SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET
Diplomski studij računarstva

Izvješće

Projekt - Strojno učenje

Rijeka, veljača 2023. Ivan Rubinić

Sadržaj

Popis slika

Poglavlje 1

Uvod

Tema projektnog zadatka iz kolegija Strojno učenje je istražiti postupke objašnjivosti (engl. *explainability*) i obrezivanja (engl. *pruning*) modela strojnog učenja. Uvodno poglavlje izvješća opisuje iz čega proizlazi potreba za svakim od spomenutih mehanizama, a nastavak izvješća pruža uvid u pregled istraženih metoda i stvorenih demonstracija kao prilog istih.

1.1 Potreba za objašnjivosti modela strojnog učenja

U doba sve učestalijeg korištenja sve kompleksnijih modela strojnog učenja u sve širem spektru zadaća i poslova, pojavila se potreba za pružanjem uvida u načine na koje modeli dolaze do svojih odluka ili predikcija. Postoji puno razloga zbog kojih je poželjno postići zadovoljavajuću razinu objašnjivosti korištenog modela strojnog učenja, a neki od glavnih su [3]:

- korištenje sustava temeljenih na strojnom učenju u osjetljivim područjima – primjerice, za sustave korištene za klasifikaciju medicinskih slika u svrhu detekcije malignih bolesti mora postojati određena vrsta objašnjivosti svake odluke (inače se smanjuje vjera u točnost i pouzdanost sustava),
- uklanjanje odluka pod utjecajem diskriminacije – kako bi korisnici imali povjerenja u činjenicu da model strojnog učenja odlučuje bez diskriminacije prema bilo kakvim značajkama, mora postojati objašnjivost svake njegove odluke kako bi se moglo potvrditi da ona ne diskriminira po bilo kojoj osnovi,

- sporovi u kojima su odluke sustava temeljenih na strojnom učenju značajni faktori zahtjevaju objašnjivost istih.

1.2 Potreba za obrezivanjem modela strojnog učenja

Potreba za obrezivanjem modela strojnog učenja javlja se iz sličnog uzroka kao i potreba za objašnjivosti istih – zbog sve kompleksnijih modela za sve šire spektre zadaća koji postaju sve zahtjevniji za uređaje na kojima se koriste. Ako se tome pribroji porast korištenja sustava strojnog učenja na mobilnim uređajima, u raznim ugradbenim sustavima, robotima i sl., obrezivanje modela strojnog učenja postaje vrlo važno. Obrezivanjem modela smanjenjujemo broj parametara istog što u konačnici dovodi do smanjenja veličine i ubrzavanja procesa donošenja odluke [4]. Također, usprkos postupcima provedenim u procesu obrezivanja modela, postignuti se rezultati originalnog, neobrezanog modela uglavnom ne narušavaju.

Poglavlje 2

Objašnjivost modela strojnog učenja

Kako je već spomenuto u uvodu, ovaj projektni zadatak uključuje istraživanje različitih metoda koje nude razna sredstva u svrhu postizanja određene razine objašnjivosti osobina korištenih modela strojnog učenja. Ovo poglavlje se dotiče same definicije objašnjivosti modela strojnog učenja, njene razlike u usporedbi s interpretabilnosti modela strojnog učenja te pregleda istraženih metoda koje pružaju sredstva za objašnjenje s pripadnim demonstracijskim primjerima. Priloženi primjeri su stvoreni koristeći model slučajne šume (engl. *random forest*) i podatke o broju iznajmljenih bicikala [5] kako bi se mogli usporediti s ilustracijama danim u literaturi.

2.1 Definicija

Objašnjivost modela strojnog učenja se može definirati kao proces objašnjavanja unutarnjih zakona modela na način razumljiv za ljude. Objašnjivost modela vodi k otkrivanju veza između ulaznih i izlaznih podataka te načina na koji promjene vrijednosti ulaznih podataka utječu na promjene izlaznih vrijednosti. [6] Objašnjivost modela strojnog učenja postiže se korištenjem raznih tehnika i metoda od kojih su neke opisane u nastavku ovog poglavlja.

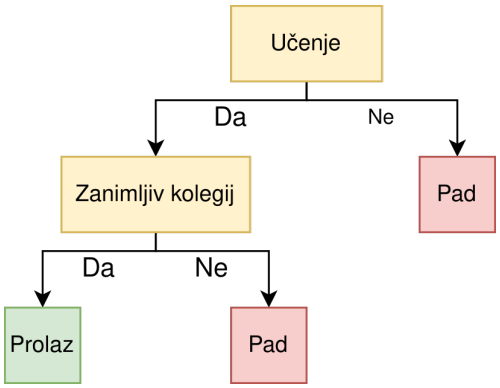
2.1.1 Objašnjivost nasuprot interpretabilnosti

Važno je razlikovati objašnjivost i interpretabilnost modela strojnog učenja. Model strojnog učenja je interpretabilan ukoliko je razumljiv za ljude u svojem izvornom

