Predikcija popularnosti pesme na Spotify platformi

Mihailo Majstorović
Fakulet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

Nemanja Radojčić
Fakulet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

Mitar Branković
Fakulet tehničkih nauka
Univerzitet u Novom Sadu
Trg Dositeja Obradovića 6
21000 Novi Sad

***Sažetak*—U današnje vreme, aplikacije poput Spotify-a su postale najpopularniji način za slušanje muzike. Zbog toga, razumevanje razloga koji čine neku pesmu popularnom je postao važan i zanimljiv problem za izučavanje. Ovaj rad se bavi upravo time. Napravljeni su modeli koji uzimaju u obzir brojne auditivne osobine pesama, poput tempa, glasnoće, akustičnosti, energije i mnogih drugih. Pored toga, još jedan cilj je bio kreiranje modela koji uzima u obzir ne samo auditivne osobine pesama, nego i metapodatke vezane za nju, poput popularnosti izvođača i godine kada je pesma izašla. Skup podataka koji je korišćen broji skoro 600 000 redova i nije balansiran, pa je uzet relativno balansiran podskup od 20 000 pesama. Modeli mašinskog učenja koji su korišćeni su linearna regresija, Random Forest algoritam i neuronska mreža. Krajnji cilj rada jeste utvrđivanje da li postoji bliska veza između popularnosti pesme i nekog njenog svojstva, kao i poređenje običnog modela sa onim koji uzima u obzir i metapodatke. U prevodu, da li je moguće predvideti popularnost samo na osnovu audio svojstava, ili su ipak za popularnost bitniji izvođači i godina nastanka. Početna pretpostavka je bila da su metapodaci značajniji, i da neće biti pronađena neka velika korelacija auditivnih svojstava i popularnosti, jer muzika ima ogroman broj izvođača, žanrova i tipova. Rezultati su pokazali da je početna pretpostavka ispravna.**

Ključne reči—auditivna svojstva; pesma; popularnost; metapodaci;

I. UVOD

Sve manje ljudi kupuje i skida muziku sa interneta, a većina prelazi na razne muzičke platforme i koristi aplikacije koje imaju ogroman broj dostupnih pesama, među kojima je i Spotify, kao najpopularnija od svih. Iz tog razloga, odlučili smo da radimo sa Spotify skupovima podataka. Sa preko 450 miliona korisnika mesečno i 80 miliona pesama, ima najviše dostupnih i prikupljenih podataka koji su nam korisni.

Ovaj rad može biti koristan umetnicima koji mogu dobiti uvid u to šta čini pesmu popularnom, pa se fokusirati na te aspekte prilikom stvaranja muzike. Sem toga, vlasnici aplikacije i njeni kreatori potencijalno mogu poboljšati svoje algoritme preporuke, tako da imaju što više korisnika koji često koriste platformu.

Rad pokriva nekoliko tehnika mašinskog učenja i upoređuje dobijene rezultate, prednosti i mane svakog od njih. Glavni izazov je bio obrada korišćenih skupova podataka, njihovo povezivanje i formatiranje. Napravljen je novi skup podataka povezivanjem skupa podataka pesama i izvođača, tako da bude potpuniji i bolje balansiran sa tačke gledišta popularnosti.

U ostatku rada će biti detaljnije opisani korišćeni skupovi podataka, izazovi i rešenja. Naredno poglavlje se bavi srodnim radovima iz kojih su izvučene korisne smernice, ali i nedostaci koje smo se trudili da izbegnemo. Treće poglavlje za temu ima korišćene skupove podataka i njihovu pripremu za obučavanje i testiranje modela. Četvrto poglavlje se bavi metodologijama koje su korišćene za implementaciju rešenja. U petom poglavlju se opisuju dobijeni rezultati, iz kojih se izvlače zaključci, detaljnije predstavljeni u šestom poglavlju.

II. SRODNA ISTRAŽIVANJA

U projektu Rutgera Nijkampa[1] istražena je veza između audio osobina pesme i njene popularnosti na primeru skupa od 1000 pesama dobijenih putem *Spotify API-ja*. Metode korišćene u istraživanju uključuju primenu linearne regresije i korelacijske analize. Dobijeni su podaci o akustičnosti, trajanju, energiji, instrumentalnosti, živosti, glasnoći, tempu, pozitivnosti i drugim atributima pesama.

Evaluacija rešenja sprovedena je uz pomoć R^2 metode, koja je pokazala da se na osnovu audio svojstava pesama ne može sa velikom uspešnošću predvideti njihova popularnost, s obzirom na meru R^2 od 20.2%. Međutim, ovaj rad predstavlja dobro startno mesto za dalju predikciju popularnosti pesama.

Kao komentar na ovaj rad, navodi se da je skup podataka bio premali, te da su u obzir uzeta samo audio svojstva pesama. Unatoč tome, ovaj rad može biti koristan u budućim projektima koji uključuju analizu audio osobina pesama.

Sem toga, može biti koristan u oblasti muzičke industrije i marketinške strategije za promovisanje novih pesama. Na primer, analizom audio osobina pesama mogu se prepoznati ključni faktori koji doprinose njihovom uspehu, poput tempa ili živosti, te ih koristiti u marketinškim kampanjama za promociju novih pesama.

