מטלת סיכום – חלק 2

דוח מסכם: חיזוי דירוג סרטים ב-IMDb

שמות הסטודנטים: כפיר זיסו ויעל בוקריס

תעודות זהות: _____

תאריך הגשה:29.05.2026

---

1. הקדמה ותיאור הנתונים

1.1מטרת הפרויקט

הפרויקט עוסק בבניית מודל רגרסיה מבוסס למידת מכונה לחיזוי הדירוג הממוצע של סרטים (averageRating)בפלטפורמת IMDb. החיזוי מבוצע על בסיס מאפיינים שונים של הסרט הזמינים במלואם טרם יציאתו לאקרנים (כגון אורך, ז'אנר, שפה, מדינת הפקה, מאפייני קאסט ושנת יציאה). מטרת המחקר היא לזהות אילו גורמים תורמים ביותר לחיזוי הדירוג, להשוות מתודולוגית בין אלגוריתמים שונים, ולבחון את הגינות (Fairness) המודל המופק בין חתכים שונים של אוכלוסיית הדאטה.

2.1 תיאור מערך הנתונים

מערך הנתונים המקורי מכיל רשומות רבות הכוללות מידע מבני וטקסטואלי. כחלק מתהליך הניקוי והכנת הנתונים, הוסרו רשומות שבהן משתנה המטרה (averageRating) היה חסר. מערך הנתונים הסופי שעליו אומנו והוערכו המודלים כולל 115,560שורות (סרטים) ו 13-עמודות. המשתנים הנומריים והקטגוריאליים עברו סיווג והכנה ייעודית בתוך פונקציית העיבוד prepare_data טרם הזרמתם ל Pipeline-האימון.

2. פונקציית העיבוד המרכזית

ארכיטקטורת פונקציית prepare_data

בהתאם לדרישות הטכניות, הוגדרה פונקציה אחידה, דטרמיניסטית ורובוסטית בשם prepare_data(df). הפונקציה מקבלת DataFrame במבנה הגולמי ומבצעת את השלבים הבאים:

1. המרות טיפוסים וניקוי ערכים חסרים קשיחים :משך הסרט (runtimeMinutes) ושנת היציאה (startYear) מומרים לטיפוסים נומריים. שנים שאינן בטווח הגיוני (קטנות מ-1000 או גדולות מ-9999) מוחלפות ב NaN-לשם טיפול בהמשך הצינור.
2. סטנדרטיזציה של טקסטים :עמודות ה genres-וה lead_actors_ids-עוברות ניקוי תווים מיוחדים (כגון סוגריים מרובעים וגרשיים) וטקסטים המייצגים ערכים חסרים במאגר (כמו 'N\\'מומרים למחרוזות ריקות.)

3. סינון מדינות ושפות (Domain Knowledge): עמודות המדינה והשפה ממופות אל מול רשימות מוגדרות מראש של הישויות הנפוצות והרלוונטיות ביותר בתעשייה, תוך איחוד שמות זהים) למשל, איחוד 'United States' ו.('USA'-
4. הסרת משתני Data Leakage: עמודות ה numVotes-BoxOffice I-מוסרות מהדאטה באופן מוחלט, שכן הן אינן ידועות בזמן יציאת הסרט לאקרנים..

3. הנדסת מאפיינים(Feature Engineering)

במסגרת תהליך הנדסת המאפיינים (Feature Engineering), יוצרו 6 פיצ'רים חדשים המבוססים על ידע ספציפי מעולם הקולנוע:

מוטיבציה	נוסחה / מקור	שם המאפיין
סרטים מרובי-ז'אנר עשויים להיות מורכבים יותר ולקבל דירוגים שונים	ספירת מספר הז'אנרים שסרט שייך אליהם (split על ',' בעמודת genres)	genre_count
סרטים שיצאו אחרי ייסוד IMDb זוכים לכיסוי דירוגים שונה	1 אם startYear ≥ 1990, אחרת 0 (שנת ייסוד IMDb)	is_post_imdb
נוכחות שחקנים מובילים משפיעה על ציפיות ועל דירוג	ספירת שחקנים ראשיים מתוך רשימת שחקני A-list	num_a_list_actors
סרטים בשפה האנגלית מגיעים לקהל רחב יותר	1 אם השפה היא English, אחרת 0	is_english
הפקות אמריקאיות שולטות בפלטפורמה ועשויות להיות בעלות הטיית דירוג	1 אם מדינת ההפקה היא USA, אחרת 0	is_usa
אורך הסרט קשור לסוג הסרט ולציפיות הצופים	קטגוריזציה של אורך הסרט: 1=קצר (<60'), 2=סטנדרטי (60–120'), 3=ארוך (>120')	runtime_category_num
ייצוג קטגוריאלי המאפשר למודל לזהות השפעת ז'אנר ספציפי	עמודה בינארית לכל ז'אנר עיקרי (Drama, Comedy, Thriller, Horror, Sci-Fi, Documentary וכו')	One-Hot על genres

בסך הכול הונדסו 6 מאפיינים חדשים, בנוסף לקידוד One-Hot של הז'אנרים. מאפיינים אלה משולבים עם המאפיינים המספריים המקוריים (runtimeMinutes, numVotes) ואחרים במטריצת הפיצ'רים הסופית.

4. בחירת מודלים ומתודולוגיה

1.4 המודלים שנבחרו

נבחרו שני מודלים בעלי גישות שונות מהותית:

– Elastic Net Regression מודל לינארי עם רגולריזציה משולבת (L1 + L2), המאפשר גם בחירת מאפיינים (sparsity) וגם יציבות בנוכחות מולטי-קולינאריות.

– Random Forest Regressor מודל אנסמבל לא-לינארי, המבוסס על עצי החלטה. נבחר במקום עץ החלטה בודד מכיוון שהוא מפחית משמעותית את שונות החיזוי (variance reduction) באמצעות מיצוע על מאות עצים, ומסוגל ללכוד אינטראקציות בין מאפיינים. הבחירה משלימה את המודל הלינארי באופן טבעי.

2.4 חלוקת אימון–בדיקה Cross-Validation-

הנתונים חולקו ל-80% אימון ו-20% בדיקה (test set) באקראי עם seed קבוע. על מערך האימון בוצע 10 fold Cross-Validation-לכל מודל, כך שכל קיפול שימש פעם אחת כוולידציה ותשע פעמים כאימון. כל שלבי העיבוד (imputation, scaling, encoding) שולבו ב Pipeline-והופעלו בתוך כל קיפול בנפרד.

1.5השוואת ביצועים

הטבלה הבאה מסכמת את ביצועי שני המודלים על ה-10 (fold CV)-ממוצע ± סטיית תקן:

מדד	Elastic Net	Random Forest
RMSE (ממוצע ± סטיית תקן)	*1.1267 (0.0486)*	*1.1205 (0.0493)*
MAE (ממוצע ± סטיית תקן)	*0.8636 (0.0357)*	*0.8579 (0.0368)*
R^2 (ממוצע ± סטיית תקן)	*0.2368 (0.0223)*	*0.2452 (0.0233)*
היפר-פרמטרים אופטימליים שנבחרו(Best Params)	*alpha: 0.001, l1_ratio: 0.3*	*max_depth: 10, min_samples_leaf: 4*

