Osnove statističkog programiranja

Ak. god. 2023./2024.

Neki naslov

Dokumentacija

Julijana Kolarec, Lucija Topolko

siječanj 2024., Zagreb

Nastavnik: *prof.dr.sc. Damir Pintar*

Sadržaj

1. Prediktivni modeli primjenom strojnog učenja

Kako bismo bolje razumjeli što film čini uspješnim ili neuspješnim, provele smo analizu dobivenog skupa podataka primjenom strojnog učenja. Cilj nam je bio razviti model koji može čim točnije predviđati uspjeh filma na temelju njegovih karakteristika.

1.1 Priprema podataka

Iz dobivenih podataka izbacile smo retke kojima su nedostajali neki podaci. Takvih je redaka bilo 1261. Također, uklonile smo stupce koji su sadržavali jedinstvene ili skoro jedinstvene vrijednosti (*movie_title*, *movie_imdb_link*, *plot_keywords*, *genres*). Još smo izbacile tekstualne stupce koji su bili prekorelirani s nekim numeričkim stupcem. Na primjer, *actor_1_name* je prekoreliran s *actor_1_facebook_likes*.

```
columns <- c('duration', 'director_facebook_likes', 'actor_1_facebook_
    likes','actor_2_facebook_likes', 'actor_3_facebook_likes', 'num_user_
    for_reviews', 'num_critic_for_reviews', 'num_voted_users', 'cast_
    total_facebook_likes', 'movie_facebook_likes', 'facenumber_in_poster'
    , 'color', 'title_year', 'language', 'country', 'content_rating', '
    aspect_ratio', 'gross', 'budget', 'imdb_score')

mldata <- data[,columns]

mldata <- na.omit(mldata)
```

Preostale nenumeričke stupce pretvorile smo u tip integer.

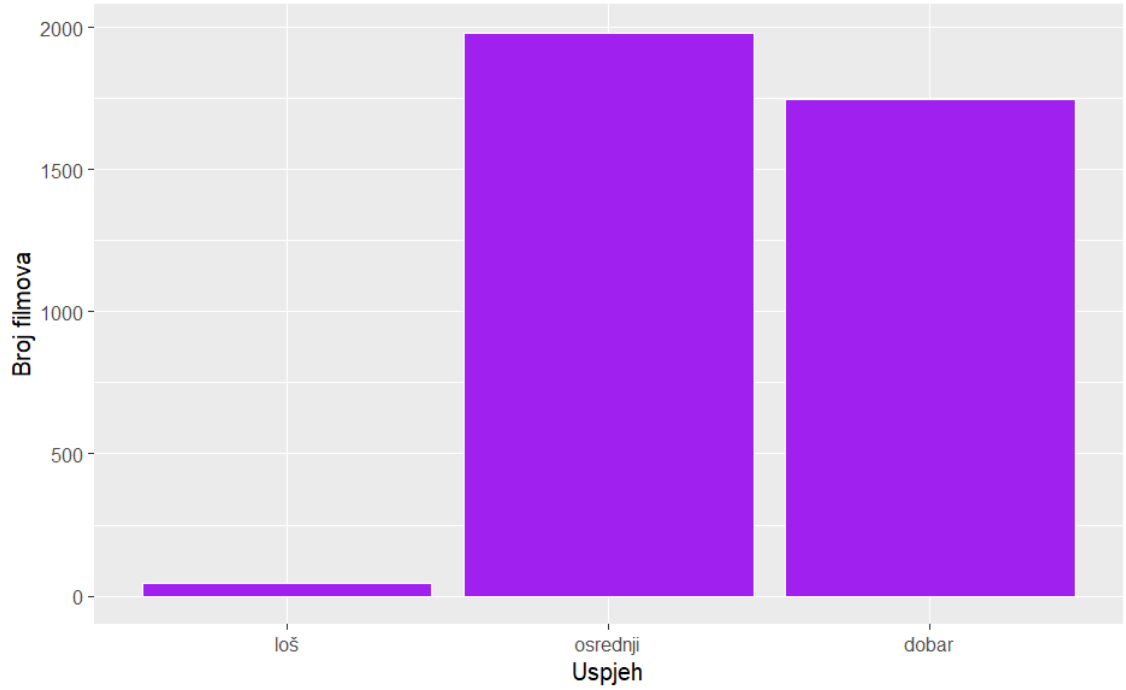
```
label_encode <- function(column) {
  as.integer(factor(column, levels = unique(column)))
}
```

