Višestruka linearna regresija

Tim:

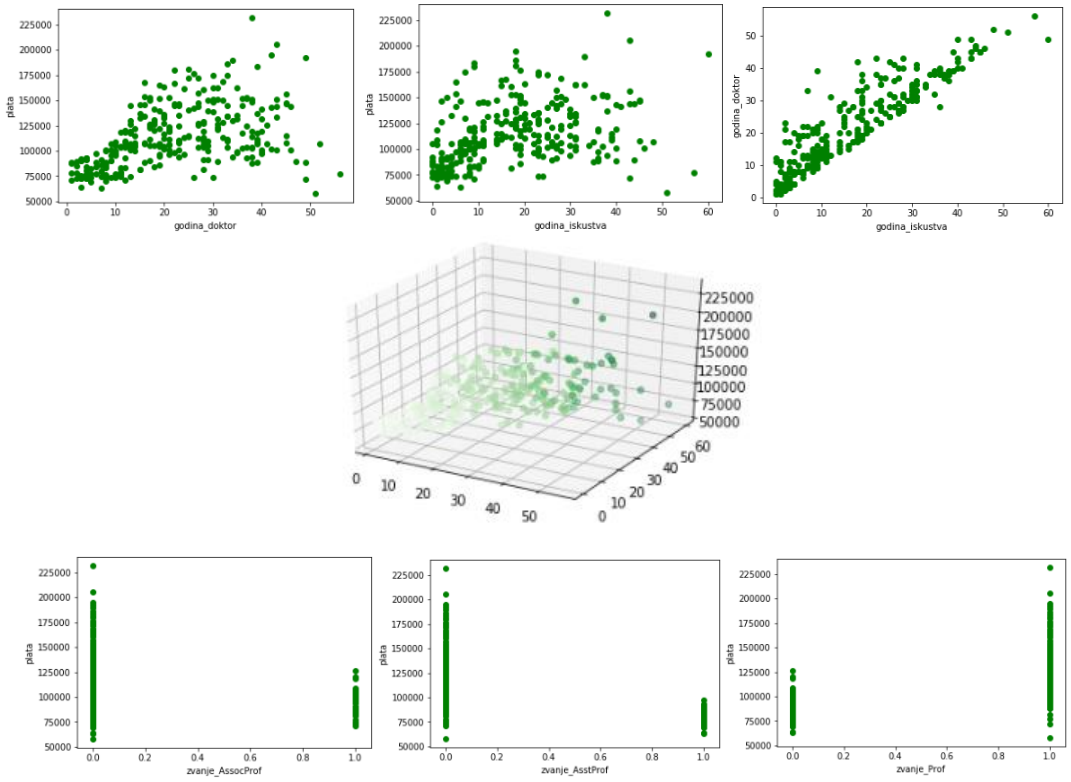
- Mihajlo Kušljić
- Nikola Zubić
- Filip Mladenović

Zadatak:

Prediktovati platu nastavnog osoblja u SAD na osnovu više atributa (zvanje, oblast, godina_doktor, godina_iskustva, pol).

Analiza podataka:

Analizom 2D i 3D grafika, došli smo do zaključka da postoje određeni *outlayeri*.



Korišćena je metoda *one-hot encoding* za kodovanje kategoričkih obeležja - zvanje i oblast; pol je izbačen iz razmatranja zato što smo smatrali da to obležje ne bi trebalo da utiče na visinu nečije plate.

Normalizacija je primenjena na numerička obeležja (godina_doktor i godina_iskustva).

Primenjene metode:

- KNN regression: Za predikciju plate radnika koristi se prosek vrednosti plate za k najsličnijih radnika u trening skupu podataka. Za izbor mere sličnosti korišćen je validacioni skup podataka. Praćena je ostvarena *RMSE* greška za euklidsko rastojanje, Menhetn rastojanje, kosinusnu sličnost i Pirsonovu korelaciju. Kao najbolja metrika pokazalo se Menhetn rastojanje. Za parametar k izabrana je vrednost 10 pomoću unakrsne validacije. Ostvarena *RMSE* greška nad *test_preview* skupom podataka iznosi **22379.09506**.

