בניית מודל חיזוי דירוג סרטים

מגישות: תפארת בלוקה וזיו נגד

1. תיאור הדאטה המעובד וטבלת הפיצ'רים:

הגרפים מופיעים בנספח א'.

הדאטה מבוססת על מאגר סרטים שנבנה בחלק הראשון של המטלה.
המטרה היא לחזות דירוג סרטים ממוצע עוד לפני יציאתו לאקרנים תוך שימוש בשדות הידועים טרם הרצת הסרט. בחלק זה נתאר את תהליך העיבוד והנדסת הפיצ'רים שבסופו קיבלנו כ115,560 סרטים ו18 פיצ'רים.

תאים 2-4:

תחילה טענו את קובץ הנתונים dataset.csv לתוך Dataframe והדפסנו את גודל הדאטה כדי לבדוק כפילויות של שורות. לאחר מכן, בדקנו את כמות ואחוז הערכים החסרים בכל עמודה על מנת לזהות אילו משתנים דורשים טיפול בהמשך. בנוסף, סיננו רק סרטים שיש להם ערך קיים במשתנה המטרה. הפרדנו בין משתנה המטרה (דירוג הממוצע של הסרט), לבין שאר המשתנים (ההפרדה נדרשת כדי שהמודל ילמד לחזות את הדירוג ולא ישתמש בו בטעות כקלט).

תא 5- בדיקה ויזואלית של המשתנים:

בתא זה ביצענו ניתוח ראשוני של חלק מהעמודות בדאטה באמצעות גרפים.
בדקנו את התפלגות שנות היציאה, אורך הסרטים, הז'אנרים הנפוצים ומשתנים נוספים.
מטרת הניתוח היא להבין את מאפייני הדאטה לפני בניית הפיצ'רים ולזהות דפוסים שיכולים לעזור בשלב Feature Engineering.

גרף	מה נבדק	ממצא מרכזי	מסקנה
Release Year Distribution	התפלגות שנות היציאה של הסרטים	רוב הסרטים במאגר יצאו משנות ה־90 וה־2000 ואילך	שנת היציאה עשויה להשפיע על הדירוג ולכן נשמרה כמשתנה במודל
Runtime Distribution	התפלגות אורך הסרטים בדקות	רוב הסרטים באורך 80–120 דקות, עם חציון של כ־95 דקות	אורך הסרט עשוי להשפיע על הדירוג ושימש גם ליצירת פיצ'רים חדשים
Top 10 Genres	עשרת הז'אנרים הנפוצים ביותר	קיימת שונות משמעותית בשכיחות הז'אנרים במאגר	הז'אנר הוא משתנה חשוב לצורך חיזוי הדירוג
Top 10 Languages	עשר השפות הנפוצות ביותר	אנגלית היא השפה הדומיננטית בפער גדול	השפה עשויה להשפיע על הדירוג ולכן נבנה הפיצ'רis_english
Top 10 Countries	עשר מדינות המקור הנפוצות ביותר	ארצות הברית היא מדינת המקור הנפוצה ביותר	מדינת המקור עשויה להשפיע על הדירוג ונבנו פיצ'רים גיאוגרפיים
Mean Rating by Genre	דירוג ממוצע לפי ז'אנר	קיימים הבדלים ברורים בין הז'אנרים בדירוג הממוצע	מחזק את החשיבות של הז'אנר כחלק מהמודל
Average Rating Distribution	התפלגות משתנה המטרהaverageRating	רוב הסרטים מדורגים בטווח 5–7, ודירוגים קיצוניים נדירים יחסית	המודל צפוי להתקשות יותר בחיזוי סרטים בעלי דירוגים קיצוניים

טבלת פיצ׳רים:

Feature	סוג	נוסחה / אופן חישוב	מוטיבציה
movie_age	נומרי	2025 - startYear	הבדל בין דירוג סרטים ישנים לסרטים חדשים.
num_genres	נומרי	מספר הז׳אנרים של הסרט	מבדיל בין סרט ז׳אנר אחד לסרט עם מספר ז׳אנרים.
runtime_category	קטגוריאלי	Short / Medium / Long לפי אורך הסרט	חלוקת אורך הסרט לקטגוריות מאפשרת למודל לזהות דפוסים שקשה ללמוד ממשתנה רציף.
is_english	בינארי	אם English מופיעה בשפות הסרט אז 1	אנגלית היא השפה הדומיננטית ביותר בדאטה.
runtime_per_genre	נומרי מורכב	runtimeMinutes / num_genres	פיצ׳ר Ratio המשלב אורך סרט ומספר ז׳אנרים.
is_usa is_europe is_india is_east_asian	בינארי	1 אם הסרט מארה״ב 1אם הסרט מצרפת/גרמניה/איטליה/ספרד 1 אם הסרט מהודו 1 אם הסרט מיפן/סין/דרום קוריאה	-ארה״ב היא מדינת המקור הנפוצה ביותר. -מייצג סרטים אירופאים. -הודו הייתה המדינה השנייה בשכיחותה בדאטה. -מייצג את תעשיית הקולנוע במזרח אסיה.
Genre_Drama,Comedy, Documentary, Horror, Action, Romance, Thriller, Crime	בינארי	1 אם הסרט שייך לאחד הז׳אנרים	מבוסס על ניתוח ז׳אנרים

