Predikcija cene smart telefona

Darie Čolak
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
colakdarie@gmail.com

Miloš Josipović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
milos.josipovic2000@gmail.com

Aleksandar Stevanović
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad
a.stevanovic98@gmail.com

***Apstrakt*—Tržište pametnih telefona predstavlja izuzetno raznoliku sliku kako po cenama tako i po karakteristikama samih uređaja. Pored osnovnih komponenti koje čine pametne telefone, kao što su procesor, memorija, kamera i slično, važan faktor koji značajno utiče na cenu jeste brend. Brendiranje nije samo puka marketinška strategija ono nosi sa sobom i troškove reklamiranja, ali takođe predstavlja i statusni simbol u društvu. Na primer, lideri u industriji pametnih telefona kao što su Samsung i Apple imaju privilegovanu poziciju da postavljaju veće cene za svoje proizvode zbog visokog stepena poželjnosti kod potrošača. Međutim, ovakva situacija stavlja potrošače u nezavidan položaj jer je teško oceniti da li zaista dobijaju najbolje vrednosti za uloženi novac. Uvođenjem objektivnog modela za procenu cena pametnih telefona, kupci bi imali jasniju sliku o adekvatnosti tražene cene. Ovaj model bi se oslanjao na širok spektar podataka o mobilnim telefonima, koje danas možemo naći na internetu. Za potrebe ovog istraživanja, korišćeni su već prikupljeni podaci koji sadrže informacije o karakteristikama svakog uređaja, uključujući i cene. Priprema podataka za obučavanje modela podrazumevala je eliminisanje atributa koji bi mogli nepovoljno uticati na proces učenja, kao i prilagođavanje pojedinih atributa u format koji je pogodan za algoritme mašinskog učenja. Nakon ove pripreme, skup podataka je podeljen na skup za obuku modela i skup za testiranje modela radi evaluacije performansi. U okviru ovog istraživanja, razmatrani su različiti modeli za predikciju cena pametnih telefona, uključujući Linearnu regresiju, Random Forest regresiju, SVM regresiju, XGB regresiju, Decision Tree i Logistic regresiju. Za vizualizaciju i analizu rezultata korištene su Python biblioteke koje su podržane za ovu svrhu, pružajući dublji uvid u performanse modela i relevantnih algoritama nad skupom podataka.**

Ključne reči—pametni telefoni; karakteristike; brend; regresija;

I. Uvod

U današnjem digitalnom dobu, mobilni telefoni su obavezna stavka za veliki broj ljudi, preciznije 91% ljudi na svetu poseduje mobilni telefon a preko 83% ljudi na celom svetu ima pametni telefon. Tržište pametnih telefona je izuzetno dinamično i konkurentno. Danas postoji oko 250 različitih brendova mobilnih telefona a u 2017. godini ih je bilo čak 750 [1]. Proizvođači svoje uređaje prilagođavaju različitim tržištima na osnovu ekonomskog i društvenog faktora, što predodređuje koliko će biti uspešna njegova prodaja u datom području. Sa sve većim brojem proizvođača, modela i karakteristika, potrošači se često suočavaju sa izazovom odabira optimalnog uređaja koji zadovoljava njihove potrebe i budžet. Cene pametnih telefona variraju u zavisnosti od brenda, performansi, dizajna i drugih faktora, što čini proces procene njihove vrednosti složenim i često subjektivnim.

Uvođenje objektivnog metoda za predikciju cena pametnih telefona može značajno olakšati potrošačima proces donošenja odluke o kupovini, kao i proizvođačima i prodavcima pružiti relevantne smernice pri određivanju adekvatnih cena. Inspirisan uspehom sličnih modela za predikciju cena pametnih telefona, u ovom radu će biti predstavljeno jedno rešenje za objektivno određivanje cena pametnih telefona.

Model će biti razvijen na osnovu obimnih podataka o karakteristikama pametnih telefona, poput procesora, RAM memorije, internog skladištenja, kamere, ekrana, baterije i drugih specifikacija koje utiču na njihovu cenu. Takođe, u obzir će biti uzeti trendovi tržišta, prethodne cene sličnih modela kao i faktori kao što su reputacija brenda i dostupnost na tržištu.

Ovim istraživanjem će se pokušati dati odgovor na pitanje koje komponente mobilnih telefona najviše utiču na njegovu cenu i poređenje različitih modela za predikciju cene pametnog mobilnog telefona.

Iako postoje izazovi u prikupljanju pouzdanih podataka i validaciji modela, očekuje se da će predloženo rešenje pružiti korisne smernice kako potrošačima, tako i tržištu pametnih telefona u celini. Detaljniji opis podataka, metodologije i rezultata biće izložen u nastavku rada.

