דו״ח סחר אלקטרוני

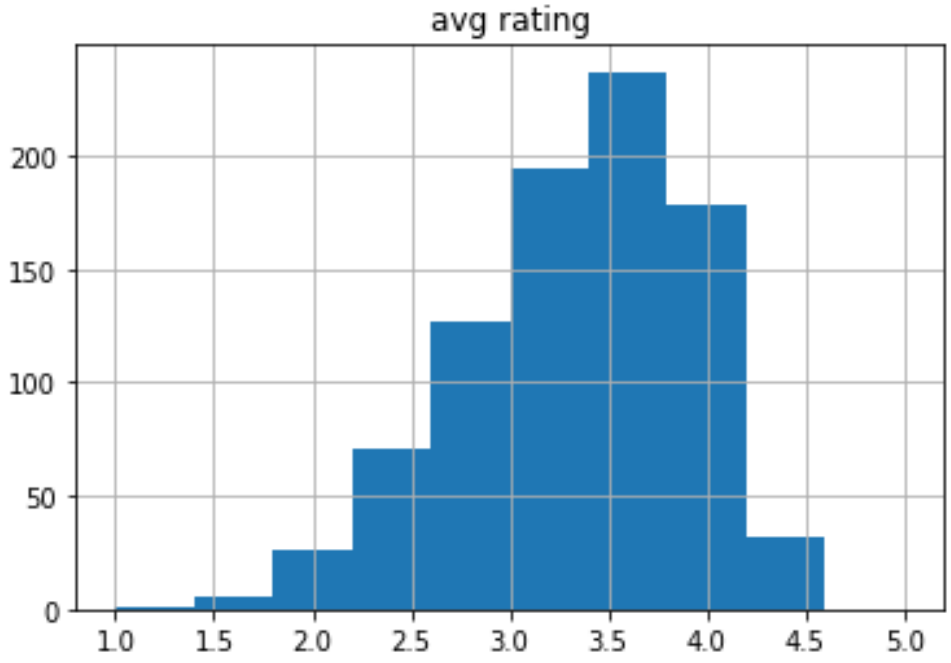
מבוא:

התעסקנו בפרויקט עם אוסף של נתונים שריכזו את העדפותיהם של אנשים שונים לסרטים שונים.
המידע שהיה בדאטה סט שניתן לנו היישויות העיקריות שאותן חקרנו היו האנשים והסרטים.
על האנשים ידענו פרטים מאוד בסיסיים (מין, עיסוק, גיל והמיקוד שלהם).
על הסרטים ידענו קצת יותר פרטים, כגון שמות הסרטים, האתרים שלהם, והז'אנרים השונים שאליהם השתייכו הסרטים.
הדאטה סט שקיבלנו הכיל 943 אנשים ו-1682 סרטים, ומכאן שיכולנו לקבל עד כ- 1.5 מיליון דירוגים של משתמשים לסרטים.
בפועל, מספר הדירוגים הקיימים בדאטה סט שקיבלנו הוא 100,000 (ומכאן שמו ml-100k).
ניתן להבין אם כן כי מאגר הנתונים שקיבלנו לא היה שלם , וכי היינו צריכים להשלים בעצמנו את הדירוגים החסרים, כלומר לממש אלגוריתם למערכת המלצה.

בשלב הראשון, ביצענו data exploration , כלומר רצינו "ללכלך את הידיים" ולהבין אילו נתונים קיבלנו ולנתח אותם.
בשלב השני, כתבנו מודל בסיסי יחסית שנותן לאדם מסוים, המלצות לסרטים על סמך נתונים כלליים על הסרטים שאינם קשורים לנתונים האישיים של האדם עצמו (המלצה לא אישית).
בשלב השלישי, פיתחנו אלגוריתם למערכות המלצה על סמך המלצה אישית. השווינו בין המודלים השונים שהוצגו בכיתה באמצעות מטריקת ה MAE והמלצנו לעבוד עם המודל שנתן את ה MAE הנמוך ביותר מבין כל המודלים שחקרנו.
בשלב הרביעי, של הפרויקט, השתמשנו גם בטכניקות מ Deep Learning ובנינו מערכת לחיזוי רייטינג לפי neural collaborative filtering שהוצג בכיתה. שינינו את הפרמטרים השונים של המודל ובדקנו מי מהם מספק את התוצאה הכי טובה (גם כאן השתמשנו במטריקת MAE).
בשלב האחרון של הפרויקט, שאפנו לקחת את כל מה שגילינו ולנסות לבנות מודל משודרג אף יותר שמשתמש גם בטכניקות מ Deep Learning וגם על סמך מידע אישי.

