Iby and Aladar Fleischman
Faculty of Engineering
Tel Aviv University

הפקולטה להנדסה
ע"ש איבי ואלדר פליישמן
אוניברסיטת תל-אביב

דו"ח מסכם פרויקט בקורס מבוא ללמידת מכונה למדעים דיגיטליים להייטק

רוני לופטין 315536912

חמדת לבל 313469710

יוני 2022

תקציר מנהלות:

במסגרת הפרויקט קיבלנו דאטה המכילה 21 פיצ'רים אודות סשנים (Sessions) של משתמשים באתר קניות באינטרנט (E-commerce), כך שכל סשן יכול להסתיים ברכישה או לא. רובם של הפיצ'רים היו בעלי משמעות אך היו מספר פיצ'רים שמשמעותם לא הייתה ידוע לנו.
מטרת הפרויקט שלנו היא לבנות מערכת המנבאת מה הסיכוי של משתמש מסוים, עם מאפיינים מסוימים, לבצע רכישה בזמן הגלישה באתר. במהלך הפרויקט נתבקשנו לבצע מספר שלבים לעיבוד הדאטה. בסופו של דבר נתבקשנו לבחור מודל שבעזרתו ננבא את ההסתברות לביצוע רכישה באתר.
מבין כל המודלים שהרצנו המודל שבחרנו הוא Random-Forest שקיבלן ציון AUC של 0.96 כאשר שני הפיצ'רים המשפיעים ביותר על זיהוי רכישות הם **'PageValues'** ו-**'D'**.

פירוט תהליך העבודה:

חלק 1 - אקספלורציה (ר' נספחים):

ראשית ייבאנו את הדאטה (את החלק של הtrain), אשר כל החלק של העיבוד המקדים יעשה עליו.
חילקנו את הפיצ'רים לשלוש רשימות (ר' נספחים): פיצ'רים נומריים, פיצ'רים קטגוריאליים, ופיצ'רים נומריים אותם נרצה להציג בסקאלת log בשלב האקספלורציה. הפיצ'רים הלא ידועים חולקו בין קבוצות אלו.
יצרנו מספר פונקציות על מנת שנוכל לעבוד עם הנתונים בצורה נוחה (פונקציות אלו לא שינו את הנתונים).
יצרנו עבור כל פיצ'ר נומרי היסטוגרמה על מנת לבחון כיצד הוא מתפלג. מבין כל הפיצ'רים, נראה כי פיצ'ר B ככל הנראה מתפלג נורמלית, מה שיועיל לנו בהמשך תהליך העיבוד המקדים.
לאחר מכן יצרנו עבור כל פיצ'ר נומרי היסטוגרמה ועבור כל פיצ'ר קטגוריאלי דיאגרמת עמודות - לערכים בעלי אותו לייבל (purchase = 0 או purchase= 1), על מנת לנסות להבין מהם הפיצ'רים שיכולים להיות משמעותיים. זיהינו זאת ע"י חוסר תאימות בהתפלגות הלייבל. הפיצ'רים שסימנו כמשמעותיים הם: **'closeness_to_holiday' 'D' 'BounceRates','ExitRates',**.
בנוסף יצרנו תרשימי עוגה עבור חלק מהפיצ'רים הקטגוריאליים כדי לבחון את הפרופורציה של כל קטגוריה.
משני סוגי הגרפים האלו זיהינו עבור הפיצ'רים **'A'** ו-**'device'** שחלק מהקטגוריות בהם מקבלות מספר תצפיות נמוך מאוד ולכן החלטנו לאחד חלק מהקטגוריות בפיצ'רים אלו על מנת שיוכלו לתרום לנו יותר מידע.

חלק 2 - עיבוד מקדים:

טיפול באאוטליירים:

'אאוטליירים' - נתונים חריגים, יכולים להוות בעיה רצינית עבור מודלים של למידת מכונה, הם משפיעים על המודל ויכולים לגרום לו לחזות דברים לא מדויקים. ראשית, השתמשנו ב-Box-plot כדי לבחון האם קיימים אאוטליירים בדאטה, וראינו שרוב הפיצ'רים מכילים נתונים חריגים. שנית, השתמשנו ב-Scatter-plot על מנת לראות באופן ויזואלי בכל פיצ'ר עד איזה ערך רוב הנתונים מקובצים ומהיכן הנתונים "מתחילים" להיות חריגים (ר' נספחים).

החלטנו להסיר את הערכים החריגים ע"פ הדרכים הבאות:

1. עבור הפיצ'ר **'B'** המתפלג נורמלית (אומת ע"י qq-plot) (ר' נספחים): הסרנו ערכים שסוטים ממרחק של 3 סטיות תקן מהממוצע.
2. עבור שאר הפיצ'רים שאינם מתפלגים נורמלית: בתחילה זיהינו "לפי העין" עבור כל פיצ'ר מהו הערך ממנו הנתונים מתחילים להיות חריגים והסרנו את כל התצפיות שהיו בעלות ערכים הגדולים מערך זה. בשלב מאוחר יותר הבנו שדרך זו אינה מדויקת ולכן בדקנו מהו ערך האחוזון ה-98% (הערך העליון שמחזיק 98% מהתצפיות) והסרנו את כל התצפיות בעלות ערכים הגדולים מערך זה (סה"כ 938 תצפיות הוסרו).

לבסוף בשלב הרצת המודלים חשדנו כי להסרת תצפיות רבות עלולה להיות השפעה שלילית על תוצאות המודלים ולכן שבנו לחלק זה והגדלנו את האחוזון ל-99.5% כדי לאבד פחות תצפיות (סה"כ 260 תצפיות הוסרו).

נורמליזציה:

הפיצ'רים בדאטה אינם מנורמלים, הם מסודרים על סקאלות שונות ובעלי טווחים שונים. הנרמול הוא שלב הכרחי מכיוון שישנם מודלים הרגישים לכך, לכן כדי למקסם את תוצאות המודלים - יש צורך לנרמל אותם(להמיר את כל הערכים להיות בין 0 ל1). עבור פיצ'ר 'B' המתפלג נורמלית השתמשנו ב-StandartScalre כדי שיהיה לנו ממוצע - 0 ואת אותה שונות וסטיית תקן - 1. עבור שאר הפיצ'רים השתמשנו ב-MinMaxScaler.

ערכים חסרים:

לא נוכל להפעיל מודלים על דאטה המכילה ערכים חסרים ולכן זהו שלב חשוב בעיבוד המקדים. בנינו טבלה המתארת את מספר ואחוז הערכים החסרים עבור כל פיצ'ר (ר' נספחים), ראינו שרוב הערכים החסרים של הפיצ'רים נעו בין 1% ל7% מה שהדגיש לנו את החריגות של 2 פיצ'רים:

1. 'D' – 98% ערכים חסרים: בתחילה הסקנו שעקב האחוז הגבוה של ערכיו החסרים תרומתו של פיצ'ר זה תהיה אפסית וכן מילוי ערכיו החסרים יכול להביא לתוצאות לא מדויקות, ולכן החלטנו להסירו. בהמשך לאחר שראינו שביצועי המודלים שלנו אינם טובים, שבנו לשלב זה והחלטנו לפעול בצורה שונה ולהשלים את ערכיו החסרים באמצעות Knn-Imputer אשר משלים את הערכים החסרים לפי ממוצע משוקלל של חמשת השכנים הקרובים ביותר.
2. 'total duration' – 45% ערכים חסרים: החלטנו לא להשלים את ערכיו החסרים ולבחון האם ישנה קורלציה גבוהה בינו לבין פיצ'רים אחרים או בינו לבין הלייבל, על מנת לקבל החלטה מושכלת יותר כיצד נכון לטפל בו.

שאר הפיצ'רים טופלו ע"פ הדרכים הבאות:
- עבור פיצ'רים בעלי פחות מ30 ערכים חסרים, הסרנו את התצפיות בעלות הערכים החסרים.
- פיצ'רים בעלי ערכים נומריים – ערכים חסרים הוחלפו בערך החציוני.
- פיצ'רים בעלי ערכים קטגוריאליים – ערכים חסרים הוחלפו בערך הנפוץ ביותר.
- בפיצ'רים 'A' ו-'device' – ערכים חסרים הומרו ל"other".

התמודדות עם משתנים קטגוריאליים:

מודלי למידת מכונה לא יכולים להתמודד עם משתנים קטגוריאליים אלא אם הופכים אותם למשתנים נומריים.
הדאטה מכילה פיצ'רים קטגוריאליים משלושה סוגים: ערכים קטגוריאליים המיוצגים ע"י מספרים (למשל 'device'), מילים (למשל 'user_type') וערך בוליאני (למשל 'Weekend').
בתחילה לא טיפלנו בפיצ'רים מהסוג הראשון מתוך מחשבה שהמודלים יודעים להתמודד עם ערכים מספריים. בשלב מאוחר יותר הבנו שקטגוריות המיוצגות ע"י מספרים גדולים יותר ייתפסו כחשובות יותר עבור חלק מהמודלים וזוהי טעות כי לא בהכרח ישנה חשיבות לסדר בין הקטגוריות.

