דוח מסכם- פרויקט בקורס 'מבוא ללמידת מכונה'

שם
עומרי שהם
עדו לוי

נספח א': פירוט אחריות כל שותף ותרומתו לעבודה
נספח ב': ויזואליזציה המוזכרת בדו"ח זה
נספח ג': שאר ויזואליזציות

תקציר מנהלים

בפרויקט זה המטרה הייתה לייצר את מודל החיזוי הטוב ביותר על מנת לחזות האם סשן גלישה באתר מסוים יסתיים ברכישה \ לא יסתיים ברכישה. קיבלנו סט נתוני אימון עם עמודת לייבלים, וסט נתוני מבחן עליו ירוץ המודל הסופי ותבוצע הפרדיקציה. יש לציין כי לרשותינו עמד סט נתוני אימון "מלוכלך" שכלל מספר עמודות אנונימיות (חלקן מספריות וחלקן קטגוריאליות), ערכים חסרים וגם חוסר תאימות של הנתונים איתם התמודדנו בשלבי העיבוד המקדים (data tidying).
לאחר ביצוע התהליך של עיבוד הנתונים והשלמת ערכים חסרים, תוך השמטת ערכי קיצון והתמודדות עם משתנים קטגוריאליים, ביצענו גם נרמול של הנתונים על מנת שנוכל "להאכיל" את המודל בהם. מעבר לכך, ביצענו הכלה של העיבוד המקדים על סט הולידציה ששימש אותנו לטובת הערכת המודלים, וכמובן גם החלנו את העיבוד המקדים על סט המבחן. כל זאת על מנת להריץ את המודלים, לבחון ולהעריך אותם, כאשר בסופו של דבר מצאנו שהמודל המיטבי הינו Random Forest בעל ציון AUC = 0.945.

Data Exploration

שלב זה הינו שלב מהותי בהתהוותו של המחקר, מטרתו המרכזית היא בחינת הנתונים ממספר זוויות שונות על מנת להבין את הסיפור שעומד מאחוריהם, גיבוש דרכי פעולה וחשיבה על הצורה המיטבית לביצוע ויזואליזציה של הנתונים. בנוסף, בשלב זה התרכזנו במחשבה וביצוע של ניקוי הנתונים, אילו ערכים אנחנו מחליפים \ משלימים ובאיזו צורה על מנת שלא לגרום להטייה מסוימת שלא בכוונת תחילה.

<u>נתונים כללים:</u>
סט האימון בגודל של 10,479 שורות ו-23 פיצ'רים (מתוכם אחד הוא הלייבל).
הלייבל בינארי- 0 מסמל סשן שלא הסתיים ברכישה ו-1 מסמן סשן אשר הסתיים ברכישה, מצאנו כי אחוז הרכישות מתוך סט האימון עומד על 15.47%.
מתוך 23 עמודות ישנן עמודות 8 קטגוריאליות ו-15 עמודות מספריות.

במסגרת שלב זה הצגנו היסטוגרמה של המשתנים המספריים על מנת להבין כיצד כל עמודה מתפלגת (איור 1 בנספח ב').
בהמשך, שלב משמעותי היה מילוי ערכים חסרים. גילינו שהפיצ'ר total_duration הינו בעל ערכים חסרים בכ-50% מהשורות הקיימות בסט הנתונים. בשלב מאוחר יותר, פעלנו על מנת להשלים את הערכים בו מתוך העבודה שהוא הסכום של העמודות admin_page_duration, info_page_durtaion, product_page_duration. מסקנות אלו נתמכו בעזרת גרף עמודות (איור 2 בנספח ב'.(
במסגרת העיסוק בערכים החסרים, בדקנו את קיום הקורלציה בין הערכים החסרים שבפיצ'רים (איור 3 בנספח ב'). בביצוע בדיקה זו שאפנו לגלות האם ישנה סיבה מסוימת לקיום ערכים חסרים בין עמודות, גילוי שכזה, היה גורם לנו להבין שיש משמעות מסוימת להשלמת הערכים החסרים בעמודות אלו. מצאנו, כי העמודות בעלות הקרולוציה הגבוהה ביותר הן exit_Rates ו- page_Value, כאשר הקורלציה בין החוסרים עומדת על 0.6%, שהיא כשלעצמה קורלציה לא מאוד גבוהה. נתון זה יכול ללמד אותנו כי ישנו קשר בין אחוז העזיבות של האתר מעמוד מסויים לבין כמות הדפים ששווים כסף (עמודים אשר יש פוטנציאל רכישה בהם, לדוגמא עמוד הגדרות משתמש אינו שווה כסף). בהם ביקר המשתמש.

במעבר על עמודות המתארות את משך הגלישה בעמודים השונים (3 פיצ'רים), ראינו כי בחלק מהרשומות מצויין הזמן בדקות (כיתוב minutes), אך מתוך בחינה על ערכים רבים היה לנו מוזר לראות שישנם הרבה מאוד שורות בעלות זמנים מאוד ארוכים לשהייה בעמוד. על כן, אנו חושבים כי ייתכן וקיימת טעות בהצגת הנתונים כפי שקיבלנו אותם (אנו מאמינים כי הנתונים האמיתיים מדברים על שניות). מכיוון שבהמשך העבודה אנו ננרמל את הנתונים החלטנו כי אין צורך לבצע המרות לשניות או ליחידת זמן אחרת.

כמו כן, הצגנו גרף קורלוציה בין העמודות השונות (איור 4 נספח ב'). בתחילה, הערכנו כי תרשים זה ישמש אותנו להורדות מימדים, תחת ההנחה כי אם ישנן עמודות בעלות קורלציה גבוהה, ניתן להשאיר אחת מהן בלי לפגוע במידע שהעמודה השנייה תורמת לנו. עם זאת, בהמשך העדפנו להשתמש ב- Feature imortance לטובת הורדת המימדים, אותו ביצענו בשלב הרצת המודלים. זאת, על מנת לקבל את המודל המיטבי.
נתון מעניין שמצאנו הוא מתוך בחינה שביצענו עבור החודש שבו ישנן הכי הרבה רכישות באופן יחסי ומצאנו כי חודש נובמבר הוא החודש בו מתבצעות הכי הרבה רכישות (איור 5 נספח ב'). קל לבחון כי סוף שבוע מהווה

28.6% (יומיים מתוך 7 ימים), היינו מצפים שאחוז הרכישות בסוף השבוע יהיה לפחות כך , ברם מצאנו שהוא מעט פחות מזה (איור 6 נספח ב').

Preprocessing

בשלב זה תחילה הורדנו את פיצ'ר D אשר ברובו היו ערכים חסרים וכעת נותרנו עם 21 פיצ'רים. בהמשך, זיהינו קשר בין מספר עמודים למשך הזמן עמוד (בהתאמה- כאשר ישנן 6 עמודות כאלו, 3 זוגות). ראינו, כי קיימות שורות שבאחת העמודות קיים ערך חסר ואילו בעמודה המתאימה לה קיים ערך 0. בהתאם לכך, הבנו שאפשר למלא את האחת בעזרת השניה. אם באחת העמודות הערך הוא 0, אזי שבעמודה המתאימה לה גם כן צריך להיות 0. זאת, מתוך ההנחה כי לא ניתן לבקר יותר מ-0 דק' עבור סשן שסוג מסוים של עמוד היה ב-0 עמודים, ולהיפך. לצורך זה כתבנו פונקציה המשלימה את ערכים אלו (find_ziro_and_replace), זאת על מנת לא לאבד ערכים רלוונטים.

לאחר מכן, ראינו לנכון לכתוב פונקציה אשר מוחקת שורות בהן יש יותר מ-5 ערכי nan כערך סף לשורה אשר אינה מייצגת את סט הנתונים (בעזרת מחקנו 54 שורות). מצד אחד לא רצינו למחוק יותר מדי שורות, ומצד שני הבנו ששורה בעלת 5 ומעלה ערכים חסרים (המהווים כ-25% מהדאטה) הינה שורה שכנראה לא מייצגת בצורה טובה את המציאות.
כעת, התייצבנו אל מול הדילמה כיצד נכון למלא ערכים חסרים בעמודות מספריות. התמקדנו בעמודות pages ו- duration, יצרנו דאטה חדש ובו כל עמודה היא בוליאנית, עבור שורה שבה יש ערך חסר מילאנו 1 ולהיפך 0, לאחר מכן בדקנו את הקשר בין עמודות אלו. מצאנו שאין קשר מלבד עמודות num_of_info_pages_nan ל-info_page_duration_min_nan. כלומר, את יתר העמודות אפשר להשלים בצורה מספרית כלשהו (ממוצע, חציון , קבוע) ואת העמודות בעלות הקשר נדרש להשלים בצורה שונה.
בהמשך, ניסינו להשלים את עמודות אלו וכן עמודות נוספות בעזרת KNN אך נתקלנו בקושי בשלב מאוחר יותר בהחלת סט האימון על הטסט. זאת, משום שלא רצינו לבצע KNN על סט המבחן כדי לא "ללמוד" אותו מצד אחד, ומצד שני לא רצינו לבצע תהליך עיבוד מסוים על סט האימון ותהליך עיבוד אחר על סט המבחן, דבר שהיה עלול לגרום לסטייה באמינו המודל וביצועיו. לבסוף, לאחר ניסיונות רבים השלמנו את עמודות אלו לפי הערך השכיח.
בהמשך במסגרת Feature engineering יצרנו עמודה חדשה בעזרת עמודת month אשר מציגה את עונות השנה, שינינו את עמודת weekend לעמודה בוליאנית לטובת המשך העבודה עם סט הנתונים.

