דו"ח פרויקט למידת מכונה

1. תיאור הדאטה המעובד ורשימת הפיצ'רים

על מנת ליצור דאטה איכותית , ביצענו את השלבים הבאים :

- **ניקוי וסינון -** ביצענו ניקוי מקיף של ערכים שגויים תוך שימוש בביטויים רגולריים על מנת להבטיח את תקינותם ואחידותם של הנתונים, תוך המרת ערכים בלתי תקינים כמו N\\ או N\A לערכים חסרים בפורמט אחיד - NaN.
- **העשרת הדאטה** - משכנו ממאגרי הנתונים של IMDB את טבלת צוות ההפקה ומיזגנו את נתוני הבמאים לדאטה הגולמית ע"י מזהה סרט ייחודי (tconst)
- **הנדסת פיצ'רים -** על בסיס הנתונים הנקיים, יצרנו משתנים מורכבים.
  הקפדנו על מניעת זליגת מידע (Data Leakage) על ידי חישוב דינמי של פיצ'רים אלו בתוך Pipeline, כך שהם מתבססים אך ורק על נתוני עבר זמינים.
- **התמודדות עם ערכים חסרים -** הסרנו שורות בהן חסר מידע קריטי (שנת יציאה או דירוג סרט).
  עבור נתונים חסרים בפיצ'רים המורכבים, השתמשנו בממוצעי העבר הגלובליים, ועבור יתר המשתנים המספריים ביצענו השלמת ערכים על בסיס החציון, תוך הבטחה שהחישוב מתבצע בהתאם לנתונים הזמינים בנקודת הזמן הרלוונטית ועל בסיס נתוני האימון בלבד.

פירוט עמודות הדאטה המעובד:

#	שם העמודה	סוג משתנה	תיאור המשתנה
עמודות שבחרנו להוסיף מהדאטה הגולמית	tconst	String	מזהה ייחודי עבור כל סרט במערכת IMDb.
	primaryTitle	String	השם הרשמי של הסרט.
	startYear	Float	שנת היציאה הרשמית של הסרט לאקרנים.
	runtimeMinutes	Float	משך הסרט הכולל בדקות.
	genres	List if strings	רשימת הז'אנרים אליהם משויך הסרט.
	actors	List if strings	מזהי השחקנים הראשיים שמופיעים בסרט.
	averageRating	Float	**משתנה המטרה** : הדירוג הממוצע של הסרט ב-IMDb
עמודות שבחרנו להוסיף על מנת להעשיר את הדאטה	directors	List if strings	מזהי הבמאים שבימו את הסרט.

פירוט הפיצ'רים:

שם הפיצ'ר	שם הפיצ'ר בקוד	סוג המשתנה	מוטיבציה	נוסחה
שנת יציאה	startYear	נומרי	תפיסת מגמות כרונולוגיות ושינויים בטעם הקהל לאורך העשורים.	ערך השנה המקורי.

כמות הז'אנרים אליהם משתייך הסרט	num_genres	נומרי / בדיד	סרטים המשלבים מספר ז'אנרים עשויים לפנות לקהל רחב יותר, אך מצד שני, ריבוי ז'אנרים עלול להעיד על חוסר מיקוד עלילתי.	ספירת מספר הז'אנרים ברשימת ה-genres לכל סרט. len(genres)
אורך הסרט – האם האורך שלו קצר מדי או ארוך מדי באופן חריג	is_too_short	בינארי	סרטים קצרים מאוד עשויים להיתפס כחסרים מבחינה עלילתית.	1 אם runtimeMinutes<75, אחרת 0.
	is_too_long	בינארי	סרטים ארוכים מאוד דורשים קשב מתמשך של הצופה . הצופה עלול להתעייף או לאבד עניין מהסרט.	1 אם runtimeMinutes>150, אחרת 0.
האם הבמאי חדש (ללא היסטוריה)	is_new_director	בינארי	בידול במאים מתחילים שאין לגביהם מידע היסטורי מבוסס.	1 אם לבמאי אין סרטי עבר, אחרת 0.
האם הקאסט חדש (כל השחקנים חדשים)	is_new_actors	בינארי	בידול צוות שחקנים ללא היסטוריה קולנועית בדאטה.	1 אם לכל השחקנים אין סרטי עבר, אחרת 0.
יחס אורך הסרט בהשוואה לחציון הז'אנר שלו	relative_runtime_by_genre	נומרי מורכב	ז'אנרים שונים הם בעלי סטנדרט שונה של אורך. סרט אימה של 120 דקות נחשב לארוך מאוד, בעוד סרט דרמה באותו אורך נחשב לקצר. פיצ'ר זה מנרמל את אורך הסרט ביחס למקובל בז'אנר שלו, מה שמאפשר למודל לזהות סרטים שחורגים מהסטנדרט המצופה	חלוקת אורך הסרט בחציון של הסרטים משנים קודמות לפי ז'אנר runtime /median
ציון הדירוג הממוצע לסרטי העבר של הבמאי	director_past_avg	נומרי מורכב	משקף את המוניטין המקצועי של הבמאי כפי שהוא משתקף בדירוגי העבר, ומאפשר למודל ללמוד את השפעת איכות העבר על הפקת סרטים עתידיים	ממוצע הaverageRating של סרטי הבמאי בשנים שקדמו ל-startYear. במידה ואין סרטי עבר המודל ישים NaN.
ניסיון עבר של השחקנים (מקסימום)	max_actors_past_experience	נומרי מורכב	מדד ניסיון של הקאסט - שימוש בערך המקסימלי מבין 5 השחקנים.	מקסימום סרטי העבר של שחקני המפתח.
ממוצע דירוג עבר של במאי לפי ז'אנר	director_genre_past_avg	נומרי מורכב	נותן מדד לכמה הבמאי מקצועי ביחס לז'אנר של אותו סרט	ממוצע הaverageRating של סרטים מאותו ז'אנר בהם השתתף הבמאי. במידה ואין סרטי עבר המודל ישים NaN.