לבסוף טיפלנו בפיצ'רים הקטגוריאליים בדרך הבאה:

1. הפיצ'ר 'closeness_to_holiday' נשאר ללא שינוי מאחר וישנה חשיבות לסדר הקטגוריות בו.
2. עבור הפיצ'ר 'Weekend' השתמשנו בEncoding – הקצאת מספר עבור כל קטגוריה.
3. עבור שאר הפיצ'רים הקטגוריאליים השתמשנו בDummy Variable – כל קטגוריה בתוך הפיצ'ר הופכת לעמודה והמשתנה מקבל 1 אם הערך שלו הוא הקטגוריה ו 0 אם לא.

יצירת פיצ'רים חדשים:

הרעיון המרכזי שהנחה אותנו בשלב זה הוא יצירה של פיצ'רים שנראים לנו מעניינים ויתרמו לנו מידע נוסף, תוך ידיעה שלא בהכרח נשתמש בהם במודל הסופי בעקבות קורלציה נמוכה או כי לא יהיו משמעותיים מספיק.
הפיצ'רים החדשים:

1. 'Is_close': מקבל את הערך 1 אם הערך של 'closeness to holiday' גדול או שווה ל0.8 ו-0 אחרת.
2. 'Mean_dur': סוכם את זמן השהייה הממוצע בעמוד עבור כל סוגי העמודים.
3. 'Total_pages': סך העמודים שהמשתמש ביקר בהם בסשן.

לאחר מספר נסיונות ראינו שהפיצ'רים החדשים לא תורמים לביצועי המודל והחלטנו לא להשתמש בהם במודל הסופי (ר' נספחים).

ממדיות הבעיה:

ניכר כי ממדיות הבעיה היא גדולה מידי , ממדיות גדולה עלולה ליצור בעיית overfitting. בהתחלה, ככל שיהיו לנו יותר פיצ'רים כך ביצועי המודל יעלו, אך בשלב מסוים נגיע ל-overfitting וביצועי המודל ירדו.
על מנת לזהות את הממדיות של הבעיה גדולה מידי יצרנו טבלת קורלציה (ר' נספחים), קורלציה גבוהה בין שני פיצ'רים מעידה על פיצ'רים 'מתואמים' שמספקים מידע דומה ולכן ניתן להוריד אחד מהם.
בשלב זה, חזרנו לפיצ'ר 'total duration', וראינו שיש לו קורלציה גבוהה (0.96) עם הפיצ'ר 'product page duration', ובהתחשב באחוז הגבוה של ערכיו החסרים ובעובדה שיש לו קורלציה נמוכה יותר עם 'purchase', החלטנו להסירו.
בנוסף ראינו שיש קורלציה גבוהה (0.91) בין הפיצ'רים 'Exit_Rates' ל-'Bounce_Rates' והחלטנו להסיר את 'Bounce_Rates' מכיוון שיש לו קורלציה נמוכה יותר עם הלייבל.

לאחר מכן ביצענו מודל PCA - מודל זה מרכיב את קבוצת הפיצ'רים מחדש על ידי קומבינציה לינארית תוך שמירה על אחוז מוגדר מהשונות המוסברת. בהתחלה כאשר הפעלנו אותו על כלל הפיצ'רים כך שישמור על 99% מהשונות המוסברת, ראינו שביצועי המודלים נמוכים והחלטנו לנקוט בגישה שונה – הפעלנו את המודל בכל פעם בנפרד על תת קבוצה של משתני דמי שהגיעו מאותו הפיצ'ר כך שישמור על 95% מהשונות המוסברת.
(כאשר הגדרנו למודל לשמור על 99% מהשונות המוסברת עבור כל תת קבוצה, ראינו כי המודל לא מקטין את הממדיות ולכן הקטנו את אחוז השונות המוסברת) ובסה"כ נשארנו עם 50 פיצ'רים (ר' נספחים).
יש לציין שהפיצ'רים החדשים שהמודל הרכיב אינם אינפורמטיביים עבור הלקוח לעומת פיצ'רים עם שמות ומשמעות וזהו חיסרון בשימוש במודל זה אך העובדה שביצענו אותו בכל פעם על קבוצת פיצ'רים שקשורים לאותו משתנה מקורי כן מאפשרת לשמור על חלק מהמידע.

חלק 3+4 – הרצת והערכת המודלים:

על מנת לקבוע את ההיפר-פרמטרים עבור המודלים השונים נעזרנו בשתי פונקציות: grid-search ו-random-search (אשר מוצאות את סט הפרמטרים האופטימלי מתוך כלל האפשרויות שניתנו להן או באופן רנדומלי או ע"י מעבר על כלל השילובים האפשריים). בנוסף עבור חלק מהמודלים בחנו את ההשפעה של ערכים שונים של פרמטר כלשהו על ציוני המודל באמצעות גרף על מנת לקבוע מהו הערך האופטימלי.

מודלים בסיסיים:

1. **KNN**: בנינו תרשים ההשפעה של מספר השכנים על ה-AUC של ה-train וה-valid (ר' נספחים).
בחרנו 151 שכנים, תחילה קראנו על כך שהקונבנציה היא שורש ריבועי של מספר התצפיות ולאחר הרצת grid-search המודל התכנס סביב 151 . מספר השכנים הוא אי זוגי כדי לקבל כלל הכרעה מדויק יותר. לא ניתנה חשיבות למרחק של השכנים על כן בחרנו uniform בפרמטר זה. בחרנו מרחק מנהטן. בסופו של דבר

קיבלנו ציון של 0.74 על ה-train ו 0.72 על ה-valid, כלומר לא נמצאים ב-overfit (ר' נספחים). מבין המודלים זהו המודל שקיבל את הציון הנמוך ביותר, ייתכן וסוג הדאטה אינו מותאם למודל זה.

2. **Logistic-Regression**: קבענו את קריטריון העצירה ל-0.01, והקטנו את C מתחת לערך ברירת המחדל ל-0.1 על מנת שתהיה יותר רגולריזציה. בסופו של דבר קיבלנו ציון 0.95 על ה-train ועל ה-valid, כלומר אנחנו לא נמצאים ב-overfit (ר' נספחים), בנוסף ניסינו להסביר את תרומתם של הפיצ'רים למודל לפי ערך מקדמיהם במשוואת הרגרסיה (ר' נספחים).

<u>מודלים מתקדמים:</u>

1. **רשת נוירונים (ANN)**: הגדלנו את מקדם ה- penalty ל-0.02 שיהיה גבוה מערך ברירת המחדל, כמו כן הגדלנו את מספר האיטרציות מערך ברירת המחדל ל-1250. השתמשנו בפונקציית אקטיבציה identity. וקצב למידה נקבע להיות adaptive. בסופו של דבר קיבלנו ציון של 0.94 על ה- train ועל ה- valid, כלומר לא נמצאים ב- overfit (ר' נספחים).

2. **Random-Forest**: אכלסנו את היער ב-100 עצים וקבענו מקסימום לעומק כל עץ 6. הגדלנו את מינימום מספר התצפיות לכל חלוקה ל-3 ואת מינימום התצפיות לעלה ל- 2. קבענו את מספר הפיצ'רים שיבואו לידי ביטוי כשורש הריבועי של סך הפיצ'רים. שילוב כלל הפרמטרים הסופי נקבע על ידי grid-search ועל ידי תרשים ההשפעה של max_depth על ה-AUC של ה-train וה-valid. התוצאות שקיבלנו הן 0.97 ל-train ו-0.96 ל-test . תחילה קיבלנו AUC של 1 ב-train, מה שזיהינו כ- overfitting. פתרנו זאת על ידי שינוי של ההיפר-פרמטרים של המודל, בעיקר תחת משחק בין ה-max_depth ל-n_estimators ועל ידי הכנת תרשים ההשפעה שתואר קודם לכן (ר' נספחים). בנוסף הראנו גרף של חשיבות הפיצ'רים(ר' נספחים).

לדעתנו התוצאות של המודל הינן אופטימליות בטרייד-אוף שבין מקסום ה-AUC של ה-valid לבין הפחתת ה-overfit. בנינו confusion matrix למודל (ר' נספחים), המטריצה ממחישה את מספר הפרדיקציות בהן סיווגנו נכון, ואת הפרדיקציות בהן טעינו בסיווג. המטריצה מחולקת ל4 משבצות במיקומים הבאים:
*שמאל למעלה - **True Positive**:* כלומר חזינו purchase וצדקנו.
*שמאל למטה - **False Negative**:* כלומר חזינו non purchase וטעינו, *זוהי טעות מסוג שני.*
*ימין למעלה - **False Positive**:* כלומר חזינו purchase וטעינו, *זוהי טעות מסוג ראשון.*
*ימין למטה - **True Negative**:* כלומר חזינו non purchase וצדקנו.

למודל שלנו יש ערך נמוך של ***False-Positive***, משמע מעטות הפעמים בהם המודל חוזה purchase וטועה. בנוסף, למודל יש ערך גבוה של ***True-Negative***, שמשמעותו הוא שהמודל מצליח לחזות בצורה נכונה non purchase – שהם המשתמשים שגלשו באתר ולא ביצעו רכישה.
אם בוחנים את כלל המקרים בהם בוצעה רכישה בפועל (real p, הצד השמאלי של המטריצה) - המודל חוזה בצורה נכונה יותר מחצי מהמקרים, אך עדיין נרצה לקבל ערך ***False-Negative*** נמוך יותר.

