UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

ZAKLJUČNA NALOGA

UPORABA PODATKOVNEGA RUDARJENJA ZA NAPOVEDOVANJE CEN NEPREMIČNIN

URŠKA MIKOLJ

2018
ZAKLJUČNA NALOGA
MIKOLJ

UNIVERZA NA PRIMORSKEM
FAKULTETA ZA MATEMATIKO, NARAVOSLOVJE IN
INFORMACIJSKE TEHNOLOGIJE

Zaključna naloga

Uporaba podatkovnega rudarjenja za napovedovanje cen nepremičnin

(Use of Data Mining for Real Estate Price Prediction)

Ime in priimek: Urška Mikolj
Študijski program: Računalništvo in informatika
Mentor: doc. dr. Branko Kavšek

Koper, julij 2018

Ključna dokumentacijska informacija

Ime in PRIIMEK: Urška MIKOLJ

Naslov zaključne naloge: Uporaba podatkovnega rudarjenja za napovedovanje cen nepremičnin

Kraj: Koper

Leto: 2018

Število listov: 64 Število slik: 17 Število tabel: 4

Število prilog: 3 Število strani prilog: 12 Število referenc: 39

Mentor: doc. dr. Branko Kavšek

Ključne besede: napovedovanje cen nepremičin, podatkovno rudarjenje, Weka

Izvleček:

V zaključni nalogi je predstavljen primer uporabe podatkovnega rudarjenja za napovedovanje cen nepremičnin. Pri tem so uporabljeni podatki za nepremičnine v Moskvi in okolici. Naloga se osredotoča predvsem na vprašanje ali je mogoče s podatkovnim rudarjenjem dobiti dovolj dobre napovedi za cene nepremičnin. Za izbrane podatke o nepremičninah je opisan celoten postopek podatkovnega rudarjenja po CRISP-DM metodologiji: od razumevanja problema in podatkov, predobdelave, izgradnje modelov, testiranja ter interpretacije. Za izgradnjo modelov in testiranje je uporabljen odprtokodni program Weka, kjer so uporabljeni različni regresijski algoritmi za izgradnjo napovednih modelov. Iz dobljenih rezultatov lahko zaključimo, da je podatkovno rudarjenje dober pristop za napovedovanje cen nepremičnin. Kot najprimernejše algoritme za ta primer bi lahko izpostavili algoritme M5P, RandomTree, RandomForest in REPTree, izmed katerih smo najboljše rezultate dobili z algoritmom RandomForest. Napovedni modeli predstavljeni v okviru te naloge bi bili primerni za približno oceno vrednosti nepremičnine. Za dejansko uporabo pa bi morali napake še malo zmanjšati, kar bi lahko dosegli z uporabo boljših podatkov, ki bi vsebovali manj manjkajočih in nepravilnih vrednosti.

Key words documentation

Name and SURNAME: Urška MIKOLJ

Title of final project paper: Use of Data Mining for Real Estate Price Prediction

Place: Koper

Year: 2018

Number of pages: 64 Number of figures: 17 Number of tables: 4

Number of appendices: 3 Number of appendix pages: 12 Number of references: 39

Mentor: Assist. Prof. Branko Kavšek, PhD

Keywords: real estate price prediction, data mining, Weka

Abstract:

The final project paper presents an example of the use of data mining for real estate price prediction used on data for real estate in Moscow and surroundings. The paper focuses primarily on the question whether is it possible to get good enough predictions with the use of data mining. Here the whole proces is described using CRISP-DM methodology - from problem and data understanding, data processing, model building, testing to interpretation of obtained results. For model building and testing we use open source program Weka where different regression algorithms are used to build prediction models. At the end the results show that data mining is a good approach for predicting real estate prices. As the most appropriate algorithms for this case M5P, RandomTree, RandomForest and REPTree are selected, where RandomForest had the best results. Models presented in this paper could be used for predicting approximate real estate value. However models used for actual prediction of real estate prices should have smaller errors. That can be achieved using better data that has less missing and inaccurate values.

Kazalo vsebine

Kazalo tabel

Kazalo slik

Kazalo prilog

Seznam kratic

$WEKA$	Programsko orodje za podatkovno rudarjenje in strojno učenje (ang. Waikato Environment for Knowledge Analysis)
CSV	Format vrednosti ločenih z vejico (ang. Comma Separated Values)
$ARFF$	Format, ki ga uporablja WEKA (ang. Attribute Relation File Format)
KDD	Odkrivanje zakonitosti v podatkih (ang. Knowledge Discovery in Databases)
$SEMMA$	Vzorčenje, raziskovanje, spreminjanje, modeliranje, ocenjevanje (ang. Sample, Explore, Modify, Model, Assess)
$CRISP-DM$	Industrijski standard za podatkovno rudarjenje (ang. Cross Industry Standard Process for Data Mining)

1 Uvod

Podatkovno rudarjenje se uporablja za iskanje določenih vzorcev in trendov v podatkih ali napovedovanje različnih vrednosti, predvsem takrat, ko so na voljo velike količine podatkov [34]. Tako je uporabno na zelo različnih področjih npr. v marketingu za izboljšanje tehnik oglaševanja in prodaje, v zavarovalništvu za ugotavljanje morebitnih prevar, v bančništvu za odločanje pri dodeljevanju kreditov, v medicini za enostavnejše postavljanje diagnoz in še na mnogih drugih področjih [22].

V tem primeru bo predstavljena uporaba podatkovnega rudarjenja za napovedovanje cen nepremičnin glede na ostale znane podatke o posamezni nepremičnini. Podatki uporabljeni za izgradnjo modelov bodo opisovali nepremičnine v Moskvi in njeni okolici. Pri tem bodo na voljo podatki o osnovnih lastnostih nepremičnine (npr. površina, število sob, nadstropje) in različni podatki o območju kjer se nepremičnina nahaja (npr. število prebivalcev, število trgovin v okolici, število športnih objektov, oddaljenost do centra mesta). Ker cena nepremičnine ni odvisna samo od teh lasnosti, ampak tudi od širšega gospodarskega dogajanja, bodo zraven uporabljeni še makroekonomski podatki o stanju v ruski ekonomiji na dan prodaje oziroma nakupa nepremičnine (npr. višina bruto domačega proizvoda, cena nafte).

Napovedovanje cen nepremičnin se velikokrat uporablja kot primer za regresijo, zato lahko pričakujemo dobre rezultate. Za napovedovanje cen bodo v programu Weka [32] uporabljeni različni regresijski algoritmi: IBk [1], M5P [24] [31], M5Rules [18] [24] [31], REPTree [34], RandomTree [34], RandomForest [5] in LinearRegression [34] ter za primerjavo še ZeroR [34] in DecisionStump [34]. Cilj naloge pa je na ta način ugotoviti, ali je podatkovno rudarjenje dejansko uporabno za napovedovanje cen nepremičnin in kako natančne so tako dobljene napovedi.

V drugem poglavju bo predstavljeno podatkovno rudarjenje in strojno učenje ter nekaj primerov uporabe. V tretjem poglavju bo opisano stanje na ruskem nepremičninskem trgu. V četrtem poglavju bo opisana metodologija dela s podatki, ki vključuje razumevanje in predobdelavo podatkov. Tu bo predstavljeno tudi programsko orodje Weka, ki bo uporabljeno za namene podatkovnega rudarjenja. V petem poglavju bodo opisani algoritmi za izgradnjo napovednih modelov ter načini za njihovo ocenjevanje in testiranje. V šestem poglavju bodo predstavljeni dobljeni rezultati in interpretacija rezultatov. V sedmem poglavju bodo podani zaključki in smernice za nadaljnje delo.

2 Podatkovno rudarjenje

Podatkovno rudarjenje je proces odkrivanja novih smiselnih povezav, vzorcev in trendov s preučevanjem velikih količin podatkov z uporabo tehnologij za prepoznavanje vzorcev kot tudi s statističnimi in matematičnimi tehnikami [29]. Namen podatkovnega rudarjenja je predvsem v pridobivanju novega znanja iz podatkov [17].

Pomen podatkovnega rudarjenja postaja vedno večji, saj zelo hitro narašča količina najrazličnejših podatkov od velikih količin podatkov na spletu, različnih znanstvenih podatkov do baz podatkov, ki jih o svojih strankah hranijo zavarovalnice, banke in velike trgovine. Ker so takšne količine podatkov zelo velike, jih je potrebno še dodatno obdelati, da iz njih dobimo določene smiselne povezave, vzorce in podobno. Pri tem si lahko pomagamo prav s podatkovnim rudarjenjem.

2.1 Podatkovno rudarjenje in strojno učenje

Poleg podatkovnega rudarjenja je potrebno omeniti še strojno učenje, saj se oba pojma velikokrat uporabljata hkrati. Strojno učenje se ukvarja s tem kako se računalniki lahko učijo glede na podatke, ne da bi bili pred tem eksplicitno programirani [16]. Lahko bi rekli, da je strojno učenje nekakšna tehnična osnova za podatkovno rudarjenje, saj se veliko tehnik strojnega učenja uporablja tudi pri podatkovnem rudarjenju. Kljub temu med njima obstajajo razlike, saj je cilj strojnega učenja predvsem napovedovanje, cilj podatkovnega rudarjenja pa iskanje novih vzorcev v podatkih.

Strojno učenje se največkrat uporablja za probleme, ki so težko rešljivi na način, da bi bili vnaprej programirani. Tako je uporabno za prepoznavanje različnih predmetov na slikah, za prepoznavanje obraza ter tudi pisave in govora. Koristno je pri filtriranju elektronske pošte in generiranju priporočil za glasbo, filme, različne izdelke in podobno. Prav tako pomaga pri zaznavanju prevar kot so goljufive bančne transakcije, pri personalizaciji oglaševanja in še na različnih drugih področjih [20].

Kot je vidno iz primerov uporabe je strojno učenje usmerjeno v praktično reševanje problemov npr. prepoznavanje določenega predmeta na sliki ali filtriranju elektronske pošte. Za razliko od podatkovnega rudarjenja pri strojnem učenju ni toliko poudarka na iskanju nekega novega znanja v podatkih. Ne glede na to, pa so tehnike, ki jih uporablja strojno učenje zelo uporabne tudi pri podatkovnem rudarjenju.

2.2 Primeri uporabe podatkovnega rudarjenja

Podatkovno rudarjenje je uporabno na različnih področjih, kjer imamo na voljo večje količine podatkov iz katerih želimo pridobiti neko novo znanje oziroma določene vzorce, trende in podobno.

Lahko ga uporabljajo trgovine za analizo prodaje in tako ugotovijo kateri izdelki so pogosto kupljeni hkrati ter druge nakupovalne navade kupcev. To pripomore tudi k izboljšanju oglaševanja. Podobno je z bankami, kjer si lahko na ta način pomagajo pri odločitvi, katerim strankam bodo odobrili kredit. Uporabljajo ga lahko operaterji pri predvidevanju za katere stranke obstaja možnost, da bi operaterja zamenjali. Tem strankam nato ponudijo kakšne dodatne ugodnosti. Različna podjetja se lahko s pomočjo podatkovnega rudarjenja odločijo katerim strankam bodo ponudili določene nove izdelke ali storitve. Uporabno je tudi za odkrivanje različnih prevar kot na primer za odkrivanje goljufivih bančnih transakcij ali prevar v zavarovalništvu.

Uporaba podatkovnega rudarjenja pa ni samo poslovno usmerjena. Podatkovno rudarjenje je uporabno še na veliko področjih npr. v biologiji za raziskovanje genoma različnih rastlin, v geologiji za iskanje možnih lokacij nafte in zemeljskega plina, v kmetijstvu za napovedovanje pridelka glede na vremenske pogoje, v kemiji za napovedovanje strukture različnih organskih spojin, ipd. Zelo uporabno je v medicini, kjer lahko pomaga pri diagnosticiranju pacientov ter tudi pri ugotovljanju uspešnost delovanja različnih načinov zdravljenja in zdravil. Koristno je tudi v industriji za preverjanje kakovosti izdelkov in za odkrivanje morebitnih napak. Uporabno naj bi bilo tudi za ugotavljanje trendov terorističnih aktivnosti. V tem primeru pa bo preizkušeno kako uporabno je za napovedovanje cen nepremičnin.

2.3 Metodologije podatkovnega rudarjenja

Podatkovno rudarjenje se velikokrat uporablja pri projektih na najrazličnejših področjih, kot je bilo omenjeno že v prejšnjem poglavju. Tako se pojavlja tudi potreba po določenih standardih, ki bi zagotovili večjo uspešnost, razumljivost in preverljivost vseh takšnih projektov.

Ena izmed metodologij je KDD (ang. Knowledge Discovery in Databases). Podatkovno rudarjenje je samo ena izmed faz v procesu odkrivanja znanja v podatkih, ki se nanaša predvsem na uporabo različnih modelov za iskanje vzorcev, povezav ali trendov v podatkih. Celoten proces odkrivanja znanja v podatkih pa je veliko širši. Zajema celoteno dogajanje od pridobivanja podatkov, njihovega urejanja in obdelave, uporabe podatkovnega rudarjenja do evalvacije in predstavitve pridobljenega znanja [3] [13] [16].

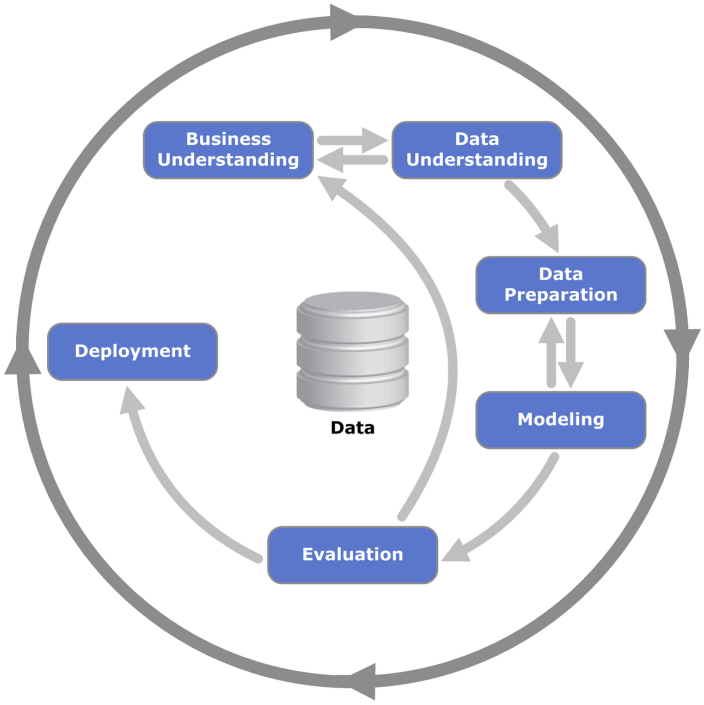
Druga takšna metodologija je SEMMA. Zasnoval jo je SAS Institute z namenom,

da jo uporabi v programu za podatkovno rudarjenje SAS Enterprise Miner. Metodologija je sestavljena iz posameznih korakov, ki opisujejo faze v projektu podatkovnega rudarjenja. Vsaka črka v imenu predstavlja eno izmed faz: sample, explore, modify, model, asses (vzorčenje, razliskovanje, spreminjanje, modeliranje, ocenjevanje) [3] [22].

Nekoliko novejša je metodologija CRISP-DM, ki bo uporabljena tudi v tem primeru za analizo podatkov o nepremičninah in podatkovno rudarjenje, zato bo podrobneje opisana v nadaljevanju [28] [33].

2.3.1 CRISP-DM metodologija

CRISP-DM [22] [28] [33] predstavlja strukturiran pristop k planiranju in izvedbi faz projektov za podatkovno rudarjenje. Razvila ga je skupina petih podjetij: SPSS, Teradata, Daimler AG, NCR Corporation in OHRA. Na CRISP-DM lahko gledamo kot na metodologijo ali procesni model. CRISP-DM kot metodologija vključuje opise posameznih faz projekta, in nalog, ki so vključene v posamezno fazo. Kot procesni model pa ponuja pregled življenjskega cikla podatkovnega rudarjenja. Celoten cikel je sestavljen iz 6 faz, ki so prikazane tudi na sliki 1:



Slika 1: Faze CRISP-DM procesnega modela [10]

1 **Razumevanje problema** (ang. Business understanding) - Ta faza se osredotoča na razumevanje projektnih ciljev in zahtev. Njen namen je natančno definirati problem, ki ga želimo rešiti ter pripraviti načrt izvedbe.

2 **Razumevanje podatkov** (ang. Data understanding) - Faza se začne z zbiranjem podatkov in nadaljuje z njihovim preučevanjem. Pri tem so nam v pomoč tabele, grafi ter različna vizualizacijska in statistična orodja. Takšne analize lahko pomagajo pri razumevanju problema definiranega v prejšnji fazi in tudi že pri postavljanju določenih hipotez. Poleg tega lahko v tej fazi že ugotovimo kje v podatkih so manjkajoče vrednosti in napake, kar je uporabno za naslednjo fazo.

3 **Priprava podatkov** (ang. Data preparation) - Je najpomembnejša faza, ki običajno zahteva tudi največ časa. Vključuje različne aktivnost npr. izbiro podatkov, čiščenje podatkov zaradi manjkajočih in nepravilnih vrednosti, generiranje novih atributov, ipd. Cilj faze je, da dobimo končne urejene podatke, ki bodo uporabljeni za modeliranje.

4 **Modeliranje** (ang. Modeling) - V tej fazi je potrebno najprej izbrati tehnike in modele, ki so najprimernejši za podan problem in podatke. Nato izbrane modele uporabimo na podatkih. Ta postopek se običajno izvaja večkrat z različnimi modeli in parametri, dokler ne dobimo optimalnih rezultatov.

5 **Vrednotenje** (ang. Evaluation) - Ta faza je namenjena ovrednotenju modelov dobljenih v prejšnji fazi. Ugotoviti je potrebno ali modeli ustrezajo zastavljenim kriterijem. Če modeli še niso dovolj dobri lahko ponovimo prejšnjo fazo.

6 **Uporaba** (ang. Deployment) - Izdelava modelov ponavadi še ne pomeni zaključka projekta. Tudi če je namen modela dobiti neko znanje iz podatkov, je potrebno dobljeno znanje organizirati in predstaviti na način, da se bo lahko tako novo pridobljeno znanje uporabilo. Ta faza se razlikuje pri različnih projektih. Lahko je zelo preprosta in je na koncu potrebno izdelati le poročilo o novo pridobljenih ugotovitvah ali pa zelo kompleksna in zahteva dejansko implementacijo procesa podatkovnega rudarjenja.

Po metodologiji CRISP-DM bodo analizirani tudi podatki o nepremičninah. Ta metodologija je izbrana, saj je najbolj pregledna in vključuje vse faze od razumevanja problema do uporabe. Tu je poudarek tudi na razumevanju problema in razumevanju podatkov pred samim modeliranjem, kar je koristno, saj lahko tako dobimo boljši vpogled v podatke in tudi bolje razumemo problem, ki ga rešujemo. Ostale faze pa so večinoma primerljive s fazami metodologij KDD in SEMMA. Sicer je CRISP-DM kot procesni model zelo prilagodljiv. Omogoča prehajanje med fazami naprej in nazaj vkolikor je to potrebno. Poleg tega omogoča, da damo nekaterim fazam večji poudarek kot drugim, saj so v nekaterih primerih določene faze bolj pomembne, druge malo manj, kar je odvisno od posameznega projekta.

3 Ruski nepremičninski trg

V tem poglavju bo podan opis problema t.j. določanje cen nepremičnin na ruskem nepremičninskem trgu. Ruski nepremičninski trg, ima tako kot drugi nepremičninski trgi, svoje posebnosti. Nanj vplivajo določeni lokalni dejavniki ter splošno gospodarsko stanje v Rusiji in širše. Da bi razumeli cene nepremičnin in njihovo spreminjanje je potrebno razumeti tudi dejavnike, ki vplivajo nanje. Zato bo tukaj za boljše razumevanje predstavljeno kateri tipi nepremičnin obstajajo, kateri dejavniki so najpomembnejši za cene nepremičnin in kako se cene spreminjajo odvisno od gospodarskih dejavnikov.