2. <u>פונקציית prepare_data()</u>

פונקציית prepare_data היא פונקציה עיקרית בפרויקט - היא מקבלת DataFrame גולמי ומחזיר DataFrame נקי ומוכן לחיזוי.

שלבי הפונקציה:

<u>שלב מקדים- הוספת עמודת הבמאים</u>

במידה והדאטה שנקבל לא כוללת את עמודת הבמאים , נוסיף אותה באופן יזום.

<u>שלב 1 - פונקציות הניקוי</u>

תחילה בחרנו לבנות פונקציית ניקוי ייעודית עבור כל עמודה :

- clean_year() - בדיקה ע"י ביטוי רגולרי האם הערך בעמודת שנה הוא ביטוי בעל 4 ספרות
- clean_runtime() - בדיקה האם הערך הוא מספר בלבד (ללא אותיות)
- clean_rating() - בדיקה האם הערך הוא מספר בין 1.0 ל10.0
- clean_genres() - הופכת מחרוזות מופרדות בפסיקים לרשימות תקינות, מנקה רווחים מיותרים ומסננת ערכים לא תקינים
- clean_actors() - בדיקה האם הערכים עומדים במזהי IMDb תקינים של שחקנים בפורמט (nm ולאחר מכן 7 או 8 ספרות). היא מחלצת מתוך הערך רק ערכים תקינים ומחזירה אותם כרשימה.
  אם הערך חסר או שלא נמצא אף מזהה תקין, מוחזרת רשימה ריקה [].
- clean_director() – בדיקה האם הערכים עומדים במזהי IMDb תקינים של במאים בפורמט (nm ולאחר מכן 7 או 8 ספרות). היא מחלצת מתוך הערך רק ערכים תקינים ומחזירה אותם כמחרוזת מופרדת בפסיקים.
  אם הערך חסר או שלא נמצא אף מזהה תקין, מוחזר NaN.

<u>שלב 2 - הוספת הפיצ'רים הפשוטים</u>

לאחר הניקוי, הפעלנו את add_static_features להוספת פיצ'רים בסיסיים

כלומר פיצ'רים שמחושבים מהסרט עצמו בלבד, ללא תלות בסרטים אחרים:

- num_genres - ספירת מספר הז'אנרים
- is_too_short - האם הסרט קצר מ-75 דקות
- is_too_long - האם הסרט ארוך מ-150 דקות

הערה : הפיצ'רים המורכבים לא מחושבים כאן אלא בתוך ה Pipeline מה שמבטיח שהחישובים יתבצעו על **סט האימון בלבד**. (בנקודה זו עדיין לא פיצלנו לסט אימון ובדיקה)

<u>שלב 3 – הסרת שורות</u>

בחרנו למחוק שורות שחסר בהן מידע קריטי :

שורות ללא משתנה מטרה (averageRating) - מאחר שמטרת הפרויקט היא חיזוי דירוג הסרט, רשומות שאינן כוללות ציון יעד אינן מספקות מידע לשיפור המודל או להערכת ביצועיו.

שורות ללא שנת יציאה (startYear) - מאחר שמרבית הפיצ'רים המורכבים שבנינו (כגון דירוג עבר של במאי או ותק שחקנים) מתבססים על ציר זמן היסטורי, רשומה ללא שנת יציאה היא חסרת משמעות לוגית.

בחרנו להסיר רשומות אלו במקום לבצע השלמת ערכים , וזאת כדי להימנע מהטיות בחישובים ההיסטוריים ומהסתמכות על נתונים סינתטיים שעלולים להוביל לזליגת מידע.

<u>שלב 4 - הסרת עמודות לא רלוונטיות</u>

הוסרו עמודות שאינן משמשות כפיצ'רים או שמהוות זליגת מיגע

- numVotes, BoxOffice - ידועים לאחר יציאת הסרט

- Language, Country, budget, plot – הפיצ'רים לא מבוססים על עמודות אלה

3. <u>השוואת שני המודלים</u>

בשלב זה ביצענו השוואה בין שני מודלים בעלי תפיסות למידה שונות **ElasticNet** ו **RandomForest** . על מנת להבטיח את אמינות התוצאות, השתמשנו **10-Fold Cross-Validation** המאפשר לנו לבחון את יציבות המודל על פני תת-קבוצות שונות של הנתונים ולמנוע תלות במקריות של פיצול יחיד.

