|  |  |
| --- | --- |
| פרויקט גמר | תקציר  האם ניתן למצוא דרך לנבא או לחזות את הדירוג של סרט באתר www.IMDB.com  Dror nir 303159834  BIG-DATA |

**מבוא:**

אתר IMDB הינו אתר המכיל מידע אודות סרטים, שחקני קולנוע, תכניות טלוויזיה, משחקי וידאו ועובדים בתעשיית הקולנוע והטלוויזיה.

האתר בבעלות חברת אמזון.

באתר אפשר למצוא מידע על סרטים, שחקנים, שם הבמאי, תקציר העלילה, מידע כספי אודות הסרט ועוד דירוגים עבור כל סרט שנעשה ע"י הגולשים והאתר.

IMDB הוא אתר חופשי, אך יש להירשם אליו בחינם על מנת לקבל גישה לכל הנתונים באתר. משתמשים רשומים באתר יכולים לשלוח בדואר אלקטרוני עדכונים ושינויים, לקרוא ולכתוב הודעות בפורומים, וכמו כן לדרג את הסרטים לפי טעמם - דירוג הגורם לסרט להיות פופולרי יותר באתר שהפך בר סמכה כיום בנוגע לסרטים.

באתר יש סולם דירוג, המאפשר למשתמשים רשומים לדרג כל סרט באתר בציון של 1 עד 10. כל משתמש רשאי להצביע לכל סרט פעם אחת בלבד, כאשר הצבעה נוספת תבטל את ההצבעה הקודמת. בשנת 2006 נוספה האפשרות להצביע גם לתוכניות טלוויזיה.

ההצבעות מסוננות ומשוקללות בדרכים שונות על מנת ליצור ממוצע משוקלל. שיטת החישוב:

דירוג משוקלל = X + average imdb score X 7.1

שאלת המחקר:

"האם ישנו מדד להצלחה של סרט?"

קודם כל, מהי הצלחה של סרט?

האם אפשר לצמצם אותה למספר כלשהו כגון רווח בקופות או מספר הלייקים שהסרט קיבל בפייסבוק?

אולי בכלל מדובר בכמות האנשים שנכנסו לאתר ודירגו את הסרט (בלי להסתכל האם הביקורת משבחת או לא).

בכל מקרה, בשביל שיהיה גורם בקרה, החלטתי שדירוג הסרט באתר מהווה קריטריון להצלחתו.

עתה, אסתכל על גורמים אחרים בטבלה ואבחן את מידת הקורלציה בינם לבין הדירוג המקורי של האתר.

הכנת מסד הנתונים:

השתמשתי במסד הנתונים "movie metadata" מאתר [www.kaggle.com](http://www.kaggle.com).

מסד הנתונים הכיל רשומות אודות כ 5000 סרטים כשכל רשומה מכילה מידע על הסרט כפי שנזכר לעיל.

בנוסף, הסתכלות על זוג סרטים משני ז'אנרים שונים לאו דווקא יכולה לייצר לנו השוואה אמינה. לעיתים קרובות, סרט דוקומנטרי לא יופק באותה עלות כמו סרט אקשן וכנראה שכמות הלייקים שהשחקן הראשי בסרט האקשן קיבל תהייה גבוהה מכמות הלייקים שהשחקן הראשי קיבל בסרט הדוקומנטרי.

מהסיבות הנ"ל, לא אסתכל על מסד הנתונים כפי שהוא אלא אחלק אותו לתתי מאגרים לפי הז'אנר של כל סרט.

עמודת הז'אנר בנויה בצורה כך שסרט יכול להיות משויך למספר ז'אנרים.

לכן, גם פה חילקתי את הטבלה המקורית לתתי טבלאות עפ"י הז'אנר,

עבור כל ז'אנר הכנתי מס' טבלאות נוספות:

* קורלציה לפייסבוק - הוספתי עמודה שסכמה את מספר הלייקים שקיבלו השחקנים, הסרט עצמו והבמאי בפייסבוק.
* קורלציה למספר הביקורות שנעשו ע"י גולשים ומשתמשים רשומים באתר - הוספתי עמודה שסכמה את מספר הביקורות שנעשו ע"י משתמשים רשומים באתר ביחד עם מספר הביקורות שנעשו ע"י משתמשים רנדומליים (הספירה מתייחסת למספר הביקורות שנעשו, בלי להתייחס האם הביקורת שלילית או חיובית)
* קורלציה לרווח הנטו (profit) של הסרט - הוספתי עמודה בשם profit ובה את הרווח שעשה הסרט פחות התקציב שלו. ואת העמודה נרמלתי.
* קורלציה לרווח הברוטו של הסרט (gross)
* קורלציה לתקציב של הסרט

לאחר מכן השמטתי שורות כפולות ושורות בהם היה מחסור במידע ( למשל עבור סרט שלשחקן הראשי או לבמאי בו אין עמוד פייסבוק ולכן גם אין לייקים).