נבחן את המדדים הנוספים אותם ניתן להסיק מה-confusion matrix, בהינתן ה-threshold הדיפולטיבי לסיווג שהינו 50% (לסף זה ישנה משמעות כלכלית-עסקית כיוון שאם הוא יהיה נמוך אז יותר דוגמאות, גם כאלה שאינן נכונות - יסווגו כ-purchase, ואם הסף יהיה גבוה אז הרבה דוגמאות שהיו צריכות להיות מסווגות כ-purchase 'יתפספסו'. ולכן סף בגובה 50% אינו בוודאות מתאים למקרה זה):
למודל שלנו יש שיעור **specificity** גבוה לעומת ה-sensitivity, כלומר, אנחנו מכסים יותר ***True-Negative*** מאשר ***True-Positive***, ויהיו לנו פחות "אזעקות שווא" כלומר המודל שלנו יחזה יותר טוב לקוחות שגלשו באתר ולא ביצעו רכישה. בנוסף למודל שלנו שיעור precision גבוה שאומר שברוב הפעמים שהמודל חוזה

רכישה הוא צודק. בדקנו את ה-confusion matrix של המודל שנבחר לעומת המטריצות המתקבלות ממודלים אחרים, שחלקם היו בפחות overfitting (למשל Logistic-Regression) וראינו שה-**accuracy** וכמו כן גם ה-**precision** במודל הנבחר גבוהים יותר. עובדה זו חיזקה בעינינו את הבחירה במודל זה.
הרצנו k-fold cross validation על כלל המודלים, בחרנו K=5 שזו הבחירה הסטנדרטית. עבור כך, איחדנו את הדאטה של ה-train וה-valid משום שk-fold עושה חלוקה פנימית בצורה עצמאית. ברובם קיבלנו גרף דיי אחיד עבור כל ה- folds, מה שמעיד על סטיית תקן נמוכה.

<u>סיכום:</u>
ניכר שהמודל שלנו יכול להבטיח סיכויי חיזוי טובים ולעזור לאתר להגדיל את רווחיו ע"י זיהוי משתמשים שביקרו באתר ולא ביצעו רכישה. בממוצע 92 מתוך 100 לייבלים יהיו נכונים. במודל שלנו שיעור ה-specificity גבוה, אך זה בא על חשבון ה-sensitivity של המודל, בממוצע רק 53 מתוך 100 רכישות יזוהו מראש ככאלו. אם נתייחס לtrade-off בין שני מדדים אלו, לדעתנו במקרה זה מדד ה-specificity הוא בעל חשיבות כלכלית רבה יותר, מאחר וכדאי לאתר לזהות משתמשים שגלשו בו ולא ביצעו רכישה וזאת על מנת לאפשר לו "לטרגט" אותם באמצעות פרסומות/מיילים וכו' ולגרום להם לשוב לאתר, לבצע את הרכישה ובכך להגדיל את רווחיו.

חלוקת עבודה:

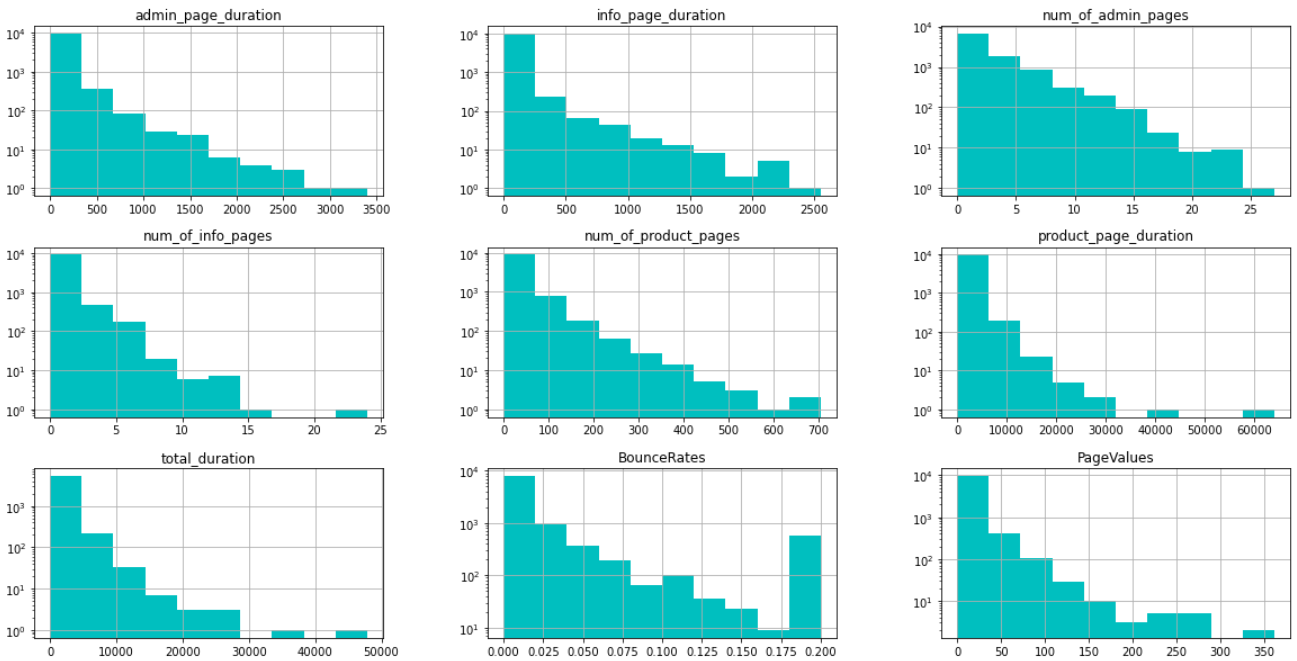
לאורך **כל** שעות העבודה על הפרוייקט, שתינו היינו נוכחות ועבדנו ביחד או במקביל.

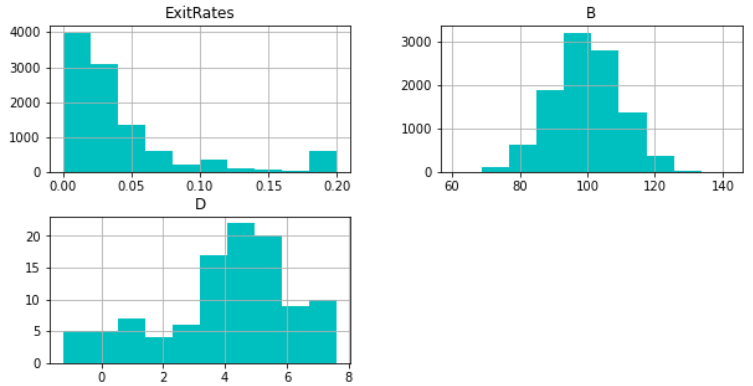
נספחים:

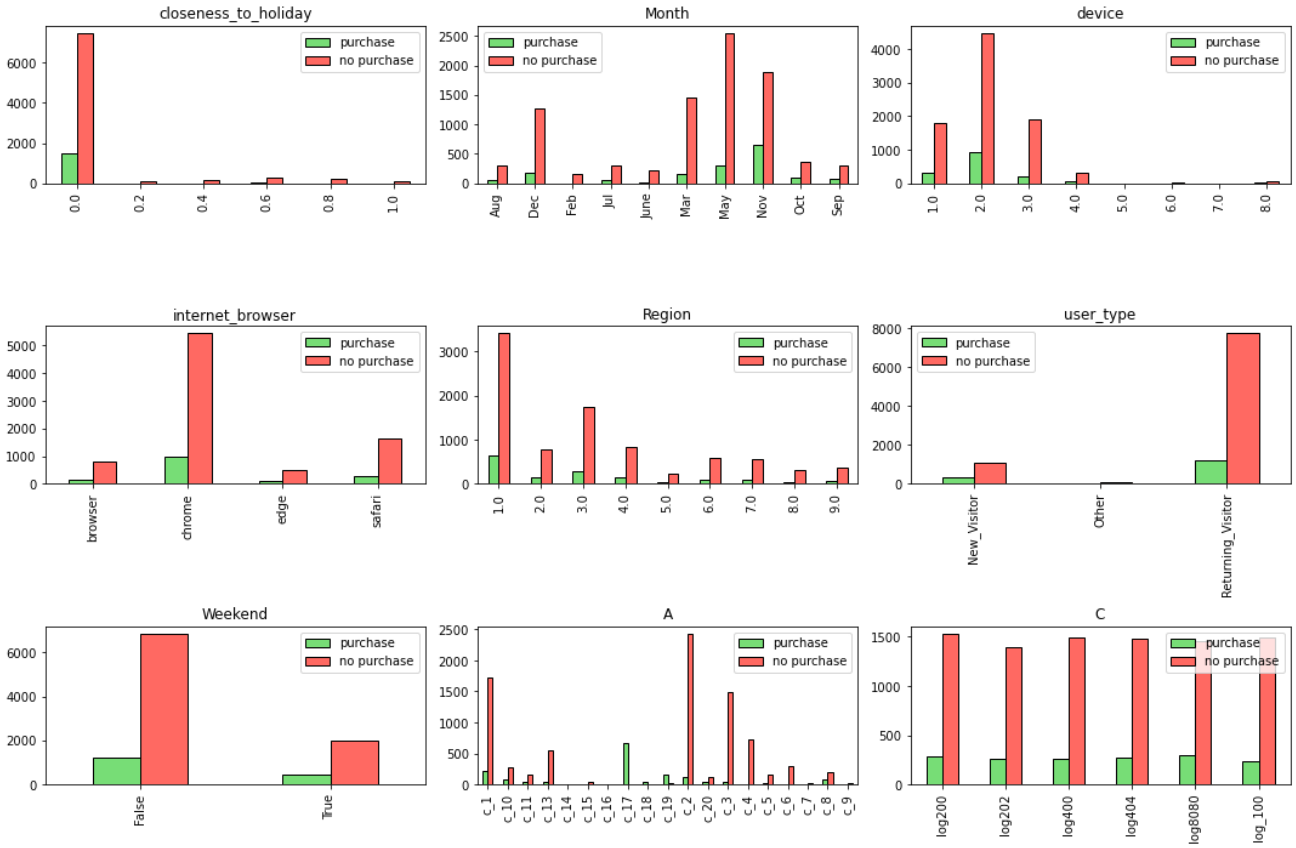
קבוצות הפיצ'רים השונות:

