Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.9
DOI: https://doi.org/10.24867/27BE24Grahovac

ПРЕДИКЦИЈА ДУЖИНЕ БОРАВКА ПАСА И МАЧАКА У ПРИХВАТИЛИШТУ ЗА ЖИВОТИЊЕ

PREDICTING LENGTH OF STAY FOR DOGS AND CATS IN AN ANIMAL SHELTER

Ана Граховац, *Факултет техничких наука, Нови Сад*

Област – РАЧУНАРСТВО И АУТОМАТИКА

Кратак садржај – *У овом раду описан је поступак анализе и обраде података о усвојеним псима и мачкама из прихватилишта за животиње. Упоређивани су различити модели машинског учења на проблему класификације животиња по предвиђеној дужини боравка у прихватилишту.*

Кључне речи: *анализа и истраживање података, машинско учење, класификација*

Abstract – *This paper explores the process of analyzing and processing data on adopted dogs and cats from an animal shelter. Various machine learning models were compared for the classification of animals based on the anticipated length of stay in the shelter.*

Keywords: *data science, machine learning, classification*

1. УВОД

Већина прихватилишта за животиње суочава се са проблемом превелике попуњености. Као резултат тога, азили често немају ресурса да приме незбринуте животиње, или чак пруже адекватну негу онима који се већ у њима налазе. Главни циљ овог рада јесте да повећа шансе за удомљавање животиња из азила, кроз предвиђање дужине њиховог боравка у истом и анализу карактеристика животиња које доводе до брзог удомљавања. Познавање ових информација омогућило би азилима да направе бољу организацију и расподелу расположивих ресурса, као и да кроз различите акције побољша шансе за удомљавање животиња са већом предвиђеном дужином останка.

Скупови података су анализирани са циљем да се утврди повезаност различитих атрибута са дужином боравка у азилу. Дужине останка груписане су у временске периоде, затим су тренирани следећи модели за класификацију: каскадни модел - бинарне класификације, *Gradient Boosting*, *Random Forest*, *XGBoost* и вештачка неуронска мрежа. Резултати су упоређивани коришћењем просечне прецизности, одзива и F1 мере.

2. ПРЕГЛЕД СТАЊА У ОБЛАСТИ

Жеља да се помогне животињама у проналаску сталног дома, као и за растерећењем прихватилишта, подстакла је бројна истраживања у овој области.

Старији радови највећим делом се баве чишћењем и статистичком обрадом података са циљем да се идентификује које су то карактеристике паса и мачака које доводе до лакшег и бржег удомљавања; радови [1, 2] неки од њих. Поменута истраживања се слажу у чињеници да старост паса и мачака једна од кључних карактеристика: најбрже су се усвајале младе животиње, док се са старошћу линерано повећава и дужина боравка. Раса и величина се такође показала као битна, међутим више код паса него код мачака, док су код мачака боја и шаре имале већи утицај. Код мачака је примећена већа потражња за мужјацима, док код паса пол није имао утицаја на брзину удомљавања.

Временом су развој и добијање на популарности метода машинског учења довели до комплекснијих радова, те су, поред анализе пожељних особина за љубимца, све чешћи и покушаји да се на основу података о псу или мачки унапред процени исход или дужина њиховог задржавања у азилу.

Бредли и Рацердан су 2021. спровели истраживање [3] са циљем да повећају шансе за удомљавање животиња кроз две фазе где се прва бави предвиђањем дужине боравка у азилу. Предикција је рађена коришћењем логистичке регресије, вештачке неуронске мреже, *Random Forest* и *Gradient Boosting* алгоритама. Као универзално најбољи показали су се *Gradient Boosting* (макро F1 0.58) и *Random Forest* (макро *recall* 0.65).

Аутори рада [3] су уз табеларне податке из азила, укључили и текстуални опис мачке или пса над којим је вршена анализа сентимента. Сама процена дужине боравка вршена је помоћу следећих модела: логистичка регресија, стабла одлучивања, *Random Forest*, *Gradient Boosting*, и вештачка неуронска мрежа. Најбоље резултате дао је *Gradient Boosting*.

3. ТЕОРИЈСКЕ ОСНОВЕ

У овом поглављу биће представљене теоријске основе алгоритама машинског учења који су коришћени за процену дужине боравка паса и мачака у прихватилишту за животиње.

