בס"ד

דוח עבודת למידת מכונה - חיזוי דירוגי סרטים

מגישים: מאיה הלוי 212074736 | עידן וידר 315143305

1. תיאור הדאטה המעובד וטבלת הפיצ'רים

המשימה היא לחזות את הציון הממוצע של סרטים ב-IMDb (averageRating) על בסיס מטה-דאטה הזמינה לפני שחרורם. סט הנתונים המקורי כולל 133,884 סרטים. לאחר העשרת נתונים, הנדסת פיצ'רים ומניעת דליפת מידע, יצרנו 18 פיצ'רים.

שם הפיצ'ר	סוג	טכניקה	תיאור ומוטיבציה
startYear	נומרי	גולמי	שנת היציאה לאקרנים
runtimeMinutes	נומרי	גולמי	משך הסרט בדקות
num_actors	נומרי	גולמי	מספר השחקנים הראשיים
num_genres	נומרי	גולמי	מספר הז'אנרים של הסרט
screen_time_per_actor	נומרי	**Ratio**	runtime/num_actors. תופס 'אינטימיות' הפקה - סרט ארוך עם מעט שחקנים שונה מסרט אנסמבל
director_oscar_wins	נומרי	היסטורי	זכיות אוסקר של הבמאי לפני יציאת הסרט
lead_actors_oscar_wins	נומרי	צבירה	סך זכיות אוסקר של השחקנים לפני הסרט
lead_actors_oscar_wins_max	נומרי	Aggregation	מקסימום זכיות - מזהה נוכחות כוכב
actors_recent_movies_sum	נומרי	צבירה	סך סרטי השחקנים ב-5 שנים שקדמו
actors_recent_movies_max	נומרי	Aggregation	מקסימום פעילות שחקן בודד
has_director	בינארי	Indicator	1 אם זהות הבמאי ידועה
is_sequel	בינארי	Text	1 אם סרט המשך (לפי הכותרת)
missing_actors	בינארי	**MNAR**	1 אם אין מידע על שחקנים. חוסר אינו אקראי - מאפיין תיעודי/אינדי
missing_language	בינארי	**MNAR**	1 אם אין מידע על שפה - מסמן הפקות נישתיות
missing_country	בינארי	**MNAR**	1 אם אין מידע על מדינה - סרטים נשכחים/קטנים
has_oscar_winner	בינארי	Indicator	1 אם שחקן זכה באוסקר לפני הסרט
actors_very_active	בינארי	**Binning**	1 אם יש שחקן עם 5+ סרטים ב-5 שנים. סף המבדיל כוכבים מבוקשים
decade	קטגוריאלי	Binning	העשור שבו יצא הסרט
genres	Multi-label	רשימה	רשימת הז'אנרים (מ-28 אפשריים)
Country	Multi-label	רשימה	רשימת מדינות ההפקה
Language	Multi-label	רשימה	רשימת שפות המקור
director_name	טקסט	High-card	שם הבמאי (Top 30 הנפוצים)

2. פונקציית prepare_data ושלביה

פונקציית prepare_data מהווה את שלב ההכנה היחיד שמתבצע מחוץ ל-Pipeline. היא אחראית באופן בלעדי על ניקוי הדאטה, הסרת עמודות מדליפות, והנדסת הפיצ'רים. אין בה imputation, scaling או encoding - אלה מתבצעים בתוך ה-Pipeline בלבד, כדי למנוע data leakage.

שלבי הפונקציה

1. הסרת עמודות שמדליפות מידע עתידי: averageRating (המטרה), numVotes (מצטבר אחרי השחרור), BoxOffice (הכנסות לאחר השחרור).
2. הסרת עמודות לא תורמות: plot (50% חסר, טקסט חופשי), primaryTitle (לאחר חילוץ is_sequel), budget (87% חסר).
3. העשרה ממקורות חיצוניים: imdb_enrich_cache (מדינה/שפה/במאי), oscar_wins_cache (זכיות אוסקר מ-actor_movies_cache), Wikidata (היסטוריית סרטים).
4. הנדסת 18 הפיצ'רים, תוך הקפדה שכל פיצ'ר היסטורי מחושב רק על אירועים שקדמו לשנת יציאת הסרט.