Kroz ovaj rad, istraživači će imati priliku da se upoznaju sa kompleksnošću procesa predikcije cena pametnih telefona, kao i da steknu uvid u primenjive metode za objektivno određivanje cena na tržištu koje je često podložno promenanma i trendovima.

II. Srodna istraživanja

U radu [2] fokus je na predikciji cena pametnih telefona primenom tehnika mašinskog učenja. Korišćeni su podaci o ključnim karakteristikama pametnih telefona i njihovim cenama kako bi se razvio model koji bi mogao približno predvideti cenu novog pametnog telefona sa razumnom tačnošću. Analizirane su tri vrste klasifikatora: Random Forest Classifier, Support Vector Machine (SVM) i Logistic Regression. Na osnovu dobijenih rezultata, zaključeno je da su SVM i Logistic Regression imali najveću tačnost od 81%, a zatim je Logistic Regression korišćen za predikciju cena pametnih telefona. Ovi rezultati pružaju korisne uvide kako kupcima da donesu informisanu odluku prilikom kupovine telefona, kao i proizvođačima da odrede odgovarajuću cenu za ponuđene karakteristike.

Najvažniji aspekt istraživanja u radu [3] je primena hedonističke analize u predviđanju cena pametnih telefona, koristeći različite tehnike mašinskog učenja kao što su OLS regresija, random forest regresija i XGBoost regresija. Ovo istraživanje ističe ključne varijable koje utiču na cene pametnih telefona u Indoneziji, pri čemu su RAM, interna memorija, gustina piksela, NFC, veličina i težina identifikovani kao značajni prediktori cena. Osim toga, istraživanje je pokazalo da specifikacije kamere (prednje i zadnje) nemaju značajan uticaj na cene pametnih telefona, dok baterija i radio imaju negativan uticaj na cene. Kroz upotrebu hedonske linearne regresije, istraživanje je potvrdilo da sve eksplanatorne varijable značajno utiču na cene pametnih telefona, pri čemu se oko 75% varijabilnost icena može objasniti sa 12 eksplanatornih varijabli u modelu. Važno je napomenuti da je odsustvo višestruke kolinearne varijabilnosti potvrđeno, što dodatno validira rezultate analize.

U radu [4] Ritika Singh istražuje složen skup podataka o pametnim telefonima radi razumevanja ponašanja potrošača i segmentacije tržišta. Kroz detaljnu analizu podataka, identifikovani su ključni faktori koji utiču na preferencije potrošača pri kupovini pametnih telefona. Metode klasterovanja su primenjene kako bi se identifikovali segmenti potrošača na osnovu njihovih karakteristika i preferencija. Rezultati su pružili dublji uvid u materiju, što može biti od koristi za prilagođavanje marketinških strategija i proizvodnje prema specifičnim potrebama segmentiranih grupa potrošača.

Rad “Mobile Phone Price Prediction Based on Supervised Learning Algorithms” [5] autora A. Maesya, Y. Yanfi i Lukas, istražuje metode regresije za predviđanje cena mobilnih telefona na osnovu njihovih karakteristika. Koristeći linearnu regresiju i random forest regresiju, istražuje se koje varijable značajno predviđaju cenu i razvijaju se modeli za predviđanje cene. Istraživanje se oslanja na podatke sa Kaggle platforme koji sadrže 145 cena i karakteristika mobilnih telefona. Nalazi pokazuju da algoritmi linearne regresije i random forest regresije pružaju relativno dobre predikcije cena mobilnih telefona sa MAPE rezultatima ispod 10% i R2 rezultatima iznad 95%. Random forest je pokazao nešto bolje rezultate od linearne regresije u predviđanju cena.

III. Opis skupa podataka

Početni skup podataka je preuzet sa interneta [6] ali je sužen kako bi bio relevantan za istraživački rad. Kako nije bilo moguće pribaviti cenu svakog mobilnog telefona, time je skup podataka sužen jer telefoni bez cena nisu od interesa za ovaj istraživački rad. Predstavljena cena u skupu podataka je izražena u američkim dolarima (USD) na dan izlaska telefona u prodaju i to su cene u zvaničnim kanalima prodaje za nove telefone. Ovako pribavljen skup podataka se sastojao od podataka za 1513 telefona.

