[1]

Multimodalna predikcija za tabelarne podatke sa tekstualnim poljima zasnovana na transformerima

Aleksandar Mićić, dr Nemanja Ilić

***Sadržaj* — Tema ovog rada je korišćenje metoda i modela mašinskog i dubokog učenja u rešavanju zadatka predikcije na strukturiranim tabelarnim podacima koji uključuju tekstualna polja. Svrha rada je poboljšanje rezultata metoda koji su se pokazali najbolji u radu sa tabelarnim podacima (ansambli stabala odlučivanja / regresije), uključivanjem metoda koji su se pokazali najbolji u radu sa sekvencama i tekstom (transformer modeli dubokog učenja zasnovani na mehanizmu pažnje). Takođe, nekoliko klasičnih metoda mašinskog učenja i obrade teksta će se koristiti za referenciranje i poređenje.**

***Ključne reči* — DistilBERT, Duboko učenje, Linearna regresija, Mehanizam pažnje, Obrada prirodnog jezika transformerskom neuralnom mrežom, PCA, Random forest regresija, Transformerske neuralne mreže, Učenje u ansamblu, XGBoost**

I. UVOD

CILJ ovog rada je poboljšanje rezultata metoda koji su se pokazali najbolji u radu sa tabelarnim podacima (ansambli stabala odlučivanja / regresije), uključivanjem metoda koji su se pokazali najbolji u radu sa sekvencama i tekstom (transformer modeli dubokog učenja zasnovani na mehanizmu pažnje). U radu će se konktekstualni embeding vektori teksta generisani modelom dubokog učenja koristiti kao dodatna obeležja za model koji

[1] Aleksandar Mićić, Računarski fakultet, Srbija (email: amicic521m@raf.rs).
Dr Nemanja Ilić, Računarski fakultet, Srbija (email: nilic@raf.rs).

podrazumeva ansambl stabla regresije.
Metodološki, u izradi rada su na analitičko-sintetiči način analizirani metodi, problemi i potencijalna rešenja za dizajn algoritama koji se mogu koristiti u posmatranom tekstu, a sa empirijske tačke pokazan je kvalitet predloženog modela. Transformerski model koji će se koristiti je DistilBert model koji se zasniva na bidirekcionim transformerima.
Mnogi tabelarni podaci sadrže ne samo numerička i kategorička već i polja u slobodnoj formi. Skup podataka prikupljen je sa sajta za oglašavanje polovnih automobila. Prikupljeni tabelarni podaci su različitih oblika: numeričke vrednosti, kategoričke vrednosti, kao i tekst u slobodnoj formi. Ovakve vrste podataka u ovom radu nazivamo multimodalnim podacima. Da bi se veštačka inteligencija koristila za razumevanje podataka, potrebno je uspešno interpretirati multimodalne podatke. Numeričke tabelarne vrednosti mogu biti korišćene u različitim modelima, a tekstualne vrednosti se mogu procesirati tehnikama obrade prirodnog jezika (NLP tehnikama) [1].
U ovom radu je pokazano kako se mogu dizajnirati algoritmi za obradu multimodalnih podataka. Da bi se podaci bolje razumeli, svako numeričko obeležje je prvo iscrtano na grafiku. Grafički prikaz je pomogao u otkrivanju anomalija u podacima i u odabiru metoda regresije. Neka obeležja su u visokoj korelaciji jedna sa drugima, i takve bi bilo redudantno uzimati u analizu, stoga su takva obeležja eliminisana. Za otkrivanje obeležja koja su u visokoj korelaciji jedna sa drugima, korišćena je matrica korelacije.
Ovako sređeni podaci sa obeležjima numeričkih i kategoričkih vrednosti su prosleđeni različitim modelima mašinskog učenja, da bi se dobili bazni rezultati koje ćemo u ovom radu pokušati da poboljšamo. Način na koji ćemo probati da poboljšamo rezultate je generisanje novih obeležja u vidu embeding vektora, na osnovu teksta opisa oglasa. Opis oglasa je tekst u slobodnoj formi, pa se ne može proslediti modelu mašinskog učenja u sirovoj formi, već se prethodno mora obraditi. Nova obeležja ćemo izgenerisati na dva načina: primenom klasičnih NLP tehnika i primenom transformerskog modela, a zatim uporediti rezultate, kako između ove dve tehnike međusobno, tako i sa baznim modelom.
NLP je oblast veštačke inteligencije koji se bavi razumevanjem prirodnog jezika. NLP tehnike uključuju pretprocesirajne teksta, razbijanje teksta na reči i podreči koje nose značenje, i reprezentacija teksta u obliku vektora koji se mogu kasnije koristiti za mašinsko učenje.
Drugi metod ekstrakcije obeležja je primena modela zasnovanog na bidirekcionim transformerima. Danas transformerske neuralne mreže postaju sve popularnije i pokazale su se kao izuzetno efikasne u obradi prirodnog

jezika. Najpoznatija kompanija koja se bavi primenom transformera u obradi prirodnog jezika je OpenAI i njihovi jezički modeli, kao što su GPT-2, GPT-3 i ChatGPT, uspešno rešavaju probleme obrade prirodnog jezika [2]. Problemi koje ovi modeli rešavaju se odnose na generisanje koda u nekom programskom jeziku iz teksta na prirodnom jeziku, sumiranje velikog teksta u jednom pasusu, ali može da se podesi i da igra šah, kao i za druge primene [3]. U svim ovim problemima, ekstrakcija obeležja igra ključnu ulogu u obradi tekstualnih podataka. Transformeri su se pokazali izuzetno efikasni u ekstrakciji obeležja zbog svog mehanizma pažnje.

