Machine Learning Poject

Classification Model -

making a purchase while browsing an online shopping site (E-Commerce)

תקציר מנהלים

בהינתן מידע על משתמשים באתר קניות באינטרנט (E-Commerce), נתבקשנו לבנות מודל החוזה מה הסיכוי של משתמש מסוים, לבצע רכישה בזמן הגלישה באתר וזאת על סמך 10,479 תצפיות שלכל אחת מהן צוין האם בוצעה רכישה או לא. תחילה הצגנו את הנתונים כפי שהם, חקרנו ועיבדנו אותם בכדי שנוכל להשתמש בהם לבחינת מודלים שונים. לאחר מכן, עבור כל אחד מארבעת המודלים שבחרנו, מצאנו את הפרמטרים המיטביים, אימנו כל מודל ומצאנו את המיטבי על ידי הערכה באמצעות ROC curve ו-Confusion matrix. לבסוף, המודל הטוב ביותר בו מצאנו לנכון להשתמש, היה Random Forest בעזרתו הפקנו תחזיות לסט המבחן.

חלק ראשון – אקספלורציה

לפני ביצוע כל פעולה, נבחן את השטח ונחקור את הפיצ'רים כדי להסיק מסקנות ולתכנן את צעדינו הבאים. לשם כך נתמקד בכמה נושאים שנחקור בהמשך: נתונים חריגים, ערכים חסרים, נרמול, ממדיות, התמודדות עם משתנים קטגוריאליים ומניפולציה מתמטית על פיצ'רים קיימים.

לאחר הצצה ראשונית, גילנו מספר פרטים חשובים:

1. סוגי הפיצ'רים: מסוג מספרי ואובייקט. הבנו כי סוג הפיצ'ר לא בהכרח מייצג אותו בצורה מהימנה על כך ניתן את הדעת בהמשך. בנוסף נציין כי ישנם כמה פיצ'רים אנונימיים עליהם אין מידע מקדים-"A", "B", "C" ו-"D".
ערכי עמודה "C" הינם בצורת log, יתכן כי זוהי תצורת יומן- תיעוד אוטומטי המופק וחתום בזמן של אירועים הרלוונטיים למערכת מסוימת. עמודה "A" בעלת קידומת בצורת _c ולאחר מכן מספר ועל כן תסווג כקטגוריאלית.
2. הפיצ'רים- 'product_page_duration', 'info_page_duration' הינם נומריים אך בצירוף עם יחידות זמן אלו גורמים לפיצ'ר להיות מסוג אובייקט.
3. כמה מן המשתנים הנומריים- 'Region', 'device', הם קטגוריאליים כאשר כל ערך מספרי מייצג קטגוריה. דבר זה עלול לגרום להטיה במודל שכן יש "משקל" לקטגוריה ללא בסיס רציונלי, למה קטגוריה X צריכה לקבל את הערך 5 וקטגוריה Y את הערך 3? נרצה להתייחס בהמשך בטיפול שלנו במשתנים קטגוריאליים.
בנוסף, שמנו לב שלפיצ'רים "A", 'internet_browser' ו "C" נראה כי יש קטגוריות רבות ועל כן הם חדשים.
4. שמנו לב לערכים חסרים במספר עמודות.

בדקנו כמה קטגוריות יש לכל פיצ'ר חשוד. פיצ'ר 'C' מכיל 84 קטגוריות, פיצ'ר 'A' מכיל 94 קטגוריות ו- 'internet_browser' מכיל 126 קטגוריות. כאשר מירב הקטגוריות מופיעות פעם אחת. שמנו- לב, כי לגבי עמודה 'internet_browser', כל דגימה מתחילה בשם הדפדפן ולאחריו גרסת הדפדפן. אנו מניחים כי הגרסה אינה משמעותית ביחס להשפעה של קטגוריות רבות לטיפול.

כבר בשלב זה, **נבצע התאמות שיקלו עלינו בהבנת הפיצ'רים ויקלו עלינו בביצוע במודלים העתידיים.**
ראינו כי הגיוני יותר לבצע עיבוד-מקדים לחלק מהנתונים בשלב זה מפני שמיקום הפעולה היה כך הגיוני יותר. הפחתנו את מספר הקטגוריות בפיצ'ר internet_browser לכארבע קטגוריות על פי סוג הbrowser ללא התייחסות לגרסה. בנוסף, נהפוך את המשתנים C ו-A לייצוג מספרי תוך התייחסות אליהם כקטגוריאליים. וכמובן הוצאת הרעש מהפיצ'רים 'product_page_duration', 'info_page_duration' והפיכתם לנומריים.

<u>התפלגות</u>
נרצה לבצע **ויזואליזציות** בכדי להבין טוב יותר את הנתונים. ראשית נבחן **התפלגות כל פיצ'ר** (נספח 1.1, 1.2). ראינו כי פרט לעמודה "B" הנראית כמתפלגת נורמלית שאר הפיצ'רים אינם מתפלגים נורמלית. ישנם מספרי פיצ'רים בעלי 2 'פיקחים' שאולי נוכל לחלק אותם לאוכלוסיות שונות ולמצוא התפלגות נורמלית לכל אוכלוסייה.
ראינו זאת בפיצ'רים - "ExitRates","BounceRates", "C", "D". ננסה לבחון את ההשערה באמצעות המשתנים הקטגוריאליים הנתונים לנו, ובפרט אלו בעלי 2 קטגוריות. אך קודם לכן, החלטנו להתבונן במשתנים הקטגוריים שלנו (נספח

1.3) לפני כל פעולה. לאחר מספר ויזואליזציות בהן פילגנו את הפיצ'רים הבעיתים שלנו בעזרת המשתנים הקטגוריאליים (נספח 1.4) לא מצאנו קומבינציה שתאשש את ההשערה ולכן לא נוכל להניח ולהתייחס כעת לעמודות אלה כמתפלגות נורמלית. בבדיקה, גילינו "על הדרך" כי בפיצ'רים 'Region', 'Month', 'user_type' ו-'device' יש קטגוריות בעלות נתח קטן מסך הנתונים (מתחת ל-5%) דבר היכול לגרום לבעיות כמו over-fitting ועוד.

<u>ערכים חריגים</u>

נרצה לבחון את כמות **הערכים החריגים (outliers)** בכל פיצ'ר באמצעות ויזואליזציה "BOXPLOT"(נספח 1.5, 1.6). נציין כי ויזואליזציה זו מתבססת על כך שהפיצ'רים מתפלגים נורמלית, דבר שלא קורה ברוב הפיצ'רים שלנו, אך זה יכול לתת לנו תחושה כלשהי לגבי הערכים החריגים של הפיצ'רים הנומריים. לכן, נוכל להגדיר על ידי הוויזואליזציה רק את הערכים החריגים של פיצ'ר 'B' בוודאות. מהוויזואליזציה הבחנו כי פיצ'רים הקשורים ל-Page ו-duration בעלי ממוצע אפס לכן יש הרבה ערכים "חריגים". כעת נעשה את אותו התהליך אך נבדוק איך נראה ה-"BOXPLOT" של כל פיצ'ר לפי התווית שלו(נספח 1.7). ראינו כי לרוב הפיצ'רים הקשורים ל-Page ו-duration , הממוצע של 'רכש' גבוה מזה של 'לא רכש'. לגבי הפיצ'רים 'ExitRates' ו- 'BounceRates', אנו יכולים לראות של-'לא רכש' יש ממוצע גבוה יותר מזה של 'רכש'. 'ExitRates' ו- 'BounceRates' גבוהים יותר פירושם סיכוי גבוה יותר לצאת מדף האתר ולסיים את ה-session ללא רכישה, לכן זה גם עושה שכל. נתייחס לכך בהמשך בשלב של חשיבות כל פיצ'ר.

<u>בדיקת קורלציות</u>

בשלב זה ננסה **לבדוק קורלציות בין פיצ'רים** בעזרת ויזואליזציה ולא בהכרח בעזרת מטריצת קורלציות. התחלנו בלבחון את פיצ'ר 'Weekend'. רצינו לראות אם ישנה השפעה לפיצ'ר על התווית. קו המחשבה שלנו הוא שאנשים פנויים יותר בסופי השבוע ולכן הם ירכשו יותר בסופי שבוע. הצגנו זאת בתרשים עוגה(נספח 1.8), וציפינו שפחות מ5/7 מסך הרכישות יהיו במהלך השבוע ויותר מ2/7 בסוף השבוע. בסוף גילנו כי ההשערה הייתה שגויה ואילו בפועל קורה ההפך.
כעת לקחנו מספר קבוצות של 3 פיצ'רים בעלי פוטנציה לקשר פרבולי וניסנו למצוא אותו בעזרת גרף תלת מימדי (נספח 1.9), אך לצערנו לא מצאנו מסקנות חדשות שלא ידענו ורלוונטיות בהן נוכל להשתמש בהמשך (ראו פירוט המסקנות במחברת).

<u>פיצ'רים קטגוריים</u>

בכדי להטיב עם המודל, אחד הדברים הכי חשובים ובאותה העת בעייתיים, איתם אנו מתמודדים הם פיצ'רים קטגוריים(נספח 1.3). ראשית, נסתכל על הפיצ'רים הלא אנונימיים שלנו ונבחן מהו **שיעור הרכישה באחוזים לכל קטגוריה**(נספח 2.0).
הפיצ'רים שניקח הם - Month, device, internet_browser ו- Region. הבחנו כי לקטגוריות שונות שיעורי רכישה שונים ובחלקם יש הבדלים משמעותיים. נוכל להשתמש בעובדה זו בהמשך כאשר נרצה לטפל במשתנים הקטגורים.

<u>ערכים חסרים</u>

כעת נרצה לבחון את התפלגות הערכים החסרים, ראשית נבדוק כמה ערכים חסרים יש במספרים ובאחוזים(נספח 2.1).
ראינו כי לעמודה 'D' יש מספר חריג של ערכים חסרים(כ-99%) וכן לעמודה total_duration כמעט 50% מהערכים החסרים.
החלטנו כי לעמודות בעלות מעל 4% ערכים חסרים נשים דגש נוסף במילויים. בנוסף, בכדי לחשוב איך לטפל בחסרים, ניצור מטריצות קורלציות לכל הדאטה ואחת רק לפי עמודת התווית(נספח 2.2). **ראינו כי קיים קשר חזק בין הפיצ'רים הבאים:**

1. 'ExitRates' – 'BounceRates' - קורלציה של 0.91
2. 'product_page_duration' – 'num_of_product_page' – קורלציה של 0.86
3. 'proudct_page_duration – 'total_duration' – קורלציה של 0.99

חשבנו על מספר דרכים למלא את הערכים החסרים- דרך קורלציה מתמטית בין משתנים בעלי קורלציה גבוהה, דרך קשר לוגי ביניהם ולסיום מילוי לפי מדד סטטיסטי כלשהו. נרחיב על כך בחלק העיבוד המקדים.

<u>חלק שני - העיבוד המקדים</u>

הנחנו כי רק פיצ'ר 'B' מתפלג נורמלי ולכן החלטנו לטפל **בערכים החריגים** שלו בלבד. לגבי שאר העמודות, החלטנו לא לעסוק בערכים אלה משתי סיבות. ראשית, משום שהם אינם מתפלגים נורמלית, קשה להבחין בערכים חריגים. שנית,

יצאנו מנקודת הנחה כי בחלק מהעמודות, שימור ערכי הקצה יתכן ויטיב עימנו, ויתכן כי ישנם מודלים שידעו כיצד להשתמש בערכים החריגים, על כך נפרט בהמשך. בנוסף, נציין כי יכולנו להמיר את ההתפלגות של עמודות הpage וה-duration לנורמליות, אבל מהסיבות לעיל החלטנו לא לעשות זאת. השתמשנו בשיטת ה'IQR' למציאת ערכי הקצה(הקצוות מהם IQR מסווג כערך חריג) והשמה לערכים החריגים את ערך הקצה, מלמעלה ומלמטה בהתאמה.

לאחר מכן עברנו לטיפול בערכים החסרים בשלוש דרכים-

- **דרך קורלציה מתמטית בין משתנים,** תחילה השלמנו את הערכים החסרים במשתנים הקורלטיביים שמצאנו לפי מטריצת הקורלציות. בכך שהרצנו מודל רגרסיה לינארית בין כל זוג פיצ'רים(נספח 2.4) ומצאנו את משוואת הקו הישר שלהם. בעזרת המשוואה יכלנו למלא כל ערך חסר בעזרת הערך התואם של הזוג שלו. בנוסף מיותר לציין שנתנו קדימות לקורלציה גדולה יותר, קודם מילאנו לפי זוג מספר 3 ורק אז לפי זוג מספר 2 (בזוגות לעיל) את הערכים החסרים. הערכים להם שתי העמודות היו בעלות ערכים חסרי, מילאנו את ערכים אלו בעזרת הדרך השלישית (ערך חציוני) אך מספר ערכים אלה היה מועט.
- **בעזרת קשר לוגי בין המשתנים,** בהנחה כי קיים קשר חיובי בין משך הזמן בעמוד לכמות העמודים ב-session, נמלא את הערכים חסרים בעמודות אלו. תחילה ניקח כל זוג עמודות הנוגעות לאותו העמוד. לכל סוג עמוד, נחשב את משך הזמן הממוצע לכל מספר עמודים וכן במקרה ההפוך, ניקח את מספר העמודים שהממוצע שלהם הוא הקרוב ביותר. למשל אם ידוע שמשך הזמן הוא 32, נמלא את ערך מספר העמודים החסר עם הערך הממוצע שקרוב ביותר אליו בצורה אבסולוטית למשל 5. במקרה בו יש שני ערכים חסרים לאותו זוג נמלא את החציון (זה קרה במקרים בודדים לכן לא משפיע).
- לאחר השלמת כמעט כל הערכים החסרים של הפיצ'רים הקשורים במשך ובמספר העמודים, נותר להשלים את 'total duration'– עליו הנחנו כי הוא סכימה של הפיצ'רים הנוגעים במשך הזמן בעמוד.
- **דרך ערך חציוני/שכיח,** כעת את שאר הערכים החסרים למעט פיצ'ר D שיש לו 99% ערכים חסרים נשלים בצורה הבאה: משתנים מספריים נשלים בעזרת חציון ואילו קטגוריאליים נשלים בעזרת הקטגוריה השכיחה ביותר. בכך נדאג לא לשנות את החציון של העמודות ואופיין ע"י השמת ערך אחר.

כעת בדבר המשתנים הקטגוריים, לא ייתכן שניצור משתני דמה למשתנים עם ריבוי קטגוריות לכן צריכים להיות יצירתיים. החלטנו לנקוט בגישה ייחודית של **נרמול**. נרמלנו בשיטה שהזכרנו לעיל, ניתן לקטגוריה עם שיעור הרכישות הגבוה ביותר את הערך 1 וכל קטגוריה אחרת תקבל את שיעור הרכישות שלה חלקי שיעור הרכישות המקסימלי. עשינו זאת על הפיצ'רים הבאים: 'month' , 'device' ו- 'region'. בחרנו אותם מתוך המשתנים הקטגוריים משיקולי אי אנונימיות כלומר אנחנו יודעים מה הם אומרים ומשיקול נוסף של ריבוי קטגוריות, בפיצ'רים עם מעט קטגוריות נטפל בדרך אחרת.
לאחר שעשינו את הטרנספורמציה גילנו שנוצר **עיוות**, כלומר בקטגוריות קטנות יכולים להיות שיעורי רכישה קיצוניים לשני הצדדים, למשל הבחנו בקטגוריה אחת ב'device' עם 100% שיעורי רכישה מכיוון שהייתה שם רק דגימה אחת. לכן בדקנו את התפלגות כל הקטגוריות במשתנים הקטגוריים והחלטנו **לאחד קטגוריות לפי ערך סף,** כלומר כל הקטגוריות שהן מתחת ל5% יאוחדו לאחת וכך נמנע מהטיה. חזרנו שוב על התהליך וקיבלנו תוצאות מספקות. כעת נותר לנו לטפל בשני המשתנים הקטגוריים– 'user_type' ו-'internet_browser'. מכיוון שלפיצ'רים אלה מעט קטגוריות - 4 ו3 בהתאמה, נטפל בהם **בשיטת משתני הדמה.** כמו כן הבחנו קודם שלקטגוריית 'other' בפיצ'ר 'user_type' מספר קטן מאוד של ערכים. בשביל למנוע בעיות בפגיעה בפונקציות בחלק הוולידציה, למשל שלקטגוריה זו יהיה אף נציג החלטנו לתת לה את ערך הקטגוריה השכיחה יותר בפיצ'ר- 'Returning_visitor'.
לאחר מכן יצרנו משתני דמה וכך הוספנו עוד 3 מימדים לסט הנתונים שלנו.

