הפקולטה להנדסה
ע"ש איבי ואלדר פליישמן
אוניברסיטת תל אביב

דו"ח מסכם - פרויקט חיזוי רכישות באתר אינטרנט

קורס 0560182401 מבוא ללמידת מכונה

מדעים דיגיטליים להייטק

צוות הקורס:

מרצה: מר דור בנק

מתרגל: מר אילן וסילבסקי

מגישה: מור שנבסקי

תקציר מנהלים

בפרויקט זה נדרשתי לחזות את קיומה של רכישה ע"י משתמש בסיום סשן באתר אינטרנט מסוים (e-commerce) באמצעות משתנים כמו זמן הגלישה באתר לפי סוגי הדפים, שיעורי הנטישה והיציאה, ערך ממוצע לדף וכן משתני זמן ומאפיינים של משתמשים. כצעד ראשון והחלק העיקרי בפרוייקט, ביצוע חקירה ועיבוד של הנתונים, ההתפלגות וההתנהגויות שלהם וכן מתאם עם משתנים נוספים בסט. במסגרת העיבוד השתמשתי במתודות של הסרת ערכים חריגים, נרמול נתונים (באמצעות מיצוע ושימוש בנורמליזציה MinMax), בנית פיצ'רים חדשים לביטוי קשרים בין משתנים וקידוד של משתנים קטגוריאליים והפיכתם למשתני דמי. השתמשתי בטכניקות שונות (PCA, filter method) להורדת המימדיות של סט הנתונים עד להחלטה על סט הנתונים אשר מקסם את תוצאות המודלים מבלי להסתכן בoverfitting. לאחר ביצוע העיבוד ונרמול סופי של הסט בחרתי בארבעה מודלים להרצה לשם ביצוע חיזוי וקלסיפיקציה, תוך בחירת היפר-פרמטרים בטכניקת GridSearch. הערכת המודלים נעשתה בטכניקת Confusion Matrix וכן ולידציה ע"י K-Fold Cross Validation מתוך מטרה להגיע לתוצאות חיזוי הגבוהות ביותר. המודל הנבחר שימש לאחר מכן לביצוע תחזית על סט test ותוצאותיה נשמרו בהתאם. לבסוף גיבשתי כמה המלצות מעשיות להגדלת הסיכוי לביצוע רכישה באתר והמלצות לכיווני חקירה נוספים.

אקספלורציה

סט הנתונים מכיל קובץ ובו נתוני אימון (train) ובו 10,479 סשנים של משתמשים באתר קניות באינטרנט. לכל סשן יש 22 פיצ'רים, המכילים מידע אודות מאפייני הגלישה והמשתמש (פירוט הפיצ'רים וחלוקה לתתי קבוצות בטבלה 2.1 בנספח). כל סשן מסתיים בהחלט של המשתמש האם לבצע רכישה (1) או שלא (-0), וזהו הlabel שישמש לאימון המודלים בטרם ביצוע פרדיקציה על סט המבחן (test). רוב הפיצ'רים מכילים את כל הנתונים והערכים החסרים בהם אינם עולים על 7% מסך הנתונים. העמודות היחידות בהן חסרה כמות משמעותית של מידע הן total duration שאותו אשלים בהמשך באמצעות שאר העמודים ועמודה D החסרה כ-99% מהנתונים. שני משתנים (זמן לעמודי אינפו ולעמודי מוצר) היו מסוג object שלא תאם את המהות שלהם (מספר) וכן מחרוזת minutes מה שהקשה על הניתוח ולכן הפכתי אותם לסוג המתאים (float) והסרתי את המחרוזת המיותרת.

תרשים 1.1 בנספח מציג את ההתפלגות בין משתמשים שביצעו רכישה לבין אלו שבחרו שלא. ניתן לראות כי מדובר ביחס של 15%-85% ועל כן מדובר בסט נתונים שאינו מאוזן. בכדי להתגבר על כך אבצע חלוקה של הנתונים לאימון וולידציה בעת הרצת המודלים לחיזוי. ניתן לראות כי עבור משתמשים שכן ביצעו רכישה, פרקי הזמן בהם שהו באתר, כמו גם מספר העמודים שביקורו וערך הדפים גבוהים משמעותית לעומת אלו שלא. העמודה היחידה שבה יחס זה משתנה היא D ויתכן כי היא תוכל לתת מידע לגבי החלטות של משתמשים שלא לבצע רכישה.

תרשים 1.2 מציג את התפלגות המשתנים, עבור משתנים נומריים וקטגוריאליים. סט הנתונים אינו מנורמל וניתן לראות זאת לפי ההתפלגויות. הפיצ'רים של פרקי הזמן ומספר העמודים, וגם עבור אחוזי הנטישה והיציאה, מתפלגים בצורה שמזכירה התפלגות פואסונית לזמני המתנה וניתן לראות כי קיימים בהם ערכים חריגים בעיקר בזנב הימני של ההתפלגות. משתנים D,B מתפלגים בצורה שנראית נורמלית וגם הם מכילים חריגים. בעת ניתוח הקשרים בין המשתנים אסיר את התצפיות החריגות בהתאם להתנהגות של כל פיצ'ר ואנרמל חלק מהנתונים לשם הכללתם במודל. מההתפלגויות ניתן גם להסיק כי במילוי ערכים חסרים עדיף להשתמש בחציון על פני ממוצע כיוון שהוא אינו מושפע בצורה קיצונית מערכים חריגים והינו אומדן טוב יותר בהתפלגויות שאינן נורמליות.

תרשים 1.3 מציג את הקורלציה (המתאם) בין הפיצ'רים השונים. כצפוי, פרקי הגלישה בעמודים השונים כמו גם מספרי העמודים מתואמים אחד עם השני ולכן בעיבוד המקדים הפיצ'רים האלה יעובדו כסט אחד ובאותו הקשר. הקורלציה העיקרית היא בין משך הגלישה הכולל באתר למשך הגלישה בעמודי מוצר. גם זה הגיוני כיוון שאלו מהווים את רוב העמודים באתרי e-commerce ומשתמשים מבלים בהם את רוב זמנם באתר.

קיימים מספר פיצ'רים שביניהם לבין תווית הרכישה הסופית קיימת קורלציה חיובית כגון PageValues ומשך זמן לעמוד ומספר עמודים. היינו מצפים לקשר זה שכן ככל שמשתמש בילה יותר זמן באתר ואסף יותר פריטים לרכישה כך גדלים

הסיכויים שהוא יבצע את הרכישה כיוון שהוא יותר "מושקע" בתהליך (זאת אומרת ככל שמדובר ברכישה משמעותית יותר מבחינת היקף המוצרים וערכם קטן הסיכוי שהמשתמש ישאיר את הסל ויעזוב את האתר מבלי לבצע טרנזקציה).

הנתון המפתיע הוא הקורלציה הגבוהה והשלילית שקיימת בין עמודה D לליבל ואחקור קשר זה בהמשך. בהמשך לפסקה לעיל נראה גם קורלציה חיובית עם שיעורי הנטישה והיציאה וקורלציה שלילית עם פרקי הזמן לעמוד, מספר העמודים וערך ממוצע לעמוד. עמודה זו כנראה מייצגת פיצ'ר שגורם לעזיבה של משתמשים באמצע התהליך ולהחלטה שלא לבצע טרנזקציה.

עיבוד מקדים

עיבוד הנתונים נעשה באמצעות חלוקת הפיצ'רים לתת קבוצות עם מאפיינים דומים ומתוך ההקשר שלהם (פירוט הפיצ'רים בכל קבוצה מופיע בנספח כטבלה 2.1):

1. <u>זמנים ומספרי עמוד</u> - בתחילה בחרתי להסיר את הערכים החריגים כדי להימנע מהטיות בהמשך במילוי ערכים חריגים או נרמול. הסרתי את הערכים שנמצאים מעל לאחוזון ה-99% לאחר הצגת הפיצ'רים בתרשים boxplot (תרשים 2.2 בנספח). ערכים אלו היוו 5% מסך הסשנים בסט הנתונים. לאחר מכן מילאתי ערכים חסרים לפי ההנחות הבאות- אם משך זמן היה ריק אך מס' העמודים הרלוונטי שווה ל-0 אז הנחתי כי לא הייתה גלישה. גם במקרים בהם מס' העמודים ריק אך משך הזמן היה 0 הנחתי כי לא הייתה גלישה ומילאתי את הערכים החסרים ב-0. את שאר הערכים מילאתי בחציון. לאור ההנחה שאתר אינטרנט מורכב משלושה סוגי עמודים אלו בלבד, השלמתי ערכים חסרים של סה"כ משך הגלישה ע"י סכימה של שלוש העמודות של פרקי הזמן (מפיצ'ר זה בחרתי שלא להוריד תצפיות חריגות כיוון שהוא מתואם בצורה כמעט מושלמת עם עמודי מוצר ומכיוון שבמודלים הסופיים בחרתי שלא לכלול אותו לאור זאת). לבסוף, החלטתי ליצור משתנים חדשים אשר מייצגים את ממוצע זמן הגלישה לסוג עמוד בהנחה של נרמול סה"כ הזמן לעמוד יתן ייצוג טוב יותר ויאפשר השוואה טובה יותר בין הפיצ'רים. שימוש בזמן ממוצע לעמוד אכן הוריד את הקורלציה בין הפיצ'רים אך גם את הקורלציה עם החלטת הרכישה הסופית.
2. <u>שיעורי נטישה ויציאה וערך ממוצע</u> - תרשים 2.3 מציג את הפיזור של שיעורי הנטישה והיציאה. נתונים אלה נמצאים בקורלציה מושלמת ובפיזור לינארי כמעט מושלם. בחרתי להסיר מס' תצפיות חריגות שאינן תואמות את הלינאריות שלהם (מוקפות בכתום בגרף). הערך הממוצע לדף מתפלג כמו פרקי הגלישה ובחרתי לא להסיר ממנו חריגים לאור הזניחות שלהם. לאחר מכן מילאתי את כל הערכים החסרים של הפיצ'רים בקבוצה זו בחציונים שלהם כיוון שהיה מדובר במספר לא משמעותי. בתרשים 2.4 ניתן לראות את הקורלציה בין הפיצ'רים. ניתן להסביר את הקורלציה הגבוהה בין שיעורי נטישה ויציאה בכך שהם מודדים כמעט את אותו הדבר אבל בצורה שונה, האחד את הנטישה מהאתר לאחר ביקור בעמוד אחד בלבד והשני את הנטישה מעמוד ספציפי וכנראה באתר זה מדובר באותו עמוד "בעייתי" (יתכן ויהיה צורך לבצע שינויים בעמוד הנחיתה או עמוד הבית של האתר). בשל המתאם הגבוה בין שיעורי הנטישה והיציאה ולאור העובדה שהאחרון מתואם בצורה יותר גבוהה עם הלייבל בחרתי להשתמש רק בו במודלים. ערך ממוצע לעמוד גם הוא פיצ'ר חשוב לחיזוי.
3. <u>תקופות וזמן</u> - בתרשים 2.5 ניתן לראות שיותר רכישות מתבצעות במהלך השבוע לעומת סופ"ש אך בעיני מדובר בהשוואה לא נכונה כיוון שבמהלך השבוע יש 5 ימי עבודה ואילו סופ"ש מתפרס על פני יומיים בלבד. שהבחירה אם לבצע רכישה או לא לא משתנה בהרבה בין סופ"ש לשאר ימות השבוע והלקוחות מתנהגים בצורה דומה לאורך השבוע. הייתי מצפה לראות אחוזי רכישה גבוהים יותר בסופ"ש אך יכול להיות שהנחה זו מתבדה בשל הנוחות שבביצוע רכישות אונליין וחוסר במאמץ פיזי כך שהצרכנים אדישים להאם מדובר בסופ"ש או לא. פיצ'ר סופ"ש יקודד כ-1 במידה ומדבור בסופ"ש ו-0 אחרת וערכים חסרים קיבלו ערך 0 בהתאם לערך הנפוץ ביותר. בבחינת החודשים (תרשימים 2.6-2.7) נמצא שהחודשים הפופולריים יותר לאתר הם מאי, נובמבר ומרץ כאשר יש חודשים כמו ינואר ואפריל בהם אין מספר כניסת (יתכן שזה נובע ממאפיינים של האתר ועצמו ויתכן כי זו הטיה בסט הנתונים אך אין לי דרך לדעת). מבחינת יחס, לקוחות נוטים יותר להחליט על ביצוע רכישות בחודשים מרץ ויוני. המרתי את המשתנה הקטגוריאלי בערכים שבין 1-12 (ערך גבוה יותר מסמל חודש מתקדם יותר בשנה קלנדרית) ומילאתי ערכים חסרים לפי החודשים הפופולריים ביותר - מאי ונובמבר. בניסיונות עיבוד קודמים ביצעתי גם המרה של החודשים לרבעונים ולמשתני דמי בניסיון להתחקות