מדדי Elastic Net

R^2: 0.2368 ± 0.0056
RMSE: 1.1293 ± 0.0099
MAE: 0.8637 ± 0.0076

מדדי ה RandomForest

R^2: 0.2956 ± 0.0073
RMSE: 1.0849 ± 0.0099
MAE: 0.8240 ± 0.0077

מניתוח המדדים עולה כי מודל ה **Random Forest**-השיג תוצאות עדיפות על פני ה **Elastic Net**-בכל המדדים שנבדקו:

- מודל ה RandomForest השיג R^2 של כ-0.295 לעומת 0.237 בElasticNet.

העלייה בערך זה מלמדת כי הפיצ'רים המורכבים שבנינו (כמו ממוצע דירוג עבר לפי במאי או יחס אורך הסרט לחציון הז'אנר) מקיימים יחסי גומלין מורכבים ולא-ליניאריים עם דירוג הסרט, אשר מודל ליניארי מתקשה למצות.

- השיפור ב RMSE (ירידה ל-1.08) וב MAE (ירידה ל-0.82) מדגיש את יכולתו של המודל מבוסס העצים לצמצם את טעות החיזוי הממוצעת. הפער המצומצם בין מדדים אלו בשני המודלים מעיד על כך שהשגיאות אינן נובעות מחריגות קיצוניות (Outliers) אלא משגיאה ממוצעת יציבה.
- סטיית התקן הנמוכה בשני המודלים (סביב 0.005 - 0.01) מעידה על כך שהמודלים יציבים מאוד ואינם מושפעים מפיצול מקרי של הנתונים. יציבות זו מחזקת את המסקנה כי תהליך ה Pipeline שבנינו והפיצ'רים שפיתחנו מספקים נתונים עקביים ומהימנים לאורך כל שלבי העיבוד.

4. <u>ניתוח חשיבות הפיצ'רים</u>

- **Elastic Net:** הפיצ'ר החשוב ביותר הוא **ממוצע דירוגי העבר של הבמאי בז'אנר הסרט** ,בפער משמעותי משאר המשתנים. הדבר מצביע על כך שהצלחות קודמות של הבמאי באותו ז'אנר הן גורם מרכזי בחיזוי דירוג הסרט. בנוסף, גם **התאמת אורך הסרט לאורך המקובל בז'אנר וממוצע דירוגי העבר של הבמאי** נמצאו כבעלי השפעה חיובית משמעותית. שני המשתנים הנוספים - **האם כל השחקנים חדשים** ו**האם הסרט קצר באופן חריג** –תרמו אף הם לחיזוי, אך בעוצמה נמוכה יותר. כל חמשת הפיצ'רים המובילים במודל קיבלו השפעה **חיובית** על הדירוג החזוי.

- **Random Forest:** גם במודל זה ,**ממוצע דירוגי העבר של הבמאי בז'אנר הסרט** הוא הפיצ'ר החשוב ביותר בפער גדול, מה שמדגיש את חשיבות הניסיון וההצלחה הקודמת של הבמאי בתחום הרלוונטי. אחריו מופיעים **ממוצע דירוגי העבר של הבמאי** ו**התאמת אורך הסרט לאורך המקובל בז'אנר** ,בדומה למודל Elastic Net בנוסף, המודל מצא חשיבות מסוימת גם ל**שנת היציאה של הסרט** ול**ניסיון המצטבר של השחקנים** .במודל זה ניתן למדוד רק את עוצמת החשיבות של הפיצ'ר ולא את כיוון ההשפעה שלו.

שני המודלים מסכימים באופן ברור על שלושת הפיצ'רים החשובים ביותר :**הצלחת הבמאי בז'אנר הסרט בעבר** , **הצלחתו הכללית בעבר** ,ו**התאמת אורך הסרט לנורמה המקובלת בז'אנר** .הסכמה זו מחזקת את המסקנה כי ניסיון והצלחה קודמת של הבמאי הם הגורמים המשמעותיים ביותר בחיזוי דירוג הסרט.

עם זאת, קיימים גם הבדלים. המודל הelastic מעניק חשיבות גבוהה יותר למאפיינים בינאריים הקשורים לחדשות הקאסט ולאורך חריג של הסרט, בעוד שמודל random forest מייחס חשיבות גבוהה יותר לשנת היציאה ולניסיון המצטבר של השחקנים.

הבדל זה נובע ככל הנראה מכך שהמודל הelastic מזהה קשרים ישירים ופשוטים יותר, בעוד שמודל random forest מסוגל לזהות גם קשרים מורכבים ולא-לינאריים.