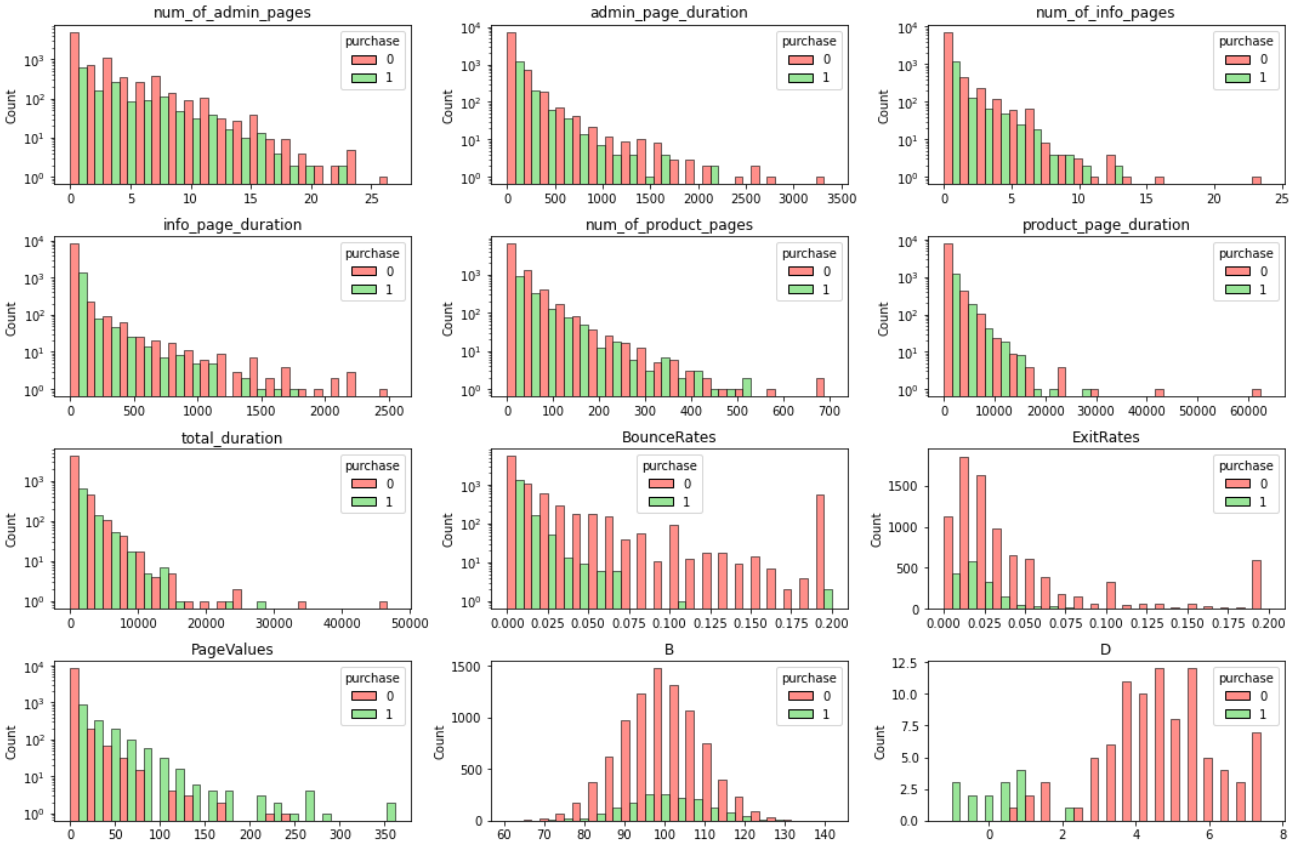
Numeric Features	Categorical Features	Log List
'num_of_admin_pages'	'closeness_to_holiday'	'admin_page_duration'
'admin_page_duration',	'Month'	'info_page_duration'
'num_of_info_pages'	'device'	'num_of_admin_pages'
'info_page_duration'	'internet_browser'	'num_of_info_pages'
'num_of_product_pages'	'Region'	'num_of_product_pages'
'product_page_duration'	,'user_type'	'product_page_duration'
'total_duration'	'Weekend'	'total_duration'
'BounceRates'	'A'	'BounceRates'
'ExitRates'	'C'	'PageValues'
'PageValues'		
'B'		
'D'		

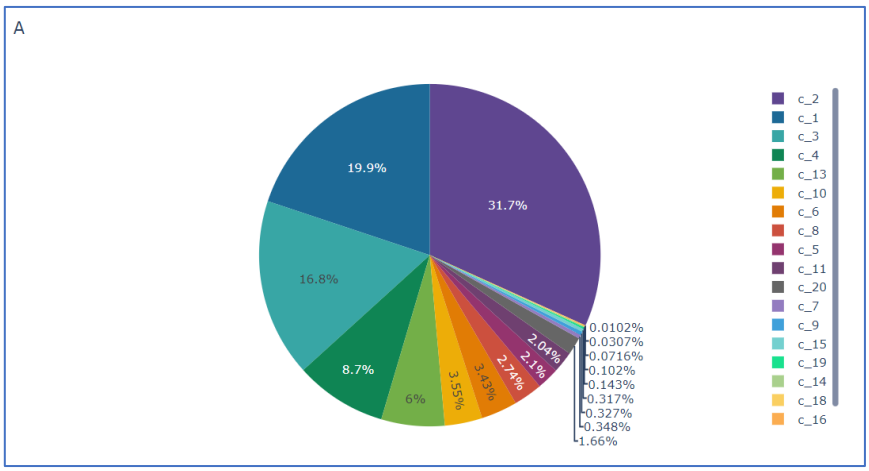
אקספלורציה:

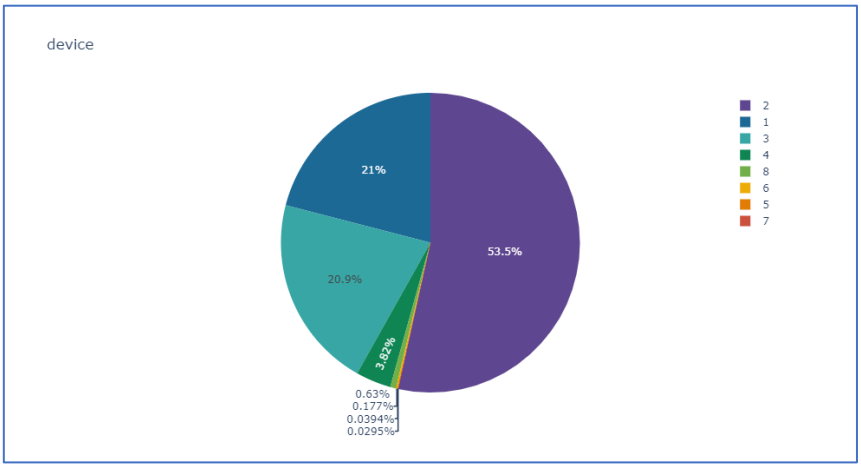


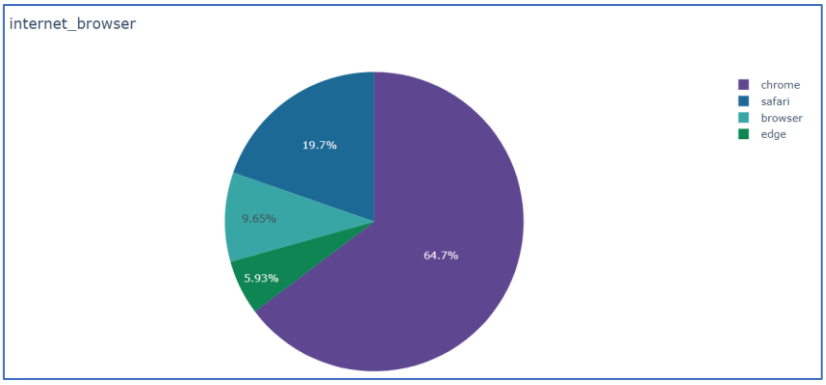

ExitRates
B
D

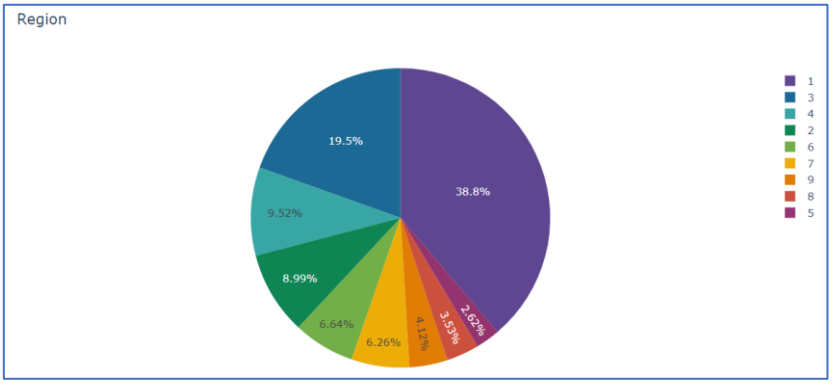

closeness_to_holiday
Month
device
internet_browser
Region
user_type
Weekend
A
C
purchase
no purchase
browser
chrome
edge
safari
New_Visitor
Other
Returning_Visitor
False
True
log200
log202
log400
log404
log8080
log_100

