[1]

Polovni telefoni

Ime Prezime, index, e-mail

Uvod

Kupovina i prodaja korišćenih telefona i tableta nekada su se obavljale na samo nekoliko online tržišnih sajtova. Međutim, tržište polovnih i obnovljenih uređaja značajno je poraslo tokom poslednje decenije.Ovaj rast se može pripisati porastu potražnje za polovnim telefonima i tabletima koji nude značajne uštede u poređenju sa novim modelima. Obnovljeni i korišćeni uređaji i dalje pružaju ekonomične alternative kako potrošačima tako i poslovnim korisnicima koji žele da uštede novac prilikom kupovine. Postoje mnoge druge prednosti koje su povezane sa tržištem korišćenih uređaja. Polovni i obnovljeni uređaji mogu se prodavati sa garancijama, a mogu se osigurati i dokazom o kupovini. Treće strane prodavci/platforme, poput A1,Telenora, MTS-a, itd., nude atraktivne ponude za obnovljene uređaje. Povećanje trajnosti uređaja kroz trgovinu polovnih uređaja takođe smanjuje njihov ekološki uticaj i pomaže u recikliranju i smanjenju otpada.

I. Baza podataka

Baza podataka sadrzi 15 atributa(kolona) i 3454 podataka(redova)(baza podataka je dimenzija 3454x15).Ti atributi su device_name (ime brenda),os (operativni sitem), screen_size (velicina ekrana u cm), 4g, 5g, front_camera_mp (rezolucija prednje kamere u mega pikselima), rear_camera_mp (rezolucija zadnje kamere u mega pikselima), internal_memory(interna memorija u gigabajtima),ram(kolicina operativne memorije u gigabajtima),battery(energetski kapacijtet baterije u mAh), weight(tezina u gramima), release_year(godina proizvodnje),days_used(koliko je uredjaj dana koriscen), normalized_new_price(normalizovana cena novog uredjaja), normalized_used_price(cena polovnog uredjaja).

Cilj nam je da odredimo cenu polovnog uredjaja (zavisna promenjiva) koristeci neki od regresionih modela.

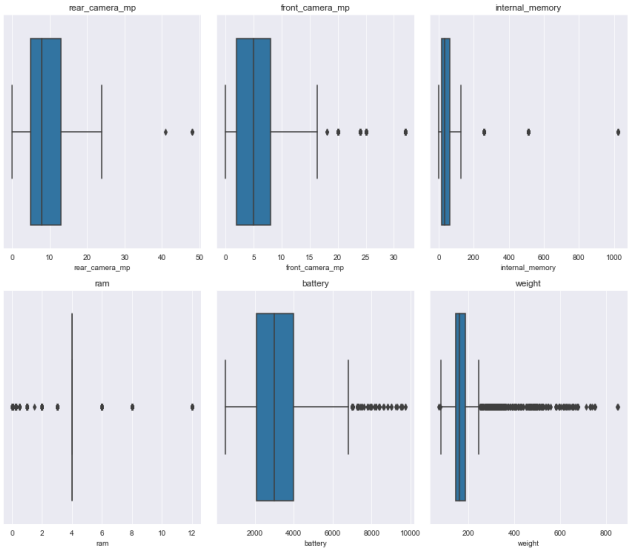
II. Analiza podataka

A. Tip podataka

Atributi kategorickog tipa koji su prisutni u nasoj bazi su 4g ,5g, ime brenda i operativni sistem.Ostali atributi su numerckog tipa. Atributi 4g i 5g imaju 2 moguce vrednosti(da ili ne), pa su njihove vrednosti zamenjene sa 0 i 1 respektivno.Za atribute ime brenda i operativni sitem smo primenili ’one hot encoding’ zato sto sto ne postoji ordinalni odnos izmedju njihovih vrednosti, pa ’integer encoding’ nije dovoljan. Ime brenda ima 34 unikatne vrednosti,stoga imacemo 34 nove kolone za svaku marku telefona koja je prisutna u nasoj bazi.Operativni sitem ima 4 unikatne vrednosti (android, ios, windows, other), stoga imacemo 4 nove kolone koje ce predstavljati odgovarajuci operativni sistem telefona.Posle prevodjenja svih kategorickih obelezja u numericka, dimenzija baze je 3454x48.