Podaci koji su obrađivani su:

- Pun naziv modela
- Brend mobilnog telefona
- Operativni sistem
- Dijagonala ekrana izražena u inčima
- Rezolucija ekrana izražena u formatu V x Š u pikselima
- Veličina baterije u mAh
- Tip baterije
- Količina RAM memorije izražena u GB
- Datum najave telefona
- Težina izražena u gramima
- Količina skladištene memorije izražena u GB
- Podržanost snimaka u 720p rezoluciji
- Podržanost snimaka u 1080p rezoluciji
- Podržanost snimaka u 4K rezoluciji
- Podržanost snimaka u 8K rezoluciji
- Podržanost snimaka u 30fps (okvira po sekundi)
- Podržanost snimaka u 60fps (okvira po sekundi)
- Podržanost snimaka u 120fps (okvira po sekundi)
- Podržanost snimaka u 240fps (okvira po sekundi)
- Podržanost snimaka u 480fps (okvira po sekundi)
- Podržanost snimaka u 960fps (okvira po sekundi)
- Cena izražena u američkim dolarima (USD)

Preuzeti skup podataka je sadržao nevalidne podatke u smislu netačnih naziva modela, neispravnih odnosno nepotpunih vrednosti ali i dupli podaci i takvi podaci su uklonjeni kako ne bi pravili problem algoritmima.

Brend	Količina
Xiaomi	264
Oppo	213
Samsung	206
Vivo	168
Realme	157
Huawei	137
Honor	124
LG	75
OnePlus	47
Lenovo	43
Sony	37
Apple	22
Google	19

Tabela 1: Raspodela telefona prema brendu

U tabeli 1 je prikazana raspodela mobilnih telefona prema brendovima nakon uklanjanja nevalidnih podataka.

Kako bi modeli za učenje bili u mogućnosti obraditi priložene podatke, bilo je neophodno izvršiti prilagođavanje sirovih podataka korištenim modelima.

Prvenstveno je izvršena modifikacija kolone za rezoluciju ekrana. Pomnožene su vrednosti za visinu i širinu i tako je dobijen ukupan broj piksela koji ekran poseduje. Uvrštena je nova kolona sa brojem piksela a izbačena prethodna kolona za rezoluciju. Primenjena je sledeća formula:

$$p = p\_w * p\_h$$

gde je p ukupan broj piksela ekrana, p_w je broj piksela po širini, a p_h broj piksela po visini ekrana.

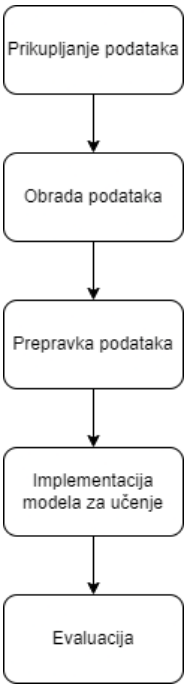
Sledeća izmena se odnosi na operativni sistem gde je izvršeno grupisanje u tačno određene verzije, na način da su podverzije uvrštene u matične verzije operativnih sistema.

Izvršeno je filtriranje uređaja radi poboljšanja preciznosti obučenih modela i to tako da su izbačeni oni telefoni koji su teži od 450 grama i skuplji od 2000 američkih dolara (USD).

Specifikacije kamera su sadržale veliki broj polja pa je odlučeno da se boolean vrednosti tih polja pretvore u numeričke i to tako da 0 označava FALSE a 1 TRUE. Tom modifikacijom bilo je moguće kreirati i primeniti novi atribut vezan za rang kamere sabiranjem vrednosti. Maksimalna moguća vrednost je 10 a najmanja 0. Ostale kolone su uklonjene.

Pored sve ovoga kolene koje su sadržale podatke tekstulane vrednosti enkodovani su u numeričke vrednosti. Ovo je korisno jer većina algoritama mašinskog učenja radi sa numeričkim podacima.

Takvi podaci su potom podeljeni na dva skupa, prvi koji je skup za obuku i iznosi 70% podataka (mobilnih telefona) i drugi koji je skup za testiranje obučenih modela od 30% svih podataka. Iz skupa za testiranje je uklonjena kolona za cenu.



Dijagram 1: Tok istraživanja [5]

Prikupljanje podataka, obrada kao i prepravka podataka je predočena u prethodnom delu, a implementacija modela za učenje i evaluacija će biti izloženi u narednim poglavljima.

IV. METODOLOGIJA

U metodologiji istraživanja primenjivani su različiti regresivni algoritmi kako bi se analizirao skup podataka i modelirale kompleksne veze između varijabli.

