

פרויקט למידת מכונה



מגיש:

דניאל וולקוביץ (תז:207257668 )

תוכן עניינים

[תקציר מנהלים 3](#_Toc111985335)

[פירוט שלבי הפרויקט המקדימים להרצת המודלים 3](#_Toc111985336)

[עיבוד מקדים 3](#_Toc111985337)

[עיבוד סוג הערכים 3](#_Toc111985338)

[הסקת תובנות סטטיסטיות 3](#_Toc111985339)

[הסרת חריגים 3](#_Toc111985340)

[טיפול במידע חסר 3](#_Toc111985341)

[נרמול ערכים 3](#_Toc111985342)

[התמודדות עם משתנים קטגוריאליים 4](#_Toc111985343)

[דוגמא להנדסת פיצ'רים - רשומה D 4](#_Toc111985344)

[הורדת ממדים 4](#_Toc111985345)

[Filter 4](#_Toc111985346)

[PCA 4](#_Toc111985347)

[Forward Selection 4](#_Toc111985348)

[סיכום 4](#_Toc111985349)

[הרצת המודלים 4](#_Toc111985350)

[הערכת המודלים 5](#_Toc111985351)

[הערכה כללית 5](#_Toc111985352)

[confusion-matrix לרגרסיה לוגיסטית 5](#_Toc111985353)

[אחרית דבר 5](#_Toc111985354)

[נספח 1 – מטריצת קורלציות 6](#_Toc111985355)

[נספח 2 – תרשימי חשיבויות הפיצ'רים 6](#_Toc111985356)

[Naïve-Bias 6](#_Toc111985357)

[Logisticregression 6](#_Toc111985358)

[Ada-boost 6](#_Toc111985359)

[KNN 7](#_Toc111985360)

[נספח 3 – KNN Train Evaluation Report 7](#_Toc111985361)

[נספח 4 – ROC AUC CURVE 7](#_Toc111985362)

[נספח 5 – Confusion matrix 7](#_Toc111985363)

# תקציר מנהלים

בפרויקט זה, ביצעתי ניתוח של מידע מפיצ'רים שונים אשר מעידים על אופיו והתנהגותו של משתמש המבקר באתר. זאת במטרה לייצר תחזיות לסיכוי שתבוצע רכישה ע"י משתמש כלשהו. בהתאם להנחיות, תהליך הסקת המסקנות מהמידע כלל: עיבוד מקדים, נרמול הנתונים, שימוש בשיטות להורדת מימדים, אימון מודלים שנלמדו במהלך הקורס, ביצוע אופטימזציות למודלים, והערכות ביצועיהם. בשלבי העיבוד המקדים והורדת הממדים החלטתי לשים דגש על שימור מידע עסקי, לצד שימוש בפוטנציאל של אלגוריתמים לעיבוד המידע באופן אופטימלי. היתרון הוא שלאחר תהליך הלמידה, ניתן להבין מה הם הגורמים המשפיעים ביותר על סיכויי הרכישה. כתוצאה מכך, בעל עסק e-commerce יוכל להבין מה הם הפרמטרים שעליו להשקיע בהם על מנת שסיכויי המכירה שלו יהיו גבוהים יותר.

# פירוט שלבי הפרויקט המקדימים להרצת המודלים

## עיבוד מקדים

בשלב זה של הפרויקט ניתנה התייחסות נפרדת לכל פיצ'ר, כאשר לכל אחד מהם בוצעה הסקת תובנות סטטיסטיות על אופיו. בהתאם לתובנות ביצעתי את שלבי העיבוד המקדים השונים: עיבוד סוג הערכים, הסרת חריגים, טיפול במידע חסר, נרמול ערכים והתמודדת עם משתנים קטגוריאליים.

### עיבוד סוג הערכים

היו פיצ'רים שבהם ערכים שאינם מספריים (למשל: ערכים בוליאניים ומחרוזות), והחלפתי אותם בערכים מספריים מתאימים.

