UDC: 004.45 Info M str. 25-31

PREPOZNAVANJE SLIKA POMOĆU DUBOKIH NEURONSKIH MREŽA ZA PREDVIĐANJE RENTIRANJA STAMBENIH OBJEKATA
IMAGE RECOGNITION WITH DEEP NEURAL NETWORKS FOR PREDICTION OF RENTING RESIDENTIAL OBJECTS

Damir Pajaziti

REZIME: U ovom radu se istražuje prepoznavanje slika pomoću dubokih neuronskih mreža za predviđanje rentiranja stambenih objekata. Cilj istraživanja je eksperimentalna evaluacija algoritama mašinskog učenja, sa poređenjem rezultata modela na inicijalnom skupu podataka i na proširenom skupu koji je oplemenjen novim atributima koji su rezultat prepoznavanja objekata na slikama stambenih objekata. Za potrebe istraživačkog rada, napravljen je projekat, koristeći programske jezike *Python*, *C#* i *TSQL*. Projekat se bavi predviđanjem klase zainteresovanosti zakupaca stambenih objekata u gradu *Nju Jork* u *Sjedinjenim Američkim Državama* nad podacima iz 2016. godine, korišćenjem algoritama mašinskog učenja. Moguće klase zainteresovanosti zakupaca su: niska, srednja i visoka. Koristeći prethodno trenirane modele dubokih neuronskih konvolucionih mreža za prepoznavanje objekata na slikama, proširuje se inicijalni skup podataka, sa novodetektovanim objektima kao novim atributima. Nakon pripreme podataka za prediktivno modelovanje, vrši se predikcija izlazne klase pomoću algoritama mašinskog učenja za klasifikaciju. Nakon toga se upoređuju rezultati predviđanja na inicijalnom i proširenom skupu podataka. U cilju poboljšanja uspešnosti predviđanja modela, eksperimentalno se evaluiraju modeli kreirani tehnikama za: podešavanje parametara algoritama, balansiranje podataka prema izlaznoj klasi kao i tehnikom za pronalaženje najbitnijih atributa. U zaključku su istaknuti ključni rezultati i opažanja, preporuke za primenu algoritama kao i pravci budućeg razvoja.

KLJUČNE REČI: mašinsko učenje, konvolucione neuronske mreže, prepoznavanje objekata na slikama, rentiranje stambenih objekata

ABSTRACT: In this paper we research image recognition with deep neural networks for prediction of renting residential objects. The aim of this research is experimental evaluation of machine learning algorithims, with comparation of model results on initial data set and on extended data set, which is enriched with new attributes that are the result of object recognition on images of residential buildings. For the needs of research work, a project in *Python*, *C#* and *SQL* programming languages was made, which deals with predicting the interest class of tenants of housing in the city of *New York* in the *United States* on data from 2016., using machine learning algorithms. Possible tenant interest classes are: low, medium and high. Using pre-trained models of deep neural convolutional networks for image recognition, the initial data set is extended, with newly recognized objects as new attributes. After preparing the data for predictive modeling, the output class is predicted using machine learning algorithms for classification. In order to improve the accuracy of model prediction, models were evaluated experimentally and created by techniques for: fine tuning algorithm parameters, balancing data according to the output class and with technique for finding best K attributes. The results of the prediction on the initial and extended data sets are then compared. In conclusion, results and proposals for future application and development of machine learning algorithms are proposed.

KEY WORDS: machine learning, convolutional neural networks, recognizing objects in images, renting residential objects

1. UVOD

Prema [4], „ulaganje u nekretnine je mukotrpan posao, ali se isplati onome ko je dovoljno istrajan“. Uspešni investitori moraju imati potrebne veštine da pronađu odgovarajuće stambene objekte, da ih evaluiraju i pronađu kupce za njih. Pomoć u analizi rentiranja nekretnina može se pronaći i u nauci o podacima. Novi analitički alati sa prediktivnim mogućnostima mogu da dramatično utiču na budući razvoj urbanih sredina u industriji nekretnina, menjajući proces i tok poslovanja [5].

U ovom radu se istražuju načini i metode korišćenja veštačke inteligencije, za podršku u odlučivanju prilikom rentiranja stambenih objekata. Korišćenjem algoritama mašinskog učenja u sinergiji sa dubokim neuronskim mrežama, eksperimentalno se pravi model koji bi trebalo da bude sposoban da uči na osnovu prethodnih iskustava (u ovom istraživanju iskustva su podaci o rentiranju stambenih objekata) i da uspešno klasifikuje zainteresovanost klijenata za stambene objekte. S obzirom da su modeli koji se koriste u istraživanju pripremani i trenirani uz ljudsku pomoć, koriste se algoritmi koji spadaju u nadgledani tip mašinskog učenja (*eng. Supervised learning*).

Prema [2], neki od glavnih indikatora kvaliteta stambenih objekata su: vizuelni uticaj enterijera i eksterijera, lokacija, položaj i opremljenost, dok se u literaturi [3] navode sledeći kriterijumi: visina cene rentiranja, visina dodatnih troškova, blizina obrazovnih ustanova, dostupnost lekova kao i starost objekta.

Jedna od glavnih ideja ovog istraživanja je otkrivanje dodatnih informacija sa slika stambenih objekata, korišćenjem dubokih konvolucionih neuronskih mreža *DCNN (eng. Deep Convolutional Neural Networks),* i uključivanje istih u kreiranju konačnog modela predviđanja. Nakon toga se upoređuje uspešnost modela koji sadrži nove informacije dobijene prepoznavanjem objekata na slikama i inicijalnog modela koji ne sadrži nove informacije. Obogaćivanjem inicijalnog skupa atributa podataka, pokušava se doći do poboljšanja performansi ali i značajnih opažanja prilikom izvođenja eksperimenata nad algoritmima nadgledanog učenja.

