פרויקט אופנת ספורט – Recommender System

אולגה פישקינה

נעמי לוי

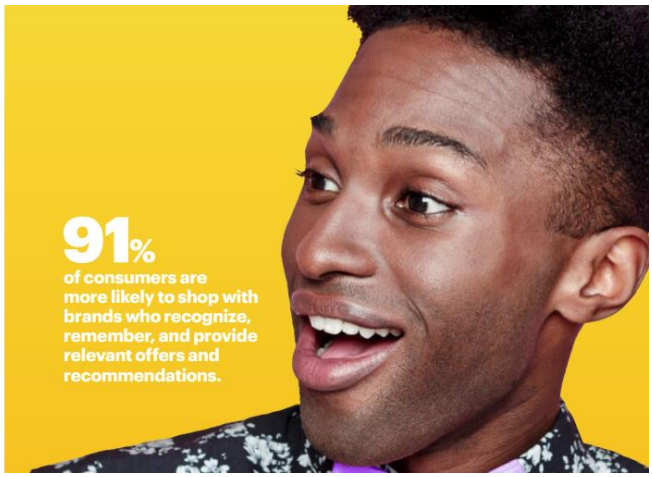
עידן סבאגי

תקציר

מערכות המלצה פוגשות אותנו בכל אספקט של החיים, בין אם בחוויית הצפייה בנטפליקס, האזנה למוזיקה בספוטיפיי וכלה בקניות שלנו באונליין, אפילו בהזמנות של קניות מרשת הסופרים. במאמר זה נמחיש עד כמה ניתן ליישם את אותם מערכות המלצה ברשת אופנה בה הקניות מתבצעות בסניפים בלבד. בתקופה האחרונה (עולם הפוסט קורונה בתקווה), אנו עדים לשינויים בהרגלי הצריכה בכל העולם ובפרט בעולם הקמעונאות אשר משנה את פניו במהירות עצומה. עליית שיעורן של הקניות המקוונות שהואצה בעקבות מגפת הקורונה, מציבה קשיים גם בפניי רשתות גדולות ומבוססות, ולא מותירה לאותן רשתות ברירה אלא להשתפר במגוון תחומים. בין היתר שיפור חווית הקנייה במובן הרחב, ובפרט לשם המאמר שלנו מכוון הוא היפר פרסונליזציה על מנת למשוך את הקונה לרשת באמצעות הצעות ערך שיווקיות ממוקדות ומותאמות אישית – קרי קופונים מותאמים אישית בהתבסס על מערכת המלצה – מטרת המחקר שביצענו.

