Predviđanje popularnosti ugostiteljskog objekta korišćenjem geoprostorne i društvene analitike

Emina Turković
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
emina.turkovic@uns.ac.rs

Tamara Ranković
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
tamara.rankovic@uns.ac.rs

Milijana Đorđević
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
milijanadjordjevic@uns.ac.rs

***Apstrakt*—Pronalaženje optimalne lokacije, u današnje vreme, je nesumnjivo važna tema prilikom uspostavljanja marketinških strategija nekog prodajnog objekta. Takođe, kako otvaranje novog prodajnog objekta iziskuje značajnu investiciju, kompanijama je u interesu da znaju koje lokacije bi privukle veći broj mušterija i samim tim obezbedile veću finansijsku dobit. Brzi rast servisa baziranih na lokaciji u poslednjim godinama, doveo je do visoke dostupnosti podataka koji opisuju mobilnost korisnika i popularnost objekata. Fokus ovog rada bio je na prikupljanju podataka o restoranima na teritoriji grada Njujorka, kako bi se na osnovu tih prikupljenih informacija izvršila procena uspeha njihovih poslovnih objekata. Pored geografskih atributa lokala i njegove okoline, u obzir su uzeti socijalni atributi, koji su oslikavali mišljenja korisnika društvenih mreža. Ciljni atribut koji predstavlja popularnost lokacije je klasni i vrednosti koje može imati su neutral, popular i unpopular. Skup podataka podeljen je na trening, validacioni i test skup. Za potrebe ovog rada korišćeni su sledeći klasifikatori: K-Nearest Neighbor, Logistic Regression, Support Vector Classification, Naive Bayes, Decision Tree i Random Forest Classifier. Model koji se pokazao kao najuspešniji jeste Random Forest Classifier, što pokazuje da su određene osobine lokacija (npr. udaljenost od centra i prosečna cena nekretnina) imale pozitivan uticaj na popularnost objekata. Problem slabog razlikovanja između nepopularnih i neutralnih objekata izdvojio se kao najveći, ali metode Naive Bayes, Decision Tree i Random Forest Classifier su bile najuspešnije u njegovom prevazilaženju.**

Ključne reči—prodajni objekat; optimalna lokacija; geoprostorna analitika; klasifikatori; servisi bazirani na lokaciji

I. UVOD

Odabir optimalnog mesta za prodajni objekat je jedan od važnih aspekata strateškog poslovnog planiranja – ukoliko se pažljivo isplanira, može umnogome uticati na poboljšanje prosperiteta samog prodajnog mesta. Marketing, široka ponuda i kvalitet jesu aspekti u koje kompanija ili vlasnik treba da ulaže, ali ukoliko lokacija nije adekvatno odabrana – položaj objekta može direktno ugroziti poslovanje, bez obzira na značajne investicije. U suprotnom, može u velikoj meri pospešiti i doprineti ostvarivanju profita. Zato se ovaj faktor ne sme zanemariti i potrebno je unapred predvideti kako će lokacija uticati i da li će podstaći uspešno poslovanje.

Prilikom procene lokacija javlja se velika količina parametara. Recimo, prirodno je da će lokacije koje dnevno poseti veliki broj ljudi biti podložnije za uspeh poslovnog objekta od lokacija koje su posećene u manjoj meri. Takve parametre treba uočiti, izdvojiti, a zatim ih uvesti u proračun i analizu.

Domen rada pripada oblasti urbanog računarstva koje ima za cilj rešavanje problema uzrokovanih povećanjem gustine populacije u gradskim sredinama. Time se podstiče kontinuirano poboljšanje kvaliteta života, kao i bolje prostorno planiranje budući da je geografski položaj od primarne važnosti sa stanovišta savremenih trgovinskih i komercijalnih ekosistema u današnjim gradovima. Iz perspektive urbanog računarstva, dinamičniji i informacijama bogatiji podaci mogu se akumulirati razvojem mobilnih uređaja i interneta. Budući da su danas informacije svuda oko nas, od raznih statistika pa sve do informacija ostavljenih od strane samih korisnika (komentari, ocene, posećene lokacije i tako dalje) - mnogo je lakše skupiti podatke i vršiti predikcije. Ako se pravilno analiziraju, ovi korisnički podaci mogu poslužiti kao bogat izvor inteligencije za određivanje optimalnog smeštaja prodajnog objekta.

Ovaj rad daje prikaz jednog rešenja problema procene uspeha poslovnog objekta na osnovu njegovog geografskog položaja. Fokus je na prikupljanju što više raznovrsnih informacija o restoranima na teritoriji grada Njujorka, a zatim se njihovom analizom i transformacijom identifikuju atributi na osnovu kojih je moguće vršiti predikciju njihove popularnosti. Pored toga, rad prikazuje rezultate primene različitih modela nad istim podacima i poredi algoritme primenjene za rešavanje ovog problema.

U nastavku rada biće opisano kako su se podaci sakupljali i prilagođavali navedenom problemu. U drugom poglavlju prikazan je pregled odabrane literature, odnosno naučnih radova, slične tematike. U trećem poglavlju opisan je način formiranja skupa podataka (formiranje atributa), njihovo prikupljanje i procesiranje (priprema za obučavanje modela). Nakon toga, biće predstavljena metodologija koja je korišćena za rešavanje problema predviđanja uspeha prodajnog objekata. Ovo poglavlje uključuje opis svih algoritama klasifikacije koji su upotrebljeni u radu. U petom poglavlju prikazani su i diskutovani rezultati formiranih modela, kao i moguća unapređenja. Na kraju, u poglavlju 6, nalazi se sumarizacija zaključaka dobijenih na osnovu rezultata iz prethodnog poglavlja.

II. PREGLED POSTOJEĆE LITERATURE

U radu [1], cilj je predviđanje popularnosti i ranga maloprodajnih objekata u Teheranu korišćenjem geoprostorne analitike. Rad je u obzir uzeo šest parametara: konkurenciju unutar oblasti, popularnost područja, udaljenost lokacije od centra grada, koliko je lokacija pristupačna (u smislu saobraćaja), komplementarnost posmatranog objekta s drugim objektima u oblasti i entropiju oblasti (raznolikost oblasti - računa se preko kategorija objekata u oblasti). Podaci o lokaciji za svaku maloprodajnu radnju dobavljeni su pomoću *Foursquare Places API*-ja (*Application Programming Interface*) [4]. Zaključeno je da ukoliko je prodavnica smeštena u oblasti sa većom popularnošću, boljom dostupnošću, manjom udaljenošću od centra, sa većom entropijom, više komplementarnih poslova i manje konkurencije, ona je na boljem geografskom položaju i može se očekivati više kupaca. Za nas je takođe važno što se među klasifikacionim algoritmima izdvojio *K-Nearest Neighbor*, kao i *Bayesian Classification*.

