Predviđanje popularnosti žanrova video igara

(uz implementaciju odabranih klasifikacionih algoritama)

Žarko Blagojević RA 44/2018

Asistent: Stefan Anđelić

Motivacija

Video igre svojom popularnošću i rasprostranjenošću uzimaju sve više vremena (i prostora, ako ste ljubitelj fizičkih kopija) u svakodnevici ljudi. Ogromna popularnost koju uživa ovaj vid zabave sa sobom donosi i ogromne zarade kompanijama koje proizvode popularne video igre. S tim u vidu, kompanijama su značajna istraživanja tržišta video igara, kako bi se bolje upoznali s tim šta karakteriše igre koje su trenutno najpopularnije.

Karakteristike video igre u su velikoj meri određene žanrom kojem igra pripada. Popularnost žanra može u mnogome doprineti većoj prodaji same igre. Od velike važnosti za predviđanje uspešnosti video igre jeste upravo da li pripada žanru koji ljudi trenutno vole da igraju.

S ovim u vidu, odluke koje kompanije donose o tome kakve igre će praviti u budućnosti mogu zavisiti upravo od žanra koji će u budućnosti biti popularan, te je predviđanje popularnosti žanrova video igara od velikog značaja za ove kompanije.

Ideja

Skup podataka o prodajama video igara korišćen u projektu sadrži podatke o prodajama pojedinačnih video igara na raznim svetskim tržištima (SAD, EU, Japan...) kao i podatke o žanru kojem igra pripada, izdavaču igre i konzoli za koju je igra izdata. Popularnost video igre karakterišu podaci o globalnim prodajama, gde je najpopularnija igra ona koja ima najveće globalne prodaje. Skup podataka preuzet sa sledeće Kaggle stranice.

Pošto su podaci orijentisani ka pojedinačnim video igrama, potrebno ih je pretvoriti u podatke orijentisane na žanrove video igara, i to izdeljene po godinama, kako bi se mogle vršiti predikcije. Neophodno je uvideti koji podaci su od važnosti za predviđanje, kao i koji se dodatni preditkori mogu kreirati ne bi li predviđanje bilo bolje.

Nakon transformacije sledi obučavanje klasifikacionih algoritama sa ciljem poređenja njihove sposobnosti da predvide koji žanr će biti najpopularniji za datu godinu.

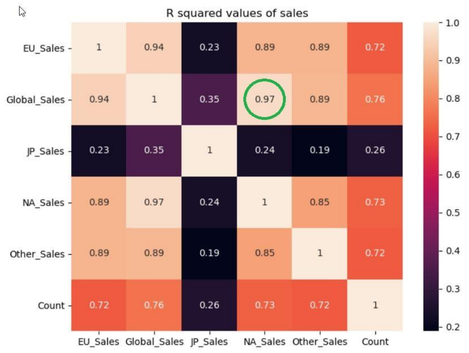
Metodologija

Nakon reorganizovanja skupa podataka da odgovara žanrovima i godinama, za **svaku godinu** stvoreni su sledeći podaci:

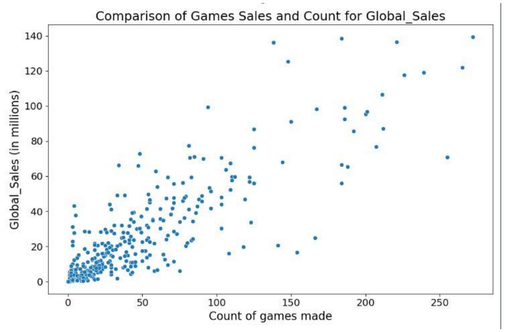
- Broj prodaja po žanru u SAD, EU, Japanu, ostatku sveta i u globalu
- Broj proizvedenih igara po žanru
- Najpopularniji žanr u prethodnih N godina – N = 3
- Najpopularniji žanr pre N godina – N = {1, 2, 3}
- Najpopularniji žanr za datu godine – onaj čije su globalne prodaje najveće

Odabir prediktora, kao i kreiranje novih prediktora važan je i težak posao. Postoje brojni načini da se podaci iz skupa podataka i njihova međusobna relacija metrički ispitaju.

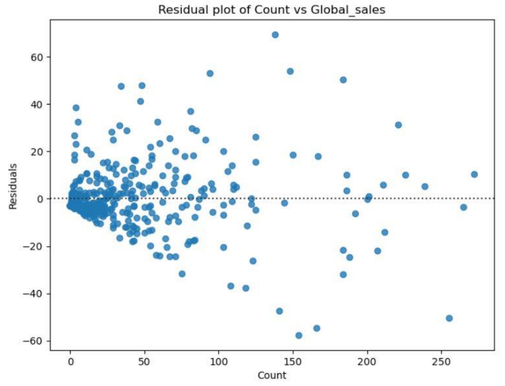
Nakon analize toplotne karte (eng. heatmap) R^2 skora između prodaja utvrđeno je da varijacije u globalnim prodajama najbolje objašnjavaju varijacije u prodajama u SAD (97%). Samo ona će od podataka o prodajama biti izabrana kao prediktor.



Bitno je uvideti da li broj igara proizvedenih u nekom žanru za datu godinu korelira globalnim prodajama tog žanra.



Na grafiku iznad postoji izvesna korelacija između ove dve vrednosti, ali daleko je od odlične. O tome svedoči i grafik reziduala ispod, gde je umesto nasumične raštrkanosti tačaka oko centra grafika, prisutna veća gustina tačaka oko 0 x ose.



Pristupi

Za klasifikaciju su korišćeni sledeći algoritmi:

- Logistic Regression – OneVsRest (modul scikit learn)
- Logistic Regression – OneVsRest (implementacija u projektu)
- Gaussian Naive Bayes – GaussianNB (modul scikit learn)
- Gaussian Naive Bayes – NaiveBayes (implementacija u projektu)
- Support Vector Machine - (modul scikit learn)
- Random Forrest Classifier - (modul scikit learn)

Rezultati

Ogroman skup podataka za pojedinačne igre se transformacijom u skup podataka orijentisan na podatke o žanrovima i indeksiran godinama, sveo na svega 35 semplova (za 35 godina), reprezentativnost podataka i rezultata je diskutabilna.

Pored standardnog pristupa, gde se skup podataka deli na trening (~70%) i test podatke (30%), izvršena je unakrsna validacija izostavljanjem jednog sempla (Leave-one-out Cross Validation - LOOCV) koja je česta za skupove podataka sa malim brojem semplova.

Classification algorithm	Accuracy (LOOCV)	Accuracy (Standard)
Logistic Regression - sklearn	0.62	0.75
Logistic Regression - mine	0.64	0.75
GaussianNB- sklearn	0.59	0.67
NaiveBayes - mine	0.78	0.67
Support Vector Machine - sklearn	0.59	0.83
Random Forrest Classifier - sklearn	0.59	0.75

Zaključak

Iako je početni skup podataka bio izuzetno bogat za pojedinačne igre, za problem koji je ovaj projekat pokušao da reši podaci jednostavno nisu bili dovoljni.

Da je početni skup podataka sadržao tačne datume kada su igre izdate, a ne samo godine izdavanja tada bi se podaci mogli finije vremenski podeliti (ne samo po godinama, već i po kvartalima ili drugoj vremenskoj odrednici). Samim tim obučavanja bi bila načinjena nad većim brojem podataka, a predviđanja bi bila orijentisana skorijoj budućnosti.