**חלק 1 – בדיקת קורלציה בין דירוג IMDB לבין מדדים אחרים מהטבלה:**

הסבר על התהליך:

עבור כל אחת מהטבלאות קורלציה שציינתי:

עבור כל ז'אנר, לקחתי את הטבלת סרטים מהז'אנר הנ"ל.

בכל טבלה ביצעתי שני סינונים – חיסור שורות בהן חסר מידע, חיסור שורות בהם הנתונים עליהם אעשה קורלציה מול דירוג IMDB גבוהים או נמוכים במידה מסוימת[[1]](#footnote-1).

עתה, אם לאחר הסינונים קיבלתי טבלה בעלת יותר מ 100 שורות אז ביצעתי שלושה חישובים:

1. K-means
2. Pca
3. רגרסיה לינארית

* בקורלציית פייסבוק בדקתי את דירוג IMDB מול מספר הלייקים הכולל
* בקורלציית מספר הביקורות בדקתי את דירוג IMDB מול מספר הביקורות הכולל
* בקורלציית התקציב בדקתי את דירוג IMDB מול התקציב של הסרט
* בקורלציית הרווח נטו בדקתי את דירוג IMDB מול נרמול של עמודת profit
* בקורלציית הרווח ברוטו בדקתי את דירוג IMDB מול העמודה gross

תוצאות:

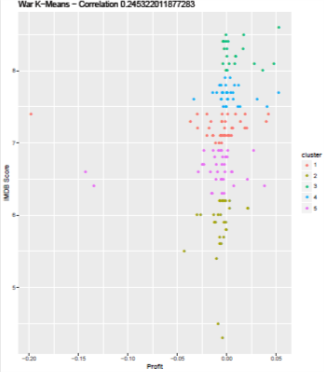
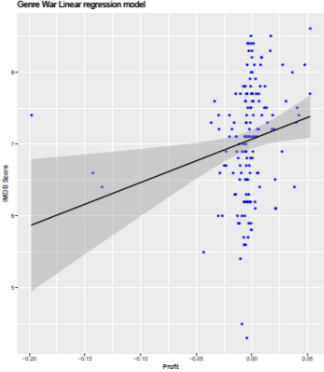
בגרף הבא ניתן לראות את התפלגות הקורלציה בבדיקות שערכתי.

עמודות ירוקות = אם התוצאה התקבלה באלגוריתם k-means לאחר pca.

עמודות אדומות = אם התוצאה התקבלה באלגוריתם k-means.

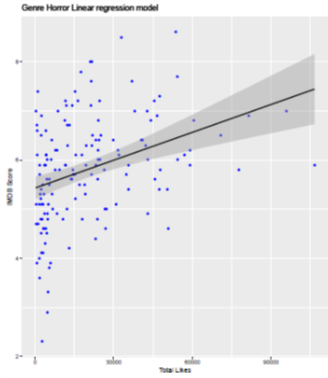
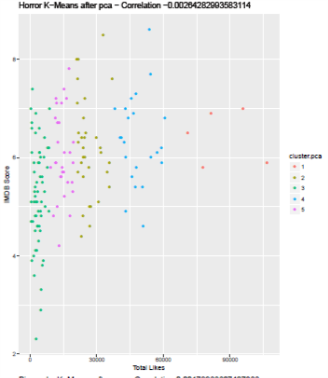
בגרפים הבאים ניתן לראות את הקורלציה המקסימלית שהתקבלה ואת גרף הרגרסיה הלינארית מאותה בדיקה.

הקורלציה התקבלה בבדיקה מול רווח נטו (profit) בז'אנר War



בגרפים הבאים ניתן לראות את גרף הקורלציה המינימלית שהתקבלה ואת גרף הרגרסיה הלינארית מאותה בדיקה.

הקורלציה התקבלה בבדיקה מול מספר הלייקים בפייסבוק בז'אנר Horror



מסקנות:

הרגרסיה הלינארית לרוב מציגה גרף עם שיפוע חיובי,

לפי הרגרסיה הלינארית בבדיקה מול רווח נטו – לרוב אפשר יהיה להסיק שסרט שהרווח שלו גבוה יקבל דירוג IMDB גבוה.

אך הפיזור של כל גרף הוא רחב מה שגורם לקורלציה נמוכה.

