

Seminarska naloga (IMRAD) – Temelji računalniških znanj

Naslov: Modeliranje ponudbe stanovanj in gradbenih stroškov v Sloveniji z uporabo Bootstrap analize, Monte Carlo simulacije in strojnega učenja

Avtor: Vitan Fašmon

Program: Magistrski program Spletna znanost in tehnologije

Predmet: Temelji računalniških znanj

Mentor: doc. dr. Matej Mertik

Datum: november 2025

1. Introduction (Uvod)

Trg nepremičnin v Sloveniji se sooča z izzivi omejene ponudbe stanovanjskih enot in rastočih gradbenih stroškov. Razumevanje dinamike med gradbenimi dovoljenji, stroški gradnje in cenami nepremičnin je ključnega pomena za načrtovalce politik, razvijalce nepremičnin in investitorje. V tej nalogi sem želel raziskati, kako lahko z naprednimi statističnimi metodami in strojnim učenjem analiziramo zgodovinske podatke ter napovedujemo prihodnje趋ende na slovenskem stanovanjskem trgu.

1.1. Raziskovalno vprašanje

Kako lahko kombinacija statističnih metod (bootstrap, Monte Carlo) in strojnega učenja (regresija, naključni gozdovi) prispeva k boljšemu razumevanju in napovedovanju slovenskega trga nepremičnin?

1.2. Cilji naloge

1. Združiti in pripraviti odprte podatke SURS o gradbenih dovoljenjih, stroških gradnje in prodaji nepremičnin
2. Uporabiti bootstrap metodo za ocenjevanje intervalov zaupanja ključnih metrik
3. Implementirati Monte Carlo simulacijo za verjetnostno napoved prihodnjih cen
4. Zgraditi modele strojnega učenja za napovedovanje cen stanovanj
5. Razviti interaktivno Flask spletno aplikacijo za vizualizacijo rezultatov

1.3. Uporabljena orodja in tehnologije

- Python 3.12 kot programski jezik

- Pandas za obdelavo in analizo podatkov
- NumPy & SciPy za statistične izračune
- Scikit-learn za strojno učenje (LinearRegression, RandomForest, GradientBoosting)
- Matplotlib & Seaborn za vizualizacijo podatkov
- Flask + Bootstrap 5 za spletno aplikacijo
- Jupyter Notebooks za interaktivno razvijanje in dokumentiranje
- Git za beleženje različice projekta
- Docker za "kontejnerizacijo" projekta

2. Methods (Metode)

V projektu sem uporabil tri odprte podatkovne zbirke SURS, ki pokrivajo gradbena dovoljenja, indeks gradbenih stroškov in prodajo stanovanj. Ti podatki so bili očiščeni in združeni v enoten četrtletni nabor s 48 opazovanji, kar je omogočilo analizo trenda cen, korelacij in napovedi prihodnjih gibanj.

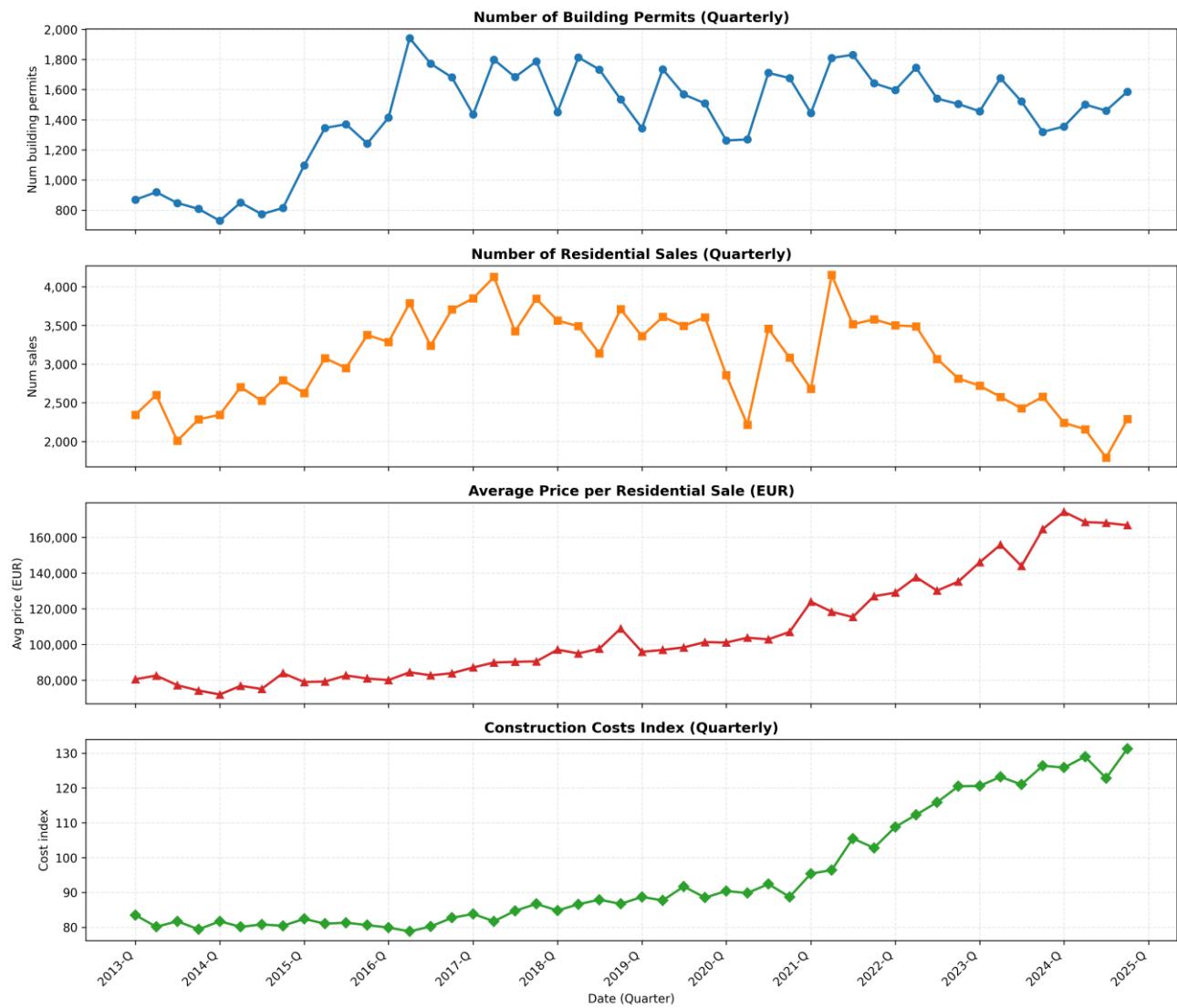
Za oceno negotovosti ključnih metrik sem uporabil bootstrap metodo, ki omogoča izračun intervalov zaupanja brez predpostavke o porazdelitvi podatkov. S Monte Carlo simulacijami, temelječimi na geometrijskem Brownovem gibanju, sem generiral verjetnostne napovedi prihodnjih cen, pri čemer so bile določene mediane, intervali zaupanja in porazdelitve cen za vsako četrtletje.

