Čo robí filmy obľúbenými a neobľúbenými?

Maximilián Zivák, Dávid Daniš

Zima 2023

1 Úvod

V tomto projekte sa pozrieme ktoré faktory robia filmy obľúbenými. Keďže existuje veľa možností ako na túto otázku odpovedať, rozhodli sme sa pre preskúmať ich viac. Zvolili sme 3 modely a 3 testovacie metriky. Najprv sme testovali random forest regressor, teda množinu rozhodovacích stromov ktorých výsledok sa spriemeruje, na les sme využili metriku MDI(Mean Decrease in Impurity), táto meria priemernú zmenu vo variancii (impurity) keď prechádzame z rodiča do potomka. Keďže táto metóda má určité obmedzenie, použili sme aj permutation importance, najpv zrátame scoring function na fitnutom datasete a potom postupne permutujeme premenné, permutation importance je rozdiel medzi score na fitnutom datasete a score po permutácií. Túto metódu sme použili aj na jednoduchej lineárnej regresii. Na koniec sme ako "sanity check"fitli XGBoostRegressor, na ktorom sme použili metriku gain, teda zlepšenie score po pridaní jednej premennej.

2 Dáta

Ako dáta sme použili len tsv súbory z IMDb Non-Commercial Datasets [1], rozmýšľali sme aj nad API (IMDB, letterboxd) ale dostať k nim prístup nie je úplne jednoduché. Z datasetov sme použili 4:

- title.basics.tsv.gz - základné dáta o filmoch, uložené v basics.tsv
- title.crew.tsv.gz - dáta o režiséroch a spisovateľoch každého filmu, uložené v crew.tsv
- title.ratings.tsv.gz - dáta o hodnoteniach filmov, uložené v ratings.tsv
- name.basics.tsv.gz - dáta o jednotlivých ľuďoch pracujúcich na filmoch , uložené v people.tsv

Tieto boli spracované do output/data-40000.csv, a následne použité na analýzu. data-40000.csv obsahuje 40000 samplov z 4 častí datasetu rozdeleného podľa kvartilov počtu hodnotení filmov.