U radu [2], cilj je odabir optimalne lokacije za novi prodajni objekat. U obzir su uzeta tri ugostiteljska lanca (McDonald's, Starbucks i Dunkin' Donuts) na području grada Njujorka. Autori su formulisali problem pronalaska takvih lokacija kao problem rangiranja, tako da na osnovu skupa lokacija i njenih karakteristika treba identifikovati one koje će potencijalno privući najveći broj mušterija. Skup podataka formiran je na osnovu Foursquare API-ja i Twitter streaming API-ja odakle su dobavljene informacije o samim objektima i o prijavljivanju korisnika u tim objektima. Svaka oblast predstavljena je svojom geografskom širinom, dužinom i ima određeni radijus. Za svaku oblast vezan je skup atributa koji se može podeliti u dve grupe:

1. Geografski atributi: *Density* (koliko se objekata nalazi u datoj oblasti), *Neighbors Entropy* (mera raznovrsnosti tipova objekata u oblasti), *Competitiveness* (mera broja objekata istog tipa u oblasti), *Quality by Jensen* (mera uticaja okolnih tipova objekata na kvalitet lokacije)
2. Atributi bazirani na mobilnosti korisnika: *Area popularity* (ukupan broj prijavljivanje korisnika u oblasti), *Transition density* (gustina tranzicija među objektima unutar oblasti), *Incoming flow* (prelasci iz objekata van oblasti u objekte unutar oblasti), *Transition quality* (mera potencijalnih mušterija iz okolnih objekata).

Identifikovani su atributi na osnovu kojih se može vršiti predikcija pogodnosti lokacije i primećeno je da kombinacija geografskih i atributa mobilnosti daje bolje rezultate u odnosu na slučaj kada su u obzir uzeti samo geografski atributi. Međutim, kako razlika u rezultatima nije velika, u našem radu biće korišćeni samo geografski atirbuti.

U radu [3], cilj je odabir optimalnog položaja za smeštanje maloprodajnih radnji zbog njihovog značaja za uspeh preduzeća. Predložen je metod za transfer znanja iz većih u manje pametne gradove, na osnovu skupova podataka prikupljenih iz otvorenih izvora podataka u nekoliko velikih gradova u Kini. U ovom radu razmatrana su eksplicitna mišljenja iz recenzija korisnika, kao i implicitna mišljenja iz urbanih regionalnih podataka. Komentari i ocene za svaki region prikupljeni su sa sajta *www.dianping.com*, a saobraćajni indeks je dobijen sa sajta *www.nitrafficindex.com*. Korišćene su karakteristike transakcija sa pametnih kartica u pet gradova, kao i podaci sa najvećeg onlajn sistema za nekretnine u Kini – *www.sofun.com*. Atributi su: *Overall Satisfaction* (ukupno zadovljstvo), *Service Quality*, *Environment Class*, *Consumption Cost* (nivo potrošnje), *Bus stop density* (broj autobuskih stanica), *Smart card balance*, *Real Estate features* (cene nekretnina), *Traffic index features* (karakteristike saobraćaja), *Competitivness features* (broj susednih mesta istog tipa), *Points of interest* i *Quality by Jensen features* (komplementarnost objekta s drugim objektima u oblasti). Postoji korelacija između ocena sa drustvenih mreža i popularnosti radnje što implicira da ako korisnici mobilnih uređaja daju veće ocene za okolinu prodavnice, njena popularnost je veća. Takođe, postoji pozitivna korelacija između uspeha prodavnice i karakteristika vezanih za saobraćaj (kao što su odlazak i dolazak autobusa, gustina autobuskih stanica). Između popularnosti i stanja na pametnoj kartici postoji negativna korelacija.

U tabeli 1 prikazano je koje sličnosti i razlike ovaj rad ima sa radovima iz navedene literature.

TABELA I. SLIČNOSTI I RAZLIKE SA RADOVIMA IZ LITERATURE

Tabela 1	
[1]	- parametri koji su korišćeni u radu zajedno s postupkom njihove obrade i transformacije - Foursquare API kao jedan od izvora podataka - algoritmi klasifikacije: K-Nearest Neighbor, Logistic Regression, Support Vector Classification (SVC), Bayesian Classification, Decision Trees i Random Forests
[2]	- navedeni geografski atributi (ali bez atributa vezanih za mobilnost) - kod nas, problem će biti formulisan kao klasifikacioni problem, a ne problem rangiranja
[3]	- atributi vezani za nekretnine i saobraćaj, ali ih nećemo dobavljati sa istih izvora - nijedan algoritam u radu neće biti korišćen u našem projektu

III. OPIS SKUPA PODATAKA

Ovo poglavlje detaljno opisuje korake koji su doveli do formiranja ciljnog skupa podataka nad kojim je dalje vršena predikcija popularnosti objekta. Prvi zadatak bio je definisanje željenih atributa, što je potom uticalo na odabir adekvatnih izvora podataka. Za uspešno formiranje atributa bilo je neophodno dobaviti podatke o svim restoranima na teritoriji grada Njujorka. Po prirodi sakupljenih podataka oni se mogu podeliti u dve kategorije: geografski atributi lokala i njegove okoline i socijalni atributi, zasnovani na aktivnostima korisnika društvenih mreža. Nakon prikupljanja podataka, usledila je njihova transformacija u finalne atribute, kao i agregacija u jedan skup podataka koji predstavlja ulaz algoritmima mašinskog učenja.

A. Atributi skupa podataka

Skup podataka sastoji se iz jedanaest atributa, od čega njih šest pripada grupi geografskih atributa, dok se ostali svrstavaju u kategoriju socijalnih. Geografski atributi pružaju uvid u tip objekta od interesa, kao i u osobine njegove okoline, gde se okolina definiše kao skup značajnih lokacija u neposrednoj blizini. Tačna odabrana mera neposredne blizine opisana je u

narednom delu poglavlja. Socijalni atributi fokusirani su na mišljenja posetilaca odabranih lokala iskazana kroz nekoliko vidova interakcije na društvenim mrežama.

Za potrebe geografskih atributa definisana je okolina koja uključuje sve objekte koji se nalaze na rastojanju manjem od zadatog radijusa u odnosu na trenutni posmatrani restoran. Atributi za čije je sračunavanje značajna informacija o okolini su gustina, konkurentnost i entropija [2].