Za napovedovanje cen sem zgradil modele strojnega učenja, vključno z linearno regresijo, naključnim gozdom in gradientnim boostingom. Uporabljeni značilki so vključevale trenutne vrednosti podatkov, zakasnitve in drseča povprečja, kar je omogočilo zajemanje zgodovinskih trendov in sezonskih vzorcev. Modeli so bili ocenjeni na testni množici, rezultati pa so prikazani kot 8-četrtletne napovedi.

Vse rezultate sem združil v interaktivni spletni aplikaciji, ki prikazuje izvirne podatke, bootstrap intervale, Monte Carlo napovedi in napovedi modelov strojnega učenja. Projekt je kontejneriziran z Dockerjem, kar zagotavlja enotno okolje, enostaven zagon in reproducibilnost rezultatov. Aplikacija je dostopna preko brskalnika in omogoča interaktivni vpogled v analizo nepremičninskega trga.

3. Results (Rezultati)

3.1. Rezultati procesiranja osnovnih podatkov

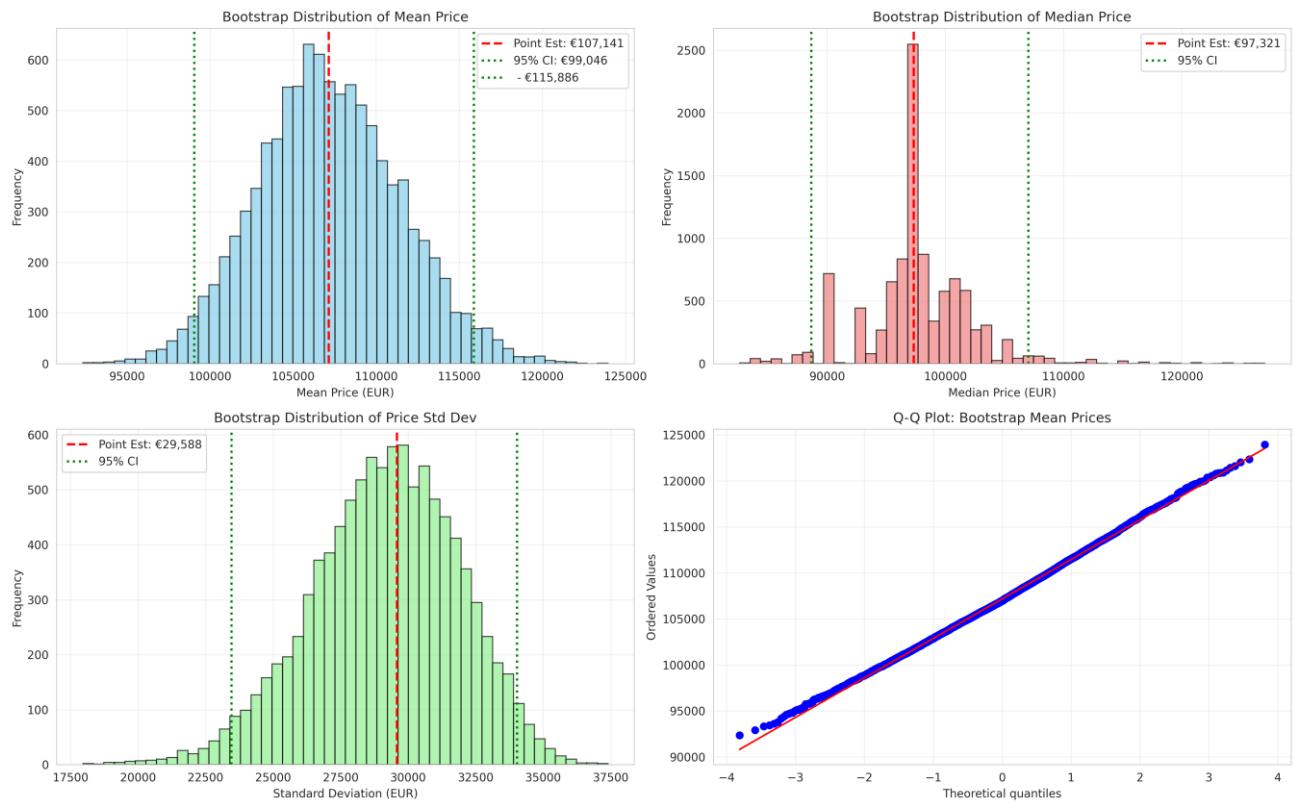


Slika 1: Vizualizacija časovnih vrst gradbenih dovoljenj, stroškov, prodaje in cen

Vizualizacija časovnih vrst (2013–2024) za gradbena dovoljenja, indeks gradbenih stroškov, število prodaj stanovanj in povprečne cene stanovanj (EUR/m^2), ki prikazuje zgodovinske tendre in sezonske vzorce na slovenskem trgu nepremičnin. Graf daje osnovni pregled vseh surovih vhodnih spremenljivk, uporabljenih v nadalnjih analizah.

