SLOVENSKÁ TECHNICKÁ UNIVERZITA V BRATISLAVE

Fakulta informatiky a informačných technológii

Projekt 1:

Vplyv dĺžky filmu na jeho hodnotenie

Matej Skyčák a Tomáš Socha

Akademický rok 2021/2022

Existuje prepojenie medzi dĺžkou filmu a jeho obľúbenosťou?

Určite ste si niekedy všimli, že kultové filmy ako napríklad Schindlerov list (195 minút), Titanic (194 minút) alebo klasika zo života talianskej mafiánskej rodiny Krstný otec (175 minút) sú pomerne dlhé a trvajú okolo troch hodín prípadne ešte viac.

My sme sa v tomto projekte zamerali hlavne na tento atribút filmov, a teda zisťovali sme, či existuje nejaká súvislosť medzi dĺžkou filmu a jeho obľúbenosťou medzi ľuďmi.

Dáta, s ktorými sme pracovali

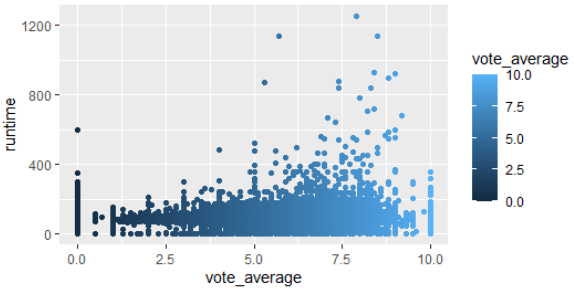
Pracovali sme s množinou dát s názvom „The Movies Dataset“, ktorý je v podstate zbierkou 45000 filmov spolu s ich 24 atribútmi zo stránky TMBD (The Movie Database,zdroj: https://www.kaggle.com/rounakbanik/the-movies-dataset?select=movies_metadata.csv). Z týchto atribútov sme nevyužili všetky, ale zamerali sme sa predovšetkým na dĺžku filmu (runtime) a priemerne hodnotenie od používateľov stránky TMBD (vote_average).

Pôvodný dataset bol síce obšírny, ale mnoho filmov sme museli vyfiltrovať, pretože príliš vybočovali z priemeru alebo nemali pre naše skúmanie hodnotu. Niektoré filmy nemali žiadne hodnotenia, alebo ich mali príliš málo, čo by mohlo prípadné prepojenie skresľovať. Preto sme stanovili ako minimum 100 recenzií, čo nie je zrovna veľa, ale dáta nie sú z až tak populárnej stránky a tak sa množstvo recenzií nepohybuje v státisícoch prípadne miliónoch. Filmy s najviac recenziami majú len niečo cez 10000 hodnotení.

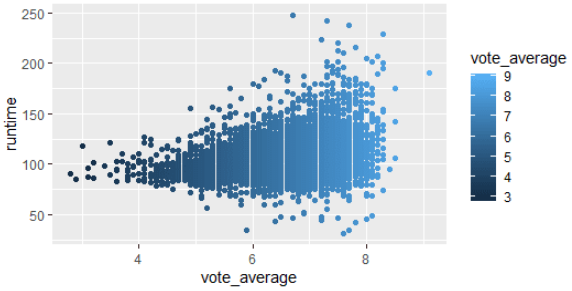
Taktiež sme vyfiltrovali aj filmy, ktoré boli až príliš dlhé a neboli ani tak filmy, ale skôr mini-série. Práve z tohto dôvodu sme dali horný limit pre dĺžku filmu 250 minút. Na druhú stranu sa v množine dát, ale objavovali aj takzvané krátke filmy, ktoré boli zvyčajne nejaké študentské projekty alebo filmy, ktoré sa prehrávali jedine ako predohry ku skutočným celovečerným filmom. Preto spodný limit bol stanovený na 30 minút, keďže veľa filmov zo začiatkov minulého storočia bolo skutočne krátkych a nechceli sme ich vynechať.