Regresivni algoritmi:

- *RandomForestRegression*
- *XGBRegression*
- *SVMRegression*
- *DecisionTree*
- *LinearRegression*
- *LogisticRegression*

Ovi algoritmi omogućavaju nam da precizno modeliramo i predviđamo vrednosti ciljnih promenljivih na osnovu ulaznih karakteristika, kao i da istražimo uzročne veze između nezavisnih i zavisnih varijabli u skupu podataka. Kombinacija ovih algoritama pružila je dublje razumevanje analiziranih fenomena i omogućila nam da donosimo informisane odluke zasnovane na rezultatima istraživanja.

- *RandomForestRegressor* je algoritam ansambla stabala odlučivanja koji radi tako što konstruiše više stabala odlučivanja tokom obuke i kombinuje njihove rezultate kako bi postigao bolje predikcije. Svako stablo se gradi na podskupu podataka i koristi se slučajni izbor osobina pri svakom čvoru. Konačna predikcija se dobija agregacijom predikcija svih

stabala.[7]

- *XGBRegression* (XGBoost Regression): XGBoost je efikasan algoritam za ansambl stabala odlučivanja koji koristi gradijentno pojačavanje (gradient boosting) kako bi se postigle visokokvalitetne predikcije. Radi iterativno dodajući nove modele, fokusirajući se na instance koje su pogrešno klasifikovane. XGBoost često postiže bolje rezultate u odnosu na klasične metode mašinskog učenja. [8]

- *SVMRegression* (Support Vector Machine Regression): SVM Regression je tehnika koja koristi metode mašinskog učenja za predikciju kontinuiranih vrednosti. Cilj je pronaći optimalnu liniju (ili hiperravan) koja najbolje odvaja podatke u prostoru funkcija koje opisuju podatke. Ova linija se koristi za predikciju vrednosti novih instanci. [9]

- *Decision Tree* (Stablo odlučivanja) je algoritam mašinskog učenja koji se sastoji od strukture stabla u kojem svaki unutrašnji čvor predstavlja test na atributu, svaka grana predstavlja rezultat testa, a svaki list stabla predstavlja klasu ciljnog atributa. Ova jednostavna struktura omogućava interpretaciju modela i dobru sposobnost generalizacije. [10]

- *Linear Regression* (Linearna regresija) je osnovni algoritam za regresiju koji modeluje linearnu vezu između nezavisnih promenljivih i zavisne promenljive. Cilj je pronaći linearnu funkciju koja najbolje odgovara podacima, minimizujući kvadratne greške između predikcija modela i stvarnih vrednosti. [11]
- *Logistic* regresija koristi linearni model koji kombinuje težine atributa kako bi se formirala linearna kombinacija ulaznih podataka.

A. *Priprema hiperparametara*

Koristeći *grid search* metod, identifikovani su hiperparametri koji pružaju najbolje rezultate. S obzirom na veliki broj modela predikcije za koje je bilo neophodno pripremiti hiperparametre, potrebno je voditi računa o dimenzionalnosti pretrage. Kako bi se efikasno upravljalo ovom dimenzionalnošću, posebna pažnja je posvećena odabiru ključnih parametara:

- *XGBoost: n_estimators, max_depth, learning_rate, subsample, colsample_bytree, reg_alpha, reg_lambda, gamma*
- *Decision Tree: max_depth, min_samples_split, min_samples_leaf*
- *Random Forest: n_estimators, max_depth, min_samples_split, min_samples_leaf* [12]

B. *Metod evaluacije*

U procesu modeliranja regresije, ključni korak je evaluacija performansi modela kako bismo procenili koliko dobro model odgovara podacima i kako se ponaša u predviđanju ciljne promenljive. Ova evaluacija je od suštinskog značaja za procenu korisnosti i pouzdanosti modela u stvarnom svetu. Evaluacija našeg regresivnog modela izvršena je na nekoliko načina:

- *Prilagođeno* R^2 - prilagođeni R^2 koji uzima u obzir broj prediktivnih promenljivih i ukupan broj opservacija, pružajući korigovanu procenu adekvatnosti modela. [13]
- *RMSE* (Root Mean Squared Error) - kvadratni koren prosečne kvadratne greške između stvarnih i predviđenih vrednosti. Ova metrika nam daje ideju o prosečnoj veličini greške u predikcijama.
- *MAE* (Mean Absolute Error) - prosečna apsolutna greška, koja meri prosečno odstupanje između stvarnih i predviđenih vrednosti.
- *MSE* (Mean Squared Error) - prosečna kvadratna greška, koja meri prosečno kvadratno odstupanje između stvarnih i predviđenih vrednosti.

V. REZULTATI

Radi lakšeg razumevanja i interpretacije dobijeni rezultat biće vizualizovani preko tabela i dijagrama.

