Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.9
DOI: https://doi.org/10.24867/30BE07Milijevic

PREDIKCIJA DOBITNIKA FILMSKE NAGRADE OSKAR UPOTREBOM MAŠINSKOG UČENJA

USING MACHINE LEARNING TO PREDICT OSCAR WINNERS

Jelena Milijević, *Fakultet tehničkih nauka, Novi Sad*

Oblast – RAČUNARSTVO I AUTOMATIKA

Kratak sadržaj – *У раду је вршена предикција добитника награде Оскар за најбољи филм и за глумачко остварење. Изтраживање ове теме проистиче из њеног значаја за filmsku industriju. Korišćeni algoritmi su: Logistička Regresija, SVM, Random Forest, Bagging, XGBoost і Неуронска Мрежа. За евалуацију модела коришћена је 10-унакрсна валидација. У првој предикцији најбоље се показао Random Forest са тачношћу 91,59%, а у другој XGBoost са 84,39%.*

Ključne reči: *Mašinsko učenje, Logistička regresija, SVM, Random Forest, Bagging, XGBoost, Neuronska mreža*

Abstract – *The study focused on predicting the winners of the Oscar award for Best Picture and for acting achievements. This research is motivated by its significance to the film industry. The algorithms used include: Logistic Regression, SVM, Random Forest, Bagging, XGBoost, and Neural Networks. Model evaluation was conducted using 10-fold cross- validation. In the first prediction, Random Forest demonstrated the best performance with an accuracy of 91.59%, while in the second prediction, XGBoost achieved an accuracy of 84.39%.*

Keywords: *Machine Learning, Logistic Regression, SVM, Random Forest, Bagging, XGBoost, Neural Network*

1. UVOD

U dinamičnom svetu filmske industrije, Oskar nagrada predstavlja vrhunac priznanja za izuzetna dostignuća u kinematografiji. Efikasno predviđanje pobednika Oskara može imati značajan uticaj na filmsku industriju, unapređujući strategije produkcije, marketinga i distribucije, te zadovoljavajući očekivanja publike i stvaralaca.

Ovaj rad istražuje mogućnost primene metodologije mašinskog učenja u predviđanju pobednika Oskara u kategoriji za najbolji film, kao i za glumačko ostvarenje. Glavni izazov je kompleksnost podataka, koji obuhvataju atribute poput žanra, kritika publike, režije, i prethodnih nagrada. Raznovrsnost podataka zahteva pažljivu analizu i odabir relevantnih atributa za uspešno modelovanje.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Kovačević, red.prof.

U ovom radu je rađeno prikupljanje podataka, kao i eksplorativna analiza tih podataka. Analizirajući obimne podatke o filmovima, uključujući žanr, režiju, kao i prethodne nominacije i osvojene nagrade, primenjeni su različiti algoritmi: Logistička regresija, Random Forest, XGBoost, Veštačka neuronska mreža, Support Vector Machine i Random Forest sa Bagging-om. Ovaj ceo postupak je urađen i za predviđanje dobitnika nagrede za glumu. Takođe su primenjeni isti ovi algoritmi na podatke kao što su godina rođenja glumca, starost, prethodne nominacije i osvajanja ove i drugih nagrada.

Kako bi se modeli evaluirali korišćena je 10-ostruka unakrsna validacija. Takođe za svaki model je računata tačnost, preciznost, odziv, F1 mera i matrica konfuzije.

2. PREGLED STANJA U OBLASTI

Jedan od prvih radova vezanih za predikciju uspeha filma bio je rad iz 2000. godine. Ovaj rad [1] nudi detaljnu analizu i metodologiju za predikciju filmskih zarada, ističući važnost različitih faktora pre i nakon izlaska filma, uključujući i uticaj Oskar nominacija. Autori su koristili regresione modele za predikciju logaritmovanih vrednosti domaćih zarada filmova.

Iain Pardoe [2] se bavio statističkom analizom predikcije dobitnika Oskara u četiri glavne kategorije (najbolji film, režija, glumac i glumica u glavnoj ulozi) od 1938. do 2004. godine. Za predikciju je korišćen Multinomial Logit Model (MNL).

U sledećem radu Iain Pardoe i Dean K. Simonton [3] se fokusiraju na korišćenje modela diskretnog izbora za predikciju pobednika Oskara u četiri glavne kategorije. Pored MNL modela koriste se i ML(Mixed Logit Model) modeli.