- Single kernel regression: Po učitavanju podataka izvršen je *one-hot encoding* za kategorička obeležja. Za predikciju plate radnika korišćena je kernel regresija, uzimajući u obzir sva obeležja. Zbog relativno malog broja primera u trening skupu korišćen je Gausov kernel koji svakom primeru daje neku pozitivnu težinu, kako bi se izbegle situacije kada su težine svih primera jednake nuli, jer ne postoje dovoljno bliski primeri u test skupu. Faktor lambda (propusni opseg kernela) određen je primenom unakrsne validacije i iznosi 0,01. Izbor metrike za rastojanje vršen je poređenjem dobijenih *RMSE* grešaka nad validacionim skupom podataka za različite metrike. Od isprobanih metrika (euklidsko rastojanje, Menhetn rastojanje, kosinusna sličnost, Pirsonova korelacija) najbolje rezultate dalo je Menhetn rastojanje. Nad *test_preview* skupom podataka sa kernel regresijom pomoću Gausovog kernela i Menhetn rastojanja ostvarena je *RMSE* greška od **19445.78583** - Ovaj pristup pokazao se kao najbolji za kernel regresiju.

- Multiple kernel regressions: Drugi pristup je bio uprosečavanje predikcija kernela koji rade nad pojedinačnim obeležjima radnika. Za svako obeležje korišćen je Gausov kernel, ponovo da bi se izbegle situacije da prilikom formiranja predikcije nema dovoljno sličnih primera u obučavajućem skupu. Za kategorička obeležja korišćeno je Hemingovo rastojanje, a za nekategorička korišćeno je Menhetn rastojanje. Konačna procena dobijena je uprosečavanjem procena plate za svako obeležje. Uz ovaj pristup nad *test_preview* skupom podataka ostvarena je *RMSE* greška od **22316.22732**.

- **Linearne kombinacije predikcija kernela**: Još jedan pristup predikciji bio je formiranje linearne kombinacije predikcija kernela koji rade nad pojedinačnim obeležjima. Kernel za svako obeležje implementiran je na isti način kao u prethodnom pristupu. Za određivanje njihovog značaja u konačnoj predikciji korišćen je algoritam gradijentnog spusta. Međutim, po završetku treniranja težine su uniformno raspoređene po svim kernelima (procene za svako obeležje imaju isti značaj/težinu), tako da se ovaj pristup ne razlikuje mnogo od uprosečavanja. Ostvarena *RMSE* greška nad *test_preview* skupom podataka iznosi **22316.23002**.

Izabrana metoda:

Korišćen je pristup gde je *Elastic Net* regresija implementirana sa više hiperparametarskih vrednosti po uzoru na rad: *Regularization Paths for Generalized Linear Models via Coordinate Descent* (Rob Tibshirani *et al.*).

Koristio se pristup višestruke regresije uz *Elastic Net* regularizaciju što znači da su u metodi predikcije sve težine pomnožene sa ulaznim X vrednostima.

Kao izlaz dobija se Y koji predstavlja predikciju.

Formalno, *Elastic Net* regularizacija u kontekstu višestruke regresije rešava sledeći problem:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2+\lambda P_\alpha(\beta)\right]$$

gde je β_0 intercept, a β skup ostalih koeficijenata/težina modela koje je cilj minimizovati. rho_factor (λ) koji iznosi 0.1 je korišćen kao kompromis između *Ridge-regression* i *Lasso-regression penalty*-ja, te na osnovu njega se računaju λ_{lasso} i λ_{ridge}.

Na kraju, penalizacija za elastic-net je:

$$\begin{aligned} P_\alpha(\beta) &= (1-\alpha)\frac{1}{2}||\beta||_{\ell_2}^2+\alpha||\beta||_{\ell_1} \\ &= \sum_{j=1}^{p}\left[\tfrac{1}{2}(1-\alpha)\beta_j^2+\alpha|\beta_j|\right] \end{aligned}$$

Dobijeno rešenje nad test_preview je: **15035.59835**.