פונקציית prepare_data :

פונקציית prepare_data אחראית על הכנת הנתונים לפני כניסתם למודל. להלן שלבי הפונקציה:

שלב 1 : הסרת משתנים היוצרים Data Leakage
בשלב הראשון הוסרו עמודות שעלולות לגרום לData Leakage . BoxOffice וnumVotes אינן זמינות בזמן יציאת הסרט לאקרנים. AverageRating הוא משתנה המטרה ולכן אסור להשתמש בו כפיצ׳ר. ועמודת tcost היא מזהה טכני ואינה מכילה מידע חיזוי.

שלב 2 : ניקוי משתנים נומריים
עמודות כמו startyear, runtimeminutes, budget הומרו לערכים נומריים וטופלו גם ערכים חריגים. לדוגמא- הגבלנו את שנת היציאה לטווח שנים שבין 1900-2025 . ערכים מחוץ לטווחים הוחלפו בNan.

שלב 3: ניקוי משתנה הז׳אנרים
עמודת genres הופיעה במספר פורמטים שונים. לכן בוצע תהליך ניקוי: הסרת תווים מיותרים, המרת רשימות טקסט למחרוזת אחידה, הסרת ערכים חסרים וערכים לא תקינים. מטרת השלב הייתה ליצור ייצוג אחיד של הז׳אנרים לצורך בניית פיצ׳רים בהמשך .

שלב 4 : ניקוי שפות
בעמודת Language הוסרו ערכים כגון: Not Found / None / nan / \N לאחר הניקוי נשמרו רק שפות תקינות לצורך יצירת פיצ׳רים הקשורים לשפת הסרט .

שלב 5 : ניקוי מדינות מקור
בעמודת Country בוצע ניקוי של תווים מיוחדים וערכים שאינם מייצגים מדינת מקור אמיתית, כגון: Earth, International, Not Found. שלב זה אפשר להשתמש במדינות המקור בצורה עקבית בתהליך הנדסת הפיצ׳רים .

שלב 6 : Feature Engineering
פירוט מופיע בטבלה מעלה. פיצ׳רים אלו נועדו לספק למודל מידע שלא קיים ישירות בדאטה הגולמי ולשפר את יכולת החיזוי שלו .

שלב 7 : הסרת עמודות שאינן נדרשות
לאחר יצירת הפיצ׳רים החדשים הוסרו עמודות מקור שכבר אינן נדרשות: plot , primaryTitle, Country , Language , budget , startYear , genres , lead_actors_ids . חלק מהעמודות הוסרו משום שהוחלפו בפיצ׳רים חדשים וחלקן משום שהכילו טקסט חופשי או אחוז גבוה מאוד של ערכים חסרים. לדוגמה, עמודת budget הוסרה מכיוון שכ־87% מערכיה היו חסרים .

חלוקת הנתונים לסט אימון וסט מבחן:

בוצעה הפרדה בין משתנה המטרה לבין משתני הקלט. תחילה נשמרו רק סרטים בעלי דירוג קיים, ולאחר מכן הוגדרו: Y משתנה המטרה (averageRating) , X כלל משתני הקלט.
לאחר מכן בוצעה חלוקה לסט אימון וסט מבחן באמצעות train_test_split, כאשר 80% מהנתונים הוקצו לסט האימון ו20% מהנתונים הוקצו לסט המבחן, Random State=42 שימש לשחזור תוצאות זהה בהרצות חוזרות. פונקציית prepare_data הופעלה בנפרד על סט האימון והמבחן, ולא על כלל הדאטה לפני החלוקה. זה נועד למנוע Data Leakage ולהבטיח שהמידע מסט המבחן לא ישפיע על תהליך בניית המודל.

השוואת שני המודלים עם מדדי ביצועים:

טבלת התוצאות מופיעה בנספח ב'

בניית Pipeline ותהליך ה־Preprocessing לאחר השלמת שלב ניקוי הנתונים והנדסת הפיצ'רים, נבנה תהליך עיבוד נתונים (Pipeline) שמטרתו להכין את הנתונים בצורה אחידה עבור המודלים השונים. הפיצ'רים חולקו לשלוש קבוצות עיקריות: 1. משתנים נומריים עבורם בוצעה השלמת ערכים חסרים באמצעות החציון (Median Imputation), ולאחר מכן בוצע Standard Scaling.

