Predviđanje cene polovnih telefona

Anja Delić, IN1-2019, delicc.anja@gmail.com

I. UVOD

Tržište korišćenih i obnovljenih uređaja je u protekloj deceniji doživelo ogromnu ekspanziju, zbog činjenice da predstavlja ekonomski isplativu alternativu kako za potrošače tako i za kompanije koje žele da uštede prilikom kupovine. Međutim, jedan od glavnih izazova koji se javljaju prilikom kupovine ovakvih uređaja jeste neizvesnost u pogledu njihove cene. Različiti faktori, poput starosti uređaja, stanja, kvaliteta obnove i drugih, mogu uticati na cenu, što čini kupovinu ovakvih uređaja rizičnom za potrošače. Sa druge strane, prodavci korišćenih i obnovljenih uređaja takođe se suočavaju sa izazovom formiranja adekvatnih cena, budući da je teško proceniti stvarnu vrednost uređaja u datom trenutku. To može dovesti do previsokih cena koje odbijaju potencijalne kupce, ili do preniskih cena koje ne pokrivaju troškove prodavca. Rešavanje problema postavljanja adekvatne cene za ovakve uređaje nije važno samo za pojedinačne potrošače i prodavce, već i za životnu sredinu. Prodaja korišćenih i obnovljenih uređaja pomaže u smanjenju otpada i reciklaži, čime se smanjuje negativan uticaj na životnu sredinu. Stoga, ovaj problem ima značajne posledice za društvo u celini. Zbog svega navedenog, sledeći izveštaj ima za cilj da prikaže analizu dostupnih podataka o cenama polovnih mobilnim telefona koja obuhvata prikaz obeležja koja direktno utiču na cenu kao i primenu različitih algoritama s ciljem što boljeg predviđanja iste.

II. ANALIZA PODATAKA

Baza podataka obuhvata 3454 uzorka i 15 obeležja. Svaki uzorak predstavlja skup različitih obeležja koje opisuju jedan model telefona.

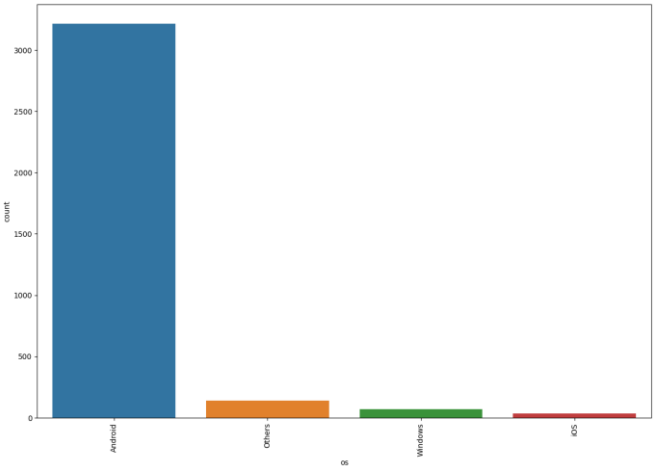
Obeležja su: marka telefona, operativni sistem, veličina ekrana, informacija da li postoji 4G i 5G, rezolucija prednje i zadnje kamere izraženu u megapikselima, količina interne i ram memorije koju uredjaj sadrži u gigabajtima, energetski kapacitet baterije u miliamper satima, težina telefona u gramima, godina kada je model izašao, broj dana korišćenja telefona, i normalizovane cene novog i polovnog modela. Baza podataka sastoji se od 5 kategoričkih i 10 numeričkih obeležja. U kategorička obeležja spadaju: marka telefona, operativni sistem, informacija o postojanju 4G i 5G, kao i godina kada je model izašao.

