

Список літератури

1. Ситуація з розповсюдження вірусу в реальному часі (дані університету Джонса Хопкінса) (англ.). Дата звернення 19 квітня 2020. (англ.)
2. Набір даних Novel Corona Virus 2019 Dataset [Електронний ресурс]. Режим доступу – <https://www.kaggle.com/sudalairajkumar/novel-corona-virus-2019-dataset#novel-corona-virus-2019-dataset>
3. NV.UA [Електронний ресурс]. Режим доступу – <https://nv.ua/ukr/world/countries/koronavirus-v-bilorusi-skilki-zarazhenih-i-tempi-poshirennya-ostanni-novini-50085697.html>
4. Humdata.org [Електронний ресурс]. Режим доступу – <https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>

УДК 004.82: 004:85

*Сніжинський М. В., студент 2 курсу спеціальності 122 «Комп'ютерні науки»
Нескородєва Т. В., к.т.н., доцент, доцент кафедри інформаційних технологій*

АНАЛІЗ ДАНИХ ПРО РЕЙТИНГ ФІЛЬМІВ НА ПЛАТФОРМІ ОЦІНЮВАННЯ IMDB ЗАСОБАМИ МОВИ R

Донецький національний університет імені Василя Стуса, м. Вінниця

У сучасному світі жодна людина не уявляє свого життя без перегляду фільмів, особливо зараз це заняття стало популярним у зв'язку із ситуацією у світі. При виборі фільму для перегляду люди часто користуються основним критерієм – рейтингом. Рейтинг у фільмів буває двох видів: від глядачів та від критиків, людей, які професійно аналізують та розбираються в даній сфері. Але як залежить рейтинг фільму глядачів від рейтингу критиків, чи є між ними схожість? Чи впливає рейтинг на популярність та успіх фільму? Дати відповідь на це питання ми зможемо провівши дослідження на основі даного набору даних.

Його актуальність обумовлена наступними факторами: зростання ринку фільмів, необхідність визначення факторів, що суттєво і несуттєво впливають на популярність фільмів. Аналіз проведемо за допомогою пакету R. Для дослідження був використаний набір даних IMDB Movie Data, що містить 1000 спостережень по 10 показниках:

1. Rank – порядок ранжування фільму.
2. Title – Назва фільму.
3. Genre – Жанр фільму.
4. Description – короткий опис фільму.
5. Director – ім'я режисера фільму.
6. Actors – Список основних зірок фільму, розділений комою.
7. Year – рік виходу фільму.
8. Runtime – Тривалість фільму в хвилинах.
9. Rating – Рейтинг фільму 0-10 (від глядачів).
10. Votes – Кількість голосів.

11. Revenue(Millions) – Дохід від фільмів у мільйонах.

12. Metascore – Сукупне середнє значення балів критиків. Значення становлять від 0 до 100. Більш високі бали представляють позитивні відгуки.

Побудувавши графік залежності рейтингу глядачів та рейтингу від критиків (рис.1), ми зробили висновок, що оцінки фільмів, які знаходяться вище регресійної прямої є більш об'єктивними та вказують на більшу популярність фільмів. Видно, що в цілому оцінки від глядачів та оцінки від критиків співпадають. Ми також можемо побачити, що більш популярні фільми мають оцінки більше 5 серед глядачів та більше 50 у критиків.