אז אחרי שטיפלנו בנתונים שאלנו את עצמנו- **האם הממדיות של הנתונים גדולה מדי?**
נשארנו עם כ-25 פיצ'רים עבור 10,000 תצפיות לכן לא נראה כי ממדיות הנתונים גדולה מדי אך יתכן כי עם פחות פיצ'רים נוכל לקבל תוצאות דומות. בשלב זה, השתמשנו בשתי שיטות ובחרנו את האחת שתעניק לנו ביצועים טובים יותר -

1. <u>השיטה הראשונה: PCA-</u>

נרצה להקטין את המימדים כדי **לכסות 99% מהשונות** בנתונים, ראינו כי מספר הפיצ'רים המשמרים לפחות 99% מהשונות הוא 2. מימד נמוך עלול לגרום ל- under-fitting ולגרום לביצועים גרועים בסיווג ועל כן החלטנו **לא להשתמש בשיטה** זו

להפחתת המימדים. נציין כי כבר טיפלנו ברוב הפיצ'רים הקטגוריים, שחלקם הפכנו למשתני דמה נזכור זאת מכיוון שהערך המוסף הכולל של כל תכונה אינו נתפס בגלל הפיצול למשתני דמה או הטרנספורמציות.

2. <u>השיטה השנייה: בחירת פיצ'רים ידנית-</u> בחרנו להוריד את העמודות הבאות: 'D', 'C', 'A' ו- 'page_duration'.

- *עמודה 'D'* - אנו מאמינים ריבוי הערכים החסרים שלה ולמרות שיש לה מול מתאם גבוה עם עמודת התווית, מהווה בעיה מהותית ועל כן נוריד אותו.
- *עמודה 'A' ועמודה 'C'*- עמודות אנונימיות בעלות ריבוי קטגוריות כאשר רוב הקטגוריות של העמודות מופיעות רק פעמים בודדות. לא נוכל להפוך אותם למשתני דמה בגלל ריבוי הקטגוריות ובנוסף לא נוכל לעשות עליהם את הטרנספורמציה שעשינו קודם משיקולי אנונימיות, על כן נוריד אותן גם.
- *עמודת 'product_page_duration'*- לעמודות 'product_page_duration' ו-'total_duration' קורלציה חזקה , גם לאחר שהשתמשנו בעובדה זו כדי למלא את הערכים החברים של כל עמודה (נספח 2.5,2.6). בכדי למנוע מולטיקוליניאריות, נרצה להסיר אחת מן העמודות. מצד אחד, ל-product_page_duration יש מתאם חלש יותר עם עמודת הרכישה בפער קטן. אנו שואפים לקבל מתאם גבוה יותר עם עמודת התווית מכיוון שהמטרה שלנו היא לסווג לפיה. לעמודה זו יש גם ערכים גדולים יותר מעמודות ה-duration האחרות, מה שמעניק לה אולי משקל רב יותר בתהליך הרכישה באופן הגיוני. מצד שני, ל-'total_duration' היה מספר רב של ערכים חסרים, אלו שמילאנו באמצעות עמודת משך הזמן(duration) האחרות כולל ה-product_page_duration. לכן, הוא כולל כעת את כל עמודות ה-duration ובמיוחד את product_page_duration. דבר זה נותן ל-'total_duration' ערך מוסף מקיף יותר. לכן, אנחנו נוטים ללכת עם שמירת עמודת ה-total_duration והסרת עמודת ה-product_page_duration.
- *בנוסף נציין כי יכלנו להסיר מאותם שיקולים את אחת מעמודות* 'BounceRates' – 'ExitRates'. אך בגלל שיש להן חשיבות מבחינת המשמעות שלהן, החלטנו להשאיר את שתיהן. אנחנו סבורים כי מדדים אלה משמעותיים ובשימוש רחב בגoogle analytics .

<u>חלק שלישי- הרצת המודלים</u>

לבחירת הפרמטרים האידיאליים השתמשנו ב-grid_search_cv, תוך בחינת מירב האופציות למיקסום תוצאות המודל, והשארת הערכים הדיפולטיביים. להלן הדגמים(נספח 2.6) שבהם השתמשנו ומדוע השתמשנו בהם:

- <u>K-Nearest-Neighbors- AUC= 0.838:</u> עבור מודל זה, בחרנו לעשות סטנדרטיזציה(נרמול) של הנתונים מראש כדי למנוע מקנה המידה של כל פיצ'ר להשפיע על המודל באופן לא פרופורציונלי. במודל זה בחרנו את הפרמטרים האופטימליים על ידי ה grid search כאשר שאר המודלים השתמשנו באלה הדיפולטיבים. הפרמטרים שנבחרו הם **'algorithm'** = 'auto', **'n_neighbors'** = 20, **'weights** = 'distance'. יצא לנו במודל המודל האימון AUC = 0.838 ובמודל ולידציה AUC = 1, זהו פער גדול שיכול להעיד על overfitting, כמובן שלא ניתן יותר מידי להסתמך על כך כי זה תלוי דאטה.
- <u>Logistic Regression- AUC=0.886:</u> עבור מודל זה, בחרנו לעשות סטנדרטיזציה של הנתונים מכיוון שהשתמשנו בעונש l2, המושפע מקנה המידה של התכונות. במודל זה בחרנו את הפרמטרים האופטימליים על ידי ה grid search כאשר שאר המודלים השתמשנו באלה הדיפולטיבים. הפרמטרים שנבחרו הם: **'penalty'** = l2, **'n_jobs'** = -1, **'C'** = 21. יצא לנו במודל המודל האימון ROC = 0.898 ובמודל הוולידציה 0.886, זהו פער מזערי שנותן לנו תוצאות כמעט זהות בין מודל האימון לוולידציה, הדבר שאנו שואפים לו.
- <u>Multi-Layer Perceptron (ANN) - AUC=0.908:</u> עבור מודל זה, בחרנו לנרמל את הנתונים כדי לסייע בהתכנסות מהירה יותר וכדי להפחית את ההסתברות להגיע למינימום מקומי גרוע. אכן מצאנו שנורמליזציה של הנתונים לפני הרצת MLP הפכה את הביצועים שלו לעקביים משמעותית. במודל זה בחרנו את הפרמטרים האופטימליים על ידי ה grid search כאשר שאר המודלים השתמשנו באלה הדיפולטיבים. הפרמטרים שנבחרו הם: **'hidden_layer_sizes'** =(40,) **'epsilon'** = 0.0001. יצא לנו במודל האימון ROC = 0.949 ובמודל הוולידציה 0.917, זהו פער קטן שנותן לנו תוצאות מאוד דומות בין מודל האימון לוולידציה, הדבר אנו שואפים לו.
- <u>Random Forest- AUC= 0.92:</u> ידוע בתוצאות מעולות ונמצא בשימוש תדיר בתעשייה. כאן, בחרנו לא לנרמל את הנתונים, מכיוון שמודל זה אינו מושפע מקנה המידה. מודל זה ביצע את הטוב ביותר מבין המודלים ובעל ביצועים טובים יותר על ה-train(0.95) ועל ה-validation ועל כן נרצה להמשיך איתו להמשך הפרויקט. הפרמטרים שנבחרו הם: **'max_depth'**= 7, **'max_features'**='log2', **'min_samples_split '**= 4. אנו מאמינים שמכיוון ש-Random Forest מסוגל לסיווג לא ליניארי, הוא יכול להתמודד ביעילות עם חריגים על-ידי הפרדתם מהנתונים, ויכול להתמודד עם מערך הנתונים הגדול שלנו על-ידי יצירת יותר עצים ועצים עמוקים יותר.

לאחר הרצת המודלים ומציאת ציוני ה-AUC, כעת נוכל לנסות להסביר מי **עמודות המפתח הכי המשפיעות** על המודל. כל אלה עושות שכל מבחינה רציונלית, כפי שנפרט.

אנו יכולים לראות (נספח 2.7) כי **4 הפיצ'רים החשובים ביותר בקירוב הם:**

1. 'PageValues' (55%) - הגיוני ש-'PageValues' הוא העמודה החשובה ביותר בגלל המשמעות שלה. אם לדף כלשהו יש ערך דף גבוה, נניח כי יש יותר סיכוי שה session יסתיים בקנייה. אם ניזכר, ראינו כי לעמודה זו יש קורלציה יחסית גבוהה עם עמודת התווית, וכן כי יש הבדל ניכר בין הממוצעים לפי התווית של העמודה (בחלק של הערכים החריגים), דבר המחזק את חשיבותו למודל. התוצאה הפתיעה אותנו, ציפינו להשפעה גדולה אך לא להשפעה כה גבוהה של כמעט 50%.
2. 'ExitRates' (9%)- נניח כי בעמוד בעל ערך גבוה מאוד, בממוצע רוב המשתמשים יצאו מהאתר ללא רכישה או בעמודים כמו עמוד הקנייה עם רכישה. בנוסף הגיוני שהוא יהיה השני, משום שהוא ועמודת 'PageValues' בעלי קורלציה גבוהה.
3. total_duration (7%) – עמודה זו קשורה במשך הזמן הולל שהמשתמש מבלה באתר. יותר זמן באתר, יכול להעיד על רצינות המשתמש לגבי המוצר שיכולה להסתיים בקניה.
4. 'Month' (6%)- חודשים עם חגים ארוכים, חודש עם מכירות מיוחדות ועוד משפיעים על המכירות.

חלק רביעי - הערכת המודלים

בחלק זה, הערכנו את המודל באמצעות:

1. K-Fold Cross Validation- (נספח 2.9) בנינו פלט ROC על כל K-Fold עבור כל אחד מדגמי הריצה. נציין כי ביצענו את תהליך העיבוד המקדים על כל train ו-validation, דבר המשפר את האינדיקציה לתוצאות התחזיות על קובץ ה-test.

מהתבוננות בתרשים(נספח), אנו יכולים לראות שיש לנו טווח של 11% של דיוק בין הביצועים של הדגם הגבוה ביותר לנמוך ביותר(במודל שבחרנו). באופן כללי, תוצאות הדיוק שלנו די טובות, הנמוכה ביותר היא כמעט 87%, בעוד שהתוצאה הטובה ביותר שלנו היא כמעט 99% מהדיוק. ניתן לראות שתוצאות המודלים בעלות שונות גבוהה, יתכן שהדבר מעיד על התאמה של כל מודל לדאטה השונה, נדון בכך בהמשך.

2. confusion matrix- (נספח 2.8)המודל שלנו יכול להבחין בין התוויות בצורה טובה יחסית. היה לנו קצת ' _FALSE NEGATIVE '- כלומר המודל שלנו חזה שמשתמשים לא יקנו בסוף הsession כאשר הם בפועל קנו. לכן, המודל יכול להתאים לאתרים שמשקיעים במשאבים רבים כאשר הם מאמינים שיש סיכוי טוב לרכישה.

לאחר שראינו את ציוני ה-AUC בכל אחד מהמודלים, שאלנו את עצמנו **האם למודלים שלנו יש Over-fitting ?**

במהלך הפרויקט, דאגנו למזעור הסיכוי ל-overfitting ע"י צמצום הממדיות, הסרת קטגוריות קטנות, טיפול בחריגים ומילוי כל הערכים החסרים. בנוסף, שימוש בK-Fold ובממוצע תוצאותיו. בשלב הרצת המודלים (ללא ה-K-Fold) ההפרש ה-AUC בין ה-train ל-validation הוא בגדר הרגיל. ולכן הסבירות ל-overfitting יותר נמוכה. בחרנו בנוסף לבדוק זאת דרך אימון המודל שלנו על נתונים שונים בכל פעם(train ו-validation). בדרך זו, אנו יכולים לראות אם המודל מתאים יותר מדי את עצמו לנתונים הספציפיים. קיום מגוון רחב של ביצועים עשוי להעיד על overfitting. בכל קבוצה של K-Fold, אנו מבצעים את העיבוד המקדים בכל פעם מחדש. זה מאפשר לנו לראות אם המודל באמת מתאים יותר מדי ולהבין טוב יותר מה יהיה ממוצע ה-AUC של סט המבחן. לכן, הסבירות ל-overfitting עלולה להיות יותר גבוהה.

חלק חמישי – ביצוע תחזיות

לאחר שבחרנו במודל Random Forest, נרצה לבצע תחזיות על סט המבחן. נדגיש כי סט המבחן עבר עיבוד מקדים דומה לסט האימון ובהתאמה אליו. התחזיות בצורת הסתברות לכל תחזית אשר יצאנו לקובץ אקסל.

סיכום

בפרויקט זה נתבקשנו להציג מודל אשר חוזה בצורה מיטבית האם תבוצע או לא רכישה בזמן הגלישה באתר קניות באינטרנט. כפי שראינו בהורדת הממדים, בחרנו לנכון להשתמש רק בבחירת הפיצ'רים הידנית ולא בPCA משום זו מבוססת הגיון ומשאירה לנו את רוב המימדים. בנוסף, המודל הטוב ביותר שמצאנו לנכון היה Random Forest בעזרתו הפקנו תחזיות לסט המבחן. נשים לב כי בחיזוי עתידי, על בעל האתר לשים דגש במיוחד על הפיצ'רים 'PageValues', 'ExitRates', total_duration ו- 'Month' שכן להם משקל כבד בקבלת ההחלטה האם בסופו של דבר, המשתמש יבצע רכישה.

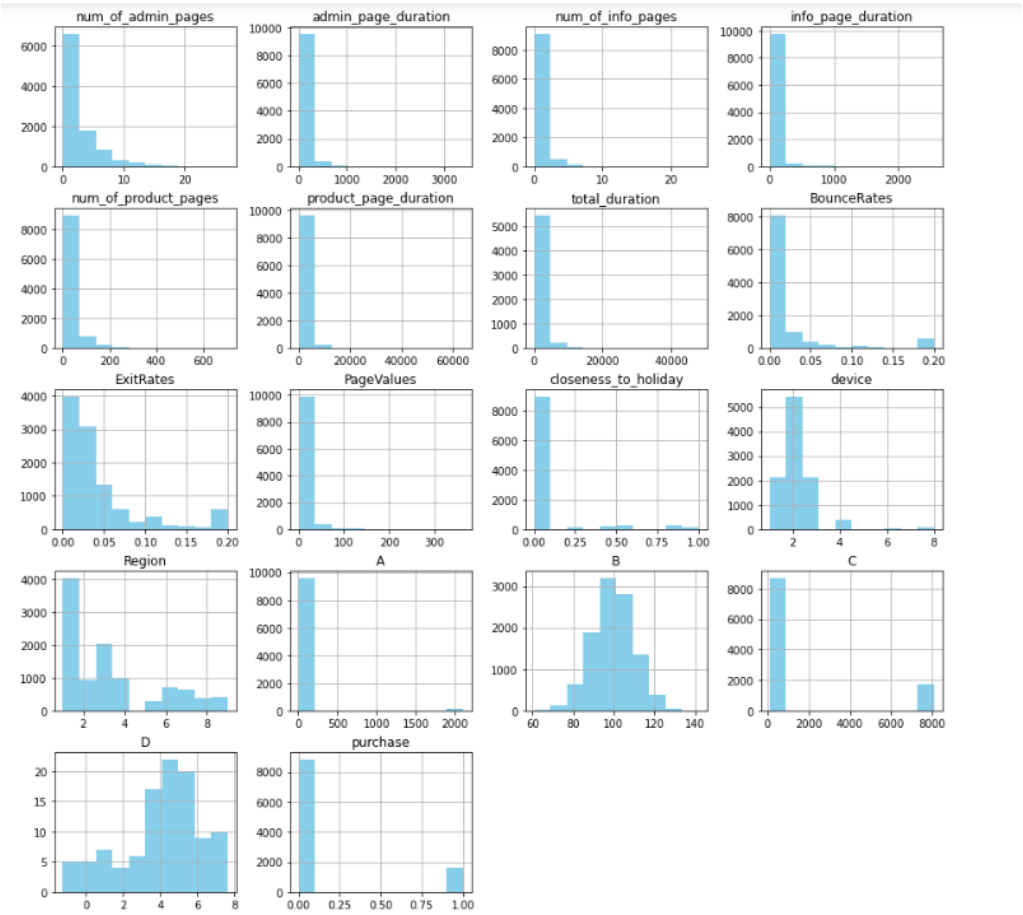
נספחים-

נספח : אחריות כל חבר צוות

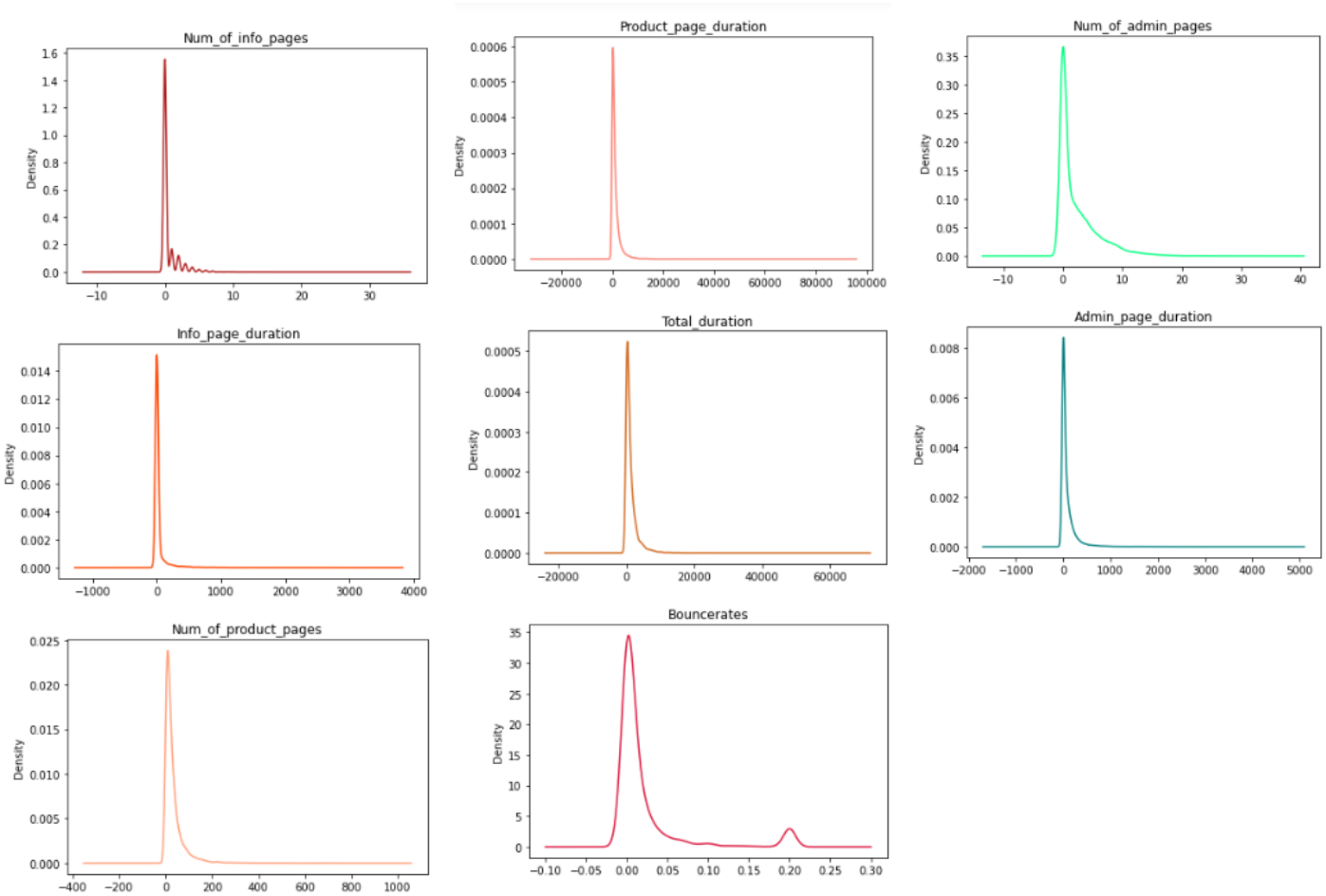
בנספח זה נתאר כיצד חולקה העבודה בין חברי הקבוצה: קארין אגם ונועם דותן.