2. משתנים בינאריים עבורם בוצעה השלמת ערכים חסרים באמצעות הערך 0, המייצג היעדר של המאפיין.
3. משתנה קטגוריאלי עבורו בוצעה השלמת ערכים חסרים באמצעות הקטגוריה "Other", ולאחר מכן בוצע One-Hot Encoding לצורך המרתו למשתנים מספריים המתאימים למודלים.
לבסוף, כל תהליכי העיבוד שולבו באמצעות ColumnTransformer, אשר הפעיל את פעולות העיבוד המתאימות לכל קבוצת משתנים. לאחר שלב זה התקבלו 18 פיצ'רים סופיים ששימשו לאימון המודלים.

מודל 1 – ElasticNet

המודל הראשון שנבחר לפרויקט הוא ElasticNet, מודל רגרסיה לינארית המשלב בין רגולריזציית L1 לבין רגולריזציית L2.
לצורך בחירת הפרמטרים האופטימליים בוצע Grid Search בשילוב 10-Fold Cross Validation. במסגרת החיפוש נבדקו מספר ערכים עבור הפרמטרים Alpha ו־L1 Ratio, ונבחר השילוב שהשיג את ביצועי החיזוי הטובים ביותר.
הפרמטרים שנבחרו היו: L1 Ratio = 0.15 , Alpha = 0.0005 .
לאחר בחירת הפרמטרים בוצעה הערכה נוספת באמצעות 10 Fold Cross Validation .
ניתן לראות כי המודל הצליח להסביר כ־22.5% מהשונות בדירוגי הסרטים. בנוסף, סטיות התקן הנמוכות בין ה־folds מעידות על יציבות יחסית של המודל.

מודל 2 – Random Forest

המודל השני שנבחר הוא Random Forest Regressor. מודל זה מבוסס על אוסף של עצי החלטה, כאשר כל עץ מאומן על דגימה שונה של הנתונים. לצורך כיוונון הפרמטרים בוצע Randomized Search, אשר בדק מספר שילובים שונים של: מספר העצים , עומק העצים , מספר הדוגמאות המינימלי בעלים, מספר המשתנים הנבדקים בכל פיצול.
השילוב שנמצא אופטימלי הוא: 'model__n_estimators': 300, 'model__min_samples_leaf': 10, 'model__max_depth': None, 'model__max_features': 'sqrt'. לאחר בחירת הפרמטרים בוצעה הערכה באמצעות 10 Fold Cross Validatio . נבחר Random Forest ולא Decision Tree משום שהוא מפחית את הסיכון ל־Overfittingבאמצעות שילוב של מספר עצים במקום עץ בודד. בנוסף, הוא יציב יותר לשינויים בנתונים ומספק ביצועים טובים יותר ברוב בעיות הרגרסיה. לכן נבחר כמודל מבוסס העצים בפרויקט.

ניתוח התוצאות

ניתן לראות כי Random Forest השיג תוצאות טובות יותר בהשוואה ל־ElasticNet במדדים שנבדקו.
בהשוואה בין המודלים ניתן לראות כי Random Forest השיג את ערך ה־RMSE הנמוך ביותר, כלומר את שגיאת החיזוי הממוצעת הנמוכה ביותר. בנוסף, הוא השיג גם את ערך ה־MAE הנמוך ביותר ואת ערך ה־R^2 הגבוה ביותר. תוצאות אלו מצביעות על כך שמודל Random Forest מצליח ללכוד טוב יותר את הקשרים

המורכבים הקיימים בין מאפייני הסרט לבין דירוגו. מכיוון שמדובר בבעיה שבה קיימים קשרים שאינם בהכרח לינאריים, יתרונו של Random Forest בולט במיוחד.

ניתוח חשיבות הפיצ׳רים (Feature Importance)

הגרף מופיע בנספח ג׳.

לאחר בחירת המודלים בוצע ניתוח חשיבות פיצ׳רים במטרה להבין אילו מאפיינים השפיעו בצורה המשמעותית ביותר על חיזוי דירוג הסרטים.
עבור מודל ElasticNet נמדדה החשיבות באמצעות ערכי המקדמים (Coefficients) מכיוון שבוצע Standard Scaling למשתנים הנומריים, ניתן להשוות בין גדלי המקדמים באופן ישיר. סימן המקדם מציין את כיוון ההשפעה: מקדם חיובי מעיד על קשר חיובי לדירוג, בעוד שמקדם שלילי מעיד על קשר שלילי.
עבור מודל Random Forest נמדדה החשיבות באמצעות מדד MDI (Mean Decrease in Impurity), אשר בוחן עד כמה כל פיצ׳ר תרם לשיפור איכות הפיצולים בעצי ההחלטה. מאחר שמדד זה אינו מספק כיוון השפעה, כיוון ההשפעה הוערך באמצעות הקורלציה בין הפיצ׳ר לבין משתנה המטרה.