Imperativ visoke preciznosti predviđanja se eksperimentalno traži pomoću nekoliko metoda optimizacije modela i pripreme podataka, kao što su: podešavanje parametara algoritama, traženje optimalnog broja najvažnijih atributa i balansiranje skupa podataka u odnosu na izlaznu klasu predviđanja.

U odeljcima 3, 4 i 5 će biti predstavljeni načini organizacije, razvoja, struktura projekta kao i metode istraživanja i sam skup podataka nad kojim se vrši istraživanje.

2. RELEVANTNA PRETHODNA ISTRAŽIVANJA

Prilikom istraživanja najpre je revidirana javno dostupna literatura u potrazi za prethodnim sličnim istraživanjima. U ovom poglavlju su predstavljene ideje nekoliko autora koji su pokušali da na osnovu specifičnih faktora predvide kvalitet i cenu stambenih objekta.

Prema [7] bitne informacije o zainteresovanosti klijenata se mogu prikupiti analitikom na samom veb sajtu. Podaci koje korisnici ostavljaju samim pretragama ili kliktanjem na veb stranici mogu biti jako važni za analizu. Neka od prethodnih istraživanja iz oblasti mašinskog učenja su se bavila sličnim problemima, kao što su procene dobrih investicija ili predviđanje cena stambenih objekata na osnovu lokacije, enterijera, eksterijera, ekonomskih faktora na određenim područjima i drugih faktora koji se mogu analizirati iz dostupnih skupova podataka.

U literaturi [8] autori su pokušavali da pronađu dobre ponude na osnovu razlike procenjene i stvarne cene stambenog objekta. Oni su koristili sledeće algoritme mašinskog učenja: *KNN*, *Multi-layer perceptron*, *Ensambles of regression trees* i *SVR*. Najbolju ocenu predviđanja su dobili korišćenjem ansambla stabla odlučivanja (*eng. Ensambles of regression trees*).

U literaturi [9] autor se bavio istraživanjem predviđanja cena stambenih objekata na prostoru *Sjedinjenih Američkih Država*, koristeći skup podataka sa repozitorijuma *Univerziteta u Kaliforniji u Ervajnu* (*eng. University of California, Irvine* – skraćeno *UCI*). Eksperimentalno su evaluirani sledeći algoritmi: *Random Forest*, *Multiple regression*, *SVM*, *Gradient Boosting*, *Neural Networks* i *Ensamble learning* (*bagging*). Najveću ocenu postigao je ansambl algoritam *Slučajnih Šuma* (*eng. Random Forest*).

Drugačiji pristup za predviđanje cena rentiranja koriste autori u literaturi [10], gde su pokušali da korišćenjem prostornog modelovanja objasne lokalnu varijaciju cena zakupnine stana. Smatraju da se iznajmljivanje može predvideti na osnovu prostorne varijacije, gde se na određenom tržištu nalaze uglavnom stambeni objekti sa višom cenom i suprotno od toga, mesta gde se pretežno nalaze objekti sa nižom cenom.

Kao što prethodno pomenuti autori iz literature smatraju [3] i [4], vizuelni efekat stambenog objekta je jako važan indikator kvaliteta. U ovom istraživanju koristimo veštačke neuronske mreže za prepoznavanje objekata na slikama koje bi trebalo da donesu nove informacije o stambenim objektima.

U literaturi [11] autori su eksperimentalno evaluirali različite arhitekture dubokih konvolucionih neuronskih mreža za prepoznavanje i detekciju objekata nad skupovima *Pascal voc 2007*, *Pascal voc 2012* i *Micorosoft COCO*, koristeći kao osnovni model *VGG16* koji je prethodno treniran na skupu *ImageNet*. Autori su koristili sledeće metode:

- *Region Proposal Based Frameworks: (R-CCN(Alex), R-CCN(VGG16), SPP-net(ZF), Fast R-CCN, SutffNet30, NOC, Faster R-CCN (ResNet101), R-FCN(ResNet101), FPN, Mask R-CNN, Mask(ResNeXt101+FPN), Mask (ResNet101+FPN))*
- *One Step Frameworks: (YOLO, YOLOv2, SSD 300, SSD 512)*

Na osnovu dobijenih rezultata, primećeno je da metode *One Step Frameworks* rade dosta brže od metoda *Region Proposal Based Frameworks*. Takođe, metoda *SSD 512* se pokazala kao jedna od vodećih kada je u pitanju preciznost. Ispred ostalih metoda istakle su se: *R-FCN(ResNet101)* i *Faster R-CCN (ResNet101)* kao jedne sa najvećom uspešnosti detektovanja. Na skupu *Microsoft COCO*, najbolje se pokazala metoda *Mask(ResNeXt101+FPN)*.

Kada je u pitanju *SSD* (*Single Shot Detector*) metoda, do sličnih zaključaka je došao i autor [12] na skupu za prepoznavanje vozila, gde *SSD* metoda može da radi na približno 24 *FPS (eng. Frames Per Second),* dok metoda *Faster R-CNN* može da radi na približno 9 *FPS*.

YOLO metoda za detektovanje objekata u realnom vremenu, koju su predstavili [13], u njihovom istraživanju je eksperimentalno evaluirana i upoređena sledećim metodama: *Fastest DPM*, *R-CNN Minus R*, *Fast R-CCN*, *Faster R-CNN VGG16* i *Faster R-CNN ZF*. Najbolje rezultate su dobili sa *YOLO VGG16* metodom, i sa ocenom srednje prosečne preciznosti (*eng. Mean Average Precision* – u daljem tekstu *maP*) 66,4% i 21 *FPS-a*. Autori napominju da ova metoda daje kompetitivne rezultate samo u detektovanju objekata u realnom vremenu.

Metoda *Faster R-CCN(ResNet101)* koja se ne izvršava u realnom vremenu, u istraživanju iz literature [12], postigla je visokih 98% mAR detektovanja na skupu podataka sa saobraćajnim znakovima, nakon 50000 iteracija treniranja, Autor navodi da preciznost dodatno može da se poboljša ukoliko se poveća ukupan broj iteracija treniranja.