Mehanizam pažnje omogućava transformerima da dodele težine određenim delovima ulazne sekvence, u zavisnosti od relevantnosti. Ovo omogućava modelu da se fokusira na bitnije reči i fraze, a da umanji nerelevantne informacije. Dodatno, ovde ćemo koristiti bidirekcione transformere koji su u stanju da čitaju celu sekvencu odjednom i nauče kontekst i levo i desno od svake reči. Na taj način se uče međusobne zavisnosti između udaljenih reči u sekvenci. Korišćenjem ovih mehanizama, transformeri mogu da izvuku obeležja svesna konteksta što dovodi do poboljšanja performansi.

Konkretan transformerski model korišćen u ovom radu je DistilBERT. DistilBert je destilovana verzija popularnog BERT (eng. Bidirectional Encoder Representations from Transformers) modela. Predstavljen je od strane Hugging Face istraživača 2019. godine [4].

BERT (Bidirectional Encoder Representations from Transformers) je unapred treniran jezički model zasnovan na transformerima, za obradu prirodnog jezika [5]. BERT je napravljen i objavljen 2018. od strane Jakoba Devlina i njegovih kolega iz Google-a. BERT je treniran unapred za dva zadatka: modelovanje jezika i predviđanje sledeće rečenice. Posle treninga, koji je računarski zahtevan, BERT može biti podešen na manjem skupu podataka da optimizuje performanse na specifičnim zadacima [6].

BERT se zasniva na bidirekcionim transformerima koji uče kontekstualne veze između reči u sekvenci. Tradicionalni modeli procesiraju sekvencu u jednom smeru, sleva nadesno ili zdesna nalevo, dok je BERT bidirekcion. U ovom modelu, određene reči se nasumično maskiraju i model uči da predvidi maskirane reči u zavisnosti od konteksta. Ovaj pristup, zajedno sa predikcijom sledeće rečenice omogućavaju da BERT postigne visoke performanse prilikom rešavanja NLP problema.

DistilBERT je unapređena verzija BERT-a. DistilBERT je manji, brži, manje memorijski zahtevan transformerski model treniran destilovanjem BERT baze. Destilovanje se zasniva na treniranju manjeg modela, tzv. studentskog modela (DistilBERT), koji pokušava da imitira predikcije većeg, kompleksnijeg, tzv.

učiteljskog modela (BERT). Ovaj proces omogućava studentskom modelu da nauči značajne šablone i reprezentacije koje je veći model naučio, pritom zadržavajući visoku efikasnost. Procesom destilacije i kompaktnijom arhitekturom, DistilBERT čini praktično rešenje koje nudi dobre performanse koje ne idu na uštrb preciznosti.

DistilBERT koristimo za generisanje 768 novih obeležja. Treniranje modela mašinskog učenja nad ovolikim brojem obeležja može da bude izuzetno vremenski, a i memorijski zahtevno. Zato je odlučeno da se dobijena obeležja redukuju i za redukciju je izabrana PCA (eng. Principal Component Analysis) tehnika. Cilj PCA tehnike je da redukuje broj dimenzija skupa podataka uz očuvanje najvažnijih informacija. PCA identifikuje glavne komponente koje opisuju najveću varijansu podatka. Tako glavna komponenta opisuje najveći deo varijanse, zatim sledeća komponenta i tako redom.

Prikupljena i generisana obeležja će zatim biti korišćena za obučavanje različitih modela regresije da bi se dobile odgovarajuće predikcije cena, koje će biti upoređene da bi se empirijski ispitala upotreba tekstualnih obeležja u predikciji cene.

Regresija u mašinskom učenju je tehnika nadgledanog učenja kojom se obučavaju modeli da razumeju relaciju između nezavisnih promenljivih, ili obeležja, i zavisnih promenljivih, odnosno ishoda. Koristi se za predikciju kontinualnih vrednosti. Tako treniran model može da se koristi za predviđanje rezultata nad nepoznatim podacima, ili da popuni rezultate za podatke koji nedostaju [7].

Linearna regresija je statistička procedura za računanje vrednosti zavisne promenljive od nezavisnih promenljivih povlačenjem odgovarajuće linearne funkcije kroz skup podataka. Cilj linearne regresije da pronađe regresionu liniju koja najbolje opisuje podatke. Najučestaliji metod nalaženja regresione linije je metod najmanjih kvadrata. Ovaj metod računa koja je regresiona linija najbolja minimizacijom sume kvadrata vertikalnih devijacija između svake tačke i regresione linije. Pošto se devijacije kvadriraju pa sumiraju, ne poništavaju se pozitivne i negativne vrednosti.

Random forest regresija je algoritam koji koristi kombinaciju više nasumičnih stabala regresije, istreniranih na podskupu podataka. Korišćenje većeg broja stabala daje veću stabilnost algoritmu i smanjuje varijansu, zato što stabla nisu u korelaciji. Tako dobijamo bolje rezultate jer greške u nekim stablima se ne prenose na druga stabla. Random forest regresija je često korišćen algoritam zbog svoje sposobnosti da radi sa velikim i raznovrsnim podacima [8].

XGBoost (eng. Extreme Gradient Bosting) je algoritam mašinskog učenja koji kombinuje učenje u ansamblu sa gradijentnim pojačanjem, što dovodi do boljih

performansi. Umesto da se simultano pokrene grupa nezavisnih stabala nasumično, svakom narednom stablu se podešavaju težine na takav način da se kompenzuju slabosti (rezidualne greške) prethodnog stabla. Ovo se naziva gradijentno pojačanje [9].