Značajka koju predviđamo je *imdb_score*. To broj zaokružen na jednu decimalu, pa smo za bolje rezultate uspjeh filma podijelile u tri skupine: loš, osrednji i dobar,

a stupac imdb_score smo zbog prekoreliranosti uklonile.

```
mldata$score <- ifelse(mldata$imdb_score < 3.33, "los", ifelse(mldata$
    imdb_score < 6.66, "osrednji", "dobar"))
```

Graf 1.1 prikazuje omjer broja filmova po uspjehu. Filmova koji su ocijenjeni kao loši znatno je manje od ostalih. Točnije, loših je filmova 43, osrednjih 1981, a dobrih 1746.



Slika 1.1: Podjela filmova po uspjehu

1.2 Modeli s nebalansiranim skupom podataka

Kako bismo razvile model za predviđanje uspješnosti, skup podataka podijelile smo u omjeru 80:20. Na temelju 80% gradile smo model, a na 20% ga testirale. Prvi model koji smo razvile koristi metodu potpornih vektora.

```
Model <- train(score ~ ., data = training_set,
               method = "svmPoly",
               na.action = na.omit,
               preProcess=c("scale","center"),
               trControl= trainControl(method="none"),
               tuneGrid = data.frame(degree=1,scale=1,C=1)
)
```

Model radi s uspješnošću od 72.8%.

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš        0        0     0
  osrednji   8      339   141
  dobar      0       54   205

Overall Statistics

               Accuracy : 0.7282
                 95% CI : (0.6948, 0.7599)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4518

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity             0.00000          0.8626       0.5925
Specificity             1.00000          0.5791       0.8653
Pos Pred Value              NaN          0.6947       0.7915
Neg Pred Value          0.98929          0.7915       0.7111
Prevalence              0.01071          0.5261       0.4632
Detection Rate          0.00000          0.4538       0.2744
Detection Prevalence    0.00000          0.6533       0.3467
Balanced Accuracy       0.50000          0.7208       0.7289
```

Slika 1.2: Metoda potpornih vektora - rezultati

Drugi model koji smo razvile koristi metodu slučajne šume. Rezultati su nešto bolji, uspješnost je 78.3%.

```
Model_rf <- randomForest(score ~ ., data = training_set, ntree = 500,
    importance = TRUE)
```

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš        0        0     0
  osrednji   7      325    86
  dobar      1       68   260

Overall Statistics

               Accuracy : 0.7831
                 95% CI : (0.7518, 0.8122)
    No Information Rate : 0.5261
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5677

 Mcnemar's Test P-Value : 0.0177

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity             0.00000          0.8270       0.7514
Specificity             1.00000          0.7373       0.8279
Pos Pred Value              NaN          0.7775       0.7903
Neg Pred Value          0.98929          0.7933       0.7943
Prevalence              0.01071          0.5261       0.4632
Detection Rate          0.00000          0.4351       0.3481
Detection Prevalence    0.00000          0.5596       0.4404
Balanced Accuracy       0.50000          0.7821       0.7897
```

Slika 1.3: Metoda slučajne šume - rezultati

Iako su ovi rezultati na prvi pogled donekle zadovoljavajući, nijedan od ovih modela nije predvidio da će ijedan film biti loš. To je očekivani rezultat jer podaci nisu nimalo balansirani - loših filmova je znatno manje pa ih je i puno teže predvidjeti. Produkciji filma bilo bi najkorisnije imati model koji može predvidjeti neuspjeh filma, a ovi modeli to ne uspijevaju pa smo ih odbacile.

1.3 Modeli s balansiranim skupom podataka

S ciljem poboljšanja točnosti predviđanja loših filmova, podatke smo balansirale. Nastojale smo broj loših i dobrih filmova približiti broju osrednjih filmova (Graf 1.4).

```
oversample <- ovun.sample(score~., data = over, method = "both", N =
    3932)$data
```

Slika 1.4: Podjela filmova po uspjehu - balansirani podaci

Nad novim smo podacima ponovo testirale naše modele. Ovaj je put metoda potpornih vektora postigla uspješnost od 75.6% (Slika 1.5), a metoda slučajne šume visokih 90.9% (Slika 1.6).

