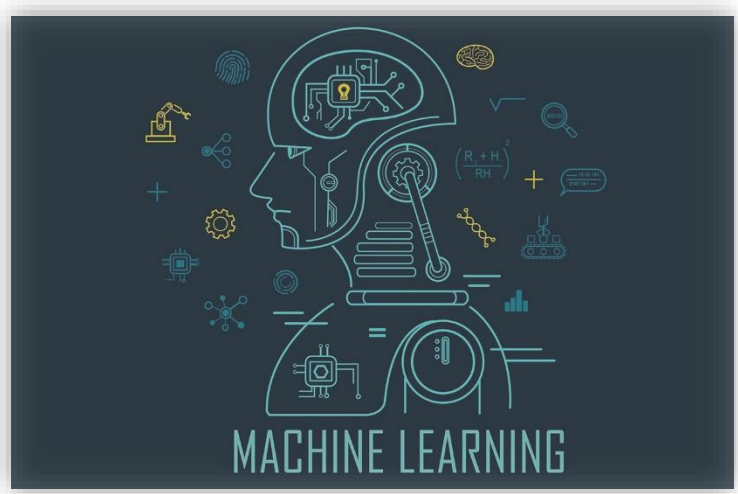
אוניברסיטת תל אביב

הפקולטה להנדסה

מדעים דיגיטליים להיי- טק

<u>פרויקט למידת מכונה - דוח המסכם</u>



מרצה: מר דור בנק

מתרגל: מר אילן וסילבסקי

מגישים:

ניצן כהן

ינאי איתם ברדוש

09.06.2022

סמסטר ב', תשפ"ב

תחילה קיבלנו שני קבצי דאטה, כאשר אחד הוא אימון והשני מבחן. ביצענו על שניהם עיבוד מקדים, כאשר בעיקר חקרנו את קובץ האימון ויישמנו על קובץ המבחן. לאחר מכן, אימנו את קובץ האימון בארבעה סוגי מודלים שונים כאשר פיצלנו את דאטה האימון לtrain וvalidation. לבסוף, בחרנו את המודל שהשיג את הביצועים הטובים ביותר לשם הפרדיקציה.

<u>הסתכלות ראשונית על המידע</u>

ראשית, ייבאנו את כל החבילות הדרושות עבור ביצוע המודלים והפרויקט. לאחר מכן, הסתכלנו ממבט-על על המידע שקיבלנו, בדקנו קורלציות בין כל המשתנים (**ראו נספח 1**)- מתוך כך ראינו כי משתנה PageValues יחסית מתואם עם purchase. בנוסף, ראינו כי עמודה D מתואמת בקשר שלילי חזק מאוד עם משתנה purchase. ניתן לראות גם כי משתני הduration pages וה-num of pages למיניהם די מתואמים אחד עם השני.

רצינו לראות איך העמודות המספריות מתפלגות, ושמנו לב כי יש הרבה יותר אי-רכישות מאשר רכישות (**ראו נספח 2).** ראינו גם כי יש הרבה משתנים אשר מתפלגים דומה להתפלגות חי-בריבוע כלומר בתחילת ההתפלגות ישנם הרבה ערכים שקרובים לאפס וככל שהערך עולה מספר התצפיות יורד.

<u>ארגון המידע</u>

<u>התמודדות עם עמודות מספריות</u>

1. תחילה, בנינו פונקציה אחת שמותאמת לכל עמודות הpage למיניהם בשביל מילוי ערכים חסרים על פי השלבים הבאים:

א. בעמודות הduration החסרות מילאנו את עמודות הnum שמתואמות להם באפס. לדוגמא אם ישנה שורה של ערך חסר בעמודה 'page_duration_**product**', אז בעמודה 'num_of_**product**_pages' מילאנו אפס.

ב. הסרנו מהערכים שבעמודות ה-duration את המילה 'minutes' כדי להפוך אותה לעמודה מספרית.

ג. הנחנו כי אם בעמודת num יש ערך אפס, משמע המשתמש לא היה באף עמוד, אז בעמודת duration שמתאימה לה יהיה ערך אפס.

ד. הנחה נוספת שהנחנו היא שסכום עמודות הduration של שלל סוגי הדפים, צריך להיות שווה לtotal_duration.

ה. בעקבות הנחה 4, ובנוסף שהחלטנו למלא את עמודת 'product_page_duration' אחרונה, מילאנו את הערכים החסרים שלה כהפרש בין total_duration ל-'info_page_duration' ו-'admin_page_duration'.

ו. כעת, מילאנו ערכים חסרים בעמודות הduration לפי חציון הduration שמותאם ל-num שלו. לדוגמא, עבור תצפית שעמודת num_of_product = 5, ועמודת 'product_page_duration' בעלת ערך חסר. ערך ה- 'product_page_duration' תמולא על פי החציון של 'product_page_duration' כאשר num_of_product = 5.

ז. יצרנו רשימה של חציוני duration, אשר המיקום ברשימה הוא גם הערך שיהיה בעמודת הnum של העמוד. לדוגמא, במקום האפס ברשימה נמצא החציון של עמודת הduration כאשר ה-num=0, במקום האחד ברשימה נמצא החציון של עמודת הduration כאשר ה-num=1 וכך הלאה.

ח. הנחנו כי כאשר עמודת ה-duration שווה לאפס, המשתמש לא גלש בעמודים מסוג זה ולכן מספר העמודים יהיה אפס.

ט. בהמשך מילאנו את הערכים החסרים שבעמודות הnum בעזרת הרשימה שיצרנו בסעיף 7. ראשית, בדקנו איפה הערך הספציפי שאנחנו עובדים איתו נמצא ברשימת החציונים. כלומר אם אנחנו עובדים עם ערך duration ששווה 4.7, אז נצמצם את הבדיקה שלנו לחציוני duration בין 4-5. שנית, חישבנו את הערך הממוצע של החציוני duration המתאימים ובדקנו האם ערך הduration הספציפי שלנו

מעל/מתחת הממוצע. אם הוא מעל, שייכנו אותו לחציון הnum הגבוה יותר ואם הוא מתחת אז שייכנו אותו לחציון הnum הנמוך יותר.
למשל, בדוגמא עם ערך duration ששווה 4.7, ערך זה ממוקם בין החציונים שבמיקומים 3-4, ומאחר ו4.7 גדול מ4.55 (ממוצע של 3.1 ו-6) נשייך ערך זה לחציון שבמיקום 4, כלומר נמלא את הערך החסר ב-4 (האינדקס המותאם לו).

Num of pages/index	0	1	2	3	4	5	6	7	8
Median duration	0	2	2.3	3.1	6	9.5	11.2	20.4	40.7

י. מקרה קצה שהבחנו בו היה כי כאשר ערך הduration הספציפי שלנו קטן מכל החציונים, בפרט ערך זה קטן מהחציון של כל התצפיות כאשר num=1 ולכן נמלא ערך num זה באחד. נזכיר כי אנחנו מניחים שרק אם ה-duration=0 אז נמלא את הnum באפס.

יא. מקרה קצה נוסף- לאחר סיום מילוי הערכים, מי שנשאר לו ערך חסר אז הduration שלו היה גבוה מידי כדי להיות בין חציוני הרשימה ולכן מילאנו את הnum שלו בnum המקסימלי שברשימה.

2. כעת, לאחר שסיימנו להריץ את הפונקציה ומילאנו את הערכים החסרים, גילינו כי יש 3 מקרי קצה ספציפיים מאוד ומילאנו אותם לפי ממוצע הערכים הקרובים אליהם.

3. בעמודת total_duration מילאנו את הערכים החסרים לפי הנחה בסעיף 4 בסכום הבא: 'admin_page_duration'+'product_page_duration'+'info_page_duration' .

4. בעמודות 'BounceRates','ExitRates',PageValues','device','B' מילאנו את הערכים החסרים לפי החציון שלהם.

5. שמנו לב כי בעמודה 'D' יש המון ערכים חסרים. רצינו להציג זאת ויזואלית **(ראו נספח 3)** ולכן יצרנו משתמש דמי שיגיד האם קיים/לא קיים ערך D בתצפית. קל לראות מהגרף שיצרנו כי יש יותר מידי ערכים חסרים. מהסיבה שכל מילוי אפשרי של העמודה יכול להפריע לחיזוי, בחרנו להסיר עמודה זו מהדאטה.

6. לגבי עמודה 'id', אנו מניחים כי אין קשר בינה לבין חיזוי הרכישה. אנו מניחים זאת מהסיבה שעמודה זו ממוספרת לפי סדר כרונולוגי של מספר השורות, כלומר אין לה קשר לרכישה/לא רכישה ולכן בחרנו להסיר אותה מהדאטה.