Duboka neuralna mreža, DNN (eng. Deep neural network) je vrsta veštačkih neuralnih mreža koja se sa stoji od višestrukih slojeva međusobno povezanih čvorova, koji se nazivaju neuroni. Slojevi dubokih neuralnih mreža se tipično sastoje od ulaznog sloja, relativno velikog broja skrivenih slojeva i izlaznog sloja. Ulazni sloj prima inicijalne podatke, skriveni sloj vrši izračunavanja i transformacije nad ulaznim podacima, dok izlazni sloj generiše izlaz odnosno predikcije. Ovakva duboka arhitektura, u kojoj je svaki sloj uči kompleksnije i apstraktnije odlike, omogućava da DNN stekne bolju predstavu o podacima, što se naročito pokazalo efikasnim u NLP polju.

Na kraju, rezultati svih ovih tehnika su predstavljeni i upoređeni u sekciji Rezultati, i rad je kompletiran diskusijom u sekciji Zaključak.

II. TEORETSKI PREGLED

Sveobuhvatna tema ovog rada je mašinsko učenje, i njegov podskup, duboko učenje. Mašinsko učenje je oblast, deo veštačke inteligencije, koja se bavi konstrukcijom algoritama koji mogu da, upotrebom podataka, uče i izvršavaju određene zadatke [10].

Duboko učenje može da koristi i označene ali i neoznačene skupove podataka za obučavanje. Duboko učenje može da koristi nestrukturirane podatke, kao što su tekst ili slika, i da odredi karakteristike na osnovu kojih podatke može da razvrsta u kategorije. Duboko učenje se koristi za automatsko izvlačenje korisnih informacija iz podataka, uz minimalnu ljudsku intervenciju. Klasični algoritmi mašinskog učenja tipično više zavise od ljudske intervencije da bi učili i bolje strukturirane podatke.

Neuralne mreže se sastoje od čvorova koji su podeljeni u slojeve: ulazni čvor, jedan ili više skrivenih slojeva i izlazni sloj. Svaki čvor iz svakog sloja je povezan sa svim čvorovima iz susednog sloja. Rezultat čvorova iz jednog sloja predstavlja ulaz za čvorove sledećeg sloja. "Duboko" u dubokom učenju predstavlja dubinu u smislu broja slojeva neuralne mreže [10].

Regresija u mašinskom učenju je tehnika nadgledanog učenja kojom se obučavaju modeli da razumeju relaciju između nezavisnih promenljivih, ili obeležja, i zavisnih promenljivih, odnosno ishoda. Koristi se za predikciju kontinualnih vrednosti. Tako treniran model može da se koristi za predviđanje rezultata nad nepoznatim podacima, ili da popuni rezultate za podatke koji nedostaju [7].

Linearna regresija je statistička procedura za računanje vrednosti zavisne promenljive od nezavisnih promenljivih povlačenjem odgovarajuće linearne funkcije kroz skup podataka. Na primeru ovog rada, želimo da uočimo zavisnost cene automobila od pređene kilometraže [11].

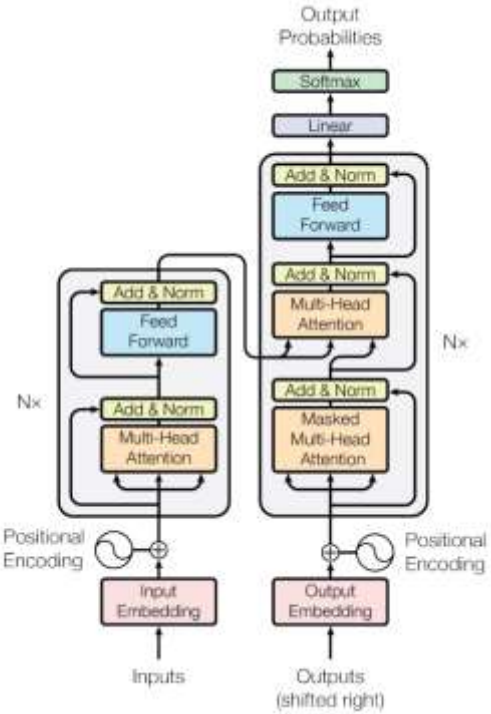
Random forest regresija je algoritam koji koristi kombinaciju više nasumičnih stabala regresije, istreniranih na podskupu podataka. Korišćenje većeg broja stabala daje veću stabilnost algoritmu i smanjuje varijansu, zato što stabla nisu u korelaciji. Tako dobijamo bolje rezultate jer greške u nekim stablima se ne prenose na druga stabla. Random forest regresija je često korišćen algoritam zbog svoje sposobnosti da radi sa velikim i raznovrsnim podacima [8].

XGBoost (eng. Extreme Gradient Bosting) je biblioteka koja implementira gradijentno pojačanje. Gradijentno pojačanje vodi poreklo iz rada Greedy Function Approximation: A Gradient Boosting Machine, od Friedman-a [12]. U osnovi ove tehnike se nalazi stablo regresije. Međutim, pojedinačna stabla imaju veliku varijansu; mala promena ulaznih podataka vodi do radikalno drugačije strukture stabla. Zato se stabla koriste kao ansambl, odnosno šuma, kao na Slici 4. Razlika između Random forest i ovog algoritma je u tome što XGBoost koristi gradijentno pojačanje. Umesto da se simultano pokrene grupa nezavisnih stabala nasumično, svakom narednom stablu se podešavaju težine na takav način da se kompenzuju slabosti (rezidualne greške) prethodnog stabla. Ovo se naziva gradijentno pojačanje [9].

Sekvenca-u-sekvencu modeli (poznatiji kao Seq2Seq eng. sequence-to-sequence), pretvaraju jednu sekvencu u novu sekvencu, gde se najčešće radi o sekvencama teksta. Imaju primenu u mašinskom prevođenju, sumiranu teksta i titlovanju slika. Pre toga, prevođenje se obavljalo reč po reč, ne uzimajući uobzir širi kontekst rečenice. Seq2seq uzima u obzir ne samo trenutnu reč već i susedne reči i time se značajno doprinelo kvalitetnijem prevođenju.

