PRIMENJENI ALGORITMI

Priprema podataka za regresioni model

Priprema podataka

- Podaci (data setovi) su često puni nedostataka i njihovo uklanjanje i prilagođavanje je neophodno kako bi se dobio dobar prediktivni model.
- U realnoj situaciji podaci se prikupljaju sa interneta iz različith izvora, unose se ručno ili se mere. Sve to može dovesti do nenamerne greške koja onda može ugroziti kvalitet dobijenog modela.
- Neki podaci se nakon analize mogu pokazati kao beznačajni za model ili nekompletni, te se mogu izbaciti.
- Analizom se može utvrditi i da su neke 2 nezavisne promenljive previše međusobno vezane, do te mere da mogu „zbuniti“ model ako se zajedno koriste.

Priprema podataka – analiza Data frejma

- Bitan korak pre bilo kakve izmene podataka je upoznavanje sa njima. Podaci o kolonama iz Data frame-a se mogu uzeti pozivom funkcije **display(describe(df)).**

Row	variable Symbol	mean Union…	min Any	median Union…	max Any	nunique Union…	nmissing Union…
1	Cena	2564.58	0	1400.0	32000		12
2	Stanje		Nov motor		Polovan motor	2	4
3	Tip		Chopper / Cruiser		Touring	11	
4	Godiste	2002.24	1756	2003.0	2029		
5	Kilometraza	27212.3	0	21000.0	969922		7
6	Kubikaza	436.397	1.0	349.0	1832.0		
7	kW	37.0735	1	25.0	175		
8	KS	50.3888	1	34.0	238		
9	BrojCilindara	2.98439	-1	1.0	3241		
10	Boja		Braon		Zlatna	8	3841
11	Ostecenje		Nije ostecen		Ostecen	2	3

- Ovde se prikazuju statistički podaci o svakoj koloni, gde se vidi minimalna, maksimalna vrednost, srednja vrednost, broj nedostajućih podataka i slično..
- Ukoliko imamo kategoričke promenljive (Tip automobila), možemo izlistati distribuciju po kategorijama sa **display(countmap(df[!, :NazivPromenljive]))**

Priprema podataka– nedostajući podaci

- Prilikom analize data frejma može se ustanoviti da neki podaci nedostaju delimično ili potpuno.
- Ukoliko u nekoj koloni nedostaje više od 60% podataka, može se razmisliti o potpunom uklanjanju te kolone. **select!(df, Not(:NazivKolone))**
- Ukoliko su podaci u nekoj koloni ključni za model, možemo izbaciti sve redove iz frejma gde ti podaci nedostaju. **dropmissing!(df, [:Cena])**

Priprema podataka – zamena nedostajućih

- Ako imamo prazna polja u kolonama sa kategoričkim podacima (primer: Boja: bela, crna, siva), možemo ih zameniti vrednošću iz te kolone koja se najčešće pojavljuje funkcijom **mode()**.

- Funkcija **mode()** će vratiti podatak iz kolone koju prosledimo koji se u njoj najčešće pojavljuje. Primer: Ako imamo podatke o 50 automobila, a 30 je sive boje onda će mode za kolonu boja biti **siva**.

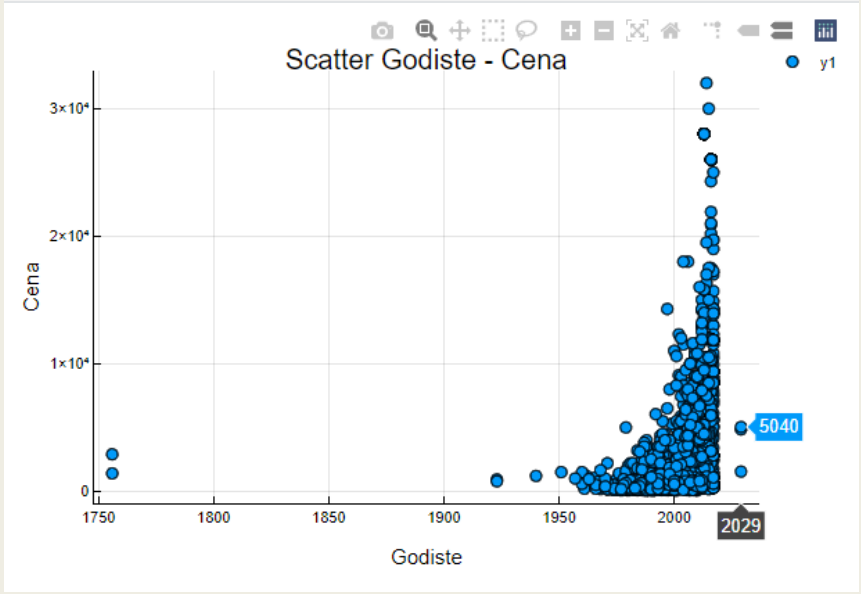
- Ako imamo prazna numerička polja i želimo ih popuniti umesto izbaciti to možemo uraditi tako što ćemo tu umetnuti srednju vrednost iz cele kolone.

- Naravno pored ovakvih, primitivnih metoda, nedostajuće vrednosti se mogu umetnuti koristeći i različite kompleksne prediktivne metode.