Tema istraživanja drugog relevantnog rada[2], čiji su autori Kai Midlbruk i Kjan Šeik je predikcija koje pesme će

biti na top listama, koristeći podatke dobijene iz *Spotify API-ja*. Korišćene su četiri različite metodologije: logistička regresija, neuronske mreže, *Random Forest* algoritam i *Support Vector Machine*.

Podaci su izvučeni iz *Spotify API-ja*, a nakon filtriranja uzeto je samo top 100 pesama svake godine od 1985. do 2018. godine. Uzeti su i nasumično odabrani podaci pesama koje nisu bile hit, kako bi se podaci izbalansirali. Skup podataka uključuje 27 karakteristika kategorisanih po informaciji o numeri, izvođaču, albumu i karakteristikama analize zvuka. Evaluacija rešenja obuhvata korišćenje oko 40 računara, sa prosečnim vremenom obuke jednog dana.

Rezultati pokazuju da su svi algoritmi bili precizni, pri čemu je *Support Vector Machine* dao najpreciznije rezultate sa preciznošću od ~89%. *Random Forest* i *Support Vector Machine* su se pokazali kao dosta precizni algoritmi za predikciju popularnosti pesama.

Ovaj rad ima veliku korist za projekat koji planira da koristi iste algoritme, kao i podatke o izvođačima i albumima koje su takođe prikupljene.

Istraživanje pruža korisne uvide u to kako se može predvideti popularnost pesama koristeći različite klasifikacione algoritme. Takođe, pokazalo se da je moguće postići preciznost od ~89%, što je prilično visok nivo tačnosti. Ovi rezultati mogu biti od koristi ne samo u muzičkoj industriji, već i u drugim oblastima gde je potrebno predvideti popularnost proizvoda. Važno je napomenuti da su podaci korišćeni u istraživanju izbalansirani, što može biti ključno u predviđanju popularnosti pesama u realnom svetu. S obzirom na to da se među atributima nalaze i podaci o izvođačima i albumima, ovo istraživanje može pružiti dragocenu pomoć u planiranju i kreiranju muzičkih projekata.

Tema trećeg relevantnog projekta[3] je predikcija popularnosti pesama na osnovu *Kaggle* dataseta. Cilj ovog projekta bio je da se predvidi popularnost pesama na osnovu skupa podataka sa platforme *Kaggle*, koji sadrži informacije o 116191 pesama, 32105 izvođača i 17 karakteristika za svaku numeru.

Zavisna varijabla korišćena za merenje popularnosti bila je *popularity_score*, sa vrednostima između 0 i 100 na temelju ukupnih nedavnih slušanja. Metodologija koja je primenjena u ovom projektu uključivala je pregled podataka, istraživačku analizu podataka i primenu linearnih i logističkih regresija za predviđanje popularnosti pesama. Prvo je izvršena linearna regresija, ali nije dala dovoljno dobre rezultate (treniranje je izvršeno nad 20% pesama), pa je izvršen *undersampling* kako bi se više dalo na značaju nezavisnim karakteristikama koje su bile korisnije od ostalih. Nakon toga, ponovljena je linearna regresija i R^2 vrednosti su bile bolje. Konačno, nad drugim modelom je izvršena logistička regresija, što je dovelo do dosta preciznijih rezultata.

Obe regresije su dale dobre rezultate, a posebno je iznenađujuća bila uspešnost logističke regresije. Kao zaključak, logistička regresija se pokazala kao algoritam koji može dati dobre rezultate, ako se znaju koje su karakteristike pesme najpogodnije za što bolju predikciju. Autor projekta koristi slične atribute kao prvi navedeni projekat, ali na kraju predlaže da se uzmu u obzir i atributi izvođača, jer se pokazalo da popularnost pesme zavisi i od same izvođačke karijere.

III. Opis skupa podataka

Skupovi podataka koji su korišćeni su sa veb sajta *kaggle.com*. U pitanju su dva kompatibilna skupa, od kojih jedan sadrži informacije o pesmama[4], a drugi se odnosi na izvođače[5].

Atributi koje sadrži skup vezan za pesme koji su korisni za rad i predstavljaju ulaze u model:

- *release_date* – datum objave pesme na platformu;
- *acousticness* – akustičnost;
- *duration_ms* – dužina trajanja pesme u milisekundama;
- *loudness* – glasnoća;
- *speechiness* – prisustvo reči u pesmi;
- *liveness* – verovatnoća da je pesma snimana uživo;
- *danceability* – pogodnost pesme za ples;
- *tempo* – tempo;
- *energy* – energičnost;
- *valence* – pozitivnost;
- *instrumentalness* – prisustvo instrumentalnih delova u pesmi;
- *key* – muzička lestvica;
- *mode* – da li je u pitanju dur ili mol;
- *time_signature* – takt;
- *explicit* – da li pesma sadrži eksplicitne reči;
- *popularity* – popularnost pesme.