Mehanizam pažnje (eng. Attention Mechanism) je predložen u radovima “Neural Machine Translation by Jointly Learning to Align and Translate” [13] i “Effective Approaches to Attention-based Neural Machine Translation” [14]. Mehanizam pažnje omogućava modelu da se fokusira na relevantne delove rečenice.

Koncept transformera je predložen u radu Guglovih istraživača pod nazivom “Attention is all you need” [15], iz 2017. godine.



Slika 1. Arhitektura transformera sa svim elementima.

U njemu je predložejna transformerska arhitektura koja se bazira na mehanizmu pažnje. Sastoji se od kodera i dekodera, koji su dizajnirani kao višeslojni stekovi, prikazano na Slici 1.

Koder se sastoji od šest uniformnih slojeva. Svaki sloj se sastoji od dva podsloja i to: višestruki sloj pažnje i potpuno povezana neuralna mreža sa propagacijom unapred. Oko svakog od podslojeva je rezidualna konekcija, što znači da se ulaz sabira sa izlazom podmreže i normalizuje po formuli "(1)".

$$y = \mathrm{LayerNorm}(x + \mathrm{Sublayer}(x)) \tag{1}$$

Dekoder se takođe sastoji od šest identičnih slojeva. Pored dva podsloja koja ima i koder, dekoder ima i treći podsloj višestruke pažnje. Identično se koristi rezidualna konekcija kao kod kodera zajedno sa normalizacijom. Podsloj pažnje ima modifikaciju, maskiraju se naredne reči u sekvenci tako da dekoder

ne može da gleda unapred, već dekoder predviđa sledeću reč samo na osnovu prethodnih reči.

Ulazni podaci za transformer su upiti (sa oznakom Q) i ključevi (sa oznakom K) dimenzije dk i vrednosti (sa oznakom V) dimenzije dv. Konkretna pažnja se u radu naziva "Skalirani skalarni proizvod pažnje" (eng. "Scaled Dot-Product Attention") i računa se po formuli "(2)".

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Osnovni mehanizam pažnje je skalarni proizvod upita i ključeva. Pošto proizvod ima tendenciju da raste sa dimenzijom upita i ključeva, transformer skalira taj proizvod deljenjem sa korenom dimenzije dk.

BERT (Bidirectional Encoder Representations from Transformers) je unapred treniran jezički model zasnovan na transformerima, za obradu prirodnog jezika [5]. BERT je napravljen i objavljen 2018. od strane Jakoba Devlina i njegovih kolega iz Google-a. BERT je treniran unapred za dva zadatka: modelovanje jezika i predviđanje sledeće rečenice. Posle treninga, koji je računarski zahtevan, BERT može biti podešen na manjem skupu podataka da optimizuje performanse na specifičnim zadacima [6].

Ključna inovacija kod BERT-a je primena bidirekciono treniranih transformera na modelovanje jezika. Za razliku od direkcionih modela, koji čitaju tekst sekvencijalno, sleva nadesno ili zdesna nalevo, transformer čita celu sekvencu reči odjednom. Zato se smatra bidirekcionim, iako bi bilo preciznije nazvati ga nedirekcionim. Ova osobina omogućava modelu da nauči kontekst reči zasnovan na rečima levo i desno od reči. Rezultati BERT-a pokazuju da ovako treniran jezički model može da razume dublji smisao jezičkog konteksta i toka nego jednosmerni jezički modeli [16].

PCA (Principal component analysis) je tehnika za redukciju dimenzija koja pronalazi glavne ose koje predstavljaju skup podataka. Ove ose omogućavaju redukciju dimenzija uz zadržavanje maksimalne količine informacija. Ovo se postiže linearnom transformacijom podataka na novi koordinatni sistem u kome se (najveći deo) varijacije podataka može predstaviti manjim brojem dimenzija od polaznog skupa podataka. Mnoge studije koriste prve dve glavne komponente da grafički predstave podatke u dve dimenzije, što omogućava vizuelnu identifikaciju klastera blisko povezanih tačaka na grafu. PCA ima primenu u mnogim oblastima kao što su genetika, mikrobiologija i atmosferske

nauke [17,18].

III. PRIKUPLJANJE, OBRADA I ANALIZA PODATAKA

Podaci su prikupljeni sa sajta polovniautomobili.com. Ovaj sajt je izabran kao najposećeniji sajt na kome se oglašavaju polovni automobili u Srbiji [19] i kao sajt sa najvećim brojem strukturiranih parametara i filtera što olakšava prikupljanje podataka o automobilima.

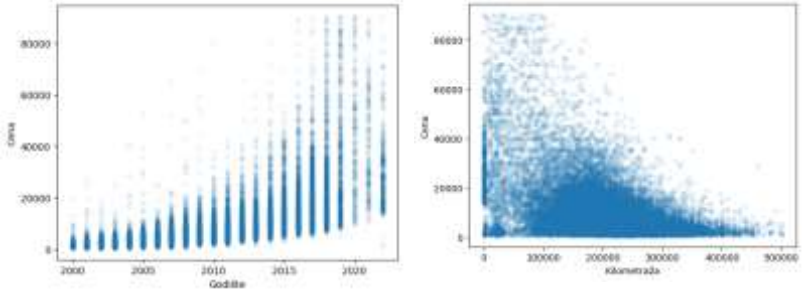
U cilju prikupljanja podataka, napravljen je web crawler koji je prošao kroz oglase i prikupio podatke o automobilima. Za pravljenje web crawlera korišćena je Pyhton biblioteka BeautifulSoup [20].