השוואה בין המודלים

קיימת הסכמה ברורה בין המודלים לגבי שני הפיצ׳רים החשובים ביותר: genre_Documentary השפעה חיובית ו- genre_Horror השפעה שלילית.
ממצא זה מחזק את ההנחה שלז׳אנר הסרט יש השפעה משמעותית על דירוגו.
עם זאת, קיימים גם הבדלים בין המודלים. ElasticNet נתן משקל רב יותר למאפייני ז׳אנר, בעוד ש Random Forestזיהה חשיבות גבוהה יותר למשתנים נומריים כמו movie_age ו־runtimeMinutes הסיבה לכך היא ש־Random Forest מסוגל ללמוד קשרים לא לינאריים ואינטראקציות מורכבות בין משתנים, בעוד ש ElasticNet מבוסס על קשר לינארי.
בנוסף, הפיצ׳ר is_east_asian הופיע רק במודל ElasticNet מהפיצ׳רים בעוד ש- runtime_per_genre הופיע רק במודל Random Forest . הדבר מעיד שכל מודל זיהה דפוסים מעט שונים בתוך הנתונים.
לסיכום, שני המודלים מסכימים כי ז׳אנר הסרט הוא אחד הגורמים המשמעותיים ביותר לחיזוי דירוגו, אך Random Forest הצליח לנצל בצורה טובה יותר גם מידע מספרי מורכב שנוצר במהלך תהליך הנדסת הפיצ׳רים.

5.ניתוח שגיאות (Error Analysis)

5.1 חילוץ 20 ה-outliners- Random Forest: הטבלה מצורפת בנספח ד׳.

5.2. ניתוח איכותני- 5 סרטים ספציפיים

Kurz (2023)

דירוג: 1.2, חזוי: 7.14
דוקומנטרי קצר שעוסק בנושא שנוי במחלוקת. המודל טעה כי למד ש-Documentary דירוג גבוה, אך הסרט קיבל vote bombing מכוון קהל עוין דירג 1/10 בכוונה. המאפיין הייחודי: הנושא הפוליטי של הסרט משך קהל עוין שלא ניתן לזהות מהמאפיינים הקיימים בדאטה.

Justin Bieber’s Believe (2013)

דירוג: 1.6, חזוי: 7.19
דוקומנטרי מוזיקלי על Justin Bieber. המודל לא יכול לדעת שהאמן עצמו מקוטב קהל ששנא אותו דירג 1/10 בכוונה. המאפיין הייחודי: vote bombing שיטתי על אמן ספציפי, לא על איכות הסרט עצמו.

Proud American (2008)

דירוג: 1.1, חזוי: 6.55

סרט דרמה פטריוטי אמריקאי שקיבל קיטוב פוליטי חריף. המודל ראה Drama רגיל וחזה ממוצע לא יכול לדעת על קיטוב פוליטי לפני יציאה. המאפיין הייחודי: דירוגי 1 ו-10 ביחד הורידו את הממוצע ל-1.1.

Of Sound Mind (2015)

דירוג: 9.6, חזוי: 5.15

סרט רומנטי-מתח עם קהל מעריצים קטן וספציפי מאוד. המודל ראה Thriller וחזה נמוך לא יכול לדעת שיש קהל נישה נלהב. המאפיין הייחודי: מעט מאוד דירוגים, כולם 10/10 מקהל מחויב.

Bukunja Tekunja Mitti (2012)

דירוג: 9.1, חזוי: 4.13

סרט אסייתי בז'אנר Action/Horror עם קהל מעריצים נלהב שלא מיוצג ב-IMDb. המודל ראה Horror וחזה נמוך. המאפיין הייחודי: סרט מחוץ לתרבות המערבית קהל ה-IMDb לא מייצג אותו.

מסקנות מניתוח השגיאות

בניתוח הסרטים שעבורם המודל ביצע הערכת יתר (Overprediction) . נמצא כי רבים מהם שייכים לז'אנרים דוקו ודרמה. במקרים אלו המודל חזה דירוגים גבוהים יחסית, בעוד שבפועל הסרטים קיבלו דירוגים נמוכים באופן חריג. ממצא זה מרמז כי המודל למד מהדאטה שסרטים מסוג דוקו ודרמה נוטים לקבל דירוגים גבוהים יותר מהממוצע.

בניתוח הסרטים שעבורם המודל ביצע הערכת חסר (Underprediction) נמצא כי רבים מהם שייכים לז'אנרים Horror־ Thriller . במקרים אלו הסרטים קיבלו בפועל דירוגים גבוהים, אך המודל חזה עבורם דירוג נמוך משמעותית.