3.1 Dejavniki, ki vplivajo na cene nepremičnin

Podobno kot na drugih nepremičninskih trgih cene nepremičnin niso odvisne samo od posameznih lastnosti kot so npr. površina, število sob, leto izgradnje. V veliki meri na ceno vplivajo tudi ekonomski dejavniki v državi, ki so lahko specifični za Rusijo ali pa so posledica globalne ekonomske situacije. Tega se moramo zavedati tudi pri napovedovanju cen nepremičnin, kjer je potrebno upoštevati tako lastnosti nepremičnin kot tudi vpliv različnih drugih dejavnikov.

Dejavnike, ki vplivajo na cene nepremičnin lahko razdelimo v dve skupini: lokalne in globalne. Lokalni dejavniki so odvisni predvsem od lastnosti posamezne nepremičnine, in vplivajo na to, da so cene nepremičnin različne saj je npr. neka nepremičnina na boljši lokaciji, druga je novejša ali ima večjo kuhinjo. Ti dejavniki so neodvisni od časa, saj je stanovanje v prvem nadstropju vedno nekoliko cenejše od stanovanja v medetaži, podobno je stanovanje v zgradbi zgrajeni iz opeke dražje kot v zgradbi iz panelov. Kot pomembnejše dejavnike lahko tu omenimo starost nepremičnine, ki je povezana z materialom izgradnje, velikost nepremičnine, število sob, velikost kuhinje in lokacijo.

Druga skupina dejavnikov so globalni dejavniki. Ti so povezani z makroekonomskimi parametri, kot so raven gospodarskega razvoja in poslovanja v mestu, raven dohodkov in na splošno življenjski standard v mestu. Če primerjamo splošno ravnen cen v nekem mestu, z ravnjo cen v drugegih mestih, lahko ugotovimo, da je razmerje med cenami podobnih stanovanj v različnih mestih približno sorazmerno glede na to kakšen gospodarski status ima mesto. Tako so cene nepremičnin v Moskvi višje od cen v Sankt Peterburgu. Podobno lahko posplošimo na raven celotne države.

3.2 Posebnosti ruskega nepremičninskega trga

Ruski nepremičninski trg ima nekatere svoje posebnosti. Tukaj bo nekoliko več povedano o nepremičninskem trgu za Moskvo in okolico. V nadaljevanju bo predstavljeno kakšni so glavni tipi nepremičnin v Moskvi in kateri dejavniki imajo največji vpliv, ko se kupci odločajo za nakup nepremičnine.

3.2.1 Različni tipi nepremičnin

Največji delež nepremičnin predstavljajo stanovanja, zato bo tu poudarek predvsem na tej vrsti nepremičnin. Seveda so tukaj še hiše, poslovni prostori in različni drugi objekti, vendar jih je v primerjavi s stanovanji veliko manj.

Obstaja več tipov stanovanjskih objektov. Zanimivo pri tem je, da so v določenih obdobjih gradili določene tipe stavb, ki so bile med seboj zelo podobne. Tako je tudi zelo enostavno ločiti kateremu obdobju pripada posamezna stavba. Najmanj zaželene so petnadstropne stavbe imenovane tudi 'hruščovke' po Nikiti Hruščovem, zgrajene med leti 1950 in 1960. Stanovanja v teh stavbah so zelo majna, imajo slabo izolacijo, stare napeljave in nepraktično razporeditev prostorov. V zadnjem času jih zaradi dotrajanosti rušijo in nadomeščajo z novimi modernejšimi stanovanjskimi objekti.

Nekoliko boljše od njih so 'brežnjevke' poimenovane po naslednjem vodji Leonidu Brežnjevu, ki so bile zgrajene v obdobju med leti 1960 in 1980. Te stavbe imajo od 9 do 16 nadstropij, kvaliteta gradnje je tukaj malo boljša, vključno z boljšo razporeditvijo prostorov in malo večjo bivalno površino. Težava pa je še vedno izolacija in to, da je pozimi hladno in poleti vroče zaradi gradnje iz panelov.

Naboljše so 'stalinke' poimenovanje po Stalinu, ki imajo 4 do 7 nadstropij. Zgrajene so bile nekoliko prej v obdobju med leti 1930 in 1950. Kljub večji starosti so takšne stavbe zgrajene veliko bolje z uporabo dobrih gradbenih materialov. Značilnost stanovanj v teh stavbah so visoki stropi, večja bivalna površina in dober razpored prostorov. Nekatere izmed stavb imajo tudi lepo zunanjost. Večina izmed njih se nahaja v centru mesta. Slabost teh stanovanj je odsotnost dvigala in stara napeljava, ki jo je potrebno zamenjati. Zaradi starosti je velikokrat potrebna celotna obnova. Kljub temu je veliko takšnih stanovanj obnovljenih in dosegajo visoke cene tudi v primerjavi z novimi stanovanji.

Modernejše stavbe imajo do 21 nadstropij. Večinoma so zgrajene iz panelov, kar omogoča hitro gradnjo. Stanovanja v takšnih stavbah so večja in imajo boljšo razporeditev, vendar na splošno niso kaj dosti različna od prejšnjih različic. Nekatere nadstandardne stavbe imajo lepši izgled, različne nestandardne razporeditve prostorov, notranjo garažo, fitnes in podobno. V zadnjih letih se Moskva hitro modernizira, zato predvsem izven centra nastajajo velika stanovanjska naselja. V središču Moskve

so po vzoru drugih velikih mest zgradili tudi poslovni center poimenovan Moskva City, ki ga sestavlja več nebotčnikov.

Poleg centra ima Moskva tudi veliko obrobje, kjer se večinoma nahajajo stanovanjska naselja. Kakšen je izgled modernega obrobja Moskve je lepo prikazano na sliki 2. Tu so vidne različne nove stavbe, ki se od starejših ločujejo po modernejšemu izgledu in večjemu številu nadstropij. Med njimi je predvsem na desni še veliko starejših 5-nadstropnih 'hruščovk'. Z leve strani sta ob cesti vidni še dve 'brežnjevki'.

Slika 2: Nepremičnine na obrobju Moskve [35]

Za razliko od stanovanjskih objektov so hiše v Moskvi zelo redke. Kljub temu ima veliko prebivalcev, ki si to lahko privoščijo izven mesta 'dačo'. To je ponavadi manjša lesena hiša z vrtom. Tja se lastniki odpravijo čez vikend, kjer vrtnarijo in preživljajo prosti čas, vendar le malokdo dejansko živi tam, saj bi zaradi slabe situacije v prometu potrebovali več ur, da pridejo v službe, ki so ponavadi v centru.

3.2.2 Pomembni dejavniki pri nakupu nepremičnin

Pri nakupu nepremičnine je pomembnih več različnih dejavnikov. Zelo pomembna je kakovost izgradnje in praktičen razpored prostorov. Najbolj cenjene so monolitne stavbe in stavbe grajene iz opek, še posebej tiste, kjer imajo stanovanja visoke strope, večje sobe in kuhinjo, balkon ter novo napeljavo. Tako so najbolj zaželena stanovanja predvsem v novih stavbah in v starejših 'stalinkah'. V nasprotju z njimi je pri stavbah

grajenih iz panelov kakovost gradnje veliko slabša, vendar so stanovanja v takšnih stavbah cenovno najbolj ugodna.

Poleg kakovosti izgradnje same stavbe je precej pomembna tudi njena lokacija. Predvsem v smislu dobrih prometnih povezav, saj je v Moskvi situacija v prometu zelo slaba. Tako se nekateri prebivalci vozijo po več ur v in iz služb, ki so ponavadi v centru mesta. Temu se lahko vsaj delno izognejo z nakupom nepremičnine, kjer so prometne povezave malo boljše. Za družine z otroki je pomembna tudi bližina šol in vrtcev ter bližina različnih ustanov, kjer otroci obiskujejo obšolske aktivnosti npr. glasbene šole. Bližina drugih ustanov kot so npr. bolnišnice ni tako pomembna.

Eden od najmanj pomembnih dejavnikov je okolica nepremičnine. Zato ni tako pomembno, če okolica nepremičnine ni urejena in primanjkuje parkirnih prostorov, ni igrišč za otroke in podobno, vse dokler je kvaliteta same nepremičnine zelo dobra. Simbolično vrednost ima tudi novost nepremičnine, saj že samo dejstvo, da je nepremičnina nova odtehta vse ostale slabosti.

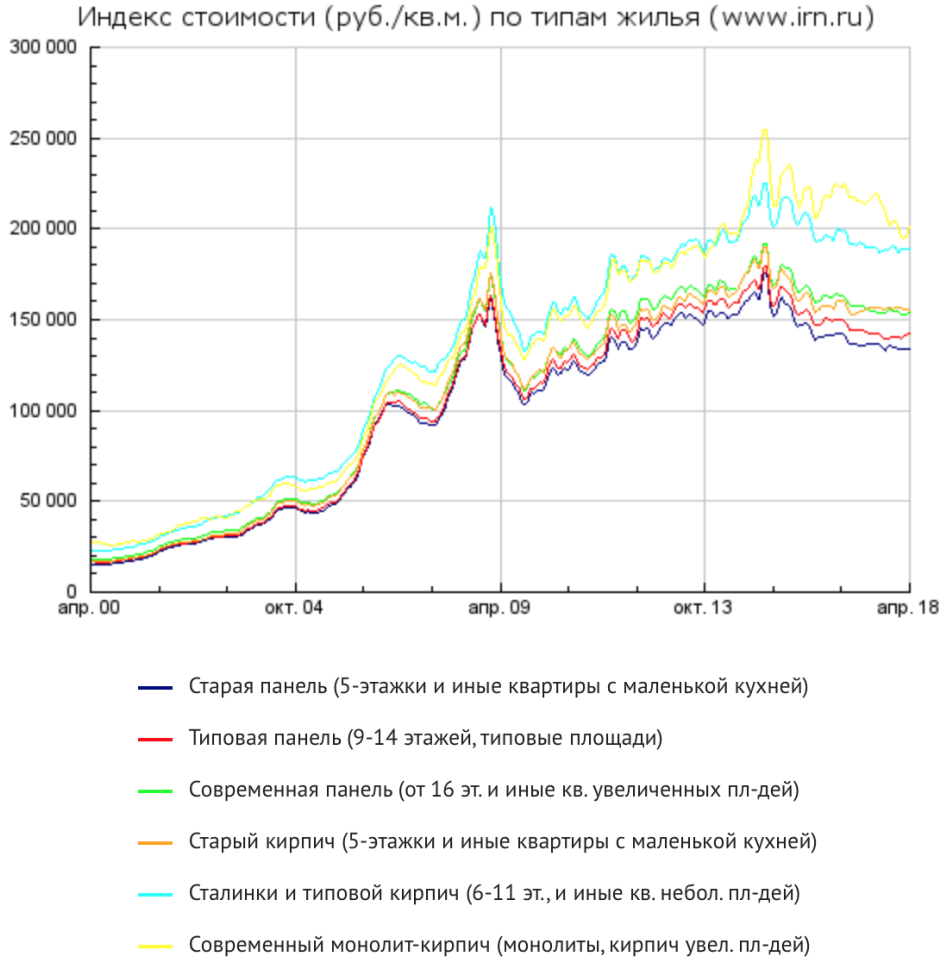
Najpomembnejši dejavnik pri izbiri nepremičnine pa je predvsem cena. Nove nepremičnine, ki imajo največ prednosti, so ponavadi tudi najdražje, medtem ko imajo starejše nepremičnine svoje slabosti kot so manjša bivalna površina, slabši razpored prostorov in stara napeljava. Tako je pri nakupu nepremičnine potrebno doseči nekakšen kompromis med prednostmi in slabostmi. Dobra stran ruskega nepremičninskega trga pa je, da ponuja zares veliko različnih možnosti za nakup nepremičnin od cenejših do zelo dragih.

3.3 Gibanje cen na ruskem nepremičninskem trgu

Cene na ruskem nepremičninskem trgu so večinoma odvisne od gospodarske situacije in se zelo spreminjajo. To je vidno tudi iz grafa na sliki 3, ki prikazuje spreminjanje povprečnih cene v rubljih za kvadratni meter stanovanja v Moskvi od leta 2000 do leta 2018. Iz grafa je vidno, da je cena za kvadratni meter med leti 2000 in 2009 strmo naraščala. Leta 2009 sledi padec, med leti 2010 do 2014 cena spet začne naraščati. Po letu 2015 pa cena ponovno nekoliko pada.

Graf z različnimi barvami prikazuje tudi kako se spreminja cena za različne tipe nepremičnin. Najdražje so moderne nepremičnine iz monolita in opeke prikazane z rumeno, sledijo 'stalinke' in tipne nepremičnine iz opeke s svetlo modro, starejše nepremičnine iz opeke z oranžno, moderne nepremičnine iz panelov z zeleno, tipne nepremičnine iz panelov z rdečo in najcenejše stare nepremičnine iz panelov. Kot je vidno iz grafa je cena nepremičnine odvisna tudi od tipa gradnje, saj se povprečna cena med najdražjimi in najcenejšimi nepremičninami razlikuje tudi za več kot 50.000 rubljev za kvadratni meter. Kljub temu se pri vseh tipih nepremičnin trendi padanja in naraščanja

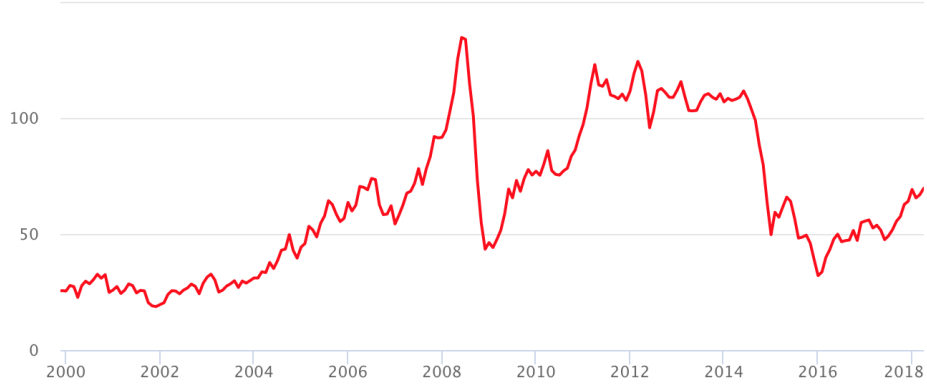
povprečnih cen spreminjajo približno enako. Iz tega lahko ugotovimo, da so cene nepremičnin v veliki meri odvisne od gospodarske situacije v državi, ki posledično vpliva na celoten nepremičninski trg.



Slika 3: Povprečna cena v rubljih za kvadratni meter za različne tipe nepremičnin v Moskvi od leta 2000 do leta 2018 [38]

Razlogov za takšno gibanje cen je več. Eden od najpomembnejših dejavnikov pa je cena nafte [19], saj je naftna industrija v Rusiji ena izmed največjih na svetu. Rusija ima tudi največje zaloge in je največji izvoznik zemeljskega plina. Je druga na svetu po zalogah premoga in osma po zalogah nafte ter ena izmed največjih proizvajalcev nafte [23]. Zato je razumljivo, da nafta predstavlja enega izmed ključnih dejavnikov, ki vplivajo na rusko ekonomijo. Cena nafte posledično vpliva tudi npr. na izvoz, višino bruto domačega proizvoda in tudi na vrednost rublja. Podobni vplivi se kažejo tudi na drugih področjih, v tem primeru pri cenah na nepremičninskem trgu. Povezave med enimi in drugimi trendi so vidne, če primerjamo grafa na slikah 3 in 4. Ko cena nafte narašča ali pada se podobno spreminjajo tudi cene nepremičnin, s tem da je pri

nepremičninah opazen zamik nekaj mesecev. Podobno lahko opazimo, da če je cena nafte visoka to dobro vpliva na ceno nepremičnin, v primeru da začne cena nafte padati začnejo kmalu padati tudi cene nepremičnin. Na sliki 3 sicer ni videti tako velikega padca cen v rubljih po letu 2014, kot je viden na sliki 4 pri padcu cen nafte, vendar se padec cen dejansko zgodi in bi ga videli, če bi namesto rubljev prikazali ceno v evrih ali dolarjih. Razlog za to je, da je s ceno nafte pada tudi vrednost rublja glede na evro ali dolar. V primeru, da bi graf za povprečne cene nepremičnin na sliki 3 prikazali v dolarjih ali evrih, bi bil ta graf, če zanemarimo kaj prikazuje, po izgledu skoraj identičen grafu za ceno nafte na sliki 4.



Slika 4: Cena nafte v USD za sodček od leta 2000 do leta 2018 [39]

Na ekonomijo v državi poleg nafte vpliva tudi širše dogajanje npr. svetovna gospodarska kriza leta 2008 ali sankcije zaradi posredovanja Rusije v Ukrajini leta 2014. Obstaja pa še veliko drugih dejavnikov, ki vplivajo na ekonomijo v državi in posledično na ceno nepremičnin. Lahko pa rečemo, da se dobro stanje na gospodarskem področju v državi odraža tudi na drugih področjih, v tem primeru pri ceni nepremičnin.

3.4 Napovedovanje cen nepremičnin za Moskvo in okolico

Ker nas zanima napovedovanje cen nepremičnin za Moskvo in okolico, bo tukaj predstavljen primer napovedovanja cen za nepremičnine na tem območju, ki je že v uporabi. Na spletni strani `https://www.irn.ru/price/` je mogoče uporabiti storitev analitičnega centra IRN.RU in oceniti približno tržno vrednost nepremičn, natančneje stanovanj v Moskvi in njeni okolici. Ocena temelji na uporabi multidimenzionalnih matrik indeksov za cene nepremičnin in matrik za ocenjene popravke. Takšen način vrednotenja je razširitev modela primerjalne tržne analize [37].

S primerjalno tržno analizo lahko ocenimo vrednosti nepremičnine, če poznamo vrednosti podobnih nepremičnin. Za to so ponavadi uporabljeni podatki o nedavno prodanih nepremičninah. V primeru da teh podatkov ni, se lahko uporabijo tudi podatki o oglaševanih nepremičninah, čeprav so ti manj natančni, saj ni nujno, da bo nepremičnina prodana za oglaševano ceno [9].

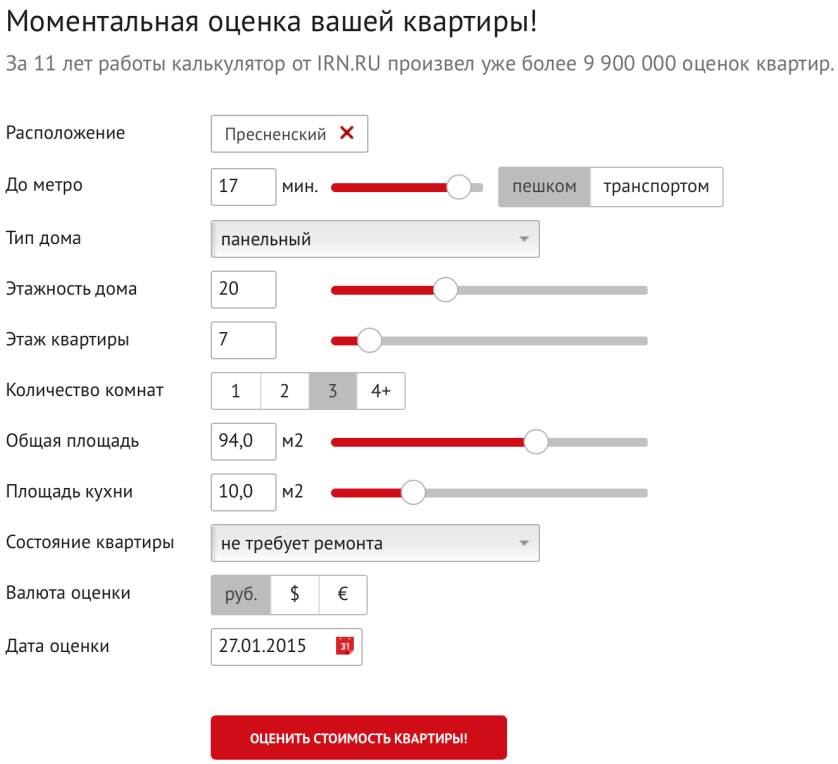
Da bi na ta način ocenili vrednost stanovanja je potrebno samo vnesti zahtevane podatke, kot je to prikazano na sliki 5. To so: lokacija nepremičnine (lahko izberemo glede na najbližji metro, rajon v Moskvi ali kraj če, je nepremičnina v okolici Moskve), oddaljenost v minutah do metroja (peš ali s transportom), tip nepremičnine glede na izgradnjo (nepremičnina grajena iz opeke, panelov, monolit, 'stalinka' ali elitni razred), število vseh nadstropij, nadstropje stanovanja, število sob, celotna površina stanovanja, površina kuhinje, stanje (v kakšnem stanju je stanovanje npr. če potrebuje obnovo) in izbran datum prodaje nepremičnine.

