תוכן עניינים:

Deep Learning – NFC Recommendation System

מבוא:

פרויקט זה נבנה כחלק מקורס "למידה עמוקה" ומטרתו היא ליישם את הכלים שלמדנו בקורס. הפרויקט הנ"ל משתמש בשיטות שונות הקשורות לניתוח נתונים וללמידה עמוקה בפרט, בפרויקט זה נבנה ארכיטקטורה מתאימה למודל של "למידה עמוקה" כאשר מטרתנו הינה ליצור מערכת המלצות עבור המשתמשים, נרצה לבדוק עד כמה הפריט (מטבלת הפריטים) "מתאים" למשתמש, כלומר נבדוק את האינטרקציה בין המשתמש לפריט ועל בסיס המידע הזה, נוכל למצוא את הפריטים המומלצים ביותר לאותו משתמש. מערכת המלצות היא פתרון טכנולוגי שמטרתו לספק המלצות אישיות למשתמשים, על פי הפרופיל האישי שלהם, מיקום, היסטוריית רכישה וכדומה. באמצעות המודל שנבנה, ניתן יהיה לשפר את חוויית הקניה ולהגדיל את התמריצים לרכישה באמצעות הצגת המוצרים הרלוונטיים ביותר לכל לקוח.

למעשה, מערכת המלצות היא טכניקה בתחום למידת המכונה שמטרתה לספק המלצות אוטומטיות ומותאמות אישית למשתמשים. היא פועלת על פי נתונים סטטיסטיים ומודלים מתמטיים לזיהוי רגישויות והעדפות של המשתמשים. מערכות המלצה נמצאות בשימוש נרחב בפלטפורמות דיגיטליות שונות, כמו אתרי קניות מקוונים, פלטפורמות תוכן ועוד.

מערכות המלצה המבוססות על רשתות נוירונים (Neural Recommendation Systems) הן פרדיגמה חדשה בעולם המערכות המלצה. כשנשתמש ברשתות נוירונים למשימת המלצה, אנו מאמצים את היכולות המתקדמות של למידת המכונה לזיהוי תבניות בנתונים מורכבים. בזכות השימוש ברשתות נוירונים, נוכל לשפר את איכות ההמלצות וכמו כן נוכל לספק למשתמשים המלצות שמבוססות גם על פרטי הפרופיל האישי שלהם.

מערכת ההמלצה שאבנה עבור פרויקט זה מבוססת על נתוני רכישות מוצרים של חברת הביגוד H&M מאתר Kaggle הנתונים מכילים שלושה סוגים של טבלאות:

- טבלת "מוצרים" המכילה מידע על המוצרים השונים שנמכרים על ידי החברה.
- טבלת "לקוחות" המכילה פרטים אישיים של לקוחות החברה.
- טבלת "טרנזקציות" המכילה מידע על רכישות שונות שבוצעו על ידי הלקוחות.

בנוסף, יש תיקיית תמונות המכילה את התמונות השונות של כל מוצר בבסיס הנתונים.

חילקנו את העבודה בפרויקט ל-5 חלקים:

חקירת בסיס הנתונים – על מנת להבין היטב מה אני עושה וכיצד לבנות את המודל שלי בצורה הטובה ביותר, ארצה להבין את הנתונים, החל מסוגם, סטטיסטיקות בסיסיות ובעיקר למקסם את היעילות של המודל.

הכנת סט הנתונים למודל של "למידה עמוקה" – ארצה ליצור סט נתונים אחד המתאים לארכיטקטורה של מודל של למידה עמוקה, ארצה לבנות פיצ'רים חדשים המתארים את האינטראקציה בין המשתמשים לפריטים וכמו כן להתאים טכנית את הנתונים (ניקוי נתונים, חריגים, קידוד וכו')

בניית ארכיטקטורה – אבנה ארכיטקטורה מתאימה, כזו שתביא למקסום הדיוק של המודל, לאחר מכן אעשה Fine Tuning על מנת לבחור את ההיפרפרמטרים הטובים ביותר, לאחר מכן אבדוק את תוצאות המודל על פי המדדים המתאימים.

בניית מערכת ההמלצה – ארצה ליצור פונקציה שמחשבת עבור כל משתמש את אחוז האינטרקציה שלו עם המוצרים השונים והמערכת בעצם תוציא פלט של חמשת המוצרים המומלצים ביותר עבור המשתמש.

לבסוף אסכם את הפרויקט ואעלה מסקנות ותובנות מהליך בניית הפרויקט.

Deep Learning – NFC Recommendation System

פרק 1 – חקירת סט הנתונים:

כאמור לפני שנעמיק בבניית הארכיטקטורה המתקדמת של מודלי הלמידה העמוקה, חשוב לבצע חקירה מדויקת על הטבלאות הזמינות. חקירת הטבלאות מהווה שלב חשוב ביישום כלי המלצה, מאחר והיא מאפשרת לנו להבין את מבנה הנתונים, הרלוונטיות של השדות השונים והקשרים ביניהם. החקירה הזו מאפשרת לנו על בסיס היסטוריה הרכישות הקודמת של הלקוחות לזהות דפוסים שונים והרגלי קנייה של הלקוחות מה שיוביל ליעילות בהמלצת מוצרים ולהגדיר את המשימה המובנית למערכת המלצה הנוספת.

טבלת המוצרים:

- טבלת מוצרים:
  - טבלת המוצרים מכילה מידע אודות 105,544 מוצרים מרשת ההלבשה H&M
  - בטבלה קיימות 25 עמודות המכילות פרטי מידע שונים אודות המוצרים הקיימים ברשת הבגדים H&M.
- להלן פירוט העמודות:
  - מזהה מוצר (מפתח, ערך מספרי ייחודי)
  - קוד מוצר (int קטגוריאלי)
  - שם מוצר (שם המוצר – string)
  - סוג מוצר – (int קטגוריאלי)
  - שם סוג מוצר – (string קטגוריאלי)
  - שיוך לקבוצת מוצרים - (string קטגוריאלי, לדוגמא תחתונים, חולצות)
  - מראה גרפי מספרי – (int קטגוריאלי)
  - מראה גרפי תיאור (string קטגוריאלי, לדוגמא פסים, מודפס, שקוף)
  - קוד קבוצת גוונים (int קטגוריאלי)
  - תיאור קבוצת גוונים (string קטגוריאלי, לדוגמא כחול כהה , אדום בהיר, אפור כהה)
  - מזהה ערך צבע נתפס (int קטגוריאלי)
  - צבע נתפס (string קטגוריאלי, לדוגמא בהיר, כהה, שקוף)
  - קוד צבע עיקרי (int קטגוריאלי)
  - שם צבע עיקרי (string קטגוריאלי לדוגמא אדום, כחול, ירוק)
  - קוד מחלקה (int קטגוריאלי)
  - תיאור מחלקה (string קטגוריאלי, גברים, תינוקות, נשים)
  - אינדקס קוד (string קטגוריאלי)
  - אינדקס תיאור (string קטגוריאלי, לדוגמא, ספורט, מידות תינוק, תכשיטי נשים)
  - קוד מדור (int קטגוריאלי)
  - שם מדור (string קטגוריאלי, לדוגמא תכשיטי נשים, מוצרי תינוקות, גרביי גברים)
  - מספר קבוצת בגדים (int קטגוריאלי)
  - שם קבוצת בגדים (string קטגוריאלי)
  - תיאור המוצר (string תיאור המוצר בטקסט)

** ניתן להבין רק מהעמודות שקיימת כפילות בייצוג הנתונים (חלק מיוצגים כמספרים וחלק כטקסט למען הנוחות, נשאיר רק את העמודות הקטגוריאליות שמיוצגות בטקסט)

** כמו כן, המחיר של המוצרים, נתון בטבלת הטרנזקציות (עליה נרחיב בהמשך) לכן משם ניקח את עמודת המחיר עבור הפריטים

Deep Learning – NFC Recommendation System

חקירת הנתונים בטבלת המוצרים:

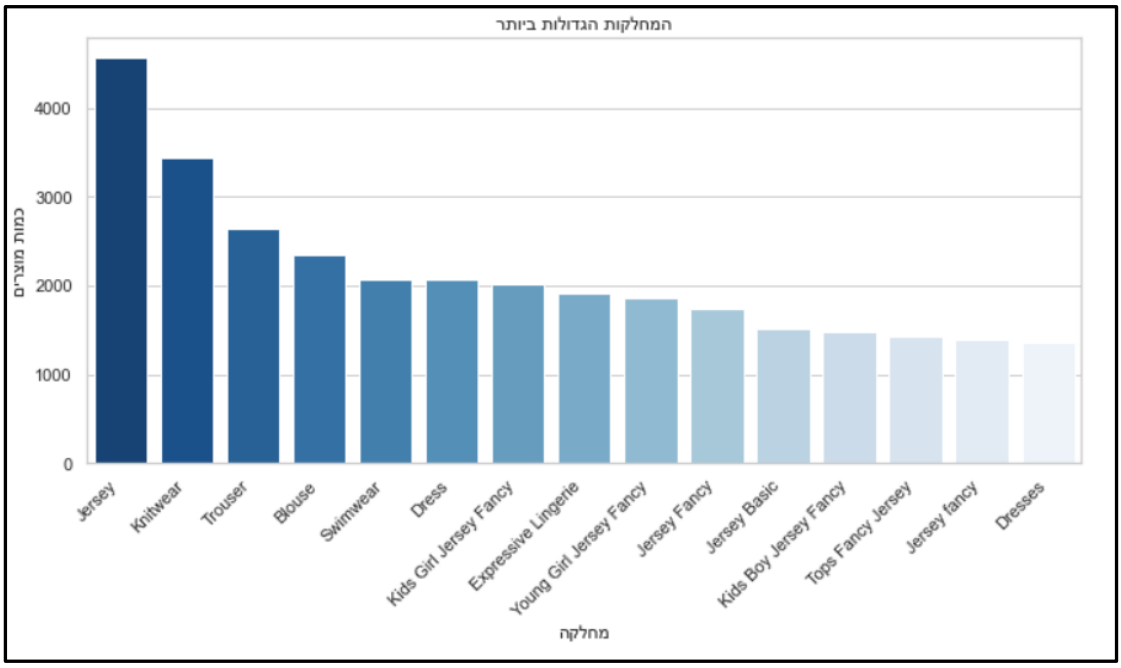
לאחר הכנת הטבלה נשארתי עם טבלה המכילה 104,547 רשומות (מוצרים שונים) ו-14 עמודות.