Atributi *acousticness*, *speechiness*, *liveness*, *danceability*, *energy*, *valence* i *instrumentalness* su brojne vrednosti između 0 i 1. Što je veća vrednost atributa, to je on izraženiji u pesmi. Obeležje *release_date* je bilo potrebno dodatno formatirati jer su negde zabeleženi tačni datumi, a negde samo godine izlaska pesme. Iz tog razloga, za svaku pesmu je uzeta samo godina. Treba napomenuti da je najskoriji datum neke pesme 16. april 2021. godine, kada je verovatno i nastao skup podataka. Atributi *mode* i *explicit* su binarne vrednosti 0 ili 1, dok je *loudness* atribut izražen u decibelima, uzimajući vrednosti od -60 do 5.38. Ciljna promenljiva, *popularity*, uzima vrednosti od 0 do 100, što je veći broj u pitanju, pesma je popularnija. Ostali muzički atributi koriste različite brojne opsege, koji im omogućavaju da dobro predstave razlike u tim osobinama. Oni neće biti detaljno objašnjeni jer nisu cilj ovog rada, i uzimaju fokus od onog zaista bitnog.

Sva svojstva koja predstavljaju ulaze u model su pre njegovog treniranja normalizovana, tako da budu u opsegu od 0 do 1. Pored njih, skup podataka sadrži artibute poput jedinstvenog identifikatora pesme, njenog naziva, imena umetnika i njegov jedinstveni identifikator. Svi oni su tekstualnog tipa, pa su izbačeni iz razmatranja. Jedinstveni identifikator izvođača je korišćen da se pesme mogu povezati sa atributima vezanim za umetnika iz drugog skupa podataka koji je naveden. On sadrži:

- *id* – jedinstveni identifikator izvođača;
- *followers* – broj pratilaca;
- *genres* – žanrovi;
- *name* – ime izvođača;
- *popularity* – popularnost izvođača.

Od ovih atributa, nama je interesantna popularnost izvođača, koja, kao i kod pesama, uzima vrednost od 0 do 100. Koristeći identifikator izvođača kao zajedničko obeležje oba skupa podataka, napravljen je nov skup koji sadrži listu pesama sa svim relevantnim obeležjima iz prvog skupa, zajedno sa popularnošću izvođača.

Jedna dilema koja se javila jeste šta da se radi u situacijama gde pesmu izvodi više izvođača, različite popularnosti. To se ne dešava često, pa je izabrano rešenje da se uzme samo najpopularniji umetnik. Razlog za ovo je što nismo upoznati sa algoritmom koji računa popularnost izvođača na *Spotify* platformi. Ako bismo sami smislili algoritam i primenili ga na popularnost izvođača, uneli bismo nekonzistentnost u podatke, i odstupanje od načina na koji *Spotify* to radi. Takođe, popularnost izvođača je usko povezana sa popularnošću njihovih pesama, pa bismo uneli nepoželjnu promenu i ispravljali realnu sliku popularnosti.

Oba skupa podataka su jako veliki, sa skoro 600 000 pesama i preko 1 000 000 izvođača, pa je nov sveden na ukupno 20 000 pesama. Njihov odabir nije izvršen nasumično, jer je skup podataka loše balansiran. Veliki broj pesama u njemu ima popularnost 0, pa se prilikom kreiranja novog skupa podataka vodilo računa o tome da se biraju pesme koje će učiniti skup ravnomernim u pogledu popularnosti. Takođe, nešto više od 700 pesama ima popularnost preko 80, što je jako mali broj, pa su te pesme namerno sve odabrane. Podela na obučavajući i test skup je izvršena u odnosu 80:20, što je standardna podela.

Pored ovog skupa, napravljen je još jedan manji podskup od oko 3500 pesama, koji je najbolje izbalansiran u pogledu popularnosti, i služi tome da vidimo koliko disbalans velikog skupa podataka negativno utiče na rezultate, ako je to uopšte slučaj. Ispod se nalazi tabela koja bliže opisuje razlike u 3 već pomenuta skupa podataka.

TABLE I. STATISTIKA POPULARNOSTI SKUPOVA PODATAKA

	Originalni skup pesama	***Skup ~ 20 000 pesama***	***Skup ~ 3500 pesama***
Broj pesama	586 672	19937	3681
Srednja vrednost	27.57	41.49	48.64
Standardna devijacija	18.37	24.13	27.23
Minimalna vrednost	0	0	0
25%	13	21	25
50%	27	42	51
75%	41	62	74
Maksimalna vrednost	100	100	100

IV. METODOLOGIJA

Način izrade ovog projekta se sastoji iz prikupljanja skupova podataka, njihovog spajanja, obrade, eksplorativne analize i izvođenja određenih zaključaka. Na osnovu svega toga, izabrani modeli mašinskog učenja pomoću kojih će se vršiti predikcija popularnosti su:

- Linearna regresija;
- *Random Forest* algoriram;
- Veštačka neuronska mreža.

Na osnovu predikcija ovih modela, možemo izvući određene zaključke, menjati ulaze u model i videti kako koji atribut utiče na rezultate prilikom predikcije.

A. Eksplorativna analiza podataka

Pošto je popularnost ključno obeležje, koje se pokušava predvideti, sortiranjem skupa podataka od najviše do najmanje popularnih pesama, primećene su sledeće stvari u prvih 50 redova skupa podataka:

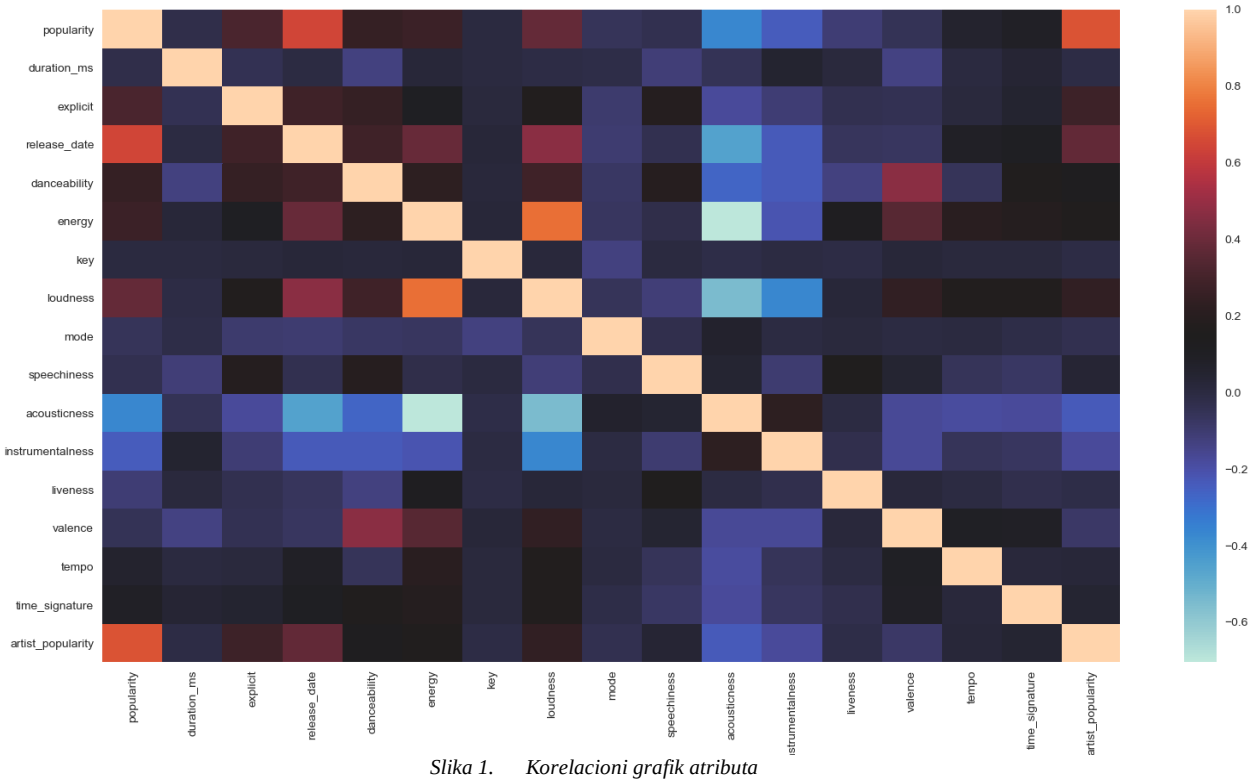
- Sve pesme su izašle 2019. godine ili kasnije, sa jednom iz 2016. i jednom iz 2013. godine;
- *danceability* obeležje uglavnom ima vrednosti veće od 0.5, sem 5 slučajeva gde je vrednost između 0.4 i 0.5, što je opet relativno visoko;
- slična situacija je i sa *energy* atributom, vrednosti su uglavnom iznad 0.5;
- takt je uglavnom četiričetvrtinski;

- Najmanje popularan izvođač ima vrednost 78, dok svi ostali imaju 80 ili više;
- Kod ostalih atributa nije primećena nikakva značajnija zakonitost.

Iz ovih pronalazaka se mogu izvući sledeći zaključci:

- Pesme koje su skoro izašle imaju veću verovatnoću da postanu popularne, što ima smisla jer su ljudi zainteresovani da uvek čuju nešto novo;
- Više se slušaju pesme koje su pozivitnije, energičnije, brže i imaju veći potencijal za igru i ples;
- Podaci o četiričetvrtinskom taktu ne daju previše informacija, jer oko 85% pesama iz skupa podataka ima taj takt. U najpopularnijih 50 pesama to nije jedini tip takta, pa je verovatnije da je ovo proizvod visokog prisustva četiričetvrtinskog takta , nego vredan podatak koji nam može koristiti;
- Što su popularniji izvođači, to ima više publike koja želi da sluša njihovu muziku. U najpopulatnijih 50 pesama nemamo nijedan takozvani *one hit wonder*, odnosno hit nekog malo poznatog izvođača kome je to jedina popularna pesma.

Na sledećoj slici se mogu videti međusobne korelacije atributa. Sa desne strane se nalazi legenda boja koja predstavlja brojeve. Što je broj bliži jedinici, to je veća korelacija između 2 atributa. Naravno, ta korelacija može biti i negativna, što znači da ako jedan atribut ima veliku vrednost, drugi će imati malu, i najčešća je kod kontrastnih atributa.



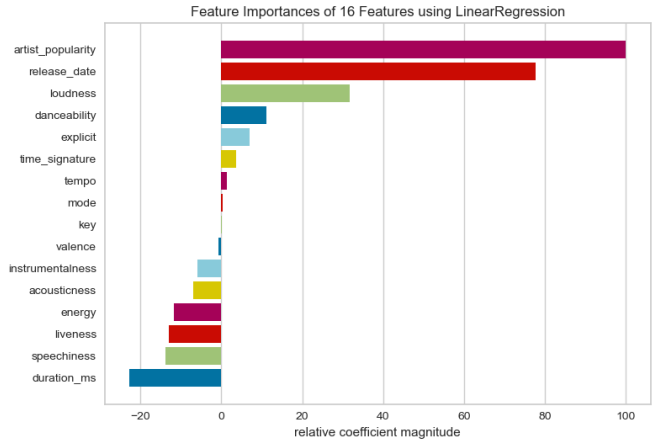
Slika 1. Korelacioni grafik atributa

Na grafiku možemo primetiti sledeće:

- Najviši nivo korelacije je između atributa *loudness* i *energy*, koji i jesu relativno srodni;
- Najmanja korelacija je kod *energy* i *acousticness* atributa, koji i jesu kontrastni;
- Atributi *popularity*, *artist_popularity i release_date* su svi međusobno u visokoj korelaciji, što se i pretpostavljalo da će biti slučaj. Pesme koje postanu popularne čine umetnika popularnim, novije pesme se slušaju više od starijih, izvođač koji je popularan ima širu publiku koja želi da čuje njegovu novu muziku, što povećava šanse da pesma postane popularna;
- Ostali atributi koji imaju skroman nivo korelacije sa popularnošću jesu *loudness, energy, danceability i explicit*, što se uklapa u trend današnje moderne muzike.

B. *Linearna regresija*

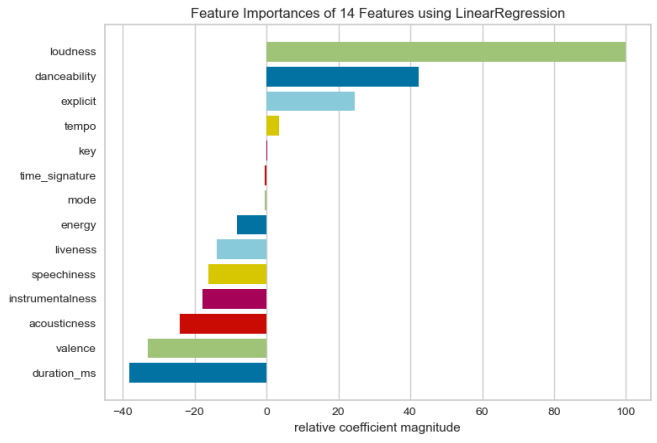
Linearna regresija je popularna i efikasna metoda za prediktivne modele u mašinskom učenju zato što je jednostavna za razumevanje i interpretaciju. Pored toga, pruža dobar balans između pristrasnosti i varijanse. Sem toga, linearna regresija je brz model koji se može brzo trenirati i primeniti, što ga čini pogodnim za aplikacije u realnom vremenu.



Slika 2. Relativna važnost atributa koristeći linearnu regresiju

Na slici iznad možemo primetiti dominantnost popularnosti izvođača i datuma izlaska prilikom predviđanja popularnosti pesme. Pored toga, malu ulogu imaju *loudness, danceability* i *explicit.* Sve je u skladu sa korelacionim grafikom sa slike 1. Može se primetiti da neki atributi imaju vrednosti u minusu, što nam ukazuje da oni mogu imati negativan uticaj na rezultate, pa se može probati sa uklanjanjem istih.

Zbog ubedljive dominantnosti 2 atributa, koji predstavljaju metapodatke pesme, želeli smo da dobijemo uvid u važnosti atributa kada se izbace popularnost izvođača i datum izlaska pesme, odnosno da se fokusiramo samo na auditivna i muzička svojstva.



Slika 3. Relativna važnost bez popularnosti izvođača i datuma

Kao što vidimo, rangiranje se promenilo za nijansu, ali ne znatno. Koeficijenti koji su dobijeni linearnom regresijom su prikazani u sledećoj tabeli:

TABLE II. KOEFICIJENTI DOBIJENI LINEARNOM REGRESIJOM

Naziv atributa	***Vrednost koeficijenta***
duration_ms	-11.27673789
explicit	4.07937789
release_date	42.61902789
danceability	6.12220288
energy	-7.78059869
key	0.14382038
loudness	18.18732128
mode	0.38087886
speechiness	-7.24846571
acousticness	-4.49300893
intrumentalness	-3.55018406
liveness	-7.13320273
valence	-0.38078356
tempo	0.76176502
time_signature	1.78113613
artist_popularity	54.43978467

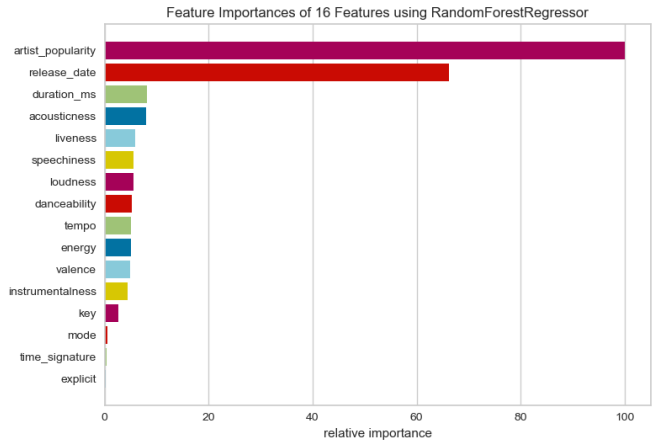
Može se primetiti da su koeficijenti srazmerni vrednostima koje su dobijene za važnost atributa.

C. *Random Forest algoritam*

Random Forest je algoritam mašinskog učenja za klasifikaciju, regresiju i druge zadatke predikcije. On se sastoji od velikog broja stabala odlučivanja koja se treniraju na različitim podskupovima skupa za obuku, uz dodatak nasumičnosti u procesu kreiranja stabala. Svako stablo u *Random Forest*-u je konstruisano na osnovu nasumično odabranih karakteristika iz skupa za obuku. Kada se vrši predikcija za novi ulazni podatak, svako stablo vrši predikciju na osnovu svojih karakteristika, a konačna predikcija se izračunava agregiranjem pojedinačnih predikcija svakog stabla.

Glavni razlog korišćenja ovog algoritma jeste otpornost na takozvani *overfitting*, kao i visoka tačnost i sposobnost da odredi relevantne atribute, kada ih ima dosta, a neki od njih nisu značajni, što je i slučaj u ovom radu.

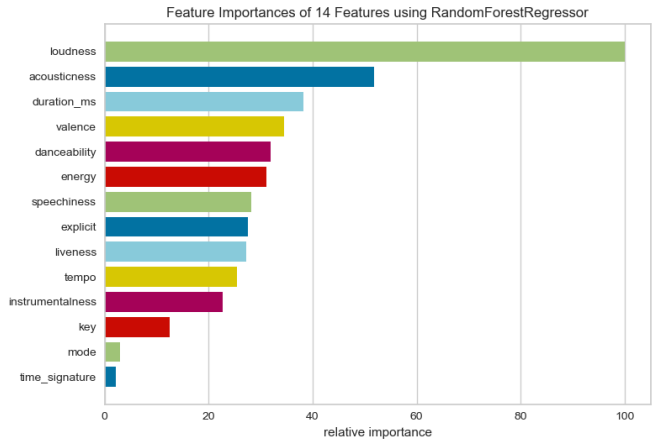
Kao i kod linearne regresije, želeli smo da vidimo relativnu važnost atributa kod ovog modela, i kako se on poredi sa linearnom regresijom.



Slika 4. Relativna važnost atributa koristeći Random Forest

Situacija na vrhu je slična kao i kod prethodnog modela, s tim što ovde popularnost umetnika i datum objave pesme postaju još dominantniji. Zanimljivo je da su sledeća dva najbitnija atributa *duration_ms* i *acousticness*, koji su kod prethodnog modela bili u velikom minusu. Razlog za ovo je što *Random Forest* predstavlja stablo i relativna važnost ne može biti negativna. Kod prethodnog modela dobijamo informacije da što su veći atributi poput *duration_ms i acousticness*, predviđa se manja popularnost pesme. Za Random Forest algoritam, ovo je važna informacija, što dokazuje i grafik iznad.

Ako isključimo 2 najznačajnija atributa, dobijamo sledeći rezulat:



Slika 5. Relativna važnost bez popularnosti izvođača i datuma

U ovoj situaciji *loudness* je duplo dominantniji od svih ostalih atributa. Iza njega je *acousticness*, sa kojim je u kontrastu, dok su *mode* i *time_signature* ubedljivo najmanje značajni. Još jedna stvar koja se može uočiti da isključenjem 2 najznačajnija atributa, većina ostalih uzimaju vrednosti od 20 do 40 na grafiku, ali sa svim svojstvima uključenim, padaju na ispod 10. To samo pokazuje ogromnu dominantnost metapodataka nad auditivnim i muzičkim osobinama pesme.

D. Veštačka neuronska mreža

Neuronske mreže su algoritam mašinskog učenja inspirisan radom ljudskog mozga. Ove mreže se sastoje od velikog broja umetnutih čvorova, poznatih kao neuroni, koji obrađuju i prenose informacije. Dobar su model za predikciju jer su u mogućnosti da „uče“, pa mogu otkriti složene veze između ulaznih i izlaznih podataka. Sposobne su da modeluju složene funkcije, pa su pogodne za situacije kada postoji veliki broj atributa. Jedina mana je što zbog složenosti, njihovo treniranje može zahtevati veliku količinu podataka, vremena i računarskih resursa.

Neuronska mreža koja je napravljena za potrebe ovog rada ima jedan ulazni sloj sa 16 neurona, jer je to broj atributa koji predstavljaju ulaze u model, jedan skriveni sloj od 8 neurona i izlazni sloj koji se sastoji iz samo jednog neurona. Ovo se pokazalo kao najbolje rešenje, nakon eksperimentisanja sa više unutrašnjih slojeva, menjajući broj neurona unutar njih. Mreža je pokretana na 100 epoha, uz sigmoidnu aktivacionu funkciju na ulaznom i skrivenom sloju. Eksperimentisano je i sa *ReLU* i *softmax* aktivacionim funkcijama, ali je sigmoidna dala za nijansu bolje rezultate.

Funkcija koja meri performanse neuronske mreže je srednja apsolutna greška, međutim, dodato je još nekoliko odabranih mera koje bliže opisuju rezultate dobijene ovim modelom, detaljnije objašnjenih u narednom poglavlju.

V. REZULTATI

Korišćenjem prethodno opisanih modela dobili smo ono očekivano i prvobitno pretpostavljeno. Sva 3 modela daju dobre rezultate i sa relativno visokom uspešnošću predviđaju popularnost pesama, kada su im na raspolaganju svi atributi. Međutim, ako se izuzmu popularnost izvođača i godina objave pesme, stvari se drastično menjaju. U nastavku će biti navedeno nekoliko tabela koje to i ilustruju.

Svaki model je vršio predikciju više puta, dok su podaci bili promešani na različite načine, podela na obučavajući i test skup je ostajala u odnosu 80:20. Za svaki od modela je uzet prosečan rezultat svih predikcija.

TABLE III. LINEARNA REGRESIJA - REZULTATI

	RMSE	***MAE***	***R^2***
Svi atributi	14.07	11.14	65.39%
Svi atributi sem popularnosti izvođača i godine izlaska	20.44	16.85	27.77%
Samo popularnost izvođača i godina izlaska	15.23	11.86	59.99%
Svi atributi – skup od ~ 3500 pesama	14.23	10.82	73.10%

S obzirom da je popularnost mera od 0 do 100, srednja apsolutna(*MAE – mean absolute error*) i korenska srednja kvadratna greška(*RMSE – root mean square error*) su relativno dobre. R^2 mera, koja uzima vrednosti 0-100% je takođe prilično visoka i iznosi 65.39%, što znači da model dobro objašnjava varijabilnost ciljne promenljive i ima dobre prediktivne performanse. Ova mera raste i do 73,1% kada se podaci još bolje balansiraju, što je slučaj sa manjim skupom od ~ 3500 pesama.

TABLE IV. RANDOM FOREST – REZULTATI

	RMSE	***MAE***	***R^2***
Svi atributi	13.11	9.89	70.67%
Svi atributi sem popularnosti izvođača i godine izlaska	18,99	15.37	37.62%
Samo popularnost izvođača i godina izlaska	15.04	11.14	60.89%
Svi atributi – skup od ~ 3500 pesama	13.15	9.97	73.89%

Kada je u pitanju *Random Forest* algoritam, dobijaju se još bolji rezultati. Greške su za nijansu manje, a razlike u brojkama velikog skupa podataka od ~ 20 000 i malog od ~ 3500 pesama je znatno niža u odnosu na model linearne regresije. Ovaj algoritam, koji je zasnovan na stablima odlučivanja (engl. *decision tree*) ne pati od nebalansiranog skupa podataka kao prethodni, niti mu je potrebna normalizacija podataka. Čak ima niže greške kod lošije balansiranog skupa. R^2 vrednost je još bolja, pa možemo zaključiti da je *Random Forest* algoritam, za ovaj konkretan problem i skup podataka, pogodniji model u odnosu na linearnu regresiju.

Ostala je još analiza rezultata neuronske mreže. U ovom slučaju, sem srednje apsolutne i korenske srednje kvadratne greške, posmatraće se kolike su razlike u pravim i pogađanim vrednostima popularnosti pesme, odnosno u koliko slučajeva se te dve vrednosti razlikuju za manje od X, gde bi X predstavljao neki realan broj. Naravno, računa se apsolutna vrednost razlike. Sa druge strane, analizira se da li postoje slučajevi gde neuronska mreža jako puno promaši sa predikcijom, pa želimo da otkrijemo šta je uzrok toga.

Mera kojom se nećemo baviti jeste preciznost, koja je jako niska iz prostog razloga što ne radimo predviđanje klasifikacionog tipa, pa svaka mala razlika u realnoj i prediktovanoj vrednosti znatno smanjuje preciznost. Sem toga, radimo sa spektrom brojeva od 0 do 100, što dodatno smanjuje ovu meru.

TABLE V. NEURONSKA MREŽA - REZULTATI

	RMSE	***MAE***
Svi atributi	13.19	10.02
Svi atributi sem popularnosti izvođača i godine izlaska	20.02	16.08
Samo popularnost izvođača i godina izlaska	13.86	10.23
Svi atributi – skup od ~ 3500 pesama	13.41	9.97

Vidimo da su vrednosti slične kao kod prethodna dva modela, sa 2 sitne razlike. Neuronska mreža se malo lošije pokazuje kada izostavimo 2 glavna atributa. Sa druge strane, rezultati kada u model ulaze samo 2 najznačajnija svojstva su bolji nego kod prethodnih modela.

TABLE VI. NEURONSKA MREŽA – REZULTATI SUBJEKTIVNIH MERA PERFORMANSI

X = razlika prediktovane i prave vrednosti	***X <= 5***	***X <= 10***	***X <= 15***	***X <= 20***	***X >= 30***
Svi atributi	34.63%	62.06%	78.98%	89.31%	3.48%
Svi atributi sem popularnosti izvođača i godine izlaska	16.42%	33.90%	51.27%	65.69%	13.16%
Samo popularnost izvođača i godina izlaska	28.66%	56.37%	76.55%	88.28%	3.63%
Svi atributi – skup od ~ 3500 pesama	32.56%	59.16%	77.74%	88.87%	3.25%

Kada su u pitanju subjektivne mere preciznosti i nepreciznosti, dobijamo rezultate iz tabele iznad. Prikazane procentualne vrednosti pesama čija je razlika realne i prediktovane popularnosti manja ili veća od navedene granice u kolonama. U skoro 90% slučajeva, razlika je manja od 20, što je solidan rezultat, mada se razlika manja od 10 može tretirati kao precizna predikcija, s obzirom da je skala od 0 do 100.

Poslednja kolona daje i najzanimljivije rezultate. 3.48% predikcija promaši za više od 30. Nakon uočavanja pojedinačnih pesama za koje se to dešava, vidi se jasan razlog, a to su upravo 2 najdominantnija atributa, popularnost izvođača i datum objave. Dešava se da popularni umetnici izbace novu pesmu koja ne postane hit. Zbog velike korelacije popularnosti pesme, izvođača i godine objave, model predviđa

veliku slušanost pesme, pa potpuno promaši. Sa druge strane, nekada se desi situacija da je i dalje relativno popularna pesma koja je izašla davno. Model u ovom slučaju, zbog ranije godine objave, predviđa nisku popularnost. Postoje i situacije gde se novi izvođač tek probije na scenu nekom pesmom, koja naglo odskače od njegovih drugih pesama po broju slušanja, ali sam izvođač nije toliko poznat, pa model opet potceni popularnost.

VI. ZAKLJUČAK

Nakon analize skupa podataka, primena transformacija, metoda mašinskog učenja i analize rezultata, možemo da izvučemo određene zaključke:

- Ako uzmemo u obzir samo auditivne i muzičke osobine pesme, teško je predvideti njenu popularnost;
- Kada se taj skup atributa proširi metapodacima poput popularnosti izvođača i godine objave pesme, dobijamo znatno bolje i preciznije predikcije;
- *Random Forest* algoritam se pokazao za nijansu bolje od linearne regresije za naš konkretan problem;
- Neuronske mreže su dale dobar uvid u situacije gde model najviše greši i razloge zbog kojih se to dešava;
- Šanse da pesma bude popularna drastično rastu uz skoriji datum objave i popularnost samog umetnika, što je i bila početna pretpostavka.

Zaključci i dobijeni rezultati su dobra vest za sve mlade umetnike koji imaju strast prema muzici. Niko na početku karijere nije bio popularan, ali svojim radom i jedinstvenim stilom, sve je moguće. Današnja muzika je raznovrsnija nego ikada, sa gomilom žanrova. Svaki od njih ima svoj deo publike koji uživa slušajući ga, sa jedinstvenim sklopom auditivnih, instrumentalnih i muzičkih osobina.

Blaga prednost u mogućnosti da postane popularna se daje pesmama koje su malo brže, glasnije, pozitivnije i izazivaju na ples. Postoji više mesta i situacija gde se mnogo češće pušta takva muzika, nego nešto sporije i tužnije. Naravno, i takve pesme imaju svoje slušaoce i fanove.

Literatura

[1] Rutger Nijkamp, University of Twente, "Prediction of product success: explaining song popularity by audio features from Spotify data" [Online]. Available: https://essay.utwente.nl/75422/1/NIJKAMP_BA_IBA.pdf

[2] Kai Middlebrook and Kian Sheik, Department of Math & Statistics, University of San Francisco, "Song hit prediction: Predicting bilboard hits using Spotify Data" [Online]. Available: https://arxiv.org/pdf/1908.08609.pdf

[3] Matt Devor, "Predicting Spotify Song Popularity – Capstone Project for Galvanize Data Science Immersive" [Online]. Available: https://github.com/MattD82/Predicting-Spotify-Song-Popularity/blob/master

[4] Lehak Narnauli, "Spotify Datasets" [Online]. Available: https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets?select=tracks.csv

[5] Lehak Narnauli, "Spotify Datasets" [Online]. Available: https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets?select=artists.csv

[6] Ibrahim Abayomi Ogunbiyi, "Top Evaluation Metrics for Regression Problems in Machine Learning" [Online]. Available: https://www.freecodecamp.org/news/evaluation-metrics-for-regression-problems-machine-learning/

[7] Alison Salerno, "Predicting Song Popularity" [Online]. Available: https://medium.com/analytics-vidhya/predicting-song-popularity-71bc3b067237

[8] Sruthi E.R., "Understand Random Forest Algorithms With Examples" [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/