ניתוח השגיאות הראה כי עיקר הטעויות הגדולות של המודל נובעות ממקרים חריגים שאינם תואמים את המגמות הכלליות בדאטה. כאשר סרט משתייך לז'אנר שבדרך כלל מקבל דירוג גבוה אך בפועל קיבל דירוג נמוך, המודל נוטה להעריך אותו ביתר. באופן דומה, כאשר סרט משתייך לז'אנר שבדרך כלל מקבל דירוג נמוך אך בפועל קיבל דירוג גבוה, המודל נוטה להעריך אותו בחסר. ממצא נוסף הוא שהמודלים נוטים להתכנס לכיוון הדירוג הממוצע. מכיוון שרוב הדירוגים בדאטה נמצאים בטווח של 5–7, המודלים מתקשים יותר לחזות סרטים בעלי דירוגים קיצוניים במיוחד.

5.3 השוואת המודלים

מבין 20 החריגים של כל מודל, 13 סרטים הופיעו ברשימת החריגים של שני המודלים בעוד שלכל מודל היו 7 חריגים ייחודיים .החפיפה הגבוהה מצביעה על כך שחלק מהטעויות נובעות ממאפייני הסרטים עצמם או ממידע חסר בדאטה, ולא רק מהאלגוריתם.
בנוסף, נבנה גרף פיזור המשווה את שגיאות שני המודלים. רוב הנקודות נמצאו לאורך האלכסון, כלומר כאשר מודל אחד טעה באופן משמעותי, גם המודל השני נטה לטעות באותו כיוון. לכן לא נמצא מצב שבו מודל אחד מצליח באופן עקבי במקרים שבהם המודל השני נכשל.

סוגי שגיאות ייחודים לכל מודל:

ברשימת החריגים הייחודיים ל־ Random Forest הופיעו בעיקר סרטים מז'אנרים כגון Horror, Thriller ו־ Action/Adventure . הסרטים הללו קיבלו בפועל דירוגים גבוהים מאוד (בסביבות 8.5–9.5), אך המודל חזה עבורם דירוגים נמוכים משמעותית. ממצא זה מעיד כי גם כאשר Random Forest מצליח ללמוד קשרים מורכבים בין המשתנים, הוא עדיין מושפע מהמגמות הכלליות בדאטה. בנוסף, נמצא מקרה חריג של סרט מסוג Drama שקיבל בפועל דירוג נמוך מאוד אך המודל חזה עבורו דירוג גבוה יחסית. הדבר מצביע על כך שהמודל למד את המגמה הכללית של סרטי דרמה, אך התקשה לזהות מקרה קיצוני שאינו תואם את דפוס הז'אנר.

ברשימת החריגים הייחודיים ל־ ElasticNet הופיעו בעיקר סרטים מז׳אנרים Action, Comedy, Adventure ו־ Horror. גם כאן מדובר ברוב המקרים בסרטים בעלי דירוגים גבוהים מאוד, בעוד שהמודל חזה עבורם דירוגים בינוניים בלבד. התנהגות זו תואמת את אופיו של ElasticNet כמודל לינארי. המודל נוטה לחזות ערכים הקרובים יותר לממוצע ומתקשה לזהות דפוסים מורכבים או מקרים קיצוניים. בנוסף, נמצא סרט מסוג Documentaryשקיבל בפועל דירוג נמוך מאוד, אך ElasticNet חזה עבורו דירוג גבוה. מקרה זה מחזק את הממצא שהמודל למד כי סרטי Documentary מקבלים בדרך כלל דירוגים גבוהים, ולכן התקשה לזהות מקרה חריג שאינו תואם את הדפוס המקובל.

5.4 הצעות לשיפור

בהתבסס על ניתוח השגיאות, ניתן להציע מספר פיצ׳רים חדשים שעשויים לסייע למודל לזהות טוב יותר סרטים חריגים ולהפחית את השגיאות שנמצאו.

genre_Horror × movie_age

פיצ׳ר המשלב בין השתייכות לז׳אנר Horror לבין גיל הסרט. ניתוח השגיאות הראה כי המודל התקשה במיוחד עם סרטי Horror בעלי דירוג גבוה. ייתכן שסרטי Horror ישנים שעדיין זוכים לדירוגים ולחשיפה ב־ IMDb מייצגים סרטים איכותיים או קלאסיקות בתחום, בעוד שסרטי Horror חדשים רבים מקבלים דירוגים נמוכים יותר. שילוב בין הז׳אנר לבין גיל הסרט עשוי לסייע למודל להבחין בין סוגי סרטי Horror שונים.

genre_Documentary × num_genres

פיצ׳ר המשלב בין הז׳אנר Documentary לבין מספר הז׳אנרים של הסרט. בניתוח השגיאות נמצא כי חלק גדול ממקרי ה־ Overprediction היו סרטים דוקומנטריים בעלי דירוג נמוך במיוחד. פיצ׳ר זה עשוי לסייע להבחין בין סרטים דוקומנטריים בעלי ז׳אנר יחיד, אשר עשויים לפנות לקהל מצומצם יותר, לבין סרטים המשלבים מספר ז׳אנרים ופונים לקהל רחב יותר.