3.1. Random Forest

Random Forest је алгоритам надгледаног учења. Користи се као класификатор, али и као селектор битних особина или за редукцију димензионалности.

Random Forest повећава робусност система коришћењем већег броја стабала где је циљ да стабла буду што

NAPOMENA:
Овај рад је проистекао из мастер рада чији ментор је био др Александар Ковачевић, ред. проф.

мање међусобно корелирана. Ово се постиже тако што се свако стабло одлучивања тренира на произвољном подскупу обележја и података (*bootstraping*). Када се врши предикција, свако од стабала врши предвиђање и она класа која је имала највише гласова представља коначни излаз (*aggregation*).

Обучвање великог броја независних стабала решава проблем *overfitting*-а, али уједно чини алгоритам захтевнијим за извршавање и смањује могућности његове употребе у реалном времену.

3.2. Gradient Boosting

Gradient Boosting алгоритам одликује висока предиктивна моћ и способност да ухвати комплексна правила и везе међу подацима.

Принцип *boosting* алгоритама јесте да се иницијално одабере неки једноставни модел и помоћу њега се одреди предикција. Такође је потребно изабрати *loss* функцију у складу са захтевима система. Очекивано је да ће излаз оваквог модела много одступати од жељених вредности, међутим, следећи једноставни модел се тренира над грешком излаза претходног модела. Како је у питању *Gradient Boosting* алгоритам за минимизацију грешке се користи *gradient descent*. На овај начин, нови модел надопуњује претходни, и збир ових модела даје тачнија предвиђања од било ког од модела појединачно.

Додатна предност је што је алгоритам релативно робусан на *outlier*-е или недостајуће податке. Ипак, компјутерски је захтеван и тешко интерпретабилан.

3.3. XGBoost

XGBoost (енгл. *Extreme Gradient Boosting*) представља врло ефикасну и скалабилну имплементацију *Gradient Boosted* стабла одлучивања. Овај алгоритам је паралелизацијом успео да значајно скрати време тренирања модела у односу на *Gradient Boosted* стабала. Уз то, уведене су опције за пенализацију прекомплексних модела помоћу L1 и L2 регуларизације, како би се спречио *overfitting*.

3.4. Вештачке неуронске мреже

Вештачке неуронске мреже представљају модел машинског учења инспирисан структуром и функционалношћу људског мозга.

Неуронска мрежа се састоји из вештачких неурона организованих у слојеве. На улазни слој неурона се доводи улаз у систем, док сваки следећи слој као улаз прима излаз претходног слоја неурона.

Појединачни неурон ове вредности обрађује на следећи начин: сумирају се отежињени улази и коефицијент пристрасности, та сума се даље трансформише активационом функцијом и ово представља излаз неурона. Активациона функција се користи како би ограничила излазне вредности неурона и како би се систему дало нелинеарно понашање.

Тренирање неуронске мреже се састоји у томе да се одреди оптималне вредности коефицијента тежина за сваки од неурона, тако да предвиђања мреже буду што ближа стварним вредностима. Пондерисање тежина врши се помоћу алгоритма пропагирања грешке уназад (енгл. *back-propagation*).

Вештачке неуронске мреже имају способност да апроксимирају компликоване и нелинеарне односе међу подацима, међутим, да би успеле да науче ова понашања, потребни су велики обучавајући скупови података. Уз то, оне су „*black box*“ алгоритам, те није лако интерпретирати резултате ових система.

4. МЕТОДОЛОГИЈА

У овом поглављу биће представљен процес имплементације система за предвиђење дужине боравка паса и мачака у азилу за животиње.

Прво је вршена анализе и претпроцесирање скупа података, како би се утврдило који су то атрибути који заправо утичу на исход и смањила димензионалност скупа података. Претпроцесирани подаци даље за тренирање класификационих модела.