Po finálnom vyfiltrovaní sme zostali na finálnom počte, niečo málo pod 6000 filmov.

Na obrázkoch nižšie si môžete všimnúť bodové grafy znázorňujúce rozloženie jednotlivých filmov v závislosti od ich priemerného hodnotenia (os x) a dĺžky (os y) v porovnaní s ostatnými pred a po vyfiltrovaní.



Obrázok 1 Pôvodný dataset



Obrázok 2 Upravený dataset

Overenie hypotézy

Vzťah medzi dĺžkou filmu a jeho hodnotenia sme skúmali pomocou **regresnej analýzy** v programovacom jazyku R. Za prediktor sme si zvolili atribút predstavujúci priemerné hodnotenie (vote_average) a premenná závislá od prediktora (response), bola dĺžka filmu (runtime).

Na vyfiltrované dáta sme aplikovali funkciu *lm()* na vytvorenie lineárneho regresného modelu, pomocou ktorej sme zistili koeficienty úrovňovú konštantu (intercept), sklon krivky (slope), ich štandardnú **odchýlku a reziduály**.

Následne sme dosiahnuté výsledky potvrdili pomocou **krížovej validácie** vstupných dát. Vytvorených bolo 50 podmnožín, kde každý obsahoval náhodnú vzorku 50% z celej upravenej množiny dát. Tieto hodnoty 50 podmnožín a 50% množiny dát sme zvolili z dôvodu, aby si podmnožiny neboli moc podobné, prípadne, aby sa podmnožiny a prvky v nich neopakovali. Zároveň ich potrebujeme dostatok, aby boli zastúpené rôzne hodnoty množiny dát. Tento krok bol realizovaný pomocou funkcie *sample*(), ktorá zabezpečila náhodné vybratie filmov a funkcionálu *map()*, ktorý zabezpečil najskôr vytvorenie prázdneho listu a taktiež jeho zaplnenie listami predstavujúce jednotlivé podmnožiny. Listu listov bol priradený názov *data*.

Na každý subset bol následne napasovaný lineárny **regresný model**, na základe ktorého boli zozbierané koeficienty a odchýlky. Bolo to realizované pomocou funkcionálu *map()* zavolaním funkcie *lm()* pre každú podmnožinu. Zistené hodnoty boli uložené do listu listov, ktorému sa pridelil názov *models*.

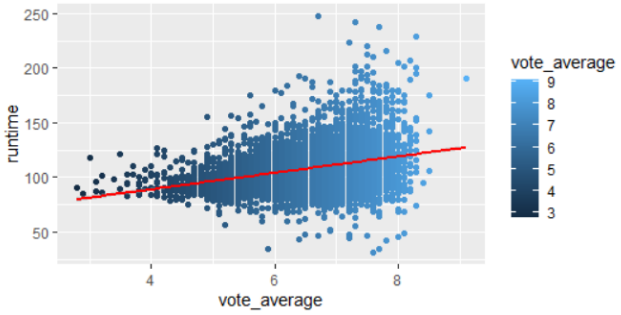
Ďalej sa vypočítala štandardná chyba koeficientov pomocou funkcie *sd()*. V nej sme použili funkcionál *map_dbl()*, ktorý vráti double vektory ,ktoré sú potrebné ako vstup do funkcie. Štandardná chyba je vypočítaná najprv pre úrovňovú konštantu a následne aj pre sklon krivky.

