ML18-Z4-code10: Izveštaj

Pristup zadatku

Zadatak je urađen u Python-u uz biblioteke Pandas i Scikit-Learn.

Najpre smo obavili istraživanje podataka (*feature engineering*) i zaključili sledeće:

1. Atribut ***dead*** (da li je učesnik preživeo udes ili ne) je suvšan pošto je ta informacija već sadržana u atributu ***injSeverity*** (stepen povreda).
2. Atributi ***airbag*** (da li je učesnik imao airbag) i ***deploy*** (da li se airbag aktivirao) su suvišni pošto atribut ***abcat*** sadrži njihovu kombinaciju.
3. Atribut ***weight*** (procenjena masa učesnika) sadrži mnogo nepreciznih vrednosti (~4k uzoraka sa masom preko 400kg).
4. Kategoričke vrednosti atributa ***injSeverity*** nisu potpuno logično sortirane (npr. *5 - nepoznato* nije ni na početku ni na kraju).
5. Ostali atributi ne sadrže previše nepreciznih vrednosti.

Na osnovu toga odlučili smo da izacimo atribute *dead*, *airbag* i *deploy*, što nije uticalo na grešku a smanjilo je dimenzionalnost problema. Atribut *weight* nismo izbacili pošto se pokazao kao relevantan, a korišćeni modeli deluju otpornim na tačke koje mnogo odstupaju.

Podaci su pre obučavanja modela normalizovani korišćenjem *min-max* normalizacije. Kategoričke vrednosti su zamenjene brojčanim korišćenjem *label encoding* pristupa. Vrednosti atributa *injSeverity* su namapirane na sledeći redosled:

1. Nepoznato
2. Bez povreda
3. Lakše telesne povrede
4. Teže telesne povrede bez invaliditeta
5. Teže telesne povrede sa invaliditetom
6. Teže telesne povrede sa smrtnim ishodom
7. Smrt

Tačnost modela utvrđivali smo korišćenjem unakrsne validacije (*RepeatedStratifiedKFold)*, a parametri su optimizovani korišćenjem grid pretrage (*GridSearchCV*).

Isprobani algoritmi

Kako bi stekli osećaj koje algoritme vredi obučavati, najpre smo isprobali sve metode ansambla za klasifikaciju uz minimalno podešavanja parametara. Najbolje su se pokazali upravo *state of the art* algoritmi *Random Forest Classifier* (*RFC*) i *Gradient Boosting Classifier* (*GBC*).

Za njih smo nastavili sa intenzivnim grid pretragama i dobili sledeće vrednosti parametara:

parametar	Vrednost za *RFC*	Vrednost za *GBC*
max_depth	9	6
max_features	3	9
min_samples_leaf	1e-5	1e-9
min_samples_split	1e-5	1e-9
n_estimators	100	275
criterion (samo za *RFC*)	entropy	/
subsample (samo za *GBC*)	/	0.9

S obzirom da su oba algoritma davala solidne rezultate odlučili smo da ih ukombinujemo u *Voting Classifier*, čije težine smo optimizovali korišćenjem unakrsne validacije.

Imali smo nameru da u *Voting Classifier* dodamo i *KNN* i *SVM*, međutim njihove performanse nisu bile dovoljno dobre da bi oni doprineli ansamblu.

Konačno rešenje

Voting Classifier sa *Random Forest Classifier* i *Gradient Boosting Classifier* sa težinama 3.5 i 6.5 dao je najbolje rezultate.

Ostvareni rezultati

Konačan *micro F1 score* na test skupu iznosi 0.565934065934066.