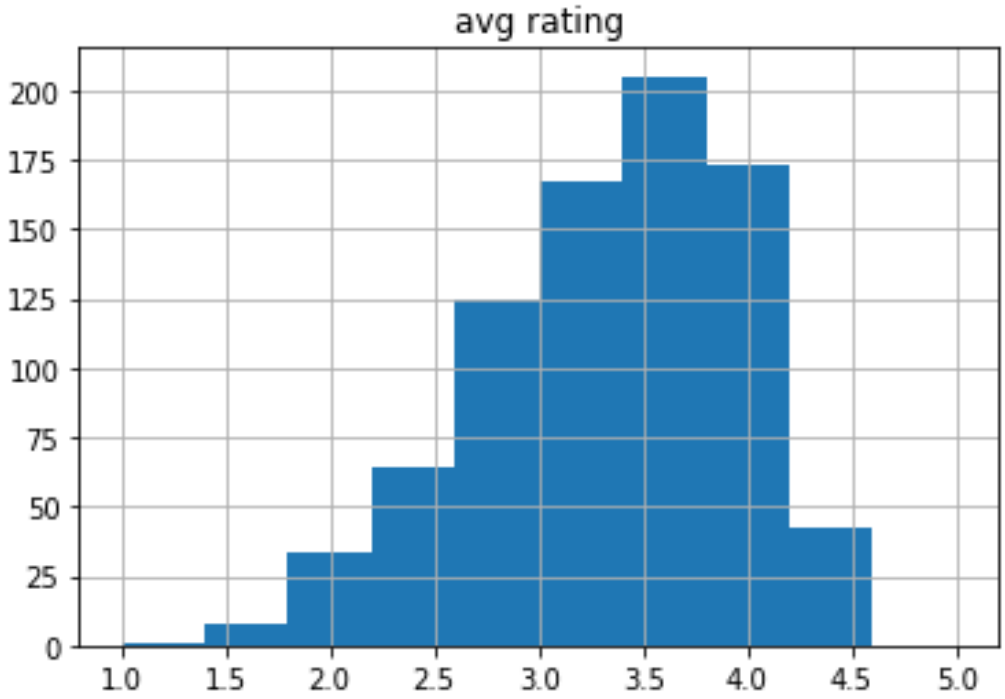
תיאור הדאטה והצגת התוצאות:

כפי שתיארנו במבוא, היו חסרים לנו נתונים באופן יחסי. ה- sparsity=94% בדאטה סט שקיבלנו, כלומר היו לנו מעט נתונים באופן יחסי שיכולנו להסתמך עליהם.
בשלב הראשון, נתבקשנו לנתח את הדאטה שקיבלנו על האוכלוסייה כולה, ולהתמקד גם בחלוקת האוכלוסייה לפי מין ולבדוק את התפלגות המידע גם באוכלוסיית הגברים וגם באוכלוסיית הנשים.
לצורך כך ניקינו מהדאטה הכללי את כל הרשומות בהן מופיעים סרטונים שדורגו מספר נמוך של פעמים (פחות מ- 0.001% מכלל הדירוגים, כלומר סרטים שדורגו פחות מ-9 פעמים לא נכנסו כלל לדאטה שחקרנו).
הממצאים שלנו על כלל האוכלוסייה הם שסרטי קומדיה ואנימציה נוטים להיות הסרטים האהובים ביותר (2 מתוך 3 הסרטים בעלי הרייטינג הממוצע הגבוה ביותר). גילינו כי האוכלוסייה נוטה לדרג באופן ממוצע סרטים בצורה מתונה ולא קיצונית (מרבית מהסרטים קיבלו רייטינג של 3-4 בערך כפי שניתן לראות בגרף)



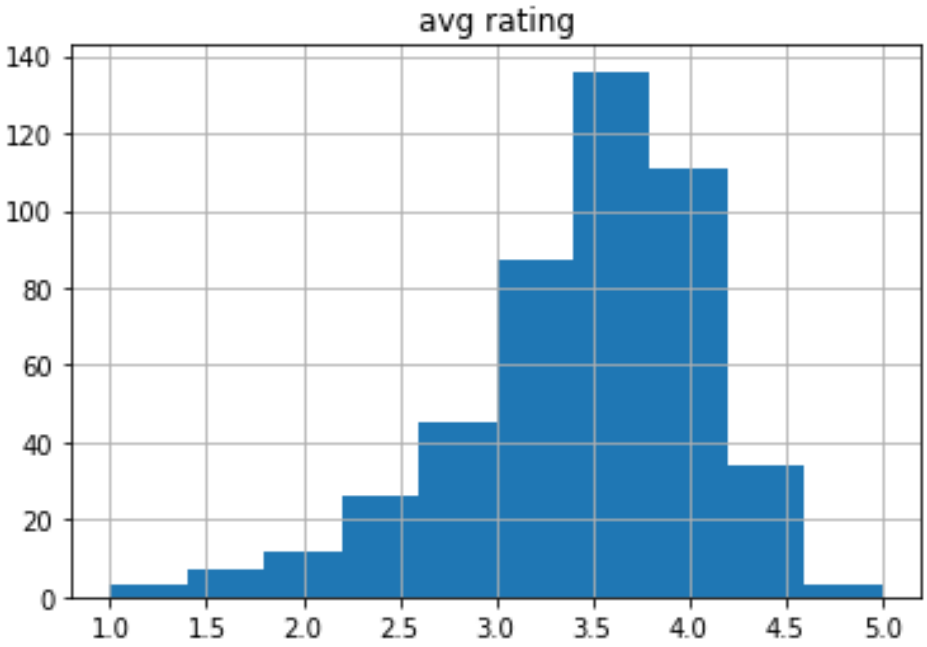
[תרשים 1 : התפלגות הדירוג הממוצע של סרטים לפי כלל האוכלוסייה]

גם אצל אוכלוסיית הגברים מתקיים ש-2 מתוך 3 הסרטים בעלי הרייטינג הממוצע הגבוה ביותר הם סרטי קומדיה, אך ניכר כי ישנם הרבה יותר סרטים שהיו בעלי דירוג ממוצע נמוך מאוד, כמתואר בגרף שלמטה



