Analiza rozkładu ocen filmów

Kacper Budnik, Aleksy Walczak, TBA, TBA

2024-12-13

Spis treści

1 Wstęp

Celem tej analizy jest zbadanie rozkładu ocen filmów. Skupimy się głównie na dwóch składowych: średniej ocenie oraz odchyleniu standardowemu ocen. Dane zostały scrapowane z portalu Filmweb. Pobrane zostały informacje o 10 000 filmach. Filmy zostały wybrane na podstawie wyszukiwarki portalu Filmweb z parametrem `Popularne: najbardziej` w dniu 8 styczeń 2025 roku. Dane (rozbite na dwie tabele) prezentują się następująco

Tabela 1: Przykładowe dane.

Title	Year	Genre	Award.Amount	Box.Office	Budget	Time..mins.
Zielona mila	1999	Dramat	30	286801374	60000000	188
Skazani na Shawshank	1994	Dramat	30	28884504	25000000	142
Forrest Gump	1994	Dramat	68	678226465	55000000	142
Leon zawodowiec	1994	Dramat	9	19569225	16000000	110
Requiem dla snu	2000	Dramat	31	7390108	4500000	102
Matrix	1999	Akcja	41	467222728	63000000	136

Tabela 2: Przykładowe dane.

Title	Reviews	Number.of.Reviews
Zielona mila	[0.6, 0.2, 0.5, 0.8, 2.0, 3.0, 9.0, 23.0, 29.0, 32.0]	1017686
Skazani na Shawshank	[0.6, 0.3, 0.4, 0.7, 1.0, 2.0, 7.0, 20.0, 30.0, 38.0]	942435
Forrest Gump	[0.6, 0.3, 0.6, 0.9, 2.0, 4.0, 10.0, 24.0, 28.0, 31.0]	1009592
Leon zawodowiec	[0.7, 0.3, 0.7, 1.0, 3.0, 6.0, 15.0, 29.0, 24.0, 20.0]	844224
Requiem dla snu	[1.0, 0.7, 2.0, 2.0, 4.0, 7.0, 16.0, 27.0, 21.0, 18.0]	682470
Matrix	[2.0, 0.9, 2.0, 3.0, 5.0, 9.0, 18.0, 26.0, 17.0, 17.0]	843868

Zmienne `Title`, `Year`, `Box.Office`, `Budget`, `Number.of.Reviews` są dobrze tłumaczone przez swoją nazwę. Zmienna `Genre` jest zmienną odpowiadającą za gatunek filmowy, a zmienna `Time..mins.` odpowiada za czas trwania filmu w minutach. Natomiast zmienna `Award.Amount` może być delikatnie myląca. Zawiera ona nie tylko informacje o zdobytych nagrodach (jak Oscary, Złote Globy, Saturny, ...) ale również nominacje do nagród. Ostatnia zmienna: `Reviews` zawiera wektor z informacjami o rozkładzie ocen w procentach, posortowanych rosnąco względem oceny. Zatem w pierwszym przypadku 0.6% osób oceniało film `Zielona mila` 1 na 10, 0.2% na 2, ..., 32% osób oceniło 10 na 10.

W danych znajduje się łącznie 9352 obserwacji brakujących. W szczególności w zmiennych `Box.Office` (3806) oraz `Budget` (5509). Ponieważ brakujące dane często występują w tych samych obserwacjach (tylko 5 posiada budżet, a nie posiada jakiejkolwiek innej zmiennej)

oraz głównym celem projektu była nauka scrapowania danych, postanowiliśmy się pozbyć tych obserwacji, otrzymując w ten sposób 4486 obserwacji zawierających wszystkie cechy.

Dodatkowo w dalszej analizie nie korzystaliśmy ze zmiennej `Title`, która jest unikatowa dla każdej obserwacji.

2 Analiza

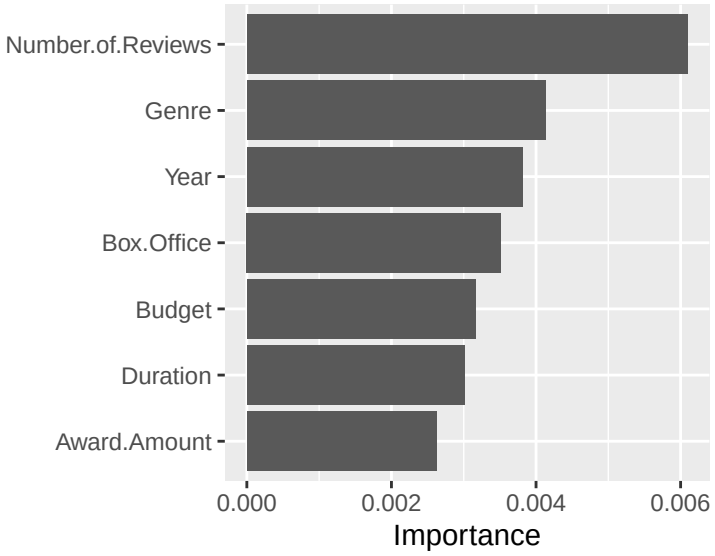
Analizę zaczniemy od podzielenia obserwacji na zbiór uczący i testowy w stosunku 2:1. W tej analizie zastosujemy trochę bardziej niestandardowe podejście. A mianowicie dla obserwacji ze zbioru uczącego wylosujemy ocenę odpowiadającą jej, z rozkładu ocen związanego z daną obserwacją. Dla takich danych stworzymy drzewo decyzyjne predykujące jedną ocenę w zależności od pozostałych danych. Taką predykcję powtórzymy wiele razy, każdorazowo losując inną ocenę. W ten sposób otrzymaliśmy 1000 drzew, każde zwracające jedną ocenę. Podejście to rozumiemy, że każde drzewo predykuje, jaką ocenę wystawiła by dana osoba poszczególnym filmom. Oczywiście nie jest to podejście doskonałe, ponieważ podczas losowania ocen w celu stworzenia modelu, losujemy te oceny niezależnie. W świecie rzeczywistym jednak każdy ma choćby swój ulubiony i nielubiany gatunek filmowy, ale wyniki uznaliśmy za zadowalające. W celu sprawdzenia jak sobie jak sobie poradził model

$$err_j = \frac{1}{10} \sum_{i=1}^{10} \left| \frac{n_{i,j}}{N} - p_{i,j} \right|$$

gdzie

- err_j – błąd podczas predykcji rozkładu j-tego filmu;
- N – liczba drzew, w naszym przypadku N =1000;
- $n_{i,j}$ – liczba drzew predykujących ocenę i dla obserwacji j;
- $p_{i,j}$ – prawdopodobieństwo z danych oznaczające, że j-ty film zostanie oceniony na ocenę i.

Zmienna ta określa z jakim średnim błędem oszacujemy rozkład danego filmu. Uśredniając teraz po wszystkich filmach otrzymaliśmy błąd na poziomie 2.7% na zbiorze uczącym oraz 1.1% na zbiorze testowym.


Number.of.Reviews
Genre
Year
Box.Office
Budget
Duration
Award.Amount
0.000
0.002
0.004
0.006
Importance