U sledećem poglavlju će biti predstavljen skup podataka koji se koristi u ovom istraživanju.

3. SKUP PODATAKA

Kada govorimo o rentiranju, potrebno je da razumemo korisnike stambenih objekata i grupe kojima oni pripadaju [6]. Tipovi zakupaca se mogu grupisati prema njihovim preferencijama. Oni mogu biti porodičnog tipa, studentskog, a dalje se mogu podeliti prema godinama, društvenom staležu, mestu rođenja i još mnogim drugim karakteristikama. U ovom istraživanju se koriste podaci koji obuhvataju sve grupe, odnosno ne postoji konkretna fokus grupa.

Skup podataka je preuzet sa internet lokacije https://www.kaggle.com/c/two-sigma-connect-rental-listing-inquiries/data i sačuvan u *Microsoft SQL Server-u,* u tabelarnom obliku. U pitanju je otvoreno takmičenje koje su organizovale kompa-

nije *Two Sigma* i *RentHop* u cilju poboljšanja usluga prilikom rentiranja stambenih objekata. Inicijalno je preuzet *train.json* fajl u kome se nalaze podaci u *JSON* formatu sa 49352 slučajeva i 15 atributa. Skup sadrži, osam nezavisnih kategoričkih (*building_id, created, description, display_address, features, manager_id, photos, street_address*), šest nezavisnih numeričkih (*bathrooms, bedrooms, latitude, listing_id, longitude, price*) i jedan kategorički zavisan atribut (*interest_level*).

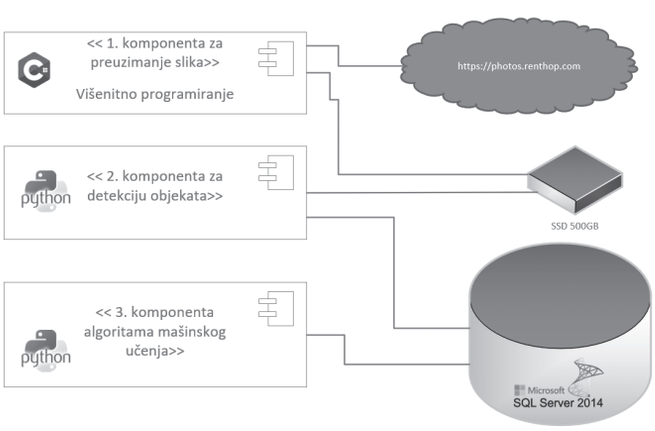
Kolona *features* nema atomske vrednosti, već može sadržati nula ili više zapisa. U te svrhe je kreirana funkcionalnost u projektnom delu rada, koja kreira nove kolone pomoću tehnike *One Hot Encoding,* gde je dobijeno ukupno 1556 novih nedeljivih atributa, koji su u prvoj normalnoj formi (1NF).

Kolona *photos* takođe nema atomske vrednosti, i može imati nula ili više zapisa za jedan slučaj. Svaki zapis predstavlja putanju do internet lokacije slike stambenog objekta. Skup podataka sadrži ukupno 331600 slika stambenih objekata. Prošireni skup podataka, sadrži dodatnih 773 atributa koji predstavljaju različite objekte koji su prepoznati na slikama stambenih objekata.

U sledećem poglavlju predstavljena je organizacija i metode istraživanja projektnog dela rada.

4. ORGANIZACIJA I METODE ISTRAŽIVANJA PROJEKTNOG DELA RADA

U skupu podataka u koloni *Images* nalaze se veb putanje slika za svaki apartman. Te slike su preuzete pomoću posebnog aplikativnog programa napisanog u *C#* programskom jeziku koji predstavlja prvu komponentu arhitekture projektnog dela rada (slika 1).



Slika 1 - *prikaz šire slike arhitekture projektnog dela rada*

Druga komponenta arhitekture (slika 1) predstavlja program napisan u *Python* programskom jeziku, koji je napravljen da obavlja funkcionalnost prepoznavanja objekata na slikama. U okviru druge komponente, koriste se biblioteke za rad sa dubokim neuronskim mrežama *OpenCV, Keras* i *TensorFlow*. Pored pomenutih biblioteka za rad sa neuronskim mrežama i slikama, koriste se i biblioteke *Numpy* i *Pandas* koje predstavljaju strukture podataka višeg nivoa i olakšavaju rad sa matricama.

Za dobijanje novih informacija iz podataka za predviđanje, koriste se duboke neuronske mreže za prepoznavanje objekata na slikama. Zbog skraćivanja vremena potrebnog za izgradnju modela za prepoznavanje objekata na slikama, u istraživanju se koristi model koji je prethodno treniran na skupu *ImageNet*[1]. Broj klasa objekata na slikama stambenih objekata nije unapred poznat, pa je potrebno korišćenje modela sa što većim spektrom klasa za predviđanje. *ImageNet* skup podataka sadrži više od 14 miliona označenih slika [1], koje se mogu mapirati na 1000 različitih izlaznih klasa. Koristi se arhitektura *VGG16* (sa težinama za *ImageNet* skup podataka) koja je u literaturi [13] ocenjena kao jedna od najuspešnijih za klasifikaciju i detekciju objekata na slikama.

Najveći deo posla odrađen je u trećoj komponenti (slika 1) gde je kreiran aplikativni kod u *Python* programskom jeziku za: pripremu podataka za prediktivno modelovanje, implementaciju eksperimentalne optimizacije modela kao i za treniranje modela algoritama i evaluaciju istih.

