Analýza hodnotení 5000+ filmov z databázy IMDB

Jakub Bahyl[1]
FMFI, mFTL - 4.ročník

(Dated: 26 June 2017)

V tejto práci prezentujem výsledky analýzy niektorých vlastností datasetu 5000+ filmov, pochádzajúcich zo stránky `imdb.com`. Analyzovaná je kvalita a predikcia hodnotení filmov natočených USA a Kanadou.

Keywords: kontingenčná tabuľka, logistická regresia, bootstrap

"If your experiment needs statistics, you ought to have done a better experiment."
Ernest Rutherford

I. INTRO

Je úspešnosť USA v tvorbe kvalitných filmov vyššia ako pre zvyšok sveta? Je možné predpovedať verejné hodnotenie filmu ešte predtým, než film ktokoľvek uvidí? Ako veľmi si môžeme byť istý kvalitou kanadských filmov?

Na účel zodpovedania týchto otázok som si zobral na pomoc verejnú databázu viac ako 5000 filmov, ktorá bola získaná *scrapovaním* stránky `imdb.com`. V datasete sa pre každý film nachádza 28 parametrov, medzi ktoré patria okrem iného aj: *názov, režisér, rozpočet, farebnosť, dĺžka filmu, IMDB hodnotenie, jazyk, rok premietania,...*

V rámci tejto práce som si z tohto veľkého datasetu vybral podmnožinu, ktorá:

- Nikde neobsahuje nevyplnené (NA) polia
- Na účel trénovania modelu obsahuje len filmy, ktoré sú v TOP 250 (veľmi kvalitné filmy s hodnotením približne aspoň $\approx 80\%$) alebo filmy s hodnotením pod 50% (podľa verejnej mienky *zlé filmy*). Na účel testovania modelu je vybraných ďalších zhruba 100 dobrých a zlých filmov. Vynechávam teda tú množinu, ktorá je z hľadiska kvality diskutabilná.

Prílohou k tomuto dokumentu je R-kový program, v ktorom možno chronologicky vidieť technickú realizáciu všetkých krokov, ktoré v práci opisujem.

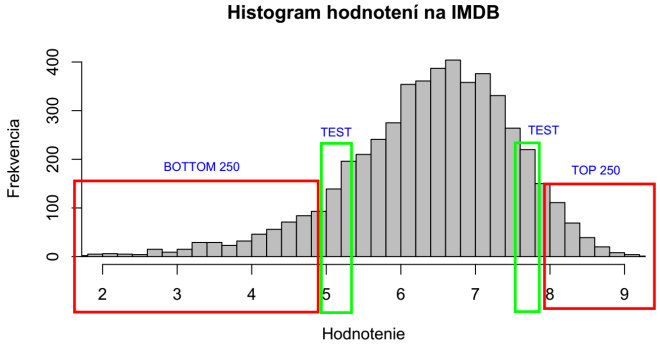

FIG. 1. Z datasetu vyberám filmy na tréning a test modelov.

II. KONTINGENČNÉ TABUĽKY

V tejto časti faktorizujem *IMDB hodnotenia* na *"good"* a *"bad"* podľa toho, či má daný film hodnotenie ≥ 8 alebo < 5. Ďalej takisto faktorizujem krajinu pôvodu filmov na *"USA"* a *"Other"* a *rok premietania* na *"old"* a *"new"* podľa toho, či bol film natočený do roku 2000 alebo neskôr. Časom uvidíme, že na tom záleží.

Jednotlivé počty filmov pre danú kvalitu a pôvod sú vizualizované nižšie (momentálne nezáleží na *roku premietania*):