לאחר מכן, התמודדנו עם ערכי קיצון (Dealing with outlires) בחנו את ההתפלגות של כלל העמודות וסיננו את הערכים הקיצוניים מה שהביא להורדה של 3.2% מהשורות. לפיצ'רים שמתפלגים נורמלית בחנו ערכי קיצון בעזרת Boxplot והורדנו אותו על ידי שימוש בפונקציה שכתבנו (remove_outlier_iqr). בעזרת כך הורדנו 1.5% מהשורות.
בהמשך ביצענו Data normalization בעזרת mmscaler על מנת שכלל הערכים יהיו באותה סקאלה, על למנוע הטיות אפשריות במודל.

בשלב הבא, התמודדנו עם משתנים קטגוריאלים בעזרת get_dummies. במהלך שלב זה הגענו ל-58 עמודות ובדקנו תחילה את הקורולוציה בין כלל העמודות. זיהינו כי ישנה קורולוציה גבוהה בין מספר עמודות (מפורט במחברת). בעקבות זאת, תחילה חשבנו כי יהיה נכון להוריד את עמודות אלו לטובת הקטנת המימדים, ברם,

בהמשך העבודה ולאחר הרצת המודלים גילינו כי מיטבי יותר יהיה להוריד את המימדים בעזרת Feature imortance.
לאחר בחינת השונות של כל פיצ'ר החלטנו שכשלב ראשוני לרדת מ58 פיצ'רים ל-34 (איור 8 נספח ב'). קו המחשבה בנקודה זו היה שאיננו רוצים לוותר על פיצ'רים בעלי שונות רבה, להיפך. זאת, משום שאנו מעדיפים נתונים מגוונים, הטרוגניים, אשר כל אחד מהם מספר לנו מגוון רחב של סיפורים, מאשר נתונים הומוגניים. על כן, ומתוך הכוונה לאמן את המודל על מציאות כמה שיותר רחבה ומורכבת, העדפנו פיצ'רים בעלי שונות גבוהה.

בשלב הבא, חילקנו את הדאטה ל-train and validation והחלנו את כל המתואר לעיל על הסט הזה, כאשר המטרה היא לבצע עליו עיבוד מקדים זהה לזה שנבצע על סט המבחן, כדי שנוכל לבחון את המודל בצורה מהימנה.

Models

במסגרת שלב זה התבקשנו לבחור 4 מודלים (2 כמודלים ראשוניים ו-2 כמודלים עיקריים) שבעזרת המודל הטוב ביותר נריץ את המודל הסופי.
המודלים שבחרנו הם:

1. KNN- `Best Params: {'n_neighbors': 350}`
2. Logistic Regression- `Best Params: {'C': 0.1, 'penalty': 'l1', 'solver': 'liblinear'}`
3. Decision Tree- `Best Params: {'criterion': 'gini', 'max_depth': 5, 'random_state': 0}`
4. Random Forest- `Best Params: {'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 2, 'n_estimators': 193, 'random_state': 0}`

את ההיפר פרמטרים עבור כל מודל בחרנו על ידי שימוש ב GridSearch מתוך ספריית Sklearn, כאשר ביצענו מספר איטרציות על כל מודל על מנת לדייק את טווח הערכים שמוביל לציון AUC מיטבי (איור 9 נספח ב').
ההיפר פרמטרים הכי טובים שמצאנו מופיעים לעיל.
עבור כל מודל חישבנו את המדדים הבאים:

1. MSE
2. AUC
3. Accuracy
4. STD
5. K-fold Cross Validation
6. Confusion Matrix
7. ROC

מתוך ה- Confusion Matrix אשר מצורפת לכל מודל שהרצנו ניתן ללמוד מספר מטריקות באמצעותן ניתן להעריך את המודל. Accuracy, Precision, sensitivity and specificity.

כמובן שיש את המתח של sensitivity-specificity tradeoff, בהמשך הצגנו את עקומת ROC curve - AUC metric שבאה לתת ציון אחד עבור המודל. זוהי מטריקה שנוח להשתמש בה שבעקרון לוקחת בחשבון את כל המדדים המוזכרים לעיל ונותנת ציון למודל.

<u>מתוך התאים במטריצה ניתן ללמוד-</u>

TP -(1,1)- מספר הסשנים שהסתיימו ברכישה, וגם המודל סיווג אותם כבאלה.

TN -(0,0)- מספר הסשנים שלא הסתיימו ברכישה, וזיהינו אותם כבאלה שלא יסתיימו ברכישה.

FN -(0,1)- מספר הסשנים שהתחזית שלנו טעתה לגביהם, והם לא הסתיימו ברכישה, כלומר המודל חזה שכן יסתיימו ברכישה.

FP -(1,0)- מספר הסשנים שהתחזית שלנו טעתה לגביהם, והם כן הסתיימו ברכישה, כלומר המודל חזה שאלו לא יסתיימו ברכישה.

המודל עבורו קיבלנו את הביצועים הכי טובים הינו Random Forest עם ציון AUC של 0.945. גם מתוך בחינה של ה- Confusion Matrix במודל זה אל מול השוואה ליתר המטריצות אותן קיבלנו בשאר המודלים, אנו מסיקים כי זהו המודל המיטבי שמצאנו.

בנוסף, על המודל הנבחר הרצנו בדיקה של feature importance שנתנה לנו להבין את מידת החשיבות של כל פיצ'ר עבור המודל. בשלב זה, הבנו כי כמעט חצי מהפיצ'רים עליהם הרצנו את המודלים הקודמים, כמעט ולא תורמים בחשיבותם (איור 7 נספח ב').

לכן, הרצנו פעם נוספת את המודל הנבחר, הפעם עם 17 הפיצ'רים הכי חשובים, מה שבסופו של דבר הביא לשיפור המודל מקבלת ציון 0.944 לציון 0.945. ניתן לראות כי המעבר משימוש ב- 33 פיצ'רים לשימוש ב- 17 פיצ'רים לא תרם באופן משמעותי לביצועי המודל.

<u>סיכום</u>

בפרויקט זה התמודדנו עם אתגרים רבים שכללו הבנה מעמיקה של הנתונים- הייחודיות והתרומה של כל פיצ'ר, מה הדרך הנכונה להשלים ערכים חסרים בו, האם יש קורלציה בינו לבין משתנים אחרים. בנוסף, הבאנו לכדי יישום פרקטי כלים אותם למדנו במהלך השיעורים והתרגולים לאורך הקורס. מעבר לכך, היה עלינו לנוע הלוך ושוב בין השלבים השונים של הפרוייקט מספר פעמים על מנת ליישם תובנות שהבנו בשלבים מתקדמים יותר, אשר השפיעו על האופן בו ביצענו חלקים מוקדמים יותר. ראוי לציין כי היו מספר דרכי פעולה וגרפים שבסופו של דבר החלטנו להשמיט (והשקענו בהם כמה שעות יפות..), מתוך ההבנה כי אלו לא תורמים רבות למודך או שלא תורמים למי שקורא את הפרויקט.

מאוד נהננו מביצוע הפרויקט וניקח ממנו הרבה כלים להמשך הן במסגרת לימודית והן במסגרת עבודתנו.

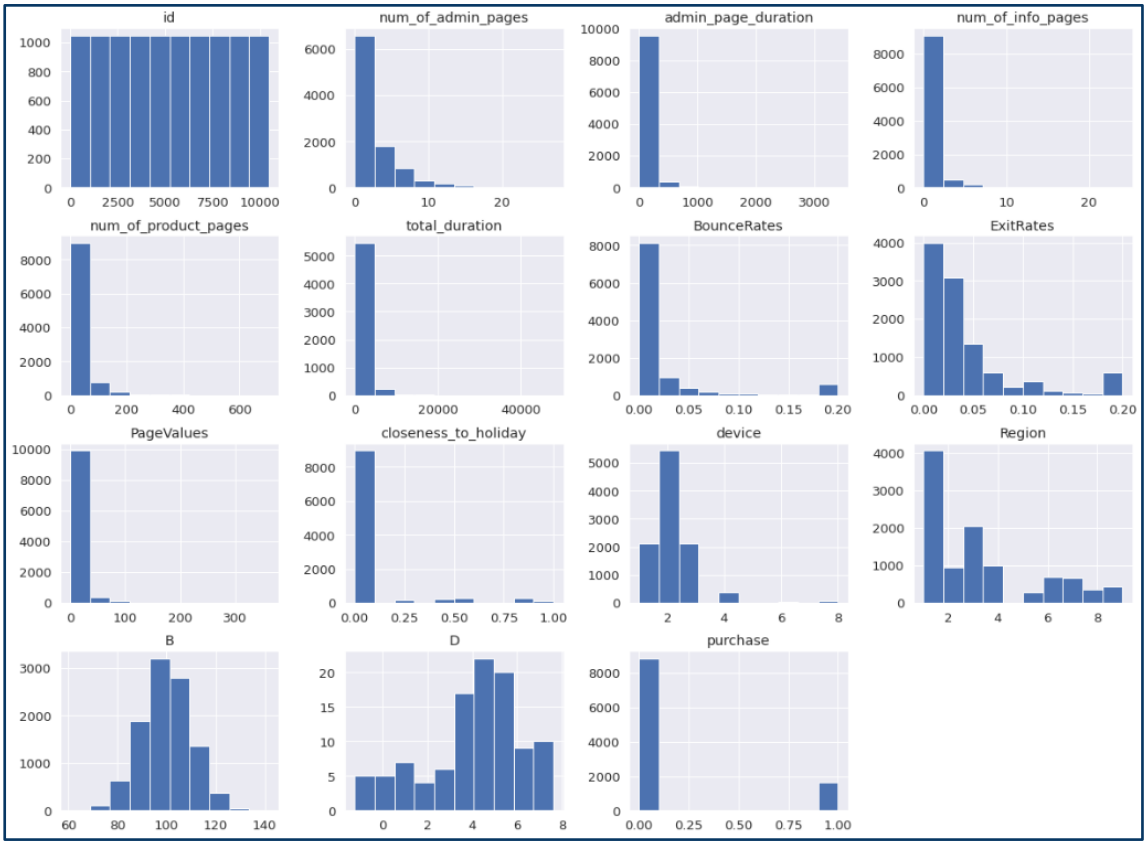
<u>נספח א'</u> :פירוט אחריות כל שותף ותרומתו לעבודה

אמל"ק- לא הייתה חלוקת עבודה ברורה.