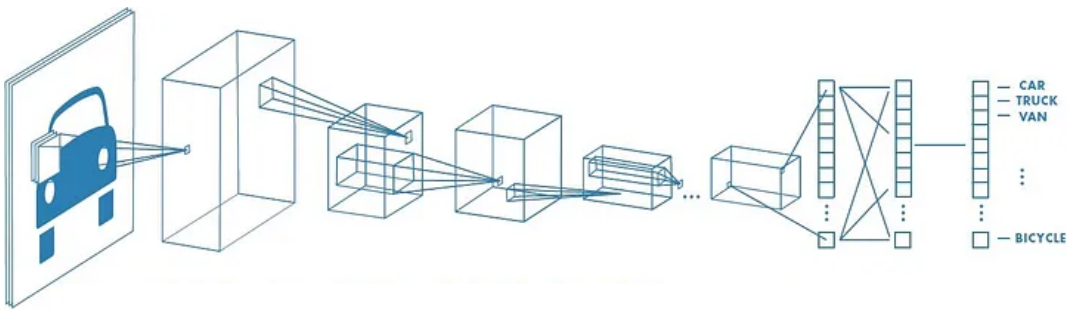
obliku, bez dodatnih metoda koje ga pojašnjavaju. Drugim riječima, ljudi mogu odluke i razloge koje su doveli do odluka interpretabilnog modela razumijeti analizom njegovih parametara ili arhitekture. [1]

S druge strane, ukoliko model nije interpretabilan, njegove odluke i zakonitosti ne mogu se pojmiti isključivo analizom parametara i arhitekture – za to su tada potrebne tehnike za objašnjivost modela strojnog učenja.

Razlika interpretabilnog i neinterpretabilnog modela je dana na slici 2.1 na stranici 4.



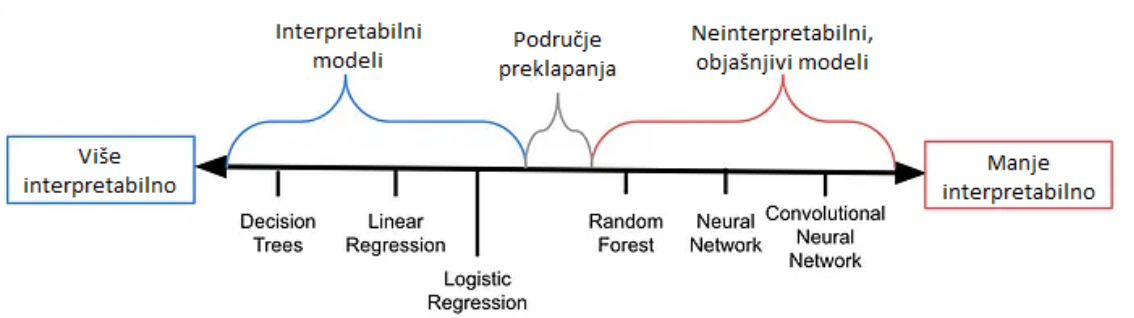
(a) Intepretabilan model – stablo odluke



(b) Neinterpretabilan model – konvolucijska neuronska mreža [7]

Slika 2.1 Iz arhitekture i parametara interpretabilnog modela posve je jasno na temelju čega model dolazi do kakve odluke. S druge strane, kako bismo isto utvrdili za neinterpretabilan model, potrebno je iskoristiti tehnike koje nude objašnjivost takvih modela.

Važno je napomenuti da su interpretabilnost i složenost modela strojnog učenja uglavnom obrnuto proporcionalni – što je model složeniji, to je manje interpretabilan i obratno – prikaz na slici 2.2 na stranici 5.



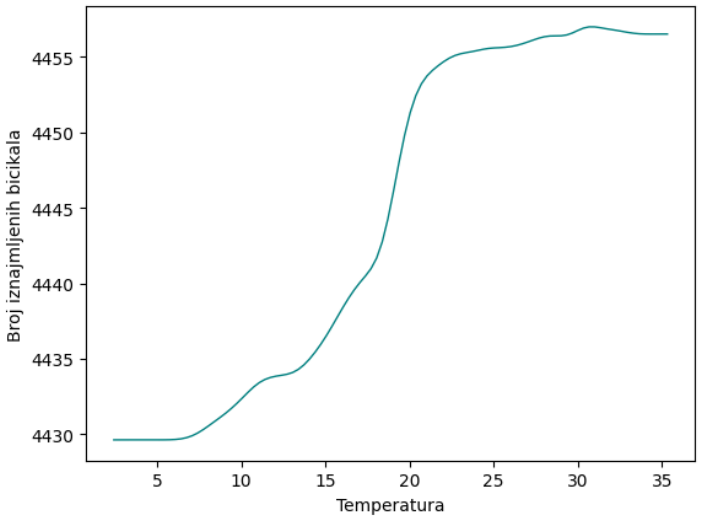
Slika 2.2 Odnos interpretabilnosti i složenosti modela. [1]

2.2 Partial Dependence – PD

PDP (engl.*Partial Dependence Plot*) prikazuje marginalni učinak jedne ili dvije značajki na predviđeni rezultat modela strojnog učenja [8]. Vrijednost PD funkcije je određena prosječnim predviđanjem modela strojnog učenja u slučaju kada se vrijednosti promatrane značajke svih primjeraka skupa podataka zamijene željenom vrijednosti. Izraz prema kojemu se dobiva vrijednost PD funkcije za željenu vrijednost promatrane značajke S glasi $\hat{f}_S(x_S) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(x_S, x_C^{(i)})$ [2], a uobičajeno se računa kroz cijelu ili dio domene promatrane značajke S.