Rad [4] predstavlja jedan od prvih pokušaja korišćenja mašinskog učenja za predikciju dobitnika Oskara. Predikcija je takođe obuhvatala četiri glavne kategorije. Metodologije koje su korišćene za predikciju su: SVM, Logistic Regression, Random Forest, Gaussian Naive Bayes, Multinomial Naive Bayes.

Rad [5] koristi mašinsko učenje za predviđanje popularnosti filmova. Primenjene metodologije u ovom radu su: Logistic Regression, Simple Logistics, J48, Naive Bayes, Multilayer Perceptron Neural Network, PART. Metode su evaluirane koristeći 10-ostruku unakrsnu validaciju.

U radu [6] se vrši evaluacija performansi klasifikacionih tehnika mašinskog učenja za predviđanje uspešnosti filma.

Metodologije primenjene u ovom radu su: Logistic Regression, Support Vector Machine, Random Forest, Gaussian Naive Bayes, AdaBoost, Stochastic Gradient Descent, Multilayer Perceptron Neural Network. Skup podataka je sadržao 755 filmova između 2012. i 2015.

U radu [7] se vrši predviđanje pobednika nagrade Oskar za najbolji film na osnovu odabranih karakteristika. Korišćena metodologija je Binary Logistic Regression.

U radu [8] vršena je predikcija ekonomskog uspeha filma korišćenjem metodologija: Random Forest, SVM i Multilayer Perceptron Neural Network. Autori su koristili skup podataka koji obuhvata 3167 filmova iz perioda između 1980. i 2019. godine. Autori su koristili 10-ostruku unakrsnu validaciju za sve eksperimente, kao i metrike poput preciznosti i F1 ocene.

3. TEORIJSKI POJMOVI I DEFINICIJE

3.1. Logistička Regresija

Logistička regresija je nadgledani algoritam mašinskog učenja koji ispunjava zadatke binarne klasifikacije predviđanjem ishoda. Model daje binarni ishod ograničen na dva moguća ishoda: da/ne, 0/1 ili tačno/netačno.

3.2. Support Vector Machine (SVM)

SVM je jedan od najpopularnijih algoritama za nadgledano učenje, koje se koristi za probleme klasifikacije i regresije. Osnovna ideja SVM-a je pronaći hiper-ravan (ili više njih) koji najbolje razdvaja podatke u različite klase.

3.3. Random Forest i Bagging

Random Forest je popularan i snažan ansambl algoritam za klasifikaciju, regresiju, i druge zadatke, koji koristi više odluka stabala za donošenje konačne odluke.

Ansambl metoda kombinuje predikcije više modela kako bi se poboljšale performanse u poređenju sa pojedinačnim modelima. Random Forest koristi tehniku poznatu kao bagging (Bootstrap Aggregating), gde više odluka stabala radi zajedno. Bagging je metoda ansambl tehnike koja kombinuje više modela kako bi smanjila varijansu i poboljšala preciznost.

3.4. XGBoost

XGBoost je optimizovana implementacija gradient boosting algoritma, dizajnirana da bude izuzetno efikasna, fleksibilna i prenosiva. Gradient Boosting je popularan algoritam za boosting. Kod gradient boostinga, svaki prediktor ispravlja grešku svog prethodnika.

XGBoost je implementacija Gradient Boosted stabala odluke. U ovom algoritmu, stabla odluke se kreiraju u sekvencijalnom obliku. Težine igraju važnu ulogu u XGBoost-u. One se dodeljuju svim nezavisnim promenljivama koje se zatim ubacuju u stablo odluke koje predviđa rezultate.

3.5. Neuronska mreža

Neuronske mreže su osnovni deo mnogih savremenih metoda mašinskog učenja i dubokog učenja. Osnovni gradivni blokovi neuronskih mreža su neuroni. Svaki neuron prima ulaze, obrađuje ih i šalje rezultat napolje. Takođe bitna stavka kod neuronskih mreža su aktivacione funkcije. To je funkcija koja određuje izlaz neurona na osnovu njegovog ulaza.

3.6. One-hot-encoding i TF-IDF

One-Hot Encoding je tehnika za pretvaranje kategorijskih (diskretnih) varijabli u numerički format. Svaka kategorija se pretvara u binarni vektor sa samo jednim aktivnim (1) bitom, dok su svi ostali bitovi nulti (0).

TF-IDF (Term Frequency Inverse Document Frequency) je tehnika za pretvaranje tekstualnih podataka u numerički format, koja meri značaj svake reči u dokumentu u odnosu na sve druge dokumente u korpusu.