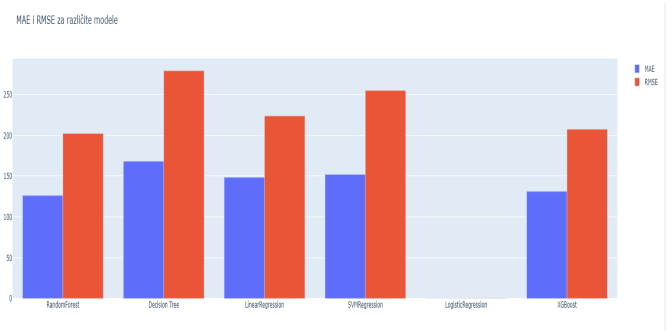
Na osnovu rezultata iz tabele 2, možemo izvesti nekoliko zaključaka. *Random Forest* i *XGBoost* modeli pokazuju bolje performanse u poređenju sa drugim modelima, s obzirom na više vrednosti prilagođenog R^2 i niže vrednosti srednjih apsolutnih grešaka (*MAE*), srednjih kvadratnih grešaka (*MSE*) i korena srednjih kvadratnih grešaka (*RMSE*). To ukazuje na to da su ovi modeli bolje u stanju da objasne varijacije u podacima i daju preciznije i zadovoljavajuće predikcije. Decision Tree model, iako poznat po svojoj jednostavnosti, nije se pokazao adekvatnim za naše podatke. Decision Tree model pokazuje najlošije performanse s negativnim prilagođenim R^2. Linearna regresija pokazuje umerene performanse s prilagođenim R^2 vrednostima oko 0.20, što znači da model objašnjava samo mali deo varijacije u podacima. Greške su takođe visoke u poređenju sa *Random Forest* i *XGBoost* modelima. *SVM* regresivni model takođe pokazuje loše performanse s negativnim prilagođenim R^2 vrednostima i visokim vrednostima grešaka. Logistička regresija se čini da nije adekvatan model za ove podatke, na osnovu rezultata.

Model	***adj. R^2***	***MAE***	***MSE***	***RMSE***
XGBoost	0.31052	131.43923	43097.2315	207.5987
RandomForest	0.34434	126.4472	40983.077	202.44277
Decision Tree	-0.24830	168.400	78028.170	279.3352
LinearRegression	0.19750	148.6890	50162.1602	223.9691
SVMRegression	-0.040734	152.2778	65053.7020	255.056272
LogisticRegression	-0.329872	0.291390	0.29139	0.53980

Tabela 2: Rezultati

A. Analiza grešaka

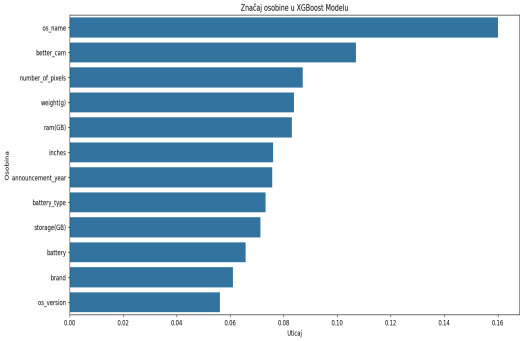
Na osnovu vrednosti srednje apsolutne greške (*MAE*) i korena srednje kvadratne greške (*RMSE*) prikazano na dijagramu 2, *Random Forest* model pokazuje najmanje vrednosti od 126.45 i 202.44 redom. Ovo sugeriše da je Random Forest model najpreciznij u predviđanju ciljne promenljive u poređenju sa ostalim modelima. Iako *XGBoost* model ima blago veću grešku u poređenju sa Random Forest-om, i dalje pokazuje solidne performanse sa MAE od 131.44 i *RMSE* od 207.60. Ovo ukazuje na njegovu stabilnost i pouzdanost u predikciji. Linearna regresija i *SVM* regresija modeli pokazuju slične vrednosti grešaka, sa *MAE* oko 148 i *RMSE* oko 224 i 255 redom. Ovi rezultati ukazuju na njihovu srednju tačnost u predikciji ciljne promenljive. Decision Tree model ima najveće vrednosti grešaka sa *MAE* od 168.40 i *RMSE* od 279.34. To sugeriše na preprilagođavanje ovog modela podacima. Logistička regresija, iako nije regresioni model već model klasifikacije, pokazuje izuzetno niske vrednosti grešaka (*MAE* od 0.29 i *RMSE* od 0.54), što ukazuje na njenu neadekvatnost za ovaj tip zadatka.