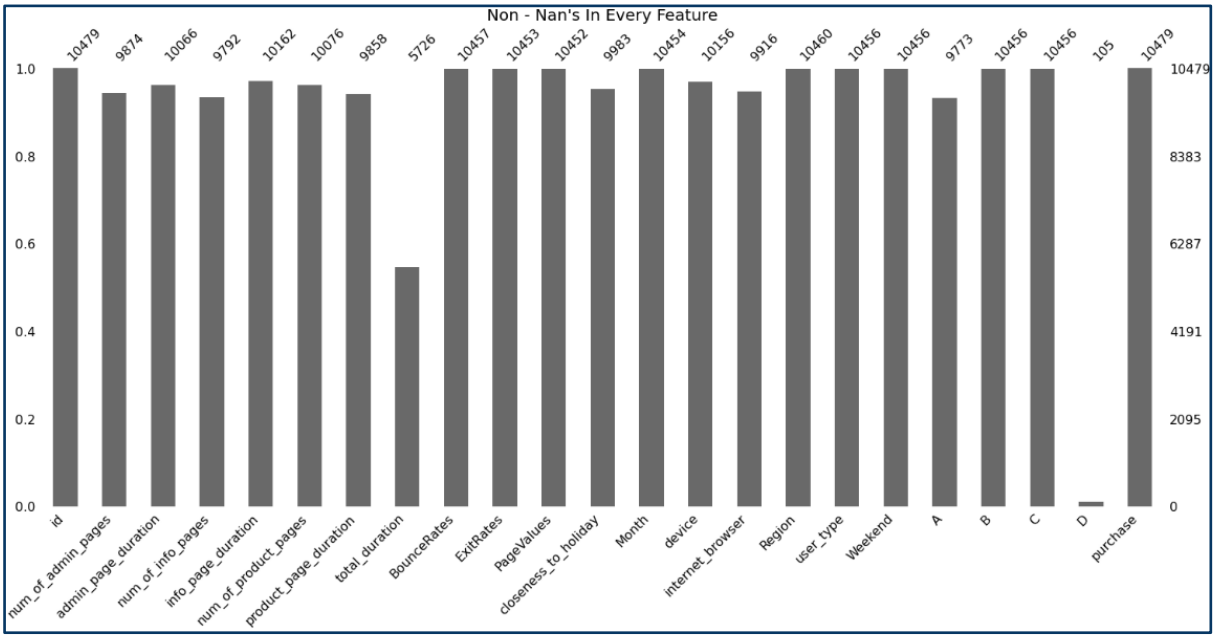
- מאחר ואנחנו חברים קרובים מזה כ-10 שנים, לומדים ביחד במהלך כל התואר, היה לנו טבעי שגם את הפרוייקט הזה נבצע ביחד.
- לא הייתה חלוקה מסוימת של אחראויות עבור כל אחד מאיתנו במהלך ביצוע פרוייקט זה.
- שיטת העבודה הייתה שעברנו על שלבי הפרוייקט ביחד, תכננו ביחד את תכנית הפעולה, וגם את הביצוע בפועל עשינו יחדיו.

נספח ב': ויזואליזציה המוזכרת בדו"ח זה

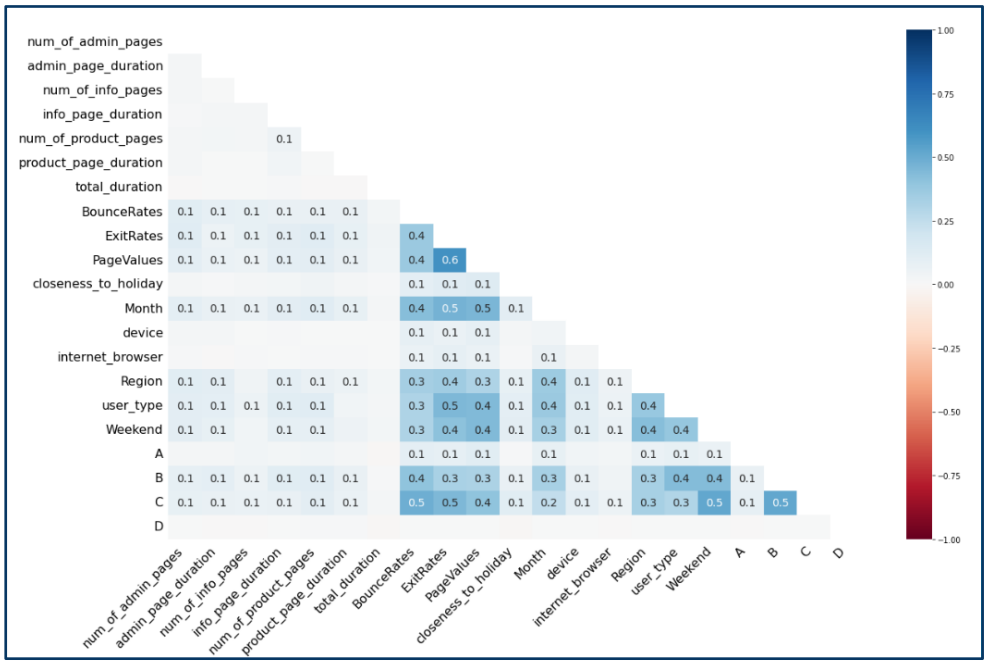
איור 1



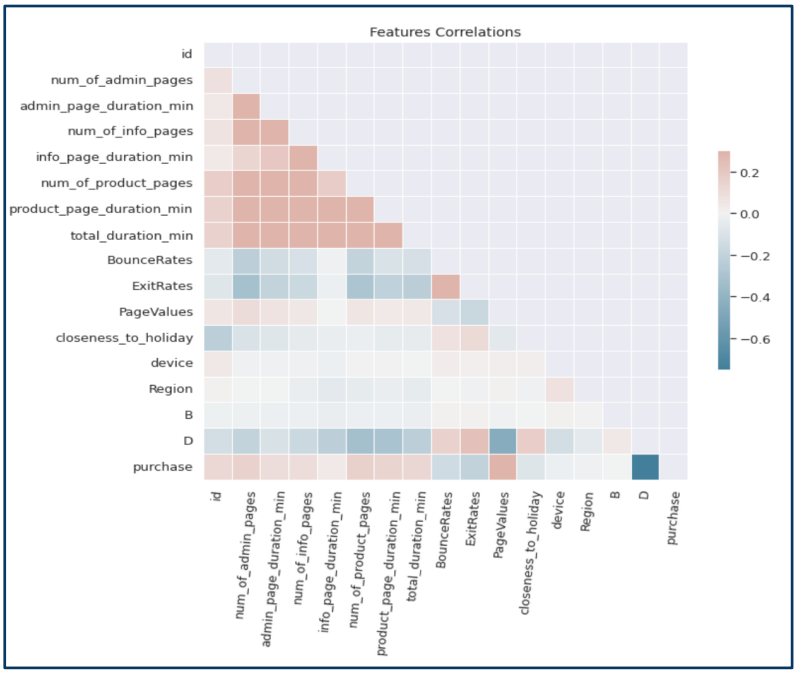
איור 2


Non - Nan's In Every Feature
10479
9874
10066
9792
10162
10076
9858
5726
10457
10453
10452
9983
10454
10156
9916
10460
10456
10456
9773
10456
10456
105
10479
1.0
0.8
0.6
0.4
0.2
0.0
10479
8383
6287
4191
2095
0
id
num_of_admin_pages
admin_page_duration
num_of_info_pages
info_page_duration
num_of_product_pages
product_page_duration
total_duration
BounceRates
ExitRates
PageValues
closeness_to_holiday
Month
device
internet_browser
Region
user_type
Weekend
A
B
C
D
purchase