<u>התמודדות עם עמודות קטגוריאליות</u>

קודם כל, ביצענו כמה גרפים **(ראו נספחים 4,5)** כדי לזהות את הקטגוריות הנפוצות בחלק מהעמודות. לאחר מכן, יצרנו פונקציה שמותאמת לכל העמודות הקטגוריאליות שבה היא מחליפה את ערכיהם למספרים על פי השכיחות שלהם ואז ממלאה את הערכים החסרים בחציון העמודה. כאשר הערך הכי נפוץ מקבל את הערך אחד, אחריו שתיים והלאה. לדוגמא בעמודת ה'internet_browser' ראינו כי הדפדפן chrome_89 הכי נפוץ, ולכן הערך החדש שלו יהיה אחד.

שמנו לב כי בעמודת 'closeness_to_holiday' , העמודה מספרית אך לא רציפה ולכן התייחסנו אליה כאל עמודה קטגוריאלית. רצינו למלא את הערכים החסרים לפי החציון של העמודה בהתאם לחודש של התצפית, כי הנחנו שהקרבה לחג תלויה בחודש שבו אנחנו נמצאים. כאשר בדקנו את החציונים של כל חודש, גילינו כי חציון העמודה בכל חודש הוא אפס, ולכן מילאנו את הערכים החסרים באפס (לא עשינו זאת על פי ממוצע משום שהיינו מקבלים ערכים לא שלמים).

בעמודת 'Region' הבנו כי למספרים יש משמעות, לכן לא היינו צריכים לעשות להם טרנספורמציה ומילאנו את ערכיהם בחציון העמודה.

בעמודת 'Weekend' ראינו כי עמודה זו היא מסוג בוליאני, אז החלפנו את ערכיה לאפס ואחד כאשר אפס מייצג False ואחד מייצג True ומילאנו את הערכים החסרים עם חציון העמודה.

<u>התמודדות עם ערכים קיצוניים</u>

הבחנו כי עמודה 'B' היא היחידה שמתפלגת דומה להתפלגות נורמלית (**ראו נספח 8**), ולכן בעמודה זו בחרנו להשתמש בשיטת IQR אשר מורידה ערכים קיצוניים משני קצוות ההתפלגות.

בנוסף, ניתן לראות מהגרפים **(ראו נספחים 9,10)** והשונויות שיצרנו, כי לעמודות 'ExitRates' ו-'BounceRates' יש שונות נמוכה, והרבה מידע בקצוות התפלגות הנתונים. השערתנו היא במידה ונוריד מהן ערכים קיצוניים תפגע יכולת החיזוי של המודלים.

<u>בדיקה ראשונית להשערת העמודות החשובות לחיזוי</u>

באמצעות **נספחים 11-16** הצגנו את ההשערות שלנו לחיזוי העמודות החשובות בהן נרצה לשים דגש.

בנספח 11 הבחנו בהתפלגות הדומה של שתי העמודות: 'product_page_duration' ו- 'total_duration'. **בנספח 12** מוצגות כמות הרכישות אל מול אי הרכישות בכל חודש. נקודה מעניינת היא כי החודש עם מספר הרכישות הגבוה ביותר, אינו החודש עם אי ביצוע הרכישות הגבוה ביותר. **בנספח 13**, ניתן לראות את הקשר בין 'ExitRates' ל'BounceRates' אל מול רכישות. כאשר ערכי שתי העמודות הללו קטן מ0.035 יש יותר סיכוי שתבוצע רכישה. **בנספח 14**, מתואר הקשר בין 'Weekend' אל מול רכישות, ניתן לראות כי בזמן סופ"ש יש יותר סיכוי שתבוצע רכישה. **בנספח 15**, מתואר הקשר בין 'PageValues' ל'total_duration' אל מול רכישות. כאשר ערכי שתי העמודות עולה (בעיקר PageValues) אז יש יותר סיכוי שתבוצע רכישה. **בנספח 16**, בדקנו את התפלגות עמודה 'B' בגרף מסוג היסטוגרמה, שאיפשר להסיק כי ההתפלגות של העמודה נורמלית בקירוב.

<u>הורדת מימדים</u>

בחרנו להשתמש בשתי שיטות להורדת מימדים. האחת, שיטת **PCA** – שאינה תלויה בפרמטרים. נרמלנו את הנתונים ולאחר מכן ביצענו את השיטה פעם אחת לפני ביצוע המודלים. השנייה, שיטת **Forward Selection**- שכן תלויה בפרמטרים.

לאחר מכן, בנינו פונקציה כללית אשר תכריע איזה שיטה הפחתת מימדים טובה יותר עבור מודל- PCA/Forward Selection תוך התחשבות במדד MSE.

<u>ROC K-Fold</u>

פונקציה זו מקבלת מערך דו מימדי עם כל הפיצ'רים, מערך חד מימדי של ה-labels ואת המסווג שתלוי במודל. תחילה חילקנו את הדאטה ל-10 חלקים. אימנו את הדאטה בלולאה ויצרנו Roc Curve לכל פיצול דאטה. חישבנו את הממוצע של הציונים ואת סטיית התקן שלהם והצגנו זאת בגרף.

<u>מודלים</u>

<u>רגרסיה לוגיסטית</u>

תחילה, ביצענו את שיטת הורדת המימדים-Forward Selection, (**ראו נספח 17**). בהמשך, בדקנו איזה שיטת הורדת מימדים תביא ל-MSE קטנה יותר. ראינו כי ה-MSE בשיטת ה-Forward Selection הייתה נמוכה יותר. אחרי כן, פיצלנו את הדאטה לtrain ולvalidation כך שהvalidation מהווה 20% מהדאטה, ויצרנו פונקציה שבוחרת את ה-C האופטימלי ברגרסיה (**ראו נספח 18**), שמייצגת את אחד חלקי כוח ההענשה. לאחר מכן, יצרנו מילון של היפר-פרמטרים ובעזרת GridSearchCV לקחנו את ההיפר-פרמטרים האופטימליים.

ההיפר-פרמטרים שבחרנו לבדוק הם: **(1) Penalty**- פרמטר זה אחראי על סוג ההענשה של הרגרסיה. **(2) solver**- פרמטר זה אחראי על סוג האלגוריתם שדרכו אנחנו מגיעים למשקולות האופטימליות. **(3) Max_iter**- פרמטר זה אחראי על מקסימום מספר האיטרציות. **(4) Fit_intercept**- פרמטר זה בודק האם חותך המשוואה שייך אליה או שזה רעש.

הפרמטרים שנבחרו ע"י הGridSearchCV (<u>בהתאמה</u>): **(1)** סוג ההענשה שנבחר הוא l1 כלומר סוג הענשה לפי שיטת lasso. **(2)** סוג האלגוריתם שדרכו אנחנו מגיעים למשקולות האופטימליות הוא מסוג 'saga'. **(3)** מקסימום האיטרציות = 100 **(4)** Fit_interscept = False.

C = 0.1 (היישום בוצע ע"י הפונקציה שבוחרת את ה-C האופטימלי ברגרסיה).

לאחר מכן, בנינו פונקציה אשר מציגה את ה confusion matrix (**ראו נספח 19,20**).

1. ערכי מטריצת **הtrain** הם: TN=6874, TP=496, FP=144, FN=808. כלומר, צדקנו כאשר סיווגנו לאי-רכישה ב6874 תצפיות וצדקנו כאשר סיווגנו לרכישה ב496 תצפיות ולכן **דיוק מטריצה האימון שלנו היא 0.88.**
2. ערכי מטריצת **הvalidation** הם: TN=1745, TP=105, FP=34, FN=197. כלומר, צדקנו כאשר סיווגנו לאי-רכישה ב1745 תצפיות וצדקנו כאשר סיווגנו לרכישה ב105 תצפיות ולכן **דיוק מטריצה ה-validation שלנו היא 0.88.**

ניתן להסיק מהconfusion matrix כי ברוב המקרים צדקנו כשחזינו אי-רכישה. בנוסף, יש לשים לב שהריבוע שמתייחס לכך שחזינו נכונה רכישה אינו כהה, משום שאנחנו מתייחסים במטריצה לגודל אבסולוטי ולא יחסי וזאת בהתאם לכך שאין הרבה רכישות בדאטה. מפני שTP גדול משמעותית מFP אנחנו מסיקים כי המודל שלנו פעל בצורה טובה.

ניתן לראות מנספחים **21,22** כי ממוצע הROC-KFOLD לרגרסיה הלוגיסטית הוא 0.89 על הtrain ו0.88 על הvalidation.