Podaci su očišćeni od anomalija i uklonjene su kolone koje su u velikoj korelaciji jedne sa drugima. Za pronalaženje suvišnih kolona, korišćena je matrica korelacije.

IV. IMPLEMENTACIJA MODELA PREDIKCIJE

Cena koju želimo da predvidimo je kontinualna vrednost. To znači da je ovo problem regresije i potrebno je odabrati model regresije za predviđanje cene. Odluka je probati sa nekoliko modela regresije i uporediti ih.

Kada se podaci predstave vizuelno (Slika 2), vide se rastući / opadajući trendovi za koje je prvi model koji se tipično proba model linearne regresije. Pored toga, linerna regresija ima kratko vreme obučavanja.



Slika 2. Primer raspodele cene i godišta / kilometraže.

Random forest regresija je metoda mašinskog učenja u ansamblu. Koristi stabla regresije koja zajedno dolaze do rezultata predikcije. Prednost ovog modela je da se stabla regresije generišu paralelno i da su relativno nekorelisana, što daje dobre rezultate jer model ne podleže greškama pojedinačnih stabala. Ovaj model je izabran zato što može da se nosi sa velikim brojem kolona i konzistentno pokazuje dobre performanse u radu sa tabelarnim podacima [8].

XGBoost je još jedna tehnika koja se zasniva na stablima regresije, s dodatkom da se koristi gradijentno pojačanje gde se, u svakoj narednoj iteraciji algoritma, dodeljuju težine zasnovane na metrikama koje imaju za cilj da minimizuju. U našem slučaju minimizuje se funkcija troška. Korišćena je maksimalna dubina stabla od 6 čvorova.

Konkretna neuralna mreža u ovom radu je sačinjena od pet dense slojeva sa po 25 neurona, i jednim dense slojem sa jednim neuronom na izlazu. Dense slojevi su slojevi u kojima je svaki neuron povezan sa svakim neuronom iz prethodnog sloja.

Korišćena funkcija aktivacije je ReLU (eng. rectified linear unit). Optimizacioni algoritam koji je korišćen je Adam (eng. adaptive moment estimation). Korišćena funkcija troška je srednja kvadratna greška (eng. mean Square Error, skraćeno MSE).

Za analizu sekvenci, biće korišćen DistilBERT i klasična metoda analize teksta.

Iako mnoge neuralne arhitekture mogu da modeluju tekst, predtrenirani transformeri su najdominantniji u polju prirodne obrade jezika. Ovaj model je unapred treniran na nenadgledani način nad velikim kolekcijama teksta, a u ovom radu je korišćen za generisanje dodatnih obeležja iz teksta koja će koristiti modeli regresije. Pošto je BERT prvi pokazao uspešnost unapred treniranih transformera [1], upravo ćemo njega primeniti za analizu sekvence. Odnosno preciznije, unapređenu verziju BERT-a, DistilBERT.

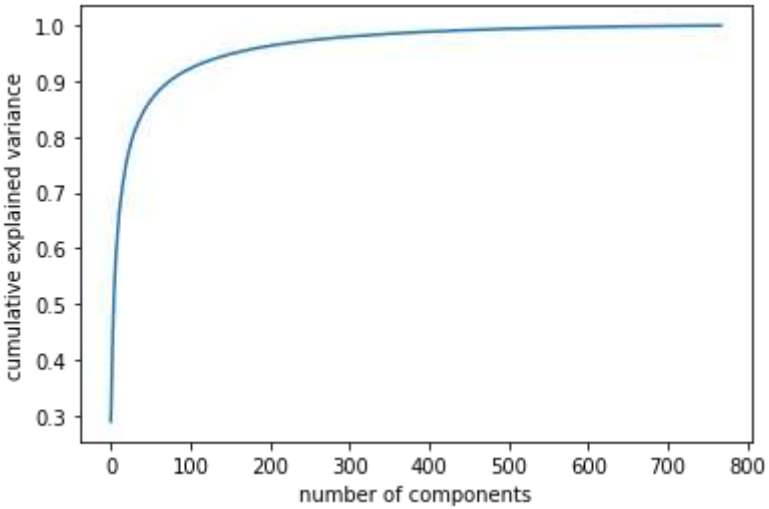
Podatke je potrebno prethodno obraditi da bi bili spremni za DistilBERT model. DistilBERT radi sa engleskim rečima. Opis oglasa je na srpskom, pa je morao biti preveden na engleski. Za prevod je korišćena Python biblioteka *googletrans*. Sledeći korak je tokenizacija. Rečenice se moraju razbiti na reči i podreči u formatu sa kojim BERT može da radi. Te reči se nazivaju tokeni. Tokeni pomažu u razumevanju značenja teksta. Takođe se uklanjaju zaustavne reči, a to su reči koje ne nose nikakvo značenje. Primeri zaustavnih reči su “a”, “the”, “is” i slično. Uklanjaju se iz teksta da bi se smanjio šum i broj obeležja [21]. Tokenizacijom dobijamo rečenice u formi liste tokena. Elementi dobijene liste će formirati kolone koje će model koristiti. Sve rečenice, pa samim tim ni dobijene liste, nisu iste dužine, a da bi BERT mogao da ih sve obradi moraju biti iste dužine. Zato se kraće liste dopunjavaju nulama do dužine najdužeg niza. Dobijeni podaci su niz vektora dužine 768 embedinga, i mogu se dalje koristiti za obučavanje.

Nakon dopunjavanja, potrebno je zamaskirati dopunjene elemente, da ih BERT ne bi uzeo u razmatranje. Maskiranje se vrši postavljanjem nula na odgovarajuća polja.