מטווח התוצאות שהתקבל קשה להסיק קורלציה מסוימת ולא ניתן לקבוע כי מדד מסוים בטבלה משפיע על דירוג IMDB.

למרות זאת, אציין שברוב הפעמים קל לראות כי אלגוריתם k-means ללא pca נתן קורלציה גבוהה יותר מאשר עם pca.

וזה בשל החלוקה ששניהם בחרו:

ב k-means לאחר pca החלוקה למרכזי כובד אנכיים,

ב k-means לפני pca החלוקה למרכזי כובד אנכיים ואופקיים,

יכול להיות שהחלוקה של השני נכונה יותר ולכן לרוב נקבל שמה קורלציה גבוהה יותר,

אך התוצאה הנמוכה מלכתחילה מקשה על הסקת המסקנות.

**חלק 2 – ניסיון לנבא את דירוג IMDB:**

הסבר על התהליך:

את עמודת דירוג הIMDB ערכתי בצורה הבאה:

דירוג IMDB < 5.75 == 1

5.75 < דירוג IMDB < 6.5 == 2

6.5 < דירוג IMDB < 7.5 == 3

7.5 < דירוג IMDB == 4

החלוקה אפשרה לקבץ סרטים לתתי קבוצות בעלות אותו הגודל (~300-400 סרטים בכל אחת).

גם פה השתמשתי במסד הנתונים המחולק (לפי ז'אנרים). כאשר כל טבלה מחולקת ל - Training(60%) ול - Testing(40%).

במידה ובטבלה (טבלה המכילה סרטים מאותו ז'אנר כאשר עמודת דירוג הIMDB שלה כפי שתיארתי לעיל) היו לפחות 36 סרטים[[2]](#footnote-2) , אז

ביצעתי עבורה מודל Random forest ומודל Decision tree .

מתוך שתי תוצאות המודלים, לקחתי את המודל שתוצאת החיזוי שלו גבוהה יותר ואותה השוויתי לתוצאת המודל שערכתי לאחר מכן,

אותו מודל שהתחיל לאחר מכן רץ על העמודות שמכילות את מספר הלייקים ,מספר הביקורות ,תקציב הסרט, רווח נטו ורווח ברוטו

ובודק האם שתיים מהן יכולה לספק לי מודל מדויק יותר.

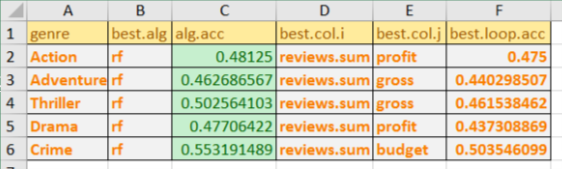
תוצאות:

בטבלה הבאה ניתן לראות סיקור כללי על תוצאות חלק 2 בעבודה:

* בעמודה "best.alg" המתארת את הניבוי המדויק יותר:

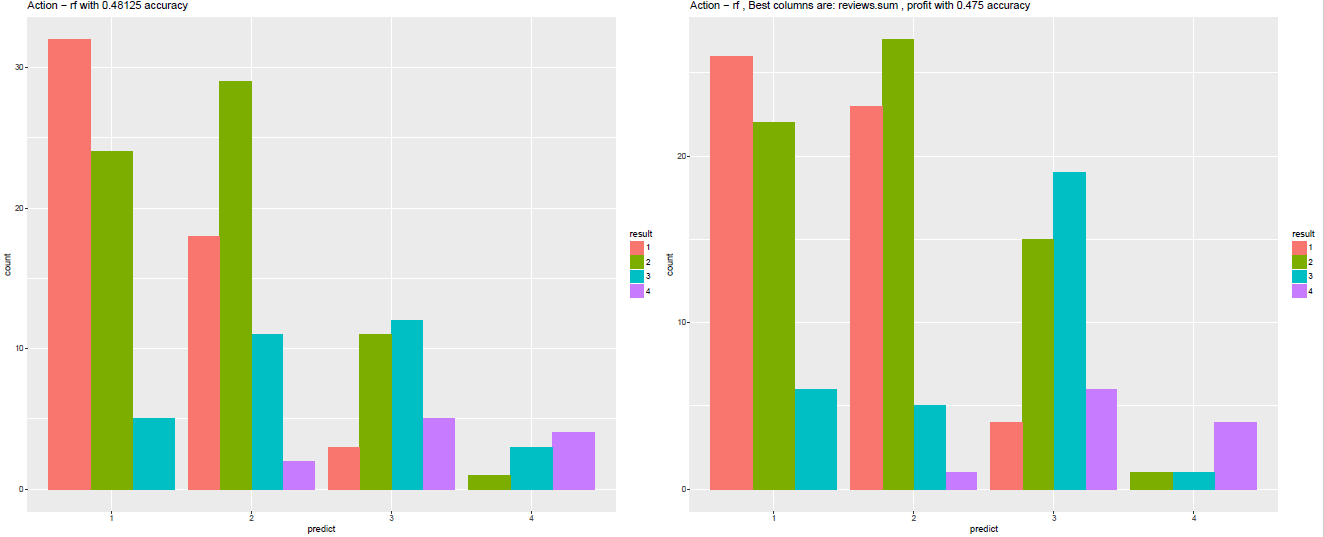
"rf" אומר שהמודל המדויק יותר היה Random forest.

