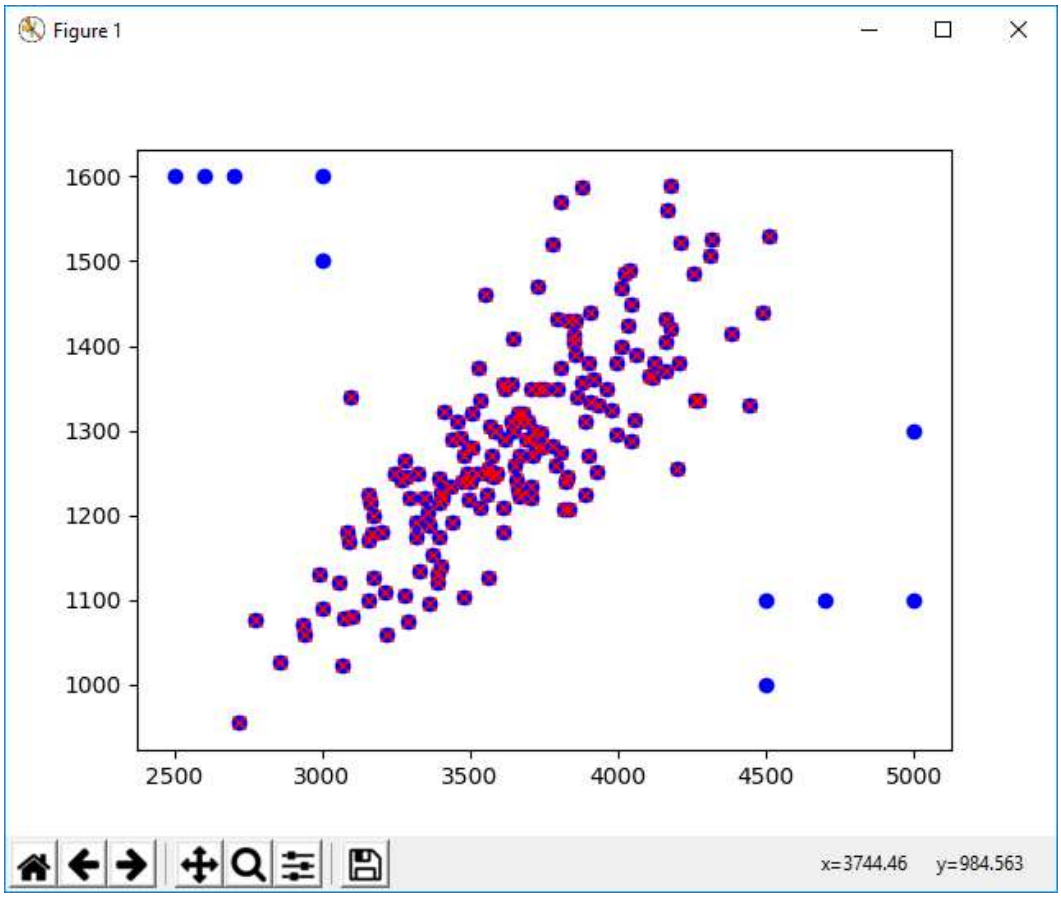
ML18-Z1-code10: Izveštaj

Pristup zadatku

Zadatak je urađen u Python-u, uz biblioteke Pandas, NumPy i Matplotlib.
Najpre smo plotovali zadate podatke. Odmah smo primetili prisustvo tačaka visokog uticaja i zaključili da bi bilo dobro uzeti ih u obzir. S obzirom da tih tačaka ima dosta i da su grupisane, pomislili smo da nisu tu greškom i da ih možda treba posebno obraditi (npr. korišćenjem više modela ili polinoma). Ipak, s obzirom da smo ograničeni na jednostruku linearnu regresiju, odlučili smo se da ih izbacimo iz trening skupa podataka kako ne bi loše uticali na model.
Podaci su pre obučavanja modela normalizovani. Prilikom primene modela podaci se takođe normalizuju, ali se pre računanja greške denormalizuju kako bi greška bila interpretabilna.