Gustina predstavlja ukupan broj objekata na rastojanju manjem od zadatog radijusa. Ako je P skup svih mogućih lokala, $dist(p,l)$ rastojanje dva lokala, a r radijus, formalni zapis je sledeći:

$$density(r)_l = |\{p \in P : dist(p,l) < r\}| \quad (1)$$

Motivacija za odabir atributa zasniva se na činjenici da veći broj objekata u određenoj oblasti implicira veći broj ljudi koji će se tu naći u određenom trenutku i potencijalno posetiti posmatrani restoran.

Atribut konkurentnosti koristi informacije o kategoriji lokala i meri broj objekata iste kategorije u zadatoj okolini. Kako svaki objekat može imati više kategorija, mere za svaku pojedinačnu kategoriju se sumiraju. Broj konkurentnih lokala potencijalno može imati i pozitivan i negativan uticaj na popularnost objekta. Razlog je taj da oblast sa velikim brojem objekata istog tipa privlači određenu ciljnu grupu posetilaca, ali to može dovesti i do smanjenja mušterija jednog lokala u korist nekog drugog. Atribut se formira po sledećoj formuli, gde je γl skup kategorija objekta l, $N_{\gamma l}(l,r)$ broj svih lokala iste kategorije u radijusu r, a $N(l,r)$ broj svih lokala u zadatom radijusu:

$$competitivness(r)_l = -\frac{N_{\gamma l}(l,r)}{N(l,r)} \quad (2)$$

Entropija neke oblasti pruža informaciju o raznovrsnosti tipova objekata i što je njena vrednost veća to je veći i broj tipova kojima lokacije pripadaju. Ako G predstavlja skup svih mogućih tipova, a γ jedan konkretan tip iz skupa G, pri čemu su $N(l,r)$ i $N_\gamma(l,r)$ broj svih objekata odnosno broj objekata tipa γ u okolini objekta l na rastojanju manjem od r, onda se entropija oblasti može definisati na sledeći način:

$$entropy(r)_l = -\sum_{\gamma \in G} \frac{N_\gamma(l,r)}{N(l,r)} \times \log \frac{N_\gamma(l,r)}{N(l,r)} \quad (3)$$

Kako veći broj ljudi u tranzitu u nekoj oblasti može dovesti do novih potencijalnih posetilaca objekta i kako stanice javnog prevoza predstavljaju mesto okupljanja ljudi koje uključuje periode čekanja na ili u blizini stanice, formiran je atribut koji meri broj stanica javnog prevoza u okolini objekta. Ako $d_{min,\,l}$ predstavlja minimalno rastojanje neke stanice javnog prevoza od objekta, a $N(l,r)$ ukupan broj stanica na rastojanju manjem od r, atribut se formira na sledeći način [1]:

$$stops(r)_l = \frac{\log_2 (N(l,r)+1)}{\log_2 (d_{min,\,l})} \quad (4)$$

Centar grada sadrži značajne objekte i za lokalnu populaciju i za turiste, što tu tačku čini oblašću visoke koncentracije potencijalnih posetilaca lokala. Pretpostavka je da što je lokal bliži centru grada, veći broj ljudi je u njegovoj neposrednoj blizini i ima veću šansu da bude popularan. Za objekte u neposrednoj blizini centra nekoliko stotina metara može napraviti veliku razliku u prilivu posetilaca, dok za objekte koji su već značajno udaljeni ta distanca ne predstavlja presudan faktor posetiocima koji dolaze iz centra grada, doneta je odluka da se atribut ne računa samo kao rastojanje centra i objekta, već upotrebom sledeće formule, tako da s predstavlja geografske koordinate centra, a d_s rastojanje te tačke i lokala [1]:

$$dist(s)_l = \frac{1}{\log (d_s)} \quad (5)$$

Poslednji geografski atribut jeste prosečna cena nekretnina u oblasti kojoj pripada restoran. Kako [3] navodi, istraživanja zaključuju da cena nekretnina utiče na kupovnu moć određene oblasti, stoga je i broj posetilaca lokala ili frekvencija njihovih poseta potencijalno veća.

Sem geografskih osobina lokala, na njegovu popularnost utiču i iskustva bivših posetilaca. Veliki broj ljudi, naročito pri prvoj poseti, na društvenim mrežeama pronalazi informacije koje su ostavili drugi korisnici, a koje im služe kao indikator kvaliteta usluga koje restoran pruža. Kako bi se takav faktor uzeo u obzir, formirani su atributi koji kvantifikuju mišljenja dostupna na društvenim mrežama i obuhvataju ukupnu ocenu lokala, broj ocena, broj komentara sa akcentom na sentiment tih komentara što je iskazano kroz atribute procenta pozitivnih i broja negativnih komentara. Za procenat pozitivnih komentara odrađena je dodatna transformacija tako da je na sve vrednosti primenjen kvadratni koren kako bi se napravila jasnija distinkcija između lokala koji imaju mali procenat pozitivnih komentara.

Svi navedeni atributi predstavljaju ulaz u prediktor, dok je ciljni atribut mera popularnosti lokala iskazana kroz broj njegovih posetilaca. Što je veći broj zabeleženih poseta, to je lokal popularniji. Način prikupljanja tih informacija detaljno je objašnjen u narednom delu poglavlja, a oslanja se na društvene mreže putem kojih korisnici prijavljuju svoje prisustvo u nekom lokalu.

B. Prikupljanje podataka

Kako bi opisani atributi bili formirani, prikupljeni su podaci iz više izvora podataka, a to su:

- *Foursquare Places API*
- *Transitland Stops API* [5]
- *Google Geocoding API* [6]
- Veb stranica sa cenama nekretnina [7]

Sa *Foursquare API*-ja inicijalno su dobavljene informacije o svim restoranima na teritoriji grada Njujorka. Spisak atributa koji se mogu dobaviti dostupan je na njihovoj stranici [8], a od ponuđenih dobavljeni su: identifikator lokala, geografska širina i dužina, broj ocena i komentara, ukupna ocena lokala, lista svih komentara i popularnost. Upit za pretragu lokacija sastavljen je tako da zahteva samo lokale odgovarajuće kategorije koji pripadaju ciljnom geografskom području. Prvobitno je kao geografsko područje odabran grad Njujork, međutim kako API ograničava broj rezultata pretrage nisu se mogli dobaviti svi potrebni lokali. Problem je prevaziđen upotrebom skupa podataka koji sadrži cene nekretnina po delovima grada i za svaki deo grada formiran je poseban upit, čime su restorani uspešno dobavljeni. Mana ovog pristupa je ta da rezultujući skup ima veliki broj duplikata jer je API prepoznao pripadanje jednog lokala većem broju delova grada. Svi duplikati po identifikatoru lokacije su uklonjeni, čime je broj rezultata sa 12474 smanjen na 2252.