* בעמודות "best.col.i" ו- "best.col.j":

הכוונה היא לעמודת ה-i ועמודת ה-j שנתנו את הניבוי הכי מדויק.

* התאים שמסומנים בירוק הגיעו לדיוק גבוהה יותר במודלים (אלגוריתמים Random forest ו-decision tree אל מול ניבוי מעמודות אחרות בטבלה).
* קשר ברור בין מספר הביקורות מול מדדים אחרים, ניתן לראות שהוא מנבא את דירוג IMDB בדיוק של 46% בממוצע.
* לעיתים בהרצת הפרויקט היה ניתן לראות שאלגוריתם Decision tree מנבא יותר טוב מאלגוריתם Random forest אך לרוב האלגוריתם השני נתן ניבוי יותר מדויק.

בגרפים הבאים ניתן לראות את הניבוי המדויק יותר מהאלגוריתמים Decision tree ו- Random forestמול ניבוי מעמודות אחרות בטבלה

* את כל הגרפים ניתן לראות בקובץ "part2\_plots.pdf",אך נסתכל על שתיים לדוגמא מהם:
* הגרף הימני מציג את הניבוי הכי מדויק שהושג מהעמודות של מספר הביקורות באתר והרווח נטו (reviews.sum,profit).
* הגרף השמאלי מציג את הניבוי שהתקבל ע"י אלגוריתם Random forest.
* שני הניבויים מתייחסים לז'אנר Action.
* אפשר לראות כי לא היו תוצאות בניבויים שעבור דירוג IMDB של 4, נתנו דירוג IMDB של 1.

כמו כן, לא היו תוצאות בניבויים שעבור דירוג IMDB של 1, נתנו דירוג IMDB של 4.

* בכל הבדיקות בהן חיפשתי עמודות אחרות שייתנו ניבוי מוצלח יותר – העמודה שמכילה את מספר הביקורות באתר הייתה ב100% מהתוצאות.

כלומר, ייתכן וקיים קשר בין דירוג IMDB לבין מספר הביקורות של גולשים באתר IMDB.

**מסקנות וסיכום:**

חלק 1:

בניסיון למצוא קורלציה בין מדדים בטבלה לבין דירוג IMDB בעזרת אלגוריתמים k-means ו-k-means לאחר pca לא נמצאה קורלציה גבוהה מ 24% אפשר לראות זאת בגרף של רווח נטו בז'אנר סרטי מלחמה תחת k-means.

למרות שחלוקת מסד הנתונים המקורי לז'אנרים העלה את הקורלציה (לפני החלוקה לתתי טבלאות לא נמצאה קורלציה מעל 9%),

עדיין לא ניתן להעריך את מדד IMDB במידה מספקת.

מכך עולה כי למרות שהסתכלתי על כל קבוצת סרטים מאותה הז'אנר ולמרות שפעולת האלגוריתמים חילקה את הסרטים לקבוצות (מרכזים) עפ"י המדדים:

* מספר הלייקים בפייסבוק "facebook.crr.pdf"
* מספר הביקורות באתר "reviews.corr.pdf"
* תקציב הסרט "budget.corr.pdf"
* רווח נטו (profit) "profit.corr.pdf"
* רווח ברוטו של הסרט(gross) "gross.corr.pdf"

לא היה ניתן להעריך את דירוג הסרט באתר IMDB לפי אותם אלגוריתמים.

בכל זאת, אלגוריתם k-means נתן קורלציה גבוהה יותר לפני שימוש באלגוריתם pca מאשר אחרי השימוש בו.

את הקורלציה הגבוהה יותר אפשר לראות כמעט בכל החישובים וייתכן והיא קשורה לעובדה שלפני השימוש באלגוריתם pca המרכזים "מסודרים" בצורה אנכית וגם אופקית ואילו לאחר השימוש המרכזים "מסודרים" בצורה אנכית בלבד.

