Višestruka Regresija

Tim:

- Marko Vukotić SW-71/2018
- Miloš Popović SW-24/2018

Zadatak:

Napraviti predikciju plata nastavničkog osoblja u SAD na osnovu više ulaza. Potrebno je naći regresioni model koji se najbolje uklapa u skup podataka.

Analiza podataka:

Postoje atributi uz pomoć kojih se prediktuje plata:

- **Zvanje:** Nastavno zvanje: Profesor, Vanredni Profesor i Docent
- **Oblast:** Teorijska i Primenjena
- **Broj godina:** Koliko godina je prošlo od dobijanja doktorata nastavnik
- **Iskustvo:** Broj godina radnog staža
- **Pol:** Muško/Ženski

Takođe kolone **zvanje, oblast i pol** imaju kategoričke podatke.

Ovo smo rešili **One Hot Encoding**-om. Od N različitih parametara uzimali smo N-1 za model osim za *male i female* (Iz nekog razloga bolje rezultate smo dobijali).

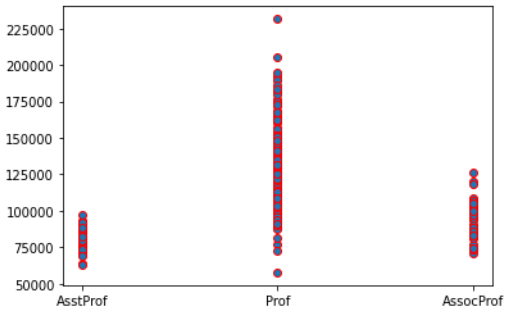
```
pol_one_hot = pandas.Series(list(train_data.pol))
oblast_one_hot = pandas.Series(list(train_data.oblast))
zvanje_one_hot = pandas.Series(list(train_data.zvanje))

pol_one_hot_data = pandas.get_dummies(pol_one_hot)
oblast_one_hot_data = pandas.get_dummies(oblast_one_hot)
zvanje_one_hot_data = pandas.get_dummies(zvanje_one_hot)

train_data["Female"] = pol_one_hot_data.Female
train_data["Male"] = pol_one_hot_data.Male


train_data["oblast_a"] = oblast_one_hot_data.A
train_data["oblast_b"] = oblast_one_hot_data.B


train_data["assoc_prof"] = zvanje_one_hot_data.AssocProf
train_data["asst_prof"] = zvanje_one_hot_data.AsstProf
train_data["prof"] = zvanje_one_hot_data.Prof
```

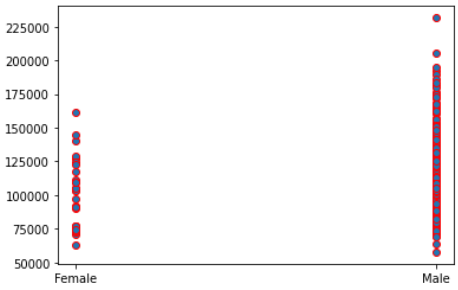
Female	Male	oblast_a	oblast_b	assoc_prof	asst_prof	prof
1.0	0.0	1.0	0.0	0.0	1.0	0.0
0.0	1.0	0.0	1.0	0.0	0.0	1.0
0.0	1.0	0.0	1.0	0.0	0.0	1.0
0.0	1.0	0.0	1.0	0.0	0.0	1.0
0.0	1.0	1.0	0.0	0.0	0.0	1.0
...	...	...	...	...	...	...

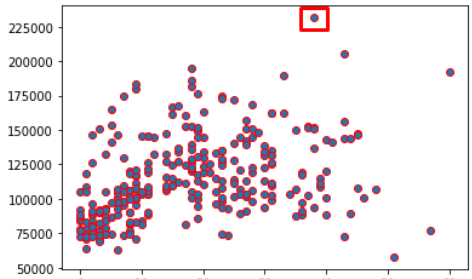
Kolone nakon One Hot

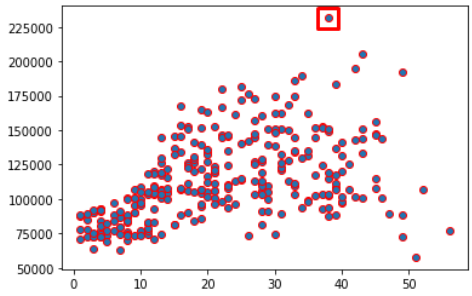
Na ovim slikama postoje *Outlieri* (tačke visokog uticaja). Da bi smo poboljšali ponašanje naseg sistema uklonili smo na trening skupu tačke visokog uticaja. Metoda je napisana da uklanja sve tačke koje se nalaze iznad 200,000.

Tokom analiziranja *slika ispod* videli smo neke tačke koje dosta odskaču od proseka za isti broj godina radnog staža i odlučili smo da ih uklonimo metodom `remove_outliers`.







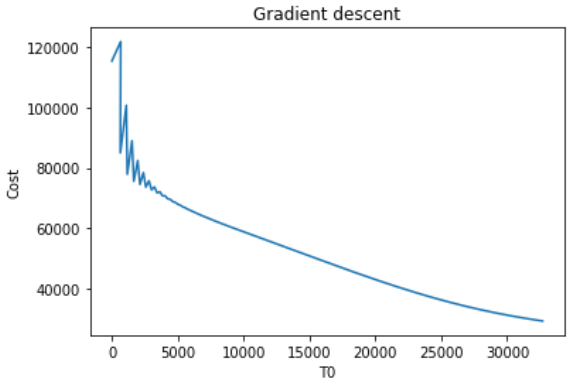


Metoda

Metoda sa kojom smo radili je *Gradijentni Spust* (Gradient Descent).

Gradijenti: Urađen za svaki od parametara.

Inicijalni parametri za *Gradient* su dobijeni empirijskim pokretanjem sa različitim learning rate-om.



Primećujemo da je na početku prevelik learning rate međutim kasnije nam je potreban veći learning rate pa smo ručno pokretali sa različitim learning rate-ovima i predefinisanim parametrima.

Funkcija:

_t_musko_0 * x_musko +
_t_zensko_0 * x_zensko +
_t_god_iskustva_0 +
_t_god_iskustva_1 * x_god_iskustva +
_t_god_iskustva_2 * x_god_iskustva ^ 2 +
_t_godina_doktor_0 +
_t_godina_doktor_1 * x_godina_doktor +
_t_godina_doktor_1 * x_godina_doktor ^ 2 +
_t_asst_prof_0 * x_asst_prof +
_t_assoc_prof_0 * x_assoc_prof +
_t_oblast_a_0 * x_oblast_a

"_t" – Parametri
"_x" – Ulazne Vrednosti