Višestruka regresija za predikciju cijene automobila u Srbiji

Teodor Vidaković

Jul 2025

1 Eksplorativna analiza podataka

Dataset sadrži informacije o oglasima polovnih automobila u Srbiji. Skup podataka ima ukupno 1756 primjera i 10 kolona, pri čemu ne postoje nedostajuće vrednosti. Korišćene kolone uključuju: `Marka`, `Grad`, `Godina proizvodnje`, `Karoserija`, `Gorivo`, `Zapremina motora`, `Kilometraža`, `Konjske snage`, `Menjač` i ciljnu promenljivu `Cena`.

1.1 Osnovne statistike

Nakon uklanjanja duplikata, ostaje 1353 validnih zapisa. Osnovne statistike za numeričke promjenljive prikazane su u Tabeli 1.

Tabela 1: Osnovne statistike numeričkih promenljivih

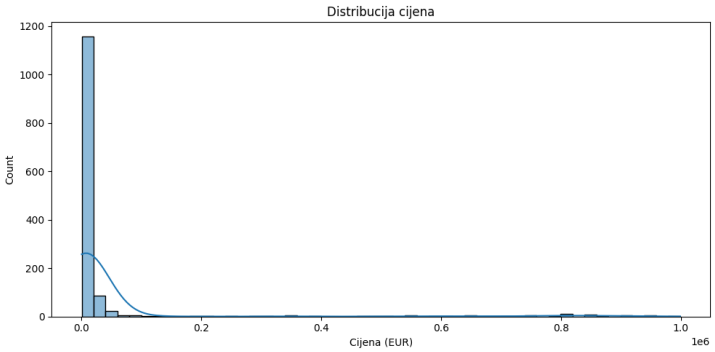
Atribut	**Min**	**(25%)**	**(50%)**	**(75%)**	**Max**	**Srednja**
Cijena [EUR]	1 000	3 000	5 900	13 500	999 000	45 027
Godina proizvodnje	1894	2006	2010	2015	2025	2010
Zapremina motora [cm^3]	163	1 398	1 600	1 984	5 461	1 742.6
Kilometraža [km]	1 000	165 563	209 000	256 000	552 000	206 512
Konjske snage	45	92	116	150	620	129.5

Vidimo da su cijene veoma varijabilne, pri čemu se maksimalna cijena kreće i do 999 000 EUR, što ukazuje na prisustvo outlier-a. Većina automobila ima pređeno između 150 000 i 300 000 kilometara.

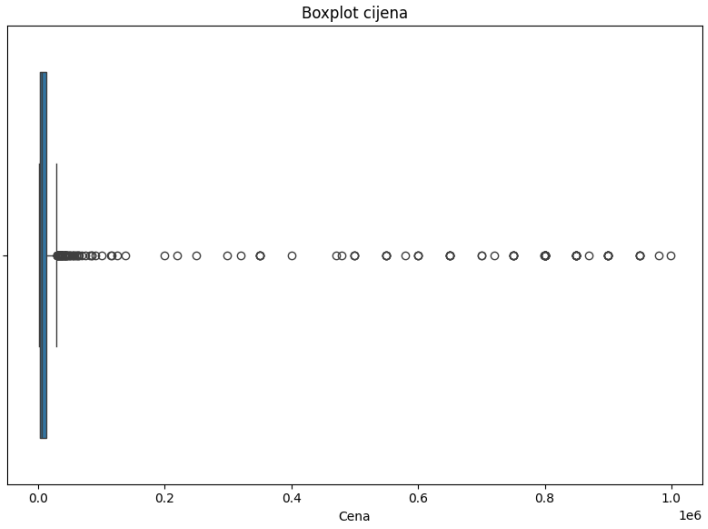
1.2 Vizuelizacije

U cilju grafičke analize podataka, korišćeni su sledeći prikazi:

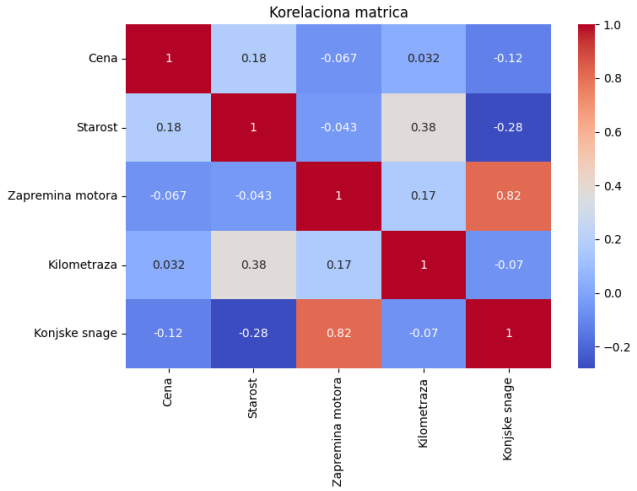
- **Histogram distribucije cijena** (Slika 1)
- **Boxplot cijena** (Slika 2)
- **Korelaciona matrica numeričkih promenljivih** (Slika 3)
- **Scatterplot: Cijena vs. Starost vozila** (Slika 4)