1. מבוא

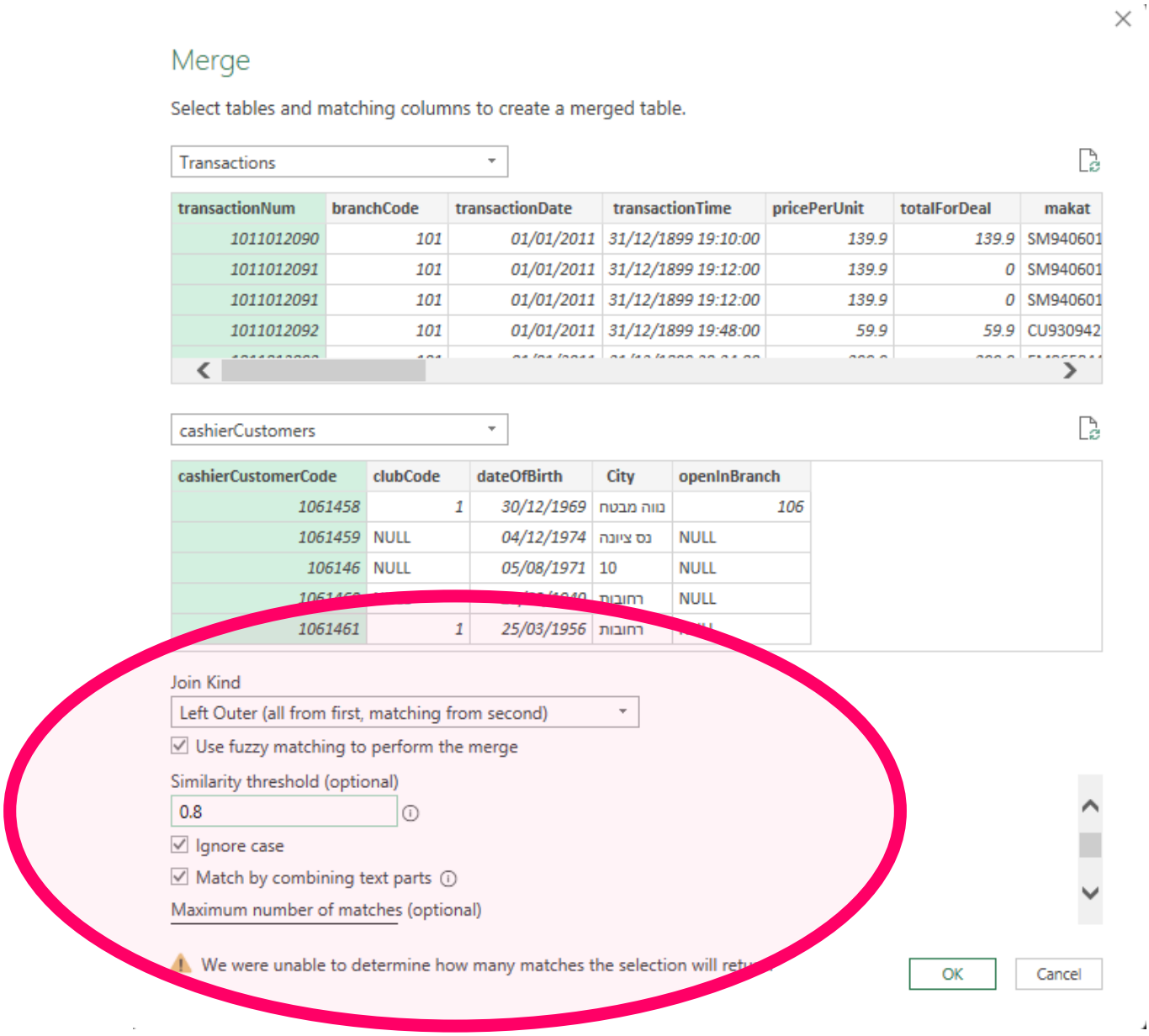
למרות נתונים המתפרסמים על זינוק בעלייה בקניות האונליין, ניתן לומר כי רוב הקניות עוברות לרשת. יחד עם זאת למרות הצלחת האונליין המסחר הפיזי ממשל לא הולך להיעלם אלא להשתנות ולהשתרג. ענקית מהמסחר באונליין Amazon הודיעה כבר על הרצון להקים סניפים פיזיים בתצורה של מודלים חדשים. ברוב המחקרים שנערכו בשנים האחרונות ישנה סלידה של לקוחות מרשתות שלא עושות את ההתאמה ללקוח, בין אם באימיילים ושיווק לא רלוונטיים. הלקוח מודל שנת 2022 מצפה ומוכן שילמדו את הרגלי הצריכה שלו על מנת לקבל הצעות והנחות מותאמות אישית, ולא די בשינוי הכותרת בהודעת ה-SMS או המייל[1].



[1] https://www.forbes.com/sites/blakemorgan/2020/02/18/50-stats-showing-the-power-of-personalization/?sh=48dc30142a94

2. הנתונים

בשלב זה על מנת להפיק את המירב מסט הנתונים שקיבלנו נאלצנו לעשות התאמות רבות, הכוללות בין היתר התאמה של מספר הטרנזקציה עם מספר הלקוח. התאמה זו לא הייתה כל כך פשוטה וכן נאצלנו לבצע התאמה של מספר הטרנזקציה עם מספר הלקוח בהתאמה של 80% שכן הבנו כי מספר הטרנזקציה כולל בתוכו את מספר הלקוח וכן מספר סניף כלשהו.



לאחר התאמה הכרחית זו, יכולנו לגשת ולנקות ולטייב את הנתונים.

אתגר נוסף שצלחנו היה לתרגם את הערים של הלקוחות שכן על מנת להשתמש בעיר המגורים, לטובת סגמנטציה של הלקוחות, נאלצנו לתרגם את מרבית היישובים המופיעים בדאטה, אשר נכתבו בשדה חופשי ולא מתוך שדה בחירה. לדוגמה לקוח שגר בתל אביב נרשם כמי שגר ב: ת"א, תא, תל אביב, תל-אביב, תל אביב וכו'.