שם העמודה (מקורי)	שם עמודה (תרגום)	מספר ערכים חסרים	סוג נתונים
Article_id	מזהה מוצר	0	Int
Prod_name	שם המוצר	0	Object
Product_type_name	שם סוג המוצר	0	Object
Product_group name	שיוך לקבוצת מוצרים	0	Object
Graphical_appearance_name	תיאור מראה גרפי	0	Object
Colour_group_name	תיאור קבוצת גוונים	0	Object
Perceived_colour_value_name	שם צבע נתפס	0	Object
Perceived_colour_master_name	שם צבע עיקרי	0	Object
Department_name	שם מחלקה	0	Object
Index_name	תיאור אינדקס	0	Object
Index_group_name	תיאור קבוצת אינדקסים	0	Object
Section_name	שם מדור	0	Object
Garment_group_name	שם קבוצת בגדים	0	Object
price	מחיר	0	Float

לאחר מכן, שיחקתי מעט עם הטבלה בניסיון ללמוד אותה וסטטיסטיקות שונות ממנה.

Deep Learning – NFC Recommendation System

להלן כמה דוגמאות:



איור 1 - המחלקות הגדולות ביותר

ספירת המוצרים לכל מחלקה (כאן אציג את 5 המחלקות עם מספר המוצרים הגדול ביותר)

מחלקה	כמות מוצרים
Jersey	4560
Knitwear	3441
Truser	2538
Blouse	2342
Swimwaer	2075

Deep Learning – NFC Recommendation System

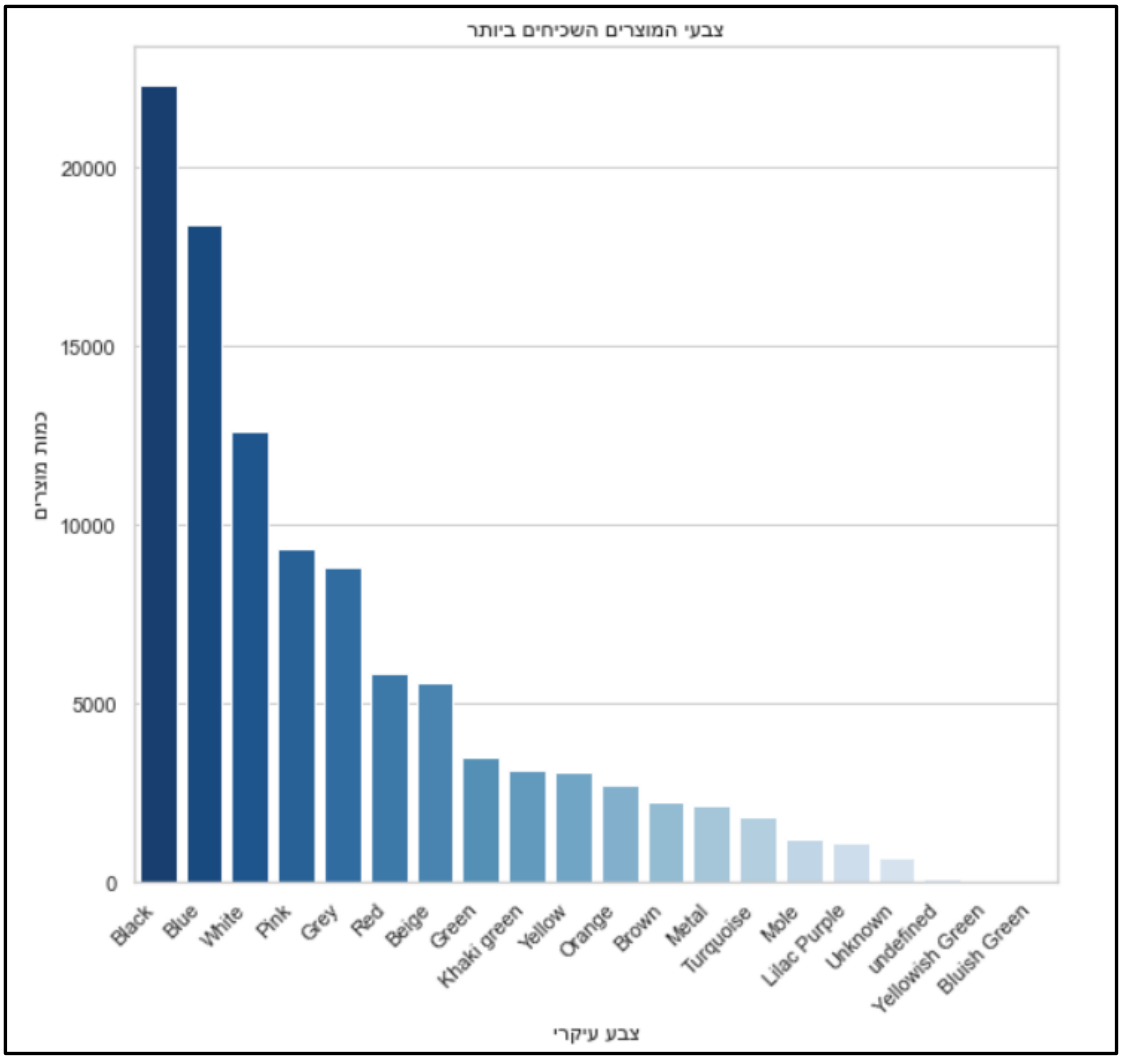
עניין אותי לבדוק כיצד הפריטים מחולקים בין המדורים (נציג את שלושת הפריטים הראשונים)

מדור	כמות חולצות (jersey)
Mama	1103
Womens Casual	1023
H&M+	889
Womens Trend	869
Womens Tailoring	676

מדור	כמות סריגים (Knitwear)
Womens Everyday Collection	925
Man suits & Tailoring	420
Contemporary Smart	367
Womens Casual	355
Womens Trend	341
Contemporary Casual	329
H&M+	179
Mama	140
Men Edition	140
Contemporary street	63

מדור	כמות מכנסיים (Truser)
Everyday Collection	910
Womens Tailoring	467
Womens Casual	320
Contemporary Smart	205
Contemporary Casual	235
Man suits & Tailoring	225

לאחר מכן, הסתכלתי קצת על הצבעים:

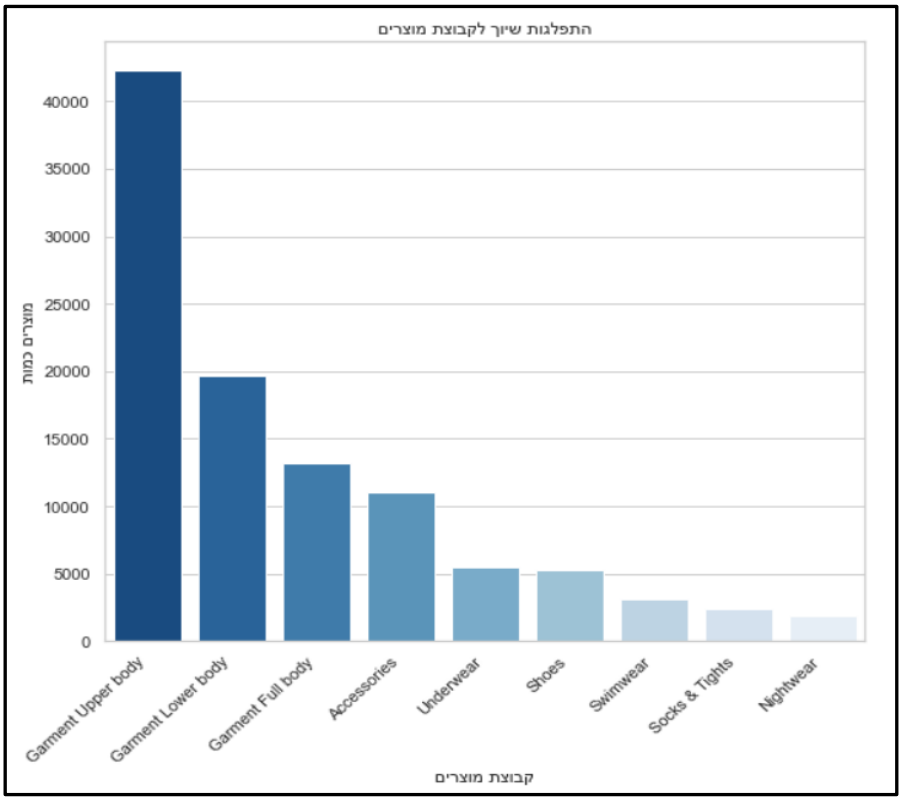


איור 2 – הצבעי המוצרים השכיחים ביותר

ניתן לראות כאן שצבעי המוצרים השכיחים ביותר הם: שחור, כחול, לבן, ורוד ואפור כאשר שחור וכחול הם הצבעים השכיחים ביותר.