Isti API iskorišćen je za dobavljanje objekata u blizini svake od prikupljenih lokacija. Upit je formiran tako da se kao rezultat vrate svi lokali na rastojanju manjem od 300 metara od trenutnog objekta. Rastojanje je odabrano po ugledu na [2] koji navodi da se 50% tranzicija posetilaca iz jednog lokala u drugi odigrava među objektima koji su na rastojanju između 200 i 300 metara. Od atributa okolnih lokala preuzeti su identifikator i lista kategorija, kao i geografska širina i dužina radi provere ispravnosti rezultata upita.

Informacije o stanicama javnog prevoza u okolini objekata dobavljene su sa *Transitland API*-ja, formirajući upit koji vraća sve stanice na rastojanju manjem od 300 metara. Sačuvani atributi su identifikator stanice, kao i geografska širina i dužina. Motivacija za preuzimanje geografskih koordinata ista je kao i u slučaju okolnih lokala.

Skup podataka koji sadrži informacije o prosečnim cenama nekretnina ručno je formiran zapisivanjem podataka sa gorenavedene stranice. Svaki red skupa sadrži polje naziva dela grada i prosečne cene nekretnine izražene u dolarima. Ukupan broj delova grada, odnosno podataka u skupu, je 130.

C. Transformacije podataka

Dobavljene podatke bilo je neophodno transformisati u oblik pogodan za predikciju. Postojeći skupovi su agregirani, izračunati su željeni atributi i rešeni su problemi dupliranih i nedostajućih vrednosti.

Analizom atributa otkriveno je da 19% restorana nema vrednost ocene. Prvobitno su te vrednosti popunjene srednjom vrednošću ocena objekata kojima je ona dodeljena. Ponovnim uvidom zaključeno je da su nedostajuće vrednosti vezane za restorane kojima je ukupan broj ocena nula. Pristup popunjavanja srednjom vrednošću ocene se kroz proces obučavanja pokazao kao neefikasan jer se takvim lokacijama davala neopravdano visoka ocena. Stoga, ta vrednost je zamenjena vrednošću nula.

Za potrebe formiranja atributa udaljenosti lokacije od centra grada bilo je potrebno identifikovati geografske koordinate lokacije koja to predstavlja. Kako su obuhvaćeni lokali na teritoriji grada Njujorka, kao centar grada odabran je *Times Square*, čija su geografska širina i dužina bili argumenti formule koja računa vrednosti atributa.

Atributi vezani za dobavljene komentare objekta zahtevali su upotrebu modela koji je sposoban da vrši predikciju sentimenta i klasifikuje komentare na pozitivne i negativne. Odabran je pretrenirani binarni klasifikator ponuđen kroz *AllenNLP* biblioteku [9]. U pitanju je *Long Short-Term Memory* (LSTM) sa *GloVe* embedding-om. Predikcija je izvršena za svaki komentar svake lokacije nakon čega je bilo moguće kreirati atribute čije formule kao ulaz koriste broj pozitivnih ili negativnih komentara.

Spajanje skupova vršeno je nad skupom koji sadrži informacije o lokacijama i skupom koji čuva cene nekretnina. Inicijalno je skup lokacija sadržao informaciju o delu grada kom objekat pripada, što je i polje u skupu cena, ali je takav pristup doveo do velikog broja lokacija koje nisu bile uparene ni sa jednom cenom zbog nepoklapanja u nazivima. Kako bi se problem prevazišao, za svaku lokaciju iz prvog i svaki deo grada iz drugog skupa određen je poštanski kod koji će poslužiti kao atribut za spajanje skupova. Lokacijama je poštanski kod dodeljen upotrebom *Google Reverse Geocoding API*-ja prosleđivanjem geografske širine i dužine, čime su dobavljene detaljne informacije o adresi. Izazovniji zadatak predstavljao je skup cena jer on poseduje samo naziv dela grada. Kako bi se od te tačke došlo do geografskih koordinata nad kojima se može primeniti isti pristup kao nad lokacijama, iskorišćen je *Google Geocoding API* kom su se prosleđivali naziv dela grada spojen sa nazivom grada i države, a kao odgovor dobijale procenjena geografska širina i dužina. Po obavljenom procesu skupovi su spojeni što je uzrokovalo pojavu 214 duplikata i 40 nedostajućih vrednosti za cenu. Problem duplikata javio se zbog toga što više delova grada ima isti poštanski kod. Rešenje je bilo da se kao finalna cena odabere srednja vrednost svih cena koje su spojene sa određenom lokacijom. Nedostajuće vrednosti su posledica poštanskih kodova koji se ne poklapaju ni sa jednim ponuđenim iz skupa cena. Takvim lokacijama dodeljena je srednja vrednost cene za lokacije koje poseduju vrednost atributa.

Poslednja transformacija izvršena je nad ciljnim atributom, popularnošću. *Foursquare API* nudi atribut koji uzima vrednost između nula i jedan i predstavlja meru posećenosti datog objekta u prethodnih šest meseci [8]. Kako je problem predikcije postavljen kao klasifikacioni problem, vrednosti su transformisane tako da pripadaju jednoj od sledećih klasa: nepopularan, neutralan, popularan. Granice za mapiranje su ravnomerno raspoređene, tako da je svaki objekat sa vrednošću manjom ili jednakom 0,33 nepopularan, većom od 0,66 popularan, dok svi ostali restorani pripadaju klasi neutralnih.

D. Eksplorativna analiza

Radi detaljnog uvida u formirani skup podataka izvršena je njegova eksplorativna analiza. Kao rezultat ovog koraka odrađeno je balansiranja skupa podataka prema ciljnom atributu, uklonjeni su redundantni atributi čiji su koeficijenti

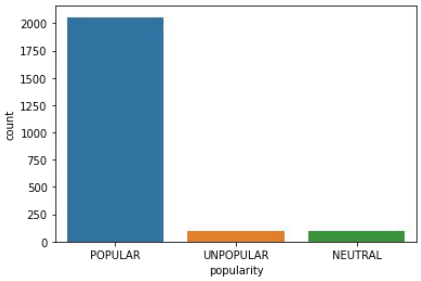
korelacije visoki i istražen je uticaj pojedinačnih atributa na popularnost lokacije.