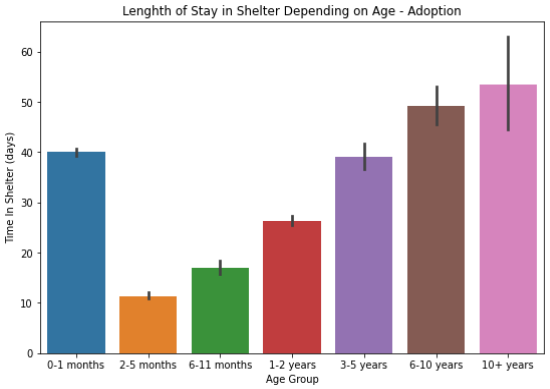
4.1. Експлоративна анализа и претпроцесирање

„*Austen Animal Center*“ скуп података је преузет са *Kaggle*-а [5]. Он садржи информације о животињама које су примљене и отпуштене из азила за животиње у периоду од 2013. до 2018. године. Скуп података има преко 79 хиљада редова описаних 41 атрибутом. Ови атрибути говорили су о старости животиње по уласку и изласку из азила и датуме, по који пут је примљена у азил и у ком стању, типу животиње, боји и раси, месту где је пронађена, да ли је стерилисана и на крају који је њен крајњи исход и колико дуго се задржала у азилу.

Велики број атрибута имао је поновљене информације које су већ садржане у другим атрибутима, те је било потребно избацити сувишне.

Како се преко 94 процента скупа односио на псе и мачке, остатак је избачен у даљем раду са подацима. Уз то, овај рад ће фокусирати само на животиње чији је исход био удомљавање.

Експлоративном анализом потврђено је да дужина боравка у азилу расте са старости животиње, али је постојало знатно одступање за мачиће и кучиће испод два месеца старости. Ово се може објаснити чињеницом да су животиње у том узрасту и даље сувише младе за удомљавање. У складу, пси и мачке су на основу старости категорисани у групе (слика 1).



Слика 1 *Боравак у зависности од старосне групе*

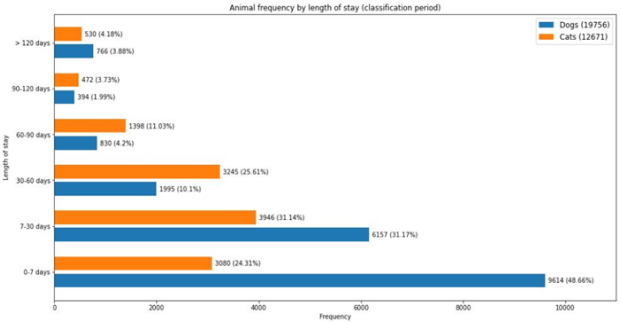
Такође је уочена значајна разлика у броју прихваћених животиња у различитим месецима. Ова разлика нарочито је упадљива код мачака и претпоставка је да то повезано са већим бројем новорођених мачића у периоду април-јун. Ово прати и повећан број усвајања мачака са помереним врхунцем на месеце мај-јул, те ови атрибути могу бити од значаја код предикције усвајања.

Код мачака, обележја која садрже информације о раси и боји су због превеликог броја различитих вредности ручно мапирана на обележја о боји, шарама и дужини крзна са знатно мањим бројем вредности.

Скуп података о псима проширен је другим скупом података „*Dog Breeds*“ [6] који је скупу придодао информације о величини паса, као и неким општим карактерним особинама расе.

4.2. Припрема података за класификационе моделе

Обележје *time_in_shelter_days*, искоришћено је да дужине боравка групишу у следеће категорије: животиње које су у азилу остале од 0-7 дана, 7-30 дана, 30-60 дана, 60-90 дана, 90-120 дана и 120+ дана (слика 2). Овим је циљно обележје трансформисано у одговарајући облик за класификацију.



Слика 2 *Заступљеност паса и мачака по класама*

Како би се подаци из ових скупова података припремили за моделе машинског учења, неопходно је извршити нормализацију. За нумеричке податке употребљена је *Min-Max* нормализација, док је за категоричке коришћен *One Hot Encoding*.

За скуп мачака, изузетак је направљен за месец прихватања у азил, старосну групу и дужину длаке, који су нормализовани на више начина и потом међусобно упоређени. Први начин нормализације за сваки од наведених је помоћу *One Hot Encoding*-а, док ће алтернативни начини бити описани у наставку.

Како истраживања [7] показују да је цикличне податке попут временских одредница често корисно нормализовати помоћу вредности синуса и косинуса, ово употребљено као други начин нормализације месеца уласка у азил. Старост паса и мачака је додатно нормализована прво узимањем средње вредности старости групе, а друго додељивањем свакој групи, од најмлађе до најстарије, броја од 0 до 6, и у оба случаја додата бинарна колона *is_kitten*. Нумеричка вредност је затим нормализована коришћењем *Min-Max* нормализације.