A. *Nedostajuće vrednosti*

Pre svega je prilikom analize baze podataka utvrđeno da ne postoje ponovljeni uzorci. Nedostajuće vrednosti pojavljuju se u okviru obeležja i to njih 179 u *rear_camera_mp*, 2 u *front_camera_mp*, 4 u *internal_memory*, 4 u *ram*, 6 u *battery* i 7 u *weight*. S obzirom da je skup podataka relativno mali, ne bi bilo ispravno dodatno ga smanjivati. Nakon dubljeg proučavanja baze podataka, nedostajuće vrednosti bivaju zamenjene medijanom. Medijana predstavlja vrednost skupa podataka koja se nalazi tačno u sredini, tako da je 50% podataka ispod te vrednosti, a 50% podataka iznad nje. Drugim rečima, medijana predstavlja srednju vrednost niza podataka kada su poređani po veličini. Korisna je kada skup podataka sadrži *outlier*-e, odnosno vrednosti koje značajno odstupaju od ostalih, što je upravo i razlog zbog kojeg je ova mera izabrana.

B. *Analiza pojedinačnih obeležja*

Dalje, analizom pojedinačnih obeležja zaključeno je da je najkorišćeniji operativni sistem *Android*, dok je *iOS* na poslednjem mestu, što se jasno vidi sa slike 1.



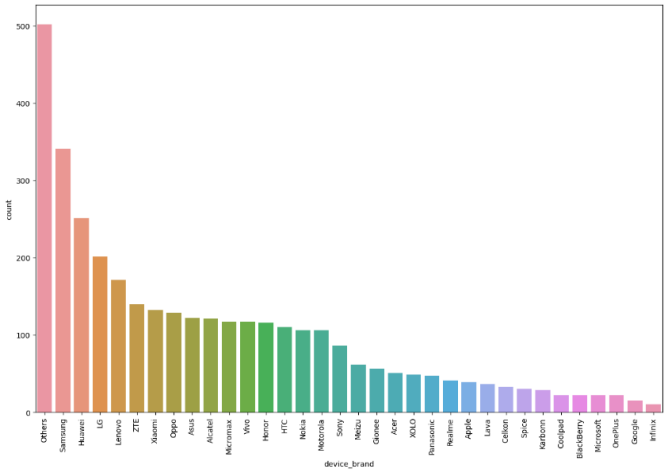
Sl. 1. Prikaz zastupljenosti operativnih sistema

Takođe, zaključeno je da su u skupu podataka najzastupljeniji mobilni telefoni označeni markom *Others*, koje prate *Samsung*, *Huawei* i ostali (Sl. 2.).

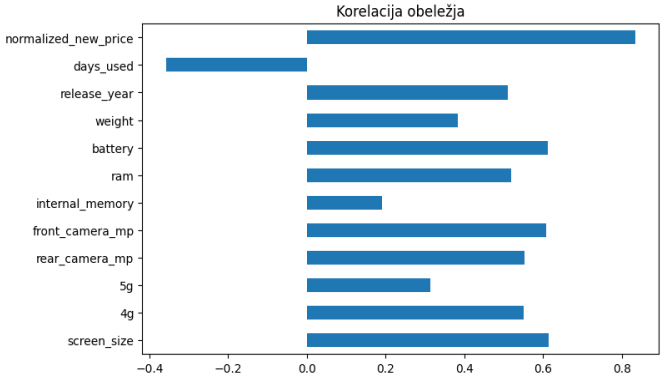
C. *Korelacija među obeležjima*

Daljom analizom dolazi se do uvida u korelacije, međusobne odnose ostalih obeležja u odnosu na cenu polovnih telefona. Jasno se vidi uticaj svakog od obeležja na cenu, pri čemu pozitivna korelacija označava da kada raste cena polovnih telefona tada raste i vrednost tog obeležja, dok negativna označava da ukoliko cena raste vrednost tog obeležja opada i obrnuto. Najveću pozitivnu korelaciju očekivano ima nova cena telefona, za kojom su i kapacitet baterije i veličina ekrana, dok negativnu

korelaciju ima jedino obeležje koje označava dane korišćenja uređaja.