### הסקת תובנות סטטיסטיות

לצורך הסקת התובנות מכל פיצ'ר, השתמשתי בשלושה תרשימים קבועים:

1. **היסטוגרמה:** תרשים שמטרתו לייצר ויזואליזציה של התפלגות הפיצ'ר, כך שהוא מאפשר לזהות התפלגויות מוכרות, ומאפיינים מובהקים של התפלגות המידע.
2. **Boxplot**: תרשים שמטרתו לאפשר המחשה של כמות הנתונים החריגים ואופי התפלגותם.
3. **Correlation to labels:** תרשים שמטרתו להציג את מקדם ההתאמה בין הפיצ'ר לבין נתוני ביצוע הרכישה. ככלל תרשים זה פחות שימושי מהשאר, אך סייע מהותית בניתוח פיצ'ר D (וראו מטה).

### הסרת חריגים

בחרתי שלא להסיר את שורות המידע אשר מכילות מידע חריג, בכדי לא לאבד מידע רב. במקום זאת, החלפתי את המידע החריג בכל פיצ'ר ב-NULL, ולאחר מכן עדכנתי את הערכים פעם נוספת בשלב הטיפול במידע החסר (וראו מטה). במשתנים קטגוריאליים וסטטיסטיים, לא ביצעתי הסרת חריגים מכיוון שלא ניתן לבצע הסרה שכזו.

השתמשתי במספר שיטות להסרת חריגים, ולכל פיצ'ר נבחרה שיטה התואמת את אופי התפלגותו:

1. **טרנספורמציית לוג:** ביצוע לוג מתמטי על המידע, היתרון הוא שיחסיות הערכים נשמרת ובמקביל הערכים מצטופפים וכך אין צורך בהחלפת הערכים החריגים במידע אחר. החיסרון הוא שכאשר רק חלק מן הערכים המקוריים במידע קטנים מאחד אזי המידע מתעוות. ישנם פעמים שבכדי להתמודד עם פיצ'רים שבהם אפסים רבים, ביצעתי הוספה של 1 לכלל הערכים ורק לאחר מכן ביצעתי את הטרנספורמציה.
2. **IQR:** אלגוריתם קבוע אשר קובע אילו ערכים נחשבים לחריגים ע"י אחוזונים. החיסרון הוא שישנם פעמים בהם המדד גס מידי ומייצר יותר מידי חריגים, דבר אשר מוביל לאיבוד מידע.
3. **אחוזונים**: במקרים בהם ראיתי ש-IQR הוא מדד גס מידי, השתמשנו באחוזון ספציפי אשר היה נראה מתאים למידע.
4. **מדד סטיות תקן בהתפלגויות נורמליות**: כאשר נתקלתי בהתפלגות אשר קרובה להתפלגות נורמאלית, ביצעתי עליה מניפולציית מדד סטיות תקן (לדוגמא פיצ'ר B).

### טיפול במידע חסר

בהתאם להתפלגות כל אחד מן הפיצ'רים לאחר שלב הסרת החריגים, עדכנתי את הערכים החסרים באחד מן המדדים הבאים: בממוצע, בחציון או ב-mode. לרוב במשתנים נומריים עדכנתי את המידע החריג בעזרת הממוצע או החציון, ואילו במשתנים קטגוריאליים בעזרת mode.

### נרמול ערכים

נרמלתי את כל המשתנים הנומריים לתווך של בין 0 ל-1. כיוון שבשלב זה כבר טיפלתי בחריגים של כל פיצ'ר, לא ראיתי סיבה להשתמש בשיטות מתוחכמות של נרמול, וביצעתי נרמול פשוט.

### התמודדות עם משתנים קטגוריאליים

מעבר להתאמות המתוארות לעיל שביצעתי עבור משתנים קטגוריאליים, השתמשתי בשיטת one-hot-encoding לטובת התאמת המידע למודלים החישוביים. ישנם מקומות בודדים (לדוגמא: בפיצ'רים A ו-user\_type) בהם איחדתי קטגוריות ספציפיות שנדמו דומות או מיותרות לכדי קטגוריה אחרת.

יש לציין כי עפ"י התפלגות פיצ'ר A הסקתי שהוא קטגוריאלי.