Nakon preuzimanja skupa podataka, i pratećih slika sa internet lokacija, zadatak treće komponente (slika 1) je priprema podataka za prediktivno modelovanje. Problem nedostajućih podataka rešen je metodom umetanja nedostajućih vrednosti srednjim vrednostima. Za pripremu dobijenih informacija o prepoznatim objektima na slikama, koristi se metoda zamene redova i kolona (*eng. Pivoting*). Zbog preduslova rada sa numeričkim atributima, kategorički atributi su transfomisani u numeričke metodama *One Hot encoding* i *Label encoding,* gde prva metoda predstavlja metodu kreiranja novog *boolean* atributa za svaku klasu i označava njenu prisutnost, a druga metoda jednostavno menja klase u kategoričkom atributu sa jedinstvenim mapiranim numeričkim vrednostima.

U okviru treće komponente (slika 1) se evaluiraju tri osnovna algoritama mašinskog učenja: *KNN, Decision Tree* i *Naive Bayes*. U cilju poboljšanja performansi predviđanja, koriste se i sledeća četiri algoritma ansamba, koji spadaju u metode grupnog odlučivanja [16]: *XGBoost, Gradient Boosting, Adaboost* i *Random Forest*. Ideja je da se grupnim glasanjem poveća ukupna ocena celog ansambla kao jednog konačnog prediktora.

Metodom unakrsne validacije vrši se evaluacija uspešnosti modela različitih prediktora, a kao mere evaluacije višeklasne klasifikacije koriste se: *Accuracy*, *Precision Macro* i *F1 Macro*.

U narednom poglavlju biće predstavljeni rezultati eksperimenata i evaluacija kreiranih modela za predviđanje.

5. EKSPERIMENTALNA EVALUACIJA I DISKUSIJA KREIRANIH MODELA

Najpre su prikazane formule koje su potrebne za razumevanje prethodno pomenutih mera za evaluaciju: *Accuracy*, *Precision macro* i *F1 macro*.

$$\boldsymbol{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

1 ImageNet – javno dostupni skup podataka sa slikama za istraživanje, organizovanih u *WordNet* hijerarhiji [1]

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp+fn}$$

$$F1 = 2\,\frac{Precision * Recall}{Precision + Recall}$$

$$\boldsymbol{Precision\ macro} = \frac{\sum_{i=1}^{n} Precision_i}{n}$$

$$\boldsymbol{F1\ macro} = \frac{\sum_{i=1}^{n} F1_i}{n}$$

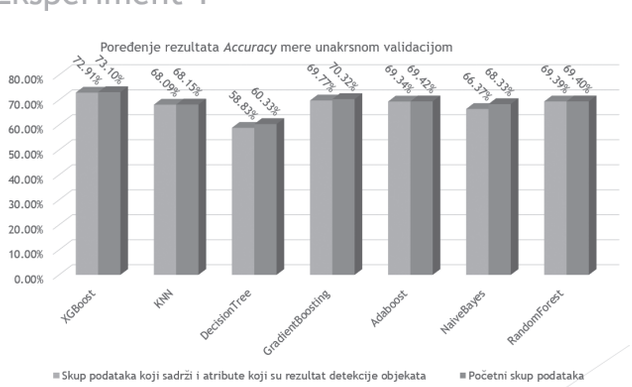
gde su: ***tp*** slučajevi za koje se predviđa da postoji efekat i istina je da efekat postoji, ***tn*** slučajevi za koje se predviđa da efekat ne postoji i istina je da efekat ne postoji, ***fp*** slučajevi za koje se predviđa da efekat postoji a nije istina da efekat postoji, ***fn*** slučajevi za koje se predviđa da efekat ne postoji a istina je da efekat postoji.

U narednom tekstu će biti predstavljeni rezultati evaluacije pet eksperimenata i jednog merenja. U svim eksperimentima modeli su evaluirani merama *Accuracy*, *Precision macro* i *F1 macro.* U eksperimentima 1,2,3 i 5, modeli su validirani metodom unakrsne validacije, dok se u četvrtom eksperimentu koristi po jedan trening i test skup podataka.

5.1 *Prvi eksperiment*

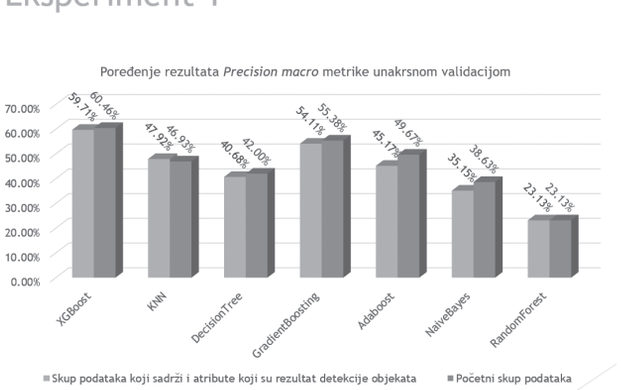
U prvom eksperimentu kreirano je sedam modela klasifikatora na inicijalnom skupu podataka i na proširenom skupu koji uključuje novokreirane atribute.

U prvom eksperimentu, koristeći meru *Accuracy* (slika 2) *XGBoost* algoritam je pokazao najbolju ocenu 73.1% na inicijalnom skupu podataka, dok je najlošiji rezultat pokazao algoritam *Decision Tree.* Kod svih sedam algoritama se može primetiti da su bolje ocene dobijene na inicijalnom skupu.