Na slici 1 a) može se videti dijagram koji prikazuje broj lokacija prema popularnosti. 2057 lokacija je popularno, 102 su neutralne, a 93 su nepopularne. Kako se algoritmi ne bi obučavali na lokacijama koje su dominantno popularne, što na kraju može dovesti do nekvalitetno formiranog modela, izvršeno je uklanjanje primeraka iz klasa popularnih i neutralnih lokacija kako bi svaka klasa bila podjednako zastupljena, sa po 93 predstavnika, čime skup podataka sada umesto 2252 broji 279 objekata. Slika 1 b) prikazuje broj lokacija prema popularnosti nakon balansiranja.

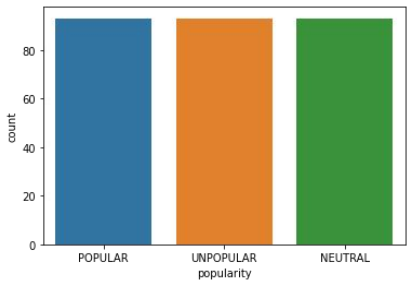
Za kombinacije svaka dva atributa izračunati su koeficijenti korelacije kako bi se otkrili atributi koji su jako linearno zavisni, sa ciljem odstranjivanja suvišnog atributa. Pojava koja se želela izbeći je ta da više atributa nosi sličnu informaciju i samim tim daje veću težinu istoj informaciji prilikom obučavanja. Slike 2 a) i 2 b) prikazuju matrice koeficijenata korelacije sa kojih se primećuju sledeće kombinacije visoko linearno zavisnih kombinacija atributa:

- *rating* i *total ratings*
- *rating* i *total tips*
- *total tips* i *total ratings*
- *density* i *entropy*
- *total ratings* i *negative tips*

Atributi *density*, *total tips* i *total ratings* su isključeni iz skupa podataka čime je rešen problem prenošenja iste informacije kroz više atributa.



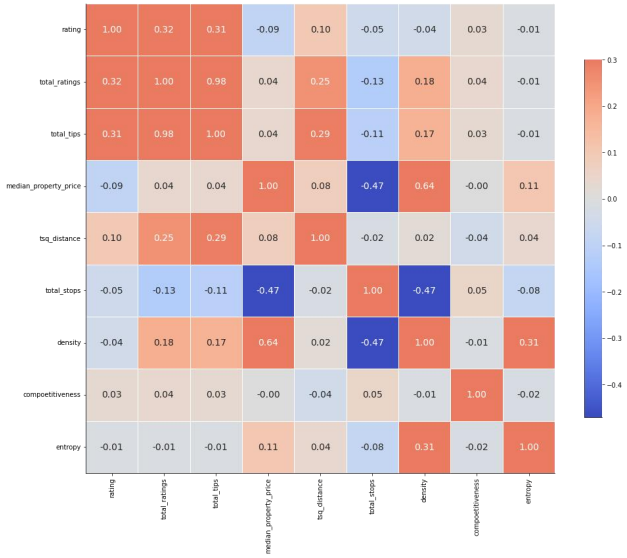
a)



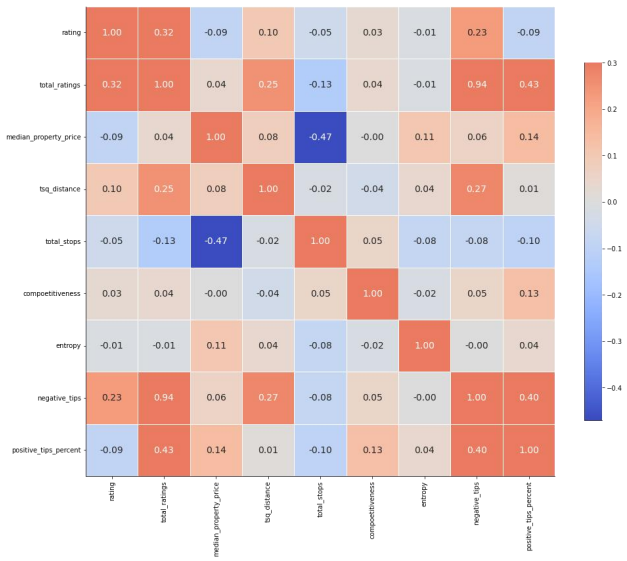
b)

Slika 1. Broj lokacija prema popularnosti a) pre balansiranja b) nakon balansiranja

Radi daljeg uvida u odnose među atributima analizirani su X-Y dijagrami sa kojih se najuočljivije primećuje da su popularne lokacije one sa velikim brojem negativnih komentara, što na prvi pogled može biti kontraintuitivno, ali kako su koeficijenti korelacije broja negativnih i ukupnog broja komentara visoki, sledi da bi isto pravilo važilo i da je posmatran atribut ukupnog broja komentara. Činjenica da su popularne lokacije one koje imaju veći broj negativnih komentara sledi iz toga da one generalno imaju veliki broj komentara u odnosu na one koje su nepopularne i neutralne, koje odlikuje osobina da ili nemaju komentare i ocene ili je taj broj znatno manji. Slika 3 daje prikaz opisanog šablona koji važi za broj komentara lokacije i njenu popularnost. Analizirani su i X-Y dijagrami ostalih kombinacija atributa, ali se sa njih nisu mogli uočiti nikakvi šabloni niti jake zavisnosti.