מתוך ביצועי מודל אלו, אנו מסיקים כי המודל שלנו הוא לא overfitted, משום שהtrain_score גדול מהvalidation_score. ניסינו להימנע מoverfitting כשהורדנו את מימדיות הדאטה, כלומר הקטנו את מספר הפיצ'רים תוך התחשבות במדד MSE ושימוש בcp mallow בשיטת forward selection. יש לציין כי מספר הפיצ'רים שנבחרו במודל הוא 9 שזהו מספר די נמוך, ולכן השונות של המודל יחסית נמוכה. בנוסף השתמשנו ברגולריזציה ושיטת ענישה אשר גם אמורים לעזור בהפחתת הoverfitting.

<u>מסווג נאיב בייס</u>

אנו מניחים כי התצפיות הן בלתי תלויות אחת בשנייה. תחילה, ביצענו את שיטת הורדת המימדים- Forward Selection, (**ראו נספח 23**) ולאחר מכן בדקנו איזה שיטת הורדת מימדים תביא ל-MSE קטנה יותר. מצאנו כי ה-MSE בשיטת ה-Forward Selection הייתה נמוכה יותר. בהמשך פיצלנו את הדאטה לtrain ולvalidation כך שהvalidation מהווה 20% מהדאטה.

ניתן להבחין מנספחים **25,24** כי ממוצע הROC-KFOLD לבייס הוא 0.85 על הtrain ו0.83 על הvalidation.

לדעתנו מודל זה לא overfitted משום שזהו מודל אשר חוזה לפי הסתברות, ובנוסף הtrain_score גדול מהvalidation_score. בנושא סיבוכיות זמן הריצה, הבחנו כי מודל זה מחשב את ההסתברות של כל פיצ'ר, לכן ככל שיש יותר פיצ'רים הסיבוכיות עולה. כדי להקטין את בעיית החישוביות של המודל הורדנו את מימדיות הדאטה, ובכך הקטנו את מספר הפיצ'רים תוך התחשבות במדד MSE ושימוש בcp mallow בשיטת forward selection. יש לציין כי מספר הפיצ'רים שנבחרו במודל הוא 4 שזהו מספר נמוך, ולכן הסיבוכיות של המודל יחסית נמוכה.

<u>Random Forest</u>

בחרנו שלא להשתמש בשיטת הורדת מימדים משום שבמודל זה, ככל שיש יותר עצים וככל שהם יותר עמוקים המודל יחזה יותר טוב. בהתאם לכך, לא ראינו בעייתיות שיכולה להיווצר ממספר הפיצ'רים. פיצלנו את הדאטה לtrain ולvalidation כך שהvalidation מהווה 20% מהדאטה. בהמשך, יצרנו מילון של היפר-פרמטרים ובעזרת GridSearchCV לקחנו את ההיפר-פרמטרים האופטימליים.

ההיפר-פרמטרים שבחרנו לבדוק הם: **(1) n_estimators**- מספר העצים ביער, **(2) criterion**- סוג הפונקציה שמודדת את השגיאה בחיזוי, **(3) Max_depth**- מקסימום עומק עץ, **(4) Min_samples_leaf**- מינימום תצפיות בעלה, **(5) Min_samples_split**- מספר תצפיות מינימלי שצריך לבצע עבורן עוד פיצול בעץ

הפרמטרים שנבחרו ע"י הGridSearchCV (<u>בהתאמה</u>): **(1)** n_estimators=200 , **(2)** criterion = 'entropy' כלומר הוא מודד את המידע המעיד על אי-סדר הפיצ'רים עם הlabel. **(3)** Max_depth=200, **(4)** Min_samples_leaf=5, **(5)** Min_samples_split=3.

לאחר מכן, הצגנו בגרף את חשיבות הפיצ'רים (**ראו נספח 26**). ניתן לראות כי בפער משמעותי הפיצ'ר 'PageValues' החשוב ביותר לביצוע הפרדיקציה כפי ששיערנו **(מנספח 15).** ממצא זה הגיוני משום שה'PageValues' זוהי עמודה אשר מייצגת את ערך הדף, וכאשר ערך הדף גבוה גדל הסיכוי לביצוע רכישה. עמודה חשובה נוספת אשר חזינו זוהי עמודה 'product_page_duration' **(מנספח 11)**. ניתן להסיק כי היא חשובה משום שככל שמשך השהייה גדלה בעמוד מסוג זה, כך הלקוח יותר מעוניין במוצר והסיכוי לרכישה גדל (במידה והלקוח אהב את המוצר כמובן). עמודה זו מתואמת חזק עם עמודת 'total_duration', ולכן גם 'total_duration' עמודה די חשובה.

ניתן לראות מנספחים **28,27** כי ממוצע הROC-KFOLD לבייס הוא 0.93 על הtrain ו0.90 על הvalidation.

מתוך ביצועי מודל אלו, אנו מסיקים כי המודל שלנו הוא לא overfitted, משום שהtrain_score גדול מהvalidation_score.

<u>Neural Networks</u>

במודל זה בחרנו שלא להשתמש בשיטת הורדת מימדים. PCA בחר לקחת 18 משתנים, וזה רק הבדל של 2 משתנים מהדאטה ללא הורדת מימדים ולכן השיטה לא יעילה. בחרנו שלא להשתמש בForward selection מטעמי סיבוכיות הריצה. לאחר מכן פיצלנו את הדאטה לtrain ולvalidation כך שהvalidation מהווה 20% מהדאטה. כעת, יצרנו מילון של היפר-פרמטרים ובעזרת GridSearchCV לקחנו את ההיפר-פרמטרים האופטימליים.

ההיפר-פרמטרים שבחרנו לבדוק הם: **(1) Hidden_layer_sizes**- משתנה זה קובע את מספר השכבות ומספר הפרמטרים בכל שכבה, **(2) Solver**- שיטת אופטימיזציה למשקלים, **(3) Activation**- סוג הפונקציה שנמצאת בשכבות.

הפרמטרים שנבחרו ע"י הGridSearchCV (<u>בהתאמה</u>) : **(1)** Hidden_layer_sizes- 3 שכבות של 100 משתנים, **(2)** Solver='adam'. **(3)** Activation ='logistic'.

ניתן לראות מנספחים **29,30** כי ממוצע הROC-KFOLD לבייס הוא 0.90 על הtrain ו0.89 על הvalidation.

מתוך ביצועי מודל אלו, אנו מסיקים כי המודל שלנו הוא לא overfitted, משום שהtrain_score גדול מהvalidation_score.

סיכום

בפרויקט זה הרצנו ארבעה סוגי מודלים: רגרסיה לוגיסטית, Neural Networks, מסווג נאיב בייס, Random Forest. בהתאם לביצועים של כל מודל, לזמן הסיבוכיות ול-overfitting, בחרנו במודל Random Forest אשר השיג את הביצועים הטובים ביותר לפי ROC K-FOLD. יחסית למורכבות המודל ולביצועים שלו, זמן הריצה תקין ולכן בחרנו בו בעבור הפרדיקציה.

נספח אחראיות כל שותף ותרומתו לעבודה

תחילה, נדגיש כי את הפרויקט ביצענו בסביבת עבודה Visual Studio Code תוך כדי שיתוף פעולה מלא אשר התאפשר באפליקציית הZoom, כך שחלקים נרחבים מכתיבת הקוד עשינו בצורה משותפת ובזמן אמת. עם זאת, חלקים קטנים חולקו בינינו כדי למזער את העומס:

ניצן כהן:

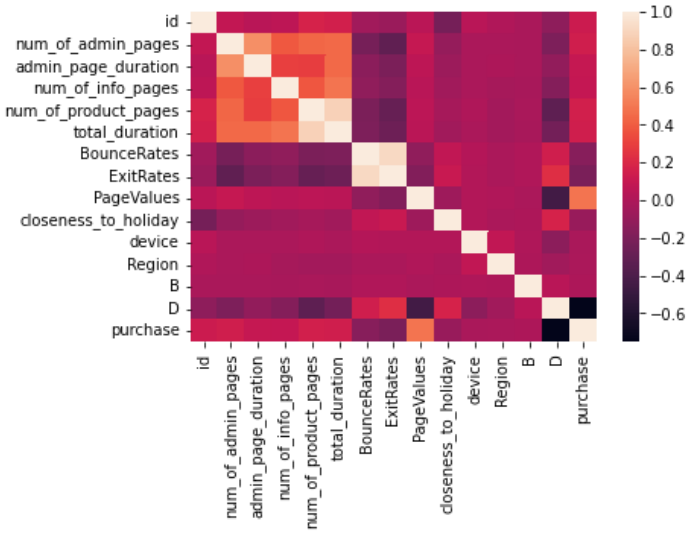
1. ניצן החליט על הצגת הוויזואליזציה שבחלק ביצוע המודלים.
2. ניצן רשם את ההערות שבmarkdown בחלק ביצוע המודלים.