[תרשים 2 : התפלגות הדירוג הממוצע של סרטים לפי אוכלוסיית הגברים]

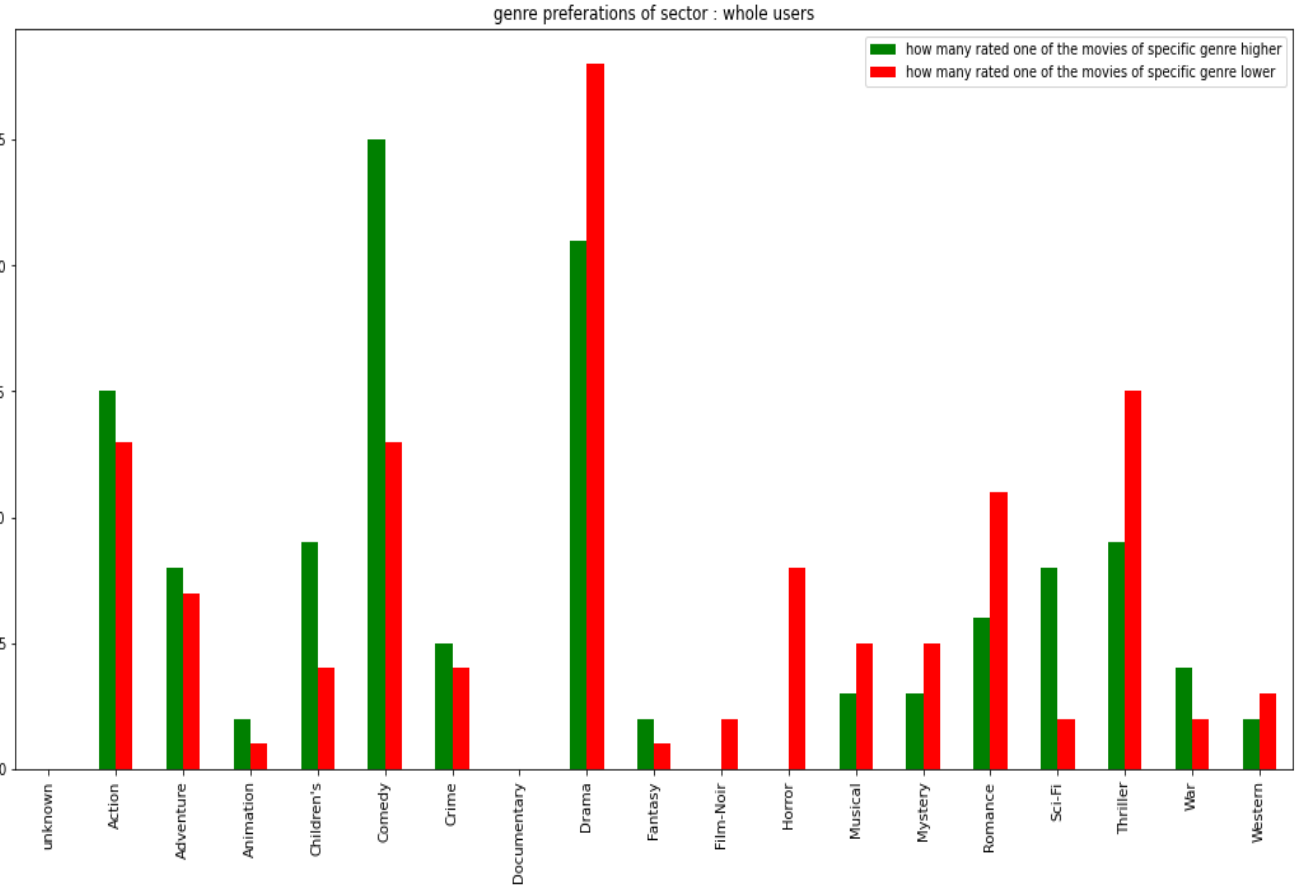
בניגוד לאוכלוסייה הכללית ולאוכלוסיית הגברים, אצל אוכלוסיית הנשים 2 מ-3 הסרטים בעלי הדירוג הגבוה ביותר הם סרטי דרמה.