ייתכן והפעלה של סיווג נוסף על כל טבלה היה נותן לנו קורלציה גבוהה יותר.

כלומר, ייתכן ומספר הלייקים בפייסבוק קשור לכמות הפרסום שסרט משקיע,

או שייתכן ומספר הלייקים אומר שהשחקנים מפורסמים ולכן התקציב של הסרט צריך להיות גבוהה יותר בשביל לממן זאת,

באופן דומה אפשר לקשר את כמות הביקורות באתר לכמות הלייקים שסרט קיבל מאחר וישנו קישור באתר IMDB לעמוד של אתר IMDB בפייסבוק (לחיצה על כפתור שעושה לייק בפייסבוק לאתר IMDB או לעמוד של הסרט בפייסבוק).

ולכן, במידה והיה מופעל עוד סיווג לאחר החלוקה לפי ז'אנרים – יכול להיות שהאלגוריתמים היו "מוצאים" מרכזים מדויקים יותר והקורלציה הייתה גבוהה יותר.

חלק 2:

החלוקה לז'אנרים הפיקה דיוק גבוה יותר (לפני החלוקה הוא עמד על כ17%).

לאחר החלוקה של דירוג IMDB לערכים חדשים (1-4 לפי החלוקה "בחלק 2") תוצאת הניבויים הייתה בממוצע 49% (Random forest, Decision tree) ו-46% (ניבוי דרך בידוד עמודות אחרות בטבלה).

בכל הניבויים בהם נעזרתי בעמודות אחרות מהטבלה, עמודת מספר הביקורות באתר הייתה אחת מהשתיים שנבחרו.

מכך עולה שבמידה ונתון מספר הביקורות של סרט (reviews.sum), ניתן לנבא את הדירוג IMDB שלו בדיוק של כ-40%.

נראה כי יש קשר בין מספר הביקורות באתר IMDB לדירוגו באתר, ייתכן ואנשים שמעלים ביקורות של סרטים באתר נוטים לדרג את הסרטים גם כן.

לכן כאשר מסתכלים על מספר הביקורות אפשר למצוא קשר לדירוג IMDB של סרט מסוים. אך לא ניתן לבסס זאת עפ"י הבדיקות שערכתי אלא על הסקה מהתוצאות ביחד עם שיקול דעת אישי.

בכל הניבויים בהם השתמשתי ב Random forest וב-Decision tree המודל Random forest היה יותר מדויק.

ולכן, במידה ונשתמש באלגוריתם Random forest נוכל לנבא את דירוג IMDB של סרט בדיוק של כ-50%.

סוף דבר:

מניסיון למצוא קורלציה בעזרת k-means לא נותרתי עם התאמה ממשית, אי אפשר להגיד על פי 24% קורלציה כי ישנו קשר ממשי בין אף אחד מהמדדים שבדקתי מול דירוג IMDB.

מהמודל Random forest ניתן להשיג תוצאות גבוהות יותר.

ועדיין, הבדיקה הכי טובה שהושגה עמדה על ניבוי של 50%.

לבסוף - ייתכן ומספר הביקורות באתר נגרר ממידת השפעה שסרט מסוים יצר, לכן ניתן לראות שבמידה וסרט קיבל הרבה ביקורות – לרוב הוא גם יקבל דירוג IMDB גבוה יותר. זאת בהנחה שדירוג IMDB של סרט הוא אכן מדד להצלחתו.

יכול להיות שההחלטה על דירוג IMDB בתור מדד להצלחה של סרט היא טעות.

ואם הייתי מחליט שרווח של סרט או מידת העניין שסרט מסוים יצר (מספר הביקורות + מספר הלייקים הכולל) הם המדדים להצלחתו אז התוצאה הייתה גבוהה יותר.

1. בהסתכלות על העמודה איתה נבחן קורלציה מול דירוג IMDB,

   חילקתי את טווח הערכים של אותה עמודה ל20,

   אם נתון מסוים נמצא במקטע לבדו – נקטין את הטווח הכולל כך שלא יכיל את אותו מקטע.

   בצורה הזו יצא שהתעלמתי מהנקודות ש"ברחו" משאר הנקודות (outliers) וקיבלתי מדגם מצומצם יותר אך מדויק יותר. [↑](#footnote-ref-1)
2. במידה והיו פחות סרטים, נוצר מצב בו ישנם ערכי דירוג IMDB (1/2/3/4) שלא היה אף סרט בדירוג הזה והתקבלה תוצאת שגיאה מהמודלים. [↑](#footnote-ref-2)