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš      374       69    34
  osrednji   7      256   109
  dobar      0       68   261

Overall Statistics

               Accuracy : 0.7564
                 95% CI : (0.7308, 0.7806)
    No Information Rate : 0.343
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6352

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity              0.9816          0.6514       0.6460
Specificity              0.8708          0.8522       0.9121
Pos Pred Value           0.7841          0.6882       0.7933
Neg Pred Value           0.9900          0.8300       0.8316
Prevalence               0.3234          0.3336       0.3430
Detection Rate           0.3175          0.2173       0.2216
Detection Prevalence     0.4049          0.3158       0.2793
Balanced Accuracy        0.9262          0.7518       0.7791
```

Slika 1.5: Metoda potpornih vektora - rezultati s balansiranim podacima

```
Confusion Matrix and Statistics

          Reference
Prediction loš osrednji dobar
  loš      381        0     0
  osrednji   0      330    44
  dobar      0       63   360

Overall Statistics

               Accuracy : 0.9092
                 95% CI : (0.8913, 0.925)
    No Information Rate : 0.343
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8637

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: loš Class: osrednji Class: dobar
Sensitivity              1.0000          0.8397       0.8911
Specificity              1.0000          0.9439       0.9186
Pos Pred Value           1.0000          0.8824       0.8511
Neg Pred Value           1.0000          0.9216       0.9417
Prevalence               0.3234          0.3336       0.3430