Taktiež sa zistila štandardná chyba reziduálov. Najskôr však bolo potrebné zistiť štandardnú chybu štvorcov vzdialeností residuálov pre jednotlivé subsety. Aj tu bol použitý funkcionál *map_dbl()* a štandardný vzorec na výpočet sumy štvorcov vzdialeností reziduálov. Získané hodnoty boli uložené do vektora typu double s názvom ***RSS***. Následne sa nad každou hodnotou vykonala funkcia, ktorá vypočítala štandardnú chybu pre jeden reziduál. Funkcia vypočítala odmocninu z vyššie získanej hodnoty predelenej veľkosťou danej podmnožiny, ktorej patrila. Pre zabezpečenie vykonania tejto operácie pre všetky podmnožiny bol znovu použitý funkcionál *map_dbl()*. Výsledné hodnoty boli uložené ako vektor typu double pod názvom *rse* (skratka pre residual standard error).

Jednym z krokov krížovej validácie bola extrakcia β koeficientov a reziduálov z každého modelu. Toto bolo zabezpečené pomocou 3 príkazov. V prvom bol vytvorený list s názvom *listoffunctions,* ktorý obsahoval funkcie *coef()* a *residuals(*). Tento list bol vytvorený s dôvodu umožnenia použitia funkcionálu *map()* čo zabezpečilo zjednodušenie a sprehľadnenie kódu. Pomocou druhého príkazu sa vytvorila pomocná funkcia *f()*, ktorej úlohou bolo namapovať pre každý vstup všetky funkcie z premennej *listoffunctions*. A posledný príkaz sa použil vytvorenie nového listu *extracteddata,* do ktorého boli uložené všetky koeficienty a reziduály pre každú vygenerovanú podmnožinu. Túto premennú však nebolo nutné použiť, pretože sme štandardnú chybu reziduálov zistili vďaka vyššie opísanému postupu pomocou funkcionálu *map_dbl()*.

Dosiahnuté výsledky

Aplikovaním lineárneho regresného modelu na vyfiltrovaný dataset sme dostali hodnoty **koeficientov** rovnajúce sa: $\beta_0 = 58.738$ (intercept) a $\beta_1 = 7.525$ (slope), čo značilo stúpajúcu úmernosť medzi nami skúmanými atribútmi.

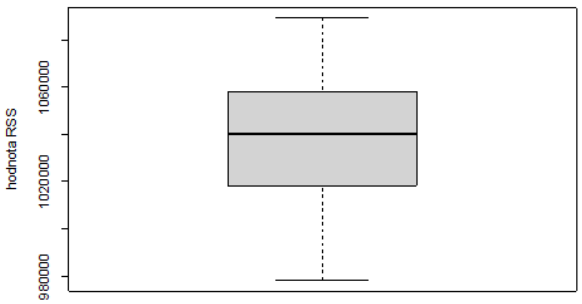


Obrázok 3 Aplikovanie lineárneho regresného modelu na vyfiltrovanú množinu dát

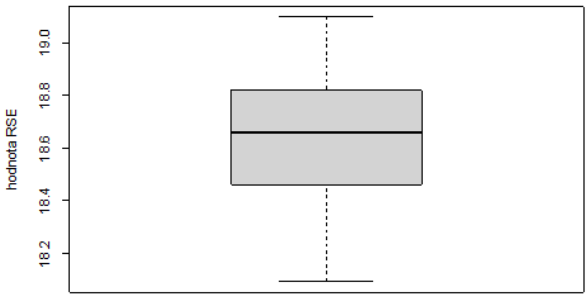
Štandardná chyba jednotlivých koeficientov predstavovali hodnoty pre $\beta_0 = 1.8243$ a pre $\beta_1 = 0.2819$, čo predstavovalo primerané čísla vzhľadom na veľkosti koeficientov. Veľkosť štandardnej chyby reziduálu sa zastavila na čísle 18.63. Na základe tohto môžeme prehlásiť dataset ako dobrú reprezentatívnu vzorku.

Namerané hodnoty bolo ešte potrebné potvrdiť **krížovou validáciou** (cross-validation), pomocou ktorého sme zistili, ako moc budú analyzovaný model ovplyvňovať nezávisle vybrané vzorky dát. Hodnoty štandardných chýb namerané spriemerovaním týchto hodnôt každej z 50 podmnožín predstavovali hodnoty pre koeficienty $\beta_0 = 1.735$ a pre $\beta_1 = 0.2890$. Ako si môžete všimnúť, namerané čísla chýb pomocou krížovej validácie sa líšia len v desatinách, respektíve stotinách, čo je základom k dobrej stabilite modelu.