Pri tem je potrebno izbrati samo prej omenjene glavne dejavnike, ki najbolj vplivajo na ceno nepremičnine. Ostali manj pomembni dejavniki kot npr. prisotnost ali odsotnost balkona, vrsta talnih oblog, višina stropov, vplivajo na ceno stanovanja med 1 % in 2 % in so tako vključeni v oceni napake. Za večino primerov naj bi bila napaka med 3 % in 5 %, kar je sprejemljivo tudi pri pravi oceni nepremičnine s strani strokovnjaka. Napaka je lahko večja za luksuzne nepremičnine, stanovanja v centru mesta in druge nestandardne opcije [37].

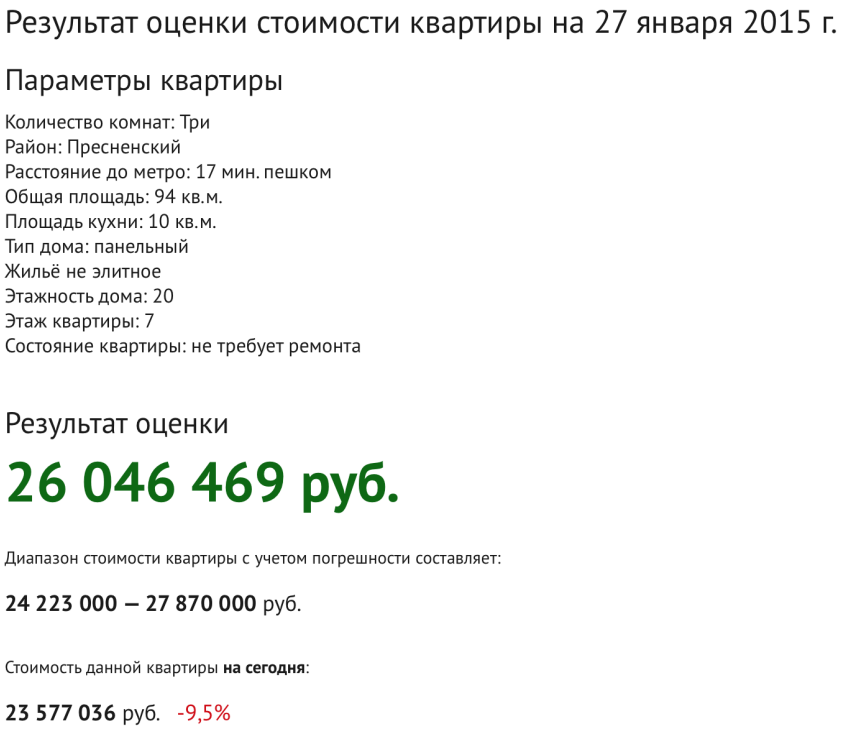
Na slikah 5 in 6 je prikazana ocena vrednosti za eno izmed nepremičnin v podatkih. Nepremičnina se nahaja v rajonu Presnenskij, od metroja je oddaljena 17 minut, zgrajena je iz panelov, vseh nadstropij je 20, stanovanje se nahaja na 7. nadstropju, celotna površina stanovanja meri 94 m^2, površina kuhinje je 10 m^2, nepremičnina ne potrebuje obnove, izbran datum prodaje je 27. 1. 2015. Nepremičnina je bila ocenjena na 26.046.469 rub. Upoštevajoč napako je nepremičnina v rangu od 24.223.000 rub do 27.870.000 rub. Dejanska cena v podatkih za nepremičnino je bila 27.928.732 rub, kar je za 1.882.263 rub več kot ocenjena vrednost. To predstavlja približno 6,7 % napako, kar je malo več od pričakovane 5 % napake.

Podobne storitve za oceno vrednosti nepremičnin lahko najdemo tudi na drugih spletnih straneh. Večina jih deluje tako, da glede na vnešene podatke (npr. površina, število sob, lokacija) v bazi podatkov (npr. baza podatkov o preteklih prodajah ali baza oglasov za nepremičnine) poiščejo podobne nepremičnine in tako določijo ceno. Nekateri izmed njih so zelo enostavni, drugi so malo bolj kompleksni in poleg primerjave cen podobnih stanovanj uporabljajo še različne modele, da izboljšajo natančnost.

Iz tega lahko sklepamo, da je uporaba podatkovnga rudarjenja za napovedovanje cen nepremičnin smiselna, saj so podobni postopki že v uporabi in je njihova natančnost dovolj dobra. S podatkovnim rudarjenjem bi najverjetneje dobili podobne rezultate.



Slika 5: Ocena vrednosti stanovanja



Slika 6: Rezultati ocene vrednosti stanovanja

4 Metodologija dela s podatki

Pred samim izvajanjem podatkovnega rudarjenja in uporabo različnih algoritmov je potrebno dobro razumeti problem, ki ga želimo rešiti in podatke, ki bodo za to uporabljeni. Poleg samega razumevanja podatkov je potrebno tudi to, da podatke uredimo in jih tako pripravimo na naslednjo fazo, kjer bodo za napovedovanje cen nepremičnin na podatkih uporabljeni različni modeli.

Če upoštevamo CRISP-DM procesni model bodo v tem poglavju združene prve tri faze: razumevanje problema, razumevanje podatkov in priprava podatkov. Večina informacij o splošnem stanju na ruskem nepremičninskem trgu, ki se nanašajo na fazo razumevanja problema, je bilo podanih že v prejšnjem poglavju. Tukaj pa bo poudarek predvsem na razumevanju podatkov ter na njihovem urejanju in pripravi za naslednjo fazo.

4.1 Programsko orodje Weka

Za namene podatkovnega rudarjenja bo v tem primeru uporabljen program Weka (Waikato Environment for Knowledge Analysis) [32]. Program vključuje različna orodja za predobdelavo podatkov, klasifikacijo, regresijo, združevanje v skupine, asocicijska pravila in vizualizacijo. Glavni del programa so v javi implementirani različni algoritmi za uporabo pri podatkovnem rudarjenju in strojnem učenju.

Program sestavlja več delov: Explorer, Expreimenter, KnowledgeFlow, SimpleCLI in Workbench. Explorer omogoča nalaganje podatkov, vizualizacijo in uporabo različnih algoritmov. Experimenter je namenjen za preizkuševanje in primerjavo algoritmov, pri tem je možna primerjava enega ali večih različnih algoritmov z izbranimi parametri. KnowledgeFlow omogoča načrtovanje konfiguracij za tokovno procesiranje podatkov. SimpleCLI je namenjen pisanju in izvajanju ukazov iz ukazne vrstice. Workbench pa je nekakšna kombinacija vseh štirih. Za namene te naloge bo zadoščala uporaba Explorerja in Expreimenterja.

Program Weka bo uporabljen predvsem v naslednji fazi, kjer bodo za napovedovanje cen nepremičnin iz podatkov uporabljeni različni napovedni modeli. Za to bo uporabljen Experimenter. Program pa je omenjen že na tem mestu, saj lahko v Explorerju pogledamo grafe za različne atribute, ugotovimo koliko je manjkajočih vrednosti

za posamezen atribut, kakšen je tip atributa, kakšna je povprečna vrednost zanj in podobno, kar koristi pri razumevanju in kasneje pri predobdelavi podatkov.

4.2 Opis podatkov

Podatki, ki bodo uporabljeni za napovedovanje cen nepremičnin za Moskvo in okolico so dobljeni s spletne strani `https://www.kaggle.com/c/sberbank-russian-housing-market`. Na omenjeni spletni strani so dostopni še mnogi drugi nabori podatkov za podatkovno rudarjenje in strojno učenje.

Podatki so bili zbrani s strani največje ruske banke Sberbank, najverjetneje iz podatkov o hipotekah in predstavljajo transakcije od avgusta 2011 do junija 2015. Vsebujejo obširen nabor podatkov o lastnostih nepremičnin. Poleg tega vsebujejo tudi podatke o makroekonomskih dejavnikih, ki opisujejo splošne stanje v ruski ekonomiji in finančnem sektorju za posamezen datum.

Cilj je na podlagi teh podatkov zgraditi takšen model oziroma modele, ki bodo omogočali napovedanje cen za nepremičnine. To bi bilo zelo uporabno tudi za banke, saj bi takšni modeli omogočali veliko enostavnejše določanje vrednosti nepremičnin za hipoteke. Trenutno je v Rusiji pri najemu hipotekarnega kredita potrebna ocena nepremičnine s strani usposobljenega ocenjevalca, ki nepremičnino pregleda in oceni. Uporaba napovednih modelov za ocenjevanje, pa bi vse skupaj poenostavila.

4.3 Struktura podatkov

Podatki vsebujejo veliko različnih atributov, vseh skupaj je 390. Atribute lahko razdelimo v dve skupini. Prvih 290 atributov opisuje podatke o nepremičnini in njeni okolici. Drugih 100 atributov opisuje različne makroekonomske indikatorje na dan nakupa oziroma prodaje nepremičnine.

Podatki za posamezno nepremičnino vsebujejo osnovne podatke o nepremičnini, kot so celotna površina, bivanjska površina, nadstropje, najvišje nadstropje, material, leto izgradnje, število sob, površina kuhinje in stanje. Poleg tega so podani podatki o lokaciji, kjer se nepremičnina nahaja: ime rajona (eden izmed 146 rajonov v Moskvi), število prebivalcev, delež zelenih površin, delež industrijskih površin, število različnih ustanov in objektov (vrtci, šole, izobraževalni centri, univerze, športni objekti, kulturni objekti, nakupovalni centri, uradi), prisotnost različnih vrst industrije, sestava prebivalstva po spolu in po letih, število zgrajenih stavb po materialu in po letih. Za vsako nepremičnino posebej so zapisane tudi oddaljenosti (do zelenih površin, industrije, vrtca, šole, železniške postaje, metroja, avtobusnega terminala, parka, večje ceste, nuklearnega reaktorja, stadiona, bazena, fitnesa, cerkve, muzeja, gledališča, itd.).

Podani so še podatki o številu različnih objektov ter povprečne cene za račun v kavarni ali restavraciji glede na različne oddaljenosti (0.5, 1, 2, 3 in 5 km) .

Makroekonomski podatki vsebujejo zelo različne podatke npr. podatke o rasti bruto domečega proizvoda, inflaciji, višina indeksa za izmenjavo med rubljem in dolarjem, uvozni in izvozni kapital, višina obrestne mere, število hipotekarnih posojil, povprečen dohodek na prebivalca, povprečna mesečna plača, delež nezaposlenih, število porok in ločitev, rast prebivalstva, povprečna življenjska doba, povprečna cena najemnine za stanovanja glede na število sob, število obiskov pri zdravniku, število obiskov gledališča, delež prebivalstva, ki se redno ukvarja s športom, itd.

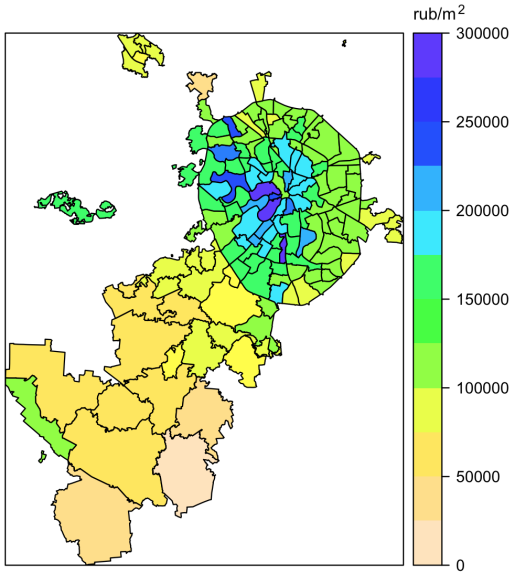
4.4 Razumevanje podatkov

Pred predobdelavo podatkov je dobro podatke malo pregledati tako, da jih bolje razumemo. Na ta način si lahko tudi poenostavimo naslednjo fazo predobdelave in urejanja podatkov, saj lahko že tukaj ugotovimo kje so določene manjkajoče vrednosti, napačne vrednosti, kakšne vrednosti imajo določeni atributi, kateri izmed atributov so pomembnejši in podobno.

4.4.1 Uporaba različnih grafov

Podatke lahko najenostavneje pregledamo v obliki tabele, takšne kot so. Tu lahko pogledamo katere stolpce oziroma atribute imajo. Podobno kot je to opisano prej pri strukturi podatkov. Poleg tega lahko za atribute pogledamo katere vrednosti imajo. Pri tem je najlažje, da si za posamezen atribut uredimo podatke po velikosti, najprej v eno in potem še v drugo smer. Tako lahko ugotovimo tudi kakšne so mejne vrednosti, kjer največkrat pride do napak. Obenem lahko ugotovimo tudi približno koliko je manjkajočih vrednosti za posamezen atribut. Ta način je zelo enostaven, vendar je primeren le za nekakšen osnoven vpogled v podatke.

Veliko boljši način pa je uporaba različnih grafov, histogramov in podobnega. Tako si podatke lahko predstavljamo bolj vizualno, kar nam je včasih lahko v veliko pomoč. Takšen primer je na sliki 7, kjer vidimo kakšna je povprečna cena za kvadratni meter v posameznem rajonu v Moskvi. Iz tega lahko ugotovimo, da je cena odvisna tudi od lokacije oziroma rajona, kjer se nepremičnina nahaja. Najvišje cene za kvadratni meter so bližje centru, v okolici pa je cena nižja. Če bi izpisali imena rajonov, bi lahko ugotovili točno v katerih rajonih so nepremičnine najdražje in kje so najcenejše. S preučevanjem samih podatkov tega ne bi mogli ugotoviti na tako enostaven način. Podobne grafe bi lahko naredili še npr. za povprečno leto izgradnje ali povprečno število nadstropij za posamezen rajon.



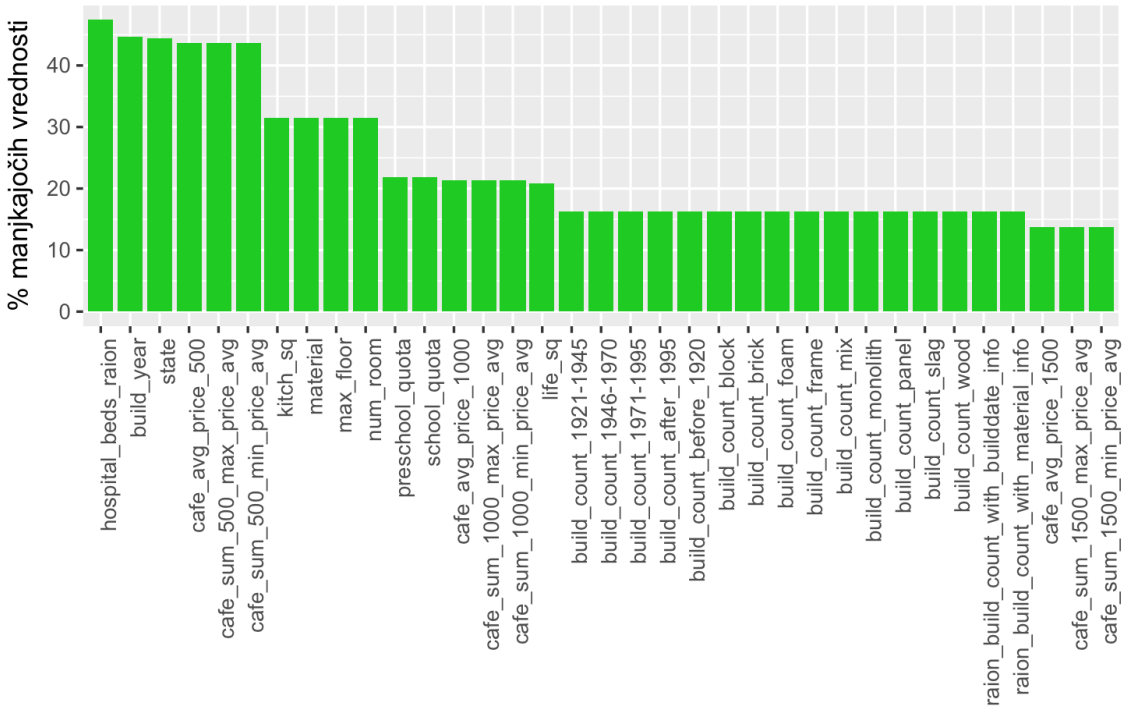
Slika 7: Povprečne cene nepremičnin v rubljih za kvadratni meter za Moskvo

Različni grafi so zelo uporabni, saj tako dobimo boljši vpogled v podatke. Lahko ugotavljamo kakšna so razmerja med različnimi atributi in kakšne so vrednosti za posamezne atribute. Poleg tega lahko že iz grafov opazimo tudi določene nepravilnosti v podatkih. Zato je koristno, da si pri preučevanju podatkov pomagamo tudi z različnimi grafi. To je dobro storiti še posebej za pomembnejše atribute in za atribute, kjer lahko pričakujemo nepravilne vrednosti. Pri tem si lahko pomagamo z Weko, kjer lahko pogledamo osnovne grafe za atribute. Primeri za to so na slikah v nadaljevanju. Ker je v Weki nabor grafov omejen, so ostali grafi narejeni v programskem jeziku R [30].

4.4.2 Manjkajoče vrednosti

V podatkih so tudi manjkajoče vrednosti. Razlog za to je, da včasih ni mogoče dobiti vseh podatkov, lahko se npr. zgodi, da ni mogoče dobiti podatka kdaj je bila nepremičnina zgrajena ali podatka v kakšnem stanju je nepremičnina.

Ker je v podatkih kar nekaj manjkajočih vrednosti, lahko to predstavlja težavo in moramo zato preveriti kje se manjkajoče vrednosti pojavljajo. Histogram na sliki 8 prikazuje delež manjkajočih vrednosti za vse atribute, ki imajo več kot 10 % manjkajočih vrednosti. Glede na to, da je vseh atributov 290, ima le 35 atributov več kot 10 % manjkajočih vrednosti, lahko ugotovimo, da v podatkih ni toliko manjkajočih vrednosti. Kljub temu pa je iz histograma vidno, da imajo velik delež manjkajočih vrednosti ravno nekateri pomembnejši atributi kot so: leto izgradnje (build_year), stanje (state), površina kuhinje (kitch_sq), material (material), najvišje nadstropje (max_floor), število sob (num_room), bivalna površina (life_sq).



Slika 8: Atributi z največjim deležem manjkajočih vrednosti

Po drugi strani je zelo malo manjkajočih vrednosti pri različnih podatkih o območju kjer se nepremičnina nahaja. Razlog za to je, da so pri pripravi podatkov najverjetneje poznali lokacijo nepremičnine in so na podlagi tega dobili ostale podatke. Podobno je s podatki o makroekonomskem stanju, kjer je prav tako zelo malo manjkajočih vrednosti.