И на крају, дужина длаке је додатно нормализована као нумеричке вредности 0, 0.5 и 1 за вредности *Shorthair*, *Medium Hair* и *Longhair*, ретроспективно.

Са слике 2 уочљиво је да класе нису једнако заступљене у коришћеном скупу података. Обучавање класификационих модела на небалансираном скупу може за последицу имати лошије перформансе модела, те ће бите коришћен SMOTE алгоритам.

4.3. Предиктивни модели

Претпроцесирани подаци о мачкама и псима даље се користе за обучавање класификационих модела за предвиђање дужине боравка у азилу. Како је приликом експлоративне анализе утврђено да исти атрибути немају једнаку важност за усвајање паса и за усвајање мачака, за сваку врсту ће бити одвојено тренирани модели.

Прву групу класификатора представљају ансамбл модели стабала одлучивања. Овде су испробани следећи алгоритми: *Random Forest, Gradient Boosting* и *XGBoost*. Приликом тренирања ових модела подаци су подељени на тренинг и тест скупове, а затим је вршено обучавање модела. Хипер-параметри модела оптимизовани су путем поступка *Grid Search Stratified K Fold Cross Validation* (k = 10).

Други приступ јесте вештачка неуронска мрежа. Процес проналажења оптималне архитектуре и хипер-параметара одвијао итеративно, где је прво креиран веома једноставан модел, да би се затим он усложњавао до проналаска модела који даје најбоље резултате. Овакав модел добијен је са два потпуно повезана слоја од 64 и 32 неурона са *ReLu* активационом функцијом, где сваки слој има *dropout* од 0.5 и примењује L2 регуларизацију, и излазним слојем од 6 неурона (број класа) са *softmax* активационом функцијом. Класа чији неурон има највећу активацију у излазном слоју узима се као резултат предикције.

Како су дубоке неуронске мреже због комплексности архитектуре склоне претренирању, при обучавању је праћена вредност функције грешке на валидационом скупу. Уколико би ова грешка да расла у три узастопне епохе, обучавање.

На скупу података о псима уочено да је заступљеност сваке наредне класе оквирно двоструко мања од претходне, што се може видети са слике 2. Ово је подстакло идеју да се тренира више модела за бинарну класификацију са релативно уравнотеженим класама. Први у низу модел треба да предвиди само да ли ће пас остати мање или више од 7 дана, следећи да ли ће остати 7-30 дана или више од 30 дана. Ова процедура се наставља до последњег модела који треба да предвиди да ли животиња остаје 90-120 дана или више од 120 дана. Као модели бинарне класификације коришћени су претходно описани ансамбл модели. Најбоље резултате овај приступ постигао је када је свих 5 подмодела било *Gradient Boosting* алгоритам са следећим вредностима хипер-параметара.

Скуп података подељен је на тренинг (80%), валидациони (10%) и тест (10%) скуп. Валидациони и тест скуп стратификовани, са циљем да одржи једнака заступљеност класа као у оригиналном скупу.

Финална евалуација је извршена над истим тест скупом за све приступе. За поређење су искоришћене

прецизност, одзив и F1 мера, где је највећи значај дат макро усредњеним вредностима ових мера.

5. РЕЗУЛТАТИ

Најбоље резулате при предвиђању дужине боравка паса у азилу досегао је *Random Forest* алгоритам без аугментације скупа података (*macro avg. precision*: 0.63, *macro avg. recall*: 0.29, *macro avg. F1 score*: 0.3). Неуронска мрежа показала се као алгоритам са најнижом просечном прецизношћу због врло малог броја тренинг примерака најмање заступљених класа.. Просечна прецизност каскадног модела (0.45) боља је од неуронских мрежа, а лошија *Gradient Boosting* алгоритама (0.53 и 0.51).

Како је било више верзија скупова података за мачке, као најбољи показао се скуп података у коме су месец пристизања мачке у азил и дужина длаке *One Hot* кодирани, док је старосне групе мачака претворене нумеричке вредности од 0 до 6, од најмлађе до најстарије групе редом, и потом *Min-Max* нормализоване и додата колона *is_kitten*.