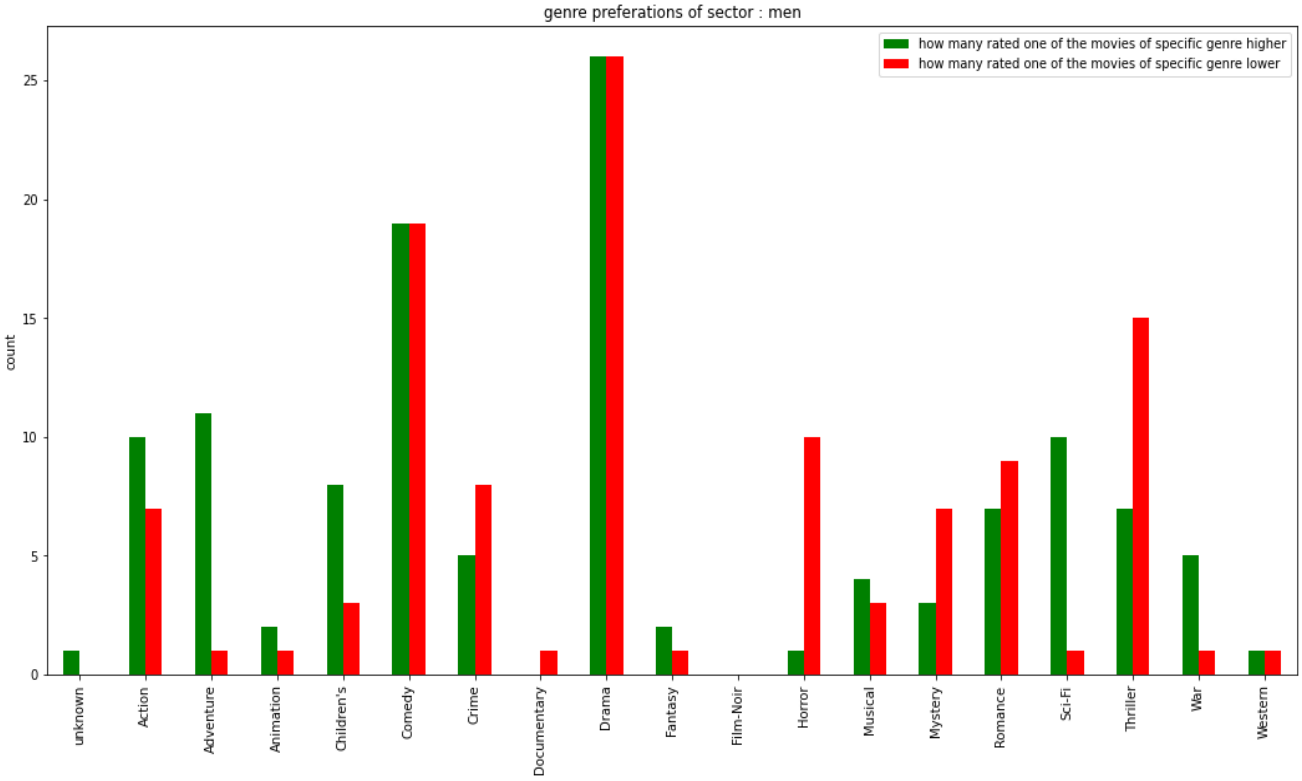
[תרשים 3: התפלגות הדירוג הממוצע של סרטים לפי אוכלוסיית הנשים]

בהמשך, נתבקשנו לחקור את התפלגות הקטגוריות של הסרטים הטובים ביותר והגרועים ביותר (לפי דירוג ממוצע). בחרנו לקחת את העשירון העליון והתחתון של הסרטים לפי דירוג הממוצע. לאחר מכן בחרנו לקחת מתוך כל אחת מהקבוצות הללו, רק את הסרטים שדורגו ע"י יותר מ- 0.0005% מהאוכלוסייה, על מנת לקבל נתונים אמינים יותר שמתבססים על דירוג של כמות נכבדת יותר של אנשים ולא על מספר אנשים מצומצם.

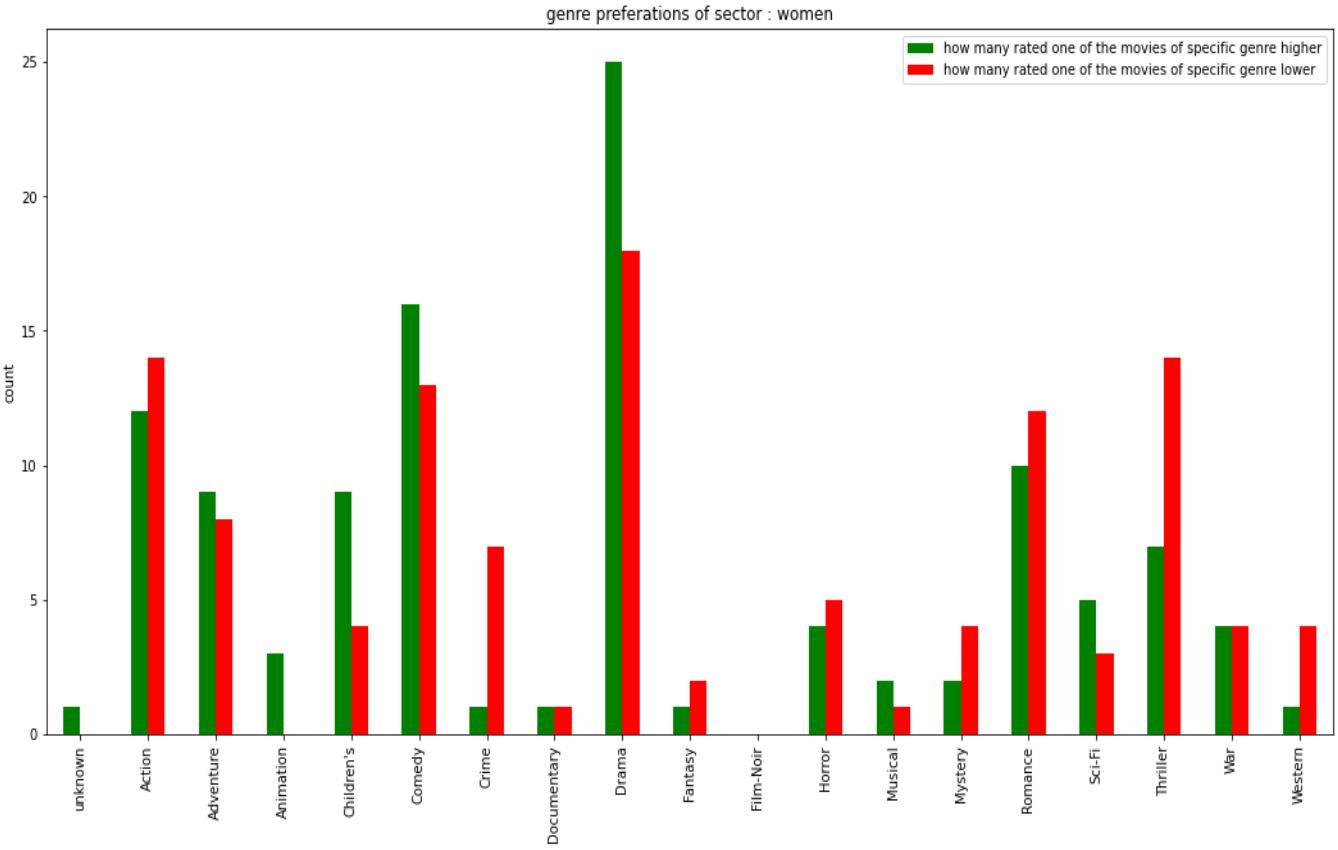
להלן התוצאות שאספנו על כלל האוכלוסייה, אוכלוסיית הגברים ואוכלוסיית הנשים:



[תרשים 4: התפלגות הקטגוריות של סרטים בעלי דירוג גבוה ונמוך ביותר עבור כלל האוכלוסייה]



[תרשים 5: התפלגות הקטגוריות של סרטים בעלי דירוג גבוה ונמוך ביותר עבור אוכלוסיית הגברים]



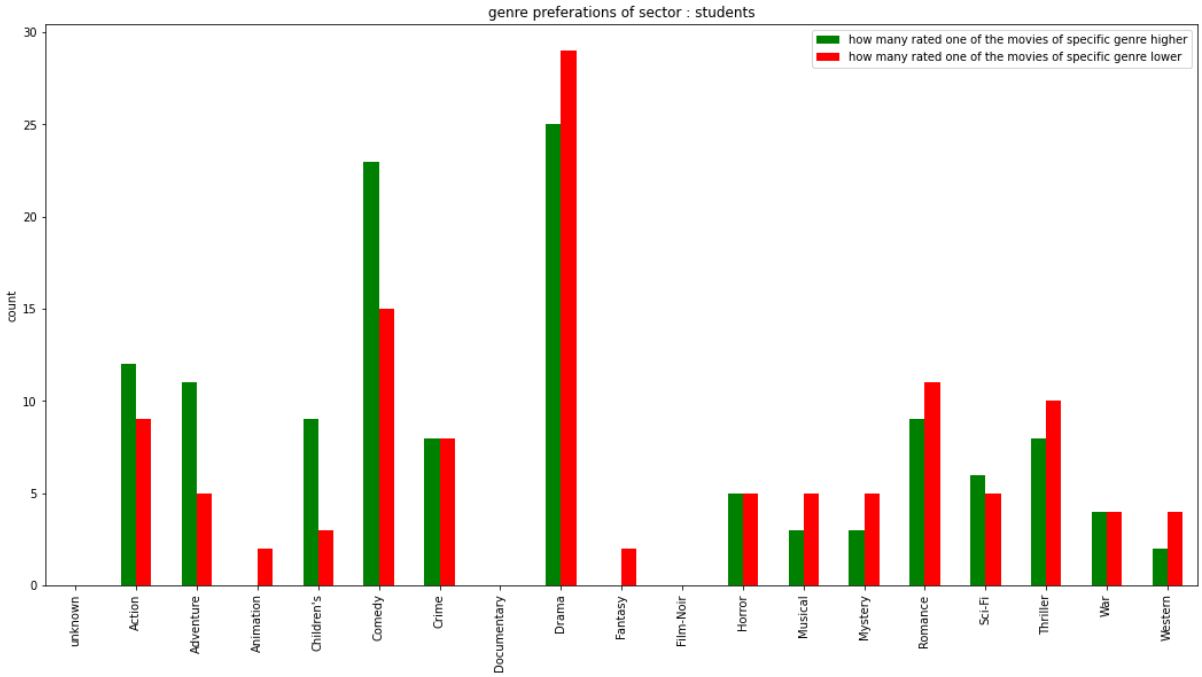
[תרשים 6: התפלגות הקטגוריות של סרטים בעלי דירוג גבוה ונמוך ביותר עבור אוכלוסיית הנשים]

כמו כן, בדקנו מי הם שלושת הסרטים בעלי הפערים הגבוהים ביותר בדירוג הממוצע בין אוכלוסיית הגברים ואוכלוסיית הנשים. התוצאות שקיבלנו מחזקות את התמונה שהתקבלה מתרשימים 4-6. 3 הסרטים בעלי הפער הגדול ביותר הם כולם סרטי דרמה - נשים דירגו סרטים אלו דירוג גבוה ואילו הגברים היו קצת יותר ביקורתיים כלפיהם. ראו תמונה מצורפת למטה.

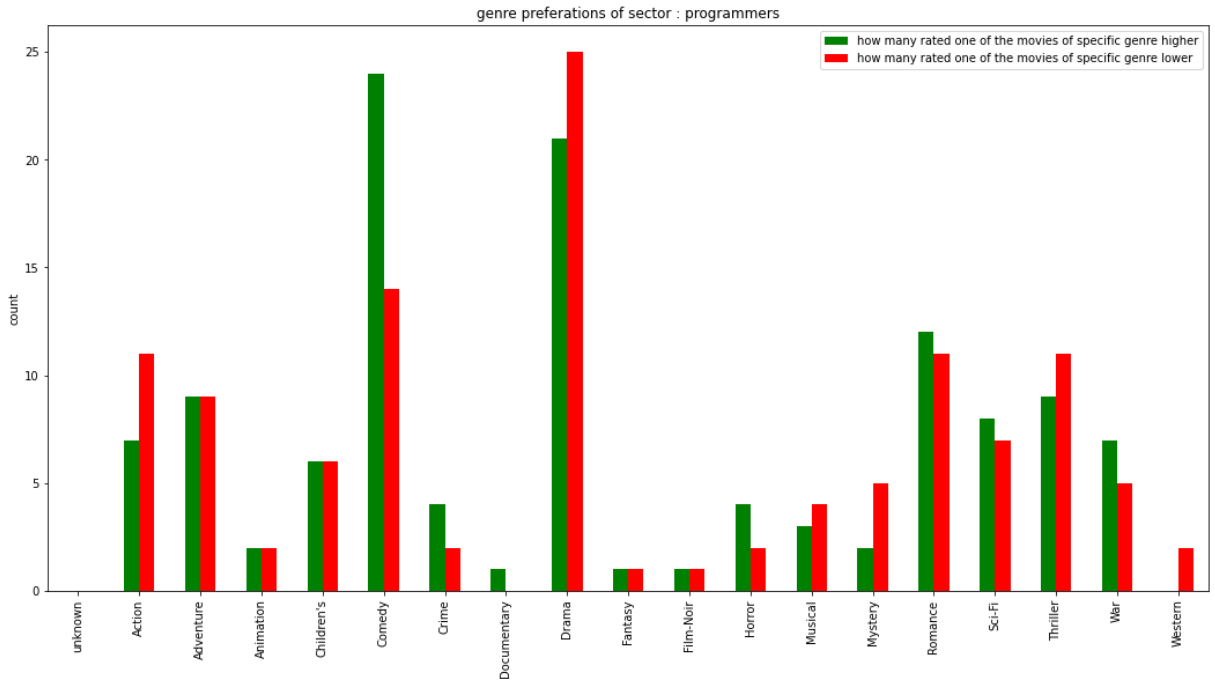
movie id	women avg rating	movie title_x	men avg rating	avg difference	movie title_y	release date
883	5.000000	Telling Lies in America (1997)	3.083333	1.916667	Telling Lies in America (1997)	01-Jan-1997
1451	5.000000	Foreign Correspondent (1940)	3.785714	1.214286	Foreign Correspondent (1940)	01-Jan-1940
318	4.632911	Schindler's List (1993)	4.406393	0.226519	Schindler's List (1993)	01-Jan-1993

[תמונה 7: טבלת הסרטים בעלי הפער הגדול ביותר בדירוג הממוצע באוכלוסיות הגברים והנשים]

חקרנו גם את ההעדפות של אוכלוסיות הסטודנטים והמתכנתים לז'אנרים השונים.



[תרשים 8: התפלגות של ז'אנרים של סרטים טובים וגרועים ביותר בקרב הסטודנטים]



[תרשים 9: התפלגות של ז'אנרים של סרטים טובים וגרועים ביותר בקרב המתכנתים]

להלן התובנות משלב ניתוח המידע:

1. אופי התפלגות הדירוג הממוצע של הסרטים לפי סקטורים - ניתן לראות כי בתרשימים 1-2 המצורפים עבור כלל האוכלוסייה ואוכלוסיית הגברים שההתפלגות היא נורמלית ואילו אצל אוכלוסיית הנשים (תרשים 3) התפלגות הדירוג הממוצע לסרטים אינה בדיוק נורמלית (הפעמון גבוה מידי ביחס לסביבה שלו). מכאן ניתן להבין כי קיים פחות דטא על אוכלוסיית הנשים. לו היה קיים יותר מידע על דירוגי הנשים, היינו מקבלים התפלגות נורמלית בקירוב.
2. סרטי דרמה מותירים חותם משמעותי מאוד על הגברים לטובה ולרעה. קיימים לא מעט דירוגים על סרטי דרמה בעלי ממוצע דירוג גבוה ונמוך במיוחד מצד הגברים.
3. לפי תרשימים 4-6 : סרטים דוקומנטרים לא הותירו חותם משמעותי על הצופים (לטוב ולרע). הטענה מתקיימת גם לגבי סרטי File-Nor.
4. לפי תרשימים 1-3 : הצופים נוטים שלא לצאת מגדרם. מעט מאוד סרטים קיבלו דירוגים גבוהים או נמוכים במיוחד.
5. הסטודנטים אמביוולנטים מאוד לגבי מספר ז'אנרים לא מבוטל לפי תרשים 8.
6. רק אוכלוסיית הנשים אוהבת ברובה לצפות בסרטי דרמה.

דיון בתוצאות עבור כל שלב ואלגוריתמים בהם השתמשנו:

<u>שלב ראשון:</u> בשלב זה התעסקנו בחקירת הנתונים שהוצבו בפנינו. רצינו לחקור ולראות כיצד אוכלוסיות שונות מדרגות סרטים שונים ולכן חילצנו היסטוגרמה שמציגה את התפלגויות הדירוג הממוצע של סרטים (משום שהוא הכי משקף את הקבוצה שראתה את הסרטים).
גילינו כי אופן ההתפלגות מצביע על אופן הדירוג של אוכלוסיה מסוימת (לדוגמא גילינו כי הנשים במאגר יותר הקצינו את עמדותיהן ביחס לגברים). ראינו את ההעדפות של האוכלוסיות השונות לז'אנרים שונים ע"י ספירת מספר הדירוגים שניתנו לסרטים מז'אנר מסוים שהיו בעלי דירוג ממוצע גבוה ונמוך יותר מתוך דירוגים שניתנו ע"י אוכלוסיה מסוימת.
כשראינו שישנו מספר גדול של דירוגים לסרטים שהשתייכו לז'אנר מסוים שדורגו באופן מסוים ע"י כמה אוכלוסיות יכולנו למצוא מכנה משותף בין האוכלוסיות. לדוגמא ניתן לראות בתרשימים 8 ו-9 של המתכנתים והסטודנטים בהתאמה, שאופן דירוג סרטים של ז'אנרים מסוימים דומה מאוד במרבית הז'אנרים.
השווינו בין הסרטים האהובים ביותר לפי דירוג ממוצע לפי אוכלוסיות הנשים והגברים, וגם מנתונים אלו הצלחנו לגזור העדפות לז'אנרים מסוימים של סקטור. כך למשל גילינו שגברים נוטים לאהוב יותר סרטי קומדיה ואנימציה, ואילו נשים נטו לאהוב יותר סרטי דרמה.
לסיכום בשלב זה התמקדנו בחקירת הדפוסים וההעדפות של אוכלוסיות מסוימות.

<u>שלב שני -</u> בשלב זה התמקדנו במתן המלצות על סמך מדדים שאינם מתייחסים לאדם שמבקש אותן. חקרנו 2 מודלים להמלצות לא אישיות - המלצה רנדומלית, והמלצה מבוססת ממוצע רייטינג לסרט. מצאנו כי המלצה שמבוססת על ממוצע רייטינג היא מדויקת יותר משום שה precision and recall, לפיהן מדדנו את הדיוק של המודלים היו גבוהים יותר במודל השני שתואר. מצב זה חזר על עצמו גם כאשר המלצנו לכלל האוכלוסייה על סמך ממוצע הדירוגים שנאסף מכלל האוכלוסייה, וגם כשחישבנו זאת על אוכלוסיית הגברים.

שאלה 3:

בשאלה 3 מימשנו מודל לחיזוי rating לסרט עבור user עפ״י המודלים matrix factorization, user to user, item similarity ו- item content.
התוצאות שהתקבלו עבור כל אחד מהמודלים הם:

matrix factorization model:
MAE: 0.8523414113387942
train time: 6.27s
user to user model:
MAE: 3.4894869063150638
train time: 0.082963s
item similarity model:
MAE: 3.258036293061598
train time: 0.055805s
item content model:
MAE: 3.3564270475826468
train time: 0.030361s

ניתן לראות כי עפ״י התוצאות המודל הכי יעיל הוא matrix factorization והמודל הכי פחות יעיל הוא item content.
משך האימון הקצר ביותר היה עבור item content.

שאלה 4:
מודל neural collaborative filtering עם שכבת hidden אחת
התוצאות שקיבלנו הם:

MAE: 3.5296
train time: 30s

בנוסף בנינו מודל neural collaborative filtering עם פרמטרים שונים עבור שלושת האפשרויות הבאות:
אפשרות ראשונה:

number of layers: 1
size of a layer: 20
optimizer: Adagrad
loss function: cosine similarity
activation function: relu

התוצאות שהתקבלו במודל זה הם:

MAE: 3.4743
train time: 40s

אפשרות שניה:

number of layers: 20
size of a layer: 20

optimizer: SGD
loss function: binary_crossentropy
activation function: softmax

התוצאות שהתקבלו במודל זה הם:

MAE: 2.5296
train time: 40s

אפשרות שלישית:

number of layers: 30
size of a layer: 30
optimizer: Adam
loss function: mse
activation function: selu

התוצאות שהתקבלו במודל זה הם:

MAE: 0.7303
train time: 206s

ניתן לראות כי את התוצאות הכי טובות קיבלנו עבור האפשרות השלישית.
מס׳ השכבות הגדול ביותר ולכן הכי מדויק ,אך נאלצנו לבצע חישוב ארוך יותר.

שאלה 5:

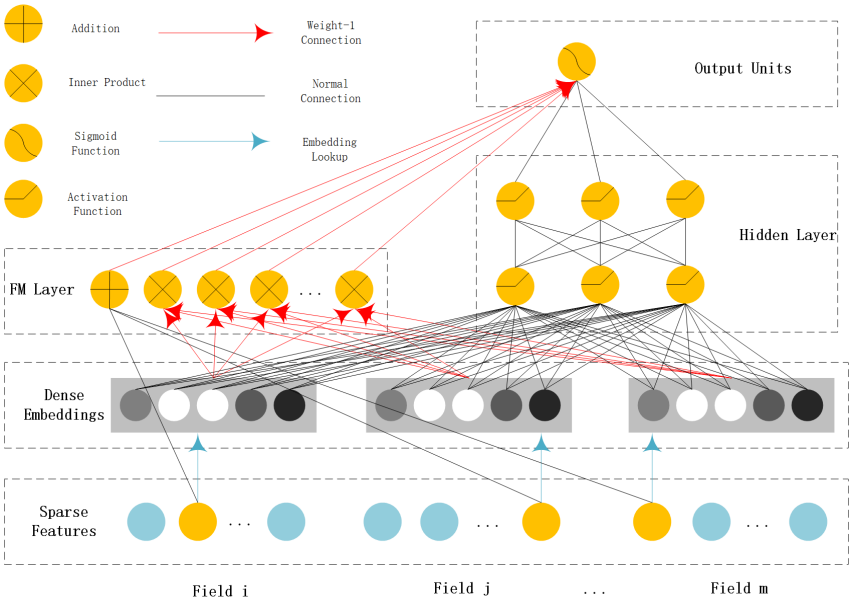
בשאלה זו , התבקשנו לנצל מידע נוסף שיש לנו ב- dataset של movielens - כלומר שימוש בנתונים שקימים כמו מין הצופה, גיל , מאפייני סרט וכדומה.
החלטנו בעבודה זו להתמקד בשני מאפיינים עיקריים שנתונים לנו על המשתמשים שהם הגיל והתעסוקה שלהם בחיים (Occupation)

לשם כך השתמשנו בשני מודלים עיקריים:

מודל **DeepFM:** מכונת פקטוריזציה המבוססת רשת נוירונים לחיזוי CTR.
מודל זה משלב בתוכו את הכוח של מכונות פקטוריזציה להמלצה ולמידה עמוקה. בשונה ממודלים אחרים כמו לדוגמא מודל ה- Wide and Deep של גוגל, מודל זה מקבל כקלט משותף את כל האלמנטים הנדרשים ללמידה.
מודל זה הוקם מהרעיון שההתנהגות מאחורי מה שגורם ליוזר ללחוץ/ לבחור במוצר מסוים מאוד חשובה לתהליך הלמידה.
למשל אנשים נוטים להוריד אפליקציות של הזמנת מזון בשעת הצהריים שבה נהוג לאכול.
או למשל שבני עשרה אוהבים לשחק דווקא משחקי יריות ומשחקי RPG.
כלומר, בהחלט יש חשיבות רבה לפרטים האלה אודות המשתמשים, ושכן הם יכולים להיות רבים מאוד, ומסובכים להצבה בפונקציות חיזוי.

מודל זה איפשר לנו בעזרת API נוח ופשוט להכניס את המידע של ה- train שכולל בתוכו את המאפיינים שהזכרנו (גיל ותעסוקה) ולהשתמש בהם כחלק מתהליך הלמידה והיכולת לחזות דירוגים.

באיור ניתן לראות כיצד המודל מקבל את כל ה- **sparse features** הנדרשים כשכבה אחת משותפת ושהוא משתמש במספר קבוע של שכבות.



המודל השני:**Item Popularity, turicreate**

אלגוריתם שמנבא דירוג של סרטים על סמך הפופולריות שלהם (ממוצע דירוג גבוה, לאחר שסיננו מראש סרטים שדורגו מעט פעמים) מהספרייה של **turicreate.**

למודל זה הוספנו "תפנית קטנה בעלילה", במקום לאמן את האלגוריתם על ה- train הנתון כפי שהוא היה, הוא עבר מניפולציה.

רצינו לבדוק השערה אפשרית שייתכן ששיפור בחיזוי הדירוגים כאשר היוזרים הנבדקים הם מטווח גילאים זהה ותעסוקה דומה.

כלומר, קשישים נהנים מסוג מסוים של סרטים , וסטודנטים נהנים מסוג אחר של סרטים וכדומה.

לכן הכנסנו בכל פעם לאלגוריתם המדובר מקטעים שונים של ה- data לאחר שעברו סינון לטווח גילאים מסוים ומקצוע.

ובדקנו אותם על ה- test עם סינון זהה.

התוצאות איששו את ההנחה, שכן בסופו של דבר באמת קיבלנו תוצאות לא רעות מהמודל.

תוצאות שהתקבלו עבור DeepFM:

MAE: 0.8567

תוצאה שהתקבלה עבור train ו- test מסוננים על סמך יוזרים בעלי מאפיינים משותפים באלגוריתם השני: (סטודנטים בגילאים 20 -40)

MAE: 0.8353056745299932

תוצאות בטווח של 0.5 - 0.9 המשיכו להתקבל עבור קבוצות בכל מיני גילאים ומקצועות.

לסיכום, 2 המודלים תיפקדו כראוי ואנחנו רואים בשניהם אופציות טובות לחיזוי תוצאות של דירוגים.
עם זאת הגענו למסקנה שעבור קבצי מידע מספיק גדולים שמכילים המון מאפיינים שונים כנראה שנעדיף את האלגוריתם הראשון, שככל שיש לו יותר מידע, הוא יודע לבצע חיזוי טוב יותר והוא יכול להשתפר עם הזמן.

המודל השני מכריח קיצוץ של מידע מאוד ספציפי - במקרה הזה עבור 2 מאפיינים בלבד, זה לא דבר שנוכל לשמור על המגמה שלו בכזאת פשטות כאשר נרצה להשתמש בכמות רבה יותר של מאפיינים שונים וראינו עד כמה המאפיינים האלה חשובים לחיזוי.