Prednosti PDP-a su intuitivnost i lakoća implementacije. Unatoč tome, PDP ima velik problem kod značajki koje nisu neovisne – primjerice, za računanje PDP-a na slici 2.3 je ovisnost temperature o godišnjem dobu potpuno zanemarena te su za računanje PDP-a za vrijednost visokih temperatura i primjerci iz skupa podataka koji opisuju zimske dane ravnopravno uključeni u izračun prosječne vrijednosti predikcije što stvara iskrivljenu interpretaciju. Nadalje, budući da PDP podrazumijeva usrednjavanje izračunatih vrijednosti, nasuprotni utjecaji će se međusobno poništiti te tako mogu biti nezamijećeni. Još jedan nedostatak PDP-a je nemogućnost vizualiziranja utjecaja više od dvije značajke istovremeno. [2]

Primjer PDP-a je prikazan na slici 2.3. Pregledom PDP-a utvrđuje se da porastom temperature raste i broj iznajmljenih bicikala.



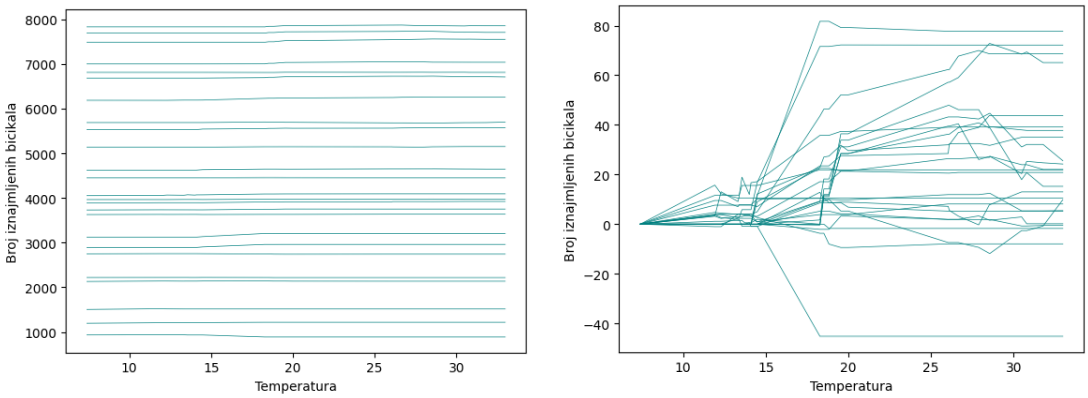
Slika 2.3 Partial Dependence Plot za značajku koja opisuje temperaturu.

2.3 Individual Conditional Expectation – ICE

ICE prikazi su ekvivalenti PDP-u, ali prikazuju svaki primjerak skupa podataka zasebnom krivuljom. Drugim riječima, za svaki primjerak skupa podataka $\{(x_S^{(i)}, x_C^{(i)})\}_{i=1}^N$ se prikazuje krivulja $\hat{f}_S^{(i)}$ uz neizmijenjene vrijednosti $x_C^{(i)}$ i vrijednosti $x_S^{(i)}$ iz domene značajke S. [2]

Prednosti ICE prikaza su, kao i kod PDP-a, intuitivnost i lakoća izračuna, ali i mogućnost otkrivanja nasuprotnih utjecaja koji kod PDP-a zbog usrednjavanja izračunatih vrijednosti mogu proći nezamijećeno. Uz nedostatke PDP-a koji se tiču zanemarivanja ovisnosti između značajki i broja značajki koje je moguće istovremeno prikazati, ICE prikazi uglavnom nisu pregledni zbog broja primjeraka unutar skupa podataka.

Primjer ICE-a je prikazan na slici 2.4. Pregledom ICE-a utvrđuje se da nema primjeraka unutar skupa podataka koje se ponašaju značajno drugačije od prosjeka te da PDP dobro ilustrira prosječnu situaciju. Slika 2.4 prikazuje i Centered ICE prikaz kojim se često jednostavnije uspoređuju primjerci skupa podataka s obzirom na promatranu značajku.



(a) Individual Conditional Expectation Plot za značajku koja opisuje temperaturu.

(b) Centered Individual Conditional Expectation Plot za značajku koja opisuje temperaturu.

Slika 2.4 Individual Conditional Expectation Plot.

2.4 Accumulated Local Effects – ALE

ALE opisuje kako značajke utječu na predikciju modela strojnog učenja u prosjeku. Za razliku od PD-a i ICE-a, ALE u izračun uzima samo primjerke s realnim kombinacijama značajki: [2]

- PD – što model u prosjeku predviđa kada se promatrana značajka za svaki primjerak postavi na željenu vrijednost,
- ALE – kako se predviđanja modela mijenjaju u uskom području značajke oko željene vrijednosti za primjerke skupa podataka unutar tog područja.