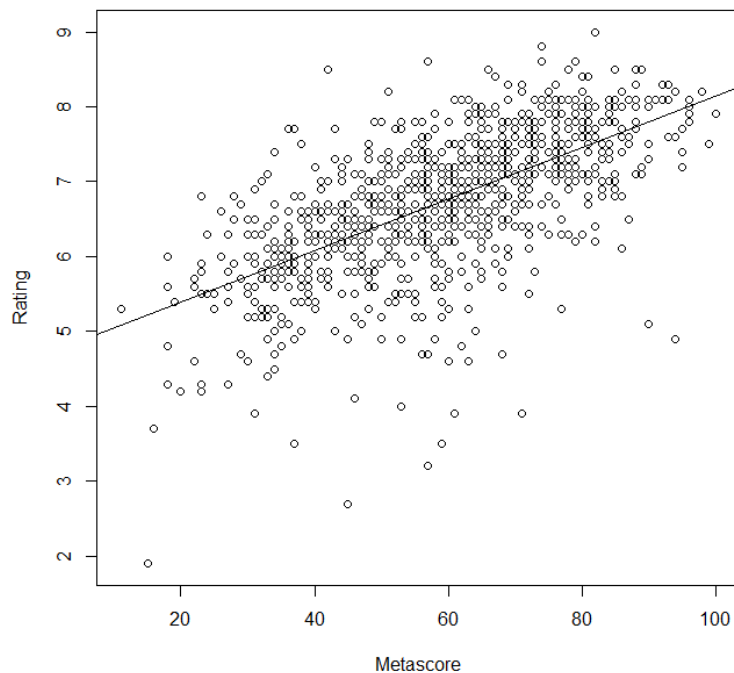


Рисунок 1 – Залежність рейтингу фільмів глядачів від рейтингу фільмів критиків

Щоб з'ясувати варіацію рейтингу фільмів глядачів від рейтингу критиків, побудуємо багатофакторну регресійну модель, у якій предиктором буде виступати Rating (рейтинг глядачів). В якості відгука – Metascore.

На рисунку 2 відображений звіт даної регресійної моделі.

Логістична модель дає змогу передбачити популярність фільму від оцінок. Параметр дисперсії для прийнятої сімейства гаусових рівний 0.5259682. Також звернемо увагу що нульове відхилення більше за залишкове приблизно у 1,6 рази. Також AIC:2058,9.

Використавши predict() видно, що фільми які будуть отримувати оцінки вище 5 серед глядачів та вище 50 серед критиків будуть користуватися більшою популярністю.

```

> summary(glm.fit)

Call:
glm(formula = Rating ~ Metascore)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5485  -0.3586   0.0813   0.4460   2.3546

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.701914   0.084745   55.48  <2e-16 ***
Metascore    0.034369   0.001379   24.92  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.5259682)

Null deviance: 817.79  on 935  degrees of freedom
Residual deviance: 491.25  on 934  degrees of freedom
(64 observations deleted due to missingness)
AIC: 2058.9

Number of Fisher Scoring iterations: 2

```

Рисунок 2 – Звіт підгонки регресійної моделі

Зі звіту можемо побачити, що у більшості випадків оцінки фільмів глядачів співпадають із оцінками від критиків. Також видно, що якщо оцінка від глядачів більша 5 та оцінка від критиків більша 50, то це гарантує популярність фільму.

Отже, побудована логістична модель визначила залежність якості оцінок і показала їх прямий вплив на популярність фільмів. Тобто, якщо фільм має оцінки більше 5 від глядачів та більше 50 від критиків, то це гарантує успіх фільму. За допомогою пакету R були побудовані графіки та дані, що яскраво ілюструють дану залежність.

Список літератури

1. *IMDB Movie Rating*: <https://clck.ru/NAG6X>
2. *Набір даних IMDB Movie Data [Електронний ресурс]. Режим доступу – <https://www.kaggle.com/PromptCloudHQ/imdb-data>*
3. Джеймс Г., Уиттон Д., Тибиширани Р. *Введение в статистическое обучение с примерами R. Изд. Второе, испр. Пер с англ. С.Э. Мاستицкого – М. ДМК Пресс, 2017. -456с.*

УДК 004.82: 004:85

*Сорока О. С., студент 2 курсу спеціальності 122 «Комп'ютерні науки»
Нескородєва Т. В., к.т.н., доцент, доцент кафедра інформаційних технологій*

АНАЛІЗ ДАНИХ ПРО РІВЕНЬ КРИМІНАЛЬНОСТІ В ОКРЕМИХ МІСТАХ УКРАЇНИ

Донецький національний університет імені Василя Стуса

Рівень злочинності в Україні є гострим питанням, так як цей показник впливає на добробут та злагодженість роботи держави. Високий рівень створює