Sl. 2. Prikaz zastupljenosti marki telefona



Sl. 3. Prikaz korelacije među obeležjima

III. MODELI I REZULTATI

A. *Priprema podataka*

Pre korišćenja različitih metoda za predviđanje cene polovnih telefona, nužno je da se podaci pripreme. Priprema pre svega podrazumeva pretvaranje svih kategoričkih obeležja i to na način da se za svaku vrednost pojednog kategoričkog obeležja pravi onoliko novih obeležja koliko ima jedinstvenih vrednosti u okviru kategoričkog, pri čemu se u nazivu navodi o kom tipu kategoričkog obeležja se radi. Svaka vrednost u okviru novog obeležja označava se sa 0 ili 1, u zavisnosti od toga da li je uzorak pripadao tom kategorickom obeležju.

Dalje, izvršena je podela na trening, test i validacioni skup podataka pri čemu trening podaci sadrže 80% a test i validacioni po 10% uzoraka iz celokupnog skupa podataka, a uzorci su nasumično izabrani.

Zatim je izvršena i standardizacija vrednosti numeričkih obeležja.

B. *Primena algoritama mašinskog učenja*

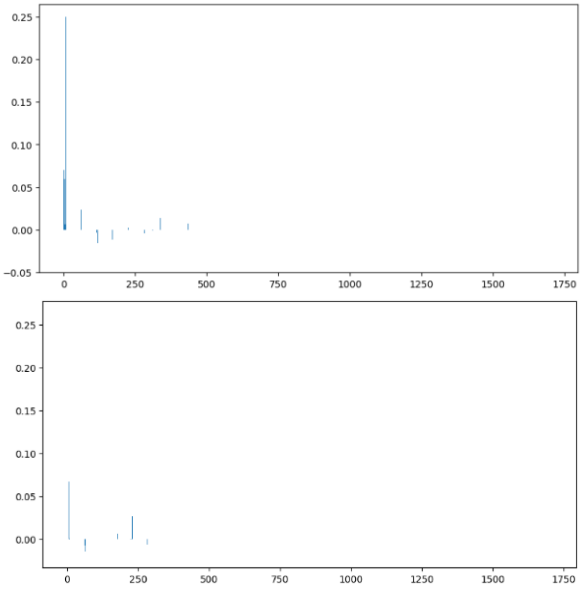
U daljoj analizi testirano je nekoliko algoritama mašinskog učenja kako bi se pronašao najadekvatniji za predviđanje cene polovnih mobilnih telefona pri čemu se boljim modelom smatra model koji ima veću vrednost parametra R-kvadrat koji predstavlja statističku meru koja se koristi za procenu kvaliteta prilagođavanja regresionog modela na podatke. Značajan je jer se može jednostavno interpretirati i omogućava procenu koliko dobro se model uklapa u podatke i koliko dobro objašnjava varijabilnost zavisne promenljive. Vrednost R-kvadrata se kreće u opsegu od 0 do 1, pri čemu vrednost 0 ukazuje na to da nezavisna promenljiva ne objašnjava varijabilnost zavisne promenljive, a vrednost 1 ukazuje na to da nezavisna promenljiva u potpunosti objašnjava varijabilnost zavisne promenljive.

Prvi korišćen model je linearna regresija sa hipotezom $y = b0+b1*x1+b2*x2+..+bn*xn$. Sledeći korišćen model je *Ridge* a zatim i *Lasso* regularizacija. Kod pomenutih modela prvo se analizira koja je optimalna vrednost hiperparametra *alpha* koji se koristi za kontrolu jačine regularizacije kako ne bi došlo do natprilagođenja modela. Povećavanjem ovog hiperparametra povećava se jačina regularizacije i smanjuje kompleksnost modela. Kod *Lasso* regularizacije ovaj hiperparametar vrši dodatno i selekciju obeležja.