Над овим скупом најбоље резултате *Gradient Boosting* алгоритам без аугментације скупа података (*macro avg. precision*: 0.71, *macro avg. recall*: 0.35, *macro avg. F1 score*: 0.37). Исту прецизност има и *Random Forest*, али има мању макро F1 меру и просечни одзив. Као и код предвиђања дужине боравка паса у азилу, неуронска мрежа се и у случају мачака показала као алгоритам са најмањом просечном прецизношћу, поново због јако малог броја тренинг примерака најмање заступљених класа.

За све моделе над којим је извршено SMOTE узорковање порастао је одзив, а опала прецизност.

Матрице конфузије резултата указују на то да примерци класе 7-30 дана и класе 30-60 дана најчешће бивају замењени, и за податке мачака и паса.

Због сличности приступа решавања и циљног обележја овог и истраживања [3], може се направити паралела између крајњих резултата ових радова. Иако аутори не користе исти скуп података, моделе обучавају над скупом података који обухвата псе и мачке истовремено и имају другачије класе (0-8, 8-42, 42-365 и 365+ дана) утврђено је да су добијени резултати релативно слични. И у овом раду најбоље се показао *Random Forest* алгоритам, код кога су просечни одзив и F1 мера бољи него код модела приказаних у овом раду, док је најбоља просечна прецизност 0.59 лошија од најбоље просечне прецизности за псе (0.63) као и за мачке (0.71) приказане у овом раду.

6. ЗАКЉУЧАК

У раду је предложен систем за предвиђање дужине боравка паса и мачака у прихватилишту за животиње.

Са овим циљем анализиран је скуп података о удомљеним животињама из *Austin Animal Center* азила, додатно проширен подацима који садрже карактеристике паса. Након претпроцесирања података и одабира атрибута, креирани су одвојени модели за класификацију паса и мачака по дужини боравка. Испробани су *Random Forest, Gradient Boosting, XGBoost*, и вештачка неуронска мрежа. За класификацију паса је креиран и каскадни модел бинарних класификација. Најбољи резултати за скуп података са псима постигнути су уз помоћ *Random Forest* алгоритма чија је просечна прецизност била 0.63, док за скуп података са мачкама највећу просечну прецизност од 0.71 има *Gradient Boosting*.

Примећено је да модели код обе врсте животиња примерке чија дужина боравка спада у групу 0-7 дана често мешају са примерцима чија дужина боравка спада у групу 7-30 дана и обрнуто. Развој система у будућности могао би се усмерити на идентификацију додатних предиктора који ће омогућити тачније раздвајање ове две групе. Проширивање скупа описом карактера животиње и њеном фотографијом, такође би повећали прецизност одређивања удомљивости. Када би се уз то укључили и подаци из више установа на различитим локацијама, могли би се пратити и различити трендови усвајања у зависности од локације, што би дало могућност релокације животиње у установе где би имале веће шансе за удомљавање.

7. ЛИТЕРАТУРА

[1] W.P. Brown, J.P. Davidson, M.E. Zuefle, „Effects of phenotypic characteristics on the length of stay of dogs at two no kill animal shelters“, *Journal of Applied Animal Welfare Science*, 16(1), 2-18, 2013.

[2] W.P. Brown, K.T. Morgan, „Age, breed designation, coat color, and coat pattern influenced the length of stay of cats at a no-kill shelter“. *Journal of Applied Animal Welfare Science*, 18(2), 169-180, 2015.

[3] J. Bradley, S. Rajendran, „Increasing adoption rates at animal shelters: A two-phase approach to predict length of stay and optimal shelter allocation“, *BMC Veterinary Research*, 17, 1-16, 2021.

[4] A. Zadeh, K. Combs, B. Burkey, J. Dop, K. Duffy, Nosoudi, „Pet analytics: Predicting adoption speed of pets from their online profiles“, *Expert Systems with Applications*, 204, 117596, 2022.

[5]https://www.kaggle.com/datasets/aaronschlegel/austin-animal-center-shelter-intakes-and-outcomes (приступљено у новембру 2023)

[6] https://www.kaggle.com/datasets/yonkotoshiro/dogs-breeds (приступљено у новембру 2023)

[7] https://towardsdatascience.com/ml-intro-5-one-hot-encoding-cyclic-representations-normalization-6f6e2f4ec001 (приступљено у новембру 2023)

Кратка биографија:

Ана Граховац рођена је у Новом Саду 1999. године. Мастер рад на Факултету техничких наука из области Рачунарство и аутоматика – Интелигентни системи одбранила је 2023. године.

контакт: grahovac.ana99@gmail.com