גישה נוקשה להשלמת מדינה ושפה

נתקלנו באתגר מתודולוגי משמעותי: השדה region בקובץ title.akas של IMDb מציין טריטוריית הפצה ולא מדינת הפקה. סרט הוליוודי המופץ ב-100 מדינות יופיע עם 100 שורות. הגישה הנאיבית של 'המדינה השכיחה ביותר' הובילה לתוצאות שגויות - למשל סרט שתויג כהפקה לטבית בטעות.

הפתרון: גישה שמרנית הלוקחת מידע רק משורות עם סימון מפורש של מקור (isOriginalTitle=1 או types='original'). אם אין סימון אמין - מחזירים None, והפיצ'ר missing_country/language מקבל 1. עיקרון מנחה: עדיף NaN מפורש מאשר ערך משוער שגוי.

בנוסף, יישמנו STRICT WHITELIST: כ-160 מדינות, כ-140 שפות, ו-28 ז'אנרים מוגדרים מראש. כל ערך שאינו ברשימה נדחה. ישויות היסטוריות (Soviet Union, Yugoslavia) נשמרות בנפרד מהמדינות של היום, כדי שהמודל ילמד את ההקשר ההיסטורי.

3. מניעת Data Leakage - הגנה דו-שכבתית

Data Leakage היא דליפת מידע מסט המבחן לסט האימון, המובילה להערכת ביצועים מנופחת. נקטנו הגנה דו-שכבתית.

שכבה ראשונה: הסרת משתנים עתידיים

הסרנו את כל המידע שלא היה זמין לפני שחרור הסרט: averageRating (המטרה), numVotes (מצטבר עם הזמן), ו-BoxOffice (נמדד אחרי השחרור). גם בפיצ'רים שיצרנו הקפדנו - director_oscar_wins סופר רק זכיות שקדמו לשנת הסרט, ואותו עיקרון חל על שאר הפיצ'רים ההיסטוריים.

שכבה שנייה: Pipeline בתוך Cross-Validation

השכבה הקריטית עוסקת בשלבים שלומדים מהנתונים: SimpleImputer לומד חציון, StandardScaler לומד ממוצע וסטיית תקן, OneHotEncoder לומד קטגוריות. אם פעולות אלה נעשות על כל הדאטה לפני הפיצול ל-folds - סט המבחן 'משפיע' על מה שהמודל לומד.

הפתרון: כל שלבי העיבוד עטופים ב-Pipeline. בקריאה ל-cross_validate, Scikit-Learn מבצע את הלמידה מחדש בכל fold - על נתוני האימון של אותו fold בלבד. סט המבחן מקבל את הטרנספורמציה אך לא משפיע עליה. כך כל מדד RMSE/MAE/R² בדוח הוא הערכה 'נקייה' של ביצועים על נתונים חדשים.

ColumnTransformer מנתב כל קבוצת פיצ'רים לטיפול מתאים: נומריים → median imputer + scaler; קטגוריאליים → most_frequent + OneHotEncoder; multi-label → MultiLabelTopN(25); director → OneHotEncoder(max_categories=30). הכל בנפרד בכל fold.

4. השוואת שני המודלים (Fold CV-10)

הערכנו שני מודלים שונים מהותית - Elastic Net (רגרסיה ליניארית עם רגולריזציה L1+L2) ו-Random Forest (אנסמבל עצים). היפר-פרמטרים נבחרו ב-Grid Search (5-Fold), וההערכה הסופית ב-Fold CV-10.

מודל	RMSE	Std±	MAE	Std±	R²	Std±
Elastic Net	1.1229	0.0140	0.8582	0.0111	0.2474	0.0126
Random Forest	**1.0873**	0.0153	**0.8290**	0.0118	**0.2944**	0.0124

ניתוח התוצאות

Random Forest עולה על Elastic Net בכל שלושת המדדים: שיפור של 3.2% ב-RMSE ו-19% יחסית ב-R². ההפרש מובהק - גדול בהרבה מסטיות התקן. שני המודלים יציבים מאוד (סטיית תקן 1.2%-1.4% מהממוצע), מה שמעיד על למידת דפוסים אמיתיים ללא overfitting.

מדוע Random Forest טוב יותר?

ראשית, יכולת ללכוד יחסים לא-ליניאריים: עצים מבצעים פיצולים בסף (כמו runtime > 120) ותופסים תופעות שלא ניתנות לתיאור במקדם ליניארי. למרות שהוספנו PolynomialFeatures ל-Elastic Net, זה מוגבל לאינטראקציות מדרגה 2 בלבד.