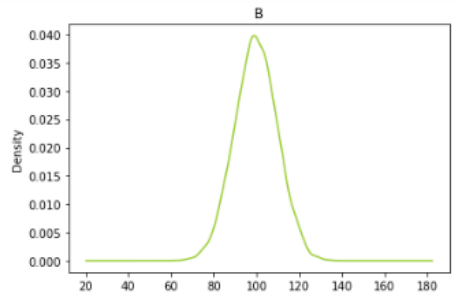
שם חבר הצוות	קארין אגם	נועם דותן
החלק אותו לקח חבר הצוות	תחילה עשינו שיחת זום והחלנו לסקור ביחד את הנתונים ומשמעותם של הפיצ'רים לפי החומר היבש שסופק לנו. לאחר מכן חילקנו את האחריות הראשונית באופן הבא:	
	אחראית על בניית פונקציות הפרי פרוססינג.	אחראי על בניית ויזואליזציות ראשוניות
	לאחר בניית מודל ראשוני נפגשנו ובנינו ביחד מודל "בזק"(ביצוע מודל כמה שיותר מהר) של רגרסיה לינארית שייתן לנו תחושה לגבי המודל. בהמשך הדרך לאחר ה"אב-טיפוס" הראשוני למודל, היינו נפגשים בין פעמים לשלוש בשבוע ועובדים ביחד על המודל ובין לבין כל אחד היה עושה השלמות משלו לבדו-	
חלוקה לפונקציות לפי נושאים-	אחראית על ויזואליזציות מקיפות, חלק מפונקציות השלמת חסרים, חלק ערכים חריגים ועוד	אחראי על חלק מפונקציות השלמת הערכים החסרים, הנרמול ועוד
	כשהגענו לשלב הסופי כשהיה מודל מוכן היינו עושים שיחות זום ועוברים ביחד על המודל ומוסיפים הסברים ופרשנויות למה שעשינו בצורה משותפת.	
	את הורדת המימדים עשינו יחד וכן את שלבים 3 ו-4. בכל שלב כל אחד ניסה לחקור את הפרמטרים באופן הכי טוב לבד ואז התכנסנו לדון בתוצאות.	
	שלב 5- פרדיקציה	pipeline
	לבסוף כל אחד הכין חצי מתקציר המנהלים ולבסוף סקרנו אותו ביחד.	
מבחינת הסברים-	עשתה את החלק של הורדת הממדיות, הסבר על הפונקציות והסיכום	היה אחראי על החלק של האקספלורציה, ויזואליזציות והפרי פרוססינג.
	כמובן שכל הדברים והחלטות נעשו בשיתוף פעולה מלא, ועברנו ביחד כל פעם על הדברים שעשינו בצורה עצמאית והחלטנו מה להשאיר ומה להוריד.	

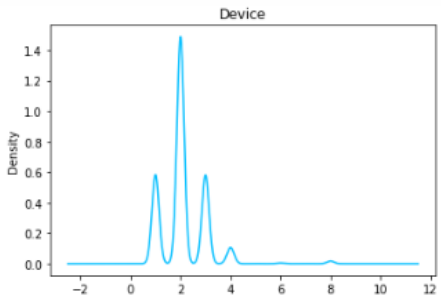
נספחים-נספח 1.1

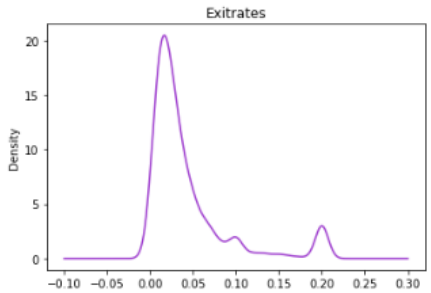


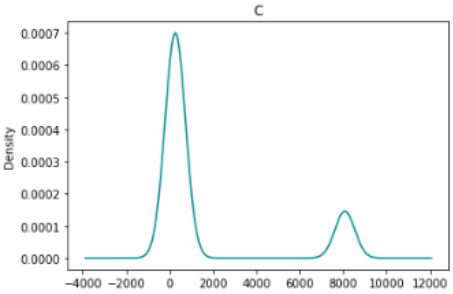
נספח 1.2

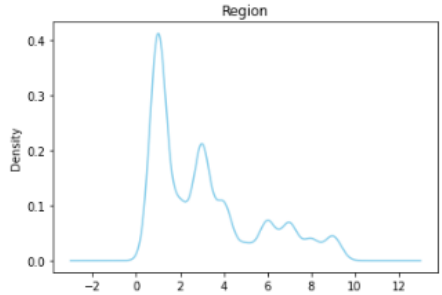


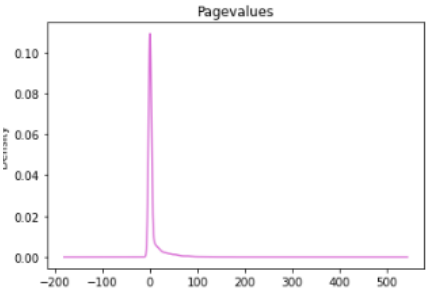

B
Density

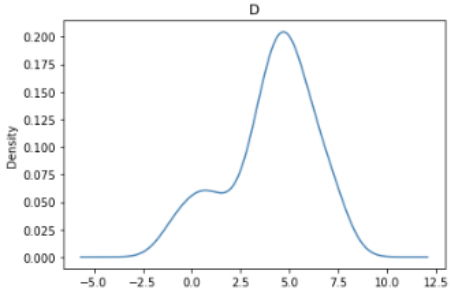

Device
Density

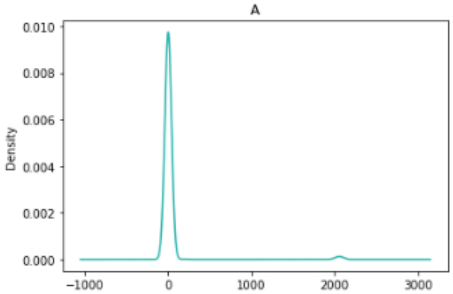

Exitrates
Density

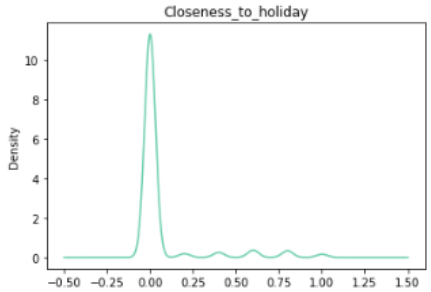

C
Density


Region
Density

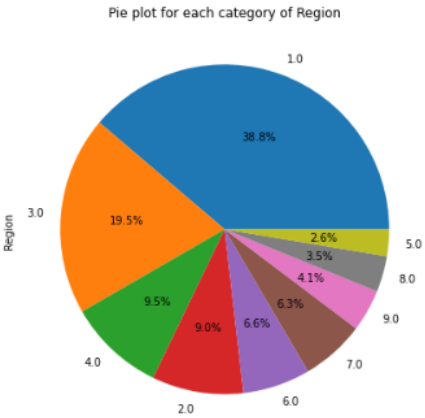

Pagevalues

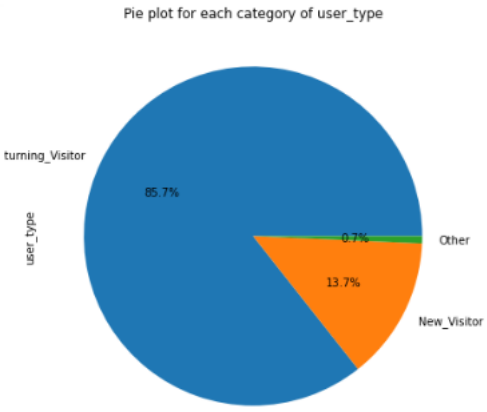

D
Density

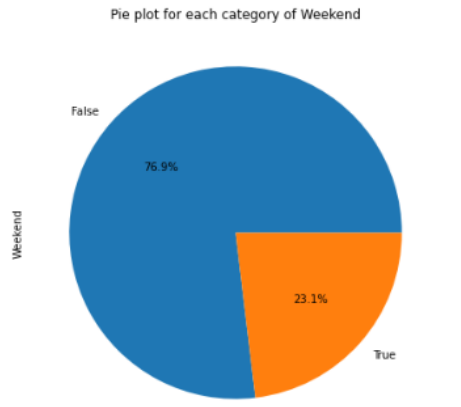

A
Density

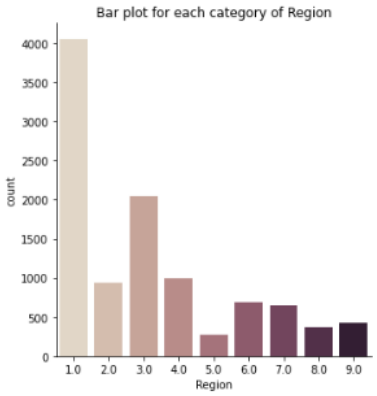

Closeness_to_holiday
Density

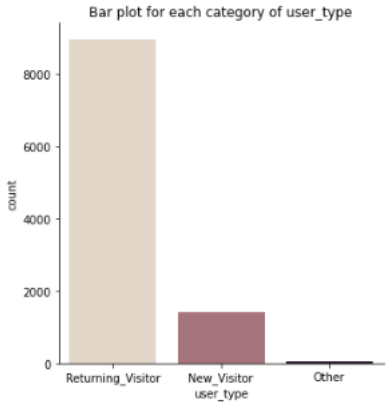
נספח 1.3

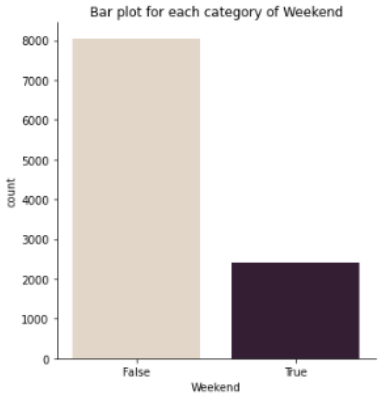

Pie plot for each category of Region
1.0
38.8%
3.0
19.5%
2.6%
5.0
3.5%
8.0
4.1%
9.0
6.3%
9.5%
9.0%
6.6%
7.0
4.0
2.0
6.0
Region

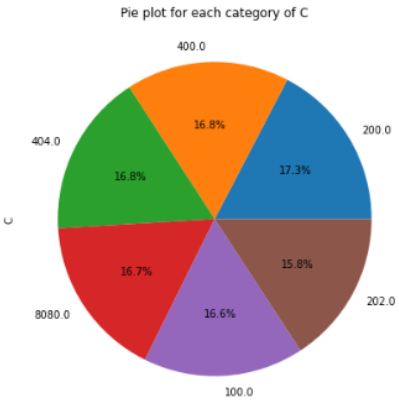

Pie plot for each category of user_type
turning_Visitor
85.7%
0.7%
Other
13.7%
New_Visitor
user_type

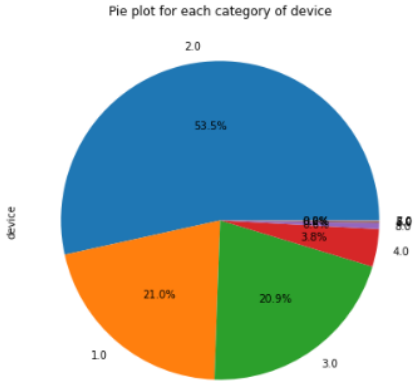

Pie plot for each category of Weekend
False
76.9%
23.1%
True
Weekend

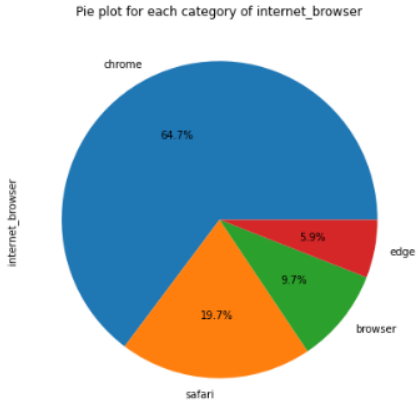

Bar plot for each category of Region
count
Region

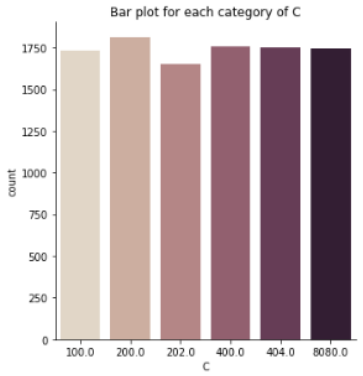

Bar plot for each category of user_type
count
Returning_Visitor
New_Visitor
Other
user_type

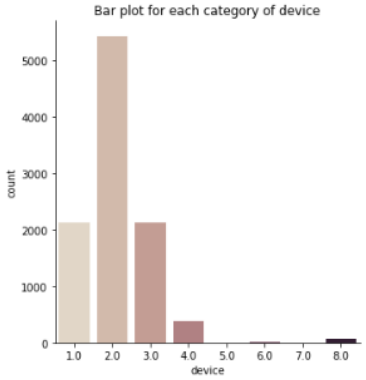

Bar plot for each category of Weekend
count
False
True
Weekend

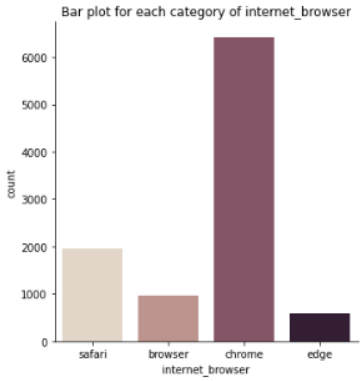

Pie plot for each category of C
400.0
16.8%
200.0
17.3%
404.0
16.8%
16.7%
15.8%
8080.0
16.6%
202.0
100.0
C

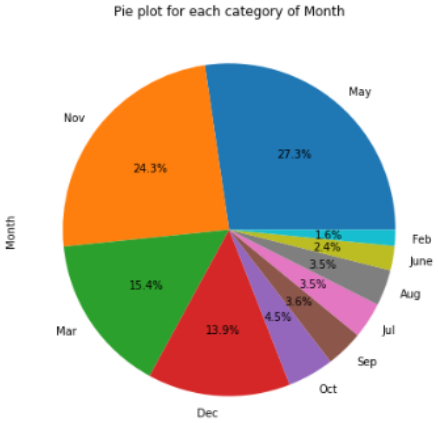

Pie plot for each category of device
2.0
53.5%
3.8%
4.0
21.0%
20.9%
1.0
3.0
device

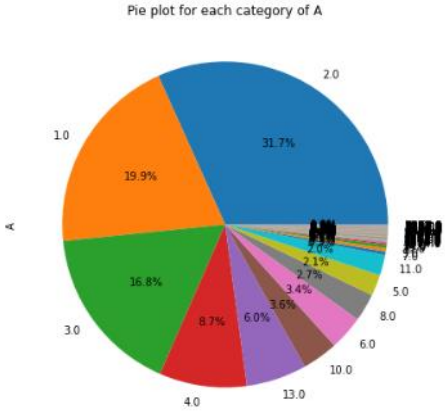

Pie plot for each category of internet_browser
chrome
64.7%
5.9%
edge
9.7%
19.7%
browser
safari
internet_browser