Na kraju je potrebno pomoću ovako dobijenih embedinga izgenerisati nova obeležja. Dobijeni rezultati se dodaju u polaznu tabelu kao nove kolone. Međutim, ovako dobijamo 768 novih kolona, što drastično povećava vreme treniranja modela regresije i ovoliko veći broj kolona preti da unese šum u rezultate. Da bismo rešili taj problem, upotrebićemo PCA tehniku za redukciju dimenzija. Takođe, biće zanimljivo uporediti rezultate sa i bez redukcije dimenzija.

PCA (eng. *Principal component analysis*) je tehnika za redukciju dimenzija koja pronalazi glavne dimenzije koje najbolje opisuju skup podataka.

Koliko dimenzija treba izabrati da bismo najbolje opisali podatke? Odgovor na ovo pitanje možemo dobiti iscrtavanjem odnosa kumulativne objašnjene varijanse (cumulative explained variance) i broja komponenti, odnosno dimenzija koje opisuju podatke (Slika 3).



Slika 3. Graf kumulativne objašnjene varijanse i broja komponenata.

Funkcija označava koliki procenat varijanse je objašnjen određenim brojem komponenata. Primera radi, sa 100 komponenata objašnjeno je preko 90% varijanse. Ne postoji formula po kojoj se može naći idealan broj dimenzija. Iz tog razloga će biti testiran model sa 1, 10, 50 i 100 dimenzija i uporediti rezultate.

Zanimljivo bi bilo uporediti BERT sa klasičnom metodom obrade prirodnog jezika. Pretprocesiranje obuhvata akcije čišćenja teksta od nepotrenih reči i

simbola. Sastoji se od:

A. Uklanjanja znakova interpunkcije.

B. Uklanjanja zaustavnih reči.

C. Konverzija svih reči u mala slova - da se ne bi neka reč tretirala kao nova samo zbog velikog početnog slova.

D. Određivanje korena reči - cilj je da se zbog padeža ili glagolskog oblika neka reč ne protumači kao više različitih reči. Postoji više načina da se odredi koren reči u engleskom jeziku, ovde je korišćen Porter Stemmer algoritam, razvijen od strane Martina Portera [22].

Za procesiranje teksta i formiranje TF-IDF vektora korišćena je biblioteka *sklearn*. Broj obeležja vektora je ograničen na 500 i dobijena matrica sa 500 kolona se koristi za trening.

V. REZULTATI

Algo.	Trening rezultat	Test rez.	Trening rezultat DistilBert	Test rezultat DistilBert	Tren. rez. NLP	Test rez. NLP
Lin. reg.	0.716	0.711	0.763	0.748	0.763	0.753
Rand. forest	0.909	0.905	0.951	0.917	0.964	0.913
XGB	0.916	0.917	0.937	0.918	0.946	0.916
DNN	0.877	0.878	0.903	0.884	0.929	0.864

Tabela 1. Rezultati predikcije cena bez redukcije dimenzija.

Uticaj opisa embedovan DistilBert-om je najveći kod linearne regresije (Tabela 1), gde se tačnost povećala za čak 3.7%. Kod random forest, XGBoost i duboke neuralne mreže, tačnost je samo neznatno povećana, za 1.2%, 0.1% i 0.6%, redom.

Kada se uporede rezultati embedovanja opisa korišćenjem DistilBert-a i klasične NLP tehnike, najveća razlika je kod duboke neuralne mreže, gde se DistilBert ispostavio za čak 2% bolji. Random forest i XGBoost model su dali

neznatno bolje rezultate primenom DistilBert analize opisa i to za 0.4% i 0.2%, redom, dok je linearna regresija dala bolje rezultate klasičnom primenom NLP tehnike i to za 0.5%.

PCA1		PCA10		PCA50		PCA100	
TR	TE	TR	TE	TR	TE	TR	TE
0.715	0.711	0.718	0.713	0.727	0.722	0.729	0.724
0.908	0.904	0.963	0.911	0.970	0.909	0.971	0.907
0.929	0.929	0.947	0.917	0.954	0.916	0.955	0.915
0.904	0.884	0.902	0.900	0.907	0.897	0.905	0.888

Tabela 2. Rezultati predikcije cena nakon redukcije dimenzija. Algoritmi odozgo nadole: Linearna regresija, *Random forest*, XGBoost, Duboka neuralna mreža.

Posmatrajmo rezultate nakon redukcije dimenzija, date u Tabeli 2. Zanimljivo je da je različit uticaj kolona dobijenih iz opisa zavisno od korišćenog modela regresije.

Linearna regresija daje bolje rezultate što se veći broj dimenzija koristi, a najbolje kada se koriste sve dimenzije.

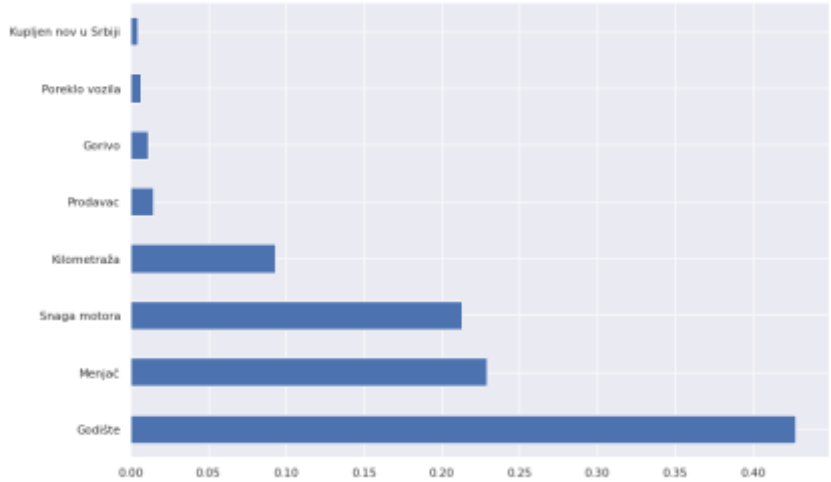
Random forest regresija daje najbolje rezultate za PCA10, i to 91.1%, i tačnost se postepeno smanjuje kada se broj dimenzija povećava ili smanjuje.

