פרוייקט למידת מכונה

דו"ח מסכם

גיא חן	יעל פליישר
315007187	204758189

09/06/22

תקציר מנהלים

הפרויקט שלנו עסק בבעיית קלסיפיקציה שמטרתה לחזות האם משתמש אינטרנטי יבצע רכישה באתר מסוים. תחילה ביצענו מחקר ראשוני על סט הנתונים – ניסינו ללמוד על יחסים והשפעתם של הפיצ'רים השונים אחד על השני ועל ביצוע הרכישה הסופית. כדי לקבל גישה ויכולת עיבוד לכל הנתונים, כך שנוכל לאמן עליהם את המודלים בצורה הטובה ביותר, ביצענו המרה של כלל הנתונים הקטגוריאלים למספרים בדרכים שונות כאשר מיטב המאמצים הוקדשו לשימור כמה שיותר אינפורמציה רלוונטית ויצירת כמה שפחות ממדים נוספים. כמו כן, הסרנו ערכים קיצוניים ומילאנו ערכים חסרים בהתאם לקורלציות בין פיצ'רים שונים ובהתאם לערכים הסטטיסטיים של כל עמודה. לבסוף, פיצלנו את הדאטה שלנו לסט אימון ולסט ואלידציה, ונרמלנו את הנתונים שלנו, כך שכל הפיצ'רים ינועו על סקאלה אחידה והמודל שלנו יוכל ללמוד מהם ללא תלות ביחידות המידה השונות. בשלב בחינת המודלים ביצענו תחילה רידוד נוסף של הממדים על ידי PCA וראינו שפעולה זו מצמצמת ממד אחד בלבד. הרצנו שני מודלים פשוטים (KNN, רגרסיה לוגיסטית) ושני מודלים מורכבים (NN ,SVM), כאשר עבור כל מודל מצאנו את היפר הפרמטרים האידיאליים עבורו באמצעות שיטת K-Fold Cross Validation (מיושם דרך grid search), ולאחר מכן בחנו ציון AUC באמצעות סט הואלידציה. מסקנתנו העיקרית מההרצות הללו הייתה שניתן לפצל את המודל שלנו באופן לינארי על ידי מודל לוגיסטי, ושמודל ה-NN משפר מאוד את ביצוע הפרדיקציה. לפני שבנינו מודל סופי, השתמשנו ברגרסיה לוגיסטית שבאמצעותה מצאנו את הפיצ'רים החשובים ביותר במודל שלנו (נספח 5). לאחר שנשארנו רק עם שמונת הפיצ'רים החשובים ביותר, ביצענו אופטימיזציה למודל רשתות נוירונים והשגנו מודל סופי. הבחירה נעשתה תוך בקרה על כך שהמודל אינו מורכב מדי (באמצעות סט האימון), ובו בזמן נותן תוצאות מקסימאליות במדד AUC על סט הואלידציה. הצלחנו להגיע לציון 0.93 על סט הואלידציה (נספח 6).

חקירת הדאטה וטיפול במשתנים שאינם מספריים

התחלנו את הפרויקט בחקירה של הדאטה – בדקנו כמה תצפיות יש לנו, כמה פיצ'רים יש לנו, מהו סוג האובייקט של כל פיצ'ר, וכמה ערכים חסרים יש לנו מכל פיצ'ר. כאשר ניסינו לצפות בסטטיסטיים שונים של פיצ'רים שונים, הדבר הראשון ששמנו לב אליו הוא שיש לנו פרספקטיבה מוגבלת על הדאטה, שכן לא ניתן היה להציג סטטיסטיים עבור משתנים קטגוריאליים, או למשתנים כמו Info_page_duration, אשר לכל תצפית מספרית הוצמד סטרינג שצריך לנקות.

לאחר שקיבלנו רושם על התפלגותם של הפיצ'רים השונים באמצעות היסטוגרמות, החלטנו לטפל תחילה בפיצ'רים הקטגוריאלים בדאטה, כדי שנוכל לקבל "גישה" ליחסים בין הפיצ'רים השונים. נקטנו גישה לפיה אם הרמות השונות של הפיצ'ר לא נותנות הבחנה בסיכוי לרכישה, אז השונות של הפיצ'ר לא עוזרת בניבוי רכישה ואולי עדיף לוותר עליו. המרת פיצ'ר קטגוריאלי לערכים מספריים בעזרת one-hot encoding משמעותה הרחבת מימדים ומכך רצינו להימנע, במיוחד כאשר מדובר במשתנה בעל רמות רבות. להלן האופן בו טיפלנו בפיצ'רים הקטגוריאליים:

1. **דפדפן האינטרנט ופיצ'ר C –** משתנה בעל מספר עצום של רמות. בשלב הראשוני, לקחנו כל דפדפן, ובדקנו האם תתי הסוגים שלו מצליחים לחזות בצורה שונה את הלייבל. הגענו למסקנה שתתי הסוגים לא נותנים לנו מידע בנוגע לחיזוי הלייבל, ולכן צמצמנו את רמות המשתנה לסוגים העיקריים של הדפדפנים ללא ההבחנה לתתי סוגים – Safari, Chrome, Edge, Browser. כשהשוונו בין ההסתברויות לביצוע רכישה בין הדפדפנים השונים, גם ראינו שאין הבחנה בין הדפדפנים, ואף אחד מהם לא מעיד על סיכוי גבוה יותר או נמוך יותר לבצע רכישה. לכן החלטנו להיפטר מהפיצ'ר לחלוטין. בפיצ'ר C ביצענו תהליך דומה, והגענו למסקנה שהרמות השונות של משתנה C אינן תורמות לנו, לכן הסרנו גם אותו.
2. **פיצ'ר A –** פיצ'ר נוסף עם רמות רבות (96 רמות). הצעד הראשון שעשינו הוא בדיקת שכיחות ההופעה של כל אחת מהרמות. ראינו שלאחר הטופ 10, השכיחות להופעת כל ערך הופכת להיות קטנה בצורה משמעותית, לכן החלטנו להתייחס לכל אותם ערכים נדירים כקטגוריה אחת בשם C_Others. לאחר מכן בדקנו את אחוזי הרכישה של כל אחת מהרמות השונות, והצלחנו לבודד את ההתנהגות של כל הרמות לשלוש קבוצות שונות שמייצגות אחוזי רכישה שונים. לאחר ניתוח שהשקענו בו מחשבה רבה (נספח 1) החלטנו לקבץ את הרמות השונות ל-2 קבוצות מובחנות ולא לשלוש, זאת מתוך מחשבה שכך נוכל לוותר על one-hot encoding, להפוך את המשתנה לבינארי, וכך להימנע מ"קללת המימדיות". בגדול, הסקנו שהמידע שנאבד הוא לא גורלי עבורנו. בשלב האחרון, לקבוצת ערכי C אחת נתנו את הערך 0, ולשאר נתנו 1, כך שכל קבוצה מייצגת אחוזי רכישה שונים.

3. **פיצ'ר ה-weekend –** פיצ'ר בוליאני המציין האם הרכישה התקיימה בסוף השבוע או לא, אותו המרנו לפיצ'ר נומרי עם 0 ו-1.
4. **User_type –** פיצ'ר נוסף שדרש טיפול אינטנסיבי. שמנו לב שכל רמה של סוג יוזר נותנת אחוז שונה לרכישה, אבל לא יכולנו שלא לשים לב לכך שהרמה Others היא רמה עם מספר תצפיות זניח (72 תצפיות). החלטנו לפצל את התצפיות ברמה Others לקטגוריות האחרות ובכך להפוך גם את משתנה "סוג המשתמש" לבינארי, במקום להשתמש ב one-hot encoding ולהוסיף ממדים. את הפיצול לקטגוריות האחרות עשינו בצורה שמטרתה הייתה לשמר את אחוזי הרכישה של הקטגוריה Others – לתצפיות באותה קטגוריה היה 18% לבצע רכישה, אז החלטנו לקחת 65% מהערכים בקטגוריה באופן אקראי ולהפוך אותם למשתמש חוזר, רמה עם 13.8% רכישה, ואת 35% מהערכים הנותרים להפוך למשתמש חדש, רמה עם 25.77% רכישה (ראו נספח 2). בדרך זו יכולנו לשמר את אחוז הקניה של הרמה המקורית. חשבנו שגם אם הדרך הזאת לא אידיאלית, מכיוון שמדובר במספר תצפיות קטן מאוד שנמצא ברמה הזאת, הנזק לא יכול להיות גדול, והעדפנו לחסוך במימדים.
5. **Month Column – feature engineering**

ראשית בדקנו כרגיל את אחוזי הרכישה של כל רמה. ראינו שכל רמה אכן חשובה לנו – כל חודש ייצג אחוזי רכישה שונים, בתחילת השנה ישנם אחוזים נמוכים שבאופן כללי עולים ככל שמתקדמת השנה. החלטנו שאנחנו רוצים לשמור על המידע של הפיצ'ר והבנו כי הפיכת כל חודש לערך מספרי בצורה סטנדרטית תפגע במשמעותו של הפיצ'ר, שכן מדובר בפיצ'ר מחזורי. למשל – אם היינו הופכים את החודש דצמבר למספר 12 ואת החודש פברואר למספר 2 באופן נאיבי, המשמעות הייתה שדצמבר ופברואר רחוקים מאוד אחד מהשני, כשבפועל הם מאוד קרובים כשמדובר בחודשי השנה. על מנת לפתור את הבעיה, השתמשנו בפונקציות המחזוריות סינוס וקוסינוס. ראשית, נרמלנו את ערכי החודשים שלנו בין 0 לπ2, המייצגים מחזור של הפונקציות סינוס וקוסינוס, ולאחר מכן הפעלנו עליהם את הפונקציות. הסיבה שהזדקקנו גם לסינוס וגם לקוסינוס הייתה כדי שכל חודש יקבל קואורדינטה ייחודית במעגל שבנינו (נספח 3).

6. **פיצ'ר D –** עמודה עם 105 ערכים ו-10,347 ערכים חסרים. בגלל הכמות האדירה של הערכים החסרים והיחס הקטן של הערכים שאינם חסרים, החלטנו שהעמודה הזאת לא יכולה לתת לנו תועלת. על אף שמעט הערכים שכן קיימים היו בקורלציה גבוהה עם הלייבל, מדובר בכמות תצפיות קטנה שעליה מבוססת הקורלציה, ולא רצינו להחיל הנחות על כל כך הרבה ערכים חסרים (על מנת למלא אותם). לכן ויתרנו על העמודה.
7. **פיצ'רים המייצגים זמן גלישה –** העברנו את המילה "דקות" מהערכים עצמם לכותרת הפיצ'ר, כך שנוכל לשנות את התכנים של העמודות לערכים נומרים. הבנו שכל העמודות מחושבות בדקות מכיוון שהסכום של כל עמודות הזמן תמיד הסתכם למשך הזמן הכולל באותה הסקלה.

כעת כשכל הפיצ'רים נומרים, הייתה לנו נגישות לכל המידע שקיים לנו. הסתכלנו על הקורלציות השונות הקיימות בדאטה. ראינו שכל הפיצ'רים המייצגים זמנים שונים בסוגי העמודים השונים (Admin, Info, Product) הם קורלטיביים אחד עם השני, וכך גם הפיצ'רים המייצגים מספרי עמודים בסוגי עמודים שונים, זה מידע שהשתמשנו בו בהמשך להשלמת חסרים, וגם להסרת עמודה.

<u>טיפול בערכי קיצון ובערכים חסרים</u>

הסרת תצפיות עם משתנים חסרים רבים – תחילה, הסרנו תצפיות עם מעל 50% ערכים חסרים. הרציונל היה הסרת תצפיות שנותנות "יותר רעש מסיגנל", במובן הזה שנצטרך ל"נחש" להן יותר ערכים מאשר מה שהם נותנות לנו. מכיוון שהסרנו רק 15 תצפיות בדרך הזאת, חשבנו שזה מהלך נכון שלא מאבד לנו הרבה דאטה.

טיפול בערכים קיצוניים – החלטנו לעשות זאת לפני הטיפול המלא בערכים החסרים מתוך מחשבה שלא נרצה למלא ערכים חסרים באמצעות חישובים שמסתמכים על תצפיות קיצוניות. נקודה נוספת הייתה שאם היינו מאתרים ערכים קיצוניים אחרי מילוי הערכים החסרים, אז היינו מסתמכים על ערכים ששיערנו על מנת לזהות מי קיצוני, וזה היה פוגע בדיוק של הניתוח – העדפנו להחליט מי קיצוני רק על סמך המידע שאנחנו יודעים בוודאות ולא צריכים לחזות.

- **עמודה B –** הסתכלנו על המשתנים שלנו שוב דרך היסטוגרמות, וראינו שהתפלגות עמודה זו נראית נורמלית. ביצענו מבחן השערות שבמסגרתו השערת האפס שלנו הייתה שההתפלגות היא נורמלית, והשערת האפס לא נדחתה. כלומר,

ההתפלגות אכן נורמלית. לקחנו את כל השורות בהן ציוני התקן בערך מוחלט של עמודה B היו גדולים מ-3, והחלטנו להסיר אותן (28 תצפיות).

- **PageValue, Total_Time_Duration -** השמטת ערכים חריגים במיוחד בפיצ'רים אלו שאינם חסומים מלמעלה (נספח 4). ראינו שמעט מאוד תצפיות קיבלו ערכים גבוהים ביותר. מכיוון שאין הגבלה לעד כמה אותם ערכים יכלו להיות גדולים, חשבנו שכדאי להשמיט אותם, זאת על מנת שלא ישפיעו בצורה דרסטית על ממוצעים שנחשב למילוי ערכים חסרים, או על מודלים שנבחן בהמשך. במקרה של משך הזמן הכולל, עלתה לנו גם מחשבה שבסשנים מסוימים, עלולים לקרות מצבים שיוזר משאיר את הסשן פתוח כאשר הוא בכלל אינו נמצא, ובגלל זה יכול להיווצר ערך רעש חריג שלא נרצה בדאטה שלנו.

מילוי הערכים החסרים בעמודות הפיצ'רים השונים שמודדים זמן – שמנו לב שבכל סשן, הזמן בעמודי המוצר, עמודי היוזר, ועמודי האדמין תמיד מסתכם לזמן הכולל. לכן יכולנו למלא את הערכים הריקים האלה כאשר התצפית החמיצה רק אחד מערכי הזמן הללו. במקרה שהזמן הכולל בתצפית מסוימת היה שווה לאפס, ידענו למלא את הערכים החסרים של שאר הזמנים לאפס גם כן. כך הצלחנו למלא ערכים חסרים בלי לבצע שום אומדן ובצורה מיטבית.

מילוי ערכים חסרים בעמודה A – ראינו שלעמודה זו יש קורלציה גבוהה עם סוג המשתמש, לכן החלטנו לחזות את ערכי A בעזרת העמודה הזאת. חישבנו את החציון של עמודה A תחת הרמה "משתמש חדש" ותחת רמת "משתמש חוזר" ומילאנו את ערכי ה-A המתאימים בהתאם לסוג היוזר.

עמודות info_dur, info_page_num – במקור, תכננו להיעזר בקורלציות של הפיצ'רים ולחזות עם אחד את השני. במבט מעמיק, ראינו שבגלל שרוב הערכים בinfo_dur הם אפס, החלוקה לקטגוריות על סמך הפיצ'ר info_page_num (תצפיות עם מספר דפים גדול/קטן – ביחס לממוצע), מניבה את אותו החציון של משך הזמן (0), גם עבור תצפיות בקטגוריית מספר דפים גדול, וגם עבור קטגוריית מספר דפים קטן. לכן, החלטנו למלא את הערכים בצורה פשוטה - הכנסנו עבור מספר הדפים את הממוצע עבור ערכים חסרים, ואת החציון עבור משך הזמן. עבור משך הזמן הכנסנו את החציון כיוון שאינו סטטיסטי המושפע מערכים קיצוניים ומשך הזמן יכול להגיע לערכים גבוהים כי אינו חסום מלמעלה (קיים הסיכון שדיברנו עליו מוקדם יותר, יוזרים שאינם על המחשב אך הסשן פתוח).

מילוי העמודה closeness_to_holiday – בבחינה של עמודה זו ראינו שאחוז גבוה מהערכים שלה הם 0 (89.7%). במילוי הערכים החסרים ב-0, אחוז זה עלה ל-90.2%. זהו הבדל לא מאוד משמעותי ולכן החלטנו למלא את הערכים החסרים בצורה זו.

מילוי עמודת סוג המכשיר – בעמודה זו בחרנו למלא את הערכים החסרים בשכיח, זאת מכיוון שמספר ההופעות שלו הוא הגבוה ביותר ובפער גדול מהערכים האחרים. כמו כן, מספר הערכים החסרים לא היה גדול, כך שהנחנו שלא יהיה הבדל משמעותי בהתחשב בפער הנתון.

עמודות num_of_admin_pages and admin_page_duration – זיהינו קורלציה גבוהה בין עמודות אלו, לכן התאפשר לנו למלא אותן בצורה "סימטרית" על פי חלוקה של כל עמודה לשתי תת קטגוריות. תחילה חישבנו את חציון העמודה " num_of_admin_page" תחת הערכים הקטנים מהחציון בעמודה admin_page_duration ותחת הערכים הגבוהים מהחציון. כמו כן חישבנו את חציוני העמודה admin_page_duration באותו אופן כאשר החלוקה היא על פי העמודה השנייה. לאחר מכן מילאנו את הערכים החסרים בכל עמודה בחציונים שלה בהתאם לחלוקות. לבסוף מילאנו ערכים חסרים נותרים באמצעות החציון בעמודה admin_page_duration ובעזרת הממוצע בעמודה num_of_admin_pages.

עמודות num_of_product_pages, total_duration , product_page_duration – זיהינו קורלציה מאוד גבוהה (0.99) בין עמודות ה- product_page_duration" ו - total_duration. כמו כן, קורלציה גבוהה (0.85) בין עמודות num_of_product_pages ו- total_duration. חשבנו שההסבר לקורלציות אלו הוא שכאשר לקוח נמצא באתר, סביר להניח שהדבר שמעניין אותו הוא המוצרים עצמם, והם מהווים את עיקר חווית הגלישה. לכן, עמודי המוצר מסבירים כמעט את כל השונות של הזמן הכולל, ומאפשרים פרדיקציה כמעט מוחלטת. מכיוון שהקורלציות כל כך גבוהות, יכולנו ליצור קו רגרסיה איכותי עבור כל עמודה על בסיס נתוני עמודה שנמצאת בקורלציה גבוהה איתה וכך ליצור ניבויים עבור הערכים החסרים.

תחילה יצרנו מודל לינארי המבוסס על הערכים הקיימים בעמודות product_page_duration_minutes ו-num_of_product_pages. במודל זה השתמשנו לניבוי הערכים החסרים בעמודה product_page_duration_minutes. לאחר מכן מילאנו את הערכים החסרים הנותרים בעמודה product_page_duration_minutes באמצעות חציון העמודה. תהליך דומה ביצענו גם עבור חיזוי הערכים החסרים בעמודות total_duration and num_of_product_pages, כמובן בהתבסס על הקורלציות המתאימות. לבסוף, החלטנו לוותר על הפיצ'ר product_page_duration_minutes ולשמר את הפיצ'ר של הזמן הכולל, זאת על בסיס הקורלציה הגבוהה ביניהם והנחה כי הזמן הכולל מייצג היטב את זמן ההימצאות בעמודי מוצר. השארנו את total_duration מכיוון שאת חיזוי הערכים שלו ביצענו עם הקורלציה הטובה ביותר.

<u>סטנדרטיזציה</u>

בשלב זה נדרשנו לבצע טרנספורמציה לנתונים שלנו כך שיהיו באותה יחידת מידה. בחרנו בסטנדרטיזציה באמצעות ציוני תקן כיוון שהאלטרנטיבות האחרות עלולות לגרום עיוותים בדאטה עקב רגישות לערכים קיצוניים, למשל MinMax Scaling.

- את החלוקה לסט אימון וואלידציה ביצענו לפני הסטנדרטיזציה, ביחס של 80/20, ודאגנו לאמן את הנירמול רק על סט האימון, ולהחיל אותו על שני הסטים.

<u>בחירת המודל</u>

בשלב הראשון, התחלנו עם הרצת PCA. הרשנו לעצמנו לשמר 99% מהשונות ולא להוריד מימדים רבים מכיוון שבעקבות בחירות שעשינו בשלב העיבוד המוקדם, מספר הפיצ'רים ההתחלתי שלנו לא היה גדול במיוחד. בסך הכל ירדנו מ-18 ל-17 ממדים. נציין כי בשלב זה לא השתמשנו ב-forward selection מכיוון שהוא תלוי מודל והבנו שזמן הריצה יהיה ארוך מדי. תכננו להשתמש בסינון מהסוג הזה אחרי שנדע איזה מודל מתאים לדאטה שלנו. המודל הראשון שהרצנו הוא KNN, מודל פשוט ומהיר שמטרתו הייתה בעיקר לראות שאנחנו "בכיוון הנכון". הגדרנו את השורש של גודל המדגם בתור עוגן להיפר-פרמטר מספר השכנים והרצנו grid search על ערכים קרובים לאותו עוגן. למודל הטוב ביותר הצגנו את עקומות ה-ROC השונות שנוצרו באמצעות K-Cross-Validation. כמו כן, הצגנו ציון AUC לסט הואלידציה ולסט האימון (עבור סט האימון – כדי לבחון overfitting). את אותו התהליך עשינו עבור רגרסיה לוגיסטית, SVM, ורשתות נוירונים. בחרנו ערכי היפר-פרמטרים ראשוניים על סמך מוסכמות שחקרנו על כל מודל ולאחר מכן ביצענו ניסוי וטעיה עד שהגענו לביצועים הטובים ביותר. עבור רשתות נוירונים הצגנו גם Confusion Matrix בהתאם לדרישה, ודיברנו על החוזקות והחולשות של המודל. לאחר שעשינו את כל התהליך, ראינו שהמודל שלנו עובד טוב במיוחד עבור רגרסיה לוגיסטית ועבור רשתות נוירונים עם פונקציית הפעלה לוגיסטית. מכך הסקנו שהדאטה שלנו מופרד ליניארית בצורה טובה והחלטנו שזה צריך להיות הפוקוס שלנו.

- עוד ביסוס לכך שהדאטה מופרד ליניארית בצורה טובה היה שמודל SVM עם Kernel ליניארי נתן תוצאות טובות (עם זמן ריצה איטי). בנוסף, מודל העצים לא נתן ביצועים כל כך מרשימים, דבר שיכול לרמוז על דאטה שדורש הפרדה ליניארית.

ציוני AUC על סט הואלידציה בשלב ה-PCA: KNN (0.83), SVM(0.89), LR (0.89), NN (0.92).

החלטנו לנצל את העובדה שהדאטה שלנו מנורמל לאותה הסקלה, והשתמשנו ברגרסיה לוגיסטית על הנתונים המקוריים (טרום-PCA) על מנת למצוא את הפיצ'רים החשובים ביותר עבורינו. עשינו זאת באמצעות המשקולות שכל פיצ'ר קיבל – ככל שערכה המוחלט של המשקולת גדול יותר, כך נקבע שהפיצ'ר יותר חשוב לנו במודל (אם הפיצ'רים לא היו באותה הסקלה, המשקולות לא היו יכולות להביע חשיבות של פיצ'רים). התוצאות הטובות ביותר על סט הואלידציה התקבלו כאשר השתמשנו ברגולאריזציה של לאסו (AUC 0.90), שמאפשרת איפוס פיצ'רים לחלוטין. בחרנו את אותם הפיצ'רים שלא אופסו (נספח 5) והחלטנו שאלו יהיו הפיצ'רים שנשתמש בהם במודל הסופי. למעשה זהו ה-feature selection שעשינו, שהוביל אותנו לבחירת 8 פיצ'רים בודדים. כעת כשהיו לנו את הפיצ'רים שאנחנו רוצים, הרצנו שוב grid search על רשתות נוירונים, והפרמטרים הטובים ביותר נתנו לנו את הAUC המקסימלי שהצלחנו לקבל – 0.93 על סט הואלידציה. לכן החלטנו שרשתות נוירונים יהיו המודל שלנו. כעת כשההחלטה הייתה כבר בידינו, השתמשנו בכל הדאטה שיש לנו (train+validation) על מנת למצוא מודל סופי של רשתות נוירונים – גם כאן השתמשנו בgrid search, והחלטנו על מודל עם היפר פרמטריים סופיים:

```
{'activation': 'logistic',
 'alpha': 0.001,
 'hidden_layer_sizes': (55, 55, 55),
 'learning_rate': 'constant',
 'solver': 'adam'}
```

ציון AUC ממוצע שחושב באמצעות K Cross Validation על כלל הדאטה - 0.905.

<u>בניית pipeline וביצוע הפרדיקציה הסופית</u>

בשלב האחרון של הפרויקט, המרנו את מרבית קטעי הקוד הנדרשים לעיבוד הנתונים לכדי פונקציות, כך שנוכל להריץ אותם באופן רציף ומסודר על קובץ האימון, וכמובן להחיל את השינויים הנדרשים על סט המבחן. חשוב לציין שבסט המבחן השלמנו ערכים חסרים על פי הערכים התואמים לכל עמודה מסט האימון (כמו ממוצע, חציון, שימוש במודל לינארי וכו..), זאת כדי לאפשר פרדיקציה טובה גם עבור מדגם קטן במידת הצורך וכמובן מתוך הנחה כי סט האימון שלנו מייצג את האוכלוסייה בצורה טובה. נקודה חשובה נוספת הייתה הסטנדרטיזציה באמצעות StandardScaler- דאגנו לאמן אותה באמצעות סט האימון בלבד ולהחיל אותה על שני הסטים.

לאחר שסיננו את הפיצ'רים הסופיים שהחלטנו לעבוד איתם בשני הסטים, אימנו את המודל הסופי שלנו על סט האימון וביצענו פרדיקציה סופית על סט המבחן. על פי החיזוי הסופי ניתן לראות מה ההסתברות של כל משתמש המופיע בסט המבחן שלנו לבצע רכישה באתר.

<u>סיכום</u>

בפרויקט עסקנו בפתרון בעיית קלסיפיקציה בינארית בעולם האמיתי. הייתה לרשותנו כמות דאטה עצומה, שהתבססה על מאפיינים שונים (פיצ'רים) של סשנים שונים (תצפיות) באתר קניות. בכל סשן, היה פרט אחד מרכזי שעניין אותנו במיוחד – האם בוצעה רכישה או לא. כל מטרתנו בפרויקט הייתה להשתמש בכל המידע שעומד לרשותנו, על מנת ללמוד לחזות באילו תנאים מתקיימת רכישה ובאילו תנאים לא – מה מאפיין רכישה? מה מונע אותה? יצאנו לדרך עם רצון לבנות את המודל שיצליח להשיב על השאלות הללו בצורה מיטבית, וקיווינו שלא רק המודל יצליח ללמוד את החוקיות הזאת, אלא גם אנחנו באופן אישי. השתמשנו בטכניקות מגוונות על מנת להמיר משתנים שונים לצורה נומרית, ולמדנו את הדאטה שלנו מכל מיני היבטים – כיצד כל פיצ'ר מתפלג? איזה פיצ'ר משתנה עם איזה פיצ'ר? מי הם הערכים הקיצוניים של כל פיצ'ר? באמצעות הלמידה הזאת, הצלחנו להשלים ערכים חסרים שונים בצורה מיטבית, להבין מי אלו התצפיות עליהן לא נרצה להסתמך, ואפילו לקבל אינטואיציות אישיות מסוימות לגבי הדאטה. לבסוף, לאחר שהכנו את הדאטה שלנו בצורה שראינו לנכון, נרמלנו את סקאלת הפיצ'רים וסקרנו מודלים שונים. באמצעות אותה סקירה, הצלחנו להגיע לתובנות משמעותיות – הבנו כי הדאטה שלנו ניתנת להפרדה ליניארית בצורה טובה, הבנו מי המודל המתאים ביותר לביצוע המשימה אליה התבקשנו, ואפילו הצלחנו להבין מי אלו הפיצ'רים המשמעותיים לחיזוי רכישה בעצמנו, מידע שייחסנו לו חשיבות רבה. לאחר שבחרנו את המודל הטוב ביותר עבורנו (רשתות הנוירונים) וביצענו אופטימיזציות על היפר פרמטרים שונים (באמצעות grid search), הצלחנו להגיע לתוצאת AUC מקסימלית של 0.93 על סט הואלידציה.

נספחים

חלוקת עבודה:

נשמח לציין כי כל העבודה נעשתה בשיתוף, התדיינות והתייעצות. בתחילת כל שלב עשינו חקר משותף וקיבלנו יחד החלטות נדרשות – איך לטפל בכל פיצ'ר, כיצד למלא ערכים חסרים, באילו מודלים כדאי שנשתמש וכו'. להלן החלוקה מבחינת כתיבת העבודה:

1. המרה של פיצ'רים שאינן נומרים:
   - גיא – id ,D ,A ,C ,internet_browser
   - יעל – month ,user_type.
2. טיפול בערכים חסרים:
   - גיא – מחיקת תצפיות עם רוב חסר, השלמת עמודות זמנים על ידי משלימים (באשר אפשר להסיק את הערכים על סמך הזמנים האחרים, ללא פרדיקציה), sin/cos_month,closeness_to_holiday , מילוי עמודות נותרות בערך החציוני.
   - יעל – עמודה A, product_page_duration_minutes ,num_of_product_pages ,total_duration_minutes
   - בשיתוף – admin_pages ,info_pages
3. נורמליזציה – יעל
4. צמצום מימדים ובניית מודלים:
   המודלים המוצגים – גיא
   מודל עצים (בחרנו לא להכניס) - יעל
5. המרת קטעי הקוד לפונקציות, ובניית pipeline סופי בהתאם להחלטותינו משלב המחקר ולמודל הסופי - יעל
6. כתיבת דוח מסכם – עבודה משותפת.

נספח 1 – ניתוח עמודת A:

analyzing the results:

c_2 (3101), c_1 (1944) and c_3 (1644) are the most dominant categories by far, so we would want to lose minimal information about them.

c_2 (21.7%) has distinguished purchase percentage compared to c_1 (10.8%) and c_3 (9.06%).

after looking at the other categories, we noticed that most categories have similar information compared to either c_2 or c_1/c_3, so we could use it for grouping.

The only category with considerable observations with different behaviour is c_4 (850,14.47%).

We considered grouping the data to three values: c_2, c_1 (because it is more dominant then c_3) and c_4, but after giving it some thought, we realized no other categories should be labeled as c_4 (it would make more sense to label them as c_2 or c_1), so the c_4 grouping would include only c_4 itself.

If we could avoid the c_4 grouping, we could avoid using dummy variables altogether and protect ourselves from the curse of dimensionality, we could make the A column into a binary column based on 2 groupings. Taking all of that into consideration, we decided to use the following grouping:

1. c_2 grouping - includes c_2 (3101, 21.7%), c_10 (347, 20.17%), c_others(294, 20.41%), c_8 (268, 27.61%), c_5 (205, 22.93%), c_11 (199, 18.09%).

- c_5 gives slightly higher percentage than c_2, c_11 gives slightly lower percentage than c_2, so intuitively, it cancels out.
- c_10 and c_others gives a lower percentage (than c_2), and c_8 gives a higher percentage, while it does not cancel out (because c_8 is more dominant) we thought that considering the low number of observations, there won't be a high price to pay, information-wise.

2. c_1 grouping - includes c_1 (1944,10.8%), c_3(1644,9.06%), c_4(850,14.47%), c_13(586, 6.14%), c_6(335, 11.34%).

- c_4 and c_13 roughly cancel out. c_13 pulls harder (4.6% lower than c_1) but has less observations, while c_4 pulls less (3.6% higher) but has more obvervations.
- c_3 and c_6 don't exactly cancel out, but we thought it is negligible.

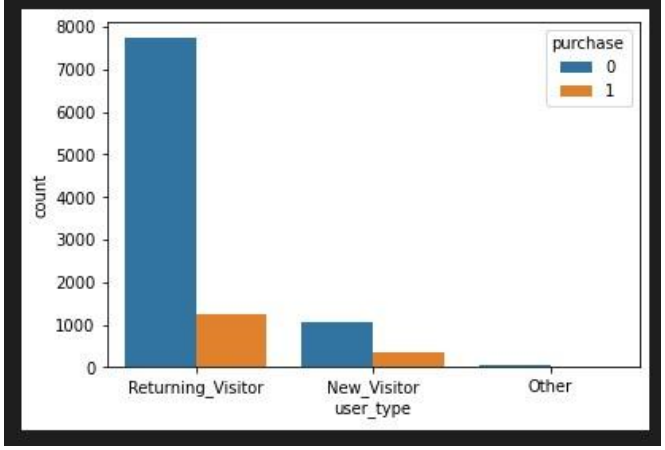
c_2 grouping - numeric value 1

c_1 grouping - numeric value 0

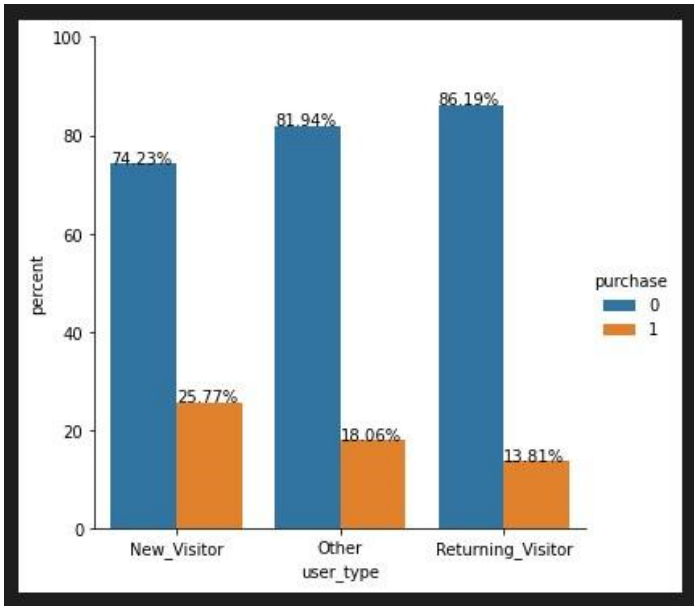
נספח 2 – ניתוח עמודת User_Type:

יחסים מספריים

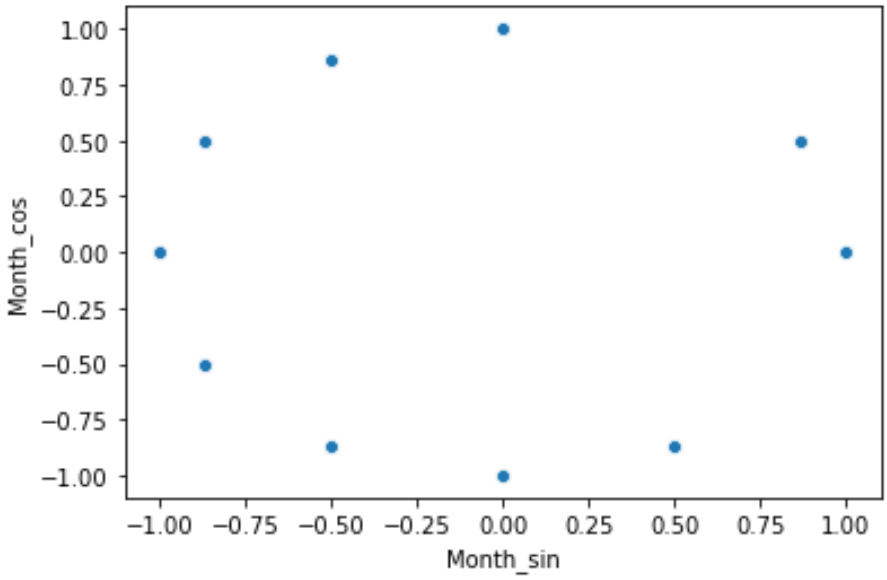
```
Returning_Visitor    8956
New_Visitor          1428
Other                  72
Name: user_type, dtype: int64
```



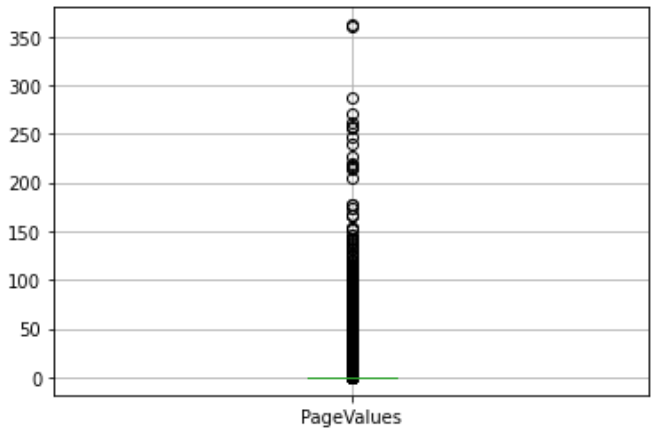
אחוזים

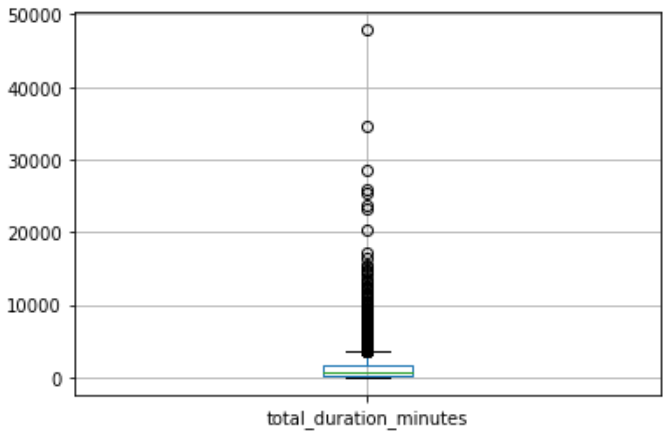


נספח 3 – מעגל ייצוג לחודשי השנה

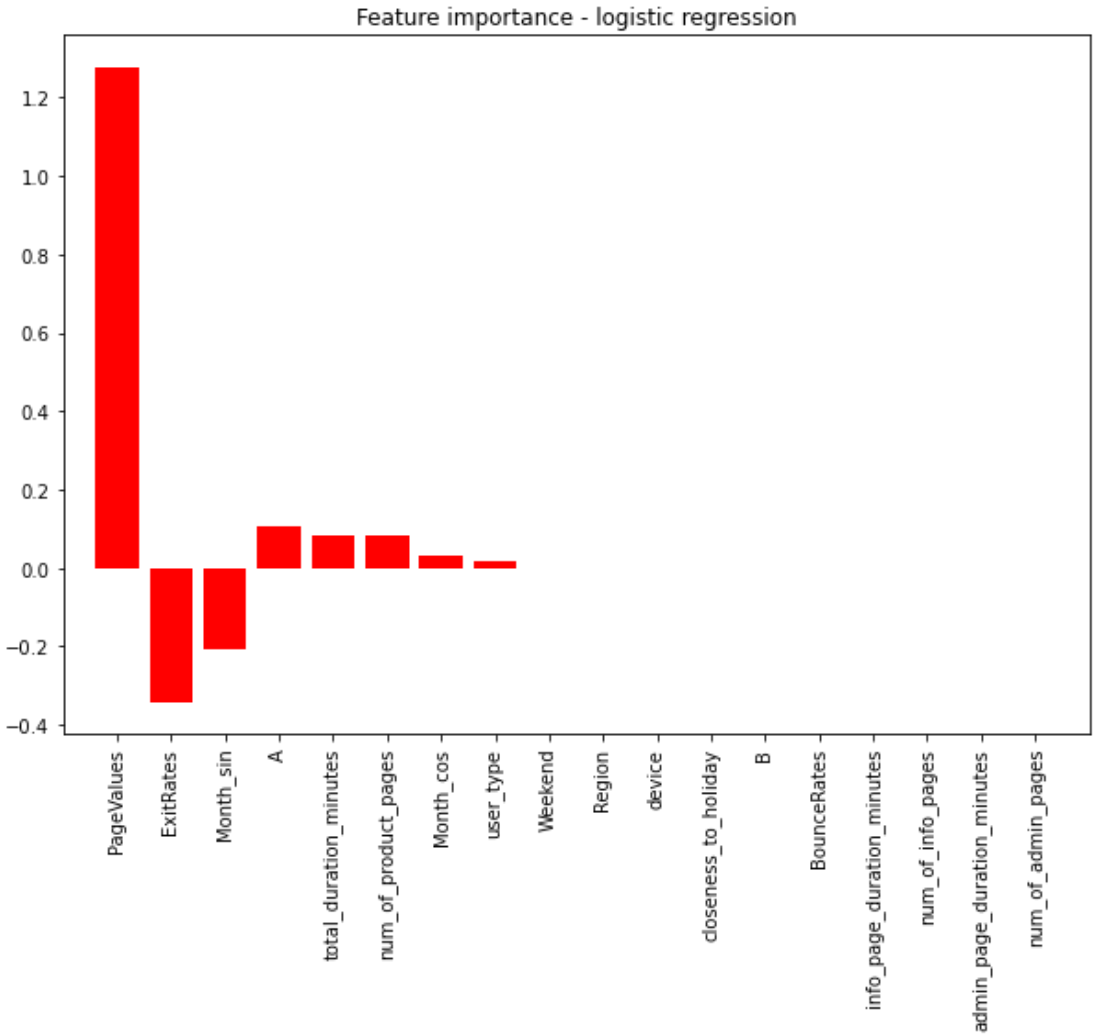


נספח 4 – הסרת ערכי קיצון





נספח 5 – הפיצ'רים המשמעותיים ביותר על מנת לחזות רכישה



נספח 6 – ציון AUC סופי על סט הואלידציה