### דוגמא להנדסת פיצ'רים - רשומה D

באמצעות תרשים Correlation to labels שהוזכר לעיל מצאתי כי אמנם פיצ'ר זה כולל רק מספר מועט של תצפיות, אך התצפיות הללו מסווגות את המידע באופן מובהק בהתאמה ל-labels. כלומר, היה ניתן לראות כי מעל ערך מסוים של D, ה- labelהתואם הוא בוודאות 1, ומתחת לו הוא בוודאות 0. בהתאם לממצא זה איתרתי את גבול הסיווג בין הקבוצות באמצעות ממוצע בין ה-IQR (העליון והתחתון בהתאמה) של שתי הקבוצות, ועיצבתי את הפיצ'ר מחדש. ביצעתי השמה של ערך 1 עבור ערכי label של 1, השמה של ערך (-1) עבור ערכי label של 0, ואילו עבור התצפיות החסרות ביצעתי השמה של הערך 0.

## הורדת ממדים

### Filter

בחלק זה, הורדתי פיצ'רים באופן נאיבי. למשל, הגעתי למסקנה שהפיצ'ר total-duration רק סוכם את שאר הפיצ'רים מסוג duration, ואינו מכיל מידע נוסף על פני המידע שיתר הפיצ'רים הללו מספקים. על כן בחרתי להוריד פיצ'ר זה.

בנוסף, יצרתי מטריצת קורלציות (ראו נספח 1) בין הפיצ'רים הנומריים השונים וכן חישבתי את הקורלציות שלהם ל-label. הסקתי שיש מתאם גבוה בין פיצ'רים המייצגים את כמות מספר דפים מסוג X שביקרו בהם, לבין משך זמן שביקרו בדפים מאותו הסוג (דבר אשר מתיישב עם ההיגיון). כיוון שלפיצ'רים מסוג duration קורלציה נמוכה ל-label מאשר הקורלציה של כמות הדפים ל-label, ומכיוון שלא נדמה לי כי הם מספקים מידע שונה, בחרתי להוריד את כלל הפיצ'רים מסוג duration.

### PCA

רציתי להשתמש באיכויות של PCA, אך חששתי מאובדן היכולת לפרש את התוצאות ומאובדן ההבנה העסקית של הפיצ'רים. בעיני בפרויקט אשר עוסק בניתוח נתוני מכירות, המשמעות של הסקת תובנות לגבי חשיבות הפיצ'רים חשובה לא פחות מהיכולת לבצע חיזויים נכונים. ללא עיבוד עסקי של תוצאות המודל בעל עסק לא יוכל לדעת במה עליו להתמקד כדי לשפר את נתוני המכירות שלו. לשם כך, החלטתי לפעול ע"י פשרת ביניים כפי שיתואר מטה.

חילקתי את הפיצ'רים לארבעה קטגוריות כדלהלן:

1. **Website-data:** כולל מידע שנוגע לשימוש באתר עצמו, כגון: סוגי הדפים בהם ביקר המשתמש והציונים השונים שסופקו משהייתו באתר.
2. **Timing of visit data:** כולל מידע שנוגע לתזמון של ביקור המשתמש באתר, כגון: האם סוף השבוע, קירבה לחגים וחודש הביקור.
3. **User-details:** כולל פרטים על המשתמש עצמו, כגון: סוג המכשיר, סוג המשתמש, סוג הדפדפן ואזור בו המשתמש נמצא.
4. **Unknown:** כולל את המשתנים הבלתי ידועים A, B, C מכיוון שבכל מקרה איננו יודעים את משמעותם והם אינם משפיעים על הבנה עסקית.

לא ביצעתי PCA על פיצ'ר D, כדי לשמר את הערך המסווג שיצרנו לו (כאמור לעיל).

### Forward Selection

לאחר ביצוע PCA, בחרתי להשתמש גם ב-Forward Selection. זאת בשל ניסוי וטעיה בשלב הערכת המודלים, ראיתי כי לאחר שימוש בשני השיטות קיבלתי תחזיות טובות יותר.