ALE vrijednost se računa prema izrazu

$$\hat{\tilde{f}}_{j,ALE}(x) = \sum_{k=1}^{k_j(x)} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[f(z_{k,j}, x_{\setminus j}^{(i)}) - f(z_{k-1,j}, x_{\setminus j}^{(i)}) \right]$$

koji se može podijeliti na tri dijela:

1. **e**ffect – crveni dio izraza – razlika u predikcijama (kako bi se u obzir uzeo učinak na predikciju isključivo promatrane značajke) nad primjercima kojima je značajka j postavljena na vrijednost z_k, odnosno z_{k-1} koja označava granicu trenutnog područja,

2. **l**ocal – zeleni dio izraza – izračun prosječne razlike u predikcijama svih primjeraka skupa podataka unutar promatranog područja (dakle samo onih realnih primjeraka – razlika u odnosu na PD),
3. **a**ccumulated – plavi dio izraza – zbrajanje učinka kroz sva definirana područja.

Vrijednosti dobivene navedenim izrazom se centriraju kako bi prikazivale odstupanje od prosječne predikcije – $\hat{f}_{j,ALE}(x) = \hat{\tilde{f}}_{j,ALE}(x) - \frac{1}{n}\sum_{i=1}^{n} \hat{\tilde{f}}_{j,ALE}(x_j^{(i)})$. [2]

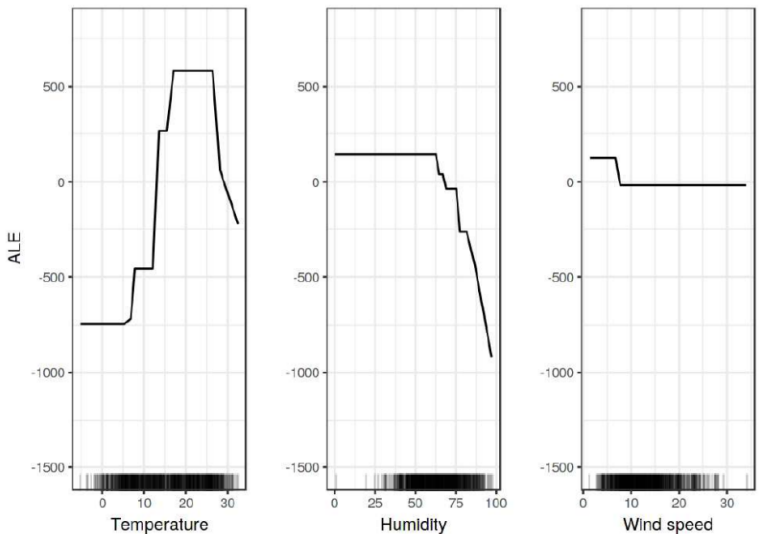
U konačnici, ALE vrijednost se treba shvatiti kao učinak značajke pri određenoj vrijednosti u usporedbi s prosječnom predikcijom modela. Primjerice, ALE vrijednost 15 za $x_j = 11$ znači da je pri takvoj vrijednosti značajke x_j predikcija za 15 veća u odnosu na prosječnu predikciju modela.

ALE prikazi donose nekoliko prednosti:

- rade i kada postoji međudjelovanje kod značajki skupa podataka,
- izračun je brži u odnosu na PD prikaze,
- interpretacija ALE prikaza je jasna i intuitivna.

Unatoč velikim prednostima, nedostaci ALE prikaza se manifestiraju u vidu određivanja broja korištenih područja prilikom izračuna te mnogo složenije implementacije u odnosu na PD i ICE prikaze. [2]

Primjer ALE prikaza je dan na slici 2.5. Pregledom ALE prikaza može se utvrditi da temperatura ima najveći utjecaj, a brzina vjetra najmanji utjecaj na prosječnu predikciju modela.



Slika 2.5 ALE prikaz za značajke koje opisuju temperaturu, relativnu vlažnost i brzinu vjetra. [2]

2.5 Međudjelovanje značajki (engl. *feature interaction*)

Međudjelovanje između dvije značajke je promjena u predikciji modela koja nastaje zbog specifične kombinacije vrijednosti značajki nakon uzimanja individualnih doprinosa značajki u obzir. [2]

Primjerice, ukoliko je cijena automobila određena značajkama potrošnje i snage (prikaz 2.6):

- uz izostanak međudjelovanja značajki potrošnje i snage, ukupna cijena automobila će biti zbroj početne cijene automobila te individualnih doprinosa cijeni temeljem potrošnje i snage (gornji dio prikaza 2.6),
- uz međudjelovanje značajke potrošnje sa značajkom snage, ukupna cijena automobila će biti zbroj početne cijene automobila, individualnih doprinosa obje značajke i doprinosa koje nosi njihovo međudjelovanje u određenoj kombinaciji – npr. brz auto s malom potrošnjom doprinosi cijeni s 5000 EUR (donji dio prikaza 2.6).

Razina međudjelovanja značajki se računa na jednostavan način koristeći H-

Potrošnja	Snaga	Cijena
Niska	Mala	23.500 EUR
Niska	Velika	27.500 EUR
Visoka	Mala	20.000 EUR
Visoka	Velika	24.000 EUR

(a) Cijena automobila bez međudjelovanja značajki potrošnje i snage.