Uporedan prikaz vrednosti parametra R-kvadrat za svaki od pomenutih modela prikazan je u tabeli ispod, gde se vidi da je od navedena tri po ovom parametru najadekvatniji *Lasso*. S druge strane, manje koeficijente ima *Rigde* regresija što je u prvu ruku neočekivano. Ovakva pojava proizilazi iz činjenice da se pri traženju optimalnog modela akcenat stavljao na najvećoj vrednosti R-kvadrat hiperparametra. Ukoliko bi se prednost dala tačnosti, vrednost koeficijenata bi značjno opala, što se vidi sa Sl. 4.

Tabela 1: Upoređivanje parametra R-kvadrat.

Vrsta modela	R-kvadrat
Linearna regresija	0,836
Ridge regularizacija	0,850
Lasso Regularizacija	0,851



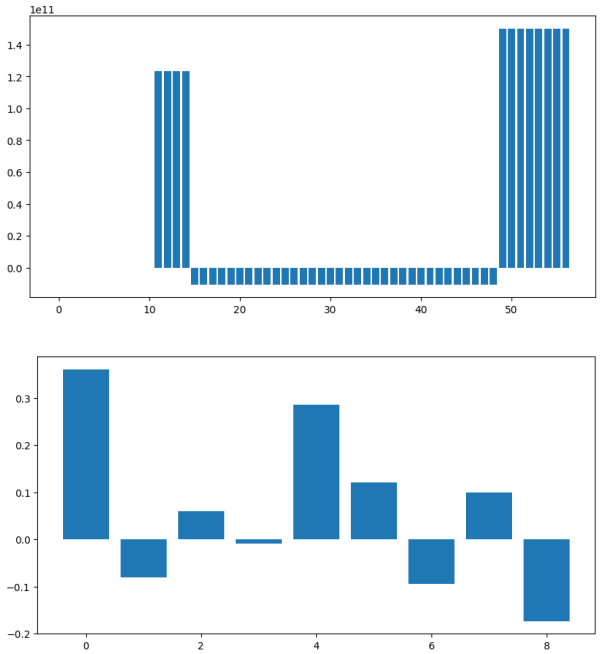
Sl. 4. Uporedan prikaz koeficijenata *Lasso* regularizacije s različitim parametrima *alpha*

C. *Primena PCA za redukciju dimenzionalnosti*

PCA (*Principal Component Analysis*) je tehnika koja se koristi za smanjenje dimenzionalnosti. Ova tehnika koristi se za izdvajanje bitnih karakteristika i transformaciju višedimenzionalnih podataka u manje dimenzije, što, u opštem slučaju olakšava njihovu analizu i vizuelizaciju.

Jedan od najbitnijih parametara predstavlja *n_components* koji označava broj glavnih karakteristika koje će biti izdvojene iz originalnih podataka. Nakon analize i posmatranja međusobnih zavisnosti obeležja s ciljnim obeležjem, uočava se da 9 obeležja imaju primetan uticaj na ciljno obeležje, što ukazuje na to da bi optimalna vrednost parametra *n_components* bila upravo 9, što se eksperimentalno i pokazalo kao tačno.

Nakon primene PCA tehnike, nisu uočene značajne promene kod pomenutih modela, čak su kod *Ridge* i *Lasso* regularizacije uočena pogoršanja u pogledu parametra R-kvadrat. Linearna regresije, s druge strane, ima značajan napredak u pogledu raspodele koeficijenata.



Sl. 5. Uporedan prikaz koeficijenata linearne regresije pre i posle primee PCA

Tabela 2: Upoređivanje parametra R-kvadrat posle PCA

Vrsta modela	R-kvadrat
Linearna regresija	0,828
Ridge regularizacija	0,839
Lasso Regularizacija	0,844

IV. LITERATURA

[1] Sergious Theodoridis and Konstantinos Koutroumbas, “Pattern recognition,” Elsevier academic press, 2003.