B. Nedostajuci podaci

Podaci koji nedostaju su prednja kamera(5.18% podataka), zadnja kamera(0.057% podataka), interna memorija(0.11% podataka), ram memorija(0.11% podataka), baterija(0.17% podataka).Izbacivanje ovih podataka ne bi bilo ispravno resenje posto bismo imali manje podataka, i potencijalno bismo izgubili znacajne informacije.Da bismo odlucili da li cemo podatke zameniti medianom ili srednjom vrednoscu moramo se prvo upoznati sa distribuciom podataka i utvrditi da li postoje outlieri.Na osnovu slike 1 mozemo utvrditi da outlieri postoje.Takodje,distribucija ovih podataka je asimetricna.Najbolja praksa je da kada su outlieri prisutni nedostajuce podatke zamenimo sa medianom, jer je mediana robustnija i manje osetljiva na outliere nego srednja vrednost.Nedostajace vrednosti svakog atributa smo zamenili,tako sto smo svaki atribut grupisali po njemu respektivnom brendu i odredili medianu te grupe.Atribut zadnja kamera je i dalje imala nedostajece vrednosti, jer za brend ‘Celkon’ ne postoje podaci o zadnjoj kameri.Preostale nedostajuce podatke za atribut zadnja kamera smo zamenili uopstenom medianom.



Slika 1. Boxplot sa outlierima za atribute zadnja kamera,prednja kamera,interna memorija, ram memorija, baterija i tezina.

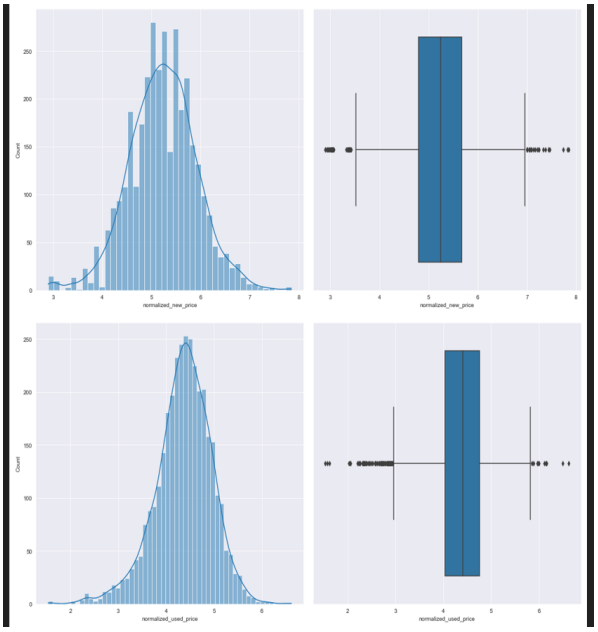
C. Outlieri i distribucija podataka

Atribute smo predstavili funkcijom boxplot, gde smo uvideli njihove interkvartalne opsege, medianu i outliere.Takodje, smo za svaki atribut odredili njegovu distribuciju.Na slici 1 su vec prikazani interkvartalni opsezi, mediana i outlieri za nedostajuce vrednosti.

Na slici 2 vidimo da normalizovana cena koriscenih uredjaja i normalizovana cena novih uredjaja imaju normalnu(simetricnu) distribuciju.

[1]

Ostali atributi imaju asimetricnu distribuciju.Na osnovu izracunatih interkvartalnih opsega outlieri su izbaceni kako bi povecali linernost u nasim podacima i izvrsili bolju predikciju.



Slika 2.Hisplot sa distribuciom i boxplot sa outlierima za atribute normalizovana cena novih uredjaja i normalizovana cena koriscenih uredjaja.

