Predlog projekta iz SIAP-a

Predviđanje popularnosti prodajnog objekta korišćenjem geoprostorne analitike

Turković Emina	Ranković Tamara	Đorđević Milijana
E2 - 44/2021	E2 - 53/2021	E2 - 93/2021

1. Definicija problema

Cilj projekta je predviđanje uspeha prodajnog objekta na osnovu njegovog geografskog položaja. Fokus je na prikupljanju podataka o *Starbucks* prodajnim objektima na teritoriji grada Njujorka. Potrebno je prikupiti što više raznovrsnih informacija o oblasti u kojoj se svaka prodavnica nalazi kako bi se njihovom analizom i transformacijom identifikovali atributi na osnovu kojih je moguće vršiti predikciju popularnosti.

2. Motivacija

Domen projekta 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. 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.

3. Relevantna 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.

- **Tema rada:**
  Cilj je predviđanje popularnosti i ranga maloprodajnih objekata u Iranu korišćenjem geoprostorne analitike. Model koji se dobija primenom algoritama mašinskog učenja na podatke o lokaciji već postojećih objekata koristi se za predikciju uspeha prodavnica u Teheranu.

- **Metodologija:**
  1. Tradicionalni algoritmi klasifikacije/regresije: SVC, Decision Trees, Logistic Regression, Bayesian Classification, K-Nearest Neighbor i Random Forests.
  2. Learn-to-rank algoritmi: RankNet, LambdaMART i MART

- **Skup podataka:**
  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 izvučeni su pomoću *Foursquare API*-ja.

- **Evaluacija:**
  Klasifikacioni algoritmi evaluirani su pomoću Precision, Recall i F-mere. Za learn-to-rank algoritme koristi se precision@k metrika kao i nDCG@k (*Normal Discounted Cumulative Gain* – standardna *Information Retrieval* metrika koja upoređuje relativnu poziciju svake stavke u dobijenim rezultatima s njenim stvarnim rangom).

- **Najvažniji rezultati:**
  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. Među klasifikacionim algoritmima, izdvojio se K-Nearest Neighbor (Precision 0.899, Recall 0.899 i F-mera 0.911). Poređenjem po nDCG@k metrici, learn-to-rank algoritmi daju preciznije rezultate od klasifikacionih algoritama. MART je imao najveću preciznost u rangiranju i predviđanju (nDCG@k 0.854), a odmah za njim LambdaMART (nDCG@k 0.8275) i RankNet (nDCG@k 0.823). Takođe, po ovoj metrici, Bayesian Classification pokazao se kao najprecizniji klasifikacioni algoritam (nDCG@k 0.706).

- **Zaključak:**
  U našem projektu uzećemo u obzir podatke (odnosno parametre) koji su korišćeni u radu zajedno s postupkom njihove obrade i transformacije. Koristićemo *Foursquare API* kao jedan od izvora podataka. Takođe, algoritmi klasifikacije koji su primenjeni i analizirani u ovom radu biće upotrebljeni i kod nas.

[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).

- **Tema rada:**
  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.

- **Metodologija:**
  U slučaju kada je samo jedan atribut uzet u obzir, sračunata je njegova numerička vrednost, a zatim su oblasti rangirane prema istoj. Kada se predviđanje vršilo nad kombinacijom više atributa, korišćeni su sledeći modeli nadgledanog učenja:
  1. Regresija: Support Vector Regression, M5 Decision Trees, Linearna regresija sa regularizacijom
  2. Learning-to-rank: RankNet

- **Skup podataka:**
  Skup podataka formiran je na osnovu *Foursquare Venue 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)

- **Evaluacija:**
  Rešenje je evaluirano unakrsnom validacijom. Korišćena je nasumična sub-sampling metoda tako da se u svakom eksperimentu odabere 33% oblasti kojima pripadaju objekti brenda za potencijalne nove lokacije. Ostali podaci su bili deo trening skupa. Upotrebljene metrike su NDGC@k i Accuracy@k (k = x%).