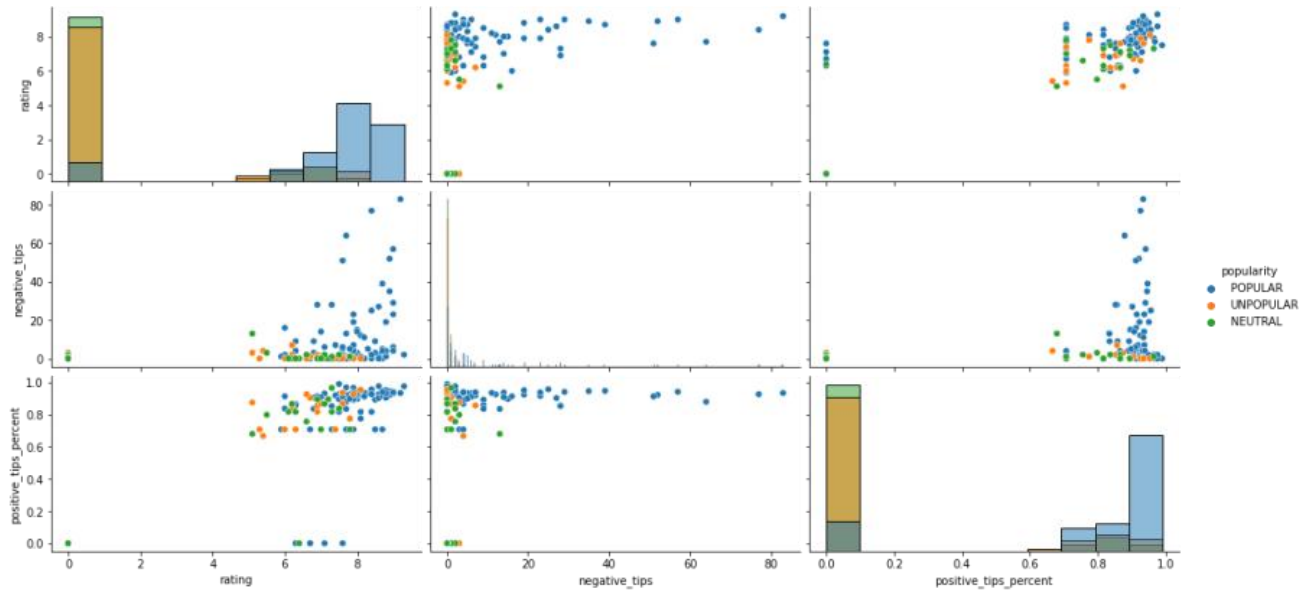


a)



b)

Slika 2. Matrice koeficijenata korelacije a) pre izbacivanja *total_tips* i *density* atributa b) nakon dodavanja *negative_tips* i *positive_tips_percent* atributa



Slika 3. X-Y dijagram

IV. METODOLOGIJA

Za implementaciju korišćena je biblioteka *Scikit-learn* [10], softverska biblioteka otvorenog izvornog koda za kodiranje u *Python*-u. Ova biblioteka sadrži brojne algoritme za probleme regresije, klasifikacije i klasterovanja, i napisana je na način koji je kompatibilan sa drugim korišćenim *Python* bibliotekama kao što su *NumPy* [11] i *SciPy* [12].

Po ugledu na rad iz literature [1] biće korišćeni sledeći klasifikatori:

- KNN (K-Nearest Neighbor)
- Logistic regression
- SVC (Support Vector Classification)
- Naive Bayes
- Decision Tree
- Random Forest Classifier

Za svaki od modela urađena je optimizacija hiperparametara nad validacionim skupovima podataka. Optimizacija se vrši tako što se odabere raspon potencijalno pogodnih vrednosti, nakon čega se treniraju modeli za svaku od mogućih kombinacija hiperparametara. Vrednost koja je dala najbolje rezultate nad validacionim skupom uzima se kao konačna vrednost hiperparametra. O svakom od modela i optimizacijama njihovih hiperparametara biće više reči u nastavku poglavlja.

A. K-Nearest Neighbor

Ovo je jedan od najosnovnijih i najjednostavnijih metoda klasifikacije i prirodno je da bude jedan od prvih izbora za klasifikaciju kada postoji malo prethodnog znanja o distirbuciji podataka. To je metod nadgledanog mašinskog učenja, koji vrši klasifikaciju podataka na osnovu blizine. Da bi se performanse ovog modela maksimizovale potrebno je optimizovati hiperparametre. Kod KNN-a imamo samo jedan hiperparametar, a to je *n_neighbors* (broj suseda). Nakon optimizacije, za parametar *n_neighbors* odabrana je vrednost 6.

B. Logistic Regression

Ovo je tehnika nadgledanog učenja koja se koristi za rešavanje klasifikacionih problema. U osnovi, to je linearna regresija sa nelinearnom aktivacionom funkcijom koja konvertuje kontinualni izlaz u klase. Logistička regresija se podrazumevano ne može koristiti za klasifikacione zadatke koji imaju više od dve labele klasa (eng. *multiclass classification*), zbog čega zahteva dodatnu modifikaciju kako bi podržala takve probleme. *Scikit-learn* pruža nekoliko različitih vrednosti hiperparametra za problem optimizacije (eng. *solver*). Svaki *solver* traži težine koje će minimizovati funkciju cene (eng. *cost function*) i tako rešava problem optimizacije. Od toga, samo *solver*-i „newton-cg“, „sag“, „saga“ i „lbfgs“ mogu raditi s *multiclass classification* problemima. Hiperparametar *penalty* predstavlja kaznu modelu i to rezultira smanjenjem koeficijenata manje doprinosnih varijabli ka nuli. Pošto je potrebno obezbediti kompatibilnost između vrednosti *solver*-a i hiperparametara *penalty*, odabrana je norma kazne „l2“. Još jedan hiperparametar koji je optimizovan, jeste *C* koji predstavlja inverznu jačinu regularizacije (manje vrednosti određuju jaču regularizaciju). Nakon optimizacije, vrednost hiperparametra *C* iznosila je 0.1, dok je za *solver* odabran algoritam „lbfgs“.

C. Support Vector Classification

Ovo je model nadgledanog učenja sa povezanim algoritmima koji se koriste za klasifikacionu i regresionu analizu. U ovom algoritmu svaki podatak prikazujemo kao tačku u n-dimenzionalnom prostoru, nakon čega se klasifikacija obavlja pronalaženjem hiper-ravni (eng. *hyper-*

plane) koja dalje razdvaja klase. Hiperparametri koji se tipično optimizuju kod SVC-a jesu: *C* (dodaje kaznu za svaku pogrešno klasifikovanu tačku podataka), *kernel* (metoda koja se koristi za formiranje izlaza od ulaznih podataka) i *gamma* (parameter za nelinearni kernel čija vrednost može biti „auto“ i „scale“). Najbolje rezultate nad validacionim skupom dala je „sigmoid“ *kernel* funkcija, uz *C*=1 i *gamma*=scale.

D. Naive Bayes

Ove metode predstavljaju skup algoritama nadgledanog učenja zasnovanih na primeni Bajesove teoreme sa „naivnom“ pretpostavkom o uslovnoj nezavisnosti između svakog para karakteristika (eng. *features*) date vrednosti varijable klase. Hiperparametar koji je optimizovan jeste *var_smoothing* koji predstavlja korisnički definisanu vrednost koja se dodaje varijansi distribucije. Najbolji rezultat prilikom optimizacije ostvaren je za vrednost 1e-05.

