis, brend i kategorija proizvoda.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bila dr Jelena Slivka, docent.

Nad obeležjima brend i kategorija proizvoda izvršene su sve prethodne metode pretprocesiranja, kao i pretvaranje u numeričke vrednosti *SKLearn One-hot* encoding metodom.

2.2. Inicijalni RNN model

Inicijalni model je bila rekurentna neuronska mreža, koja je kao ulaz primala skup podataka, nad kojim je izvršeno samo pretvaranje kategoričkih obeležja brenda i kategorije proizvoda u numeričke vrednosti tih obeležja, kao što je opisano u prethodnom poglavlju. Rekurentna neuronska mreža (RNN) je tip veštačke neuronske mreže koja se obično koristi u obradi prirodnog jezika. RNN je dizajnirana tako da prepoznaje sekvencijalne karakteristike podataka i da koristi paterne za predviđanje najverovatnijeg sledećeg scenarija [5]. Naš inicijalni model je imao *Embedding* slojeve, po jedan za opis, naziv, brend, kategoriju i stanje proizvoda, 2 GRU (*Gated Recurrent Unit*) sloja, a za aktivaciju je korišćena linearna funkcija. GRU je varijanta LSTM-a, koja takođe rešava problem nestajućeg gradijenta, ali je jednostavnija i brža od LSTM-a (*Long Short Term Memory*) [6]. Ova arhitektura je dizajnirana po uzoru na rad [7]. Kao mera za testiranje ovog modela korišćen je RMSLE.

2.3. *Random Forest* algoritam

U radu je isproban i *Random Forest* algoritam, nad skupovima podataka različitih veličina i sa različitim vrednostima parametara. Performanse ovog algoritma testirane su pomoću mere tačnosti MAE (*Mean Absolute Error*). U radu [8] je predložena ova mera performansi za *Random Forest* algoritam i, nakon analize rezultata, utvrđeno je da je tačno tvrđenje iz rada [8] i da *Random Forest* daje loše rezultate pri rešavanju predikcije cena proizvoda. Iz tog razloga odlučeno je da se pažnja posveti rešavanju problema korišćenjem modela neuronske mreže, analizi i pretprocesiranju skupa podataka.

2.4. *BagOfWords* i *Word2Vector*

Većina modela opisanih u ovom radu vrši predikciju samo na osnovu opisa proizvoda u tekstualnom obliku, ali su isprobani i modeli koji u obzir uzimaju i stanje, status, kategoriju i brend, kako bi se utvrdilo da li se dodavanjem ostalih obeležja može poboljšati rezultat.

U narednim modelima, opisi svakog proizvoda iskazani su kao vreću reči (eng. *Bag of words*) ili vektore reči (eng. *Word2Vector*) i njihova reprezentacija predstavljala je ulaze u neuronsku mrežu.

Sledeće unapređenje modela bilo je to da *Word2Vector* reprezentacija opisa proizvoda predstavlja ulaz za LSTM sloj neuronske mreže. Modeli sa *softmax* aktivacionom funkcijom sortiraju proizvode u 12 ili 20 različitih kateogrija cena, kao što je urađeno i u radu [9], dok modeli sa linearnom aktivacionom funkcijom pokušavaju da pogode tačnu cenu za svaki testni proizvod. Kod modela sa *softmax* aktivaciom, kategorije su napravljene tako da su relativno balansirane. Za evaluaciju ovih modela korišćena je tačnost (*Accuracy*) kada je u pitanju *softmax* aktivacija, a *Percent error* i RMSLE kada je u pitanju linearna aktivaciona funkcija.

Nakon pretprocesiranja skupa podataka opisanog u prethodnom poglavlju, izvršeno je enkodiranje obeležja, tako da budu pogodni za slanje na ulazni sloj novih modela. U početku je najveći fokus bio na opisu proizvoda, jer je bila pretpostavka da će on najviše uticati na rezultate predikcije. Opis svakog proizvoda predstavljen je na dva različita načina.

Za implementaciju "vreće reči" napravljen je vektor, veličine rečnika svih različitih reči koje se pojavljuju u koloni opis proizvoda iz skupa podataka. Zatim je svaki opis predstavljen kao vektor veličine jednakoj broju svih različitih reči koje se pojavljuju u svim opisima. Za svaki vektor je zabeležemo prisustvo te reči u odgovarajućem opisu, 1 ako je reč prisutna, 0 ako nije.

2.5. Jednostavna troslojna neuronska mreža

Model 1 je jednostavna troslojna neuronska mreža kojoj je kao ulaz prosleđena prethodno opisana vreća reči. Ovaj model na izlazu ima *Softmax* aktivacionu funkciju i zbog toga mera je za evaluaciju ovog modela tačnost modela (*Accuracy*). Cilj modela je da svaki proizvod svrsta u jednu od 12 kategorija cena. Ovaj model je isproban na malom trening skupu podataka od 2.000 primera i sa malim brojem epoha i dobijena je tačnost od 23,7% nad trening skupom i 16,5% nad test skupom. Ovakvi loši rezultati su i bili očekivani jer se treniranje vršilo sa malim brojem primera, što u startu nije dobro kada se koriste neuronske mreže. Takođe, trebalo je jako puno vremena za njegovo treniranje zbog veoma dugačkih vektora koji su predstavljali ulaze, pa je odlučeno da se koristi *Word2Vector* enkodiranje opisa proizvoda, jer proučavajući literaturu, posebno rad [9] ustanovljeno je da se na taj način mogu poboljšati performanse.

U modelu 2, korišćena je neuronska mreža sa istom arhitekturom kao i u modelu 1, razlika je što je ulazni sloj predstavljao vektore enkodiranih vrednosti opisa proizvoda, koje su dobijene korišćenjem GloVe fajla (glove.6b.100d), kako je predloženo u radu [9]. GloVe fajl sadrži 100 dimenzionalne vektore pretreniranih vrednosti za reči engleskog jezika dobijenih na osnovu učestalosti pojavljivanja te reči. Za svaki proizvod formiran je vektor vrednosti, gde svaka vrednost odgovara jednoj reči iz opisa proizvoda i predstavlja prosek vrednosti za tu reč iz GloVe fajla. Model je bilo moguće trenirati i sa većim skupom podataka, a da pritom ne dođe do prevelikog gubitka vremena u treniranju. Najbolji rezultat je dobijen nad skupom od 50.000 primera i iznosi 19,18% tačnosti nad trening skupom i 16,46% tačnosti nad validacionim skupom.

2.6. Neuronska mreža sa LSTM slojem

Nakon toga, odlučeno je da se sačuva redosled reči u opisu proizvoda, kako bi se i to uzelo u obzir prilikom predikcije cene. Da bi ovo bilo postignuto, jednostavna neuronska mreža zamenjena je neuronskom mrežom sa LSTM slojem, na koji se šalju pretrenirane, enkodirane *Word2Vector* vrednosti opisa proizvoda kao ulazi. Ovaj model je odabran proučavajući literaturu [5], gde se navodi da je pogodan za rešavanje ovakve vrste problema predikcije, a da takođe rešava i problem nestajućeg Gradienta[1] (sprečavanje napredovanja procesa obučavanja neuronske mreže tako što gradijenti prilikom treniranja dobijaju previše malu vrednost). Ovaj model se bazira na

[1] https://ayearofai.com/rohan-4-the-vanishing-gradient-problem-ec68f76ffb9b , preuzeto 08.09.2019.

tome da se opisi proizvoda šalju na LSTM sloj gde je *Word2Vector* enkodirana vrednost za svaku reč korišćena za svaki korak u LSTM. Ovaj proces se ponavlja za prvih X reči svakog opisa. Utvrđeno je da je tačna tvrdnja u radu [9] i da se dobija približna tačnost modela i kada se proces ponavalja za maksimalnih 415 reči i kada se ponavlja za 72 reči, a dolazi se do znatne uštede vremena. Iz tog razloga, odlučeno je da vrednost ovog parametra bude 72. Izlaz poslednjeg LSTM koraka ide na *Softmax* izlazni sloj i izlaz je vektor dužine 12, gde svaki izlaz odgovara jednoj grupi cena.

Model 4 ima istu arhitekturu kao i model 3, samo su umesto pretreniranih *Word2Vector* vrednosti, korišćene *Word2Vector* enkodirane vrednosti trenirane nad našim skupom podataka.

Od reči iz opisa proizvoda formiraju se vektori enkodiranih vrednosti reči, korišćenjem fajla kreiranog po uzoru na GloVe fajl, ali se u ovom fajlu za svaku različitu reč iz svih opisa proizvoda kreira vektor vrednosti u zavisnosti od njene učestalosti pojavljivanja u svim opisima. Svaka vrednost dobijenog vektora odgovara jednoj reči iz datog opisa i predstavlja prosek vrednosti za tu reč iz kreiranog fajla. Ovako dobijeni vektori predstavljaju ulaze za LSTM neuronske mreže.

2.7. Neuronska mreža sa LSTM slojem i potpuno povezanim slojem

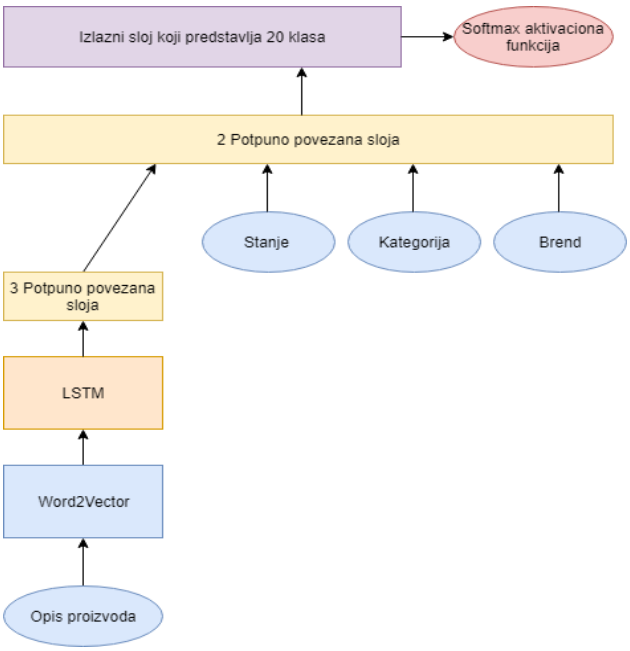
Zatim je odlučeno na osnovu pročavanje rada [4] da se prilikom predikcije cene proizvoda, osim njegovog opisa, uzmu u obzir i kategorija, naziv brenda i stanje kako bi se poboljšale performance modela. Kod modela 6 opisi proizvoda se šalju na LSTM sloj na isti način kao i kod modela 3 i 4, samo što izlaz poslednjeg LSTM koraka, zajedno sa One-hot enkodiranim vrednostima za brend, kategoriju i stanje proizvoda predstavlja ulaz u potpuno povezan sloj, koji ima *Softmax* aktivacionu funkciju i predstavlja vektor dužine 20, gde svaki izlaz odgovara jednoj grupi cena. Broj klasa je povećan kako bi se poboljšala uspešnost našeg modela.

Model 7 je potpuno iste arhitekture kao i model 6, samo je aktivaciona funkcija linearna.

2.8. Neuronska mreža sa LSTM slojem i 3 potpuno povezana sloja

Naredno unapređenje modela predstavlja zamenu postojećeg potpuno povezano sloja sa 3 potpuno povezana sloja, kako bi se poboljšali rezultati predikcije. Testirali smo model 8 i 10 sa *Softmax* aktivacionom funkcijom, gde kod modela 10 postoje 12 klasa, a kod modela 8 postoji 20 klasa, koje predstavljaju rezultat predikcije. Model 9 je testiran sa linearnom aktivacionom funkcijom. Modeli koji vrše klasifikaciju su se pokazali kao najbolji od svih isprobanih modela sa *Softmax* aktivacijom, a takođe je model 9 imao RMSLE koji je najbolji u istraživanju i spao bi u polovinu najboljih rezultata ostvarenih na *Kaggle*-ovom takmičenju Arhitektura jednog od modela sa najboljim rezultatima je prikazana na slici 1.

Modeli 11 (*Softmax* aktivaciona funkcija, 12 klasa), 12 (*Softmax* aktivaciona funkcija, 20 klasa) i 13 (linearna aktivaciona funkcija), imaju arhitekturu kao i prethodno navedeni modeli, samo što se pri predviđanju ne uzimaju u obzir naziv brenda, kategorija i stanje proizvoda, već samo opis proizvoda. Testiranjem, došlo se do zaključka da ova 3 modela daju lošije rezultate od prethodnih.



Slika 1. Arhitektura modela sa najboljim performansama

3. EVALUACIJA I REZULTATI

Za glavnu metriku evaluacije uzet je RMSLE (*Root Mean Square Error)* (slika 2), jer daje veću kaznu za podcenjivanje cene, nego za veću procenu cene. Korišćenje ovog metoda evaluacije zahtevano je i na *Kaggle*-ovom takmičenju i korišćeno u radovima [1][9], pa je odlučeno da se i za verifikaciju modela opisanih u ovom radu koristi ovaj metod evaluacije za modele koje imaju linearnu aktivacionu funkciju, dok je za evaluaciju neuronske mreže koja ima softmax aktivacionu funkciju korišćena tačnost (*Accuracy*).

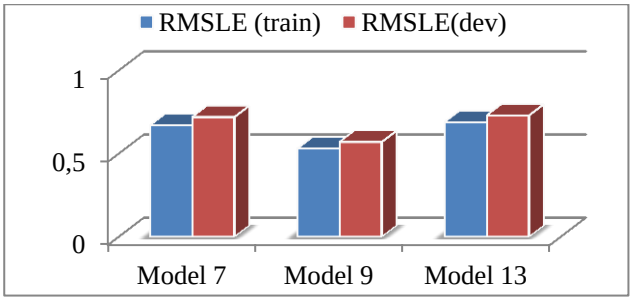
$$\varepsilon = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i + 1) - \log(a_i - 1))^2}$$

- ε – vrednost RMSLE
- n – ukupan broj observacija u javnom/privatnom skupu podataka
- p_i – predikcija cene proizvoda
- a_i – stvarna cena proizvoda
- $\log(x)$ – prirodni logaritam od x

U tabeli 3 su prikazani najbolji rezultati predikcije modela sa linearnom aktivacionom funkcijom, dok je na slici 4 prikazan dijagram ovih rezultata.

Tabela 3. *Najbolji rezultati predikcije modela sa linearnom aktivacionom funkcijom*

	RMSLE (train)	RMSLE (dev)
Model 7	0,67	0,72
Model 9	0,53	0,57
Model 13	0,69	0,73



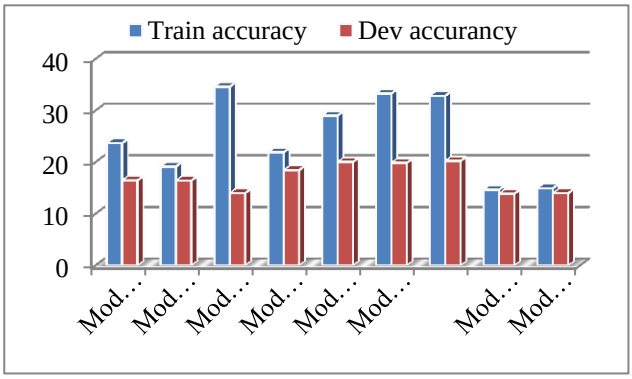
Slika 4. *RMSLE za modele sa lineanom aktivacionom funkcijom*

U tabeli 5 su prikazani najbolji rezultati predikcije modela sa *softmax* aktivacionom funkcijom, dok je na slici 4 prikazan dijagram ovih rezultata.