ניתוח מתודולוגי והנדסי של התוצאות

1. השוואת ביצועים וקביעת המודל המוביל: מודל ה Random Forest Regressor-נקבע כמודל האופטימלי והמוביל בפרויקט זה. הוא השיג ציון R^2 גבוה של 0.2452 וציון RMSE נמוך וטוב יותר של 1.1205. היתרון של ה Random Forest-נובע מיכולתו המובנית של אלגוריתם מבוסס עצים ללמוד קשרים לא-ליניאריים מורכבים ואינטראקציות חבויות בין הפיצ'רים הקטגוריאליים (לדוגמה, שילוב של ז'אנר ספציפי עם שפת מקור ומשך הסרט) בצורה גמישה, בעוד שהמודל הליניארי (Elastic Net), על אף ששודרג עם הרחבה פולינומית, נותר מוגבל יחסית ביכולת ההכללה שלו על מבנה הנתונים הספציפי הזה.
2. ניתוח כושר ההסבר (R^2): מודל ה Random Forest-מסביר 24.52% מהשונות של דירוג הסרטים באוכלוסייה. בעולם של חיזוי מדדים אמנותיים ותלויי טעם אנושי (Human-centric metrics), כושר הסבר הקרוב ל-25% נחשב להישג הנדסי מכובד ומקובל לחלוטין. המודל נדרש לנבא ציון של יצירה קולנועית ***טרם*** שהיא נחשפת לעולם, ללא גישה למשתנים איכותיים מכריעים כגון: איכות הבימוי, רמת התסריט, עריכה, או הבאז החברתי שנוצר בזמן ההפצה. התוצאות מראות שהמודלים הצליחו למצות את מקסימום המידע מתוך הפיצ'רים המבניים הקיימים.
3. יציבות ורובוסטיות הצינור: סטיות התקן שחושבו על פני 10 ה Folds-הן נמוכות במיוחד (בסביבות 0.02–0.04) בשני המודלים. הדבר מעיד על כך שהמדדים יציבים ועקביים מאוד על פני חלוקות שונות של הדאטה, (Low Variance) ומספק הוכחה מתודולוגית לכך שצינור העיבוד (Pipeline) המורכב שנבנה מונע Data Leakage בצורה הרמטית ומבטיח יכולת הכללה אמיתית.

2.5ניתוח שגיאות – 10 הסרטים עם ה Overprediction-הגדול ביותר

הטבלה הבאה מציגה את 10 הסרטים שהמודל היה אופטימי מדי בחיזוייו (חזה דירוג גבוה משמעותית מהדירוג האמיתי):

שם הסרט	דירוג אמיתי	דירוג חזוי	שגיאה
Horseplay: The History of Horse Riding	1.0	7.19	6.19
Justin Bieber: Always Believing	1.0	7.13	6.13
Rape of the Soul	1.7	7.83	6.13
Ramo Trip	1.4	7.40	6.00
Kurz	1.2	7.16	5.96
At the Back of the Screen	1.5	7.34	5.84
Zwischen Glück und Krone	1.2	7.01	5.81
Jasenovac: Istina	1.4	7.18	5.78
Tribalism Is Killing Us	1.3	7.05	5.75
Obama in NC: The Path to History	1.5	7.21	5.71

3.5ניתוח שגיאות – 10 הסרטים עם ה Underprediction-הגדול ביותר

שם הסרט	דירוג אמיתי	דירוג חזוי	שגיאה
Son-nim-1 Cheo-beon-jjae I-ya-gi	9.1	4.29	-4.80
Bukunja Tekunja Mitti: The Cannibals	9.1	4.30	-4.79
Ungleich	9.1	4.35	-4.74
Il Santo Veleno	9.4	4.98	-4.41
Oscuro by Rene Potter	8.6	4.19	-4.40
Of Sound Mind	9.6	5.21	-4.382
Puli Vachindi Meka Chachindi	9.5	5.17	-4.32
Out 2 Kill	9.4	5.09	-4.30
Pulse of the Indigo	9.3	5.00	-4.29
Colour Palette	9.1	4.84	-4.25

ניתוח איכותי של ה-Outliers-חמישה סרטים ספציפיים

1. Justin Bieber: Always Believingשגיאה: +6.1286 (סרט תיעודי מוזיקלי משנת 2012 העוסק בכוכב הפופ ג'סטין ביבר. המודל זיהה שמדובר בז'אנר דוקומנטרי (המקבל משקל חיובי גבוה מאוד במודל) ובשפה האנגלית, ולכן חזה לו ציון גבוה של 7.12. בפועל, הסרט קיבל את הציון

המינימלי 1.0 עקב תופעה חברתית של – "Review Bombing" מתקפת הצבעות המונית של אנטי-מעריצים ברשת, המטים את הציון ללא קשר לאיכות הסרט.