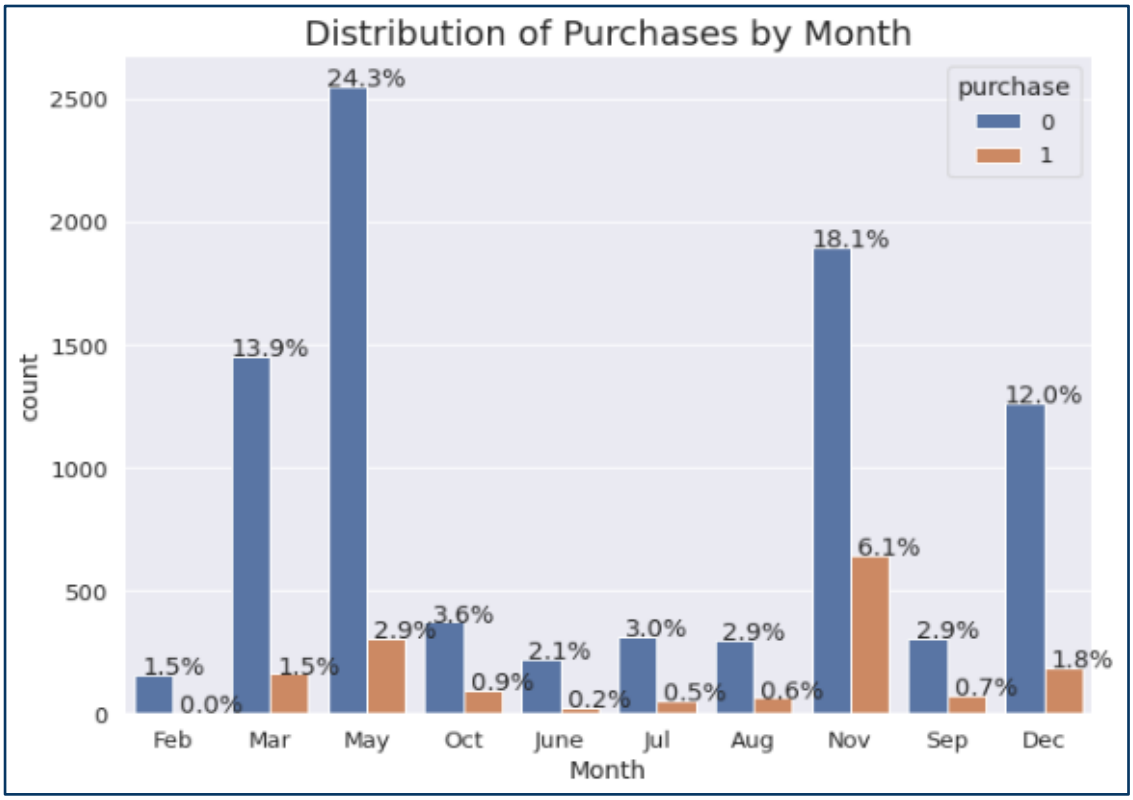
איור 3


num_of_admin_pages
admin_page_duration
num_of_info_pages
info_page_duration
num_of_product_pages
product_page_duration
total_duration
BounceRates
ExitRates
PageValues
closeness_to_holiday
Month
device
internet_browser
Region
user_type
Weekend
A
B
C
D
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

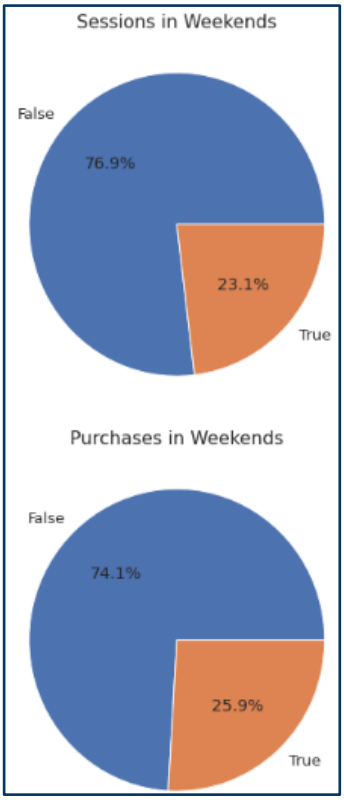
איור 4


Features Correlations
id
num_of_admin_pages
admin_page_duration_min
num_of_info_pages
info_page_duration_min
num_of_product_pages
product_page_duration_min
total_duration_min
BounceRates
ExitRates
PageValues
closeness_to_holiday
device
Region
B
D
purchase
0.2
0.0
−0.2
−0.4
−0.6

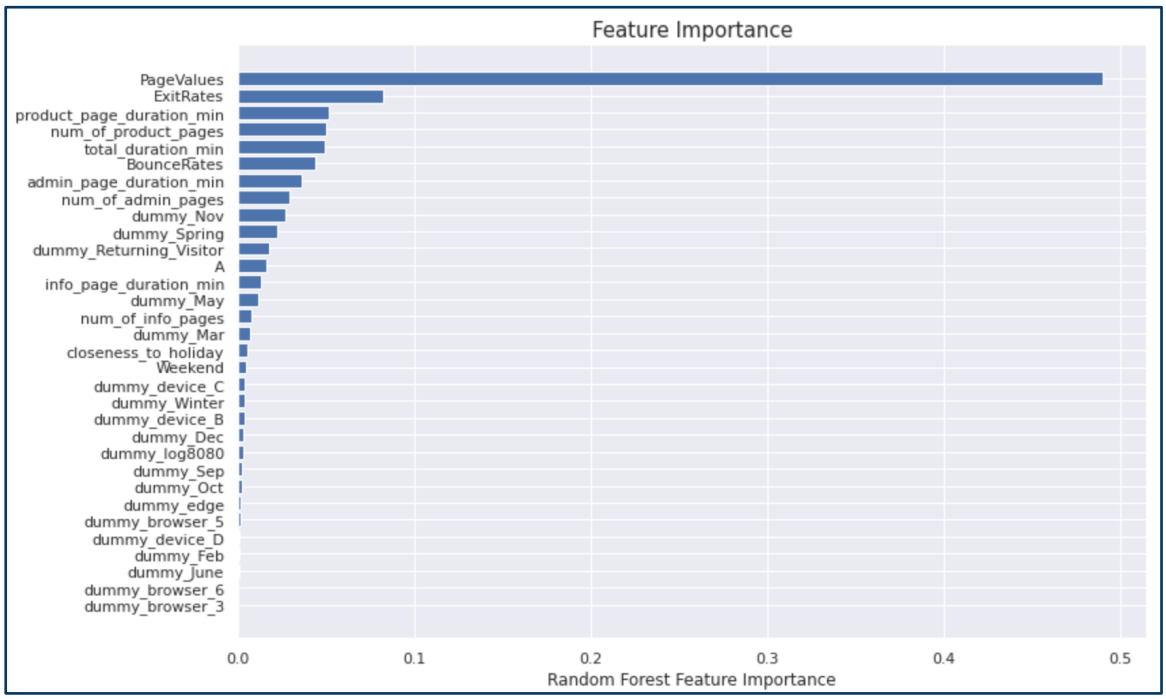
איור 5


Distribution of Purchases by Month
purchase
0
1
count
2500
2000
1500
1000
500
0
1.5%
0.0%
13.9%
1.5%
24.3%
2.9%
3.6%
0.9%
2.1%
0.2%
3.0%
0.5%
2.9%
0.6%
18.1%
6.1%
2.9%
0.7%
12.0%
1.8%
Feb
Mar
May
Oct
June
Jul
Aug
Nov
Sep
Dec
Month

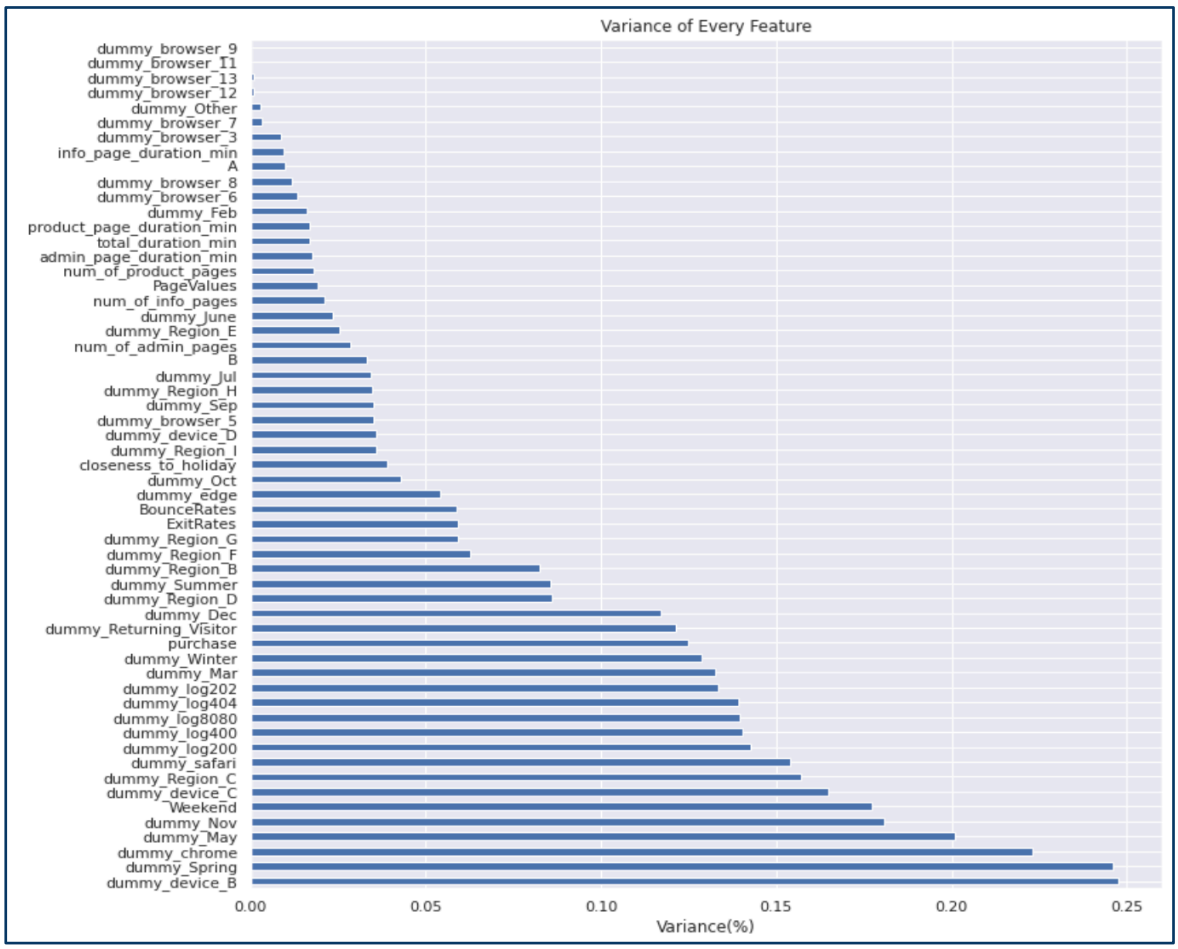
איור 6


Sessions in Weekends
False
76.9%
23.1%
True
Purchases in Weekends
False
74.1%
25.9%
True

