Višestruka linearna regresija

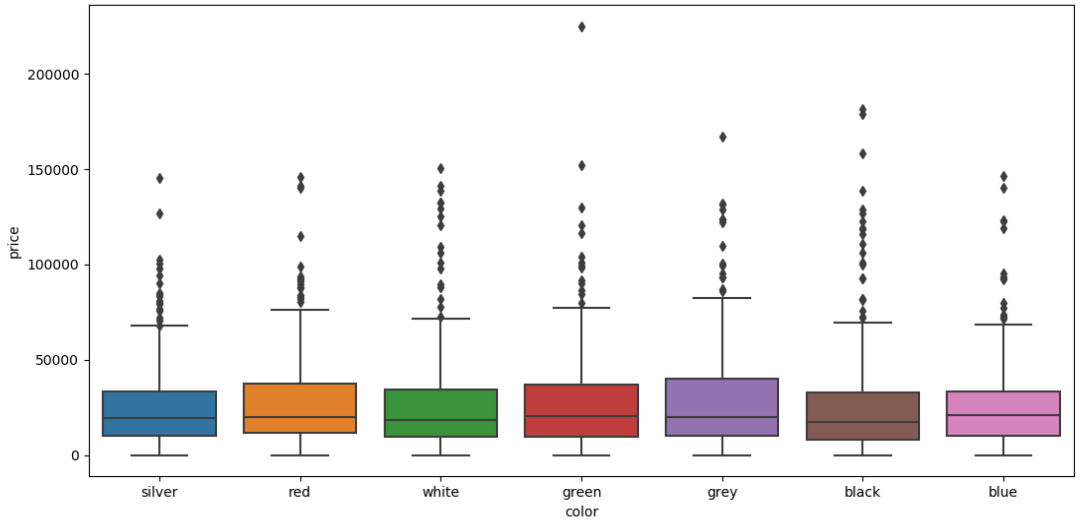
Tim 6_23

- Anastasija Samčović, SW44/2019
- Strahinja Popović, SW51/2019
- Srđan Đurić, SW63/2019

Obrada podataka

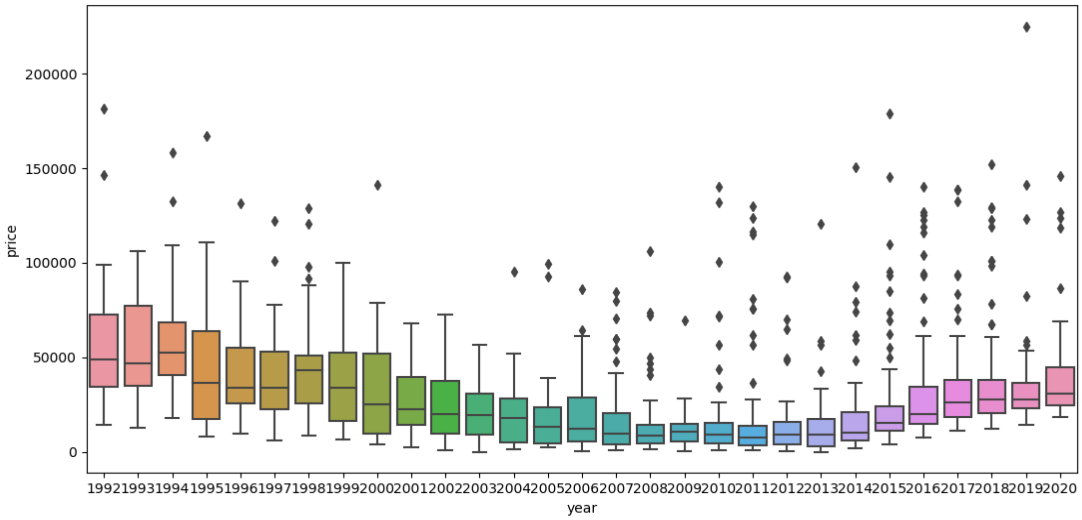
Prvi korak u rešavanju datog problema bila je vizualizacija i analiza datog trening skupa podataka po obeležjima. Koristili smo box plot metod za vizualizaciju odnosa datog obeležja i cene vozila. Nakon analize dobijenih prikaza, utvrdili smo da od postojećih 8 obeležja, 2 obeležja treba zanemariti. Obeležja koja su zanemarena:

- Color(Slika br. 1) – Utvrđeno da je približno ista vrednost cene za svaku boju
- Transmission – Korelirano obeležje sa Fuel obeležjem



Slika br. 1 Box plot prikaz odnosa Color i Price obeležja

Analizom podataka, utvrđeno je da automobili koji su proizvedeni pre 2000-e godine i oni koji su proizvedeni nakon 2016-e godine imaju više cene u odnosu na automobile u godinama između(Slika br. 2). Pretpostavka je da su automobile koji su proizvedeni pre 2000-e godine i imaju visoku cenu zapravo oldtajmeri.



Slika br. 2 Uticaj Year obeležja na Price obeležje

Pokušali smo da kategorička obeležja(Fuel, Make, Category) obradimo pomoću Label Encoding-a, ali pošto su obeležja nominalna(nemaju prirodni poredak) ova tehnika nije davala zadovoljavajuće rezultate. Nakon toga smo pokušali da uklonimo određene opservacije. Definisali smo outlier-e:

- Automobili čija je cena manja od 300
- Automobili čija je cena veća od 140000 i čija je marka Toyota

Dobili smo gore rezultate nakon uklanjanja outlier-a nego pre njihovog uklanjanja, ovaj pristup je zbog toga odbačen. Treći pristup je bio pokušaj kreiranja novog obeležja: oduzeli smo godinu proizvodnje od trenutne godine i dobili starost automobila. Prilikom primene ovog pristupa dobili smo identične rezulate, te je i ovaj pristup zanemaren. Za obradu kategoričkih obeležja smo na kraju odlučili da koristimo **One-Hot Encoding**.

Za preostala obeležja - numerička(Year, Mileage, Engine Size) smo koristili standardizaciju. Primenili smo standardizaciju kako bi neutralisali velike razlike opsega vrednsoti između obeležja.

Testirani algoritmi

Skup podataka korišćen za testiranje predstavlja 30% datog trening skupa, a skup podataka korišćen za treniranje modela predstavlja 70% datog trening skupa.

Neparametarski pristup:

- Primenili smo KNN algoritam za koji je dobijem RMSE veći od 26000
- Neparametarski pristup smo odbacili, zbog činjenice da neparametarski algoritmi ne rade dobro kada je broj obeležja veći od 4

Parametarski pristup:

- Primenili smo Elastic Net, Lasso i Ridge algoritme(rezultati su prikazani u Tabela 1)

Parametarski pristup	**RMSE(70% trening, 30% test)**
Elastic Net	~20650.96
Lasso	**~ 15553.49**
Ridge	~22114.38

Tabela 1 Prikaz dobijenih rezultata

Izabrano rešenje

Naše konačno rešenje je Lasso algoritam sa jednačinom drugog stepena, uz pretprocesiranje podataka One-Hot Encoding-om i standardizacijom. Parametri za koje je naš model pokazao najbolji odnos rezultata i brzine izvršavanja su:

- EPOCHES(Maksimalan broj iteracija) = 9000
- LR(Learning rate) = 0.21
- ALPHA = 0.3