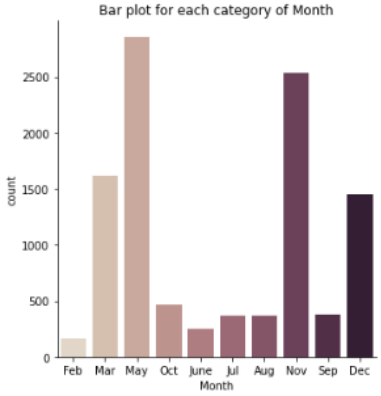

Bar plot for each category of C
count
C
100.0
200.0
202.0
400.0
404.0
8080.0

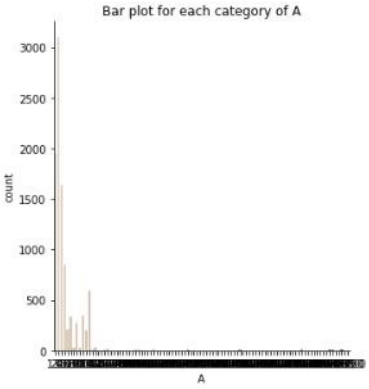

Bar plot for each category of device
count
device
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0


Bar plot for each category of internet_browser
count
safari
browser
chrome
edge
internet_browser

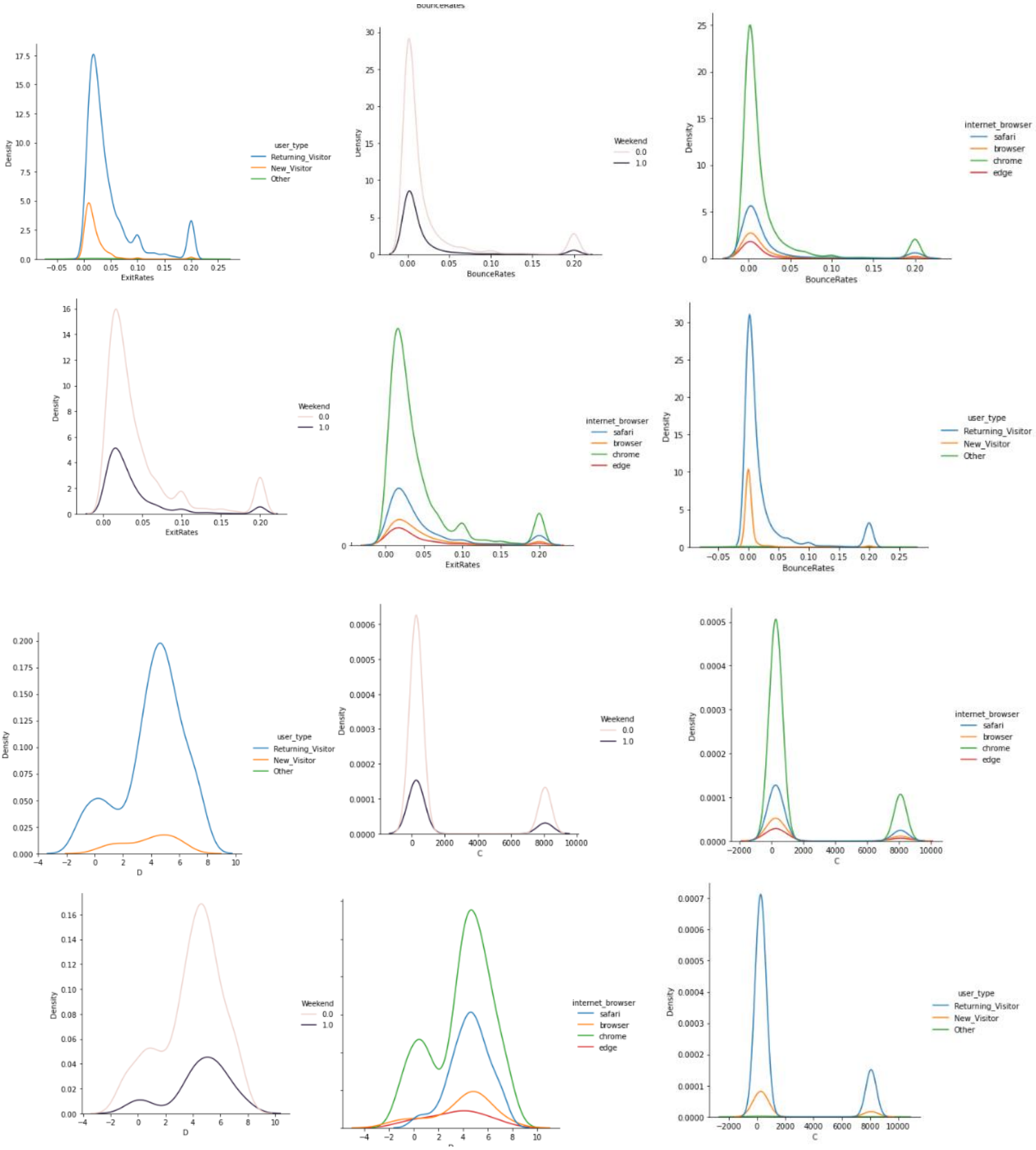

Pie plot for each category of Month
May
Nov
27.3%
24.3%
1.6%
Feb
2.4%
June
3.5%
Aug
3.5%
3.6%
Jul
15.4%
4.5%
Mar
13.9%
Sep
Oct
Dec
Month


Pie plot for each category of A
2.0
1.0
31.7%
19.9%
2.0%
2.1%
2.7%
11.0
3.4%
5.0
16.8%
3.6%
8.0
6.0%
8.7%
6.0
3.0
10.0
13.0
4.0
A

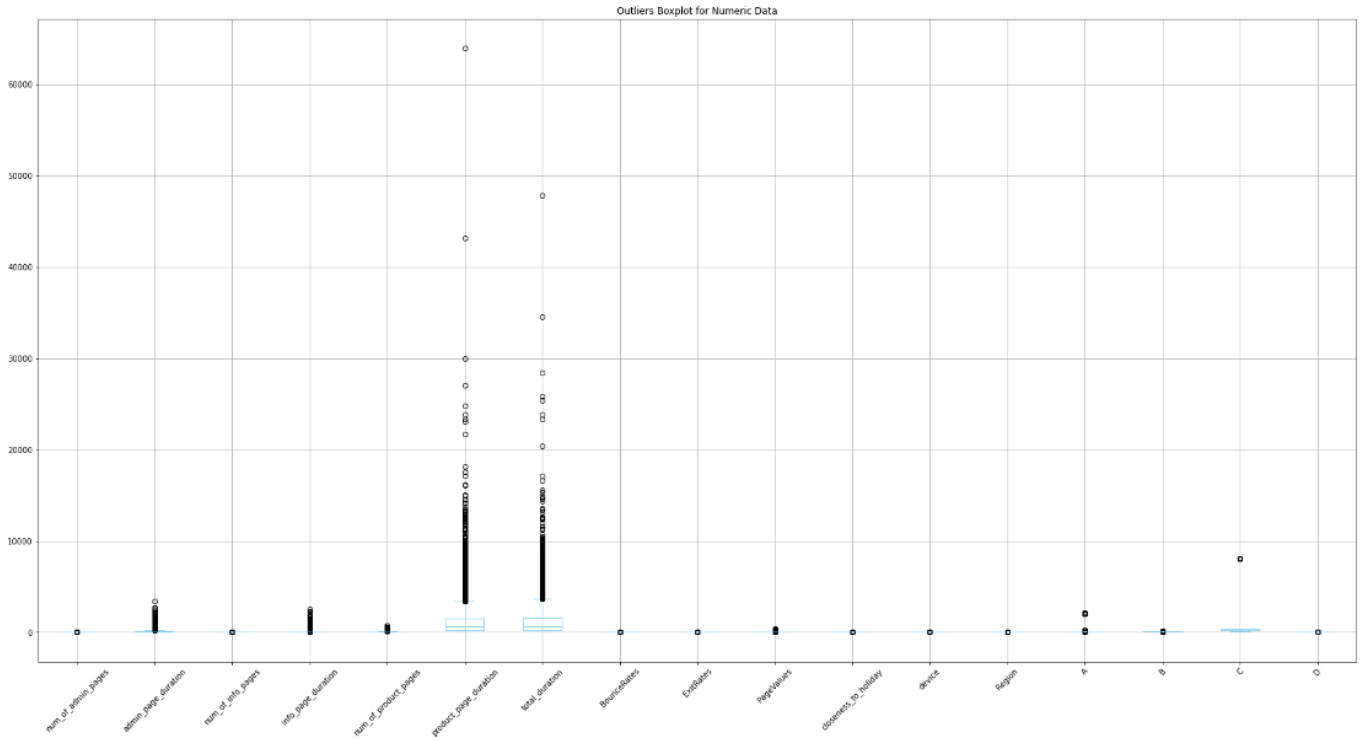

Bar plot for each category of Month
count
Feb
Mar
May
Oct
June
Jul
Aug
Nov
Sep
Dec
Month


Bar plot for each category of A
count
A

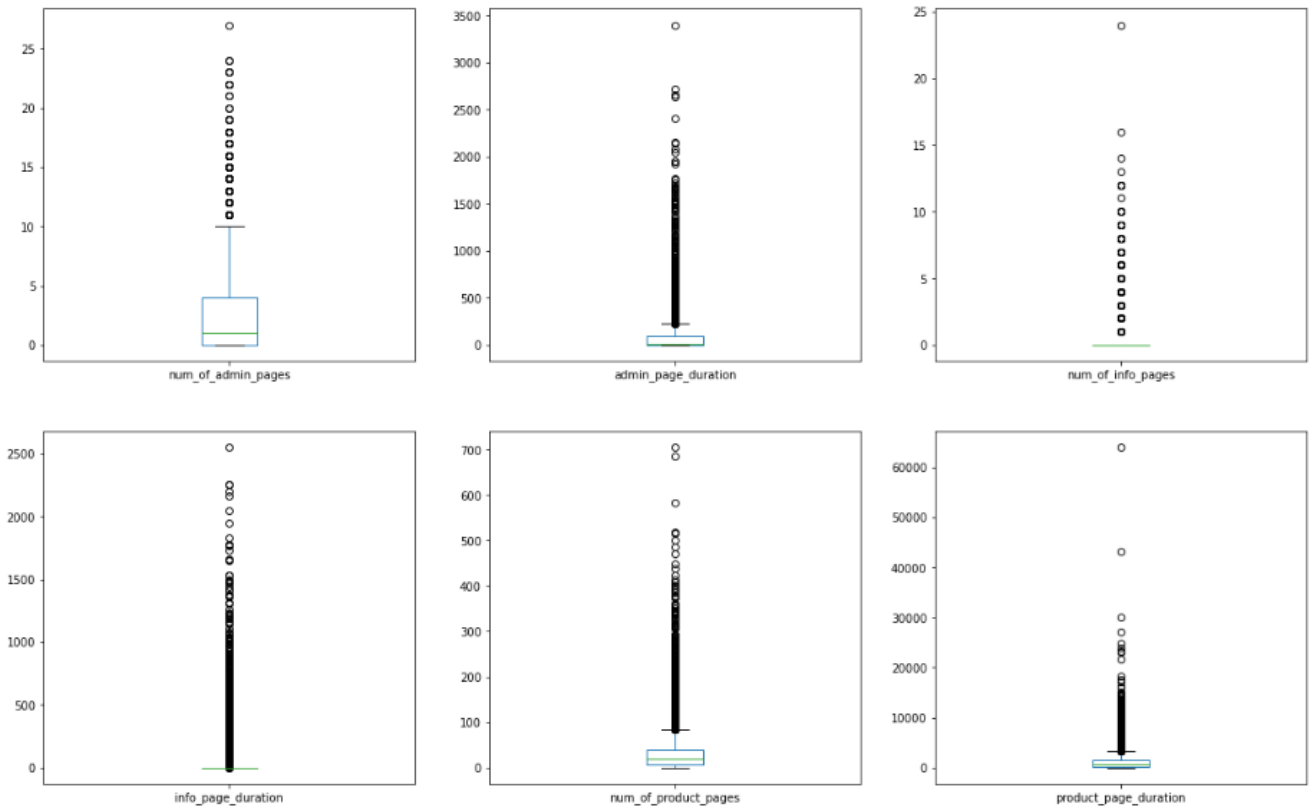
נספח 1.4

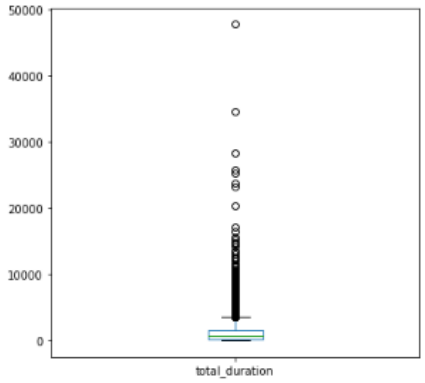

Density
ExitRates
user_type
Returning_Visitor
New_Visitor
Other
BounceRates
Weekend
0.0
1.0
internet_browser
safari
browser
chrome
edge
D
C