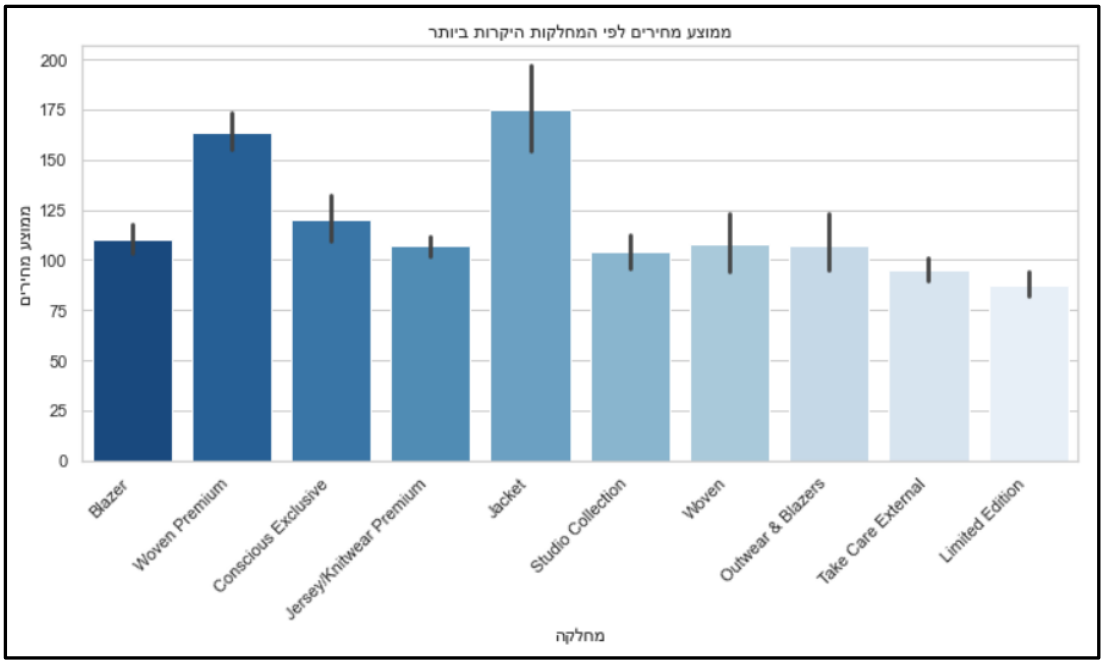
Deep Learning – NFC Recommendation System

מכאן רציתי להבין מהי החלוקה לקבוצות המוצרים כלומר איזו מחלקה היא הגדולה ביותר



איור 3 – התפלגות מוצרים לפי מחלקות

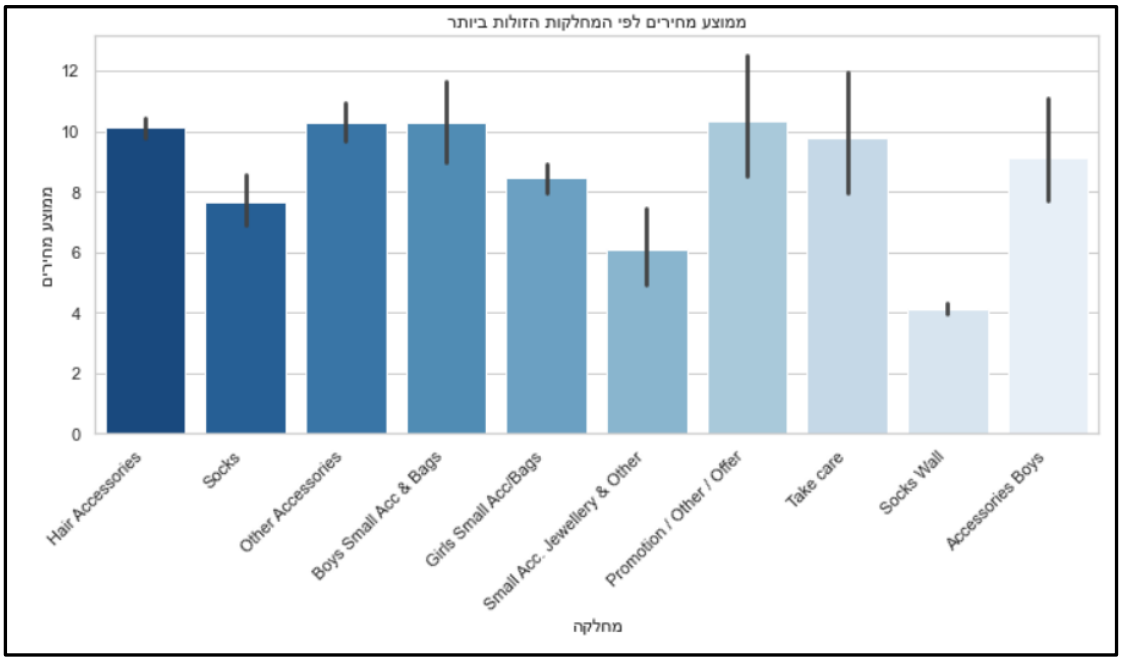
לאחר מכן רציתי לעשות השוואת מחירים בין המחלקות השונות:



איור 4 – ממוצע מחירים לפי מחלקה (היקרות ביותר)

ניתן לראות שז׳קטים, בגדי נשים (פרימיום) ובלייזרים הם המוצרים היקרים ביותר.

בדקתי גם את המחלקות הזולות ביותר:



איור 4 – ממוצע מחירים לפי מחלקה (הזולות ביותר)

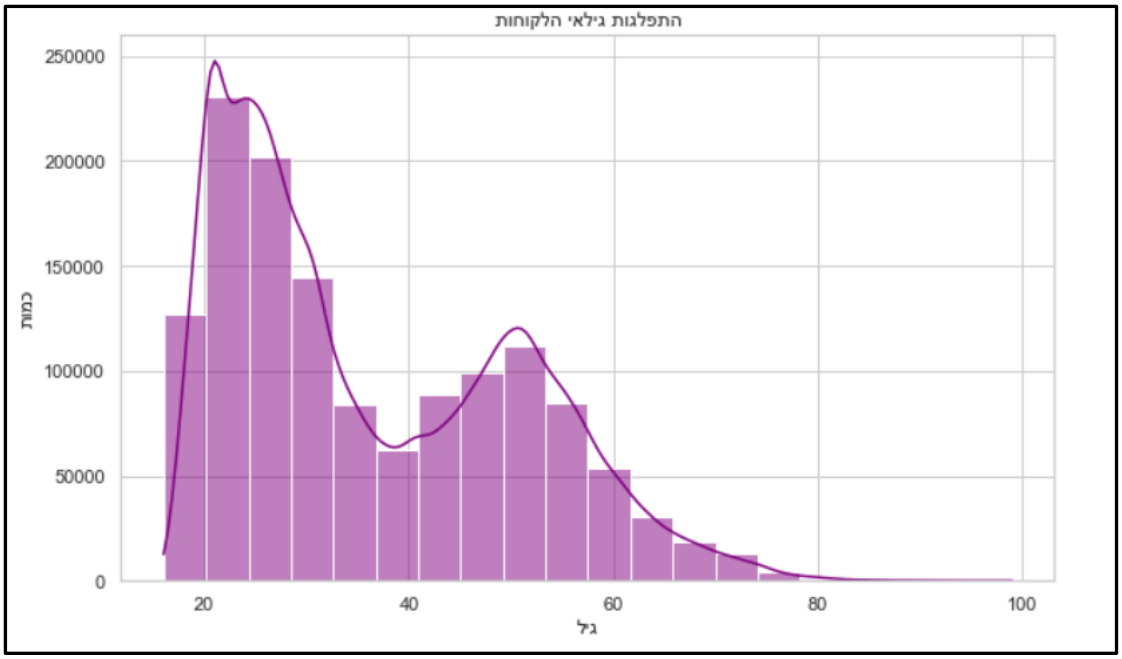
ראיתי שגרביים ותכשיטים קטנים (accessories) הם המוצרים הזולים ביותר.

<u>טבלת לקוחות:</u>

לאחר שהבנתי פחות או יותר מה הולך עם טבלת המוצרים, עברתי לחקירת טבלת הלקוחות

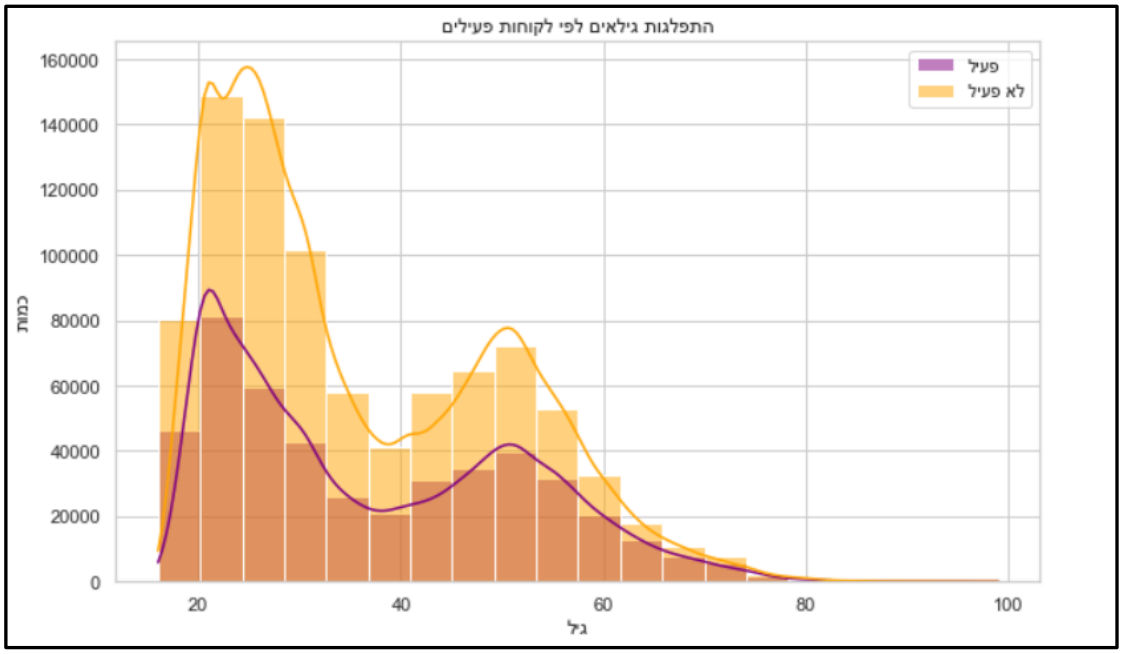
- <u>טבלת לקוחות:</u>
  - טבלת הלקוחות מכילה מידע אודות 1,371,979 לקוחות ברשת M&H
  - בטבלה יש סה"כ 7 עמודות
- <u>להלן פירוט העמודות:</u>
  - מזהה לקוח (מפתח, ערך ייחודי, string)
  - גיל (int)
  - מנוי להודעות (string, קטגוריאלי)
  - חבר מועדון (string, קטגוריאלי)
  - מיקוד (string)
  - משתמש פעיל (בוליאני)
  - FN (בוליאני)

בטבלת הלקוחות יש לנו 1,371,980 רשומות, מתוכם 1,272,491 יש חבר מועדון ו- 464,404 הינם לקוחות פעילים. ממוצע גילאי הלקוחות הוא 36.4 (הגיל החציוני הוא 32) כאשר הלקוח המבוגר ביותר הוא בן 99 והצעיר הוא בן 16. מטה מופיע גרף המציג את התפלגות גילאי הלקוחות.