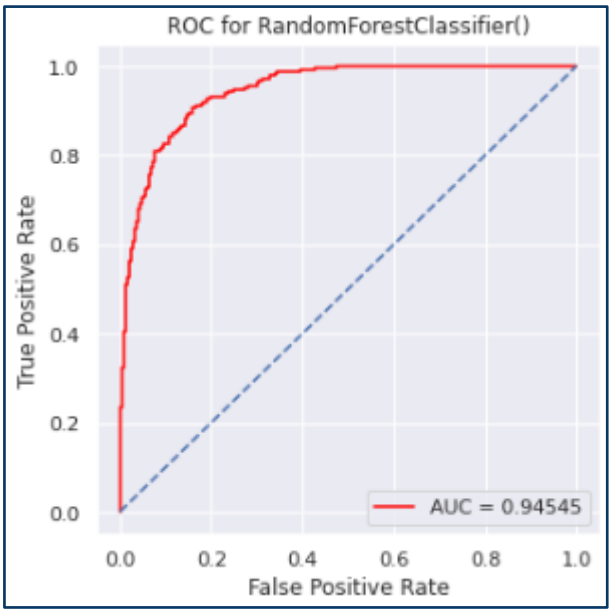
איור 7

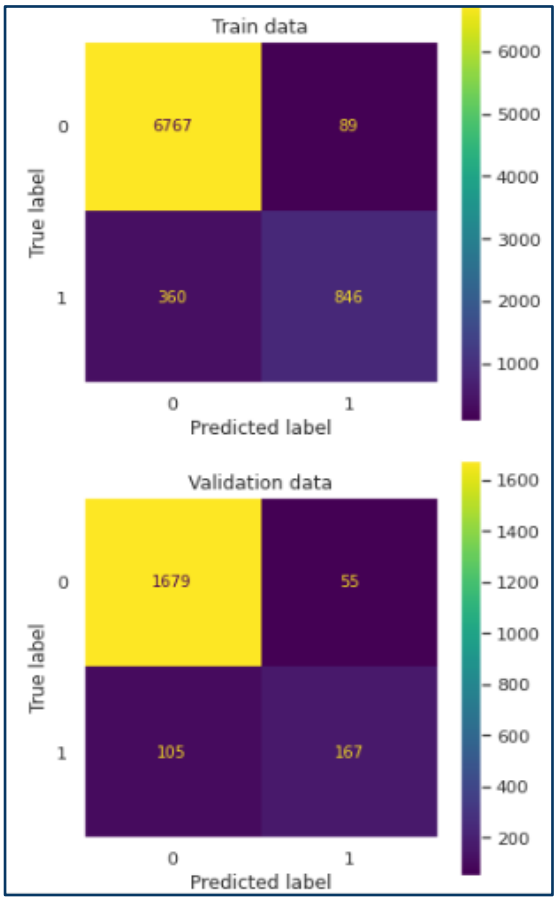

Feature Importance
PageValues
ExitRates
product_page_duration_min
num_of_product_pages
total_duration_min
BounceRates
admin_page_duration_min
num_of_admin_pages
dummy_Nov
dummy_Spring
dummy_Returning_Visitor
A
info_page_duration_min
dummy_May
num_of_info_pages
dummy_Mar
closeness_to_holiday
Weekend
dummy_device_C
dummy_Winter
dummy_device_B
dummy_Dec
dummy_log8080
dummy_Sep
dummy_Oct
dummy_edge
dummy_browser_5
dummy_device_D
dummy_Feb
dummy_June
dummy_browser_6
dummy_browser_3
0.0
0.1
0.2
0.3
0.4
0.5
Random Forest Feature Importance

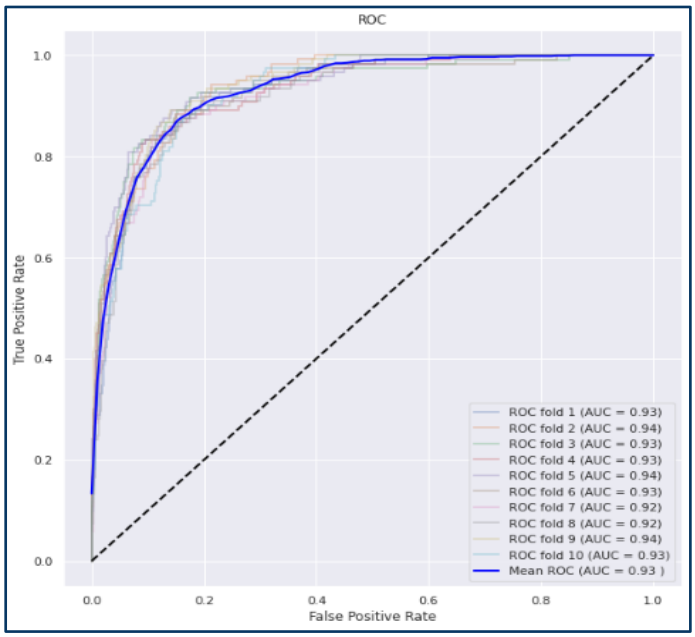
איור 8



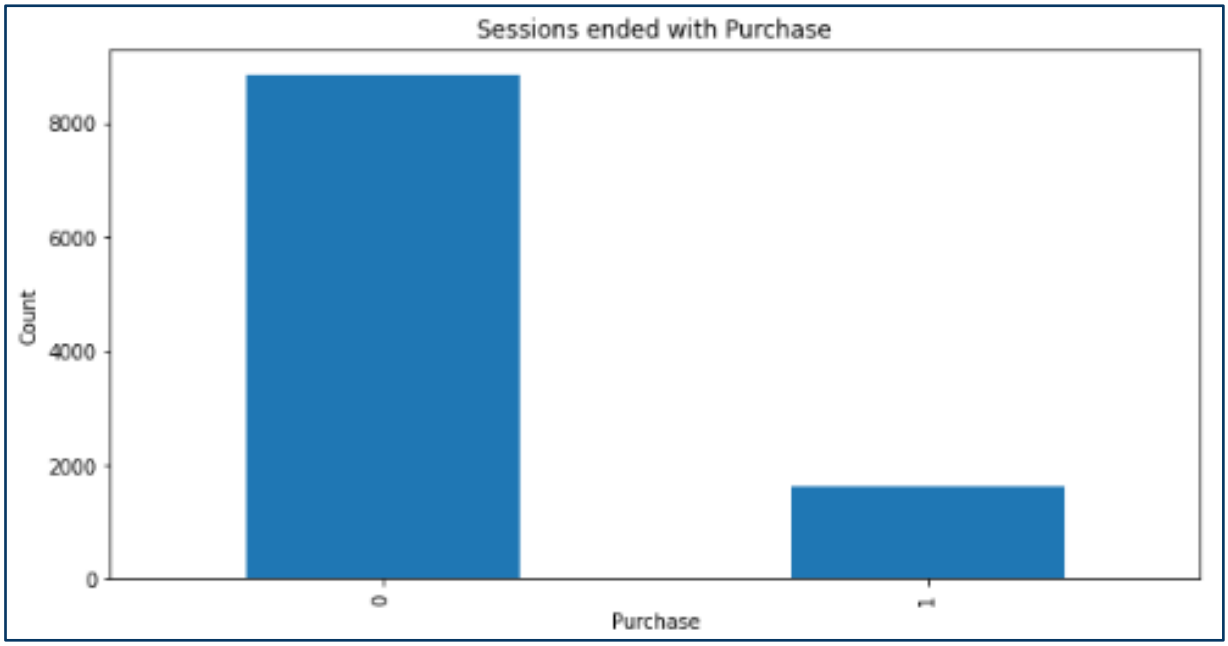
איור 9 - המודל הנבחר

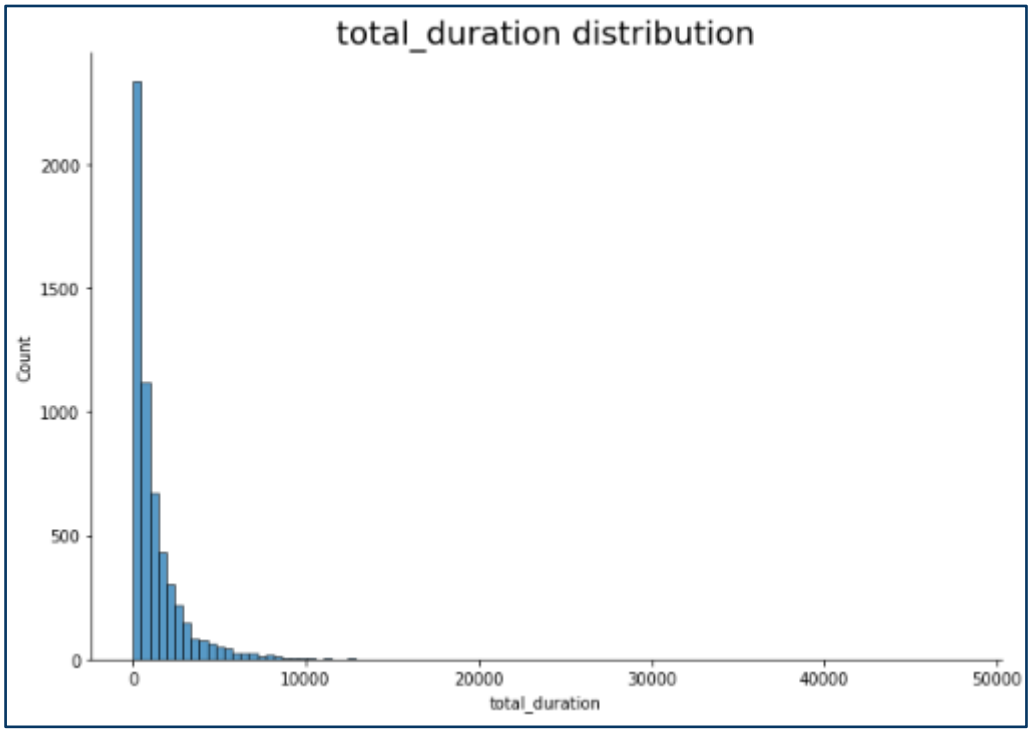


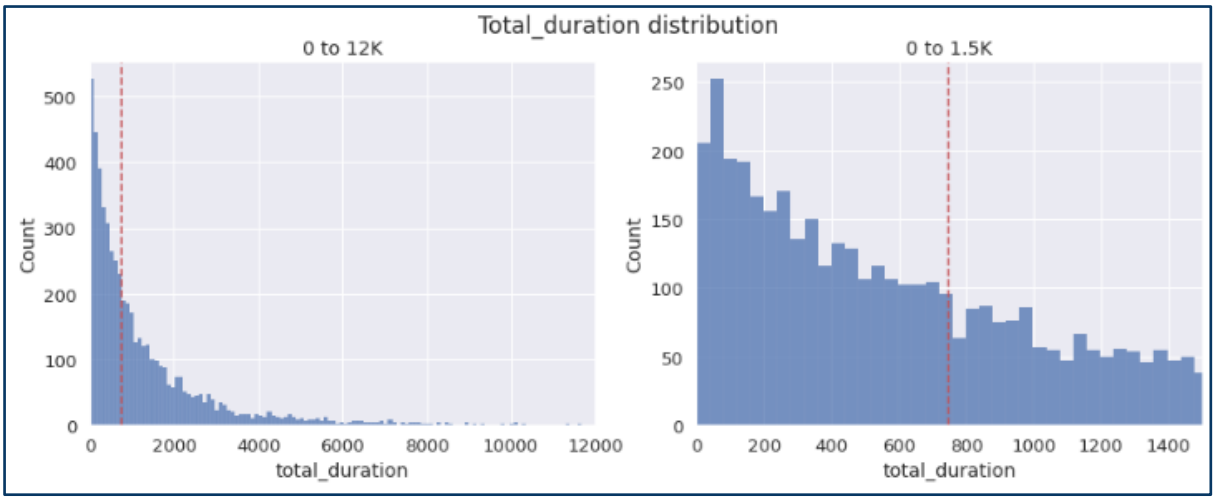

Train data
True label
Predicted label
0
1
6767
89
360
846
6000
5000
4000
3000
2000
1000
Validation data
True label
Predicted label
0
1
1679
55
105
167
1600
1400
1200
1000
800
600
400
200

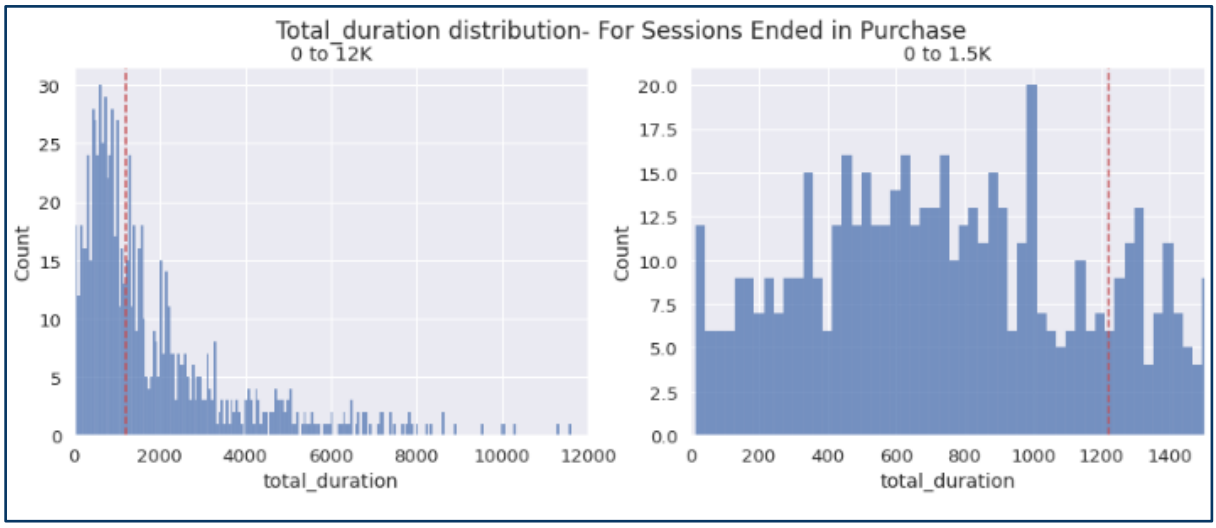

ROC
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 1 (AUC = 0.93)
ROC fold 2 (AUC = 0.94)
ROC fold 3 (AUC = 0.93)
ROC fold 4 (AUC = 0.93)
ROC fold 5 (AUC = 0.94)
ROC fold 6 (AUC = 0.93)
ROC fold 7 (AUC = 0.92)
ROC fold 8 (AUC = 0.92)
ROC fold 9 (AUC = 0.94)
ROC fold 10 (AUC = 0.93)
Mean ROC (AUC = 0.93)

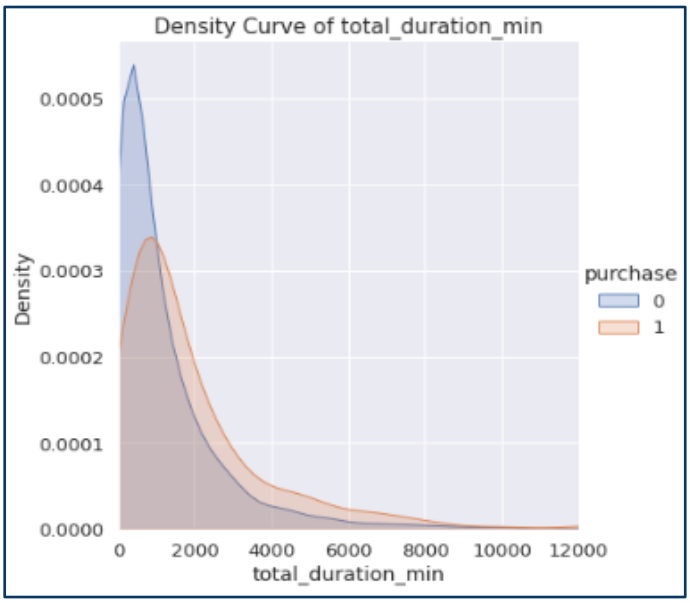
<u>נספח ג':</u> שאר הויזואליזציות

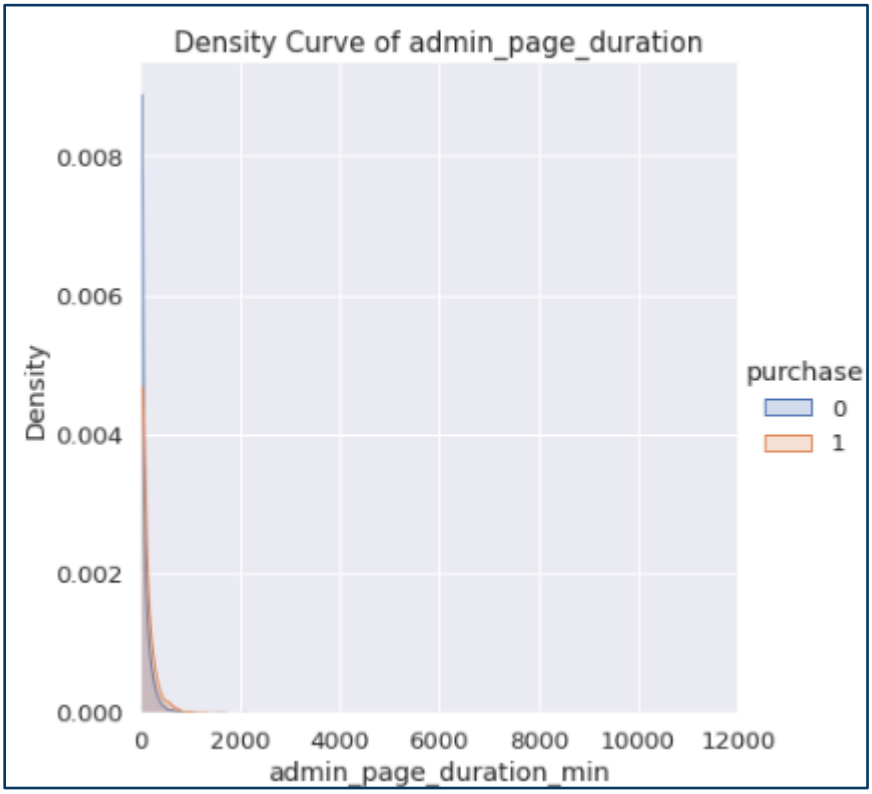


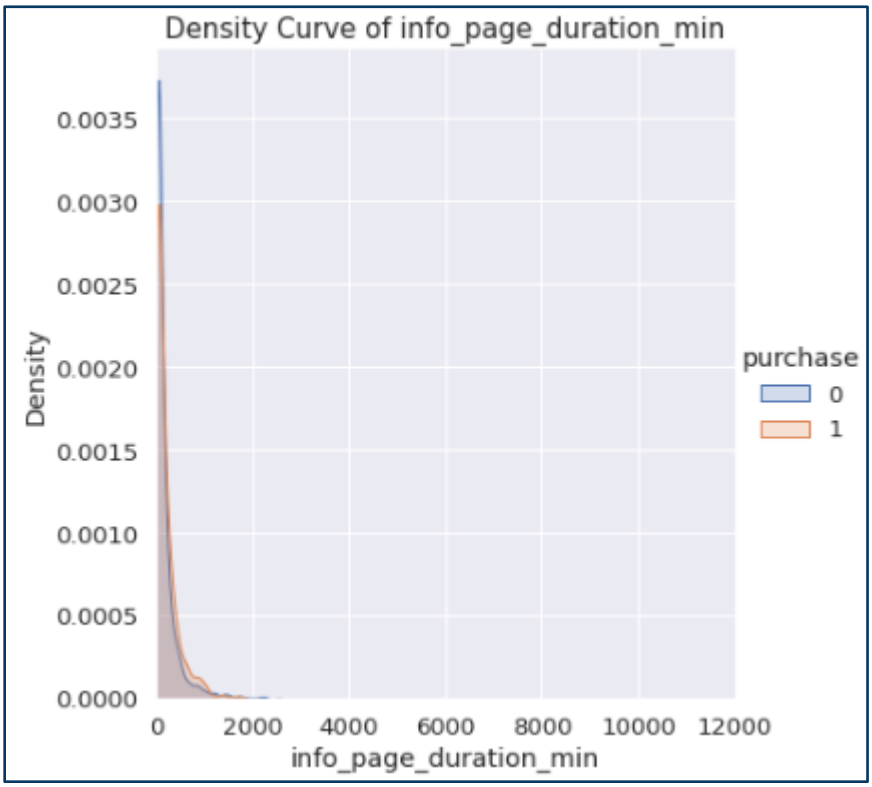


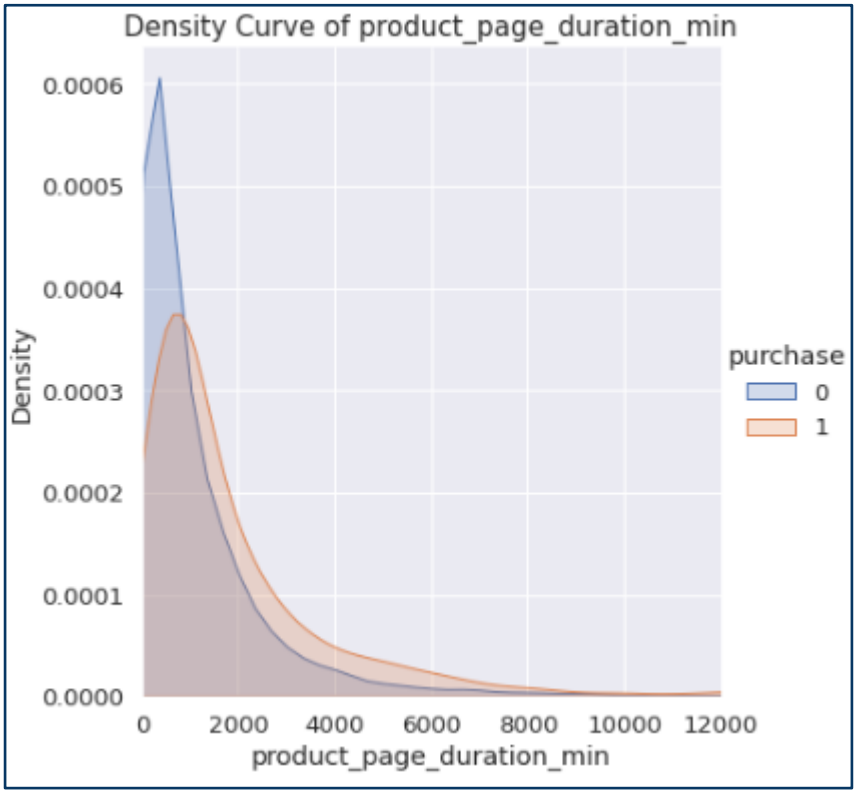

Total_duration distribution
0 to 12K
0 to 1.5K
Count
total_duration

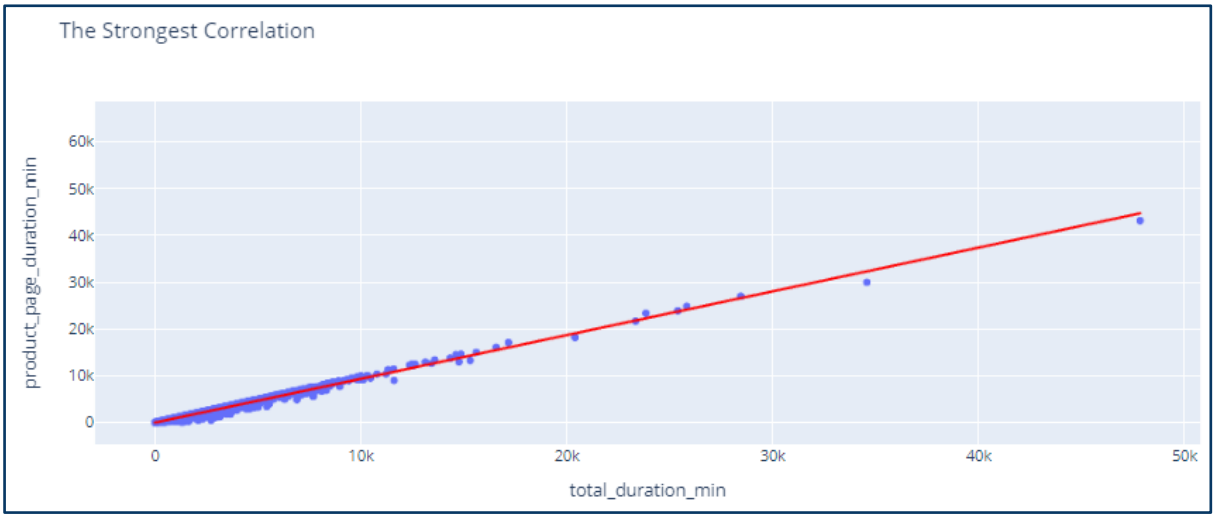

Total_duration distribution- For Sessions Ended in Purchase
0 to 12K
0 to 1.5K
Count
total_duration

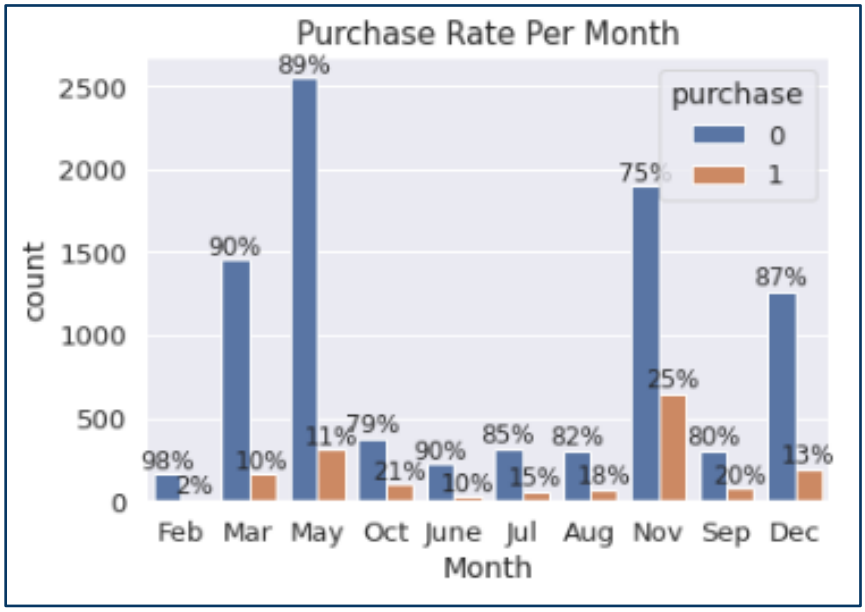

Density Curve of total_duration_min
Density
total_duration_min
purchase
0
1

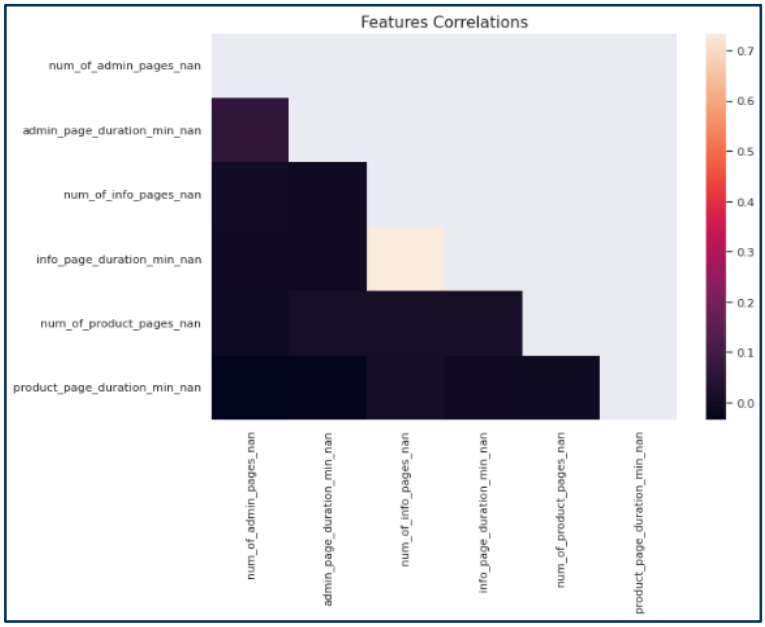

Density Curve of admin_page_duration
Density
0.008
0.006
0.004
0.002
0.000
0
2000
4000
6000
8000
10000
12000
admin_page_duration_min
purchase
0
1

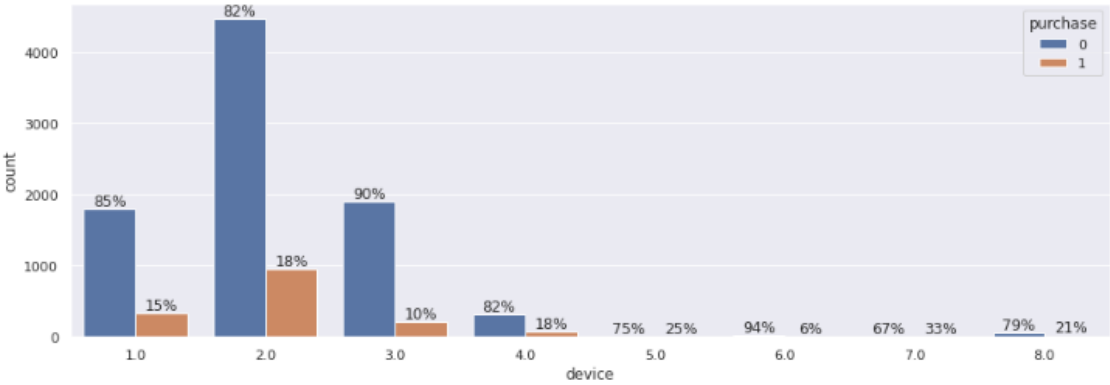

Density Curve of info_page_duration_min
Density
0.0035
0.0030
0.0025
0.0020
0.0015
0.0010
0.0005
0.0000
0
2000
4000
6000
8000
10000
12000
info_page_duration_min
purchase
0
1

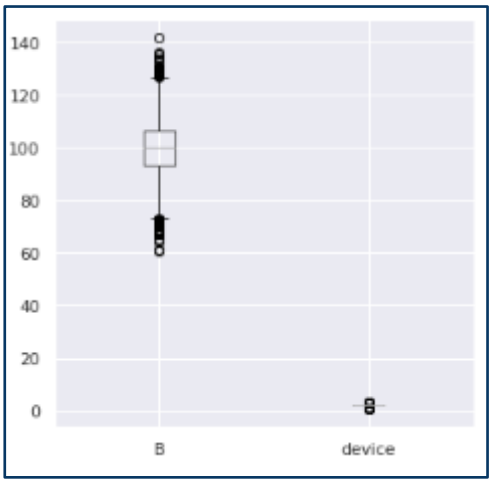

Density Curve of product_page_duration_min
Density
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
2000
4000
6000
8000
10000
12000
product_page_duration_min
purchase
0
1

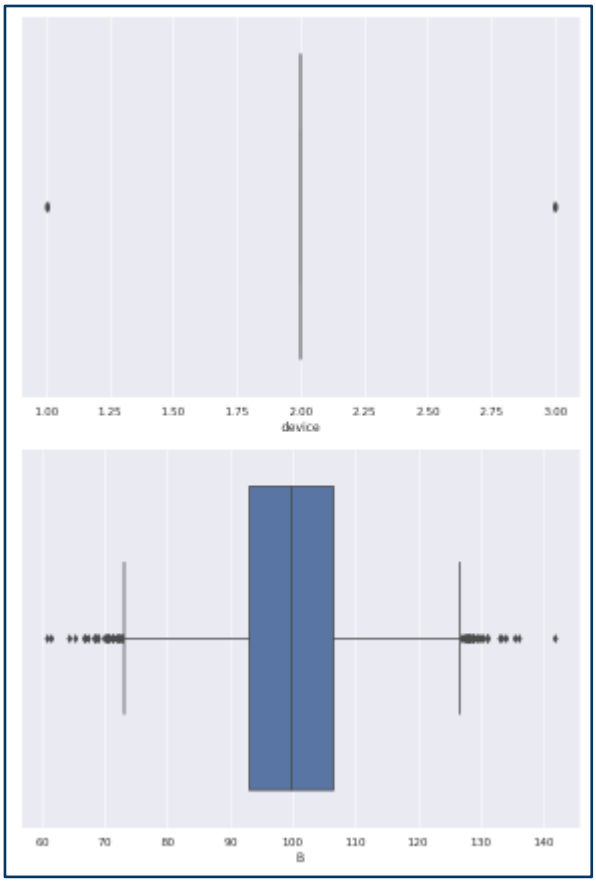

The Strongest Correlation
product_page_duration_min
60k
50k
40k
30k
20k
10k
0
0
10k
20k
30k
40k
50k
total_duration_min

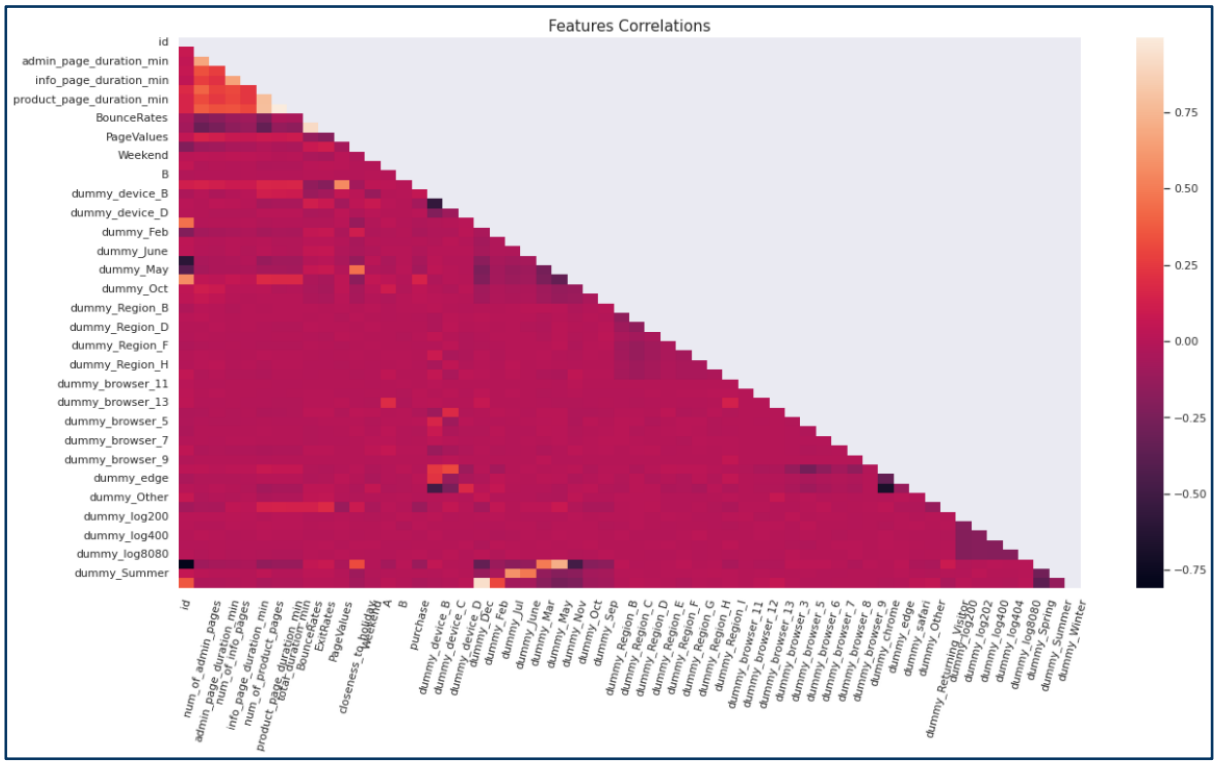

Purchase Rate Per Month
purchase
0
1
count
2500
2000
1500
1000
500
0
98%
2%
90%
10%
89%
11%
79%
21%
90%
10%
85%
15%
82%
18%
75%
25%
80%
20%
87%
13%
Feb
Mar
May
Oct
June
Jul
Aug
Nov
Sep
Dec
Month


Features Correlations
num_of_admin_pages_nan
admin_page_duration_min_nan
num_of_info_pages_nan
info_page_duration_min_nan
num_of_product_pages_nan
product_page_duration_min_nan
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0


purchase
0
1
count
4000
3000
2000
1000
0
85%
15%
82%
18%
90%
10%
82%
18%
75%
25%
94%
6%
67%
33%
79%
21%
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
device


140
120
100
80
60
40
20
0
B
device


1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
device
60
70
80
90
100
110
120
130
140
B


Features Correlations
id
admin_page_duration_min
info_page_duration_min
product_page_duration_min
BounceRates
PageValues
Weekend
B
dummy_device_B
dummy_device_D
dummy_Feb
dummy_June
dummy_May
dummy_Oct
dummy_Region_B
dummy_Region_D
dummy_Region_F
dummy_Region_H
dummy_browser_11
dummy_browser_13
dummy_browser_5
dummy_browser_7
dummy_browser_9
dummy_edge
dummy_Other
dummy_log200
dummy_log400
dummy_log8080
dummy_Summer
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75