Potrošnja	Snaga	Cijena
Niska	Mala	22.500 EUR
Niska	Velika	35.000 EUR
Visoka	Mala	15.000 EUR
Visoka	Velika	25.000 EUR

Potrošnja	Snaga	Utjecaj
Niska	Mala	-1.000 EUR
Niska	Velika	7.500 EUR
Visoka	Mala	-5.000 EUR
Visoka	Velika	1.000 EUR

(b) Cijena automobila uz međudjelovanja značajki potrošnje i snage te doprinosi pojedinih kombinacija međudjelovanja.

Slika 2.6 Međudjelovanje značajki.

statistiku. Moguće je izračunati razinu međudjelovanja jedne značajke sa svim ostalim značajkama ili međudjelovanje dvije značajke.

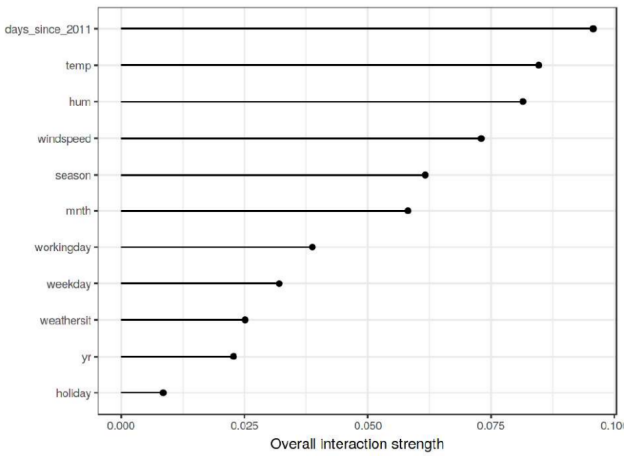
Ukoliko značajka j ne međudjeluje s niti jednom drugom značajkom, predikcija modela se može izraziti kao zbroj PD funkcija od kojih jedna ovisi samo o značajki j, a druga o svim ostalim značajkama:

$$\hat{f}(x) = PD_j(x_j) + PD_{-j}(x_{-j})$$

. U sljedećem koraku se promatra razlika između dobivenih izlaza iz modela i pretpostavljene funkcije ukoliko značajka j ne međudjeluje s niti jednom drugom značajkom što otkriva razinu međudjelovanja. [2] Na sličan način se dobiva razina međudjelovanja između para značajki.

Prednosti ove metode su što se detektiraju sve vrste međudjelovanja, a rezultat je broj unutar intervala $[0, 1]$ što metodu čini usporedivom između više različitih modela i značajki. Unatoč tome, metoda je računski zahtjevna te uključuje procjenu marginalnih distribucija što doprinosi nestabilnosti rezultata. [2]

Primjer prikaza međudjelovanja značajki sa svim ostalim značajkama je dan na slici 2.7.



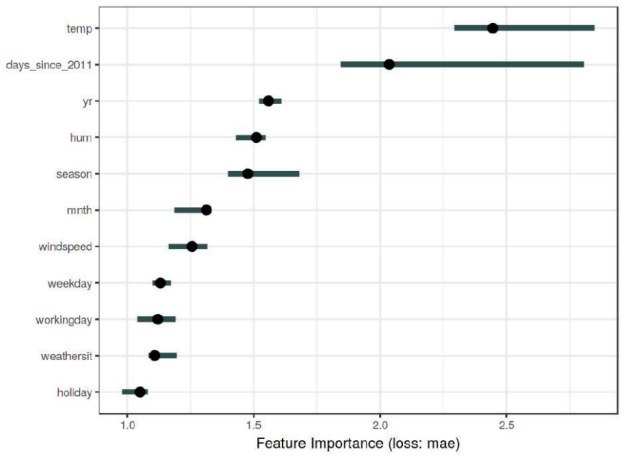
Slika 2.7 Prikaz međudjelovanja pojedinih značajki sa svim ostalim značajkama iz skupa podataka za procjenu dnevnog broja iznajmljenih bicikala. [2]

2.6 Važnost značajki (engl. *feature importance*)

Razina važnosti značajke je povećanje u greški modela strojnog učenja nakon permutiranja vrijednosti značajke čime se postiže učinak uništenja veze između značajke i stvarnog rezultata. Način na koji se razina važnosti pojedine značajke računa je promatranje greške modela strojnog učenja koja nastaje nakon spomenute permutacije – što je značajka važnija, to je greška modela nakon permutacije njenih vrijednosti veća. [2]

Prednosti ove metode su što se interpretira na jednostavan način, usporediva je kroz različite vrste problema te u obzir uzima sve vrste interakcija s ostalim značajkama. Najveća zamjerka je što nije potpuno jasno treba li se računati na testnom ili trening skupu podataka te što postoje razlike u rezultatima s obzirom na način permutiranja vrijednosti značajki unutar dostupnog skupa podataka. [2]

Primjer prikaza važnosti značajki je dan na slici 2.8 te se iz njega može zaključiti da je u skladu s navedenim postupkom najveća važnost od strane modela strojnog učenja pridana značajki koja opisuje temperaturu.