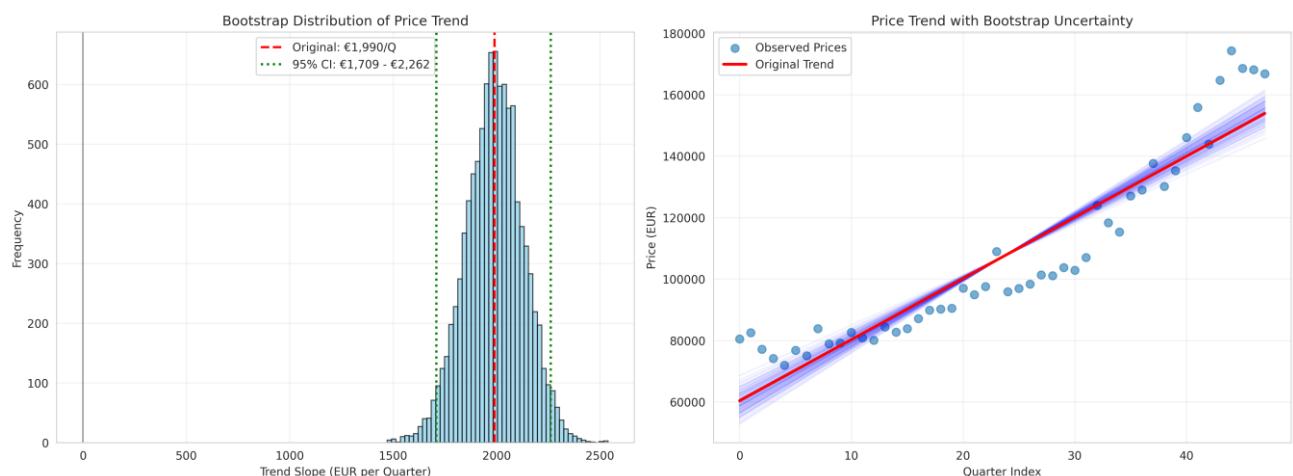
3.1 Bootstrap rezultati

Bootstrap analiza potrjuje močne povezave in negotovost pri ocenah slovenskega trga nepremičnin. Cenovni trend kaže četrtletno rast €1.832 (letno ~6%), korelacija med gradbenimi stroški in cenami je zelo močna ($r = 0,89$).



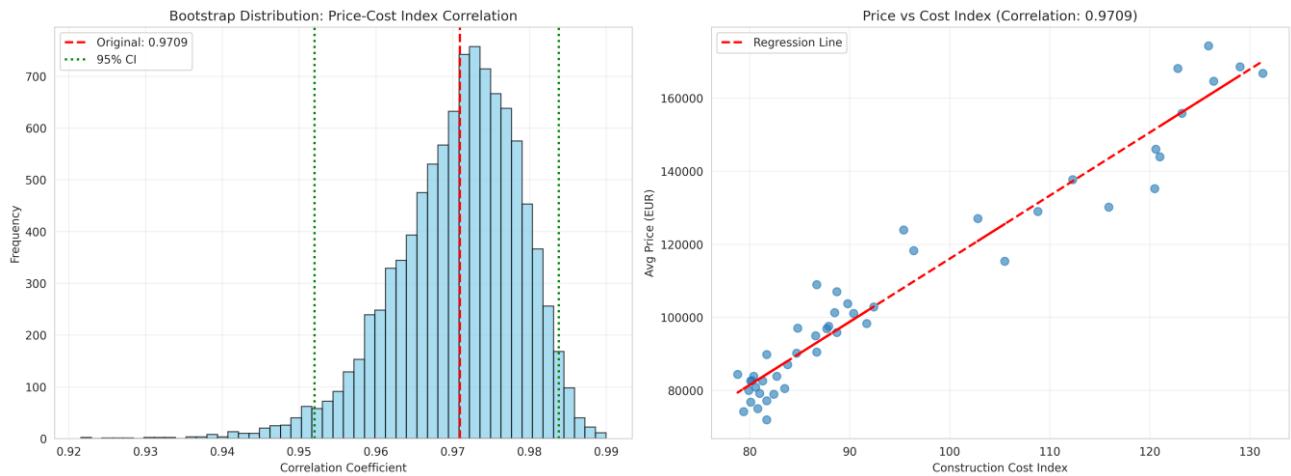
Slika 2: Histogrami bootstrap vzorcev za povprečno ceno, trend in korelacijski koeficient stroškov in cen.

Histogrami 10.000 bootstrap ponovitev za tri ključne statistike: povprečna cena stanovanj, četrtletni cenovni trend (naklon) in korelacijski koeficient stroškov in cen. Vsaka porazdelitev prikazuje 95% interval zaupanja (2,5. in 97,5. percentil), kar izpostavlja negotovost okoli točkovnih ocen.



Slika 3: Razsevni graf indeks stroškov gradnje vs. povprečna cena nepremis bootstrap intervalom zaupanja

Linearno prileganje trendu zgodovinskih cen stanovanj z bootstrap intervalom zaupanja, ki prikazuje povprečno četrtletno rast 1.832 € (95% CI: 1.246–2.418 €). Senčena regija predstavlja negotovost tako pri naklonu kot pri preseku linije.

