Druga domača naloga

Tine Markočič

Maj 2023

1 Uvod in analiza podatkov

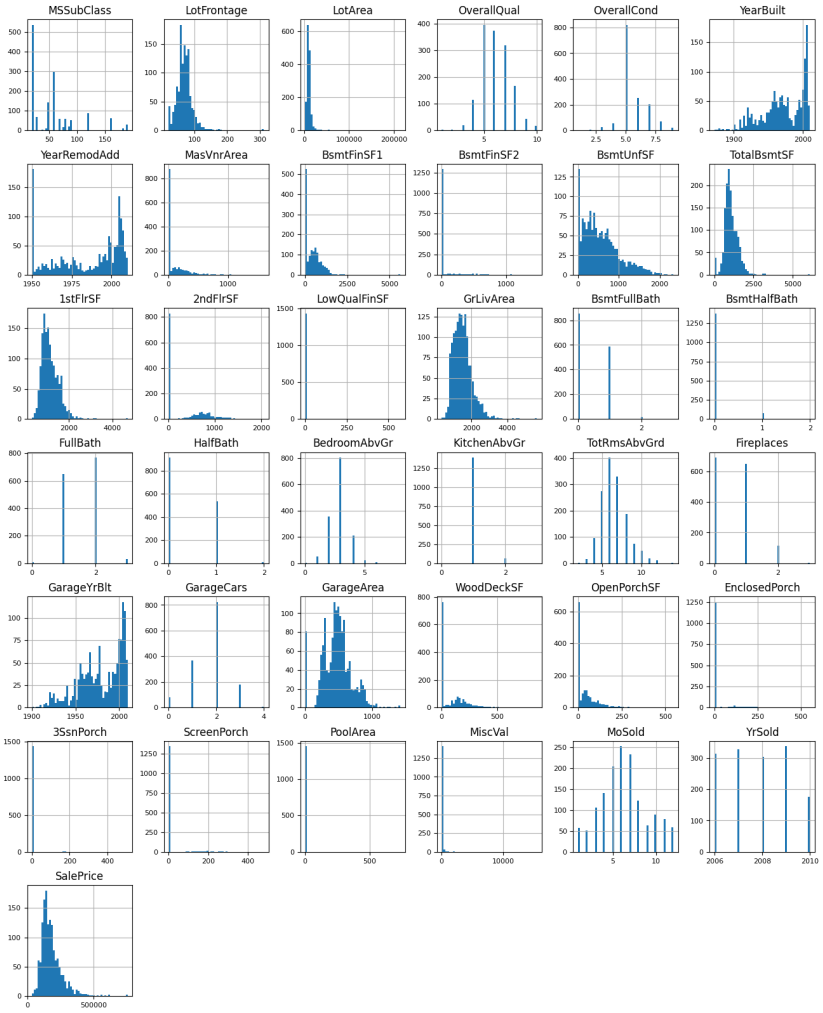
Cilj domače naloge je bil analizirati podatke o hišah na Ameriškem nepremičninskem trgu in čim bolje napovedati njihove cene. Podatki so razdeljeni na učno in testno množico v razmerju 50:50, s tem da učna množica vsebuje še stolpec s cenami, ki jih je potrebno za testne podatke napovedati. Ob prvem pogledu opazimo, da so podatki numerični in kategorični, prav tako pa je pri nekaterih atributih veliko manjkajočih vrednosti.

V datoteki *data_description.txt* je opisano, kaj predstavlja posamezen stolpec v tabeli in katere vrednosti se v njem nahajajo. Ob pregledu atributov si lahko ustvarimo sliko, kateri bi lahko igrali večjo vlogo pri ocenjevanju cene hiše.

Gotovo bo pametno vključiti naslednje atribute:

- 'LotArea' - celotna velikost zemljišča - večje kot je zemljišče, višja cena?
- 'Neighborhood' - soseska v kateri se neprimičnina nahaja - boljša soseska pomeni višjo ceno
- 'OverallQual' - ocenjena splošna kvaliteta hiše - gotovo želimo višjo kvaliteto, hkrati zanjo pričakujemo višjo ceno
- 'OverallCond' - stanje hiše - podobno kot prejšnja alineja
- 'YearBuilt' - leto izgradnje - iz tega lahko ocenimo veliko drugih spremenljivk (stanje električne/vodovodne napeljave, potrebo po obnovitvi,...)
- 'GrLivArea' - velikost površine za bivanje - podobno kot 'LotArea'
- 'Garage Area' - velikost garaže - ZDA je ena izmed držav z največjim deležem motornih vozil na prebivalca, zato bo kupec potreboval dovolj veliko garažo
- 'SaleType' - odvisno od kupca kakšno finančno breme je pripravljen prevzeti

Sedaj si lahko podrobneje ogledamo histograme numeričnih spremenljivk.

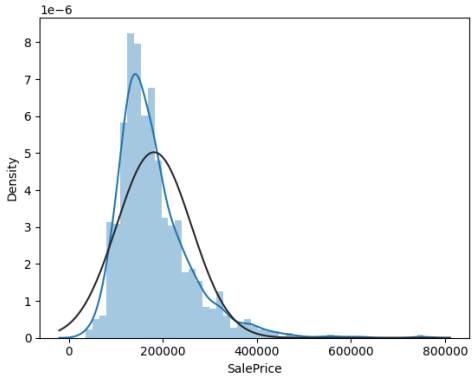


Slika 1: Histogrami numeričnih spremenljivk

- pri 'LotFrontage' in 'LotArea' prevladujejo majhna števila,
- 'OverallQual' in 'OverallCond' sta največji pri vrednostih 5,6 in 7 - zelo malo takih z nizkimi in zelo visokimi vrednostmi,
- Z leti narašča tako število zgrajenih hiš kot tud število zgrajenih garaž, YearRemod ima najvišjo vrednost pri 1950, kar pomeni da so tam verjetno združene vse hiše pred letom 1950 - in niso bile nikoli kasneje obnovljene,
- večina hiš ima klet, ki pa v večini primerov ni končana,

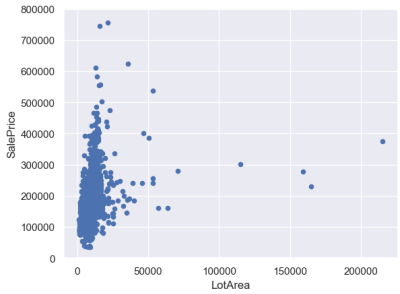
- večina hiš ima eno ali dve kopalnici in nima kopalnice v kleti, ter med 4 in 9 sob,
- največ hiš nima ognjišča, sledijo hiše z enim, redko katere imajo 2,
- večina garaž sprejme 2 avta, ter so velike okoli 500 kvadratnih čevljev,
- večina hiš je bila prodanih v pomladnih oziroma poletnih mesecih, med leti ni velike razlike, rahel upad v 2010 (recesija),
- porazdelitvi prodajnih cen so najbolj podobni histogrami 'LotFrontage', 'GrLivArea', 'TotalBsmtSF' in '1stFlrSF'.

Preden začnemo iskati povezave med atributi in ceno, si najprej oglejmo napovedno spremenljivko 'SalePrice'.



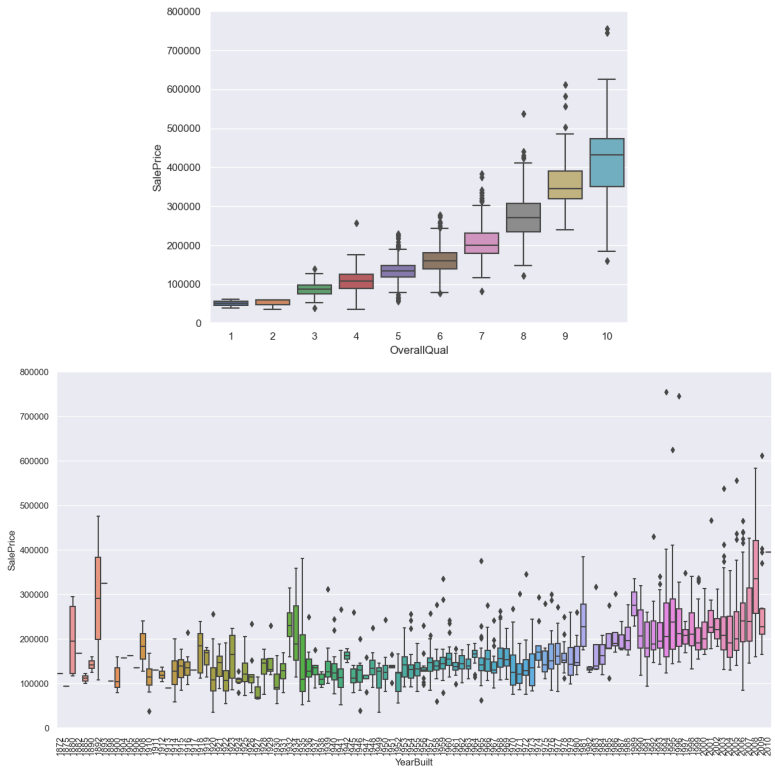
Slika 2: Histogram cene zemljišč

Spremenljivka nima ničelnih vrednosti, kaže odstopanje od normalne porazdelitve. Ima opazen koeficient asimetrije (je očitno asimetrična) 1.882876, ter opazno sploščenost 6.536282.



Slika 3: Povezava med ceno in velikostjo zemljišča

Sedaj lahko preverimo če smo na začetku res izbrali dobre atribute. Kot vidimo na zgornjem grafu je cena nepremičnine neodvisna od velikosti zemljišča, saj se opazi vse mogoče range cene, pri isti velikosti zemljišča, cena za večje zemljišče ni občutno večja. Morda velikost zemljišča in cena nista tako zelo korelirani kot bi intiutivno mislili.



Slika 4: Boxplota kvalitete hiše in leta izgradnje

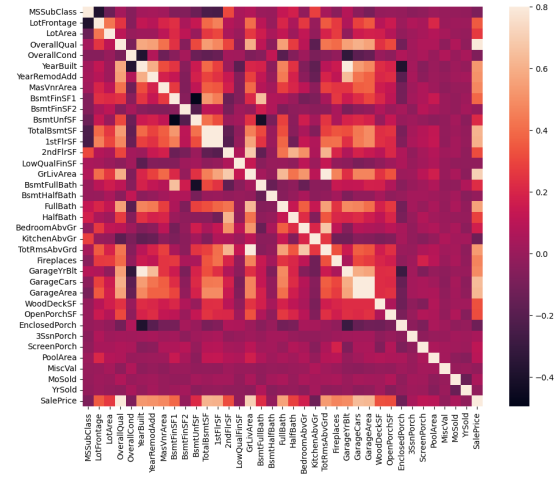
Boljšo korelacijo vidimo med ceno hiše in njeno splošno kvaliteto. Višja kot je ocena hiše višja je njena cena. Opažamo najmanj kvadratičen, morda celo rahlo eksponenten trend.

Tudi med letom gradnje in ceno hiše je opaziti rahlo korelacijo. Imamo večji pas znotraj katerega je večina cen, ki počasi z leti (verjetno z rahlim kvadratičnim trendom) narašča. Seveda se v vseh letih pojavljajo tudi odstopanja.

V podatkih imamo še veliko več drugih spremenljivk, zato bo najbolje, da se na začetku posvetimo vsem hkrati in nato po potrebi zožimo izbor.

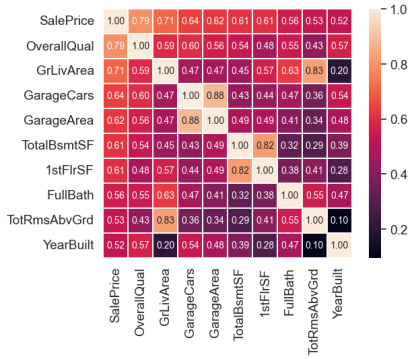
2 Izbira in priprava atributov

Sedaj, ko poznamo osnovne odnose med nekaterimi atributi in napovedno spremenljivko, lahko pogledamo širšo sliko. S korelacijsko matriko lahko združimo skupaj vse numerične atribute in vidimo koliko so odvisni med seboj.



Slika 5: Korelacijska matrika - vsi numerični atributi

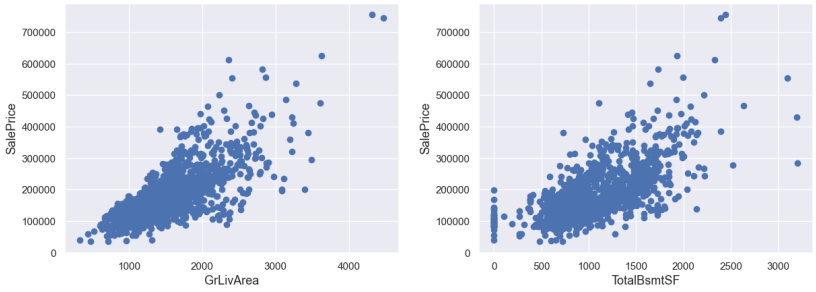
V matriki svetlejša polja predstavljajo zelo močno pozitivno korelacijo med spremenljivkami. Predvsem bodeta v oči dva bela kvadratka na diagonali. V obeh primerih gre za spremenljivki, ki pri napovedovanju cene povesta dvakrat isto stvar. Ta problem bomo v nadaljevanju poskusili rešiti. Naprej lahko vidimo da v podatkih nimamo nikjer izrazite negativne korelacije. Zadnji stolpec (ali zadnja vrstica) prikazuje katere spremenljivke najbolj korelirajo s ceno hiše oz. imajo pri napovedovanju te največji vpliv.



Slika 6: Korelacijska matrika - 10 najpomembnejših atributov

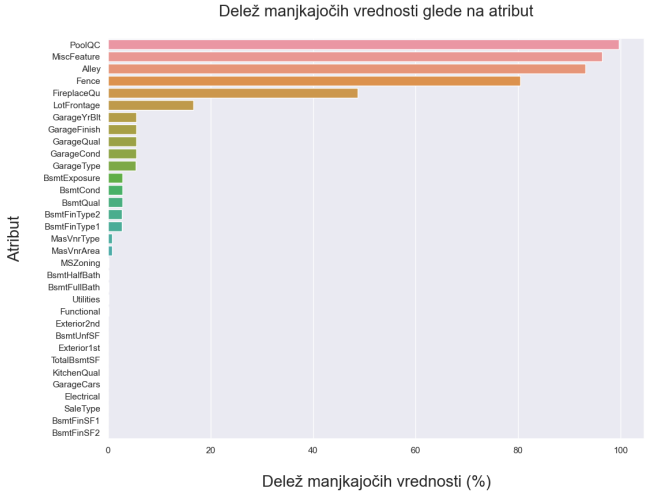
Za natančnejšo analizo izpišemo 10 najvplivnejših v novo korelacijsko matriko.

Opazimo lahko, da je med najbolj koreliranimi desetimi več spremenljivk, ki opisujejo podobne lastnosti - 'GrLivArea' in 'TotRoomsAbvGrd' predstavljajo zelo podobno stvar - bivalni prostor, nadalje 'GarageCars' in 'GarageArea' prav tako opisujeta podoben parameter - velikost garaže. Nazadnje 'TotalBsmtSF' in '1stFlrSF' predstavljata površino hiše, ki je verjetno pri obeh zelo podobna. V podatkih imamo tako imenovano multikolinearnost. Odločimo se, da bomo izločili te duplikate, da ne bi imeli težav, ter tako ohranili le 'SalePrice', 'OverallQual', 'GrLivArea', 'GarageCars', 'TotalBsmtSF', 'FullBath', 'YearBuilt'.



Slika 7: Izločene odstopajoče vrednosti

Naprej pri spremenljivkah 'GrLivArea' in 'TotalBsmtSF' opazimo nekaj odstopajočih podatkov, zato jih odstranimo, da tej našega modela ne bodo preveč zmedli.

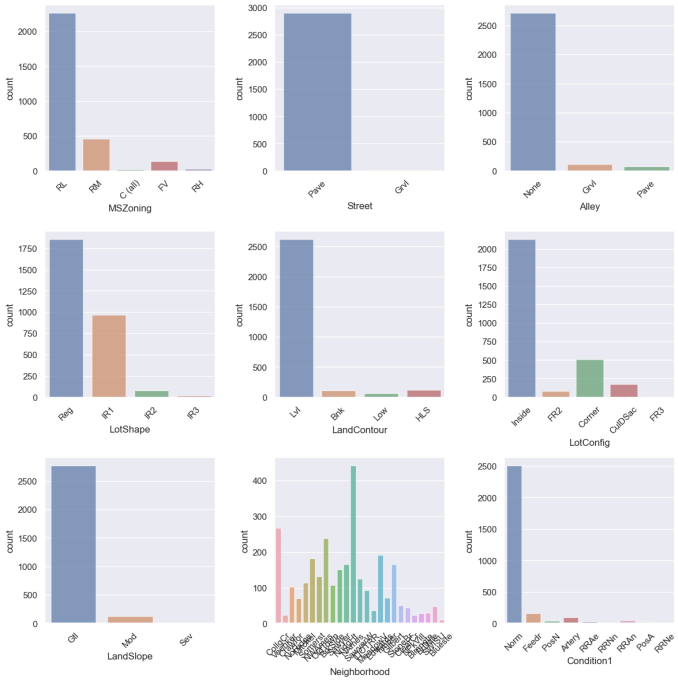


Slika 8: Odstotek manjkajočih vrednosti

Na tej točki se zdi smiselno združiti učno (brez stolpca 'SalePrice') in testno množico, saj je pred nami nov problem, in sicer manjkajoče vrednosti. Te so prisotne tako v učnih k v testni množici in tako bomo z združevanjem ubili dve muhi na en mah.

Vidimo, da je pri nekaterih spremenljivkah manjkajočih celo več kot 90% podatkov, kljub temu pa teh atributov ni potrebno odstraniti, saj lahko vsebujejo pomembne vrednosti, ki bodo ločile med cenami hiš. Vrednost spremenljivke 'PoolQC' lahko namreč izredno poveča vrednost nepremičnine.

Problem bomo rešili z imputacijo vrednosti, pri tem si bomo pomagali z opisno datoteko, ki je omenjena v uvodu. Kjer si z opisno datoteko ne bomo mogli pomagati, bomo uporabili mediano ali najpogostejšo vrednost. Imputacija za posamezen atribut je zakomentirana v kodi. Na koncu odstranimo stolpce 'Utilities', 'MasVnrType', 'MasVnrArea', saj se nam ne zdijo uporabni.

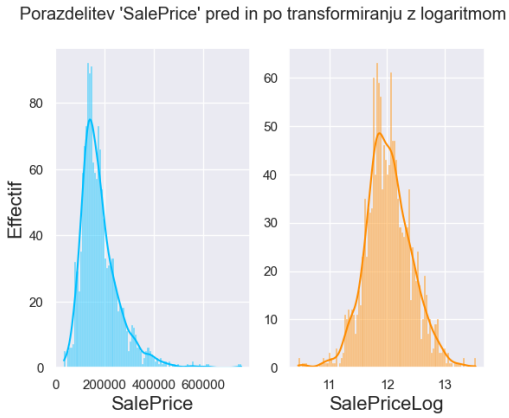


Slika 9: Histogrami kategoričnih spremeljivk

Na koncu si ogledamo še kategorične atribute. Na sliki 9 so histogrami prvih 9 kategoričnih stolpcev. Čeprav izgleda, da nekateri ponujajo večjo informacijo od ostalih, je težko določiti, katere bi zares lahko izpustili. Tudi spremenljivka 'Street', ki izgleda popolnoma neuporabna, saj ima skoraj vedno samo eno vrednost, pa lahko v tistih nekaj primerih naredi razliko. Tako se torej odločimo obdržati vse kategorične atribute. Za regresijo jih transformiramo v nove stolpce, tako imenovane "dummy variables".

3 Strojno učenje

Podatki za učenje so že skoraj pripravljeni, rešiti moramo še problem asimetričnosti napovedne spremenljivke. To storimo tako, da jo logaritmiramo. Primerjavo porazdelitev vidimo na spodnjem grafu.



Slika 10: Korelacijska matrika

Logaritmirana porazdelitev izgleda precej bližje normalni kot nelogaritmirana, kar je bolj ugodno za večino modelov strojnega učenja, zato bomo modele učili na logaritmiranih podatkih.

Modele ovrednotimo s prečnim preverjanjem, pri čemer jih enkrat ocenjujemo z metodo RMSE, drugič pa z R^2.

Učimo naslednje modele:

- $DummyRegressor()$
- $LinearRegression()$
- $ARDRegression(iteracije = 30)$
- $BayesianRidge(iteracije = 30)$
- $Ridge()$
- $Lasso(\alpha = 0.001)$
- $ElasticNet(\alpha = 0.001)$
- $SVR()$
- $RandomForestRegressor()$
- $XGBRegressor()$
- $LGBMRegressor()$

Po ovrednotenju s prečnim preverjanjem so se za najboljše tri izkazali modeli Lasso, ElasticNet in LGBMRegressor. Te modele natreniramo na učnih podatkih ter jih poženemo na testnih. Rezultate zabeležimo v csv datoteko in naložimo na kaggle.

4 Zaključek

Na testni množici je najmanjšo napako priredil model LGBMRegressor, in sicer 0.13186. Blizu zadaj pa sta bila tudi ostala dva modela. Napako bi se zagotovo dalo še veliko zmanjšati. Regresijskim modelom bi lahko optimizirali hiperparametre, lahko bi vzeli ansambel več različnih modelov, poskusili bi lahko tudi z meta učenjem. Morda bi lahko pametneje izbrali atribute, saj so verjetno nekatere spremenljivke še vedno veliko odvisne med sabo in tako negativno vplivajo na napoved cene. Lahko bi pomagalo tudi dodatno izračunati kakšen nov atribut.