Priprema podataka – podaci koji štrče

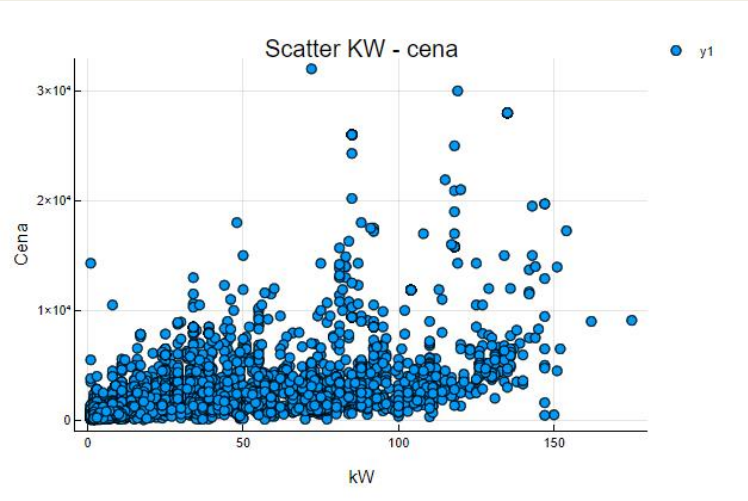
- Pored podataka koji nedostaju, mnoge probleme nam mogu napraviti i podaci koji su nelogični ili previše odskaču u odnosu na ostale koje imamo.
- Ako podatak previše odskače od odstalih, možemo posumnjati da je rezultat nevalidnog merenja, unosa ili je previše ekstreman.
- Kako bismo lakše uočili podatke koji štrče ili su nelogični možemo nacrtati grafik.

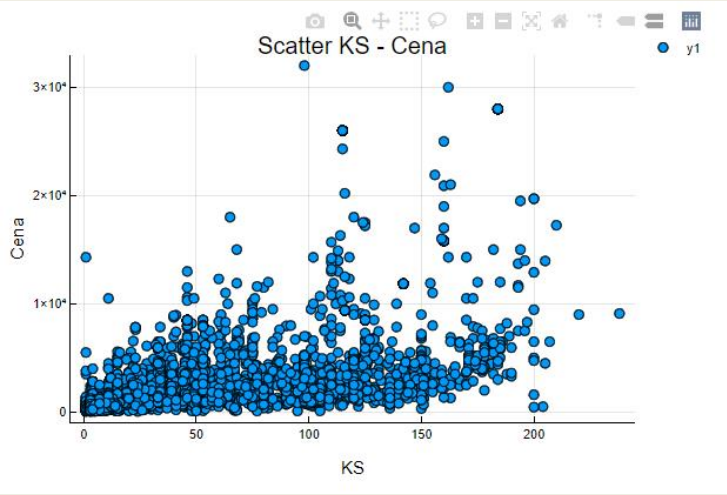
- Na ovom grafiku vidimo uticaj godišta na cenu automobila, takođe vidimo da su neki automobili iz 2029. i 1750. godine što nije realno.



Priprema podataka - korelacija i multikolinearnost

- Koristiti podatke iz neke kolone kao nezavisnu promenljivu pri predviđanju ima smisla ako su podaci u korelaciji sa zavisnom promenljivom (onom koju procenjujemo).
- Pored određivanja Perasonovog koeficijenta, to možemo utvrditi i crtajući grafik zavisnosti između njih.
- Ako međutim primetimo da postoji jaka korelacija između neke 2 nezavisne promenljive, onda jednu moramo izbaciti kako bismo održali tačnost modela.





K-fold kros validacija

- Da bi se izbegao slučaj da nije dobra podela na skupove za obuku i testiranje, uvodi se K-fold kros validacija
- Ona podrazumeva podelu skupa podataka na K delova i testiranje svakog dela u odnosu na ostatak. Nakon toga može da se traži prosečna greška za K testiranja
- Na ovaj način dobija se pouzdanija mera o kvalitetu testiranja, tj. o njegovoj grešci
- Primer: Ako imamo skup A sa 100 podataka i K = 5, tada se skup A deli na 5 skupova A1, A2, A3, A4, A5 sa po 20 elemenata i vrši se 5 testiranja, pri čemu se jedan skup uzima za test skup, a unija ostala 4 za skup za obuku

Primer 1: K-fold kros validacija

- Primer 1: Neka je dat skup tačaka u ravni. Napisati skript kojim se određuje zavisnost vrednosti y koordinate u zavisnosti od x koordinate tačke. Pri tome:
  - *Testiranja izvršiti K-fold kros validacijom za k = 5*
  - *Pronaći prosečnu absolutnu grešku za svako od 5 testiranja*
  - *Pronaći prosečnu absolutnu grešku za tih 5 testiranja*

Primer 2: Procena cene motora automobila

- Primer 2: U datoteci **„motor-cena.csv“** nalaze se podaci o cenama motora koji su skinuti sa sajta polovniautomobili.com. Koristeći tehnike pripreme podataka i linearnu regresiju napraviti model za procenu **cene motora.**

  -Ukoliko u nekoj koloni nedostaje manje od 40% podataka, zameniti ih odgovarajućim tehnikama zamene.

  -Proceniti kvalitet modela na osnovu greške.

Zadatak 1: Predviđanje odobravanja kredita

- Zadatak 1: U datoteci **„trening-krediti.csv“** nalaze se podaci o klijentima banke i da li im je odobren kredit. Koristeći tehnike pripreme podataka i logističke regresije istrenirati model koji predviđa da li će nekom klijentu biti odobren kredit. Model testirati podacima iz „**test-krediti.csv“**.

  -Ukoliko u nekoj koloni nedostaje manje od 40% podataka, zameniti ih odgovarajućim tehnikama zamene.

  -Proceniti kvalitet modela.