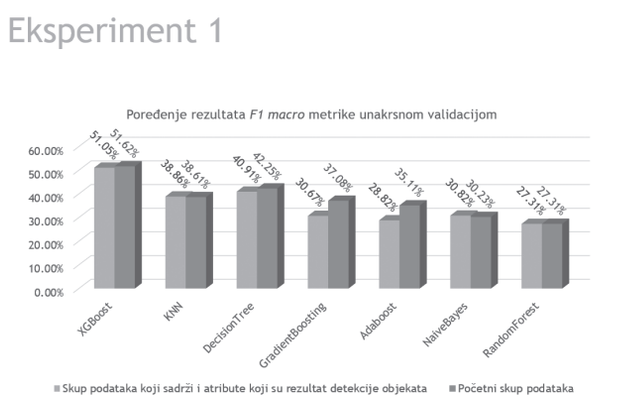
Slika 2 - *eksperiment 1 - mera evaluacije Accuracy*

Na slici 3, unakrsno su validirani rezultati na prethodno pomenuta dva skupa podataka, pomoću mere *Precision macro*. I kod ove mere *XGBoost* je pokazao najviše ocene na inicijalnom skupu podataka 60.46% a na proširenom 59.71%, dok je najlošije ocene pokazao algoritam *Random Forest*. Takođe se može primetiti da su svi algoritmi, osim *KNN-a,* imali veće ocene na inicijalnom skupu. Najveću razliku ocena na inicijalnom skupu i skupu koji sadrži proširene podatke na slici 3, pokazali su algoritmi *Naive Bayes* i *Adaboost*.



Slika 3 - *eksperiment 1 - mera evaluacije Precision macro*

Kod *F1 macro* mere ocenjivanja u prvom eksperimentu (slika 4), najvišu ocenu je pokazao ansambl algoritam *XGBoost* sa 51.62% na inicijalnom skupu i 51.05% na proširenom skupu, dok je najlošiju ocenu pokazao ansambl algoritam *Random Forest*.



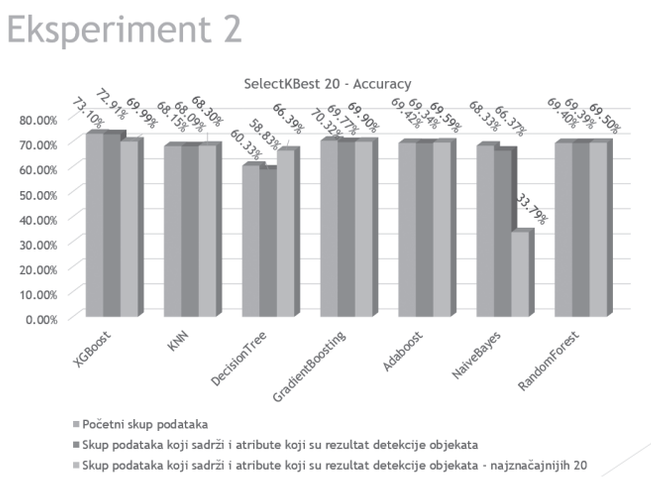
Slika 4 - *eksperiment 1 - mera evaluacije F1 macro*

5.2 DRUGI EKSPERIMENT

U drugom eksperimentu koristi se prošireni skup podataka iz prvog eksperimenta, koji je skraćen metodom *Select K Best,* gde je odabrano 20 najvažnijih atributa. Eksperimentalno su evaluirani modeli nad skupom sa 20 najbitnijih atributa i upoređeni sa rezultatima algoritama iz prvog eksperimenta.

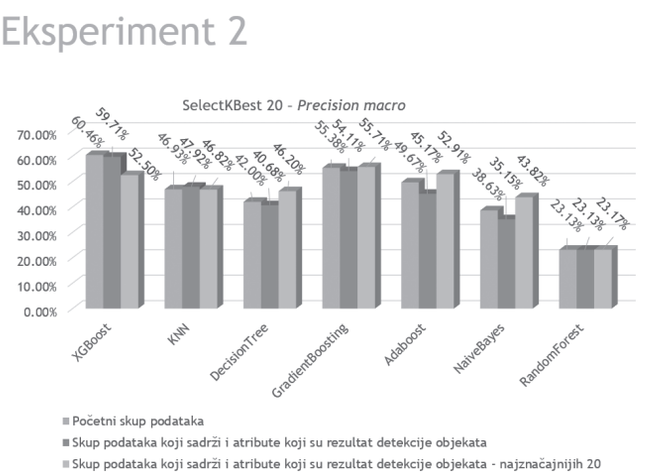
Na slici 5, prikazani su rezultati unakrsne validacije korišćenjem mere evaluacije *Accuracy*, gde je *XGBoost* zadržao najvišu ocenu na inicijalnom skupu podataka sa 73.1%. Može se primetiti da je algoritam *Naive Bayes* pokazao najveći pad performansi u odnosu na rezultate iz prvog eksperimenta (oko 2 puta lošije performanse). Algoritam *Decision Tree* je imao značajna poboljšanja u odnosu na prvi eksperiment, gde je re-

zultat povećan sa 60.33% na 66.39%, gde je u ovom slučaju pokazao da radi bolje sa manje kompleksnim modelom koji sadrži 20 najvažnijih atributa.



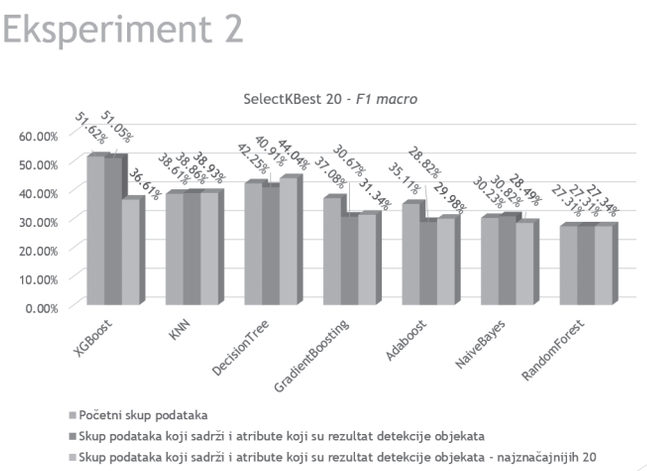
Slika 5 - *eksperiment 2 - mera evaluacije Accuracy*

U drugom eksperimentu kod mere *Precision macro* (slika 6), algoritam *XGBoost* je zadržao vodeću poziciju sa ocenom 60.46% na inicijalnom skupu podataka. Međutim kod drugih modela, došlo je do značajnih poboljšanja. Algoritam *Decision Tree* zabeležio je skok sa 42% na 46.2%, algoritam *Adaboost* sa 49.67% na 52.91%, algoritam *Naive Bayes* sa 38.63% na 43.82%. Iako su navedeni modeli poboljšali performanse, *XGBoost* i dalje ima najvišu ocenu.