4. METODOLOGIJA

Struktura sistema korišćenog za predikciju dobitnika Oskara za najbolji film je identična za predviđanje dobitnika Oskara za glumačko ostvarenje. Sam sistem započinje prikupljanjem podataka sa različitih izvora i sajtova, čime se formira osnovni skup podataka za analizu.

Nakon što su podaci prikupljeni, sprovedena je eksplorativna analiza podataka (EDA) kako bi se identifikovali ključni obrasci, odnosi i potencijalni problemi u podacima, uključujući nedostajuće vrednosti. Kako bi se smanjio uticaj ekstremnih vrednosti i postigla bolja normalizacija podataka, primenjene su transformacije poput korenovanja i logaritmovanja na odabranim numeričkim atributima. Takođe, za pretvaranje tekstualnih podataka u numeričke vrednosti korišćene su metode One-Hot Encoding i TF-IDF, što je omogućilo efikasniju obradu i analizu podataka.

Nakon eksplorativne analize podataka izvršena je podela skupa podataka na trening i test skup u odnosu 80:20. Budući da se radi o nebalansiranom skupu podataka, gde postoji značajno veći broj nominovanih u odnosu na one koji su osvojili Oskara, posebna pažnja posvećena je stratifikaciji tokom podele.

Trening modela je sproveden koristeći šest različitih algoritama: Logistička Regresija, SVM, Random Forest, Random Forest uz Bagging, XGBoost i neuronska mreža. Ulaz u svaki model je obrađen skup podataka sa numeričkim vrednostima, a izlaz je klasifikacija koja označava da li je osvojen Oskar ili ne.

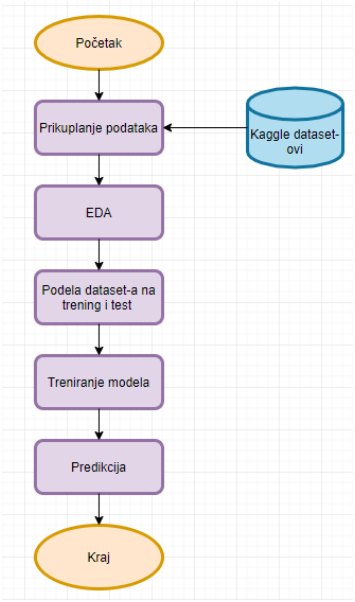
Svaki od ovih modela je obučen na trening skupu, a zatim evaluiran koristeći 10- ostruku unakrsnu validaciju, što je omogućilo robustan uvid u performanse modela. Za svaki model izračunate su ključne metrike, uključujući tačnost, preciznost, odziv, F1 meru i matricu konfuzije.

5. EKSPERIMENTI

U ovom radu su všene dve predikcije, jedna za predikciju dobitnika Oskara za glumu, a druga za glumačko ostvarenje. Korišćena su dva skupa podataka.

5.1. Skupovi podataka

Za prvi skup podataka izabrani su filmovi koji su bili nominovani od 1961. do 2021. godine. Ukupno su preuzeta tri seta podataka o nominovanim filmovima, koji su kasnije spojeni u jedan set. Set [9] je poslužio kao osnova za formiranje konačnog skupa podataka zbog raznolikosti atributa kojima raspolaže.



Slika 1. *Arhitentura sistema*

Nedostajući podaci za ceremoniju iz 2021. godine pronađeni su u okviru drugog seta podataka [9]. Spajanjem dva seta dobijen je novi set podataka od ukupno 342 uzorka. Iz trećeg preuzetog skupa podataka [10], upotrebljena je duration kolona koja sadrži informaciju o trajanju filma Jedini vid modifikacije ovog skupa podataka je preimenovanje labele original_title u Film kako bi se moglo uspešno izvršiti spajanje sa prethodnim setom po nazivu filma.
Drugi skup podataka sadrži sve nominovane glumce i glumice u glavnim i sporednim ulogoma periodu od 1961. do 2021. godine. Ukupno su preuzeta dva seta podataka sa podacima o nominovanim glumcima i glumicama, koji su kasnije spojeni u jedan set. Set [9] je poslužio kao osnova za formiranje konačnog skupa podataka zbog raznolikosti atributa kojima raspolaže. Nedostajući podaci za ceremoniju iz 2021. godine pronađeni su u okviru drugog seta podataka [9]. Spajanjem dva seta dobijen je novi set podataka od ukupno 1185 uzorka.

5.2. Eksperiment 1 – Logistička Regresija

Opisana su dva modela jedan koji se koristi za predikciju dobitnika Oskara za najbolji film a drugi za glumaćko ostvarenje.
Vrednosti hiper-parametara za prvi model su: max_iter=2000, random_state=26, C=0.1, penalty='l2', solver='liblinear', class_weight='balanced' fit_intercept= False.
Vrednosti hiper-parametara za prvi model su: max_iter=8000, random_state=26, C=100, penalty='l2', solver=lbfgs, class_weight='balanced' fit_intercept= False.

5.3. Eksperiment 2 – SVM

Za optimizaciju parametara oba modela korišćen je Grid Search. Za prvi model najbolje su se pokazali parametri C=1, gamma=0.1 i kernel='sigmoid. Dok za drugi model najbolje su se pokazali parametri C=1, gamma=0.01 i kernel='sigmoid'.

5.4. Eksperiment 3 – Random Forest

Hiper-parametri za oba modela su se podešavala ručno.
Vrednosti hiper-parametara za prvi model su: n_estimators=100, criterion='gini', max_depth=1, bootstrap=False, max_features=None, min_samples_leaf= 1, random_state=42, min_samples_split=2.
Vrednosti hiper-parametara za drugi model su: n_estimators=300, criterion='entropy', max_depth=4, bootstrap=False, max_features=0.6, min_samples_leaf= 2, random_state=42, min_samples_split=4, class_weight= 'balanced'.

5.5. Eksperiment 4 – Random Forest sa Bagging-om

Hiper-parametri za oba modela su se podešavala ručno.
Za prvi model koristio se RandomForestClassifier sa hiper-parametrima: n_estimators=10, criterion='gini', max_depth=1, min_samples_split=2, min_samples_leaf= 1, max_features=None, bootstrap=True. Zatim je RandomForestClassifier korišćen kao bazni model u BaggingClassifier-u sa 10 instanci i random_state =42.
Drugi model koristi BalancedRandomForestClassifier sa hiper-parametrima: n_estimators=200, criterion='entropy', max_depth=1, min_samples_split=2, min_samples_leaf= 1, max_features=0.6, bootstrap=True, sampling_strategy= 'auto'. Zatim je taj BalancedRandomForestClassifier korišćen kao bazni model u BaggingClassifier-u sa 10 instanci i random_state =42.

5.6. XGBoost

Hiper-parametri za oba modela su se podešavala ručno.
Vrednosti hiper-parametara za prvi model su: n_estimators=100, reg_alpha=0, random_state=42, learning_rate=0.1, random_state=42, max_depth=2, min_child_weight=1, gamma=0.1, reg_lambda=1, colsample_bytree=0.8, subsample=0.8.
Vrednosti hiper-parametara za drugi model su: n_estimators=100, reg_alpha=1, random_state=42, learning_rate=0.1, random_state=42, max_depth=4, min_child_weight=2, gamma=0.1, reg_lambda=2, colsample_bytree=0.8, subsample=0.8.

5.6. Neuronska mreža

Hiper-parametri za oba modela su se podešavala ručno
Prva model se satoji iz tri sloja. Prvi sloj ima 128 neurona, drugi 64 i poslednji 1. U prva dva sloja kao aktivaciona funkcija navedena je ReLU, a u poslednjem Sigmoid aktivaciona funkcija.
Drugi model je, s druge strane, složeniji, jer se obučava na većem skupu podataka. Sadrži četiri ključna sloja, a između svakog para slojeva umetnuti su Dropout slojevi koji smanjuju overfiting. Prvi sloj je sloj sa 512 neurona, koji koristi ReLU aktivaciju. Drugi sloj je sloj sa 256 neurona, koji takođe koristi ReLU aktivaciju i L2 regularizaciju sa parametrom 0.001. Treći sloj je sloj sa 128 neurona, koji koristi ReLU aktivaciju i L2 regularizaciju sa parametrom 0.001. Čeetvrti sloj je sloj sa jednim neuronom i sigmoid aktivacijom.
Obe neuronske mreže koriste za optimizator adam, a za funkciju gubitka binary_crossentropy.

6. REZULTATI I DISKUSIJA

Tabela 1. *Rezultati modela za najbolji film*

Modeli	10-unakrsna validacija
Logistička Regesija	87,92%
SVM	87,45%
Random Forest	**91,59%**
RF sa Bagging-om	**91,56%**
XGBoost	90,49%
NN	89,78%

Analizom dobijenih rezultata utvrđeno je da najbolje performanse za predikciju za najbolji film imaju Random Forest i Random Forest sa Bagging-om, takođe i Logistička regresija, a najgore SVM model.

Tabela 2. *Rezultati modela za glumu*

Modeli	10-unakrsna validacija
Logistička Regresija	**83,75%**
SVM	83,54%
Random Forest	82,81%
RF sa Bagging-om	81,96%
XGBoost	**84,39%**
NN	81,16%

Analizom dobijenih rezultata utvrđeno je da najbolje performanse za predikciju za glumu imaju XGBoost i Logistička regresija, takođe i SVM, a najgore neuronska mreža.

Eksplorativnom analizom podataka utvrđeno je da filmovi koji su osvojili Oskara često imaju i nominaciju za najboljeg režisera. Takođe kao što je i očekivano da najviše nominovanih filmova za nagradu Oskar i onih koji su osvojili tu nagradu pripadaju žanru drama.

Eksplorativnom analizom podataka utvrđeno je da veću šansu imaju glumci da osvoje Oskara ako je film u kojem su glumili nominovan za Oskara. Takođe kao što je i očekivano da najviše nominovanih glumaca za nagradu Oskar i onih koji su osvojili tu nagradu su glumili u filmovima koji pripadaju žanru drama. Što se tiče odnosa nagrade Oskar sa drugim nagradama, veću verovatnoću da osvoje Oskara imaju glumci ako su osvojili Zlatni globus za dramu nego za komediju.

Rad [8] fokusira se na predikciju ekonomske uspešnosti filma, dok ovaj rad predviđa osvajanje Oskara za najbolji film. Iako oba koriste slične tehnike mašinskog učenja (SVM, Random Forest, MLP), ključna razlika leži u vrsti podataka i ciljevima. Rad [8] koristi filmove iz perioda 1980–2019. i uključuje ekonomske faktore poput budžeta i prihoda, dok ovaj rad obuhvata širi vremenski period (1961–2021) i fokusira se na predviđanje nagrada. Random Forest je bio najuspešniji u oba rada, ali su MLP i SVM pokazali različite performanse u zavisnosti od prirode problema.

7. ZAKLJUČAK

U ovom radu predstavljen je sistem za predikciju dobitnika nagrade Oskar u kategorijama za najbolji film i za glumačko ostvarenje.

Jedan od ključnih zaključaka jeste da prisustvo nominacija i osvajanja drugih prestižnih filmskih nagrada, poput BAFTA, SAGA, DGA, PGA i Zlatnog globusa, značajno povećava šanse za osvajanjem Oskara, što ih čini jednim od najvažnijih faktora.

Najbolji rezultati za predikciju najboljeg filma došli su od Random Forest-a i Random Forest-a sa Bagging-om, dok je SVM bio najlošiji. Za predikciju glumačkih nagrada, najbolji su bili XGBoost i Logistička regresija, a najlošija neuronska mreža.
Kako bi se poboljšalo rešenje, potrebno je razmotriti proširenje skupova podataka, uključujući informacije o ceremonijama Oskara pre 1961. i posle 2021. godine, kao i dodavanje novih parametara poput opisa ili finansijskih informacija.

8. LITERATURA

[1] Jeffrey S. Simonoff, Ilana R. Sparrow, Predicting movie grosses: Winners and losers, blockbusters and sleepers, 2000.
[2] Iain Pardoe, "Just how predictable are the Oscars?", 2005.
[3] Iain Pardoe, Dean K. Simonton, Applying discrete choice models to predict Academy Award winners, 2008.
[4] Predicting the 85th Academy Awards: Stephen Barber, Kasey Le, Sean O'Donnell December 13, 2012
[5] Prediction of Movies popularity Using Machine Learning Techniques: Muhammad Hassan Latif, Hammad Afzal, National University of Sceinces and technology, Pakistan, August 2016.
[6] Performance evaluation of seven machine learning classification techniques for movie box office success prediction: Nahid Quader, Md. Osman Gani, Dipankar Chaki , December 2017.
[7] Predicting the "Best Picture" Oscar Award Winner: Paul Ables, 2018.
[8] Revisiting predictions of movie economic success: random Forest applied to profits: Thaís Luiza Donega e Souza, & Marislei Nishijima, Ricardo Pires, Mart 2023.
[9] https://www.kaggle.com/datasets/matevaradi/oscar-prediction-dataset(pristupljeno u januaru 2024.)
[10] https://zenodo.org/records/4244691(pristupljeno u januaru 2024.)

Kratka biografija:

Jelena Milijević rođena je u Beogradu 1998. god. Master rad na Fakultetu tehničkih nauka iz oblasti Računarstva i automatike odbranila je 2024.god.
kontakt:
jelenamilijevic98@gmail.com