		Kvalita	
		"bad"	"good"
Krajina	"Other"	43	66
	"USA"	259	148

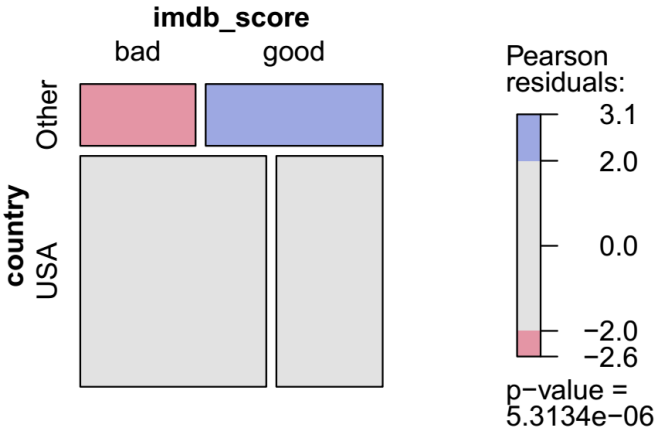

FIG. 2. Kontingenčná tabuľka kvality a pôvodu filmov

Z uvedeného vyvstáva otázka: *Produkuje USA percentuálne viac "zlých" filmov, než priemerne ostatné krajiny?* Obrázok aj odds ratio $(43/66)/(259/148) \doteq 0.37 < 1$ nasvedčujú, že natáčanie filmovou produkciou USA nepôsobí prívetivo na pravdepodobnosť kvality filmu.

Pearsonov χ^2 test homogenity (p-value je vidno na FIG.2.) a D-test (p-val$\sim 10^{-90}$) taktiež jasne ukazujú, že nejde o náhodu a teda že v relatívnych jednotkách generuje USA horšie filmy, než iné krajiny.

Avšak, s veľkou pravdepodobnosťou model prepadol Simpsonovmu efektu. Keby by sme filmy pôvodne rozdelili sofistikovanejšie na rôzne skupiny (podľa žánru,

obsadenia hercov, roku natočenia...), mohla by analýza dopadnúť inak. Príklad jedného takéhoto delenia je už spomenutá faktorizácia *roku premietania* ma *"old"* a *"new"*. Týmto vznikajú dve tabuľky a dve vizualizácie rozdelenia:

Nové		**Kvalita**	
		"bad"	"good"
Krajina	"Other"	30	43
	"USA"	208	72

Staré		**Kvalita**	
		"bad"	"good"
Krajina	"Other"	13	23
	"USA"	51	76

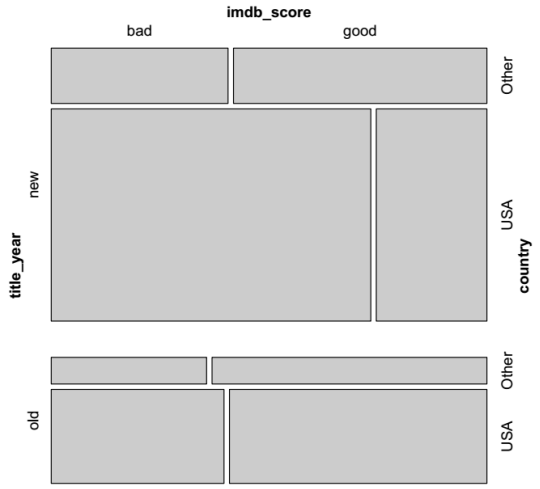

FIG. 3. Dve kontingenčné tabuľky pre "nové" a "staré" filmy

V takomto rozdelenom prípade je odds ratio pre nové filmy rovné 0.24 a pre staré filmy 0.84. Woolfovým testom overíme, že takýto dramatický rozdiel medzi odds ratios nie je náhoda (p-val$\sim 10^{-2}$). Podobne dopadne aj Breslow-Day test (p-val$\sim 10^{-2}$) , ktorý takisto ako ten Woolfov testuje, či sa odds ratios viacerých tabuliek rovnajú. Jeho testovacia štatistika vyzerá ale trošku inak:

$$T = \sum_{ijk} \frac{(O_{ijk} - E_{ijk})^2}{E_{ijk}} \sim \chi^2_{K-1}$$

Keďže testy rovnosti odds ratios boli zamietnuté, nemôžeme použiť ani CMH test na našu 3D tabuľku, aby sme zistili, či hodnoty 0.24 a 0.84 nie sú len štatistickou odchýlkou od hodnoty =1. Samostatné vyhodnotenie D-testu aplikovaného na odds ratio pre staré filmy (0.84) dopadol tak, že hypotézu o rovnosti =1 nezamietol. Záverom teda je, že distribúcia dobrých a zlých starých filmov môže byť homogénna (USA voči iným krajinám).

III. LOGISTICKÁ REGRESIA

V tejto časti sa budem venovať druhej otázke z úvodu: *"Je možné predpovedať verejné hodnotenie filmu ešte predtým, než film ktokoľvek uvidí?"*

Vytvoríme model logistickej regresie, ktorý bude predpovedať pravdepodobnosť p_x, s akou sa film s danými parametrami zaradí do kvalitných (*"good"* $\mapsto$ 1) alebo nekvalitných (*"bad"* $\mapsto$ 0). Z pripraveného datasetu vyberám len filmy vzniknuté v USA, pretože hodnoty *rozpočtov* v tabuľke sú uvedené vždy v lokálnej mene, s čím by sa ťažko pracovalo, ak by sme pracovali aj s krajinami inými, ako USA.