3. שיטות

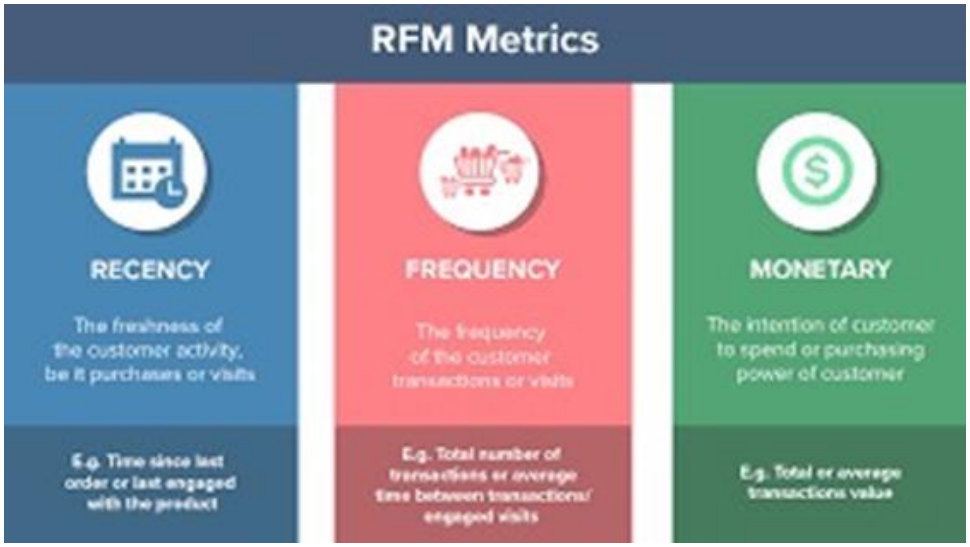
3.1 בניית מערכת המלצה למשתמש

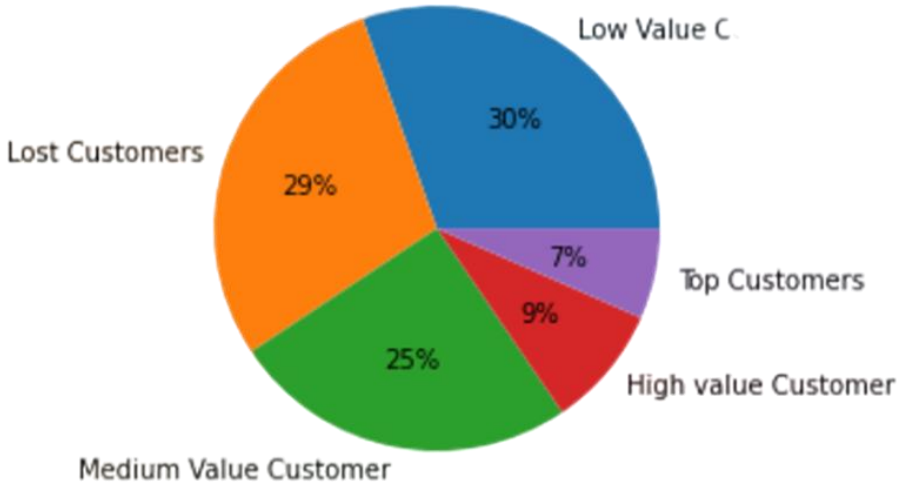
3.1.1 איסוף הדאטה: קיבלנו קובץ דאטה בדוגמת טבלאות אקסל המכילות נתונים משנים 2011-2014 על רכישות של לקוחות וכן פירוט המוצרים שנרכשו ומהיכן ומתי וכמו כן פירוט על הלקוחות עצמם שכן בשל העובדה שהיו נתונים לנו רק גיל הלקוח (חושב מתאריך הלידה שנמסר), מקום מגוריו, האם חבר מועדון או לא, התמקדנו בנתונים של הרכישות על מנת למצוא לקוחות דומים ועל ידי כך להמליץ לאותם לקוחות לרכוש מוצרים בהתאם.

3.1.2 עיבוד הדאטה: בשל העובדה שהיה ברשותנו מידע רב מכדי להגיע לתוצאות בזמן סביר, חילקנו את השורות לכ-1000 שורות בכל פעם.

3.2 סגמנטציה של הלקוחות:

3.2.1 RFM SCORE: סיווג לקוחות בעולם הקניות מתחיל בראש ובראשונה ב-RFM SCORE, המסווג את הלקוחות לפי יכולת הקנייה, שכיחות הקניות והסכומים. בתהליך זה אפיינו את הלקוחות לקשת של ערכים, מלקוחות אבודים ועד לקוחות עם ערך גבוהה.