E. Decision Tree

Ovo je neparametarski (eng. *non-parametric*) metod nadgledanog učenja koji se koristi za klasifikaciju i regresiju, a kod koga je model odluka i njihovih mogućih posledica u obliku drveta (eng. *tree*). Cilj je kreiranje modela koji predviđa vrednost ciljne varijable učenjem jednostavnih pravila odlučivanja koja su izvedena iz karakteristika podataka. Kod stabala odlučivanja, kriterijumi podele (mere čistoće) biraju određeni atribut nad kojim će se podela izvršiti ili granicu koju treba postaviti da bi podela imala smisla (hiperparametar *criterion*) [13]. Od hiperparametara koji utiču na kriterijum zaustavljanja, optimizovan je *max_depth* koji predstavlja maksimalnu dubinu stabla. Za kriterijum podele odabran je *Gini Index* (*criterion* = „gini“), dok je za maksimalnu dubinu stabla odabrana vrednost 7.

F. Random Forest Classifier

Ovo je algoritam nadgledanog mašinskog učenja za klasifikaciju i regresiju, koji funkcioniše tako što konstruiše mnoštvo stabala odlučivanja za vreme treniranja. Za klasifikacione zadatke, izlaz ovog algoritma biće klasa koju bira većina stabala. Hiperparametri koji su optimizovani su:

- *n_estimators* = 30 (broj stabala u šumi)
- *max_depth* = 30 (maksimalna dubina stabla)
- *max_features* = 3 (broj karakteristika koje treba uzeti u obzir pri traženju najbolje podele)
- *min_samples_split* = 20 (minimalni broj uzoraka potreban za podelu unutrašnjeg čvora)
- *min_samples_leaf* = 5 (minimalni broj uzoraka koji je potreban da bi se čvor proglasio listom)

V. Rezultati i diskusija

Zadatak algoritama bio je da izvrši klasifikaciju po atributu *popularity* (na neutralne, popularne i nepopularne). Prilikom optimzacije modeli su isprobavani nad različitim skupovima atributa iz skupa podataka (eng. *feature selection*), pa tako svaki model koristi različite atirbute u zavisnosti od toga za koje atribute daje najbolje rezultate (tabela 2).

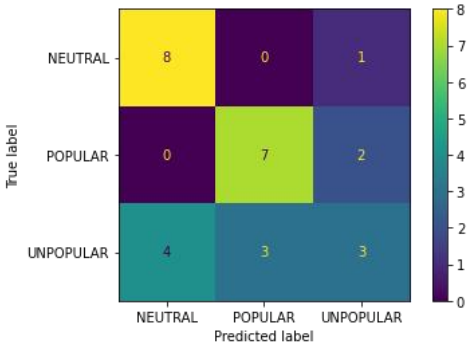
Tabela II. Za svaki od modela prikazan je skup atributa koje oni koriste u zavisnosti od rezultata

	Atributi							
	rating	median property price	tsq distance	total stops	competitiveness	entropy	negative tips	positive tips percent
KNN		✓	✓	✓	✓	✓		✓
Logistic Regression	✓	✓	✓	✓	✓	✓		✓
Support Vector Classification		✓	✓	✓	✓	✓		✓
Naive Bayes			✓	✓	✓	✓	✓	✓
Decision Tree		✓	✓	✓	✓	✓		✓
Random Forest Classifier		✓	✓	✓	✓	✓		✓

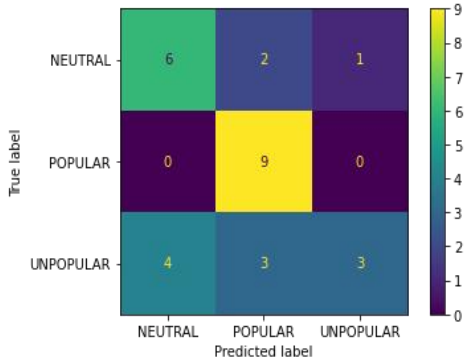
A. Rezultati

Radi verifikacije skup podataka je podeljen na trening, validacioni i test skup u razmeri 80:10:10. Pri evaluaciji algoritama korišćene su Precision, Recall i F-mera, koje su standardne validacione metrike za probleme u oblasti *data mining*-a (po ugledu na rad [1] iz literature).

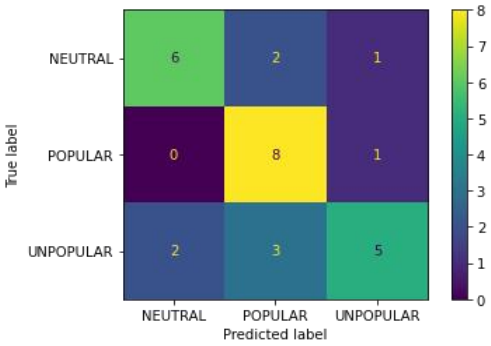
U nastavku rada redom su prikazane matrice konfuzije (eng. *confusion matrix*) za svaki od modela: KNN (slika 4), Logistic Regression (slika 5), SVC (slika 6), Naive Bayes (slika 7), Decision Tree (slika 8) i Random Forest Classifier (slika 9).



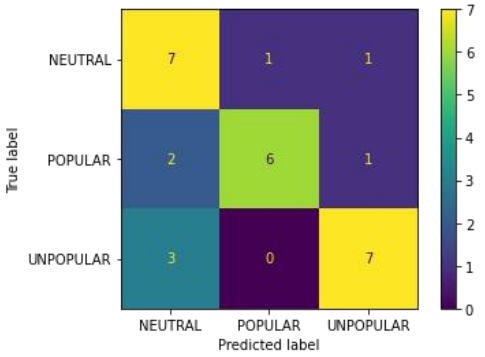
Slika 4. Matrica konfuzije za KNN klasifikator



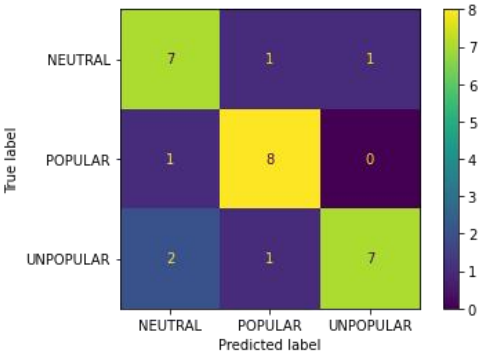
Slika 5. Matrica konfuzije za Logistic regression klasifikator



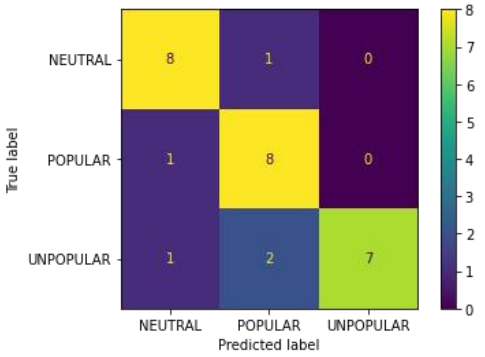
Slika 6. Matrica konfuzije za SVC klasifikator



Slika 7. Matrica konfuzije za Naive Bayes klasifikator



Slika 8. Matrica konfuzije za Decision Tree klasifikator



Slika 9. Matrica konfuzije za Random Forest Classifier klasifikator

Kao glavni problem prilikom predviđanja izdvojilo se slabo razlikovanje neutralnih i nepopularnih mesta, zbog toga što i jedna i druga klasa imaju slične osobine u skupu podataka (npr. mali broj komentara i odsustvo ocena). Bez obzira na to, dobre rezultate u raspoznavanju nepopularnih i neutralnih objekata dali su Naive Bayes, Decision Tree i Random Forest Classifier.

