Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Aleš Koncilja

Napovedovanje vrednosti nepremičnin iz podatkov Evidence trga nepremičnin

magistrsko delo

magistrski program druge stopnje
računalništvo in informatika

Mentor: doc. dr. Dejan Lavbič

Ljubljana, 2018

Zahvala

Zahvaljujem se mentorju doc. dr. Dejanu Lavbiču za vso strokovno pomoč in usmerjanje pri izdelavi magistrskega dela. Zahvala gre tudi Mateju za strokovno pomoč in vsem prijateljem za lepo preživeta študijska leta. Zahvaljujem se tudi Janji za vso podporo in spodbudne besede. Posebna zahvala pa gre mojim staršem, ki so me skozi celoten študij podpirali in mi stali ob strani.

Aleš Koncilja, 2018

Staršem.

Kazalo

Seznam uporabljenih kratic

kratica	angleško	slovensko
SQL	structured query language	strukturirani povpraševalni jezik
CSV	comma-separated values	vrednosti, ločene z vejicami
GURS	The Surveying and Mapping Authority of the Republic of Slovenia	Geodetska uprava Republike Slovenije
ETN	Real Estate Market Record	Evidenca trga nepremičnin
REN	Real Estate Register	Register nepremičnin
SURS	Statistical Office of the Republic of Slovenia	Statistični urad Republike Slovenije
PCA	principal component analysis	metoda glavnih komponent
OLS	ordinary least squares	navadni najmanjši kvadrati
MAE	mean absolute error	povprečna absolutna napaka
MSE	mean square error	povprečna kvadratna napaka
RMSE	root mean square error	koren povprečne kvadratne napake
BSI	The Bank of Slovenia	Banka Slovenije
MGRT	Ministry of Economic Development and Technology	Ministrstvo za gospodarski razvoj in tehnologijo
CHAID	chi-squared automatic interaction detector	avtomatski interaktivni detektor chi-kvadrat

kratica	**angleško**	**slovensko**
CART	classification and regression trees	klasifikacijska in regresijska drevesa
KNN	k-nearest neighbors	k-najbližjih sosedov
LOESS	local regression	lokalna regresija
NAMEA	national accounting matrix including environmental accounts	nacionalna računovodska matrika z okoljskimi računi
ECB	European Central Bank	Evropska centralna banka
IQR	interquartile range	interkvartilni obseg
VIM	variable importance measure	pomembnost spremenljivke

Povzetek

Naslov: Napovedovanje vrednosti nepremičnin iz podatkov Evidence trga nepremičnin

Trgovanje z nepremičninami (oddajanje, prodajanje) poteka vsak dan, zato je napovedovanje vrednosti nepremičnin zelo pomembno. Cilj magistrske naloge je bil razviti napovedni model vrednotenja nepremičnin s podatkovnim rudarjenjem, ki napoveduje vrednost nepremičnine (pogodbeno najemnino oz. ceno za stanovanja) na podlagi podatkov iz različnih virov. Pomemben faktor pri pripravi zbirk podatkov je bil vključevanje podatkov, ki posredno vplivajo na vrednost nepremičnin. Izhodiščno zbirko podatkov ETN smo razširili z dodatnimi podatki in ustvarili dve novi zbirki podatkov – najeme in kupoprodaje stanovanj. Nad zbirkama smo izvajali postopke čiščenja (odstranjevanje osamelcev, vstavljanje manjkajočih vrednosti). Izvedli smo tudi izbor pomembnih atributov. Z metodama za napovedovanje (linearna regresija, naključni gozdovi) smo iz zbirk podatkov zgradili napovedne modele za napovedovanje vrednosti nepremičnin ter jih ovrednotili.

Pri napovedovanju pogodbenih cen za stanovanja smo z naključnimi gozdovi dosegli najnižjo povprečno absolutno napako (MAE) 10.986,15 €, kar je boljše kot z linearno regresijo, kjer je MAE 14.496,75 €. Obe metodi presežeta MAE 25.424,58 € ničelnega modela. Tudi pri napovedovanju pogodbenih najemnin za stanovanja smo z naključnimi gozdovi dobili boljše rezultate (MAE je 63,74 €) kot z linearno regresijo (MAE je 81,20 €), kar je boljše od ničelnega modela (MAE je 95,15 €).

Napovedni model vključuje stanje trga in predstavlja alternativo trenu-

tnemu GURS vrednotenju, ki temelji na kompleksnih modelih vrednotenja.

Ključne besede

evidenca trga nepremičnin, podatkovno rudarjenje, napoved, vrednost nepremičnine, čiščenje podatkov

Abstract

Title: Forecasting the value of Real Estate from Real Estate Market Records

Real Estate trading (renting, selling) is carried out every day, so the forecasting the value of Real Estate is very important. The aim of the master's thesis was to develop a forecast model for Real Estate valuation with data mining, which forecast the value of Real Estate (contract rent or price for apartment) based on data from various sources. An important factor in the preparation of data sets was the integration of data that indirectly affects the value of Real Estate. We extended the baseline REMR data set with additional data and created two new data sets - renting and buying apartments. We carried out cleaning procedures on these data sets (removal of outliers, imputation of missing values). We also carried out a feature selection. Using forecast methods (linear regression, random forests), we made data from the data sets forecast models for forecasting the value of Real Estate and evaluated them.

When forecasting contract prices for apartments, random forest defects reached the lowest mean absolute error (MAE) of € 10,986.15, which is better than with linear regression, where the MAE is € 14,496.75. Both methods exceed the MAE of € 25,424.58 of the null model. Also in forecasting contractual rents for apartments, random forests have obtained better results (MAE is € 61.57) than with linear regression (MAE is € 81.20), which is better than the null model (MAE is € 95.15).

The forecast model includes the state of the market and represents an alternative to the current MGRT evaluation, based on complex evaluation

models.

Keywords

Real Estate Market Record, data mining, forecast, value of Real Estate, data cleaning

Poglavje 1

Uvod

Trgovanje z nepremičninami (t.j. prodaja in nakup nepremičnin oz. oddajanje in najemanje nepremičnin) poteka vsak dan. To ima pomembno vlogo v gospodarskem razvoju in temeljnih potrebah ljudi. Vsak, ki želi nepremičnino prodati ali jo kupiti oz. oddati ali najeti, opravi predhodno cenitev oz. vrednotenje le-te na podlagi lastnosti nepremičnine. Natančno napovedovanje vrednosti nepremičnin je torej zelo pomembno. V Sloveniji se z množičnim vrednotenjem nepremičnin ukvarja GURS [1]. Ta na podlagi osnovnih podatkov o nepremičnini s pomočjo različnih modelov določi vrednost nepremičnine. Za vsako vrsto nepremičnine je zgrajen lasten model, tako kot to določa *Uredba o določitvi modelov vrednotenja nepremičnin* [2]. Model za vrednotenje stanovanj vsebuje naslednje podatke: lokacijo, ki določa vrednostno cono in raven, leto izgradnje stavbe, obnove oken, fasade, strehe in inštalacij, vrednost stavbe in zemljišč pod stavbo, uporabno površino, točke in vrednostne faktorje za lastnosti stavbe ter faktor za oddaljenost od linijskih objektov. V praksi se pogosto izkaže, da je takšno vrednotenje lahko le približek dejanski vrednosti nepremičnine. Hkrati pa je tak pristop nepriročen, saj se modeli ne prilagajajo avtomatsko na druge dejavnike, kot so npr. dvig trošarin, povprečna mesečna plača itd. Takšne vplive na ceno nepremičnin skušajo upoštevati s pomočjo posebnih parametrov posameznega modela vrednotenja, ki pa jih je potrebno prilagajati ročno.

Podobno trdijo tudi v [3], kjer so zapisali, da so matematični izračuni cen nepremičnin neodvisni od delovanja trga, torej ponudbe in povpraševanja, ki deluje v skladu s tržnimi zakonitostmi. V primeru postopka vrednotenja posamezne nepremičnine, le-to oceni strokovnjak (cenilec). Cenilec zbere podatke o že izvedenih poslih primerljivih nepremičnin ocenjevani nepremičnini, ki na podlagi zbranih podatkov in informacij o stanju nepremičnine določi vrednost - pogodbeno najemnino ali ceno.

V delu se bomo posvetili izgradnji napovednega modela vrednotenja nepremičnin s podatkovnim rudarjenjem nad podatki iz Evidence trga nepremičnin (ETN) [4]. Pri izgradnji modela bomo upoštevali tudi druge podatke, ki pomembno vplivajo na vrednost nepremičnin oz. nepremičninski trg in tako upoštevali tudi stanje na trgu. Takšne podatke in dejavnike (npr. povprečna plača v občini, poseljenost, življenjski standard, razmere na trgu dela, cene naftnih derivatov...), bomo pridobili iz različnih virov, npr. SI-STAT [5], Google... Upoštevanje takšnih podatkov pri izgradnji napovednega modela bo še ena od izboljšav trenutnemu načinu vrednotenja na GURS - množičnemu vrednotenju nepremičnin s pomočjo modelov vrednotenja, ki so se izkazali za nepriročen pristop [3]. ETN vsebuje tri vrste podatkov, in sicer podatke o kupoprodajnih poslih delov stavb, kupoprodajnih poslih zemljišč in najemnih poslih delov stavb. Zaradi razlik med strukturo podatkov o poslih zemljišč in delov stavb, ter tudi med najemnimi in kupoprodajnimi posli, se bomo najprej posvetili obliki zapisa podatkov ter analizo pričeli z najemnimi posli stanovanj. Poskušali bomo predlagati metodologijo za pripravo podatkov, primernih za izgradnjo napovednega modela, ki bo sestavljena iz več metod in postopkov, npr. zajem smiselnih podatkov iz različnih virov, analizo podatkov, čiščenje podatkov, izbor pomembnih atributov za napovedovanje vrednosti nepremičnin. Napovedni model bomo gradili z različnimi tehnikami napovedovanja, in sicer na podlagi podatkov že izvedenih poslov nepremičnin iz ETN z upoštevanjem stanja trga, t.j. drugih podatkov. Napovedne modele bomo tudi testirali in napovedi ovrednotili z različnimi metrikami.

Preostanek magistrske naloge je sestavljen iz 8 poglavij. V uvodnem poglavju smo podali pregled obstoječih del na področju napovedovanja vrednosti nepremičnin, v tretjem poglavju pa smo opisali postopke zbiranja podatkov iz različnih virov ter združitev podatkov v dve novi zbirki podatkov. Poglavje 4 opisuje uporabljeno metodologijo za analizo podatkov, iskanje in odstranjevanje osamelcev, vstavljanje manjkajočih vrednosti ter izbor najpomembnejših atributov za razlago odvisne spremenljivke. V nadaljevanju smo opisali uporabljene metode za napovedovanje, postopek testiranja in metrike za ocenjevanje napovednih modelov. Šesto poglavje predstavi rezultate testiranja napovednih modelov in oceno uspešnosti posameznih napovednih modelov. Na koncu sledijo sklepne ugotovitve in zaključek, kjer smo našteli tudi možne izboljšave in nadgradnje.

Poglavje 2

Pregled sorodnih del

V Sloveniji poznamo dve vrsti vrednotenja nepremičnin, in sicer posamično ter množično vrednotenje, kot je opisano v [6, 7]. Posamično vrednotenje pomeni, da strokovna oseba (cenilec) oceni eno nepremičnino na podlagi informacij trga in informacij o nepremičnini. Pri množičnem vrednotenju pa se ocenjuje več nepremičnin. Množično vrednotenje v Sloveniji izvaja GURS, ki s pomočjo različnih in kompleksnih modelov določi vrednost nepremičnine, kot to opisujejo v [2]. Modeli zajemajo več dejavnikov, ki vplivajo na vrednost nepremičnine. Ta način vrednotenja se izkaže za dobrega, vendar je za razvoj modelov potrebno veliko znanja in dela, obenem pa je potrebno faktorje, ki jih upoštevajo modeli, nujno ves čas posodabljati. Matematični izračuni cen nepremičnin so neodvisni od delovanja trga, torej ponudbe in povpraševanja, ki deluje v skladu s tržnimi zakonitostmi, kot to trdijo v [3].

V preteklosti je bilo že izvedenih nekaj poskusov napovedovanja s podatkovnim rudarjenjem nad podatki ETN [6], vendar trenutno še vedno ni razvitega sistema, ki bi lahko s podatkovnim rudarjenjem nad podatki o že sklenjenih poslih (ETN) ocenil vrednost nepremičnine. Mi pa želimo s podatkovnim rudarjenjem izboljšati trenutni GURSov način vrednotenja z modeli in odkriti parametre, ki pomembno vplivajo na vrednost nepremičnine. Uporabiti želimo tudi parametre, ki vplivajo na kvaliteto življenja v posamezni občini, kot to opisujejo v [8, 9].

Podatkovno rudarjenje je primeren pristop za napovedovanje vrednosti nepremičnin, ki iz množice podatkov razvije enega ali več modelov [6]. Tak pristop napovedovanja zahteva v prvi meri pripravo podatkov [10], t.j. analizo, vizualizacijo in čiščenje podatkov [11]. Čiščenje podatkov ETN so izvajali tudi drugi [12, 13]; posamezne zapise so združili, odstranili dvojnike in tudi tiste, ki po ključnih vrednostih odstopajo od večine podatkov. Podoben način so uporabili tudi v [11]; podatke so čistili v treh korakih, in sicer so najprej odstranili transakcije, ki so se pojavile večkrat, nato pa še transakcije, ki so odstopale od določenih meja in ta korak zopet ponovili.

Pred pričetkom rudarjenja je potrebno podatke analizirati in s pomočjo analiz odkriti tiste podatke, ki pomembno vplivajo na vrednost nepremičnine. Taki podatki so primerni za grajenje napovednega modela [6, 7]. Med pomembne podatke tako štejejo nekatere lastnosti nepremičnine: lokacija, občina, leto izgradnje stavbe, neto tlorisna površina stavbe, število prostorov, opremljenost... Poleg samih lastnosti stavbe, pa na vrednost nepremičnine vplivajo tudi drugi dejavniki, kot to navajajo v [14, 15]. Avtorji opisujejo, katere lastnosti so tiste, ki vplivajo na cene stanovanj v predmestju Bostona (kriminal, število sob, onesnaženost, oddaljenost od zaposlitvenih centrov). Linearna regresija je metoda, ki je bila v tem primeru bolj učinkovita kot odločitvena drevesa. Za izboljšanje učinkovitosti predlagajo odstranitev nepomembnih spremenljivk iz modela ter zmanjšanje korelacije nad spremenljivkami.

Posebno pozornost je potrebno nameniti metodam množičnega vrednotenja in njihovi uporabi tudi pri razvoju davčnega sistema. Večkratna regresijska analiza je le ena od metod, ki se uporabljajo v ta namen. V [16] so se osredotočili na značilnosti te metode in prednosti pri razvoju sistema množičnega vrednotenja. V članku opisujejo pristop z več-regresijskim modelom. Z regresijsko analizo so skušali oceniti vrednost nepremičnine s tremi lastnostmi (površina bivalnega prostora, garaže in starost nepremičnine).

Ideja o izgradnji napovednih modelov z metodami podatkovnega rudarjenja za napovedovanje vrednosti nepremičnin ni nova. Obstaja že precej

poskusov izgradnje napovednih modelov, ki pa so bili v večini izvedeni s pristopom strojnega učenja in nevronskih mrež [17]. Raziskava je pokazala uporabnost podatkov za rudarjenje, še posebej z nevronskimi mrežami in odločitvenimi drevesi. Algoritem odločitvenih dreves je v tem primeru pripeljal do najboljših rezultatov, saj je model dokaj enostaven za razumevanje in proizvaja najmanjšo napako (MAE). Poleg odločitvenih dreves so uporabili tudi najmanjše število napovedovalcev, da so prišli do rešitve. Nevronske mreže prikazujejo pomembne rezultate pri napovedih vrednosti stanovanj, ki so v nekaterih simulacijah dosegle celo 96 % [18].

Pri učenju regresijskih modelov je lahko poseben izziv, če iščemo zanimive zbirke podatkov v realnem življenju, katere omogočajo analize, ki vse pojme združujejo v en velik primer. V članku [19] so opisali celovito linearno regresijo analize podatkov o cenah stavb, ki zajema veliko regresijskih tem, vključno s prepletanjem interakcij in napovedne transformacije ter tudi praktične nasvete o oblikovanju modelov. V prvi fazi so podatke analizirali in postavili hipoteze. Izvedli so tudi t.i. raziskovalno analizo podatkov, kjer so ugotavljali povezave med atributi z razpršeno matriko količinskih spremenljivk v naboru podatkov ter odvisnost spremenljivk s ceno. Napovedne modele so gradili z linearno regresijo in rezultate sproti primerjali. Izvedli so tudi t-test za ugotavljanje povezav med spremenljivkami in na podlagi *p-vrednosti* nekatere spremenljivke odstranili. To je pripomoglo k izboljšanju ocene prilagojenega deleža razložene variance (prilagojeni R^2) in regresijske standardne napake (s).

V večih primerih se je izkazalo, da je za napovedovanje vrednosti nepremičnin s podatkovnim rudarjenjem linearna regresija zelo uporabna metoda za napovedovanje. V [10] so uporabljali in primerjali različne regresijske metode - linearno regresijo, regresijo podpornih vektorjev (SVR), k-najbližjih sosedov (kNN) ter metodo naključnih gozdov. Izvedli so tudi linearno regresijo z uporabo regularizacije. Pri napovedih so uporabili večkratno (10) navzkrižno validacijo, pogledali so srednjo absolutno razliko in njeno varianco. Ugotovili so, da je med srednjo vrednostjo in varianco potrebna korelacija,

vendar zmanjšanje obeh pomeni izboljšanje delovanja modela. Predstavili so podrobna vprašanja o napovedovanju vrednosti nepremičnin, način analiziranja podatkov o nepremičninah ter predstavili rezultate testiranja napovedi. Izkazalo se je, da sta bili v tem primeru metodi kNN in naključni gozdovi najboljša regresijska modela. Bila sta boljša tudi od linearne regresije. Razlog za to je morda njihova zmožnost, da upoštevata nelinearne interakcije med številčnimi značilnostmi in cenovnimi cilji. Ugotovili so, da je analiza kNN z naključnimi gozdovi izboljšala napoved. Podobne metode so uporabili tudi v članku [20]. Uporabili so linearno regresijo in naključne gozdove, za izboljšanje napovedi pa so izvedli prečno preverjanje, regularizacijo ter prilagajanje modelu. Metoda naključnih gozdov se je na koncu izkazala kot najboljša metoda oz. boljša kot linearni model. Tudi v članku [21] so za metodo naključnih gozdov ugotovili, da je uspešna in zelo odporna na šume. Za testiranje so uporabljali k-kratno prečno preverjanje [22]. Za optimiziranje napovednih algoritmov priporočajo regularizacijo in zmanjšanje dimenzije s PCA; oba načina sta se dobro izkazala pri zmanjšanju prevelikega prilagajanja (ang. overfit) in povečanju natančnosti.

Za masovno ocenjevanje stanovanjskih nepremičnin so kot potencialno tehniko napovedovanja uporabili metodo naključnih gozdov [23]. V empiričnih študijah z uporabo podatkov o stanovanjih, je bila omenjena metoda boljša od tehnik kot so CHAID, CART, kNN, večregresijska analiza, nevronske mreže, spodbujevalna drevesa. Nabor podatkov so sestavili tako, da so vsak objekt kategorizirali po naboru spremenljivk. Napovedne metode so primerjali med sabo tudi v [24]. Primerjali so nevronske mreže, metodo naključnih gozdov, podporne vektorje (SVM). Rezultati analize so pokazali, da je metoda naključnih gozdov boljša od drugih modelov pri napovedovanju cen stanovanj. Avtorji članka sklepajo, da lahko tehnike strojnega učenja zagotovijo uporaben nabor orodij za pridobivanje informacij o stanovanjskih trgih.

Več testov je pokazalo, da je metoda naključnih gozdov postala priljubljena tehnika za klasifikacijo, napovedovanje, preučevanje pomena spremen-

ljivk, izbiro spremenljivk in zunanje odkrivanje. Obstajajo številni primeri uporabe naključnih gozdov na različnih področjih. Eksperimentalno so v članku [25] preučili skladnost in splošnost metode. Izkazalo se je, da je 'boosting' metoda za prikaz cepljenja vozlišč najboljša tehnika. Boosting je strojni učni ansambelski meta-algoritem, ki v prvi vrsti zmanjšuje pristranskost, pa tudi variance v nadzorovanem učenju in družino algoritmov strojnega učenja, ki pretvarjajo šibke učne primerke v močne.

Razvoj uspešne metode podatkovnega rudarjenja zahteva veliko količino preučevanja in iskanja najboljših scenarijev izbora atributov, uteži ter tehnik rudarjenja nad podatki. Zato je potrebno upoštevati priporočila podatkovnega rudarjenja [26]. V prvi vrsti je za učenje potrebno podatke predstaviti, oceniti in jih optimizirati. Pri gradnji regresijskega modela so podatke za prilagajanja linije prikazali tudi s pomočjo metode LOESS [19]. S pomočjo te metode lahko razkrijemo trende in cikle v podatkih, ki jih je težko modelirati s parametrično krivuljo. Potrebno je izbrati tudi zadostno število parametrov učenja v primerjavi s številom podatkov in paziti na zasičenje. Teoretično razumevanje je ključno za oblikovanje algoritma. Priporočajo tudi uporabo večih modelov, ne le enega.

Obravnavani članki priporočajo napovedovanje z različnimi metodami, mi pa se bomo prednostno posvetili linearni regresiji in metodi naključnih gozdov. Na podlagi omejenih metod bomo s podatkovnim rudarjenjem razvili model za napovedovanje vrednosti nepremičnin. Prednosti posameznih metod bomo aplicirali na področje Slovenije ter podatke ETN in druge (podatki iz SI-STAT, razdalje občin do Slovenskih mest, cene naftnih derivatov, obresti idr.). Poleg upoštevanja drugih podatkov, bomo razmislili tudi o ločenem modeliranju vrednosti lokacije in stavbe.

Poglavje 3

Priprava in analiza podatkov

V poglavju najprej predstavimo postopek pridobivanja podatkov iz različnih virov ter jih naštejemo. Predstavimo tudi nekaj osnovnih statističnih podatkov o podatkih. V nadaljevanju smo predstavili tudi korake in postopke, s katerimi smo pridobili podatke ter način shranjevanja le-teh. V zaključku poglavja razložimo postopek združevanja podatkov iz različnih virov v eno (oz. dve) zbirko podatkov ter končno statistiko o novi zbirki. Na koncu smo predstavili še pomen podatkov v pripravljeni zbirki.

3.1 Pridobivanje in zajem podatkov

3.1.1 Podatki GURS - ETN in REN

Idejo za magistrsko nalogo o napovedovanju vrednosti nepremičnin smo črpali na podlagi poznavanja podatkov iz preteklega dela in izkušenj s podatki, še posebej iz mojega diplomskega dela. To je tudi razlog, da bo naš izhodiščni (morda glavni) vir podatkov zbirka *Evidenca trga nepremičnin* (ETN), ki jo pripravlja *Geodetska uprava Republike Slovenije* (GURS). Zbirka ETN vsebuje podatke o poslih nepremičnin, ki so bili izvedeni na področju Republike Slovenije. GURS zbira sicer več vrst podatkov, ena izmed njih je tudi evidenca o nepremičninah - zbirka *Register nepremičnin* (REN). Zbirka REN vsebuje podatke o vseh nepremičninah v Republiki Sloveniji. Več o zbir-

kah ETN in REN smo opisali v delu 3.3. V praksi vrednost nepremičnine[1] ocenjujejo ustrezno usposobljeni cenilci [27] (sodni cenilci, pooblaščeni ocenjevalci in cenilci za potrebe bank, zavarovalnic in posredovanja pri prometu z nepremičninami). Cenilci pogosto zajamejo podatke o podobnih poslih nepremičnin (ETN) in podatke o nepremičninah (REN) ter na podlagi primerjav parametrov ocenjevane nepremičnine z ostalimi podatki določijo vrednost nepremičnine. Zato smo se odločili, da bomo kot osnovo za našo zbirko podatkov izbrali zbirki ETN in REN.

Tako smo se najprej posvetili pridobitvi podatkov iz zbirk ETN in REN, ki so bili shranjeni v več tabelah podatkovne baze Microsoft (ang. Microsoft SQL Server) (MSSQL):

- `posliKPP` - podatki o kupoprodajnih poslih (ETN),
- `delistavbKPP` - podatki o sestavnih delih kupoprodajnih poslov - deli stavb (ETN),
- `zemljiscaKPP` - podatki o sestavnih delih kupoprodajnih poslov - zemljišča (ETN),
- `posliNP` - podatki o najemnih poslih (ETN),
- `delistavbNP` - podatki o sestavnih delih najemnih poslov - deli stavb (ETN),
- `sifranti` - šifranti (ETN),
- `coordinates2` - koordinate lokacij,

[1]Ocenjevanje vrednosti nepremičnin [27] je podajanje mnenja o vrednosti nepremičnine na podlagi strokovnega tehtanja, ki temelji na objektivnosti, pravilni presoji, znanju, podatkih in izkušnjah. Neformalno ocenjevanje je ocenjevanje na podlagi intuicije. Takšno ocenjevanje uporablja velik del udeležencev na trgu nepremičnin in daje neko splošno oceno. Formalne cenitve, ki jih izvajajo strokovnjaki – cenilci, pa temeljijo na rezultatih metodičnega zbiranja in analiziranja tržnih podatkov. Pri svojem delu uporabljajo različne standarde, na koncu pa izdajo pisno cenitev, ki vsebuje opis celotnega postopka cenitve.

- `renStavbe` - podatki o stavbah (REN),
- `renDeliStavb` - podatki o delih stavb (REN),
- `renZemljisca` - podatki o zemljiščih (REN).

Za delo s podatki v programskem jeziku *Python* smo podatke iz podatkovne baze *MSSQL* shranili v tekstovni obliki kot zapis *csv*. Vsaka *csv* datoteka je tako vsebovala podatke le iz ene tabele. Za kupoprodajne posle smo imeli na voljo podatke od leta 2007, za najemne posle pa od leta 2013 naprej. Tabela 3.1 prikazuje število zapisov po posamezni tabeli.

Tabela 3.1: Število zapisov po posamezni tabeli.

ime tabele	število zapisov
`posliKPP`	333.133
`deliStavbKPP`	350.891
`zemljiscaKPP`	530.522
`posliNP`	185.704
`deliStavbNP`	216.922
`sifranti`	341
`coordinates2`	541.445
`renStavbe`	27.377
`renDeliStavb`	45.783
`renZemljisca`	69.731

3.1.2 Podatki SURS

Ob pregledu sorodnih del smo v nekaterih prispevkih odkrili tudi nabor atributov, ki so jih uporabili pri napovedovanju [15]. To je bilo izhodišče za črpanje idej za razširitev našega nabora podatkov. Idejo za dodatne atribute smo iskali tudi v člankih, ki opisujejo kvaliteto življenja v različnih krajih po Sloveniji [8, 9]. Podatke, ki so jih navajali v člankih, smo za območje Slovenije našli na portalu *Statističnega urada Republike Slovenije* (SURS).

Podatkovna baza SI-STAT [5] omogoča dostop do statističnih podatkov iz različnih virov na enem mestu. Podatki, ki smo jih pridobili na podlagi članka [15]:

- obsojeni polnoletni in mladoletni po občinah stalnega prebivališča,
- število otrok na vrtec in število otrok na vzgojitelja in pomočnika vzgojitelja, občine,
- osnovne skupine prebivalstva po spolu,
- povprečne mesečne plače po občinah,
- dovoljenja za gradnjo stavb - število stavb, njihova gradbena velikost in stanovanja v njih, glede na vrsto stavbe, po občinah,
- gradbena dovoljenja - izbrani kazalniki, po občinah,
- NAMEA emisije v zrak (SKD 2008[2]),
- nastali, zbrani in odloženi komunalni odpadki, občine,
- dokončana stanovanja po številu sob in površini, po občinah,
- ocena dokončanih stanovanj - izbrani kazalniki, po občinah,
- ocena stanovanjskega sklada, stanovanja po letu zgraditve po občinah Slovenije.

Podatki, ki smo jih pridobili na podlagi člankov [8, 9]:

- cestna vozila konec leta (31. 12.) glede na vrsto vozila in starost,
- delovno aktivno prebivalstvo (brez kmetov) po občinah prebivališča in občinah delovnega mesta,
- delovno aktivno prebivalstvo (brez kmetov), medobčinski delovni migranti ter indeks delovne migracije,

[2]Standardna klasifikacija dejavnosti - SKD 2008.

- delovno aktivno prebivalstvo po občinah delovnega mesta,
- delovno aktivno prebivalstvo po občinah prebivališča,
- indeksi cen in število transakcij stanovanjskih nepremičnin po vrstah stanovanjskih nepremičnin, četrtletno,
- podjetja po občinah,
- povprečne mesečne plače po dejavnostih (SKD 2008), občine,
- prebivalstvo,
- delovno aktivno prebivalstvo po občinah delovnega mesta, doseženi izobrazbi,
- delovno aktivno prebivalstvo po občinah prebivališča, doseženi izobrazbi,
- gostota naseljenosti,
- osnovni podatki o razvezah zakonskih zvez (absolutni podatki in kazalniki),
- osnovni podatki o sklenitvah zakonskih zvez (absolutni podatki in kazalniki),
- prebivalstvo po izbranih starostnih skupinah,
- prebivalstvo, staro 15 ali več let, po izobrazbi,
- dovoljenja za gradnjo stavb - število stavb, njihova gradbena velikost in stanovanja v njih, glede na vrsto stavbe,
- naseljena stanovanja, prebivalci in gospodinjstva po uporabni površini,
- ocena dokončanih stanovanj - izbrani kazalniki,
- ocena dokončanih stanovanj po številu sob in površini,

- otroci, vključeni v vrtec, po občini stalnega prebivališča,
- stanovanjski standard,
- vrtci po številu otrok in občini zavoda,
- število naseljenih stanovanj po uporabni površini in številu prebivalcev,
- umrli,
- obsojeni polnoletni in mladoletni po občinah stalnega prebivališča.

3.1.3 Drugi podatki

Nekatere podatke smo pridobili tudi po lastni presoji:

- cene naftnih derivatov,
- indeksi cen življenjskih potrebščin,
- obrestne mere Evropske centralne banke,
- razdalje med občino in slovenskimi mesti.

Zgoraj naštete podatke smo pridobili iz različnih virov; s portala *Ministrstva za gospodarski razvoj in tehnologijo Republike Slovenije* [28] smo s skripto 3.1 v programskem jeziku *JavaScript* pridobili podatke za *cene naftnih derivatov*.

Koda 3.1: Skripta v programskem jeziku JavaScript za zajem podatkov o cenah naftnih derivatov.

```
var datum = "";
var cenaBencin = 0;
var cenaNafta = 0;
var td = "";

$("table.contenttable tr").each(function () {
  var tr = $(this).find("td");

  td = $(tr[0]);
  datum = td.find("span").text();
```

```
  td = $(tr[1]);
  cenaBencin = td.find("span").text();
  td = $(tr[2]);
  cenaNafta = td.find("span").text();

  console.log(datum + ";" + cenaBencin + ";" + cenaNafta);
```

Podatke za *obrestne mere Evropske centralne banke* smo pridobili s portala *Banke Slovenije* (BSI) [29]. Podobno kot zgoraj, smo tudi za pridobitev teh podatkov uporabili skripto 3.2 v jeziku *JavaScript*:

Koda 3.2: Skripta v programskem jeziku JavaScript za zajem podatkov o obrestnih merah.

```
var rows = "";
var row_csv = "";

$("table tr").each(function () {
  if ($(this).has("th").length > 0) {
      // header
      rows = $(this).find("th");
    }
    else {
      // ni headeer
      rows = $(this).find("td");
    }

    row_csv = "";
    for (var i = 0; i < rows.length; i++) {
      row_csv += rows[i].innerHTML + ";";
    }

    row_csv = row_csv.slice(0, -1);
    console.log(row_csv);
});
```

Za vsako nepremičnino imamo na voljo tudi podatek o lokaciji (katastrska občina, občina, naselje, ulica). Po našem mnenju je lokacija dovolj natančno določena že s parametri: katastrska občina, občina in naselje, zato smo podatek o ulici izpustili. V poglavju 2 smo zaznali, da na vrednost nepremičnine poleg lokacije vpliva tudi oddaljenost do večjih krajev, kjer ljudje običajno najdejo zaposlitev ali pa tam opravijo nakupe in druga opravila (zato smo se odločili, da dodamo tudi te podatke). Za pridobitev omenjenih podatkov

smo razvili skripto v programskem jeziku *JavaScript*, ki za izračun razdalj uporablja *Maps JavaScript API* preko storitve *Distance Matrix Service* [30]. Najprej smo na oblačni upraviteljski konzoli Google ustvarili novo storitev in pridobili ključ *Distance Matrix API* za avtentikacijo klica servisa. S klicem, prikazanem na kodi 3.3, smo pridobili razdaljo med občino in mestom. Za iskanje najkrajše razdalje in najhitrejše poti med občino in mestom, smo dopustili tudi možnost vožnje z avtom po avtocesti ter vožnje po cestah, za katere je potrebno plačilo cestnin. Primer odgovora storitve za najkrajšo razdaljo med krajema je na sliki 3.1.

Koda 3.3: Skripta v programskem jeziku JavaScript za pridobitev razdalje in potovalnega časa med občino in mestom.

```
var service = new google.maps.DistanceMatrixService();
service.getDistanceMatrix(
  {
    origins: ["Cerkno, Slovenia"],
    destinations: ["Ljubljana, Slovenia"],
    travelMode: google.maps.TravelMode.DRIVING,
    avoidHighways: false,
    avoidTolls: false
  },
  callback
);
```

Pri pridobivanju podatkov o razdaljah med kraji smo upoštevali tudi omejitve uporabe storitev Google na največ 2.500 zahtevkov na dan. Zato smo dodatno razvili mehanizem, ki omogoča sprotno shranjevanje rezultatov. Na ta način smo lahko izvedli postopek pridobivanja razdalj v več delih. Rezultate smo shranjevali v format zapisa *csv*, podobno kot smo to počeli za podatke ETN in REN.

3.2 Nova zbirka podatkov

Po končanem zbiranju različnih podatkov je sledilo še združevanje le-teh. Podatke smo najprej prebrali iz datotek *csv* (50 različnih) ter razvili mehanizem za združevanje podatkov iz različnih datotek. Naš cilj zbiranja podatkov je

Slika 3.1: Odgovor storitve 'Distance Matrix Service' za razdaljo med Cerknim in Ljubljano.

bil narediti novo zbirko podatkov s širokim naborom atributov, ki smo jo v nadaljevanju dela še izboljšali; odstranili osamelce, poiskali manjkajoče vrednosti ter iskali najpomembnejše atribute za napovedovanje najemnin oz. cen nepremičnin. Več o izboljšavah zbirke podatkov je zapisano v poglavju 4.

3.2.1 Združevanje podatkov v eno zbirko podatkov

Po branju podatkov smo pričeli z medsebojno povezavo podatkov. Cilj je bil združiti podatke iz različnih virov in datotek v eno zbirko podatkov, nad katero smo nato izvajali analize in na podlagi teh analiz z različnimi pristopi izboljšali zbirko podatkov, ki bo naše izhodišče za napovedovanje vrednosti nepremičnin z metodami podatkovnega rudarjenja.

Najprej smo vzpostavili relacije med podatki v zbirki ETN. Vsem atributom, ki predstavljajo šifrante v podatkih o nepremičninah in poslih smo nastavili vrednosti in tako razširili atribut za šifrant npr. atributu *Sifrant_VrstaKupProdPravPosla* instance razreda `PosliKPP` smo nastavili vrednost ustreznega objekta tipa `Sifranti`. Vsakemu poslu smo nastavili seznam ključev nepremičnin, vsaki nepremičnini pa smo nastavili tudi podatke o koordinatah ter inicializirali atribut `DodatniPodatki`, ki predstavlja nabor podatkov SURS in ostali, ki smo jih našteli zgoraj. Vsak objekt tipa `Dodatni Podatki` vsebuje nabor ključev, ki predstavlja po en dodaten podatek. Tako vzpostavljene relacije med podatki so nam predstavljale pomemben korak do našega cilja, t.j. združitev velikega nabora podatkov v eno zbirko podatkov, ki jo je mogoče zapisati v eno datoteko. Na začetku tega koraka smo podatke deloma analizirali, saj lahko vsak posel vsebuje več različnih tipov nepremičnin, kot je opisano v poglavju 3.3. Poseben poudarek na analizi smo dali za:

- porazdelitev števila kupoprodajnih poslov po vrsti kupoprodajnega pravnega posla,
- porazdelitev števila najemnih poslov po vrsti najemnega pravnega po-

sla,

- porazdelitev števila delov stavb na najemni posel,
- porazdelitev števila delov stavb na kupoprodajni posel,
- porazdelitev števila nepremičnin po vrsti oddane površine.

Več o zgoraj naštetih analizah smo zapisali v poglavju 4.1. Sklenili smo, da bomo najprej izvedli celoten postopek (priprava zbirke podatkov, čiščenje zbirke, napovedovanje) za najemne posle tipa 'oddajanje na prostem trgu' (tip 1), ki vsebujejo po eno nepremičnino - del stavbe tipa 'stanovanje' (tip 2). Razvili smo funkcijo `makeLongCSV_NP`, katere namen je iz nabora podatkov ustvariti zbirko podatkov, ki vsebuje glavo z imeni atributov in zapise z vsemi vrednostmi po atributih. Najprej smo sestavili nabor tistih najemnih poslov, ki so tipa 1, vsebujejo le eno nepremičnino ter nabor tistih nepremičnin, ki so tipa 2 - torej zbirko najemnih poslov, ki je bil sklenjen le za najem enega stanovanja. Več o tej odločitvi smo že pisali v poglavju 4.1. V naslednji fazi smo razvili funkcionalnost, ki iz podatkov najprej sestavi seznam imen vseh atributov, ki so na voljo. Nato smo s podobnim postopkom prehoda čez vse posle sestavili še seznam vseh vrednosti, ki predstavljajo en najemni posel na prostem trgu za oddajo enega stanovanja. Vsak tak seznam predstavlja eno vrstico v *csv* oz. en zapis (ang. observation) v zbirki podatkov oblike *csv*. Naša zbirka podatkov *najemni posli stanovanj* (glej tabelo 3.2) je na koncu postopka združevanja podatkov vsebovala 50.359 zapisov in 922 atributov.

Ko smo ustvarili zbirko podatkov za najemne posle stanovanj, smo se lotili še kupoprodajnih poslov. Na podlagi analiz porazdelitve podatkov za kupoprodajne posle smo se odločili, da naredimo novo zbirko podatkov za kupoprodajne posle tipa 'prodaja na prostem trgu', ki vsebuje le en del stavbe - stanovanje, posel pa lahko vsebuje tudi parkirni prostor ali garažo ali zemljišča. Tudi za kupoprodajne posle smo razvili funkcijo `makeLongCSV_KPP`, katere namen je iz nabora podatkov ustvariti zbirko podatkov, ki vsebuje glavo z imeni atributov in zapise z vsemi vrednostmi po atributih. Najprej smo sestavili nabor tistih kupoprodajnih poslov, ki vsebujejo eno stanovanje,

Tabela 3.2: Porazdelitev podatkov v zbirki *najemni posli stanovanj*.

atributi	število atributov
atributi o poslih iz ETN	17
atributi o delih stavb iz ETN	30
atributi iz REN	74
atributi za šifrante	84
atributi za koordinate	7
atributi iz SURS	526
atributi iz drugi (Google in ostali)	165
ostali atributi	19

lahko pa tudi parkirni prostor ali garažo ali zemljišča. V naslednji fazi smo že razvito funkcionalnost za najemne posle prilagodili še za kupoprodajne posle, in sicer tako, da smo iz podatkov sestavili seznam imen vseh atributov in za vse izbrane posle še seznam vrednosti, ki predstavljajo en kupoprodajni posel za stanovanje. Ta lahko vsebuje garažo ali parkirni prostor ali zemljišča (skupna površina zemljišč). Vsak tak seznam predstavlja eno vrstico v *csv* oz. en zapis (ang. observation) v zbirki podatkov oblike *csv*. Naša zbirka podatkov *kupoprodajni posli stanovanj* (glej tabelo 3.3) je na koncu postopka združevanja podatkov vsebovala 79.841 zapisov in 914 atributov.

Tabela 3.3: Porazdelitev podatkov v zbirki *kupoprodajni posli stanovanj*.

atributi	število atributov
atributi o poslih iz ETN	19
atributi o delih stavb iz ETN	35
atributi iz REN	74
atributi za šifrante	70
atributi za koordinate	7
atributi iz SURS	526
atributi iz drugi (Google in ostali)	165
ostali atributi	18

3.3 Razlaga (posameznih) podatkov

3.3.1 Podatki ETN

Evidenca trga nepremičnin (ETN) je javna zbirka podatkov o sklenjenih kupoprodajnih in najemnih pravnih poslih z nepremičninami [31]. Podatke o sklenjenih pravnih poslih v ETN mesečno posredujejo: davčna uprava, notarji, nepremičninske družbe ter upravne enote in občine. V ETN se evidentirajo dosežene pogodbene cene in najemnine na slovenskem trgu nepremičnin. Evidenca trga nepremičnin vsakomur omogoča vpogled v tržne cene pri odločanju za nakup ali prodajo nepremičnine. ETN omogoča neposredno spremljanje in primerjavo doseženih kupoprodajnih cen po različnih vrstah nepremičnin, časovnih obdobjih in območjih. Osnovni namen evidence je sistematično spremljanje in analiziranje tržnih cen in najemnin nepremičnin, za potrebe množičnega vrednotenja nepremičnin in periodičnih poročil za zagotavljanje javne preglednosti slovenskega nepremičninskega trga. Izhodiščna zbirka podatkov ETN vsebuje podatke o poslih nepremičnin, ki so lahko kupoprodaje ali najemi. Kupoprodaje sestavljajo ali deli stavb ali zemljišča ali pa deli stavb in zemljišča. En kupoprodajni posel lahko vsebuje nič ali več zemljišč, nič ali več delov stavb, vsebuje pa vsaj eno nepremičnino. Primer kupoprodajnega posla je prodaja hiše z ločeno garažo ter več zemljišči na katerih stojita oba dela stavbe. Najemne posle sestavljajo le deli stavb, kjer en najemni posel lahko vsebuje en ali več delov stavb. Primer najemnega posla je najem stanovanja. Zbirka ETN vsebuje različne vrste poslov za kupoprodaje in za najeme posebej. Tudi za vsak tip sestavnih delov poslov (nepremičnine; deli stavb in zemljišča) zbirka vsebuje različne vrste posameznih nepremičnin.

3.3.2 Podatki REN

Register nepremičnin (REN) [32] je evidenca, ki vsebuje podatke o vseh nepremičninah v Sloveniji. V Registru nepremičnin so zbrani podatki o:

- zemljiščih, evidentiranih v zemljiškem katastru,
- stavbah in delih stavb, evidentiranih v katastru stavb,
- lastnikih,
- vseh ostalih nepremičninah, ki še niso evidentirane v zemljiškem katastru in katastru stavb.

Za posamezne nepremičnine so poleg podatkov zemljiškega katastra in katastra stavb v REN zbrani še drugi podatki. Podatki, ki se vodijo v Registru nepremičnin, so prevzeti iz obstoječih javnih evidenc (zemljiškega katastra, katastra stavb, centralnega registra prebivalstva ...) in dopolnjeni s podatki popisa nepremičnin. Vzdrževanje podatkov se izvaja na osnovi prevzema sprememb iz javnih evidenc, s terenskimi ogledi in meritvami, z uporabo aeroposnetkov in drugih metod inventarizacije, pa tudi na osnovi podatkov, ki jih posredujejo lastniki, uporabniki nepremičnin.

3.3.3 Ostali podatki

Ostale podatke (cene naftnih derivatov, indeksi cen življenjskih potrebščin, obrestne mere ECB, razdalje med občino in slovenskimi mesti) smo v večini pridobili iz portala SI-STAT - podatki SURS in jih predstavili v poglavju 3.1.3. Nekateri podatki so vezani na čas (leto, polletje, kvartal, mesec, šolsko leto) ali občino ali oboje. Podatke SURS smo pridobili s portala prek iskalnih obrazcev in jih shranili v *csv* dokumente. Vrednosti za podatke SURS so številčne.

Poglavje 4

Metodologija

Glavni cilj naloge je bil razviti napovedni model vrednotenja nepremičnin s podatkovnim rudarjenjem, ki napoveduje vrednost nepremičnine na podlagi podatkov iz različnih virov: podatki iz GURS - ETN in REN, podatki iz SI-STAT in drugi. Napovedni model predstavlja alternativo trenutnemu vrednotenju na GURS, ki temelji na kompleksnih modelih vrednotenja.

V prvem delu smo se posvetili zbiranju (glej poglavje 3) in obdelavi podatkov. Podatke ETN in REN smo pridobili iz podatkovne baze *MSSQL* s poizvedovalnim jezikom *SQL*, na različnih portalih pa smo poiskali nekatere smiselne podatke, npr. portal SI-STAT. Poiskali smo tudi druge podatke, ki bi lahko vplivali na vrednost nepremičnin, npr. cena naftnih derivatov, obresti, razdalje do mest idr. Vse podatke smo shranili v obliki zapisa *csv*.

Nato smo s programskim jezikom *Python* podatke iz različnih virov združili in shranili v ustrezno obliko (glej poglavje 3.2). Podatke smo analizirali in vizualizirali ter deloma tudi prečistili. Drugi del čiščenja podatkov smo izvedli v programskem jeziku *R*, kjer smo izvedli postopek odkrivanja vrednosti, ki izstopajo od ostalih - odkrivanje osamelcev (ang. outliers detection), izvedli postopek vstavljanja vrednosti (ang. imputing values) ter odkrivanja pomembnih atributov (ang. feature selection) za napovedovanje vrednosti.

V tretjem delu smo z metodami podatkovnega rudarjenja nad podatki, ki smo jih predhodno pripravili, izdelali modele napovedovanja vrednosti

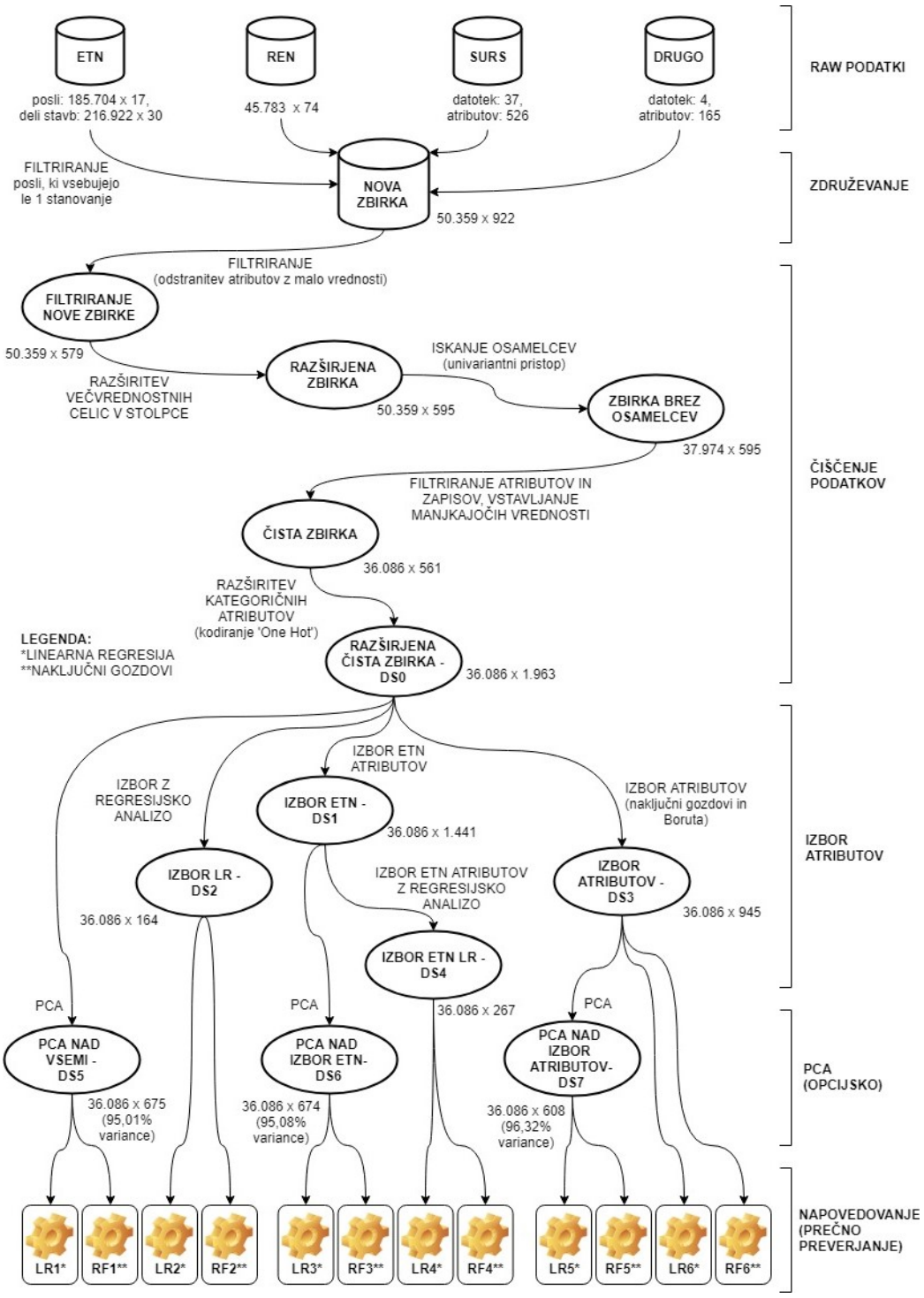
nepremičnin. Posvetili smo se dvema metodama napovedovanja, linearni regresiji (glej poglavje 5.1) in naključnim gozdovom (glej poglavje 5.2). Napovedne modele smo razvili v programskem jeziku *Python* s pomočjo znanih knjižnic kot so `NumPy` [33], `scikit-learn` [34],...

Vsi koraki razvoja so med sabo povezani. Med posameznimi koraki smo tudi prehajali naprej in nazaj. Četrti korak razvoja pa je bil še posebej tesno povezan s tretjim korakom. Tu smo izvedli testiranje, t.j. ocenjevanje/vrednotenje napovednega modela. Za ugotavljanje učinkovitosti napovednih modelov smo implementirali metodo za evalviranje napovednih vrednosti, ki uporablja pristop prečnega preverjanja (ang. cross validation). Pri vrednotenju napovednega modela smo vrednotili efektivno vrednost napake (RMSE), povprečno absolutno napako (MAE), delež razložene variance (R^2), kot so to ovrednotili tudi ostali [18, 23, 24] ter prilagojen delež razložene variance (ang. adjusted R^2). Na podlagi sprotnega računanja napak in učinkovitosti posamezne metode, smo se sproti odločali o naslednjih korakih razvoja. Na koncu smo rezultate testiranja združili in naredili povzetek testiranja napovedovanja. Celoten potek razvite metodologije prikazujeta slika 4.1 za najeme stanovanj in slika 4.2 za kupoprodaje stanovanj.

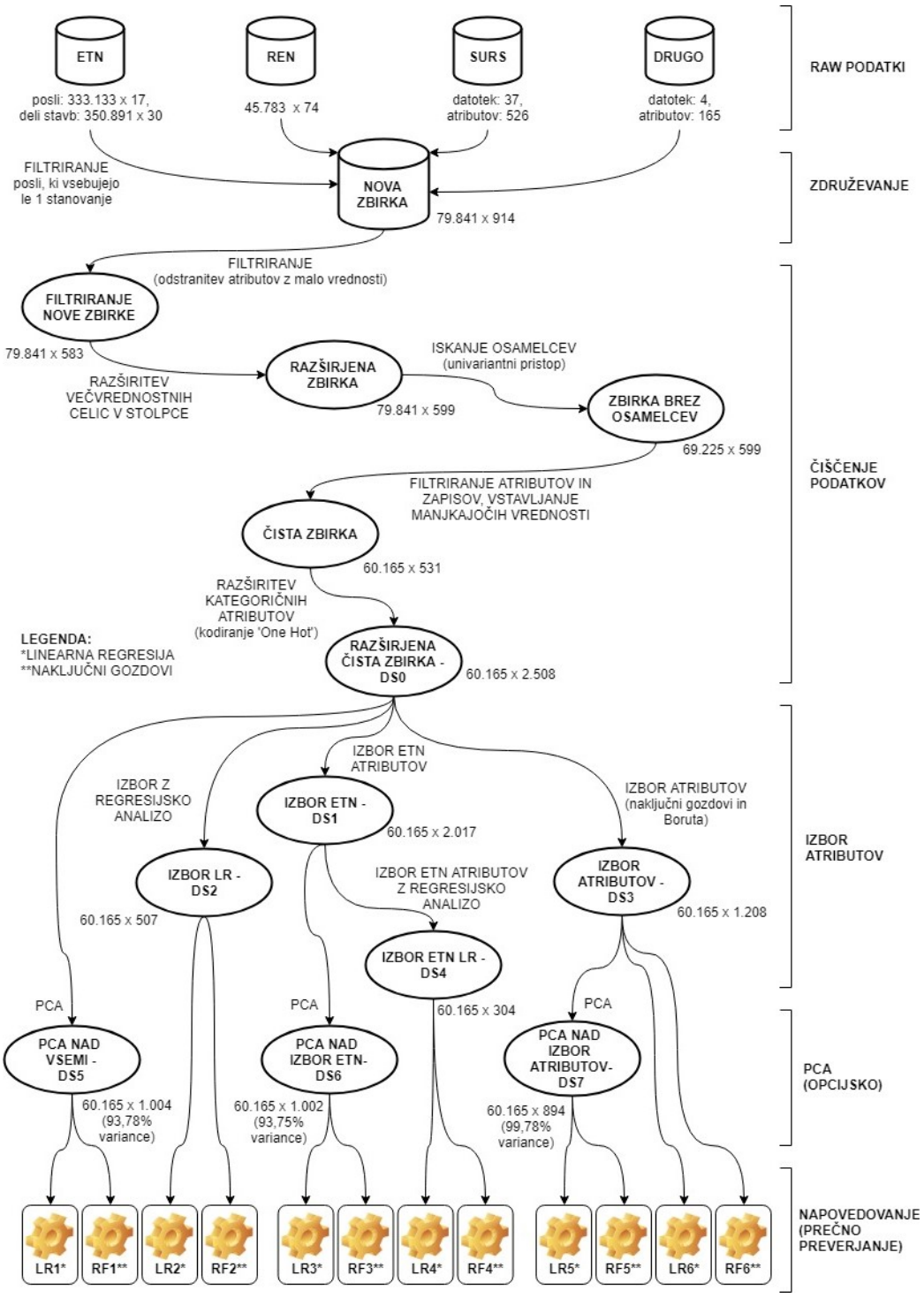
4.1 Analiza podatkov

V tem poglavju predstavimo postopek analize podatkov, s katerim smo podrobneje spoznali podatke, ki smo jih pridobili. Opisali smo tudi pristope za iskanje ekstremnih vrednosti in reševanje problema manjkajočih vrednosti v podatkih. Na koncu poglavja opišemo še uporabo metod za izbor ključnih atributov na primeru naše zbirke podatkov.

Po zaključenem razvoju postopka branja podatkov iz različnih virov smo izvedli analizo nad podatki. Vsak posel vsebuje tudi podatek o vrsti pravnega posla. Cilj tega dela naloge je bil najti in razumeti podatke, ki so potrebni za naše delo. Vsaka nepremičnina je predstavljena z več parametri, npr. površina in lokacija. V kolikor se odločimo za najem ali nakup stanovanja,



Slika 4.1: Diagram poteka za najeme stanovanj.



Slika 4.2: Diagram poteka za kupoprodaje stanovanj.

sta omenjena parametra pomemben faktor pri višini pogodbene najemnine ali cene stanovanja. V kolikor je stanovanje v Ljubljani in je prostorno, lahko pričakujemo, da bo višina pogodbene najemnine ali cena stanovanja visoka. Poleg površine in lokacije, na višino pogodbene najemnine ali cene vplivajo tudi drugi dejavniki. V našem delu bomo subjektivne dejavnike, kot so prijaznost sosedov ali prodajalke v bližnjem kiosku, izpustili. Osredotočili se bomo na podatke, ki so objektivni.

Naša zbirka podatkov vsebuje največ podatkov za *prodaje na prostem trgu* (glej tabelo 4.1).

Tabela 4.1: Porazdelitev števila kupoprodajnih poslov po 'vrsti kupoprodajnega pravnega posla'.

vrsta kupoprodajnega pravnega posla	število podatkov	delež podatkov [%]
prodaja na prostem trgu	302.366	92,65
prodaja na javni dražbi (prostovoljna)	4.768	1,46
prodaja na javni dražbi v izvršilnem postopku ali stečaju	7.573	2,32
prodaja med družinskimi člani ali povezanimi fizičnimi in pravnimi osebami	7.551	2,31
finančni najem (lizing)	4.086	1,25

Podobno analizo smo izvedli tudi nad podatki za najemne posle. Za najemne posle ugotovimo, da naša zbirka podatkov vsebuje največ podatkov za vrsto najema *oddajanje na prostem trgu* (glej tabelo 4.2).

Analizirali smo tudi porazdelitev nepremičnin po posameznem poslu in ugotovili, da največ najemnih poslov (91,24 %) vsebuje le en del stavbe.

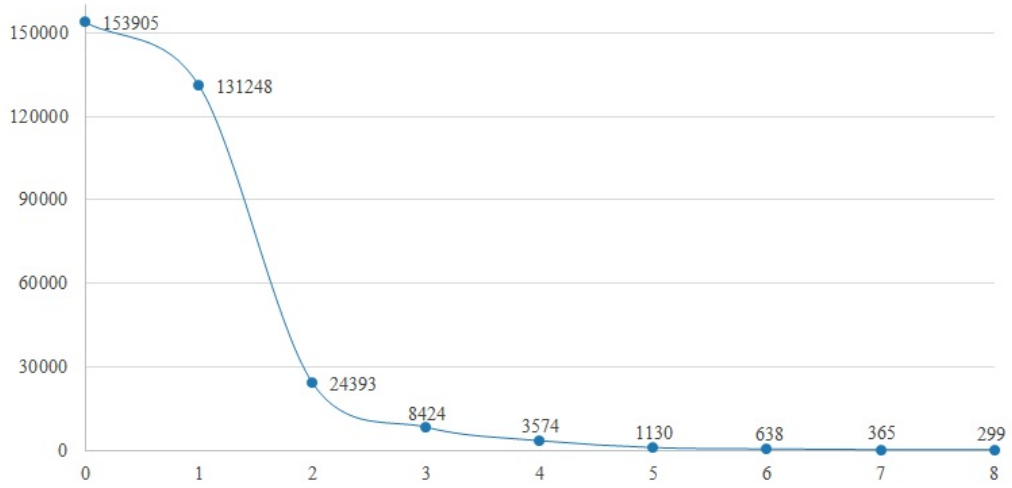
Za kupoprodajne posle je potrebno upoštevati, da je lahko posel sestavljen tudi iz delov stavb in zemljišč. Iz analize porazdelitve števila delov stavb na

Tabela 4.2: Porazdelitev števila najemnih poslov po 'vrsti najemnega pravnega posla'.

vrsta najemnega pravnega posla	število podatkov	delež podatkov [%]
oddajanje na prostem trgu	130.179	71,65
oddajanje družinskim članom ali povezanim fizičnimi in pravnim osebam	9.639	5,31
neprofitno oddajanje denacionaliziranih stanovanjskih nepremičnin na podlagi upravne ali sodne odločbe	628	0,35
drugo odplačno oddajanje	19.729	10,86
neprofitno oddajanje stanovanjskih nepremičnin v lasti države in občin za najemnino, določeno na podlagi zakona	21.520	11,84

posel (glej sliko 4.3) smo ugotovili, da največ poslov vsebuje en (46,20 %) ali dva (39,40 %) dela stavbe. Veliko je npr. takšnih poslov, ki se sklenejo za nakup stanovanja s parkiriščem, kar je posel z dvema deloma stavbe. Na tej točki smo se odločili, da se bomo najprej posvetili najemnim poslom, saj (najemni posli) lahko vsebujejo le dele stavb, medtem ko lahko kupoprodajni posli poleg delov stavb vsebujejo tudi zemljišča ali le zemljišča. Sklenemo tudi, da bomo za najemne posle uporabili le tiste posle, ki vsebujejo le en del stavbe. Le-ti predstavljajo pretežni del podatkov za najemne posle. O podatkih ETN smo se posvetovali tudi s predstavniki GURSa. Izvedeli smo, da so podatki o najemnih poslih precej nezaupljivi in vsebujejo veliko napak, saj je bila ob vnosu podatkov v zbirko kontrola podatkov slaba. V začetku so podatke vnašali iz obrazcev, sedaj pa podatke vnaša vsak sam - fizična oseba

odda fizični osebi, vnašajo lahko tudi poslovni subjekti. Kakovost vnesenih podatkov je praviloma odvisna od osebe, ki podatke sporoča. Na GURSu poudarjajo, da so podatki o najemnih poslih zelo vprašljivi, še posebej vrednost najemnine. Pred vnosom podatkov v zbirko, vsak kupoprodajni posel na GURSu ročno pregledajo oz. prečistijo na način, da pred vnosom podatke preveri za to odgovorna oseba. Pri najemnih poslih pa se preverjanje podatkov ne izvaja na tak način. Preverjanje izvedejo le za pisarne in lokale.



Slika 4.3: Porazdelitev števila delov stavb za kupoprodajne posle.

Pregled števila zapisov za posamezno vrsto podatkov je za kupoprodajne posle na voljo v tabeli 4.3 in za najemne posle v tabeli 4.4.

Iz grafa 4.4 o porazdelitvi števila dela stavb za najemne posle po vrsti oddane površine opazimo, da je največ takšnih najemnih poslov, ki so bili sklenjeni za najem stanovanja. Na podlagi vseh ugotovitev o podatkih zaključimo, da uporabimo tiste najemne posle, ki so bili sklenjeni za le en del stavbe tipa *stanovanje* in pričnemo s pripravo nove zbirke podatkov za najemne posle (glej poglavje 3.2.1).

Tabela 4.3: Porazdelitev podatkov za kupoprodajne posle.

podatek	število podatkov	delež podatkov [%]
posli	326.344	21,76
deli stavb	339.347	22,62
zemljišča	507.943	33,86
posli, ki vsebujejo le del stavbe	106.005	7,07
posli, ki vsebujejo le zemljišče	153.905	10,26
posli, ki vsebujejo del stavbe in zemljišče	66.434	4,43

Tabela 4.4: Porazdelitev podatkov za najemne posle.

podatek	število podatkov	delež podatkov [%]
posli	181.749	32,49
deli stavb	208.225	37,22
posli, ki vsebujejo le en del stavbe	169.439	30,29

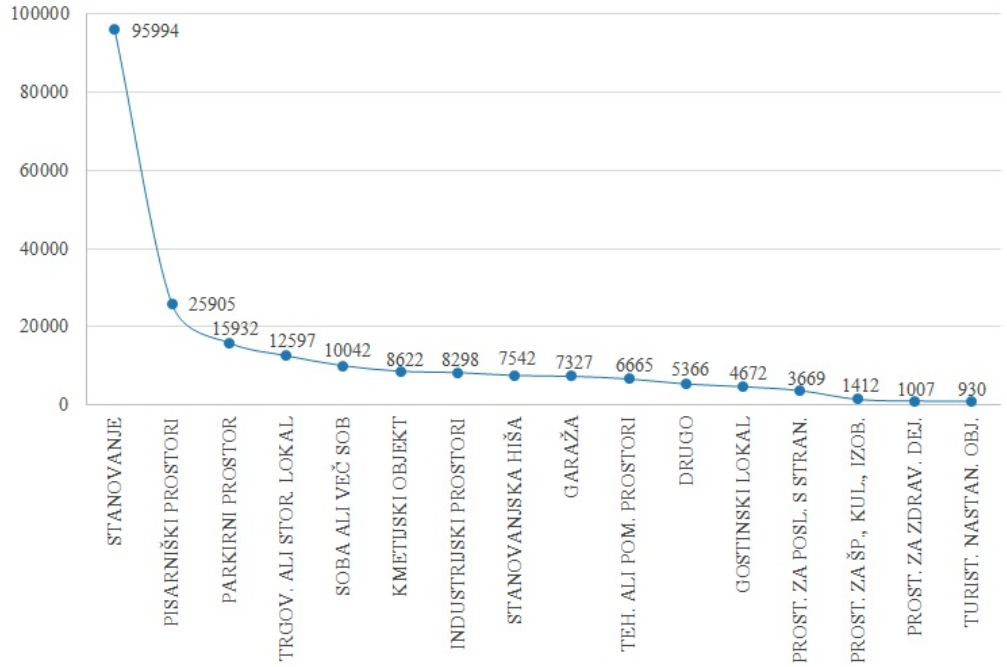
4.1.1 Analiza podatkov - najemni posli stanovanj

Domensko znanje o podatkih je pri pripravi zbirke podatkov pomembno, prav tako pa tudi poznavanje podatkov, s katerimi operiramo. Zato smo analizirali tudi novo zbirko podatkov za najemne posle za stanovanja. Atribute za podrobnejšo analizo smo izbrali po predhodnem posvetu z domenskim ekspertom - poznavalcem podatkov o nepremičninah. Tudi tu smo bili opozorjeni, da imajo podatki o najemnih poslih nizko stopnjo zaupanja. Za izbrane atribute smo izvedli začetno analizo in podatke zbrali v tabeli 4.5. Zbirka podatkov je v začetku vsebovala 50.359 zapisov.

Za atribut *pogodbena najemnina vseh oddanih površin* je iz grafa porazdelitve vrednosti na sliki 4.5 mogoče razbrati, da se vrednosti najemnin gibljejo

Tabela 4.5: Statistični podatki analiziranih atributov za najemne posle stanovanj.

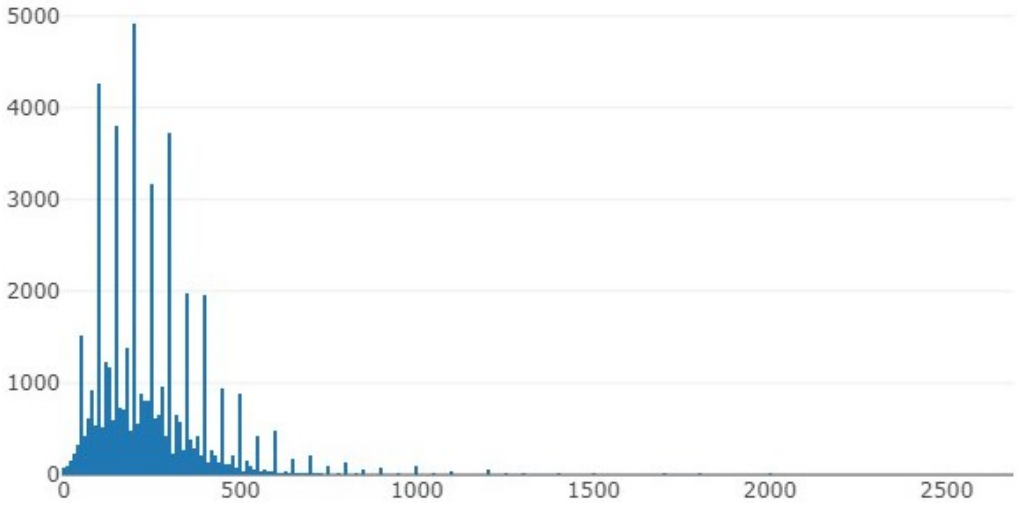
	povpr. vred.	std. odklon	min vred.	maks vred.	sred. vred.	manjkajočih vred. [%]
pogodbena najemnina vseh oddanih površin [€]	256,96	335,56	0,00	60.225,00	210,00	0,00
datum sklenitve pogodbe	28. 11. 2013	/	24. 11. 204	01. 11. 2017	28. 2. 2014	0,00
obratovalni stroški [ne/da]	0,16	0,37	0,00	1,00	0,00	0,00
čas najema [nedoločen/ določen]	1,29	0,45	1,00	2,00	1,00	0,00
trajanje najema [mesec]	19,90	495,76	0,00	84.011	12,00	28,60
opremljenost oddane površine [ne/da]	0,79	0,41	0,00	1,00	1,00	0,00
oddana površina [m^2]	58,10	310,82	0,00	43.090,00	49,00	0,00
število sob	2,16	1,76	0,00	55,00	2,00	5,10
površina dela stavbe [m^2]	302,38	2.525,81	0,00	28.500,00	56,60	0,60
uporabna površina dela stavbe [m^2]	80,17	255,77	0,00	22.161,40	48,80	1,20
povprečna mesečna bruto plača [€]	1.572,38	190,44	829,11	2.348,26	1.542,38	10,90
cena neosvinčenega bencina 95-oktanski [€/liter]	1,267	0,127	0,887	1,444	1,332	6,300
indeksi cen življenjskih potrebščin [%]	100,00	0,57	98,60	101,60	100,10	1,60



Slika 4.4: Porazdelitvi števila delov stavb za najemne posle po 'vrsti oddane površine'.

med 0 € in 60.225 €. Vrednosti so v glavnem porazdeljene do 590 €, kar kaže na porazdelitev z izrazitim desnim repom. Prvi kvartil za najemnine je pri 140 €, tretji pa pri 320 €.

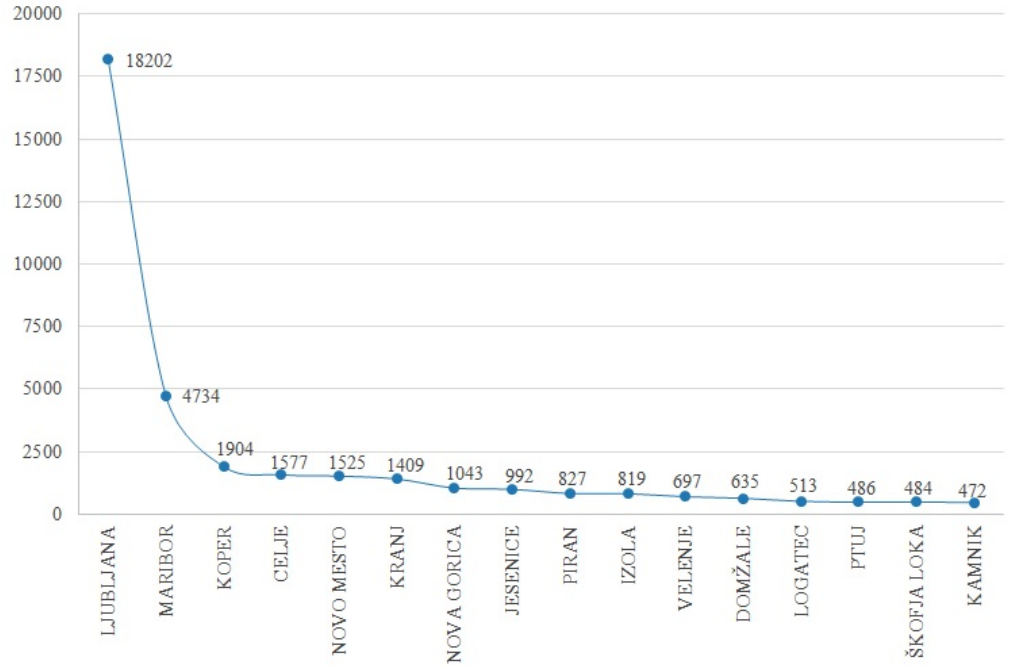
Iz podatkov je mogoče razbrati, da so na voljo podatki za *datum sklenitve pogodbe* do novembra 2017, srednja vrednost pa je februar oz. marec 2014. Večina podatkov se nahaja med letoma 2012 in 2017, kar kaže na porazdelitev z izrazitim levim repom. Vrednosti za atribut *obratovalni stroški* vsebujejo podatek o vključenosti obratovalnih stroškov, kjer vrednosti zavzemajo le dve vrednosti - ali najemnina vključuje obratovalne stroške ali ne (84 %). Iz povprečne vrednosti (0,16) lahko ugotovimo, da je večina vrednosti izrazito levo, kar nakazuje na to, da večina poslov 'ne vključuje obratovalnih stroškov'. Porazdelitev podatkov za atribut *čas najema* nakazuje na večji delež podatkov (72 %) za vrednost, ki pomeni *določen čas*, kar potrjuje tudi



Slika 4.5: Porazdelitev vrednosti po pogodbeni najemnini vseh oddanih površin za najemne posle stanovanj.

srednja vrednost, ki je 1. Iz statistike za atribut *trajanje najema* ugotovimo, da je kar 28 % zapisov brez podatka o trajanju najema. *Čas najema* in *trajanje najema* sta atributa, ki sta med sabo povezana. V primerih, ko gre za posel sklenjen za nedoločen čas, podatka o trajanju najema nimamo. Za posle, sklenjene za določen čas, pa ta podatek imamo. Več o manjkajočih vrednostih v zbirki podatkov smo opisali v podpoglavju 4.3. Srednja vrednost za *trajanje najema* nakazuje, da je največ poslov za določen čas sklenjenih za obdobje 12 mesecev. Pričakovana pa je porazdelitev podatkov za atribut *občina* (glej sliko 4.6), kjer opazimo, da je kar 36 % poslov sklenjenih v občini Ljubljana. Iz grafa je razvidno tudi, da je največje število poslov razporejenih po mestih, kar smo tudi pričakovali.

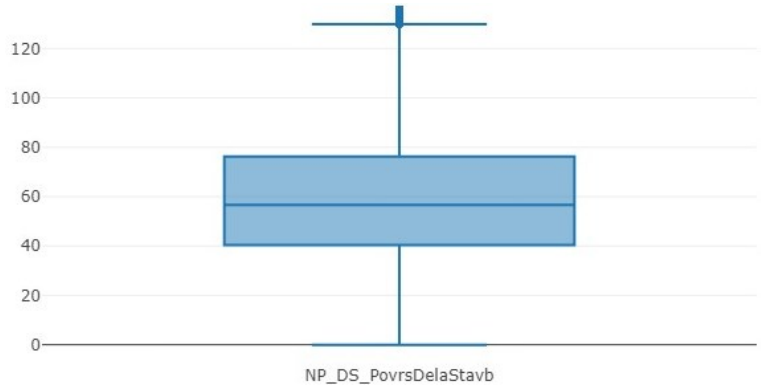
Iz statističnih podatkov za atribut *opremljenost oddane površine* je mogoče razbrati povprečno vrednost (0,79) in srednjo vrednost (1). Večina podatkov (79 %) nakazuje na dejstvo, da so stanovanja v glavnem opremljena. Z analizo atributa *oddana površina* smo ugotovili, da je stanovanje v povprečju manjše od 58,10 m^2, kar potrjuje tudi srednja vrednost 49 m^2 (glej tabelo 4.5). Večina vrednosti se nahaja v intervalu od 10,00 m^2 do 100,00 m^2, kar



Slika 4.6: Porazdelitev vrednosti po občinah za najemne posle stanovanj.

kaže na porazdelitev z izrazitim desnim repom. Iz statističnih podatkov (srednja vrednost (2) in povprečna vrednost (2,16)) za atribut *število sob* smo ugotovili, da je največ stanovanj dvosobnih, v večini pa so stanovanja eno, dvo ali trisobna. Povprečna vrednost za atribut *površina dela stavbe* je 302,38 m^2, srednja vrednost pa 56,6 m^2. Razlika je precejšnja, kar pomeni veliko razpršenost vrednosti, na kar opozarja tudi standardni odklon (2.525,81 m^2). Z grafa porazdelitve vrednosti (glej sliko 4.7) razberemo, da je prvi kvartal podatkov pri 40,40 m^2, tretji kvartal pa pri 76,20 m^2. Porazdelitve podatkov kažejo na desni rep porazdelitve podatkov za ta atribut.

Tudi vrednosti za atribut *uporabna površina stanovanja* so zamaknjene v levo in v glavnem zajemajo vrednosti do 200 m^2. Iz statistične analize atributa *povprečna mesečna bruto plača* (glej sliko 4.8) smo ugotovili, da je bilo največ poslov sklenjenih v času, ko je bila povprečna mesečna bruto plača v občini, kjer se stanovanje nahaja, 1.765 €. V povprečju sklepanja



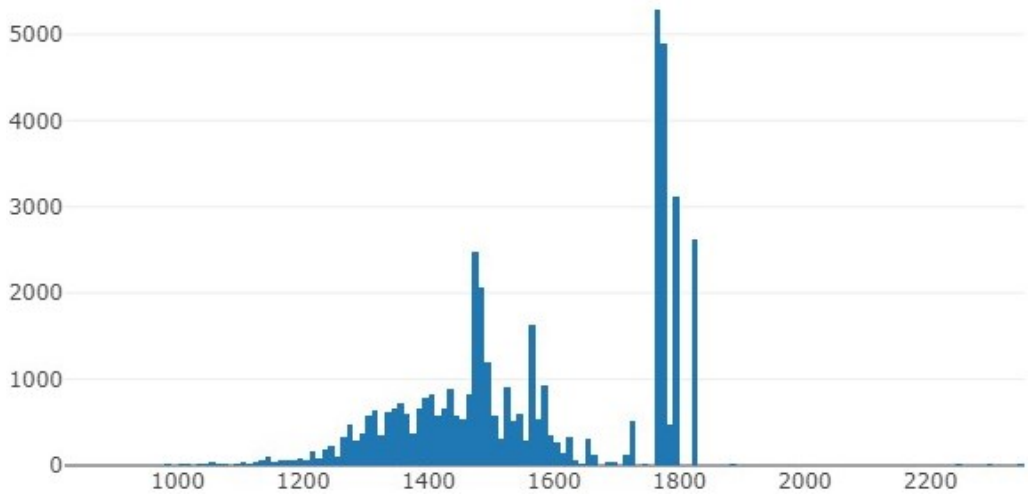
Slika 4.7: Porazdelitev vrednosti za atribut 'površina dela stavbe'.

poslov se je plača gibala okoli 1.572,38 €, največ poslov pa se je sklenilo, ko je bila povprečna bruto plača v občini med 1.431,49 € in 1.766,16 €.

V analizo zanimivih atributov smo uvrstili tudi atribut *cena neosvinčenega bencina 95-oktanski*. Iz statistike podatkov za atribut opazimo, da se je cena bencina v podatkih gibala med 0,887 €/liter in 1,444 €/liter, v povprečju pa je bila cena 1,267 €/liter. Največ poslov je bilo sklenjenih v času, ko se je cena gibala okoli 1,360 €/liter oz. v območju od 1,153 €/liter do 1,376 €/liter.

Zadnji izmed izbranih atributov za analizo pa je bil atribut *indeksi cen življenjskih potrebščin*. Atribut prikazuje mesečno gibanje drobnoprodajnih cen izdelkov in storitev. Podatke za ta atribut razumemo tudi kot merilo inflacije po klasifikaciji ECOICOP[1]. Povprečje za indeks cen življenjskih potrebščin je 100,00 %, standardni odklon pa 0,57 %, kar nakazuje na nizko stopnjo gibanja vrednosti. Največ poslov je bilo sklenjenih, ko je bil indeks cen med 99,70 % in 100,30 %. Srednja vrednost nakazuje na rahlo povišano inflacijo kar pomeni, da so se najemnine najpogosteje sklepale, ko je bil zaznan rahel dvig inflacije.

[1]ECOICOP - Evropska klasifikacija individualne potrošnje glede na namen (ang. European Classification of Individual Consumption according to Purpose).



Slika 4.8: Porazdelitev vrednosti za atribut 'povprečna mesečna bruto plača'.

4.1.2 Analiza podatkov celotne nove zbirke

Z analizo izbranih atributov ETN smo ustvarili podrobnejšo predstavo o tem, kakšne podatke imamo v naši zbirki podatkov in kakšne vrednosti vsebujejo. V naslednjem koraku smo se posvetili še analizi celotne zbirke podatkov najemnih poslov za stanovanja. Razvili smo skripto v *Pythonu*, ki za vsak atribut posebej vrne statistiko o vrednostih - število vrednosti, število manjkajočih vrednosti ter delež za vsako posebej. S pregledom statistike smo ugotovili, da atributi, ki vsebujejo malo vrednosti niti niso uporabni (atribut brez vrednosti doda k celotni informaciji zelo malo). Zato smo postavili mejo 40 % kot minimalni delež podatkov, ki jih mora atribut vsebovati, da ga obdržimo v naši zbirki podatkov. Med atributi, ki so bili pod mejo 40 % deleža podatkov je bil npr. tudi atribut *delovno aktivno prebivalstvo po občini prebivališča in delovnega mesta* za Ankaran. Ankaran je občina, ki je bila ustanovljena šele leta 2011 in je zanjo še zelo malo podatkov. Podatki za to občino obstajajo šele od leta 2011, posli iz območja občine pred 2011 so evidentirani pod občino Koper, od katere se je občina Ankaran odcepila. Ker naši izhodiščni podatki o koordinatah vsebujejo malo podatkov,

je bila med neuporabne podatke uvrščena tudi informacija iz datoteke *coordinates*. (Po razmisleku o podatkih za koordinate smo zaključili, da nam informacija o koordinatah niti ne koristi. Lokacija z natančnostjo zemljepisna širina in dolžina je preveč natančen podatek za napovedovanje vrednosti nepremičnin.) Ob pregledu statistike smo ugotovili, da imamo na voljo zelo malo podatkov REN za najemna stanovanja, zato smo podatke REN iz naše zbirke podatkov izpustili. Vključitev podatkov REN bi zagotovo pripomogla k izboljšanju napovedi vrednosti nepremičnin, tako da bomo idejo o razširitvi podatkovne zbirke s podatki REN uvrstili med možne izboljšave za nadaljnje delo. Atribut *datum sklenitve pogodbe* smo razširili z dodatnimi atributi, t.j. leto, mesec, kvartal in leto-mesec sklenitve. V kolikor smo z razširitvijo atributa naš nabor izboljšali, smo ugotavljali v poglavju 4.4. S filtriranjem podatkov, smo našo zbirko podatkov o najemih stanovanj iz 922 atributov zmanjšali na 579 atributov. Zbirka je vsebovala 50.359 zapisov. Ugotovitve o podatkih smo aplicirali še na zbirko kupoprodaj za stanovanja; zbirka kupoprodaj za stanovanja je vsebovala 79.841 zapisov, iz začetnih 914 atributov pa smo zbirko skrčili na 583 atributov.

4.2 Metode za iskanje in odstranjevanje osamelcev

Eden ključnih korakov priprave podatkov za napovedovanje v magistrskem delu je tudi čiščenje podatkov, kamor sodita iskanje in odstranjevanje osamelcev oz. ekstremnih vrednosti. Osamelec [35] je vrednost, ki se glede na opredeljena merila bistveno razlikuje od drugih vrednosti. Definicija osamelca ni absolutna, pojem je natančno določen šele z izbiro ustreznih kriterijev v posameznem raziskovanju. Kadar obravnavamo osamelce, je pomembno predvsem določiti:

- ali je vrednost osamelca napačen podatek ali pa gre za sicer izstopajoč, vendar pravilen podatek,

- ali je osamelec – če gre za pravilen podatek in če raziskovanje izvajamo na podlagi slučajnega vzorca – reprezentativen. Osamelec je reprezentativen, če v populaciji obstaja (vsaj približno) tolikšno število podatkov podobnega obsega, kot je faktor preračuna na populacijo (populacijska utež).

Za osamelce velja, da so njihove vrednosti neobičajno nizke ali visoke vrednosti - odklon od povprečja je velik v primerjavi z variabilnostjo. V primeru, da osamelcev ne prepoznamo in z njimi ne ravnamo ustrezno, lahko izkrivljajo napoved in vplivajo na točnost. To velja še zlasti za regresijske modele [36]. Če ekstremnih vrednosti ne odstranimo, se lahko ocene in napovedi močno pristransko spremenijo. Verjetno je sicer, da so podatki za posel pravi, lahko pa je bila pri vnosu podatka storjena napaka, ki jo uvrščamo med napake pri zbiranju podatkov oz. vnosu/manipulaciji podatkov. Odločili smo se, da zapise za posle, ki vsebujejo vrednosti, zaznane kot osamelec, izpustimo. Pristopov za iskanje osamelcev je več, npr. univarianten pristop, vizualizacija, filtriranje... [37].

4.2.1 Razširitev atributa "prostori stanovanja"

Postopek iskanja, zaznavanja in odstranjevanja osamelcev smo izvajali v programskem jeziku *R*. Podatke iz zbirke podatkov, ki smo jo pripravili v prejšnjem koraku, smo najprej prebrali iz datoteke *csv* ter pričeli z analizo podatkov. Tekom celotnega postopka iskanja in odstranjevanja osamelcev smo izvajali različne analize podatkov. Ugotovili smo, da atribut *prostori stanovanja* vsebuje vrednosti tipa besedilo (ang. string), vrednosti pa so našteti prostori stanovanja, ki so med seboj ločeni s pokončno črto '|'. Metode za napovedovanje ne delujejo nad vhodnim naborom podatkov v obliki besedila, zato smo omenjene podatke razširili v več stolpcev.

Z uporabo funkcije smo sestavili unikatni seznam vrednosti (drvarnica, klet. shramba, zaprt balkon...) v posameznem stolpcu, vsaka izmed teh pa je predstavljala po en binarni stolpec, razširjen iz stolpca *prostori stanovanj*.

Tako smo ustvarili novo matriko s 16 stolpci (16 različnih unikatnih vrednosti), ki je razširila atribut *prostori stanovanj*. Z razširitveno matriko smo nato zbirko podatkov o najemnih poslih stanovanj še razširili, vrednosti za atribut *prostori stanovanj* pa odstranili iz zbirke podatkov.

4.2.2 Iskanje osamelcev

Nekatere atribute smo odstranili že ob pripravi podatkovne zbirke, saj smo ugotovili, da pri nekaterih manjka veliko vrednosti (glej poglavje 4.1.2). Nekateri so bili le identifikatorji, zato smo odstranili tudi tiste. Za iskanje ekstremnih vrednosti v naši zbirki smo uporabili univarianten pristop, ki za dano zvezno spremenljivko kot osamelce označi vse tiste vrednosti, ki ležijo izven določenega območja, npr. `1.5 * IQR`. Konstanta IQR (ang. Interquartile Range) je razlika med prvim in tretjim kvartalom. Prvi kvartal je vrednost, kjer je 25 % vrednosti v seznamu manjših od te vrednosti, tretji kvartal pa vrednost, kjer je manjših 75 % vrednosti oz. 25 % vrednosti večjih od te vrednosti. Najprej smo poiskali tiste posle, ki izstopajo glede na *datum sklenitve posla* in preverili kakšna je osnovna statistika (glej tabelo 4.6) za atribut *datum sklenitve pogodbe - leto*:

Tabela 4.6: Statistika za atribut 'datum sklenitve pogodbe - leto'.

minimum	1. kvartal	srednja vrednost	povprečje	3. kvartal	maksimum
204	2013	2014	2013	2015	2017

Iz statistike ugotovimo, da je minimalna letnica osamelec, ki je verjetno nastala zaradi napačnega vnosa podatkov. Za iskanje osamelcev smo najprej razvili funkcijo, ki v seznamu vrednosti poišče osamelce po univariantnem pristopu. Najprej pridobimo seznam unikatnih vrednosti za atribut *datum sklenitve pogodbe - leto* ter število pojavitev vrednosti v zbirki podatkov kot prikazuje tabela 4.7.

Iz statistike pojavitve vrednosti po vrednostih ugotovimo, da je kar nekaj

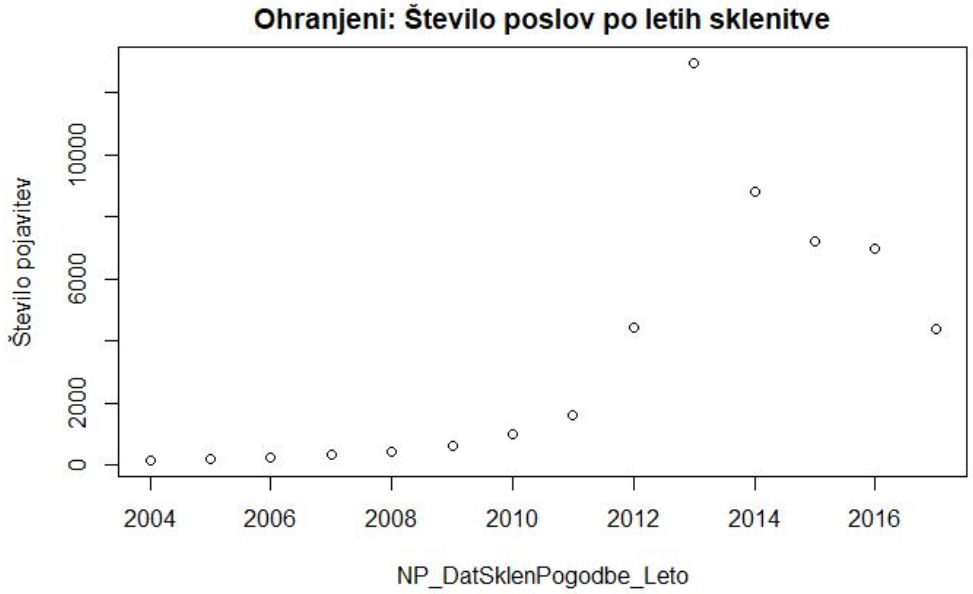
Tabela 4.7: Število pojavitev nekaterih vrednosti v zbirki podatkov.

NP_DatSklenPogodbe_Leto	total
204	1
205	1
1911	1
1912	3
1913	7
1914	2
1955	1
1960	1
1967	1
1968	1

takšnih poslov, pri katerih je bil verjetno ob vnosu podatkov vnesen napačen datum sklenitve pogodbe. Očitna osamelca sta letnici '204' in '205'.

Pred končnim odstranjevanjem smo najprej poiskali primerno vrednost za `factor_IQR`. Najprej smo preverili, ali je za `factor_IQR` vrednost 1,8 smiselna, kjer smo najprej odstranili 2.972 vrednosti in po pregledu odstranjenih vrednosti ugotovili, da je bilo odstranjenih kar nekaj takšnih vrednosti, ki so povsem ustrezne in jih po našem mnenju ne moremo uvrstiti med osamelce. Preverili smo še, katere vrednosti odstranimo z univariantnim pristopom po *datumu sklenitve pogodbe* in ne po *letnici*, vendar smo tudi na ta način odstranili relevantne podatke, t.j. tiste posle, ki so bili sklenjeni v zadnjih 10 letih. Naš cilj je bil odstraniti le tiste posle, ki so 'stari'. Poiskali smo primerno mejo za leto pri odstranjevanju starejših poslov. Iz podatkov smo ugotovili, da je bilo pred letom 2004 sklenjenih le 910 poslov. Po odstranitvi poslov sklenjenih pred 2004, je naša zbirka vsebovala še 49.449 poslov, porazdelitev poslov pa prikazuje slika 4.9.

Osamelce smo iskali le znotraj podatkov ETN, saj smo ostalim podatkom zaupali. Analiza je pokazala, da vrednosti ne odstopajo od večine in osamelcev nismo zaznali. Dodaten razlog za tako odločitev je bilo tudi dejstvo, da



Slika 4.9: Porazdelitev najemnih poslov za stanovanja po letnici sklenitve pogodbe po odstranitvi osamelcev za ta atribut.

so nas na GURSu opozorili, da so v podatkih (lahko) napake. Še posebej so nas opozorili, da je največ napak v podatkih za najemne posle, najmanj zaupanja vreden podatek pa je atribut, ki ga napovedujemo, t.j. *pogodbena najemnina vseh oddanih površin*. Zato smo precej pozornosti namenili temu atributu. Ekstremne vrednosti za ta atribut smo odstranjevali ročno, in sicer na podlagi analize vrednosti za atribut v odvisnosti od atributa *obratovalni stroški*. Z analizo atributa smo ugotovili, da je zbirka vsebovala 48 poslov z najemnino nižjo od 20 € in vključenimi stroški, 137 poslov pa stroškov nima vključenih, a je najemnina nižja od 20 €. Kot osamelce smo označili tiste posle, ki so vsebovali obratovalne stroške in je bila najemnina nižja od 20 €.

Po priporočilu GURSa smo za atribut *pogodbena najemnina vseh oddanih površin* iskali osamelce po vsaki občini posebej, z uporabo univariantne metode. Ugotovili smo, da vrednosti 1,5 in 2,2 za `factor_IQR` nista primerni. Izkazalo se je, da je za iskanje osamelcev za ta atribut najprimernejša vre-

dnost za `factor_IQR` 2,0. S takšnim pristopom smo odstranili 933 poslov. Največ, 520 poslov je bilo odstranjenih za občino Ljubljana. Ob pregledu podatkov smo ugotovili, da so ohranjeni podatki smiselni. Prav tako ocenjujemo, da smo s takšnim pristopom odstranili le nesmiselne podatke. Z vrednostma 1,8 in 2,0 za `factor_IQR` smo zaznali sicer zelo podobne osamelce.

Z univariantnim pristopom in sprotno analizo posameznega atributa, smo poiskali ekstremne vrednosti še za naslednje atribute:

- trajanje najema,
- oddana površina v m^2,
- leto izgradnje stavbe,
- število sob v stanovanju,
- površina stanovanja v m^2,
- uporabna površina stanovanja v m^2.

Za atribute, ki označujejo lokacijo nepremičnine (naselje, šifra katastrske občine, občina), osamelcev nismo iskali, saj so ti podatki kategorični. Več o kategoričnih atributih smo opisali v poglavju 4.4. Na koncu celotnega postopka iskanja osamelcev smo iz zbirke podatkov dejansko odstranili tiste posle, za katere smo ugotovili, da vsebujejo vsaj en osamelec. Takšnih poslov je bilo 10.494. Po odstranitvi vseh osamelcev je naša zbirka podatkov za najeme stanovanj v tej fazi priprave podatkov vsebovala 37.974 poslov.

Zgoraj opisane postopke iskanja in odstranjevanja osamelcev smo aplicirali še nad podatki o kupoprodajah stanovanj, le da smo meje tekom postopkov nekoliko prilagajali podatkom, tako da smo ohranili smiselne vrednosti. Iz začetnih 79.841 zapisov, smo jih na koncu ohranili še 69.225. Torej smo odstranili približno enako število poslov kot pri najemnih poslih.

4.3 Metode za vstavljanje manjkajočih vrednosti

Skozi analize, ki smo jih izvajali nad podatki najemnih stanovanj, smo označili nekaj atributov kot nepotrebne, še posebej na podlagi pomena posameznih atributov. Tako smo v fazi vstavljanja manjkajočih vrednosti odstranili naslednje atribute: vrsta najemnega pravnega posla, ime katastrske občine, številka stavbe, številka dela stavbe, ulica, hišna številka, številka stanovanja ali poslovnega prostora, prostori stanovanja, vrsta oddane površine. Tudi v tej fazi priprave podatkov smo se ukvarjali s sprotno analizo podatkov in odstranjevanjem nekaterih atributov ter zapisov. V začetku postopka vstavljanja manjkajočih vrednosti smo razvili metodo, ki izračuna statistični pregled manjkajočih vrednosti po atributih. Ugotovili smo, da kar 359 od skupno 595 atributov v zbirki podatkov vsebuje manjkajoče vrednosti. Manjkajoče vrednosti v zbirki podatkov so predstavljale 0,84 % vseh podatkov. Ugotovili smo tudi, da je veliko takšnih atributov, ki vsebujejo le nekaj 100 manjkajočih vrednosti. Na podlagi ugotovitev smo odstranili tiste atribute, ki vsebujejo več kot 1.200 manjkajočih vrednosti. Z odstranitvijo le-teh smo zbirko podatkov zmanjšali za 25 atributov (zbirka podatkov pa je še vedno vsebovala 334 atributov, ki so vsebovali manjkajoče vrednosti). S tem pristopom smo velik delež manjkajočih vrednosti izpustili/odstranili, tako da je zbirka podatkov vsebovala le še 0,25 % manjkajočih vrednosti, kar je predstavljalo 53.160 posameznih manjkajočih vrednosti.

Iz podatkov smo ugotovili, da mlajši občini Ankaran in Mirna predstavljata kar 13,71 % vseh manjkajočih vrednosti (t.j. 7.289 vrednosti) in le 97 sklenjenih poslov, kar predstavlja 0,26 % vseh poslov. Opazimo tudi, da manjkajoče vrednosti za občini predstavljata 4,50 % vseh manjkajočih vrednosti, glede na začetno zbirko podatkov, t.j. po odstranitvi osamelcev. To je bil razlog, da smo iz naše zbirke podatkov odstranili zapise za ti dve občini in tako delež manjkajočih vrednosti znižali na 0,22 %. Posli sklenjeni pred letom 2011 v naši zbirki podatkov predstavljajo 4,70 % podatkov in kar

39,70 % manjkajočih vrednosti v podatkih. S ponovno analizo podatkov po atributih ugotovimo, da posli, ki so bili sklenjeni pred letom 2011 predstavljajo zelo velik delež manjkajočih vrednosti (94,08 %). Zato odstranimo vse posle (1.778 zapisov), ki so bili sklenjeni pred letom 2011. Z odstranitvijo teh poslov je zbirka podatkov vsebovala še 36.099 zapisov. V primerjavi z začetno zbirko podatkov smo do te faze čiščenja, število manjkajočih vrednosti uspeli znižati na le 2.749, kar predstavlja le še 0,01 % vseh vrednosti v celotni zbirki podatkov. Analiza zbirke podatkov po odstranitvi je bila za nas zelo pozitivna in je prikazana v tabeli 4.8.

Tabela 4.8: Statistika manjkajočih vrednosti po odstranitvi nekaterih atributov in poslov sklenjenih pred 2011.

atribut	število manjkajočih vrednosti
DelavnoAktivniPoObciniDelovnegaMestaIzobrazba_Osnovnosolska_ali_manj	1
DelavnoAktivniPoObciniDelovnegaMestaIzobrazba_Visjesolska_visokosolska	1
StOtrokNaVrtecInVzgojitelja_Stevilo_otrok_na_vrtec	5
VrtciPoSteviluOtrokInObciniZavoda_Stevilo_otrok	5
VrtciPoSteviluOtrokInObciniZavoda_Stevilo_vrtcev_in_enot	5
NP_DS_LetoIzgradnje	142
NP_DS_PovrsDelaStavb	232
ObrestneMereEvropskeCentralneBanke_Mejni_depozit	261
NP_DS_UporPovrsDelaStavb	402
NP_DS_StevSob	1.135

Za občino Osilnica zbirka podatkov vsebuje le en posel, ki vsebuje kar 5 manjkajočih vrednosti. 12 poslov je bilo sklenjenih za občino Rečica ob

Savinji in občino Jezersko, ki pa skupaj vsebujejo 13 manjkajočih vrednosti. Po odstranitvi poslov sklenjenih za stanovanja v občinah Osilnica, Rečica ob Savinji in Jezersko, število manjkajočih vrednosti močno skrčimo, kot prikazuje tabela 4.9. Iz celotne zbirke podatkov nam je tako ostalo le še 5 atributov, ki vsebujejo manjkajoče vrednosti. Za nadaljevanje postopka vstavljanja vrednosti smo zgradili zbirko podatkov iz atributov (556 atributov), ki ne vsebujejo manjkajočih vrednosti.

Tabela 4.9: Statistika manjkajočih vrednosti po odstranitvi nekaterih atributov in poslov.

atribut	število manjkajočih vrednosti
NP_DS_LetoIzgradnje	142
NP_DS_PovrsDelaStavb	232
ObrestneMereEvropskeCentralneBanke_Mejni_depozit	261
NP_DS_UporPovrsDelaStavb	402
NP_DS_StevSob	1.134

V nadaljevanju postopka vstavljanja manjkajočih vrednosti smo za 5 atributov (glej tabelo 4.9), ki so še vsebovali manjkajoče vrednosti, uporabili metode za vstavljanje, ki na podlagi zgrajenega modela iz zbirke podatkov brez manjkajočih vrednosti, predlaga vrednosti na mesto manjkajočih vrednosti. Za vstavljanje manjkajočih vrednosti za posamezni atribut smo uporabili metodo `mice`, ki na podlagi izbrane metode za grajenje modela predlaga oz. napove dejanske vrednosti za manjkajoče vrednosti. Multivariantno vstavljanje z verižnimi enačbami `mice` (ang. Multivariate Imputation by Chained Equations) [38] je knjižnica v okolju R, ki vsebuje napredne funkcije za manjkajoče vrednosti. Vstavljanje se izvaja v dveh korakih - grajenje modela in generiranje končnih vrednosti. Funkcija `mice` naredi več popolnih kopij podane zbirke podatkov (df) od katerih za vsako izvede različno vstavljanje manjkajočih podatkov. Funkcija `complete` vrne vrednosti za atribute, ki so

manjkajoči. Najprej smo vstavljanje manjkajočih vrednosti izvedli za atribute iz zbirke ETN: leto izgradnje, površina stanovanja, uporabna površina stanovanja, število sob.

Postopek smo izvajali za vsak atribut posebej ter sproti tudi analizo in preverjanje napovedanih manjkajočih vrednosti. Z analizo in preverjanjem le-teh smo ugotovili, da so napovedane vrednosti smiselne. Za napovedane vrednosti za atribut *površina stanovanja* smo npr. vrednosti primerjali tudi z ostalimi atributi za površino (oddana površina, uporabna površina) in atributi za prostore stanovanja. Ob tem smo ugotovili, da za stanovanja, kjer manjka vrednost za atribut *površina dela stavbe* in atribut *število sob*, manjka tudi vrednost za atribut *uporabna površina stanovanja* (glej tabelo 4.10).

Tabela 4.10: Pregled napovedanih vrednosti za 'površina stanovanja' v primerjavi z ostalimi atributi.

napovedi PovrsDelaStavb	NP_DS_OddanaPovrs [m^2]	NP_DS_UporPovrs DelaStavb [m^2]
38,6	35,00	/
35,3	33,60	/
110,8	110,95	/
76,6	76,60	/
66,2	54,50	/
80,6	83,30	/
25,1	29,60	/
60,2	55,20	/
81,9	85,60	/
51,5	56,00	/

Med postopkom vstavljanja podatkov, smo poskusili tudi z gradnjo modela nad atributi, ki po našem mnenju najbolj (smiselno) vplivajo na atribut *površina stanovanja*. Poleg iskanega atributa, smo med atribute za gradnjo modela uvrstili še:

- oddana površina stanovanja,

- opremljenost oddane površine,
- dejanska raba stanovanja,
- lega stanovanja,
- prostori stanovanja.

Ko smo napovedane vrednosti za manjkajoče vrednosti po modelu iz izbranih atributov primerjali z ostalimi atributi in napovedjo opisano zgoraj, smo ugotovili, da so napovedane vrednosti po modelu iz izbranih atributov manj smiselne, kot po modelu iz vseh atributov (glej tabelo 4.11).

Tabela 4.11: Pregled napovedanih vrednosti za 'površina stanovanja' po modelu iz izbranih atributov v primerjavi z ostalimi atributi.

napovedi PovrsDelaStavb 2	napovedi PovrsDelaStavb	NP_DS_OddanaPovrs [m^2]
36,4	35,0	35,00
75,0	33,0	33,60
102,9	107,7	110,95
93,4	82,4	76,60
69,9	12,7	54,50
66,4	82,7	83,30
34,0	29,5	29,60
58,1	55,0	55,20
83,0	85,6	85,60
55,4	68,1	56,00

Enak postopek vstavljanja manjkajočih vrednosti smo uporabili tudi za iskanje manjkajočih vrednosti za ostale atribute iz zbirke ETN. Na koncu nam je ostal le še atribut *obrestne mere Evropske centralne banke (ECB)*, ki je vseboval 261 manjkajočih vrednosti. Tudi nad tem atributom smo napovedali manjkajoče vrednosti z uporabo metodo `mice`, ter napovedane vrednosti tudi analizirali. Ugotovili smo, da so napovedane vrednosti za obresti v veliki večini zelo smiselne in zajemajo tudi nihanje obrestnih mer,

saj za obdobje 8. 4. 2009 - 13. 4. 2011 ni bilo na voljo podatkov. Ker smo imeli že v samem začetku na voljo podatke za obresti le do 16. 3. 2016, smo za obdobje po tem datumu določili kar vrednost z dne 16. 3. 2016. Zato ocenjujemo, da je napoved manjkajočih vrednosti za atribut *obrestne mere ECB* zelo smiselna.

S pregledom vrednosti po atributih smo ugotovili, da nekateri atributi vsebujejo besedilne vrednosti. Takšni atributi so: občina, naselje, dejanska raba stanovanja, lega stanovanja v stavbi. Ugotovili smo tudi, da nekateri atributi vsebujejo pomensko enake vrednosti, vendar vsebujejo šumnike ali pa posebne znake. Zato smo se odločili, da skladno z našimi predpostavkami za možen nastanek težav, zaradi posebnih znakov v vrednostih za omenjene atribute, te vrednosti popravimo. S tem posegom nismo spremenili dejanski pomen vrednosti. V ta namen smo razvili funkcijo, ki nad podanim nizom z uporabo regularnih izrazov izvede zamenjavo znakov. Nato za (vse štiri) naštete atribute, vrednosti atributov v zbirki podatkov ustrezno popravimo.

Za konec postopka vstavljanja manjkajočih vrednosti, smo izvedli še nekaj statističnih izračunov in primerjav za oceno uspešnosti tega koraka. Zbirka podatkov je na začetku vsebovala 188.391 manjkajočih vrednosti, kar predstavlja 0,84 % vseh podatkov v zbirki podatkov. Skozi celoten postopek smo manjkajoče vrednosti v zbirki uspeli izničiti. Prečiščeno zbirko podatkov za najemne posle stanovanj smo zapisali v *csv* datoteko, ki smo jo uporabljali pri nadaljnjem delu. Zbirka je na koncu vsebovala 561 atributov in 36.086 najemnih poslov za stanovanja (zapisov). Tako smo pripravili zbirko podatkov, ki jo je mogoče uporabiti za grajenje napovednih modelov in napovedovanje vrednosti za *pogodbene najemnine stanovanj*.

Opisane postopke za vstavljanje manjkajočih vrednosti v zbirki najemnih poslov za stanovanja smo aplicirali še nad zbirko kupoprodajnih poslov za stanovanja. Zbirka je v začetku vsebovala kar 2,09 % manjkajočih vrednosti celotne zbirke, t.j. 853.261 vrednosti. Na koncu pa je zbirka vsebovala 531 atributov in 60.165 kupoprodajnih poslov za stanovanja (zapisov). Pripravljeno zbirko podatkov je že bilo mogoče uporabiti za grajenje napovednih

modelov in tako napovedovanje vrednosti za *pogodbene cene stanovanj.*

4.4 Izbor atributov

Po končani fazi priprave zbirke podatkov za najemne posle stanovanj je sledila še faza iskanja najpomembnejših napovedovalnih spremenljivk (ang. feature selection), ki pojasnjujejo večji del variance odzivne spremenljivke. Iskanje najpomembnejših napovedovalnih spremenljivk je ključnega pomena za prepoznavo in izgradnjo uspešnih napovednih modelov [36]. Poiskali smo najpomembnejše spremenljivke in na ta način učinkovito zmanjšali nabor atributov.

4.4.1 Metoda naključnih gozdov

Metoda naključnih gozdov je zelo učinkovita za iskanje nabora odvisnih spremenljivk, ki najbolje razlagajo varianco v odzivni spremenljivki. Za določitev atributov, ki najbolje razlagajo varianco smo uporabili metodo `cforest` [39] iz knjižnice `party` v programskem jeziku *R* kot je prikazano na primeru 4.1. Algoritem naključnih gozdov smo opisali v poglavju 5.2. Implementacija algoritma naključnih gozdov v funkciji `cforest` se razlikuje od referenčne implementacije v `randomForest`[2] glede na uporabljene osnovne učne primerke in uporabljeno shemo agregacije.

[2] `randomForest` izvaja Breimanov algoritem naključnih gozdov za klasifikacijo, ki se lahko uporablja v nenadzorovanem načinu za iskanje osamelcev ali ocenjevanje bližine med podatkovnimi točkami.

Koda 4.1: Procedura v programskem jeziku R, ki z uporabo metode 'cforest' in 'varimp' poišče najučinkovitejše napovednike.

```
library(party)

# fit the random forest
cf1 <- cforest(NP_PogodbNajemnVsehOddanPovrs ~ . , data=data_NP,
       control=cforest_unbiased(mtry=2,ntree=5))

# get variable importance, based on mean decrease in accuracy
varimp_OnMeanDecrease <- varimp(cf1)

sorted_varimp_OnMeanDecrease <- sort(varimp_OnMeanDecrease,
                    decreasing = TRUE)
```

Ob klicu funkcije podamo tudi nekatere parametre:

- `mtry` - število naključno izbranih spremenljivk, ki je pritrjen na kvadratni koren števila vhodnih spremenljivk,
- `ntree` - število dreves.

Shema agregacije deluje s povprečenjem uteži spremenljivk (atributov), pridobljenimi iz vsakega drevesa (`ntree`) in ne s povprečjem napovedi neposredno kot v `randomForest`. Metodi smo podali tudi podatek za simbolični opis primernega modela (`formula`) in podatkovni okvir, ki vsebuje vse spremenljivke v modelu (`data`). Simbolični opis pomeni, da je naš odziv modela atribut *pogodbena najemnina vseh oddanih površin* definiran kot vsi ostali atributi v podanih podatkih. Z metodo `varimp` [40] iz zgrajenega modela pridobimo pomen spremenljivk v modelu, ki temelji na povprečnem zmanjšanju natančnosti. Končni rezultat predstavlja urejen seznam spremenljivk v padajočem vrstnem redu po pomembnosti - od najbolj do najmanj pomembne spremenljivke. Med najpomembnejše atribute je metoda *cforest* uvrstila kar nekaj atributov iz dodatnih podatkov, npr. *delovno aktivno prebivalstvo po občini prebivališča in delovnega mesta* ter *razdalja do slovenskih mest*, kar nakazuje na to, da je bila razširitev zbirke podatkov z dodatnimi podatki (SURS, razdalje med občinami in mesti,...) smiselna. Iz zbirke ETN sta najvišje uvrščena atributa *uporabna površina stanovanja* in *datum sklenitve pogodbe*.

4.4.2 Algoritem Boruta

Za določanje ali je spremenljivka pomembna ali ne, smo uporabili metodo *Boruta* [41]. Metoda *Boruta* je algoritem za izbiro spremenljivk, ki deluje s poljubno klasifikacijsko metodo, katere izhod je pomembnost spremenljivke (ang. variable importance measure) (VIM). Metoda *Boruta* privzeto uporablja naključne gozdove. Metoda izvaja iskanje od zgoraj navzdol za ustrezne spremenljivke, tako da primerja pomembnost originalnih atributov s pomembnostjo doseženo naključno, ocenjeno z uporabo njihovih permutiranih kopij. Nepomembne spremenljivke za stabilizacijo poskusa odstranjuje postopoma. *Boruta* iterativno primerja pomembnost atributov s pomembnostjo senčnih atributov, ustvarjenih z mešanjem prvotnih. Atributi, ki imajo slabše pomembnosti kot sence prvotnih, se zaporedoma izpuščajo. Po drugi strani pa so atributi, ki so bistveno boljši od senc, sprejeti v potrditev (ang. confirmed). Algoritem se ustavi, ko ostanejo le potrjeni atributi ali ko doseže vrednost `maxRuns`, kot največje število ciklov pomembnosti. Če bi želeli razrešiti atribute, ki so ostali kot začasni (ang. tentative), vrednost za ta parameter lahko povečamo. Lahko pride tudi do drugega scenarija in nekateri atributi ostanejo brez odločitve. Takšne atribute se označi kot začasne (ang. tentative). S povečanjem atributa `maxRuns` ali znižanjem `pValue` bi atribute lahko razjasnili, v nekaterih primerih pa njihove pomembnosti nihajo preveč za *Borutsko konvergenco*. Namesto tega se lahko uporabi funkcijo `TentativeRoughFix`, ki bo izvedla drugačen, šibkejši test za izdelavo dokončne odločitve ali pa se jih preprosto obravnava kot neodločene pri nadaljnji analizi. Za izvedbo izbire atributov z metodo *Boruta*, smo ob klicu metode `Boruta` podali tudi simbolični opis modela, ki ga je treba analizirati. Podali smo tudi zbirko podatkov in nastavili sled na opcijo *2*, kar pomeni, da poroča odločitev takoj, ko je upravičeno poročanje o vsakem pomembnem viru.

Iz rezultata metode smo uporabili vrednost *'finalDecision'* (faktor treh vrednosti, ki vsebuje končni rezultat izbire funkcij: potrjen, zavrnjen ali začasen) (glej tabelo 4.12) ter izbrali le tiste atribute, ki so bili potrjeni ali

začasni. To je bil naš cilj iskanja pomembnih atributov z metodo *Boruta*.

Tabela 4.12: Porazdelitev atributov po faktorju treh vrednosti kot rezultat metode *Boruta*.

faktor treh vrednosti	število atributov
potrjen	310
začasen	110
zavrnjen	140

Iz porazdelitve po faktorjih treh vrednosti kot rezultat metode *Boruta* (glej tabelo 4.12) ugotovimo, da je 140 takšnih atributov, ki so v naši zbirki podatkov, vendar niso primerni za napovedovanje. Primernih atributov je 310, ostalih 110 pa je začasnih in jih lahko razumemo kot pogojno sprejete. Ugotovimo tudi, da je večino atributov iz zbirke ETN označenih kot *sprejeti*. Zelo zanimiva ugotovitev je tudi, da so vsi atributi, ki označujejo razdalje med občino lokacije stanovanja in mesti, uvrščeni v množici *potrjeni* ali *začasni*, nobenega takšnega atributa pa ni v množico *zavrnjeni*. Med zavrnjenimi sta npr. atributa, ki smo ju razširili iz atributa *datum sklenitve pogodbe*, t.j. *datum sklenitve pogodbe - leto* in *datum sklenitve pogodbe - kvartal*. Tudi ta porazdelitev atributov po faktorjih treh vrednosti kaže na to, da je bila naša razširitev zbirke podatkov z dodatnimi podatki (SURS, razdalje med občinami in mesti,...) smiselna.

4.4.3 Združitev izbora atributov

Izbor najpomembnejših napovedovalnih atributov po pomembnosti smo izvedli z dvema različnima pristopoma, z metodo *naključnih gozdov* (glej poglavje 4.4.1) in metodo *Boruta* (glej poglavje 4.4.2). Iz vsake metode posebej smo dobili rezultate, ki pa jih je bilo potrebno združiti. Odločili smo se, da bomo rezultate metod združili kot presek obeh in iz preseka vzeli približno 100 najpomembnejših atributov. Najprej smo se posvetili analizi rezultatov

po metodi *naključnih gozdov* in ugotovili, da po tej metodi med 120 najpomembnejših atributov v zbirki, kar 13 od skupno 40 atributov iz zbirke ETN (urejeni od najpomembnejšega do najmanj pomembnega):

- uporabna površina stanovanja,
- datum sklenitve pogodbe,
- datum začetka najema,
- čas najema,
- dejanska raba stanovanja,
- leto izgradnje stavbe,
- število sob,
- šifra katastrske občine,
- ali je podatke posredovala nepremičninska agencija,
- površina stanovanja,
- oddana površina stanovanja,
- opremljenost oddane površine,
- občina.

V preseku 120 najpomembnejših atributov po metodi naključnih gozdov in potrjenih atributov po metodi Boruta, je le 86 atributov. Preverili smo tudi presek najpomembnejših 120 atributov po metodi naključnih gozdov s potrjenimi in začasnimi atributi po metodi Boruta - v tem preseku je 109 atributov. Ob preverjanju smiselnosti atributov smo ugotovili, da so v preseku atributov obeh metod tisti atributi, ki smo jih našteli že zgoraj in so po naši oceni tudi smiselni. Presek potrjenih atributov po metodi Boruta s 150 najpomembnejšimi atributi po metodi naključnih gozdov pa vsebuje 106 atributov. 133 atributov je vsebovanih v preseku potrjenih in začasnih atributov

po metodi Boruta s 150 najpomembnejšimi atributi po metodi naključnih gozdov. Ob primerjavi vseh štirih scenarijev združitve izbranih atributov in preučitvi, katera združitev vsebuje najbolj smiselne atribute smo zaključili, da je najbolj smiselna združitev atributov iz preseka 120 najpomembnejših atributov po metodi naključnih gozdov ter potrjenih in začasnih atributov po metodi Boruta. Tako naš končni izbor najpomembnejših atributov vsebuje 109 atributov, kjer je 13 atributov iz ETN.

Z apliciranjem zgornjih postopkov nad zbirko kupoprodajnih poslov stanovanj za izbor najpomembnejših atributov, smo izbrali 112 atributov, med katerimi je le 10 atributov iz zbirke ETN.

4.4.4 Zaključna faza priprave podatkovne zbirke

Algoritma *linearna regresija* in *naključni gozdovi* v programskem jeziku *Python* za napovedovanje na vhodu pričakujeta številčne vrednosti. Ker nekateri atributi v naši zbirki podatkov vsebujejo tudi besedilne vrednosti, je bilo potrebno izvesti določene korake v smeri priprave podatkov. Vse t.i. kategorične atribute, ki vsebujejo več kot dve različni vrednosti, smo razširili v več stolpcev, podobno kot smo to storili z atributom *prostori stanovanja*. Atributi, ki smo jih razširili:

- občina,
- naselje,
- dejanska raba stanovanja,
- lega stanovanja v stavbi,
- posredovanje nepremičninske agencije,
- šifra katastrske občine.

Takšen pristop razširitve atributov se imenuje *kodiranje 'One Hot'* (ang. One-Hot Encoding) [42]. *Kodiranje 'One Hot'* je metoda predstavitve, ki

vzame vsako vrednost kategorije in jo spremeni v binarni vektor (glej sliko 4.10) velikosti enaki številu unikatnih vrednosti v kategoriji, kjer so vsi stolpci enaki nič, razen stolpec kategorije. Takšen pristop je eden od najpogostejših za obravnavo kategoričnih podatkov in je tudi zelo pogost pri besedilnih modelih. S takšnim pristopom se je dimenzija naše podatkovne zbirke močno povečala, iz 561 atributov smo podatkovno zbirko razširili na 1.963 atributov.

ID	Gender
1	Male
2	Female
3	Not Specified
4	Not Specified
5	Female

ID	Male	Female	Not Specified
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	1
5	0	1	0

Slika 4.10: Primer kodiranja 'One Hot'.

Slabost tega pristopa je izredno povečanje trajanja učenja algoritmov napovedovanja in časa dela, poveča se tudi poraba pomnilnika. Druga slabost je, da se lahko model preveč prilagodi podatkom (ang. overfitting). Tretji problem pa je dodajanje novih vrednosti, ki niso bile v podatkih za učenje, kar je lahko problematično v domenah, kjer se podatki neprestano spreminjajo.

Poglavje 5

Uporaba metod napovedovanja

Obstajata dve vrsti algoritmov podatkovnega rudarjenja za napovedovanje: regresijski in klasifikacijski algoritem [43]. Prvi napoveduje stalne vrednosti (zvezne), medtem ko drugi napoveduje diskretne vrednosti (razrede). Npr. napovedovanje vrednosti hiše v evrih je regresijski problem, medtem ko je napovedovanje ali je tumor maligni ali benigni, problem klasifikacije. Problem, ki ga rešujemo mi, je torej regresijski. Regresijski problem je tisti, pri katerem iz več atributov, ki opisujejo nek primer, napovedujemo vrednost ciljne zvezne spremenljivke. Regresijske napovedne modele gradimo iz učnih podatkov, ki vsebujejo atribute kot opis enega atributa (atributa, ki ga napovedujemo).

5.1 Linearna regresija

Linearna regresija je metoda, ki se uporablja za iskanje razmerja med odvisno spremenljivko in nizom neodvisnih spremenljivk [44]. Izraz 'linearnost' se v algebri nanaša na linearno razmerje med dvema ali več spremenljivkami. Enačbo te premice, ki se najbolje prilagaja vhodnim podatkovnim točkam, lahko zapišemo kot: `y = mx + b`, kjer je `b` pristranskost (lahko tudi začetna vrednost) in `m` naklon linije.

Linearni regresijski algoritem nam v bistvu daje najbolj optimalno vre-

dnost za pristranskost in naklon (v dveh dimenzijah). Isti koncept se lahko razširi tudi na primere, kjer imamo več kot dve spremenljivki. To imenujemo *linearna regresija za več spremenljivk* (ang. Multiple Linear Regression). Če npr. želimo napovedati vrednost hiše, ki temelji na podatku o lokaciji, številu spalnic, povprečnem dohodku prebivalcev v občini, letu izgradnje idr., je odvisna spremenljivka odvisna od več neodvisnih spremenljivk. Regresijski model, ki vključuje več spremenljivk pa lahko predstavimo kot: `y =` b_0 `+` m_1b_1 `+` m_2b_2 `+` m_3b_3 `+ ... +` m_nb_n. To je enačba hiper ravnine. Linearni regresijski model v dveh dimenzijah je ravna črta, v treh dimenzijah je ravnina in v več kot treh dimenzijah je hiper ravnina.

V naši nalogi je eden izmed ciljev tudi napovedati vrednost nepremičnine iz več različnih spremenljivk, torej je naš problem regresijski. Zato smo napovedni model gradili po principu linearne regresije za več spremenljivk. Za izgradnjo linearnega modela smo uporabili knjižnico za podatkovno rudarjenje `scikit-learn` [45].

Za izvajanje linearne regresije smo razvili funkcijo, ki iz podanih podatkov zgradi linearni model in izvede napovedovanje vrednosti. Za izvedbo omenjenega postopka potrebujemo podatke v 4 različnih zbirkah:

- učna množica podatkov neodvisnih spremenljivk - `X_train`,
- učna množica podatkov za spremenljivko, ki jo napovedujemo - `X_test`,
- testna množica podatkov neodvisnih spremenljivk - `y_train`,
- testna množica podatkov za spremenljivko, ki jo napovedujemo - `y_test`.

Več o razdelitvi podatkov smo opisali v poglavju 5.3. V prvem koraku ustvarimo model linearne regresije in nad modelom izvedemo učenje modela s podatki za učenje. Napovedi nad zgrajenim linearnim modelom pridobimo s klicem funkcije za napovedovanje `predict` nad testnimi podatki.

V nekaterih primerih gradnje napovednega modela smo ob ocenjevanju napovedi zaznali, da so bili nekateri napovedni modeli zgrajeni nad spremenljivkami, ki niso primerne. S funkcijo `OLS` smo zgradili statistični model in

s preverjanjem statistične značilnosti s pomočjo *p-vrednosti* zaznali, kateri atributi so tisti, ki so nepomembni. Več o postopku odstranjevanja nepomembnih atributov smo zapisali v poglavju 5.4. Rezultate napovedovanja in postopke testiranja smo predstavili v poglavju 6.

Testiranje smo izvajali nad različnimi podmnožicami zbirke podatkov. Za čim boljše razumevanje modela, izogibanje prevelikemu prileganju modela na podatke in razlago čim več variance s čim manj atributi oz. komponentami, smo uporabili *metodo glavnih komponent* (ang. Principal Component Analysis) (PCA), kjer so podrobnosti opisane v poglavju 5.5. Osnovno implementacijo linearne regresije smo nadgradili z apliciranjem metode *PCA* nad podatki, pred samim izvajanjem linearne regresije. V prvi fazi smo podatke pretvorili v decimalna števila ter izvedli standardizacijo podatkov [46]. Metoda *PCA* se izvaja nad skaliranimi podatki, zato je potreben tudi ta korak. Z uporabo `StandardScaler` smo standardizirali podatke na lestvico enote (srednja vrednost = 0 in varianca = 1), kar je tudi zahteva za optimalno delovanje mnogih algoritmov strojnega učenja. V *PCA* nas zanimajo komponente, ki maksimizirajo varianco. Nad skaliranimi podatki izvedemo metodo *PCA*, kateri nastavimo še število komponent, ki jih želimo (več o izbiri komponent v poglavju 5.5). Model prilagodimo nad učnimi podatki ter uporabimo pretvorbo nad učnimi in testnimi podatki. Nad temi podatki izvedemo napoved, kot je opisano zgoraj. Rezultate napovedovanja z metodo linearne regresije nad analizo glavne komponente smo predstavili v poglavju 6.

5.2 Naključni gozdovi

Naključni gozdovi (ang. Random Forest) [47] so fleksibilen in enostaven algoritem za uporabo v strojnem učenju. Napovedne točnosti tega algoritma so tipično med najboljšimi, saj tudi brez hiperparametrskih nastavitev daje dobre rezultate. Je eden izmed najpogosteje uporabljenih algoritmov, saj je precej preprost. Naključni gozdovi je nadzorovani učni algoritem, ki ustvari

gozd in ga naredi naključnega. 'Gozd', ki ga gradi, je ansambel odločitvenih dreves (ang. Decision Trees), ki se večino časa uči z metodo 'vreč' (ang. bagging). Splošna ideja vreče je, da kombinacija učnih modelov poveča skupni rezultat. Naključni gozd gradi več odločitvenih dreves in jih združuje skupaj, da dobi natančnejše in stabilnejše napovedi. Zelo velika prednost naključnega gozda je, da jo je mogoče uporabiti za klasifikacijske in regresijske probleme, ki tvorijo večino sedanjih sistemov strojnega učenja.

Medtem ko rastejo drevesa, naključni gozdovi dodajajo modelu dodatno naključnost. Namesto, da išče najpomembnejši atribut (ang. feature) med deljenjem vozlišča, išče najboljši atribut med naključno podmnožico atributov. Rezultat ima za posledico veliko raznolikost, ki na splošno privede do boljšega modela. Zato v naključnih gozdovih algoritem za delitev vozlišča upošteva samo naključno podmnožico funkcij. Drevesa se lahko zgradijo celo bolj naključno z dodatno uporabo naključnih pragov za vsak atribut, namesto iskanja najboljših pragov (kot običajna odločitvena drevesa). Primer odločitvenih dreves je npr. raziskovanje kam na potovanje. Vprašamo prijatelja, zberemo nasvete in podatek ali je bilo v redu. To je tipičen primer odločitvenega drevesa. Prijatelj je ustvaril pravila za vodenje naše odločitve glede na to, kaj bi moral priporočiti na podlagi odgovorov. Nato sprašujemo vse več in več prijateljev, da nam svetujejo in jim postavljamo nova vprašanja, da lahko dobimo nova priporočila. Nato izberemo kraje, ki nam jih prijatelji najbolj priporočajo. To pa je tipičen pristop algoritma naključnih gozdov. Druga velika kakovost naključnega algoritma je, da je zelo enostavno izmeriti relativni pomen vsakega atributa glede na napoved.

Napovedni model smo zgradili z metodo naključnih gozdov, ki rešuje regresijski problem. Podobno kot pri linearni regresiji, smo postopali tudi v tem primeru; za izgradnjo linearnega modela smo uporabili znano knjižnico za podatkovno rudarjenje `scikit-learn`, ki vsebuje implementacijo naključnih gozdov [48].

Za izvajanje metode naključnih gozdov smo razvili funkcijo, ki iz podanih podatkov zgradi napovedni model ter izvede napovedovanje vrednosti. Po-

dobno kot pri linearni regresiji, smo tudi v tem primeru potrebovali podatke v 4 različnih zbirkah.

Podrobnosti o razbitju podatkov smo opisali v poglavju 5.3. V prvem koraku gradnje ustvarimo model regresije naključnih gozdov in nad modelom izvedemo učenje modela s podatki za učenje. Napovedi nad zgrajenim modelom naključnih gozdov pridobimo s klicem funkcije za napovedovanje `predict` nad testnimi podatki. Model zgradimo s 100 drevesi (hiperparameter 'n_estimators'), ki jih algoritem gradi, preden sprejme največ glasov ali vzame povprečje napovedi. V splošnem večje število dreves povečuje učinkovitost in naredi napoved stabilnejšo, vendar tudi upočasnjuje izračun. Skozi postopek testiranja smo ugotovili, da se kvaliteta napovedi z gradnjo s 100 drevesi stabilizira, zato smo ostali na takšnem obsegu gradnje dreves. Parameter 'random_state' naredi izhod modela ponovljiv. Model bo vedno prinesel enake rezultate, če bo imel določeno vrednost parametra in če operira z enakimi podatki. Parameter 'oob_score' (imenovan tudi oob vzorčenje) je prečna validacija metod naključnega gozda. Pri tem vzorčenju se približno tretjina podatkov ne uporablja za usposabljanje modela in se lahko uporabi za vrednotenje njegove učinkovitosti. Ti vzorci se imenujejo 'vzorci vreč'. Pri naključnih gozdovih se prevelika prilagoditev načeloma ne bo zgodila, saj obstaja dovolj dreves in klasifikator ne bo preoblikoval drevesa. Glavna omejitev naključnih gozdov je, da lahko veliko število dreves naredi algoritem počasen in neučinkovit za napovedovanje v realnem času. Na splošno so ti algoritmi hitri za učenje, vendar pa počasnejši za napovedovanje. Natančnejša napoved zahteva sicer več dreves, kar ima za posledico počasnejši model. V večini aplikacij v realnem času pa je algoritem naključnih gozdov dovolj hiter. Algoritem naključnih gozdov se uporablja na različnih področjih, kot so bančništvo, finance, medicina, e-trgovina. Naključni gozdovi so v večini primerov hitri, enostavni in prilagodljivi, čeprav imajo svoje omejitve.

Z namenom zmanjšanja dimenzije atributov smo uporabili metodo *PCA*. Tudi v tem primeru smo postopali podobno kot pri linearni regresiji.

Rezultate napovedovanja z metodo naključnih gozdov smo predstavili v

poglavju 6.

5.3 Testiranje napovednih modelov - prečno preverjanje

Posvetili smo se tudi procesu izvajanja in testiranja napovedovanja. Napovedovanje smo izvajali z dvema metodama, ki rešujeta regresijske probleme - linearno regresijo in naključne gozdove. Podatke smo najprej ločili v dve množici; prva je ciljna spremenljivka, ki jo želimo napovedovati in druga, ki je vsebovala vse ostale atribute, ki jih model uporablja za napovedovanje. Za razdelitev podatkov v učno in testno množico smo uporabili *prečno preverjanje* (ang. Cross Validation) oz. *k-kratno prečno preverjanje* (ang. K-Folds Cross Validation) [49]. Podatke smo razdelili v `k` različnih podmnožic (ang. fold). Metoda naključno razdeli podatke v k (10) delov. Za vsak model zgradi svoj model na `k-1` podmnožici podatkovnega nabora in nato preizkusi model na *k-ti* podmnožici. V vsakem koraku shranimo podatke evalvacije. Postopek ponovimo *k-krat*, dokler vsaka od k podmnožic ne služi kot testni nabor podatkov. Povprečje k zabeleženih napak se imenuje *napaka prečne validacije* in služi kot merilo uspešnosti za model. S tem dosežemo, da postopek nad podatki večkrat ponovimo, kar pomaga pri pravilnem preverjanju učinkovitosti modela. S tem pristopom razdelimo podatke na učno in testno množico, nad katerima izvajamo grajenje napovednih modelov in napovedovanje vrednosti. Za izvajanje k-kratnega prečnega preverjanja smo uporabili implementacijo `KFold` iz znane knjižnice `scikit-learn`.

Pripravili smo tudi funkcijo, ki nad učnimi in testnimi podatki zažene obe metodi za napovedovanje. V vsakem koraku izvedemo še ničelno hipotezo, s katero želimo postaviti minimalno kvaliteto napovedi, več o tem smo pisali v poglavju 5.7.

5.4 Regresijska analiza

Za opis razmerja med nizom neodvisnih spremenljivk in odvisno spremenljivko uporabimo regresijsko analizo [50]. Rezultat regresijske analize je regresijska enačba, kjer koeficienti predstavljajo razmerje med vsako neodvisno in odvisno spremenljivko. Enačbo se lahko uporabi tudi za izdelavo napovedi. Regresijski model nam pomaga razumeti, kako so spremembe v napovednih vrednostih povezane s spremembami v odzivu - napovedovani vrednosti. Po učenju regresijskega modela se lahko prepričamo ali imamo nepristranske ocene in nato razložimo statistične rezultate.

Za vsak koeficient v regresijski analizi imamo na voljo podatek *p-vrednost*, ki nam pove, katera razmerja v modelu so statistično značilna in kakšna je narava razmerij [51]. Koeficienti opisujejo matematično razmerje med vsako neodvisno in odvisno spremenljivko. *p-vrednosti* za koeficiente kažejo, ali so razmerja statistično pomembna. Za vsako neodvisno spremenljivko *p-vrednost* preizkuša ničelno hipotezo, ali spremenljivka nima korelacije z odvisno spremenljivko.

p-vrednost se uporablja pri preizkušanju hipotez, ki nam pomaga podpreti ali zavrniti ničelno hipotezo. Manjša kot je *p-vrednost*, močnejši je dokaz, da zavrnemo ničelno hipotezo. Pri testiranju hipotez primerjamo *p-vrednost* z vrednostjo *alfa*, ki se jo izbere, ko se opravi test [52]. Stopnjo *alfa* določi raziskovalec in je povezana z ravnjo zaupanja. Stopnjo *alfa* dobimo tako, da od 100 % odštejemo našo stopnjo zaupanja, npr. če želimo, da bi bili prepričani v 95 %, določimo stopnjo *alfa* 5 % (100 % - 95 %). Izračunano *p-vrednost* nato primerjamo z izbrano stopnjo *alfa* 0,05 in se odločimo, ali bomo hipotezo zavrnili ali ne. Za majhne *p-vrednosti* (najbolj pogosto $\leq$ *0,05*) zavrnemo ničelno hipotezo. Torej, manjša je *p-vrednost*, bolj pomembni (ang. significant) so rezultati.

Pri izgradnji modela smo uporabili metodo `OLS` (ang. Ordinary Least Squares) [53]. *OLS* pomeni *običajni najmanjši kvadrati*, metoda *najmanjših kvadratov* pa pomeni, da poskušamo prilagoditi regresijsko linijo, ki bi zmanjšala kvadratni odmik od regresijske črte. Z metodo `OLS` zgra-

dimo model nad odvisno in neodvisno spremenljivko. Na podlagi *p-vrednosti* za posamezno spremenljivko iz podatkov o statistiki modela se odločimo, katera spremenljivka je pomembna (sprejemljiva stopnja napake je *0,05*); pomembna spremenljivka je tista, za katero velja, da je *p-vrednost* za spremenljivko $\leq$ *0,05*.

5.5 Metoda glavnih komponent - PCA

Metoda glavnih komponent (ang. Principle Component Analysis) (PCA) je eden od najpomembnejših algoritmov na področju podatkovne znanosti in je ena izmed najbolj priljubljenih metod za zmanjševanje dimenzionalnosti, ki se trenutno uporablja. PCA je nenadzorovana tehnika, ki je poleg zmanjšanja dimenzionalnosti podatkov namenjena tudi izločanju šuma in prikazu podatkov v dvo ali tro-dimenzionalnih vizualizacijah. PCA je linearna transformacija, ki podatke iz atributnega prostora, kjer so atributi med sabo lahko odvisni, preslika v prostor medsebojno neodvisnih atributov. Osnovni postopek za gradnjo glavnih komponent ne upošteva vrednosti razredne spremenljivke. V postopku konstrukcije te transformacije, podatke centriramo, izračunamo kovariančno matriko in poiščemo njene lastne vektorje in vrednosti. Lastni vektorji določajo nov koordinatni sistem, v katerega preslikamo podatke, lastne vrednosti pa deležu pojasnjene variance podatkov. Glavne komponente so torej tiste z največjimi lastnimi vrednosti. Tipično nas zanima le nekaj glavnih komponent, s katerimi na primer razložimo 90 % variance. Nove spremenljivke - komponente so urejene od najpomembnejše - to je tiste, ki pojasnjuje kar največ razpršenosti osnovnih podatkov - do najmanj pomembne - tiste, ki pojasnjuje najmanjši del razpršenosti opazovanih spremenljivk [54]. Cilj te metode je poiskati nekaj prvih komponent, ki pojasnjujejo čim večji del razpršenosti analiziranih podatkov. Metoda glavnih komponent torej zmanjša razsežnost podatkov, pri tem pa poizkuša izgubiti čim manj informacij.

Za izračun glavnih komponent smo uporabili funkcijo `PCA` iz znane knjižnice

`scikit-learn`. Število komponent smo določili s pomočjo grafa *scree plot*, in sicer smo mejo določili kot število atributov, katerih lastna vrednost je višja od *1*. Več o določitvi meje smo zapisali v poglavju 5.6, o sami implementaciji metode pa v podpoglavju 5.1 za linearno regresijo in podpoglavju 5.2 za naključne gozdove.

5.6 Določitev števila komponent za PCA

Število komponent za zmanjšanje dimenzije atributov z metodo *PCA* smo določili na podlagi izračuna lastnih vrednosti posameznega atributa ter iz grafa *scree plot*, ki prikazuje lastne vrednosti po atributih. Za izračun lastnih vrednosti smo uporabili knjižnico `NumPy`. Najprej smo izračunali kovariančno matriko nad standardiziranimi podatki neodvisnih spremenljivk s funkcijo `cov`. Iz kovariančne matrike smo s funkcijo `linalg.eig` izračunali lastne vrednosti in lastne vektorje.

Tabela 5.1: Prikaz izpisa lastnih vrednosti in deleža za vsako komponento za najpomembnejših 10 atributov iz zbirke najemni posli stanovanj.

Component	Total eigenvalue	% of Variance	Cummulative %
1	4,75	0,33	0,33
2	4,30	0,30	0,63
3	3,88	0,27	0,90
4	3,60	0,25	1,15
5	3,41	0,24	1,38
6	3,33	0,23	1,62
7	3,27	0,23	1,84
8	3,23	0,22	2,07
9	3,22	0,22	2,29
10	3,18	0,22	2,51
...	...	...	...
250	2,00	0,14	48,20
...	...	...	...

Z metodo *PCA* smo želeli model poenostaviti, da bi ga bolje razumeli in s čim manj komponentami razložiti čim več variance. Iz podatkov o lastnih vrednostih (glej tabelo 5.1) smo ugotovili, da 50 % variance razložimo z več kot 250 komponentami (odvisno od testne podmnožice podatkov, glej sliko 4.1 in 4.2), mi pa smo stremeli k razložitvi vsaj 90 % variance. Na podlagi grafa *scree plot* in izpisa s podatki o lastnih vrednostih (glej tabelo 5.1) smo se odločili, da bomo za vsako zbirko podatkov, nad katero bomo testirali napovedovanje, določili drugačno število komponent, in sicer smo mejo določili pri lastni vrednosti *1*. Zato smo se odločili, da za število komponent pri *PCA* vzamemo takšno število, kot je atributov z lastno vrednostjo višje od *1*.

5.7 Postopek evalvacije

V fazi testiranja in evalvacije napovedovanja smo preverili, ali je naš model učinkovit. Pri večini vrednotenj modelov želimo izračunati mero uspešnosti, ki nam pove ali je model učinkovit. Da bi se lažje odločili, ali je določen rezultat dober ali slab, smo izvedli primerjavo z ničelnim modelom (ang. null model), ki nam pove, kakšna je učinkovitost zelo preprostega napovednega modela [55]. Ničelni model je model zelo preproste oblike, ki ga poskušamo preseči. Dva najbolj značilna izbora ničelnih modelov sta model, kjer je ena sama konstanta (enaka napoved za vse situacije) ali samostojni model (ne zapisuje nobenega pomembnega razmerja ali interakcije med vhodi in izhodi). Ničelne modele uporabljamo za določitev spodnje meje učinkovitosti. Na primer, pri klasifikacijskem problemu bi ničelni model vedno vrnil najbolj priljubljeno kategorijo (ker je to najlažje za uganiti, saj je le-ta najmanj pogosto napačna izbira). Pri modelu zveznih vrednosti je ničelni model pogosto povprečje vseh vrednosti (ker ima to najmanj kvadratno odstopanje od vseh rezultatov). Čeprav je ničelni model preprost, je težko narediti tako dober kot tudi najboljši ničelni model. Vedno moramo biti prepričani, da je ničelni model, ki ga primerjamo, najboljši možen ničelni model. Mi smo za ničelni

model izbrali model, ki napoveduje vrednost nepremičnine na podlagi lokacije - občine in napove vrednost kot povprečno vrednost pogodbene najemnine oz. cene nepremičnine v občini. Ničelni model je za nas predstavljal najnižjo stopnjo kvalitete napovedi, ki pa smo jo želeli z gradnjo napovednih modelov preseči.

Pri merjenju kvalitete napovedi, smo uporabili različne metrike; ovrednotili smo povprečno kvadratno napako (MSE), efektivno vrednost napake (RMSE), povprečno absolutno napako (MAE), delež razložene variance (R2) ter prilagojen delež razložene variance (adjusted R2), kot so ovrednotili tudi ostali [18, 23, 24]. Omenjene metrike merjenja kvalitete napovedi smo računali v vsakem koraku k-kratnega prečnega preverjanja. Vmesne rezultate smo na koncu združili kot skupno povprečno oceno. Kvaliteto napovedi smo merili za naše napovedne modele (napovedovanje z linearno regresijo in naključnimi gozdovi) in rezultate primerjali z rezultati napovedi ničelnega modela.

MAE

MAE je povprečje absolutne napake (ang. Mean Absolute Error) med predvidenimi vrednostmi in opazovano vrednostjo [56]. MAE je linearna ocena, kar pomeni, da imajo vse posamezne napake v povprečju enako težo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{5.1}$$

RMSE

Mera RMSE je efektivna vrednost napake (ang. Root Mean Square Error) [57, 56], ki predstavlja vzorčni standardni odklon napake med predvidenimi in opazovanimi vrednostmi (imenovani napaka (ang. residuals)). Meri R^2 oz. adjusted R^2 merita količini variance v ciljni spremenljivki, ki jo lahko razložimo z našim modelom. To je dober pokazatelj, v nekaterih primerih pa smo bolj zainteresirani za količinsko napako v isti merski enoti spremenljivke.

V takih primerih moramo izračunati metriko, ki je povprečje preostalih modelov. Problem je, da so napake pozitivne in negativne, njihova porazdelitev pa mora biti precej simetrična. To pomeni, da bo njihovo povprečje vedno nič. Zato moramo poiskati druge metrike za kvantifikacijo povprečnih napak, na primer tako, da povprečimo kvadrate napak količine:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{5.2}$$

To je kvadratni koren povprečja kvadratnih napak, pri čemer je ocenjena vrednost v točki i, ki je opazovana vrednost v i in je velikost vzorca. RMSE ima enako mersko enoto kot y. Ta kvadratna metoda prispeva k doseganju bolj robustnih rezultatov, ki preprečujejo preklic pozitivnih in negativnih vrednosti napak oz. ta metrika v celoti prikazuje verjetni obseg napak. Izogiba se tudi uporabi absolutnih vrednosti napak, kar je v matematičnih izračunih zelo nezaželeno. Ko imamo več vzorcev, je rekonstruiranje porazdelitve napak z uporabo RMSE bolj zanesljivo in v primerjavi s povprečno absolutno napako (MAE), RMSE daje večjo težo in kaznuje velike napake. Na splošno pa je vrednost RMSE višja ali enaka (če so vse napake enake) kot MAE.

MSE

MSE je povprečna kvadratna napaka (ang. Mean Square Error) [57], ki je preprosto števec enačbe RMSE in je manj uporabljena metrika. Podobno kot za RMSE, se napake kvadrira in v glavnem bolj prizadene velike napake. To pomeni, da tudi če naš model zelo dobro pojasnjuje veliko večino razlik v podatkih, z nekaj izjemami, bodo te izjeme povečale vrednost RMSE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.3}$$

R^2

R^2 je delež razložene variance (ang. R-Squared) in je najpogosteje uporabljena metrika [56]. Omogoča, da razumemo odstotek variance v ciljni spremenljivki, ki jo razlaga model oz. nam pove, kako izbrana neodvisna spremenljivka razlaga variabilnost odvisne spremenljivke.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{5.4}$$

V imenovalcu ulomka je vsota kvadratov napak, ki bi jo naredili, če bi model napovedoval s povprečno vrednostjo odvisne spremenljivke učne množice. Za zelo dober model gre člen z ulomkom proti vrednosti 0, ocena R^2 pa zato proti vrednosti 1. Slabši modeli pa bodo imeli napako le malo manjšo od napake napovedi s povprečno vrednostjo. Ocena R^2 bo takrat šla proti *0*. Delež razložene variance ima zato pričakovano vrednost med *0* in *1*. Visoke vrednosti za R^2, torej takšne blizu *1*, so zelo zaželene, vendar so lahko uporabni tudi modeli z R^2 blizu *0*. Na primer, z modelom, ki napoveduje tečaj neke valute z $R^2 = 0{,}1$ bi obogateli, model z isto točnostjo za napoved količine padavin na m^2 pa nam skoraj ne koristi. Zato si je potrebno pomen vrednosti R^2 razlagati domeni primerno.

Adjusted R^2

R^2 prikazuje, kako dobro se napovedi prilagajajo krivulji ali liniji [58]. Tudi prilagojeni R^2 (ang. Adjusted R-Squared) kaže, kako dobro se napovedi prilagajajo krivulji ali liniji, vendar se prilagodi glede števila zapisov in spremenljivk v modelu. Če v model dodamo še več neuporabnih spremenljivk, se bo prilagojeni R^2 zmanjšal. Če dodamo več uporabnih spremenljivk, se bo prilagojeni R^2 povečal. Prilagojeni R^2 bo vedno manjši ali enak R^2. Pri delu z vzorci je potreben le R^2 oz. če imamo podatke iz celotne populacije, R^2 ni potreben.

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n - 1)}{n - k - 1} \right] \tag{5.5}$$

n je število zapisov v vzorcu podatkov, k pa število neodvisnih spremenljivk v modelu. Glavna razlika med R^2 in prilagojenim R^2 je ta, da R^2 predpostavlja, da vsaka posamezna spremenljivka razlaga spremembo odvisne spremenljivke. Prilagojeni R^2 pa pove odstotek razlike, ki ga razlagajo le neodvisne spremenljivke, ki dejansko vplivajo na odvisno spremenljivko.

Poglavje 6

Rezultati evalvacije in diskusija

Poglavje rezultatov je razdeljeno na dve podpoglavji. Vsako podaja rezultate napovedi iz različnih vrst podatkov. V prvem podpoglavju se osredotočamo na najemne posle za stanovanja, kjer se primerja uspešnost napovednih modelov, zgrajenih z linearno regresijo in naključnimi gozdovi iz različnih naborov podatkov. Uspešnost napovednih modelov je podana z metrikami uspešnosti napovedi, ki smo jih opisali v podpoglavju 5.7. Podobno kot za najemne posle stanovanj v prvem delu, smo v drugem podpoglavju predstavili še metrike uspešnosti napovedi za kupoprodajne posle stanovanj.

Napovedovanje *pogodbenih najemnin* za najeme stanovanj smo izvajali nad različnimi podmnožicami podatkov, kot je prikazano na sliki 4.1:

- razširjena čista zbirka - DS0: zbirka vsebuje vse podatke iz zbirke ETN in podatke iz različnih virov,
- izbor ETN - DS1: podmnožica vsebuje le atribute ETN iz celotne zbirke podatkov DS0,
- izbor LR - DS2: podmnožica vsebuje izbrane pomembne atribute iz celotne zbirke podatkov DS0 izbrane s pomočjo regresijske analize, ki smo jo opisali v poglavju 5.4,
- izbor atributov - DS3: podmnožica vsebuje izbrane atribute, ki najbolj vplivajo na odvisno (napovedovano) spremenljivko. Atribute smo iz

celotne zbirke podatkov izbrali s postopkom izbora atributov, ki smo ga opisali v poglavju 4.4,

- izbor ETN LR - DS4: podmnožica vsebuje izbrane pomembne atribute iz zbirke podatkov DS1 izbrane s pomočjo regresijske analize, ki smo jo opisali v poglavju 5.4,
- PCA nad vsemi - DS5: z metodo *PCA* smo zmanjšali dimenzijo zbirki DS0 na izbrano število komponent,
- PCA nad izbor ETN - DS6: z metodo *PCA* smo zmanjšali dimenzijo zbirki DS1 na izbrano število komponent,
- PCA nad izbor atributov - DS7: z metodo *PCA* smo zmanjšali dimenzijo zbirki DS3 na izbrano število komponent.

Tudi napovedovanje *pogodbenih cen* za kupoprodaje stanovanj smo izvajali nad različnimi podmnožicami (kot pri najemih stanovanj) kot je prikazano na sliki 4.2.

6.1 Napovedovanje pogodbenih najemnin za stanovanja

Prva zbirka podatkov, ki smo jo pripravili, vsebuje podatke o najemnih poslih za stanovanja (DS0) (glej sliko 4.1). Zbirka vsebuje podatke iz GURS podatkovne zbirke ETN, dodatnih podatkov iz SURS ter drugi. Napovedne modele smo gradili z metodo linearne regresije in naključnih gozdov, ki rešujeta regresijske probleme. Testiranje smo izvajali z metodo k-kratnega prečnega preverjanja in v vsakem koraku merili napovedno uspešnost posamezne metode z različnimi metrikami uspešnosti: MAE, MSE, RMSE, R^2, prilagojeni R^2. Za določanje spodnje meje učinkovitosti napovednih modelov, smo v primerjavo vključili še ničelni model, ki je predstavljal povprečno vrednost *pogodbene nejemnine* za stanovanja po občinah. Postopek priprave

zbirke podatkov za najeme stanovanj in testnih podmnožic je prikazan na sliki 4.1.

6.1.1 Uspešnost napovedi nad podmnožico DS2

Prva podmnožica podatkov DS2 (glej sliko 4.1) nad katero smo izvedli testiranje napovedovanja, je vsebovala izbrane pomembne atribute iz celotne zbirke podatkov DS0 (1.963 atributov) (glej sliko 4.1). Pomembne atribute smo izbrali s pomočjo regresijske analize, kot smo opisali v poglavju 5.4. Podmnožica je vsebovala 36.086 zapisov in 164 atributov. Iz tabele 6.1 rezultatov ugotovimo, da sta oba napovedna modela boljša kot ničelni model, torej sta oba modela napovedi učinkovita. Ugotovimo tudi, da je model zgrajen z naključnimi gozdovi (RF2) bolj učinkovit, saj je povprečna absolutna napaka (MAE) nižja za 16,30 € oz. 15,59 € v primeru efektivne vrednosti napake (RMSE). Prilagojen delež razložene variance (prilagojeni R^2) je pri naključnih gozdovih (RF2) 55 %, kar je občutno več kot pri ničelnem modelu (12 %) in boljše kot pri linearni regresiji (LR2), ki je 38 %.

Tabela 6.1: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS2.

	linearna regresija (LR2)	naključni gozdovi (RF2)	ničelni model
MAE	83,16	66,86	95,15
MSE	11.206,44	8.145,81	15.015,58
RMSE	105,83	90,24	122,52
R^2	0,38	0,55	0,17
prilagojeni R^2	0,38	0,55	0,12

6.1.2 Uspešnost napovedi nad podmnožico DS3

Druga podmnožica podatkov DS3 (glej sliko 4.1) nad katero smo izvedli testiranje napovedovanja, je vsebovala izbrane atribute, ki najbolje vplivajo

na odvisno spremenljivko *pogodbena najemnina vseh oddanih površin* in smo jih iz celotne zbirke podatkov DS0 (glej sliko 4.1) izbrali s postopkom izbora atributov, ki smo ga opisali v poglavju 4.4. Izbrali smo 945 atributov. Iz tabele 6.2 rezultatov testiranja napovedi nad izbranimi atributi razberemo, da je napoved nad podmnožico DS3 (glej sliko 4.1) nekoliko boljša, kot v prejšnjem primeru testiranja nad podmnožico DS2. Poprečno absolutno napako (MAE) smo zmanjšali za 2,59 € ter pri obeh napovednih modelih (linearna regresija (LR6) in naključni gozdovi (RF6)) za 1 % izboljšali prilagojen delež razložene variance (prilagojeni R^2). V primerjavi z ničelnim modelom smo tudi v tem primeru dobili boljše rezultate.

Tabela 6.2: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS3.

	linearna regresija (LR6)	naključni gozdovi (RF6)	ničelni model
MAE	80,64	64,27	95,15
MSE	10.697,21	7.677,07	15.015,58
RMSE	103,39	87,60	122,52
R^2	0,41	0,57	0,17
prilagojeni R^2	0,39	0,56	0,12

6.1.3 Uspešnost napovedi nad podmnožico DS4

V prejšnjih dveh primerih smo napovedi izvajali nad podatki iz zbirke ETN ter dodatnimi podatki, ki smo jih pridobili iz drugih virov (DS2 in DS3). V tem primeru smo preverili napovedovanje le z atributi (1.441) iz zbirke ETN (DS1) (glej sliko 4.1). Iz podmnožice podatkov DS1 iz zbirke ETN smo s pristopi, ki smo jih opisali v poglavju 5.4, izbrali pomembne atribute DS4 (glej sliko 4.1). Podmnožica je vsebovala 36.086 zapisov in 267 atributov. Iz tabele 6.3 rezultatov testiranja nad podatki iz zbirke ETN vidimo, da smo v tem primeru napoved še dodatno izboljšali. Prilagojeni delež razložene variance (prilagojeni R^2) se je pri linearni regresiji (LR4) in naključnih goz-

dovih (RF4) izboljšal, in sicer 40 % pri napovedovanju z linearno regresijo in 57 % pri naključnih gozdovih. Efektivna vrednost napake (RMSE) je pri obeh metodah skoraj enaka testiranju nad podmnožico DS3. Napovedovanje z naključnimi gozdovi (RF4) nad podmnožico DS4 (glej sliko 4.1) se izkaže za nekoliko boljše, kot pri prej omenjenih scenarijih napovedovanja. Razliko napovedne točnosti lahko razumemo na več načinov, npr. podatki iz ETN ne vsebujejo podatkov o stanju na trgu, vemo pa tudi, da vrednosti za *pogodbene najemnine* niso najbolj zaupljiv podatek. Kateremu napovednemu modelu je bolje zaupati, je na tej točki težko oceniti.

Tabela 6.3: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS4.

	linearna regresija (LR4)	naključni gozdovi (RF4)	ničelni model
MAE	81,20	63,64	95,15
MSE	10.767,97	7.620,65	15.015,58
RMSE	103,75	87,27	122,52
R^2	0,40	0,58	0,17
prilagojeni R^2	0,40	0,57	0,12

6.1.4 Uspešnost napovedi nad podmnožico DS5

S ciljem zmanjšanja dimenzije zbirke podatkov smo z metodo glavnih komponent (PCA) začetni zbirki podatkov DS0 (glej sliko 4.1) zmanjšali dimenzijo na 675 komponent (DS5) (glej sliko 4.1). Vrednost 675 komponent smo določili kot število atributov (od 1.963), ki imajo lastno vrednost višjo od *1.* Rezultati napovedovanja s takšnim pristopom so zbrani v tabeli 6.4. Iz rezultatov testiranja opazimo, da so napovedni modeli manj učinkoviti in točni, v primerjavi s prejšnjimi poskusi. Največje poslabšanje opazimo pri prilagojenem deležu razložene variance (prilagojeni R^2), vendar sta kljub temu oba napovedna modela boljša od ničelnega modela. Tudi v tem primeru je napovedovanje z naključnimi gozdovi (RF1) bolj učinkovito kot napovedovanje z

linearno regresijo (LR1).

Tabela 6.4: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS5.

	linearna regresija (LR1)	naključni gozdovi (RF1)	ničelni model
MAE	84,24	77,61	95,15
MSE	11.559,22	10.161,76	15.015,58
RMSE	107,48	100,74	122,52
R^2	0,36	0,44	0,17
prilagojeni R^2	0,32	0,41	0,12

6.1.5 Uspešnost napovedi nad podmnožico DS7

Podobno kot v prejšnjem primeru, smo postopali tudi tu. Podmnožici izbranih atributov DS3 (glej sliko 4.1) smo z metodo *PCA* zmanjšali dimenzijo na 608 komponent (DS7). Model smo sicer poenostavili, vendar napovedne učinkovitosti nismo izboljšali. Naključni gozdovi (RF5) dajejo manjšo absolutno napako (MAE) kot linearna regresija (LR5). Z linearno regresijo je prilagojen delež razložene variance (prilagojeni R^2) 35 % in efektivna vrednost napake (RMSE) 107,33 €. Z naključnimi gozdovi (RF5) je prilagojen delež razložene variance (prilagojeni R^2) 46 % in efektivna vrednost napake (RMSE) 97,35 €. Oba napovedna modela sta boljša od ničelnega modela.

6.1.6 Uspešnost napovedi nad podmnožico DS6

Dimenzijo množice podatkov iz atributov ETN (1.441) (glej sliko 4.1, zbirka DS1) smo z metodo *PCA* zmanjšali na 674 komponent (DS6) (glej sliko 4.1). Z napovedovanjem nad zmanjšano podmnožico podatkov iz atributov ETN (DS7) smo dobili nekoliko boljše napovedi kot v prejšnjih dveh primerih, ko smo zmanjševali dimenzijo. Z obema napovednima modeloma smo presegli ničelni model. Najboljše rezultate dobimo z napovedovanjem z metodo na-

Tabela 6.5: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS7.

	linearna regresija (LR5)	naključni gozdovi (RF5)	ničelni model
MAE	84,41	73,27	95,15
MSE	11.525,81	9.482,19	15.015,58
RMSE	107,33	97,35	122,52
R^2	0,36	0,48	0,17
prilagojeni R^2	0,35	0,46	0,12

ključnih gozdov (RF3), kjer je povprečna absolutna napaka (MAE) 67,73 € in je nižja kot pri linearni regresiji (84,06 €).

Tabela 6.6: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS6.

	linearna regresija (LR3)	naključni gozdovi (RF3)	ničelni model
MAE	84,06	67,73	95,15
MSE	11.497,50	8.247,13	15.015,58
RMSE	107,19	90,78	122,52
R^2	0,36	0,54	0,17
prilagojeni R^2	0,34	0,52	0,12

6.1.7 Interpretacija rezultatov

Najboljše rezultate napovedi vrednosti najemnin za stanovanja dobimo z grajenjem naključnih gozdov nad podatki iz zbirke ETN (DS4), kjer je povprečna absolutna napaka 63,64 €, efektivna vrednost napake 87,27 € in prilagojen delež razložene variance 57 %. Iz podatkov vemo, da je povprečna *pogodbena najemnina* stanovanj 243,52 €, v Ljubljani, kjer je sklenjenih največ poslov pa 300,77 €. Če primerjamo povprečno absolutno napako napovedi s povprečno najemnino v Ljubljani, lahko ocenimo, da je naš napovedni mo-

del uspešen, še posebej če primerjamo rezultate napovedi z ničelnim modelom. Na podlagi predpostavke, da so izhodiščni podatki za najemne posle slabše kvalitete in vrednost *pogodbena najemnina* celo nezaupljiva, je rezultat dober. Povprečna absolutna napaka nam tako zastavlja vprašanje, kaj v resnici predstavlja napaka 63,64 € pri vrednosti *pogodbene najemnine* stanovanja. Napako lahko razložimo s pomočjo dejavnikov, ki jih pri napovedi nismo upoštevali, npr. ali je stanovanje lažje ali težje dostopno z avtom, ali stanovanje vključuje tudi parkirišče, ali je avtobusno postajališče pogoste linije javnega prevoza v neposredni bližini stanovanja ali ne, kakšna je oddaljenost nepremičnine od vrtca, šole, trgovin, pekarne, banke, pošte, zdravstvenega doma, koliko zelenih površin je okoli stanovanja itd. Prav tako lahko del napake pripišemo tudi subjektivnemu mnenju posameznika, ali mu je stanovanje všeč ali ne, ali je stanovanje v mirni soseski za nekoga pomembnejše kot bližina centra kraja. Napovedna natančnost za prilagojeni delež razložene variance 40 % (linearna regresija (LR4)) oz. 57 % (naključni gozdovi (RF4)) je za nas zadovoljiv rezultat. Če primerjamo rezultate drugih poskusov napovedi opazimo, da se tudi ostale napovedi od najboljše ne razlikujejo bistveno, saj je potrebno rezultate gledati celostno, torej kakšna je napaka napovedi v primerjavi s prilagojenim deležem razložene variance modela. Kljub pričakovanjem, da bomo s poenostavitvijo modela s *PCA* izboljšali napovedno točnost, je nismo. Modela s tem pristopom nismo poenostavili, saj rezultati niso bili boljši. Tudi dimenzije nismo zmanjšali kot smo predvidevali, saj smo 50 % variance razložili z več kot 250 komponentami (odvisno od začetne podmnožice). V splošnem lahko zaključimo, da se je metoda naključnih gozdov izkazala za boljšo izbiro in bi s takšno metodo lahko postavili izhodiščne vrednosti za *pogodbene najemnine* stanovanj.

6.2 Napovedovanje pogodbenih cen za stanovanja

Druga zbirka podatkov, ki smo jo pripravili, je zbirka podatkov o kupoprodajnih poslih za stanovanja (DS0) (glej sliko 4.2). Zbirka vsebuje podatke iz GURS podatkovne zbirke ETN in dodatni podatki SURS ter drugi. Napovedne modele smo gradili z metodama linearna regresija in naključni gozdovi, ki rešujeta regresijske probleme. Testiranje smo izvajali z metodo k-kratnega prečnega preverjanja in v vsakem koraku merili napovedno uspešnost posamezne metode z različnimi metrikami uspešnosti: MAE, MSE, RMSE, R^2, prilagojena R^2. Za določanje spodnje meje učinkovitosti napovednih modelov, smo se oprli še na ničelni model, ki je predstavljal povprečno vrednost *pogodbene cene* za stanovanja po občinah.

Postopek priprave zbirke podatkov za kupoprodaje stanovanj in testnih podmnožic je prikazan na sliki 4.2.

6.2.1 Uspešnost napovedi nad podmnožico DS2

Najprej smo testiranje izvedli nad podmnožico podatkov DS2 (glej sliko 4.2), ki je vsebovala izbrane pomembne atribute iz celotne zbirke podatkov DS0 (2.508 atributov) (glej sliko 4.2). Pomembne atribute smo izbrali s pomočjo regresijske analize, kot smo opisali v poglavju 5.4. Podmnožica je vsebovala 60.165 zapisov in 507 atributov. Iz tabele 6.7 rezultatov ugotovimo, da sta oba napovedna modela bistveno boljša kot ničelni model, torej sta oba modela napovedi učinkovita. Napovedni model, zgrajen z metodo naključnih gozdov (RF2) je sicer bolj učinkovit kot model linearne regresije (LR2), vendar pa so te razlike med metodama v tem primeru bistveno manjše kot pri najemnih poslih. Ugotovimo, da je prilagojen delež razložene variance (prilagojeni R^2) pri linearni regresiji (LR2) 77 %, pri naključnih gozdovih (RF2) pa 84 %. Povprečna absolutna napaka (MAE) je pri linearni regresiji 14.496,75 € in pri naključnih gozdovih 10.986,15 €. Tudi ničelni model je pri kupoprodajnih podatkih bistveno bolj uspešen kot pri najemnih, vendar še vedno precej manj

uspešen tako od linearne regresije kot tudi naključnih gozdov.

Tabela 6.7: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS2.

	linearna regresija (LR2)	naključni gozdovi (RF2)	ničelni model
MAE	14.496,75	10.986,15	25.424,58
MSE	442.602.649,67	303.266.445,68	1.178.330.614,72
RMSE	21.033,20	17.404,96	34.322,72
R^2	0,77	0,84	0,39
prilagojeni R^2	0,77	0,84	0,37

6.2.2 Uspešnost napovedi nad podmnožico DS3

Tabela 6.8 vsebuje rezultate testiranja nad podmnožico podatkov DS3 (glej sliko 4.2), ki smo jo pridobili z iskanjem najpomembnejših atributov kot smo opisali v poglavju 4.4 in vsebuje 1.208 atributov. Opazimo, da se v tem primeru prilagojeni delež razložene variance (prilagojeni R^2) v primerjavi s prejšnjim scenarijem testiranja nad podmnožico DS2 (glej sliko 4.2) zmanjša. Model zgrajen z metodo naključnih gozdov (RF6) je še vedno učinkovitejši od linearne regresije (LR6). Oba modela sta tudi uspešnejša od ničelnega modela.

Tabela 6.8: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS3.

	linearna regresija (LR6)	naključni gozdovi (RF6)	ničelni model
MAE	15.150,36	12.089,60	25.424,58
MSE	506.369.626,61	380.791.172,55	1.178.330.614,72
RMSE	22.495,37	19.483,63	34.322,72
R^2	0,74	0,80	0,39
prilagojeni R^2	0,73	0,80	0,37

6.2.3 Uspešnost napovedi nad podmnožico DS4

Pri najemnih poslih se je napovedovanje nad podatki iz ETN izkazalo za najučinkovitejše, v tem primeru (kupoprodaje stanovanj) pa ni tako. Pri kupoprodajah stanovanj je napovedovanje *pogodbenih cen* nad podmnožico podatkov iz celotne zbirke podatkov DS2 (glej sliko 4.2) bolj učinkovito kot napovedovanje nad podmnožico podatkov, izbranimi iz zbirke ETN DS4 (glej sliko 4.2), ki smo jih izbrali s pristopi opisanimi v poglavju 5.4. Razlika med efektivno napako v primerjavi z napovedjo nad vsemi atributi je pri linearni regresiji (LR4) 3.207,86 €, v primeru naključnih gozdov (RF4) pa 3.708,04 €. Tudi prilagojen delež razložene variance je za 7 % boljša pri napovedovanju nad izborom pomembnih atributov iz celotne zbirke podatkov DS2 (glej sliko 4.2). Bistveno večjo učinkovitost lahko pripišemo dodatnim podatkom, ki prikazujejo dejansko stanje na trgu. To je pokazatelj, da je bilo smiselno dodatno vključiti podatke iz drugih virov, kar je tudi eden izmed doprinosov magistrske naloge.

Tabela 6.9: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS4.

	linearna regresija (LR4)	naključni gozdovi (RF4)	ničelni model
MAE	17.180,70	13.921,26	25.424,58
MSE	588.546.622,74	446.842.557,54	1.178.330.614,72
RMSE	24.241,06	21.113,00	34.322,72
R^2	0,70	0,77	0,39
prilagojeni R^2	0,70	0,77	0,37

6.2.4 Uspešnost napovedi nad podmnožico DS5

Z občutnim zmanjšanjem dimenzije podatkov iz 2.508 atributov DS0 (glej sliko 4.2) na 1.004 komponent DS5 (glej sliko 4.2) smo še vedno razložili 93,78 % variance, a pri napovedih dobili nekoliko slabše napovedi kot brez

zmanjšanja dimenzije z metodo *PCA*. Poskus napovedovanja nad zmanjšano dimenzijo se je pri napovedovanju *pogodbenih cen* za kupoprodaje stanovanj izkazal za zelo dobrega v primerjavi z napovedovanjem *pogodbenih najemnin*. Poslabšanje je bilo pri linearni regresiji (LR1) minimalno, pri naključnih gozdovih (RF1) pa nekoliko večje. Še vedno pa lahko trdimo, da so napovedi tudi v tem primeru zelo dobre, saj je prilagojen delež razložene variance 80 % in povprečna vrednost napake 12.646,95 €.

Tabela 6.10: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS5.

	linearna regresija (LR1)	naključni gozdovi (RF1)	ničelni model
MAE	14.815,20	12.646,95	25.424,58
MSE	460.790.950,93	366.187.231,58	1.178.330.614,72
RMSE	21.459,53	19.126,95	34.322,72
R^2	0,76	0,81	0,39
prilagojeni R^2	0,75	0,80	0,37

6.2.5 Uspešnost napovedi nad podmnožico DS7

Dimenzijo množice izbranih atributov DS3 (glej sliko 4.2) smo z metodo *PCA* zmanjšali na 894 komponent DS7 (glej sliko 4.2) in tako razložili kar 99,78 % variance. Vrednost za prilagojen delež razložene variance se je tudi v tem primeru nekoliko znižala. Obe napovedni metodi sta se izkazali za učinkoviti, saj so rezultati bistveno boljši kot za ničelni model. Tudi v tem primeru je metoda naključnih gozdov (RF5) boljša izbira od linearne regresije (LR5).

6.2.6 Uspešnost napovedi nad podmnožico DS6

Z zmanjšanjem dimenzije iz 2.017 atributov DS1 (glej sliko 4.2) na 1.002 komponent DS6 (glej sliko 4.2) smo dosegli najboljšo napovedno učinkovitost v primerjavi s prejšnjima testnima scenarijema, ko smo z metodo *PCA* poeno-

Tabela 6.11: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS7.

	linearna regresija (LR5)	naključni gozdovi (RF5)	ničelni model
MAE	15.404,71	12.769,71	25.424,58
MSE	517.160.778,65	410.017.722,39	1.178.330.614,72
RMSE	22.734,66	20.226,49	34.322,72
R^2	0,73	0,79	0,39
prilagojeni R^2	0,73	0,79	0,37

stavljali model. Izboljšava je tako za metodo linearne regresije (LR3) kot za naključne gozdove (RF3). Prilagojeni delež razložene variance za naključne gozdove je 81 % in povprečna absolutna napaka 12.252,16 €.

Tabela 6.12: Metrike uspešnosti napovedi napovednih modelov nad podmnožico DS6.

	linearna regresija (LR3)	naključni gozdovi (RF3)	ničelni model
MAE	15.346,16	12.252,16	25.424,58
MSE	487.042.476,29	356.018.233,11	1.178.330.614,72
RMSE	22.051,96	18.856,20	34.322,72
R^2	0,75	0,82	0,39
prilagojeni R^2	0,74	0,81	0,37

6.2.7 Interpretacija rezultatov

Pri napovedovanju *pogodbenih cen* za kupoprodaje stanovanj smo najboljše rezultate dobili z naključnimi gozdovi z napovedovanjem nad podmnožico podatkov DS2 (glej sliko 4.2), ki vsebuje pomembne atribute izbrane s pomočjo regresijske analize iz podmnožice DS0 (glej sliko 4.2), ki vsebuje podatke iz ETN in dodatne podatke iz različnih virov. Prilagojeni delež razložene variance (prilagojeni R^2) je 84 %, povprečna absolutna napaka (MAE) 10.986,15

€ in efektivna vrednost napake (RMSE) 17.404,96 €. Tudi linearna regresija je bila najuspešnejša nad to podmnožico podatkov, kjer je bil dosežen prilagojeni delež razložene variance 77 %, povprečna absolutna napaka 14.496,75 € in efektivna vrednost napake 21.033,20 €. V vseh primerih sta bili obe metodi učinkovitejši od ničelnega modela, najboljše rezultate pa daje metoda naključnih gozdov. Iz zbirke podatkov o kupoprodajnih poslih stanovanj smo ugotovili, da je povprečna pogodbena cena posla za stanovanje 78.840,53 €, v Ljubljani, kjer je sklenjenih največ poslov pa 112.866,37 €. Zato lahko rečemo, da je povprečna absolutna napaka 10.986,15 € sprejemljiva in razumna. Povprečno absolutno napako 10.986,15 € lahko razložimo kot vrednost, ki zajema dejavnike, ki jih nismo uporabili ali pa jih ne moremo ovrednotiti. Takšni dejavniki so npr. kakšna je dostopnost do stanovanja, oddaljenost do avtobusne postaje priljubljene linije javnega prevoza, kakšna je oddaljenost nepremičnine od vrtca, šole, trgovin, pekarne, banke, pošte, zdravstvenega doma, koliko je zelenih površin okoli stanovanja, lega stanovanja, ali ima stanovanje pogled na dvorišče, ali na zadnjo stran stavbe, ali jutranje sonce doseže vsaj kakšno okno stanovanja, ali je sonce zgolj popoldansko, ali je stanovanje na senčni strani stavbe itd. Dejavnikov, ki jih lahko pripišemo povprečni absolutni napaki napovedi 10.986,15 € je tako kar nekaj, zato lahko zaključimo, da so naši napovedni modeli zelo uspešni. Kot smo že ugotovili, je sicer metoda naključnih gozdov pri napovedovanju bila bolj uspešna in natančna, vendar tudi linearna regresija daje dobre rezultate. Z metodo *PCA* smo zmanjšali dimenzijo, kar se je v primeru napovedovanja *pogodbenih cen* za kupoprodaje stanovanj izkazalo za dobro, saj smo dobili zelo primerljive rezultate napovedi v primerjavi z rezultati napovedi nad nezmanjšano dimenzijo podatkov. To nakazuje na to, da smo s tem pristopom uspeli zmanjšati dimenzijo in tako poenostavili model.

6.3 Ovrednotenje rezultatov nad zbirko REN

Podatke, nad katerimi smo izvajali grajenje in testiranje napovednih modelov, smo pridobili iz različnih virov. Tekom priprave zbirk podatkov smo ugotovili, da imamo na voljo le majhen delež REN podatkov, zato smo le-te izločili. Kljub temu menimo, da je potrebna primerjava dejanske in napovedane vrednosti za *pogodbeno ceno* iz ETN z *vrednostjo nepremičnine* iz REN. Za kupoprodaje stanovanj smo imeli na voljo majhen delež (17,72 %) podatkov za *vrednost nepremičnine* iz REN, ki so izračunane na podlagi modelov vrednotenja (GURS). Kljub majhnemu deležu podatkov, smo izvedli dva poskusa evalvacije, kjer smo merili povprečno absolutno (MAE) in efektivno vrednost napake (RMSE):

- primerjava dejanskih - tržnih vrednosti za kupoprodaje stanovanj (*pogodbena cena* iz ETN) z ocenjeno vrednostjo nepremičnine (*vrednost nepremičnine* iz REN),

- primerjava napovedanih vrednosti (prek naših modelov vrednotenja) za *pogodbeno ceno* stanovanja z ocenjeno vrednostjo nepremičnine (*vrednost nepremičnine* iz REN).

S prvim primerjanjem smo ugotovili, da je povprečna absolutna napaka (oz. razlika) (MAE) med tržno vrednostjo stanovanja in ocenjeno vrednostjo 17.562,09 €. Velika razlika med tema vrednostma nakazuje na to, da se tržne in ocenjene vrednosti bistveno razlikujejo, kar potrjuje tudi vrednost efektivne vrednosti napake (RMSE), ki je 71.303,51 €. S primerjanjem tržnih in ocenjenih vrednosti smo ugotovili, da je v več primerih (53,86 %) *pogodbena cena* iz ETN večja kot *vrednost nepremičnine* iz REN. Drugo primerjavo smo izvedli med ocenjenimi vrednostmi in najučinkovitejšim pristopom za napovedovanje *pogodbenih cen* za kupoprodaje stanovanj (kot smo opisali v poglavju 6.2.1). Ugotovili smo, da so razlike v tem primeru manjše; z linearno regresijo je povprečna absolutna razlika 14.004,22 €, z naključnimi gozdovi pa 13.786,54 €. Manjše razlike so tudi pri efektivni vrednosti napake;

za linearno regresijo je 69.711,96 € in za naključne gozdove 69.474,25 €. Opazimo, da so razlike med linearno regresijo in naključnimi gozdovi majhne oz. zanemarljive glede na domeno napovedovanja - kupoprodaje stanovanj.

Razlike med tržno (ETN) in ocenjeno vrednostjo (REN, GURS modeli) nakazujejo na veliko razhajanje med vrednostmi. Če predpostavimo, da *pogodbeni ceni* iz zbirke ETN zaupamo, lahko razliko 17.562,09 € razložimo kot neupoštevanje dejanskega stanja trga in subjektivnih dejavnikov, ki so sicer upoštevani ob sklenitvi kupoprodaj stanovanj. Manjša povprečna absolutna razlika 13.786,54 € med napovedanimi (naši modeli vrednotenja) in ocenjenimi vrednostmi nakazuje na to, da napovedni modeli upoštevajo stanje na trgu, vseeno pa se nekoliko približajo vrednotenju GURS. Iz tega lahko zaključimo, da naš napovedni model predstavlja alternativo modelom vrednotenja. Vprašanje, ali napovedni modeli predstavljajo izboljšavo vrednotenju GURS na podlagi modelov vrednotenja, pa prepuščamo domenskim ekspertom.

6.4 Diskusija

Iz različnih virov smo ustvarili dve novi zbirki podatkov - najemni posli stanovanj in kupoprodajni posli stanovanj. Glavni vir podatkov je bila zbirka ETN, dodatne podatke smo pridobili iz portala SI-STAT (SURS), razdalje med kraji smo pridobili prek Google storitev itd. Nad vsako zbirko podatkov smo izvedli več analiz ter uporabili različne pristope za čiščenje podatkov - poiskali in odstranili smo tiste vrednosti, ki so izstopale (osamelci), vstavili manjkajoče vrednosti ter iz vsake zbirke podatkov izbrali najpomembnejše atribute. Nad vsako zbirko podatkov smo izvedli napovedovanje z gradnjo napovednih modelov in vsak model tudi ovrednotili. Za postopek testiranja smo uporabili k-kratno prečno preverjanje, saj smo želeli nepristransko oceniti uspešnost našega modela. V vsakem koraku testiranja smo izračunali tudi različne metrike za merjenje uspešnosti in kvalitete napovednega modela. Prav tako smo rešili tudi problem nepomembnih atributov in sicer z

izračunom *p-vrednosti* za atribute in odstranitvijo tistih, ki niso pomembni. S testiranjem napovednih modelov nad različnimi nabori podatkov iz posamezne zbirke podatkov smo najprej določili najnižjo mero uspešnosti naših napovednih modelov - ničelni model, ki je bil naša mera za učinkovitost napovedi - ali je naš model napovedi uspešen ali ne.

Nad zbirko 'najemni posli stanovanj' smo napovedovali vrednost *pogodbena najemnina vseh oddanih površin* in nad zbirko 'kupoprodajni posli stanovanj' vrednost *pogodbena cena*. V obeh primerih smo torej napovedovali vrednost, ki pomeni 'denar'. Tako smo zajeli dva pogosta področja iz trga nepremičnin - najem stanovanj in prodaja stanovanj. Skozi gradnjo in testiranje napovednih modelov smo želeli ugotoviti, katera metoda daje najboljše rezultate ter ali dodatni podatki, ki smo jih pridobili, lahko pomagajo pri napovedih z gradnjo napovednih modelov.

V prvi fazi smo tako napovedovali vrednost *pogodbene najemnine* za najemne posle stanovanj. Izkazalo se je, da smo najboljše napovedne rezultate dosegli z metodo naključnih gozdov, z gradnjo napovednih modelov iz podatkov iz ETN. Doseženi prilagojeni delež razložene variance modela je 60 %, kar pomeni delež odzivne spremenljivke (pogodbene najemnine), ki je razložen z napovednim modelom. Absolutna povprečna napaka napovedi je 63,64 € in efektivna vrednost napake 87,27 €. Če se osredotočimo na najeme stanovanj v Ljubljani, kjer je povprečna najemnina stanovanja 243,52 € (zajeto iz naše zbirke podatkov) je povprečna napaka 63,64 €, sprejemljiva napaka. Poudariti je potrebno tudi dejstvo, da so podatki s katerimi smo operirali precej nezaupljivi, še posebej vrednost *pogodbena najemnina*, na kar so nas opozorili na GURSu in domenski eksperti. 63 € pri najemnini niti ni tako velika napaka, še posebej če vemo, da so posli običajno prijavljeni z nižjo najemnino kot je v resnici višina najemnine. Napako 63 € je mogoče razložiti tudi kot subjektivno mnenje, ali je stanovanje na sončni ali senčni strani stavbe, ali stanovanje vključuje tudi parkirišče (tega podatka nismo imeli na voljo), ali nam je stanovanje sploh všeč, kakšna je oddaljenost do najbližje trgovine, pekarne, šole, javnega prevoza, zdravstvenega doma ipd. Napako

lahko tako pripišemo subjektivnemu mnenju posameznika, ki pa vedno obstaja. Praktično točne napovedi z napako 0 € niti ne bi mogli doseči, sicer bi to pomenilo preveliko prileganje modela (ang. overfitting). V primerjavi z ničelnim modelom za najemne posle stanovanj lahko potrdimo, da je naš model napovedi učinkovit in zaupljiv ter tudi dovolj natančen za domeno napovedovanja vrednosti *pogodbene najemnine* za stanovanja.

Napovedovali smo tudi vrednost *pogodbena cena* za kupoprodaje stanovanj. To so dejanski nakupi stanovanj. Tudi v tem primeru smo testirali nad različnim naborom podatkov. Izkazalo se je, da so bili najboljši rezultati napovedani z modelom, zgrajenim z metodo naključnih gozdov nad podatki, ki so vsebovali izbrane pomembne atribute s pomočjo regresijske analize iz celotne zbirke podatkov za kupoprodajne posle stanovanj. Ta nabor podatkov je vseboval poleg podatkov iz ETN tudi dodatne podatke, ki smo jih pridobili iz različnih virov, kar je potrditev, da je naša ideja o vključitvi dodatnih podatkov v nabor podatkov pozitivno vplivala na kvaliteto napovedi. Prilagojen delež razložene variance je tako kar 84 %, povprečna absolutna napaka 10.986,15 € in efektivna vrednost napake 17.404,96 €. Približno absolutno napako 11.000 € pri nakupu stanovanja npr. v Ljubljani, kjer je bilo izvedenih največ kupoprodajnih poslov za stanovanja in je povprečna prodajna cena 112.866,37 €. To si lahko razlagamo kot sprejemljivo napako, ki je 'manevrski prostor' za posameznikovo subjektivno oceno stanovanja. Nekomu je stanovanje všeč, drugemu ni, bodisi zaradi lokacije, lege stanovanja v stavbi, bližine železnice ali prometnejše ceste, oddaljenosti do najbližje trgovine, pekarne ali lekarne itd. Tudi v primeru kupoprodaj stanovanj lahko trdimo, da napovedi z napako 0 € ne moremo doseči, saj bi to lahko pomenilo preveliko prileganje modela podatkom. V primerjavi z ničelnim modelom za kupoprodajne posle stanovanj lahko potrdimo, da je naš model napovedi učinkovit in zaupljiv ter tudi dovolj natančen za napovedovanje vrednosti *pogodbene cene* za stanovanja.

Za celovito evalvacijo napovednih modelov za kupoprodajne posle stanovanj, bi potrebovali več REN podatkov. Ker smo imeli na voljo le majhen

delež (17,72 %) REN podatkov, smo napovedne modele primerjali le s tistimi ocenjenimi *vrednostmi nepremičnin* iz REN, ki smo jih imeli na voljo. Postopek evalvacije kupoprodajnih stanovanj smo opisali v poglavju 6.3. Za napovedovanje vrednosti najemnin za stanovanja pa bi morali dobiti dejanske vrednosti najemnin s trga nepremičnin, za katere se sklepajo posli, a realno predvidevamo, da to ne bi bilo mogoče. Še najbližje temu bi bilo, če bi anonimno zbirali podatke o najemninah in jih primerjali z našimi napovedmi.

Poglavje 7

Sklepne ugotovitve in zaključek

7.1 Povzetek

Izhodišče za gradnjo učinkovitega in uspešnega napovednega modela je kvalitetna zbirka podatkov, ki ji tudi zaupamo. Pridobivanju in pripravi podatkov smo namenili več kot 80 % časa. Tako smo ustvarili širok nabor podatkov, ki so bili pridobljeni iz različnih virov. Širok nabor podatkov nam je omogočil, da smo lahko sestavili dve zelo široki zbirki podatkov, nad katerima smo nato izvajali postopke čiščenja podatkov in izbora pomembnih atributov. S kakovostno pripravo podatkov smo izpolnili pomemben predpogoj za gradnjo učinkovitih modelov napovedovanja vrednosti nepremičnin. Z gradnjo napovednih modelov iz različnega nabora atributov iz zbirk podatkov, smo naredili tudi primerjavo med napovednimi modeli zgrajenih brez in z dodatnimi podatki. Z izborom pomembnih atributov se je izkazalo, da so dodatni podatki, ki smo jih pridobili iz različnih virov, pomembni pri razlaganju odvisne spremenljivke - *pogodbena najemnina* oz. *pogodbena cena*. Ugotovili smo tudi, da so naši napovedni modeli uspešni, še posebej dobre rezultate dajejo z napovedovanjem *pogodbenih cen* za kupoprodaje stanovanj.

7.2 Prispevki

V okviru naše magistrske naloge smo se osredotočili na razvoj metodologije za pripravo podatkov, primernih za izgradnjo napovednega modela, ki je sestavljena iz več metod in postopkov, t.j. zajem smiselnih podatkov iz različnih virov, analiza podatkov, čiščenje podatkov (iskanje in odstranjevanje osamelcev, vstavljanje manjkajočih vrednosti), izbor pomembnih atributov za napovedovanje vrednosti nepremičnin. Eden izmed glavnih prispevkov naloge je pristop, kako s postopki zbiranja podatkov iz različnih virov, analiziranja in združevanja podatkov, ustvariti dve novi zbirki podatkov. Kljub temu, da je bil del podatkov (še posebej najemni posli) v začetku zelo neurejen in nezaupljiv, smo s postopki čiščenja ustvarili bolj zaupljivi zbirki podatkov. Zbirki vključujeta tudi podatke, ki posredno vplivajo na vrednost nepremičnin in prikazujejo razmere na trgu. Opisane pristope za čiščenje podatkov bi tako lahko aplicirali tudi nad podatke iz drugih domen. Strokovni prispevek naloge je tudi razvoj mehanizma za napovedovanje vrednosti nepremičnin z metodami podatkovnega rudarjenja nad realnimi podatki. Poleg tega smo opisali študijo izbora ključnih atributov, ki pomembno vplivajo na vrednost nepremičnine [5, 8, 9, 14, 15] in posledično na točnost napovednega modela. Z apliciranjem metod podatkovnega rudarjenja nad realnimi podatki o nepremičninah in drugimi podatki, smo prikazali možnost uporabe teh metod za namene napovedovanja vrednosti nepremičnin, ki jih je mogoče aplicirati tudi nad druge vrste podatkov o nepremičninah, npr. kupoprodaje hiš, zemljišča itd. S takšnim pristopom napovedovanja vrednosti nepremičnin je mogoče iskati tiste, že sklenjene posle, ki jim potencialno ne moremo zaupati glede na vrednost *pogodbene najemnine* oz. *pogodbene cene*.

7.3 Možnosti za nadaljnje delo

Kljub vzpodbudnim rezultatom napovednih modelov bi bilo smiselno v zbirko podatkov, še zlasti za napovedovanje *pogodbenih cen* za kupoprodaje stanovanj, vključiti tudi podatke iz zbirke REN. Vključitev podatkov REN smo

sicer prvotno predvidevali, a smo jih zaradi majhnega deleža izpustili, kot smo opisali v poglavju 4.1.2. Na podlagi podatkov iz REN bi lahko preverili, za koliko ti podatki izboljšajo učinkovitost napovednih modelov. Predlagamo tudi preveritev, kako so napovedni modeli uspešni pri napovedovanju nad svežimi podatki - podatki, ki jih še nimamo. Z večjo količino podatkov iz REN, bi lahko napovedane kupoprodajne cene za stanovanja primerjali tudi z GURS ocenami vrednosti stanovanj ter tako dodatno ocenili, kako uspešni so napovedni modeli v primerjavi z modeli vrednotenja, ki jih uporabljajo na GURSu. Razviti pristopi za čiščenje podatkov, izbor atributov ter uporaba metod napovedovanja omogočajo odprte možnosti za razširitev napovedovanja vrednosti, tudi za druge vrste delov stavb - najemne in kupoprodajne posle ter tudi za napovedovanje vrednosti *pogodbenih cen* za nakup zemljišča. Ena izmed izboljšav oz. nadgradenj bi bila tudi ta, da bi napovedne modele za napovedovanje vrednosti ponudili v uporabo kot aplikacijo oz. storitev za cenitev nepremičnin, ki na podlagi podanih parametrov napove vrednost za *pogodbeno najemnino* ali *pogodbeno ceno* stanovanja. Z implementiranjem drugih metod napovedovanja, npr. nevronske mreže, odločitvena drevesa, metoda najbližjih sosedov, bi lahko preverili ali je katera druga metoda za napovedovanje bolj primerna in učinkovita kot trenutno uporabljeni metodi. Napovedi iz večih metod napovedovanja vrednosti, bi z uporabo metode zlaganja lahko (ang. stacking) združevali napovedi večih modelov. Smiselno bi bilo raziskati možnosti za pospešitev gradnje naključnih gozdov ter možne optimizacije uporabljenih metod, še posebej za trenutni - linearno regresijo in naključne gozdove.

7.4 Zaključek

V magistrski nalogi smo razvili napovedni model, ki je sposoben napovedati vrednosti nepremičnin - za najeme stanovanj pogodbeno najemnino in za kupoprodaje stanovanj pogodbeno ceno. Napovedni model je zgrajen iz podatkov o nepremičninah in drugih podatkov, ki upoštevajo tudi stanje na

trgu. Tak pristop napovedovanja vrednosti nepremičnin predstavlja alternativo množičnemu vrednotenju nepremičnin, s katerim se ukvarja GURS, le da naš pristop upošteva stanje trga. Na GURSu, na podlagi osnovnih podatkov o nepremičnini s pomočjo različnih modelov in vrednostnih con, določijo vrednost nepremičnine. Za vsako vrsto nepremičnine je zgrajen svoj model, ki vsebuje tudi nekatere podatke o nepremičnini, npr. podatke o lokaciji. Tak pristop se je v praksi izkazal le kot približek dejanski vrednosti nepremičnine, hkrati pa je tak pristop nepriročen, saj je postopek zapleten, modeli se ne prilagajajo avtomatsko na druge dejavnike, kot so npr. dvig trošarin, povprečna mesečna plača itd. Po naših ugotovitvah, naš pristop napovedovanja vrednosti nepremičnin rešuje večino težav, ki jih pristop množičnega vrednotenja ima, hkrati pa ugotavljamo, da tak pristop z upoštevanjem nekaterih tržnih zakonitosti deluje celo bolje. Naš napovedni model poleg podatkov o nepremičninah upošteva tudi podatke iz drugih virov. Glavni vir podatkov nam je predstavljala zbirka ETN, ki smo jo razširili z dodatnimi podatki, ki posredno vplivajo na vrednost nepremičnine. Naš cilj je bil združiti podatke iz različnih virov, zgraditi novi zbirki podatkov ter razviti metodologijo, ki z različnimi pristopi omogoča čiščenje podatkov. Prikazali smo uporabo različnih tehnik za čiščenje podatkovnih zbirk - iskanje in odstranjevanje ekstremnih vrednosti ter vstavljanje manjkajočih vrednosti. Z izborom najpomembnejših atributov smo potrdili, da dodatni podatki na razlago vrednosti *pogodbene najemnine* oz. *pogodbene cene* vplivajo bistveno bolje kot nekateri podatki iz ETN.

Opisane ideje in razvite postopke za zbiranje, pripravo in čiščenje podatkovnih zbirk je mogoče aplicirati tudi na druge vrste nepremičnin, kot smo predlagali v poglavju 7.3. Dokazali smo, da lahko z gradnjo napovednih modelov z metodami podatkovnega rudarjenja precej natančno napovedujemo vrednosti nepremičnin - *pogodbene najemnine* in *pogodbene cene* za stanovanja.

Dodatek A

Opis podatkov ETN

Prilagamo opis strukture podatkov Evidence trga nepremičnin, ki smo ga pridobili prek portala *e-prostor*, ki ga upravlja GURS [59].

Tabela A.1: Šifranti.

podatek	opis podatka
FEATUREID	Tehnični atribut, ki je sestavljen iz imena servisa in naključne številke in ni primeren za kakršnokoli povezovanje.
ID	ID številka šifranta
Numerična vrednost	Numerična vrednost šifre
Opis	Opis šifre

Tabela A.2: Kupoprodajni posli.

podatek	opis podatka
FEATUREID	Tehnični atribut, ki je sestavljen iz imena servisa in naključne številke in ni primeren za kakršnokoli povezovanje.
ID Posla	Povezovalno polje, na katerega so vezani vsi podatki posameznega posla. MINUS pred nekaterimi številkami pomeni, da so bili podatki o poslu pretvorjeni iz stare strukture, ki je veljala za posle, sporočene do 30. 6. 2013.

Se nadaljuje na naslednji strani

Tabela A.2 – *Kupoprodajni posli - nadaljevanje.*

podatek	opis podatka
Vrsta kupoprodajnega pravnega posla	Vrsta kupoprodajnega posla: Podatek o vrsti sklenjenega kupoprodajnega pravnega posla po načinu prodaje.
Datum uveljavitve	Datum uveljavitve posla oz. datum vpisa posla v bazo.
Datum sklenitve pogodbe	Datum sklenitve pogodbe: Datum podpisa kupoprodajne pogodbe obeh pogodbenih strank. Če datuma podpisa obeh pogodbenih strank nista istovetna, se upošteva datum podpisa stranke, ki je podpisala pogodbo zadnja.
Pogodbena cena	Pogodbena cena: Cena, ki je navedena v pogodbi. V primeru pogodbe o finančnem lizingu nepremičnine podatek o pogodbeni ceni ne vključuje stroškov financiranja.
Vključenost DDV	Podatek o tem, ali pogodbena cena vključuje davek na dodano vrednost.
Stopnja DDV	Podatek o stopnji, po kateri je bil obračunan davek na dodano vrednost.
Datum izteka lizinga	Podatek o tem, kdaj bo v skladu s pogodbo sklenjen lizing iztekel (ekvivalent datumu prenehanja najema).
Datum prenehanja lizinga	Podatek o tem kdaj se je sklenjen lizing dejansko zaključil (ekvivalent datumu zaključka najema).
Opombe o pravnem poslu	Opombe.
Posredovanje nepremičninske agencije	Ali je v poslu med pogodbenimi strankami posredovala nepremičninska agencija?

Tabela A.3: Kupoprodajni posli - deli stavb.

podatek	opis podatka
FEATUREID	Tehnični atribut, ki je sestavljen iz imena servisa in naključne številke in ni primeren za kakršnokoli povezovanje.
ID Posla	Povezovalno polje, na katerega so vezani vsi podatki posameznega posla. MINUS pred nekaterimi številkami pomeni, da so bili podatki o poslu pretvorjeni iz stare strukture, ki je veljala za posle, sporočene do 30. 6. 2013.

Se nadaljuje na naslednji strani

Tabela A.3 – *Kupoprodajni posli - deli stavb - nadaljevanje.*

podatek	opis podatka
Šifra KO	Podatek o oznaki katastrske občine, ki je administrativna enota za vodenje podatkov zemljiškega katastra in katastra stavb. Šifra katastrske občine enolično določa katastrsko občino v Republiki Sloveniji.
Ime KO	Ime katastrske občine, kot je evidentirano v zemljiškem katastru. Ime katastrske občine je poimenovanje posamezne katastrske občine in ni enolična oznaka.
Občina	Podatek iz registra prostorskih enot o imenu občine, v kateri se nahaja pretežno število parcel, stavb oziroma delov stavb, ki so bile predmet posameznega pravnega posla.
Številka stavbe	Skupaj s šifro katastrske občine identifikacijska oznaka stavbe, ki enolično označuje stavbo.
Številka dela stavbe	Skupaj s šifro katastrske občine in številko stavbe identifikacijska oznaka dela stavbe, ki enolično označuje del stavbe. Vsaka stavba ima vsaj en del stavbe.
Parcelna številka za geolokacijo	Parcelna številka za geolokacijo.
Evidentiranost dela stavbe	Vrednost šifranta "REN status".
Naselje	Kraj: Podatek kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Ulica	Ulica: Podatek kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Hišna številka	Hišna številka: Podatek kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Dodatek HŠ	Dodatek k hišni številki: Podatek kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Številka stanovanja ali poslovnega prostora	Številka stanovanja ali poslovnega prostora.
Vrsta dela stavbe	Podatek o vrsti dela stavbe, glede na namen kupoprodaje za katerega sta se dogovorili pogodbeni stranki in za katerega velja pogodbena cena.

Se nadaljuje na naslednji strani

Tabela A.3 – *Kupoprodajni posli - deli stavb - nadaljevanje.*

podatek	opis podatka
Leto izgradnje dela stavbe	Podatek o letu, ko je bila stavba, v kateri je del stavbe, ki je predmet kupoprodajnega pravnega posla, zgrajena.
Stavba je dokončana	Podatek o tem ali je stavba dokončana ali ne.
Gradbena faza	Podatek o tem, v kateri gradbeni fazi je izgradnja stavbe, če je predmet kupoprodajnega pravnega posla del stavbe v nedokončani stavbi.
Primarni-sekudarni trg	Podatek, ali je bil prodan rabljen ali nerabljen del stavbe.
Prodana površina	Površina dela stavbe, ki je predmet kupoprodajnega pravnega posla in za katero velja pogodbena cena, izražena v m^2 brez uporabe korekcijskih faktorjev.
Prodani solastniški delež dela stavbe	Podatek, kolikšen delež dela stavbe je bil prodan, kadar predmet kupoprodajnega pravnega posla ni bila celotna stavba ali del stavbe, izražen z ulomkom (npr. 1/2, 12/40, ...).
Prodana neto tlorisna površina dela stavbe	Neto tlorisna površina dela stavbe – v m^2.
Prodana uporabna površina dela stavbe	Uporabna površina dela stavbe – v m^2.
Nadstropje dela stavbe	Številka nadstropja, v katerem se nahaja del stavbe, ki je predmet kupoprodajnega pravnega posla.
Število zunanjih parkirnih mest	Število parkirnih mest na prostem.
Atrij	Ali je stanovanje atrijsko.
Površina atrija	Površina atrija – v m^2.
Opombe o nepremičnini	Opombe.
Dejanska raba dela stavbe	Šifra in naziv dejanske rabe.

Se nadaljuje na naslednji strani

Tabela A.3 – *Kupoprodajni posli - deli stavb - nadaljevanje.*

podatek	opis podatka
Lega dela stavbe v stavbi	Izpis lege dela stavbe v stavbi.
Število sob	Število sob v delu stavbe.
Površina dela stavbe	Površina dela stavbe v m^2.
Uporabna površina dela stavbe	Uporabna površina dela stavbe v m^2.
Prostori stanovanja	Našteti so prostori stanovanja, med seboj ločeni s pokončno črto '|'.

Tabela A.4: Najemni posli.

podatek	opis podatka
FEATUREID	Tehnični atribut, ki je sestavljen iz imena servisa in naključne številke in ni primeren za kakršnokoli povezovanje.
ID Posla	Povezovalno polje, na katerega so vezani vsi podatki posameznega posla.
Vrsta najemnega pravnega posla	Vrsta najemnega pravnega posla: Podatek o vrsti sklenjenega najemnega pravnega posla po načinu najema glede na vrsto najemnine (tržna ali netržna).
Datum uveljavitve	Datum uveljavitve posla oz. datum vpisa posla v bazo.
Datum sklenitve pogodbe	Datum sklenitve pogodbe.
Pogodbena najemnina vseh oddanih površin	Pogodbena najemnina vseh oddanih površin: Podatek o višini najemnine iz pogodbe, na mesec ali na kvadratni meter oddane površine na mesec.
Obratovalni stroški	Vključenost obratovalnih stroskov: Podatek o tem, ali najemnina vključuje obratovalne stroške ali ne.
Vključenost DDV	Vključenost DDV.
Stopnja DDV	Stopnja DDV.
Datum začetka najema	Podatek o tem, kdaj začne sklenjeno najemno razmerje veljati (datum določen ob sklenitvi pogodbe).

Se nadaljuje na naslednji strani

Tabela A.4 – *Najemni posli - nadaljevanje.*

podatek	opis podatka
Datum prenehanja najema	Podatek o tem, kdaj bo sklenjeno najemno razmerje prenehalo veljati (datum določen ob sklenitvi pogodbe).
Čas najema	Podatek o tem, ali je najemno razmerje sklenjeno za določen ali za nedoločen čas (datum določen ob sklenitvi pogodbe).
Trajanje najema	Razlika med datumom prenehanja najema in datumom začetka najema, izražena na mesec natančno, če najemo razmerje ni sklenjeno za nedoločen čas.
Datum zaključka najema	Podatek o tem, kdaj je najemno razmerje dejansko prenehalo veljati (datum znan ob zaključku najemnega razmerja).
Opombe o pravnem poslu	Opombe.
Posredovanje nepremičninske agencije	Podatke posredovala nepremičninska družba.

Tabela A.5: Najemni posli - deli stavb.

podatek	opis podatka
FEATUREID	Tehnični atribut, ki je sestavljen iz imena servisa in naključne številke in ni primeren za kakršnokoli povezovanje.
ID Posla	Povezovalno polje, na katerega so vezani vsi podatki posameznega posla.
Šifra KO	Podatek o oznaki katastrske občine, ki je administrativna enota za vodenje podatkov zemljiškega katastra in katastra stavb. Šifra katastrske občine enolično določa katastrsko občino v Republiki Sloveniji.
Ime KO	Ime katastrske občine, kot je evidentirano v zemljiškem katastru. Ime katastrske občine je poimenovanje posamezne katastrske občine in ni enolična oznaka.
Občina	Podatek iz registra prostorskih enot o imenu občine, v kateri se nahaja pretežno število parcel, stavb oziroma delov stavb, ki so bile predmet posameznega pravnega posla.
Številka stavbe	Skupaj s šifro katastrske občine identifikacijska oznaka stavbe, ki enolično označuje stavbo.

Se nadaljuje na naslednji strani

Tabela A.5 – *Najemni posli - deli stavb - nadaljevanje.*

podatek	opis podatka
Številka dela stavbe	Skupaj s šifro katastrske občine in številko stavbe identifikacijska oznaka dela stavbe, ki enolično označuje del stavbe. Vsaka stavba ima vsaj en del stavbe.
Interna oznaka prostora	Najemodajalčeva interna oznaka prostora, pod katero oddaja prostor v najem; omogoči se njeno posredovanje v aplikacijo in prek XML, na obrazcu RN je ne dodajamo.
Naselje	Kraj: Podatek, kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Ulica	Ulica: Podatek, kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Hišna številka	Hišna številka: Podatek, kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Dodatek HŠ	Dodatek k hišni številki: Podatek, kot se v skladu s predpisi o evidentiranju nepremičnin vodi v registru nepremičnin.
Številka stanovanja ali poslovnega prostora	Številka stanovanja ali poslovnega prostora.
Vrsta oddane površine	Vrsta oddane površine glede na namen najema, za katero sta se dogovorili pogodbeni stranki in za katero velja pogodbena najemnina.
Opremljenost oddane površine	Podatek ali je oddana površina, ki je predmet najemnega pravnega posla, opremljena ali ne.
Mikrolokacija	Lega oddane površine v stavbi.
Izložba	Podatek o prisotnosti izložbenega okna v prostorih oddane površine.
Nakupovalni center	Podatek ali je oddana površina v nakupovalnem centru.
Oddana površina	Skupna oddana površina izražena v m^2 brez uporabe korekcijskih faktorjev, za vrsto oddane površine, za katero se dogovorita pogodbeni stranki in za katero velja pogodbena najemnina.
Oddana površina enaka delu stavbe	Podatek o tem ali oddana površina odgovarja celotnemu delu stavbe, kot je evidentiran v REN-u.

Se nadaljuje na naslednji strani

Tabela A.5 – *Najemni posli - deli stavb - nadaljevanje.*

podatek	opis podatka
Pogodbena najemnina oddane površine	Podatek o višini najemnine iz pogodbe, na mesec.
Opombe o nepremičnini	Opombe.
Leto izgradnje stavbe	Datum izgradnje stavbe.
Dejanska raba dela stavbe	Šifra in naziv dejanske rabe.
Lega dela stavbe v stavbi	Izpis lege dela stavbe v stavbi.
Število sob	Število sob v delu stavbe.
Površina dela stavbe	Površina dela stavbe v m^2.
Uporabna površina dela stavbe	Uporabna površina dela stavbe v m^2.
Prostori stanovanja	Našteti so prostori stanovanja, med seboj ločeni s pokončno črto '|'.

Literatura

[1] Vlada Republike Slovenije, UKAZ o razglasitvi zakona o množičnem vrednotenju nepremičnin (ZMVN), Uradni list 2006 (50).

[2] Vlada Republike Slovenije, Uredba o določitvi modelov vrednotenja nepremičnin, Uradni list 2011 (95).

[3] Urad za množično vrednotenje nepremičnin, GURS, Cene stanovanj v Sloveniji, Geodetski vestnik 1 (56) (2012) 196 – 206.

[4] GURS, Evidenca trga nepremičnin, (dostopano: 01.12.2017) (2017).
URL http://prostor3.gov.si/ETN-JV/

[5] STAT-RS, Podatkovni portal SI-STAT, (dostopano: 01.12.2017) (2006).
URL http://pxweb.stat.si/pxweb/Dialog/statfile2.asp

[6] J. Klemenc, Uporaba metod rudarjenja podatkov za analizo nepremičninskih transakcij v Republiki Sloveniji in izgradnjo modela za tržno vrednotenje nepremičnin, Master's thesis, Univerza v Ljubljani, Ekonomska fakulteta (2005).

[7] S. Rotar, Ocenjevanje vrednosti nepremičnin v praksi, Diploma thesis, Univerza v Ljubljani, Ekonomska fakulteta (2005).

[8] M. Milič, Kje v Sloveniji se najbolje živi? Ekskluzivno: pripravili smo lestvico 211 občin, (dostopano: 01.12.2017) (2015).
URL https://mojefinance.finance.si/8839372

[9] M. Milič, Kje v Sloveniji se najbolje živi? Pripravili smo lestvico 211 občin, (dostopano: 01.12.2017) (2017).
URL https://mojefinance.finance.si/8853962

[10] N. Pow, E. Janulewicz, L. D. Liu, Applied Machine Learning Project 4: Prediction of real estate property prices in Montreal, (dostopano: 01.12.2017) (2014).
URL http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf

[11] A. Caplin, S. Chopra, J. V. Leahy, Y. Lecun, T. Thampy, Machine Learning and the Spatial Structure of House Prices and Housing Returns, Social Science Research Network, 2008.

[12] J. Kranjc, Analiza trga stanovanj v slovenskih mestnih občinah v obdobju od januarja 2007 do junija 2015, Diploma thesis, Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo (2016).

[13] E. Samec, Analiza najemnega trga stanovanj v slovenskih mestnih občinah, Master's thesis, Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo (2016).

[14] M. Lichman, UCI Machine Learning Repository, (dostopano: 05.12.2016) (2013).
URL https://archive.ics.uci.edu/ml/datasets/Housing

[15] N. Shaganti, Exploring Boston Housing Data, (dostopano: 01.12.2017) (2015).
URL https://tuxdoc.com/download/exploring-boston-housing-data_pdf

[16] B. Božić, D. Milićević, M. Pejić, S. Marošan, The use of multiple linear regression in property valuation, Geonauka 1 (1) (2013) 41 – 45.

[17] R. D. Jaen, Data Mining: An Empirical Application in Real Estate Valuation, in: FLAIRS Conference, Pensacola Beach, Florida, USA, AAAI Press, 2002, pp. 314–317.

[18] I. S. H. Bahia, A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study, International Journal of Intelligence Science 3 (4) (2013) 162 – 169.

[19] I. Pardoe, Modeling Home Prices Using Realtor Data, Journal of Statistics Education 16 (2) (2008) .

[20] R. E. Lowrance, Y. LeCun, D. Shasha, Predicting the Market Value of Single-Family Residential Real Estate, Ph.D. thesis, Department of Computer Science, New York University (2015).

[21] T. Lasota, T. Łuczak, M. Niemczyk, M. Olszewski, B. Trawiński, Investigation of Property Valuation Models Based on Decision Tree Ensembles Built over Noised Data, in: Proceedings of the 5th International Conference on Computational Collective Intelligence. Technologies and Applications, Vol. 8083, 2013, pp. 417–426.

[22] J. W. Hujia Yu, Real Estate Price Prediction with Regression and Classification, CS 229 Project Final Report, 2016.

[23] E. A. Antipov, E. B. Pokryshevskaya, Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics, Expert Systems with Applications 39 (2) (2012) 1772 – 1778.

[24] V. Masías, M. Valle, F. Crespo, R. Crespo, A. Vargas Schüler, S. Laengle, Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile, in: AMSE Conference Santiago/Chile, At Santiago, 2002, pp. 314–317.

[25] A. Verikas, A. Gelzinis, M. Bacauskiene, Mining data with random forests: A survey and results of new tests, Pattern Recognition 44 (2) (2011) 330 – 349.

[26] P. Domingos, A Few Useful Things to Know About Machine Learning, Commun. ACM 55 (10) (2012) 78–87.

[27] SLONEP, Ocenjevanje vrednosti nepremičnin, (dostopano: 28.08.2018) (2007).
URL https://www.slonep.net/storitve/izmere-in-cenitve/ocenjevanje-vrednosti

[28] Ministrstvo za gospodarski razvoj in tehnologijo, Cene naftnih derivatov, (dostopano: 10.09.2018) (2018).
URL http://www.mgrt.gov.si/si/delovna_podrocja/notranji_trg/nadzor_cen_naftnih_derivatov/cene_naftnih_derivatov/

[29] Banka Slovenije, Obrestne mere Evropske centralne banke, (dostopano: 10.09.2018) (2017).
URL https://www.bsi.si/statistika/obrestne-mere/obrestne-mere-evropske-centralne-banke

[30] Google, Distance Matrix Service, (dostopano: 14.08.2018) (2018).
URL https://developers.google.com/maps/documentation/javascript/distancematrix

[31] GURS, Evidenca trga nepremičnin (ETN), (dostopano: 14.08.2018) (2018).
URL http://www.gu.gov.si/fileadmin/gu.gov.si/pageuploads/ETN/gradiva/Brosura_etn_splet_GIS_080602.pdf

[32] GURS, Register nepremičnin (REN), (dostopano: 14.08.2018) (2018).
URL http://www.gu.gov.si/fileadmin/gu.gov.si/pageuploads/GRADIVA/PUBLIKACIJE/zlozenke/REN_zlozenka.pdf

[33] NumPy developers, NumPy, (dostopano: 10.09.2018) (2018).
URL http://www.numpy.org

[34] scikit-learn developers, scikit-learn, (dostopano: 10.09.2018) (2018).
URL http://scikit-learn.org

[35] S. Rudi Seljak, Statistično urejanje podatkov - splošna metodološka pojasnila, (dostopano: 22.08.2018) (2016).
URL http://www.stat.si/dokument/8905/StaticticnoUrejanjePodatkovMPsplosna.pdf

[36] S. Prabhakaran, Tutorials on Advanced Stats and Machine Learning With R, (dostopano: 22.08.2018) (2016-17).
URL http://r-statistics.co

[37] J. Brownlee, How to Identify Outliers in your Data, (dostopano: 31.08.2018) (2013).
URL https://machinelearningmastery.com/how-to-identify-outliers-in-your-data/

[38] S. van Buuren, C. Groothuis-Oudshoorn, mice: Multivariate Imputation by Chained Equations in R, Journal of Statistical Software 45 (3).

[39] T. Hothorn, cforest, (dostopano: 22.08.2018) (2018).
URL https://www.rdocumentation.org/packages/party/versions/1.3-0/topics/cforest

[40] T. Hothorn, varImp, (dostopano: 22.08.2018) (2018).
URL https://www.rdocumentation.org/packages/party/versions/1.3-0/topics/varimp

[41] Boruta, (dostopano: 22.08.2018) (2018).
URL https://www.rdocumentation.org/packages/Boruta/versions/6.0.0/topics/Boruta

[42] Y. Hadar, Using categorical data in machine learning with python, (dostopano: 22.08.2018) (2017).
URL https://blog.myyellowroad.com/using-categorical-data-in-machine-learning-with-python-from-dummy-variables-to-deep-category-66041f734512

[43] S. Robinson, Linear Regression in Python with Scikit-Learn, (dostopano: 29.08.2018) (2018).
URL https://stackabuse.com/linear-regression-in-python-with-scikit-learn/

[44] D. Frossard, Linear Regression with NumPy, (dostopano: 29.08.2018) (2016).
URL https://www.cs.toronto.edu/~frossard/post/linear_regression/

[45] scikit-learn developers, sklearn linear model LinearRegression, (dostopano: 31.08.2018) (2017).
URL http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

[46] M. Galarnyk, PCA using Python (scikit-learn), (dostopano: 31.08.2018) (2017).
URL https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

[47] N. Donges, The Random Forest Algorithm, (dostopano: 31.08.2018) (2018).
URL https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd

[48] scikit-learn developers, sklearn ensemble RandomForestRegressor, (dostopano: 31.08.2018) (2017).
URL http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

[49] A. Bronshtein, Train/Test Split and Cross Validation in Python, (dostopano: 01.09.2018) (2017).
URL `https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6`

[50] J. Frost, When Should I Use Regression Analysis?, (dostopano: 01.09.2018) (2017).
URL `http://statisticsbyjim.com/regression/when-use-regression-analysis/`

[51] J. Frost, How to Interpret P-values and Coefficients in Regression Analysis, (dostopano: 01.09.2018) (2017).
URL `http://statisticsbyjim.com/regression/interpret-coefficients-p-values-regression/`

[52] S. Glen, P-Value in Statistical Hypothesis Tests: What is it?, (dostopano: 01.09.2018) (2014).
URL `http://www.statisticshowto.com/p-value/`

[53] A. Bronshtein, Simple and Multiple Linear Regression in Python, (dostopano: 01.09.2018) (2017).
URL `https://towardsdatascience.com/simple-and-multiple-linear-regression-in-python-c928425168f9`

[54] M. Kovačič, On-line slovarček statističnih pojmov: Osnovna statistična analiza, (dostopano: 01.09.2018).
URL `http://www.ljudmila.org/matej/statistika/mva.html`

[55] N. Zumel, J. Mount, Practical Data Science with R, 1st Edition, Manning Publications Co., Greenwich, CT, USA, 2014.

[56] A. Swalin, Choosing the Right Metric for Evaluating Machine Learning Models - Part 1, (dostopano: 02.09.2018) (2018).
URL `https://medium.com/usf-msds/choosing-the-right-metric-for-machine-learning-models-part-1-a99d7d7414e4`

[57] F. Veronesi, Assessing the Accuracy of our models (R Squared, Adjusted R Squared, RMSE, MAE, AIC), (dostopano: 02.09.2018) (2017).
URL https://www.r-bloggers.com/assessing-the-accuracy-of-our-models-r-squared-adjusted-r-squared-rmse-mae-aic/

[58] S. Glen, Adjusted R2 / Adjusted R-Squared: What is it used for?, (dostopano: 02.09.2018) (2013).
URL http://www.statisticshowto.com/adjusted-r2/

[59] GURS, Opis strukture podatkov Evidence trga nepremičnin, (dostopano: 14.05.2016) (2015).
URL http://www.e-prostor.gov.si/fileadmin/struktura/Opis_strukture_ETN.pdf