3 Výsledky

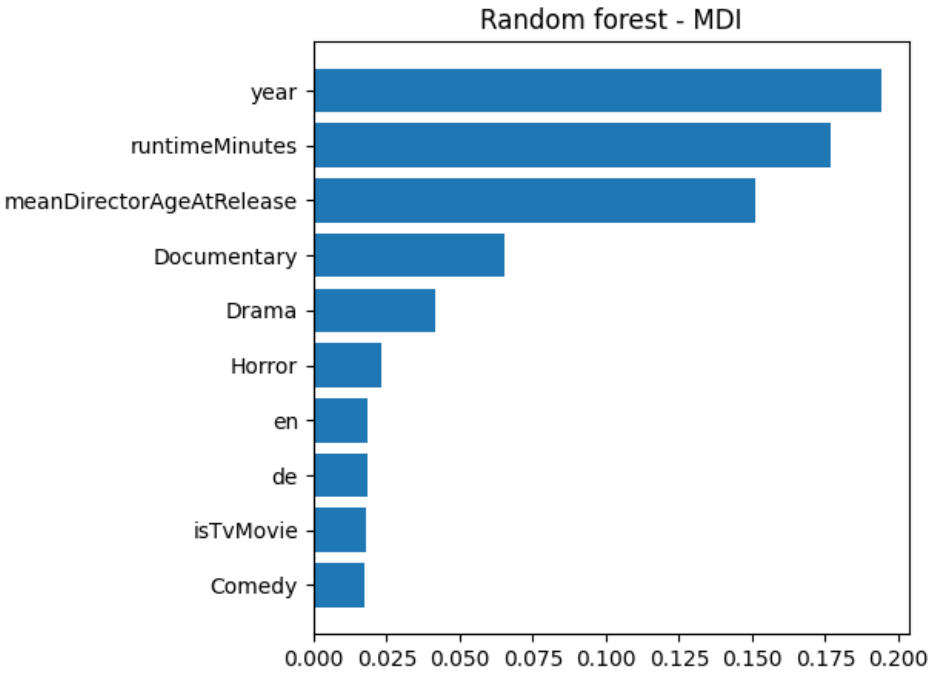
3.1 Korelačná matica

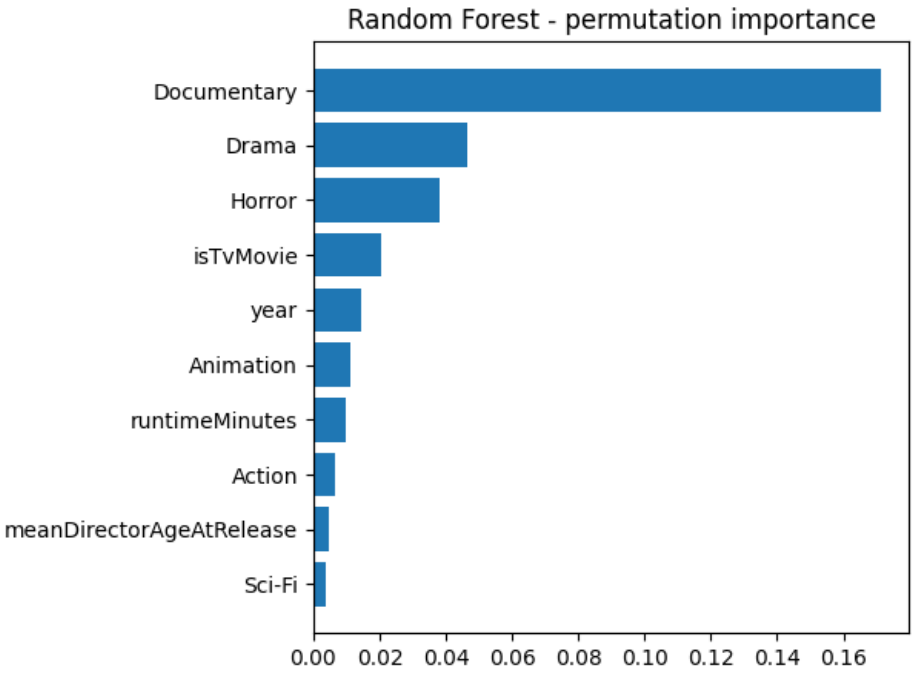
Ako najjednoduchší spôsob riešenia tohto problému sme použili jednoduchú korelačnú maticu (spearman)

```
Horror           -0.182616
Comedy           -0.111548
Thriller         -0.108114
Action           -0.105606
Sci-Fi           -0.091230
                    ...
Biography         0.101757
Drama             0.105895
runtimeMinutes    0.109611
Documentary       0.270255
averageRating     1.000000
Name: averageRating, Length: 61, dtype: float64
```

3.2 Random forest regressor

Ďalej sme natrénovali random forest regressor, tento model obsahuje metriku ktorá nám povie ktorá premenná je najdôležitejšia öut of the box", MDI (Mean Decrease in Impurity). Random forest regressor sme použili, lebo to je množina rozhodovacích stromov, teda očakávame že najdôležitejšia premenná bude na "začiatku"stromu. Bohužiaľ táto metrika dáva až moc veľký dôraz na numerické premenné[3]. Keďže náš dataset obsahuje veľa kategorických, rozhodli sme sa vyskúšať aj permutačné hodnotenie dôležitosti.

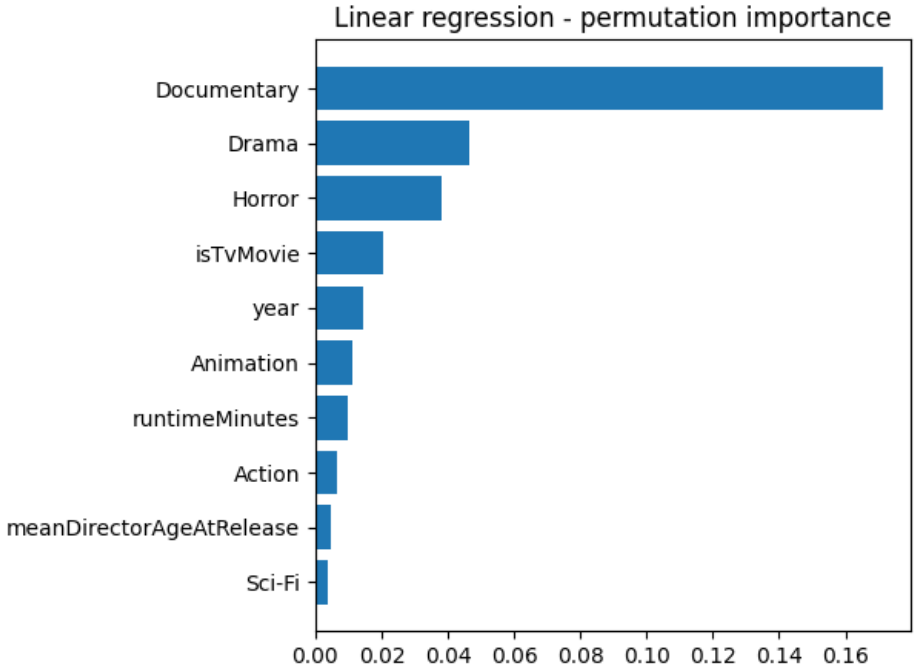




Naozaj vidno že MDI numerické premenné ako rok vydania, dĺžka filmu a priemerný vek režisérov uprednosťuje pred kategorickými.