Isprobani algoritmi

Tačkama visokog uticaja smatrali smo tačke čiji odnos x i y vrednosti previše odstupa od prosečnog odnosa. Izbacivanjem tih tačaka dobijen je skup podataka sa slike 1 (crvene tačke).



Slika 1. Skup podataka sa izbačenim tačkama visokog uticaja

Korišćena je Z-score normalizacija.
Linearna regresija implementirana je na dva načina:

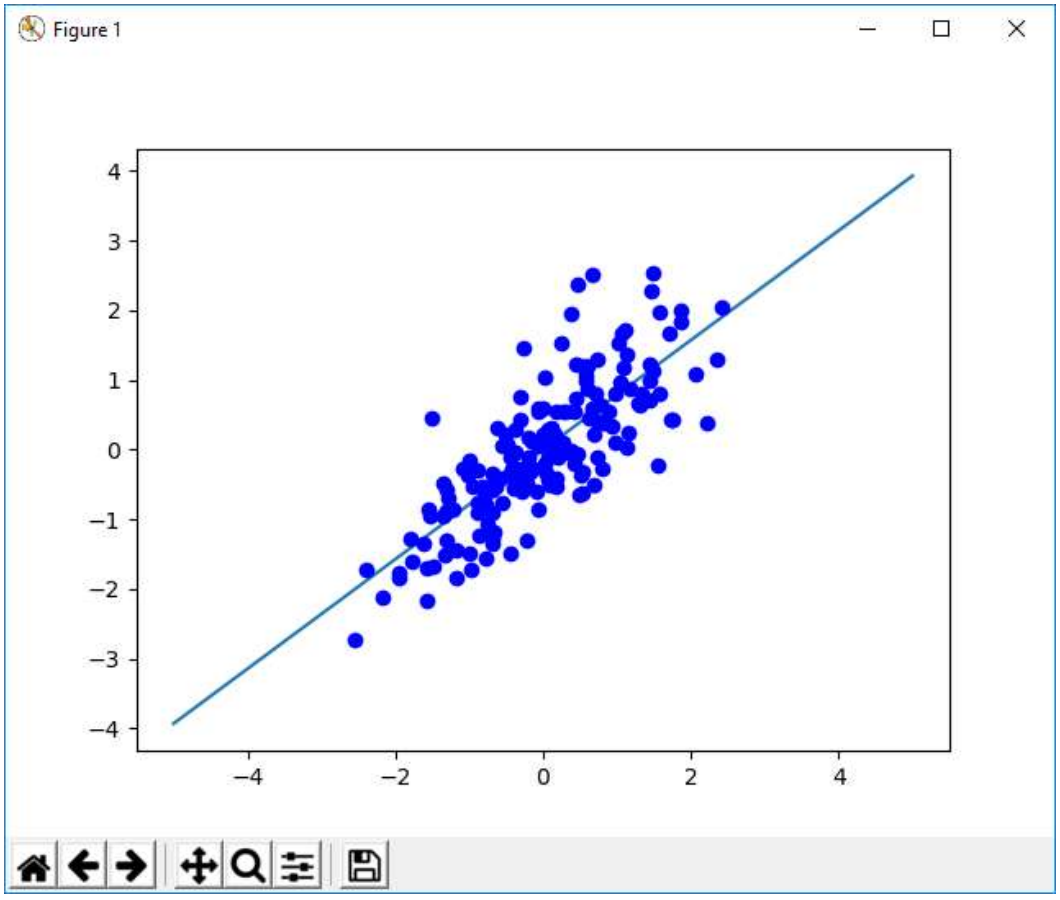
1. Batch gradient descent
2. Normal equation (closed form)

Konačno rešenje

S obzirom da closed form daje egzaktno rešenje, nema hiperparametre i trening skup je mali, odlučili smo se za tu implementaciju.

Ostvareni rezultati

Konačna greška na test skupu iznosi 68.9076227770313, a model je prikazan na slici 2.



Slika 2. Model linearne regresije