# סיכום

## הרצת המודלים

בכלל המודלים השתמשתי במתודת gridsearchcv לצורך ביצוע אופטימיזציה של הייפר-פרמטרים, כאשר האופטימיזציה נעשתה בהתאם לתוצאות מדד AUC-score. ישנם מקרים בהם ביצעתי שילוב בין המתודה לבין "היגיון ידני", וזאת בכדי לחסוך זמן ריצה בכל ריצה נתונה. כלומר עשיתי שימוש ב-grid פרמטרים רחב עם מעט משתנים, וצימצום שלו על סמך התוצאות המתקבלות. (במחברת ניתן לראות את ה- gridsהשונים בהם נעשה שימוש כהערות תחת #)

הרצתי מספר מודלים כדלהלן:

1. **Naïve-Bias:** בניגוד למודלים אחרים המודל מבוסס על נוסחאות מובהקות, ולכן לא ביצעתי אופטימיזציה עבור ההייפר-פרמטרים. לאחר עיון בתרשים חשיבות הפיצ'רים (ראו נספח 2), מצאתי שפיצ'רים מסוג website היוו הגורם המשמעותי ביותר למודל זה, אשר הביא תוצאות טובות ביותר.
2. **Logistic regression**: ביצעתי אופטימיזציה למודל בהתאם לתרשים חשיבות הפיצ'רים (ראו נספח 2). הפיצ'רים שהכי השפיעו עליו הם פיצ'רי website השונים (מבחינת קורלציה חיובית וגם שלילית).
3. **Ada-boost**: מודל שעל אף היותו מורכב יחסית, נותן יתרון בכך שהוא מספק מידע על חשיבות הפיצ'רים (בניגוד ל-SVM לדוגמא). ניתן לראות (ראו נספח 2) שמרבית הפיצ'רים הראשונים שהמודל סיווג באמצעותם הם פיצ'רים מסוג website ומסוג TIME. יש לציין לטובה שהמודל השתמש גם ב-D כמסווג משמעותי יחסית, וזאת בהתאם להנחה שהפיצ'ר צריך להשפיע רבות על החיזוי (כפי שפורט לעיל).
4. **MLP**: החולשה של המודל היא שלאחריו לא ניתן לנתח בעזרת אילו פיצ'רים ביצע למידה, והוא אמנם הביא תוצאות טובות אך לא באופן מובהק מן השאר.
5. **KNN:** לפני שביצענתי נורמליזציה על הנתונים, הוא הניב תוצאות סבירות (AUC=0.85). אך לאחר שנירמלתי את הנתונים, התוצאות שלו נהיו גרועות ומניתוח תוצאותיו ניתן לראות כי הוא סובל מאובר -פיטינג באופן מוחלט (ראו נספח 3). בפרויקט אחר הייתי שוקלים לייצר עבורו עיבוד מקדים שונה ללא נורמליזציה, אך לראייתי זה חורג ממסגרת פרויקט זה. מעיון בחשיבות הפיצ'רים, ניתן להסיק שפיצ'רים מסוג website היוו הגורם המשמעותי ביותר למודל זה (כמעט 50% מההשפעה שנמדדה).לעומת זאת, לאחר הנורמליזציה, ניתן לראות בתרשים חשיבות הפיצ'רים (ראו נספח 2) כי המודל לומד מכלל הפיצ'רים. דבר אשר מתיישב עם ההבנה שהוא סובל מאובר-פיטינג.
6. **Ensemble:** מודל אשר משתמש במודלים קיימים, ומוסיף מעליהם עוד מודל של רגרסיה לוגיסטית. בחרתי שלא להכניס את מודל MLP למודל זה, בעקבות שיקולי זמן ריצה. בנוסף לא הכנסתי את KNN למודל זה מכיוון שהוא רק היה פוגע בביצועי המודל. השכבה הנוספת מייצרת אופטימיזציה של השימוש במודלים השונים, ע"י איתור החוזקות של כל אחד מהם ושימוש בכל מודל בהתאם. המודל לא הביא תחזיות טובות מאשר הרגרסיה הלוגיסטית, ואני מסיק כי הסיבה לכך היא שככה"נ המודל נשען על הרגרסיה הלוגיסטית בצורה כמעט מוחלטת, ושכל המודלים "חזקים" בלחזות אי-רכישה, אך אינם מצליחים לחזות רכישה כהלכה. זהו המודל שהביא את התוצאות הטובות ביותר (וראו נספח 4), ובאמצעותו ביצעתי את החיזוי הסופי.