2. Horseplay: The History of Horse Riding שגיאה: +6.1918 :(סרט דוקומנטרי העוסק בהיסטוריה של רכיבה על סוסים. בדומה למקרה הקודם, המודל מעניק פרמיה משמעותית לז'אנר הדוקומנטרי ומנבא ציון של 7.19. אולם, מדובר בהפקה עצמאית, דלת-תקציב וחובבנית שזכתה לביקורות שליליות קיצוניות מהקהל המצומצם שצפה בה. המודל, הנסמך על משתנים יבשים בלבד, אינו מסוגל להעריך את איכות ההפקה בפועל.
3. Jasenovac: Istina שגיאה: +5.7757 :(סרט דוקומנטרי היסטורי-פוליטי משנת 2016 העוסק במחנה הריכוז יאסנובאץ. סרטים מסוג זה מעוררים מחלוקות לאומיות ואידיאולוגיות עזות, מה שמוביל לגיוס המוני של גולשים המצביעים בציון הנמוך ביותר (1.0) כחלק מ"מלחמה פוליטית" על הנרטיב ההיסטורי ברשת. המודל חזה ציון ממוצע של 7.17 כי הוא אינו מודע לרגישות הגיאופוליטית של נושא הסרט.
4. Of Sound Mind שגיאה: -4.3823 :(סרט דרמה/מתח עצמאי שקיבל ב IMDb-ציון קיצוני של 9.6. המודל ניבא לו ציון שמרני של 5.21 הקרוב למרכז ההתפלגות. סרטים עצמאיים קטנים אלו זוכים לרוב למספר מצומצם מאוד של הצבעות (מדגם לא מייצג), המגיעות בעיקר מאנשי צוות ההפקה שמעניקים ציוני 10 מוטים. המודל בצדק "מגן על עצמו" ואינו נוטה לנבא ערכים קיצוניים כאלו בהיעדר סימנים מובהקים.
5. Son-nim-1 Cheo-beon-jjae I-ya-gi שגיאה: -4.8040 :(סרט דרמה קוריאני שקיבל את הציון 9.1. המודל חזה לו ציון נמוך של 4.29. המודל נוטה להפחית בערכם של סרטים שאינם דוברי אנגלית או שאינם מופקים בארה"ב. מה שהמודל לא ידע הוא שמדובר ביצירת מופת מקומית נישתית שזכתה לאהדת קהל מעריצים אדוק במדינת המקור.

5.3 זיהוי דפוסים משותפים והצעות לשיפור

- דפוס ה-Overpredictions אופטימיות יתר: רובם המוחלט של הסרטים בקבוצה זו הם סרטים דוקומנטריים (Documentaries) קצרים או פוליטיים. המודל למד שז'אנר זה זוכה להצלחה, אך הוא נכשל בהבחנה בין דוקומנטרי איכותי לבין הפקות חובבניות או סרטים שסובלים ממתקפות אידיאולוגיות (Review Bombing) שמורידות את הציון ל-1.0 באופן מלאכותי.
- דפוס ה-Underpredictions פסימיות יתר: קבוצה זו מורכבת בעיקר מסרטים זרים או סרטים עצמאיים קטנים מאוד בעלי ציונים חריגים (מעל 9.0). המודל מעניק להם תחזית שמרנית במרכז ההתפלגות כיוון שאין להם את "תו האיכות" של הפקה הוליוודית רחבה, והוא אינו מודע לכך שהציון הגבוה נובע ממדגם הצבעות קטן ומוטה של מעריצים מקומיים.

הצעות ל-3 פיצ'רים חדשים לפתרון הבעיה (לקראת חלק 3):

1. פיצ'ר מוניטין חברת ההפקה/במאי: חישוב ממוצע ציוני העבר של סרטי חברת ההפקה או הבמאי, כדי לאפשר למודל להפריד בין הפקה עצמאית חובבנית להפקה מקצועית.
2. מדד אינטראקציה תרבותית (Topic Sensitivity): שימוש בטכניקות NLP על עמודת ה plot-או כותרת הסרט כדי לזהות מילים רגישות פוליטית או דמויות ציבוריות מעוררות מחלוקת (פרוקסי לזיהוי פוטנציאל ל-Review Bombing).
3. פיצ'ר אמינות המדגם (Proxy) ל Vote Volume-טרם היציאה: (הכללת משתנה המייצג את גודל קהל היעד הנישתי (למשל, שילוב של ז'אנר * שפת מקור) כדי לאותת למודל מתי מדובר בסרט קטן שציוניו צפויים להיות מוטים על ידי קבוצה קטנה של מצביעים.