Težava glede manjkajočih vrednosti je najverjetneje nastala že pri zbiranju podatkov. To lahko vidimo iz podatkov, saj imajo vse nepremičnine od avgusta 2011 do junija 2013 podano samo celotno površino, bivanjsko površino in nadstropje. Omenjeni podatki pa nimajo podanega najvišjega nadstropja, materiala, leta izgradnje, števila sob, površine kuhinje in stanja. Za nepremičnine s poznejšimi datumi se manjkajoče vrednosti še vedno pojavljajo v podatkih, vendar jih je veliko manj.

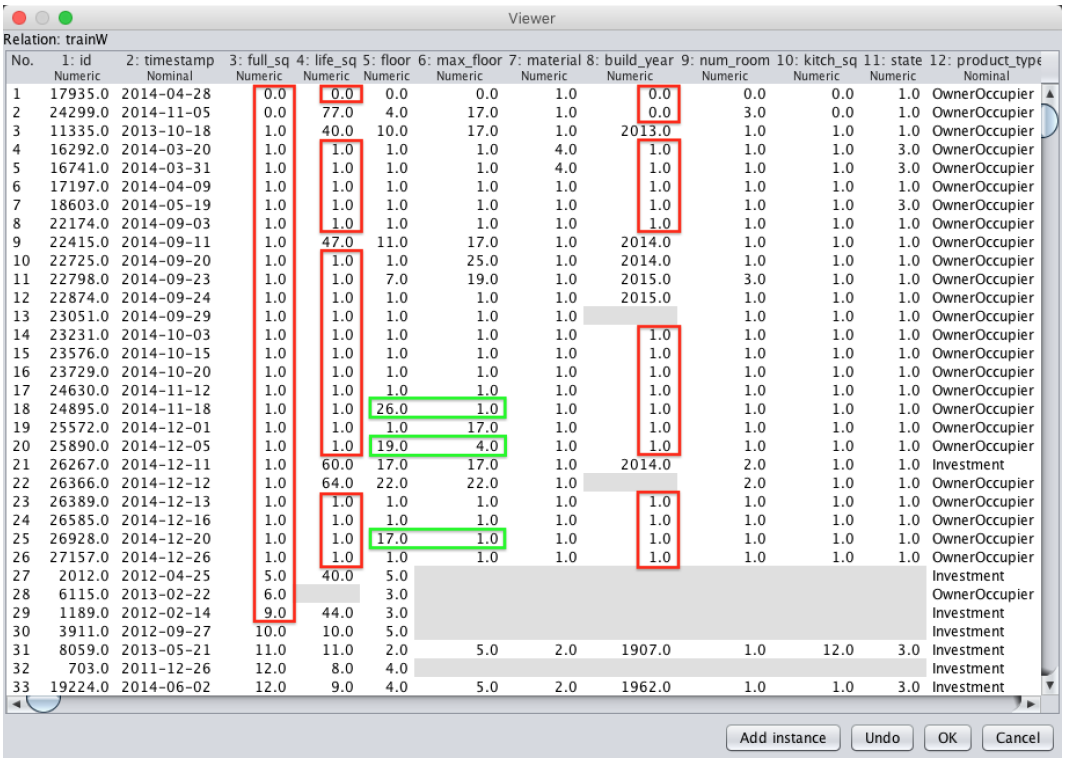
4.4.3 Nepravilne vrednosti

Ker so podatki zelo obsežni vsebujejo tudi kar nekaj nepravilnih vrednosti. Težava se, podobno kot pri manjkajočih vrednostih, pojavlja predvsem pri osnovnih podatkih za nepremičnino kot so: celotna in bivanjska površina, nadstropje, najvišje nadstropje, število sob, leto izgradnje. Pri ostalih atributih ni tako opaznih nepravilnih vrednosti.

Za ugotavljanje, kje se pojavljajo nepravilne vrednosti, je potrbno podatke dobro pregledati. Predvsem moramo biti pozorni pri najmanjših in največjih vrednostih, kjer so ponavadi napake. Na ta način lahko najdemo nepravilne vrednosti, ki odstopajo od

ostalih podatkov npr. celotna površina velikosti 1 m^2 ali površina kuhinje 100 m^2. Pri ostalih nepravilnih vrednostih pa moramo paziti predvsem na podatke, ki se ne zdijo preveč smiselni.

Najenostavnejši način za iskanje nepravilnih vrednosti je, da s primernim programom (lahko tudi z Weko, kot je primer na sliki 9) odpremo datoteko s podatki in jih za nek atribut uredimo po velikosti ter poskušamo ugotoviti ali so kakšne neskladne vrednosti. Posebej moramo paziti pri najmanjših in največjih vrednostih. To storimo predvsem za tiste atribute, kjer je večja možnost napak.



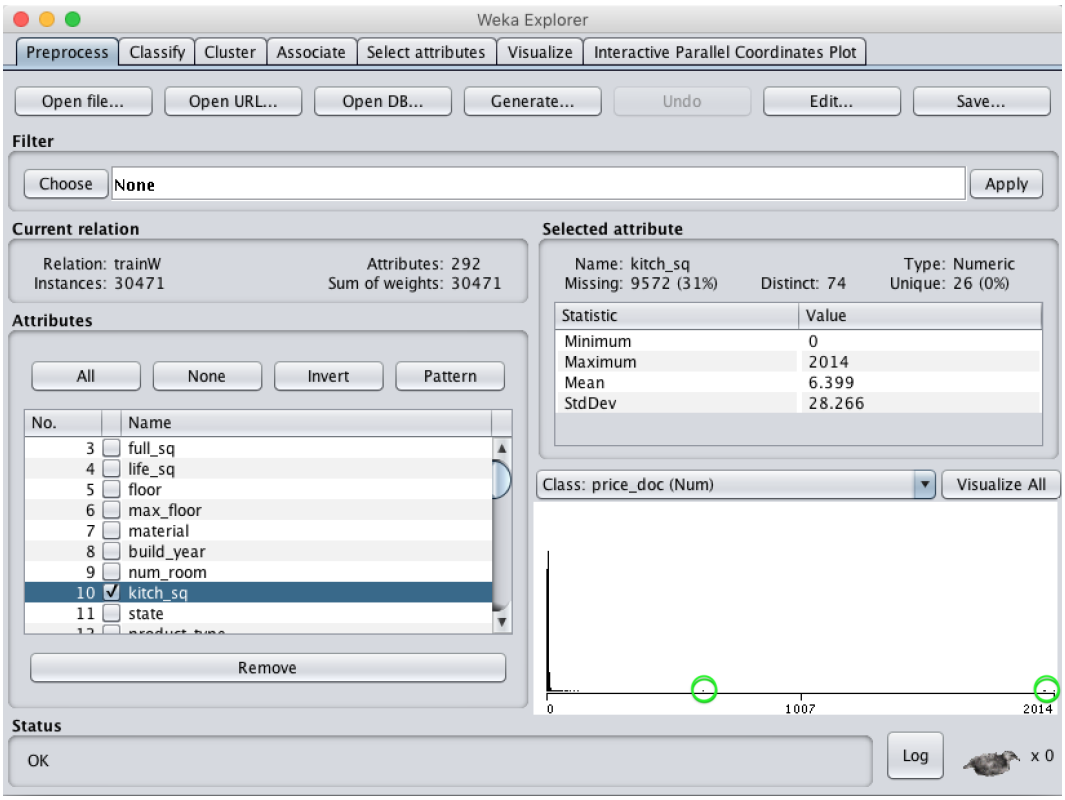
No.	1: id	2: timestamp	3: full_sq	4: life_sq	5: floor	6: max_floor	7: material	8: build_year	9: num_room	10: kitch_sq	11: state	12: product_type
	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
1	17935.0	2014-04-28	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	OwnerOccupier
2	24299.0	2014-11-05	0.0	77.0	4.0	17.0	1.0	0.0	3.0	0.0	1.0	OwnerOccupier
3	11335.0	2013-10-18	1.0	40.0	10.0	17.0	1.0	2013.0	1.0	1.0	1.0	OwnerOccupier
4	16292.0	2014-03-20	1.0	1.0	1.0	1.0	4.0	1.0	1.0	1.0	3.0	OwnerOccupier
5	16741.0	2014-03-31	1.0	1.0	1.0	1.0	4.0	1.0	1.0	1.0	3.0	OwnerOccupier
6	17197.0	2014-04-09	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
7	18603.0	2014-05-19	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	3.0	OwnerOccupier
8	22174.0	2014-09-03	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
9	22415.0	2014-09-11	1.0	47.0	11.0	17.0	1.0	2014.0	1.0	1.0	1.0	OwnerOccupier
10	22725.0	2014-09-20	1.0	1.0	1.0	25.0	1.0	2014.0	1.0	1.0	1.0	OwnerOccupier
11	22798.0	2014-09-23	1.0	1.0	7.0	19.0	1.0	2015.0	3.0	1.0	1.0	OwnerOccupier
12	22874.0	2014-09-24	1.0	1.0	1.0	1.0	1.0	2015.0	1.0	1.0	1.0	OwnerOccupier
13	23051.0	2014-09-29	1.0	1.0	1.0	1.0	1.0		1.0	1.0	1.0	OwnerOccupier
14	23231.0	2014-10-03	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
15	23576.0	2014-10-15	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
16	23729.0	2014-10-20	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
17	24630.0	2014-11-12	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
18	24895.0	2014-11-18	1.0	1.0	26.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
19	25572.0	2014-12-01	1.0	1.0	1.0	17.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
20	25890.0	2014-12-05	1.0	1.0	19.0	4.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
21	26267.0	2014-12-11	1.0	60.0	17.0	17.0	1.0	2014.0	2.0	1.0	1.0	Investment
22	26366.0	2014-12-12	1.0	64.0	22.0	22.0	1.0		2.0	1.0	1.0	OwnerOccupier
23	26389.0	2014-12-13	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
24	26585.0	2014-12-16	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
25	26928.0	2014-12-20	1.0	1.0	17.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
26	27157.0	2014-12-26	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	OwnerOccupier
27	2012.0	2012-04-25	5.0	40.0	5.0							Investment
28	6115.0	2013-02-22	6.0		3.0							OwnerOccupier
29	1189.0	2012-02-14	9.0	44.0	3.0							Investment
30	3911.0	2012-09-27	10.0	10.0	5.0							Investment
31	8059.0	2013-05-21	11.0	11.0	2.0	5.0	2.0	1907.0	1.0	12.0	3.0	Investment
32	703.0	2011-12-26	12.0	8.0	4.0							Investment
33	19224.0	2014-06-02	12.0	9.0	4.0	5.0	2.0	1962.0	1.0	1.0	3.0	Investment

Slika 9: Primeri nepravilnih vrednosti v podatkih

Na sliki 9 je primer podatkov urejenih po velikost glede na atribut full_sq. Nepravilne vrednosti so obkrožene z zeleno in rdečo barvo. Iz slike je mogoče videti, da je v podatkih veliko neprevilnih vrednosti. Celotna površina (full_sq) je v nekaj primerih 0 ali 1. Prav tako je podatek za leto izgradnje (build_year) večkrat enak 0 ali 1. Poleg tega lahko najdemo še različne nepravilnosti npr. da je vrednost za nadstropje (floor) večkrat večja od vrednosti za najvišje nadstropje (max_floor) ali vrednost za bivanjsko površino (life_sq) večja od vrednosti za celotno površino (full_sq).

Pri iskanju nepravilnih vrednosti še za ostale atribute si lahko v Weki pomagamo tudi s histogrami, ki jih program izriše za vsakega izmed atributov. Čeprav takšni histogrami niso preveč pregledni (primer je na sliki 10), lahko hitro vidimo pri katerih atributih so kakšne neskladne vrednosti, ki jih potem popravimo v podatkih. Tako

lahko v primeru na sliki 10 vidimo, da so v podatkih za površino kuhinje (kitch_sq) dve vrednosti za leto in še neka druga nepravilna vrednost, ki jih je potrebno popraviti. Podobno lahko naredimo še za ostale pomembnejše atribute.



Slika 10: Primer nepravilnih vrednosti za atribut kitch_sq

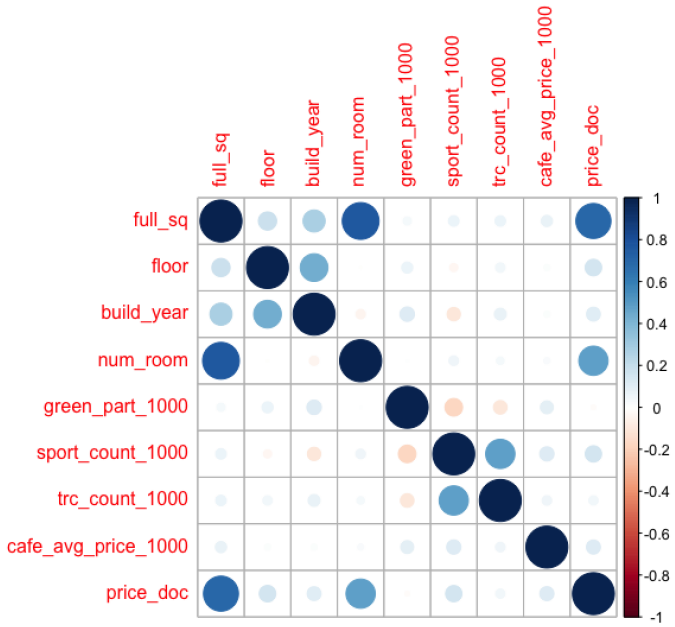
Po pregledu celotnih podatkov je mogoče ugotoviti, da je v podatkih veliko vrednosti za celotno površino, bivanjsko površino, površino kuhinje, najvišje nadstropje in leto izgradnje enako 0 ali 1. Te vrednosti si lahko razlagamo zelo različno. Pri letu izgradnje takšne vrednosti spremenimo v manjkajoče. Pri celotni in bivanjski površini si to razlagamo kot napačne vrednosti, ki jih spremenimo v manjkajoče vrednosti ali pa takšne primere izbrišemo, če so nepravilne vrednosti tudi pri drugih atributih. Pri atributu za površino kuhinje je še največ težav, saj 0 lahko predstavlja manjkajoč podatek ali podakek, da ni kuhinje. Podobno si lahko vrednost 1 razlagamo kot nepravilno vrednost ali kot dejansko vrednost, da ima nepremičnina zelo majhno kuhinjo. Veliko težav povzroča ravno vrednost 1, za katero je včasih težko ugotoviti ali predstavlja pravilno vrednost ali ne.

Poleg tega so v podatkih še različne druge nepravilne vrednosti npr. za veliko nepremičnin je vrednost za nadstropje večja kot za najvišje nadstropje, prav tako je vrednost za bivanjsko površino večkrat večja od celotne površine. Podobno je vrednost za površino kuhinje v veliko primerih enaka celotni površini, včasih je tudi večja.

4.4.4 Razumevanje vpliva atributov

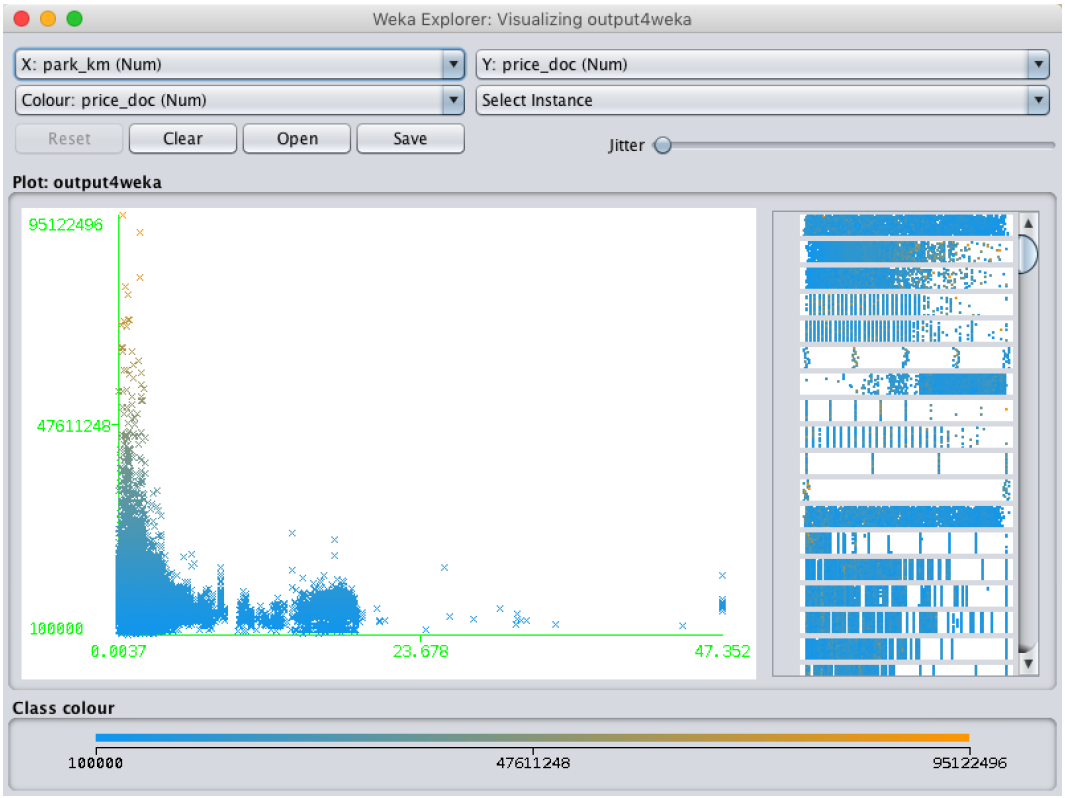
Za napovedovanje cen je pomembno, da razumemo, kako se cena nepremičnin spreminja glede na ostale dejavnike. V tem primeru jim lahko rečemo kar atributi. Pri tem nas zanima kateri atributi imajo večji vpliv, kateri manjši in kakšen je njihov vpliv na cene nepremičnin.

Pri ugotavljanju kako ostali atributi vplivajo na cene nepremičnin si lahko pomagamo tudi z grafi. Na sliki 11 je primer grafa, ki prikazuje korelacijo med nekaj izbranimi atributi. Iz grafa lahko vidimo, da ima cena nepremičnine (price_doc) največjo korelacijo s celotno površino (full_sq) in s številom sob (num_room). Manjšo korelacijo ima še z nadstropjem (floor), letom izgradnje (build_year), povprečno višino računa (cafe_avg_price_1000) in številom športnih objektov v okolici 1 km (sport_count_1000). Zanemarljivo majhno korelacijo ima z deležem zelenih površin (green_part_1000) in številom nakupovalnih središč (trc_count) v okolici 1 km. Iz grafa so vidne tudi korelacije ostalih atributov med seboj.



Slika 11: Korelacija med različnimi atributi

Povezave med različnimi atributi lahko ugotavljamo tudi iz grafov v Weki, kjer je mogoče pogledati graf za poljubna dva atributa. Primer je na sliki 12. Graf prikazuje odvisnost cene nepremičnine (price_doc) od oddaljenosti od parka (park_km). Iz grafa je vidno, so dražje nepremičnine ponavadi zelo blizu kakšnega parka. Približno polovica cenejših nepremičnin je še vedno zelo blizu parka, druga polovica pa je malo bolj oddaljena. Takšne grafe lahko pogledamo še za ostale pare atributov in ugotavljamo kakšna je povezanost med njimi.



Slika 12: Graf za atributa park_km in price_doc v Weki

4.5 Predobdelava podatkov

Podatke je potrebno še urediti in obdelati tako, da so pripravljeni za izvajanje podatkovnega rudarjenja in uporabo različnih algoritmov strojnega učenja. Ta faza je zelo pomembna, saj je od nje odvisna kakovost dobljenih napovednih modelov. Če v tej fazi na primer ne popravimo napačnih vrednosti ali vključimo atribute, ki so nepotrebni bo to vplivalo na model, ki bo zato slabši. Zato moramo v tej fazi popraviti vse morebitne nepravilne vrednosti in izbrati kateri atributi bodo vključeni v model in kateri ne.

4.5.1 Manjkajoče in nepravilne vrednosti

Glede manjkajočih vrednosti ne moremo kaj dosti storiti, lahko le izbrišemo tiste primere, ki imajo več manjkajočih vrednosti za pomembnejše atribute kot so površina, nadstropje, število sob, material izgradnje. Lahko bi tudi sami poskušali določene manjkajoče vrednosti vpisati v podatke, npr. če bi imeli podano površino ne pa števila sob bi lahko predvidevali, da ima nepremičnina s površino 60 m^2 3 sobe, s 48 m^2 pa 2 sobi. Vendar je to že preveč spreminjanja podatkov. Kljub temu pa obstaja možnost, da bi z vpisovanjem manjkajočih vrednosti dobili boljši model od tistega, ki upošteva manjkajoče vrednost.

Veliko večji vpliv na dobljen napovedni model od manjkajočih vrednosti imajo nepravilne vrednosti. Teh je v podatkih veliko, kot je bilo to ugotovljeno v prejšnjem poglavju. Največ takšnih vrednosti je ravno pri najpomembnejših atributih, to so: celotna površina, bivanjska površina, nadstropje, najvišje nadstropje, leto izgradnje in površina kuhinje.

Kot je bilo že prej omenjeno, je veliko vrednosti za leto izgradnje in površino kuhinje 0 in 1. Nekaj takšnih vrednosti je tudi pri drugih atributih. Vse takšne vrednosti je potrebno ustrezno popraviti ali spremeniti v manjkajoče vrednosti. Zelo pogosta napaka v podatkih je, da je vrednost za nadstropje večja od vrednosti za najvišje nadstropje. Podobno je včasih bivanjska površina večja od celotne površine, površina kuhinje pa enaka celotni površini. Vse to je potrebno popraviti. Primere z več napakami pa je najboljše izbrisati iz podatkov.

Zaradi nepravilnih vrednosti bo uporabljena samo približno ⅓ podatkov, kjer je tip nepremičnine OwnerOccupier. Ostali podatki, približno ⅔, kjer je tip Investment ne bodo uporabljeni, saj je veliko cen v podatkih prenizkih. Razlog za to je, da so te nepremičnine najverjetneje nove, nekatere izmed cen pa so predstavljene kot nižje zaradi nižjih davkov.

4.5.2 Osamelci

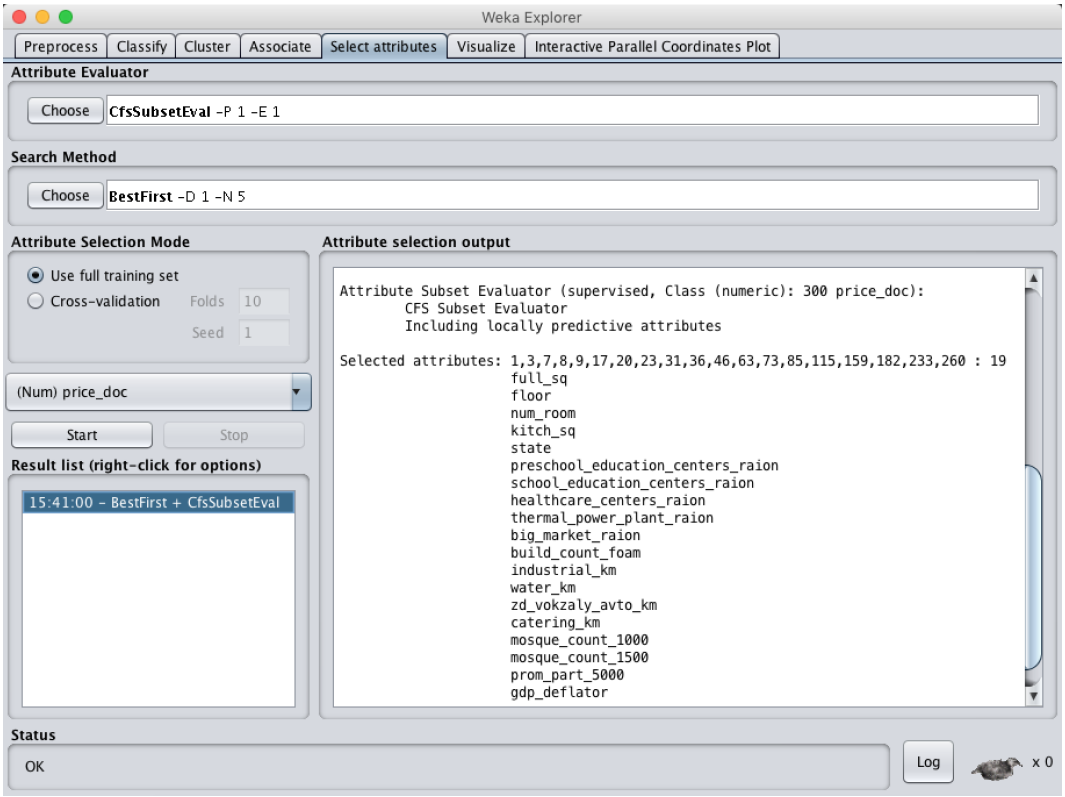
Osamelci predstavljajo vrednosti, ki preveč odstopajo od ostalih in zato lahko kasneje vplivajo na izgradnjo modela. Najdemo jih podobno kot nepravilne vrednosti, saj so to ponavadi vrednosti, ki prav tako kot nepravilne vrednosti odstopajo od ostalih podatkov. Tako so vrednosti zanje veliko večje ali veliko manjše od ostalih vrednosti. Primer za to je celotna površina nepremičnine velikosti 390 m^2, kar ni nepravilna vrednost, če imamo večjo nepremičnino ali število vseh nadstropij 40, kar je tudi možno, vendar oba podatka preveč odstopata od povprečja, kar lahko kasneje vpliva na dobljen model. Podobno je tudi s cenami nepremičnin, kjer moramo odstraniti vrednosti, ki preveč odstopajo od povprečja.

4.5.3 Izbira atributov

V podatkih je zelo veliko različnih atributov. To je dobro, saj lahko izbiramo med različnimi atributi in izberemo tiste, ki so zares pomembni. Vendar moramo paziti, da ne izberemo atributov, ki bi model poslabšali npr. da ne izberemo atributov, ki so med seboj preveč podobni, imajo preveliko korelacijo ali pa atributov, ki so za izgradnjo modela nerelavantni. Poleg tega je dobro, da omejimo število atributov, saj tako preprečimo preveliko prilagajanje modela podatkom (ang. overfitting). Druga prednost manjšega števila atributov je, da se zmanjša čas potreben za izgradnjo modelov.

V podatkih najdemo veliko atributov, ki so za cene nepremičnin nerelavantni npr. število prebivalcev posameznega rajona po starostnih skupinah, število migracij, število obiskov muzejev in gledališč, delež ljudi, ki se ukvarja s športom, ipd. Vse take atribute lahko odstranimo. Podobno lahko odstranimo atribut id in vse atribute, ki vsebujejo id-je npr. id najbližje železniške postaje, id najbližje ceste. To moramo storiti, saj bi jih Weka napačno upoštevala kot zvezne atribute in ne kot diskretne atribute. Odstranimo tudi atribut timestamp, ki pove čas nakupa oziroma prodaje nepremičnine, saj so glede na ta atribut upoštevani že makroekonomski podatki. Poleg tega lahko izločimo tudi tiste atribute, ki imajo zelo majhno varianco oziroma imajo za vse primere približno enake vrednosti, kar nam pri gradnji modela ne pomaga veliko.

Atribute za katere ne vemo, ali jih je bolje pustiti ali odstraniti iz podatkov, lahko pustimo v podatkih in jih kasneje izločimo v Weki. Weka nam lahko pomaga tudi pri izbiri atributov. Primer izbire atributov za podatke o nepremičninah je na sliki 13, kjer so bili izbrani atributi, ki imajo veliko korelacijo z atributom za ceno nepremičnine in majhno korelacijo med seboj.



Slika 13: Izbira najboljših atributov v Weki

Glede na izbrane atribute v Weki prikazane na sliki 13 lahko ugotovimo, da na ceno nepremičnine vplivajo osnovne lastnosti kot so celotna površina, nadstropje, število sob, stanje nepremičnine. Poleg tega na ceno vplivajo tudi drugi dejavniki npr. število

vrtcev, šol, bolnišnic, velikih trgovskih centrov, ki se nahajajo v rajonu. Pomembna je tudi oddaljenost od železniške postaje, oddaljenost od vode npr. reke ali jezera, ter delež industrijskih površin, število mošej v okolici, prisotnost termoelektrarne in izmed makroekonomskih podatkov deflator bruto domačega proizvoda.

Razlog zakaj so bili izbrani ravno ti atributi je v tem, da Weka izbira atribute, ki imajo veliko korelacijo s ceno nepremičnine in majhno korelacijo med seboj. Pri tem pa ni pomembno kaj ti atributi predstavljajo. Tako so bili v tem primeru izbrani tudi atributi, ki povejo kakšna je oddaljenost do cateringa ali število mošej v območju 1 km in 1.5 km. Omeniti je potrebno, da je bil izmed vseh makroekonomskih atributov izbran samo atribut za deflator bruto domačega proizvoda. To lahko pojasnimo na enak način kot prej, saj imajo makroekonomski atributi verjetno veliko korelacijo med seboj, zato je bila dovolj izbira samo enega atributa.

Pri napovedovanju cen nepremičnin bodo atributi za primerjavo izbrani na tri različne načine. V prvem primeru bodo uporabljeni vsi atributi, razen nepomembnih atributov, ki so bili odstranjeni. To je skupaj 300 atributov, ki so opisani v prilogi A. V drugem primeru bodo uporabljeni atributi, ki so bili izbrani v Weki, kot je bilo to prej opisano in prikazano na sliki 13. Tako izbranih atributov je 20 in so opisani v prilogi B. V tretjem primeru pa bodo izbrani tisti atributi, ki ustrezajo atributom za oceno stanovanja na irn.ru, kot je to prikazano v poglavju Napovedovanje cen nepremičnin za Moskvo in okolico na sliki 5. V tem primeru bodo izbrani atributi: rajon, razdalja do metroja, tip nepremičnine glede na izgradnjo, število vseh nadstropij, nadstropje stanovanja, število sob, celotna površina, površina kuhinje in stanje. Izmed makroekonomskih atributov pa bo kot v drugem primeru izbran atribut za deflator bruto domačega proizvoda. Omenjenih atributov je 11 in so opisani v prilogi C.

Z izbiro različnih atributov bomo lahko ugotovili kako na dobljene modele in rezultate poleg uporabe različnih algoritmov vpliva tudi izbira atributov. Zanimalo nas bo predvsem kako število in izbira atributov vplivajo na natančnost dobljenih modelov ter kakšna izbira atributov je optimalna.

4.5.4 Pretvorba v ARFF format

Za uporabo podatkov v Weki jih je potrebno predhodno pretvoriti v ARFF format (ang. Attribute-Relation File Format), ki ga uporablja Weka [32]. Podatke lahko odpremo tudi če so v drugih formatih npr. CSV in jih potem shranimo kot ARFF. V obeh primerih je potrebno paziti predvsem pri posebnih znakih in manjkajočih vrednostih. Posebne znake lahko iz podatkov izbrišemo, če niso pomembni, drugače uporabimo navednice. Vse manjkajoče vrednosti pa spremenimo v ?. Prav tako je pomembno, da atribut, ki predstavlja razred damo na zadnje mesto, da ga Weka avtomatsko prepozna.

5 Napovedovanje cen nepremičnin

Podatkovno rudarjenje želimo uporabiti za napovedovanje cen nepremičnin, tako da bomo lahko ceno za izbrano nepremičnino napovedali glede na znane spremenljivke (npr. površina, število sob, površina kuhinje, lokacija, itd.), nekaj podobnega kot je bilo predstavljeno v poglavju Napovedovanje cen nepremičnin za Moskvo in okolico, le da bo v tem primeru uporabljeno podatkovno rudarjenje.

Za izgradnjo napovednih modelov bodo uporabljeni različni regresijski algoritmi, ki bodo opisani v nadaljevanju. Ker želimo na koncu tudi primerjati dobljene rezultate bodo opisani tudi načini testiranja in ocenjevanja dobljenih modelov. Za oboje, tako izgradnjo modelov kot testiranje in ocenjevanje, bo uporabljena Weka.

5.1 Načini napovedovanja

Glavna cilja podatkovnega rudarjenja sta opisovanje podatkov in napovedovanje. Opisovanje se osredotoča na iskanje človeku razumljivih vzorcev za opis podatkov. Cilj napovedovanja pa je na podlagi znanih spremenljivk napovedati neznane ali prihodnje vrednosti [13]. Podatkovno rudarjenjenje pozna različne tehnike za delo s podatki npr. asociacijo, klasifikacijo, regresijo, razvrščanje v skupine, povzemanje, idr. Izbira tehnike je odvisna predvsem od podatkov in od problema, ki ga želimo rešiti. V tem primeru bo za napovedovanje cen nepremičnin uporabljena regresija.

Regresija se uporablja za napovedovanje zveznih vrednosti, ki so ponavadi numerične [16]. Ključna ideja regresije je odkriti razmerja med odvisnimi in neodvisnimi spremenljivkami [22]. V primeru nepremičnin je cena nepremičnine odvisna spremenljivka, ki jo želimo napovedati, vsi ostali atributi pa predstavljajo neodvisne spremenljivke, ki vplivajo na ceno. Tako dobljen model bo prikazoval kako različne neodvisne spremenjlivke (npr. površina, število sob, leto izgradnje, itd.) vplivajo na ceno.

5.2 Algoritmi za izgradnjo napovednih modelov

Za napovedovanje cen nepremičnin si bomo pomagali z Weko, kjer bodo za igradnjo napovednih modelov uporabljeni različni regresijski algoritmi. V tem primeru so izbrani algoritmi: IBk [1], M5P [24] [31], M5Rules [18] [24] [31], REPTree [34], Ran-

domTree [34], RandomForest [5] in LinearRegression [34]. Algoritem IBk za napovedovanje vrednosti išče najbližje sosede. Algoritmi M5P, REPTree, RandomTree in RandomForest uporabljajo različna regresijska drevesa. Algoritem LinearRegression pa je implementacija linearne regresije. Poleg omenjenih algoritmov bosta za primerjavo uporabljena še algoritma ZeroR [34] in DecisionStump [34].

5.2.1 ZeroR

ZeroR je zelo enostaven algoritem. Če ga uporabljamo pri regresiji napove povprečno vrednost. To ni preveč uporabno za izdelavo kakršnihkoli napovednih modelov, vendar lahko na ta način ugotovimo kakšna je najmanjša pričakovana natančnost, ki jo lahko primerjamo z ostalimi algoritmi [34].

5.2.2 DecisionStump

DecisionStump zgradi eno nivojsko binarno odločitveno drevo, ki ima lahko še dodatno vejo za manjkajoče vrednosti. Ker se algoritem odloča samo na podlagi enega atributa, ne moremo pričakovati, da bomo na ta način dobili uporaben napovedni model. Zato bo ta algoritem uporabljen bolj za primerjavo z drugimi algoritmi. Kljub temu pa bomo na ta način ugotovili kateri izmed atributov je najpomembnejši in kakšen je njegov vpliv na končno napoved [16].

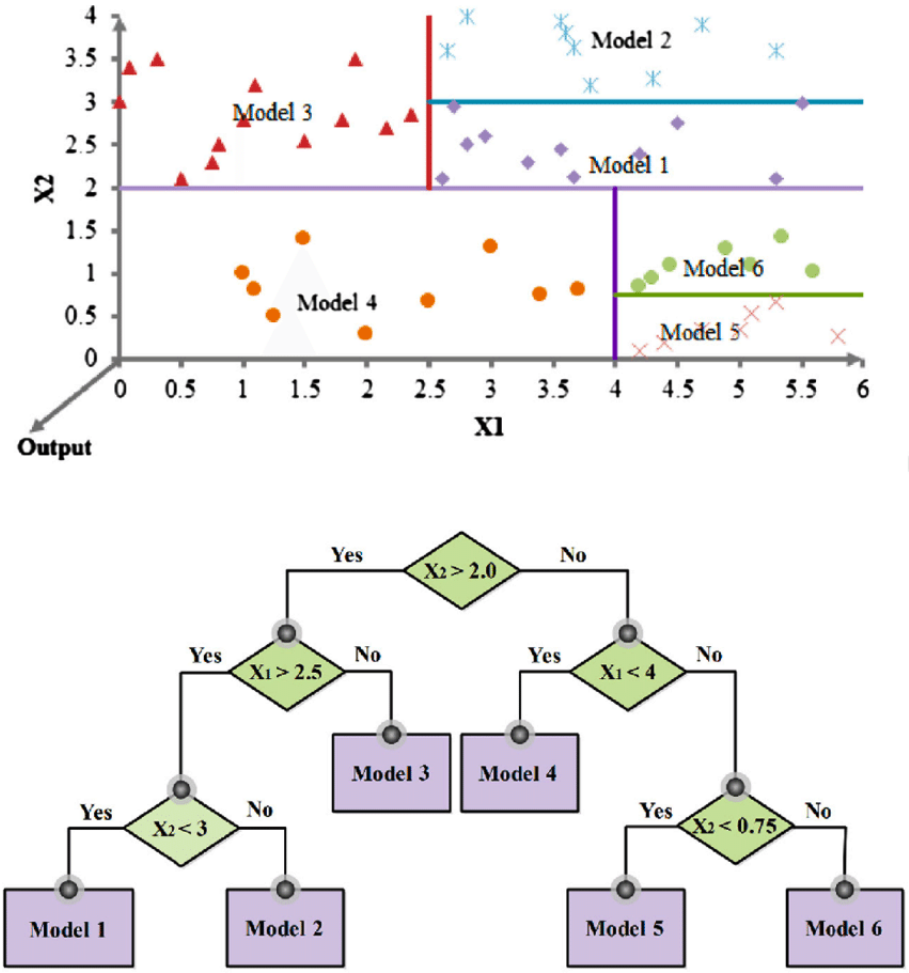
5.2.3 IBk

Algoritem IBk uporablja princip k najbližjih sosedov (ang. k-nearest neighbors). Pri regresiji vrednost za iskan primer napove glede na njegovih k najbližjih sosedov. Če je k enak 1 se primeru določi isto vrednost, kot jo ima najbližji sosed, če je k večji pa povprečno vrednost od vseh vrednosti sosedov. Pri iskanju sosedov se lahko uporabljajo različne razdalje, ki ponazarjajo kako blizu sta si posamezna primera oziroma kako podobna sta si, npr. evklidska ali manhattanska razdalja. Posebnost algoritma je, da v fazi učenja ne naredi nikakršnih pravil ali dreves kot večina ostalih algoritmov, temveč celotno učno množico shrani za kasnejšo fazo testiranja, kjer za vsak primer posebej poišče najbližje sosede [1] [16] [34].

5.2.4 M5P

M5P uporablja algoritem M5'. Najprej zgradi drevo, ki razdeli primere na več delov glede na standardni odklon med posameznimi primeri, tako da so podobni primeri skupaj. Po izgradnji drevesa za primere v vsakem vozlišču zgradi model z uporabo linearne regresije. Pri tem ne uporablja vseh atributov, ampak samo tiste, ki so v

poddrevesu vozlišča. Vsak takšen linearen model je na koncu tudi poenostavljen z odstranjevanjem neuporabnih atributov. Pri regresiji se primeru določi vrednost tako, da se gre najprej po drevesu do ustreznega lista in nato se uporabi linearni regresijski model za ta list. Lahko si predstavljamo, da algoritem primere razdeli na več delov oziroma skupin, kot je to prikazano na sliki 14, od katerih ima vsaka skupina svoj linearni model, ki ga predstavlja eden izmed listov drevesa [24] [34].



Slika 14: Algoritem M5' [4]

5.2.5 M5Rules

M5Rules za generiranje pravil uporablja M5' drevesa. Algoritem najprej zgradi M5' drevo na celotnih podatkih in nato obdrži le najboljši list iz katerega naredi pravilo, drevo pa zavrže. Po vsaki fazi izgradnje drevesa primere, ki so upoštevani z najboljšim pravilom izloči iz učnih podatkov pred izgradnjo naslednjega drevesa. Postopek nato ponavlja z gradnjo novih M5' dreves in generiranjem pravil, dokler niso vsi primeri pokriti z vsaj enim pravilom [34].

5.2.6 REPTree

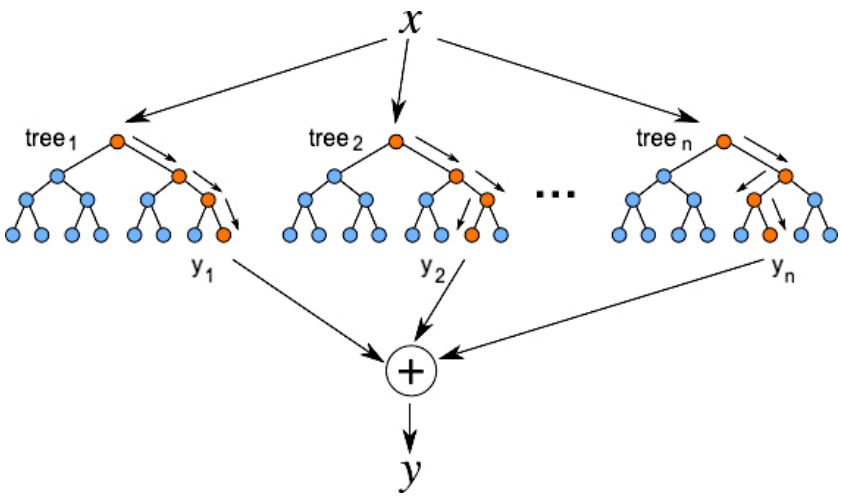
REPTree (Reduced Pruning Tree) zgradi regresijsko drevo upoštevajoč informacijski prispevek ali varianco. Učne podatke razdeli na dva dela, kjer prvi del, približno ⅔ uporabi za gradnjo drevesa, preostalo ⅓ zadrži in uporabi za rezanje zgrajenega drevesa. Algoritem je dodatno optimiziran, da je hitrejši, zato vse numerične atribute sortira samo enkrat na začetku [34].

5.2.7 RandomTree

RandomTree zgradi odločitveno drevo, tako da v vsakem vozlišču naključno izbere k atributov in izmed njih izbere najboljšega. Drevo zgradi do maksimalne velikosti in ga po izgradnji ne zmanjšuje z rezanjem. Algoritem ima tudi možnost, da izberemo vrednost za število atributov, ki bodo naključno izbrani v vsakem vozlišču [26] [34].

5.2.8 RandomForest

Algoritem RandomForest zgradi več odločitvenih dreves, od katerih je vsako zgrajeno na delu podatkov. Dejansko je za gradnjo posameznih dreves uporabljen algoritem RandomTree. Pri gradnji drevesa se v vozlišču naključno izbere k atributov, izmed katerih se za vozlišče izbere najboljšega. Tako zgrajenih dreves algoritem ne zmanjšuje z rezanjem. V primeru regresije algoritem napove povprečno vrednost od vseh vrednosti, ki jih je dobil po različnih drevesih, tako kot je prikazano na sliki 15. Prednost algoritma je, da takšen način izgradnje dreves preprečuje prekomerno prilagajanje modela podatkom. Poleg tega je končna odločitev narejena glede na rezultate dobljene pri večih drevesih, kar je boljše kot uporaba enega samega drevesa [5] [25] [34].



Slika 15: Algoritem RandomForest [7]

5.2.9 LinearRegression

LinearRegression izvaja linearno regresijo z več spremenljivkami. Ideja pri tem je, da se vrednost, ki jo želimo napovedati izrazi kot linearno kombinacijo ostalih atributov. Koeficiente se določi z metodo najmanjših kvadratov. Kot model dobimo enačbo, kjer je ciljna spremenljivka izražena z neodvisnimi spremenljivkami. Koeficienti v enačbi pa predstavljajo uteži za posamezne atribute neodvisnih spremenljivk [34].

5.3 Testiranje in ocenjevanje napovednih modelov

Pred izgradnjo modelov je potrebno določiti načine testiranja in ocenjevanja, ki bodo uporabljeni za ocenjevanje dobljenih napovednih modelov. To je zelo pomembno, saj lahko le tako ugotovimo kako uspešni dejansko so dobljeni modeli. Pri tem pa moramo izbrati tak način testiranja in ocenjevanja, ki bo za vse modele enak tako, da bomo rezultate lahko primerjali med seboj.

5.3.1 Testiranje napovednih modelov

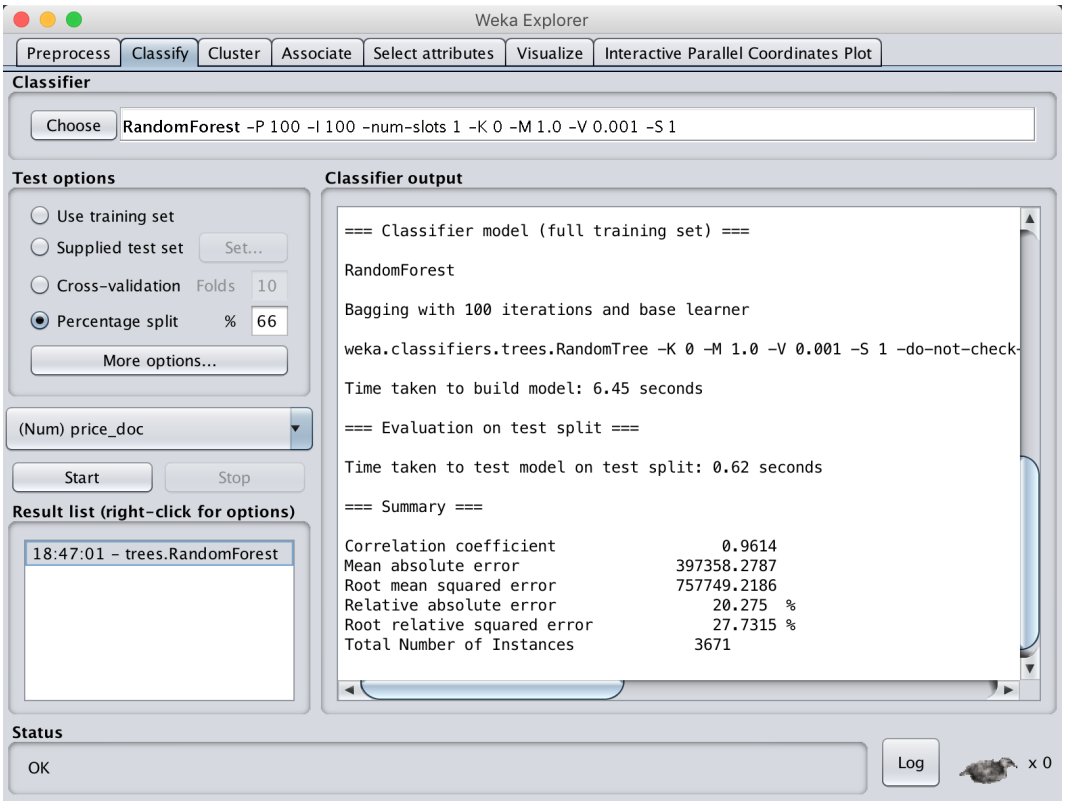
Da bi ugotovili kako dobri so dobljeni napovedni modeli, jih bomo testirali v Weki. Weka omogoča različne načine testiranja, kot je to vidno na sliki 16. Lahko uporabimo učno množico, ločeno testno množico, prečno preverjanje ali razdelitev na učno in testno množico. V tem primeru bo za hitrejše sprotno preverjanje rezultatov uporabljena razdelitev podatkov na učno in testno množico, za končne rezultate pa bo uporabljeno desetkratno prečno preverjanje.

Razdelitev podatkov na učno in testno množico

Najenostavnejša možnost za ocenjevanje modelov je razdelitev podatkov na učno in testno množico. Del podatkov, ki predstavlja učno množico uporabimo pri gradnji modela, drugi del, ki predstavlja testno množico pa uporabimo pri testiranju dobljenega modela. Navadno se uporabi ⅔ podatkov za učno množico in ⅓ za testno množico. Pomembno je, da sta množici neodvisni in da podatke iz ene množice uporabljamo samo za učenje oziroma samo za testiranje. Posebej je treba paziti tudi pri delitvi na učno in testno množico, saj morata biti obe množici dovolj veliki, da bosta model in ocenjena napaka dovolj dobra. Tukaj ni večjih težav, če imamo na voljo veliko podatkov. V primeru premalo podatkov pa se lahko zgodi, da je model slab zaradi premajhne učne množice ali da ocena napake ni dovolj natančna zaradi premajhne testne množice. Zato je takrat boljše uporabiti n-kratno prečno preverjanje, ki zmanjša določene slabosti testiranja z delitvijo na učno in testno množico [34].

N-kratno prečno preverjanje

N-kratno prečno preverjanje je zelo podobno razdelitvi na učno in testno množico. Lahko bi rekli, da je n-kratno prečno preverjanje nekakšna večkratna ponovitev razdelitve na učno in testno množico. Podatke je najprej potrebno razdeliti na n delov, ki so med seboj disjunktni. n-1 delov predstavlja učno množico in se uporabi za gradnjo modela. Preostali del predstavlja testno množico in se uporabi za testiranje zgrajenega modela. Takšen postopek se ponovi n krat, tako da vsak izmed n delov enkrat predstavlja testno množico. Ponavadi se uporablja desetkratno prečno preverjanje, kjer je n = 10, kar bo uporabljeno tudi v tem primeru. Ta postopek je uporaben predvsem takrat ko nimamo dovolj podatkov, da bi lahko imeli dovolj veliko učno in testno množico ali ko želimo tako za gradnjo modela kot testiranje uporabiti vse podatke [34].



Slika 16: Testiranje in ocenjevanje regresijskega modela v Wekinem Explorerju

5.3.2 Ocenjevanje napovednih modelov

Weka pozna več meril za ocenjevanje dobljenih modelov, kot je prikazano na sliki 16. Pri regresiji nas predvsem zanima kakšen je korelacijski koeficient (ang. correlation coefficient). To je mera, ki pove v kakšni korelaciji so napovedane vrednosti z dejanskimi vrednostmi v podatkih. Vrednosti za korelacijski koeficient so lahko med - 1 in

1, kjer 0 pomeni, da ni korelacije, 1 pomeni visoko korelacijo in - 1 obratno korelacijo. Obratne korelacije sicer pri napovednih modelih ne pričakujemo. Poleg tega nas zanimajo tudi kakšne so napake. Weka v primeru numeričnih napovedi oziroma regresije pozna štiri vrste napak. To so: povprečna absolutna napaka (ang. mean absolute error), koren povprečne kvadratne napake (ang. root mean squared error), relativna absolutna napaka (ang. relative absolute error) in koren relativne kvadratne napake (ang. root relative squared error). Način izračuna korelacijskega koeficienta in napak, ki se uporabljajo v Weki je prikazan na sliki 17, kjer so $p_1, p_2, ..., p_n$ napovedane vrednosti in $a_1, a_2, ..., a_n$ dejanske vrednsti. p_i je vrednost napovedana za i-ti primer. $\overline{a}$ je lahko povprečna vrednost za učne podatke ali povprečna vrednost za testne podatke.

Root mean-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$
Mean-absolute error	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{n}$
Root relative-squared error*	$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \overline{a})^2 + \ldots + (a_n - \overline{a})^2}}$
Relative-absolute error*	$\frac{\lvert p_1 - a_1\rvert + \ldots + \lvert p_n - a_n\rvert}{\lvert a_1 - \overline{a}\rvert + \ldots + \lvert a_n - \overline{a}\rvert}$
Correlation coefficient**	$\frac{S_{PA}}{\sqrt{S_P S_A}}$, where $S_{PA} = \frac{\sum_i (p_i - \overline{p})(a_i - \overline{a})}{n-1}$, $S_P = \frac{\sum_i (p_i - \overline{p})^2}{n-1}$, $S_A = \frac{\sum_i (a_i - \overline{a})^2}{n-1}$

*Here, $\overline{a}$ is the mean value over the training data.
**Here, $\overline{a}$ is the mean value over the test data.

Slika 17: Merila za ocenjevanje napovednih modelov [34]

Pri regresiji nas bo zanimalo kakšna je povprečna absolutna napaka, ki pove kakšno je povprečje vseh absolutnih napak (v primeru nepremičnin za koliko rubljev smo se povprečno zmotili pri napovedovanju cen) in kakšna je relativna absolutna napaka, ki to pove v odstotkih. Absolutna napaka sešteje absolutne vrednosti razlik med napovedanimi vrednostmi in dejanskimi vrednostmi, ki predstavljajo posamezne napake in jih deli s številom testnih primerov. Relativna absolutna napaka podobno sešteje absolutne vrednosti posameznih napak, vendar to vsoto deli z vsoto absolutnih razlik med posameznimi dejanskimi vrednostmi in povprečno vrednostjo za učno množico. Poleg teh dveh napak nas bo zanimal tudi korelacijski koeficient, ki bo v tem primeru povedal kakšna je korelacija med napovedano ceno nepremičnine in dejansko ceno. Ostali dve napaki nas ne bosta toliko zanimali. Razlog za to je, da ti dve napaki zaradi kvadriranja upoštevata večje napake bolj, manjše pa manj, medtem ko prej opisani napaki upoštevata vse napake enakovredno.

6 Rezultati

Pomembna faza po zaključku modeliranja je ovrednotenje dobljenih rezultatov, saj na ta način ugotovimo ali so dobljeni modeli dejansko uporabni ali ne. Rezultati nam poleg tega lahko pomagajo ugotoviti kakšna je razlika med napovednimi modeli dobljenimi z različnimi algoritmi in kateri pristop je najboljši za izbrane podatke.

V nadaljevanju bodo predstavljeni rezultati dobljeni v Weki pri izgradnji napovednih modelov za cene nepremičnin. Pri tem so uporabljeni različni regresijski algoritmi in tudi tri načini izbire atributov. Tako bomo iz rezultatov lahko ugotovili katera izbira atributov in kateri algoritem oziroma algoritmi so najboljši za gradnjo napovednih modelov v tem primeru za napovedovanje cen nepremičnin.

6.1 Natančnost napovednih modelov

Za gradnjo in testiranje napovednih modelov je bila uporabljena Weka, natančneje Experimenter, kjer so bili za gradnjo regresijskih modelov izbrani algoritmi LinearRegression, IBk, M5P, M5Rules, REPTree, RandomTree in RandomForest, ter za primerjavo še algoritma ZeroR in DecisionStump. Vsi algoritmi razen algoritma IBk so imeli privzete nastavitve. Za IBk pa je bil uporabljen k = 5 in mahattanska razdalja, obe vrednosti sta bili določeni glede na predhodne rezultate. Testiranje napovednih modelov je potekalo z 10-kratnim prečnim preverjanjem.

Dobljeni rezultati so prikazani v treh tabelah glede na atribute, ki so uporabljeni v posameznem primeru. V prvem primeru je to vseh 300 atributov razen določenih nerelavantnih atributov, ki so opisani v prilogi A. V drugem primeru so uporabljeni najboljši atributi izbrani v Weki, opisani so v prilogi B. V tretjem primeru pa so bili izbrani podobni atributi kot za oceno stanovanja na irn.ru, opisani so v prilogi C.

Algoritma ZeroR in DecisionStump nista v tabelah, ker sta bila uporabljena samo za primerjavo in so rezultati za oba algoritma v vseh treh primerih enaki. ZeroR je napovedal povprečno vrednost za ceno nepremičnine, ki je znašala 6.097.672,51 rub. DecisionStump pa je napovedal cene glede na najpomembnejši atribut, ki je bil v tem primeru atribut za celotno površino (atribut full_sq). Algoritem ZeroR je imel korelacijski koeficient enak 0,00 in vrednost za napake 1.977.725,05 rub, 2.758.334,73 rub, 100,00 %, 100,00 %. Algoritem DecisionStump je imel korelacijski koeficient 0,54 in

napake 1.682.922,89 rub, 2.319.833,02 rub, 85,13 %, 84,10 %. Vrednosti za ta dva algoritma so samo za primerjavo kakšna je najmanjša natančnost, ki jo lahko pričakujemo od drugih algoritmov. Glede na to, da so rezultati za druge algoritme veliko boljši od rezultatov za ZeroR in DecisionStump lahko ugotovimo, da so bili izbrani algoritmi uspešni pri napovedovanju cen.

	IBk (k=5)	M5P	M5Rules	REPTree	RandomTree	RandomForest	LinearRegression
Korelacijski koeficient	**0,83**	**0,95**	**0,94**	**0,93**	**0,80**	**0,90**	**0,93**
Povprečna absolutna napaka (RUB)	1.028.329,42	445.272,62	483.799,23	488.878,67	1.026.463,53	768.464,76	577.016,59
Koren povprečne kvadratne napake (RUB)	1.533.060,24	874.316,81	908.557,48	1.000.366,92	1.726.440,88	1.188.449,40	1.000.357,86
Relativna absolutna napaka (%)	**52,06**	**22,53**	**24,46**	**24,76**	**52,08**	**38,93**	**29,21**
Koren relativne kvadratne napake (%)	55,68	31,71	32,91	36,35	62,81	43,18	36,38

Tabela 1: Rezultati dobljeni pri regresiji z vsemi 300 atributi

	IBk (k=5)	M5P	M5Rules	REPTree	RandomTree	RandomForest	LinearRegression
Korelacijski koeficient	**0,84**	**0,94**	**0,93**	**0,93**	**0,93**	**0,96**	**0,84**
Povprečna absolutna napaka (RUB)	1.053.961,87	475.135,92	526.607,00	542.026,56	480.829,66	380.655,90	962.831,38
Koren povprečne kvadratne napake (RUB)	1.614.151,03	906.919,92	1.000.071,58	1.027.717,99	995.482,59	743.527,54	1.484.735,59
Relativna absolutna napaka (%)	**53,39**	**24,06**	**26,63**	**27,42**	**24,37**	**19,29**	**48,75**
Koren relativne kvadratne napake (%)	58,69	32,95	36,26	37,22	36,23	27,00	53,91

Tabela 2: Rezultati dobljeni pri regresiji z najboljšimi atributi izbranimi v Weki

	IBk (k=5)	M5P	M5Rules	REPTree	RandomTree	RandomForest	LinearRegression
Korelacijski koeficient	**0,80**	**0,94**	**0,94**	**0,94**	**0,93**	**0,96**	**0,92**
Povprečna absolutna napaka (RUB)	1.147.795,32	483.795,86	534.984,41	486.293,46	490.330,31	388.235,55	678.505,98
Koren povprečne kvadratne napake (RUB)	1.744.970,42	933.098,66	975.660,00	972.934,76	1.043.651,34	783.211,61	1.102.953,10
Relativna absolutna napaka (%)	**58,13**	**24,46**	**27,05**	**24,61**	**24,80**	**19,65**	**34,35**
Koren relativne kvadratne napake (%)	63,45	33,84	35,39	35,32	37,87	28,41	40,07

Tabela 3: Rezultati dobljeni pri regresiji z atributi za oceno stanovanja na irn.ru

Rezultati algoritmov so še kar dobri, če rezultate primerjamo glede na korelacijski koeficient, ki pove kakšna je korelacija med napovedanimi in dejanskimi vrednostmi za cene nepremičnin. Če gledamo napake, bomo ugotovili podobno. Dobre rezultate dobimo predvsem z algoritmom RandomForest v drugem in tretjem primeru, kjer je izbrano manjše število atributov. Rezultati dobljeni z ostalimi algoritmi so tudi dobri npr. z algoritmi M5P, M5Rules in REPTree. Glede na izbrane atribute pa lahko ugotovimo, da je skoraj boljša izbira manjšega števila atributov kot je to v drugem in tretjem primeru, medtem ko z izbiro večjega števila atributov, kot v prvem primeru, ne dobimo boljših rezultatov. Podrobneje bodo rezultati pojasnjeni v interepretaciji.

6.2 Čas potreben za učenje in testiranje

Dobljene napovedne modele oziroma algoritme, ki smo jih uporabili za njihovo izgradnjo lahko primerjamo tudi po temu koliko časa je bilo potrebno za učenje in testiranje. Tako lahko tudi na nek način ugotovimo kakšna je zahtevnost algoritma in dobljenega modela. V tabeli 4 je predstavljen čas v sekundah, ki je bil potreben za učenje in testiranje za določen algoritem in izbrane atribute. V tabeli tako kot prej ni algoritmov ZeroR in DecisionStump.

	IBk (k=5)	M5P	M5Rules	REPTree	RandomTree	RandomForest	LinearRegression
Čas potreben za učenje (s) - 300 atributov	0,00	18,96	120,58	1,57	0,28	15,63	96,73
Čas potreben za testiranje (s) - 300 atributov	8,92	0,01	0,01	0,00	0,00	0,10	0,00
Čas potreben za učenje (s) - atributi Weka	0,00	1,04	19,17	0,15	0,12	4,93	0,02
Čas potreben za testiranje (s) - atributi Weka	1,14	0,00	0,03	0,00	0,00	0,12	0,00
Čas potreben za učenje (s) - atributi IRN	0,00	3,63	41,51	0,09	0,09	4,13	0,92
Čas potreben za testiranje (s) - atributi IRN	0,95	0,00	0,02	0,00	0,00	0,12	0,00

Tabela 4: Čas potreben za učenje in testiranje za različne algoritme v sekundah

Iz tabele je vidno, da število atributov bistveno vpliva na čas učenja in testiranja. Še posebej veliko časa je bilo potrebno v prvem primeru, ko je bilo izbarnih 300 atributov. Poleg tega lahko vidimo, da različni algoritmi potrebujejo različno časa. Najpočasnejši je M5Rules in LinearRegression za prvi primer. M5P in RandomForest sta nekje v sredini. Nekoliko hitrejši je IBk. Najhitrejša pa sta algoritma REPTree in RandomTree. Iz tega lahko ugotovimo, da je potreben čas sorazmeren s zahtevnostjo modela. Tako sta najhitrejša REPTree in RandomTree, ki zgradita drevo. IBk, ki išče najbližje sosede

je že nekoliko počasnejši. Gradnja večih dreves pri algoritmu RandomForest, gradnja linearnih modelov pri LinearRegression ali uporaba obojega pri M5P in M5Rules, pa zahteva več časa.

6.3 Interpretacija

Kot je bilo že prej povedano so rezultati za izbrane algoritme še kar dobri. Tukaj moramo upoštevati tudi to, da so podatki vsebovali veliko manjkajočih in nepravilnih vrednosti, zato so napake lahko nekoliko večje, kot bi bile sicer.

Dobljene rezultate lahko najenostavneje primerjamo glede na korelacijski koeficient, ki pove kakšna je korelacija med napovedanimi vrednostmi za cene nepremičnin in dejanskimi vrednostmi v podatkih. Ostale vrednosti za napake so dejansko obratno sorazmerne s korelacijskim koeficientom. Tako so napake manjše za algoritme, kjer je korelacijski koeficient večji in večje, kjer je korelacijski koeficient manjši. Glede na dobljene rezultate lahko že z zelo preprostim algoritmom kot je IBk lahko dosežemo korelacijo od 0,80 do 0,84. Med ostalimi algoritmi ima korelacijo 0,80 še RandomTree v prvem primeru in 0,84 LinearRegression v drugem primeru. Korelacije za ostale algoritme so v vseh primerih večje od 0,90. Za večino algoritmov je korelacija 0,93 ali 0,94, kar je zelo dobro. Največjo korelacijo 0,96 ima algoritem RandomForest v drugem in tretjem primeru, kjer je izbrano manjše število atributov.

Poleg korelacijskega koeficienta lahko primerjamo tudi ostale napake. Pri tem nas posebej zanima povprečna absolutna napaka, ki pove kakšna je povprečna napaka ter relativna absolutna napaka, ki to pove v odstotkih. Kot lahko vidimo, so napake obratno sorazmerne s korelacijskim koeficientom, kjer je korelacijski koeficient večji so vrednosti za napake manjše. Največje napake so tako pri algoritmu IBk, ki znašajo od 52,02 do 58,13 %. Podobno kot pri korelacijskem koeficientu ima veliko napako RandomTree v prvem primeru, ki znaša 52,08 % in LinearRegression v drugem primeru, kjer napaka znaša 48,75 %. Ostali algoritmi so boljši in napaka zanje znaša od približno od 20 do 30 %. Najmanjšo napako ima za drugi in tretji primer algoritem RandomForest, kjer napaka znaša približno 19 %, v prvem primeru pa je napaka večja in znaša 38,93 %.

Ker smo v tem primeru poskušali napovedovati cene nepremičnin, je smiselno da napako izrazimo tudi na ta način. To lahko najbolje izrazimo s povprečno absolutno napako, ki pove za koliko rubljev smo se povprečno zmotili pri napovedovanju cen. Napako lahko za boljšo predstavo poleg rubljev izrazimo tudi v evrih. Pri tem je potrebno omeniti še povprečno ceno nepremičnin, ki jo napove tudi algoritem ZeroR. V tem primeru je povprečna cena znašala 6.097.672,51 rub (84.204,93 eur). Največjo napako, ki je znašala 1.147.795,32 rub (15.850,31 eur) je imel algoritem IBk v tretjem

primeru, najmanjšo napako 380.655,90 rub (5.255,76 eur) pa algoritem RandomForest v drugem primeru. Večina algoritmov je bila nekje vmes, tako kot algoritem REPTree za drugi primer, kjer je napaka znašala 542.026,56 rub (7.483,82 eur).

Če primerjamo te napake s povpečno ceno nepremičnin največja napaka za IBk predstavlja 18,82 %, najmanjša za RandomForest 6,24 %. Napaka za REPTree pa znaša 8,89 %. Tako so napake za boljše algoritme nekje od 6 % do 9 %. To je primerljivo tudi z rezultati, ki smo jih dobili za oceno stanovanja dobili na irn.ru. Te vrednosti se nekoliko razlikujejo od napake relative absolute error, saj je bila tukaj povprečna cena nepremičnine upoštevana samo enkrat, pri izračunu omenjene napake pa se povprečna cena nepremičnine upošteva za vsak posamezen primer in je tako napaka večja.

Preostalih dveh napak za koren povprečne kvadratne napake in koren relativne kvadratne napake ne bomo upoštevali, saj posameznih napak pri napovedovanju ne upoštevata enakovredno. Vrednosti za posamezne napake so tukaj kvadrirane, zato so večje posamezne napake upoštevane bolj kot manjše in so zato na koncu napake večje.

Ker smo za vse algoritme dobili dobre rezultate so razlike med njimi predvsem posledica različnega delovanja algoritmov. V tem primeru so bile napake največje za algoritem IBk, ki je napovedal ceno nepremičnine glede na 5 najbližjih sosedov. Razlog zakaj so rezultati takšni je, ker IBk išče najbližje sosede glede na vse atribute, pri tem pa ne upošteva, da so nekateri atributi pomembnejši kot drugi. Na primer algoritem bi moral upoštevati atribut za celotno površino bolj kot atribut za število nakupovalnih centrov v okolici, vendar tega ne počne. Zato ta algoritem ni najprimernejši za napovedovanje cen nepremičnin v tem primeru.

Naslednji trije algoritmi M5P, M5Rules in REPTree imajo za vse tri primere dobre rezultate. Iz tega lahko ugotovimo, da so uporabni tudi, ko imamo veliko atributov. Vsi trije atributi gradijo drevesa, M5P zgradi M5' drevo, ki ima v vsakem listu linearen regresijski model, M5Rules zgradi več takšnih dreves in iz njih dobi pravila, REPTree pa zgradi eno drevo, ki ga nato reže. Iz tega lahko zaključimo, da je uporaba dreves dober pristop za gradnjo napovednih modelov.

Takšen pristop uporabljata tudi algoritma RandomTree in RandomForest. Algoritem RandomTree zgradi drevo, kjer za vsako vozlišče izbere najboljši atribut izmed k naključno izbranih atributov, RandomForest pa zgradi več takšnih dreves. V obeh primerih se drevesa gradi do maksimalne velikosti in se jih ne reže. To je najverjetneje razlog, zakaj sta bila oba algoritma v prvem primeru, kjer je bilo veliko atributov nekoliko slabša. V drugem in tretjem primeru pa so rezultati za oba algoritma dobri, še posebej za RandomForest, ki ima od vseh algoritmov najmanjše napake.

Preostane nam še algoritem LinearRegression, ki uporablja linearno regresijo. Algoritem ima dobre rezultate tudi za veliko število atributov. Kljub temu so rezultati za drugi primer, ko je bilo izbranih malo atributov slabši od prvega primera, v tretjem

primeru pa ima ponovno boljše rezultate. Mogoče je razlog v tem, da moramo izbrati pravilne atribute ali pa izbrati vse. Zaradi tega LinearRegression ni ravno najboljša izbira, saj ne moremo vedeti ali smo dejansko izbrali najboljše atribute ali ne.

Nekaj dodatnih informacij glede algoritmov lahko dobimo tudi iz podatkov za čas, ki ga potrebujejo za učenje in testiranje. Iz rezultatov lahko ugotovimo, da nekateri algoritmi potrebujejo veliko časa, drugi algoritmi pa manj. Največ časa sta potrebovala algoritma M5Rules in LinearRegression, kar kaže tudi na njuno zahtevnost. Nekje v sredini so M5P, RandomForest in IBk. Najhitrejša pa sta REPTree in RandomTree. Če dobljene rezultate primerjamo s časom za učenje in testiranje lahko ugotovimo, da s hitrejšimi algoritmi kot sta REPTree in RandomTree dobimo primerljive rezultate kot s počasnejšimi algoritmi M5Rules in LinearRegression. Tako dejansko ni potrebno, da uporabljamo bolj zahtevne algoritme kot je M5Rules, saj na ta način ne bomo dobili veliko boljših rezultatov kot sicer.

Podobno lahko ugotovimo glede izbire atributov. Uporaba velikega števila različnih atributov ni potrebna, saj dobimo primerljive rezultate že z manjšim številom atributov. Za algoritme kot sta RandomTree in RandomForest je takšna izbira še veliko boljša. Glede same izbire atributov pa ni bilo velike razlike med rezultati za drugi primer, kjer so bili atributi izbrani v Weki in tretji primer, kjer so bili izbrani podobni atributi kot pri oceni stanovanja na irn.ru. Razlika je bila le pri algoritmu LinearRegression in manjša pri algoritmu IBk. Tako da je vseeno če sami izberemo atribute, ki se nam zdijo pomembnejši ali če izbiro prepustimo Weki.

Glede na vse dobljene rezultate in ugotovitve lahko zaključimo, da je pri izbiri atributov boljše uporabiti manjše število atributov, izberemo jih lahko sami ali z Weko. Od algoritmov je najboljše izbrati M5P, RandomTree, RandomForest ali REPTree. Najboljše rezultate pa dobimo z algoritmom RandomForest.

7 Zaključek in nadaljnje delo

Glede na rezultate lahko zaključimo, da je podatkovno rudarjenje primeren pristop za napovedovanje cen nepremičnin. Dobre rezultate smo dobili skoraj z vsemi algoritmi. Kot najprimernejše algoritme za ta primer pa bi lahko izpostavili algoritme M5P, RandomTree, RandomForest in REPTree, izmed katerih je bil najboljši algoritem RandomForest.

Omeniti je potrebno še izbiro atributov, glede katere je bilo ugotovljeno, da je veliko bolje izbrati manjše število pomembnejših atributov. To je posebej pomembno za določene algoritme kot sta RandomTree in RandomForest. Takšna izbira atributov primpomore tudi k manjši zahtevnosti napovednih modelov in krajšemu času potrebnemu za gradnjo modelov.

Napovedni modeli dobljeni v okviru te naloge, bi bili primerni za približno oceno vrednosti nepremičnine. Da bi bili modeli primerni za dejansko uporabo pri napovedovanju cen nepremičnin, bi morali napako zmanjšali še za kakšnih 5 do 10 %. To bi lahko storili zelo enostavno, če bi imeli boljše podatke z manj manjkajočimi in nepravilnimi vrednostmi. Vendar so v tem primeru podatki takšni kot so in jih naknadno ni mogoče preveč spreminjati. Zato bi morali v primeru gradnje pravih napovednih modelov zelo paziti že v fazi zbiranja podatkov.

V primeru, da bi želeli narediti podobne modele za dejansko uporabo pri napovedovanju cen nepremičnin, bi bilo pomembno predvsem, da bi zbrali oziroma dobili dobre podatke, ki imajo čim manj nepravilnih in manjkajočih vrednosti ter vsebujejo vse pomembne atribute, saj je kakovost modela v veliki meri odvisna tudi od podatkov in ne samo od izbranega algoritma. Glede algoritmov pa lahko rečemo, da bi z izbiro podobih algoritmov kot so bili izbrani tukaj, zagotovo dobili dobre rezultate.

Glede nadaljnjega dela pa bi lahko poskušali statistično primerjati modele med seboj in ugotoviti ali so razlike med njimi dejansko pomembne, saj so v veliko primerih razlike med dobljenimi modeli zelo majhne. Dodatno bi lahko poskušali modele tudi interpretirati. Tako bi lahko ugotovili kako se modeli razlikujejo glede na uporabljen algoritem, kako različni atributi vplivajo na ceno, kateri izmed njih so pomembnejši in podobno.

8 Literatura in viri

[1] D. Aha in D. Kibler, Instance-based learning algorithms. *Machine Learning* 6 (1991) 37–66.

[2] M. Alekseevsky, *The russian housing dream vs reality*, https://strelka.com/en/magazine/2017/11/01/house. (Datum ogleda: 1. 7. 2018.)

[3] A. Azevedo in M. Santos, KDD, SEMMA and CRISP-DM: a parallel overview. *IADS-DM* (2008) .

[4] A. Behnood, V. Behnood, M. M. Gharehveran in K. E. Alyamac, Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm. *Construction and Building Materials* 142 (2017) 199-207.

[5] L. Breiman, Random Forests. *Machine Learning* 45 (2001) 5–32.

[6] J. Brownlee, *How to Work Through a Regression Machine Learning Project in Weka Step-By-Step*, https://machinelearningmastery.com/regression-machine-learning-tutorial-weka/. (Datum ogleda: 1. 7. 2018.)

[7] R. Chandradevan, *Random Forest Learning - Essential Understanding*, https://towardsdatascience.com/random-forest-learning-essential-understanding-1ca856a963cb. (Datum ogleda: 1. 7. 2018.)

[8] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reintartz, C.Shearer in R. Wirth, *CRISP-DM 1.0: Step-by-step data mining guide*. SPSS Inc., 2000.

[9] *Comparative Market Analysis*, http://www.investopedia.com/terms/c/comparative-market-analysis.asp. (Datum ogleda: 1. 7. 2018.)

[10] *Cross Industry Standard Proces for Data Mining*, https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining. (Datum ogleda: 1. 7. 2018.)

[11] *Data Mining Applications*, http://www.zentut.com/data-mining/data-mining-applications/. (Datum ogleda: 1. 7. 2018.)

[12] *Examples of data mining*, https://en.wikipedia.org/wiki/Examples_of_data_mining/. (Datum ogleda: 1. 7. 2018.)

[13] U. Fayyad, G. Piatetsky-Shapiro in P. Smyth, From data mining to knowledge discovery in databases. *AI magazine* 17.3 (1996) 37.

[14] P. Giudici, *Applied data mining: statistical methods for business and industry*, John Wiley & Sons, 2005.

[15] T. Golubeva, *Understand Russia: Apartments Or Houses? How Urban Russians Live?*, https://understandrussia.com/apartments-or-houses/. (Datum ogleda: 1. 7. 2018.)

[16] J. Han, J. Pei in M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[17] D. J. Hand, H. Mannila in P. Smyth, *Principles of data mining*. MIT press, 2001.

[18] G. Holmes, M. Hall in E. Frank, Generating Rule Sets from Model Trees. *Twelfth Australian Joint Conference on Artificial Intelligence* (Springer, 1999) 1-12.

[19] Y. Kozhevnikova, *Moscow's real estate market goes from boom to bust*, https://www.inman.com/2015/10/23/moscows-real-estate-market-goes-from-boom-to-bust/. (Datum ogleda: 1. 7. 2018.)

[20] *Machine learning*, https://en.wikipedia.org/wiki/Machine_learning/. (Datum ogleda: 1. 7. 2018.)

[21] *Multiple Linear Regression Analysis*, http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis/. (Datum ogleda: 1. 7. 2018.)

[22] D. Olson in D. Delen, *Advanced Data Mining Techniques*. Springer Science & Business Media, 2008.

[23] *Petroleum industry in Russia*, https://en.wikipedia.org/wiki/Petroleum_industry_in_Russia. (Datum ogleda: 1. 7. 2018.)

[24] R. J. Quinlan, Learning with Continuous Classes. *5th Australian Joint Conference on Artificial Intelligence* (1992) 343-348.

[25] *Random forest*, https://en.wikipedia.org/wiki/Random_forest. (Datum ogleda: 1. 7. 2018.)

[26] *RandomTree*, `http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/RandomTree.html`. (Datum ogleda: 1. 7. 2018.)

[27] *Sberbank Russian Housing Market*, `https://www.kaggle.com/c/sberbank-russian-housing-market/`. (Datum ogleda: 1. 7. 2018.)

[28] C. Shearer, The CRISP-DM model: the new blueprint for data mining. *Data Warehousing Journal* 5 (2000) 13–22.

[29] G. Shmueli, N. R. Patel in P. C. Bruce, *Data mining for business intelligence: Concepts, techniques, and applications in Microsoft Office Excel with XLMiner*. John Wiley & Sons, 2011.

[30] *The R Project for Statistical Computing*, `https://www.r-project.org`. (Datum ogleda: 1. 7. 2018.)

[31] Y. Wang in I. H. Witten, Induction of model trees for predicting continuous classes. *Poster papers of the 9th European Conference on Machine Learning* (Springer, 1997) .

[32] *Weka: Data Mining Software in Java*, `http://www.cs.waikato.ac.nz/ml/weka/`. (Datum ogleda: 1. 7. 2018.)

[33] R. Wirth in J. Hipp, CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (2000) 29–39.

[34] I. H. Witten, E. Frank, M. A. Hall in C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[35] И. Варламов, Москва глазами птицы, `https://varlamov.ru/1122725.html`. (Datum ogleda: 1. 7. 2018.)

[36] Индексы рынка недвижимости, `http://www.irn.ru/index/`. (Datum ogleda: 1. 7. 2018.)

[37] Цены на квартиры в Москве и Подмосковье , `http://www.irn.ru/price/`. (Datum ogleda: 1. 7. 2018.)

[38] Цены на недвижностьи квартиры в Москве на графике, `http://www.irn.ru/gd/`. (Datum ogleda: 1. 7. 2018.)

[39] Яндекс котировкиь, `https://news.yandex.ru/quotes/`. (Datum ogleda: 1. 7. 2018.)