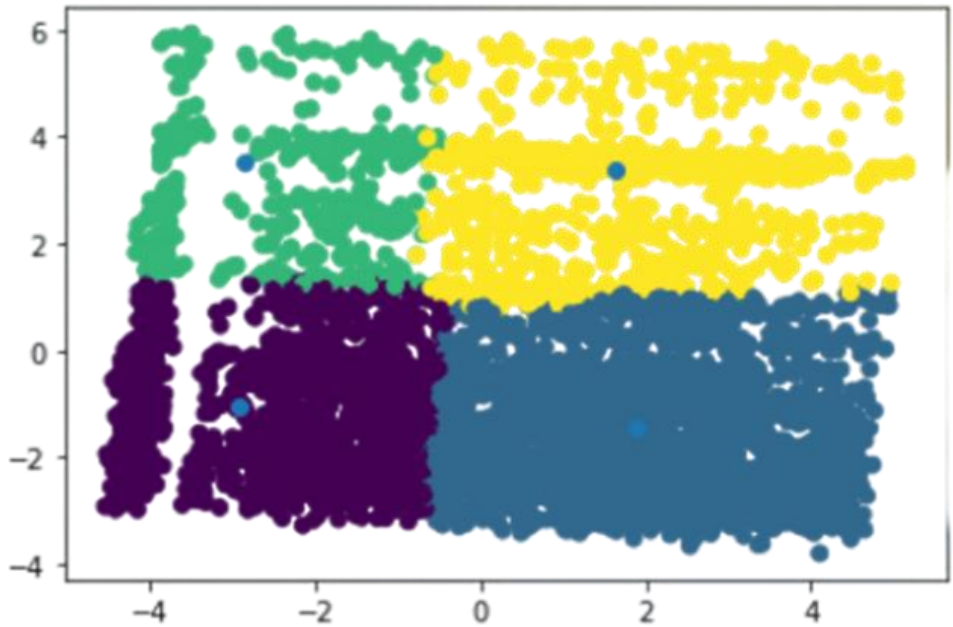




3.2.2 CLUSTERING: על מנת למצוא לקוחות דומים ככל הניתן על ידי מאפיינים דמוגרפיים נעזרנו במקור ידע חיצוני (הלשכה המרכזית לסטטיסטיקה), על מנת להשלים את המידע החסר לאותו מאפיין דמוגרפי.
שאבנו נתונים שסיפקו לנו את המעמד החברתי כלכלי של כל ישוב מגורים בישראל בחפיפה לאותן שנים של נתונים - המידע של הטרנזקציות נע בין 2011 – 2014 והמידע של הלשכה מתייחס לשנת 2013. באמצעות הנתון של דירוג כל יישוב ויישוב הצלחנו להרכיב תמונה של חלוקת הלקוחות שלנו.