ינאי איתם ברדוש:

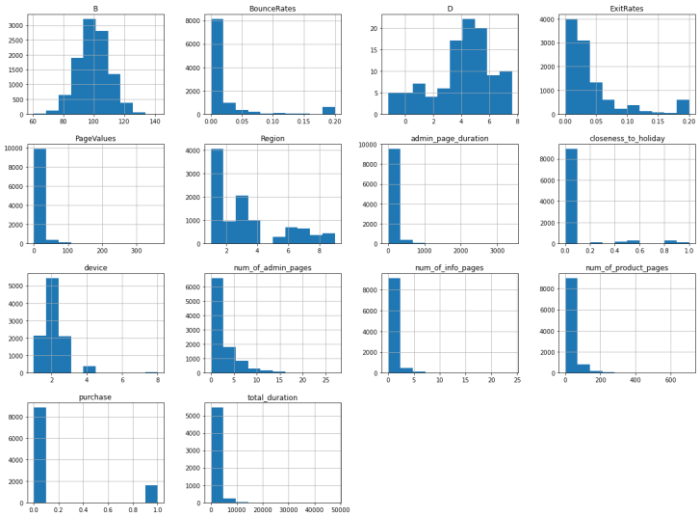
1. ינאי החליט על הצגת הוויזואליזציה שבחלק העיבוד המקדים.
2. ינאי רשם את ההערות שבmarkdown בחלק העיבוד המקדים.

נספחי ויזואליזציה

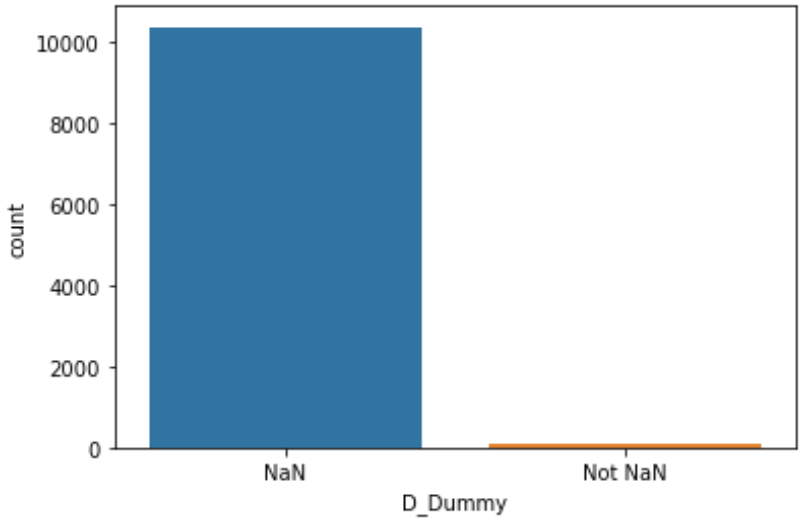
.1



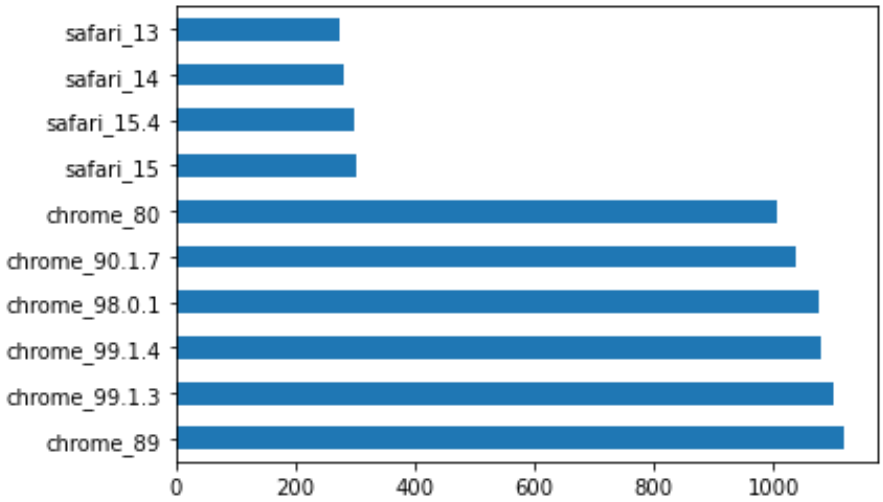
.2



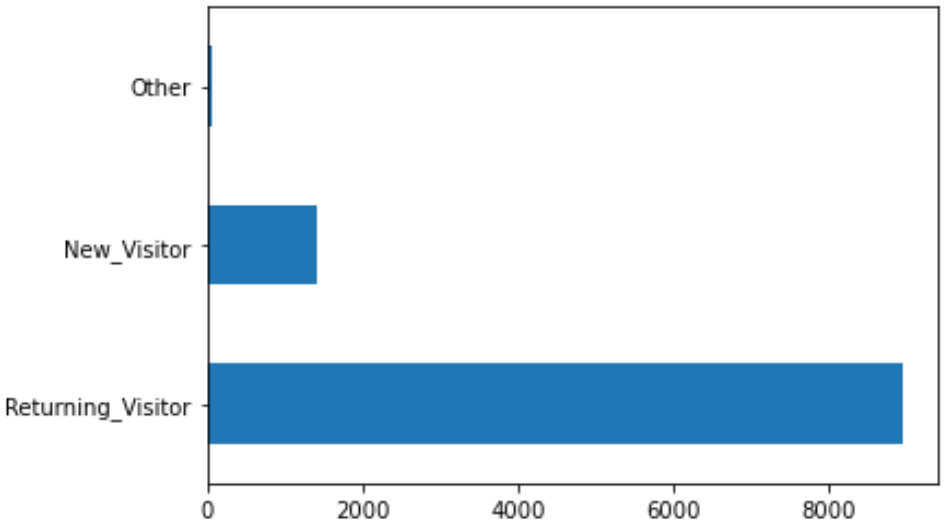
.3


10000
8000
6000
4000
2000
0
count
NaN
Not NaN
D_Dummy

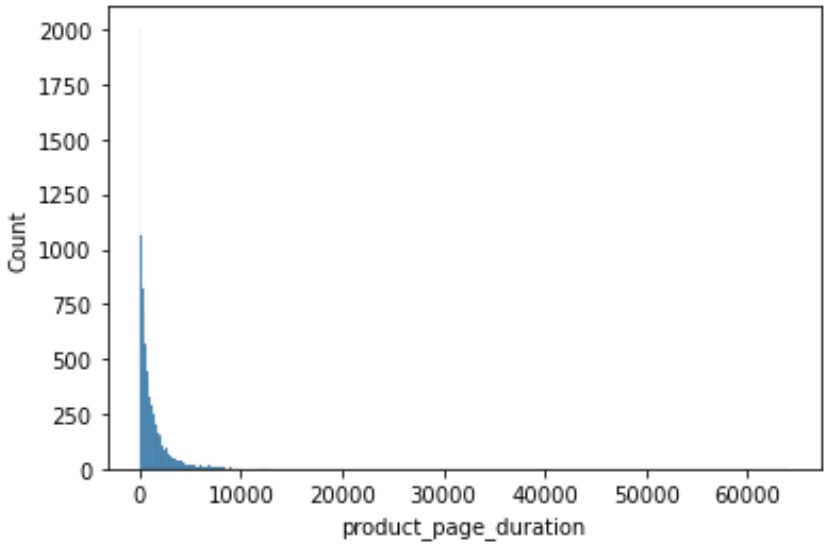
.4


safari_13
safari_14
safari_15.4
safari_15
chrome_80
chrome_90.1.7
chrome_98.0.1
chrome_99.1.4
chrome_99.1.3
chrome_89
0
200
400
600
800
1000