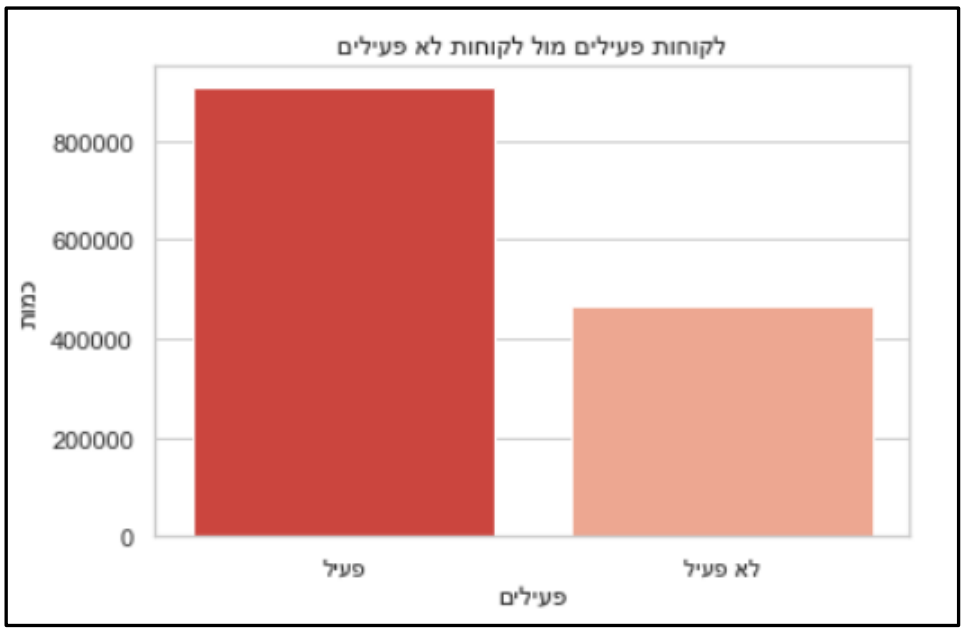
איור 5 – התפלגות גילאי הלקוחות

לאחר מכן עניין אותי לבדוק מהי התפלגות הגילאים עבור לקוחות פעילים ולא פעילים

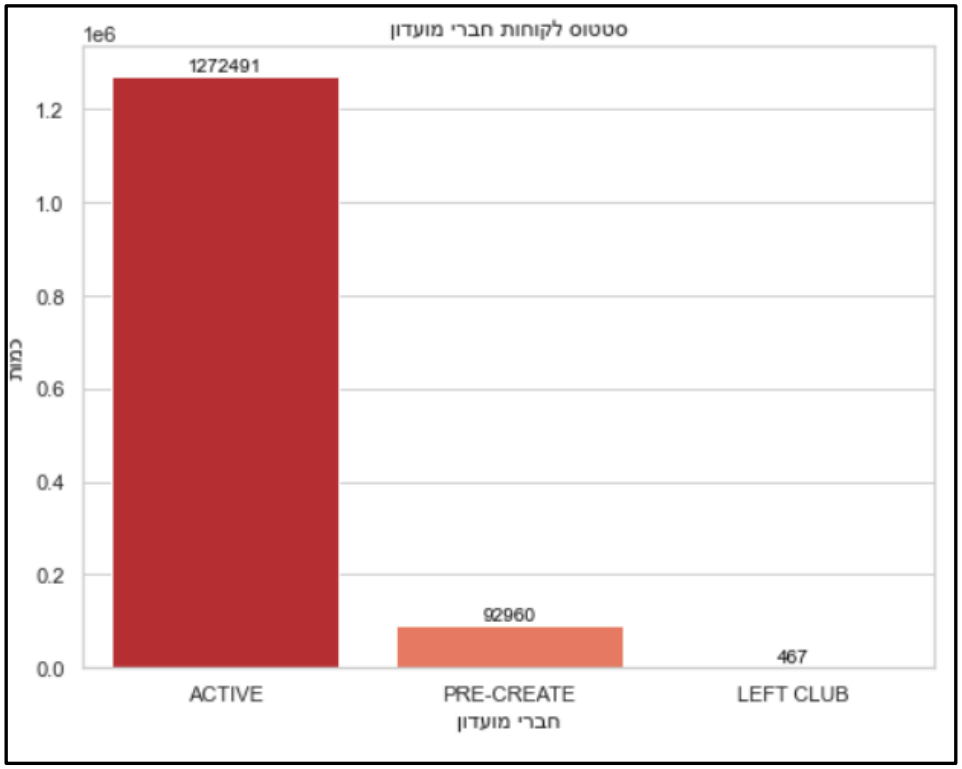


איור 6 – התפלגות גילאים לפי לקוחות פעילים

מעניין לראות שהתפלגות הגילאים של הלקוחות הפעילים ושל הלקוחות הלא פעילים פחות או יותר נראות אותו הדבר (אמנם כמות הלקוחות הלא פעילים גדולה יותר, כמעט בפי 2)

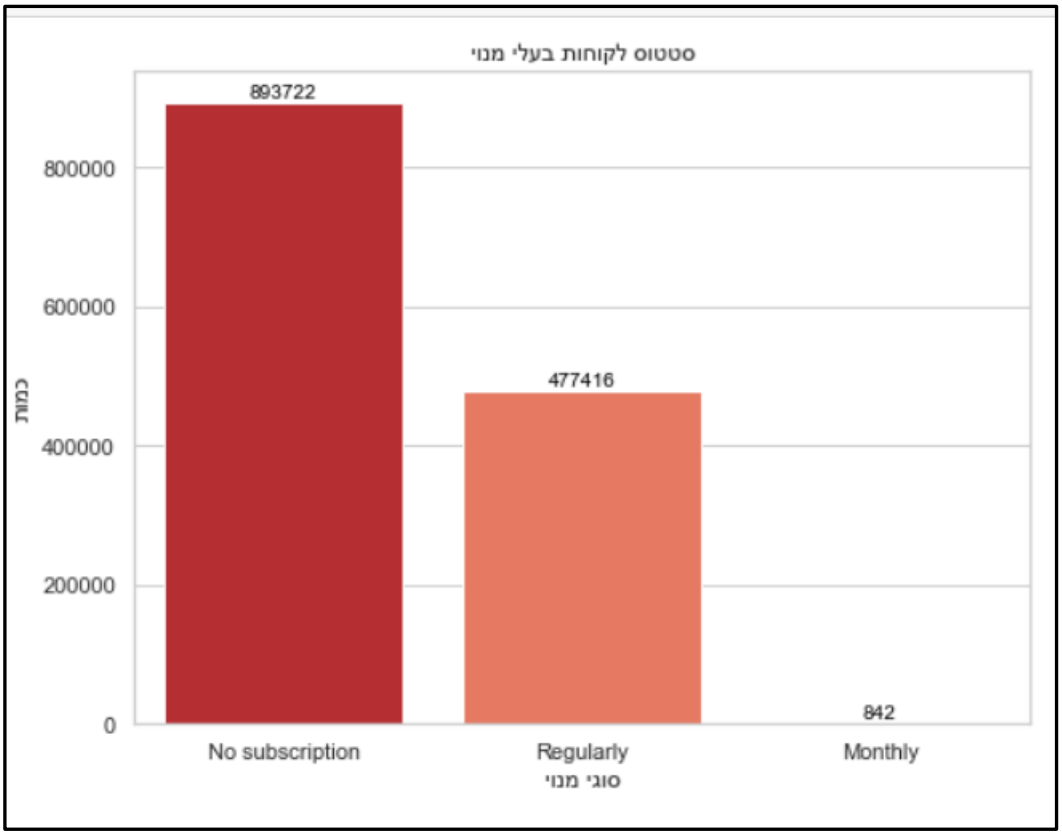


איור 7 – לקוחות פעילים מול לקוחות לא פעילים



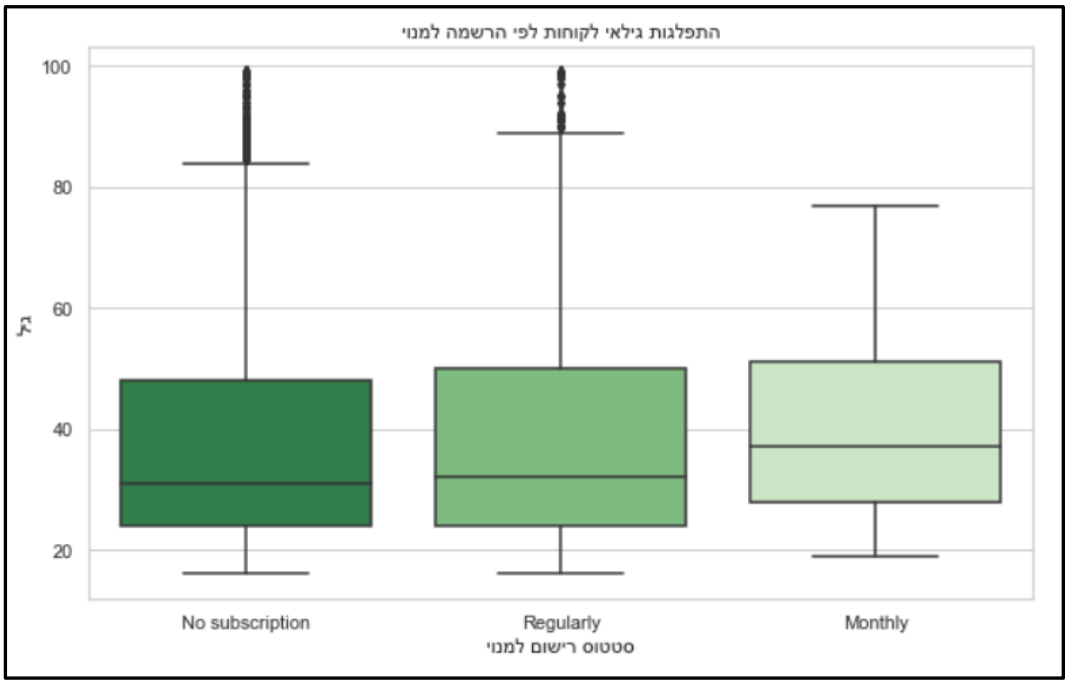
איור 8 – סטטוס לקוחות חברי מועדון

רוב הלקוחות הקיימים הינם חברי מועדון (אף על פי שאינם פעילים)



איור 9 – סטטוס לקוחות בעלי מנוי

ניתן לראות כאן שיש הרבה חברי מועדון שלא רשומים לשירות ההודעות של המועדון.