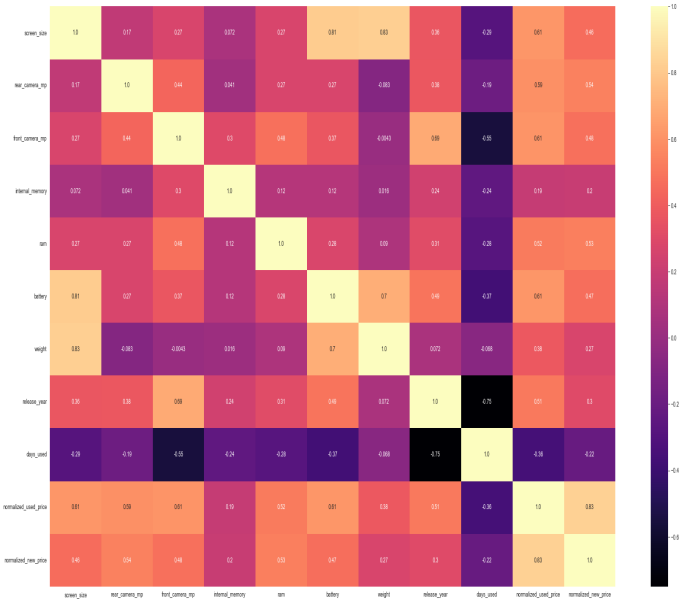
D. Korelacija atributa

Stepen korelacije se meri pomoću koeficijenta korelacije, koji može da varira od -1 do 1. Koeficijent korelacije od -1 označava jaku negativnu korelaciju, koeficijent od 0 označava da nema korelacije, a koeficijent od 1 označava jaku pozitivnu korelaciju.Na slici 3. je prikazana heat mapa pomocu koje mozemo utvrditi koeficijente korelacije izmedju nezavisnih atributa i zavisne promenjive(normalizovana cena korisnjenih uredjaja). Najveci koeficijent korelacije sa zavisnom promenjivom ima normalizovana cena novih uredjaja(0.83).Posle njega najvece pozitivne koeficijente korelacije imaju velicina ekrana(0.61), prednja kamera(0.61), kapacitet baterije (0.61), zadnja kamera(0.59), godina proizvodnje(0.51). Jedini atribut koji ima negativnu korelaciju je koliko je dana telefon koriscen(-0.36).Na slici 4. je prikazana korelacija za atribute normalizovana cena novih uredjaja i normalizovana cena koriscenih uredjaja.Vidimo vrlo dobru linearnu vezu izmedju ova dva atributa.Kada se cena novog telefona povecava, cena polovnog telefona se linerno povecava.Kod drugih atributa ne mozemo uociti dobru linernu povezanost sa normalizovanom cenom koriscenih uredjaja(primer slika 5.).

E.Generalni zaključci o podacima

Samsung je bio najčešće ponovo korišćen telefon, odmah posle brendova koji se kategorizuju kao ‘Ostali’.Najvise ljudi koristi Android operativni sistem, a najmanje njih Windows(slika 7.).