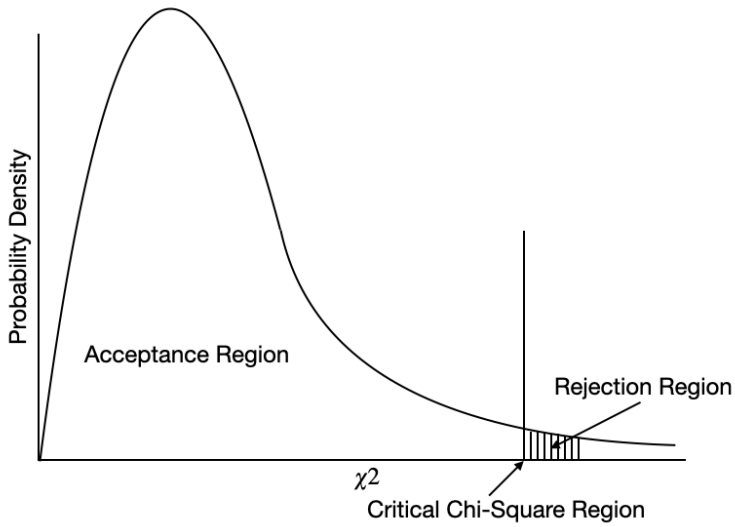
.5


Other
New_Visitor
Returning_Visitor
0
2000
4000
6000
8000

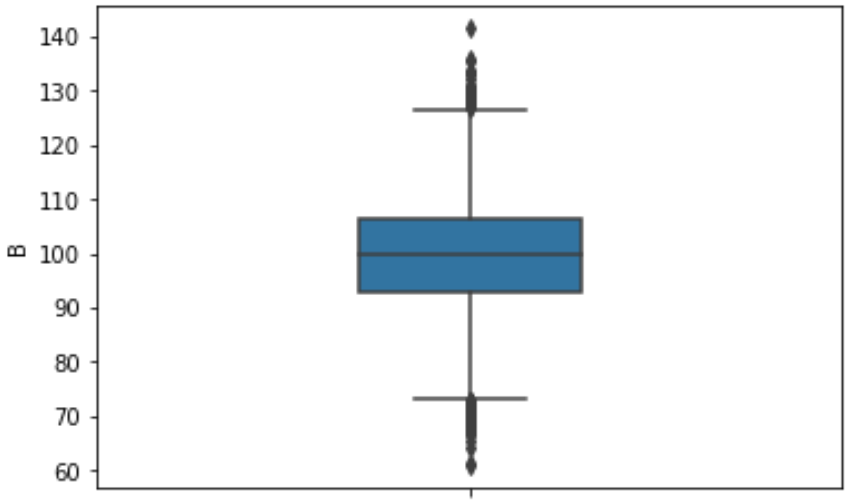
.6


2000
1750
1500
1250
1000
750
500
250
0
Count
0
10000
20000
30000
40000
50000
60000
product_page_duration

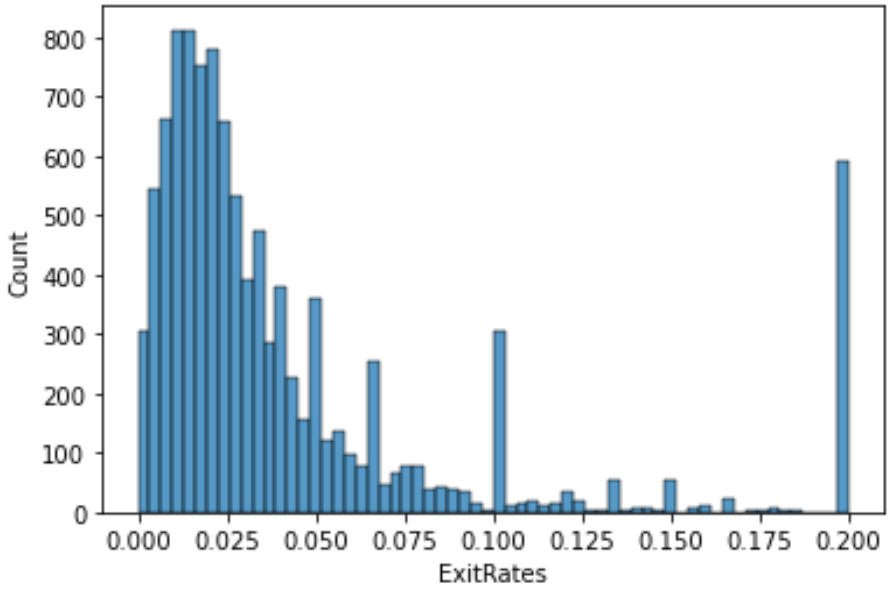
.7


Probability Density
Acceptance Region
Rejection Region
χ^2
Critical Chi-Square Region

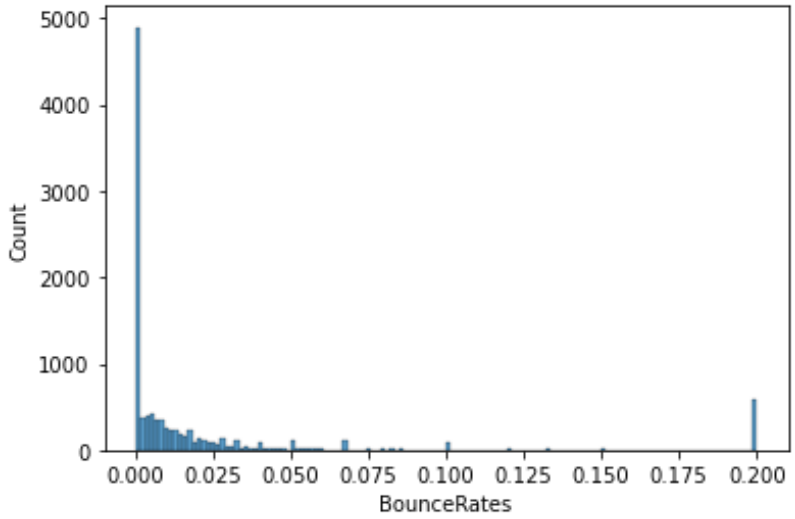
.8


140
130
120
110
100
90
80
70
60
B

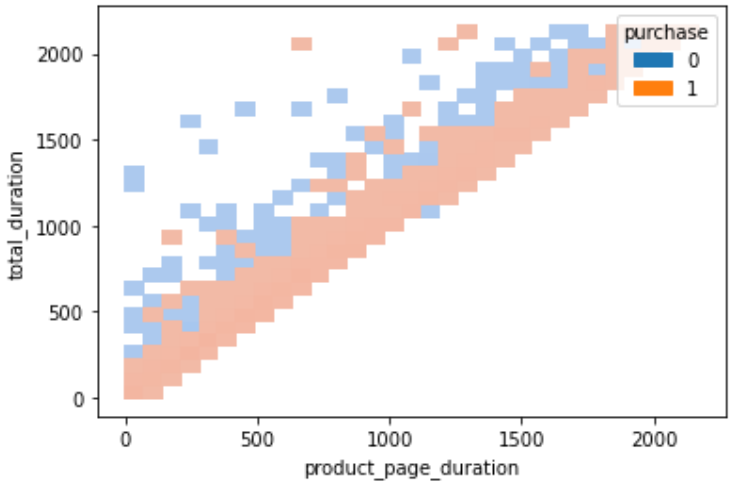
.9


Count
ExitRates
0
100
200
300
400
500
600
700
800
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

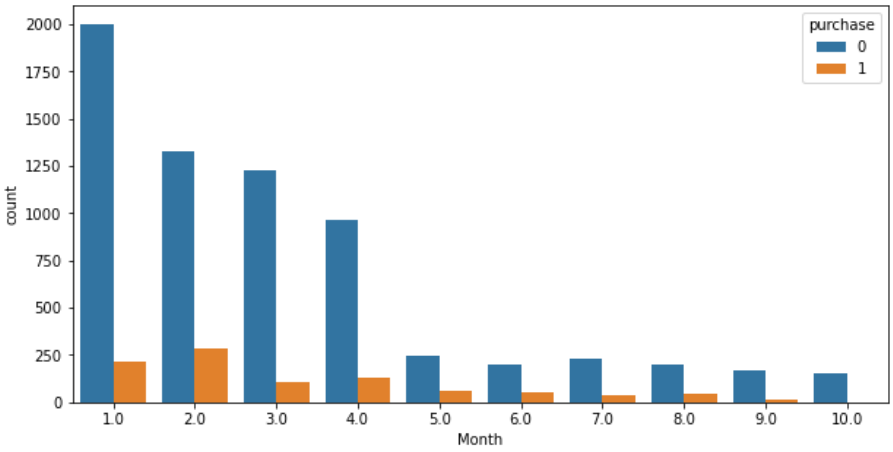
.10


Count
BounceRates
0
1000
2000
3000
4000
5000
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200

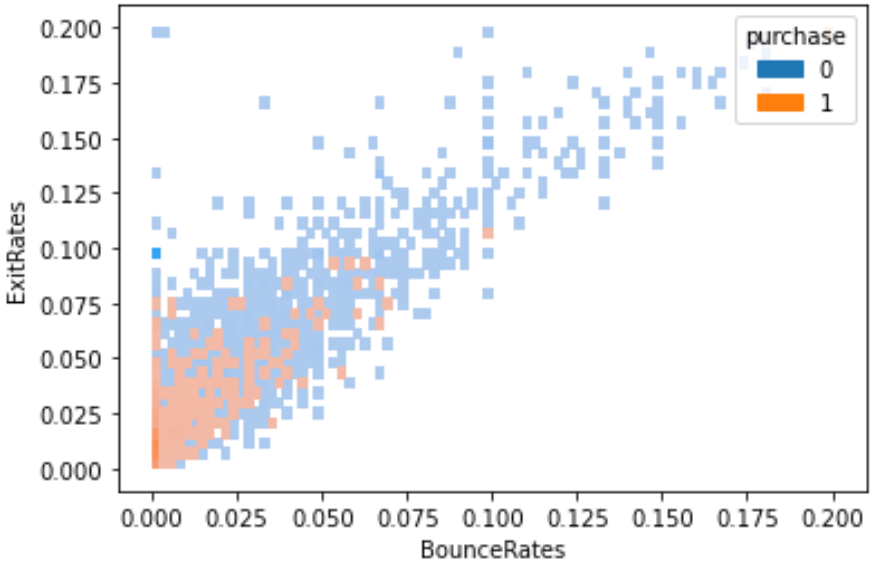
.11


purchase
0
1
total_duration
product_page_duration
0
500
1000
1500
2000

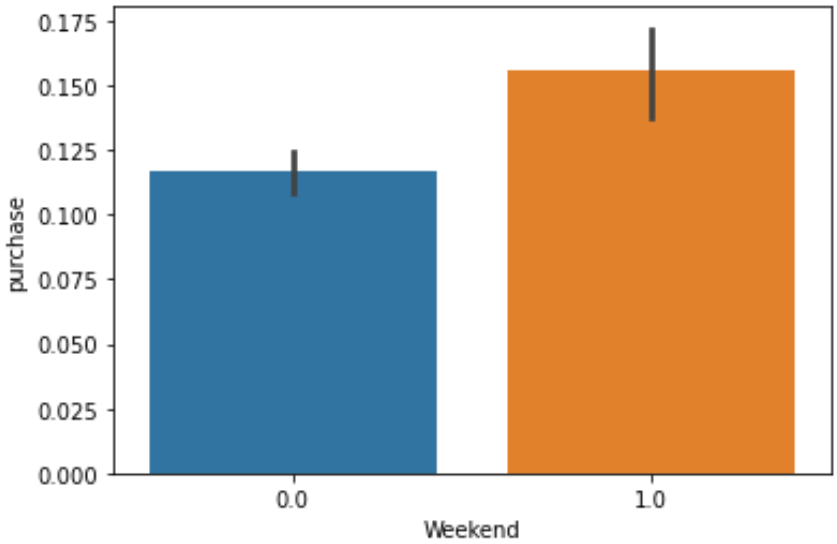
.12


purchase
0
1
count
2000
1750
1500
1250
1000
750
500
250
0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
Month

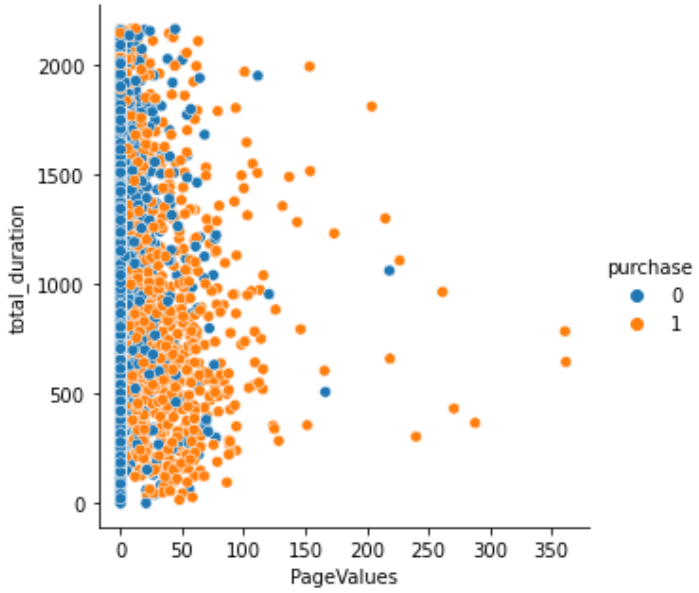
.13


purchase
0
1
ExitRates
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
BounceRates

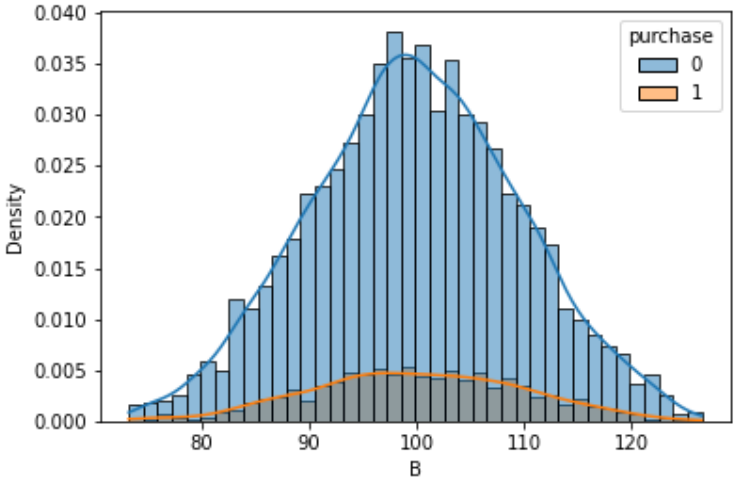
.14


purchase
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0.0
1.0
Weekend

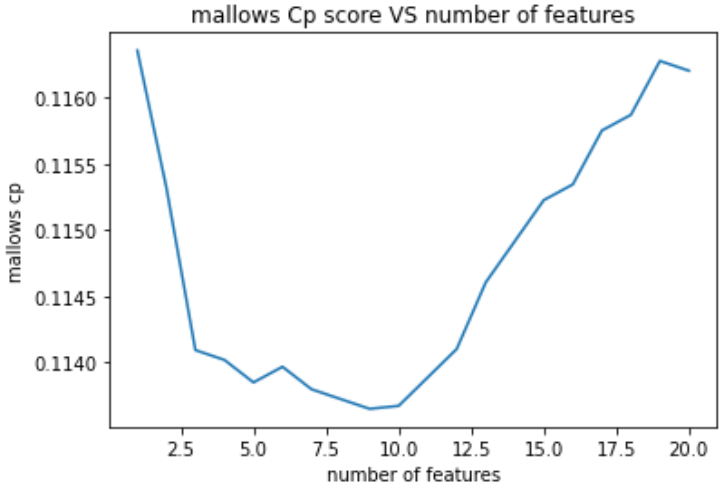
.15


total_duration
PageValues
purchase
0
1

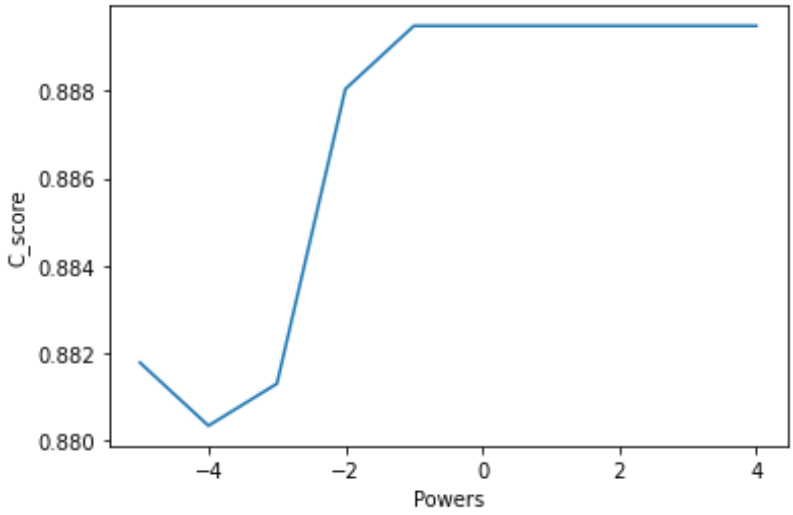
.16


Density
B
purchase
0
1

.17


mallows Cp score VS number of features
mallows cp
number of features

.18


0.888
0.886
0.884
0.882
0.880
C_score
−4
−2
0
2
4
Powers

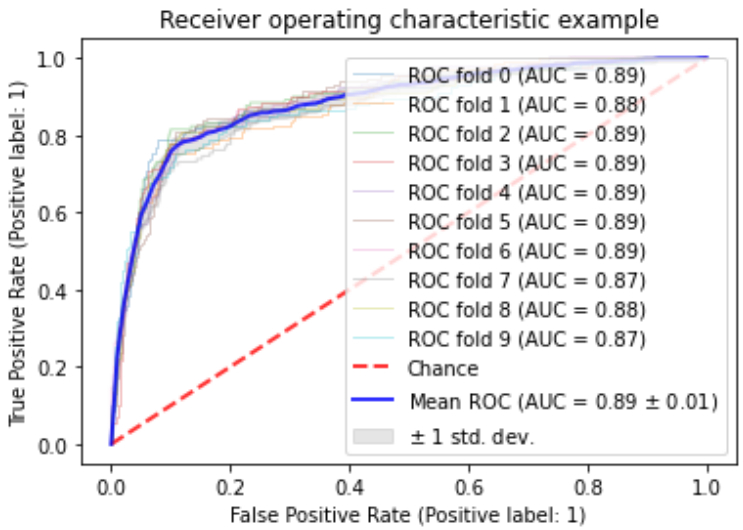
.19


no purchase
purchase
no purchase
purchase
6000
5000
4000
3000
2000
1000

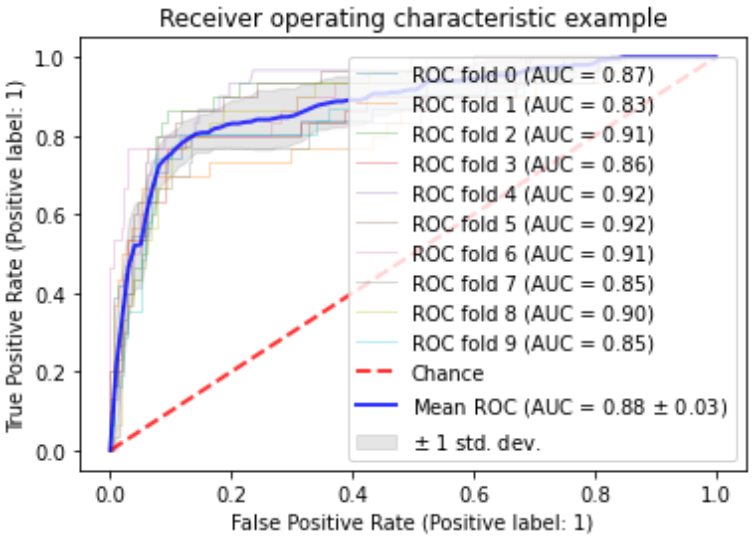
.20


no purchase
purchase
no purchase
purchase
1600
1400
1200
1000
800
600
400
200

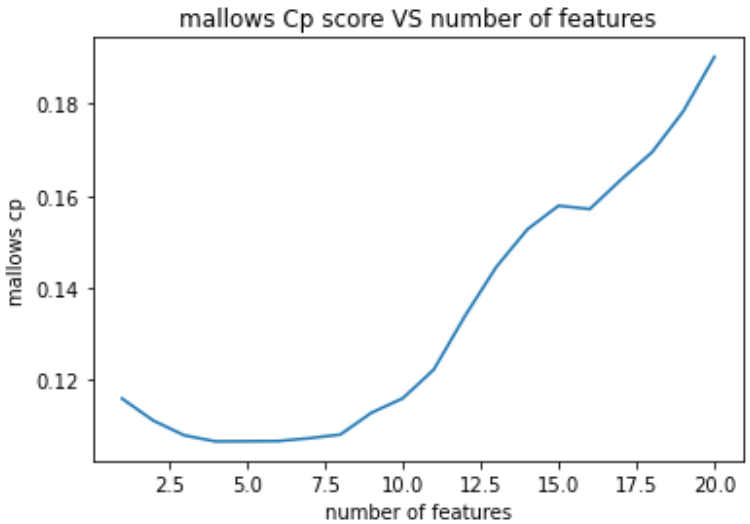
.21


Receiver operating characteristic example
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
ROC fold 0 (AUC = 0.89)
ROC fold 1 (AUC = 0.88)
ROC fold 2 (AUC = 0.89)
ROC fold 3 (AUC = 0.89)
ROC fold 4 (AUC = 0.89)
ROC fold 5 (AUC = 0.89)
ROC fold 6 (AUC = 0.89)
ROC fold 7 (AUC = 0.87)
ROC fold 8 (AUC = 0.88)
ROC fold 9 (AUC = 0.87)
Chance
Mean ROC (AUC = 0.89 ± 0.01)
± 1 std. dev.

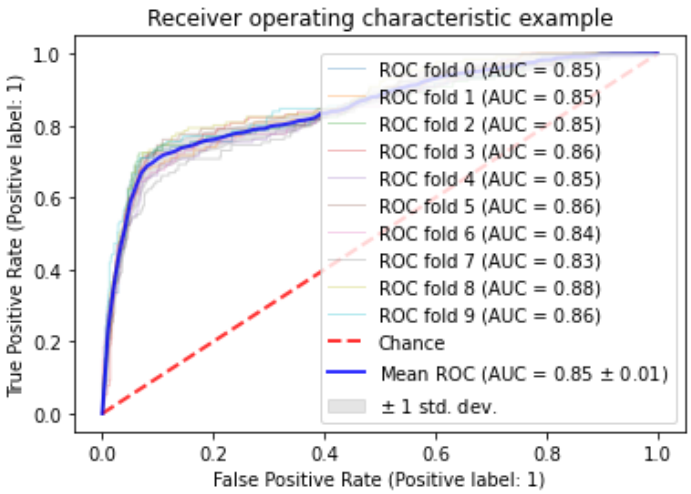
.22


Receiver operating characteristic example
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
ROC fold 0 (AUC = 0.87)
ROC fold 1 (AUC = 0.83)
ROC fold 2 (AUC = 0.91)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.92)
ROC fold 5 (AUC = 0.92)
ROC fold 6 (AUC = 0.91)
ROC fold 7 (AUC = 0.85)
ROC fold 8 (AUC = 0.90)
ROC fold 9 (AUC = 0.85)
Chance
Mean ROC (AUC = 0.88 ± 0.03)
± 1 std. dev.

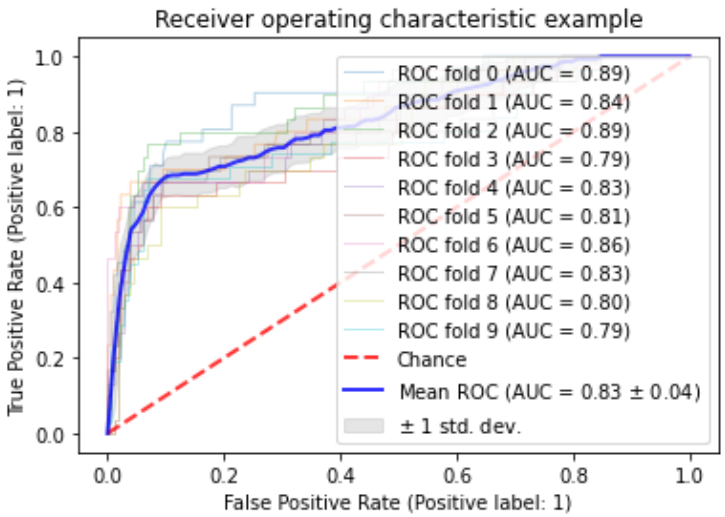
.23


mallows Cp score VS number of features
mallows cp
number of features

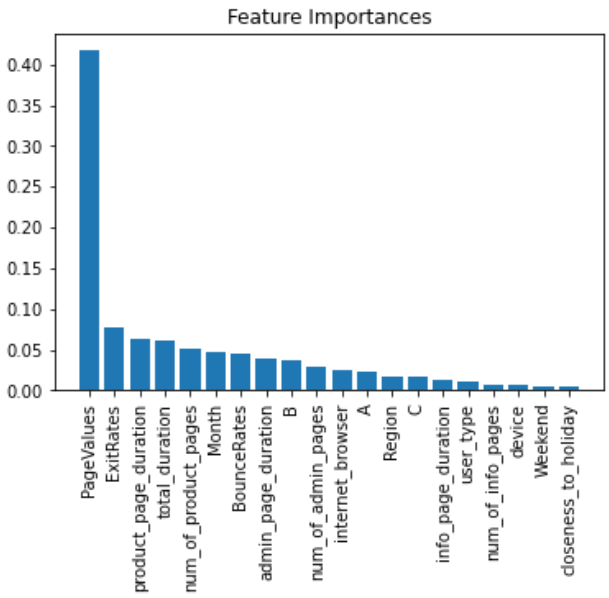
.24


Receiver operating characteristic example
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
ROC fold 0 (AUC = 0.85)
ROC fold 1 (AUC = 0.85)
ROC fold 2 (AUC = 0.85)
ROC fold 3 (AUC = 0.86)
ROC fold 4 (AUC = 0.85)
ROC fold 5 (AUC = 0.86)
ROC fold 6 (AUC = 0.84)
ROC fold 7 (AUC = 0.83)
ROC fold 8 (AUC = 0.88)
ROC fold 9 (AUC = 0.86)
Chance
Mean ROC (AUC = 0.85 ± 0.01)
± 1 std. dev.