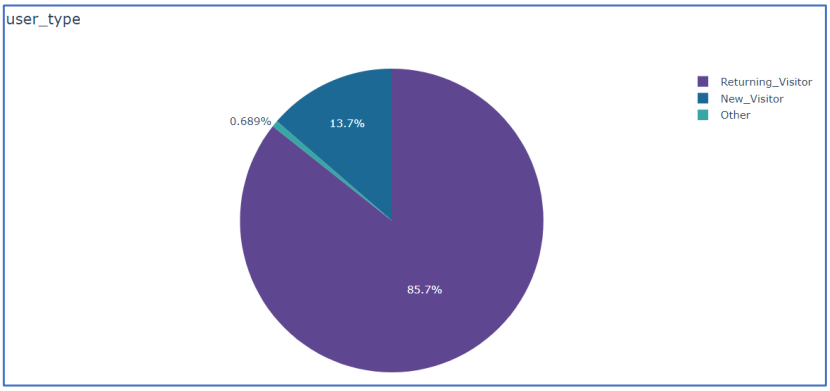

num_of_admin_pages
admin_page_duration
num_of_info_pages
info_page_duration
num_of_product_pages
product_page_duration
total_duration
BounceRates
ExitRates
PageValues
B
D
purchase
0
1
Count

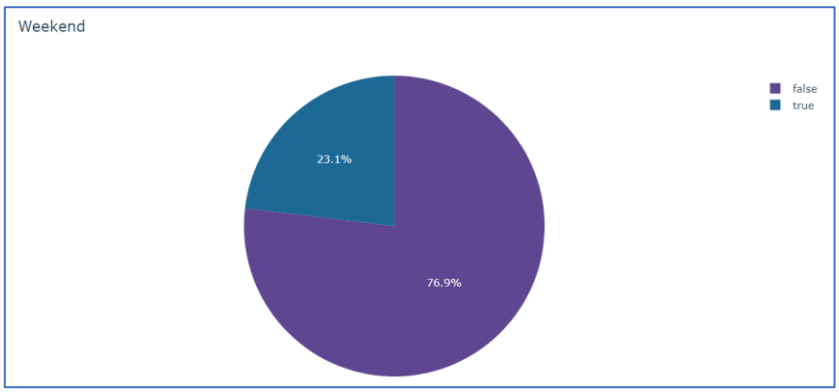

A
31.7%
19.9%
16.8%
8.7%
6%
3.55%
3.43%
2.74%
2.1%
2.04%
0.0102%
0.0307%
0.0716%
0.102%
0.143%
0.317%
0.327%
0.348%
1.66%
c_2
c_1
c_3
c_4
c_13
c_10
c_6
c_8
c_5
c_11
c_20
c_7
c_9
c_15
c_19
c_14
c_18
c_16


device
53.5%
21%
20.9%
3.82%
0.63%
0.177%
0.0394%
0.0295%
2
1
3
4
8
6
5
7


internet_browser
19.7%
9.65%
5.93%
64.7%
chrome
safari
browser
edge

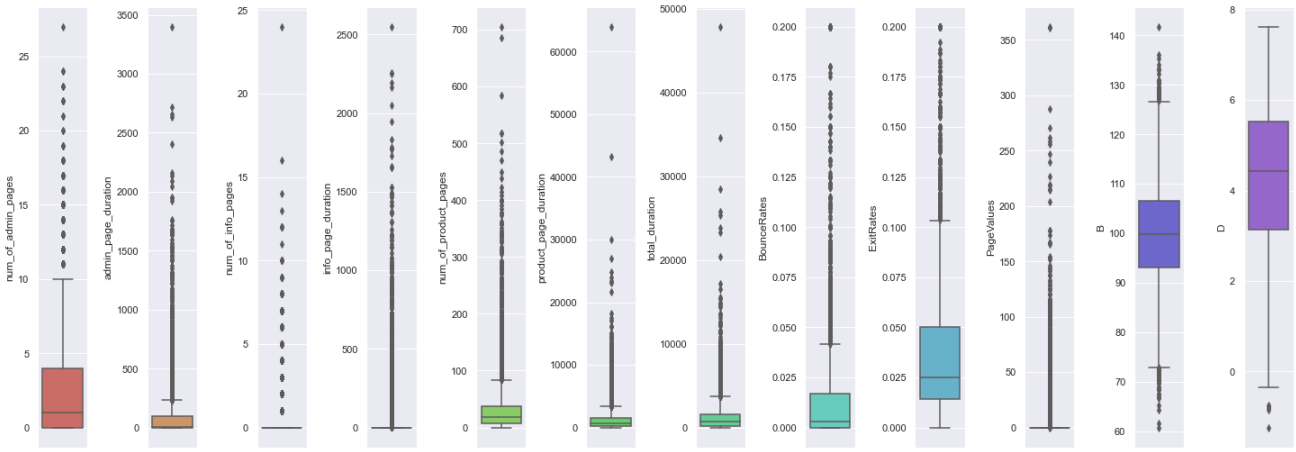

Region
19.5%
9.52%
8.99%
6.64%
6.26%
4.12%
3.53%
2.62%
38.8%
1
3
4
2
6
7
9
8
5

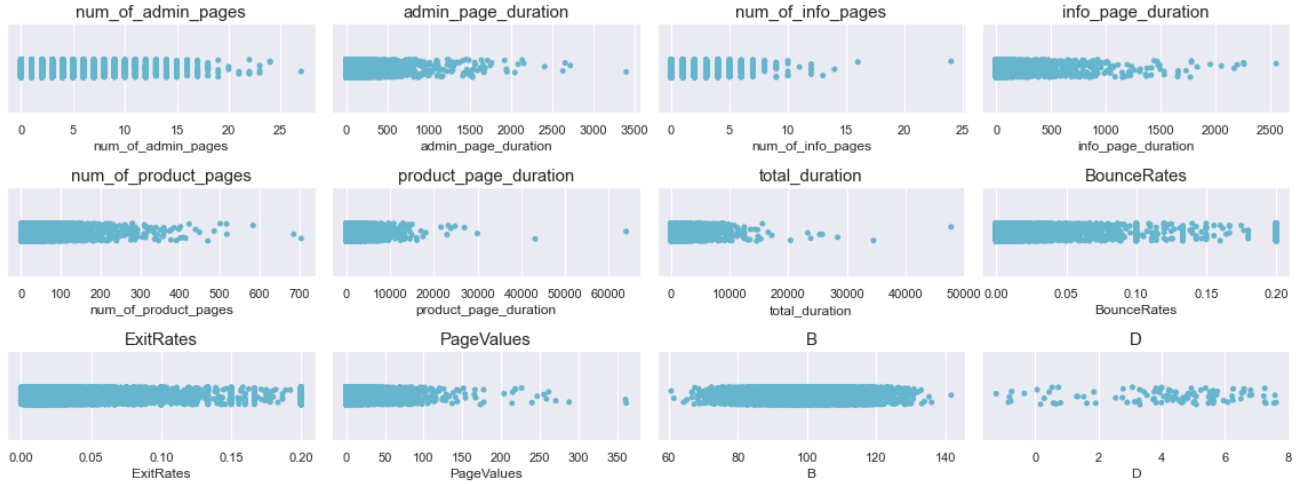

user_type
0.689%
13.7%
85.7%
Returning_Visitor
New_Visitor
Other

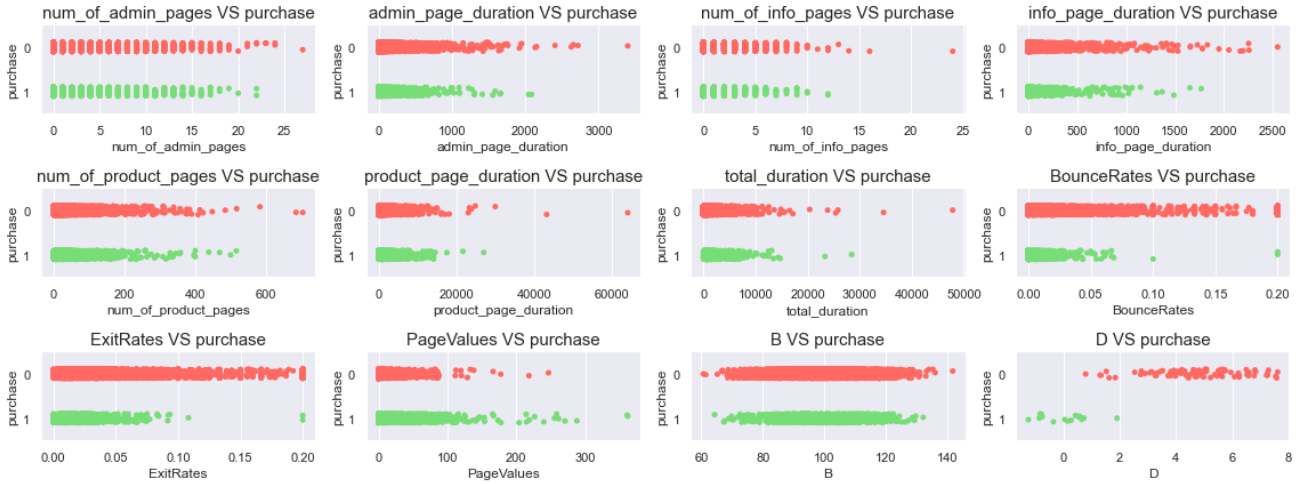

Weekend
23.1%
76.9%
false
true

עיבוד מקדים:

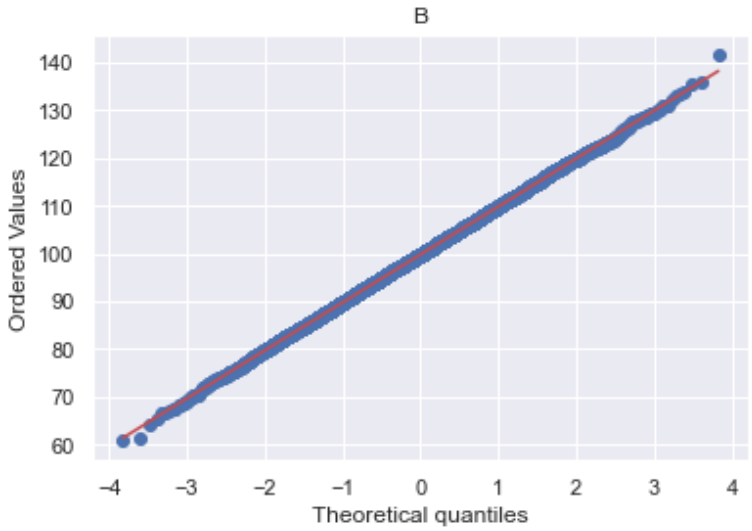
טיפול בערכים חריגים - ויזואליזציה של אאוטליירים:







בדיקת התפלגות נורמלית qq-plot:



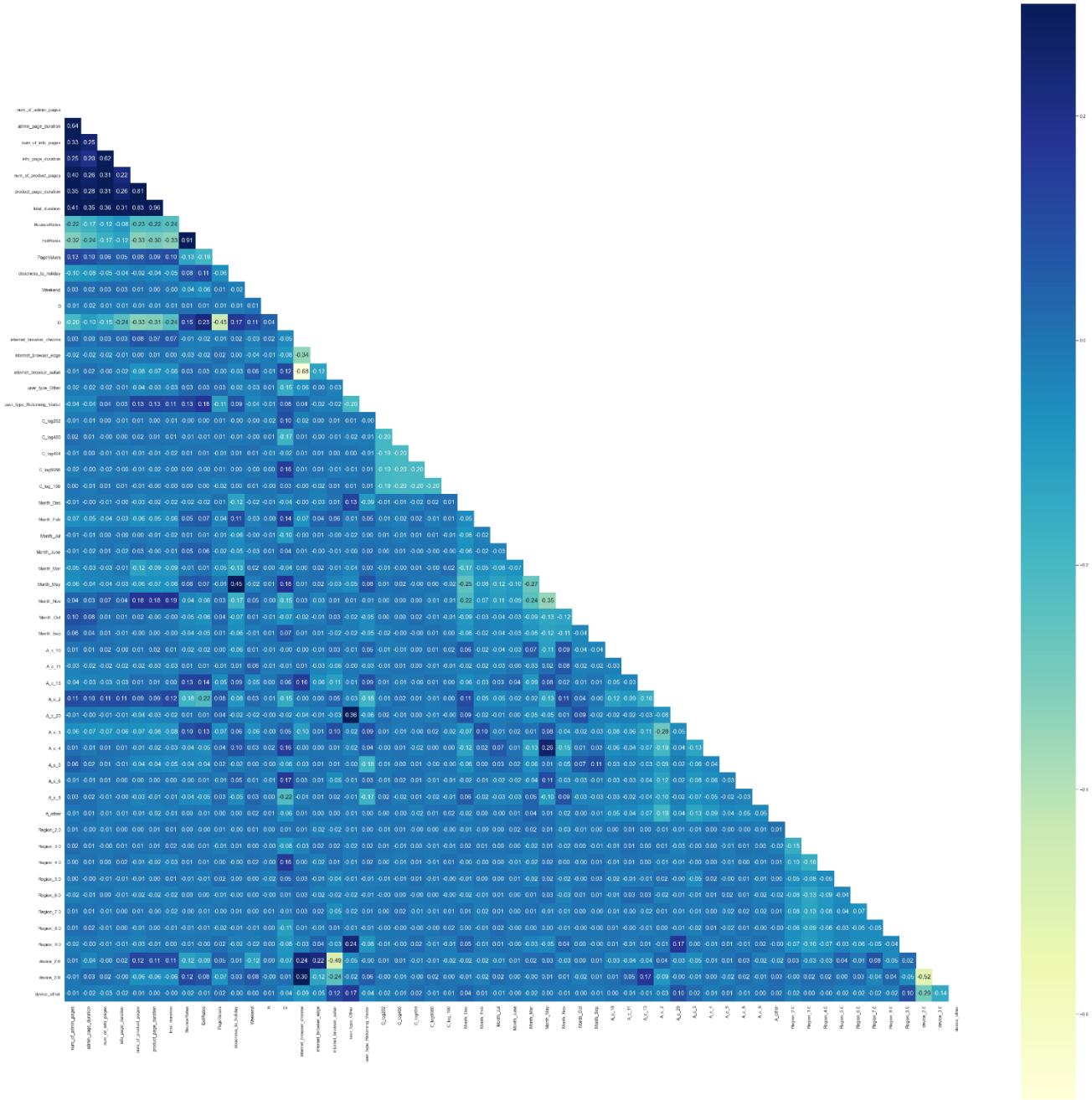
טיפול בערכים חסרים - טבלת פיצ'רים לפי אחוזי ערכים חסרים:

	nan	nan percent
D	10086	98.969679
total_duration	4612	45.255618
A	689	6.760867
num_of_info_pages	662	6.495928
product_page_duration	601	5.897360
num_of_admin_pages	585	5.740359
internet_browser	543	5.328231
closeness_to_holiday	487	4.778726
admin_page_duration	408	4.003533
num_of_product_pages	395	3.875969
device	316	3.100775
info_page_duration	306	3.002649
PageValues	27	0.264940
ExitRates	25	0.245314
Month	24	0.235502
user_type	23	0.225689
Weekend	23	0.225689
B	23	0.225689
C	23	0.225689
BounceRates	21	0.206064
Region	19	0.186439
id	0	0.000000
purchase	0	0.000000

יצירת פיצ'רים חדשים: השוואת ביצועי מודלים:

Random Forest – max_deapth = 6					
K fold mean auc	Valid auc	Train auc	סה"כ פיצ'רים	פיצ'רים חדשים	PCA
0.9555	0.96044	0.97137	53	בלי	בלי
0.9470	0.94497	0.96710	56	עם	
0.9600	0.96239	0.97577	50	בלי	עם
0.9503	0.95023	0.97119	53	עם	