Tabela 4. *Najbolji rezultati predikcije modela sa softmax aktivacionom funkcijom*

	Train accuracy	**Dev accuracy**
Model 1	23,7	16,5
Model 2	19,18	16,46
Model 3	34,65	14,04
Model 4	21,94	18,46
Model 6	29,04	20,06
Model 8	33,32	19,88
Model 10	32,86	20,26
Model 11	14,67	13,9
Model 12	14,97	14

Na osnovu ovih dijagrama dolazimo do zaključka da najbolje rezultate daju modeli 9 kada je u pitanju linearna aktivaciona funkcija sa ostvarenim RMSLE = 0.53 nad trening, odnosno RMSLE = 0.57 nad validacionim skupom podataka, što bi spadalo u prvu polovinu najboljih ostvarenih rezultata na Kaggle-ovom takmičenju, odnosno, model 10 kada je u pitanju *softmax* aktivaciona funkcija sa ostvarenom tačnošću od 32,86% nad trening skupom i 20.26% nad validacionim skupom podataka.



Slika 5. *RMSLE za modele sa lineanom aktivacionom funkcijom*

4. ZAKLJUČAK

U ovom radu su prikazani modeli za utvrđivanje adekvatne cene proizvoda na osnovu njegovih karakteristika, kao što su opis, naziv brenda, kategorija i stanje proizvoda. Skup podataka korišćen za treniranje svih modela je javno dostupan na sajtu *kaggle.com* [3].

Skup podataka je podeljen na 70% trening, 20% test i 10% validacionih podataka. Isprobani su modeli neuronske mreže sa različitim arhitekturama i sa različitim parametrima. Obučeno je 14 modela nad skupovima podataka različitih veličina u cilju utvrđivanja balansa između performansi i vremena treniranja.

Kao rešenje sa najboljim rezultatima je neuronska mreža sa LSTM slojem, čiji su ulazi enkodirane *Word2Vector* vrednosti trenirane sa našim skupom podataka. Izlaz ovog sloja ujedno predstavlja i ulaz u 3 potpuno povezana sloja mreže. Izlaz ovog dela mreže zajedno sa *One-hot encoding* reprezentacijama stanja, naziva brenda i kategorije proizvoda predstavlja ulaz u nova 3 potpuno povezana sloja sa linearnom aktivacionom funkcijom. RMSLE dobijen evaluacijom ovog modela iznosi 0.5732.

U budućnosti, modeli se mogu unaprediti analizom proizvoda čija procenjena cena se drastično razlikuje od stvarne. Takođe, jedno od unapređenja može biti analiza opisa proizvoda, kako bi se utvrdili delovi opisa koji najviše utiču na cenu, čime bi se smanjila veličina ulaza u neuronsku mrežu.

Rešenja problema predikcije cene proizvoda na osnovu njegovih karakteristika, koja su tema ovog rada, mogu pomoći kako prodavcima u određivanju cene proizvoda na tržištu, tako i kupcima u pronalaženju proizvoda po najadekvatnijoj ceni.

5. LITERATURA

[1] Su, Beichen. An Application of Recurrent Neural Network: Prediction of Items' Price by Description. Diss. UCLA, 2018.
[2] Kim, Kyoung-jae, and Hyunchul Ahn. "A recommender system using GA K-means clustering in an online shopping market." Expert systems with applications 34.2 (2008): 1200-1209.
[3] https://www.kaggle.com (pristupljeno u septembru 2019.)
[4] https://github.com/ChenglongChen/tensorflow-XNN/blob/master/doc/Mercari_Price_Suggesion_Competition_ChenglongChen_4th_Place.pdf (april 2019.)
[5] https://searchenterpriseai.techtarget.com/definition/recurrent-neural-networks (septembar 2019.)
[6] https://medium.com/mlrecipies/deep-learning-basics-gated-recurrent-unit-gru-1d8e9fae7280 (septembar 2019.)
[7] https://github.com/Sebastianvarv/MercariPriceSuggestion (maj 2019.)
[8] https://github.com/gdayal80/mercari/tree/local/Mercari%20Price%20Suggestion%20Challenge (maj 2019.)
[9] Raygada, Javier. "Product Price Suggestions for Online Marketplaces."

Kratka biografija:

Milana Bečejac rođena je u Zrenjaninu 1995. god. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnika i računarstvo odbranila je 2019. god.
kontakt: milana.becejac@gmail.com