Slika 2.8 Prikaz važnosti pojedinih značajki iz skupa podataka za procjenu dnevnog broja iznajmljenih bicikala. [2]

2.7 Globalni surogat modeli

Globalni surogat model je interpretabilan model čija je svrha aproksimirati izlaz modela koji se želi pojasniti – temeljem takvog modela može se zaključivati na koji se način ponaša izvorni, neinterpretabilni model. Nekoliko je koraka za stvaranje globalnog surogat modela (GSM) [2]:

- odabir skupa podataka na kojem će se trenirati GSM,
- izračun predikcija za odabrani skup podataka koristeći model strojnog učenja koji želimo pojasniti,
- odabir interpretabilnog modela koji će služiti za GSM i trening istog na osnovu skupa podataka i prethodno izračunatih predikcija.

Tome slijedi evaluacija nastalog GSM-a i njegova interpretacija. Ukoliko GSM izrazito dobro aproksimira izvorni model, isti se može čak i zamijeniti GSM-om.

Prednost ovog pristupa je fleksibilnost GSM-a – mogu se koristiti različiti interpretabilni modeli strojnog učenja, čak i ako se izvorni model koji se pojašnjava značajno promijeni. Ipak, za korištenje ovog pristupa potrebno je utrošiti vrijeme na trening GSM-a te rezultati mogu ovisiti o podskupu korištenih podataka.

2.8 Lokalni surogat modeli - LIME

LIME (engl. *Local Interpretable Model-Agnostic Explanations*) je implementacija lokalnih surogat modela – cilj je umjesto jednog GSM imati više lokalnih surogat modela koji objašnjavaju zasebne predikcije.

LIME se stvara u nekoliko koraka [2]:

1. odabir primjerka za kojeg se želi prikupiti objašnjenje,
2. permutacija dataseta i prikupljanje predikcija izvornog modela za nove primjerke,
3. ponderiranje novih primjeraka temeljem njihove podudarnosti s odabranim primjerkom,
4. treniranje ponderiranog, interpretabilnog modela na novostvorenom skupu podataka za učenje.

Najveća prenost LIME-a je mogućnost rada s tabularnim, tekstualnim i slikovnim podacima. Loša strana ovog pristupa je nestabilnost i nemogućnost ponovljivosti objašnjenja [2].

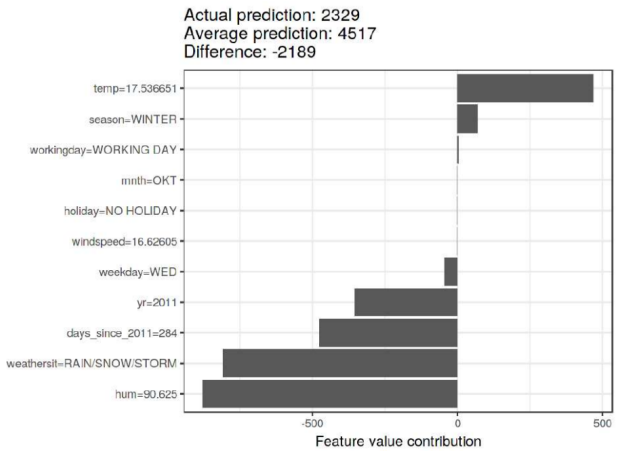
2.9 Shapely vrijednosti

Shapely vrijednost iskazuje koliko pojedina značajka doprinosi određenom rezultatu modela strojnog učenja.

Cilj Shapely vrijednosti je objasniti razliku između prosječne predviđene vrijednosti modela i predviđene vrijednosti modela, odnosno objasniti koja je značajka imala kakav utjecaj u toj razlici. [2]

Shapely vrijednost se za pojedinu značajku određuje izračunom njenog prosječnog doprinosa koristeći sve kombinacije svih ostalih značajki.

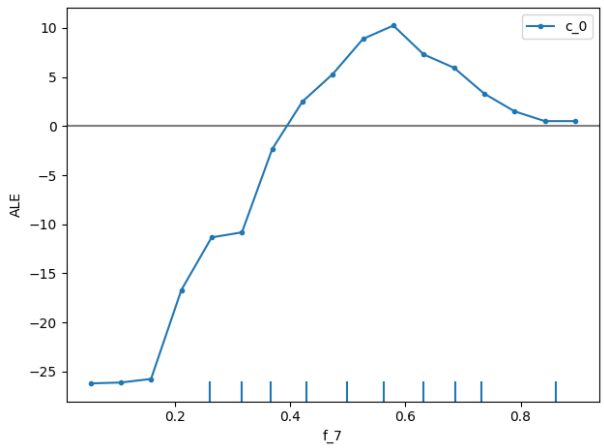
Primjer prikaza Shapely vrijednosti je dan na slici 2.9 te se iz njega može zaključiti da je najnegativniji utjecaj na predikciju modela doneslo loše vrijeme i velika vlažnost.