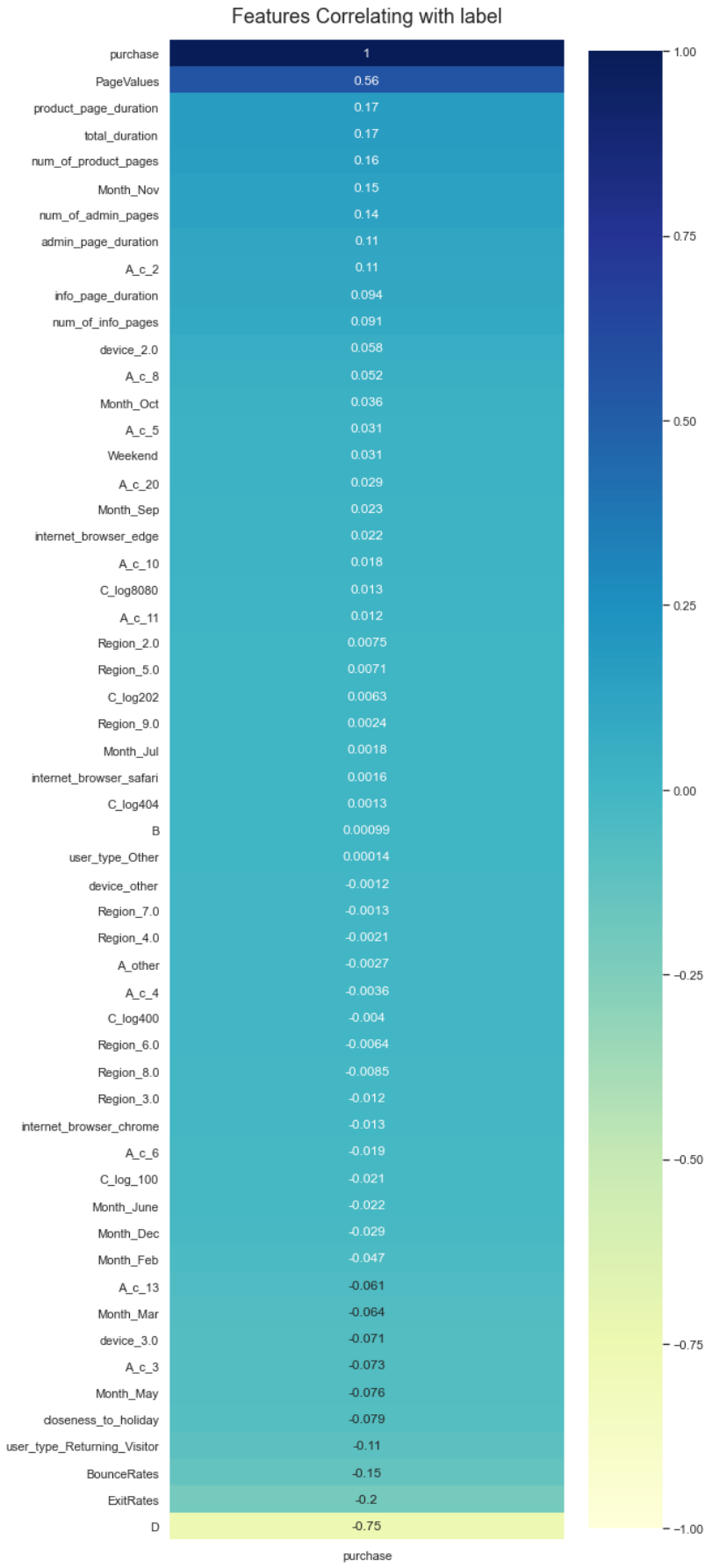
הקטנת ממדיות - טבלאות קורלציה:



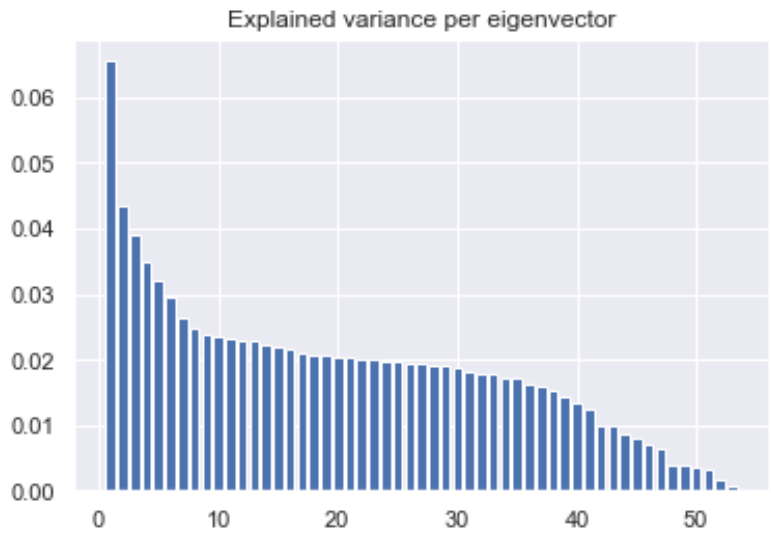
קורלציות בין כל הפיצ'רים:

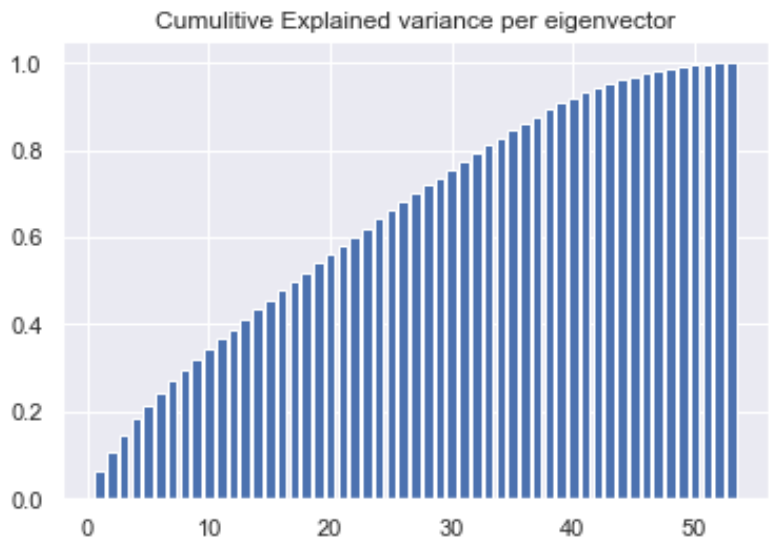
		Correlation
product_page_duration	total_duration	0.96
BounceRates	ExitRates	0.91
num_of_product_pages	total_duration	0.83
	product_page_duration	0.81
internet_browser_chrome	internet_browser_safari	0.68
num_of_admin_pages	admin_page_duration	0.64
num_of_info_pages	info_page_duration	0.62
device_2.0	device_3.0	0.52
internet_browser_safari	device_2.0	0.49
closeness_to_holiday	Month_May	0.45
PageValues	D	0.45
num_of_admin_pages	total_duration	0.41
	num_of_product_pages	0.40
num_of_info_pages	total_duration	0.36
user_type_Other	A_c_20	0.36

קורלציה של הפיצ'רים עם הלייבל:



תוצאות PCA:

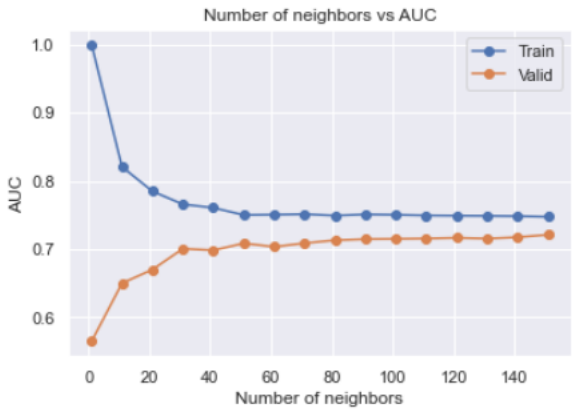

Explained variance per eigenvector
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40
50

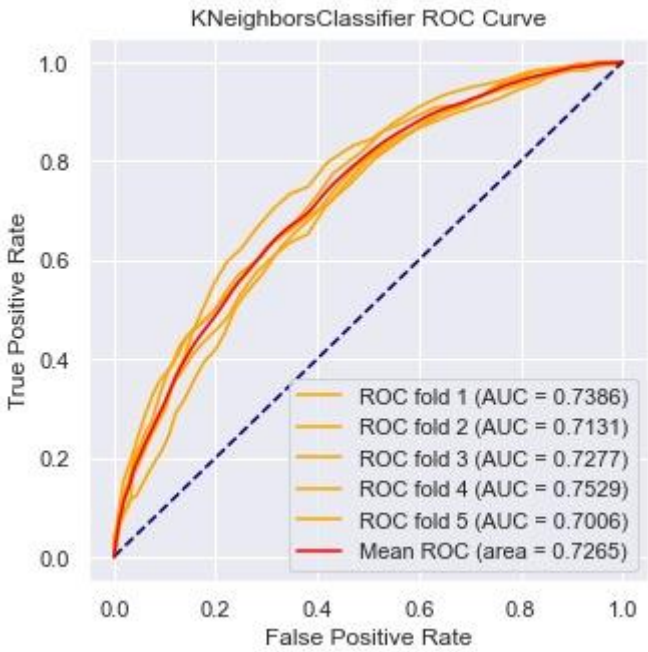

Cumulitive Explained variance per eigenvector
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50

הרצת מודלים - k-fold cross validation:

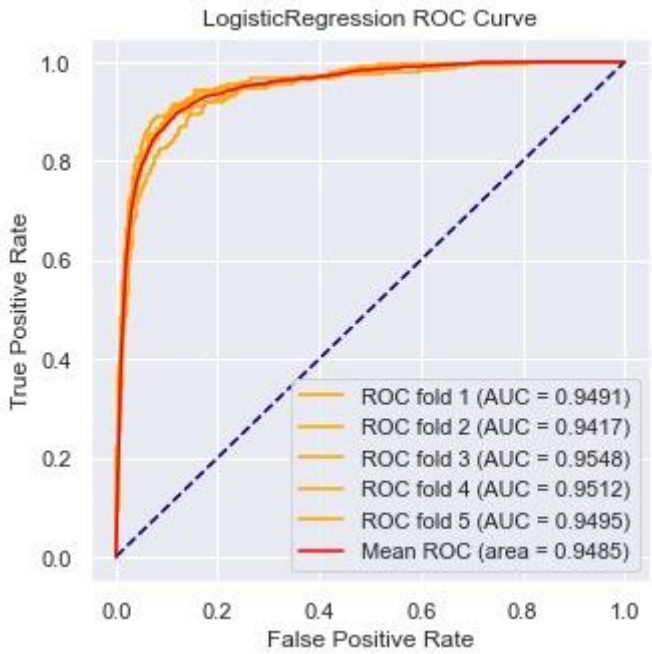
KNN:

```
>1, train: 1.000, valid: 0.565
>11, train: 0.821, valid: 0.650
>21, train: 0.785, valid: 0.670
>31, train: 0.766, valid: 0.701
>41, train: 0.761, valid: 0.698
>51, train: 0.750, valid: 0.708
>61, train: 0.750, valid: 0.703
>71, train: 0.751, valid: 0.709
>81, train: 0.749, valid: 0.713
>91, train: 0.751, valid: 0.715
>101, train: 0.750, valid: 0.715
>111, train: 0.749, valid: 0.716
>121, train: 0.749, valid: 0.717
>131, train: 0.749, valid: 0.715
>141, train: 0.748, valid: 0.717
>151, train: 0.747, valid: 0.721
```