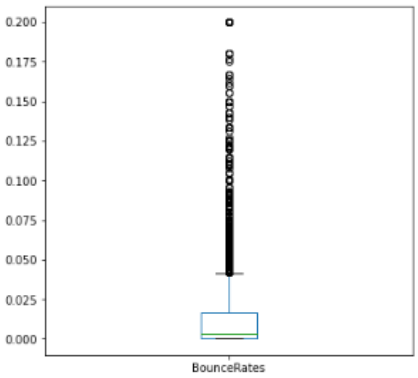
נספח 1.5

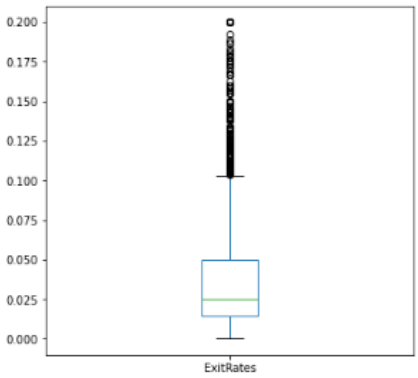


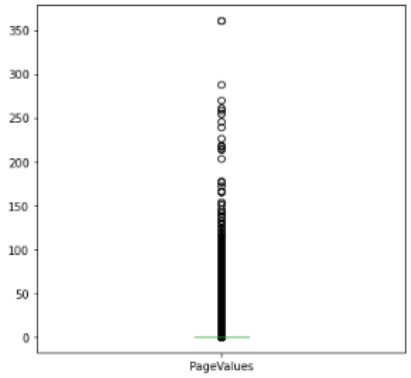
נספח 1.6

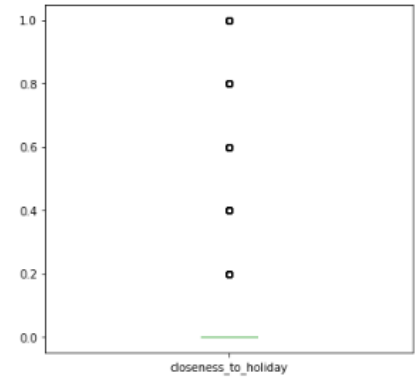


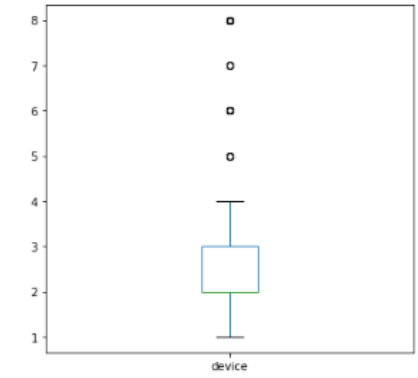

50000
40000
30000
20000
10000
0
total_duration

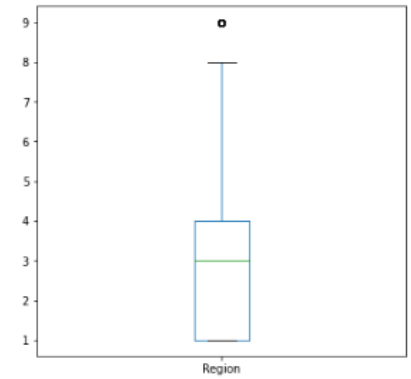

0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
BounceRates

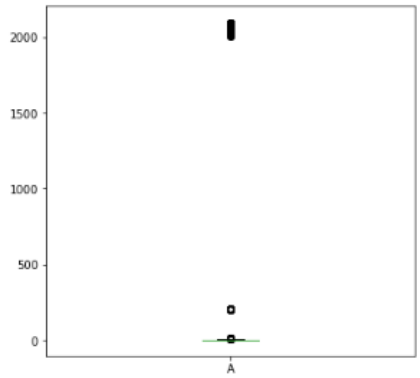

0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
ExitRates

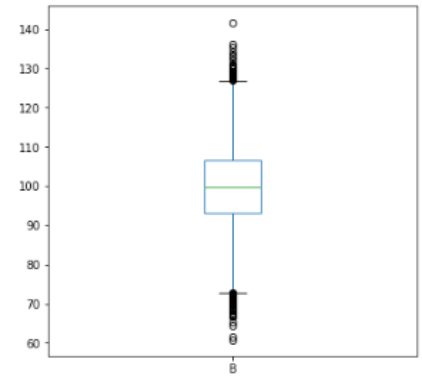

350
300
250
200
150
100
50
0
PageValues

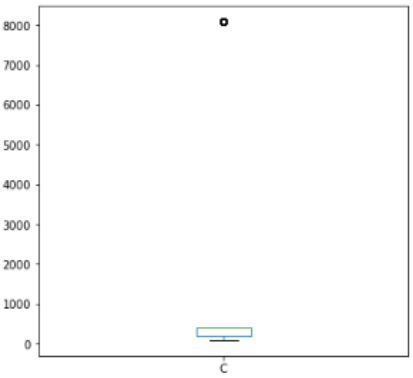

1.0
0.8
0.6
0.4
0.2
0.0
closeness_to_holiday

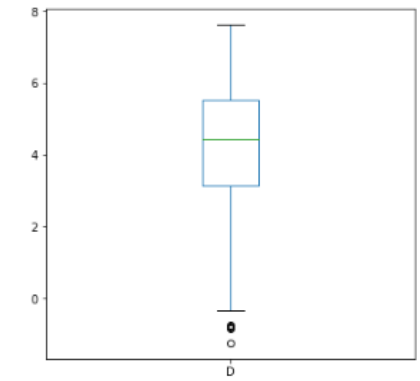

8
7
6
5
4
3
2
1
device


9
8
7
6
5
4
3
2
1
Region

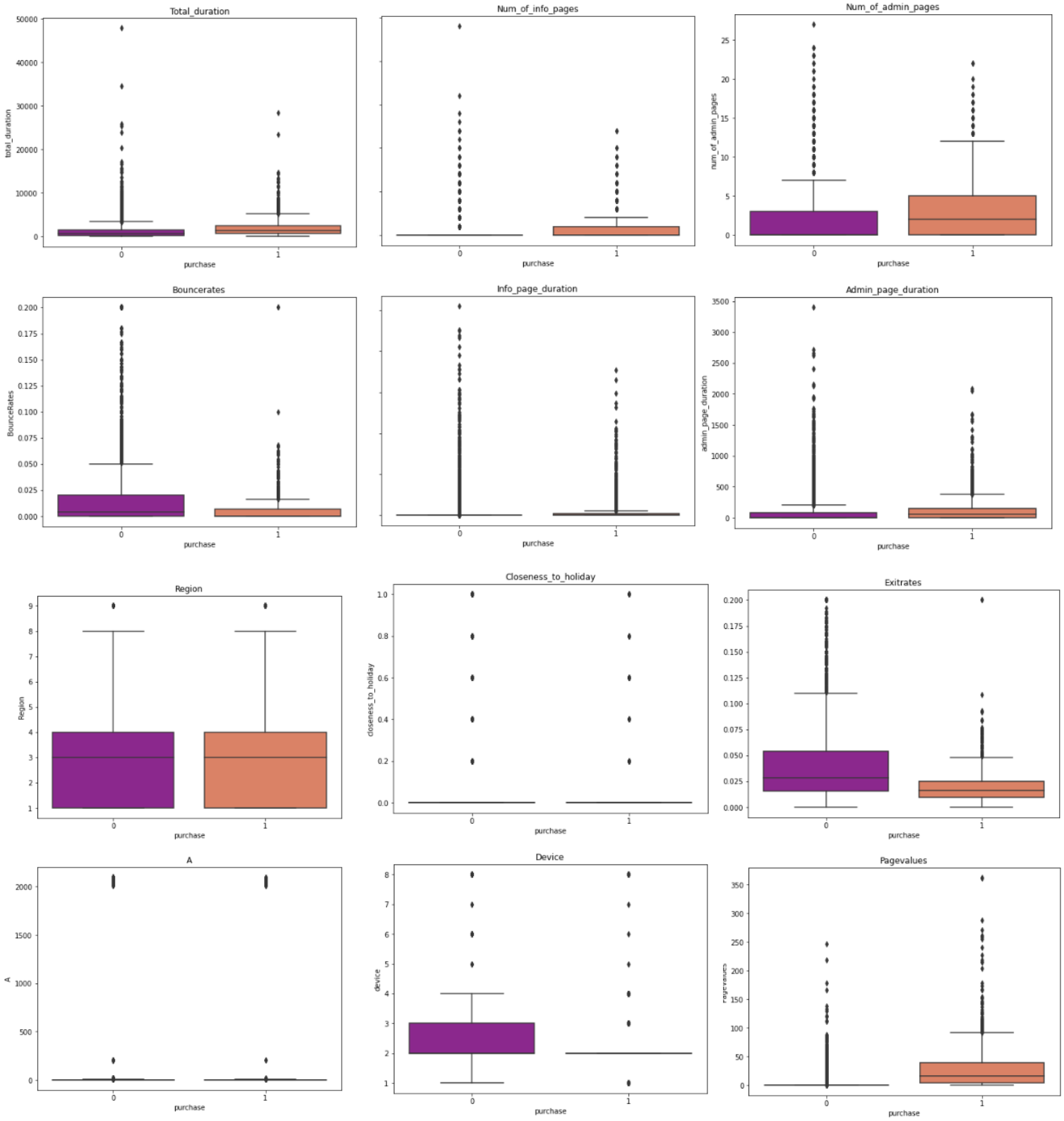

2000
1500
1000
500
0
A

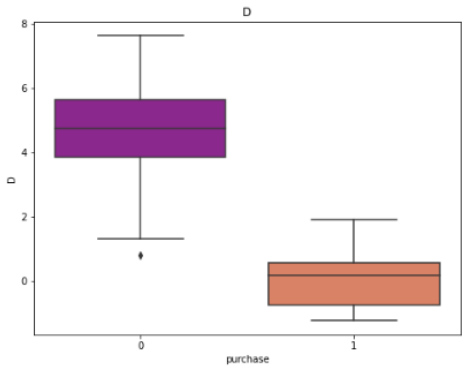

140
130
120
110
100
90
80
70
60
B

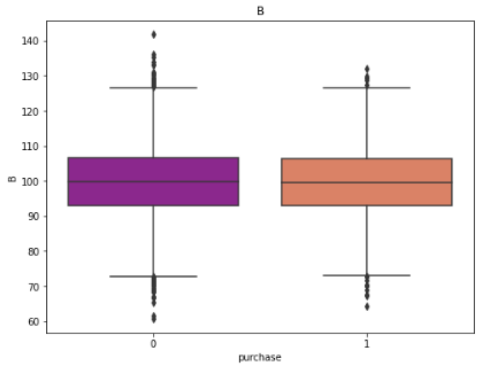

8000
7000
6000
5000
4000
3000
2000
1000
0
C

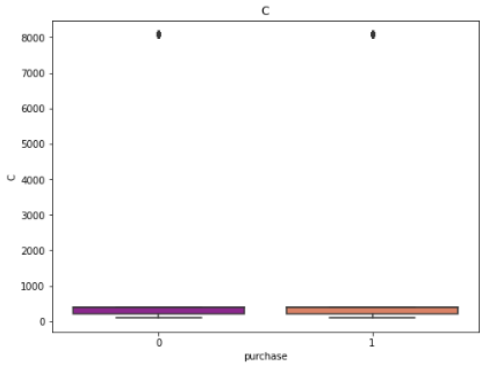

8
6
4
2
0
D

נספח 1.7

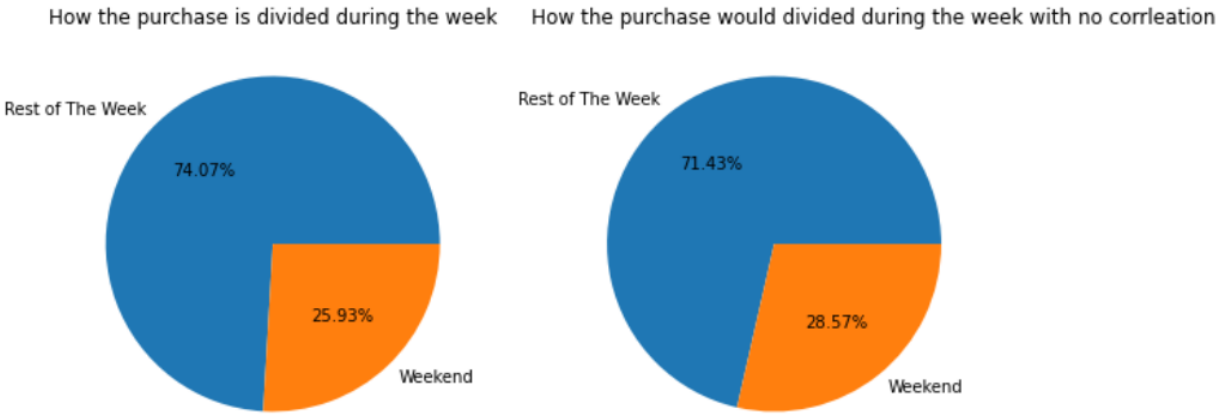

Total_duration
total_duration
purchase
Num_of_info_pages
purchase
Num_of_admin_pages
num_of_admin_pages
purchase
Bouncerates
BounceRates
purchase
Info_page_duration
purchase
Admin_page_duration
admin_page_duration
purchase
Region
Region
purchase
Closeness_to_holiday
closeness_to_holiday
purchase
Exitrates
purchase
A
A
purchase
Device
device
purchase
Pagevalues
pagevalues
purchase


D
8
6
4
2
0
D
0
1
purchase

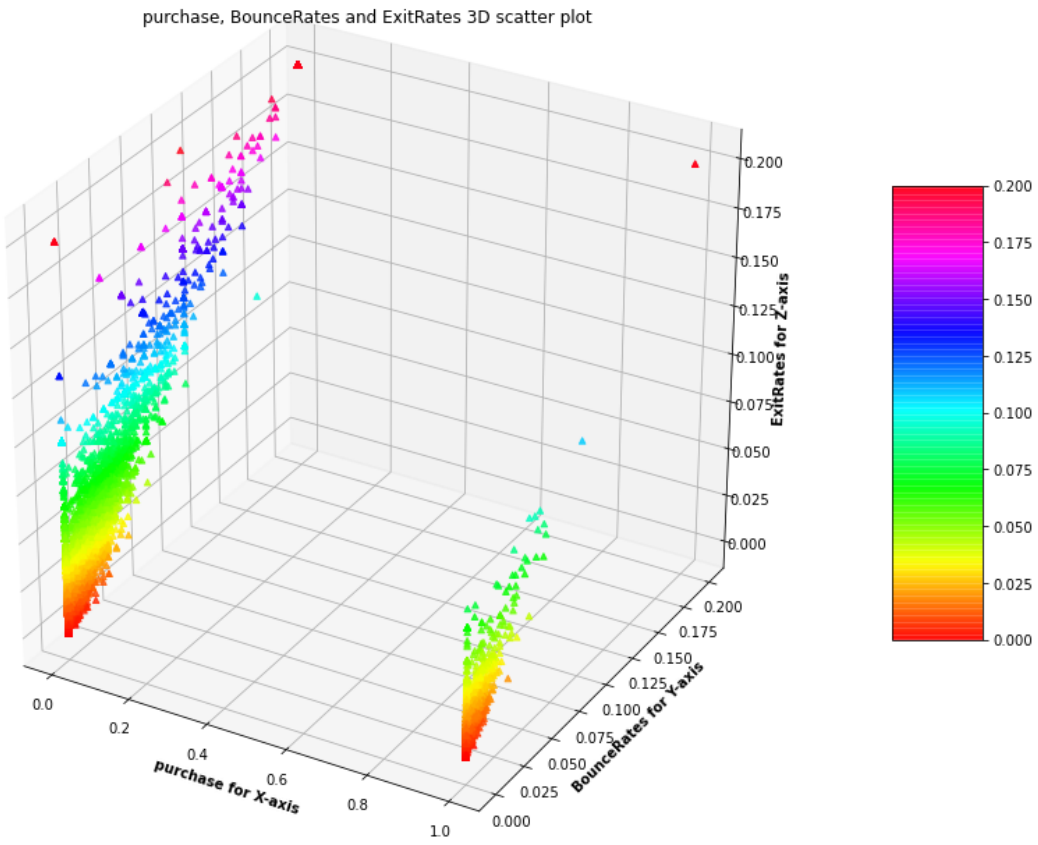

B
140
130
120
110
100
90
80
70
60
B
0
1
purchase

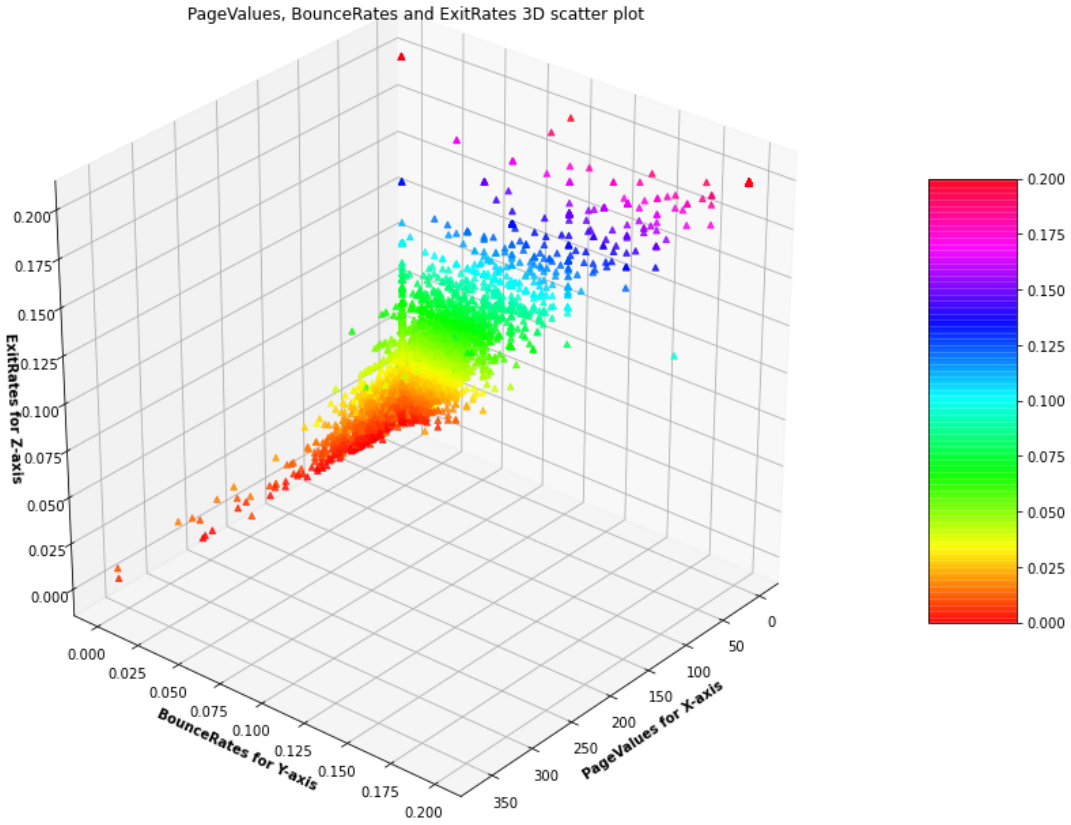

C
8000
7000
6000
5000
4000
3000
2000
1000
0
C
0
1
purchase

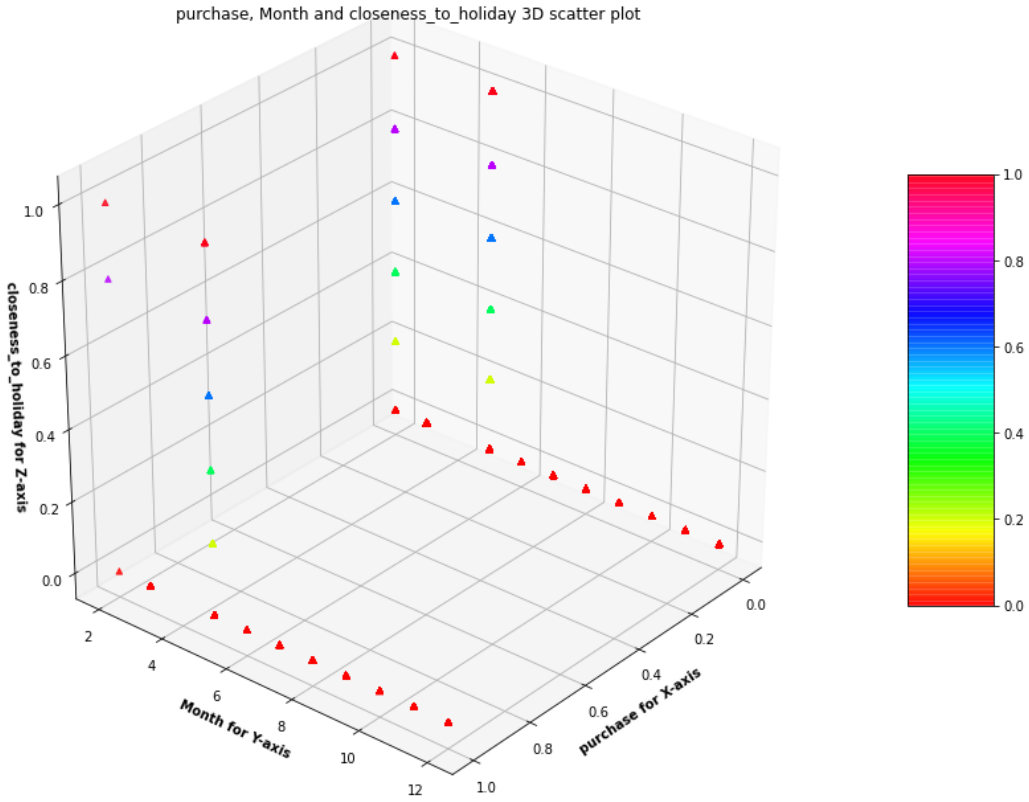
נספח 1.8

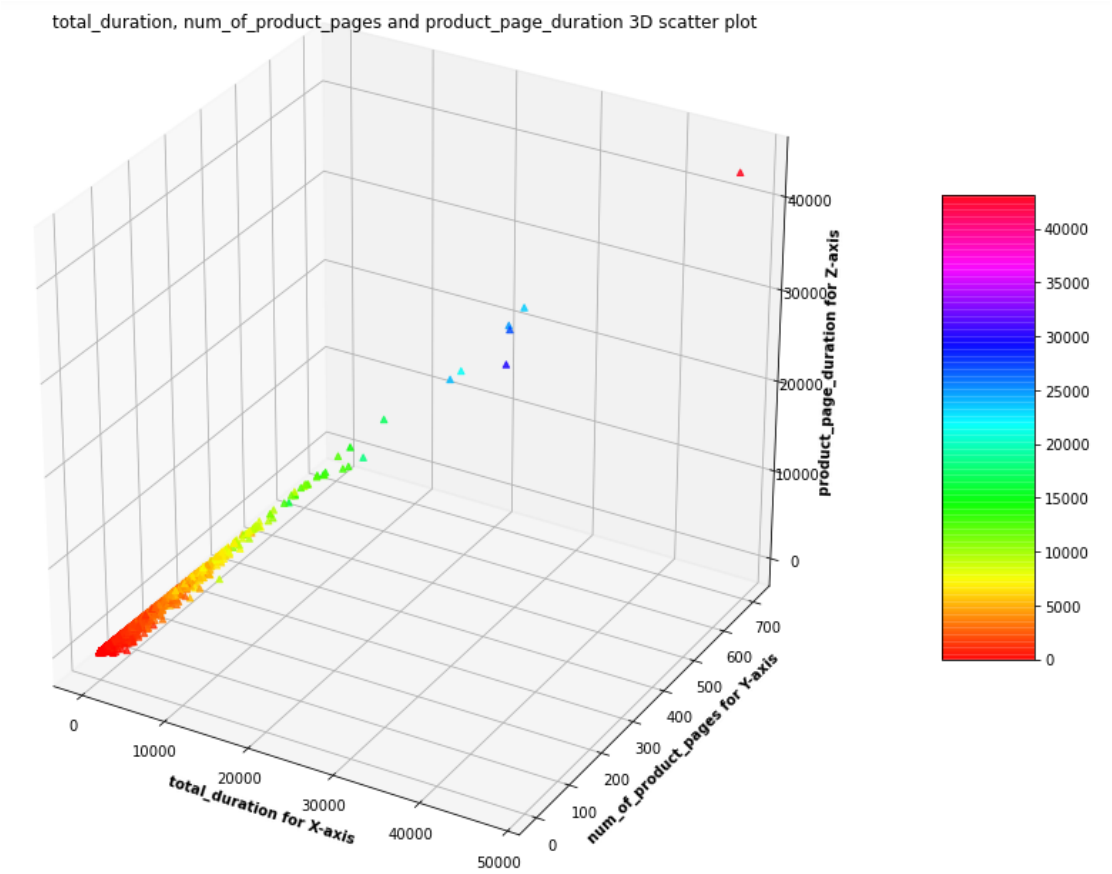


נספח 1.9

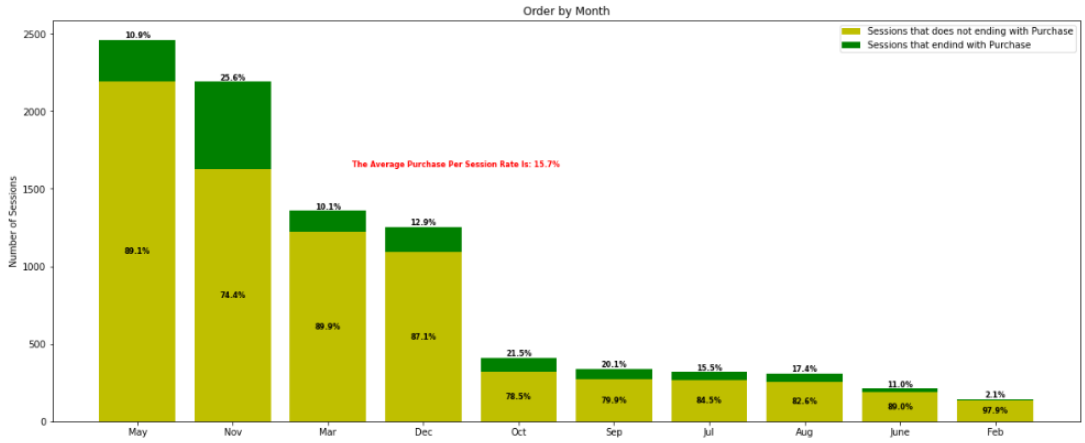


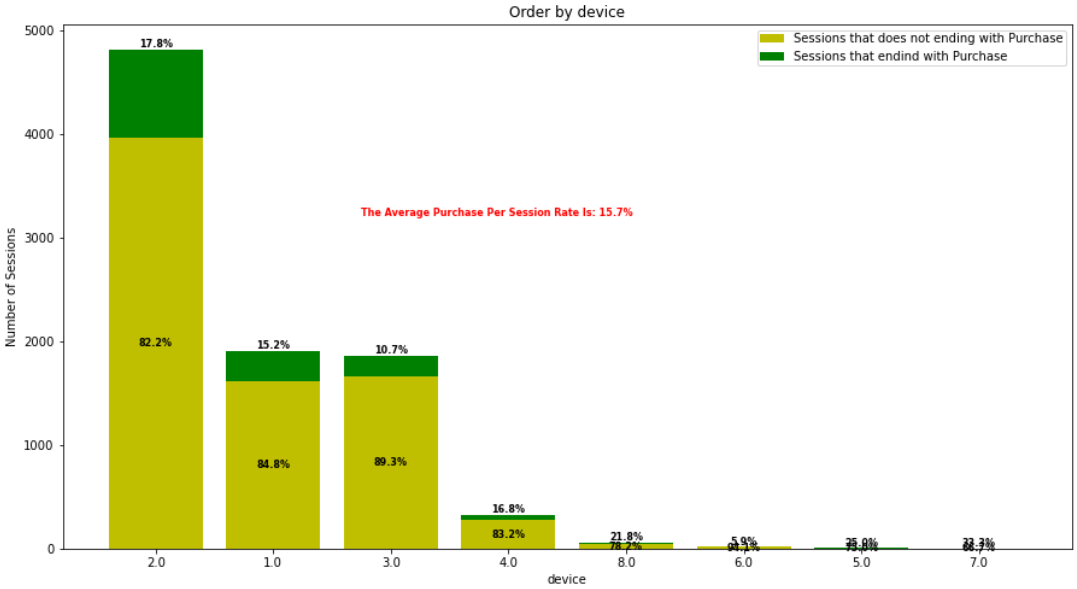

PageValues, BounceRates and ExitRates 3D scatter plot
ExitRates for Z-axis
BounceRates for Y-axis
PageValues for X-axis

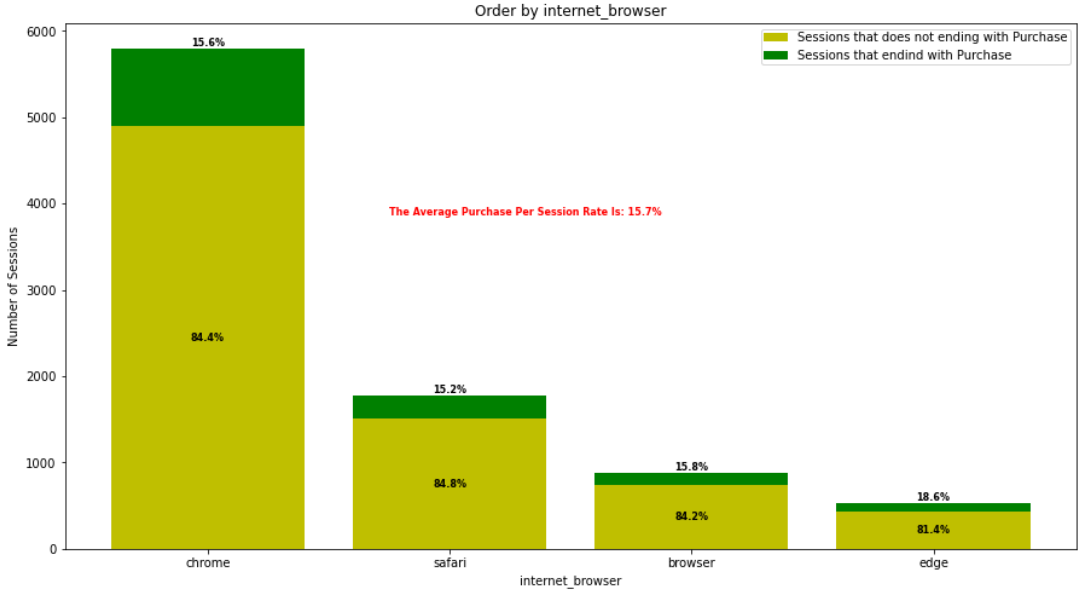

purchase, Month and closeness_to_holiday 3D scatter plot
closeness_to_holiday for Z-axis
Month for Y-axis
purchase for X-axis

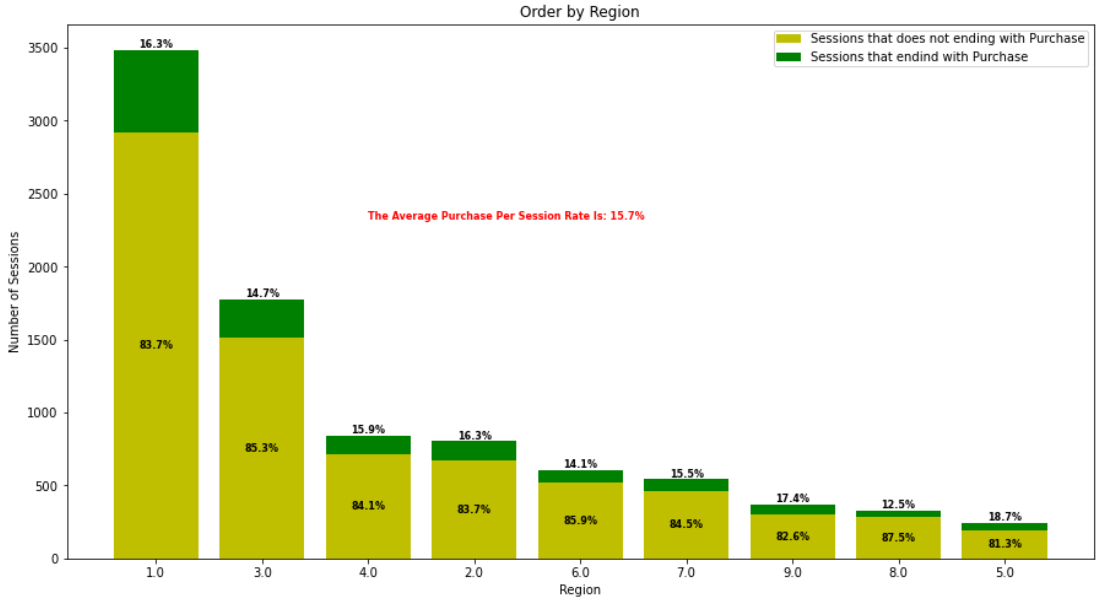

total_duration, num_of_product_pages and product_page_duration 3D scatter plot
total_duration for X-axis
num_of_product_pages for Y-axis
product_page_duration for Z-axis

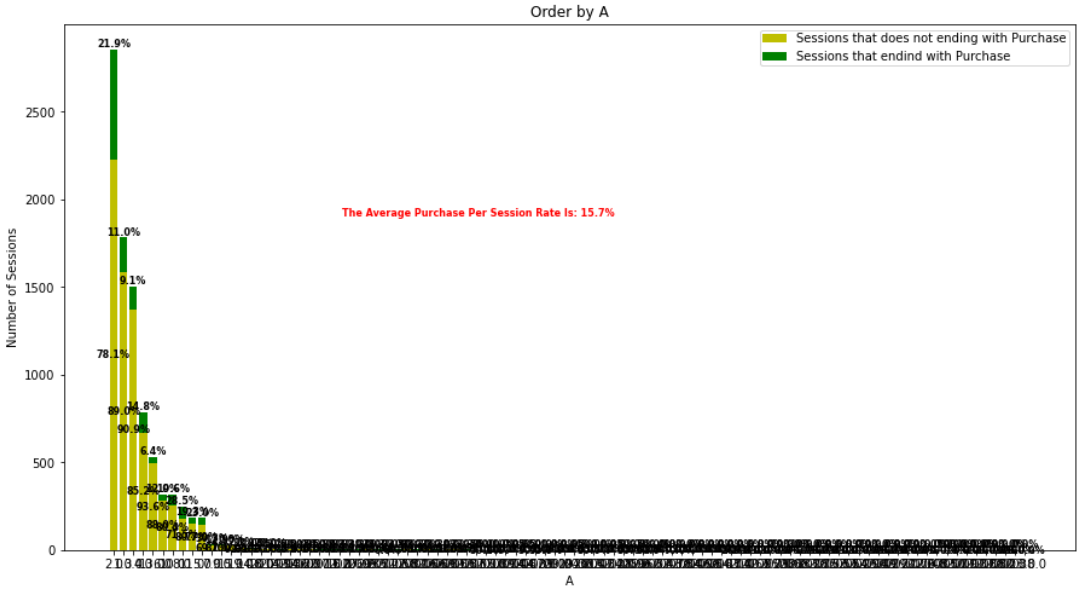
נספח 2.0


Order by Month
Sessions that does not ending with Purchase
Sessions that endind with Purchase
The Average Purchase Per Session Rate Is: 15.7%
Number of Sessions
May
Nov
Mar
Dec
Oct
Sep
Jul
Aug
June
Feb

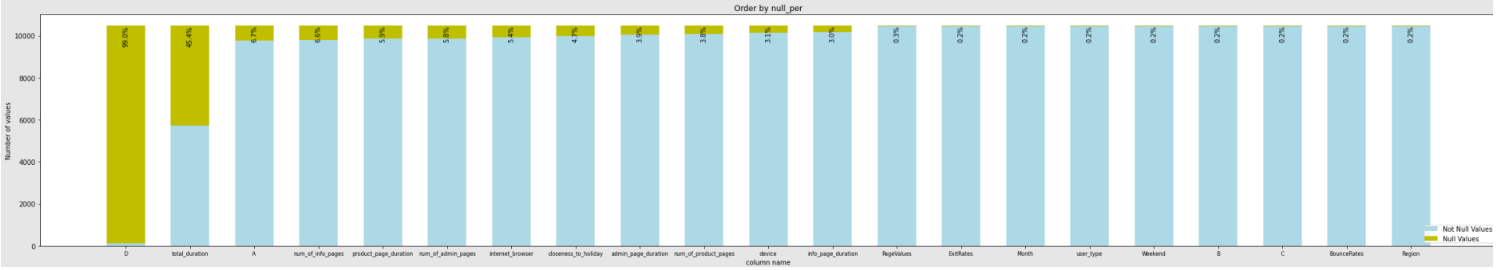

Order by device
Sessions that does not ending with Purchase
Sessions that endind with Purchase
The Average Purchase Per Session Rate Is: 15.7%
Number of Sessions
device


Order by internet_browser
Sessions that does not ending with Purchase
Sessions that endind with Purchase
The Average Purchase Per Session Rate Is: 15.7%
Number of Sessions
internet_browser
chrome
safari
browser
edge
15.6%
84.4%
15.2%
84.8%
15.8%
84.2%
18.6%
81.4%

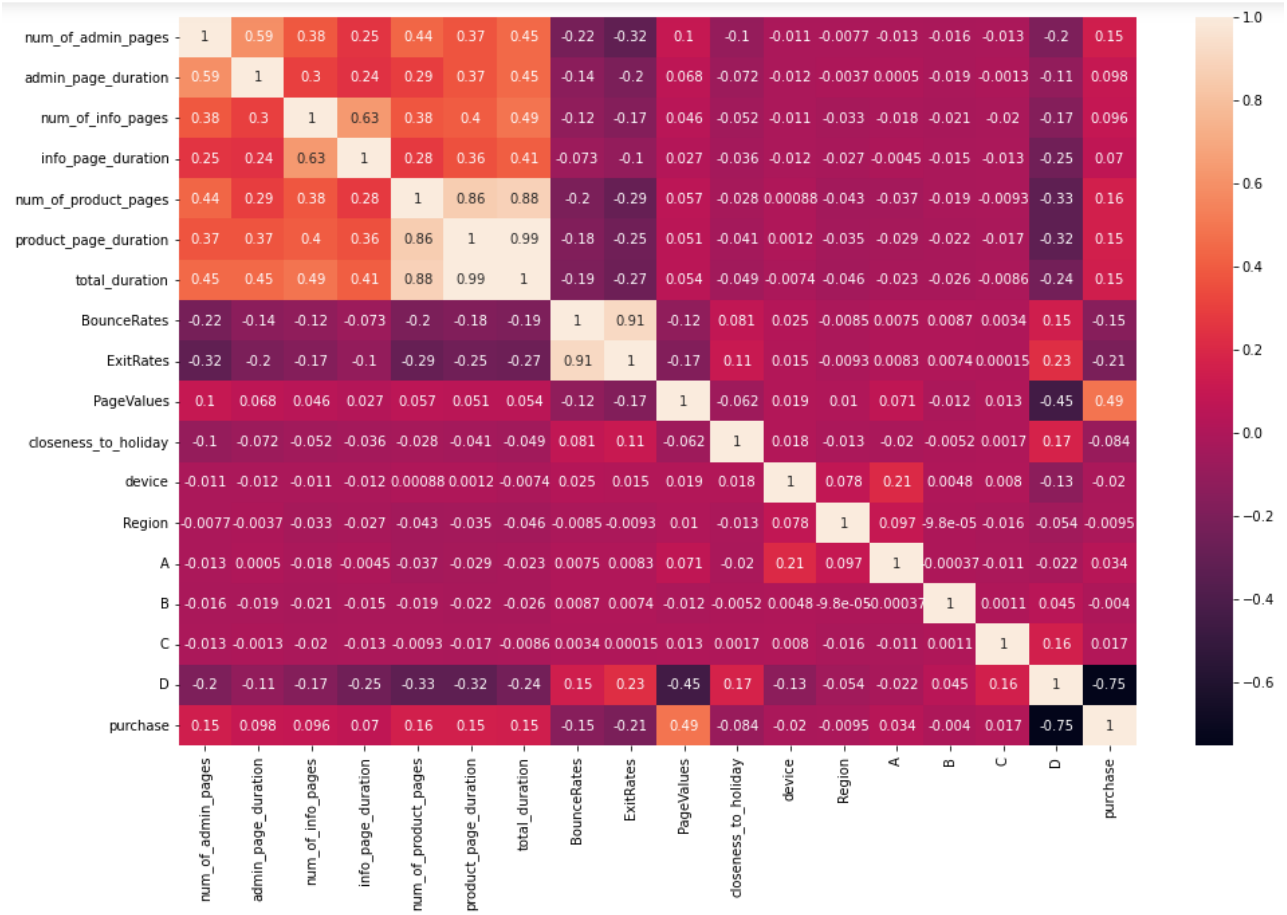

Order by Region
Sessions that does not ending with Purchase
Sessions that endind with Purchase
The Average Purchase Per Session Rate Is: 15.7%
Number of Sessions
Region
1.0
3.0
4.0
2.0
6.0
7.0
9.0
8.0
5.0
16.3%
83.7%
14.7%
85.3%
15.9%
84.1%
16.3%
83.7%
14.1%
85.9%
15.5%
84.5%
17.4%
82.6%
12.5%
87.5%
18.7%
81.3%


Order by A
Sessions that does not ending with Purchase
Sessions that endind with Purchase
The Average Purchase Per Session Rate Is: 15.7%
Number of Sessions
A
21.9%
11.0%
9.1%
78.1%
6.4%

נספח 2.1



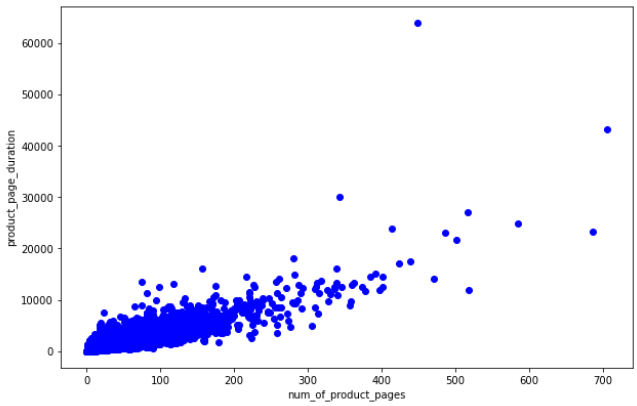
נספח 2.2

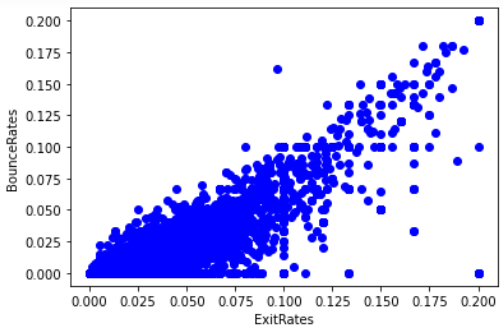


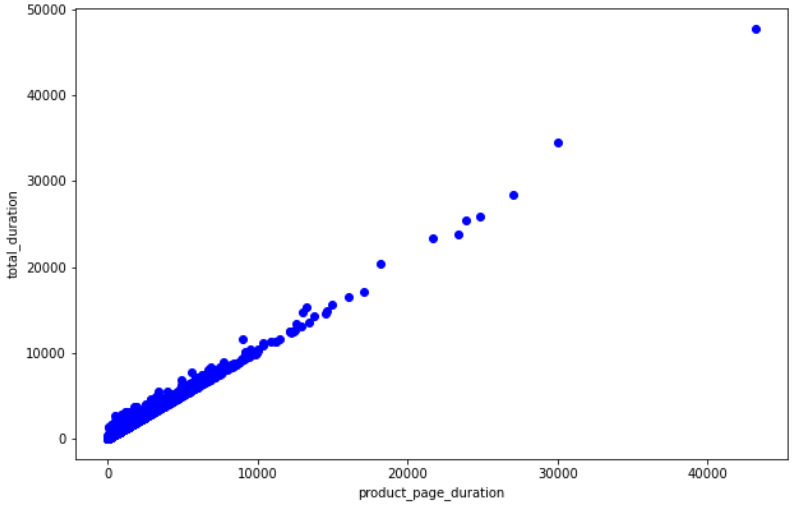
נספח 2.3

purchase	
0.486404	PageValues
0.157167	num_of_product_pages
0.152133	product_page_duration
0.145429	total_duration
0.145048	num_of_admin_pages
0.097504	admin_page_duration
0.095563	num_of_info_pages
0.070309	info_page_duration
0.034379	A
0.016940	C
0.003981-	B
0.009488-	Region
0.020196-	device
0.083926-	closeness_to_holiday
0.150683-	BounceRates
0.207804-	ExitRates
0.753238-	D

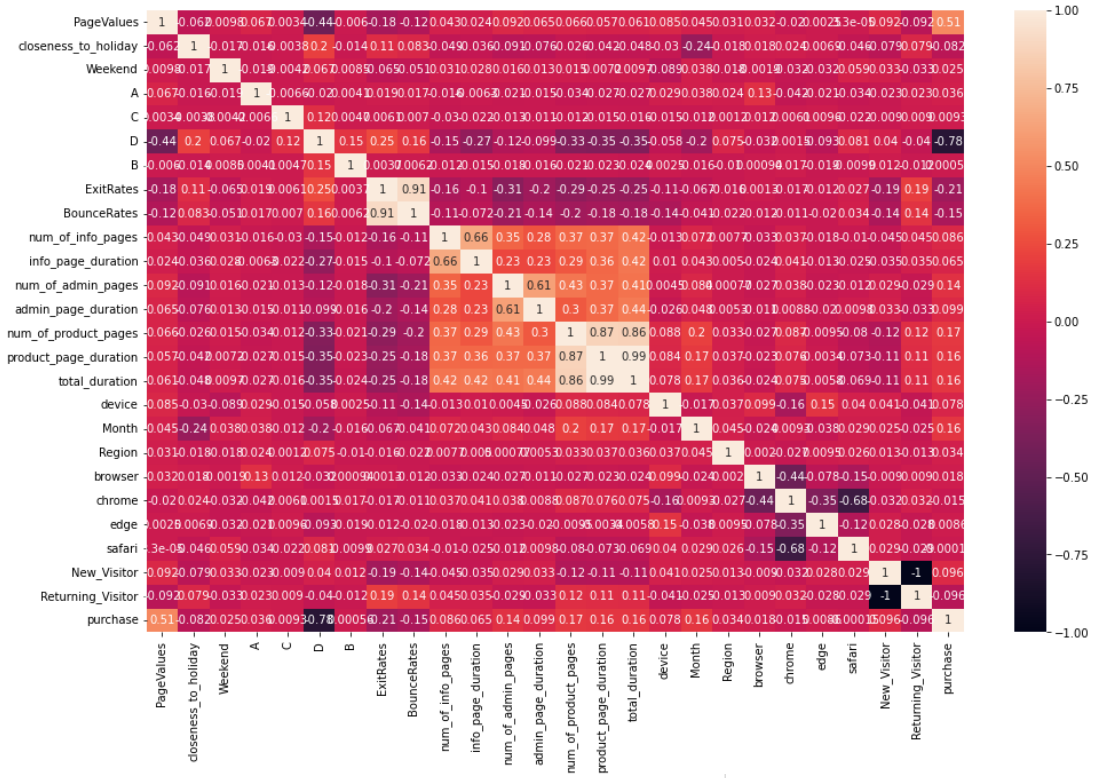
נספח 2.4





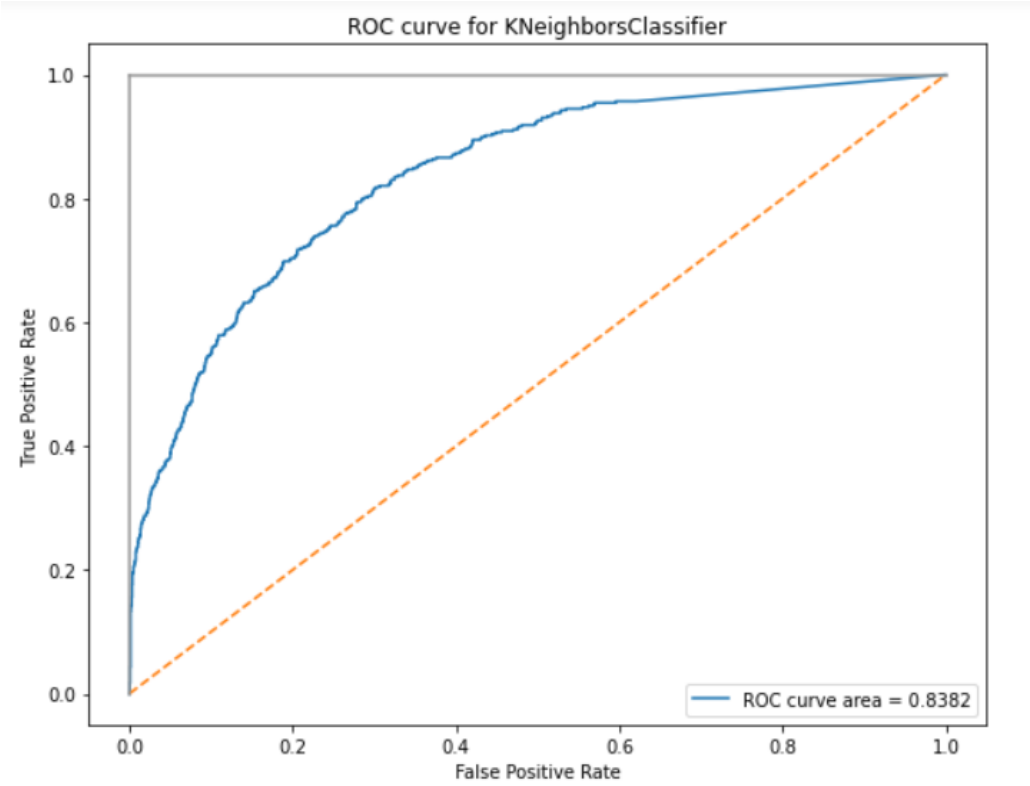


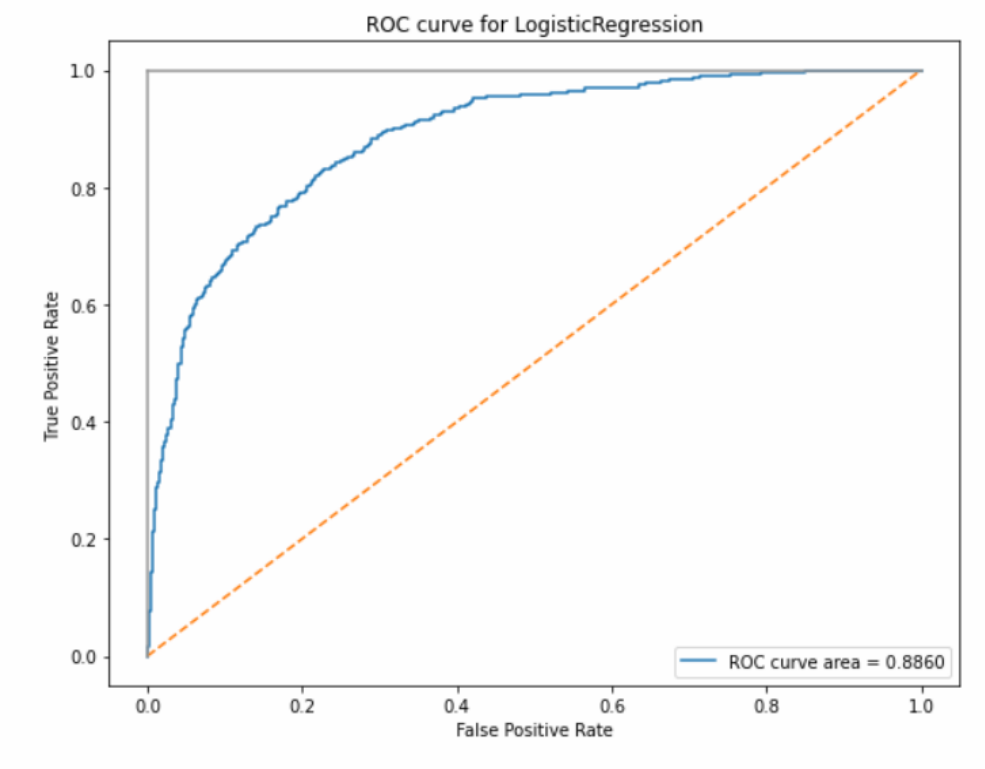
נספח 2.5

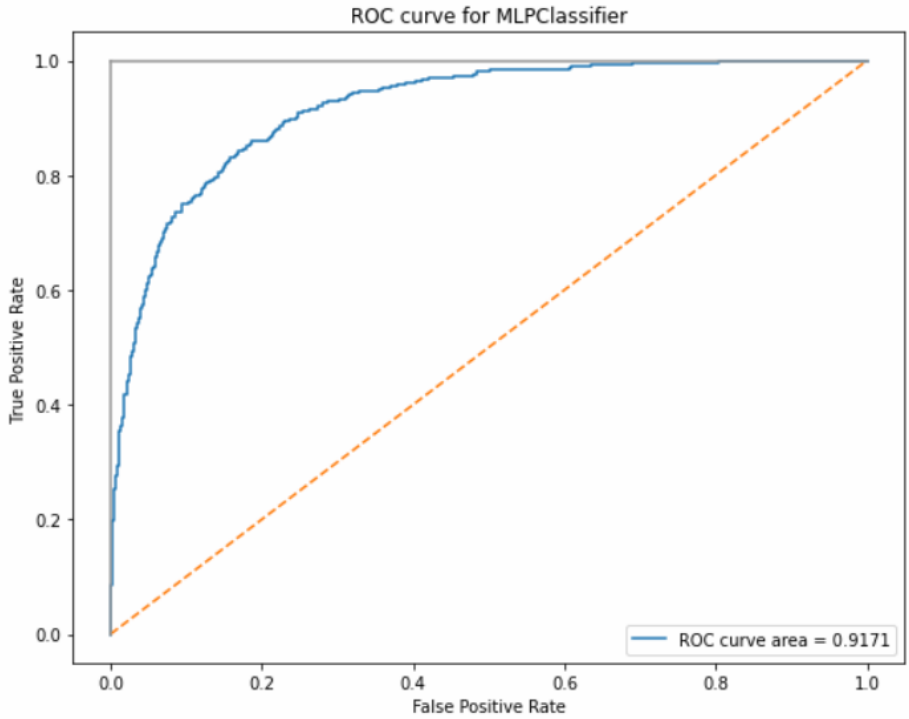


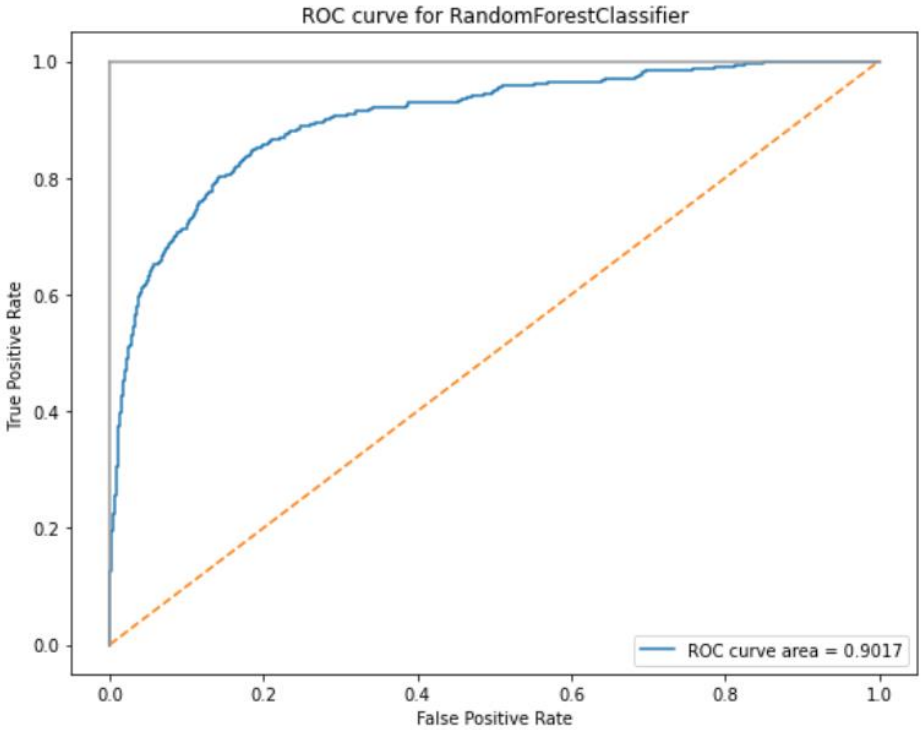
purchase	
0.505062	PageValues
0.170503	num_of_product_pages
0.163910	Month
0.163278	total_duration
0.160400	product_page_duration
0.138895	num_of_admin_pages
0.098831	admin_page_duration
0.096052	New_Visitor
0.086451	num_of_info_pages
0.078071	device
0.065203	info_page_duration
0.035616	A
0.034122	Region
0.025184	Weekend
0.018483	browser
0.009279	C
0.008579	edge
0.000562	B
0.000154-	safari
0.015344-	chrome
0.081945-	closeness_to_holiday
0.096052-	Returning_Visitor
0.150210-	BounceRates
0.206133-	ExitRates
0.782873-	D

פלטים- נספח 2.6

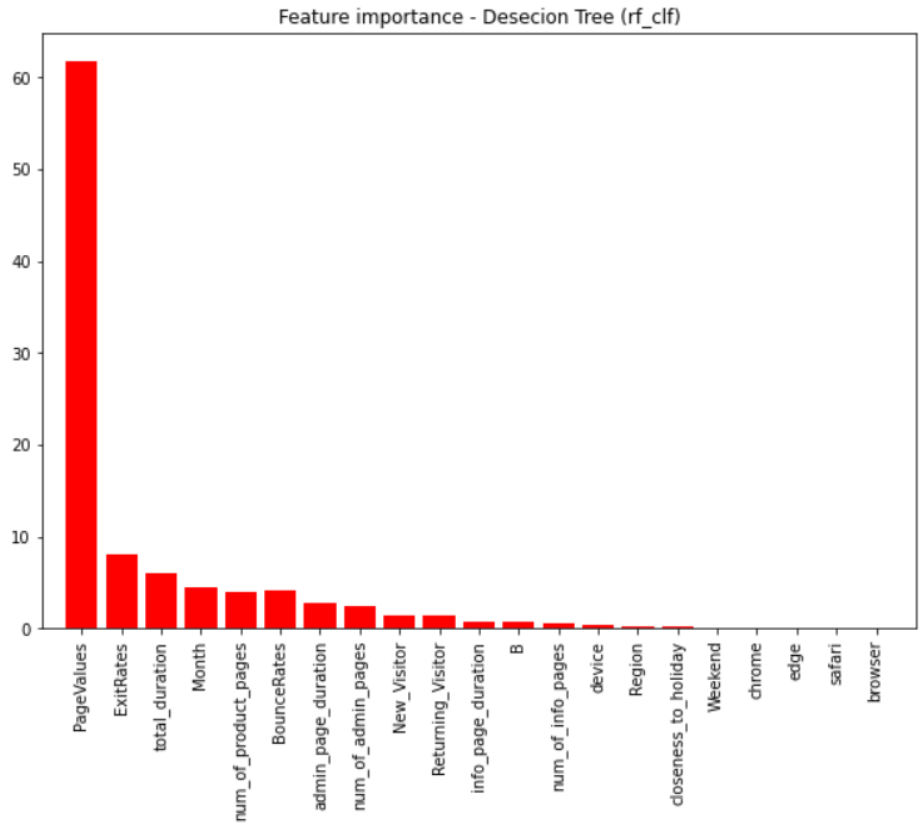




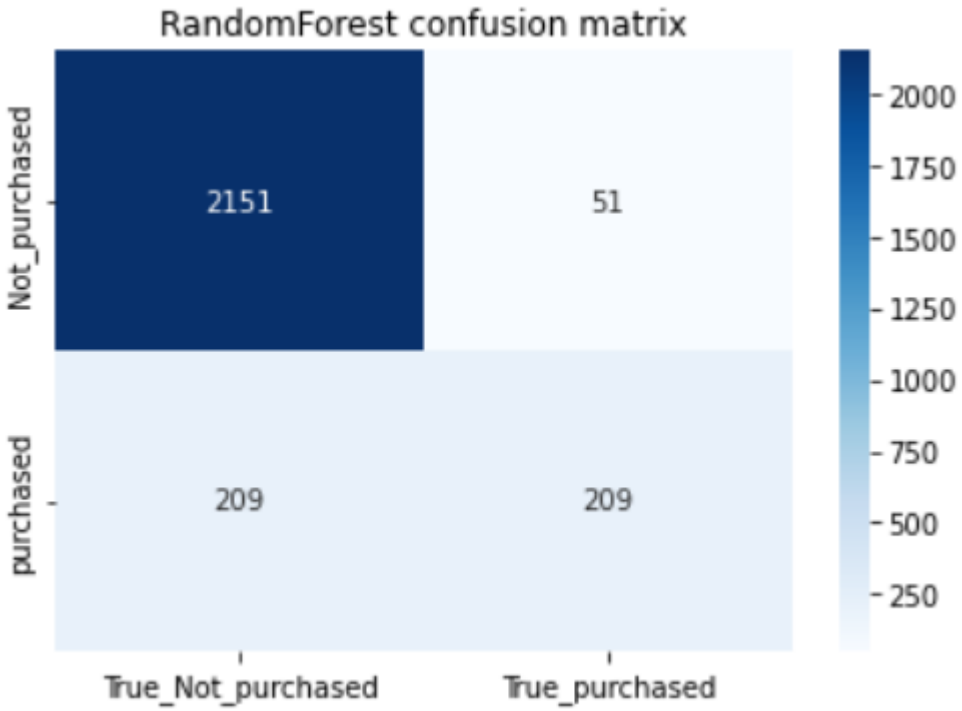

ROC curve for MLPClassifier
True Positive Rate
False Positive Rate
ROC curve area = 0.9171
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve for RandomForestClassifier
True Positive Rate
False Positive Rate
ROC curve area = 0.9017
0.0
0.2
0.4
0.6
0.8
1.0

נספח 2.7

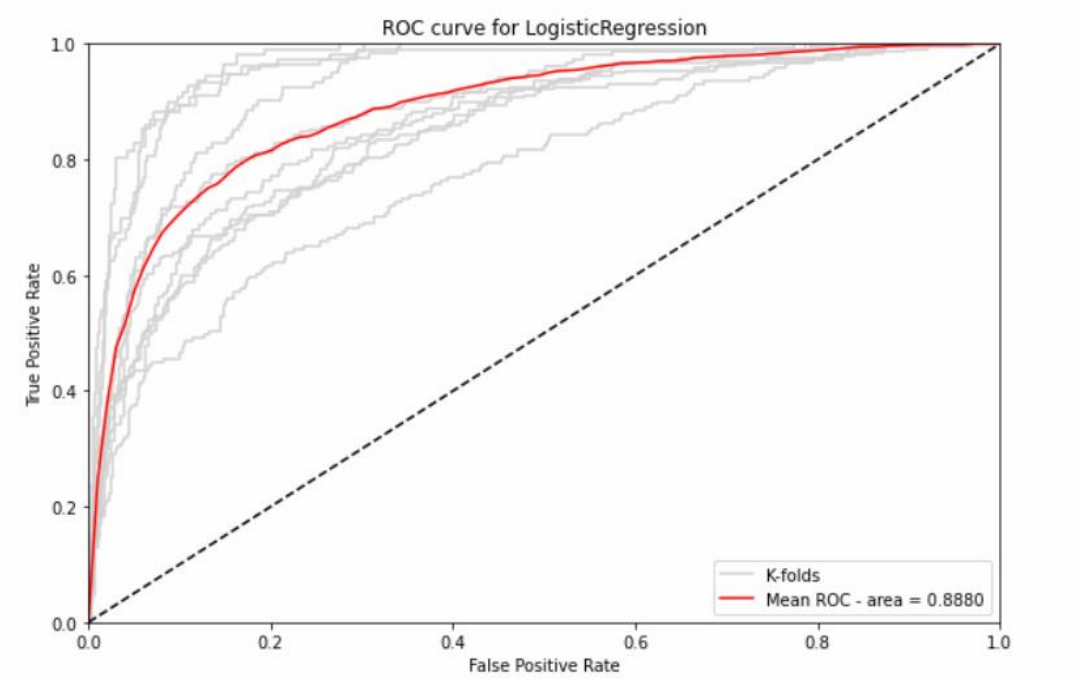


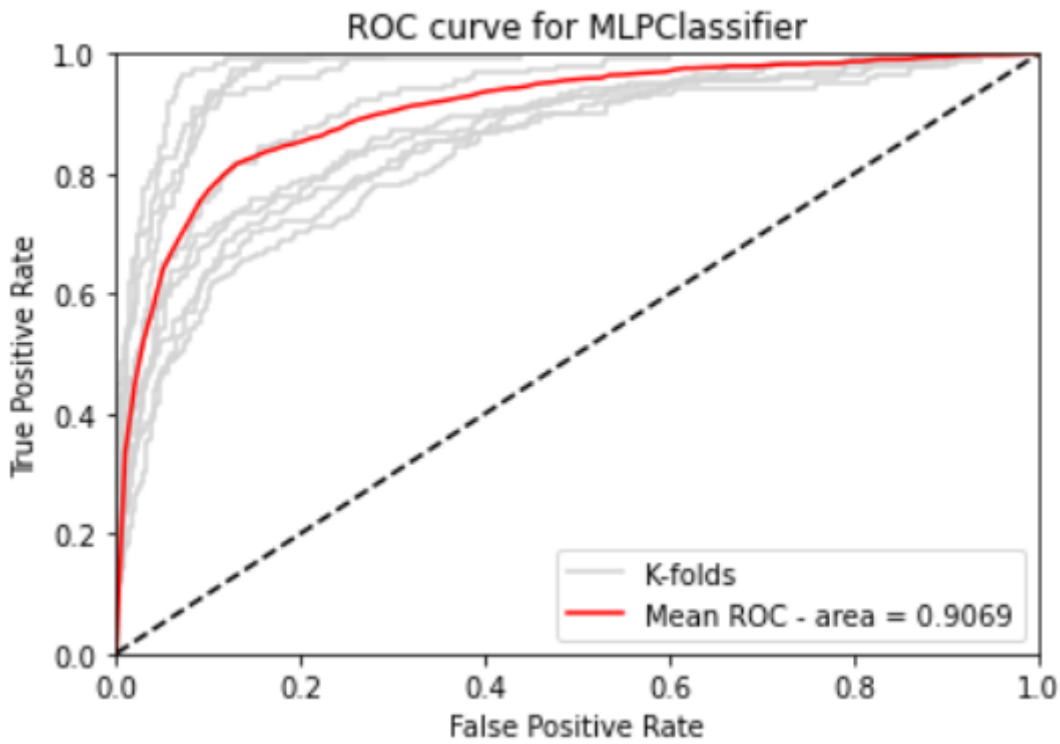
נספח 2.8

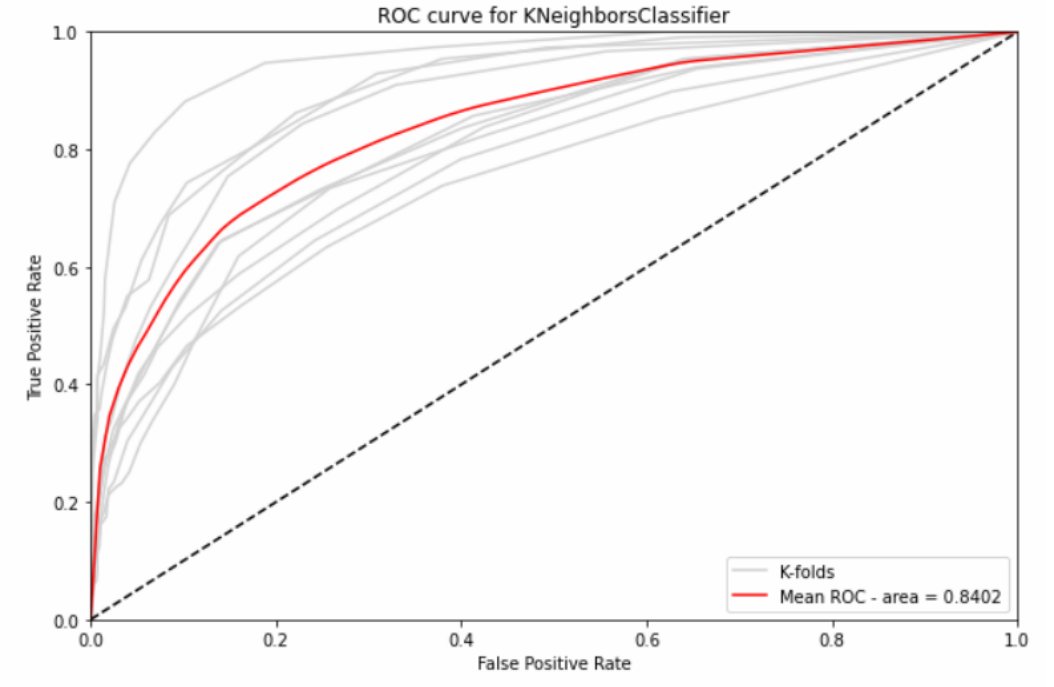


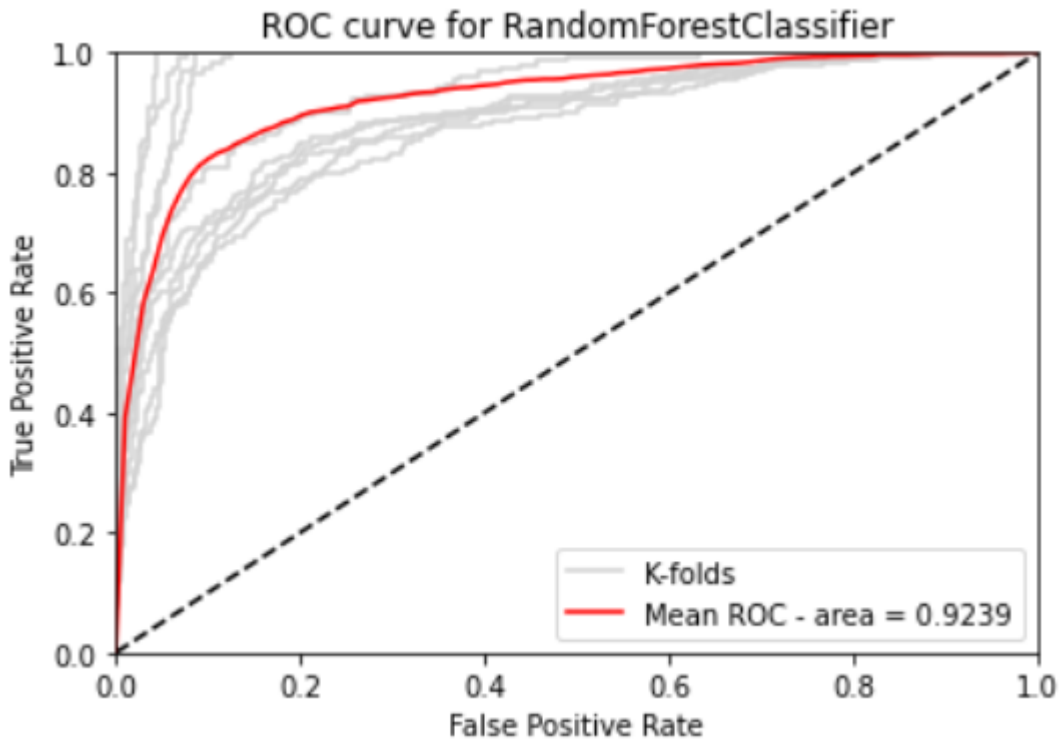
פלטים KFOLD

נספח 2.9






ROC curve for KNeighborsClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
K-folds
Mean ROC - area = 0.8402


ROC curve for RandomForestClassifier
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
K-folds
Mean ROC - area = 0.9239