איור 10 – התפלגות גילאי לקוחות לפי הרשמה למנוי

חשבתי שאולי אראה איזה שהוא הבדל בגילאים בין הלקוחות שרשומים למנוי עדכונים אך נראה שאין באמת הבדל בין הגילאים השונים לבין הרשמה למנוי עדכונים

<u>טבלת טרנזקציות:</u>

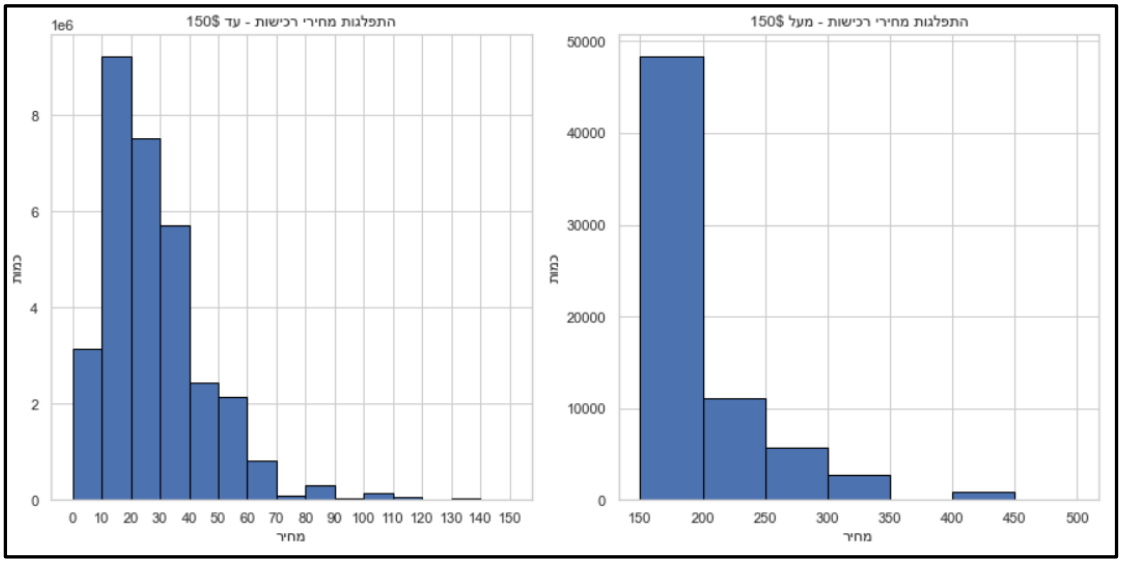
- טבלת טרנזקציות:
  - טבלת הטרנזקציות מכילה מידע אודות הרכישות שביצעו המשתמשים השונים
  - הטבלה מכילה 5 עמודות
- להלן פירוט העמודות:
  - תאריך רכישה (string)
  - מזהה לקוח (מפתח זר, ערך ייחודי, string)
  - מזהה מוצר (מפתח זר, ערך מספרי ייחודי)
  - מחיר (float)
  - ערוץ מכירות (int, קטגוריאלי)

ראשית, שיניתי את סוג עמודת התאריך, שיהיה מסוג תאריך ולא מסוג object.

הטרנזקציה הישנה ביותר היא משנת 2018 (חודש ספטמבר) כאשר האחרונה (כלומר החדשה ביותר) היא משנת 2020 כלומר יש לי תיעוד כל הרכישות התבצעו בשנתיים האחרונות. כאשר בסה"כ מדובר על 31,788,324 טרנזקציות.

H&M היא חברת fast fashion כלומר מותג בגדים זולים וטרנדיים שדוגמים רעיונות מהמסלול או מתרבות הסלבריטאים והופכים אותם לבגדים בחנויות במהירות, בעצם H&M לוקחת את המראה והאלמנטים העיצוביים מבתי האופנה המובילים ומשלבת אותם בייצור הבגדים שלהם במהירות ובזול.

ניתן לראות זאת על התפלגות המחירים בטרנזקציות השונות. הרוב המוחלט של הרכישות לא עולה על 100$

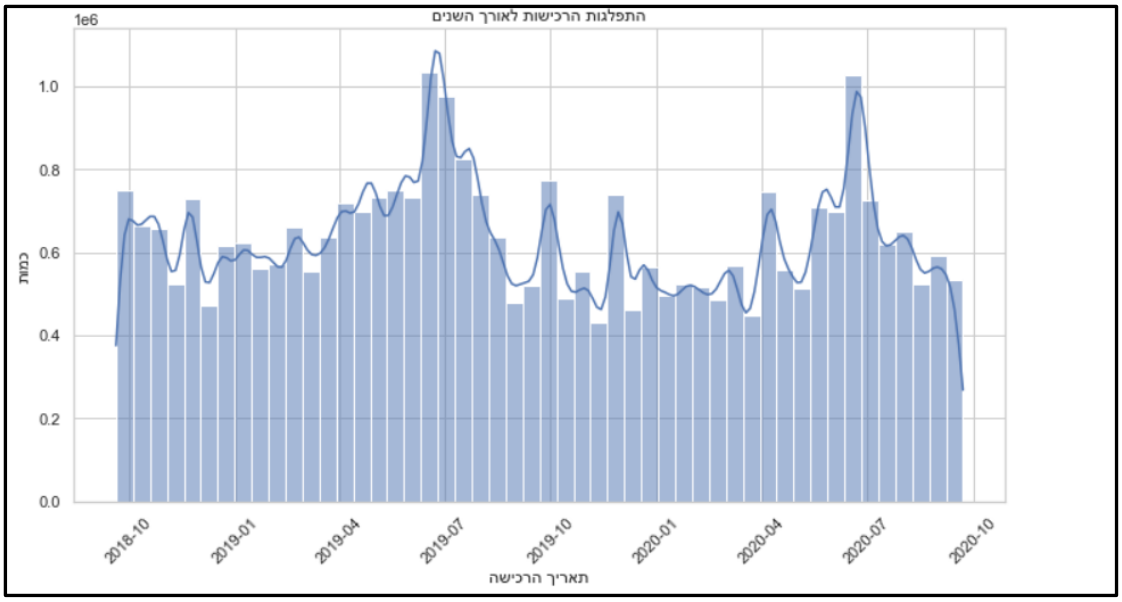


איור 11 – התפלגות מחירי הטרנזקציות

מצד שמאל, ניתן לראות את התפלגות הרכישות עד 150$ ניתן לראות שרוב הפריטים שנמכרים נעים בין 10-30, מצד ימין ניתן לראות את הכמות המועטה של רכישות במחירים גבוהים.

Deep Learning – NFC Recommendation System

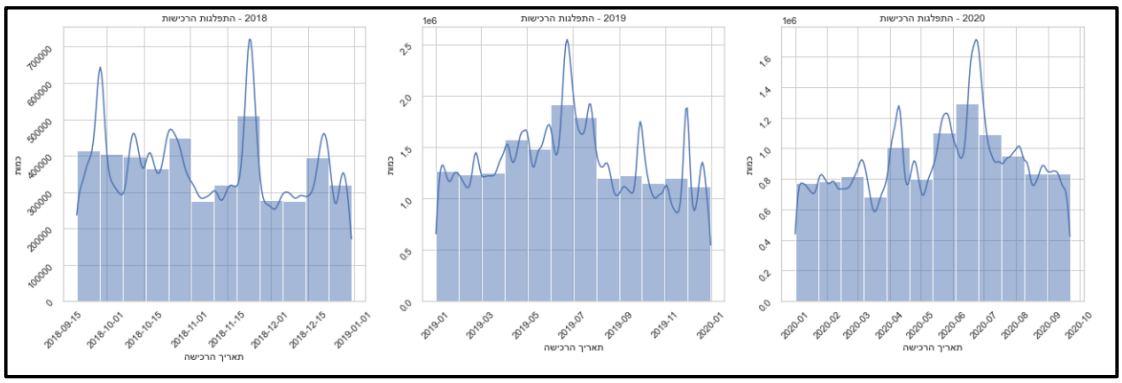
בחנתי את התפלגות הרכישות לאורך השנים (להבין אם יש איזה שהיא מגמתיות וכו')



איור 12 – התפלגות הרכישות לאורך השנים

ניתן לראות פיזור רכישות יחסית אחיד לאורך השנים עם שני "פיקים" בחודשי הקיץ (יולי אוגוסט)