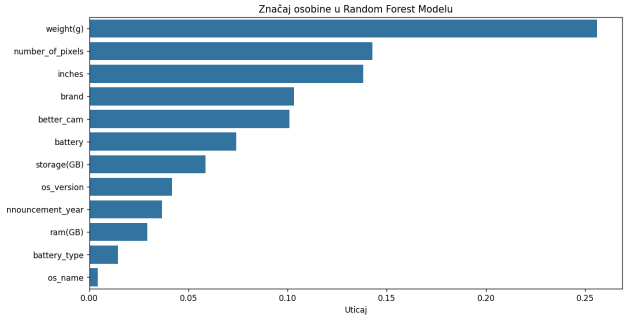
Dijagram 2: Analiza grešaka

B. Uticaj

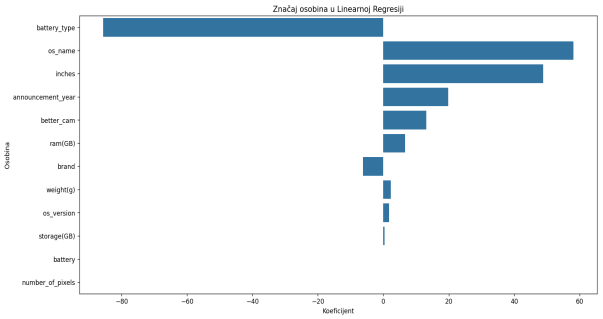
Na osnovu prikazanih dijagrama (3, 4, 5, 6), možemo zaključiti da operativni sistem, rezolucija ekrana (u pikselima) i težina telefona imaju najveći uticaj na predikciju cene mobilnih telefona, s obzirom na to da se ovi atributi ponavljaju kao ključni faktori u regresivnim algoritmima (XGBoost, Random Forest, Linearna regresija, Decision Tree). Ovo ukazuje na konzistentnost u značaju ovih atributa u svim modelima. Dijagrami SVM i Logističke regresije nisu prikazani zbog toga sto, SVM regresija ne daje direktno koeficijente kao linearna regresija, a za Logističku regresiju je potvrdjeno da je neadekvatna za ovaj tip zadatka.



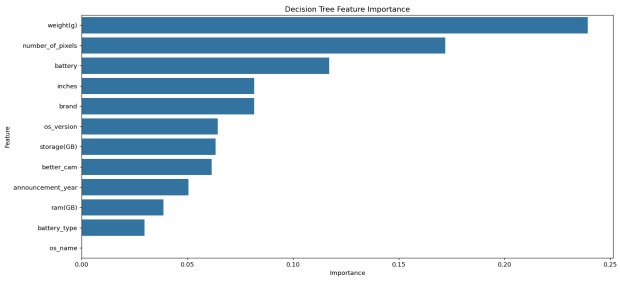
Dijagram 3: Dijagram uticaja kod XGBoost-a



Dijagram 4: Dijagram uticaja Radnom Forest-a



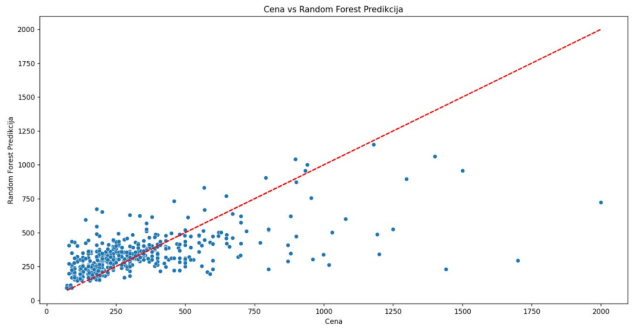
Dijagram 5: Dijagram uticaja Linearne regresije



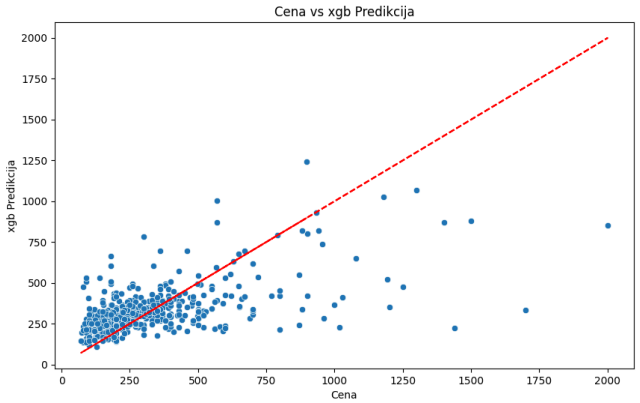
Dijagram 6: Dijagram uticaja Decision Tree algoritma

