Višestruka linearna regresija

- ## Tim sigme

Daniel Božanić SW-63/2018

Laslo Sabadi Baranji SW-51/2018

- ## Zadatak

Potrebno je predvideti platu nastavnog osoblja u SAD na osnovu atributa **zvanje**, **oblast**, koliko **godina je doktor**, koliko **godina ima iskustva** i **pol**, koristeći višestruku regresiju. Dobijeni RMSE mora da bude ispod 28500.

- ## Pristup problemu

Testiranje smo uradili nad trening podacima, odnosno trening skup smo podelili na 70% trening podatke i 30% test podatke. Odlučili smo da izbacimo atribut pol jer smatramo da ovaj atribut ne utiče na platu. Koristili smo **one-hot encoding** nad svim kategoričkim podacima, jer ovim pristupom model će tretirati sve podatke jednako i neće pretpostaviti da postoji neki poredak. Kod svakog algoritma smo normalizovali podatke koristeći **min-max** pristup. Kod neparametarskih modela koristili smo **Euclidean distance** za računanje udaljenosti. Takođe smo koristili **modified z-score** za uklanjanje outlier-a, ali na kraju smo odlučili da nećemo izbacivati outlier-e jer dovodi do gorih krajnjih rezultata.

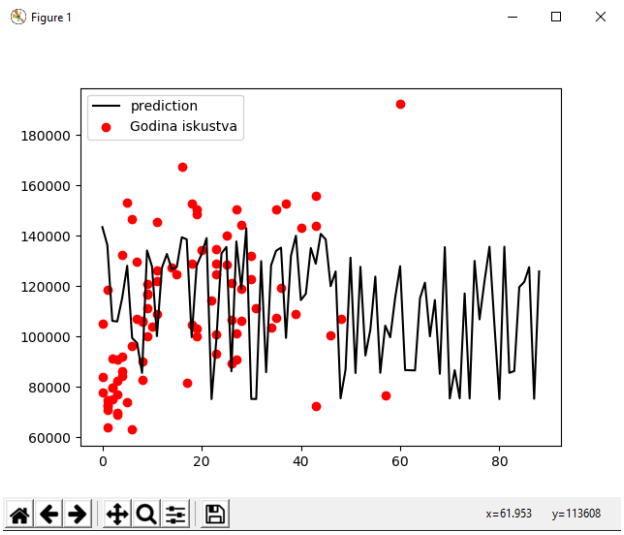
- ## Isprobani algoritmi i ostvareni rezultati

Algoritam	Parametri	RMSE nad trening podacima (70-30)
Lasso (L1) coordinate descent	L1 penalty = 0.5 Iteracije = 1000	20438.442522845628
Ridge (L2) gradient descent	L2 penalty = 0.01 Learning rate = 0.001 Iteracije = 1000	20345.626743572528
Ridge (L2) normal equation	L2 penalty = 0.5	20098.71677840043
Elastic net	L1 i L2 penalty = 0.5 Learning rate = 0.1 Iteracije = 1000	19986.77585391628
KNN	K = 35	24730.57503091165
Kernel regression (Gaussian)	Lambda = 0.1	18203.10071182054

U gornjoj tabeli se mogu videti svi isprobani algoritmi, njihovi parametri i dobijani RMSE. Parametre smo odabrali višestrukim pokretanjem i probavanjem.

- ## Konačno odabrano rešenje

Kao naše konačno rešenje smo odabrali Gausovu kernel regresiju, jer nam je taj pristup doneo najbolji rezultat.



Na X osi se nalazi godina iskustva, a na Y osi plata