Detection Rate           0.3234          0.2801       0.3056
Detection Prevalence     0.3234          0.3175       0.3591
Balanced Accuracy        1.0000          0.8918       0.9048
```

Slika 1.6: Metoda slučajne šume - rezultati s balansiranim podacima

Ovim smo rezultatima zadovoljne jer oba modela s visokom točnošću predviđaju loše filmove.

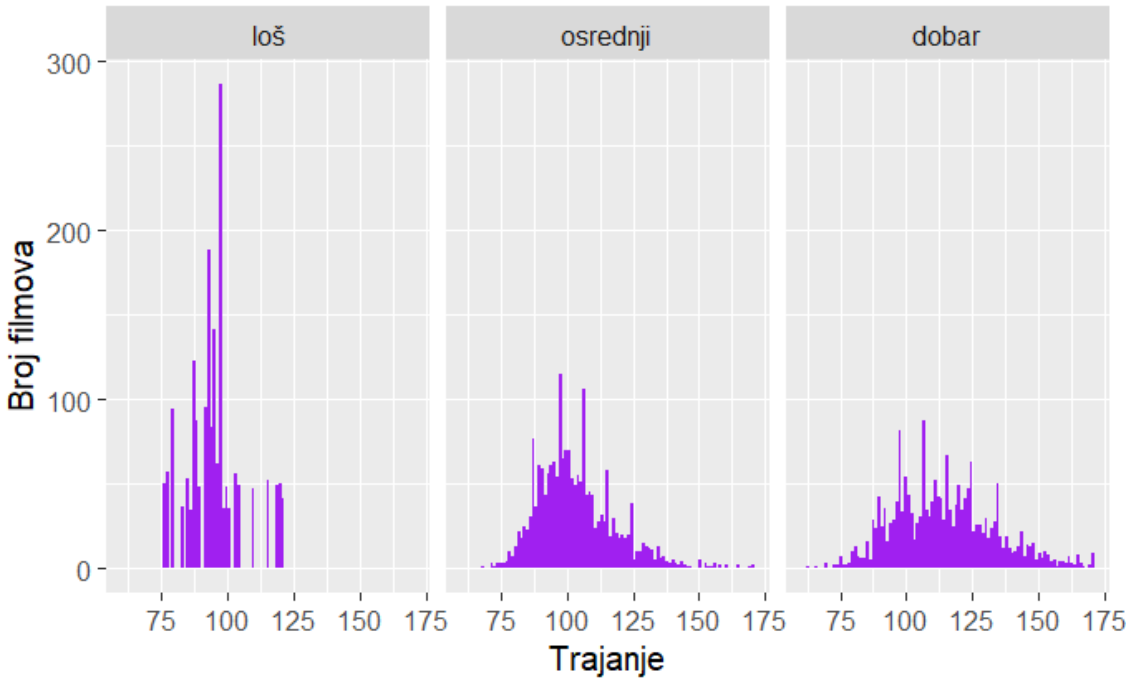
1.4 Atributi najznačajniji za predviđanje

Idući je korak u analizi bio otkriti koji atributi najviše koriste pri predviđanju uspješnosti filmova, posebice onih loših.

Analizu smo provele nad modelom koji koristi metodu slučajne šume i balansirani skup podataka jer upravo taj model daje najbolje rezultate.

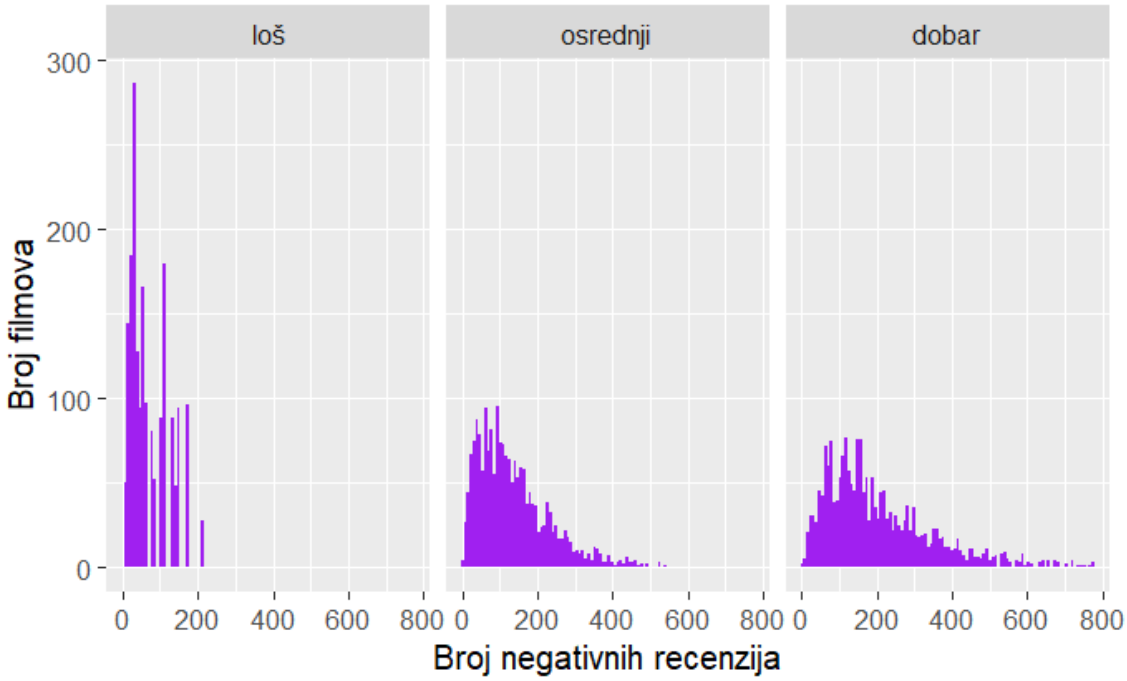
Atributi koji su u našem modelu u najvećoj korelaciji s uspješnosti su trajanje i broj negativnih recenzija.

Filmovi koji su dobili loše ocjene gledatelja najčešće traju između sat i dva sata, a većina ih traje do 100 minuta. Grafovi za ostale filmove također prikazuju da najviše filmova traje 100 ili više minuta.



Slika 1.7: Metoda slučajne šume - rezultati s balansiranim podacima

Za predviđanje uspješnih i neuspješnih filmova bio je važan i broj negativnih recenzija. Zanimljivo, filmovi koje smo klasificirale kao neuspješne imali su manji broj negativnih recenzija. Razlog tome je vjerojatno taj što se velik broj ljudi odlučio uopće ne pogledati film kad je vidio da je većina recenzija negativna. Uspješnije filmove pogleda puno više ljudi različitih mišljenja pa je očekivano da se nekima neće svidjeti.



Slika 1.8: Metoda slučajne šume - rezultati s balansiranim podacima

Atributi koji su najmanje korelirani s uspjehom filma su jezik i broj ljudi na plakatu.

Indeks slika i dijagrama