6. ניתוח הוגנות(Fairness Analysis)

1.6הוגנות לפי מדינת המקור

קבוצה	מספר סרטים	RMSE	MAE
US	22,471	0.958	0.726
Non-US	93,089	1.157	0.889

2.6הוגנות לפי עשור הוצאת הסרט

עשור	מספר סרטים	RMSE	MAE
1910s	171	0.847	0.636
1920s	785	1.038	0.790
1930s	3,121	0.848	0.646
1940s	3,292	0.839	0.646
1950s	4,602	0.898	0.696
1960s	6,208	1.006	0.796
1970s	7,969	1.033	0.812
1980s	9,631	1.075	0.839
1990s	10,213	1.079	0.844
2000s	17,555	1.147	0.878
2010s	33,605	1.178	0.893
2020s	18,405	1.244	0.948

דיון מעמיק במצאים והסברים תיאורטיים:

1. הטיית מדינת המקור :(Origin Bias) מהתוצאות עולה כי המודל מדויק משמעותית עבור סרטים אמריקאיים (US) לעומת סרטים משאר העולם ,(Non-US) עם הפרש של כ-20% ברמת השגיאה RMSE)של 0.9584 מול 1.1571 .(תעשיית הקולנוע בארה"ב פועלת תחת תבניות הפקה קשיחות ומובנות יחסית (נוסחאות הוליוודיות לאורך וז'אנרים). המודל מצליח ללמוד תבניות אלו היטב. לעומת זאת, סרטים בינלאומיים מציגים שונות תרבותית עצומה ומדורגים על ידי קהלים מקומיים בעלי דפוסי הצבעה שונים, מה שמקשה על הגעה לרמת דיוק זהה.

2. הטיית ציר הזמן :(Decade Degradation) קיים קשר מונוטוני מובהק – ככל שהסרט חדש יותר, השגיאה של המודל הולכת ומחריפה !עבור סרטים משנות ה-40 השגיאה נמוכה מאוד (RMSE = 0.8391),בעוד שעבור סרטים משנות ה-2020 היא מזנקת ל-1.2436 (הרעה של כ-

48% בדיוק!). תופעה זו מוסברת על ידי "אפקט המסננת ההיסטורית" – סרטים ישנים שקיימים כיום במאגר הם לרוב הקלאסיקות הגדולות ביותר ששרדו את מבחן הזמן, וציוניהם קונצנזואליים. לעומת זאת, בעידן המודרני ומהפכת הסטרימינג מיוצרים עשרות אלפי סרטים בשנה ברמות איכות משתנות באופן קיצוני, והצבעות הקהל מושפעות מטרנדים רגעיים ברשתות החברתיות, מה שהופך את הדירוג להרבה פחות צפוי.

7. חשיבות פיצ'רים Top 5 – לכל מודל

בהתאם לסעיף 4 במטלה, חולצו חמשת הפיצ'רים המשפיעים ביותר מתוך שני המודלים עבור Elastic Netלפי גודל המקדמים הסטנדרטיים, ועבור Random Forest לפי מדד ה Gini Feature-Importanceהמבוסס על הפחתת האי-טוהר בעצים.

להלן הטבלה הרשמית והמדויקת של הפיצ'רים המובילים וכיוון השפעתם על החיזוי כפי שהופקו בריצה:

2.7 Elastic Net – Top 5

פיצ'ר	Coefficient	כיוון
num__genre_Horror	0.4443-	שלילי (-)
num__genre_Documentary	0.4311	חיובי (+)
num__genre_Thriller	0.3407-	שלילי (-)
num__genre_Sci-Fi	0.2996-	שלילי (-)
num__runtimeMinutes	0.2420	חיובי (+)

2.7 Random Forest – Top 5

פיצ'ר	Importance	כיוון
num__genre_Documentary	0.3206	חיובי (+)
num__genre_Horror	0.2219	שלילי (-)
num__runtimeMinutes	0.1807	חיובי (+)
num__genre_Drama	0.0940	חיובי (+)
num__is_post_imdb	0.0406	שלילי (-)

3.7 פירוש הממצאים

דיון והשוואה בין המודלים

- הסכמה קונצנזואלית על המשתנים המובילים: קיים מתאם מרשים והסכמה מלאה בין המודל הליניארי (Elastic Net) למודל מבוסס העצים (Random Forest) לגבי שני הפיצ'רים המשמעותיים ביותר בחיזוי דירוג. שניהם קובעים כי הז'אנר דוקומנטרי (Documentary) הוא בעל השפעה חיובית חזקה (סרטים תיעודיים מקבלים ציונים גבוהים יותר מהקהל הממוקד

הצופה בהם), בעוד שהז'אנר אימה (Horror) הוא בעל משקל שלילי כבד) סרטי אימה נוטים לקבל באופן עקבי ציונים נמוכים בממוצע במאגר.(IMDb

- משקל אורך הסרט :(runtimeMinutes) שני המודלים מזהים שמשך סרט ארוך יותר מתואם חיובית עם ציון גבוה יותר. הדבר מתיישב עם התעשייה שבה סרטים ארוכים נתפסים לרוב כהפקות "אפיות" או דרמות מושקעות הזוכות ליחס אוהד יותר.
- נקודת השוני והיתרון של ה :Random Forest-מודל ה Random Forest-הצליח להכניס לחמישייה המובילה את הפיצ'ר ההנדסי שייצרנו is_post_imdbוהעניק לו כיוון השפעה שלילי (סרטים שיצאו בעידן הדיגיטלי שלאחר 1991 זוכים לתחזית שמרנית ונמוכה יותר מסרטים קלאסיים). יכולתו של ה Random Forest-לשלב את הפיצ'ר הזה בצורה לא-ליניארית בתוך ענפי העצים היא זו שהעניקה לו את יתרון הדיוק הדרמטי ופריצת תקרת הביצועים על פני ה-Elastic Net.

8. סיכום, מגבלות והמלצות לעתיד

1. מסקנה מתודולוגית :פרויקט זה הוכיח את החשיבות של שימוש ב Pipeline-מובנה וב Cross-Validationנקי למניעת .Data Leakage מבין המודלים שנבחנו Random Forest Regressor , נקבע כמודל הסופי והמוביל לחיזוי דירוגי סרטים, בזכות יכולתו ללמוד קשרים מורכבים בין מאפייני הסרט השונים.
2. מגבלות המודל והדאטה :המודל מוגבל קשיחה על ידי המשתנים היבשים הזמינים טרם יציאת הסרט. מניתוח השגיאות וההוגנות עולה כי הקושי המרכזי של המודל נעוץ בחיזוי סרטים מודרניים (משנות ה-2000 והלאה) וסרטים זרים ,(Non-US) שם קיים רעש תרבותי רב ודפוסי הצבעה תנודתיים ב IMDb-כמו תופעות של Review Bombing או הטיית קבוצות מעריצים קטנות

3.9 שימוש בבינה מלאכותית (AI Usage Log)

במהלך הפרויקט נעשה שימוש בכלי בינה מלאכותית (gimini) לצורך ייעוץ והכוונה. להלן נקודות ההתייעצות העיקריות:

מה שינינו בעקבות התשובה	מה שאלנו	נושא ההתייעצות
הוספנו את is_post_imdb ו-num_a_list_actors בעקבות ההמלצה.	אילו פיצ'רים הנדסיים עשויים להוסיף ערך לחיזוי דירוג סרטים?	בחירת פיצ'רים הנדסיים
בחרנו ב-MultiLabelBinarizer לפיצול לעמודות בינאריות.	איך מקודדים שדה genres שמכיל מספר ז'אנרים לסרט?	טיפול בז'אנרים רב-ערכיים
השתמשנו ב-GridSearchCV עם alpha ב-[0.001, 0.01, 0.1, 1] ו-l1_ratio ב-[0.1, 0.5, 0.9].	מהו טווח סביר ל-alpha ול-l1_ratio?	בחירת היפר-פרמטרים ל-Elastic Net
עטפנו את כל העיבוד ב-Pipeline ועברנו את ה-Pipeline ל-cross_val_score.	איך לוודא שה-Scaler לא רואה את סט הבדיקה?	מניעת Data Leakage
השתמשנו ב-seaborn barplot עם RMSE ו-MAE צמודים לכל קטגוריה.	איך להציג ויזואלית פערי ביצועים בין תת-קבוצות?	ניתוח הוגנות