נכנסתי מעט יותר לעומק והסתכלנו על ההתפלגויות לפי השנים

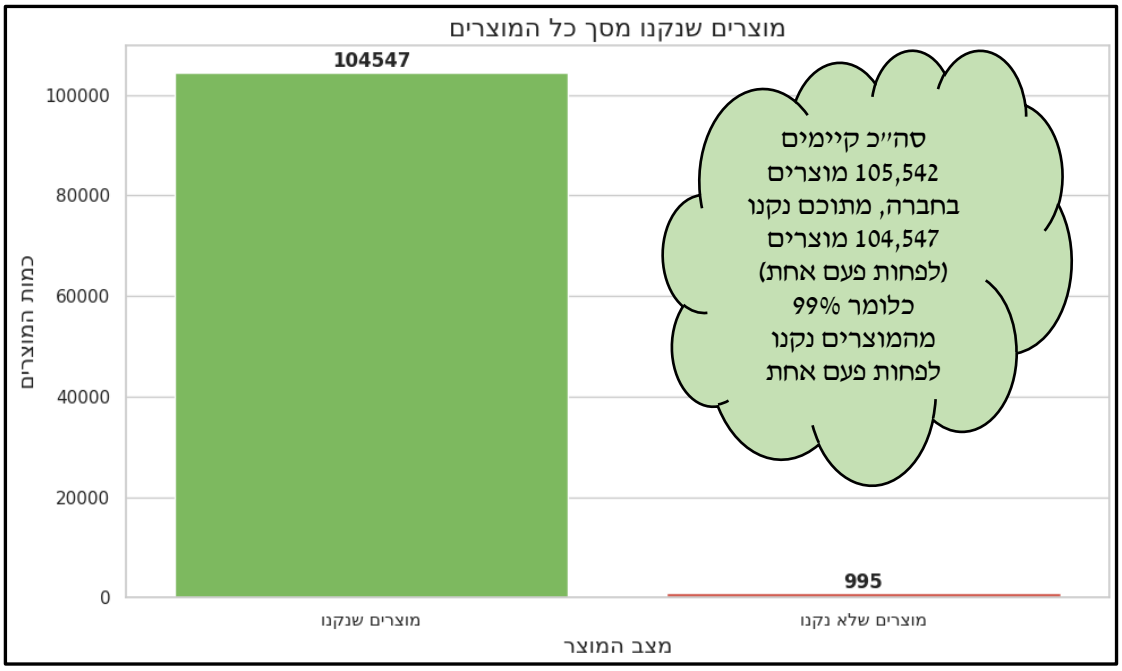


איור 13 – התפלגויות הרכישות לפי שנה

ניתן לראות שפחות או יותר ההתפלגויות בין השנים זהות

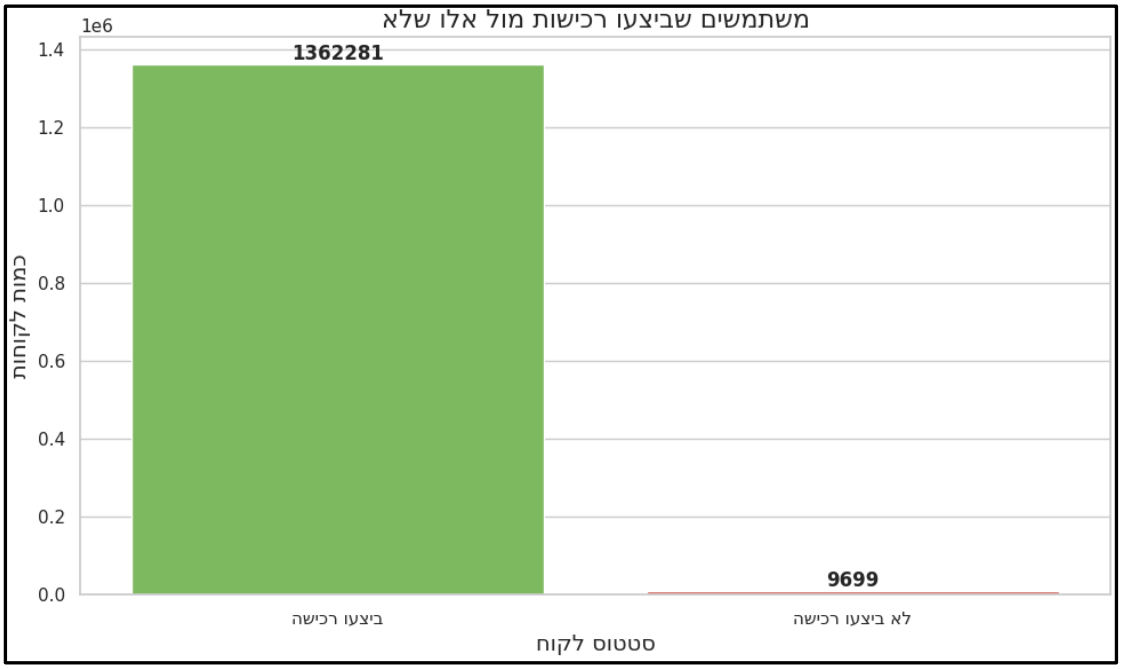
Deep Learning – NFC Recommendation System

בנוסף עניין אותי לראות כמה מוצרים נקנו מסך כל המוצרים הקיימים. גילינו שכמעט כל המוצרים נמכרו לפחות פעם אחת. יותר נכון להגיד ש-99% מהמוצרים נמכרו לפחות פעם אחת.

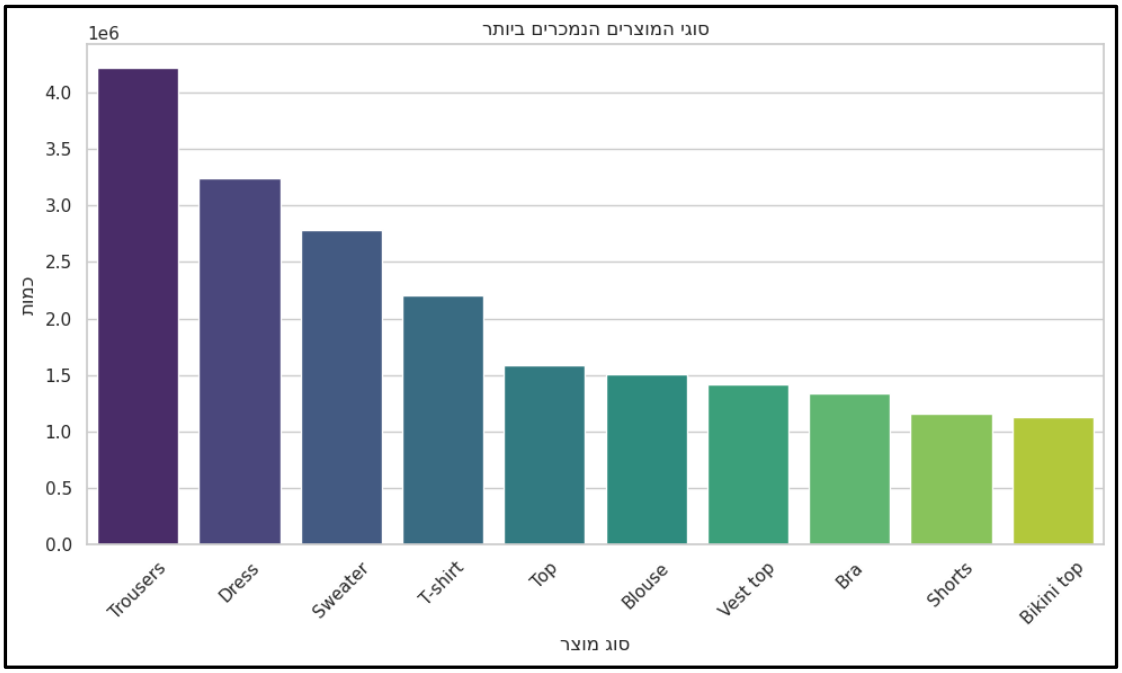


איור 14 – מוצרים שנקנו מסך כל המוצרים

ואם נסתכל על הלקוחות שביצעו רכישות מסך כל הלקוחות (1,371980) הקיימים נוכל לראות ש99.3$ מהם ביצעו רכישות אך קיימים עדיין כמעט 1,000 לקוחות שכלל לא ביצעו רכישה.

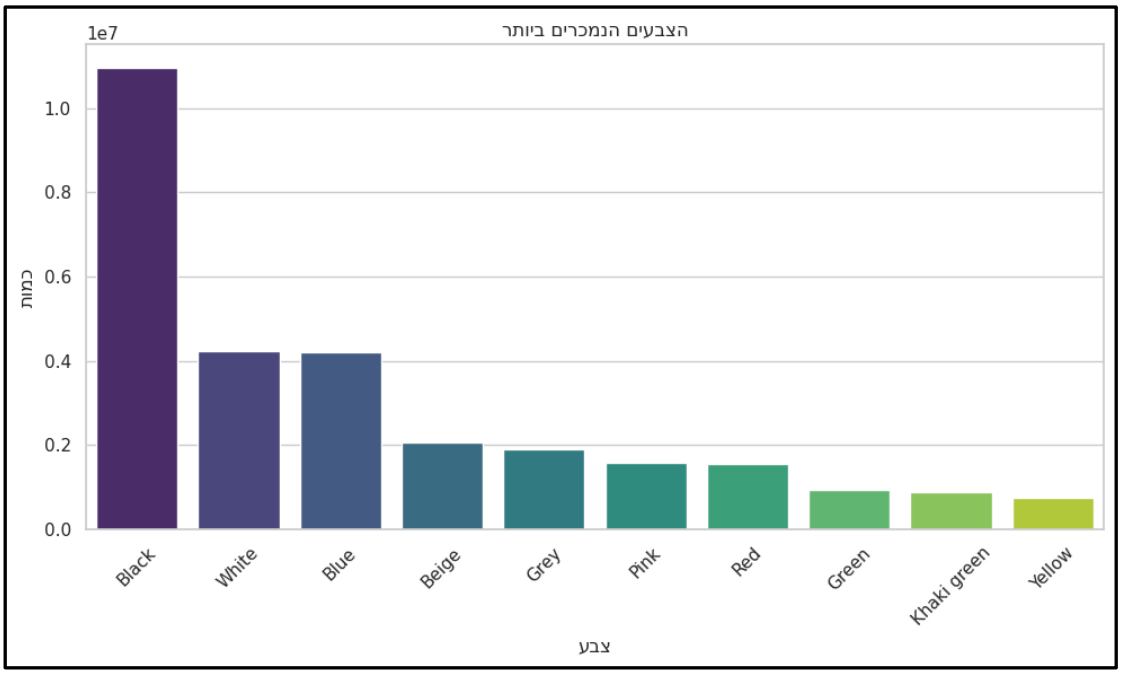


איור 15 - לקוחות שביצעו רכישה מול אלו שלא



איור 16 – סוגי המוצרים הנמכרים ביותר

על פי הגרף הנ״ל ניתן לראות שהמוצרים שנקנים הכי הרבה ברשת H&M הם מכנסיים לאחר מכן שמלות אחר כך סווטשרים וחולצות



איור 17 – הצבעים הנמכרים ביותר

כמו כן, ניתן לראות שהצבע השחור הוא הצבע הנמכר ביותר בפער די גדול משאר הצבעים, לאחר מכן לבן וכחול פחות או יותר באותה הכמות.