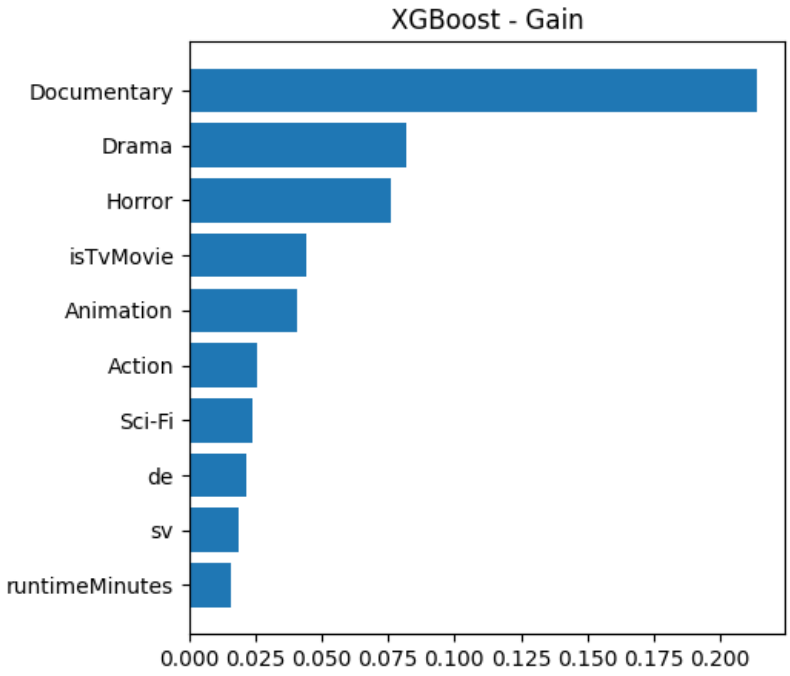
3.3 Lineárna regresia

Ďalej sme testovali permutation importance na lineárnej regresii.



3.4 XGBoost

Ako poslednú metriku sme zvolili gain v XGBoost modeli, táto metrika udáva priemerné zlepšenie loss score po pridaní určitej premennej.



3.5 Záver

Ako najdôležitejší faktor, ktorý robí film "dobrým"sa ukazuje žáner, konkrétne dokument. Tak isto netriviálnu úlohu zohráva dĺžka filmu a či je film natočený pre televíziu alebo kino. Ak chceme zistiť či tieto premenné ovplyvňujú hodnotenie pozitívne alebo negatívne môžeme nazrieť do korelačnej matice.

```
Horror           -0.182616
Animation         0.050096
isTvMovie         0.091766
Drama             0.105895
runtimeMinutes    0.109611
Documentary       0.270255
Name: averageRating, dtype: float64
```

4 Nástroje, technické výzvy etc.

4.1 Spracovanie dát

Na spracovanie dát bola použitá python knižnica pandas a langdetect [2], na zistenie v akom jazyku bol film uvedený na trh, keďže v IMDB datasetoch táto informácia vo veľa prípadoch chýba. Ako prvá technická výzva sa ukázala samotná veľkosť datasetov, počet riadkov sa pohybuje v miliónoch, pri testo-

vaní sme sa rozhodli použiť len menší subset filmov. Tiež prebehla diskusia o použití knižnice DASK, ktorá umožňuje paralelné ukladanie a spracovanie pandas dataframes. Ďalej sme museli zakódovať kategorické dáta, využili sme one-hot encoding. One-hot encoding, zakóduje kategorickú premennú do n nových binárnych premenných, kde n je počet unikátnych hodnôt ktoré môže kategorická premenná nadobúdať.

4.2 Modely

Ako modely sme sa rozhodli použiť random forest regressor, lineárnu regresiu a XGBoost regressor.

Literatúra

[1] Imdb non-commercial datasets. https://developer.imdb.com/non-commercial-datasets/.

[2] langdetect. https://pypi.org/project/langdetect/.

[3] Permutation importance vs random forest feature importance (mdi). https://scikit-learn.org/stable/auto$_e$*xamples*/*inspection*/*plot*$_p$*ermutation*$_i$*mportance.html*$sphx-glr-auto-examples-inspection-plot-permutation-importance-py$.