Medzi parametre som zahrnul:

- x_1 = *Rozpočet*: Škálujem na násobky miliónov dolárov.
- x_2 = *Dĺžka filmu*: Hodnoty posunuté, aby nula zodpovedala dĺžky filmu 2 hodiny
- x_3 = *FB likes hlavného herca*: Ide o počet lajkov, ktoré herec získali od FB užívateľov na stránke IMDB. Škálujem na násobky tisícov.
- x_4 = *Farebnosť*: 0 alebo 1 podľa toho, či je film čiernobiely alebo farebný.

Model teda vyzerá nasledovne:

$$\ln\left(\frac{p_x}{1-p_x}\right) \sim \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4\,,$$

kde vektor $\vec{\beta}$ reprezentuje silu väzby parametra na výsledok (nemá jasnú reprezentáciu). Po implementácii do Rka zisťujeme, že:

$$e^{\hat{\beta_0}} \doteq 14.8\,,\quad e^{\hat{\beta_1}} \doteq 1.0\,,\quad e^{\hat{\beta_2}} \doteq 1.1\,,\quad e^{\hat{\beta_3}} \doteq 1.1\,,\quad e^{\hat{\beta_4}} \doteq 0.1$$

Prvé číslo vyjadruje odds ratio, že film bude kvalitný pri referenčných hodnotách (budget=0, duration=2h, FBlikes=0, color=0). Zvyšné čísla sú odds ratios situácií, kedy zvýšime hodnotu i-teho parametra o jedna. Špeciálne posledný odds ratio hovorí o znížení šance filmu byť kvalitným pri prepnutí z čiernobieleho prevedenia na farebné.

Prvý test, ktorý robím je *null test*, pri ktorom sa pýtame, či sú vôbec nejaké parametre v modeli potrebné. Pearsonov χ^2 test rozdielu deviancií (viz. tabuľka nižšie) ukazuje (p-val$\to$ 0), že takáto hypotéza je extrémna.

Druhým testom je test na relevanciu parametra *rozpočet*. Odds ratio tohto parametra vyšlo v pôvodnom modeli blízke =1, čiže šanca, aby bol film kvalitný, sa len málo mení so vzrastajúcim rozpočtom. Pearsonov χ^2 test rozdielu deviancií pôvodného modelu a takéhoto submodelu *Sub1* nám v tomto prípade ukazuje (p-val$\doteq$ 0.92) , že dôležitosť parametra *Rozpočet* v modeli je signifikantne nízka.

Model dev.	**Null dev.**	**Sub1 dev.**	**Sub2 dev.**
401.3	686.6	404.4	418.3

Tretí test robím na relevanciu parametra *farebnosť*, kedy (neprekvapivo) Pearsonov χ^2 test rozdielu deviancií pôvodného modelu a takéhoto submodelu *Sub2* jednoznačne zamieta (p-val$\sim 10^{-5}$) hypotézu o tom, že vzhľadom na hodnotenie filmu je farebnosť zbytočná. Pomocou Waldovho testu vieme ukázať aj 95%-ný interval spoľahlivosti odds ratio pre *farebnosť*:

$$e^{\hat{\beta}_4} \in (0.02, 0.32)$$

Spýtajme sa teraz, ako by podľa tohto modelu dopadol môj amatérsky film, do ktorého by som neinvestoval žiadne peniaze, trval by hodinu a pol, zohnal by som herca s 20,000 FB likes a bol by to farebný film. Pravdepodobnosť, že bude kvalitný je:

$$p_{\text{moj}} = \text{logit}((1, 0, -30, 20, 1) * \hat{\beta}) \approx 24.9\%$$

Výsledok nie je ani prekvapivý ani potešujúci. Interval 95% spoľahlivosti pre tento výsledok je:

$$p_{\text{moj}} \in (15.9\%, 36.9\%)$$

Vyberme si teraz množinu filmov, na ktorej náš model otestujeme a tým vyhodnotíme jeho kvalitu. Do testovacej vzorky *dobrých* filmov vyberám zhruba 100 filmov pod TOP 250 a zhruba 100 filmov nad BOTTOM 250 (viz. FIG. 1). Validácia modelu dopadla takto:

		Skutočnosť	
		"bad"	"good"
Predpoveď	"bad"	96	40
	"good"	23	92

Pomocou dát z validačnej tabuľky vieme nakoniec určiť sensitivitu (úspešnosť klasifikovať dobrý film), specificitu (úspešnosť klasifikovať zlý film) a presnosť (úspešnosť klasifikovať pravdivo):

$$\text{sensitivity} = \frac{92}{92 + 40} \doteq 69.7\%$$

$$\text{specificity} = \frac{96}{96 + 23} \doteq 80.7\%$$

$$\text{accuracy} = \frac{96 + 92}{\Sigma} \doteq 74.9\%$$

Záverom treba dodať, že náš logistický model je určite veľmi nekvalitný, pretože najsilnejší parameter je pre neho *dĺžka filmu*, čo je v silnom rozpore s intuíciou.

Vlastný názor: *Podľa mňa je to spôsobené tým, že dlhé filmy zvyknú byť intelektuálne náročnejšie a krátke filmy slaboduchými komédiami, preto sa zvyknú známkovať vysoko resp. nízko :-). Pravdepodobne tu zohral úlohu aj Simpsonov efekt a v serióznejšej analýze by bolo potrebné zahrnúť do modelu faktorový parameter žánru filmu, pretože ako zo skúsenosti vieme, na tom veľmi záleží.*

IV. BOOTSTRAP

V tejto časti sa pokúsim zodpovedať na tretiu otázku z úvodu: *Ako veľmi si môžeme byť istý kvalitou kanadských filmov?*

Pozrime sa najskôr na histogram hodnotení 62 kanadských filmov:

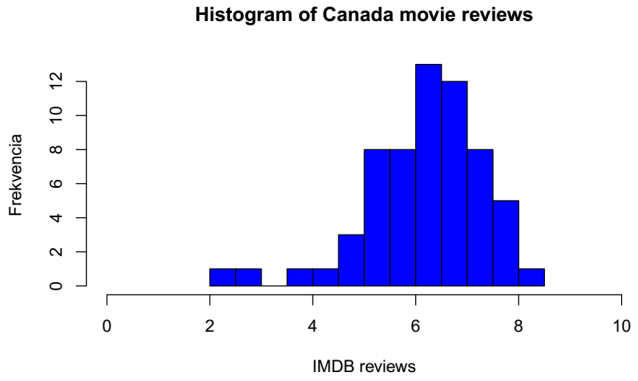

FIG. 4.

Medián takéhoto rozdelenia je rovný známke $\hat{m} = 6.2$. Aby sme mali väčší prehľad o tom, akú chybu môže mať takýto medián, vyrobíme si $N = 10,000$ ďalších datasetov kanadských známok rovnakej veľkosti, pričom dáta do nich náhodne povyberáme (aj s opakovaním) z pôvodného datasetu. Histogram mediánov potom vyzerá nasledovne:

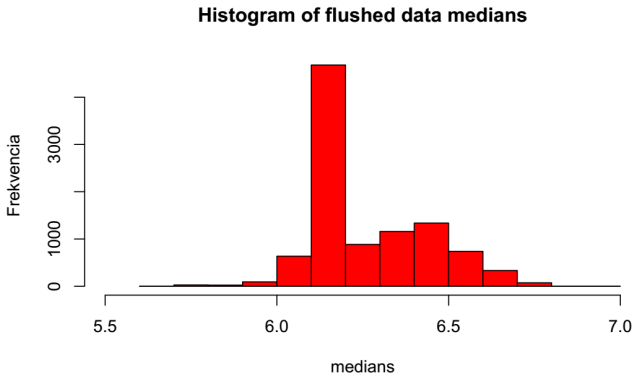

FIG. 5.

Priemerná hodnota všetkých nových mediánov je $\langle m^* \rangle \doteq 6.3$ a disperzia $\sigma^2 \doteq 0.03$. Teda 95%-ný interval spoľahlivosti pre medián hodnotení kanadských filmov je so zohľadnením vychýlenia dát:

$$m = \hat{m} - (\langle m^* \rangle - \hat{m}) \pm u_{\text{norm}}(0.975)\sigma = 6.1 \pm 0.3$$

Pre zaujímavosť, rovnaká analýza pre 3005 USA filmov dopadne s vyšším mediánom a menšou odchýlkou:

$$m_{\text{USA}} = 6.5 \pm 0.1$$