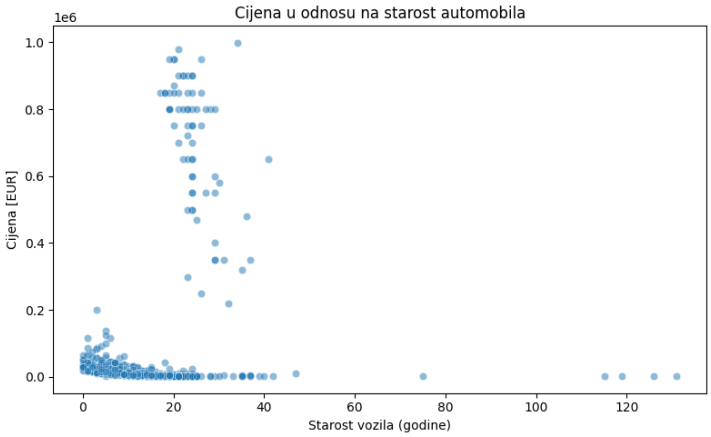
Slika 1: Distribucija cijena automobila



Slika 2: Boxplot cijena automobila



Slika 3: Korelaciona matrica numeričkih atributa



Slika 4: Raspodjela cijene u odnosu na starost vozila

1.3 Uočeni obrasci i anomalije

Na osnovu vizuelne analize identifikovani su sledeći obrasci:

- **Distribucija cijena je snažno asimetrična** (Slika 1) – većina vozila se prodaje ispod 20 000 EUR, dok manji broj primjera sa cijenama od 100 000+ EUR značajno odskače.
- **Prisustvo outlier-a potvrđeno je boxplot-om** (Slika 2) – postoje brojni ekstremi koji narušavaju raspodjelu ciljne promjenljive i negativno utiču na RMSE.

- **Starost vozila i cijena imaju blago pozitivnu korelaciju** ($r = 0.18$), što je neočekivano. Međutim, scatterplot (Slika 4) pokazuje da najskuplji automobili nisu novi, već stari — vjerovatno luksuzni, sportski ili kolekcionarski modeli. Većina novijih automobila (0–5 godina starosti) ima prosječnu ili nisku cenu, dok ekstremno visoke vrednosti dolaze od nekoliko skupocjenih vozila koja su stara 10+ godina. Ova nelinearna struktura dovodi do "lažne" pozitivne korelacije u linearnom smislu, iako bi se očekivala negativna veza između cijene i starosti.
- **Kilometraža ima vrlo slab odnos sa cijenom** ($r = 0.03$) – korisnici očigledno ne rangiraju ovaj faktor visoko pri određivanju cene.
- **Konjske snage i zapremina motora su visoko korelisani** ($r = 0.82$, Slika 3) – što je i očekivano s obzirom na tehničku povezanost tih karakteristika.
- **Kategorijske promenljive su raznovrsne i zahtevaju kodiranje**:
  - Marka: 44 vrednosti
  - Grad: 236 različitih mesta (visoka disperzija)
  - Karoserija i Gorivo: po 8 klasa
  - Menjač: 2 klase (Manuelni i Automatski)

2 Preprocesiranje podataka

Pre nego što je model treniran, podaci su prošli kroz preprocesiranje sa ciljem da se obezbjedi robusnost modela i smanji uticaj ekstremnih vrednosti.

- **Filtriranje cijena** – Na osnovu prethodne analize, identifikovani su automobili sa ekstremno visokim cijenama (iznad 100 000 EUR), koji značajno narušavaju distribuciju ciljne promenljive i utiču na RMSE. Umesto kvantilnog filtra, primenjena je jednostavna prag-filtracija: u dalji rad uključeni su samo primeri sa cijenom manjom ili jednakom **50 000 EUR**. Time je uklonjen uticaj luksuznih i atipičnih vozila.
- **Uklanjanje duplikata** – Prije treniranja modela, duplirani oglasi su uklonjeni pomoću `drop_duplicates()` metode.
- **Inženjering osobina** – Iz atributa `Godina proizvodnje` kreirana je nova numerička promjenljiva `Starost`, izračunata kao $2025 - \text{Godina proizvodnje}$. Ova osobina bolje opisuje uticaj vremena na gubitak vrijednosti vozila i zamjenjuje sirovu godinu.

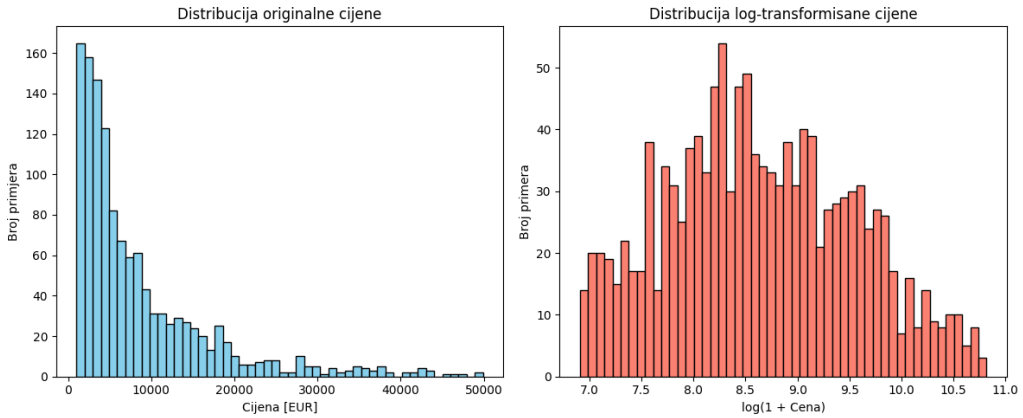
- **Obrada numeričkih podataka** – Odabrane numeričke promjenljive su:
  - `Starost`
  - `Zapremina motora`
  - `Kilometraža`
  - `Konjske snage`

  Kako bi model mogao da uoči nelinearne zavisnosti, na njih je primjenjena polinomska transformacija drugog stepena. Nakon toga, sve numeričke osobine su standardizovane (`StandardScaler`), što znači da je na njih primenjena z-score transformacija.

- **Obrada kategoričkih podataka** – Sledeće kolone su tretirane kao kategoričke:
  - `Marka`, `Grad`, `Karoserija`, `Gorivo`, `Menjač`

  Svaka od njih je kodirana tehnikom **One-Hot Encoding** kako bi se izbjegle greške pri predikciji na test skupu.

- **Log-transformacija ciljne promenljive** – Cijena automobila, kao ciljna promjenljiva, ima izraženu asimetričnu raspodjelu sa teškim desnim repom. Da bi se smanjio uticaj outlier-a i poboljšala stabilnost modela, izvršena je log-transformacija putem funkcije $\log(1 + y)$ pomoću `FunctionTransformer`. Predikcije su potom inverzno transformisane pomoću $\exp(y) - 1$.



Slika 5: Distribucija ciljne promenljive prije i poslije log-transformacije

Nakon svih koraka, numeričke i kategoričke komponente objedinjene su pomoću `ColumnTransformer`. Transformacija ciljne promjenljive vršena je korišćenjem `TransformedTargetRegressor`, čime se obezbjeđuje konzistentna transformacija pri treniranju i predikciji.

3 Podjela skupa podataka

Za treniranje i evaluaciju modela korišćena su dva režima podjele podataka, u zavisnosti od konteksta pokretanja:

- **Lokalna podjela podataka (lokalna evaluacija):** Kada se pokreće lokalno (npr. tokom eksperimentisanja), koristi se samo jedan skup podataka. U tom slučaju, podaci se dijele na trening i test skup pomoću funkcije `train_test_split` u odnosu 80:20. Da bi se očuvala reprezentativna distribucija ciljne promenljive, koristi se stratifikacija po cjenovnim opsezima.
- **Eksplicitna podela putem komandne linije:** Kada se skripta pokreće sa dva ulazna fajla (npr. `python model.py train.tsv test.tsv`), koristi se unaprijed pripremljeni trening i test skup, bez dodatne podjele unutar skripte. Ova varijanta se koristi za finalnu evaluaciju modela na test skupu koji nije korišćen u fazi treniranja.

Nakon što je trening skup definisan, primjenjena je **K-Fold unakrsna validacija** (`KFold`) sa brojem preklopa $k = 5$, uz opciju `shuffle=True` i fiksiran `random_state=42` radi reproduktivnosti. Ova validacija ima dvostruku ulogu:

- Unutar modela koristi se za automatsku selekciju optimalnog hiperparametra α.
- Paralelno se koristi za procjenu performansi modela korišćenjem metrike RMSE, preko funkcije `cross_val_score`.

Ovaj pristup omogućava robusnu procjenu generalizacione sposobnosti modela bez curenja podataka iz test skupa. Nakon toga, model se trenira na cijelom trening skupu i koristi za evaluaciju na izdvojenom test skupu.

4 Isprobani algoritmi

Za zadatak višestruke regresije testirana su tri regularizovana linearna modela iz biblioteke `scikit-learn`: `LassoCV`, `RidgeCV` i `ElasticNetCV`. Svi modeli korišćeni su unutar istog pipeline-a.

4.1 Podešavanje hiperparametara

- **LassoCV** bira optimalnu vrijednost regularizacionog parametra α iz logaritamski raspoređenog opsega od 10^{-4} do 10^{2} primjenom 5-fold unakrsne validacije.
- **RidgeCV** koristi isti opseg α vrijednosti i takođe primjenjuje 5-fold unakrsnu validaciju.
- **ElasticNetCV** osim α optimizuje i odnos između L_1 i L_2 regularizacije putem hiperparametra `l1_ratio`, pri čemu su testirane vrednosti $\{0.1, 0.5, 0.9\}$.

4.2 Rezultati

Rezultati evaluacije prikazani su kroz metriku RMSE (na originalnoj skali, nakon inverzne log-transformacije), na osnovu 5-fold unakrsne validacije i konačnog lokalnog test skupa.

Tabela 2: Uporedni rezultati modela sa 5-fold CV i lokalnim test skupom

Model	**CV RMSE (mean ± std)**	**Test RMSE**	**Izabrani hiperparametri**
LassoCV	3045.09 ± 250.22	2638.25	$\alpha = 0.00107$
RidgeCV	3146.33 ± 313.29	2717.87	$\alpha = 4.03702$
ElasticNetCV	3049.26 ± 269.87	2657.75	$\alpha = 0.00187$, $l_1 = 0.50$

4.3 Diskusija

Sva tri modela uspješno zadovoljavaju uslov RMSE ¡ 6500 EUR na lokalnom test skupu. Najbolje performanse ostvario je **LassoCV**, zahvaljujući sposobnosti selekcije relevantnih osobina putem L_1 regularizacije. Ridge je pokazao nešto lošije rezultate, dok je ElasticNet dao kompromis između prethodna dva pristupa.

5 Odabrano rešenje

Nakon eksperimentisanja sa različitim tehnikama preprocesiranja i modelima, kao najefikasnije rješenje izabran je model baziran na `LassoCV` regresiji sa sledećom konfiguracijom:

- **Preprocesiranje podataka**:
  - Numeričke promjenljive: proširene polinomskim osobinama drugog stepena (`PolynomialFeatures`) i zatim standardizovane pomoću Z-score standardizacije (`StandardScaler`).
  - Kategoričke promenljive: kodirane One-Hot encoding-om sa ignorisanjem nepoznatih vrijednosti u test skupu.
  - Inženjering osobina: uvedena nova numerička promjenljiva – starost automobila (`Starost = 2025 - Godina proizvodnje`).
- **Transformacija ciljne promjenljive**:
  - Zbog izražene asimetrične distribucije cijena sa dugim desnim repom, izvršena je log-transformacija ciljne promjenljive pomoću $\log(1 + y)$, implementirana preko `FunctionTransformer`.
  - Predikcije su inverzno transformisane sa $\exp(y) - 1$ kako bi se vratile na originalnu skalu.
- **Model**:
  - Regresor: `LassoCV` sa unakrsnom validacijom (`cv=5`) i automatskom selekcijom optimalnog hiperparametra α iz logaritamski raspoređenog skupa vrijednosti između 10^{-4} i 10^{2}.
  - Ovaj model vrši implicitnu selekciju osobina putem L1 regularizacije, što doprinosi smanjenju varijanse modela i otpornosti na overfitting.
- **Validacija i performanse**:
  - Tokom evaluacije korišćena je K-Fold unakrsna validacija (`KFold`, $k = 5$, `shuffle=True`) i lokalni test skup dobijen stratifikovanom podjelom podataka.
  - Postignuti rezultat na lokalnom test skupu:

    $$\text{Test RMSE} = \mathbf{2638.25\ EUR}$$

    što je značajno ispod zadatog praga od 6500 EUR.
  - Izabrana vrijednost hiperparametra:

    $$\alpha = \mathbf{0.00107}$$

Ova konfiguracija predstavlja optimalan kompromis između predikcione tačnosti, robusnosti modela i interpretabilnosti, te je odabrana kao finalno rešenje.

Reference

- Pedregosa et al., *Scikit-learn: Machine Learning in Python*, JMLR 2011.
- zvanična dokumentacija: `https://scikit-learn.org/stable/`
- Dataset: Oglasi automobila u Srbiji (projekat FTN, 2025)