Srednja Precission, Recall i F-mera za klasifikacione algoritme prikazana je u tabeli 3. Kao što se vidi po rezultatima, Random Forest Classifier je algoritam koji ima najveću uspešnost, sa F-merom koja iznosi 0.82. Njega prati Decision Tree, sa F-merom od 0.79, koji je nešto lošiji u razlikovanju nepopularnih i neutralnih objekata. Među ostalima, izdvaja se još i Naive Bayes sa F-merom 0.72, dok ostali nisu napravili rezultate koje bismo smatrali značajnim (iz razloga koji će biti diskutovani u nastavku), čak i nakon optimizacije hiperparametara.

TABELA III. SREDNJA PRECISION, RECALL I F-MERA ZA SVAKI OD ALGORITAMA

Tabela 3			
Klasifikacioni algoritmi	*Precison*	*Recall*	*F-mera*
KNN	*0.62*	*0.66*	*0.62*
Logistic Regression	*0.66*	*0.66*	*0.61*
SVC	*0.69*	*0.69*	*0.67*
Naive Bayes	*0.74*	*0.71*	*0.72*
Decision Tree	*0.79*	*0.79*	*0.79*
Random Forest Classifier	*0.84*	*0.83*	*0.82*

B. Diskusija

Dalja unapređenja rada mogla bi da se fokusiraju na rešavanje problema razlikovanja neutralnih i nepopularnih lokacija, budući da je to ono što najviše utiče na uspešnost modela u predviđanju. Jedno od mogućih rešenja jeste uvođenje novih atributa. Na primer, cena jela bila bi faktor čijim uvođenjem bi se dodatno razgraničila mesta kojima korisnici nisu zadovoljni, budući da mesta sa velikim cenama i malim brojem ocena spadaju u kategoriju nepopularnih lokacija. Takođe, ako uzmemo u obzir da je popularnost definisana na osnovu poseta u prethodnih šest meseci, još jedno od unapređenja na koje bi trebalo obratiti pažnju jeste filtriranje svih komentara na one koji su ostavljeni u tom periodu. Ovim postupkom sprečili bismo pogrešnu kategorizaciju objekata koji su u poslednjih šest meseci značajno promenili svoje usluge (recimo pod uticajem prethodno ostavljenih negativnih komentara).

VI. Zaključak

Kroz rad je opisan postupak vršenja predikcije popularnosti poslovnih objekata na osnovu podataka o restoranima u Njujorku dobijenih sa različitih izvora. Pored podataka o restoranima, prikupljani su i geografski podaci o lokacijama na kojima se restorani nalaze kao i komentari i ocene ostavljani za svaki od njih. Usput, analizirana je i poređena uspešnost predviđanja različitih klasifikacionih algoritama.

Popularnost objekata definisana je kroz klase popularnih, neutralnih i nepopularnih objekata. Model koji su se pokazao kao najbolji za rešavanje ovog problema jeste Random Forest Classifier sa F-merom od 0.82, a odmah za njim Decision Tree čija je F-mera bila 0.72. Ostali modeli su najviše grešili kod nepopularnih i neutralnih objekata, pošto je za njih najčešće postojala nedovoljna razlika u vrednostima atributa. Predlog za dalji razvoj je uvođenje dodatnih atributa koji bi podstakli jasniju distinkciju između navedenih klasa.

Očigledno je da su neke osobine lokacije u direktnoj korelaciji s popularnošću objekata na toj lokaciji (npr. udaljenost od centra grada), ali je cilj ovog rada bio uzeti u obzir i mnoge druge, manje očigledne faktore i na osnovu njih proceniti kako će objekat poslovati. Ovo je aktuelna tema velike važnosti samim tim što se gradovi neprestano šire i sve je veća potreba za različitim uslužnim objektima. Troškovi i ulaganja su visoki, pa je potrebno obezbediti sigurnije i isplativije planiranje. Kako su neki od modela pokazali visoke performanse prilikom predikcije, ovaj rad bi mogao doprineti minimizovanju novčanih gubitaka prilikom otvaranja novih poslovnih objekata. Sem toga, primenjeni pristupi mogli bi biti dobra polazna tačka pri rešavanju problema slične tematike.

Literatura

[1] Damavandi, H., Abdolvand, N., & Karimipour, F. (2019). Utilizing location-based social network data for optimal retail store placement. *Earth Observation and Geomatics Engineering*, 3(2), 77-91.

[2] Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V., & Mascolo, C. (2013, August). Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 793-801).

[3] Zhang, N., Chen, H., & Chen, X. (2017). Transfer learning for urban computing: A case study for optimal retail store placement. In *Proceedings of the Fourth International Workshop on Urban Computing.* Sydney, Australia: ACM.

[4] [Online]. Available: https://developer.foursquare.com/docs/places-api-overview. [Accessed: 20-Apr-2022].

[5] [Online].Available:https://www.transit.land/documentation/datastore/stops. [Accessed: 20-Apr-2022].

[6] [Online].Available:https://developers.google.com/maps/documentation/geocoding/overview. [Accessed: 20-Apr-2022].

[7] [Online]. Available: https://www.bloomberg.com/graphics/property-prices/nyc/. [Accessed: 20-Apr-2022].

[8] [Online].Available:https://developer.foursquare.com/reference/response-fields. [Accessed: 20-Apr-2022].

[9] [Online]. Available: https://docs.allennlp.org/main/. [Accessed: 20-Apr-2022].

[10] [Online]. Available: https://scikit-learn.org/stable/. [Accessed: 25-Apr-2022].

[11] [Online]. Available: https://numpy.org/. [Accessed: 25-Apr-2022].

[12] [Online]. Available: https://scipy.org/. [Accessed: 25-Apr-2022].

[13] Žitković, B. (2020). Primjena stabla odlučivanja na skupu podataka iz obrazovanja (Doctoral dissertation), *University of Zagreb. Faculty of Organization and Informatics. Department of Information Systems Development.*