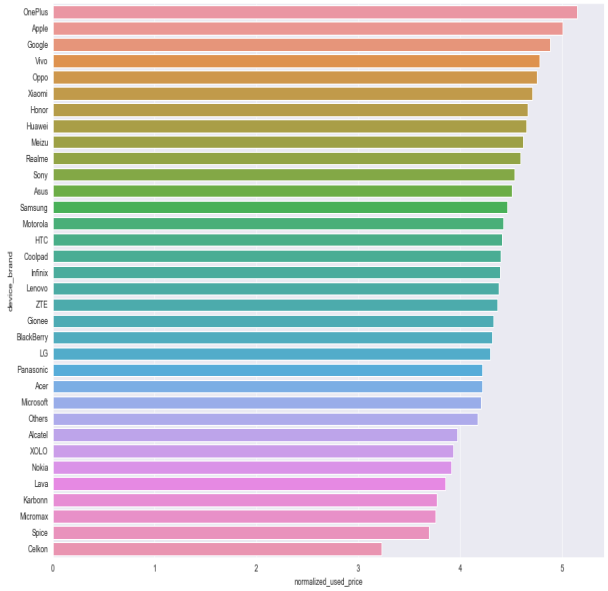
Veličina ekrana je uglavnom između 10 i 15 cm-a.Najvecu cenu imaju telefoni marke ‘One Plus’(pa slede telefoni marke ‘Apple’) dok najmanju cenu imaju telefoni marke ‘Ceikon’(slika 6.).Korisnici 4G i 5G mreže su češći.Telefoni bez 4g imaju nize cene(slika 8).Raspon megapiksela za zadnju kameru uglavnom je između 5-15 mega piksela.Raspon megapiksela za prednju kameru uglavnom je između 0-10 mega piksela.Interni memorija uglavnom iznosi između 0-100 GB.RAM memorija telefona uglavnom iznosi između 3-5 gb.Baterija traje uglavnom oko 3000 mAh.Težina telefona je oko 150-210 grama.Većina telefona je objavljena između 2013-2015 godine.Ljudi su koristili telefone u proseku između 600-800 dana, što je otprilike 2-2,5 godine.



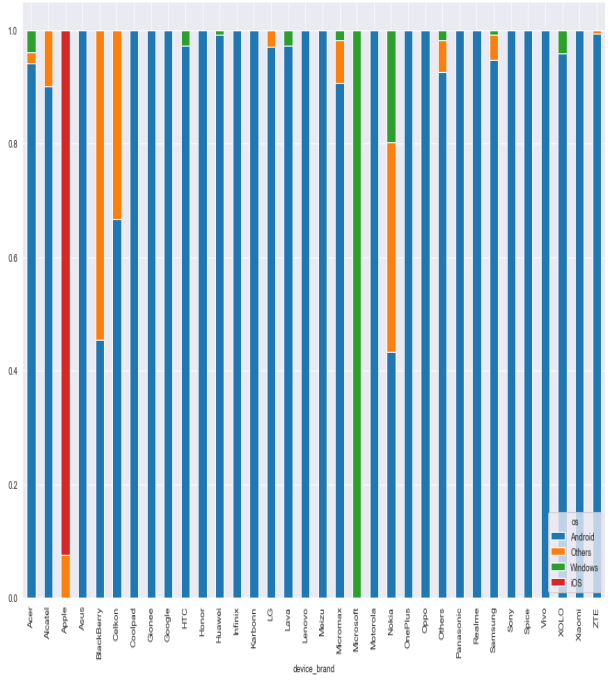
Slika 3. Heatmapa sa koeficientima korelacije izmedju atributa

Slika 4. Korelacija izmedju cene novoh telefona i cenr koriscenog telefona

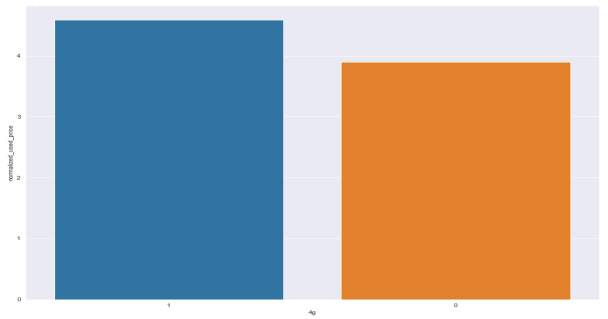
Slika 5. Korelacija izmedju tezinr telefona i cene koriscenog telefona



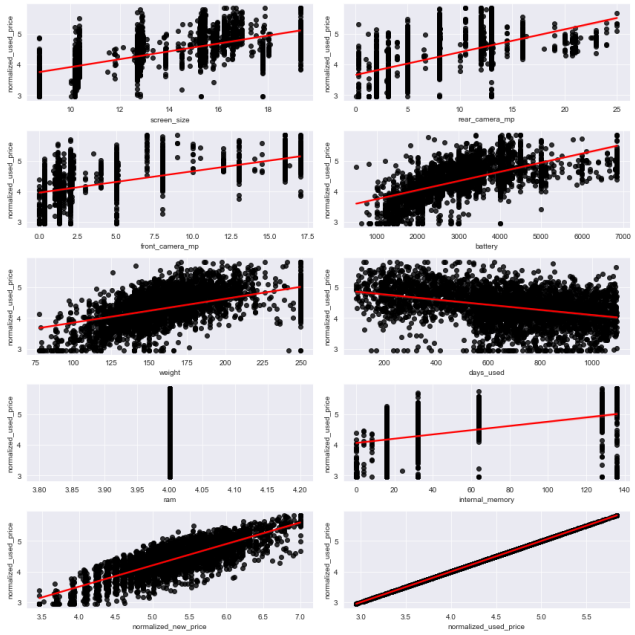
Slika 6. Cena koriscenih uredjaja i naziv brenda



Slika 7. Operativni sistem i brend telefona



Slika 8. 4g(plavi plot telefon podrzava 4g, narandzasti telefon ne podrzava 4g) i cena koriscenih telefona



Slika 9. Regresione linije za numericka obelezja

III. *Regresija*

Na osnovu slike 9. mozemo zakljuciti da dani koriscenja uredjaja imaju opadajucu(negativnu) regresionu liniju, dok ostala obelezja imaju rastucu(pozitivnu) regresionu liniju.

Kako bi se izvrsilo predvidjanje normalizovane cene polovnih telefona obuceno je 5 modela linerne regresije.Definisane su mere uspesnosti modela: srednja kvadratna greska(mse),koren srednje kvadratne greske(rmse), srednja apsolutna greska(mae) i R^2. Izvršena je podela podataka na 70% podataka za treniranje modela, 15% podataka za validaciju, dok je preostalih 15% korisceno za test skup i korišćeno tek za estimaciju performansi konačnog modela, o čemu će biti reči u nastavku. Za svaki od isteriranih modela je izvrsena evaluacija nad validacionim skupom,i izabran je model koji daje najbolje rezultate.Pre procesciranja podataka izvrsena je standardizacija obelezja, da bi se skalirale i trasformisale karekteristike na zajednicki opseg(jedna karakteristika sa vecom skalom moze dominirati nad ostalim karakteristikama) i da bi distibucija karakteristika bila vise normalizovana(Gaussian distribusion).

Prvi model regresije koji je isteran je osnovni model linearne regresije bez selekcije obelezja.Ovaj najosnovniji model je dao odlicne rezultate(slika 10. nulti model).Nakon primene linearnog regresijskog modela, izvršena je regularizacija modela korišćenjem Lasso i Ridge regresije bez selekcije obelezja. Unakrsna validacija je korišćena kako bi se izabrao optimalni parametar alpha, koji kontroliše jačinu regularizacije. Kroz analizu rezultata, primećeno je da je Lasso regresija dala bolje rezultate u odnosu na Ridge regresiju(slika 10. prvi i drugi model). Unakrsna validacija je omogućila odabir optimalne vrednosti parametra alpha, što je rezultiralo poboljšanjem performanse modela u odnosu na linearnu regresiju. Ovo ukazuje na efikasnost Lasso i Ridge regresije u smanjivanju overfittinga i poboljšanju generalizacije modela u slučajevima kada postoji više ulaznih karakteristika ili visoka multikolinearnost.

Unakrsna validacija je pokazala da je najbolja vrednost alpha parametra za Lasso regresiju 0.001, a za Ridge regresiju 10.Nakon primene regularizacije, isprobani su i drugi modeli, kao što su Decision Tree i Random Forest bez selekcije obelezja. Međutim, Decision Tree model je pokazao loše rezultate u poređenju sa ostalim modelima, što ukazuje na njegovu slabiju sposobnost za generalizaciju na novim podacima(slika 10. model tri). Sa druge strane, Random Forest model je pokazao slično dobre rezultate kao Lasso regresija(slika 10. model četri). Dodatno, izvedena je analiza "feature importance" u Random Forest modelu, koja je omogućila identifikaciju značajnih karakteristika (atributa) koje su doprinele predviđanju ciljne promenljive(slika 11.).Ovaj pristup je koristan u razumevanju uticaja pojedinačnih atributa na modeliranje.Zakljuceno je da karakteristike koje dominiraju su u opdadajucem redosledu: normalizovana cena novog uredjaja(najveci "feature importance"), velicina ekrana,interna memorija, zadnja kamera, baterija, tezina,dani koriscenja uredjaja, prednja kamera.Random Forest model se sastoji od vise stabala odlucivanja koji se treniraju na razlicitim podskupovima podataka.Konacna predikcija se dobija agregacijom predikcija svakog stabla , sto cini ovaj model stabilnim i otpornim na overfitting.Hiperparemtri koji su odabrani za Random forest regresiju dobijeni su unakrsnom validacijom i vrednosti tih parametara varira,ali najbolje su se pokazale vrednosti: “n_estimators”=150, ”max_depth”=10, “min_samples_split”=2.

U nastavku je eksperimentisano sa redukcijom dimenzionalnosti primenom PCA algoritma.Na slici 15. je predstavljena scree plot metoda objasnjenja varijanse i broja komponenti,koji je slican elbow metodi, i analizira grafikon padanja objasnjenjene varijanse sa svakom novom komponentom.Cilj je naci lakat ,tj. broj komponenti koji se nalazi pre ‘ravnog’ dela grafa.U nasem slucaju optimalan broj komponenti je 8.Ako ne specifiramo broj komponenti,primeni ce se granicni uslov,tj. da broj komponeti bude onaj koji zadrzava 95% varijablinosti uzoraka.U nasem slucaju se boje pokazao defaultni metod nego prosledjivanje broja kompontenti kao parametar pca algoritma.Performanse svakog od regresionih modela gde je koriscen pca algoritam se nalaze na slici 10(modeli 5-9).

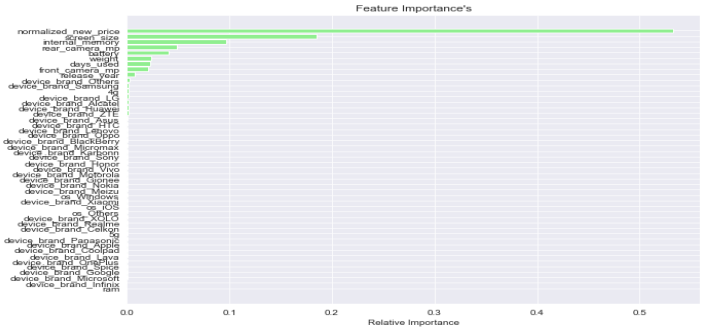
Nakon isprobavanja različitih metoda za redukciju dimenzionalnosti, analizirali smo feature importance i mutual information kako bi identifikovali atribute koji imaju najveću korelaciju sa zavisnom promenljivom. Feature importance se odnosi na ocenjivanje značaja svakog atributa u modelu, dok mutual information meri međuzavisnost između atributa i ciljne promenljive(slika 12.). Korišćenjem ovih metoda, izdvojeni su atributi koji su najviše korelisani sa zavisnom promenljivom,koje smo koristili u daljoj analizi podataka. Identifikacija ključnih atributa može biti od velike važnosti u razumevanju suštine problema i optimizaciji modela, kao i u smanjenju dimenzionalnosti podataka na suštinske informacije koje su relevantne za predikciju ciljne promenljive.Eksperimentisano je sa vredostima k prilikom selekcije broja obelezja i najbolji rezultat je bio pri selekciji 8 najbitnijih obelezja.Nakon izdvajanja 8 najbitnijih atributa na osnovu feature importance i mutual information,ponovili smo obuku već ranije koriscenih regresionih modela. Prilikom smanjenjenja broja atributa, redukcija dimenzionalnosti podataka je rezultirala poboljšanjem performansi Ridge,Lasoo i Linear regresionih modela u poređenju sa prethodno korišćenim modelima koji su koristili sve atribute(slika 10. modeli 5-9).

Ovo ukazuje na to da su izdvojeni atributi bili dovoljni za efikasniju predikciju ciljne promenljive u poređenju sa upotrebom svih dostupnih atributa.Na slici 10. su predstavljeni regresioni modeli sa izvojenih top 8 featura(modeli 10-14).

Linearna,Ridge i Lasso regresija sa izvojenih top 8 featura imaju isti skor i greske.Sva tri modela su jednako efektivni u predvidjanju zavisne varijable.Međutim, razlike u njihovim pristupima regularizaciji i selekciji funkcija mogu dovesti do razlika u njihovoj interpretabilnosti i generalizabilnosti.U nastavku cemo izvrsiti evaluaciju nad test skupom ova tri modela i odlucicemo koji je najbolji model.

	models	R2	RMSE	MSE	MAE
0	Linear regression all features	0.8329	0.2226	0.0496	0.1829
1	Ridge regression all features	0.8329	0.2226	0.0496	0.1828
2	Lasso regression all features	0.8334	0.2223	0.0494	0.1829
3	Decision Tree regression all features	0.6864	0.3050	0.0930	0.2370
4	Random Forest regression all features	0.8338	0.2220	0.0493	0.1801
5	PCA Linear regression	0.8329	0.2226	0.0496	0.1829
6	PCA Ridge regression	0.8337	0.2221	0.0493	0.1824
7	PCA Lasso regression	0.7146	0.2989	0.0846	0.2306
8	PCA Decision Tree regression	0.3451	0.4407	0.1943	0.3336
9	PCA Random Forest regression	0.6313	0.3307	0.1094	0.2530
10	Linear regression top 8 features	0.8346	0.2215	0.0491	0.1835
11	Ridge regression top 8 features	0.8346	0.2215	0.0491	0.1835
12	Lasso regression top 8 features	0.8346	0.2215	0.0491	0.1835
13	Random Forest regression top 8 features	0.8320	0.2232	0.0498	0.1786
14	Decision Tree regression top 8 features	0.6836	0.3063	0.0938	0.2376

Slika 10. evaluacija performansi nad validacionim skupom svakog koriscenog modela



Slika 11. feature importance

	MI Scores
normalized_new_price	0.603288
screen_size	0.475478
battery	0.418283
front_camera_mp	0.409601
rear_camera_mp	0.363841
internal_memory	0.352839
weight	0.317154
release_year	0.205303
4g	0.181775

1-10 Length: 48, dtype: float64

Slika 12. mutual information

IV. ZAKLJUČAK

Nakon testiranja Ridge regresionog modela sa top 8 featura na test podacima, dobili smo sledeće rezultate: R^2 skor za test podatke je 0.8417, RMSE (srednja kvadratna greška) je 0.2274, MSE (srednja kvadratna greška) je 0.0517, a MAE (srednja apsolutna greška) je 0.1777.

Nakon testiranja Linearne regresije sa top 8 featura na test podacima, dobili smo sledeće rezultate: R^2 skor za test podatke je 0.8419, RMSE (srednja kvadratna greška) je 0.2273, MSE (srednja kvadratna greška) je 0.0516, a MAE (srednja apsolutna greška) je 0.1776.

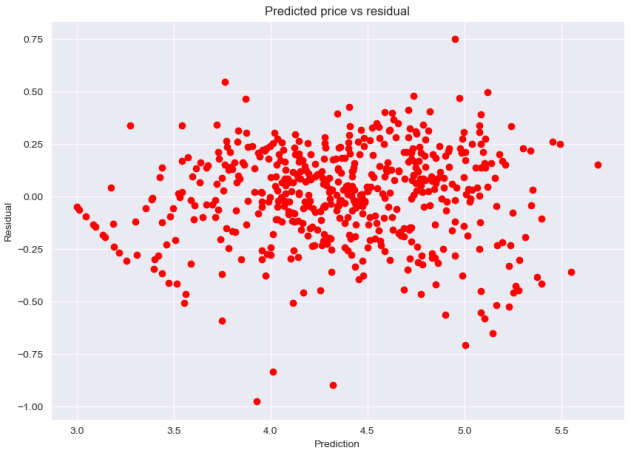
Nakon testiranja Lasso regresionog modela sa top 8 featura na test podacima, dobili smo sledeće rezultate: R^2 skor za testiranje je 0.8419, RMSE (srednja kvadratna greška) je 0.2273, MSE (srednja kvadratna greška) je 0.0516, a MAE (srednja apsolutna greška) je 0.1776. Lasso regresija se pokazala bolje na test skupu u poređenju sa validacionim skupom.Hiperparametri Lasso regresije koji su se najbolje pokazali su: alpha parametar 0.001, i selection criterion 'aic'(Akaike Information Criterio).Naši rezultati pokazuju da Lasso regresija je dala odlicne performanse na test podacima, sa visokim R^2 skorom i relativno niskim vrednostima RMSE, MSE i MAE.

Sva tri regresiona modela su se bolje na test skupu u poređenju sa validacionim skupom.Linearna regresija i Lasso regresija imaju identicne rezultate,ali cemo izabrati Lasso regresioni model kao konacni model zbog moci regularizacije i visoke interpretabilnosti.

Random forest se sa svim atributima se takodje dobro pokazao i treba izvrsiti njegovu evaluciju nad test skupom.Nakon testiranja Random Forest regresionog modela na test podacima, dobili smo sledeće rezultate: R^2 skor za testiranje je 0.8544, RMSE (srednja kvadratna greška) je 0.2182, MSE (srednja kvadratna greška) je 0.0477, a MAE (srednja apsolutna greška) je 0.1791.

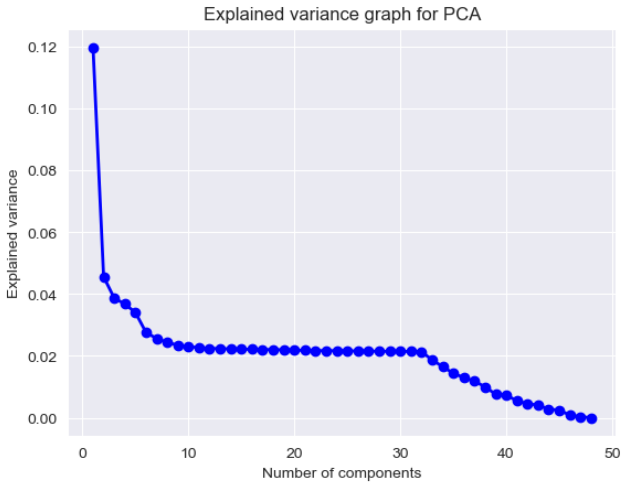
Na slikama 13. i 14. mozemo primetiti da Random Forest model ima vise scattera skoncentrisanih blize nuli i da se bolje fitovao nad podacima nego Lasso model.

Takodje, vidimo da se Random Forest model nad test podacima ima visi R^2 skor i nize vrednosti RMSE, MSE i MAE od Lasso modela i mozemo zakljuciti da je najbolji regresioni model Random Forest.



Slika 13. Rezidul(razlika izmedju posmatrane i prediktovane vrednosti) lasso modela

Slika 14. Rezidul(razlika izmedju posmatrane i prediktovane vrednosti) Random forest modela



Slika 15.Scree plot grafik zavisnosti % objasnjene varianse i broj pca komponetni