## הערכת המודלים

### הערכה כללית

הערכת המודלים לכל מודל בוצעה על בסיס שלושה כלים (אשר באמצעותם בוצעו גם אופטימיזציות כמפורט מעלה):

1. **Evaluation Report:** דו"ח אשר מעניק למודלים ציונים עפ"י מדדים שונים (מעניק פילוח נתונים דומה לזה של confusion-matrix).
2. **Feature Importance:** תרשים אשר מנתח את חשיבות הפיצ'רים השונים במודל, זאת כאשר ישנם מודלים אשר בהם לא ניתן לייצר את תרשים זה.
3. **ROC-AUC CURVE:** תרשים אשר מעניק ציונים שונים למודל באמצעות cross-validation ועל בסיס AUC ממוצע של כלל הציונים שניתנו למודל.

### confusion-matrix לרגרסיה לוגיסטית

יצרתי confusion-matrix לרגרסיה הלוגיסטית (וראו נספח 5). מהמטריצה ניתן לראות כי המודל הצליח לבצע למידה בצורה טובה, זאת מכיוון שמדדי תוצאות המודל על סט האימון הביא לתוצאות דומות על סט הולידציה. מדדי TPR, FPR, ו-AUC נתנו גם הם תוצאות דומות מאוד לשני ה-data sets. מכלל אלו ניתן להבין שמימדיות הבעיה הייתה מתאימה וכי הצלחתי לייצר מודל שמנבא תוצאות באופן מוצלח יחסית. כמו כן ניתן לראות שציוני ה-RECALL וה-PERSICISON של המודל עבור תוצאות חיוביות נמוך מאוד, אנו מניחים שהסיבה לכך היא שה ה-data set עליו המודל מתאמן איננו מאוזן ומכיל מעט מאוד תוצאות כאלו, ולכן אינו מצליח ללמוד את התוצאות הללו בצורה טובה.

### אחרית דבר

לסיום, לאורך הפרוייקט הבנתי שמודלים מגיבים בצורה קיצונית לשינויים בשלב העיבוד המקדים של הדאטה (כל אחד בצורה שונה). עם זאת, נראה לי כי ככלל ישנה מגמה של הסתמכות עיקרית על משתני website. דבר שעשוי להכווין בעל עסק כיצד עליו לנהוג בכדי לבצע מספר רב של מכירות.

# נספח 1 – מטריצת קורלציות

Shape

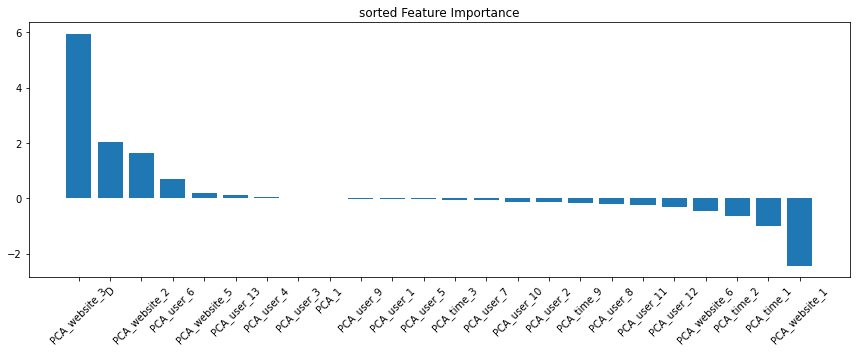
Description automatically generated with medium confidence

# נספח 2 – תרשימי חשיבויות הפיצ'רים

## Naïve-Bias



## Logisticregression



## Ada-boost

Histogram

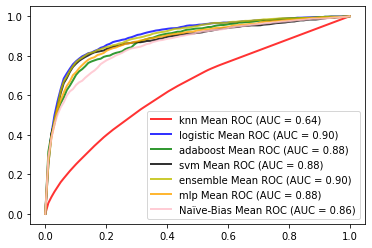
Description automatically generated

# KNNנספח 3 – KNN Train Evaluation Report

A screenshot of a computer

Description automatically generated with medium confidence

# נספח 4 – ROC AUC CURVE



# נספח 5 – Confusion matrix

