Klasifikácia skladieb do žánrov

Hlavným cieľom projektu bolo vytvoriť predikčný model žánru pesničky na základe jej charakteristík.
Za týmto účelom sme sa rozhodli pracovať s rozsiahlym datasetom z Kaggle. Ten nám poskytol základné informácie o vyše 1 000 000 piesňach aj s podrobnejšími charakteristikami. V rámci projektu sme okrem žánru používali nasledovné:

- popularity (0 - 100)
- danceability (0 - 1)
- energy (0 - 1)
- key
- loudness (0 - 1)
- mode {0, 1}
- speechiness (0 - 1)
- acousticness (0 - 1)
- instrumentalness(0 - 1)
- liveness (0 - 1)
- valence (0 - 1)
- tempo (0 - 250)
- duration_ms (2000 - 6 000 000)
- time_signature (3 - 7)

Z dôvodu rôznych rozsahov sme pred testovaním každého modelu dáta štandardne preškálovali tak, aby mali priemer 0 a disperziu 1.

Dataset bolo potrebné prečistiť a prispôsobiť naším potrebám, hlavnou úpravou bolo zredukovanie počtu žánrov. Mnohé z nich sa prelínali alebo popisovali skôr tému piesne než samotný žáner, preto sme sa pokúsili vytvoriť kategórie na mieru. Väčšinu žánrov sme následne zgeneralizovali na 13 nami vytvorených žánrov (popísaných nižšie) a niektoré sme museli vyhodiť, lebo nesedeli do žiadneho zo zvolených žanrov. Táto zmena modelom výrazne pomohla.
Súbor s touto novou klasifikáciou do žánrov je vygenerovateľný spustením generate_data.py

Po spracovaní dát sme vytvorili niekoľko štatistických modelov, s cieľom porovnať ich a zvoliť ten najlepší.

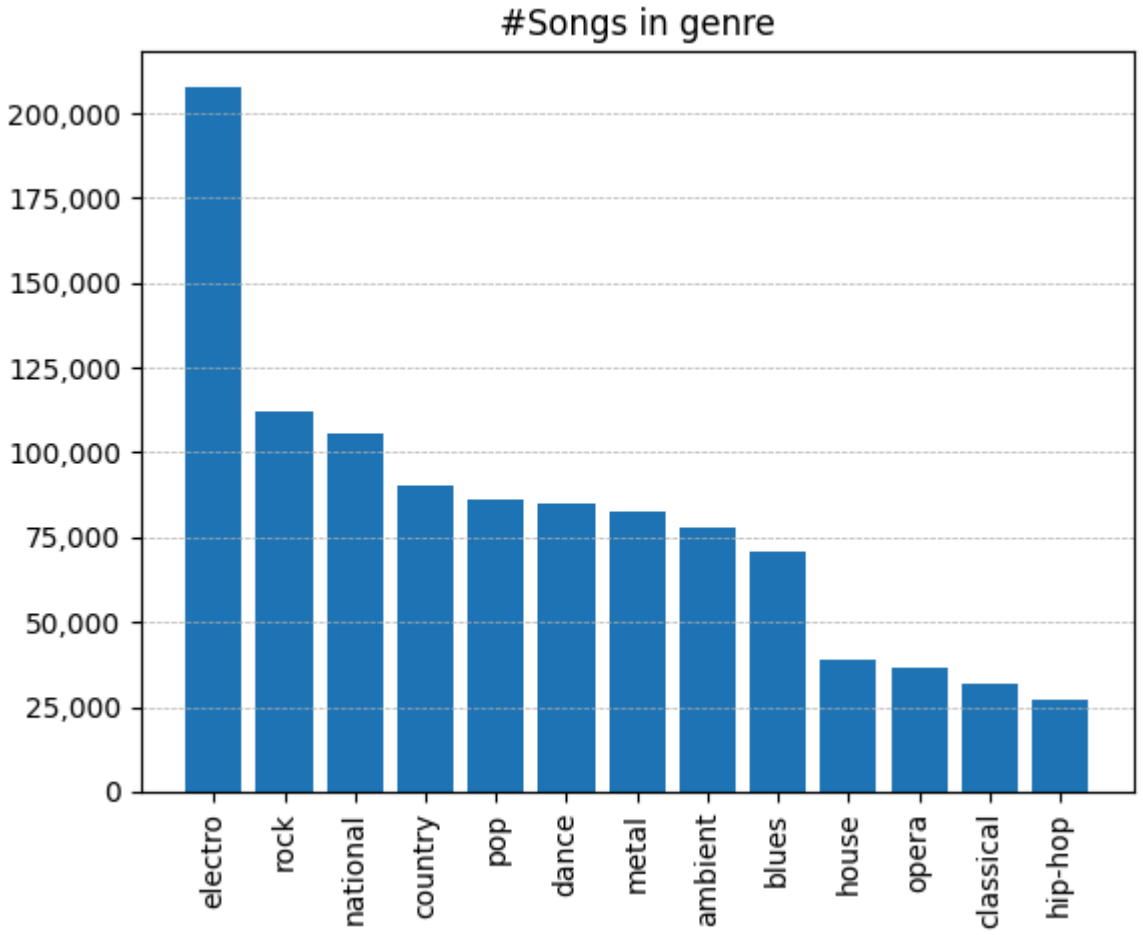
Nové žánre

Zgeneralizované žánre a ich “podžánre”

Metal: black-metal, metalcore, heavy-metal, metal, death-metal, grindcore
Rock: rock-n-roll, alt-rock, punk-rock, psych-rock, rock, hard-rock, goth, emo,
Pop: k-pop, cantopop, pop, power-pop, pop-film, indie-pop
Ambient: new-age, ambient, sleep , chill

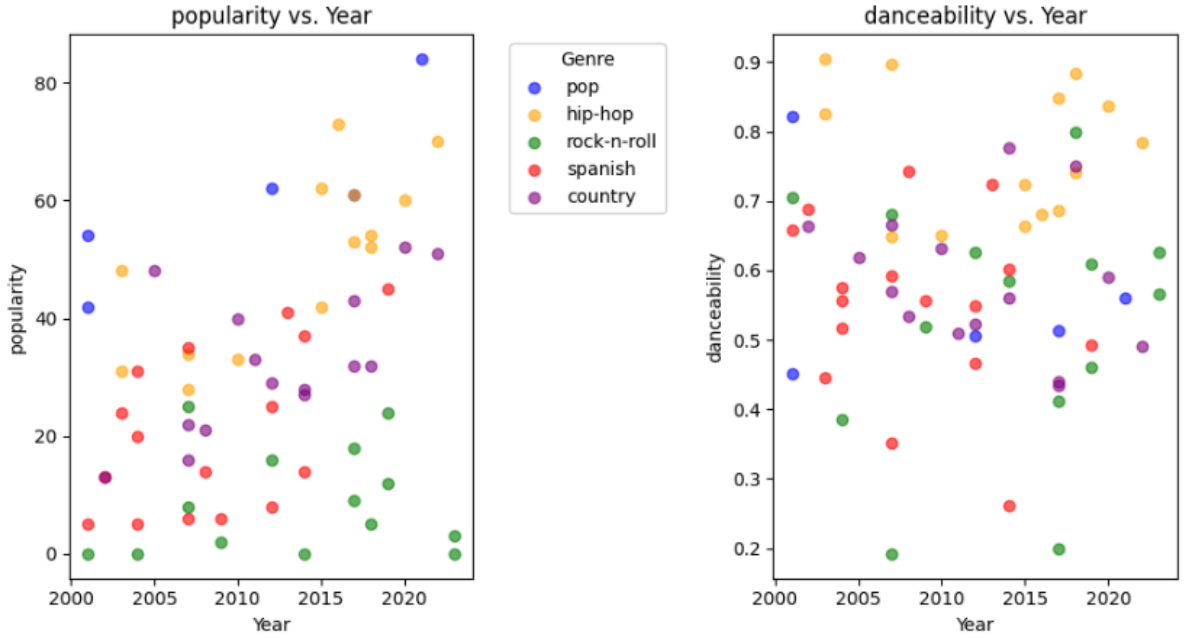
Electro:	dubstep, electronic, edm, detroit-techno, party, dance, techno, garage, disco, trance, hardstyle, electro, drum-and-bass, breakbeat, minimal-techno, dub, hardcore
Dance:	dancehall, tango, club, salsa, samba
Classical:	piano, classical
Hip-hop:	hip-hop, trip-hop
Blues:	soul, blues, jazz, afrobeat, ska
Country:	guitar, country, acoustic, folk, sertanejo
Opera:	opera, gospel
National:	indian, spanish, french, german, swedish, forro
House:	progressive-house, chicago-house , house , deep-house

Obr 1: Výsledné zastúpenie kategórií vo výslednom datasete



Celkovo najviac žánrov patrilo do electro, ak sme z týchto dát samplovali, tak sme použili stratify = “genre” na rovnaký pomer vybraných a nevybratých z každej kategórie.

Obr 2: Popularity a danceability 5-tich žánrov počas rokov



Väčšinu času sme investovali do hľadania predikčných modelov. Na toto sme si spravili spomínané podžánre, nakoľko sa nám vzdalo nepredstaviteľné nájsť model ktorý by dokázal rozlíšiť aj medzi podžánrami. Už len týchto 5 žánrov sa na grafe dosť prelína, potrebovali sme zachovať množstvo dát pre trénovanie modelov a preto sme ich zakategorizovali. Odstránili sme niektoré pôvodné žánre, ktorých bolo buď málo alebo boli príliš všeobecné. Po tejto úprave nám v datasete ostalo okolo 1 050 000 piesní.

Lineárna regresia

Ak by bolo očakávané, takýto model nie je vhodný na klasifikáciu skladieb do žánrov.
Dôvody:

- lineárna regresia nie je vhodná na klasifikáciu

Ďalšie nepríjemnosti (keby náhodou lin. regresia aj bola dobra na klasifikáciu)

- napríklad premenná interpret (artist_name), ktorá by celkom pekne mohla vravieť niečo o tom do akého žánru skladba patrí bola kategorická s príliš veľa hodnotami ~60 000. Toto zabraňovalo použitiu napríklad one-hot-encoding, kvôli príliš veľkému výstupnému data-frame.
- Použité kódovanie pre žánre bolo priradením čísla ku každému žánru. Toto je jednoznačne nevhodné napríklad z toho dôvodu, že malá zmena v hodnote spôsobí zlú predikciu (5.4 → a, 5.6 → b).

Nemalo zmysel očakávať rozumné predikcie, ale je pekne vidieť porovnanie toho, ako je táto metóda nevhodná a čo môže vzniknúť, keď sa model používa neopatrne.

Logistická regresia

Táto metóda veľmi pekne klasifikovala (aspoň oproti ostatným našim modelom).
Tiež veľmi benefitovala so zníženia počtu žánrov: pred znížením accuracy = 0.25 a po 0.37.
Taktiež bolo treba pred transformáciou žánrov zvýšiť limit na maximálny počet iterácií, inak regresia neskonvergovala.

Pri Logistickej regresii je vhodné nejakým spôsobom preškálovať vstupy, čo bolo veľmi pekne viditeľné. Pri nepreškálovaní regresia neskonvergovala na bežný počet iterácii, taktiež boli predikcie nepoužiteľné. Mali síce accuracy = 0.31, ale to bolo umelé, lebo všetky precision a recall boli 0 s výnimkou 1 žánru (electro), ktorý potiahol výsledok príliš hore.

Ďalšou zaujímavou vlastnosťou/pozorovaním je, že predikcie neboli ovplyvňované ponechaním vysoko korelovaných premennými na vstupe.

Najhoršie predikovanými žánrami sú: blues, house, pop, a national. Toto nie je nečakané vzhľadom na subjektívne celkom veľký rozdiel medzi jednotlivými “pod-žánrami”.

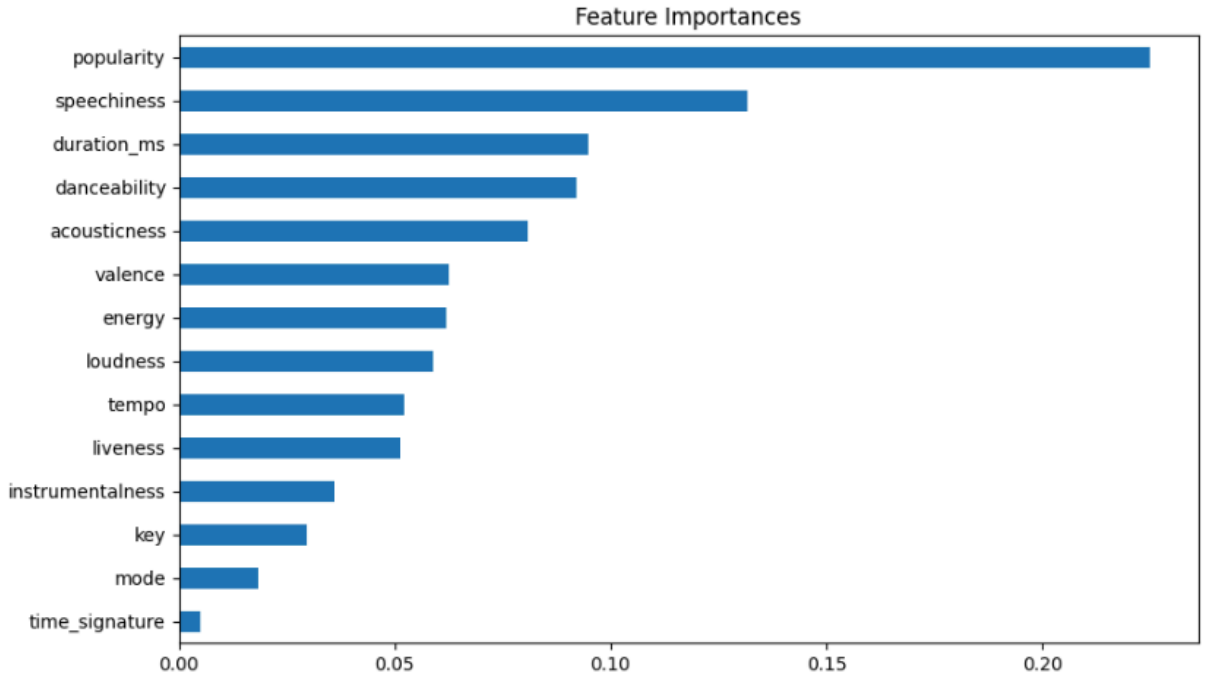
K najbližších susedov

Hľadanie parametrov bolo neprínosné. Od skúšania rôznych k po rôzne metriky vzdialenosti a sa rôzne kombinácie javili ako podobné, najlepšie vyšla menhetenská vzdialenosť s k=19.
Počiatočný pokus predikovania ztroskotal na zložitosti algoritmu K najbližších susedov.
Najúspešnejšie bolo vyselektovanie 5-tich najzaujímavejších žánrov a použitie menšieho samplu zo všetkých dát (50 000), kde bola najlepšia presnosť 0.71.
No keď sme sa snažili dáta upraviť na potenciálne lepšie výsledky iných modelov ktoré sme ešte mali na pláne, na 13-tich kategóriách a sampli s veľkosťou 100 000 sa presnosť zmenšila na 0.285.
Buď je naše kategorizovanie žánrov zlé, alebo sme mali použiť jednoducho menej dát a čisté žánre. Najpravdepodobnejšie je to spôsobené zložitejšími dátami.

Random Forest

Rôzne kombinácie parametrov prinášali asi najpodobnejšie výsledky spomedzi všetkých testovaných metód. Použitie 100 stromov bez obmedzenia hĺbky malo presnosť 0.7179 na pôvodnom datasete a na sampli (n=100 000) z upraveného 0.434 s rovnakými parametrami.

Obr.3: Vplyvy jednotlivých premenných pri použití Random Forest



Popularity má značne najvyšší vplyv, zatiaľ čo time_signature, mode (celočíselné hodnoty) vplývali najmenej, napríklad obzvlášť kvôli tomu, že väčšina piesní malo bežnú time signature 4. Instrumentalness bola skewed, čo znamená že väčšina piesní malo nulovú alebo veľmi nízku instrumentalness. Dalo by sa uvažovať napríklad nad použitím logaritmu na zlepšenie jej rozdelenia.

Neurónové siete

Experimentácia s parametrami viacvrstvovej neurónovej siete bola z hľadiska hladania parametrov podobná ako s metódou KNN. Síce sme už nemohli hrubou silou vyskúšať všetky “z nášho pohľadu vhodne” vyzerajúce parametre, ale sme len intuitívne vylepšovali čo sme mali. Každá zmena parametrov mala malý efekt na výslednú presnosť, bolo treba robiť radikálnejšie kroky. Napríklad počet neurónov bol veľmi podobný. Zastavili sme sa na 30-tich neurónoch. Presnosť sa so zvyšovaním počtu iba mierne znižovala, až hodnoty cez 1000 už spôsobovali mierny očividný overfitting.

Na urýchlenie sme použili batch_size=2048 čo nezmenilo presnosť a umožnilo nám lepšie sledovať zmeny.

Aktivačné funkcie sme skúšali tanh, sigmoid, relu aj linear, najlepšie sa hodila relu.

Na inicializácia váh siete sme tiež skúšali metódy HeNormal() (ktorá sa bežne využíva na relu) a GlorotUniform() (bežne využívaná na tanh) aj kombinovane, ale nemali efekt na presnosť.

Learning rate bol dobrý tiež rôzny od 0.05 po 0.001 sa javil rovnako.

Z pozorného sledovania zmien presnosti počas jednotlivých epoch sa dalo všimnúť, že presnosť do určitej epochy postupne stúpa, no pri presnosti maximálne 0.39, potom začala skákať znova na nižšie hodnoty a naspäť - stagnovať. Najvyšší skok bol 0.4.

Na pôvodných dátach sme ani netestovali.

Nástroje, technické výzvy a ostatné

Nástroje:

- python3
- jupyter notebook
- pandas
- numpy
- tensorflow
- scikitlearn
- seaborn
- matplotlib

Technické výzvy:

- obmedzenia RAM (OS niekedy zabíjal výpočty a iné programy)
- google collab sa tiež niekedy vzdával
- nedostatok Nvidia grafických kariet: umožnilo by nám rýchlejšie experimentovanie so zložitejšími neurónovými sieťami