Priloge

A Izbranih 300 atributov

full_sq - celotna površina
life_sq - bivanjska površina
floor - nadstropje
max_floor - število vseh nadstropij
material - material izgradnje
build_year - leto izgradnje
num_room - število sob
kitch_sq - površina kuhinje
state - stanje nepremičnine
sub_area - rajon Moskve, kjer se nepremičnina nahaja
area_m - velikost rajona
raion_popul - število prebivalcev v rajonu
green_zone_part - delež zelenih površin
indust_part - delež industrijskih površin
children_preschool - število predšolskih otrok v rajonu
preschool_quota - število mest v predšolskih ustanovah v rajonu
preschool_education_centers_raion - število predšolskih ustanov v rajonu
children_school - število šolskih otrok v rajonu
school_quota - število mest v šolah v rajonu
school_education_centers_raion - število šolskih ustanov v rajonu
school_education_centers_top_20_raion - število šol v rajonu, ki se uvrščajo med najboljših 20 šol v Moskvi
hospital_beds_raion - število postelj v bolnišnicah v rajonu
healthcare_centers_raion - število zdravstvenih ustanov v rajonu
university_top_20_raion - število univerz v rajonu, ki se uvrščajo med najboljših 20 univerz v Moskvi
sport_objects_raion - število športnih objektov v rajonu
additional_education_raion - število dodatnih izobraževalnih ustanov v rajonu
culture_objects_top_25 - prisotnost pomembnih objektov kulturne dediščine
culture_objects_top_25_raion - število pomembnih objektov kulturne dediščine v rajonu, ki spadajo med najpomembnejših 25 kulturnih objektov v Moskvi

shopping_centers_raion - število nakupovalnih središč v rajonu
office_raion - število uradov v rajonu
thermal_power_plant_raion - prisotnost termoelektrarne v rajonu
incineration_raion - prisotnost sežigalnice v rajonu
oil_chemistry_raion - prisotnost naftne industrije v rajonu
radiation_raion - prisotnost odlagališča radioaktivnih odpadkov v rajonu
railroad_terminal_raion - prisotnost železniškega terminala v rajonu
big_market_raion - prisotnost veletrgovskih centrov v rajonu
nuclear_reactor_raion - prisotnost nuklearnih reaktorjev v rajonu
detention_facility_raion - prisotnost zaporov v rajonu
raion_build_count_with_material_info - število stavb s podatki o načinu izgradnje
build_count_block - delež stavb iz blokov
build_count_wood - delež stavb iz lesa
build_count_frame - delež stavb z okvirjem
build_count_brick - delež stavb iz opeke
build_count_monolith - delež monolitnih stavb
build_count_panel - delež stavb iz panelov
build_count_foam - delež stavb iz pene
build_count_slag - delež stavb iz žlindre
build_count_mix - delež stavb z več načini izgradnje
raion_build_count_with_builddate_info - število stavb s podatki o letu izgradnje
build_count_before_1920 - delež stavb zgrajenih pred letom 1920
build_count_1921-1945 - delež stavb zgrajenih med leti 1921 in 1945
build_count_1946-1970 - delež stavb zgrajenih med leti 1946 in 1970
build_count_1971-1995 - delež stavb zgrajenih med leti 1971 in 1995
build_count_after_1995 - delež stavb zgrajenih po letu 1995
metro_min_avto - razdalja do metroja z avtom v minutah
metro_km_avto - razdalja do metroja z avtom v kilometrih
metro_min_walk - razdalja do metroja peš v minutah
metro_km_walk - razdalja do metroja peš v kilometrih
kindergarten_km - razdalja do vrtca v kilometrih
school_km - razdalja do šole v kilometrih
park_km - razdalja do parka v kilometrih
green_zone_km - razdalja do zelene cone v kilometrih
industrial_km - razdalja do industrije v kilometrih
water_treatment_km - razdalja do vode v kilometrih
cemetery_km - razdalja do pokopališča v kilometrih
incineration_km - razdalja do sežigalnice v kilometrih

railroad_station_walk_km - razdalja do železniške postaje peš v kilometrih
railroad_station_walk_min - razdalja do železniške postaje peš v minutah
railroad_station_avto_km - razdalja do železniške postaje z avtom v kilometrih
railroad_station_avto_min - razdalja do železniške postaje z avtom v minutah
public_transport_station_km - razdalja do javnega transporta peš v kilometrih
public_transport_station_min_walk - razdalja do javnega transporta peš v minutah
water_km - razdalja od vode npr. reke, jezera v kilometrih
water_1line - razdalja do prve linije do vode (150 m)
mkad_km - razdalja do MKAD (krožna avtocesta okoli Moskve)
ttk_km - razdalja do TTC (krožna cesta okoli centra Moskve)
sadovoe_km - razdalja do Sadovoye Koltso (krožna cesta v centru Moskve)
bulvar_ring_km - razdalja do Bulvarnoye Koltso (krožna cesta v centru Moskve)
kremlin_km - razdalja do Kremlja v kilometrih
big_road1_km - razdalja do najbližje večje ceste v kilometrih
big_road1_1line - razdalja do prve linije do večjih cest (100 m od avtoceste ali 250 m od MKAD) v kilometrih
big_road2_km - razdalja do druge najbližje večje ceste v kilometrih
railroad_km - razdalja do železnice v kilometrih
railroad_1line - razdalja do prve linije do železnice (100 m)
zd_vokzaly_avto_km - razdalja do železniške postaje z avtom v kilometih
bus_terminal_avto_km - razdalja do avtobusnega terminala z avtom v kilometrih
oil_chemistry_km - razdalja do naftne industrije v kilometrih
nuclear_reactor_km - razdalja so nuklearnega reaktorja v kilometrih
radiation_km - razdalja do odlagališča za radioaktivne odpadke v kilometrih
power_transmission_line_km - razdalja do električnega voda v kilometrih
thermal_power_plant_km - razdalja do termoelektrarne v kilometrih
ts_km - razdalja do elektrarne v kilometrih
big_market_km - razdalja do večje trgovine v kilometrih
market_shop_km - razdalja do različnih trgovin v kilometrih
fitness_km - razdalja do fitnesa v kilometrih
swim_pool_km - razdalja do bazena v kilometrih
ice_rink_km - razdalja do drsališča v kilometrih
stadium_km - razdalja do stadiona v kilometrih
basketball_km - razdalja so igrišča za košarko v kilometrih
hospice_morgue_km - razdalja do mrtvašnice v kilometrih
detention_facility_km - razdalja do zapora v kilometrih
public_healthcare_km - razdalja do javnih zdravstvenih ustanov v kilometih
university_km - razdalja do univerze v kilometrih

workplaces_km - razdalja do delovnih mest v kilometrih
shopping_centers_km - razdalja do trgovskih centrov v kilometrih
office_km - razdalja do uradov v kilometrih
additional_education_km - razdalja do drugih izobraževalnih ustanov v kilometrih
preschool_km - razdalja do predšolskih ustanov v kilometrih
big_church_km - razdalja do večje cerkve v kilometrih
church_synagogue_km - razdalja do cerkve ali sinagoge v kilometrih
mosque_km - razdalja do mošeje v kilometrih
theater_km - razdalja do gledališča v kilometrih
museum_km - razdalja do muzeja v kilometrih
exhibition_km - razdalja do razstave v kilometrih
catering_km - razdalja do cateringa v kilometrih
ecology - ekološka cona, kjer se nepremičnina nahaja
green_part_500 - delež zelenih površin v območju 0,5 km
prom_part_500 - delež industrijskih površin v območju 0,5 km
office_count_500 - število uradov v območju 0,5 km
office_sqm_500 - površina uradov v območju 0,5 km
trc_count_500 - število trgovskih centrov v območju 0,5 km
trc_sqm_500 - površina trgovskih centrov v območju 0,5 km
cafe_count_500 - število kavarn in restavracij v območju 0,5 km
cafe_sum_500_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 0,5 km
cafe_sum_500_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 0,5 km
cafe_avg_price_500 - povprečen račun v kavarni ali restavraciji v območju 0,5 km
cafe_count_500_na_price - število kavarn in restavracij v območju 0,5 km, kjer ni podatka o ceni
cafe_count_500_price_500 - število kavarn in restavracij v območju 0,5 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_500_price_1000 - število kavarn in restavracij v območju 0,5 km, kjer povprečen račun znaša od 500 do 1000 rub
cafe_count_500_price_1500 - število kavarn in restavracij v območju 0,5 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_500_price_2500 - število kavarn in restavracij v območju 0,5 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_500_price_4000 - število kavarn in restavracij v območju 0,5 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_500_price_high - število kavarn in restavracij v območju 0,5 km, kjer

povprečen račun znaša več kot 4000 rub
big_church_count_500 - število velikih cerkva v območju 0,5 km
church_count_500 - število cerkva v območju 0,5 km
mosque_count_500 - število mošej v območju 0,5 km
leisure_count_500 - število objektov za prosti čas v območju 0,5 km
sport_count_500 - število objektov za športne aktivnosti v območju 0,5 km
market_count_500 - število tržnic v območju 0,5 km
green_part_1000 - delež zelenih površin v območju 1 km
prom_part_1000 - delež industrijskih površin v območju 1 km
office_count_1000 - število uradov v območju 1 km
office_sqm_1000 - površina uradov v območju 1 km
trc_count_1000 - število trgovskih centrov v območju 1 km
trc_sqm_1000 - površina trgovskih centrov v območju 1 km
cafe_count_1000 - število kavarn in restavracij v območju 1 km
cafe_sum_1000_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 1 km
cafe_sum_1000_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 1 km
cafe_avg_price_1000 - povprečen račun v kavarni ali restavraciji v območju 1 km
cafe_count_1000_na_price - število kavarn in restavracij v območju 1 km, kjer ni podatka o ceni
cafe_count_1000_price_500 - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_1000_price_1000 - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša od 500 do 1000 rub
cafe_count_1000_price_1500 - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_1000_price_2500 - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_1000_price_4000 - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_1000_price_high - število kavarn in restavracij v območju 1 km, kjer povprečen račun znaša več kot 4000 rub
big_church_count_1000 - število velikih cerkva v območju 1 km
church_count_1000 - število cerkva v območju 1 km
mosque_count_1000 - število mošej v območju 1 km
leisure_count_1000 - število objektov za prosti čas v območju 1 km
sport_count_1000 - število objektov za prosti čas v območju 1 km

market_count_1000 - število tržnic v območju 1 km
green_part_1500 - delež zelenih površin v območju 1,5 km
prom_part_1500 - delež industrijskih površin v območju 1,5 km
office_count_1500 - število uradov v območju 1,5 km
office_sqm_1500 - površina uradov v območju 1,5 km
trc_count_1500 - število trgovskih centrov v območju 1,5 km
trc_sqm_1500 - površina trgovskih centrov v območju 1,5 km
cafe_count_1500 - število kavarn in restavracij v območju 1,5 km
cafe_sum_1500_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 1,5 km
cafe_sum_1500_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 1,5 km
cafe_avg_price_1500 - povprečen račun v kavarni ali restavraciji v območju 1,5 km
cafe_count_1500_na_price - število kavarn in restavracij v območju 1,5 km, kjer ni podatka o ceni
cafe_count_1500_price_500 - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_1500_price_1000 - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša od 500 do 1000 rub
cafe_count_1500_price_1500 - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_1500_price_2500 - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_1500_price_4000 - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_1500_price_high - število kavarn in restavracij v območju 1,5 km, kjer povprečen račun znaša več kot 4000 rub
big_church_count_1500 - število velikih cerkva v območju 1,5 km
church_count_1500 - število cerkva v območju 1,5 km
mosque_count_1500 - število mošej v območju 1,5 km
leisure_count_1500 - število objektov za prosti čas v območju 1,5 km
sport_count_1500 - število objektov za prosti čas v območju 1,5 km
market_count_1500 - število tržnic v območju 1,5 km
green_part_2000 - delež zelenih površin v območju 2 km
prom_part_2000 - delež industrijskih površin v območju 2 km
office_count_2000 - število uradov v območju 2 km
office_sqm_2000 - površina uradov v območju 2 km
trc_count_2000 - število trgovskih centrov v območju 2 km

trc_sqm_2000 - površina trgovskih centrov v območju 2 km
cafe_count_2000 - število kavarn in restavracij v območju 2 km
cafe_sum_2000_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 2 km
cafe_sum_2000_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 2 km
cafe_avg_price_2000 - povprečen račun v kavarni ali restavraciji v območju 2 km
cafe_count_2000_na_price - število kavarn in restavracij v območju 2 km, kjer ni podatka o ceni
cafe_count_2000_price_500 - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_2000_price_1000 - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša od 500 do 1000 rub
cafe_count_2000_price_1500 - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_2000_price_2500 - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_2000_price_4000 - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_2000_price_high - število kavarn in restavracij v območju 2 km, kjer povprečen račun znaša več kot 4000 rub
big_church_count_2000 - število velikih cerkva v območju 2 km
church_count_2000 - število cerkva v območju 2 km
mosque_count_2000 - število mošej v območju 2 km
leisure_count_2000 - število objektov za prosti čas v območju 2 km
sport_count_2000 - število objektov za prosti čas v območju 2 km
market_count_2000 - število tržnic v območju 2 km
green_part_3000 - delež zelenih površin v območju 3 km
prom_part_3000 - delež industrijskih površin v območju 3 km
office_count_3000 - število uradov v območju 3 km
office_sqm_3000 - površina uradov v območju 3 km
trc_count_3000 - število trgovskih centrov v območju 3 km
trc_sqm_3000 - površina trgovskih centrov v območju 3 km
cafe_count_3000 - število kavarn in restavracij v območju 3 km
cafe_sum_3000_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 3 km
cafe_sum_3000_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 3 km

cafe_avg_price_3000 - povprečen račun v kavarni ali restavraciji v območju 3 km
cafe_count_3000_na_price - število kavarn in restavracij v območju 3 km, kjer ni podatka o ceni
cafe_count_3000_price_500 - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_3000_price_1000 - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša od 500 do 1000 rub
cafe_count_3000_price_1500 - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_3000_price_2500 - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_3000_price_4000 - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_3000_price_high - število kavarn in restavracij v območju 3 km, kjer povprečen račun znaša več kot 4000 rub
big_church_count_3000 - število velikih cerkva v območju 3 km
church_count_3000 - število cerkva v območju 3 km
mosque_count_3000 - število mošej v območju 3 km
leisure_count_3000 - število objektov za prosti čas v območju 3 km
sport_count_3000 - število objektov za prosti čas v območju 3 km
market_count_3000 - število tržnic v območju 3 km
green_part_5000 - delež zelenih površin v območju 5 km
prom_part_5000 - delež industrijskih površin v območju 5 km
office_count_5000 - število uradov v območju 5 km
office_sqm_5000 - površina uradov v območju 5 km
trc_count_5000 - število trgovskih centrov v območju 5 km
trc_sqm_5000 - površina trgovskih centrov v območju 5 km
cafe_count_5000 - število kavarn in restavracij v območju 5 km
cafe_sum_5000_min_price_avg - najmanjši povprečen račun v kavarni ali restavraciji v območju 5 km
cafe_sum_5000_max_price_avg - največji povprečen račun v kavarni ali restavraciji v območju 5 km
cafe_avg_price_5000 - povprečen račun v kavarni ali restavraciji v območju 5 km
cafe_count_5000_na_price - število kavarn in restavracij v območju 5 km, kjer ni podatka o ceni
cafe_count_5000_price_500 - število kavarn in restavracij v območju 5 km, kjer povprečen račun znaša manj kot 500 rub
cafe_count_5000_price_1000 - število kavarn in restavracij v območju 5 km, kjer

povprečen račun znaša od 500 do 1000 rub
cafe_count_5000_price_1500 - število kavarn in restavracij v območju 5 km, kjer povprečen račun znaša od 1000 do 1500 rub
cafe_count_5000_price_2500 - število kavarn in restavracij v območju 5 km, kjer povprečen račun znaša od 1500 do 2500 rub
cafe_count_5000_price_4000 - število kavarn in restavracij v območju 5 km, kjer povprečen račun znaša od 2500 do 4000 rub
cafe_count_5000_price_high - število kavarn in restavracij v območju 5 km, kjer povprečen račun znaša več kot 4000 rub
big_church_count_5000 - število velikih cerkva v območju 5 km
church_count_5000 - število cerkva v območju 5 km
mosque_count_5000 - število mošej v območju 5 km
leisure_count_5000 - število objektov za prosti čas v območju 5 km
sport_count_5000 - število objektov za prosti čas v območju 5 km
market_count_5000 - število tržnic v območju 5 km
oil_urals - cena surove nafte
gdp_quart - bruto domači proizvod
gdp_quart_growth - realna rast bruto domačega proizvoda
cpi - indeks cen življenjskih potrebščin
ppi - indeks cen proizvodov
gdp_deflator - inflacija - deflator bruto domačega proizvoda
balance_trade - zunanjetrgovinski presežek
balance_trade_growth - trgovinska bilanca
usdrub - tečaj RUB/USD
eurrub - tečaj RUB/EUR
brent - Brent Crude v dolarjih na sodček
net_capital_export - neto uvoz/izvoz kapitala
gdp_annual - bruto domači proizvod glede na trenutne cene
gdp_annual_growth - letna rast bruto domačega proizvoda
average_provision_of_build_contract - povprečje zagotovljenih gradbenih pogodb za celotno Rusijo
average_provision_of_build_contract_moscow - povprečje zagotovljenih gradbenih pogodb za Moskvo
rts - indeks RTS
micex - indeks MICEX
micex_rgbi_tr - indeks MICEX za državne obveznice
micex_cbi_tr - indeks MICEX za podjetniške obveznice
deposits_value - vrednost depozitov

deposits_growth - rast depozitov
deposits_rate - povprečna obrestna mera za depozite
mortgage_value - vrednost hipotekarnih posojil
mortgage_growth - porast hipotekarnih posojil
mortgage_rate - povprečna obrestna mera za hipotekarna posojila
grp - bruto regionalni proizvod
grp_growth - rast bruto regionalnega proizvoda
income_per_cap - povprečen dohodek na prebivalca
real_dispos_income_per_cap_growth - rast v prihrankih na prebivalca
salary - povprečna mesečna plača
salary_growth - rast plač
fixed_basket - cena fiksne košarice proizvodov in storitev
retail_trade_turnover - promet na drobno
retail_trade_turnover_per_cap - promet na drobno na prebivalca
retail_trade_turnover_growth - rast prometa na drobno
rent_price_4room_bus - višina najemnine za 4-sobno stanovanje, višji razred
rent_price_3room_bus - višina najemnine za 3-sobno stanovanje, višji razred
rent_price_2room_bus - višina najemnine za 2-sobno stanovanje, višji razred
rent_price_1room_bus - višina najemnine za 1-sobno stanovanje, višji razred
rent_price_3room_eco - višina najemnine za 3-sobno stanovanje, ekonomski razred
rent_price_2room_eco - višina najemnine za 2-sobno stanovanje, ekonomski razred
rent_price_1room_eco - višina najemnine za 1-sobno stanovanje, ekonomski razred
apartment_build - število zgrajenih stanovanj
apartment_fund_sqm - površina vseh stanovanj
price_doc - cena nepremičnine

B Izbrani atributi Weka

full_sq - celotna površina
floor - nadstropje
num_room - število sob
kitch_sq - površina kuhinje
state - stanje nepremičnine
preschool_education_centers_raion - število predšolskih ustanov v rajonu
school_education_centers_raion - število šolskih ustanov v rajonu
healthcare_centers_raion - število zdravstvenih ustanov v rajonu
thermal_power_plant_raion - prisotnost termoelektrarne v rajonu
big_market_raion - prisotnost veletrgovskih centrov v rajonu
build_count_foam - delež stavb iz pene
industrial_km - razdalja do industrije v kilometrih
water_km - razdalja do vode npr. reke, jezera v kilometrih
zd_vokzaly_avto_km - razdalja do železniške postaje z avtom v kilometrih
catering_km - razdalja do cateringa v kilometrih
mosque_count_1000 - število mošej v območju 1 km
mosque_count_1500 - število mošej v območju 1,5 km
prom_part_5000 - delež industrijskih površin v območju 5 km
gdp_deflator - deflator bruto domačega proizvoda
price_doc - cena nepremičnine

C Izbrani atributi IRN ocena

full_sq - celotna površina
floor - nadstropje
max_floor - število vseh nadstropij
material - material izgradnje
num_room - število sob
kitch_sq - površina kuhinje
state - stanje nepremičnine
sub_area - rajon Moskve, kjer se nepremičnina nahaja
metro_min_avto - razdalja do metroja z avtom v minutah
gdp_deflator - deflator bruto domačega proizvoda
price_doc - cena nepremičnine