ML18-Z2-code10: Izveštaj

Pristup zadatku

Zadatak je urađen u *Python*-u, uz biblioteke *Pandas*, *NumPy* i *Matplotlib*.
Podatke smo najpre vizualizovali u 2D i 3D u odnosu na platu. Analizirajući grafike pomislili smo da postoje tačke visokog uticaja. Međutim, izbacivanjem istih greška je rasla pa smo zaključili da verovatno nisu u pitanju tačke visokog uticaja ako bi se posmatralo u više dimenzija.
Modeli su konfigurisani korišćenjem *cross-validation* (k=5) pristupa.

Isprobani algoritmi

Isprobani su sledeći algoritmi:
Enkodovanje kategoričkih podataka u numeričke:

1. *Label encoding*
2. *One hot encoding*

Normalizacija:

1. *z-score*
2. *min-max*

Algoritmi za regresiju:

1. Višestruka linearna regresija (za različite stepene polinoma)
   Regularizacija:
   1. *Ridge regression*
   2. *Ridge regression* + *Approximal lasso gradient descent* (*Elastic net*)
2. *KNN* (za različite vrednosti k)
3. *Weighted* KNN
4. *Kernel regressio*n (*Gaussian* i *Epanechnikov* kernel)

Konačno rešenje

Od *KNN*, *Weighted KNN* i *Kernel regression* najmoćniji je *Kernel regression*, ali je ipak najbolje rezultate dala linearna regresija stepena 2 sa *Ridge* i *Lasso* regularizacijom.
Odabir algoritama za normalizaciju i enkodovanje kategoričkih podataka nije imao značajnog uticaja.

Ostvareni rezultati

Konačna greška na test skupu iznosi 23527.0038605394.