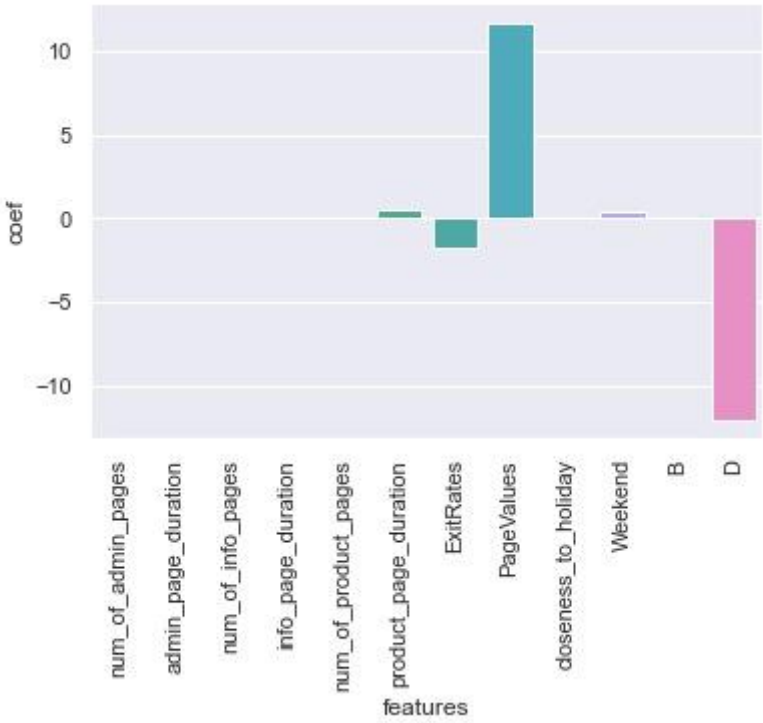




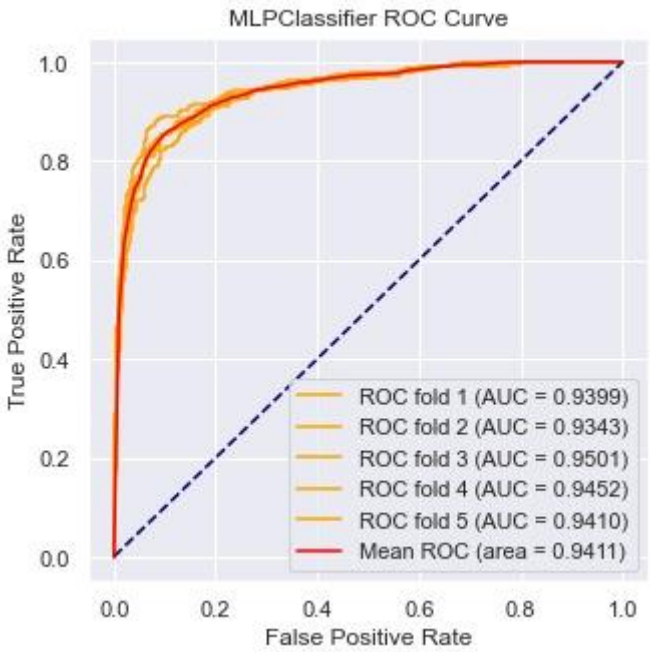
:Logistic-Regression



חשיבות הפיצ'רים לפי מודל Logistic-Regression:



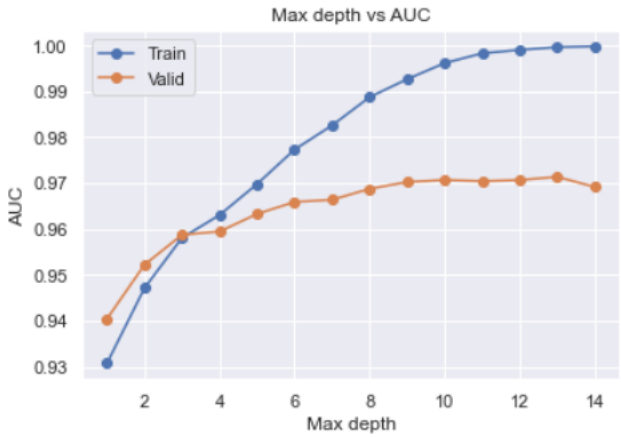
רשת נוירונים (ANN):



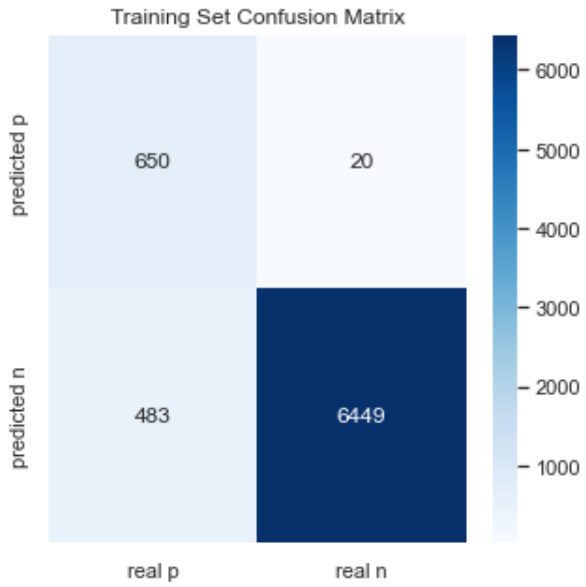
Random-Forest:

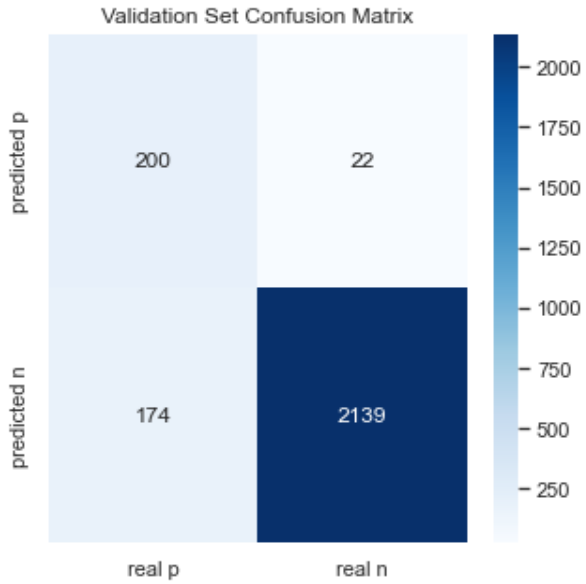
בחירת היפר פרמטרים למודל – Random Forest גרף המתאר את max_depth:

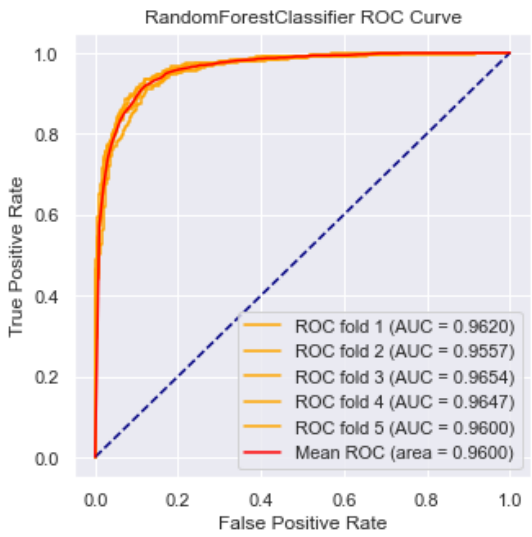
```
1 -  train: 0.931, valid: 0.940
2 -  train: 0.947, valid: 0.952
3 -  train: 0.958, valid: 0.959
4 -  train: 0.963, valid: 0.959
5 -  train: 0.970, valid: 0.963
6 -  train: 0.977, valid: 0.966
7 -  train: 0.983, valid: 0.966
8 -  train: 0.989, valid: 0.969
9 -  train: 0.993, valid: 0.970
10 -  train: 0.996, valid: 0.971
11 -  train: 0.998, valid: 0.970
12 -  train: 0.999, valid: 0.971
13 -  train: 1.000, valid: 0.971
14 -  train: 1.000, valid: 0.969
```



Confusion matrix עבור Random Forest:







חשיבות הפיצ'רים לפי מודל Random Forest: