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

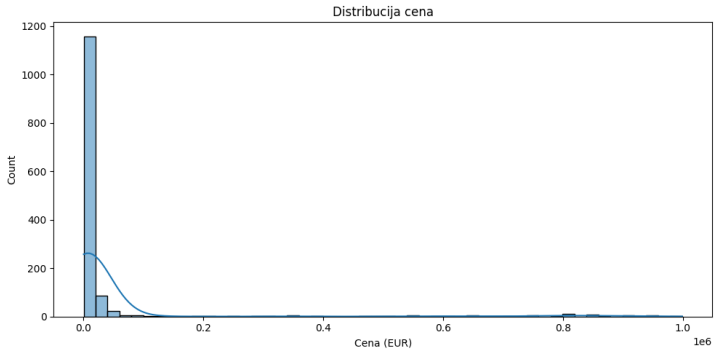
Atribut	**Min**	**(25%)**	**(50%)**	**(75%)**	**Max**	**Srednja**
Cijena [EUR]	1 000	3 000	5 900	13 500	999 000	45 027
Godina proizvodnje	1894	2006	2010	2015	2025	2010
Zapremina motora [cm^3]	163	1 398	1 600	1 984	5 461	1 742.6
Kilometraža [km]	1 000	165 563	209 000	256 000	552 000	206 512
Konjske snage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

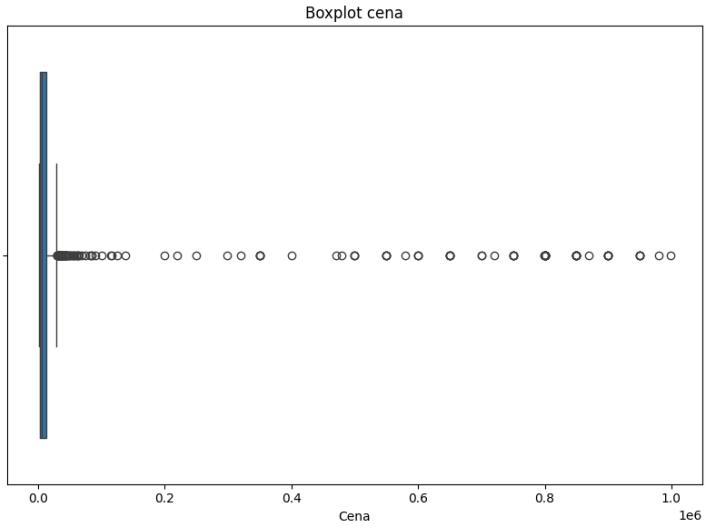
1.2 Vizuelizacije

U cilju grafičke analize podataka, korišćeni su sledeći prikazi:

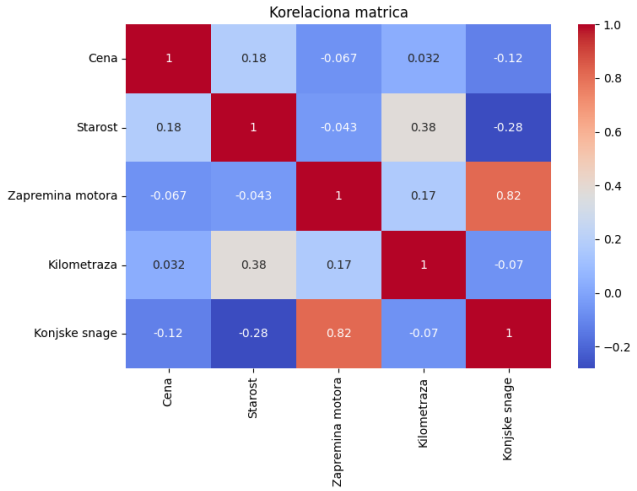
- **Histogram distribucije cena** (Slika 1)
- **Boxplot cena** (Slika 2)
- **Korelaciona matrica numeričkih promenljivih** (Slika 3)
- **Scatterplot: Cena vs. Starost vozila** (Slika 4)



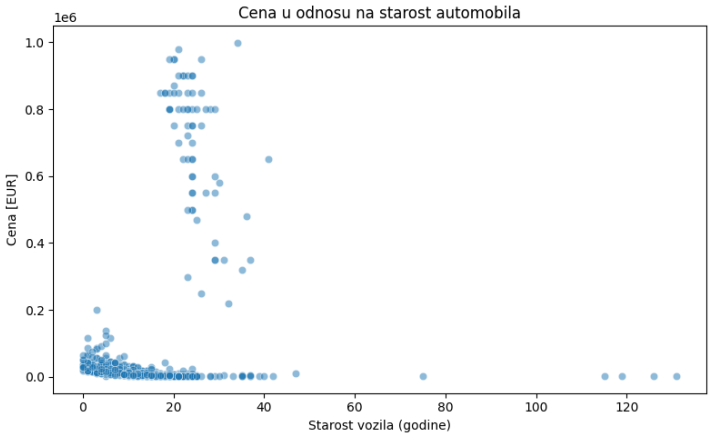
Slika 1: Distribucija cena automobila



Slika 2: Boxplot cena automobila



Slika 3: Korelaciona matrica numeričkih atributa



Slika 4: Raspodela cene u odnosu na starost vozila (2025 - godina proizvodnje)

1.3 Uočeni obrasci i anomalije

Na osnovu vizuelne analize identifikovani su sledeći obrasci:

- **Distribucija cena je snažno asimetrična** (Slika 1) – većina vozila se prodaje ispod 20 000 EUR, dok manji broj primera sa cenama od 100 000+ EUR značajno odskače. To je bio motiv za korišćenje kvantilnog filtra ($q = 0.90$).

- **Prisustvo outlier-a potvrđeno je boxplot-om** (Slika 2) – postoje brojni ekstremi koji narušavaju raspodelu ciljne promenljive i negativno utiču na RMSE.
- **Starost vozila i cena imaju blago pozitivnu korelaciju** ($r = 0.18$), što je neintuitivno. Ova pojava objašnjava se asimetričnom strukturom podataka – (Slika 4) pokazuje da je većina novih vozila luksuzna i značajno skuplja, dok su starija brojna, ali jeftinija. Time dolazi do vizuelne nelinearnosti i "lažne"pozitivne korelacije.
- **Kilometraža ima vrlo slab odnos sa cenom** ($r = 0.03$) – 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 (najčešći: Opel, VW, BMW)
  - Grad: 236 različitih mesta (visoka disperzija)
  - Karoserija i Gorivo: po 8 klasa
  - Menjač: 2 klase (Manuelni i Automatski)

2 Preprocesiranje podataka

Preprocesiranje je uključivalo sledeće korake:

- Kategoričke promenljive kodirane su pomoću One-Hot Encoding-a
- Numeričke promenljive standardizovane su pomoću `StandardScaler`
- Kreirana je nova numerička osobina – starost automobila: `Starost = 2025 - Godina proizvodnje`
- Za numeričke osobine dodati su kvadratni (polinomski) termini kako bi se omogućili nelinearni odnosi
- Target varijabla (`Cena`) je log-transformisana kako bi se smanjio uticaj velikih vrednosti

3 Podela skupa podataka

Podaci su podeljeni korišćenjem K-Fold cross-validation (n = 5 ili 10, u zavisnosti od eksperimenta). Kod eksperimenata sa kvantilnim filterom $q = 0.90$, treniralo se samo na 90% najjeftinijih automobila, dok je cilj bio generalizacija uz niži RMSE.

4 Isprobani algoritmi

Korišćen je `LassoCV` iz `scikit-learn`, zbog sposobnosti automatske selekcije osobina i regularizacije.

4.1 Podešavanje hiperparametara

`LassoCV` automatski bira vrednost regularizacionog parametra α testiranjem više vrednosti (logspace između 10^{-4} i 10^{2}). Najbolja vrednost α pronađena je unutar svakog CV fold-a.

4.2 Rezultati

Na 5-fold cross-validaciji sa $q = 0.90$, postignut je sledeći rezultat:

$$\text{RMSE (log-transformisan target)} = \mathbf{6361.27 \pm 312.45\ EUR}$$

što je ispod zadatog praga od 6500 EUR.

5 Odabrano rešenje

Finalno rešenje koristi sledeće komponente:

- Preprocesiranje: polinomske osobine + standardizacija + one-hot encoding
- Regresor: `LassoCV` sa automatskom selekcijom α
- Transformacija ciljne promenljive: $\log(1 + y)$

Ova konfiguracija pokazala se kao najefikasnija u smislu balansa kompleksnosti i performansi, omogućavajući robusnu generalizaciju uz ispunjavanje uslova zadatka.

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)