is_east_asian × genre_Comedy

פיצ׳ר המשלב בין סרטים ממזרח אסיה לבין ז׳אנר הקומדיה. בנתוח השגיאות נמצאו מספר סרטים מסוג זה שקיבלו דירוגים גבוהים במיוחד, אך לא זוהו היטב על ידי המודל הלינארי. שילוב זה עשוי לאפשר למודל לזהות דפוסים ייחודיים הקשורים לקבוצות סרטים בעלות מאפיינים תרבותיים משותפים, אשר אינם נלכדים באמצעות הפיצ׳רים הקיימים.

6. ניתוח הוגנות (Fairness Analysis)

מצורף בנספח ה׳

לאחר בחירת מודל Random Forest כמודל המוביל, בוצע ניתוח הוגנות במטרה לבדוק האם קיימות קבוצות סרטים מסוימות שעבורן המודל מציג ביצועים טובים או גרועים באופן משמעותי ביחס לממוצע הכללי.
הניתוח בוצע באמצעות השוואת מדדי RMSE ו־ MAEבין קבוצות שונות של סרטים לפי ז׳אנר, מדינת מקור ועשור יציאה.
תחילה חושבו מדדי השגיאה הכוללים של המודל,
לאחר מכן חושבו אותם מדדים עבור כל אחת מהקבוצות ונבדקו הסטיות ביחס לממוצע הכללי.

ניתוח לפי ז׳אנר

הניתוח בוצע עבור חמשת הז׳אנרים המרכזיים: Drama, Action, Horror, Documentary, Comedy.

ניתוח לפי מדינת מקור

הסרטים חולקו לשתי קבוצות, סרטים שבמקור מארצות הברית וסרטים שאינם מארצות הברית.

ניתוח לפי עשור

המודל נבחן גם עבור סרטים מעשורים שונים: שנות ה־80, 90, 2000, 2010, 2020

6.3 דיון

האם יש חתכים בהם המודל משמעותית פחות מדויק? בכמה?

כן, מהניתוח עולים מספר חתכים שבהם ביצועי המודל נמוכים מהממוצע:

- Horror: RMSE = 1.200 גבוה ב־8% מהממוצע.
- Action: RMSE = 1.206 גבוה ב־8.5% מהממוצע.
- non-US: RMSE = 1.146 גבוה ב־3.1% מהממוצע.
- 2020s: RMSE = 1.242 גבוה ב־11.7% מהממוצע.

לעומת זאת, עבור Documentary השגיאה נמוכה ב־10.7% מהממוצע, ועבור סרטים מארצות הברית השגיאה נמוכה ב־15.1%.

האם יש הסבר לכך?

Horror ו־ Action הם ז׳אנרים הטרוגניים הכוללים מגוון רחב של סרטים, החל מסרטים בעלי תקציב נמוך ועד לסרטים מצליחים ומוערכים במיוחד. כתוצאה מכך, דפוסי הדירוג שלהם פחות עקביים וקשים יותר לחיזוי. לעומת זאת, סרטי Documentary מציגים דפוסי דירוג יציבים יותר ולכן קלים יותר ללמידה.
הביצועים הטובים יותר עבור סרטים מארצות הברית עשויים לנבוע מכך שהם מהווים את החלק הגדול ביותר במאגר הנתונים. עבור סרטים שאינם אמריקאיים קיימים הבדלים תרבותיים והרגלי דירוג שונים, המקשים על המודל להכליל בצורה מדויקת.
בנוסף, סרטים משנות ה־2020 הם סרטים חדשים יחסית, ולכן מספר הדירוגים שנצבר עבורם קטן יותר והדירוגים שלהם עשויים להיות פחות יציבים.

האם זה מעלה חששות לגבי שימוש במודל בפועל?

במידה מסוימת כן. יש להתייחס בזהירות לתחזיות עבור סרטי Horror ו־ Action, סרטים שאינם אמריקאיים וסרטים שיצאו לאחר שנת 2020. בפרט, ניתוח השגיאות הראה כי המודל התקשה במיוחד בסרטי Horror מזרח־אסיאתיים בעלי דירוג גבוה, המשלבים שני חתכים שבהם ביצועי המודל נמוכים יחסית.
לעומת זאת, עבור סרטי Documentary ו־ Drama ביצועי המודל יציבים יותר ולכן ניתן לתת אמון גבוה יותר בתחזיות.

האם Missingness עשויה להיות מנבאת?

כן, ייתכן שקיים קשר בין מידע חסר לבין רמת הדיוק של המודל. סרטים בעלי מידע חסר בשדות כמו מדינה או שפה שייכים לעיתים קרובות לקבוצת הסרטים שאינם אמריקאיים, אותה קבוצה שבה המודל הציג ביצועים חלשים יותר. לכן, חוסר מידע עשוי לשמש אינדיקציה לכך שהתחזית עבור הסרט פחות אמינה, ואף לרמוז על חוסר איזון מסוים בייצוג קבוצות שונות במאגר הנתונים.

<u>מניעת Data Leakage</u>

מניעת Data Leakage הייתה עיקרון מרכזי לאורך הפרויקט.
הסרת עמודות שאינן זמינות לפני יציאת הסרט –ב תוך פונקציית prepare_data הוסרו העמודות numVotes, BoxOffice ,averageRating ו־מכיוון שהן אינן זמינות בזמן החיזוי או מהוות את משתנה המטרה עצמו.
עיבוד הנתונים בתוך ה־ Pipeline – כל שלבי ה־ Preprocessing (Imputation, Scaling ו־ Encoding) בוצעו בתוך ה־ Pipeline. כך כל טרנספורמציה נלמדה בכל Fold על נתוני האימון בלבד, ללא חשיפה לנתוני ה־ Validation.
הפרדה בין Train ל־ Test – פונקציית prepare_data הופעלה בנפרד על סט האימון ועל סט המבחן לאחר ביצוע train_test_split. סט המבחן נשמר להערכה הסופית בלבד ולא שימש במהלך האימון או כוונון ההיפר־פרמטרים.

מסקנות, מגבלות והצעות להמשך:

מסקנות

מודל Random Forest השיג ביצועים טובים יותר ממודל ElasticNet בכל מדדי ההערכה. ב־10 Cross -fold Validation התקבלו ערכים טובים יותר של RMSE, MAE ו־ R^2, וגם על סט המבחן המודל שמר על יתרון עקבי. לכן נבחר Random Forest כמודל הסופי.

ניתוח חשיבות הפיצ'רים הראה כי הז'אנר הוא המשתנה המשפיע ביותר על חיזוי הדירוג. בפרט, סרטי Documentaryנמצאו כבעלי השפעה חיובית חזקה על הדירוג, בעוד שסרטי Horror נמצאו כבעלי השפעה שלילית. בנוסף, פיצ'רים כגון movie_age ו־ runtime_per_genre נמצאו כבעלי תרומה משמעותית לביצועי המודל.

ניתוח השגיאות הראה כי שני המודלים מתקשים בעיקר עם סרטים חריגים שאינם תואמים את הדפוסים המקובלים של הז'אנר שלהם. בפרט, נמצאו סרטי Documentary בעלי דירוג נמוך במיוחד וסרטי Horror ו־ Thrillerבעלי דירוג גבוה במיוחד, אשר גרמו לשגיאות החיזוי הגדולות ביותר.

בנוסף, ניתוח ההוגנות הראה כי המודל פחות מדויק עבור סרטי Horror, Action, סרטים שאינם אמריקאיים וסרטים מהעשור האחרון. לעומת זאת, עבור סרטי Documentary ו־ Dramaהתקבלו ביצועים טובים ויציבים יותר.

מגבלות

- ערך ה־ R^2 הסופי שהתקבל הוא כ־0.26 בלבד, ולכן חלק גדול מהשונות בדירוגים אינו מוסבר על ידי הפיצ'רים הקיימים.
- מאגר הנתונים מוטה לטובת סרטים אמריקאיים, ולכן ביצועי המודל עבור סרטים שאינם אמריקאיים נמוכים יותר.
- מידע חשוב כגון במאים, שחקנים, ביקורות, פרסים וחשיפה תקשורתית אינו קיים במאגר הנתונים, למרות שעשויה להיות לו השפעה משמעותית על הדירוג.
- חלק מהשדות מכילים מידע חסר, ובפרט עמודת התקציב שלא ניתן היה להסתמך עליה באופן מלא.
- המודל מתקשה לזהות סרטים חריגים שאינם תואמים את המאפיינים הרגילים של הז'אנר שלהם.

הצעות להמשך

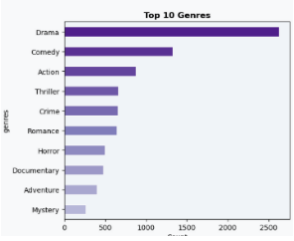
- הוספת מידע על במאים ושחקנים, אשר עשוי לסייע בזיהוי סרטים איכותיים עוד לפני יציאתם.
- יצירת פיצ'רי Interaction שהתגלו במהלך ניתוח השגיאות, כגון genre_Horror × movie_age, genre_Documentary × num_genres ,is_east_asian × genre_Comedy
- ביצוע ניתוח נוסף עבור קבוצות שבהן המודל הציג ביצועים נמוכים יחסית, כגון סרטי Horror, Actionוסרטים שאינם אמריקאיים.
- בדיקת שיטות נוספות לטיפול בערכים חסרים ובחינת השפעתן על ביצועי המודל, במיוחד עבור משתנים שבהם קיימת כמות משמעותית של מידע חסר.

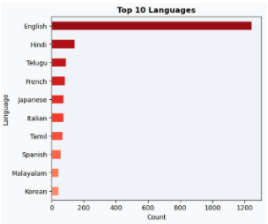
AI Usage Log

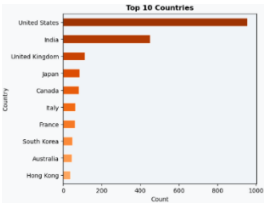
תחום	כיצד נעשה שימוש בAI	כיצד אומת
עיבוד הנתונים	התייעצות על מבנה פונקציית prepare_data והטיפול בערכים חסרים	הרצה ובדיקה על נתוני האימון והמבחן.
בניית המודלים	התייעצות לגבי בחירת פרמטרים ותהליך ההערכה	השוואת ביצועים באמצעות Cross Validation
ניתוח התוצאות	סיוע בפרשנות חשיבות הפיצ'רים, השגיאות והבדלי הביצועים בין המודלים	בדיקת התוצאות מול הגרפים והטבלאות שהופקו.
כתיבת הדוח	סיוע בניסוח הסברים, מסקנות ודיון בתוצאות.	כל הניסוחים נבדקו מול העבודה והתוצאות שהתקבלו.

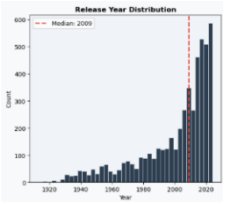
נספחים:

נספח א':

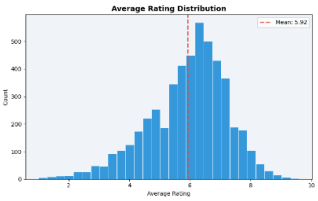


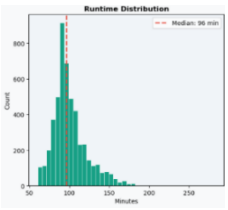






עמודה	חסרים	אחוז חסרים
budget	117,334	87.6%
BoxOffice	115,968	86.6%
Country	80,203	59.9%
Language	78,162	58.4%
plot	74,215	55.4%
averageRating	18,324	13.7%
lead_actors_ids	10,030	7.5%
numVotes	9,122	6.8%
genres	2,573	1.9%
tconst	0	0.0%
primaryTitle	0	0.0%
startYear	0	0.0%
runtimeMinutes	0	0.0%



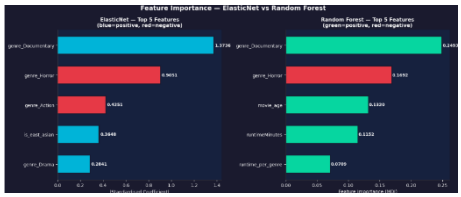


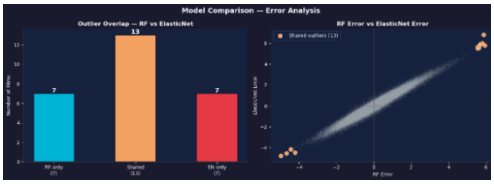
נספח ב':

Model Comparison — 10-Fold CV

Model	RMSE	MAE	R^2
ElasticNet	1.1392	0.8735	0.2252
Random Forest	1.1115	0.8481	0.2624

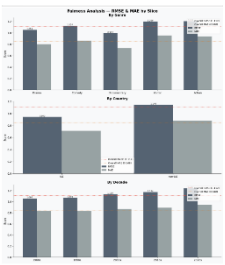
נספח ג':





נספח ד':

נספח ה':



Slice	n	RMSE	MAE	RMSE vs Avg %	MAE vs Avg %
Genre: Drama	46734	1.0509	0.8004	-5.5%	-5.6%
Genre: Comedy	26777	1.1240	0.8623	+1.1%	+1.7%
Genre: Documentary	10400	0.9922	0.7298	-10.7%	-14.0%
Genre: Horror	7995	1.2202	0.9543	+8.0%	+12.5%
Genre: Action	11110	1.2056	0.9371	+6.5%	+10.5%
Country: US	17004	0.9433	0.7116	-15.1%	-16.1%
Country: non-US	75444	1.1462	0.8789	+3.1%	+3.6%
Decade: 1960s	7698	1.0603	0.8318	-4.6%	-1.9%
Decade: 1990s	8150	1.0696	0.8324	-3.8%	-1.9%
Decade: 2000s	14058	1.1384	0.8701	+2.4%	+2.6%
Decade: 2010s	26886	1.1742	0.8909	+5.6%	+5.0%
Decade: 2020s	14691	1.2415	0.9469	+11.7%	+11.6%