Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.85
DOI: https://doi.org/10.24867/01BE48Vidovic

PREDIKCIJA CENE NEKRETNINA NA OSNOVU PODATAKA IZ OGLASA

REAL ESTATE PRICE PREDICTION USING ADVERTISEMENT DATA

Mladen Vidović, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu je predstavljen model za predikciju cene nekretnina na osnovu podataka iz oglasa. Iz oglasa, preuzetih sa veb stranice za oglašavanje, su izdvojene tehničke specifikacije nekretnine, slike nekretnine i, ukoliko su dostupne, geografske koordinate nekretnine. Geografske koordinate su upotrebljene za formiranje ocene kvaliteta lokacije. Slike su upotrebljene za obučavanje neuronske mreže za detekciju značajnih objekata na slikama. Formirana su tri skupa podataka za obučavanje prediktivnih modela. Prvi skup sadrži samo tehničke specifikacije nekretnina, drugi skup ima dodatu ocenu lokacije, a treći skup ima i ocenu lokacije i detektovane objekte na slikama iz oglasa. Za svaki skup je obučeno nekoliko regresionih modela za predikciju cene i njihove performanse su poređene. Performanse ovih prediktivnih modela, izražene kao R^2, su poređene. Najbolje performanse je imao GBT (Gradient Boosted Trees) model na skupu sa slikama i ocenom lokacije sa ostvarenom R^2 vrednošću od 0.856.*

Ključne reči: *Istraživanje i analiza podataka, mašinsko učenje, regresija, neuronske mreže.*

Abstract – *This paper presents a real estate price prediction model using advertisement data. Real estate technical specifications, images and, if available, geographical coordinates are extracted from advertisements, acquired from a real estate advertising website. The coordinates are used to form location ratings. The images are used to train a neural network for the detection of a real estate's equipment on images.Three separate datasets for training the price prediction models were formed. The first dataset contains only technical specifications of the real estates, the second dataset also contains location ratings, while the third dataset contains location ratings and objects detected on images. These datasets were used to train different regression models for predicting real estate prices. The performances of the models, represented by their achieved R^2 scores, were compared in order to establish which model had the best performances.. The best performing model was the GBT (Gradient Boosted Trees) model on the dataset with location ratings and detected objects, with an achieved R^2 score of 0.856.*

Keywords: *Data analysis, machine learning, regression analysis, neural networks.*

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, vanr.prof.

1. UVOD

Kupovina nekretnina za većinu ljudi predstavlja veliku i retku investiciju. Zbog toga je bitno pri prodaji odrediti adekvatnu cenu za nekretninu. U opštem slučaju, cenu nekretnine određuje njen vlasnik, ili agent za prodaju nekretnina. Njihova procena je zasnovana na ličnom iskustvu i interesima, i zbog toga nije objektivna i egzaktna. U ovom radu je predstavljen model za objektivno određivanje cene nekretnine na osnovu njenih karakteristika, kao i kvaliteta lokacije na kojoj se nalazi. Ulaz u model su podaci o nekretninama i njihovim lokacijama ekstraktovani iz njihovih oglasa. Oglasi su preuzeti sa veb stranice za oglašavanje. Model za predikciju cene je realizovan kao regresioni model, pri čemu je za regresiju primenjeno nekoliko različitih algoritama. Model je takođe primenjen na različite skupove podataka, sa različitim atributima, u cilju ispitivanja uticaja podataka o lokaciji, kao i slika iz oglasa, na konačne performanse modela.

Rad se sastoji iz 5 sekcija. U narednoj sekciji je dat pregled literature relevantne za ovaj rad. U trećoj sekciji su prikazani koraci realizacije rešenja predstavljenog u ovom radu. Četvrta sekcija se bavi evaluacijom rešenja i poređenjem performansi različitih prediktivnih modela. Peta sekcija predstavlja sumarizaciju rada i pravce daljeg razvoja.

2. PREGLED SLIČNIH RADOVA

Postoji veliki broj radova na temu predikcije cene nekretnine. Većina ovih radova koristi samo struktuirane tehničke specifikacije nekretnine za obučavanje prediktivnog modela. Radovi koji u skup podataka uključuju lokaciju nekretnine, ili vizualne podatke, odnosno slike, se obično fokusiraju na samo jedan od ovih skupova podataka. U radu [1], autori predstavljaju model za predikciju cene kuće na osnovu tehničke specifikacije i slika. Nad slikama su vršili ekstrakciju svojstava koristeći SURF (*Speeded Up Robust Features*) i ekstraktovana svojstva zajedno sa tehničkim karakteristikama nekretnine prosledili neuronskoj mreži za predikciju cene. Poredili su performanse mreže sa ranijim rešenjem za isti skup podataka [2] i pokazali da je dodavanje slika u obučavajući skup znatno poboljšalo performanse prediktivnog modela. Autori rada [3] su pokazali da lokacija nekretnine, relativno u odnosu na neke značajne centre, poput centra grada ili centara zaposlenja, utiče na cenu nekretnine. Zaključili su da je cena nekretnine inverzno proporcionalna njenoj udaljenosti od ovih značajnih centara. Autori rada [4] pokušavaju da pokažu koliko vrede objekti u okolini nekretnine, odnosno koliko utiču na njenu cenu. Koristeći

GIS (*Geographic Information System*), prikupili su podatke o zelenim površinama, poput parkova u radijusu od 500 metara oko nekretnine i uključili ih u skup podataka. Pokazali su da svaki tip ovakve zone pozitivno utiče na cenu nekretnine, ali da preveliki broj ovakvih zona može da dovede do prezasićenja, i negativno da utiče na cenu nekretnine. Takođe su potvrdili da objekti koji zagađuju okolinu, poput industrijskih zona negativno utiču na cenu nekretnine.

3. METODOLOGIJA

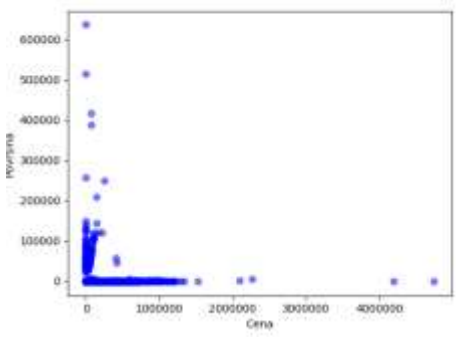
U ovoj sekciji su prikazani koraci implementacije rešenja predstavljenog u ovom radu. Dati su načini dobavljanja podataka, formiranja skupova podataka, kao i formiranje prediktivnih modela i njihova optimizacija.

3.1. Dobavljanje oglasa

Oglasi, iz kojih su izdvojeni podaci o nekretninama, su dobavljeni sa veb stranice za oglašavanje nekretnina, nekretnine.rs [5], upotrebom Scrapy biblioteke za Python programski jezik [6]. Pri prikupljanju, izdvojeni su samo oglasi za prodaju stanova u Novom Sadu i Beogradu. Dobavljeni oglasi su zapisani u datoteku u JSON formatu. Ukupno je dobavljleno 95942 oglasa i 662396 URL-ova slika, koji su kasnije upotrebljeni za dobavljanje slika.

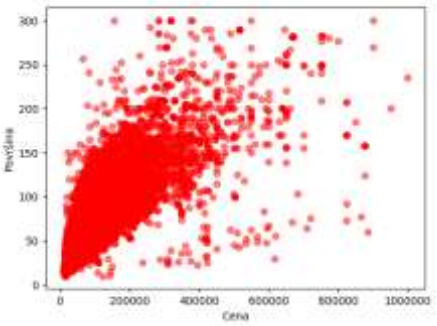
3.1. Pretprocesiranje podataka

Eksplorativnom analizom podataka uočeno je da neki oglasi nemaju navedenu cenu. Oni su neupotrebljivi za formiranje sistema, pa su uklonjeni. U nekim oglasima nije naveden sprat na kojem se stan nalazi. U ranijim radovima je ustanovljeno da sprat ima značajan uticaj na cenu stana [7], pa su oglasi bez navedenog sprata uklonjeni iz skupa podataka. Podaci o datumu objavljivanja i ažuriranja oglasa, kao i tipu oglašavača su uklonjeni iz skupa podataka, jer su vezani za sam oglas, a ne za nekretninu, pa ne bi trebalo da utiču na njenu cenu. Svaki oglas takođe sadrži komentar u obliku slobodnog teksta, koji u opštem slučaju sadrži tekstualni opis nekretnine i kontakt podatke oglašavača. Ekstrakcija podataka iz ovih komentara bi zahtevala primenu *text mining*-a, što izlazi van opsega ovog rada, tako da su slobodni komentari uklonjeni iz skupa podataka. Sledeći korak je bilo određivanje i uklanjanje šuma. Za to su pre svega posmatrane površine i cene nekretnina u oglasima, prikazane na grafiku 1.



Grafik 1. *Podaci pre uklanjanja šuma*

Uočeno je da postoje oglasi za stanove sa navedenom površinom u redu nekoliko hiljada kvadratnih metara, što nije realistično. Takođe ima veliki broj stanova sa cenom preko 1000000 evra, kao i veliki broj stanova sa veoma malom cenom, ispod 1000 evra. Čak 92 stana su imali navedenu cenu od 1 evro. Kao gornja granica za cenu, postavljeno je 1000000 evra, a donja granica je bila 10000 evra. Za površinu, gornja granica je 300m^2 a donja granica 10m^2. Na grafiku 2 se mogu videti podaci nakon uklanjanja šuma.



Grafik 2. *Podaci nakon uklanjanja šuma*

Nakon pretprocesiranja, u skupu podataka je preostalo 61942 oglasa.

3.2 Određivanje kvaliteta lokacije

Oglasi koji su sadržali geografske koordinate nekretnine su upotrebljeni za formiranje novog skupa podataka, od 18374 oglasa, za koje je određen kvalitet lokacije. Za to su, koristeći Overpass API [8] za OpenStreetMap [9], dobavljeni svi javni objekti, poput škola, restorana ili autobuskih stanica, u radijusu od 150m oko nekretnina. Nad tim detektovanim objektima je zatim primenjen *K-means* algoritam za klasterovanje, u cilju formiranja grupa nekretnina koje su slične po kvalitetu lokacije. Klasterovanje je vršeno sa različitim brojem klastera, nakon čega su dodeljeni klasteri spojeni sa cenama nekretnina, i poređene su srednje i prosečne vrednosti cena nekretnina u svakom klasteru. Ispostavilo se da se uvek formira jedan klaster sa nekretninama u čijim okolinama je detektovan mali broj objekata i taj klaster uvek ima najmanje srednje i prosečne vrednosti cene. Preostali klasteri se formiraju na osnovu tipa detektovanih objekata u okolinama nekretnina. Za konačan model odabran model sa 2 klastera, zato što model ima najveću razliku u cenama između klastera. Modeli sa većim brojem klastera imaju mali broj elemenata u nekim klasterima, što bi moglo negativno da utiče na performanse prediktivnih modela.

3.3 Detekcija objekata na slikama

Za detekciju tehničke opremljenosti nekretnina na slikama, obučena je neuronska mreža. Za potrebe formiranja obučavajućeg skupa, ručno je anotirano 7998 slika, koristeći labelImg alat za anotiranje [10], od kojih je 25% izdvojeno kao test skup za evaluiranje mreže. Na slikama su anotirani objekti za koje je smatrano da utiču na cenu nekretnine, poput kućnih aparata ili nameštaja. Za detekciju objekata, preuzeta je unapred kreirana neuronska mreža, koja je već trenirana na COCO [11] skupu podataka, koji sadrži slike svakodnevnih objekata, slične onima koje je bilo potrebno detektovati. Arhitektura preuzete mreže je Faster R-CNN ResNet101 [12]. Mreža je dodatno obučena na ručno anotiranim slikama, i promenjena da kao izlaz vrati detektovane objekte koji su smatrani bitnim za ovaj rad. Obučena

mreža je primenjena na 9862 oglasa koji su imali slike i geografske koordinate, čije slike nisu deo obučavajućeg skupa mreže i nisu anotirane ručno. Detektovani objekti na ovim slikama su dodati kao atributi ovih nekretnina, koje su zatim formirale zaseban skup podataka radi utvrđivanja uticaja detektovanih objekata na predikciju cene.

3.4 Regresija

U prethodnim sekcijama se može videti da su ukupno formirana 3 skupa podataka. Prvi skup podataka sadrži samo osnovne tehničke specifikacije nekretnina. Drugi skup podataka ima dodate ocene okoline nekretnina. Treći skup podata ima ocene okolina i objekte detektovane na slikama nekretnina. U ostatku rada će ovi skupovi podataka biti navedeni kao prvi, drugi i treći. Svaki od ovih skupova je podeljen na trening, validacioni i test skup, tako što je 80% elemenata odvojeno za trening skupove, a 20% za test skupove. Od trening skupova je još 20% elemenata odvojeno za validacione skupove. Za formiranje regresionih modela, odabrano je nekoliko različitih algoritama:

- linearna regresija [13], kao osnova za poređenje,
- linearna regresija sa lasso regularizacijom [14],
- linearna regresija sa ridge regularizacijom [15],
- linearna regresija sa elastic net regularizacijom [16],
- regresiono stablo [17],
- ansambl regresionih stabala u AdaBoost konfiguraciji [18],
- ansambl regresionih stabala u GBT (*Gradient Boosted Trees*) modelu [19].

Algoritmi su optimizovani tako što su iterativno primenjeni nad valdiacionim skupom, sa različitim parametrima. Parametri koji su dali najbolje performanse na validacionom skupu su uzeti kao optimalni. Mera performansi je R^2 vrednost.

U slučaju linearne regresije, primena regularizacija nije dovela do značajnog poboljšanja performansi, pa su ovi modeli odbačeni iz daljeg razmatranja.

4. REZULTATI

U ovoj sekciji je dat pregled ostvarenih rezultata neuronske mreže za detekciju objekata na slikama, kao i regresionih modela za predikciju cene nekretnine.

4.1 Evaluacija neuronske mreže

Neuronska mreža je evaluirana na prethodno izdvojenom test skupu slika, i ostvarila je MAP (*Mean Average Precision*) od 0.594. Najbolje performanse je imala za objekte koji su se često pojavljivali na slikama, poput radijatora centralnog grejanja, sa AP (*Average Precision*) od 0.834, ili klima uređaja sa AP od 0.842.

Pokazalo se da mreža najviše greši na retko zastupljenim objektima, poput interfona ili termoakumulacionih peći, kao i da pravi greške na objektima koji previše liče na neke druge objekte, poput umivaonika koji liče na bide. Ovo bi se moglo rešiti anotiranjem većeg broja slika za obučavajući skup podataka.

4.2 Evaluacija regresionih modela

Evaluacija regresionih modela je izvršena na test skupovima podataka, nad kojima nije vršena nikakva optimizacija modela. Drugi skup podataka je evaluiran i pre i posle dodavanja podataka o kvalitetu lokacije, a treći skup je evaluiran i pre i posle dodavanja podataka o kvalitetu lokacije i objekata detektovanim na slikama, sa ciljem ispitivanja uticaja ovih podataka na performanse prediktivnih modela. Mera za evaluaciju performansi je ostvarena R^2 vrednost. Rezultati evaluacije su dati u tabeli 1.

Tabela 1. *Ostvarene R^2 vrednosti na test skupovima*

Skup podataka	Linearna regresija	Regresiono stablo	AdaBoost	GBT
Prvi	0.69	0.71	0.741	0.759
Drugi, bez klastera	0.708	0.679	0.691	0.754
Drugi sa klasterima	0.716	0.762	0.742	0.814
Treći bez klastera i slika	0.73	0.625	0.691	0.751
Treći sa klasterima	0.734	0.735	0.775	0.851
Treći sa klasterima i slikama	0.739	0.754	0.814	0.856

Na osnovu rezultata se može zaključiti da dodavanje podataka o kvalitetu lokacije znatno poboljšava performanse modela, dok dodavanje objekata detektovanim na slikama ima manji uticaj na performanse prediktivnih modela.

Ovo može da znači da ovi podaci ne utiču u tolikoj meri na cenu, ili da su mogli biti integrisani u skup podataka na bolji način. Model sa najboljim performansama je bio GBT. Najbolju R^2 vrednost je imao za treći skup podataka, nakon dodavanja svih dostupnih atributa. U tabeli 2 je dat pregled grešaka GBT modela, sa srednjim vrednostima cena i površina primera.

Tabela 2. *Greške GBT modela na trećem skupu*

Greška modela (u evrima)	Broj primera	Srednja vrednost površine	Srednja vrednost cene
do 10000	1345	50	56650
10000-50000	539	60	95000
50000-100000	64	118	205505
100000-200000	23	145	370000

Može se videti da greška modela raste sa cenom i površinom nekretnina. To može biti posledica malog broja primera sa velikim cenama i površinama u obučavajućem skupu podataka, kao i činjenicom da na njihovu cenu utiču faktori koje nije moguće predstaviti na stranici za oglašavanje, poput većeg broja pomoćnih objekata i terasa, obezbeđenja zgrade i slično.

5. ZAKLJUČAK

U ovom radu je predstavljen model za predikciju adekvatne cene nekretnine na osnovu njenih tehničkih specifikacija, i lokacije. Podaci o nekretninama su dobavljeni iz oglasa, preuzetih sa veb stranice za oglašavanje nekretnina. Kvalitet lokacije nekretnine je određen na osnovu značajnih objekata u njenoj okolini. Skup atributa nekretnine je proširen značajnim objektima unutar nekretnine, detektovanim na slikama iz oglasa.

Za detekciju objekata na slikama obučena je neuronska mreža. Mreža je ostvarila MAP od 0.594. Analiza pogrešno detektovanih objekata je pokazala da mreža najviše greši kod tipova objekata koji su retko zastupljeni i koji previše liče na druge objekte.

Za predikciju cene nekretnine obučeno je nekoliko regresionih modela, i njihove performanse su poređene. Najbolje rezultate je dao GBT(*Gradient Boosted Trees*) model, sa ostvarenom R^2 vrednošću od 0.856, nakon uključivanja podataka o lokaciji i objekata detektovanim na slikama u skup atributa. Analiza grešaka modela je pokazala da model najviše greši kod primera sa velikom površinom, cenom, novije gradnje, kao i primera koji imaju nepotpune ili neispravno unesene tehničke specifikacije.

Performanse mreže za detekciju objekata na slikama bi se mogle poboljšati dobavljanjem većeg broja obeleženih slika za obučavanje. Detekcija značajnih objekata u većem radijusu oko okoline, zajedno sa podatkom o broju detektovanih objekata istog tipa bi mogao dati bolju reprezentaciju kvaliteta lokacije.

Neispravno popunjene tabele sa tehničkim specifikacijama nekretnine predstavljaju veliki izvor grešaka za prediktivni model. Pokazalo se da su često tačni podaci navedeni u formi slobodnog komentara u oglasu.

Ekstrakcija karakteristika nekretnina iz tih komentara bi omogućila dopunu nenavedenih podataka u tabelama oglasa, kao i ispravljanje pogrešno navedenih podataka.

LITERATURA

[1] Eman H Ahmed and Mohamed Moustafa. House Price Estimation from Visual and Textual Features. November 2016.

[2] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin. Comparative study on estimate house price using statistical and neural network model. *International Journal of Scientific & Technology Research*, 3(12):126–131, 2014.

[3] John Ottensmann, Seth Payton, and Joyce Man. Urban Location and Housing Prices within a Hedonic Model. *Journal of Regional Analysis and Policy*, 38, January 2008.

[4] Fanhua Kong, Haiwei Yin, and Nobukazu Nakagoshi. Using GIS and landscape metrics in the hedonic price modeling of the amenity value of urban green space: A case study in Jinan City, China. *Landscape and Urban Planning*, 79(3):240–252, March 2007.

[5] Nekretnine.rs. dostupno na https://www.nekretnine.rs.

[6] Scrapy | A Fast and Powerful Scraping and Web Crawling Framework.dostupno na https://scrapy.org.

[7] Stephen Conroy, Andrew Narwold, and Jonathan Sandy. The value of a floor: valuing floor level in high-rise condominiums in san diego. *International Journal of Housing Markets and Analysis*, 6(2):197–208, 2013.

[8] overpass api. dostupno na https://wiki.openstreetmap.org/wiki/Overpass_AP.

[9] Openstreetmap. dostupno na https://www.openstreetmap.org.

[10] darrenl. Tzutalin. labelimg. git code (2015)., September 2018. dostupno na https://github.com/tzutalin/labelImg.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] Giuseppe Bonaccorso. *Machine learning algorithms*. Packt, 2017.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[15] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

[17] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

[19] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

Kratka biografija:

Mladen Vidović rođen je u Loznici 14.03.1994. godine. Osnovne akademske studije na Fakultetu tehničkih nauka, smer Softversko inženjerstvo i informacione tehnologije završio je 2017. godine. Iste godine je upisao master akademske studije na Fakultetu tehničkih nauka, smer Softversko inženjerstvo i informacione tehnologije. Položio je sve ispite propisane planom i programom.