- **Najvažniji rezultati:**
  Identifikovani su atributi (kao što su *Competitivness* i *Jensen Quality* iz grupe *Geographic Features* i *Incoming Flow* i *Transition Quality* iz grupe *Mobility Features*) na osnovu kojih se može vršiti predikcija pogodnosti lokacije. Svi primenjeni modeli nadgledanog učenja dali su bolje rezultate kada su vršili predviđanje na osnovu kombinacije geografskih i atributa mobilnosti u odnosu na slučaj kada su u obzir uzeti samo geografski atributi. Na primer NDCG@10 rezultat u slučaju *Starbucks*-a porastao je sa 0.72 na 0.77 (rezultati koje su dala stabla odlučivanja i RankNet).

- **Zaključak:**
  Navedeni geografski atributi biće deo skupa podataka koji ćemo koristiti. Kako poboljšanja dobijena dodavanjem atributa vezanih za mobilnost nisu procenjena kao drastična, oni neće biti razmatrani. Takođe, problem će biti formulisan kao klasifikacioni problem, a ne problem rangiranja.

[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.

- **Tema rada:**
  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.

- **Metodologija:**
  1. Softmax regression - uopštenje Logistic Regression algoritma
  2. Multi-view Transfer Learning with Autoencoders

- **Skup 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).

- **Evaluacija:**
  Precision i Recall, nDCG@k

- **Najvažniji rezultati:**
  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.

- **Zaključak:**
  Iz ovog rada posebno ćemo obratiti pažnju na atribute vezane za nekretnine (*Real Estate features*) i saobraćaj (*Traffic index features*), ali ih nećemo dobavljati sa istih izvora. Navedeni algoritmi neće biti korišćeni u našem projektu.

3. Skup podataka

Cilj je dobaviti što opsežniji skup podataka o lokaciji i njenoj okolini. Zbog toga će primarni izvor podataka biti *Foursquare Venue API (*https://developer.foursquare.com/reference/v2-overview*)* koji nudi raznovrsne informacije o Starbucks objektima na teritoriji Njujorka. Motivacija za odabir lanca dolazi iz velikog broja prodajnih objekata, a na osnovu toga što je najviše dostupnih podataka *Foursquare Venue API*-ja vezano za teritoriju grada Njujorka, odabran je dati grad.

Ciljni atribut koji predstavlja popularnost lokacije će biti klasni i vrednosti koje može uzeti su *highly popular*, *popular*, *neutral*, *unpopular* i *highly unpopular*. Klasa kojoj lokacija pripada biće određena na osnovu broja *check in*-ova, broja komentara i ukupne ocene na *Foursquare*-u. Neki od atributa na osnovu kojih će se vršiti predikcija, a koji se mogu dobiti transformacijom podataka dobavljenih sa *Foursquare Venue API*-ja su density, entropy, competitivness, quality by Jensen (preuzeto iz literature). U obzir će biti uzeta i prosečna cena nekretnina po delovima grada (izvor podataka je: https://www.bloomberg.com/graphics/property-prices/nyc/). Nakon toga, neophodno je povezati lokaciju sa delom grada kom pripada koristeći njenu geografsku širinu i dužinu. Kako je jedan od radova [3] iz relevantne literature ukazao na značaj saobraćaja na predikciju popularnosti lokacije objekta, uz pomoć *Transitland API*-ja (https://www.transit.land/documentation/rest-api/stops) biće dobavljen broj stanica javnog prevoza za oblast kojoj lokacija pripada. Poslednji od atributa koji se trenutno planira uzeti u obzir je udaljenost lokacije od centra grada.

4. Metodologija

Po ugledu na rad iz literature [1] biće korišćeni sledeći klasifikatori: SVM (*Support-vector machine*), Decision Trees, Logistic Regression, Bayesian Classification, K-Nearest Neighbor i Random Forests.

5. Metod evaluacije

Radi evaluacije skup podataka biće podeljen na trening, validacioni i test skup u razmeri 80:10:10. Kao mera evaluacije koristiće se Precision, Recall i F-mera.