שנית, חוסר רגישות לסקאלה: RF מתבסס על פיצולים יחסיים ולכן עמיד לפיצ'רים בסקאלות שונות. שלישית, טיפול אינהרנטי באינטראקציות מסדר גבוה - פיצול לפי runtime יכול להוביל לתת-עץ שמפצל לפי genre, ללא הגדרה מפורשת. בחרנו ב-Random Forest כמודל הראשי.

5. ניתוח שגיאות (Error Analysis)

הניתוח על Random Forest. השארית: residual = actual - predicted. חילצנו 20 outliers באמצעות cross_val_predict.

5.1 חילוץ Outliers

Top 5 Overprediction (חזוי גבוה, אמת נמוכה):

שגיאה	חזוי	אמת	ז'אנר	שנה	שם הסרט
5.97-	7.17	1.20	Documentary	2023	Kurz
5.87-	7.17	1.30	Documentary	2019	Tribalism Is Killing Us
5.65-	7.15	1.50	Documentary	2010	Obama in NC
5.57-	6.87	1.30	Comedy, Doc	2009	Play in the Gray
5.50-	7.10	1.60	Documentary	2000	Hatemongers

Top 5 Underprediction (חזוי נמוך, אמת גבוהה):

שגיאה	חזוי	אמת	ז'אנר	שנה	שם הסרט
4.91+	4.19	9.10	Action, Horror	2012	Bukunja Tekunja Mitti
4.45+	4.65	9.10	Horror	2011	Son-nim Cheo-beon-jjae
4.37+	4.23	8.60	Horror	2016	Oscuro by Rene Potter
4.32+	5.08	9.40	Thriller	2023	Il Santo Veleno
4.31+	5.29	9.60	Romance, Thriller	2015	Of Sound Mind

5.2 ניתוח דפוסים

Kurz (2023): תיעודי ללא מטה-דאטה. המודל ראה Documentary ואורך סטנדרטי וחזה 7.17, אך האמת 1.20. תיעודיים נישתיים מקבלים חיזוי גנרי גבוה שלא תואם את האיכות בפועל.

Tribalism Is Killing Us (2019): תיעודי על שבטיות פוליטית - השגיאה הגדולה ב-overprediction. סרטים פוליטיים שנויים במחלוקת סובלים מ-review bombing (הצבעה אסטרטגית של מתנגדים), תופעה שהמודל לא יכול ללכוד מהמטה-דאטה.

Bukunja Tekunja Mitti (2012): סרט של Nabwana I.G.G., מייסד 'Wakaliwood' (תעשיית הקולנוע האוגנדית). השגיאה הגדולה ב-4.91+) underprediction). דוגמה מושלמת ל-cult following - הדירוג 9.10 משקף קהל קטן אך מסור עולמי.

Of Sound Mind (2015): Romance/Thriller עצמאי לא ידוע עם דירוג 9.60. סרט נישתי שזוכה להערכת קהל קטן - הדירוג מבוסס על מספר voters קטן יחסית.

דפוסים ב-6 :Overprediction מתוך 10 הם Documentary - המודל למד שתיעודיים מקבלים דירוגים גבוהים, אך לא מבחין בין איכותיים לתעמולה/פוליטיים. 6 מתוך 10 חסרי num_actors. שלושה הם תיעודיים פוליטיים עם review bombing.

דפוסים ב-9 :Underprediction מתוך 10 הם Horror/Thriller - המודל למד ש-Horror נמוך בכללי, אך לא מזהה סרטים נישתיים עם cult following. כל ה-10 הם הפקות עצמאיות של במאים לא מוכרים, עם דירוגים גבוהים מאוד (8.5-9.6) בעידן 2010-2024.

5.3 השוואת Outliers בין המודלים

12 מתוך 20 ה-60%) outliers) משותפים לשני המודלים - אינדיקציה שהבעיה בדאטה ולא במודל ספציפי. ההשוואה הכמותית:

מי טוב יותר	שגיאת RF	שגיאת EN	קבוצת הסרטים
EN ב-12/20	4.931	4.913	על 20 outliers של RF
RF ב-16/20	**4.754**	5.105	על 20 outliers של EN

מסקנה: כאשר RF טועה, EN לא מפצה (תוצאות כמעט זהות). אבל כאשר EN טועה, RF מצליח טוב יותר ב-80% מהמקרים. RF הוא מודל חזק יותר וגם עקבי יותר.