5. ניתוח שגיאות Error analysis

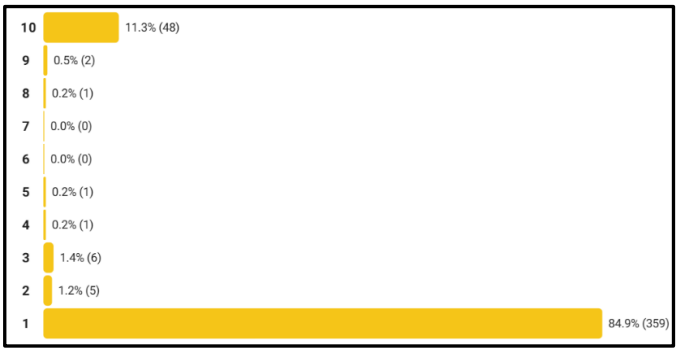
שם הסרט: Justin Bieber: Always Believing (2012) (Overprediction)

פירוט הסרט :סרט תיעודי המלווה את עלייתו המטאורית של כוכב הפופ ג'סטין ביבר לתהילה עולמית, תוך שילוב קטעי הופעות, רגעים אישיים "מאחורי הקלעים" וראיונות עם הצוות המלווה.

השערה לסיבת טעות המודל :המודל חזה לסרט ציון גבוה (7.19) כי הוא זיהה נתונים שמאפיינים הפקה מנוסה ומקצועית - כמו במאים מקצוענים , שחקנים מנוסים וז'אנרים שנחשבים לאיכותיים. בפועל, הציון שניתן לסרט היה נמוך מאוד (1.0).

הטעות של המודל קרתה כי הוא לא יכל לצפות שג'סטין ביבר היא דמות מעוררת מחלוקת, מה שגרם לכך שחלק גדול מהצופים דירגו את הסרט בציון 1 מסיבות אישיות, בעוד המעריצים דירגו בציון 10.

ניתן לראות את ההתפלגות הקיצונית בתמונה הבאה-



שם הסרט: Obama in NC: The Path to History (2010) (Overprediction)

תיאור הסרט: סרט תיעודי פוליטי המלווה את קמפיין הבחירות של ברק אובמה בצפון קרוליינה, תוך התמקדות בהיסטוריה של המאבק נגד גזענות במדינה ובבחירה בנשיא האפרו-אמריקאי הראשון.

השערה לסיבת טעות המודל: המודל חזה ציון גבוה (7.29) משום שזיהה את המעטפת של הפקה דוקומנטרית מקצועית על נושא בעל משקל היסטורי, דבר שהמודל מתרגם אוטומטית לערך של איכות הפקה גבוהה. המודל טעה כי הוא לא יכול היה לצפות שהסרט יהפוך לנושא של ויכוחים פוליטיים בדומה למקרה של ג'סטין ביבר, מדובר בדמות מעוררת מחלוקת, מה שגרם לאנשים לדרג את הסרט לפי הדעה הפוליטית שלהם על אובמה (תמיכה או התנגדות) ולא לפי רמת ההפקה המקצועית שעל פיה המודל העריך את הסרט.

שם הסרט: Son-nim-1 Cheo-beon-jjae I-ya-gi (2011) (Underprediction)

תיאור הסרט: סרט אימה קוריאני העוקב אחר משפחה שמגלה סודות אפלים ונוכחות על-טבעית בביתם החדש. הסרט בנוי על מתח פסיכולוגי גבוה ואווירה מטרידה שמחליפה את ה Jump scares-הקלאסיים המוכרים מהוליווד.

השערה לסיבת טעות המודל: המודל חזה ציון (4.38) בהתבסס על הפיצ'רים שלנו (העדר ניסיון בינלאומי משמעותי של הבמאי והשחקנים בסט האימון). המודל לא "ידע" להעריך את העלייה המטאורית באיכות הקולנוע הקוריאני, שהפך בשנים האחרונות להצלחה מסחררת אצל חובבי הז'אנר. הסיבה שהקהל העניק לו ציון גבוה (9.1) היא שהסרט מצליח לייצר חוויה רגשית ופחד אמיתי בצורה כל כך אפקטיבית, שהיא עוקפת את הצורך ב"שמות גדולים" שמופיעים בפיצ'רים שלנו. המודל הסתמך על העדר המוניטין של הצוות כסימן לחולשה, בעוד שהקהל זיהה את הייחודיות של הסיפור והביצוע הקוריאני כחוזקה יוצאת דופן.

שם הסרט: (Underprediction) Oscuro by Rene Potter (2016)

תיאור הסרט : סרט אימה פסיכולוגי המשתמש באווירה אפלה וסיפור לא ליניארי כדי לבנות מתח, תוך דגש על צילום אמנותי.

השערה לסיבת טעות המודל : המודל חזה ציון נמוך (4.08), בעוד הציון בפועל היה גבוה (8.6). המודל התבסס על פיצ'רים של ניסיון שחקנים וצוות שהיו נמוכים בסט האימון שלנו, ולכן הניח שמדובר בהפקה חובבנית. הטעות לדעתנו נובעת מחוסר היכולת של המודל להבחין באיכות של סרטי אינדי .הקהל זיהה כאן יצירה אמנותית ייחודית שהצליחה לחרוג מהמגבלות הטכניות של ההפקה, בעוד המודל שלנו "נענש" את הסרט על בסיס נתוני עבר טכניים של הצוות, מבלי להבין שבתחום האימה, סיפור חזק ובימוי יצירתי יכולים לעיתים לעקוף לחלוטין את מדדי הוותק המקצועי שהזנו למודל.

שם הסרט: (Underprediction) One Night Out (2018)

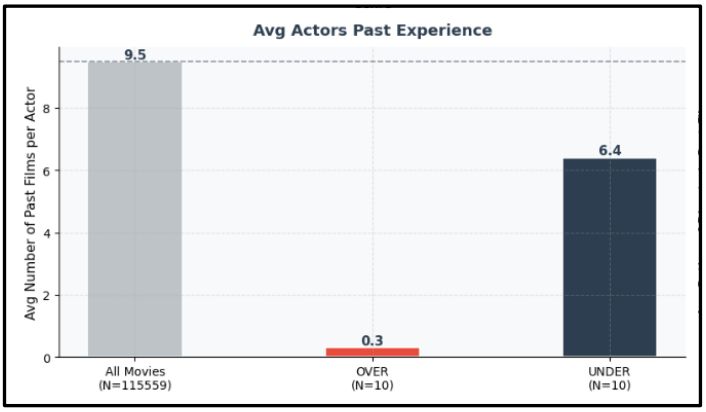
תיאור הסרט :סרט מתח הבנוי על סיפור מקורי, קצב מהיר ותפניות עלילתיות שמצליחות להפתיע את הצופה לכל אורכו.

השערה לסיבת טעות המודל :המודל חזה ציון (4.60) בעוד הציון בפועל הוא גבוה מאוד (9.0) המודל הסתמך על העדר של ניסיון מקצועי בקרב הצוות . הכישלון של המודל הוא תוצאה של חוסר יכולת למדוד את איכות התוכן. הקהל העניק ציון גבוה כי הסרט הצליח לייצר חוויית מתח אינטנסיבית ומהנה. תוכנו של הסרט והתפניות העלילתיות הן אלה שגרמו לדירוג כל כך גבוה בקרב הצופים.

<u>זיהוי דפוסים משותפים</u>

- **שנת יציאה :** ניתן לראות כי 95% מה Outliers הם סרטים שיצאו לאחר שנת 2006.
  דבר שעשוי להעיד כי המודל מתקשה יותר לזהות גורמים המשפיעים על דירוג סרטים בעידן המודרני. ככל שעובר הזמן - הציפיות של הקהל, האופן שבו סרטים מופצים (סטרימינג לעומת קולנוע) והדרך שבה הקהל מדרג אותם ב IMDb השתנו דרמטית.
  המודל לא מתאים את עצמו לעולם הקולנוע המודרני, שבו הדינמיקה של רשתות חברתיות והתפיסה החברתית לגבי דברים שונים השתנו מאוד , מה שמשנה את כללי המשחק.

- **ניסיון השחקנים :** ניתן לראות שהערכים הממוצעים של הOutliers נמוכים משמעותית ביחס לממוצע של כלל הסרטים**.** ממצא זה מצביע על כך שהמודל מציג רגישות גבוהה נסיון השחקנים. כאשר המודל נדרש לחזות דירוג של סרטים המאופיינים בצוותים ללא רקורד עשיר (שחקנים חדשים), המודל אינו מצליח לבצע הערכה מדויקת והסטייה מהדירוג האמיתי עולה.
  עם זאת , ניתן להבחין שיש הבדל גם בין סרטי הunder לover .
  אמנם שניהם מתחת לממוצע של כל הסרטים אך הunder גבוה משמעותית מהover ששם הממוצע כמעט אפסי.



בגרף ניתן לראות את ההבדל המהותי בין הOutliers לבין כל הסרטים .

בעוד שממוצע הותק של השחקן הותיק בקרב כל הסרטים עומד על 9.5 ,
ממוצע הותק של השחקן הותיק בסרטי הover והunder עומדים על 6.4 ו0.3!

השוואה בין Overpredictions לעומת Underpredictions

הבדל ז'אנר : בניתוח הOverpredictions זוהתה מגמה של רוב סרטים מז'אנר דוקומנטרי (70%).

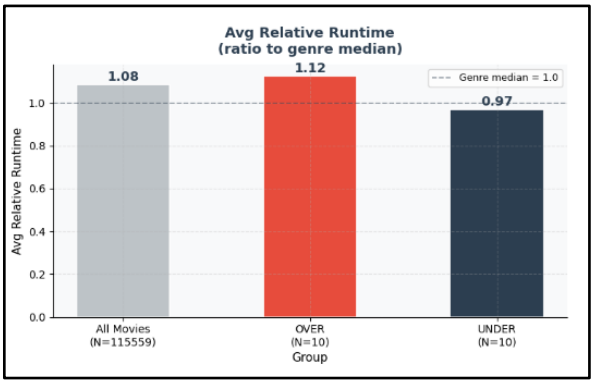
המודל נוטה לתת לסרטים הדוקומנטריים ציונים גבוהים, כנראה מתוך הנחה שסרט תיעודי הוא יצירה בעלת ערך לימודי או הפקתי גבוה. בפועל , סרטים דוקומנטריים על דמויות ציבוריות עלולים לעורר מחלוקת מה שמוביל לדירוגים נמוכים מאוד שניתנים על ידי קהל שלא בהכרח מדרג את איכות הסרט אלא מביע דעה אישית, רגשית או פוליטית.

ואילו בקבוצת הUnderpredictions ניתן לראות מחצית של סרטים מז'אנר המותחן (Thriller) והאימה(Horror).

המודל תייג את הסרטים הללו כבעלי פוטנציאל נמוך, כנראה בגלל שרובם מגיעים מהפקות אינדי, קולנוע זר או במאים פחות מוכרים אך בפועל הם מצליחים לייצר חוויה רגשית או פחד אפקטיבי מה שגרם לדירוג גבוה.

הבדל יחס אורך הסרט לחציון:

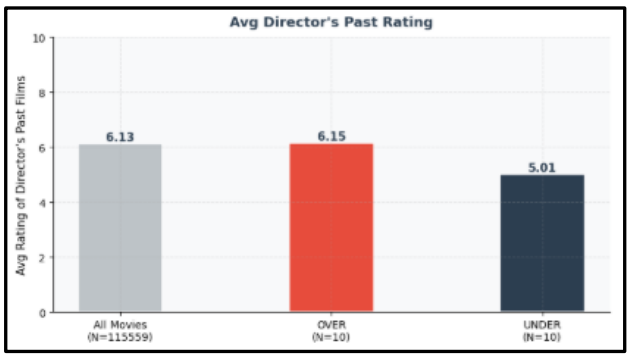
בממוצע, סרטי ה Overprediction מציגים Runtime הגבוה מהחציון עפ"י ז'אנר .לעומת זאת, סרטי ה Underprediction מציגים בממוצע Runtime כמעט שווה לחציון עפ"י ז'אנר .



ניתן לראות בגרף שסרטי הover ארוכים ב12% מהחציון

ואילו סרטי הunder נמוכים רק ב3% מהחציון

דירוג ממוצע של במאים :



דווקא מהבחינה הזו , אין הבדלים מהותיים בין ה - Overpredictions ל- Underpredictions.
ניתן לראות בגרפים שציון ממוצע דירוג של הבמאים הוא מאוד דומה .

השוואה בין המודלים

נמצא כי מתוך 20 ה Outliers-של כל מודל **10 סרטים הם משותפים לשניהם** .המשמעות היא שמחצית מהשגיאות הקיצוניות של המודלים הן "אובייקטיביות" - מדובר בסרטים בעלי מאפיינים שקשים לחיזוי עבור כל אלגוריתם (למשל, סרטים דלי נתונים או סרטים עם סנטימנט קהל קיצוני שאינו קשור לפיצ'רים טכניים).

מהשוואת השגיאות עולה שאין עדיפות מוחלטת לאף אחד מהמודלים - כל אחד מהם מתמודד טוב יותר עם סוגים שונים של סרטים.

היעדר החפיפה המלאה ב Outliers-מעיד על כך שהמודלים לומדים דפוסים שונים בנתונים.

סוגי שגיאות ייחודיים לכל מודל :

מניתוח פיצ'ר ניסיון השחקנים לפי מודל הגענו לתובנות הבאות-

- **השגיאות הייחודיות ל Random Forest**

במאים עם היסטוריה חזקה: ה Outliers הייחודיים של Random Forest מאופיינים בדירוג עבר ממוצע גבוה יותר של הבמאי (5.82 לעומת 4.72 בelastic). ההשערה היא שהמודל נותן משקל רב להצלחות העבר של הבמאי ולכן מתקשה במקרים שבהם במאי מצליח במיוחד מוציא סרט שקיבל דירוג נמוך מהצפוי, או להפך.

- **השגיאות הייחודיות ל Elastic Net**

שחקנים בעלי ניסיון רב: ב Outliers-של Elastic Net מספר סרטי העבר הממוצע של השחקנים גבוה במיוחד. (10.9 לעומת 5.2 בrandom forest). ההשערה היא הקשר בין ניסיון השחקנים לדירוג הסרט כנראה אינו לינארי. Random Forest מצליח לזהות דפוסים מורכבים יותר, בעוד Elastic Net מתקשה כאשר מדובר בקאסט ותיק במיוחד.

סרטים ישנים מאוד: ב Outliers-של Elastic Net מופיעים סרטים משנות ה-40, ה-50 וה-60, בעוד שכמעט אינם מופיעים ב Outliers של Random Forest. ההשערה היא ש Random Forest מצליח ללכוד טוב יותר השפעות תקופתיות ושינויים היסטוריים בטעם הקהל, בעוד ש Elastic Net מתקשה במקרים חריגים מתקופות רחוקות.

הצעות לשיפור

- **סכום עוקבים ברשתות החברתיות:** ראינו שרוב הטעויות של המודלים מתרכזות בשנים האחרונות. הסיבה לכך היא שהמודלים שלנו מסתמכים על נתוני עבר אבל בעידן המודרני, הצלחה של שחקן מושפעת מכמה הוא מוכר וכמה יש סביבו "הייפ" ברשתות החברתיות.
ההצעה שלנו היא לסכום את מספר העוקבים של השחקנים ברשתות נבחרות כמו אינסטגרם/ פייסבוק/ טוויטר/ טיקטוק.

- **מספר הפרסים והמועמדויות לפרסים של הבמאי / עבור כל שחקן -** הוספת נתון על זכיות או מועמדויות לפרסים יוקרתיים תספק למודל מדד ברור יותר לרמת הסטנדרט וההכרה המקצועית של הצוות.
פיצ'ר זה יאפשר למודל להבחין בין שחקן 'מנוסה' לבין שחקן 'מוערך', ובכך לשפר את יכולת החיזוי עבור הסרטים.

- **תקציב הפקה:** אחת המגבלות במודל הנוכחי היא היעדר נתוני תקציב**.**
תקציב ההפקה מהווה אינדיקטור משמעותי לעוצמת המשאבים, רמת ההפקה והחשיפה השיווקית של הסרט, ולכן הוא נחשב למשתנה בעל יכולת ניבוי גבוהה להצלחת סרט.
אולם, בחרנו להסיר משתנה זה מהמודל מכיוון כמעט כ-90% מהערכים בעמודה היו חסרים, מה שהיה פוגע משמעותית במהימנות התחזית ויוצר הטיות בתוצאות .
מה שחשוב לשים לב הוא לעשות **היוון** לסכומי התקציב, כי מליון דולר של 1990 , לא שווה מיליון דולר של 2026!

6. ניתוח הוגנות

ניתוח חתכים

- ירידה בדיוק לאורך זמן :
ניתן לראות מגמה ברורה: ככל שהסרט חדש יותר, השגיאה (RMSE) גדלה. בשנות השלושים , השגיאה הייתה נמוכה ב **26.4%** מהממוצע, בעוד שבעשור האחרון , השגיאה גדלה ל **11.9% מעל הממוצע**.
הסבר : ככל שאנחנו מתקרבים לימינו, המודל הופך לפחות מדויק. זה מחזק את הטענה שהנתונים ה"קלאסיים" (כמו ותק) כבר לא מספיקים כדי לחזות הצלחה של סרטים מודרניים.
- הבדלים בין ז'אנרים:
המודל מצטיין ב :Romance- השגיאה נמוכה ב **9.2%** מהממוצע. זה מרמז שהמודל מבין טוב מאוד את הדפוסים בז'אנר הזה (כנראה כי הוא יותר צפוי או מבוסס על נוסחאות קבועות).
המודל מתקשה ב :Action- השגיאה גבוהה **6.9%** מהממוצע. ז'אנר האקשן הוא נקודת התורפה של המודל - כנראה כי ההצלחה בו תלויה יותר בפיצ'רים שחסרים לנו (כמו תקציב, אפקטים או שחקנים מובילים) מאשר בפיצ'רים שהזנו למודל.

במה המודל פחות אמין : לא היינו משתמשים במודל הזה כדי לתת תחזית להפקות אקשן חדשות .
במקרה כזה, השילוב של ז'אנר בעייתי (Action) ותקופה בעייתית (סרטים חדשים) יוצר רמת אי-ודאות גבוהה מדי. התחזית שלו לא תהיה אמינה מספיק למקבלי החלטות.

Missingness: בשלבי הניתוח הראשוניים, עלתה השערה כי החוסר בנתוני תקציב עשוי לשמש כמנבא בפני עצמו - למשל, כאינדיקטור להפקות דלות תקציב או להסתרה מכוונת של נתונים (הפקות לא רוצות לדווח שאין להם תקציה גדול). לכן, היה חשש כי המודל יפתח הטיה (Bias) במקום ללמוד את השפעת התקציב על איכות הסרט, הוא עלול היה ללמוד שחוסר בנתונים הוא מצביע לאיכות ירודה, וזאת ללא קשר למציאות.
עם זאת, לאחר העמקה בנתונים, הבנו שהנחת העבודה הזו אינה מתבטאת במקרה שלנו. הנתונים נמשכו ממקורות שונים כחלק מעבודת הכיתה, ושיעור החוסרים הגבוה (כ-90%) מעיד כי החוסר אינו נובע מסוג הסרט, אלא משיטות איסוף נתונים (Data Scraping) לא אחידות.

לפיכך, בחרנו להסיר משתנה זה מהמודל. הסרה זו אינה מעידה על כך שהתקציב אינו משתנה חשוב, אלא שהנתונים כפי שהם משתקפים בדאטה הנוכחי אינם עומדים ברף המהימנות הנדרש לבניית מודל מדויק.

7. <u>מניעת זליגת מידע</u>

מניעת זליגת מידע הייתה שיקול מרכזי בתכנון תהליך העבודה, במטרה להבטיח שביצועי המודל ישקפו יכולת חיזוי אמיתית ולא תוצאה של חשיפת המודל למידע שאינו אמור להיות זמין לו בזמן אמת.
להלן הצעדים שנקטנו להבטחת תקינות התהליך:

- **חלוקה מוקדמת לסט אימון ומבחן:** פיצול הנתונים לסט אימון (80%) וסט מבחן (20%) בוצע בשלב ראשוני, בטרם בוצע כל עיבוד סטטיסטי או חישוב על הנתונים. פעולה זו הבטיחה שסט המבחן נותר חיצוני לתהליך הלמידה לכל אורך הדרך.

- **הקפדה על חישובים מבוססי עבר :**בעת יצירת משתנים הקשורים לזמן, כגון ותק של שחקן או במאי, הקפדנו להשתמש אך ורק במידע שהיה זמין בעת הפקת הסרט. וידאנו כי כל נקודת זמן המשמשת לחישוב מוקדמת לשנת יציאת הסרט, כדי למנוע שימוש במידע עתידי שאינו רלוונטי לזמן אמת.
- **ניהול תהליכי עיבוד (Pipeline) :** כל שלבי עיבוד הנתונים , מילוי ערכים חסרים ונרמול - בוצעו בתוך מבנה של ה Pipeline.החישובים הסטטיסטיים הנדרשים לביצוע פעולות אלו התבססו על נתוני האימון בלבד. נתוני המבחן עברו את אותן טרנספורמציות על בסיס הפרמטרים שנגזרו מסט האימון, מבלי שהתפלגותם תהיה חשופה למודל.

- **תיקוף מודל (Cross-Validation):** תהליך ה Cross-Validation בוצע באופן שבו ה Pipeline חושב מחדש בכל קבוצת אימו .גישה זו מבטיחה שהערכים הסטטיסטיים מחושבים בכל פעם מחדש רק על בסיס הנתונים הזמינים לאותה קבוצה, מה שמונע חדירת מידע חיצוני לתהליך הערכת המודל.

8. <u>מסקנות, מגבלות והצעות להמשך</u>

מסקנות:

- ממוצע דירוגי העבר של הבמאי באותו ז'אנר נמצא כפיצ'ר החשוב ביותר בשני המודלים.
- **השינוי הדינמי בשוק :** הניתוח מראה ירידה מובהקת בדיוק המודל ככל שהסרטים חדשים יותר (2024 לעומת 1984). המסקנה היא שהצלחה של סרט בעידן המודרני מבוססת פחות על "ותק" קלאסי ויותר על גורמים חיצוניים (הייפ, רשתות חברתיות).
- **ז'אנרים כסמן להצלחה :** זיהינו כי המודל מצטיין בחיזוי סרטי רומן אך כושל באקשן - זה מעיד על כך שסרטי אקשן תלויים בפרמטרים מורכבים (תקציב, אפקטים, מותג) שהמודל הנוכחי לא רואה, בעוד בז'אנרים אחרים יש נוסחה פשוטה יותר.

מגבלות:

- **פער במידע מודרני :**המודל נשען על נתונים היסטוריים ומתקשה להתמודד עם סרטים מודרניים, שם התווספו עוד מתשנים לדשים שכנראה השפיעו על הדירוג.
- **רעש בנתונים :** איכות המידע שנמשך (כמו חוסר בנתוני תקציב) הגבילה את היכולת לבנות מודל אחיד לכל התקופות.

הצעות להמשך:

- הוספת הפיצ'רים שהצענו - עוקבים ברשתות , מועמדויות וזכיות בפרסים והשלמת ערכי התקציב.
- בכל ז'אנר, גורמים שונים קובעים את ההצלחה.
  לכן המלצתינו היא לבנות מודלים נפרדים לז'אנרים שונים (למשל מודל ייעודי לאקשן), במקום מודל גנרי אחד שלא תופס את השונות בין סוגי הסרטים.

9. AI Usage Log

במהלך הפרויקט נעזרנו ב LLM ככלי עזר טכני לפתרון בעיות נקודתיות בתהליך הפיתוח והניתוח:

דיבוג וניתוח שגיאות : המודל סייע לנו באיתור שורש הבעיה בעת היתקלות בשגיאות ריצה.

חקר נתונים מעמיק : נעזרנו במודל כדי לבצע ניתוחים איכותניים על סרטים ספציפיים שבהן המודל הציג סטייה גבוהה מהממוצע.

סיוע בכתיבת קוד מורכב : נעזרנו במודל לבניית פונקציות מורכבות לטיפול בנתונים (כמו ביצוע טרנספורמציות מבוססות זמן על פיצ'רים היסטוריים או עיצוב הגרפים)

הקוד הסופי, המודלים שנבחרו והפרשנות הסופית לתוצאות הם תוצר של עבודה עצמאית שלנו, המבוססת על ניתוח עקבי של הנתונים לאורך כל שלבי הפרויקט.