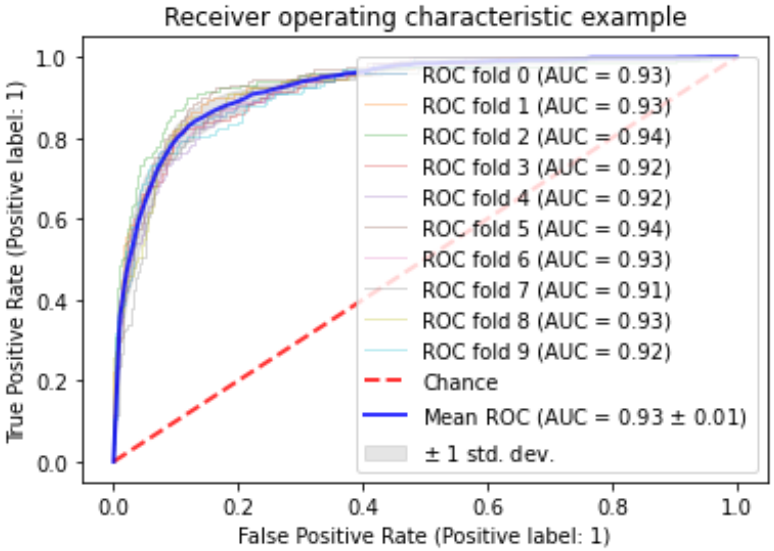
.25


Receiver operating characteristic example
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
ROC fold 0 (AUC = 0.89)
ROC fold 1 (AUC = 0.84)
ROC fold 2 (AUC = 0.89)
ROC fold 3 (AUC = 0.79)
ROC fold 4 (AUC = 0.83)
ROC fold 5 (AUC = 0.81)
ROC fold 6 (AUC = 0.86)
ROC fold 7 (AUC = 0.83)
ROC fold 8 (AUC = 0.80)
ROC fold 9 (AUC = 0.79)
Chance
Mean ROC (AUC = 0.83 ± 0.04)
± 1 std. dev.

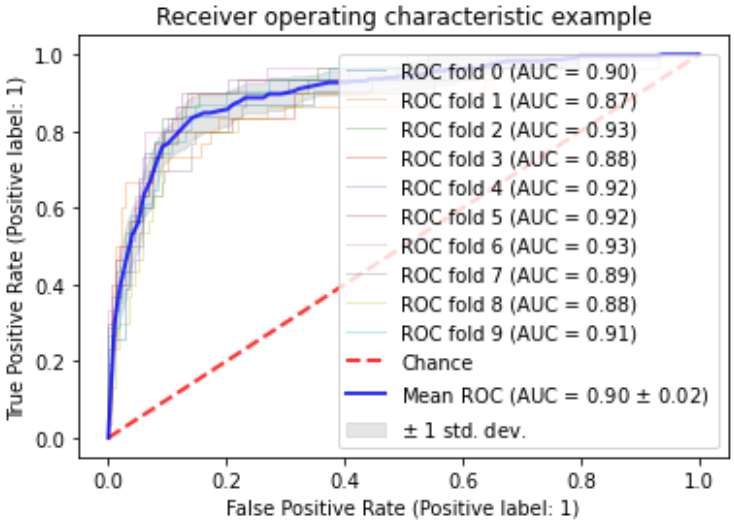
.26


Feature Importances
PageValues
ExitRates
product_page_duration
total_duration
num_of_product_pages
Month
BounceRates
admin_page_duration
B
num_of_admin_pages
internet_browser
A
Region
C
info_page_duration
user_type
num_of_info_pages
device
Weekend
closeness_to_holiday

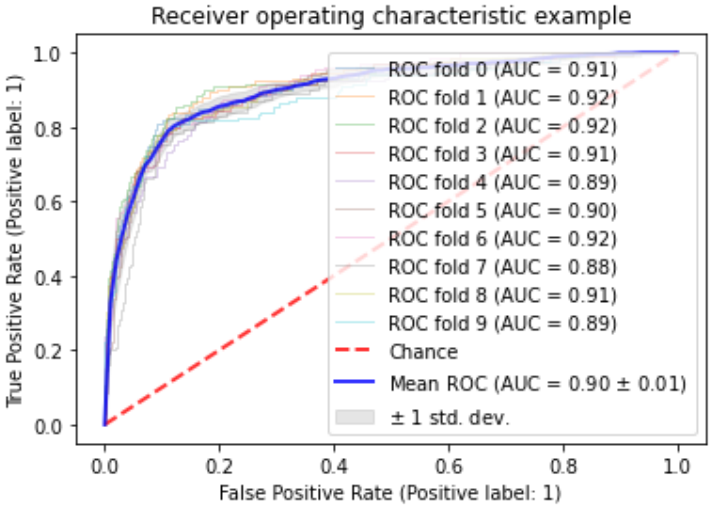
.27


Receiver operating characteristic example
ROC fold 0 (AUC = 0.93)
ROC fold 1 (AUC = 0.93)
ROC fold 2 (AUC = 0.94)
ROC fold 3 (AUC = 0.92)
ROC fold 4 (AUC = 0.92)
ROC fold 5 (AUC = 0.94)
ROC fold 6 (AUC = 0.93)
ROC fold 7 (AUC = 0.91)
ROC fold 8 (AUC = 0.93)
ROC fold 9 (AUC = 0.92)
Chance
Mean ROC (AUC = 0.93 ± 0.01)
± 1 std. dev.
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)

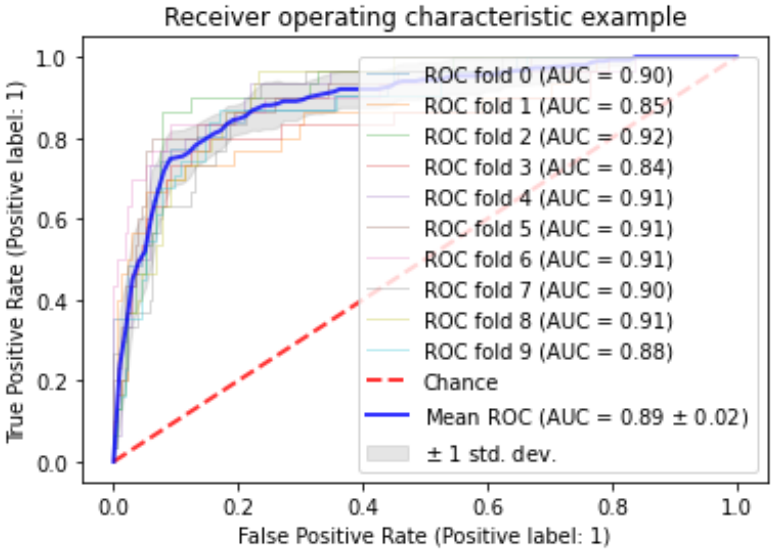
.28


Receiver operating characteristic example
ROC fold 0 (AUC = 0.90)
ROC fold 1 (AUC = 0.87)
ROC fold 2 (AUC = 0.93)
ROC fold 3 (AUC = 0.88)
ROC fold 4 (AUC = 0.92)
ROC fold 5 (AUC = 0.92)
ROC fold 6 (AUC = 0.93)
ROC fold 7 (AUC = 0.89)
ROC fold 8 (AUC = 0.88)
ROC fold 9 (AUC = 0.91)
Chance
Mean ROC (AUC = 0.90 ± 0.02)
± 1 std. dev.
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)

.29


Receiver operating characteristic example
ROC fold 0 (AUC = 0.91)
ROC fold 1 (AUC = 0.92)
ROC fold 2 (AUC = 0.92)
ROC fold 3 (AUC = 0.91)
ROC fold 4 (AUC = 0.89)
ROC fold 5 (AUC = 0.90)
ROC fold 6 (AUC = 0.92)
ROC fold 7 (AUC = 0.88)
ROC fold 8 (AUC = 0.91)
ROC fold 9 (AUC = 0.89)
Chance
Mean ROC (AUC = 0.90 ± 0.01)
± 1 std. dev.
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)

.30


Receiver operating characteristic example
True Positive Rate (Positive label: 1)
False Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
ROC fold 0 (AUC = 0.90)
ROC fold 1 (AUC = 0.85)
ROC fold 2 (AUC = 0.92)
ROC fold 3 (AUC = 0.84)
ROC fold 4 (AUC = 0.91)
ROC fold 5 (AUC = 0.91)
ROC fold 6 (AUC = 0.91)
ROC fold 7 (AUC = 0.90)
ROC fold 8 (AUC = 0.91)
ROC fold 9 (AUC = 0.88)
Chance
Mean ROC (AUC = 0.89 ± 0.02)
± 1 std. dev.