NAME OF LOCAL AUTHORITY	הפרש (אשכול 2017 פחות אשכול 2015) DIFFERENCE (CLUSTER 2017 MINUS CLUSTER 2015)	אשכול [4]2015 CLUSTER 2015[4]	דירוג [3]2015 RANK 2015[3]	אשכול [4]2017 CLUSTER 2017[4]	דירוג [3]2017 RANK 2017[3]	ערך מדד [2]2017 INDEX VALUE 2017[2]	אוכלוסיית המדד [1]2017 INDEX POPULATION 2017[1]	מחוז DISTRICT	שם רשות מקומית	סמל יישוב CODE OF LOCALITY	מעמד מוניציפלי MUNICIPAL STATUS
NEVE MIDBAR	0	1	1	1	1	-2.815	9,108	6	נווה מדבר		68
AR'ARA-BANEGEV	0	1	5	1	2	-2.535	16,988	6	ערערה-בנגב	1192	99
TEL SHEVA	0	1	2	1	3	-2.329	19,748	6	תל שבע	1054	99
KUSEIFE	0	1	8	1	4	-2.255	20,195	6	כסיפה	1059	99
MODI'IN ILLIT	0	1	7	1	5	-2.234	70,081	7	מודיעין עילית	3797	0
SEGEV-SHALOM	0	1	3	1	6	-2.215	9,897	6	שגב-שלום	1286	99
HURA	0	1	6	1	7	-2.203	20,782	6	חורה	1303	99
AL-KASUM	0	1	4	1	8	-2.138	10,014	6	אל קסום		69
LAQ'YE	0	1	9	1	9	-2.012	12,857	6	לקיה	1060	99
RAHAT	0	1	11	1	10	-1.916	66,744	6	רהט	1161	0
BETAR ILLIT	0	1	10	1	11	-1.908	54,557	7	ביתר עילית	3780	0
TEL MOND	1	8	237	9	236	1.367	12,605	4	תל מונד	154	99
DEROM HASHARON	1	8	238	9	237	1.384	32,700	4	דרום השרון		20
MODI'IN-MAKKABBIM-RE'UT	1	8	231	9	238	1.391	91,328	4	מודיעין-מכבים-רעות	1200	0
GANNE TIQWA	1	8	233	9	239	1.411	18,228	4	גני תקווה	229	99
PARDESIYYA	1	8	240	9	240	1.419	5,828	4	פרדסיה	171	99
GIV'ATAYIM	1	8	241	9	241	1.453	59,505	5	גבעתיים	6300	0
HOD HASHARON	1	8	242	9	242	1.491	60,774	4	הוד השרון	9700	0
EVEN YEHUDA	1	8	244	9	243	1.499	13,574	4	אבן יהודה	182	99
QIRYAT ONO	1	8	243	9	244	1.540	39,374	5	קריית אונו	2620	0
KEFAR WERADIM	0	9	247	9	245	1.545	5,531	2	כפר ורדים	1263	99
METAR	1	8	245	9	246	1.576	7,749	6	מיתר	1268	99
RAMAT HASHARON	0	9	249	9	247	1.606	45,909	5	רמת השרון	2650	0
SHOHAM	0	9	248	9	248	1.671	20,928	4	שוהם	1304	99
GEDEROT	0	9	246	9	249	1.705	5,180	4	גדרות		32
HAR ADAR	0	9	250	9	250	1.758	4,058	7	הר אדר	3769	99
KOKHAV YA'IR	0	9	251	9	251	1.809	8,889	4	כוכב יאיר	1224	99
OMER	1	9	252	10	252	1.900	7,613	6	עומר	666	99
LEHAVIM	1	9	253	10	253	1.989	6,392	6	להבים	1271	99
KEFAR SHEMARYAHU	0	10	254	10	254	2.068	1,811	5	כפר שמריהו	267	99
SAVYON	0	10	255	10	255	2.320	3,878	4	סביון	587	99

	RFM_Score	CityRank(1 to 255)	CustomerAge
CustomerID			
10111000	5.034483	7.625984	5.337349
10111001	5.034483	3.125984	2.409639
10111002	5.034483	3.125984	4.253012
10111003	5.034483	9.503937	5.337349
10111004	5.034483	3.125984	1.325301



3.2.3 **<u>בניית המודל</u>**: השתמשנו בסינון שיתופי (Collaborative Filtering). ההנחה הבסיסית של השיטה היא שלקוחות דומים ירכשו מוצרים דומים. בכל פעם בחרנו לקוחות מתוך הדאטה, לפי מספר הלקוח שלהם ולפי תוצאות הרצת האלגוריתם.

<u>שלב ראשון</u>: **מציאת לקוחות דומים על פי רכישות דומות (התמקדנו במספר מק"ט של כל מוצר)**: על מנת למצוא את הלקוחות הדומים ביותר השתמשנו **The Simple Matching Coefficient** על מנת לחשב את הדמיון באופן סטטיסטי. ראשית חישבנו באמצעות וקטור רכישות את המשקל שמקבל כל אטריבוט(במקרה שלנו, שני לקוחות לדוגמא A וB אטריבוט אחד יהיה כאשר שני הלקוחות רכשו את אותו מוצר, בעוד שאטריבוט שני יהיה כאשר שני הלקוחות לא רכשו את אותו מוצר ואטריבוט שלישי יהיה שאחד מהם רכש ושני לא וכן הלאה) שהצגנו במטריצה בכדי להבין את התוצאות טוב יותר.

*מטריצת רכישות לפי מק"ט

10512069	10512070	10512071	10512072	10512073	10512074	10512075	10512076	10512077	10512078
0	3	3	4	4	3	2	2	2	1
0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

CustomerID	10111090	10111091
makat		
444	0	0
999	0	0
310009	0	0
4002996	0	0
9900100	0	0
...	...	...
WM7096010L	0	0
WM7096010M	0	0
WM7096010XL	0	0
WM7096610XXL	0	0
WM8033010M	0	0

חישוב המקדמים על פי הנוסחה של SMC וקבלת רשימת לקוחות דומים: חיבור בין אטריבוטים הופכיים וחילוק שלהם בכלל האטריבוטים הניב לנו את רמת הדמיון בין הלקוחות, כאשר לדוגמא לקוחות שקיבלו תוצאת חלוקה של 7/10 יהיו בעלי דמיון של 0.7 כלומר יחסית דומים בעוד שלקוחות שקיבלו תוצאת חלוקה שלילית לא דומים כלל.