C. Poređenje rezultata i ulaznih podataka

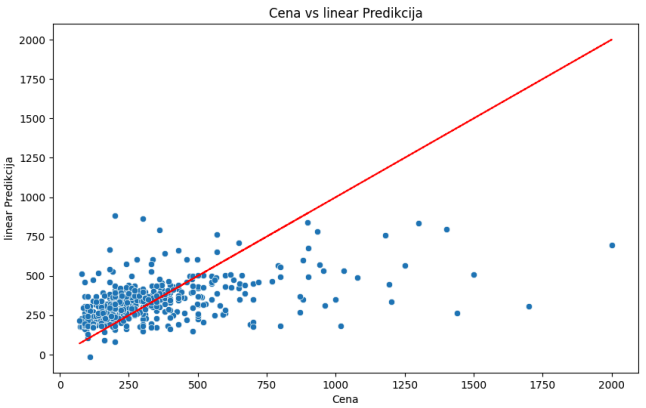
Primenom pobrojanih algoritama i uz pomoć biblioteka za predstavljanje rezultata, kreirani su dijagrami koji se tiču poređenja rezultata sa stvarnim cenama proizvoda. Ukoliko je algoritam precizan, na dijagramu će biti upisana tačka na dijagonali, odnosno biće x=y. U nastavku su prikazani dijagrami nekih algoritama koji su korišteni.



Dijagram 7: Dijagram poređenja rezultata i cena u Random Forest algoritmu



Dijagram 8: Dijagram poređenja rezultata i cena u XGBoost algoritmu

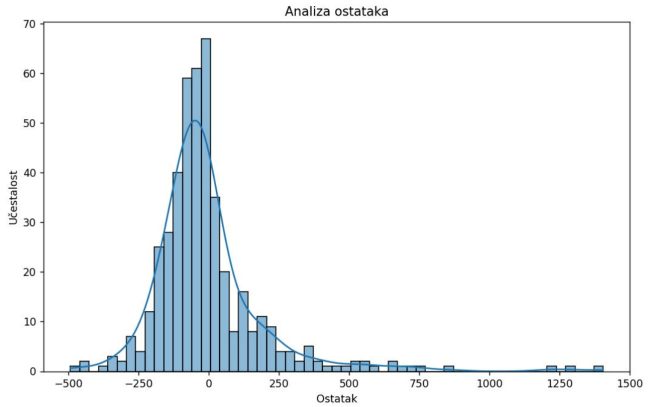


Dijagram 9: Dijagram poređenja rezultata i cena u Linear Regression algoritmu

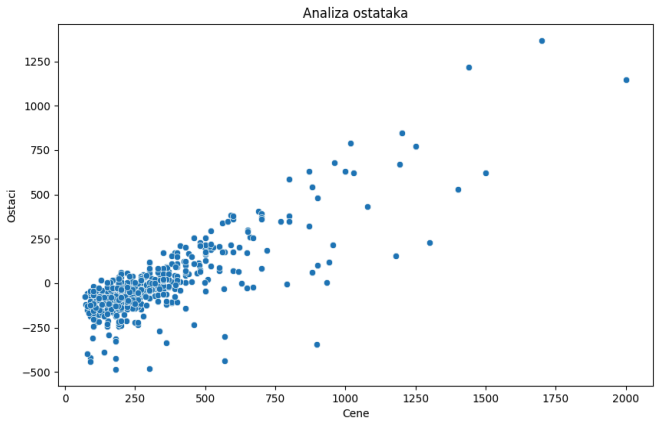
Sva tri prikazana dijagrama (dijagrami 7, 8 i 9) prikazuju približno iste rezultate predviđanja cena.

D. Analiza ostataka

Predstavljanje ostataka je još jedan vid analize rezultata dobijenih obučavanjem modela za predikciju cena pametnih telefona.



Dijagram 10: Dijagram poređenja rezultata i cena u vidu ostataka u Random Forest algoritmu



Dijagram 11: Dijagram poređenja rezultata i cena u vidu ostataka u XGBoost algoritmu

Ukoliko se oduzme predviđena cena od stvarne cene mobilnog telefona, dobija se ostatak koji je priložen na dijagramima 10 i 11. Ukoliko je ostatak 0 algoritam je stoprocentno precizan. Kako kriva na dijagramu 10 ima najviše vrednosti oko nule može se smatrati da je algoritam precizno predvideo cene mobilnih telefona. Na drugačiji način je izvršen prikaz u dijagramu 11 gde je korišten XGBoost algoritam. Što je veći broj tačaka oko nule, to je algoritam precizniji. Ostaci su izračunati po formuli:

$$r = x - y$$

gde je r ostatak, x je stvarna cena, a y je predviđena cena mobilnog telefona.

VI. Zaključak

U ovom istraživačkom radu predstavljen je model za predviđanje adekvatne cene pametnih telefona, koji uzima u obzir širok spektar tehničkih karakteristika. Podaci korišćeni u ovom istraživanju su preuzeti sa internet stranice koja se bavi tematikom analize podataka i obučavanjem modela za predikcije cena.

Analiza podataka je obuhvatila informacije o modelima, brendovima, operativnim sistemima, karakteristikama ekrana, bateriji, memoriji, težini, podršci za različite video formate i druge važne aspekte. ovi podaci su zatim obrađeni i prilagođeni za korišćenje različitim modelima mašinskog učenja kao što su Random Forest regresija i XGBoost regresija.

Rezultati istraživanja su pokazali da su modeli mašinskog učenja Random Forest i XGBoost efikasni u predviđanju cena pametnih telefona na osnovu tehničkih karakteristika. Međutim, nisu se mogli zanemariti izazovi koje su predstavili drugi modeli. Linearna regresija i SVM modeli su takođe imali svoje izazove, ali ih ne smatramo neuspehom. Naprotiv, ovi rezultati su pružili korisne uvide o granicama njihove primene i o tome gde treba usmeriti buduća istraživanja. Što se tiče Logističke regresije, iako nije bila optimalan izbor za obrađivane podatke, možemo izvući pouku da je važno prilagoditi model odgovarajućem tipu problema. Analiza uticaja varijabli je istakla ključne faktore koji utiču na varijabilnost cena, poput težine telefona, broja piksela i tipa operativnog sistema.

U pokušaju da se izbalansira model i poboljša njegova tačnost, primenjen je *oversampling* metoda na nebalansiranom skupu podataka. Cilj je bio da se obezbedi podjednak broj instanci za svaku klasu kako bi se omogućilo modelu da nauči ravnomerno reprezentaciju svih klasa i bolje se prilagodi manje zastupljenim klasama. Međutim, iako je primenjen oversampling, nisu se uspeli značajno poboljšati rezultati modela. Odnosno povećanju broja instanci manje zastupljenih klasa, model nije pokazao značajno bolje performanse. Konkretno povećavao se broj istanci *Apple* i *Google* telefona.

Osim toga, istraživanje je istaklo važnost kvalitetne pripreme podataka i odabira odgovarajućih modela za analizu, što može značajno uticati na konačne rezultate predviđanja cena pametnih telefona. Dalji rad na ovom polju bi mogao da uključi proširenje skupa podataka, uključujući primenu dodatnih tehnika mašinskog učenja za još preciznije predviđanje cena.

Literatura

[1] https://en.wikipedia.org/wiki/List_of_mobile_phone_brands_by_country

[2] Listianingrum, T., D. Jayanti, and F. M. Afendi. "Smartphone hedonic price study based on online retail price in Indonesia." *Journal of Physics: Conference Series*. Vol. 1863. No. 1. IOP Publishing, 2021.

[3] Singh, Ritika. "Exploratory data analysis and customer segmentation for smartphones." (2021).

[4] Subhiksha, S., Swathi Thota, and J. Sangeetha. "Prediction of phone prices using machine learning techniques." *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19*. Springer Singapore, 2020.

[5] Maesya, Aries, and Yanfi Yanfi. "Mobile Phone Price Prediction Based on Supervised Learning Algorithms."

[6] https://www.kaggle.com/datasets/berkayeserr/phone-prices

[7] Segal, Mark R. "Machine learning benchmarks and random forest regression." (2004).

[8] Pathy, Abhijeet, Saswat Meher, and P. Balasubramanian. "Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods." *Algal Research* 50 (2020): 102006.

[9] Yu, Hwanjo, and Sungchul Kim. "SVM Tutorial-Classification, Regression and Ranking." *Handbook of Natural computing* 1 (2012): 479-506.

[10] Wah, Yap Bee, et al. "Decision tree model for count data." *Proceedings of the World Congress on Engineering*. Vol. 1. 2012.

[11] Tranmer, Mark, and Mark Elliot. "Multiple linear regression." *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5.5 (2008): 1-5.

[12] Jain, Satendra Kumar, and Anil Kumar Gupta. "Application of Random Forest Regression with Hyper-parameters Tuning to Estimate Reference Evapotranspiration." *International Journal of Advanced Computer Science and Applications* 13.5 (2022).

[13] Leach, Lesley F., and Robin K. Henson. "The use and impact of adjusted R2 effects in published regression research." *Multiple Linear Regression Viewpoints* 33.1 (2007): 1-11.