5.4 פיצ'רים חדשים מוצעים

פיצ'ר	סוג	השגיאה שהוא פותר
title_has_political_keywords	בינארי	Overprediction של תיעודיים פוליטיים (Tribalism, Obama, Hatemongers)
country_is_niche_film_industry	בינארי	Underprediction של Bukunja/Wakaliwood (cult)
is_independent_horror	בינארי	Underprediction של אימה עצמאית (9/10 מהמקרים)
runtime_anomaly (z-score)	נומרי	אורכים חריגים ביחס לז'אנר

הפיצ'רים היו מטפלים בלפחות 8 מתוך 20 ה-is_independent_horror. outliers. לבדו מכסה 9 מתוך 10 ה-runtime_anomaly. underpredictions. מתוחכם כי הוא משווה כל סרט ל"חבריו לז'אנר".

6. ניתוח הוגנות (Fairness Analysis)

בחנו את ביצועי Random Forest על חתכים שונים. הממוצע הכללי של RMSE הוא 1.077.

6.1 חתך לפי ז'אנר

הערכה	MAE	RMSE	מס' סרטים	ז'אנר
הטוב ביותר	0.7505	0.9912	11,005	Romance
טוב מאוד	0.7352	0.9876	9,103	Documentary
טוב	0.7830	1.0294	40,964	Drama
בינוני	0.8435	1.1010	23,403	Comedy
הגרוע	0.8992	**1.1640**	9,726	Action

פער של כ-18% בין Romance (הטוב) ל-Action (הגרוע). Action קשה בגלל הגיוון - ממבלוקבאסטרים ועד B-movies. Romance ו-Documentary הומוגניים יותר.

6.2 חתך לפי מדינה (US מול Non-US)

MAE	RMSE	מס' סרטים	מדינה
0.6944	0.9228	14,827	US
0.8592	1.1210	66,065	Non-US

פער של כ-21%. מעניין שלמרות פי 4.5 יותר סרטי Non-US, המודל טוב יותר על US. הסיבה: ההוליווד הומוגנית עם פורמטים סטנדרטיים, בעוד Non-US כולל מסורות מגוונות (בוליווד, סינמה אסיאתית, אירופאית) עם דינמיקות שונות.

6.3 חתך לפי עשור

סטייה	MAE	RMSE	מס' סרטים	עשור
28%-	0.5857	**0.7750**	2,359	1940s
11%-	0.7481	0.9587	4,350	1960s
4%-	0.8114	1.0344	6,759	1980s
4%+	0.8534	1.1188	12,257	2000s
7%+	0.8756	1.1520	23,567	2010s
13%+	0.9282	**1.2161**	12,712	2020s

הממצא הדרמטי ביותר: פער של כ-57% בין שנות ה-40 (RMSE 0.7750) לשנות ה-2020 (1.2161). ככל שהסרט חדש - הביצועים גרועים יותר. הסיבה אינה כמות (ל-2010s הכי הרבה סרטים), אלא Survivorship Bias: סרטים ישנים ששרדו הם הקלאסיקות שזכרו, ולכן יציבים. סרטים חדשים כוללים את כל הספקטרום - בלוקבאסטרים, B-movies, ותוכן סטרימינג.

6.4 דיון - חששות לשימוש בפועל

המודל יחסית הוגן בז'אנר (18%) ובמדינה (21%), אך מציג בעיה משמעותית בעשור (57%). לא היינו סומכת על החיזוי עבור: סרטים חדשים (2020+), סרטי אקשן, או הפקות Non-US נישתיות. כן הייתי סומכת על: קלאסיקות (1930-1950), Romance/Documentary סטנדרטיים, וסרטי US. הפער בעשור הוא תכונה של הדאטה (survivorship bias), לא של המודל. לשימוש מסחרי מומלץ להוסיף 'אזהרת אמינות' לסרטים חדשים.

לגבי missingness כמנבא: פיצ'רי ה-(missing_actors/language/country) MNAR אכן נמצאו אינפורמטיביים. חוסר מידע מתואם עם הפקות נישתיות/עצמאיות, שמתואמות עם דפוסי דירוג מסוימים - מה שמראה שהחוסר עצמו נושא אות.

7. ניתוח חשיבות פיצ'רים

חילצנו חשיבות מ-Elastic Net (מקדמים סטנדרטיים) ומ-EN. Random Forest (MDI - Mean Decrease Impurity) כולל 148 פיצ'רים, RF כולל 142.