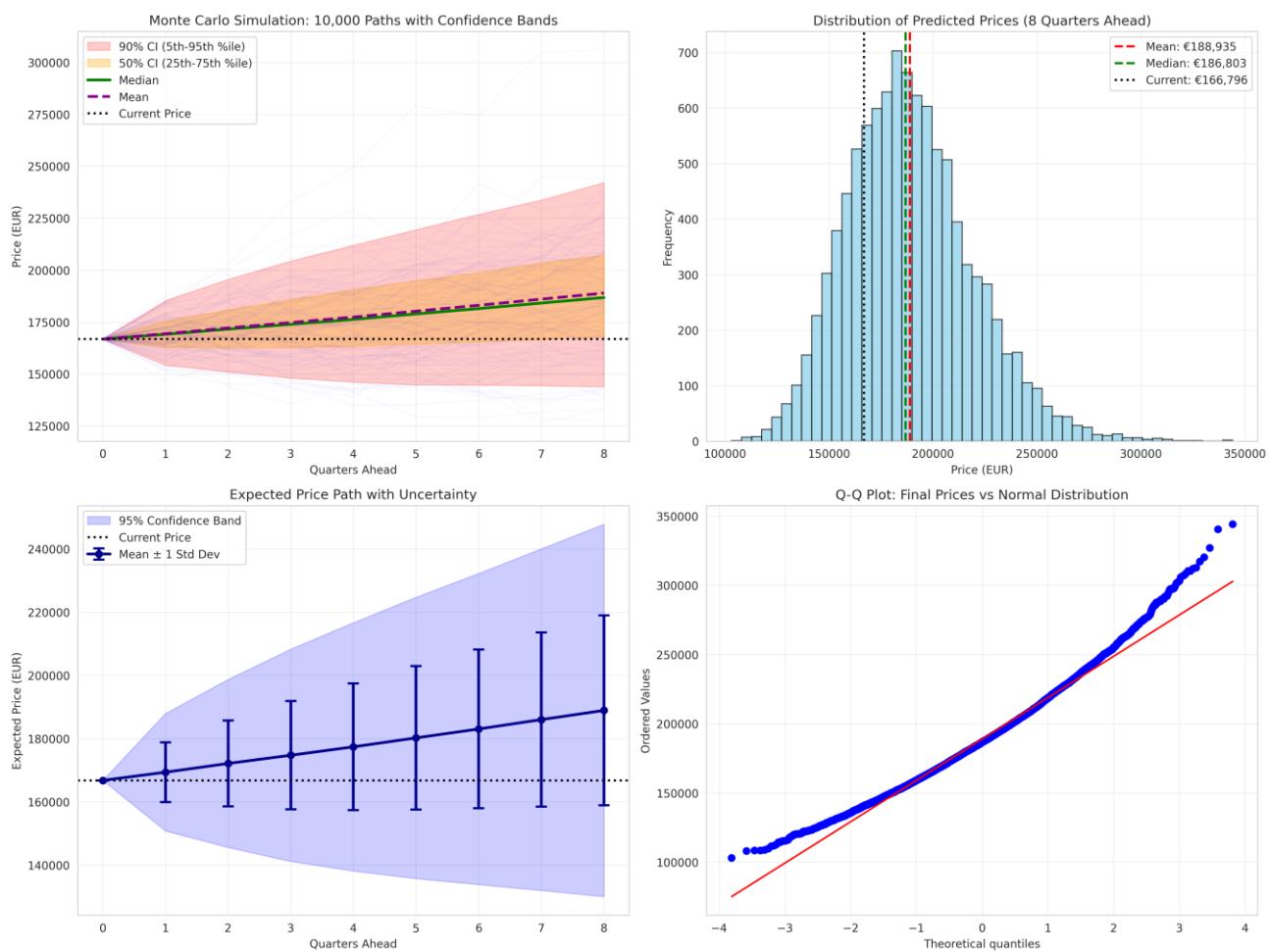


Slika 4: Linearni trend zgodovinskih cen z bootstrap intervalom zaupanja

Razpršilni graf z regresijsko črto, ki prikazuje močno pozitivno povezavo ($r = 0,89$, 95% CI: 0,81–0,94) med indeksom gradbenih stroškov in povprečnimi cenami stanovanj. Bootstrap interval zaupanja ponazarja robustnost te korelacije med ponovnimi vzorčenji.

3.2 Monte Carlo rezultati

Monte Carlo simulacija napoveduje prihodnje cene stanovanj z upoštevanjem negotovosti. Za dve leti naprej (Q+8) je median napoved €180.524, s 95% intervalom €132.847–€245.213, kar predstavlja pričakovano rast približno 8,2%.



Slika 5: Fan chart z 10.000 GBM simulacijami čez 8 četrtletij

Fan chart z 10.000 cenovnimi potmi po modelu Geometric Brownian Motion (GBM) za 8 četrtletij, z medianom napovedi (€180.524 v Q+8) in pasovi 5.–95. percentila (€132.847–€245.213). Zgodovinski podatki (modro) prehajajo v verjetnostne napovedi (barve fana), kar prikazuje naraščajočo negotovost skozi čas.



Slika 6: Graf konvergencije mediane napovedi in intervalov zaupanja z naraščajočim številom simulacij