אחר טרנדים עונתיים אך זה הביא לקורלציות יותר גבוהות עם משתנים אחרים (כמו קרבה לחגים, תרשים 2.8) ועל כן החלטתי לוותר על משתנים אלה בעיבוד הסופי. עבור ערכים חסרים בקירבה לחגים מילאתי בחציון ומעבר לכך לא השקעתי משאבים נוספים בניתוח שלו בשל המתאם הנמוך עם הלייבל והשונות הנמוכה שלו.

4. מאפייני משתמשים - בקטגורית דפדפן היה צורך להפריד קודם את שם הדפדפן מהגירסה (כיוון שהנחתתי שגירסת הדפדפן אינה רלוונטית לניתוח) וליצור קטגוריה חדשה של unknown עבור דפדפנים שסומנו כbrowser (כיוון שהמשתמש חסם כנראה את העוגיות או השתמש בVPN ולא ניתן לעקוב אחר הדפדפן) ועבור ערכים ריקים. לאחר מכן יצרתי משתני דמי באמצעות one-hot encoding. מתרשים 2.9 נראה אמנם שיש יותר סשנים של משתמשים בדפדפן chrome, אך בלא קשר לסוג הדפדפן, כ-85% מהלקוחות כן מבצעים רכישה (זאת אומרת, לצרכנים התנהגות דומה שאינה משתנה בהתאם לסוג הדפדפן). בניתוח סוג המכשיר מדובר במשתנה קטגוריאלי עם ערכים שבין 1-8. ערכים חסרים מולאו במכשיר הנפוץ. גם כאן במכשירים המובילים התנהגות המשתמשים ביחס לרכישה דומה. כיוון שמכשיר ודפדפן לעיתים קרובות קשורים זה לזה, מניתוח של תרשים 2.11 ניתן לראות שהשימוש בדפדפן chrome נפוץ במשתמשים במכשיר 2 (ולאחר מכן בדפדפן edge, שניהם נפוצים בעיקר בשימוש בdesktop) ובמכשיר 3 ואילו גולשים בsafari (שנפוץ בעיקר במכשירי אפל) משתמשים במכשיר 1. מניתוח זה, ידע מקדים ומחקר ברשת[i] אני מניחה שמכשיר אחד הוא כנראה טאבלט או סמארטפון, מכשיר 2 הוא מחשב (לפטופ/נייח) ומכשיר 3 כנראה סוג של מכשיר חכם כמו טלוויזיה. לאור זאת החלטתי ליצור משתנה חדש (browser_device_normalized) שיבטא את הקשר שבין שימוש במכשיר מסוים וגלישה מדפדפן מסוים. דירגתי את הדפדפן ואת המכשיר ונתתי להם ערכים בסדר עולה החל מ-1 לפי הקטגוריות המובילות. לאחר מכן כפלתי את הערכים בכדי להדגיש שילוב שהוא נפוץ יותר (למשל גלישה מאייפון או מלפטופ אישי). את הערכים שהתקבלו נרמלתי כיוון שאין משמעות לערך עצמו אלא רק ל"עוצמה שלו" (שילוב נפוץ יותר קיבל משקל גבוה יותר). בניתוח המשתמשים חילקתי למשתני דמי בהתאם לקטגוריות (משתמש חדש, משתמש חוזר ואחר. ערכים ריקים סווגו בקטגורית "אחר"). ניתן לראות (תרשים 2.12) שמשתמשים חוזרים נוטים יותר לבצע רכישה לעומת משתמשים חדשים וכי קיימת קורלציה מלאה ושלילית בין לקוח חדש ללקוח חוזר. בהנחה כי העדיפות הראשונה היא שימור לקוחות חדשים על פני משיכת לקוחות חדשים יהיה הגיוני להשתמש רק במשתנה זה במודלים לחיזוי (בהמשך נראה שמבחינת חשיבות פיצ'רים, לקוח חוזר אכן קיבל חשיבות גבוהה יותר). מבחינת אזור שממנו התחבר הסשן, לא ניתן להתחקות האם מדובר באזורים בין מדינות או באזורים בתוך מדינה אחת (לדוג' USA). ערכים חסרים מולאו בערך הנפוץ ביותר (אזור 1) ונראה שהמשתמשים מתנהגים בצורה דומה לגבי הרכישה (תרשים 2.14) למרות שיש אזורים פופולריים יותר שמהם מבוצעות יותר רכישות.

5. עמודות אנונימיות - עמודה A מכילה ערכי מחרוזת מהצורה 'C_num'. בניסיון להתחקות אחר משמעות הסימונים הנחתי כי כנראה מדובר באיזשהו מספור על עמודים. יצרתי פיצ'ר חדש 'A_num' שלקח את המס' של כל עמוד (תתי עמודים קיבלו את המספר של העמוד הראשי) ומילאתי ערכים חסרים ע"י מילוי של הערכים הנפוצים ביותר- 2, 1 ו-3. עמודה B מתפלגת נורמלית עם כמה תצפיות חריגות. הסרתי בצורה זהירה תצפיות מעל לאחוזון ה-99% ומתחת לאחוזון 1%. כיוון שהערכים של הפיצ'ר נעו בין 60—141 וחוסר יכולת להתחקות אחר הפיצ'ר ביצעתי נורמליזציה לפי MinMax בערכים שבין 0 ל-1 בכדי להבין האם ערך גבוה יותר תורם יותר למודל. עמודה C מכילה ערכי log עם מספרים שונים שאני מניחה שהם ערכי שגיאות HTTP[ii]. לצורך הניתוח ובהנחה כי סוג השגיאה עצמו אינו משפיע על החלטת הרכישה הסופית (תרשים 2.16), אלא רק העובדה אם הייתה שגיאה בתהליך או לא בחרתי לקודד את הערכים כ-1 עבור שגיאה ו-0 עבור אין שגיאה (פירוט הערכים, משמעות השגיאה והקידוד בטבלה 2.15 בנספח). מילאתי ערכים חסרים ביחס של 50-50 לשני הערכים. לבסוף יצרתי משתני דמי של no_error, yes_error. עמודה D מכילה מספר מועט של ערכים אך נראית כמאוד חשובה לפי הקורלציה הגבוהה שלה עם פיצ'ר הלייבל וערכי דף ופרקי הזמן לדף. נראה כי ערכים גבוהים שלה משפיעים בצורה שלילית על החלטת הרכישה. עיבודים שונים וקידודים לבדוק האם ערכים ריקים מוסיפים אינפורמציה למודל לא הובילו לתובנות משמעותיות ולכן בחרתי להסיר אותה בניתוח שביצעתי. בהיעדר אילוצי זמן ותקציב הייתי משקיעה עוד מאמצים בניסיון להתחקות אחר משמעות העמודה שכן נראה שהיא חשובה ויכולה להקפיץ את ביצועי המודלים ויכולות החיזוי באמצעות ביצוע עיבוד מקדים נכון.

בסיום העיבוד הראשוני הגעתי ל-50 פיצ'רים שונים עבור סה"כ 9339~ סשנים וניכר כי מדובר בריבוי משתנים מסבירים, מה שעלול להוביל לoverfitting של המודלים כיוון שהמשתנים תפורים לסט הנתונים הספציפי ובנויים בהתאם לו. כאן יש סכנה שהרצת מודל כזה על סט הtest תגרום לביצועים ירודים בשל תופעת הoverfitting. בנוסף, תיתכן קורלציה גבוהה בין משתנים רבים, מה שיגרום למודל חזרתי בו הרבה משתנים מביעים את אותו הדבר ואילו התרומה של כל אחד מהם לביצועי המודל ולמידע שהוא מוסיף למודל תהיה שולית בלבד וקיומו של רעש רב במודל. עיבוד מספר רב של פיצ'רים יקשה על הרצת הקוד וידרוש כוח חישובי רב ולא ניתן יהיה לבצע ניתוח של כל משתנה בפני עצמו ולהבין את החשיבות של כל פיצ'ר למודל ולחיזוי. כל אלה בעצם גורמים למה שנקרא the curse of dimensionality ועל כן אחפש לבחור בסט משתנים מצומצם יותר אשר יתן תחזית טובה יותר. לאחר הסרה ראשונית של משתנים שונים לאותו פיצ'ר נשארו 39 משתנים עליהם החלתי PCA במטרה להוריד את המימדיות. ניסיון זה לא היה מוצלח והמודל לא הצליח להוריד את המימדיות בצורה מספקת (ניסיון בקוד). לאחר ניסיון בשימוש forward selection אשר דרש כוח חישובי רב והעמיס על זמני הריצה, הגעתי למסקנה כי במסגרת הפרויקט והזמן הקיים ברשותי שימוש בהורדת מימדיות באמצעות filter method תהיה הטובה ביותר. בהסתמך על הניתוח עד כה וידע מקדים, השונות של כל פיצ'ר (שונות הנמוכה מthreshold מקובל שהוגדר, טבלה 2.17), קורלציות בין פיצ'רים ועם הלייבל (טבלה 2.18), ודירוג פיצ'רים לפי חשיבות והידע שהם תורמים למודל (תרשים 2.19) הקטנתי את מימדיות הבעיה ל-13 פיצ'רים (פירוט נוסף בקוד, רשימה סופית בטבלה 2.20) שאני מאמינה שייצגו את הסט בצורה הטובה ביותר ויוסיפו מידע בעל ערך לחיזוי.

לאחר ביצוע העיבוד המקדים נירמלתי את הנתונים באמצעות StandardScaler בכדי שיהיה ניתן לנתח אותם ולהשוות ביניהם על אותה סקאלה. את אותו העיבוד והנרמול ביצעתי גם על סט הטסט באמצעות הפונק' preprocessing.

הרצת המודלים

את סט האימון חילקתי רנדומלית ל-2 לפי יחס של 80-20 עבור אימון וולידציה (כיוון שהלייבל לא מאוזן) ובכדי להגדיל את יכולת ההכללה של המודלים. בחירת היפר-פרמטרים נעשתה באמצעות שימוש ב-GridSearch לקבל סט הפרמטרים שיביא לתוצאות אופטימליות בהתאם למודל (מלבד במודל הגאוסיאני שמחשב את ההסתברויות האפריוריות באופן עצמאי ועל כן לא ניתן לבצע בחירת פרמטרים). סיכום התוצאות והמודלים בטבלה 3.1.

1. Logistic Regression - המודל הגיע לציון של 0.89 ודיוק של 0.88. נבחרו היפר-פרמטרים:

```
{'C': 0.01, 'penalty': 'l1', 'solver': 'liblinear'}
```

מדובר ברגולריזציה (C) נמוכה שיכולה להעיד על טיב בחירת הפרמטרים.'liblinear' גם מתאים במקרה זה כיוון שמדובר על סט נתונים קטן יחסית.
הפיצ'רים החשובים ביותר להחלטה לבצע רכישה הם PageValues, month_key ואילו עבור ההחלטה שלא לבצע רכישה הפיצ'רים המשמעותיים הם ExitRates, Returning_Visitor (תרשים 3.2). מכאן שערך דף ממוצע אכן תורם לרכישה ומייצג במידה ההתחייבות המוניטרית שיש למשתמש בתהליך האתר. גם לחודש יש השפעה חיובית וזה כנראה נובע ממבצעים או סיילים בסוף עונה בחודשים מסויימים. ExitRates משפיע על ההחלטה לא לרכוש, זאת אומרת שקיים דף ספציפי שאנשים נוטים לצאת ממנו בלי לבצע רכישה יותר מאשר בדפים אחרים. גם העובדה שמדובר בלקוח חוזר משפיע על ההחלטה לא לבצע רכישה.
2. Gaussian Naïve Bayes - לא בוצעה בחירת פרמטרים. המודל הגיע לציון של 0.87 ודיוק דומה. מודל זה קיבל את הציונים הנמוכים ביותר מבין כל המודלים ויתכן שזה קורה כיוון שהמודל מניח אי תלות בין המשתנים. בפועל זה לא כך והמודל עוזר בעצן להבין שקיימת תלות בין המשתנים וכי פיצ'ר אחד ישפיע על ההתנהגות של השני ועל התוצאה בסיום התהליך (הגיוני כי מדובר בפיצ'ר של אותו משתמש).
3. Random Forest Classifier - המודל הגיע לציון של 0.92 ודיוק של 0.90. נבחרו היפר-פרמטרים:

```
{'criterion': 'entropy', 'max_depth': 200, 'max_features': 8,
'min_samples_leaf': 5, 'n_estimators': 80, 'random_state': 0}
```

הבחירה בקריטריון entropy מסייעת להבין את התרומה של כל פיצ'ר. לעץ עומק של 200 מה שמסייע להקטין את השגיאה של המודל אך מעלה את השונות בצורה ניכרת (bias-varience tradeoff).

5 הפיצ'רים החשובים ביותר לפי המודל הם PagesValues, ExitRates, month_key, avg_product_page_duration, B_normalized (תרשים 3.3). זה מחזק את ההנחה כי הזמן וההתחייבות הכספית המושקעים בסשן משפיעים בצורה רבה וחיובית על ההחלטה לבצע רכישה. משתמש שבילה זמן רב בחיפושים באתר ואסף סל קניות בערך גבוה ייטה פחות להשאיר את הסל באמצע ולעזוב את האתר מבלי להשלים את הרכישה. למבצעים או הנחות בחודשים מסויימים יש גם השפעה חיובית על רכישות, כפי שאנו כבר יודעים. ניכר שגם לפיצ'ר B המנורמל חשיבות רבה למודל אך לא הצלחתי להסביר את משמעותו.

4. MLP Classifier - המודל הגיע לציון של 0.89 ודיוק של 0.90. נבחרו היפר-פרמטרים:

```
{'activation': 'relu', 'alpha': 1, 'hidden_layer_sizes': (200, 200),
'learning_rate': 'adaptive', 'learning_rate_init': 0.01, 'max_iter': 2}
```

נראה שגם במודל זה קיבלנו רגולריזציה (alpha) לא גבוהה.

הערכת המודלים ופרדיקציה

תרשימים 4.1-4.2 מציגים confusion matrix עבור 2 המודלים עם הציונים הכי גבוהים -RandomForestClassifier וMLPClassifier. צירפתי את שתי המטריצות כיוון שהן נותנות ביצועים דומים בחיזוי נכון ויכולות לתת אינדיקציה נוספת לטיב המודלים, ולכך שהפיצ'רים מסבירים את החלטת הרכישה בצורה טובה. תוצאות המטריצה למודל הראשון:

TP= 81.04%, FP=3.89%, FN=5.92%,TN=9.15%

מכאן, ש95% מהלקוחות אשר ביצעו רכישה בסוף הסשן אכן יסווגו כבאלה שביצעו רכישה, לעומת 5% שיסווגו כלא מבצעים רכישה אך רכשו בפועל. המודל מפספס בממוצע 6.5% מהסשנים אשר מסווגים כלא ביצעו רכישה אך בעצם ביצעו רכישה בפועל. מבחינת דיוק של המודל, הוא חוזה ב-90% מהמקרים בצורה נכונה האם בוצעה רכישה או לא. מבחינה זו ביצועי המודל טובים ולרוב הוא מצליח לתת תחזית נכונה ומדוייקת.

לאחר מכן ביצעתי הערכה של המודלים ע"י K-Fold cross Validation ובנית עקומות ROC (תרשימים 4.3-4.6). הפונקציה מחלקת את הסט לאימון וולידציה ומניתוח הגרפים ניתן לראות כי מודל הRandomForestClassifier אכן נותן את הביצועים הטובים ביותר. בכל ביצוע של fold המודל מתכנס לציוני AUC דומים שהפער ביניהם לא עולה על 5%. נתון זה בנוסף לעובדה שהפער בין ציוני הbest בכל מודל לציון accuracy אינו עולה על 2% נותן אינדיקציה טובה שהצלחנו להימנע מoverfitting. הגדלת יכולת ההכללה של המודל נעשתה על ידי הרבה ניסוי וטעיה בבחירת הפיצ'רים, הסרת חריגים בצורה זהירה וניסיון להפחית קורלציה בין משתנים ככל הניתן.

סט הtest המעובד ומנורמל (ע"י StandardScaler) נשלח לביצוע תחזית באמצעות מודל RandomForestClassifier עם ההיפר פרמטרים הנבחרים. התוצאות נשמרו בקובץ CSV רלוונטי.

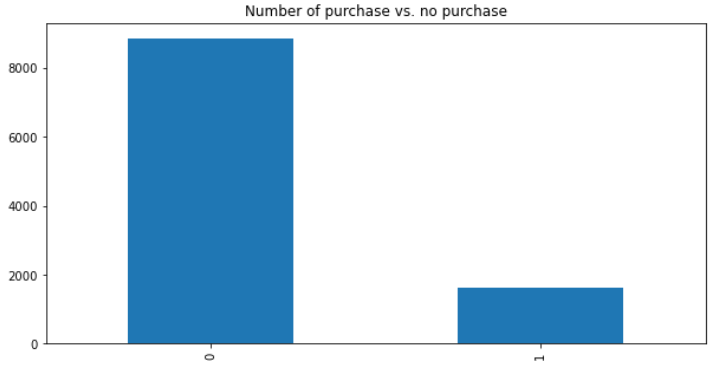
סיכום והמלצות

בהתחשב באילוצי משאבים וזמן במסגרת הסקופ של הפרויקט, נראה כי המודלים הגיעו לתוצאות חיזוי ודיוק טובות. הפער בין ציוני המודלים לדיוק נותן אינדיקציה טובה שהצלתי שהסיכון שמודלים חוטאים בoverfitting הוא קטן וכי הפיצ'רים שנבחרו אכן מצליחים לייצג בצורה טובה את הדאטה סט. גם ציוני המודלים מראים כי הם מצליחים לחזות בצורה טובה יחסית את החלטת הרכישה הסופית. מעבר ליתרון שבחיזוי תהליך רכישה בצורה נכונה, המודלים גם נתנו הערכה טובה לאילו פיצ'רים חשובים ותורמים לתהליך ביצוע הרכישה ומה ניתן לשפר. כך לדוג' אם מנהל האתר רוצה להגדיל את ההסתברות לרכישה, הוא יכול לנסות ולשפר פרמטרים אשר עשויים לגרום למשתמש לבלות יותר זמן באתר. בכדי להגדיל את הערך הממוצע לדף, ניתן לחשוב על הקטנת מספר העמודים או מעברים בין עמודים שמשתמש צריך לעבור עד להשלמת תהליך הרכישה או לנצל חודשים מסוימים בשנה למבצעים. בנוסף, על מנהל האתר לתת את הדעת לכך שיתכן שיש עמוד בעייתי אשר גורם למשתמשים לצאת מהאתר בהתאם לניתוח של שיעורי הנטישה והיציאה.

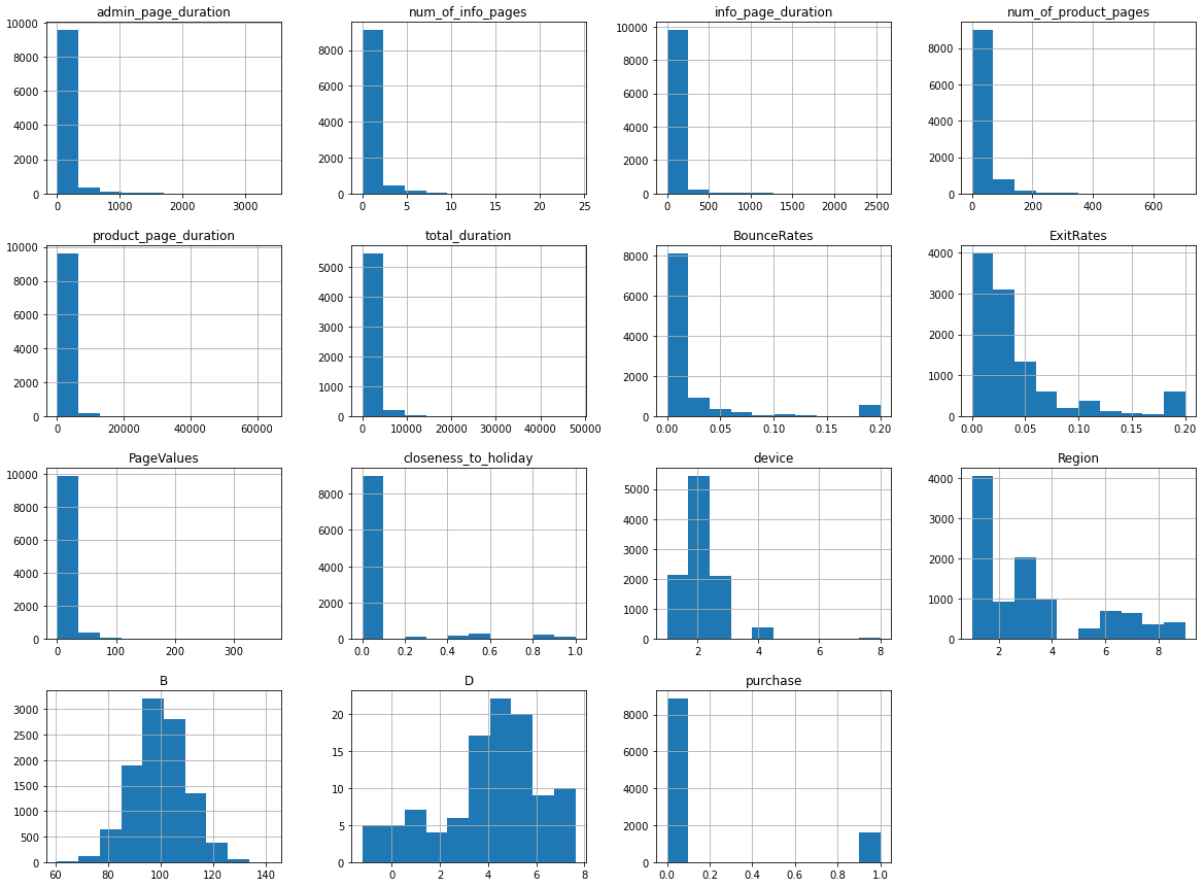
בהמשך לפרויקט ניתן להמשיך ולחקור כיוונים נוספים כמו גם את עמודות D,B שנראה כי עשויות לתרום הרבה מידע למודל, אך לא הספקתי להעמיק עוד בניתוח שלהם במסגרת הזמן הנתונה.

נספח - ויזואליזציה וגרפים

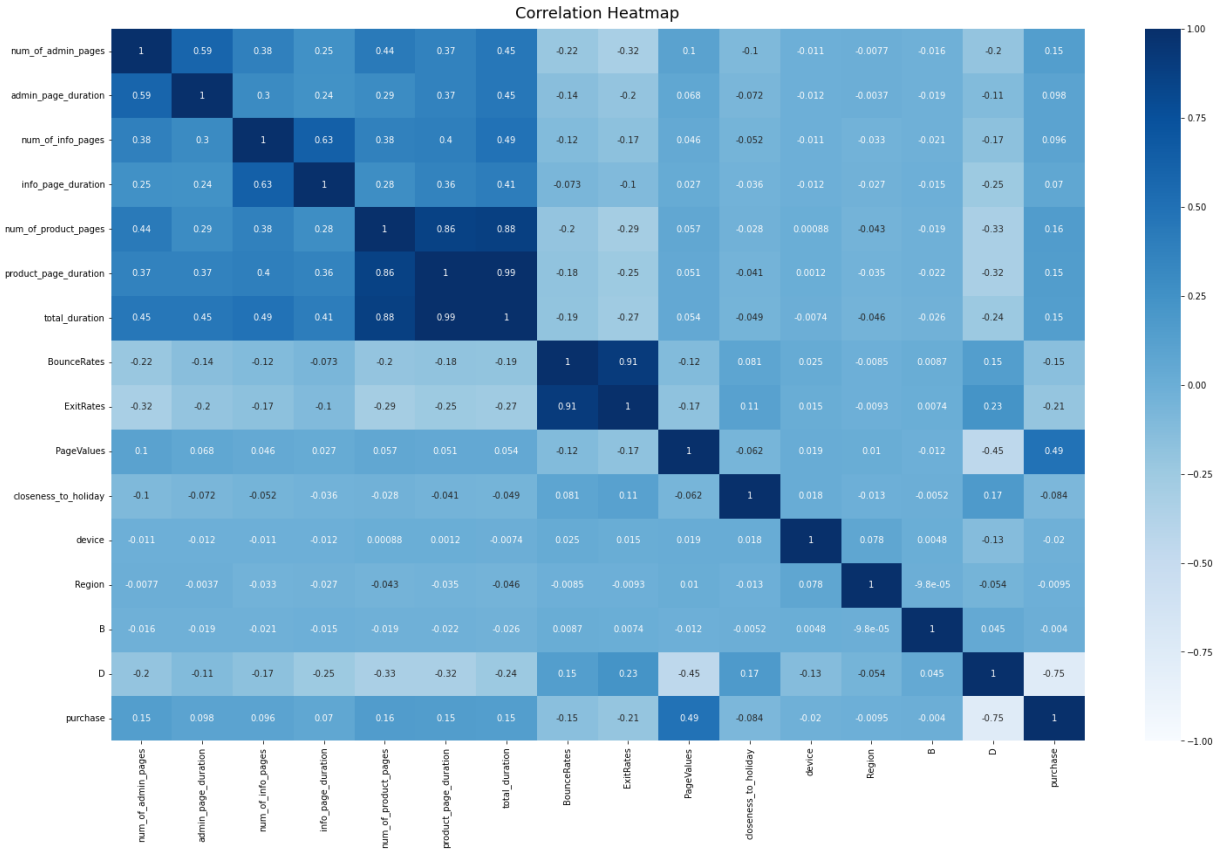
תרשים 1.1 - מספר הסשנים שהסתיימו ברכישות לעומת כאלו שלא



תרשים 1.2 - התפלגות המשתנים (נומריים וקטגוריאליים)



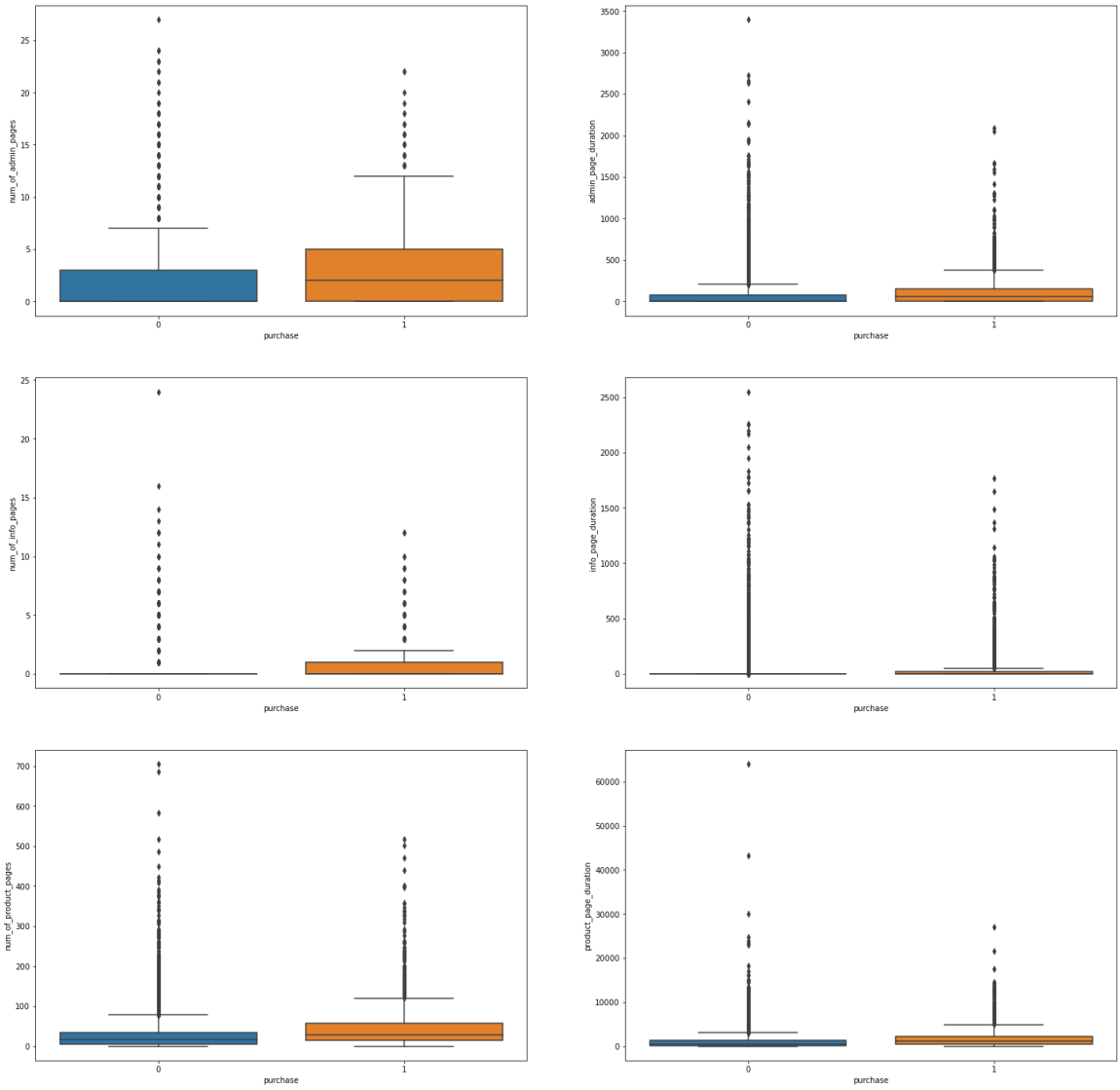
תרשים 1.3 - קורלציה (מתאם) בין המשתנים (נומריים וקטגוריאליים)



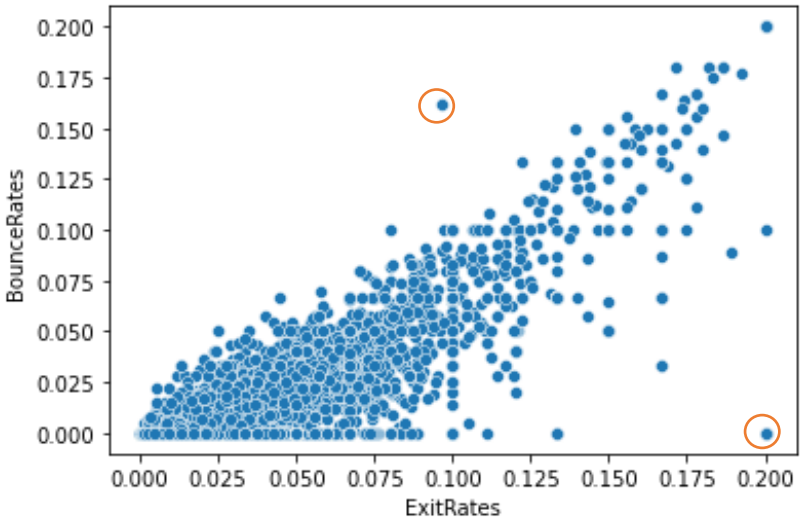
טבלה 2.1 - פירוט משתנים (פי'צרים) לפי קבוצות

תת-קבוצה	פיצ'רים כלולים
זמנים ומספרי עמוד	num_of_admin_pages, admin_page_duration, num_of_info_pages, info_page_duration, num_of_product_pages, product_page_duration, total_duration
שיעורי נטישה ויציאה וערך ממוצע	BounceRates, ExitRates, PageValues
תקופות וזמן	closeness_to_holiday, Month, Weekend
מאפייני משתמשים	Device, internet_browser, Region, user_type
עמודות אנונימיות	A, B, C, D
לייבל	purchase

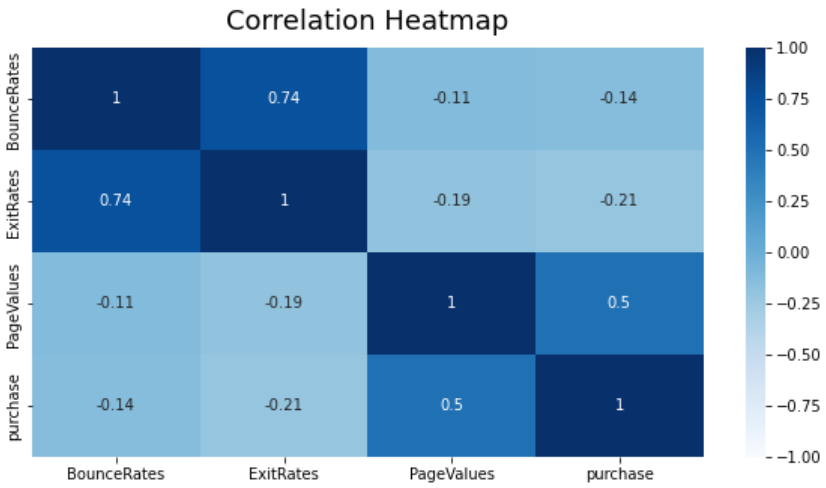
תרשים 2.2 - התפלגויות של זמנים ומספרי עמוד



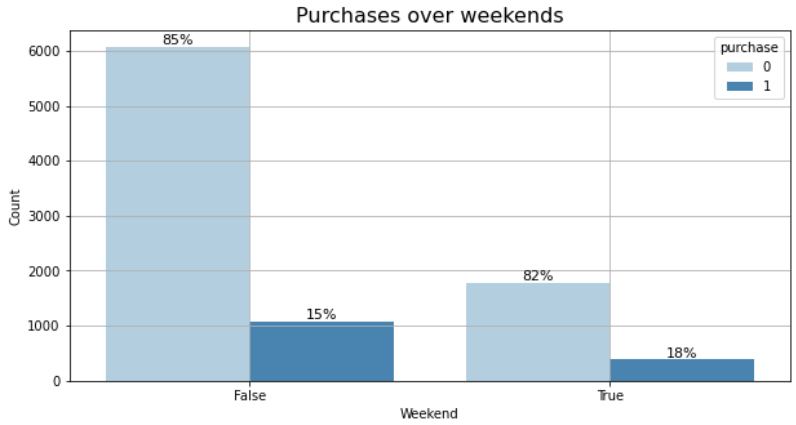
תרשים 2.3 - פיזור של BounceRates, ExitRates



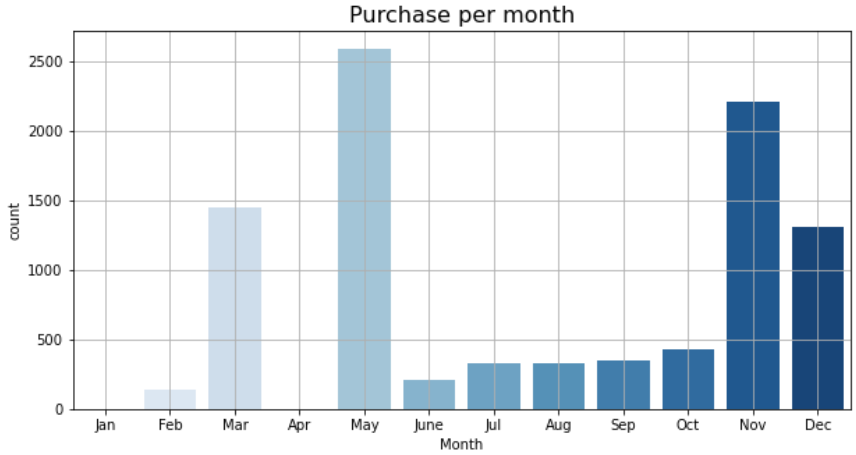
תרשים 2.4 - קורלציה בין BounceRates, ExitRates, PageValues and purchase



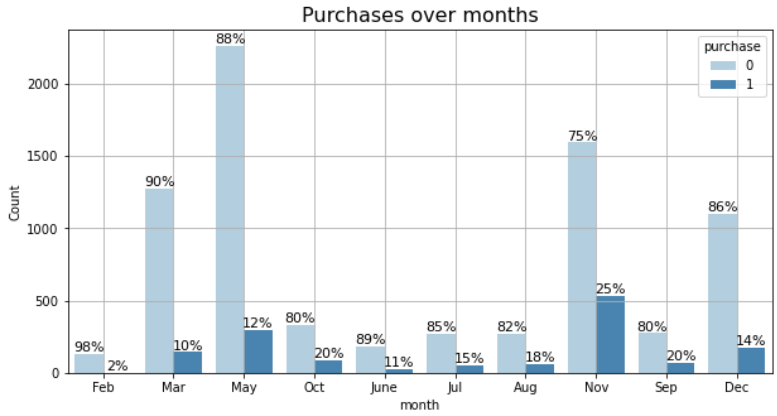
תרשים 2.5 - רכישות ואחוזי רכישות לפי סופ"ש או ימי עבודה



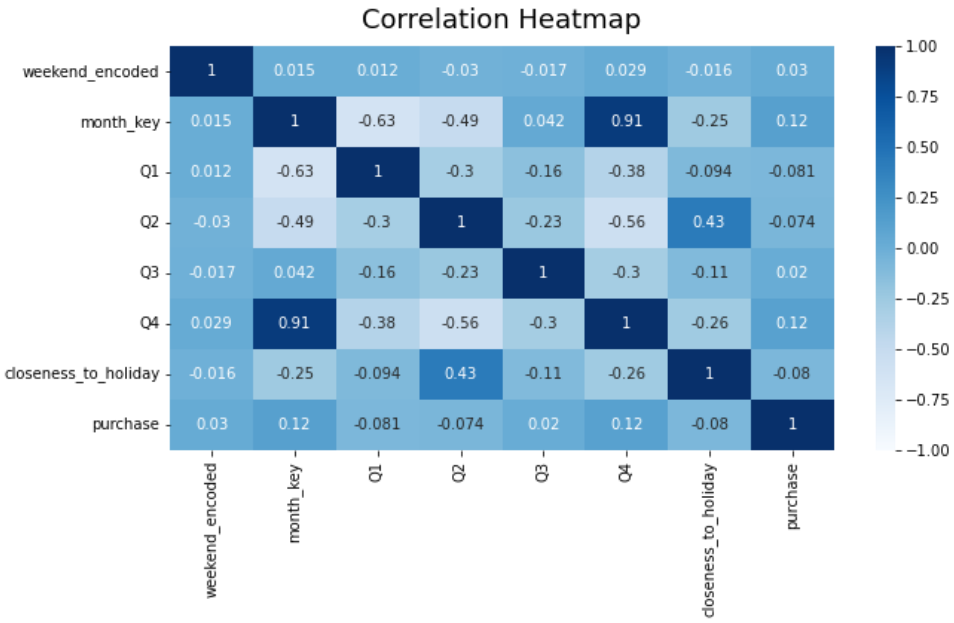
תרשים 2.6 - רכישות לפי חודש



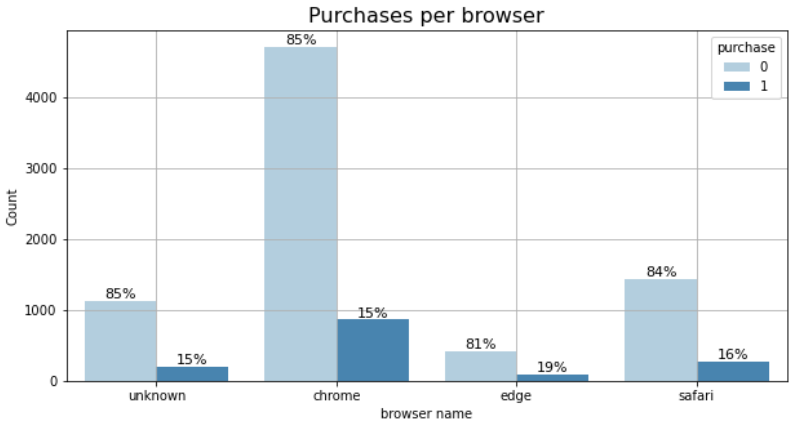
תרשים 2.7 - רכישות ואחוזי רכישות לפי חודש



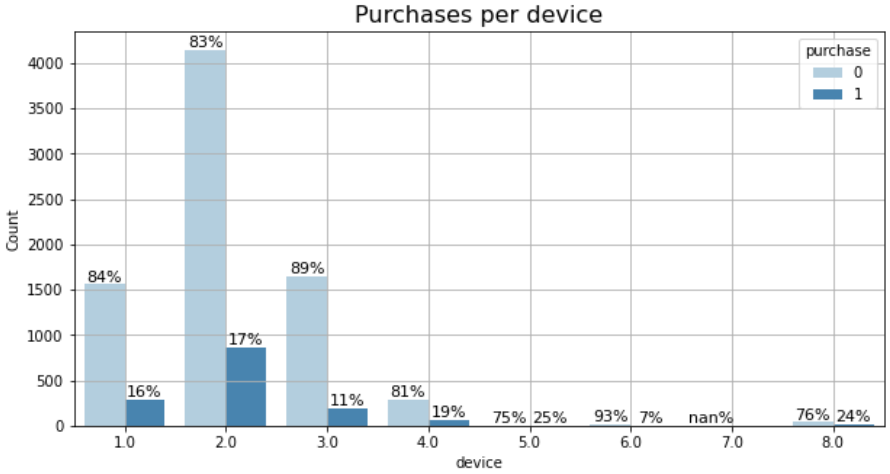
תרשים 2.8 - קורלציות בין משתני תקופות וזמן



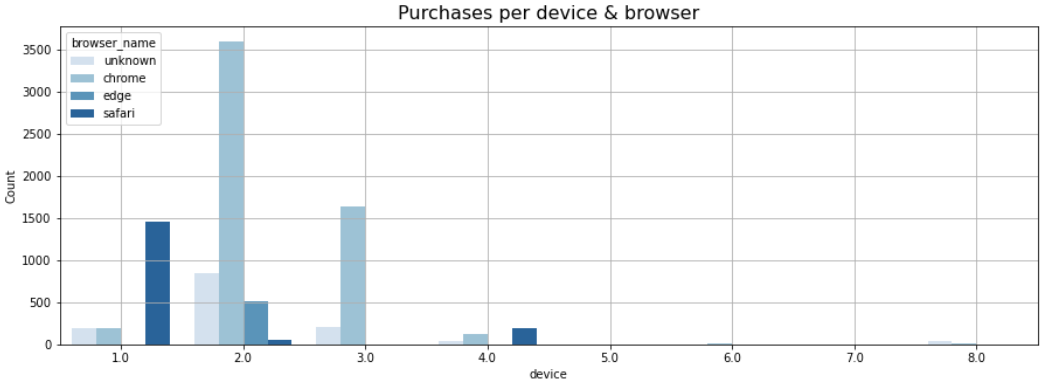
תרשים 2.9 - רכישות ואחוזי רכישות לפי דפדפן



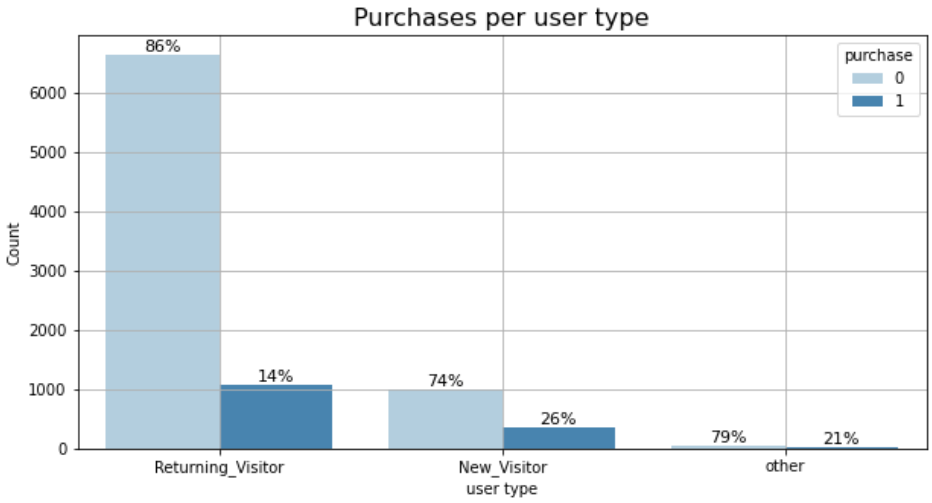
תרשים 2.10 - רכישות ואחוזי רכישות לפי מכשיר



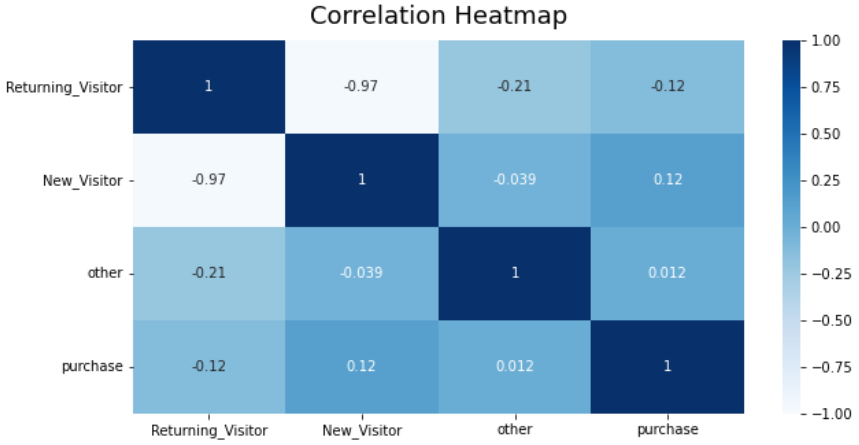
תרשים 2.11 - רכישות לפי מכשיר ודפדפן



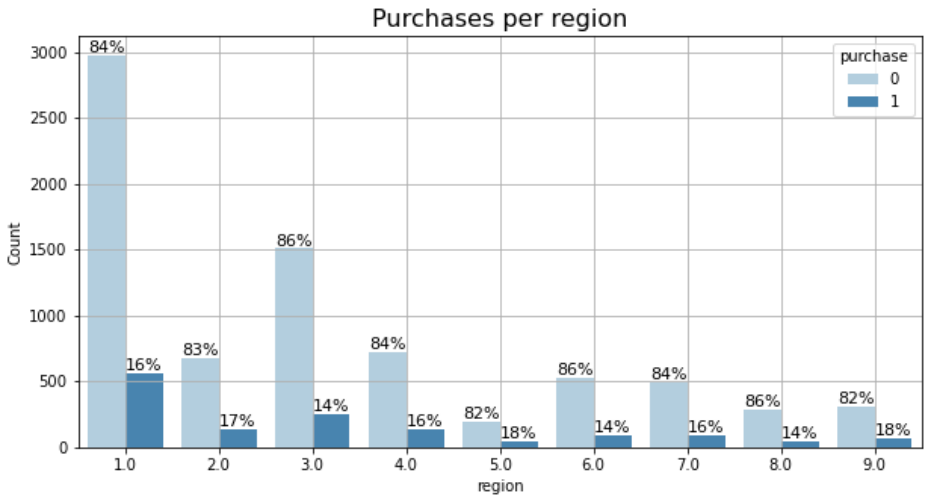
תרשים 2.12 - רכישות ואחוזי רכישות לפי סוג משתמש



תרשים 2.13 - קורלציה בין סוגי משתמשים לרכישה



תרשים 2.14 - רכישות ואחוזי רכישות לפי מכשיר

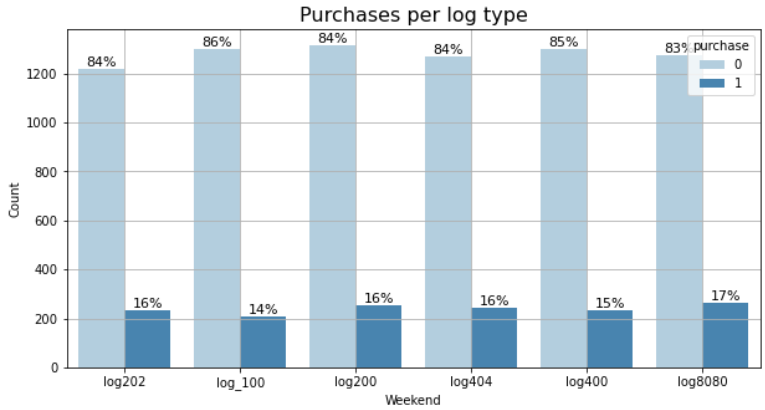


טבלה 2.15 - פירוט משמעות הודעות log והקידוד הנבחר בעיבוד המקדים

Log message	HTTP response	Code
log_100	Continue	0
log200	Ok	0
log202	Accepted	0
log400	Bad Request	1
log404	Not Found	1
log8080	Cannot Access Local Server	1

(חלק 2.6 בקוד)

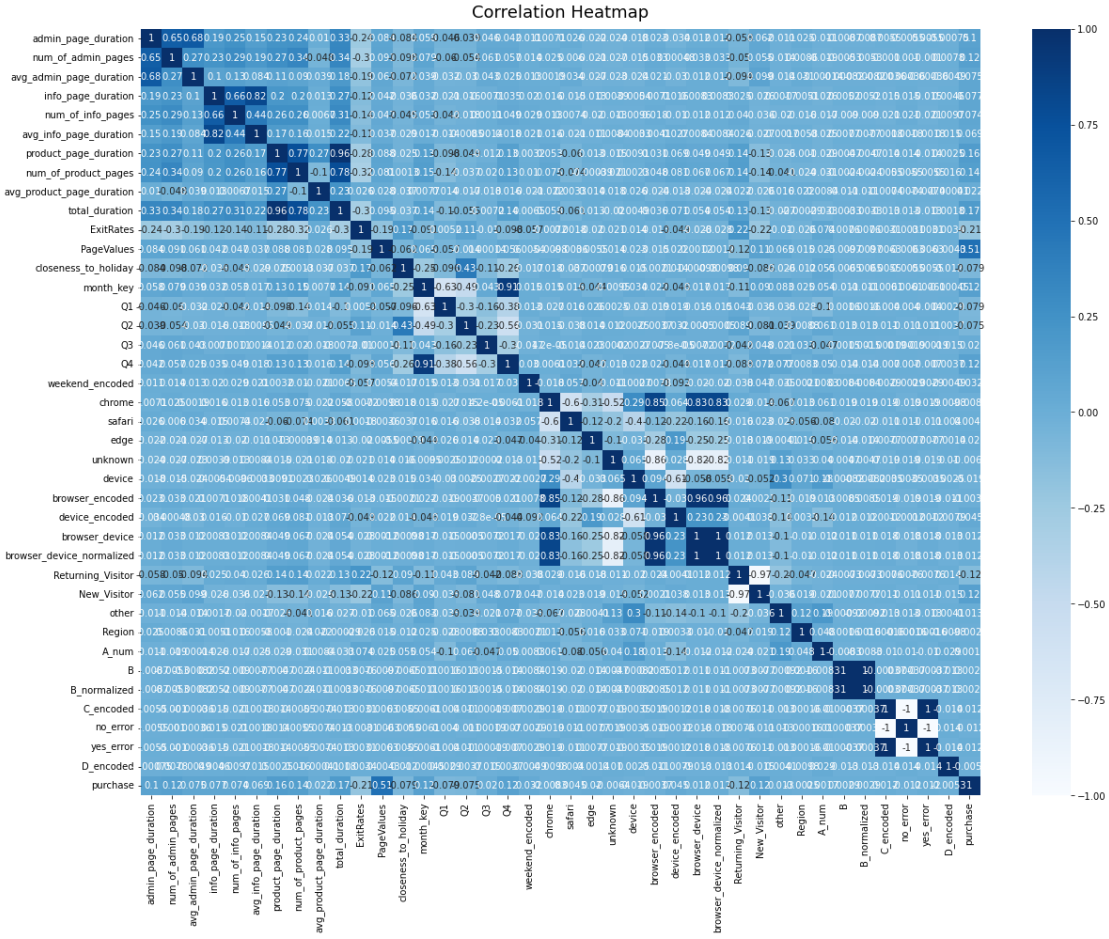
תרשים 2.16 - רכישות ואחוזי רכישות לפי הודעות log



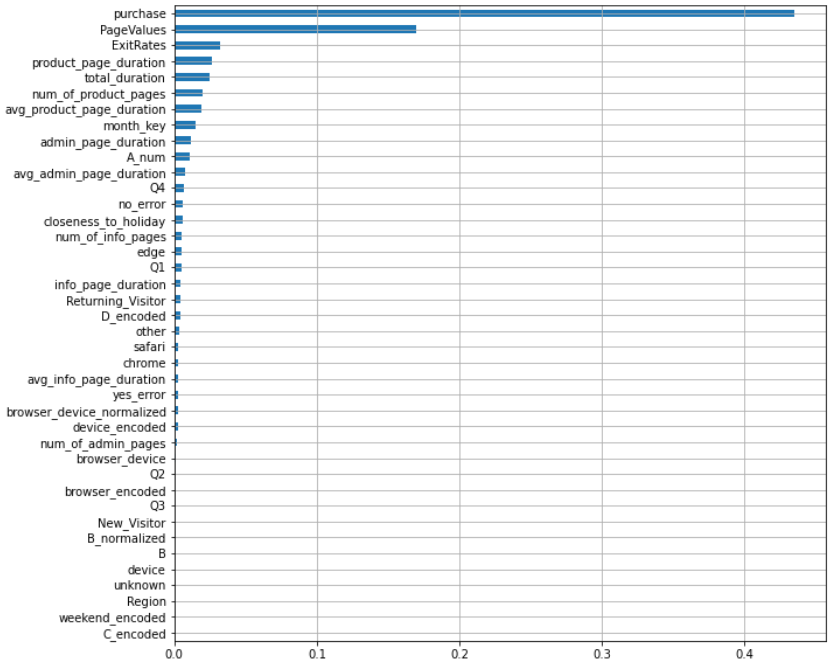
טבלה 2.17 - שונות משתנים הקטנה מ-0.16

admin_page_duration	False	edge	True
num_of_admin_pages	False	unknown	True
avg_admin_page_duration	False	device	False
info_page_duration	False	browser_encoded	False
num_of_info_pages	False	device_encoded	False
avg_info_page_duration	False	browser_device	False
product_page_duration	False	browser_device_normalized	True
num_of_product_pages	False	Returning_Visitor	True
avg_product_page_duration	False	New_Visitor	True
total_duration	False	other	True
ExitRates	True	Region	False
PageValues	False	A_num	False
closeness_to_holiday	True	B	False
month_key	False	B_normalized	True
Q1	True	C_encoded	False
Q2	False	no_error	False
Q3	True	yes_error	False
Q4	False	D_encoded	True
weekend_encoded	False	purchase	True
chrome	False	dtype: bool	
safari	True		

תרשים 2.18 - קורלציות בין פיצ'רים ולייבל



תרשים 2.19 - דירוג חשיבות פיצ'רים



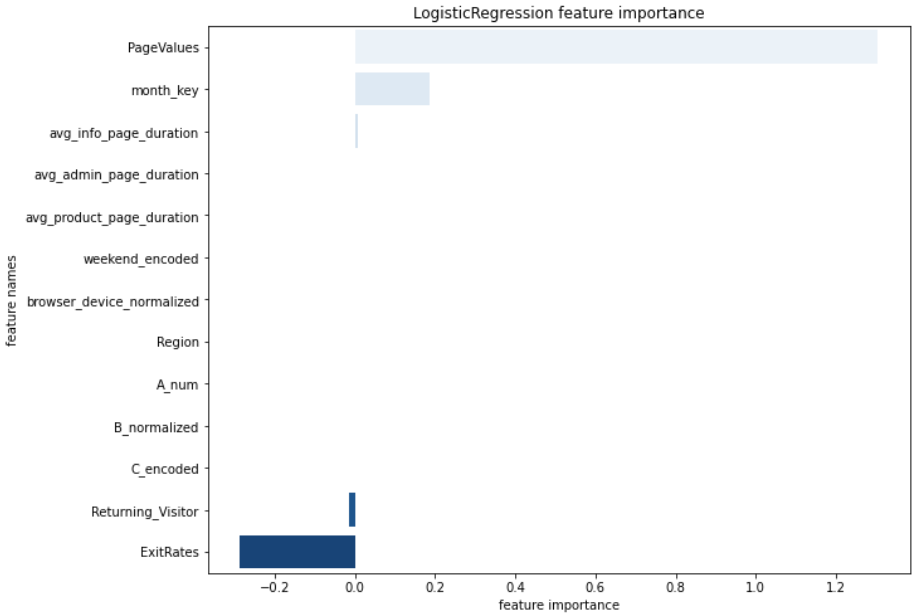
טבלה 2.20 - רשימת משתנים סופיים

['avg_admin_page_duration','avg_info_page_duration','avg_product_page_duration,' 'ExitRates','PageValues','Returning_Visitor','weekend_encoded','month_key', 'browser_device_normalized,'Region','A_num','B_normalized','C_encoded']

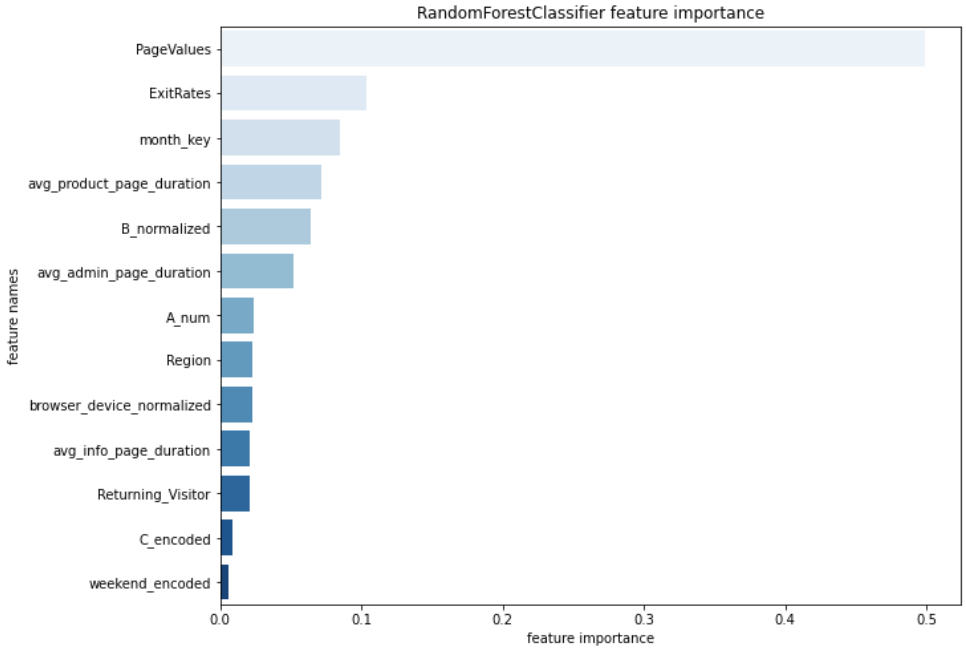
טבלה 3.1 - סיכום תוצאות המודלים והיפר פרמטרים נבחרים.

Model	Hyperparameters	Best score	Accuracy	Risk for overfitting
Logistic Regression	{'C': 0.01, 'penalty': 'l1', 'solver': 'liblinear'}	0.892	0.886	False
Gaussian Naïve Bayes	-	0.871	0.875	False
Random Forest Classifier	{'criterion': 'entropy', 'max_depth': 200, 'max_features': 8, 'min_samples_leaf': 5, 'n_estimators': 80, 'random_state': 0}	0.921	0.908	False
MLP Classifier	{'activation': 'relu', 'alpha': 1, 'hidden_layer_sizes': (200, 200), 'learning_rate': 'adaptive', 'learning_rate_init': 0.01, 'max_iter': 2}	0.896	0.902	False

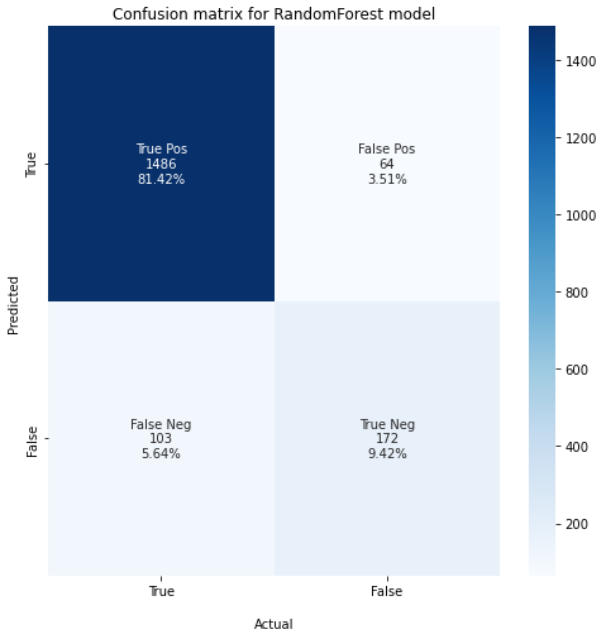
תרשים 3.2 - דירוג פיצ'רים לפי חשיבות במודל LogisticRegression



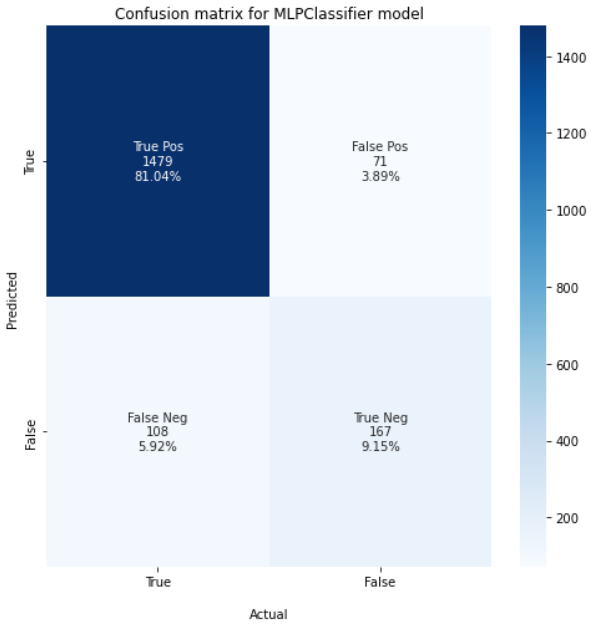
תרשים 3.3 - דירוג פיצ'רים לפי חשיבות במודל randomForestClassifier



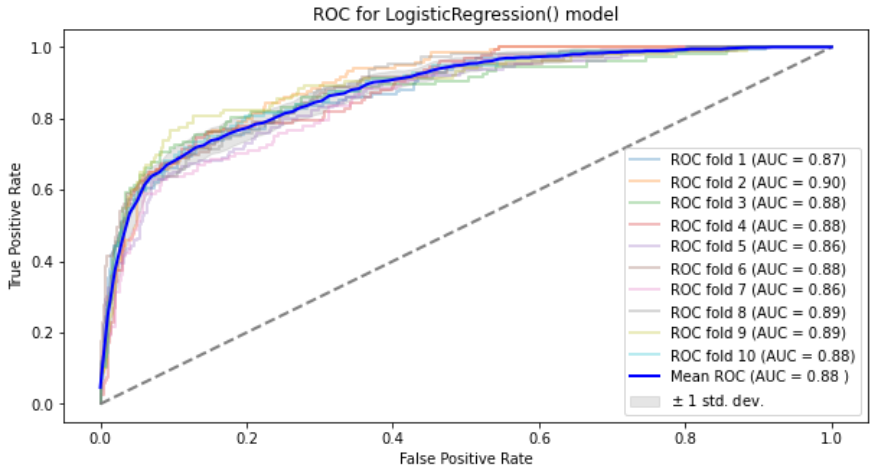
תרשים 4.1 - Confusion matrix עבור RandomForestClassifier



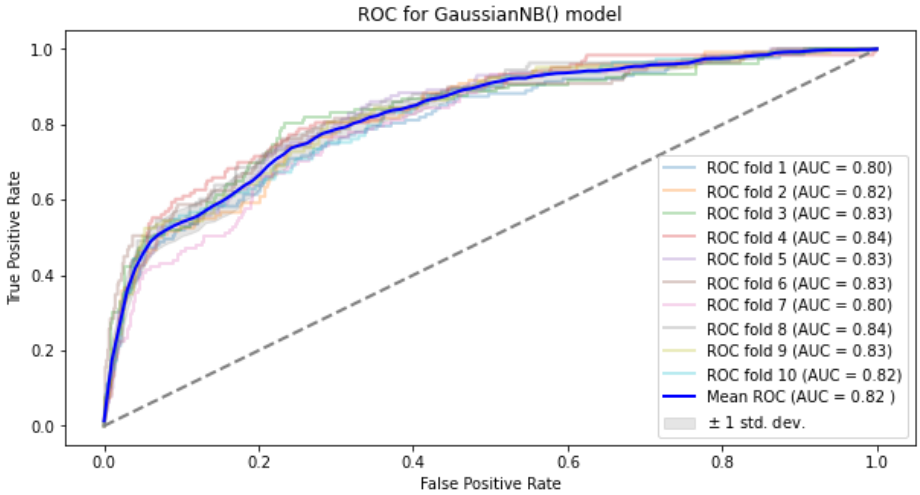
תרשים 4.2 - confusion matrix עבור MLPClassifier



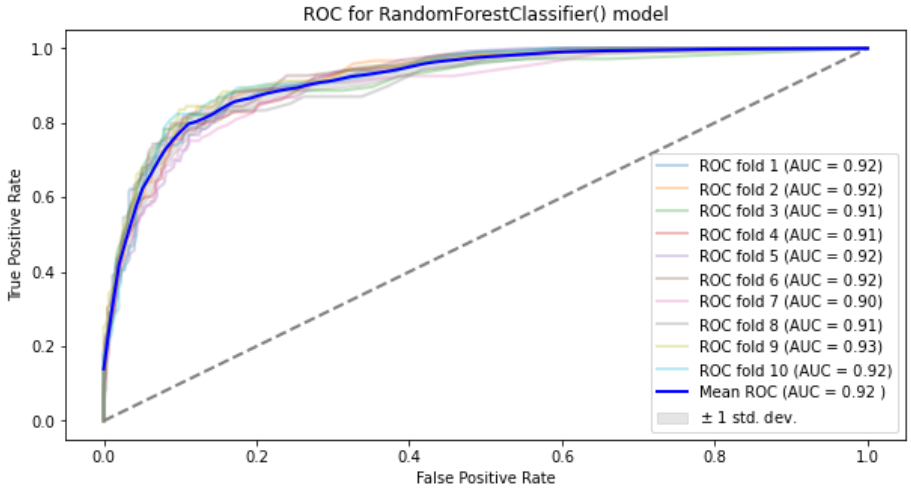
תרשים 4.3 - תרשים ROC עבור LogisticRegression



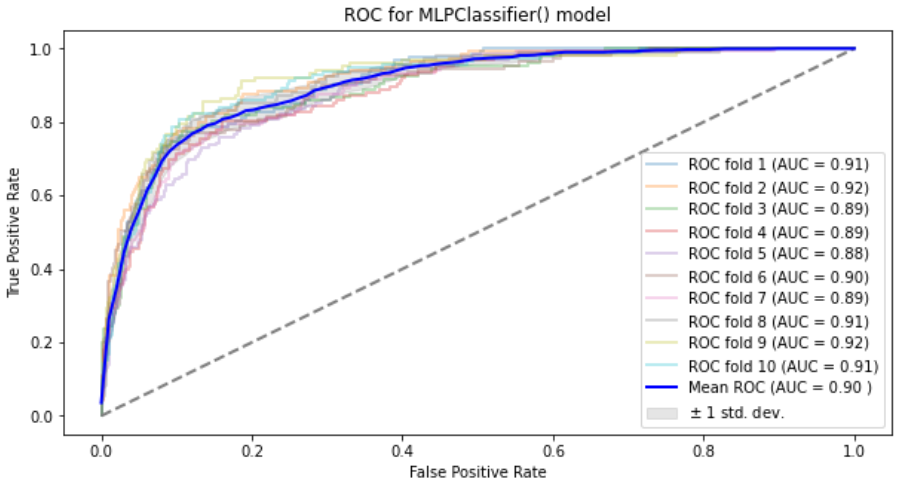
תרשים 4.4 - תרשים ROC עבור GaussianNB



תרשים 4.5 - תרשים ROC עבור RandomForestClassifier



תרשים 4.6 - תרשים ROC עבור MLPClassifier



[i] Source: device table, openRTB API specification (page 51), https://www.iab.com/wp-content/uploads/2016/03/OpenRTB-API-Specification-Version-2-5-FINAL.pdf)

[ii] Source: HTTP response status codes (https://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml)