Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/12BE37Cabrilo

PREDIKCIJA CIJENE AIRBNB SMJEŠTAJA UPOTREBOM ALGORITAMA MAŠINSKOG UČENJA

PREDICTING AIRBNB PRICES USING MACHINE LEARNING ALGORITHMS

Miljan Čabrilo, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *U ovom radu je vršena predikcija cijene Airbnb smještaja upotrebom algoritama mašinskog učenja. Podaci su preuzeti sa veb-stranice insideairbnb.com. Predikcija cijene smještaja je vršena na osnovu vrijednosti 62 atributa, koji opisuju smještaj, i na osnovu sentimenta korisničkih recenzija. Svaka korisnička recenzija je sentimentalno analizirana, a zatim je određena prosječna vrijednost sentimenta recenzija za svaki smještaj. Primijenjeno je više tehnika eksplorativne analize podataka i više algoritama za selekciju konačnog skupa atributa. Predikcija cijene smještaja je vršena i regresionim i klasifikacionim algoritmima. Korišteni su sledeći algoritmi: linearna regresija, LASSO regresija, ridge regresija, regresija potpornih vektora, Naive Bayes klasifikacija, Random Forest klasifikacija i SVM klasifikacija.*

Ključne reči: *Mašinsko učenje, Airbnb, Predikcija cijene, Regresija, Klasifikacija*

Abstract – *In this paper, the price of Airbnb accommodation was predicted using multiple machine learning algorithms. The dataset was downloaded from the insideairbnb.com website. The price prediction was based on the values of the 62 attributes, which describe the accommodation, and on the sentiment of the user reviews. Sentiment of the each user review was calculated and the average value of the review sentiment was determined for each accommodation. Multiple exploratory data analysis techniques and feature selection algorithms were applied. Both regression and classification algorithms were used. Following algorithms were selected: linear regression, LASSO regression, ridge regression, support vector regression, Naive Bayes classification, Random Forest classification SVM classification.*

Keywords: *Machine learning, Airbnb, Price prediction, Regression, Classification*

1. UVOD

Airbnb je jedna od najpopularnijih internet platformi za kratkoročno iznajmljivanje smještaja. Airbnb se ponaša kao posrednik između vlasnika i korisnika smještaja i omogućava brzo i jednostavno rezervisanje privatnih smještaja.

Vrijednost Airbnba je 2017. godine procijenjena na 31 milijardu dolara, a na njemu se nalaze oglasi za preko 5 miliona smještaja iz preko 190 zemalja.

Vlasnici definišu cijenu smještaja samostalno, a Airbnb nudi samo smjernice za kreiranje cijene u odnosu na lokaciju na kojoj se smještaj nalazi. Broj smještaja na Airbnbu raste svakodnevno, pa je za vlasnike veoma važno da odrede realnu cijenu smještaja, da bi ostali konkurentni na tržištu. Sa druge strane, korisnici moraju da utvrde da li je cijena smještaja prihvatljiva samo na osnovu podataka koje je vlasnik postavio. Cilj ovog rada jeste da se razvije pouzdan model za predikciju cijene smještaja upotrebom mašinskog učenja i sentimentalne analize, koji bi pomogao i vlasnicima i korisnicima smještaja. Upotrebom modela za predikciju cijene korisnici bi bili sigurni da je ponuđena cijena adekvatna u odnosu na kvalitet i lokaciju smještaja, a vlasnici bi bili sigurni da njihov smještaj nije potcijenjen.

Ovaj rad se sastoji od 6 poglavlja. U drugom poglavlju su navedeni prethodni radovi, slične tematike, i prodiskutovani njihovi rezultati. U trećem poglavlju je opisan skup podataka i koraci koji su sprovedeni u fazi eksplorativne analize podataka. U četvrtom poglavlju su navedeni korišteni algoritmi mašinskog učenja. U petom poglavlju su prikazani i diskutovani rezultati koje su postigli regresioni i klasifikacioni modeli. Šesto poglavlje predstavlja zaključak rada u kome su dati i mogući pravci daljeg razvoja.

2. PRETHODNA RJEŠENJA

U radu [1] autori Nikolenko, Rezaei i Rezazadeh su vršili predikciju cijene Airbnb smještaja koji se nalaze u Njujorku. Autori su smatrali da recenzije korisnika imaju veliki uticaj na cijenu smještaja. Izvršili su sentimentalnu analizu svih recenzija i za svaki smještaj su izračunali prosječnu vrijednost sentimenta recenzija. Primijenjeno je više algoritama mašinskog učenja: ridge regresija, K-means klasterovanje sa ridge regresijom, SVR regresija sa RBF kernelom, neuronska mreža i Gradien Boosting Tree regresija. Model SVR regresije je imao najmanju srednju apsolutnu grešku, najmanju srednju kvadratnu grešku i najveću R^2 mjeru.

U radu [2] autori Tang i Sangani su vršili predikciju cijene Airbnb smještaja lociranih u San Francisku i predikciju gradske četvrti u kojoj se smještaj nalazi. Za predikciju i cijene i gradske četvrti je korišten SVM klasifikator. U odnosu na cijenu, smještaji su podijeljeni u dvije grupe, jeftiniju i skuplju, medijanom cijene. Model za predikciju cijene je postigao tačnost od 81.7%,

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

a model za predikciju gradske četvrti je postigao tačnost od 42.2%. Autori su ustanovili da na tačnost modela za predikciju cijene najviše utiču atributi izvučeni iz oglasa, kao što su broj soba, broj kreveta itd. Na tačnost modela za predikciju gradske četvrti su najviše uticali atributi izvučeni iz opisa smještaja.

U radu [3] autori Yu i Wu su vršili predikciju cijene nekretnina. Koristili su relativno mali skup podataka od 1460 primjera, koji je sadržao vrijednosti 79 atributa za svaku od nekretnina. Predikciju su vršili upotrebom regresionih i klasifikacionih modela.

Od regresionih modela su koristili: LASSO regresiju, ridge regresiju, SVM regresiju i Random Forest regresiju, a od klasifikacionih modela su koristili: Naive Bayes, logističku regresiju, SVM klasifikaciju i Random Forest klasifikaciju.

Da bi smanjili dimenzionalnost skupa podataka autori su primijenili analizu glavnih komponenti. Analiza glavnih komponenti je generalno poboljšala performanse modela. Najveću tačnost od 69% kod klasifikacionih modela je postigao SVM model sa linearnim jezgrom, dok je kod regresionih modela najmanji korijen srednje kvadratne greške od 0.5269 imao model SVR regresije.

3. SKUP PODATAKA

Skup podataka preuzet je sa veb-stranice insideairbnb.com i sastoji se od dvije CSV datoteke, listings.csv i reviews.csv. Datoteka listings.csv sadrži podatke o 20.026 smještaja.

Svaki smještaj je opisan vrijednostima 62 atributa. Datoteka reviews.csv sadrži recenzije smještaja. Za svaku recenziju, pored sadržaja, postoji i informacija koja govori kome smještaju pripada.

3.1. Sentimentalna analiza

Uzimajući u obzir uticaj korisničkih recenzija na cijenu smještaja, a sa ciljem poboljšanja performansi prediktivnih modela, izvršena je sentimentalna analiza recenzija za svaki smještaj. Analiza je vršena pomoću Python programskog jezika i TextBlob biblioteke. Svakoj recenziji je dodijeljena vrijednost iz opsega [-1,1]. Gdje -1 označava izrazito negativnu recenziju, a 1 izrazito pozitivnu recenziju. Određivanje sentimenta TextBlob bibliotekom je moguće za tekstove pisane na francuskom, engleskom i njemačkom jeziku, dok je prepoznavanje jezika omogućeno za veći broj jezika. Da bi se ubrzao proces sentimentalne analize i da bi se odstranile recenzije pisane na jezicima za koje nije moguće određivati sentiment upotrebom TextBlob biblioteke izvršeno je filtriranje recenzija. Filtriranje je vršeno regularnim izrazima, a zadržane su recenzije koje sadrže samo standardne karaktere francuskog, engleskog i njemačkog jezika. Nakon filtriranja otpalo je 25.6% recenzija, odnosno od početnih 483.603 recenzija ostalo je 367.401 recenzija. Zatim je za svaku recenziju izvršeno određivanje jezika i na osnovu toga određivanje vrijednosti sentimenta.

Skup atributa, koji opisuje smještaj, je proširen novim atributom čija vrijednost predstavlja prosječan sentiment recenzija smještaja. Za svaki smještaj je izračunata prosječna vrijednost sentimenta recenzija i ta vrijednost je dodata u skup vrijednosti koji opisuje dati smještaj. Za smještaje koji nisu imali ni jednu recenziju usvojeno je da je prosječna vrijednost sentimenta recenzija 0.

3.2. Eksplorativna analiza

U prvom koraku eksplorativne analize podataka su uklonjeni svi očigledno nebitni atributi kao što su: last_review_id (jedinstvena oznaka posljednje recenzije), first_review_id (jedinstvena oznaka prve recenzije), host_id (jedinstvena oznaka vlasnika smještaja) itd. Pošto atributi zipcode (poštanski broj) i neighbourhood (gradska oblast) određuju približno istu geografsku površinu odlučeno je da se zadrži samo atribut neighbourhood.

Zatim je riješen problem nedostajućih vrijednosti. Za atribute security_deposit (sigurnosni depozit) i cleaning_fee (naknada za čišćenje) nedostajuće vrijednosti sa zamijenjene nulama. Pretpostavka je da ukoliko vlasnik ne navede ove vrijednosti da se one ne naplaćuju korisniku smještaja. Niz atributa review_score_clealiness, review_score_accuracy, review_score_location i review_score_communication predstavljaju prosječnu ocjenu korisnika za čistoću smještaja, preciznost podataka iz oglasa, lokaciju smještaja i komunikativnost vlasnika, respektivno. Niska vrijednost bilo koga od ovih atributa bi uticala negativno na cijenu smještaja, s toga je odlučeno da svi smještaji koji imaju nedostajuće vrijednosti za navedene atribute budu uklonjeni iz skupa podataka. Isti princip je primijenjen i za atribute koji opisuju broj soba, broj kupatila i broj kreveta u smještaju. Vrijednost atributa host_response_rate (označava brzinu kojom vlasnik odgovara na upite korisnika) je nedostajala u 43% slučajeva, s toga je atribut host_response_rate u potpunosti izbačen iz skupa atributa.

Kako bi se utvrdilo da li postoje atributi koji su u jakoj korelaciji izvršeno je računanje matrice korelacije. Jaka korelacija između atributa je indikator da jedan od koreliranih atributa treba ukloniti. Ustanovljeno je da su atributi availability_30, availability_60, availability_90, koji označavaju raspoloživost smještaja u periodu od 30, 60 i 90 dana, su visoko korelirani. Odlučeno je da bude zadržan samo atribut availability_60.

Nakon rješavanja problema nedostajućih vrijednosti i koreliranih atributa izvršena je transformacija nominalnih atributa u numeričke. Korišten je dummy encoding pristup. Takođe, izvršena je i transformacija atributa sa binominalnim vrijednostima u numeričke atribute sa skupom vrijednosti {0,1}.

Istraživanjem vrijednosti ciljnog atributa ustanovljeno je da on nema normalnu raspodjelu i da je raspodjela pomjerena ka vrijednosti od 110€. S obzirom da regresija pretpostavlja da su vrijednosti prediktora normalno raspodijeljene, izvršena je logaritamska transformacija cijene smještaja. Logaritamska transformacija gura vrijednosti atributa na desno čime se postiže raspodjela sličnija normalnoj raspodjeli. Pored cijene smještaja, udesno su bile pomjerene i vrijednosti atributa security_deposit i cleaning_fee. I na njih je primijenjena logaritamska transformacija.

Nakon izvršene eksplorativne analize od početna 63 atributa dobijena su 92 atributa. Uklonjeno je 30 početnih atributa, ali se ukupan broj atributa povećao zbog upotrebe dummy encoding tehnike. Ukupan broj smještaja ja smanjen na 17.725 sa početnih 20.026.

Dobijeni skup podataka je podijeljen na tri skupa trening skup (80% početnog skupa), validacioni skup i test skup (po 10% početnog skupa).

3.3. Selekcija konačnog skupa atributa

Selekcija konačnog skupa atributa je vršena upotrebom: selekcije unaprijed, selekcije unazad i selekcije na osnovu *p* vrijednosti modela linearne regresije. Unutar iteracija selekcije unaprijed i unazad skupovi atributa su poređeni R^2 mjerom modela linearne regresije. Selekcijom na osnovu *p* vrijednosti modela linearne regresije su odabrana 34 atributa sa najmanjim *p* vrijednostima.

Dobijeni skupovi atributa su međusobno upoređeni R^2 mjerom modela linearne regresije treniranog na validacionom skupu podataka. U tabeli 1 su prikazane R^2 mjere svih skupova atributa, uključujući i početni skup.

Tabela 1. *Rezultati algoritama za selekciju konačnog skupa atributa*

Skup atributa	R^2 mjera
Početni skup	0.508
Selekcija unaprijed	0.502
Selekcija unazad	0.498
p vrijednosti	0.553

4. METODOLOGIJA

Predikcija cijene smještaja je vršena i klasifikacionim i regresionim algoritmima. Odabrani su sledeći algoritmi:

- Random Forest klasifikacija,
- Naive Bayes klasifikacija,
- SVM klasifikacija sa RBF kernelom,
- linearna regresija,
- LASSO regresija,
- ridge regresija i
- SVR regresija sa RBF kernelom.

Srednja apsolutna greška (eng. Mean Absolute Error - MAE), srednja kvadratna greška (eng. Mean Squared Error - MSE) i R^2 mjera su korištene za evaluaciju regresionih modela. Za evaluaciju klasifikacionih modela korištene su sledeće mjere: tačnost, preciznost, odziv i F mjera.

Da bi se što više poboljšale performanse prediktivnih modela izvršena je optimizacija hiperparametara. Za ridge i LASSO regresiju optimizovan je regularizacioni hiperparametar *C*. Za SVR regresiju izvršena je optimizacija hiperparametra RBF kernela, kao i optimizacija hiperparametra *C*. Kod Random Forest modela optimizovana je maksimalna dubina stabla, a kod SVR regresije je izvršena optimizacija hiperparametara RBF kernela i hiperparametra *C*. Optimizacija hiperparametara je vršena nad validacionim skupom podataka.

Da bi se izvršilo obučavanje klasifikacionih modela bilo je neophodno klasifikovati podatke koji su dobijeni nakon pretprocesiranja i eksplorativne analize. Cijena smještaja određuje klasu kojoj smještaj pripada. Broj klasa i opsezi cijena koji ih definišu su određeni uzimajući u obzir iskustva iz prethodnih radova i potrebe korisnika. Povećanje opsega cijena i smanjenje broja kategorija bi dovelo do poboljšanje performansi svakog od modela. Međutim, takva predikcije bi bila veoma gruba i ne bi imala značaj za krajnjeg korisnika. Definisano je 9 cjenovnih klasa i to:

- od 20€ do 80€,
- od 80€ do 110€,
- od 110€ do 130€,
- od 130€ do 150€,
- od 170€ do 210€,
- od 210€ do 250€,
- od 250€ do 300€ i
- preko 300€.

5. ANALIZA REZULTATA I DISKUSIJA

Prikaz i analiza rezultata će biti podijeljeni u dvije cjeline. Prva cjelina se odnosi na rezultate dobijene upotrebom regresionih modela, a druga cjelina se odnosi na rezultate dobijene upotrebom klasifikacionih modela za predikciju cijene smještaja. Implementacija svih navedenih modela je vršena upotrebom RapidMiner alata.

5.1. Analiza rezultata regresionih modela

U tabeli 2 su prikazani rezultati regresionih modela. Svi regresioni modeli su imali relativno slične performanse iz čega zaključujemo da je selekcije atributa na osnovu *p* vrijednosti imala najviše uticaja na poboljšanje performansi. LASSO i ridge regresija nisu imali značajno bolje rezultate od linearne regresije, što znači da je logaritmovanje cijene smještaja uspješno umanjilo uticaj izuzetaka.

Tabela 2. *Rezultati regresionih modela*

	MAE	MSE	R^2
linearna regresija	0.237	0.114	0.549
LASSO regresija	0.213	0.107	0.565
ridge regresija	0.216	0.1042	0.566
SVR regresija	0.201	0.099	0.612

Od svih regresionih modela model SVR regresije se pokazao najbolje. Imao je najmanju srednju apsolutnu grešku, najmanju srednju kvadratnu grešku i najveću R^2 mjeru.

SVR model je najbolje uspio da modeluje nelinearne zavisnosti cijene smještaja i ostalih atributa smještaja. Primjenom optimizacije hiperparametra je uspješno spriječeno pretjerano prilagođavanje obučavajućim podacima.

5.2. Analiza rezultata klasifikacionih modela

Klasifikacioni modeli su prvo obučeni i evaluirani na skupu podataka sa atributima koji su dobijeni izborom prema *p* vrijednostima. Pošto je analiza glavnih komponenti (eng. Principal Component Anlysis – PCA) značajno poboljšala performanse klasifikacionih modela u radu [3], odlučeno je da se ona primjeni i u ovom radu.

Na skup atributa, dobijen izborom prema *p* vrijednostima, je primijenjena PCA, a zatim su obučeni modeli i evaluirani. Tabela 3 prikazuje performanse modela sa i bez PCA.

Tabela 3. *Performanse klasifikacionih modela*

	Naive Bayes	SVM	Random Forest
Tačnost prije PCA	0.39	0.44	0.58
Tačnost posle PCA	0.47	0.51	0.52

Primjena PCA je dovela do poboljšanja tačnosti kod Naive Bayes i SVM klasifikatora, a do opadanja kod Random Forest klasifikatora. U tabeli 4 prikazani su rezultati Random Forest modela bez analize glavnih komponenti.

Tabela 4. *Performanse Random Forest modela*

	Preciznost	Odziv	F mjera
[0€-80€]	0.55	0.62	0.58
[80€-110€]	0.66	0.51	0.57
[110€-130€]	0.42	0.49	0.45
[130€-150€]	0.38	0.67	0.48
[150€-170€]	0.57	0.45	0.50
[170€-210€]	0.38	0.46	0.41
[210€-250€]	0.52	0.50	0.51
[250€-300€]	0.44	0.32	0.37
300€ +	0.25	0.30	0.27

Iz tabele 4 vidimo da je Random Forest model najbolje performanse ostvario za prve dvije klase, a najlošije za posljednje dvije klase. Ukoliko uporedimo dobijene rezultate sa brojem smještaja klasa zaključujemo da su performanse modela u odnosu na klasu korelirane sa brojem smještaja u klasi. Što je manji broj smještaja u klasi to su lošije performanse modela i obratno. Za bolje diferenciranje klasi sa sličnim brojem smještaja bilo bi potrebno proširiti skup atributa sa atributima koji bolje i detaljnije opisuju smještaje (detaljniji opis namještaja i pogodnosti).

6. ZAKLJUČAK

U ovom radu obučen je niz regresionih i klasifikacionih modela za predikciju cijene Airbnb smještaja. Podaci su preuzeti sa veb-stranice insideairbnb.com. Podaci uključuju karakteristike smještaja (broj soba, broj kreveta itd), lokaciju smještaja i recenzije prethodnih korisnika.

Određivanje sentimenta recenzija je vršeno upotrebom TextBlob biblioteke. Tako dobijeni podaci su spojeni sa osnovnim skupom podataka, a zatim je izvršena eksplorativna analiza. Izbačeni su nepotrebni atributi, riješen je problem nedostajućih vrijednosti, uklonjeni su kolinearni atributi, izvršene su transformacije ciljnog atributa i transformacije nominalnih atributa u numeričke. Zatim su primijenjene tehnike za selekciju skupa atributa i to: selekcija unazad, selekcija unaprijed i selekcija na osnovu *p* vrijednosti. Najbolje se pokazao skup atributa odabran na osnovu *p* vrijednosti.

Od regresionih modela najbolje se pokazao model SVR regresije koji je imao R^2 mjeru od 0.612, srednju apsolutnu grešku od 0.201 i srednju kvadratnu grešku od 0.099. Najbolje performanse, kod klasifikacionih modela, je imao Random Forest klasifikator sa tačnošću od 0.58.

Planovi za dalji razvoj projekta:

- dobavljanje podataka o dodatnim uslugama i pogodnostima smještaja (internet, klima, doručak itd.) i njihovo uključivanje u postojeći skup podataka,
- analiziranje slika iz oglasa kako bi se utvrdila opremljenost smještaja i integracija tako dobijenih podataka u postojeći skup podataka,
- proširivanje skupa podatka udaljenostima smještaja od značajnih turističkih atrakcija i
- dobavljanje podataka o stopi kriminala, bezbjednosti i sigurnosti gradskih oblasti i uključivanje tih podataka u postojeći skup podataka.

7. LITERATURA

[1] L. Nikolenko, H. Rezaei, P. Rezazadeh, „Airbnb Price Prediction Using Machine Learning and Sentiment Analysis", Stanford, 2019

[2] E. Tankg, K. Sangani, „Neighborhood and Price Prediction for San Francisco Airbnb Listings ", Stanford, 2015

[3] H. Yu, J, Wu, „Real Estate Price Prediction with Regression and Classification", Stanford, 2016

Kratka biografija:

Miljan Čabrilo rođen je u Nevesinju 1996. godine. Master rad na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva – Računarstvo i automatika odbranio je 2020. godine.

Kontakt:
miljancabrilo@yahoo.com