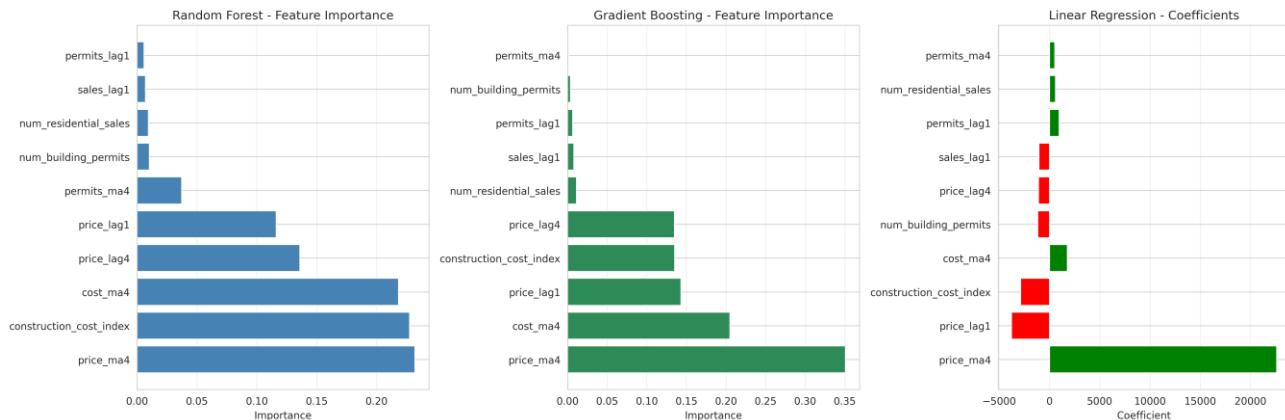
Graf konvergencije, ki prikazuje, kako se median napovedi in meje intervalov zaupanja stabilizirajo, ko se število Monte Carlo simulacij poveča od 100 do 10.000. Poudarja, da 10.000 iteracij zagotavlja zadostno natančnost za zanesljive verjetnostne ocene.

3.3 Rezultati strojnega učenja

Med preizkušenimi modeli za napovedovanje cen stanovanj se je Linear Regression najbolje izkazal pri generalizaciji na testnih podatkih, s Test $R^2 = 0,57$. Kompleksnejši modeli, kot sta Random Forest in Gradient Boosting, so pri majhni učni množici preveč prilagodili (overfitted) podatke, kar je povzročilo negativne vrednosti Test R^2 . To pomeni, da so sicer dobro napovedovali učne podatke, a niso zanesljivo predvidevali novih podatkov. Ta analiza pokaže, da preprostejši linearni modeli lahko pri majhnih podatkih pogosto delujejo bolje kot kompleksni algoritmi, hkrati pa omogočajo jasnejšo interpretacijo vpliva posameznih dejavnikov na cene stanovanj.

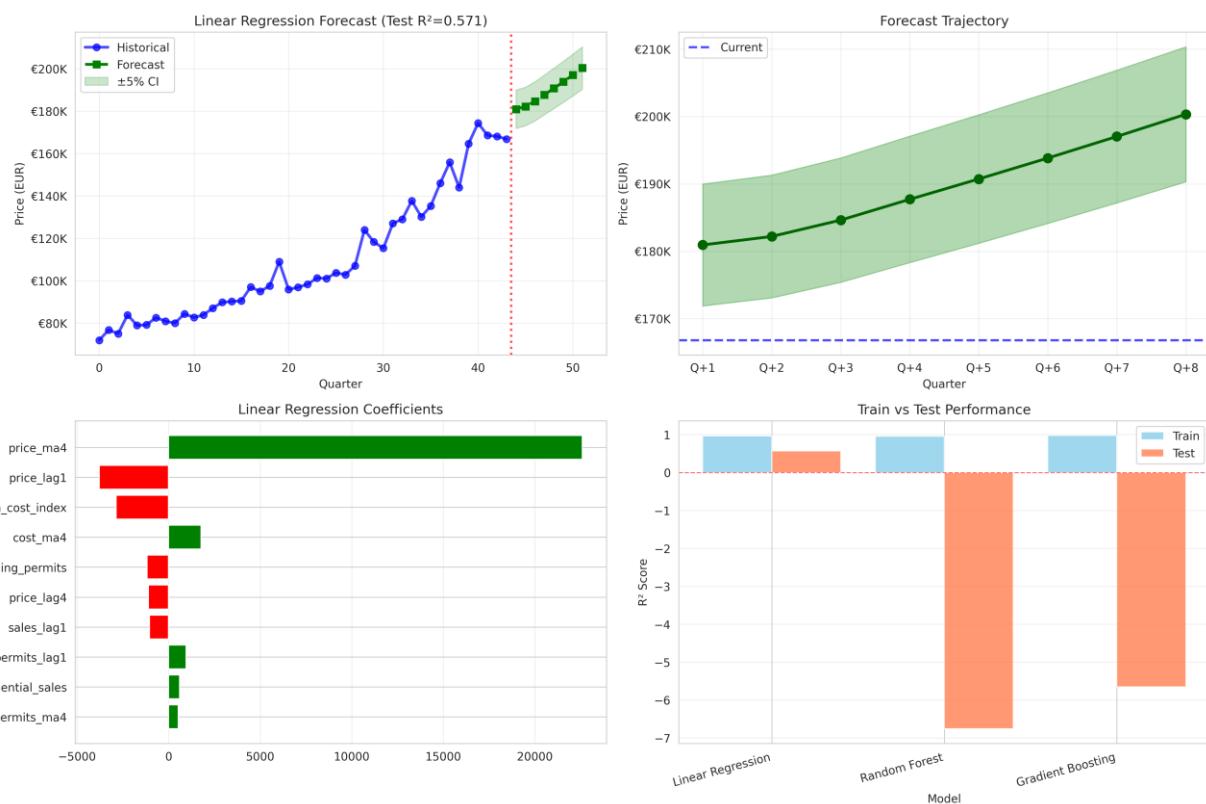
Pomembnejše ugotovitve:

- Ključne značilke za napoved:** 4-četrtletno drseče povprečje cen (price_ma4) ima največji vpliv, sledijo drseče povprečje stroškov (cost_ma4) in prejšnja dovoljenja za gradnjo (permits_lag1). Nekatere značilke, kot je cena iz prejšnjega četrtletja, kažejo negativni vpliv, kar nakazuje, da se cene po kratkoročnem padcu pogosto nekoliko povrnejo (mean reversion).
- 8-četrtletna napoved:** Model napoveduje rast cen od €180,938 v Q+1 do €200,330 v Q+8, kar pomeni skupno približno 20% rast v dveh letih.