Slika 6 - *eksperiment 2 - mera evaluacije Precision macro*

Mera evaluacije *F1 macro* u drugom eksperimentu, najvišu ocenu je imala kod algoritma *XGBoost* sa 51.62% na inicijalnom skupu podataka (slika 7). Značajna poboljšanja je pokazao algoritam *Decision Tree* koji je pokazao skok sa 42.25% na 44.04%. Najveći pad performansi je imao algoritam *XGBoost* na skupu sa 20 najbitnijih atributa (sa 51.62% na 36.61%), na osnovu čega možemo reći da u ovom eksperimentu *XGBoost* algoritam lošije radi sa manjim skupom atributa.

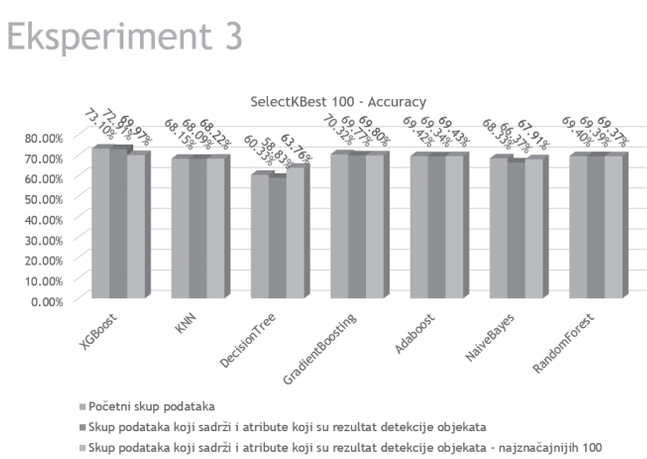


Slika 7 - *eksperiment 2 - mera evaluacije F1 macro*

5.3 *Treći eksperiment*

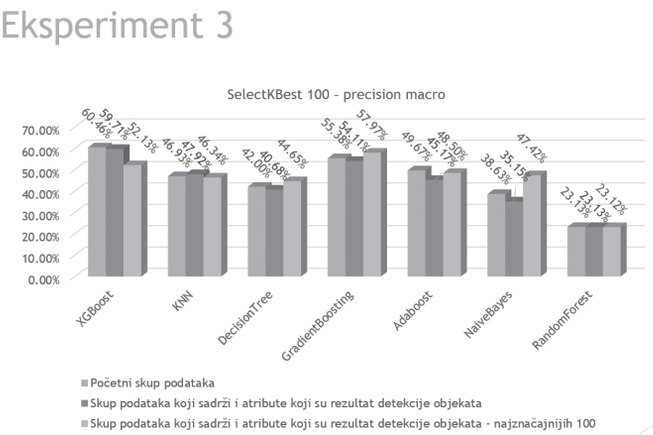
U trećem eksperimentu je takođe korišćen proširenu skup podataka iz prvog eksperimenta, koji je skraćen metodom *Select K Best,* gde je odabrano 100 najvažnijih atributa. Eksperimentalno su evaluirani modeli nad skupom sa 100 najbitnijih atributa i upoređeni sa rezultatima algoritama iz prvog eksperimenta.

Na slici 8, se može primetiti da *XGBoost* algoritam zadržava vodeću poziciju sa 73.1% kada je u pitanju mera *Accuracy*. Kod algoritama *Decision Tree* došlo je do poboljšanja performansi sa 60.33% na 63.76%, ali i bez obzira na poboljšanja, *Decision Tree* algoritam ima najlošiju ocenu u poređenju sa modelima ostalih algoritama.



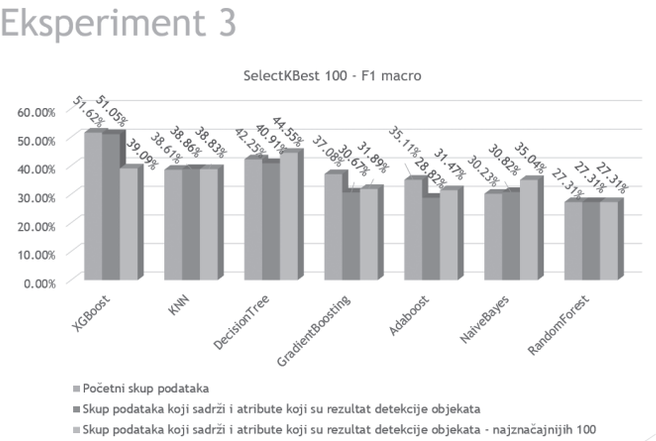
Slika 8 - *eksperiment 3 - mera evaluacije Accuracy*

U trećem eksperimentu na slici 9, evaluirani su modeli pomoću mere *Precision macro*, gde je *XGBoost* algoritam takođe zadržao najvišu ocenu na inicijalnom skupu sa 60.46%, dok je na skupu sa 100 najvažnijih atributa imao 52.13%. Sledeći algoritmi su imali značajna poboljšanja: *Decision Tree* sa 42% na 44.65%, *Gradient Boosting* sa 55.38% na 57.97% i *Naive Bayes* sa 38.63% na 47.42%.



Slika 9 - *eksperiment 3 - mera evaluacije Precision macro*

Kod mere *F1 macro* u trećem eksperimentu (slika 10), *XGBoost* je zadržao najvišu ocenu na inicijalnom skupu podataka, dok je na skupu sa 100 najvažnijih atributa imao ocenu 39.09%. Sledeći algoritmi su imali značajna poboljšanja: Decision Tree sa 42.25% na 44.55% i Naive Bayes sa 30.82% na 35.04%.

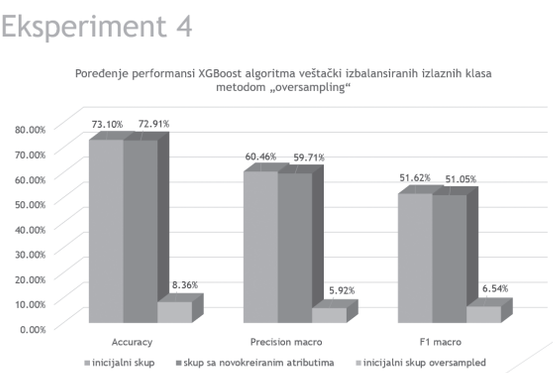


(Slika 10 - eksperiment 3 - mera evaluacije F1 macro)

5.4 *Četvrti eksperiment*

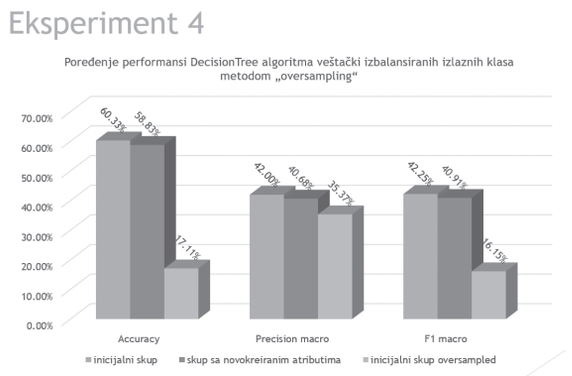
U okviru četvrtog eksperimenta upoređuju se rezultati na skupovima iz prvog eksperimenta, sa rezultatima novokreiranog skupa. Novokreirani skup je nastao korišćenjem metode *oversampling* nad skupom iz prvog eksperimenta koji u sebi sadrži novodetektovane atribute. Ideja je proveriti rezultate modela čiji je skup podataka za učenje izbalansiran, a skup za testiranje nije. U eksperimentu se upoređuju rezultati za algoritme *XGBoost* i *Decision Tree*.

Na slici 11, prikazani su rezultati za algoritam *XGBoost*, gde su dobijena značajna pogoršanja na skupu koji je kreiran metodom *oversampling*. Rezultati su pogoršani više od 7 puta kod svih mera evaluacije.



Slika 11 - *eksperiment 4 - XGBoost*

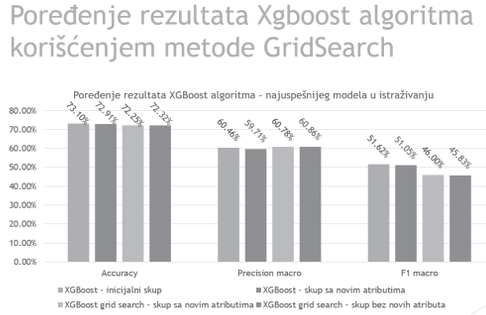
Na slici 12, prikazani su rezultati za algoritam *Decision Tree*, gde su dobijena značajna pogoršanja na skupu koji je kreiran metodom *oversampling* u svim merama evaluacije.



Slika 12 - *eksperiment 4 - Decision tree*

5.5 *Peti eksperiment*

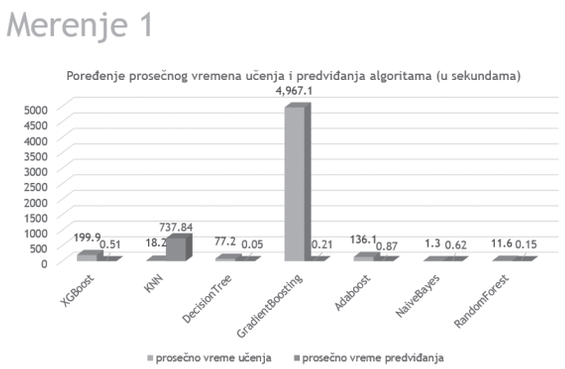
Ideja petog eksperimenta je da se tehnikom *Gread Search* kod najboljeg modela iz prethodnih eksperimenata (*XGBoost algoritam*) eksperimentalno podešavaju parametri: *min_child_weight* [1,5,10], *gamma* [0.5, 1, 1.5, 2, 5], *subsample* [0.6, 0.8, 1.0], *colsample_bytree* [0.6, 0.8, 1.0], *max_depth* [3,4,5]. *Grid Search* metoda je kao najbolju kombinaciju parametara pronašla sledeću vrednosti parametara: *min_child_weight* = 1, *gamma* = 1.5, *subsample* = 0.6, *colsample_bytree* = 0.8, *max_depth* = 5. Na slici 13, rezultati mere *Precision macro* su poboljšani sa 60.46% na 60.78% na skupu podataka sa novim atributima i sa 59.71% na 60.86% na inicijalnom skupu. U svim ostalim kombinacijama skupova i metrika, *Grid Search* tehnika, nije donela poboljšanja, a najbolji rezultati su oni koji su dobijeni na inicijalnom skupu podataka.



Slika 13 - *eksperiment 5 - XGBoost*

5.6 *Merenja*

Na slici 14, prikazana su merenja prosečnog vremena (u sekundama) za učenje i predviđanje svih algoritama. Za proces učenja najviše vremena je trebalo ansambl algoritmu *Gradient Boosting*, sa prosečnih 4967.1 sekundi, a najmanje algoritmu *Naive Bayes* 1.3 sekunde, zbog same prirode algoritma. Za proces predviđanja najviše vremena je trebalo algoritmu *KNN* 737.84, a najmanje algoritmu *Decision Tree* 0.05 sekundi. Takođe se može primetiti da u proseku za proces učenja i predviđanja algoritam *Naive Bayes* iziskuje najmanje vremena.