*דמיון בין לקוחות

```
recomendations
#similarity coefficient between customers
```

	0	1	2
0	10111090	10111091	0.706507
1	10111091	10111090	0.706507
2	10111092	10111207	0.706507
3	10111093	10111139	-0.001695
4	10111094	10111136	-0.002941
...	...	...	...
255	10512074	10511776	0.655635
256	10512075	10511776	0.499702
257	10512076	10512002	0.568766
258	10512077	10512002	0.410112
259	10512078	10111199	0.299040

בהתאם לתוצאות שקיבלנו ערכנו בדיקה של שני לקוחות מתוך הדאטה עם דמיון גדול וראינו על סמך ההרצה של האלגוריתם מה ניתן להמליץ לאיזה לקוח.

True מסמל שאכן ניתן להמליץ על אותו מוצר ללקוח הדומה ללקוח שרכש את המוצר ו False

מסמל שלא ניתן .

<u>שלב שני:</u> לאחר שקיבלנו אינדיקציה יותר מעמיקה לגבי הדמיון בין הלקוחות והקשר בין לקוחות והרכישות של כל לקוח, החלטנו לקחת 3 מודלים כאשר אחד היווה הבסיס מודל הפופולריות של המוצרים שנרכשו, השני מודל שהשתמש בכלי קוסינוס למציאת דמיון בין מוצרים והשלישי מודל שהשתמש בפירסון כאמצעי למציאת הדמיון בין המוצרים השונים. בכל מודל השווינו בין רכישות של הדאטה המקורי לבין רכישות של הדאטה המנורמלת לבין רכישות של dummy data. לאחר בחינת הprecision והrecall ולבסוף קבלת תוצאות הRMSE החלטנו להתמקד במודל עם

pearson למציאת דמיון עם הדאטה המנורמלת שכן הלה הניב לנו תוצאת RMSE משוקללת של **0.190 שעבור הדאטה שלנו זאת תוצאה טובה.**

4.<u>מסקנות ודיון:</u>

בעיה שנתקלנו בה במהלך ניתוח הנתונים ובניית מערכת ההמלצה התמקדו בעובדה שכמות הדאטה היא גדולה ולכן התוצאות שקיבלנו כהמלצה לכל לקוח היו מורכבות שכן כללו מספר רב של פריטים משום העובדה שעבור פריטים רבים יש מידות רבות מה שמביא לכפילויות של פריטים ותוצאות פחות מדויקות ממה שציפינו לקבל.

פתרון אפשרי לכך יכול להיות סינון של פריטים ממידות שונות עבור כל לקוח והתאמה מדויקת יותר מה שיביא להמלצה יותר קלה לקריאה.

<u>רעיונות נוספים לפיתוח הפרויקט:</u>

אם היה לנו יותר זמן להעמיק בעניין ולדייק את תוצאות מערכת ההמלצה היינו מוסיפים פילטר של דירוג הערים שמצאנו כאשר הוספנו את נתוני הלמ"ס ומשלבים בין רכישות דומות של לקוחות מה שהיה מתאים יותר טוב את המוצרים עבור כל לקוח.

אם היינו מקבלים מרשת קולומביה דאטה יותר ממוקד, למשל יותר פרטים על הלקוחות כמו פרופילי פייסבוק/אינסטגרם,כתובת מייל, מין הלקוח, מספר טלפון יכולנו להתמקד בפן השיווקי בצורה יותר טובה שכן שיווק ממוקד עבור כל לקוח ייעשה בדרך כלל דרך אחד או יותר מהערוצים שלעיל.(הודעות טקסט, מייל)

במהלך העבודה על הפרויקט היה לנו מאוד מעניין אך מאתגר. תחום חיזוי הקניות עבור הלקוחות בשילוב עם personalized purchasing הוא תחום מבוקש שנחקר על ידי צוותים גדולים ודרושה עבודה מעמיקה מאוד . כמו כן משום שיש מאפיינים רבים המשפיעים על הרכישות של הלקוחות ועלולות לשנות את הנתונים לעיתים תכופות יותר, קשה יותר לחזות בצורה מדויקת אך אנו ניסינו לעשות זאת בצורה המיטבית עם הדאטה שניתנה לנו.