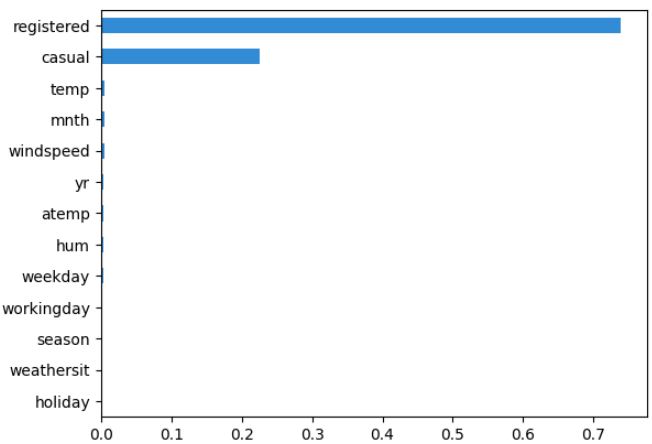
Slika 2.9 Primjer prikaza Shapely vrijednosti za predikciju iz skupa podataka za procjenu dnevnog broja iznajmljenih bicikala. [2]

2.10 Problemi

Prilikom rada na projektnom zadatku, isprobane su važnosti značajki, ALE, ICE i PD metode za objašnjivost modela strojnog učenja. Dok su rezultati vlastite implementacije PD i ICE metoda prikazane u prethodnim potpoglavljima, korištenjem paketa za izračun važnosti značajki (slika 2.11) i ALE vrijednosti (2.10) dobiveni su rezultati koji se nisu poklapali s korištenom literaturom te zato nisu uključeni u prethodna potpoglavlja.



Slika 2.10 Prikaz rezultata ALE metode – pogrešne vrijednosti na Y osi. X os sadrži skalirane vrijednosti temperatura (podijeljene s 41).



Slika 2.11 Prikaz rezultata metode važnosti značajki – pogrešne vrijednosti razine značajnosti značajki.

Poglavlje 3

Obrezivanje modela strojnog učenja

Ideja iza obrezivanja modela strojnog učenja je smanjenje broja parametara neuronskih mreža kako bi njihove predikcije bilo moguće proračunati u kraćem vremenskom roku uz manje potrebnih resursa. Obrezivanje neuronskih mreža može se zasnivati na odbacivanju utega i čvorova unutar iste [9].

3.1 Način obrezivanja modela

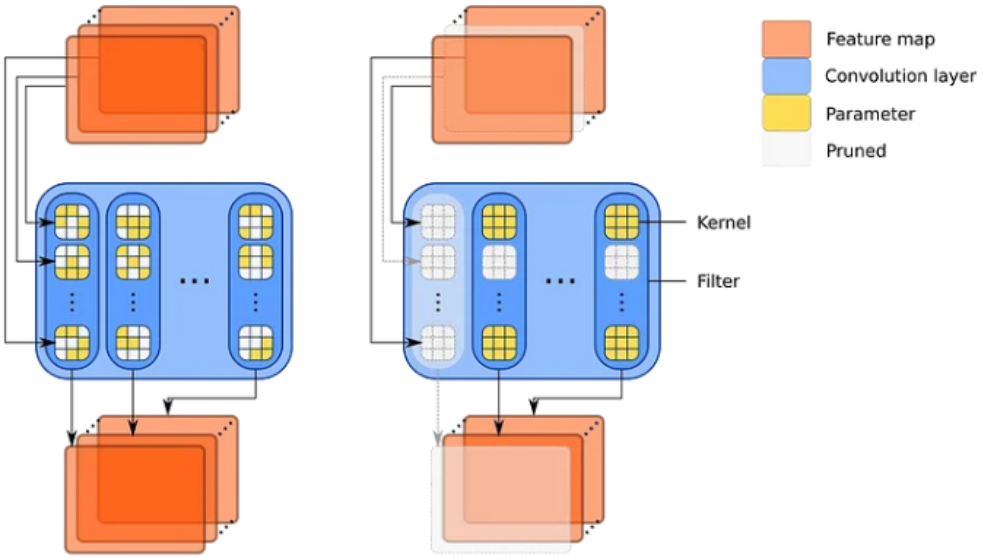
Postoje dva načina obrezivanja modela – nestrukturirano obrezivanje i strukturirano obrezivanje [10].

Nestrukturirano obrezivanje najčešće je korištena vrsta obrezivanja. Smanjenje broja parametara temelji se na odbacivanju utega neuronske mreže postavljanjem istih na vrijednost 0 prema različitim kriterijima. Najveća mana nestrukturiranog obrezivanja je što većina radnih okvira (engl. *framework*) i sklopovlja ne podržava ubrzanje raspršenih (engl. *sparse*) matrica – dakle, i uz postavljanje utega neuronske mreže na 0, odnosno otklanjanje istih, brzina izračuna predikcije neće se promijeniti. [10]

S druge strane, strukturirano obrezivanje se usredotočuje na otklanjanje neurona ili čak cijelih konvolucijskih filtera neuronske mreže. Prilikom otklanjanja konvolucijskih filtera potrebno pažnju obratiti na dimenzionalnost pojedinih slojeva kako ne bi došlo do neželjenih situacija. [10]

Slika 3.1 prikazuje razliku između strukturiranog i nestrukturiranog obrezivanja

neuronskih mreža. Nestrukturirano obrezivanje otklanja samo veze između pojedinih neurona (lijeva strana priakza), a strukturirano obrezivanje otklanja cijele konvolucijske filtere i mape značajki.



Slika 3.1 Prikaz strukturiranog i nestrukturiranog obrezivanja neuronske mreže.

