UNIVERZITET U SARAJEVU

ELEKTROTEHNIČKI FAKULTET SARAJEVO

DOMAĆA ZADAĆA 1

Zadatak 2

MAŠINSKO UČENJE

Odsjek: Računarstvo i Informatika
Datum: 18.11.2019

Studenti:

- **Mašović Haris, 1689/17993**
- **Muminović Amir, 1661/17744**

Zadatak 2 (2.5 bodova)

Prethodno odabrani najbolji model evaluirajte na setu podataka “attrition_test.csv”, koristeći paket caret i funkciju confusionMatrix(). Dokumentujte rezultate evaluacije.

Grupa koja ostvari najbolji rezultat na setu “attrition_test.csv” će dobiti maksimalni broj bodova za ovaj zadatak, ostale grupe će biti rangirane u odnosu na najbolju grupu.

Pri tome, da biste ostvarili bodove za zadatak 2, vaš model treba da ima minimalne performanse:
- Pos. Pred. Value >~ 0.65
- Balanced Accuracy >~ 0.6
- Kappa statistics >~ 0.25
- Lower bound of (donja granica) 95% CI >~ 0.80

Ukoliko model ne zadovoljava zahtijevane performanse, možete ponoviti proces iz zadatka 1. Dokumentujete promjene koje su dovele do poboljšanja performansi modela.

NAPOMENE:

- U ovom pdf-u će biti prikazan izabrani model i njegovi rezulati.
- Radi nedostatka vremena za kucanje i umjesto objašnjavanja iterativno po raznim unaprijeđenjima kada su se dobijala biti će navedeni svi načini koji su korišteni u prvom zadatku za dobijanje što boljeg modela (jer je u zadatak1 pdfu finalno rješenje) respektivno kroz historiju tj. čiščenje/analiza podataka koje je dovelo do perforamsi rezultata naših modela.

```
pred.train.dt <- predict(model_dt_c,test_im,type = "raw")
confusionMatrix(table(test_im$Attrition,pred.train.dt))
```

U nastavku su prikazani rezultati za naš izabrani model (Bootstrapping - Gbm):

```
> confusionMatrix(table(test_im$Attrition,pred.train.dt))
Confusion Matrix and Statistics

     pred.train.dt
      No Yes
  No  245   3
  Yes  29  17

               Accuracy : 0.8912
                 95% CI : (0.8498, 0.9244)
    No Information Rate : 0.932
    P-Value [Acc > NIR] : 0.9966

                  Kappa : 0.4644

 Mcnemar's Test P-Value : 9.897e-06

            Sensitivity : 0.8942
            Specificity : 0.8500
         Pos Pred Value : 0.9879
         Neg Pred Value : 0.3696
             Prevalence : 0.9320
         Detection Rate : 0.8333
   Detection Prevalence : 0.8435
      Balanced Accuracy : 0.8721

       'Positive' Class : No
```

U tabeli su prikazani rezultati izabranog modela:

	Pos pred value	Balanced Accuracy	Kappa statistics	Lower Bound CI
Model	0.9879	0.8721	0.4644	0.8498

Sve vrste tehnika/stvari koje smo koristili da bi dobili što bolje rezultate u našim modelima:

- Krenulo se nad golim podacima i popunjavajući nedostajuće vrijednosti sa MFV (Most Frequent Values) i popunjavanje kontinualnih varijabli sa Median vrijednosti
- Shavili smo da BirthDate ne možemo ostaviti kao char odnosno da moramo pretvoriti ili u kategoričku ili u numeric (izabrana je numeric)
- Nakon toga se shvatilo da postoji nekoliko varijabli koje možemo odmah otpisati i njih smo ukinuli u narednoj iteraciji (vidjeti Zadatak 1)
- U narednom koraku smo stavili aspekt na varbijale koje imaju NA vrijednosti i njih ispitali posebno tj. dodatno i njihove odnose i korelacije, ovaj put nismo odmah fillovali podatke (da bi izbjegli peglanciju, uočeno na početku, fillovanje je došlo na kraju), i na osnovu njihovih korelacija zaključili da možemo otpisati varijablu JobRole, dok ostale smo pokupili jer odgovaraju za izradu ml modela. Također u ovoj fazi smo zaključili da varijabla PercentSalaryHike ima veliki broj NA vrijednosti i da je u korelaciji sa Attrition varijablom, odnosno da vrijednosti sa kojima trebamo popuniti trebaju biti što preciznije (sa Median vrijednosti to nije davalo nekog efekta) pa smo, istraživanjem, odlučili da našu varijabli pretvorimo u kontinualnu i iskoristimo linearnu regresiju za fillovanje podataka (što je dalo bolje rezultate)
- U narednom koraku, smo odradili korelacije svih varijabli (heat maps) i ustanovili korelacije svih varijabli i koje varijable treba koristiti u odnosu na koje ne (vidjeti Zadatak 1), ključan aspekt u ovom koraku je bio skontati način kako dobiti odnos tj. koeficijent korelacija, što smo daljim istraživanjem našli da se može koristiti chi-square metoda za ispitivanje koeficijenta kategoričkih varijabli i na koji način izabrati koje varijable uzeti, a koje ne.
  Cilj je bio da nadjemo varijable koje su pogodne za ml model i da ukinemo druge varijable koje nisu potrebne (već su u korelaciji sa nekim drugim varijablama), na kraju se svelo na 19 varijabli, te ovo nam je, ujedno uzelo i najviše vremena i omogućilo da dobijemo veliki napredak u rezultatima modela.
- Nakon izbora prvog modela i testiranja nad testnim setom podataka, shvatili smo da se može još malo *tweakati* model tj. povećanje ili smanjenje po broju iteracija za koje on trenira jer je u pitanju Gbm, shodno tome u trainControl varijabli shvatili smo da najbolji izbor staviti sljedeće

  ```
  train_control <- trainControl(method="boot", number=120)
  ```

  te da nam daje najbolje rezultate koji su prikazani u gornjoj tabeli.

Shodno ovim postupkom smo završili zadatak 2 i ujedno prikazali naš najbolji model nad test_im setom podataka odnosno početni, samo bolji sa aspekta podataka, attrition-test set podataka.