XGBoost regresija je dala najbolje rezultate za PCA1, i to 92.9% i što se veći broj dimenzija koristi, tačnost opada. Ovo su bolji rezultati nego kada se ne redukuju dimenzije i to za 1.1%.

Kod duboke neuralne mreže, rezultati su varirali, odnosno nije primetan ni porast ni smanjenje tačnosti sa povećanjem broja dimenzija. Kao najbolji se ispostavio PCA10 sa 90% tačnosti i redukcija dimenzija je povećala tačnost za 1.6%.

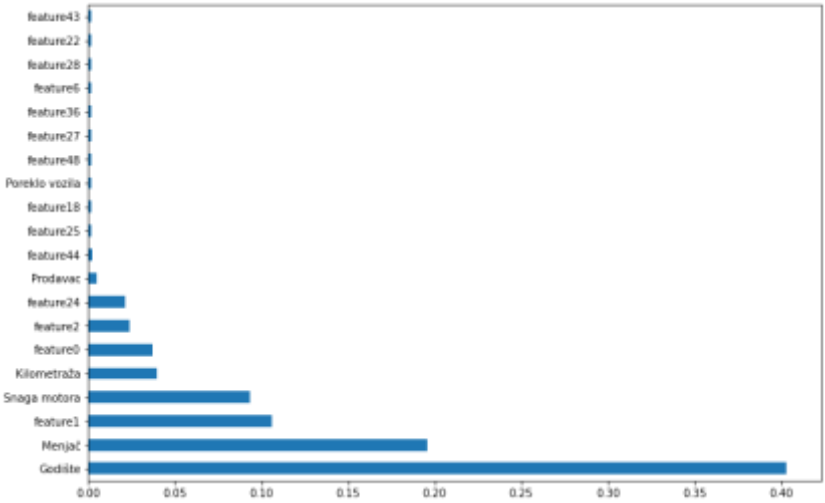
Od svih regresija, XGBoost regresija je dala najbolje rezultate u svakom od testiranih scenarija.

Na Slici 4 se nalazi grafik obeležja sa najvećim uticajem na cenu. Godište je imalo ubedljivo najveći uticaj na cenu. Nakon godišta da li je menjač automatski ili manuelni. Vrlo blizu iza je bila snaga motora. Zatim sledi kilometraža, dok su ostali parametri imali minimalan uticaj na cenu.

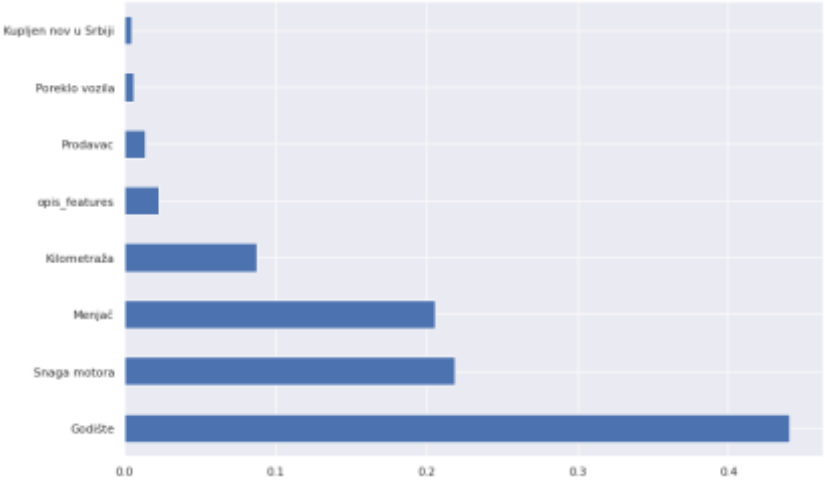


Slika 4. Obeležja koja su najviše uticala na cenu, bez analize opisa.

Zanimljivo je pogledati i uticaj opisa na cenu. Kao primer su iscrtani grafici dobijeni pomoću PCA sa 50 (Slika 5) i sa jednom komponentom (Slika 6). Kada posmatramo 50 dimenzija, ističe se jedno obeležje feature1 koje je treća po uticaju, odmah posle godišta i menjača. Interesantno je primetiti da je uticaj kolona kupljen nov u Srbiji, poreklo vozila i prodavac manji od uticaja nekih od komponenata dobijenih embedovanjem opisa. Ta činjenica ukazuje da su glavni parametri koji imaju značajan uticaj na cenu kolone godište, menjač, snaga motora i kilometraža. Po važnosti, za njima sledi opis.



Slika 5. Obeležja koje su najviše uticale na cenu, sa analizom opisa i redukcijom dimenzija PCA 50.



Slika 6. Obeležja koje su najviše uticale na cenu, sa analizom opisa i redukcijom dimenzija PCA 1

VI. ZAKLJUČAK

U ovom radu su prikupljeni podaci o 50.042 vozila, nakon čišćenja. Analizirano je koja obeležja imaju najveći uticaj na cenu i primenjeni su različiti modeli mašinskog i dubokog učenja u cilju predikcije cene automobila. Zatim su izgenerisani embeding vektori na osnovu teksta opisa koji su upotrebljeni za poboljšanje rezultata. Na kraju je primenjena PCA tehnika redukcije dimenzionalnosti.