ממוצע רכישות לפי זמנים		
ממוצע רכישות ליום	**ממוצע רכישות לשבוע**	**ממוצע רכישות לחודש**
43,408	294,336	1,271,532

סיכום:

בסיס הנתונים של H&M מכיל לא מעט נתונים אודות הפריטים, לקוחות ורכישות שנעשו ע"י הלקוחות.

בנוסף לטבלאות שבחנו כאן, קיים עוד מאגר תמונות עבור כל פריט. וכמו כן עמודת תיאור מוצר (בטבלת המוצרים)

כעת שאני מבין היטב את הנתונים, סוגם, ואילו מניפולציות אנחנו צריכים לבצע בכדי להכין אותם לעבודה עם מודל של למידה עמוקה.

פרק 2 – הכנת סט הנתונים לעבודה עם מודל של למידה עמוקה:

טבלת המוצרים:

כבר בשלב הקודם של חקירת הנתונים, סידרתי וניקיתי את הנתונים לטובת הצגת סטטיסטיקות והגרפים השונים. טבלת המוצרים הכילה 25 עמודות שונות – רובם היו עמודות קטגוריאליות. שמתי לב שישנם שהרבה עמודות הכילו את אותו המידע אך בתצורות שונות, כלומר חלק מהעמודות סיווגו את הקטגוריות לפי מספרים (לדוגמא קוד צבע מוצר: 135) ועמודה נוספת תיארה בדיוק את אותו הדבר רק במלל (לדוגמא צבע מוצר: כחול). לכן הורדתי את כפילות העמודות וכך צמצמתי את טבלת המוצרים רק ל-13 עמודות נבחרות. כל העמודות היו ללא נתונים חסרים ולכן לא הייתי צריך לבצע מניפולציות שונות עבור מילוי נתונים חסרים.

לאחר מכן, קודדתי את כל העמודות הקטגוריאליות בעזרת One Hot Encoding ובעצם יצרתי טבלה שמכילה למעלה מ-600 עמודות. הדבר בא בעוכרי, שכן הטבלה הכילה מספר רב של רשומות ומספר רב של עמודות בהמשך כאשר ניסיתי לשלב אותה עם טבלת הטרנזקציות – שמכילה למעלה מ-30,000,000 רשומות המשאבים של המחשב שלי פשוט לא הספיקו וכבר בשלב זה, חיבור הטבלאות נכשל מפאת חוסר זיכרון. על כן צמצמתי את מספר העמודות בטבלת המוצרים.

זיהיתי עמודות בעלות קורלציה גבוהה והחלטנו לוותר על חלק מהם לטובת צמצום הטבלה לדוגמא, היו 3 עמודות שמייצגות את צבע המוצר – "צבע נתפס", "צבע עיקרי" , "קבוצת צבעים" שלושת העמודות הכילו מידע בעל קורלציה גבוהה כלומר דומה במשמעות שלו ועל כן השארתי בטבלה רק את עמודת "צבע עיקרי". מהלכים דומים עשיתי עם עוד 3-4 עמודות עד שלבסוף צמצמתי את הטבלה ל-3 עמודות בלבד ולאחר קידוד מחדש היו לי סה"כ 154 עמודות (לחלק מהעמודות היה סיווג להרבה קטגוריות ולכן הכמות הגדולה של העמודות לאחר הקידוד). לעמודות הקטגוריאליות הוספנו את עמודת מזהה המוצר (מפתח) ואת עמודת שם המוצר.

Deep Learning – NFC Recommendation System

טבלת הלקוחות:
טבלת הלקוחות עברה עיבוד דומה, באופן כללי טבלת הלקוחות הינה טבלה קטנה יותר ובה רק שתי עמודות קטגוריאליות, עבור שתי עמודות אילו ביצענו קידוד בעזרת One Hot Encoding ועבור עמודת הגילאים סיווגתי אותה לקבוצות ואחר כך ביצענו עליה גם קידוד.

לדעתי עמודות "FN" ו- "Postal code" לא היו רלוונטיות יותר מדי (כלומר לא תרמו לנו מידע) ולכן החלטתי להסיר אותם מהטבלה. כך שלבסוף נשארתי עם טבלה אחת שמכילה את מזהה הלקוח, האם הוא פעיל האם הוא חבר מועדון , האם הוא מנוי לידיעות ועדכונים ועמודת קבוצת הגיל של הלקוח. לאחר הקידוד נוצרה לנו טבלת לקוחות עם 13 עמודות.

טבלת טרנזקציות:
בטבלה זו קיבלתי מידע אודות הרכישות של הלקוחות השונים, בעזרת טבלה זו גם הנדסנו תכונות חדשות שיכולות להסביר בצורה טובה יותר את האינטראקציה של המוצר והמשתמש.

תחילה חישבתי את כמות הטרנזקציות עבור כל משתמש ולאחר מכן גם את סכום הכסף אותו הוציא. משתי תכונות בניתי עוד 3 תכונות מעניינות שיכולות להסביר בצורה טובה יותר את האינטראקציה של המשתמש עם המוצר.

וכמו כן, הסרתי עמודות שלא רלוונטיות כמו "channel_id" או עמודת המחיר שעליה עשיתי מניפולציות. לאחר הנדסת התכונות החדשות עליהם נסביר בהמשך, הסרתי גם את עמודת התאריך (כיוון שיצרתי עמודות חדשות בעלות קורלציה גבוהה לתאריך, כלומר מספרות את אותו הסיפור) ביצעתי מניפולציות על עמודת התאריך, ויצרתי עמודה חדשה שמתארת את הפרשי הזמנים בימים בין הטרנזקציות העוקבות עבור על לקוח.

הנדסת תכונות חדשות:
בשלב זה בעצם לקחתי את טבלת הטרנזקציות ובעזרתה בניתי עוד מספר תכונות (Feature Engineering) חדשות על מנת לתאר בצורה טובה יותר את הלקוח (היות ולא היה לי יותר מדי מידע על הלקוחות.

יצרתי עמודה שמציגה את מספר הרכישות שהלקוח ביצע עבור כל מוצר ועמודה נוספת שמתארת את סך ההוצאות עבור כל מוצר. לאחר מכן יכולתי לחשב את מספר הטרנזקציות הכולל עבור כל לקוח, סך הכסף שכל לקוח הוציא ומספר הימים הממוצע בין הרכישות של הלקוח.

- מספר רכישות עבור כל מוצר – קיבצתי לפי מזהה לקוח ומזהה מוצר ואז ספרתי את כמות המוצרים שהלקוח רכש
- סך ההוצאות עבור כל מוצר – קיבצתי לפי מזהה לקוח ומזהה מוצר ואז סכמתי את עמודת המחיר.
- מספר רכישות כולל – כאן בעצם קיבצתי רק לפי מזהה הלקוח וסכמתי את העמודה שיצרתי קודם "מספר הרכישות עבור כל מוצר" כתוצאה מכך קיבלתי את מספר הרכישות הכולל שביצע כל לקוח
- הוצאה כוללת – באופן דומה חישבתי את ההוצאה הכוללת עבור כל לקוח
- הזמן הממוצע בין רכישות – בעזרת העמודה שיצרתי שמחשבת את הפרש הזמנים בין כל רכישה של הלקוח, חישבתי את הממוצע.
- כמות סוג המוצר - את הטבלה הזו, איחדתי יחד עם עמודת סוגי המוצרים, ואז יצרתי עמודה נוספת שמציגה את כמות הרכישות שהלקוח ביצע עבור כל סוג מוצר.

היות והיו לי המון רשומות, סיננתי את הטבלה רק ללקוחות שעשו מעל 500 רכישות.

טבלה סופית:
לאחר בניית תכונות אלו, עשיתי קרוס ג'וין (מכפלה קרטזית) בין כל סוגי המוצרים לטבלת הטרנזקציות ואז בעצם כל לקוח קיבל רשומה עם כל סוג מוצר. וכמובן כאשר הלקוח לא ביצע רכישה עם מוצר מסוים, אז אחוז האינטראקציה שלו שווה ל-0.

כעת איחדתי את כל הטבלאות (טבלת הטרנזקציות + טבלת המוצרים + טבלת לקוחות)

לטבלה הזו קראתי interaction_data , הטבלה הנ"ל מורכבת מכמעט 10 מיליון רשומות המעידות על רכישות שונות של מוצרים ע"י משתמשים שונים ומכילה 156 עמודות (בעקבות הקידוד ועמודות מרובות קטגוריות).

כעת, הינדסתי תכונות נוספות שיכלו להסביר בצורה טובה יותר את האינטראקציה בין הלקוח למוצר.

- מחיר רכישה ממוצע – לקחתי את העמודה שיצרתי קודם (סך הוצאות) וחילקתי בעמודה אחרת שיצרתי (סך כל הרכישות) כתוצאה מך קיבלתי את מחיר הרכישה הממוצע. התכונה הזו יכולה לספק לנו תובנות לגבי התנהגות ההוצאות האופיינית של הלקוח והעדפות הרכישה שלו.
- אחוז היחסי עבור כל מוצר – לקחתי את מספר הטרנזקציות עבור כל מוצר וחילקתי במספר הטרנזקציות הכולל, כך למעשה קיבלתי את האחוז היחסי עבור כל מוצר. למעשה אחוז אינטראקציה גבוה מעיד על כך שללקוח יש מעורבות חזרה עם המוצרים הללו.
- האחוז היחסי עבור כל סוג מוצר – בדומה לעמודה הקודמת, כאן פשוט חילקתי את מספר הטרנזקציות עבור כל סוג מוצר. זו גם תהיה העמודה אותה ארצה לחזות. מכיוון שאחוז גבוהה מעיד על כך שללקוח יש אינטראקציה גבוהה עם סוגי המוצרים בקטגוריה הזו.

** הסרתי את העמודה שבעזרתה חישבתי את הלייבל שלנו (מכיוון שהיא כמובן מסבירה בצורה טובה את הלייבל ואז המודל שלנו לא יהיה טוב, או יותר נכון יהיה טוב מדי)

מראש הוצאתי מתוך הטבלה 2 רשומות בכדי שאחר כך אוכל לבצע עליהם את הפרדיקציה.

לאחר מכן הגדרתי את מזהה המוצר ומזהה הלקוח כאינדקס של הטבלה והסרתי כפילויות. וכעת הטבלה הסופית שלי צומצמה לכמעט 700,000 רשומות.

בודדתי את משתנה המטרה שלנו מסט הנתונים, היות וטווח הערכים של המשתנה הזה היה מאוד קטן, בין 0 ל-0.5 (וכאשר בדקתי את המודל בפעם הראשונה גילינו שהטעות היא מאוד קטנה ואפילו אפסית, הבנתי שככל הנראה מדובר בטווח הערכים של המשתנה שאותו אני רוצה לחזות) לכן בחרתי להכפיל את כל הערכים ב100 וכך קיבלנו טווח ערכים בין 0-50.

על שאר הנתונים ביצעתי סקיילינג וכעת אני מוכן לעבוד עם מודל של למידה עמוקה.

פרק 3 – בניית ארכיטקטורה מתאימה + Fine Tuning:

חילקתי את סט הנתונים שיצרתי לסט נתונים של אימון ומבחן. את סט האימון חילקתי פעם נוספת לסט אימון וסט ולידציה.

כאמור, הכנסתי סט יחיד של נתונים לכן המודל שלנו יקבל רק אינפוט אחד של מספרים. על מנת ליצור ארכיטקטורה מתאימה יצרתי פונקציה כזו שמקבלת פרמטרים שונים לבניית המודל ובנוסף הייפרפרמטרים שונים בונה את המודל מקפלת אותו ואז מאמנת אותו, הפונקציה כוללת בתוכה גם את האופציה לעצירה מוקדמת במידה ובמהלך 7 איפוקים לא חל שינוי.

לאחר מכן בניתי גריד שמכיל סוגים שונים מכל פרמטר (בין אם זה כמות השכבות, כמות נוירונים, פונקציות אקטיבציה שונות וכו'), ראוי לציין שהפרמטרים שנבדקו נבחרו בהתאם לבעיה שאותה אנחנו רוצים לפתור.

Deep Learning – NFC Recommendation System

לדוגמא, תחילה שמשתנה המטרה שלנו היה בטווח ערכים של 0-0.5 אז כן בחרנו לבדוק פונקציית אקטיבציה של סיגמואיד, אחרי שהגדלנו את טווח הערכים בין 0-50 אז כבר לא היה מתאים להשתמש בה.

מדדנו את המודל בעזרת MAE.

```
יצירת ספרייה עם כל ההיפר פרמטרים שאני רוצה לבדוק #
param_grid = {
    #מספר הנוירונים
  'num_units': [128,256],
    #מספר השכבות
  'num_layers': [1,2,3],
    #פונקציות אקטיבציה
  'activation': ['leaky_relu','linear','relu'],
    #הסתברות לדרופ
  'dropout_rate': [0.0,0.2,0.3],
    #רגולציה
  'l1_reg': [0.005,0.1],
  'l2_reg': [0.005,0.1],
    #אופטימיזציות
  'optimizer': ['sgd', 'adagrad', 'adadelta','rmsprop'],
    #(מדובר במודל רגרסיה ולכן זו הפונקציה המתאימה) פונקציות שגיאה
  'loss': ['mse','mae'],
    # הגדרת מדד ביצוע mse
  'metrics':['mae'],
  # הגדרת מספר איפוקים
  'epochs':[20],
  'batch_size': [64, 128]

}
```

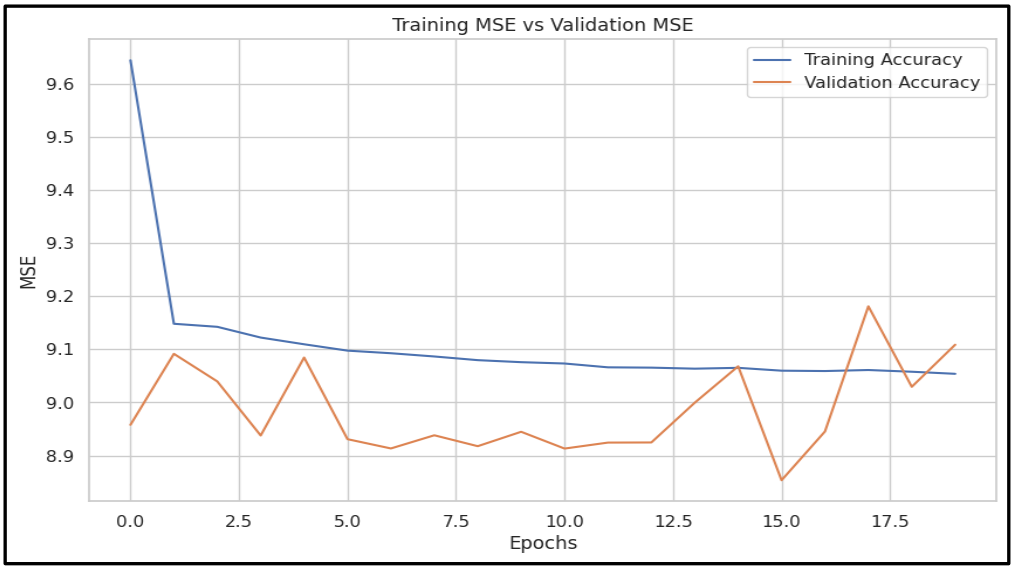
לטובת הבדיקה לקחתי מדגם של 7,000 רשומות מתוך סט הנתונים והרצתי את הבדיקה. בעצם יצרתי 3456 אופציות שונות לבניית הארכיטקטורה, כאשר בסוף בחרנו את זו עם השגיאה הקטנה ביותר.

- מספר הנוירונים שנבחר הוא 128
- המודל מכיל שכבה אחת
- פונקציית האקטיבציה היא linear
- Drop rate נקבע ל-0.3
- L1_reg = 0.005
- L2_reg = 0.005
- Optimizer = rmsprop
- Loss function = mse
- הפלט הוא כמובן output יחיד שמנבא את ההסתברות של הלקוח לקנות את סוג המוצר (ללא פונקציית אקטיבציה)

כאשר הטעות (mae) עמדה על 8.9 שהחלטנו שזה די סביר ביחס לטווח הערכים ובהקשר המשימה שלי (מערכת המלצה)

לאחר מכן אימנתי את המודל עם הארכיטקטורה וההייפר פרמטרים הללו עבור סט הנתונים המלא. עם 20 איפוקים

Training loss vs Validation loss – איור 18



Training mse vs Validation mse – איור 19

לאחר שבחנו, החלטתי להריץ את המודל על סט המבחן עם 5 איפוקים בלבד.

וכך למעשה הכנתי מודל ואימנו אותו עבור מערכת ההמלצה.

פרק 4 – יצירת מערכת ההמלצה:

על מנת ליצור את מערכת ההמלצה, יצרתי פונקציה, שלוקחת את התחזיות (פרדיקציות) עבור יוזרים ומוצרים (השתמשנו בשתי היוזרים שהחרגנו מסט הנתונים).

הפונקציה לוקחת את תוצאות התחזיות וממזגת אותם עם התוצאות המקוריות (כך שאנחנו יודעים להתאים כל תחזית ללקוח ספציפי)

לאחר מכן הפונקציה מבצעת מיון של עמודת הפרדיקציה ולוקחת את 3 התוצאות הגבוהות ביותר עבור כל לקוח. (כלומר שלושת סוגי המוצרים כי מתאימים ללקוח)

ובעצם כך אנחנו מקבלים את ההמלצות ל-3 סוגי המוצרים בעלי האינטראקציה הגבוהה ביותר עם הלקוח.

בגלל שאנחנו מדברים על סוגי מוצרים, בתוך כל קטגוריה יש מספר רב של מוצרים, לקחנו את שלושת סוגי המוצרים ופשוט יצרנו פונקציה שמציע 5 מוצרים רנדומליים מתוך שלושת הקטגוריות.

כך שלבסוף המערכת מוציאה פלט של 5 מזהיי מוצרים שהם המוצרים שעליהם אנחנו ממליצים ללקוח.

הפלט יוצא כמילון כאשר כל מזהה לקוח הוא המפתח והערכים הם רשימה של 5 המוצרים המומלצים עבור אותו לקוח.

לדוגמא עבור מזהה לקוח:
d44dbe7f6c4b35200abdb052c77a87596fe1bdcc37e011580a479e80aa940001

המערכת המליצה על המוצרים הבאים:

[611050002 ,625027006 ,794389004 ,61113003 ,850899002]

611050002

611130003

611134001

התמונות להמחשה.

פרק 5 – סיכום ותובנות:

לסיכום, הצלחתי ליצור את מערכת ההמלצה שרציתי. לקחתי בסיס נתונים, מאתר Kaggle עם 3 טבלאות, חקרתי ולמדתי אותו. (העבודה נעשתה על סביבת עבודה של Google Colab) בצעד הבא יצרתי תכונות שמאפיינות את המוצרים ותכונות שמאפיינות לקוחות ותכונות שמאפיינות את אופי הרכישות של הלקוח.

לאחר חשיבה מעמיקה, יצרתי את המשתנה שלדעתי מייצג בצורה הטובה ביותר את האינטראקציה בין המוצר ללקוח.

הכנסתי הכל לטבלה אחת אותה קודדתי והתאמתי אותה לעבודה עם מודל של למידה עמוקה.

בניתי מודל והחלטתי על ארכיטקטורה מתאימה, בדקתי אפשרויות שונות להייפרפרמטרים באמצעות גריד ולאחר מכן אימנתי את המודל על סט הנתונים המלא.

בעזרת המודל המאומן, יצרתי פונקציה שיודעת לקבל טבלה של לקוחות ולהתאים להם את 5 המוצרים בעלי האינטראקציה הגבוהה ביותר עבורם.

אמנם מערכת ההמלצות עובדת בצורה תקינה, אך יש לומר שהיא אינה מושלמת. היות ולא היה לי את כל הזמן, משאבים או ידע היא לא תמיד ממליצה באופן הכי טוב עבור הלקוח, לדוגמא, אחד הדברים הבסיסים ביותר, לדעת להמליץ על מוצר לגברים או לנשים. היות ולא היו לנו הנתונים האם הלקוח הוא זכר או נקבה אז לא יכולנו גם לחלק את המוצרים למוצרי נשים או גברים.