Ďalším krokom v našom projekte bolo vypočítanie **RSS** (Residual sum of squares), teda súčet štvorcov reziduálov, ktorý sme ďalej použili na výpočet **RSE** (Residual standard error). Naľavo na obrázku je boxplot s hodnotami RSS jednotlivých náhodne vybraných podmnožín. Obrázok napravo znázorňuje druhý boxplot, pre výpočet ktorého sme použili hodnoty RSS. Na tomto sú zachytené hodnoty RSE, čiže priemery štandardných chýb reziduálov namerané z jednotlivých ľubovoľných podmnožín.



Obrázok 4 Súčet štvorcov reziduálov náhodne vybraných podmnožín



Obrázok 5 Priemerná štandardná chyba náhodne vybraných podmnožín

Z boxplotu napravo vieme vyčítať **medián** priemernej štandardnej chyby, ten je približne na hodnote 18.6 minút. Teda z toho vieme povedať, že dokážeme dĺžku filmu na základe jeho hodnotenia od používateľov predpovedať s presnosťou okolo 18.6 minút. Čo nemusí znieť až

tak presne, ale pravdepodobne to je do veľkej miery spôsobené subjektívnym názorom používateľov, ktorý vychádza z osobného dojmu a ten je samozrejme veľmi ťažké predpovedať.

Taktiež vieme vyčítať, že maximálna priemerná chyba náhodne vybraných podmnožín pôvodnej množiny dát je približne 19.1 minút a naopak minimálna priemerná chyba je niekde v okolí 18.1 minút. **Hraničné hodnoty** sa taktiež líšia len mierne, čo je prirodzené pri zvolení rozdielnej podmnožiny vstupných dát.

Na základe rozptylu jednotlivých kvartálov boxplotu možno povedať, že nevieme jednoznačne určiť priemernú štandardnú chybu. Čím by bol rozptyl menší (jednotlivé priemery štandardných chýb by boli viacej zoskupené), s tým väčšou presnosťou by sme vedeli určiť jej hodnotu.

Zhrnutie

Na začiatku sme si vybrali množinu dát zo stránky s databázou filmov a pokúsili o overenie nami stanovenej hypotézy “Existuje nejaké prepojenie medzi dĺžkou filmu a jeho obľúbenosťou?”. Už ako z hypotézy vyplýva zamerali sme sa na dva atribúty, dĺžku filmu a jeho priemerné hodnotenie. Následne sme aplikovali na množinu dát lineárny regresný model, sme zistili, že isté prepojenie medzi dvomi atribútmi existuje.

Taktiež sme spozorovali, že medzi týmito atribútmi existuje väzba, čiže sa tam nachádza istá miera **korelácie**, pretože oba približne lineárne rastú. V prípade **kauzality** to nie je také jednoznačné. Neprišli sme na jasné logické vysvetlenie, prečo by mala dĺžka filmu závisieť priamo od jeho obľúbenosti. Možno keby sme sledovali aj rozpočet, tak by sme to vedeli logicky lepšie odôvodniť.

Ďalej sme model overili pomocou krížovej validácie. Zo získaných údajov sme spozorovali, že išlo o reprezentatívnu vzorku. Avšak model nemožno označiť za úplne stabilný hlavne kvôli vyšším hodnotám štandardnej chyby reziduálov. Čo je celkom logické, keďže predpovedať presne dĺžku filmu na základe subjektívneho názoru používateľov, znie dosť nereálne. Ale na otázku “Koľko minút by mal mať film s konkrétnym hodnotením?”, vieme odpovedať s presnosťou priemerne 18 minút.