Top 5 - Elastic Net

#	פיצ'ר	מקדם	כיוון
1	genres=Documentary	1.2381+	חיובי
2	genres=Horror	0.8754-	שלילי
3	Language=Hindi	0.6400-	שלילי
4	genres=Animation	0.5255+	חיובי
5	genres=Action	0.3971-	שלילי

תובנה: כל הפיצ'רים המובילים ב-EN הם קטגוריאליים-בינאריים. הסיבה: PolynomialFeatures פיצל את הפיצ'רים הנומריים לאינטראקציות רבות, מה שגרם לכוח החיזוי שלהם 'להתפזר'.

Top 5 - Random Forest

#	פיצ'ר	חשיבות (MDI)	סוג
1	runtimeMinutes	0.1048	נומרי
2	genres=Documentary	0.1032	קטגוריאלי
3	genres=Horror	0.1001	קטגוריאלי
4	**screen_time_per_actor**	0.0938	נומרי (הנדסי)
5	startYear	0.0705	נומרי

היכן המודלים מסכימים: שניהם זיהו את Documentary (חיובי) ו-Horror (שלילי) כקריטיים. הסכמה זו בין מודל ליניארי לעצים מעידה על תופעות אמיתיות בדאטה: תיעודיים זוכים לדירוגים גבוהים בזכות קהל מחויב, אימה סובלת משונות גדולה באיכות. הפיצ'ר ההנדסי screen_time_per_actor במקום הרביעי - הוכחה לערך הנדסת הפיצ'רים.

8. מסקנות, מגבלות והצעות להמשך

מסקנות עיקריות

- Random Forest טוב יותר ממודל ליניארי - R^2=0.2944 לעומת 0.2474 (שיפור 19% יחסי), בזכות לכידת קשרים לא-ליניאריים.
- ז'אנר הוא הפיצ'ר הדומיננטי - Documentary (חיובי) ו-Horror (שלילי) בשני המודלים. ההסכמה מחזקת את האמינות.
- הנדסת פיצ'רים תרמה: (Ratio screen_time_per_actor) במקום 4 בחשיבות, ופיצ'רי MNAR נמצאו אינפורמטיביים.
- בעיית הוגנות מרכזית בעשור - פער 57% מ-survivorship bias. ביצועים נמוכים גם על אקשן ועל סרטים פוליטיים/נישתיים.

מגבלות מהותיות

חיזוי דירוג IMDb לפני שחרור הוא משימה עם תקרה תיאורטית. הדירוג נקבע על ידי תופעות שלא ניתן לחזות מהמטה-דאטה: באז ברשתות, תגובות מבקרים, מגמות פוליטיות (review bombing), ו-cult following בקהילות נישתיות. R^2 של 0.30 על מטה-דאטה בלבד הוא תוצאה סבירה בספרות.

הצעות להמשך

5. ניתוח טקסט (NLP) על תקציר העלילה (שהוסר עקב חוסר) באמצעות embeddings סמנטיים, ללכוד את 'אופי' הסרט.
6. מדדי באז לפני שחרור: אזכורים ב-X/Twitter, חיפושי Google, צפיות בטריילרים - מנבאים עניין מוקדם.
7. ניתוח sentiment של ביקורות מוקדמות מ-Rotten Tomatoes או Metacritic לחיזוי קצר-טווח.

9. AI Usage Log

השתמשנו ב-Claude (LLM) בשלבים הבאים, בהתאם לעקרונות שקיפות אקדמית:

שלב	אופי השימוש ואימות
הנדסת פיצ'רים	התייעצות על יחסי פיצ'רים ואינדיקטורים. אימתנו ע"י בדיקת ההיגיון התחומי וניתוח feature importance.
מבנה Pipeline	עזרה במבנה ColumnTransformer ובמתודולוגיה למניעת leakage. אימתנו ע"י בדיקה שהמדדים סבירים ולא מנופחים.
דיבאג cache	זיהוי שגיאת הטיפול בשדה region של title.akas. אימתנו ידנית מול נתוני סרטים ידועים.
פירוש תוצאות	עזרה בניתוח outliers ופערי הוגנות. אימתנו ע"י בדיקה חוזרת של הסרטים הספציפיים.
ניסוח הדוח	שיפור ניסוח וארגון. כל התוכן המדעי והמסקנות הם שלנו.