Najznačajnija za cenu je bila starost vozila. Nakon nje da li poseduje automatski menjač ili ne i snaga motora. Kilometraža je imala manji uticaj dok su ostali parametri imali neznatan uticaj na cenu.

"Koliko je prešao" je pitanje koje ljudi najčešće pitaju kada razmatraju automobil. Međutim analiza je utvrdila da kilometraža, nakon nekoliko godina eksploatacije, ne govori mnogo. Velik broj automobila svih godišta ima slične kilometraže, šta više od 2006. godišta medijalne kilometraže prestaju da se povećavaju i pojavljuje se dosta vozila sa niskom kilometražom. Ova pojava ukazuje na izvesno nameštanje kilometraže, pa stoga nije mnogo relevantan faktor u proceni automobila, za razliku od uvreženog mišljenja.

Analiza tekstualnog opisa oglasa se pokazala kao uspešna i uticaj opisa na poboljšanje performansi je između 0.1%, korišćenjem XGBoost, i čak 3.6%, korišćenjem linearne regresije. Treniranje redukovanim dimenzijama se ispostavilo kao manje efikasno u slučaju linearne regresije. Korišćenje većeg broja dimenzija je imalo pozitivan uticaj na predikciju cene. Sa druge strane, random forest i XGBoost su imali bolje rezultate za manji broj dimenzija. Duboka neuralna mreža je dala različite rezultate za različit broj parametara i nije primećen porast a ni pad tačnosti u zavisnosti od broja dimenzija.

Najbolje rezultate predikcije je dala XGBoost regresija.

Kada se uporede rezultati analize opisa korišćenjem DistilBert-a i klasične NLP tehnike, najveća razlika je kod duboke neuralne mreže, gde se DistilBert ispostavio za čak 2% bolji. Random forest i XGBoost model su dali neznatno bolje rezultate primenom DistilBert analize opisa i to za 0.4% i 0.2%, redom, dok je linearna regresija dala bolje rezultate klasičnom primenom NLP tehnike i to za 0.5%.

Samostalno prikupljen i pripremljen skup podataka otvara brojne mogućnosti za analizu u kojoj bi se zadatak predikcije odnosio na atribute oglasa koji nisu cena. Na primer, bilo bi zanimljivo napraviti klasifikator koji bi predviđao da li je na datom automobilu manipulisano sa pređenom kilometražom ili ne.

ZAHVALNICA

Podaci sa sajta su vlasništvo sajta polovniautomobili.com, pa je za korišćenje njihovih podataka bilo je potrebno dobiti njihovu dozvolu. U tu svrhu, poslat je mejl sa molbom, na koji su pozitivno odgovorili, te se duboko zahvaljujemo sajtu polovniautomobili.com na dozvoli za korišćenje njihovih podatka.

LITERATURA

[1] Shi, X., Mueller, J., Erickson, N., Li, M., & Smola, A. J. (2021). Benchmarking multimodal automl for tabular data with text fields. arXiv preprint arXiv:2111.02705.
[2] Noever, D., Ciolino, M., & Kalin, J. (2020). The chess transformer: Mastering play using generative language models. arXiv preprint arXiv:2008.04057.
[3] OpenAI documentation. [Online]. Available: https://openai.com/api/
[4] DistilBert documentation. [Online]. Available: https://huggingface.co/docs/transformers/model_doc/distilbert
[5] Huggingface documentation. [Online]. Available: https://huggingface.co/transformers/v3.4.0/model_summary.html
[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[7] Dianne Castillo. (29 October, 2021). Machine Learning Regression Explained. [Online]. Available: https://www.seldon.io/machine-learning-regression-explained
[8] The Click Reader. (19 October, 2021). Random Forest Regression Explained with Implementation in Python. [Online]. Available https://medium.com/@theclickreader/random-forest-regression-explained-with-implementation-in-python-3dad88caf165
[9] Carl McBride Ellis. (17 October, 2022). An introduction to XGBoost regression. [Online]. Available: https://www.kaggle.com/code/carlmcbrideellis/an-introduction-to-xgboost-regression
[10] IBM. What is machine learning? [Online]. Available: https://www.ibm.com/topics/machine-learning
[11] Yale University. Department of Statistics and Data Science. (1997). Linear Regression. [Online]. Available: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm
[12] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
[13] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[14] Luong, M. T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.
[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. CoRR abs/1706.03762 (2017).
[16] Rani Horev. (10 November, 2018). BERT Explained: State of the art language model for NLP [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270
[17] Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.
[18] VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. " O'Reilly Media, Inc.". ISIN: 9781491912058
[19] Similarweb. [Online]. Available: https://www.similarweb.com/top-websites/serbia/

[20] BeautifulSoup documentation. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/bs4/doc/
[21] Shivani Rana (2021). [Online]. Available: https://www.kaggle.com/code/shivanirana63/beginner-s-guide-to-word-tokenization
[22] Dr. Abdul-Rahman Mawlood-Yunis. (21 February, 2002). [Online]. Available: http://people.scs.carleton.ca/~armyunis/projects/KAPI/porter.pdf

ABSTRACT

The topic of this paper is the use of machine and deep learning methods and models in solving the prediction task on structured tabular data that includes text fields. The purpose of this paper is to improve the results of methods that have proven to be the best in working with tabular data (ensembles of decision / regression trees), by including methods that have proven to be the best in working with sequences and text (transformer models of deep learning based on the attention mechanism). Also, several classical machine learning and text processing methods will be used for referencing and comparison.

Multimodal prediction for tabular data with text fields based on transformers

Aleksandar Mićić, PhD Nemanja Ilić