Model ansambla

Tim - mašinski_učenjaci

Nikola Zubić (SW-03/2016), Filip Mladenović (SW-11/2016), Mihajlo Kušljić (SW-53/2016)

Zadatak

Dostupan je deo policijskih izveštaja o saobraćajnim nesrećama u SAD u periodu 1997 - 2002. Na osnovu dostupnih podataka izvršiti procenu brzine vozila u trenutku sudara (kolona ***speed***). Kriterijum prihvatljivosti: ***Micro F1*** **> 0.51**

Analiza podataka

Opis skupa podataka:

- ***speed*** – brzina vozila u trenutku sudara (kolona koju je potrebno prediktovati):
  - 1 – 9 km/h
  - 10 – 24
  - 25 – 39
  - 40 – 54
  - 55+
- ***weight*** – procenjena masa učesnika udesa
- ***dead*** – da li je učesnik preživeo udes:
  - *alive* – preživeo
  - *dead* – nije preživeo
- ***airbag*** – da li je učesnik imao *airbag*:
  - *none* – ne
  - *airbag* – da
- ***seatbelt*** – da li je učesnik bio vezan:
  - *none* – ne
  - *belted* – da
- ***frontal*** – da li je u pitanju bio čeoni sudar:
  - 0 – ne
  - 1 – da
- ***sex*** – pol učesnika:
  - *f* – ženski
  - *m* – muški
- ***ageOFocc*** – starost učesnika
- ***yearacc*** – godina kada se dogodila nesreća
- ***yearVeh*** – godina proizvodnje vozila
- ***abcat*** – da li se aktivirao *airbag*:

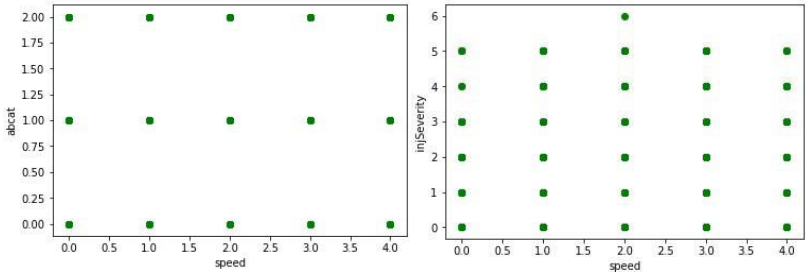
    - *unavail* – vozilo nije imalo *airbag* za tog učesnika
    - *nodeploy* – *airbag* se nije aktivirao
    - *deploy* – *airbag* se aktivirao
- ***occRole*** – tip učesnika:
    - *driver* – vozač
    - *pass* – suvozač
- ***deploy*** – da li se *airbag* aktivirao:
    - 0 – *airbag* nije dostupan za tog učesnika ili se nije aktivirao
    - 1 – *airbag* se aktivirao
- ***injSeverity*** – stepen povreda učesnika:
    - 0 – bez povreda
    - 1 – lakše telesne povrede
    - 2 – teže telesne povrede, bez invaliditeta
    - 3 – teže telesne povrede, sa invaliditetom
    - 4 – smrt
    - 5 – nepoznato
    - 6 – teške telesne povrede sa smrtnim ishodom (smrt nastupila kasnije)

Iz navedenog opisa skupa podataka može se uvideti da postoji snažna korelacija između pojedinih obeležja.
Informacija o ishodu sudara se može saznati iz ***injSeverity*** obeležja, pa se zbog toga ne mora koristiti obeležje ***dead***. ***Deploy*** i ***airbag*** su sadržani u ***abcat*** pa se i ta dva obeležja mogu izbaciti iz razmatranja. Time smo smanjili dimenzionalnost našeg problema.

Problem kategoričkih obeležja je rešen *label encodingom.* Najbolje je objasniti na primeru ***speed*** obeležja – postoji gradacija tako da se klasa sa najmanjom brzinom može obeležiti sa 0, dok ova sa najvećom sa 4.

Poređenjem dobijenih predikcija unakrsnom validacijom ustanovljeno je da je redove sa nedostajućim vrednostima, iz trening skupa, najbolje ukloniti.

Traženje *outliera* se završilo neuspešno, budući da su primeri prilično jednako raspoređeni po klasama:



Takođe, postoji slabija korelacija među obeležijima. Npr: dešava se da mlađi čovek pri manjoj brzini pogine, dok bi stariji čovek pri većoj brzini preživeo i to ne bi bili izolovani slučajevi.

Odrađena je standardizacija podataka.

Odabrani model

Korišćena je kombinacija *Ada Boost Classifier*-a, *Gradient Boosting Classifier*-a, *Decision Tree Classifier*-a i *Gaussian Naive Bayes*-a. Njihovi hiper parametri su određivani uz pomoć *GridSearchCV*-a koji se postarao da svaki pojedinačni klasifikator radi najbolje što može. Nakon toga pomenuti modeli su iskombinovani u *Soft Voting Classifier*-u čije su težine za uvažavanje klasifikacija pojedinačnih modela, takođe, podešavane uz pomoć *GridSearchCV*-a.

Rešenje je validirano unakrsnom validacijom uz pomoć *Repeated Stratified K Fold*-a gde je skup deljen na pet delova i celokupna validacija rađena u dva navrata.

Odabranom modelu su, naknadno, ručno podešavani hiper parametri. Takav model je dao najbolji rezultat koji predstavljen mikro f1 merom iznosi **0.565167390748786**.