Slika 14 - *merenje prosečnog vremena učenja i predviđanja modela*

6. ZAKLJUČAK

I u drugom i u trećem eksperimentu *Decision Tree* je pokazao da radi bolje sa manjim podacima, odnosno da metoda *Select K Best* poboljšava uspešnost algoritma, dok je algoritam *XGBoost* pokazao u drugom i u trećem eksperimentu da ima pogoršanje uspešnosti modela, što bi značilo da u ovom istraživanju bolje radi sa kompleksnijim skupovima podataka.

Rezultati testiranja pokazali su da je na inicijalnom skupu podataka *XGBoost* algoritam ostvario najviše ocene predviđanja nivoa zainteresovanosti u svim varijacijama skupova podataka i mera evaluacije. Skup podataka sa novokreiranim atributima nije uspeo da prevaziđe ocene inicijalnog skupa *Accuracy* 73.1%, *Precision macro* 60.46% i *F1 macro* 51.62%.

Korišćenjem metode *Grid Search* izvršena je pretraga najboljih parametara *XGBoost* algoritma, gde su rezultati pokazali da optimizovani parametri nisu uspešniji kod mera ocenjivanja *Accuracy* i *F1 macro*, a da su uspeli da poboljšaju ocene kod mere ocenjivanja *Precision macro* (slika 14) na inicijalnom skupu sa 60.46% na 60.86% i na skupu sa novokreiranim atributima sa 59.71% na 60.78%. U zavisnosti od poslovne potrebe, konačan model bi trebao da se bira na osnovu važnosti mere ocenjivanja, odnosno bio bi izabran onaj model koji ima najvišu ocenu za najvažniju meru evaluacije.

Merenja u toku izvršavanja eksperimenata pokazala su da algoritam *Gradient Boosting* iziskuje najviše vremena za učenje modela, u poređenju sa ostalim algoritmima i po nekoliko stotina puta više. Algoritam *Naive Bayes* je pokazao da je najbrži u fazi treniranja, jer zbog svoje prirode njegovo učenje se dešava u trenutku predviđanja. U fazi predviđanja najmanje vremena je bilo potrebno algoritmu *Decision Tree*, a najviše algoritmu *KNN*.

7. PRAVCI DALJEG ISTRAŽIVANJA

Pored već isprobanih metoda u ovom istraživanju, postoji još nekoliko ideja za budući rad koje bi mogle da poboljšaju ocene predviđanja ili bolje opišu skup podataka. Prikupljanje dodatnih sučajeva i atributa u skupu podataka bi moglo dovesti do pronalaženja novih zakonitosti u podacima koje bi eventualno poboljšale ocene predviđanja. Informacije o stopi kriminaliteta, blizina zdravstvenih i školskih ustanova bi takođe mogli biti dobri indikatori kvaliteta stambenog objekta. Dodatnim podešavanjima parametara pomoću tehnike *GridSearch* bi mogli da pronađemo novu bolju kombinaciju istih. Upotreba *GPU*-a umesto korišćenog *CPU*-a u istraživanju, bi mogla ubrzati proces učenja i predviđanja modela algoritama. Algoritmi mašinskog učenja *Natural Language Processing* (*NLP*) bi mogli pronaći nove informacije iz atributa *description* koji se nalazi u inicijalnom skupu podataka a nije korišćen u izgradnji modela.

8. LITERATURA

[1] http://www.image-net.org/ (posećeno 18.10.2020.)
[2] Agency, N, A., (2008) 721 Housing Quality Indicators (HQI) Form, Housing corporation
[3] Turner, B., (2015) The Book of Rental Property Investing, Denver: BiggerPockets Publishing LLC, Denver, USA
[4] Haight, G, T., (2005) The Real Estate Investment Handbook, New Jersey: John Wiley & Sons, USA
[5] Chaillou, S., Fink, D., Goncalves, P., (2017) Urban Tech on the Rise: Machine Learning Disrupts the Real Estate Industry, Institut Veolia, Artificial Inteligence and Robotics in the City
[6] Stojiljković, B., Jovanović, G., (2010) Potential and Importance of Multi-Family Housing Individualization, Universiti of Niš, The Faculty of Civil Engineering and Architecture, Serbia
[7] Ponniah, P., (2010) Data Warehousing Fundamentals for IT Professionals, New Jersey:John Wiley & Sons, New Jersey, USA
[8] Baldominos, A., Blanco, I., Moreno, A. J., Iturrarte, R., Bernandez, O., Alfonso, C., (2018) Identifying Real Estate Opportunities Using Machine Learning, Madrid, Spain
[9] Ravikumar, S., (2017) Real Estate Price Prediction Using Machine Learning, School of Computing National College of Ireland, Ireland
[10] Sirmans, C. F., Valente, J., Wu, S., Gelfand, A., (2005) Apartment Rent Prediction Using Spatial Modeling
[11] Zhao, Q. Zhong., Zheng, P., Xu, S. T., Wu. Xindong., (2019) Object Detection with Deep Learning: A Review, IEEE
[12] Rosenbrock, A. (2018) Deep Learning for Computer Vision with Python Starter bundle, PyImageSearch.com, 2nd edition
[13] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., (2016) You Only Look Once: Unified, Real-Time Object Detection, available: https://arxiv.org/pdf/1506.02640.pdf, University of Washington USA
[14] Rosenbrock, A. (2018) Deep Learning for Computer Vision with Python Practitioner bundle, PyImageSearch.com, 2nd edition
[15] Rosenbrock, A. (2018) Deep Learning for Computer Vision with Python ImageNet bundle, PyImageSearch.com, 2nd edition
[16] Delibašić, B., Suknović, M., Jovanović, M., (2009) Algoritmi Mašinskog Učenja za Otkrivanje Zakonitosti u Podacima, Narodna biblioteka Srbije, Fakultet Organizacionih Nauka, Beograd, Srbija

Damir Pajaziti
Kontakt: damir.pajaziti111@gmail.com
Oblast interesovanja: softversko inženjerstvo, mašinsko učenje