3.2 Kriterij obrezivanja modela

Postoje dva glavna kriterija za odlučivanje hoće li se odabrana struktura obrezati iz neuronske mreže [10]:

- magnituda utega – prilično jednostavan kriterij kod kojeg se u slučaju nestrukturiranog obrezivanja odbacuju oni utezi s magnitudom manjom od postavljenog praga, a u slučaju strukturiranog obrezivanja se filteri odbacuju najučestalije na osnovi L1 ili L2 norme,
- magnituda gradijenta – kriterij kod kojeg se o odbacivanju odlučuje temeljem metrika koje proizlaze iz gradijenta iz koraka propagacije unatrag
- kombinacija kriterija.

3.3 Redoslijed obrezivanja modela

Nakon odabira kriterija i načina obrezivanja modela strojnog učenja, preostaje definirati trenutak u kojem će se obrezivanje provoditi. Postoji nekoliko uobičajenih trenutaka u kojem se obavlja obrezivanje modela strojnog učenja [10]:

- klasični pristup – treniranje, obrezivanje, fino ugađanje (engl. *fine-tuning*) – obrezivanje i fino ugađanje se mogu iterativno ponavljati,
- varijacije klasičnog pristupa – odbacivanje sve većeg broja utega tijekom treninga; ponovno treniranje modela nakon obrezivanja,
- raspršeno treniranje – obrezivanje dijela neuronske mreže pomoću prvotno nasumične maske koja se izmjenjuje tokom procesa treniranja,
- metode temeljene na penalizaciji – primjerice LASSO regularizacija koja težine značajki koje najmanje pridonose krajnjem rezultatu približava nuli.

Poglavlje 4

Zaključak

Rad na projektnom zadatku je bio zanimljiv i pomogao mi je da naučim nekoliko novih stvari. Smatram da sam na primjeren način istražio mehanizme objašnjivosti i obrezivanja modela strojnog učenja te pružio kvalitetan pregled proučenih mehanizama kao i njihovih demonstracijskih primjera.

U sklopu istraživanja i isprobavanja mehanizama teme projektnog zadatka, demonstracijski primjeri su stvoreni koristeći skup podataka koji opisuje dnevni broj iznajmljenih bicikala u ovisnosti o nekoliko različitih značajki (poput temperature, godišnjeg doba, vlažnosti zraka, brzine vjetra itd.) [5] i model slučajne šume razvijen temeljem spomenutog skupa podataka. Sve su metode objašnjivosti modela strojnog učenja za koje postoje dostupni Python paketi (PD, ICE, ALE, važnost značajki) isprobane, a stvorena je i vlastita implementacija za PD, ICE i Centered-ICE. Postojali su problemi sa stvaranjem demonstracijskih primjera za obrezivanje modela, ALE i važnost značajki (nepodudaranje s rezultatima u literaturi).

Daljnji rad na projektu obuhvaća rješavanje postojećih problema te implementiranje opisanih mehanizama na vlastitom modelu strojnog učenja.

Bibliografija

[1] C. O'Sullivan, "Interpretable vs Explainable Machine Learning — towardsdatascience.com," https://towardsdatascience.com/interperable-vs-explainable-machine-learning-1fa525e12f48, 2020, [Pristup 27. veljača 2023.].

[2] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. , s Interneta, https://christophm.github.io/interpretable-ml-book

[3] O. G. Yalçın, "5 Significant Reasons Why Explainable AI is an Existential Need for Humanity — towardsdatascience.com," https://towardsdatascience.com/5-significant-reasons-why-explainable-ai-is-an-existential-need-for-humanity-abe57ced4541, 2020, [Pristup 27. veljače 2023.].

[4] A. VK, "What Is Neural Network Pruning And Why Is It Important Today? — analyticsindiamag.com," https://analyticsindiamag.com/what-is-neural-network-pruning-and-why-is-it-important-today/, 2022, [Pristup 27. veljača 2023.].

[5] H. Fanaee-T and J. Gama, "Event labeling combining ensemble detectors and background knowledge," *Progress in Artificial Intelligence*, pp. 1–15, 2013. , s Interneta, http://dx.doi.org/10.1007/s13748-013-0040-3

[6] "Interpretability versus explainability - Model Explainability with AWS Artificial Intelligence and Machine Learning Solutions — docs.aws.amazon.com," https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html, 2023, [Pristup 27. veljača 2023.].

[7] S. Saha, "A Comprehensive Guide to Convolutional Neural Networks – the ELI5 way — towardsdatascience.com," https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53, 2018, [Pristup 27. veljača 2023.].

[8] J. H. Friedman, "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics*, vol. 29, no. 5, pp. 1189 – 1232, 2001. , s Interneta, https://doi.org/10.1214/aos/1013203451

[9] R. Bandaru, "Pruning Neural Networks — towardsdatascience.com," https://towardsdatascience.com/pruning-neural-networks-1bb3ab5791f9, 2020, [Pristup 27. veljača 2023.].

[10] H. Tessier, "Neural Network Pruning 101 — towardsdatascience.com," https://towardsdatascience.com/neural-network-pruning-101-af816aaea61, 2021, [Pristup 27. veljača 2023.].