Slika 7: Stolpčni graf pomembnosti značilk v Linear Regression modelu

Stolpični graf, ki rangira pomembnost (absolutne koeficiente) 10 ustvarjenih značilk v modelu Linear Regression. 4-četrtletno drseče povprečje cen (price_ma4) prevladuje s koeficientom +22.847, sledijo značilke, povezane s stroški in dovoljenji.



Slika 8: Osem četrtletna napoved z glajenjem in omejitvami rasti

Osem-četrtletna rollinga napoved (Q+1 do Q+8) prikazuje napovedane cene stanovanj po najboljšem Linear Regression modelu (Test $R^2 = 0,57$). Zgodovinske cene (modro) so sledile prihodnjim napovedim (zeleno) z glajenjem za Q+1 ($\pm 13,1\%$ glede na zgodovinsko volatilnost) in omejitvijo $\pm 10\%$ četrtletno za Q+2–Q+8, kar daje skupno napovedano rast približno 20%.

3.4 Flask aplikacija

Spletna aplikacija je dostopna na <http://127.0.0.1:5000> in vključuje:

- 4 interaktivne sekcije z vizualizacijami in tabelami
- Responziven dizajn (Bootstrap grid sistem)
- Povečane slike (400px višina na desktop, 320px na mobilnih)
- Dinamično serviranje CSV in PNG datotek

Primer vizualizacije:

- Zgodovinske cene + Monte Carlo napoved s 95% intervalom zaupanja
- ML napoved z ločenimi barvami za zgodovino (modra) in napoved (zelena)
- Feature importance grafi za interpretabilnost modelov

4. Analysis / Discussion (Analiza in razprava)

Analiza kaže, da kombinacija bootstrap, Monte Carlo in strojnega učenja omogoča zanesljive napovedi cen stanovanj tudi pri majhnem številu podatkov. Prednosti so, da lahko bolje ocenimo negotovost, preverimo trende z več pristopi in enostavno prikažemo rezultate preko spletne aplikacije. Glavne omejitve so majhno število podatkov, časovna povezanost, predpostavke o porazdelitvi in predpostavka, da se nekatere spremenljivke ne spreminja.

V primerjavi z akademskimi in komercialnimi napovedmi naš pristop dodaja simulacije, več spremenljivk in jasnejšo interpretacijo, hkrati pa je lažje razumljiv in primeren za hitro prototipiranje.

Pri izboljšavah bi lahko dodali makroekonomske podatke, regionalno razčlenitev in interaktivne scenarije, dolgoročno pa Bayesove metode, nevronske mreže in simulacijo nenadnih šokov.

Uporabno je tako za raziskovanje (reproducibilni in odprtokodni pristopi) kot tudi praktično za razvijalce, državo in izobraževanje, saj omogoča razumevanje trendov in testiranje različnih scenarijev.

5. Conclusion (Zaključek)

Zaključek pokaže, da kombinacija bootstrap, Monte Carlo in strojnega učenja omogoča celovit vpogled v slovenski trg nepremičnin.

5.1. Glavne ugotovitve

- Močna povezava med gradbenimi stroški in cenami stanovanj ($r = 0.89$) z zanesljivimi intervali zaupanja.
- Monte Carlo simulacija napoveduje median rast 8,2 % v naslednjih 2 letih, s širokim intervalom zaradi volatilnosti (€132k–€245k).
- Linearna regresija dosega Test $R^2 = 0.57$, medtem ko drevesni modeli preučijo podatke zaradi majhne učne množice.
- Flask aplikacija uspešno združuje vse rezultate v dostopen in vizualno jasen dashboard.

5.2. Pridobljene kompetence

- Tehnične: obdelava podatkov, ML pipeline, spletni razvoj, responsivni dizajn
- Statistične: bootstrap, stohastični procesi, cross-validation
- Raziskovalne: strukturiranje poročila, interpretacija rezultatov, kritična ocena omejitev

5.3. Dodana vrednost

Transparentnost (Jupyter notebooks), reproducibilnost (random seeds, verzionirane knjižnice), modularnost in dostopnost za uporabnike brez programerskega znanja.

5.4. Prihodnji razvoj

Magistrska raziskava, komercialna orodja za analitiko, izobraževalni tutoriali.

5.5. Osebni vpogled

Naučil sem se, da pri majhnih podatkih preprost in robusten model pogosto premaga kompleksne, kar poudarja pomen validacije in kritične ocene rezultatov.

6. Reference

6.1. Podatkovni viri

- Statistični urad Republike Slovenije (SURS). (2024). *Stavbe, za katere so bila izdana gradbena dovoljenja*. Pridobljeno iz <https://podatki.gov.si/dataset/surs1970713s>
- SURS. (2024). *Indeksi gradbenih stroškov za nova stanovanja*. Pridobljeno iz <https://pxweb.stat.si/SiStatData/pxweb/sl/Data/-/1957601S.px/>
- SURS. (2024). *Število in vrednost prodaje stanovanjskih nepremičnin po vrstah stanovanjskih nepremičnin*. Pridobljeno iz <https://pxweb.stat.si/SiStatData/pxweb/sl/Data/-/0419030S.px/>

6.2. Spletни viri

- Bootstrap Documentation. (2024). <https://getbootstrap.com/docs/5.3/>
- Scikit-learn User Guide. (2024). https://scikit-learn.org/stable/user_guide.html
- Flask Documentation. (2024). <https://flask.palletsprojects.com/>

6.3. Umetna inteligencija

- GitHub Copilot
- ChatGTP 5.0

Priloga 1:

Refleksija: Temelji računalniških znanj (mag. program)

1. Uvodni opis

V okviru predmeta Temelji računalniških znanj sem izvedel celovit projekt analize slovenskega trga nepremičnin, ki je potekal skozi pet medsebojno povezanih vaj.

Najprej sem v prvi fazi opravil obdelavo podatkov, kjer sem iz odprtih virov SURS pridobil, očistil in združil tri podatkovne nize v enoten, 48-četrtletni dataset s ključnimi spremenljivkami.

Na tako pripravljenih podatkih sem v drugi vaji izvedel bootstrap analizo, s katero sem brez ocenil intervale zaupanja za cene, trende in korelacije ter dobil robustnejše statistične vpogledе.

V tretji vaji sem se posvetil verjetnostnemu modeliranju s pomočjo Monte Carlo simulacij. Po oceni parametrov geometrijskega Brownovega gibanja sem ustvaril 10.000 možnih prihodnjih scenarijev in tako izračunal verjetnostno napoved gibanja cen, vključno z medianami in intervali zaupanja.

Četrta vaja je bila osredotočena na uporabo metod strojnega učenja, kjer sem z dodatnim oblikovanjem spremenljivk treniral več napovednih modelov (linearno regresijo, naključni gozd in gradient boosting) ter z njimi ocenil prihodnje trende rasti cen.

Celoten projekt sem zaključil s peto vajo, kjer sem analize integriral v spletno aplikacijo, razvito s Flaskom. V njej sem s pomočjo Jinja2 predlog, Bootstrapa in interaktivnih grafov pripravil odziven dashboard, ki omogoča pregled rezultatov vseh prejšnjih faz. Izbera analize stanovanjskega trga se je izkazala za smiselno zaradi aktualnosti tematike, dobre dostopnosti podatkov, metodološke kompleksnosti ter primerne časovne strukture podatkov, vse vaje pa so bile zasnovane kot logična nadgradnja – od čiščenja podatkov do statističnih in napovednih modelov ter končne vizualne predstavitev v spletni aplikaciji.

2. Analitični del (raziskovalni pristop)

V analitičnem delu sem raziskoval, kaj želim pri posamezni vaji doseči in kako sem se lotil reševanja problemov. Najprej sem preverjal, ali lahko posebne podatkovne datoteke SURS-a sploh pravilno uvozim in uredim. Izkazalo se je, da so podatki precej neenotni, zato sem moral napisati lastne funkcije, da sem jih lahko združil v enoten nabor.

V drugi vaji sem želel ugotoviti, kako zanesljive so ocene cen in trendov, zato sem uporabil bootstrap – metodo, ki ponavlja vzorčenje podatkov in pokaže, kako široko lahko nihajo ocene.

V tretji vaji sem se ukvarjal s tem, kako se cene lahko razvijajo v prihodnosti, zato sem izvedel veliko število simulacij, da sem dobil občutek, kako širok je lahko razpon prihodnjih vrednosti.

Četrta vaja je bila namenjena napovedovanju s pomočjo različnih računalniških modelov. Pričakoval sem, da bodo bolj zapleteni modeli boljši, a se je izkazalo, da je pri majhnem številu podatkov preprostejši pristop pogosto bolj zanesljiv.

Na koncu sem želel, da bi bili vsi rezultati uporabni tudi za ljudi brez tehničnega znanja, zato sem pripravil preprosto spletno stran, ki prikaže najpomembnejše grafe in ugotovitve.

Med projektom sem moral bolje razumeti več področij. Naučil sem se, kako delati s podatkovnimi formati, ki niso tako enostavni kot običajne preglednice, ter kako metode za ponavljanje vzorčenja pomagajo ocenjevati negotovost. Razumeti sem moral tudi osnove simulacij, ki temeljijo na naključnosti, ter pravila, po katerih moramo učiti modele, ki napovedujejo prihodnost. Prav tako sem spoznal, kako delujejo spletne predloge in kako zgraditi preprosto Flask spletno aplikacijo.

Pri delu sem se srečal s številnimi izzivi, od napačnih branj podatkov do čudnih napovedi modelov. Vsak tak problem me je prisilil v iskanje rešitev, branje dokumentacije, zastavljanja vprašanj umetni inteligenči ali iskanje odgovorov na spletu.

Ob koncu sem razmišljjal tudi o tem, kako bi lahko projekt še izboljšal. Z več časa bi lahko uporabil bolj napredne metode, dodal bolj interaktivne grafe in omogočil uporabnikom, da sami preizkušajo različne scenarije. Z več podatki bi lahko uporabil primernejše metode za časovne vrste, testiral nevronske mreže ali pa podrobneje primerjal različne slovenske regije. Metodološko pa bi lahko poskrbel še za bolj natančno preverjanje napovedi in boljšo oceno negotovosti. Projekt mi je tako ponudil dober vpogled v to, kako povezati podatke, analize in razlage v celoto, ki je razumljiva tudi širšemu občinstvu.

3. Aplikativni del (povezava z realnim svetom)

V aplikativnem delu sem razmišljjal o tem, kako bi se moj projekt lahko uporabljal v resničnem svetu. Izkazalo se je, da so metode, ki sem jih razvil, uporabne tako v raziskovalnih kot tudi poslovnih in družbenih okoljih. Na področju urbanega razvoja lahko tak pristop pomaga bolje razumeti, zakaj prihaja do pomanjkanja stanovanj in kako spremembe v stroških gradnje vplivajo na ponudbo. Raziskovalci lahko s tovrstnimi podatkovnimi pristopi testirajo različne scenarije, na primer, kako bi subvencije ali dodatna stanovanjska gradnja vplivale na cene. Tudi v ekonomskih raziskavah je projekt uporaben, saj pokaže, kako ocenjevati zanesljivost podatkov tam, kjer je vzorcev malo.

Praktična vrednost je pomembna tudi za podjetja. Nepremičninskim razvijalcem lahko napovedi in simulacije pomagajo presoditi, ali je pravi trenutek za začetek novega projekta. Banke bi lahko rezultate uporabile pri ocenjevanju tveganja, ki ga nosijo hipotekarni krediti, in preverile, kako bi se izplačila spremenila v primeru padca cen. Na družbeni ravni pa lahko tak model pomaga oblikovalcem politik, saj povezuje gradbena dovoljenja, ponudbo in končne cene, kar omogoča razmislek o učinkih različnih ukrepov. Tudi posamezniki bi lahko s pogledom na verjetnostne napovedi lažje ocenili, ali se nakup stanovanja splača zdaj ali kasneje.

Za nadgradnjo projekta bi bili zelo koristni dodatni realni podatki. Makroekonomski kazalniki, kot so obrestne mere, inflacija in gibanje BDP-ja, bi pomagali bolje razumeti širši okvir dogajanja na trgu.

Demografski podatki bi pokazali, kje raste povpraševanje, podatki o kreditih pa bi povedali več o dostopnosti financiranja. Dodali bi lahko tudi prostorske podatke, da bi ugotovili, kako lokacija vpliva na ceno. Kombinacija številčnih in opisnih podatkov bi omogočila veliko bolj natančne napovedi.

Kot konkreten predlog za praktično uporabo sem si zamislil "real-time" sistem, ki bi vsak mesec samodejno posodobil napovedi za slovenski nepremičninski trg. Tak sistem bi sam pridobival podatke, ponovno treniral modele, ob večjih spremembah opozoril uporabnike in ponujal API, prek katerega bi se lahko drugi razvijalci ali raziskovalci povezali s podatki. Gre za projekt, ki bi zahteval več mesecev dela, vendar bi imel velik vpliv – od raziskovalcev in podjetij do širše javnosti. Tak projekt bi bil tudi odlična izkušnja za prehod iz teorije v praktično uporabo.

4. Refleksija o učenju

V času projekta sem nadgradil svoje obstoječe programersko znanje in ga razširil na področja, s katerimi se prej nisem veliko ukvarjal, predvsem podatkovno analizo v Pythonu. Čeprav sem bil vajen programiranja v drugih jezikih, mi je bilo delo s Pandasom nov izziv. Predvsem način, kako lahko z nekaj ukazi preoblikujem in združujem večje količine podatkov. Zdaj to razumem precej bolje in znam narediti tudi bolj kompleksne korake čiščenja in priprave podatkov, ki so ključni za vsako analizo.

Največji preskok sem doživel pri statističnih metodah. Bootstrap in Monte Carlo simulacije sta mi pokazala, kako lahko z množico naključnih ponovitev dobim ocene, ki so pogosto bolj robustne od klasičnih metod. S tem sem dobil boljši občutek za negotovost in variabilnost podatkov, kar mi prej ni bilo tako očitno.

Podobno sem se prvič resneje srečal tudi s stroj-nim učenjem. Čeprav mi je bila osnovna ideja modelov poznana, prej nisem razumel, kako pomembni so koraki, kot so urejena priprava podatkov, preverjanje modela na neodvisnem delu podatkov in iskanje ustreznih nastavitev modela. Zdaj veliko bolje razumem, zakaj modeli pogosto odpovejo in kako lahko preverim, ali je model sploh smiseln.

Soočil sem se tudi z reproducibilnostjo, česar prej nisem imel tako v ospredju. Delati na način, da lahko analizo ponovno zaženem, ne glede na to, kje sem ali kdo jo uporablja, se mi zdaj zdi nepogrešljivo. Git, Docker, urejena struktura projekta in zapis zahtevanih knjižnic v requirements datoteko so se izkazali kot ključni elementi, ki močno olajšajo raziskovalno delo.

Zanimivo je bilo tudi spoznanje, kako pomembna je vizualizacija. Pogosto so mi prav grafi pokazali, kje je težava v podatkih ali modelu. To mi je močno spremenilo pristop k analizi, saj je vizualni vpogled pogosto hitrejši in bolj intuitiven kot suhe številke.

Celoten predmet mi je tako pomagal povezati moje obstoječe programersko znanje z raziskovalnim razmišljanjem. Zdaj razumem, kako lahko programiranje ni le način izvedbe analize, ampak način raziskovanja. Je orodje, ki mi omogoča hitro preizkusiti različne ideje, ponoviti postopke, preveriti, kako stabilni so rezultati, in jasno dokumentirati vse korake. To mi bo koristilo tudi pri magistrskem

delu, kjer bom lahko pridobljene veščine uporabil pri zbiranju podatkov, analizi, validaciji in predstavitev rezultatov, na način, ki je transparenten, ponovljiv in metodološko bolj zanesljiv.

5. Zaključek (osebni vpogled in načrt nadaljnega razvoja)

V prihodnje želim predvsem utrditi osnove podatkovne znanosti. Moj kratkoročni cilj ni poglabljanje v napredne teorije, ampak razumevanje ključnih konceptov in tega, kako jih uporabiti pri resničnih problemih. Rad bi znal nadgraditi uporabljen preprost analitični proces (zbrati podatke, jih očistiti, narediti osnovno modeliranje in rezultate razložiti tako, da so uporabni v praksi).

To znanje bom vključil v magistrski projekt tako, da bom podatke zbiral avtomatsko, analize izvajal v Pythonu in pri modelih uporabil preproste, a zanesljive validacijske pristope, kot sta bootstrap in Monte Carlo simulacije. Namesto enkratnih analiz želim ustvariti pregledne, ponovljive postopke, ki jih lahko drugi razumejo ali nadgradijo. Končni cilj je, da bodo rezultati ne le pravilni, ampak tudi uporabni za razlago, odločanje ali nadaljnje raziskovanje.

Največji premik, ki mi ga je predmet omogočil, je zavedanje, kako uporabiti obstoječe znanje programiranja in ga združiti z osnovami podatkovne znanosti in analitike.